Unifying Sequential Monte Carlo with Resampling Matrices

Robert J. Webber*

Courant Institute of Mathematical Sciences
New York University
251 Mercer St, New York, NY 10012

e-mail: rw2515@nyu.edu

Abstract: Sequential Monte Carlo (SMC) is a class of algorithms that approximate high-dimensional expectations of a Markov chain. SMC algorithms typically include a resampling step. There are many possible ways to resample, but the relative advantages of different resampling schemes remain poorly understood. Here, a theoretical framework for comparing resampling schemes is presented. The framework uses resampling matrices to provide a simple description for the SMC resampling step. The framework identifies the matrix resampling scheme that gives the lowest possible error. The framework leads to new asymptotic error formulas that can be used to compare different resampling schemes.

MSC 2010 subject classifications: Primary 65C05; secondary 60J05.
Keywords and phrases: resampling, Sequential Monte Carlo.

1. Introduction

Sequential Monte Carlo (SMC) has a history that traces from the 1950’s to the present. The first examples of SMC were simulations of chain polymers in the 1950’s [15, 27]. Starting in the 1960’s, SMC was used in the quantum chemistry community to calculate the ground state energy of the Schrödinger equation [20, 12]. SMC became a standard statistical tool in the 1990’s, as the algorithm was applied to problems in Bayesian inference and signal processing [8]. In recent years, the algorithm continues to fascinate researchers who are ever developing new variations of SMC algorithms (e.g., [28, 10]).

SMC is a tool for evaluating expectations of the form

\[ E \left[ G_0(X_0) \prod_{t=0}^{T-1} G_t(X_{t-1}, X_t) f(X_{T-1}, X_T) \right] \]

where \((X_t)_{t \geq 0}\) is a discrete-time Markov chain on a sequence of state space \((E_t)_{t \geq 0}\), functions \((G_t)_{t \geq 0}\) are nonnegative, and \(f\) is real-valued. These expectations are called Feynman-Kac integrals, and they are notoriously difficult to evaluate when \(T\) is large [23]. SMC is a sampling algorithm that simulates the

*Supported by the Advanced Scientific Computing Research Program within the DOE Office of Science through award DE-SC0014205 and by NSF RTG award number 1547396
dynamics of the Markov chain \((X_t)_{t \geq 0}\) and provides random approximations for Feynman-Kac integrals that become increasingly accurate as computational effort is increased.

SMC has a wide range of applications from Bayesian statistics to rare event sampling. In Bayesian contexts, functions \((G_t)_{t \geq 0}\) are typically unnormalized likelihood ratios between prior and posterior distributions. SMC is used to estimate statistics of the posterior distribution, and the resulting algorithm is often called the particle filter \([8]\). In rare event sampling, on the other hand, SMC is used to provide estimates of rare event probabilities, and functions \((G_t)_{t \geq 0}\) bias a process \((X_t)_{t \geq 0}\) to explore regions of state space that would rarely be accessed under typical conditions \([13]\).

Despite the usefulness of SMC, practitioners are burdened with the difficult task of choosing a resampling scheme from the many options. Past analyses have provided error formulas for a few particular resampling schemes (e.g., \([5, 4, 7]\)). However, the number of resampling schemes has increased rapidly in recent years \([22]\), and more theoretical analysis is required to rigorously compare schemes. Error formulas are not available for all common resampling schemes (e.g., stratified resampling), and there remains no consensus among experts about how best to resample.

One goal of the current paper is to describe the resampling step in a unified way in order to facilitate analysis. Thus, Section 2 introduces a matrix resampling framework, inspired by work of Hu et al. \([18]\) and Whiteley et al. \([28]\). Resampling matrices provide a simple description for a great variety of resampling schemes, and any scheme in the matrix resampling framework is guaranteed to exhibit important convergence behavior. In particular, Section 2 proves unbiasedness, convergence, and an upper bound on variance for SMC estimates made using matrix resampling schemes.

Another goal of the current paper is to present a unified analysis of SMC error. Section 3 explains how error arises within the SMC algorithm and how error can be reduced by selecting an appropriate resampling scheme. The scheme that gives the lowest possible resampling error is identified. To compare the performance of resampling schemes, Section 3 also provides new asymptotic error bounds, including the first such bounds for stratified resampling and stratified residual resampling.

Technical proofs are presented in an appendix, following Section 3 and the conclusion.

2. Matrix resampling framework

The goal of the current section is to provide a matrix resampling framework that ties together diverse SMC resampling schemes. Section 2.1 provides a short overview of SMC. Section 2.2 describes the key features of the matrix resampling framework. Section 2.3 presents convergence theorems that ensure the validity of SMC estimates. Section 2.4 presents a martingale argument to show why SMC estimates are unbiased.
2.1. Overview of Sequential Monte Carlo

Sequential Monte Carlo begins by sampling initial “particles”, and then the algorithm proceeds iteratively through three main steps: reweighting, resampling, and mutation. Definition 2.1 gives an overview of these steps and the quantities that can be estimated through SMC:

Definition 2.1. Overview of Sequential Monte Carlo

1. Initialization: Independently sample \( \xi_0^{(i)} \sim \text{Law}(X_0) \) for \( 1 \leq i \leq N_0 \).
2. The algorithm proceeds iteratively for \( t = 0, 1, 2, \ldots \).
   
   (a) Reweighting: Assign weights \( w_t^{(i)} \) to each particle \( \xi_t^{(i)} \) with
   
   \[
   \begin{align*}
   w_t^{(i)} &= G_t\left(\xi_t^{(i)}\right), & t &= 0 \\
   w_t^{(i)} &= \tilde{w}_{t-1}^{(i)} G_t\left(\tilde{\xi}_t^{(i)}|\xi_{t-1}^{(i)}\right), & t &> 0
   \end{align*}
   \]

   (b) Resampling: Replace the ensemble \( \left(w_t^{(i)}, \xi_t^{(i)}\right)_{1 \leq i \leq N_t} \) with a new ensemble \( \left(\tilde{w}_t^{(j)}, \tilde{\xi}_t^{(j)}\right)_{1 \leq j \leq N_{t+1}} \), where each particle \( \tilde{\xi}_t^{(j)} \) is a copy of some particle \( \xi_t^{(i)} \) and weights \( \tilde{w}_t^{(j)} \) are defined so that

   \[
   \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} f(\xi_t^{(i)}) \approx \frac{1}{N_0} \sum_{j=1}^{N_{t+1}} \tilde{w}_t^{(j)} f(\tilde{\xi}_t^{(j)})
   \]

   for all functions \( f : E_t \to \mathbb{R} \).

   (c) Mutation: sample \( \xi_{t+1}^{(i)} \sim \text{Law}(X_{t+1}|X_t = \tilde{\xi}_t^{(i)}) \) for \( 1 \leq i \leq N_{t+1} \).

3. Estimation: To estimate quantities \( \mathbb{E}\left[ \prod_{s=0}^{t-1} G_s f \right] \), use

   \[
   \frac{1}{N_0} \sum_{i=1}^{N_t} \tilde{w}_{t-1}^{(i)} f(\tilde{\xi}_{t-1}^{(i)}, \xi_t^{(i)}) \approx \mathbb{E}\left[ \prod_{s=0}^{t-1} G_s f \right]
   \]

For notational simplicity, in expectations involving the Markov Chain \( (X_t)_{t \geq 0} \), the arguments of functions will often be omitted. For example, \( \mathbb{E}\left[ \prod_{s=0}^{t-1} G_s f \right] \) denotes \( \mathbb{E}\left[ G_0(X_0) \prod_{s=1}^{t-1} G_s(X_{s-1}, X_s) f(X_{t-1}, X_t) \right] \).

While the reweighting and mutation steps are straightforward, there are many different ways to carry out the resampling step. Outlined below are examples of resampling methods:

Example 2.1 (Sequential importance sampling). The simplest resampling scheme, sequential importance sampling [15, 27], leaves the ensemble of particles and weights completely unchanged:

\[
\left(\tilde{w}_t^{(j)}, \tilde{\xi}_t^{(j)}\right)_{1 \leq j \leq N_{t+1}} = \left(w_t^{(i)}, \xi_t^{(i)}\right)_{1 \leq i \leq N_t}
\]
In sequential importance sampling, weights $\tilde{w}_t^{(j)}$ are multiples of many functions $(G_t)_{t \geq 0}$:

$$\tilde{w}_t^{(j)} = \tilde{w}_{t-1}^{(j)} G_t \left( \xi_{t-1}^{(j)}, s_t^{(j)} \right) = \tilde{w}_{t-2}^{(j)} G_{t-1} \left( \xi_{t-2}^{(j)}, s_{t-1}^{(j)} \right) G_t \left( \xi_{t-1}^{(j)}, s_t^{(j)} \right) = \ldots$$

Consequently, some weights $\tilde{w}_t^{(j)}$ can be very large, while other weights can be very small. The imbalance in weights can potentially contribute variance to the estimates $\frac{1}{N_0} \sum_{i=1}^{N_t} \tilde{w}_t^{(j)} f \left( \xi_t^{(i)}, \xi_t^{(j)} \right)$, because the single particle with the highest weight can dominate all the others.

Alternatives to sequential importance sampling, which alleviate the imbalance in weights, include multinomial resampling and Bernoulli resampling.

**Example 2.2** (Multinomial resampling). In multinomial resampling [17], updated particles $\left( \tilde{x}_t^{(j)} \right)_{1 \leq i \leq N_t}$ are independently sampled with common distribution

$$\tilde{x}_t^{(j)} \sim \frac{\sum_{i=1}^{N_t} w_t^{(i)} \delta \left( \xi_t^{(i)} \right)}{\sum_{i=1}^{N_t} w_t^{(i)}}$$

and each updated particle is assigned an updated weight $\tilde{w}_t^{(j)} = \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)}$.

**Example 2.3** (Bernoulli resampling). In Bernoulli resampling [20], each of the original particles $\xi_t^{(i)}$ is replicated $N_t^{(i)}$ times, where the numbers $N_t^{(i)}$ are independent random variables with

$$\begin{cases} 
N_t^{(i)} = \left[ \frac{w_t^{(i)}}{\bar{w}_t} \right] + 1, & \text{with probability } \left\{ \frac{w_t^{(i)}}{\bar{w}_t} \right\} \\
N_t^{(i)} = \left[ \frac{w_t^{(i)}}{\bar{w}_t} \right], & \text{otherwise}
\end{cases}$$

Here, the floor function $[\cdot]$ is defined by $[x] = \max \{ z \in \mathbb{Z} : z \leq x \}$, the remainder function $\{ \cdot \}$ is defined by $\{ x \} = x - [x]$, and $\bar{w}_t = \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)}$ is the average of the weights. After replication, each updated particle $\tilde{x}_t^{(j)}$ is assigned an updated weight $\tilde{w}_t^{(j)} = \bar{w}_t$.

### 2.2. Extending the matrix resampling framework

Sequential importance sampling and multinomial resampling are both matrix resampling schemes. First introduced by Hu et al. [18] and Whiteley et al. [28], matrix resampling schemes involve a resampling step described by a nonnegative matrix $W_t$ with dimensions $N_t \times N_{t+1}$. The properties of this matrix are:

- The $i$th row sum equals the weight $w_t^{(i)}$ for $1 \leq i \leq N_t$.
- The $j$th column sum equals the updated weight $\tilde{w}_t^{(j)}$ for $1 \leq j \leq N_{t+1}$.
Each updated particle $\hat{\xi}_t^{(j)}$ is independently drawn from a distribution determined by the $j$th column of the resampling matrix:

$$\hat{\xi}_t^{(j)} \sim \frac{\sum_{i=1}^{N_t} w_t^{(i,j)} \delta(\xi_t^{(i)})}{\sum_{i=1}^{N_t} w_t^{(i,j)}}$$

Resampling schemes can be divided into fixed population resampling schemes, where particle numbers $(N_t)_{t \geq 0}$ are deterministic, and random population resampling schemes, where the number of particles $(N_t)_{t \geq 0}$ is random. While the matrix resampling framework is useful for describing fixed population schemes, it is necessary to extend the framework further in order to describe random population resampling schemes.

This section presents a new extension to the matrix resampling framework to random population schemes that satisfy an upper bound on the maximum possible number of particles $N_t$. In these schemes, $N_t$ can be bounded by $C_t N_0$ for each $t \geq 0$, where $(C_t)_{t \geq 0}$ is a deterministic series of constants. This assumption is often satisfied for the random population schemes used in practice. For example, in Bernoulli resampling, the random numbers $N_t$ satisfy an upper bound $N_t \leq N_0 (t + 1)$ and cannot grow in an uncontrolled way, because

$$N_{t+1} = \sum_{i=1}^{N_t} N_t^{(i)} = \sum_{i=1}^{N_t} \left( \left\lfloor \frac{w_t^{(i)}}{\hat{w}_t} \right\rfloor + 1 \right) \leq \sum_{i=1}^{N_t} \left( \frac{w_t^{(i)}}{\hat{w}_t} + 1 \right) = N_0 + N_t$$

The extended matrix resampling framework differs from the standard matrix resampling framework by including a “coffin state” $c$. The coffin state is an element of state space that particles $\hat{\xi}_t^{(j)}$ can potentially occupy, but particles in the coffin state do not affect any SMC estimates. By including a coffin state, the extended matrix resampling framework is able to reinterpret many random population schemes as schemes where the number of particles $(N_t)_{t \geq 0}$ is deterministic but the number of coffin state particles is random.

In the extended matrix resampling framework, the Markov chain $X_t$ is allowed to take values in the extended state space $E_t \cup \{c\}$. Transitions from the coffin state are described by $P \{ X_{t+1} = c | X_t = c \} = 1$. Functions defined on $E_t$ or $E_{t-1} \times E_t$ are extended to take values $f(c) = 0$ or $f(c,c) = 0$. As seen in the definition below, the extended matrix resampling framework includes a row in each resampling matrix $W_t$ governing transitions into the coffin state $c$.

**Definition 2.2.** Extended matrix resampling framework

1. Initialization: Independently sample $\xi_0^{(i)} \sim \text{Law}(X_0)$ for $1 \leq i \leq N_0$.
2. The algorithm proceeds iteratively for $t = 0, 1, 2, \ldots$.
   (a) Reweighting: Assign weights $w_t^{(i)}$ to each particle $\xi_t^{(i)}$ with

$$\begin{cases} w_t^{(i)} = G_t \left( \hat{\xi}_t^{(i)} \right), & t = 0 \\
 w_t^{(i)} = \hat{w}_{t-1} G_t \left( \hat{\xi}_{t-1}^{(i)}, \xi_t^{(i)} \right), & t > 0 
\end{cases}$$
(b) Resampling: Select a nonnegative matrix $W_t$ with dimensions $(N_t + 1) \times N_{t+1}$ and row sums $\sum_{i=1}^{N_t+1} W_t^{i,j} = w_t^{(i)}$ for $1 \leq i \leq N_t$. Independently, for $1 \leq j \leq N_{t+1}$, select $\xi_t^{(j)}$ from the distribution
\[
\xi_t^{(j)} \sim \frac{\sum_{i=1}^{N_t} w_t^{i,j} \delta\left(\xi_t^{(i)}\right) + w_t^{(N_t+1,j)} \delta(c)}{\sum_{i=1}^{N_t+1} w_t^{(i,j)}}
\]

Define the $w_t^{(j)}$ by the column sum $w_t^{(j)} = \sum_{i=1}^{N_t+1} w_t^{i,j}$.

(c) Mutation: sample $\xi_t^{(i)} \sim \text{Law} \left( X_{t+1} | X_t = \xi_t^{(i)} \right)$ for $1 \leq i \leq N_{t+1}$.

3. Estimation: To estimate quantities $E \left[ \prod_{s=0}^{t-1} G_s f \right]$, use
\[
\frac{1}{N_0} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} f \left( \tilde{\xi}_{t-1}^{(i)}, \tilde{\xi}_t^{(i)} \right) \approx E \left[ \prod_{s=0}^{t-1} G_s f \right]
\]

The extended matrix resampling framework encompasses a variety of resampling schemes. For example, Figure 1 presents resampling matrices $W_t$ that correspond to sequential importance sampling, multinomial resampling, and Bernoulli resampling. In the extended matrix resampling framework, the choice of which matrix $W_t$ to use can be made adaptively, incorporating any information, such as the values of particles $(\xi_t^{(i)})_{1 \leq i \leq N_t}$ and their weights $(w_t^{(i)})_{1 \leq i \leq N_t}$. Only the numbers $(N_t)_{t \geq 0}$ must be fixed in advance of running the SMC algorithm.

![Examples of resampling matrices $W_0$ when $N_0 = 4$, $N_1 = 6$, and particles have weights $w_0^{(1)} = 3.2$, $w_0^{(2)} = 2.4$, $w_0^{(3)} = .8$, and $w_0^{(4)} = 1.6$. A horizontal line separates the coffin state $c$.](image)

**2.3. Unbiasedness, convergence, and variance**

The matrix resampling framework leads to a series of powerful results on the unbiasedness, convergence, and variance of SMC estimates. While versions of these theorems were proved previously [5, 7, 28], this section presents results that hold more broadly and include all schemes in the matrix resampling framework.

The first of the key theorems that govern the validity of SMC estimates ensures that estimates are unbiased:
Theorem 2.1. If \( \mathbb{E} \left| \prod_{t=0}^{T-1} G_t f \right| < \infty \), then SMC estimates are unbiased:

\[
\mathbb{E} \left[ \frac{1}{N_0} \sum_{i=1}^{N_T} \tilde{w}_{t-1}^{(i)} f \left( \xi_{t-1}^{(i)}, \xi_T^{(i)} \right) \right] = \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right]
\]

Theorem 2.1 is quite general and holds without any additional assumptions. In contrast, Theorems 2.2 and 2.3 will require a mild assumption on the numbers \((N_t)_{t \geq 0}\) and on the resampling matrices \((W_t)_{t \geq 0}\):

Assumption 2.1. There exist absolute constants \((C_t)_{t \geq 0}\) such that \( \frac{N_t}{N_0} \leq C_t \) and \( \max_{1 \leq i \leq N_t+1} \tilde{w}_{t}^{(i)} \leq C_t \max_{1 \leq j \leq N_t} w_{t}^{(j)} \).

Assumption 2.1 guarantees that the number of particles does not grow too high and also that the maximum weight does not grow too high during resampling. This assumption is satisfied for all the schemes presented in the current paper, taking \( C_t = 1 \) for fixed population schemes and \( C_t = t + 1 \) for random population schemes.

The next result is a widely useful convergence theorem for SMC estimates:

Theorem 2.2. If \( \mathbb{E} \left| \prod_{s=0}^{t} G_s \right| < \infty \) for \( 0 \leq t \leq T - 1 \) and \( \mathbb{E} \left| \prod_{t=0}^{T-1} G_t f \right| < \infty \), then

\[
\frac{1}{N_0} \sum_{i=1}^{N_T} \tilde{w}_{T-1}^{(i)} f \left( \xi_{T-1}^{(i)}, \xi_T^{(i)} \right) \overset{P}{\rightarrow} \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right] \quad \text{as } N_0 \to \infty
\]

In Theorem 2.2, it is assumed the SMC algorithm is well-defined on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) for any number of starting particles \( N_0 = 1, 2, \ldots \). As \( N_0 \to \infty \), Theorem 2.2 establishes that SMC estimates converge in probability to the correct result.

Another key convergence result is a simple upper bound on the variance of SMC estimates. The upper bound leads to a clear interpretation that SMC estimates have a \( \frac{1}{\sqrt{N_0}} \) error rate when functions \((G_t)_{t \geq 0}\) are bounded.

Theorem 2.3. If functions \((G_t)_{0 \leq t \leq T-1}\) are bounded and \( \mathbb{E} \left| \prod_{t=0}^{T-1} G_t f^2 \right| < \infty \), then

\[
\text{Var} \left[ \frac{1}{N_0} \sum_{i=1}^{N_T} \tilde{w}_{T-1}^{(i)} f \left( \xi_{T-1}^{(i)}, \xi_T^{(i)} \right) \right] \leq \frac{1}{N_0} \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f^2 \right] \prod_{t=0}^{T-1} \sup G_t \sum_{s=0}^{T-t-1} C_s
\]

where \((C_t)_{t \geq 0}\) are the constants appearing in Assumption 2.1.

While antecedents of Theorems 2.2 and 2.3 appear in the SMC literature [7, 28], the versions presented here are more general with respect to possible resampling schemes or are more powerful with respect to unbounded functions \( f \). In examples outlined below, these theorems determine the convergence behavior of a diverse set of matrix resampling schemes. See also Figure 2, which provides resampling matrices for the three examples.
Example 2.4 (Adaptive resampling and parallel resampling). Two common variations on the SMC framework are adaptive resampling and parallel resampling. In adaptive resampling [23], a resampling scheme such as multinomial or Bernoulli resampling is triggered if the variation in weights exceeds a certain threshold; otherwise, sequential importance sampling is applied instead. In parallel resampling [22], resampling is applied independently on different processors in order to minimize communication costs. Theorems 2.2 and 2.3 guarantee the convergence of many adaptive and parallel resampling schemes. In particular, convergence is guaranteed even if the user decides adaptively which resampling scheme to use at the start of each resampling stage or if resampling decisions are made in parallel across different machines.

Example 2.5 (Pruning and enrichment). In the pruning and enrichment scheme [11], a lower cutoff $u_t$ and an upper cutoff $U_t$ are selected at the beginning of each resampling step. If $w_t(i) > U_t$, then the particle $\xi_t(i)$ is split into two replicas $\xi_t(j)$ and $\xi_t(k)$ with reduced weights $\hat{w}_t(j) = \hat{w}_t(k) = \frac{1}{2}w_t(i)$. If $w_t(i) < u_t$, then instead an updated particle $\hat{\xi}_t(i)$ is drawn from the distribution

$$\hat{\xi}_t(i) \sim \frac{1}{2} \delta (\xi_t(i)) + \frac{1}{2} \delta (c)$$

with weight $\hat{w}_t(i) = 2w_t(i)$. Lastly, if $u_t \leq w_t(i) \leq U_t$, the $i$th particle and weight are left unchanged, with $\left(\hat{w}_t(i), \hat{\xi}_t(i)\right) = \left(w_t(i), \xi_t(i)\right)$ for some $1 \leq j \leq N_t+1$. Theorems 2.2 and 2.3 guarantee convergence of the pruning and enrichment scheme even when cutoff values $u_t$ and $U_t$ are selected adaptively at the start of each resampling stage.

Example 2.6 (Rejection control). The rejection control scheme [24] mixes sequential importance sampling and Bernoulli resampling. In this scheme, first compute the average particle weight $\overline{w}_t = \frac{1}{N_t} \sum_{j=1}^{N_t} w_t(j)$. Then, if $w_t(i) \geq \overline{w}_t$, the $i$th particle and weight are left unchanged, with $\left(\hat{w}_t(i), \hat{\xi}_t(i)\right) = \left(w_t(i), \xi_t(i)\right)$ for some $1 \leq j \leq N_t+1$. Otherwise, if $w_t(i) < \overline{w}_t$, a particle $\hat{\xi}_t(i)$ is drawn from the distribution

$$\hat{\xi}_t(i) \sim \frac{w_t(i)}{\overline{w}_t} \delta (\xi_t(i)) + \left(1 - \frac{w_t(i)}{\overline{w}_t}\right) \delta (c)$$

with weight $\hat{w}_t(i) = \overline{w}_t$. Theorems 2.2 and 2.3 are the best known convergence results for the rejection control scheme.

Remark 2.1. Many past analyses of SMC [7, 3] have focused on SMC estimates of ratios $\mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right] / \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t \right]$. In the present analysis, the central focus is shifted toward SMC estimates of quantities $\mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right]$. This central focus has three advantages. First, estimates of $\mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right]$ are unbiased, making them simpler to analyze than estimates of ratios, which are typically biased.
Fig 2. Examples of resampling matrices $W_0$ when $N_0 = 4$, $N_1 = 5$, and particles have weights $w_0^{(1)} = 3.2$, $w_0^{(2)} = 2.4$, $w_0^{(3)} = 8$, and $w_0^{(4)} = 1.6$. In parallel resampling the resampling matrix takes a block diagonal form, with each block corresponding to a different processor. In pruning and enrichment, the cutoff values are $c_t = 1$ and $C_t = 3$.

Second, unbiased SMC estimates have not been studied in as much detail as estimates of ratios have been, despite their central importance in rare event sampling and Bayesian statistics [5, 13]. Third, convergence properties for estimates of ratios follow as a corollary of convergence properties for unbiased estimates. For more details of this relationship, refer to the discussion in the appendix.

2.4. Martingale analysis of SMC

Martingale theory provides an essential tool for the analysis of SMC [5, 7, 3]. In the current section, a martingale is used to show that SMC estimates are unbiased. In later sections, the same martingale leads to an error decomposition and asymptotic error formulas for SMC estimates. The first step in a martingale analysis is to define a filtration and a martingale sequence on that filtration. Toward this goal, fix functions $(G_t)_{0 \leq t \leq T - 1}$ and $f$ and define $\sigma$-algebras and conditional expectations as follows:

**Definition 2.3.** $\sigma$-algebras and conditional expectations

1. Introduce the filtration $(\mathcal{F}_t)_{-1 \leq t \leq T}$, where

\[
\begin{align*}
\mathcal{F}_{-1} &= \{\emptyset, \Omega\} \\
\mathcal{F}_0 &= \sigma\left(\left(G_{s0}^{(i)}\right)_{1 \leq i \leq N_0}, W_0\right) \\
\mathcal{F}_{t+\frac{1}{2}} &= \mathcal{F}_t \lor \sigma\left(\left(G_{jt}^{(j)}\right)_{1 \leq j \leq N_{t+1}}\right), \quad 0 \leq t \leq T - 1 \\
\mathcal{F}_{t+1} &= \mathcal{F}_{t+\frac{1}{2}} \lor \sigma\left(\left(G_{st+1}^{(i)}\right)_{1 \leq i \leq N_{t+1}}, W_t\right), \quad 0 \leq t \leq T - 2 \\
\mathcal{F}_T &= \mathcal{F}_{T-\frac{1}{2}} \lor \sigma\left(\left(G_{iT}^{(i)}\right)_{1 \leq i \leq N_T}\right)
\end{align*}
\]

Here, $\mathcal{G} \lor \mathcal{H}$ denotes the smallest $\sigma$-algebra containing $\mathcal{G}$ and $\mathcal{H}$.

2. Define the conditional expectations

\[
h_t(x_t) = E\left[\prod_{s=t+1}^{T-1} G_s f \left| X_s = x_t\right.\right], \quad 0 \leq t \leq T - 1
\]
with the convention that $\prod_0 G_x = 1$.

3. To keep the notation simple, write $h_T(x_T) = 1, G_T(x_T-1,x_T) = f(x_T-1,x_T)$, and $w_T^{(i)} = w_T^{(i)} G_T(\xi_T^{(i)}, \xi_T^{(i)})$ for $1 \leq i \leq N_T$.

The next theorem shows that the SMC estimate $\frac{1}{N_0} \sum_{i=1}^{N_T} w_{T-1}^{(i)} f(\hat{\xi}_{T-1}^{(i)}, \xi_T^{(i)})$ for the quantity $E[\prod_{t=0}^{T-1} G_t f]$ can be interpreted as a martingale on the filtration $\mathcal{F}_t$:

**Theorem 2.4.** If $E[\prod_{t=0}^{T-1} G_t f] < \infty$, there exists a martingale $M_t$ on the filtration $\mathcal{F}_t$ that satisfies

$$
\begin{align*}
M_{t-1} &= E\left[\prod_{t=0}^{T-1} G_t f\right] \\
M_t &= \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} h_t(\xi_t^{(i)}), \quad 0 \leq t \leq T \\
M_{t+1} &= \frac{1}{N_0} \sum_{i=1}^{N_{t+1}} w_t^{(i)} h_t(\hat{\xi}_t^{(i)}), \quad 0 \leq t \leq T - 1
\end{align*}
$$

**Proof.** For $0 \leq t \leq T - 1$,

$$
E\left[M_{t+1} \mid \mathcal{F}_{t+1}\right] = E\left[\frac{1}{N_0} \sum_{i=1}^{N_t+1} w_{t+1}^{(i)} h_{t+1}(\hat{\xi}_{t+1}^{(i)}) \mid \mathcal{F}_{t+1}\right] = \frac{1}{N_0} \sum_{i=1}^{N_t+1} w_t^{(i)} h_t(\hat{\xi}_t^{(i)}) = M_{t+1}
$$

Lines (2.1)-(2.3) use the fact that $\hat{w}_t^{(i)}$ is measurable with respect to $\mathcal{F}_{t+1}$, as well as the definitions for $w_t^{(i)}$, Law $(\hat{\xi}_{t+1}^{(i)} \mid \mathcal{F}_{t+1})$, $h_{t+1}$, and $h_t$.

Next, for $0 \leq t \leq T - 1$,

$$
E\left[M_t \mid \mathcal{F}_t\right] = E\left[\frac{1}{N_0} \sum_{j=1}^{N_t} w_t^{(j)} h_t(\hat{\xi}_t^{(j)}) \mid \mathcal{F}_t\right] = \frac{1}{N_0} \sum_{j=1}^{N_t} w_t^{(j)} h_t(\hat{\xi}_t^{(j)}) = M_t
$$

Lines (2.4)-(2.6) use the fact that $\hat{w}_t^{(j)}$ is measurable with respect to $\mathcal{F}_t$, the definition for Law $(\hat{\xi}_t^{(j)} \mid \mathcal{F}_t)$, and the fact that $\sum_{j=1}^{N_t} w_t^{(i,j)} = w_t^{(i)}$. 


Lastly, because $\xi_j \sim \text{Law}(X_0)$ for $1 \leq j \leq N_0$, $E\left[ \frac{1}{N_0} \sum_{i=1}^{N_0} w_i^{(i)} h_0 \left( \xi_i^{(i)} \right) \right] = E[G_0 h_0] = E\left[ \prod_{t=0}^{T-1} G_t f \right]$. 

Theorem 2.4 guarantees the unbiasedness of SMC estimates, confirming Theorem 2.1.

3. Unified analysis of SMC error

The current section provides a unified analysis of SMC error which facilitates comparison of different resampling schemes. Section 3.1 defines complete resampling schemes, a subset of matrix resampling schemes which will be covered in the error analysis. Section 3.2 explains how error arises within the SMC algorithm and how error can be reduced by selecting an appropriate resampling scheme. Section 3.3 identifies the matrix resampling scheme that gives the lowest possible error. Section 3.4 presents new asymptotic formulas that can be used to rigorously compare the error associated with different resampling schemes.

3.1. Complete resampling schemes

A complete resampling scheme is a matrix resampling scheme with the requirement that all the updated weights $\hat{w}_t^{(j)}$ equal the same weight $\hat{w}_t = \frac{1}{N_0} \sum_{i=1}^{N_0} w_i^{(i)}$. Complete resampling schemes, which include Bernoulli resampling and multinomial resampling, are very prominent in discussions of SMC. In fact, several previous reviews of resampling methods focused solely on complete resampling schemes \cite{6, 16}. The error analysis makes the following assumption:

**Assumption 3.1.** The resampling scheme is complete; that is, all the updated weights $\hat{w}_t^{(j)}$ equal the same weight, $\hat{w}_t = \frac{1}{N_0} \sum_{i=1}^{N_0} w_i^{(i)}$.

There are two major factors that determine the value of a resampling scheme: the computational cost of using the scheme and the accuracy of the estimates it provides. The advantage of analyzing complete resampling schemes is that all complete resampling schemes share a similar computational cost. In particular, the computational cost of an SMC algorithm is proportional to the number of non-coffin particles, and the next proposition guarantees that the number of non-coffin particles is similar for all complete resampling schemes, with a statistical range of $N_0 \pm \sqrt{N_0}$ particles:

**Proposition 3.1.** If at least one of the weights $\left( w_i^{(i)} \right)_{1 \leq i \leq N_0}$ is positive, then the number of non-coffin particles satisfies

$$
\begin{align*}
\mathbb{E} \left[ \sum_{j=1}^{N_t} 1 \left\{ \xi_j^{(j)} \neq c \right\} \mid \mathcal{F}_t \right] &= N_0 \\
\text{Var} \left[ \sum_{j=1}^{N_t} 1 \left\{ \xi_j^{(j)} \neq c \right\} \mid \mathcal{F}_t \right] &\leq N_0
\end{align*}
$$
Proof. Calculate $\sum_{j=1}^{N_t+1} P \left\{ \xi_t^{(j)} \neq c \mid F_t \right\} = \sum_{j=1}^{N_t+1} \sum_{i=1}^{N_t} \frac{w_t^{(i,j)}}{\bar{w}_t} = N_0$ and
\[
\sum_{j=1}^{N_t+1} \text{Var} \left[ \mathbb{1} \{ \xi_t^{(i)} \neq c \} \mid F_t \right] = \sum_{j=1}^{N_t+1} \frac{w_t^{(i,j)}}{\bar{w}_t} \left( 1 - \frac{w_t^{(N_t+1,j)}}{\bar{w}_t} \right)
\leq \sum_{j=1}^{N_t+1} \left( 1 - \frac{w_t^{(N_t+1,j)}}{\bar{w}_t} \right) = \sum_{j=1}^{N_t+1} \sum_{i=1}^{N_t} \frac{w_t^{(i,j)}}{\bar{w}_t} = N_0
\]

Since all complete resampling schemes share a similar computational cost, it is the accuracy of these schemes that should be the determining factor in deciding which scheme to use. The accuracy of SMC estimates made using various resampling schemes is explored in depth in the subsequent sections.

3.2. Factors contributing to SMC error

The goal of the current section is to show how each step of the SMC algorithm contributes error to SMC estimates and how this error can be reduced by selecting an appropriate resampling scheme.

The starting point for the decomposition of SMC error is the martingale introduced in Theorem 2.4.

\[
\begin{cases}
M_{-1} = E \left[ \prod_{t=0}^{T-1} G_t f \right] \\
M_t = \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} h_t \left( \xi_t^{(i)} \right), \quad 0 \leq t \leq T \\
M_{t+\frac{1}{2}} = \frac{\bar{w}_t}{N_0} \sum_{i=1}^{N_{t+1}} h_t \left( \xi_t^{(i)} \right), \quad 0 \leq t \leq T - 1
\end{cases}
\]

where $h_t(x_t) = E \left[ \prod_{s=t+1}^{T-1} G_s f \mid X_t = x_t \right]$. At time $t = -1$, the martingale is a perfect estimate $M_{-1} = E \left[ \prod_{t=0}^{T-1} G_t f \right]$. At time $t = T$, the martingale has evolved to become an imperfect estimate $M_T = \frac{\bar{w}_T}{N_0} \sum_{i=1}^{N_T} f \left( \xi_T^{(i)} \right)$. An additive decomposition of SMC error is

\[
\frac{\bar{w}_T}{N_0} \sum_{i=1}^{N_T} f \left( \xi_T^{(i)} \right) = E \left[ \prod_{t=0}^{T-1} G_t f \right] - E \left[ \prod_{t=0}^{T-1} G_t f \right] = \left( M_0 - M_{-1} \right) + \sum_{t=0}^{T-1} \left( M_{t+\frac{1}{2}} - M_t \right) + \sum_{t=0}^{T-1} \left( M_{t+\frac{1}{2}} - M_{t+\frac{1}{2}} \right)
\]

In this decomposition, SMC error is the sum of three uncorrelated error sources: initialization error, resampling error and mutation error. The first error source is initialization error, which can be written

\[
M_0 - M_{-1} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left( G_0 \left( \xi_0^{(i)} \right) h_0 \left( \xi_0^{(i)} \right) - E[G_0 h_0] \right)
\]
Initialization error is caused by random sampling of the particles \( \left( \xi_s^{(i)} \right)_{1 \leq i \leq N_0} \) during the initialization step. The mean squared initialization error can be calculated

\[
E \left| M_0 - M_{-1} \right|^2 = \text{Var} \left[ \frac{1}{N_0} \sum_{i=1}^{N_0} G_0 \left( \xi_s^{(i)} \right) h_0 \left( \xi_s^{(i)} \right) \right] = \frac{1}{N_0} \text{Var} \left[ G_0 h_0 \right]
\]

This error source is the same for all resampling schemes, with no dependence on the particular resampling scheme that is used.

Similar to initialization error is mutation error. Mutation error arises from the random sampling of particles \( \left( \xi_s^{(i)} \right)_{1 \leq i \leq N_t} \) during a mutation step. Mutation error \( M_{t+1} - M_{t+\frac{1}{2}} \) can be written

\[
\frac{\bar{w}_t}{N_0} \sum_{i=1}^{N_t} \left\{ G_{t+1} \left( \xi_t^{(i)}, \xi_t^{(i)} \right) h_{t+1} \left( \xi_t^{(i)} \right) - E \left[ G_{t+1} h_{t+1} | X_t = \xi_t^{(i)} \right] \right\}
\]

An asymptotic expansion shows how mutation error approaches a fixed asymptotic limit, regardless of which resampling scheme is used:

**Proposition 3.2.** Assume functions \( (G_t)_{0 \leq t \leq T-1} \) are bounded and assume \( \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t x^2 \right] < \infty \). Then, at each time \( 0 \leq t \leq T - 1 \) there exists a constant \( C > 0 \), independent of resampling scheme, such that

\[
\frac{1}{N_0} \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \mathbb{E} \left[ \prod_{s=0}^{t} G_s \text{Var} \left[ G_{t+1} h_{t+1} | X_t \right] \right] \leq E \left| M_{t+1} - M_{t+\frac{1}{2}} \right|^2 \leq \frac{1}{N_0} \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \mathbb{E} \left[ \prod_{s=0}^{t} G_s \text{Var} \left[ G_{t+1} h_{t+1} | X_t \right] \right] + \frac{C}{N_0}
\]

**Proof.** Define \( \nu (x_t) = \text{Var} \left[ G_{t+1} h_{t+1} | X_t = x_t \right] \). By Theorem 2.3,

\[
0 \leq \mathbb{E} \left[ \frac{\bar{w}_t}{N_0} \sum_{i=1}^{N_t} \nu \left( \xi_t^{(i)} \right) \right] - \mathbb{E} \left[ \bar{w}_t \right] \mathbb{E} \left[ \nu \left( \xi_t^{(i)} \right) \right] = \text{Cov} \left[ \frac{\bar{w}_t}{N_0} \sum_{i=1}^{N_t} \nu \left( \xi_t^{(i)} \right) \right] \leq \mathbb{E} \left[ \bar{w}_t \right] \mathbb{E} \left[ \sum_{i=1}^{N_t} \nu \left( \xi_t^{(i)} \right) \right]^\frac{1}{2} \leq \frac{C}{N_0}
\]

Moreover, Theorem 2.4 guarantees \( \mathbb{E} \left[ \bar{w}_t \right] = \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \) and

\[
\mathbb{E} \left[ \frac{\bar{w}_t}{N_0} \sum_{i=1}^{N_t} \nu \left( \xi_t^{(i)} \right) \right] = \mathbb{E} \left[ \prod_{s=0}^{t} G_s \text{Var} \left[ G_{t+1} h_{t+1} | X_t \right] \right]
\]
In summary, Proposition 3.2 demonstrates that mutation error, just like initialization error, does not depend on which particular complete resampling scheme is used.

Having discussed two sources of SMC error – initialization error and resampling error – the last error source that remains to be discussed is resampling error. Resampling error can be written

\[
M_{t+1/2} - M_t = \frac{w_t}{N_0} \sum_{j=1}^{N_{t+1}} h_t(\hat{\xi}_t^{(j)}) - \frac{1}{N_0} \sum_{i=1}^{N_t} h_t(\xi_t^{(i)})
\]

Resampling error results from random population changes during the resampling step. Resampling error exhibits quite different behavior from initialization and mutation error: the size of this error can vary significantly depending on which particular resampling scheme is used.

A tool for measuring resampling error \cite{6} is resampling variance

\[
\hat{V}_t^2[h_t] = \text{Var} \left[ \frac{w_t}{N_0} \sum_{j=1}^{N_{t+1}} h_t(\hat{\xi}_t^{(j)}) \right] \mathcal{F}_t
\]

Reducing resampling variance is a means toward increasing SMC efficiency. As illustrated in the next lemma, resampling variance can be reduced by selecting an appropriate resampling scheme:

**Lemma 3.1.** (a) Let \( h_t \in \mathbb{R}^{N_{t+1}} \) denote the vector with \( h_t^{(i)} = h_t(\xi_t^{(i)}) \) for \( 1 \leq i \leq N_t \) and \( h_t^{(N_{t+1})} = 0 \). Then, resampling variance \( \hat{V}_t^2[h_t] \) can be written as a quadratic function of the resampling matrix \( W_t \):

\[
\frac{w_t}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} \left| h_t(\xi_t^{(i)}) \right|^2 - \frac{1}{N_0^2} h_t^T W_t W_t^T h_t
\]

Consequently, minimizing resampling variance is a concave minimization problem.

(b) Consider a resampling matrix \( W_t \) containing a sequence of columns \( c_{j_1}, c_{j_2}, \ldots, c_{j_K} \). Then, replacing the columns \( c_{j_1}, c_{j_2}, \ldots, c_{j_K} \) with \( K \) identical columns \( \frac{1}{K} \sum_{k=1}^{K} c_{jk} \) either increases resampling variance or leaves resampling variance unchanged.
Proof of Lemma 3.1. Resampling variance $\hat{V}_t^2[\hat{h}_t]$ can be written as

$$\text{Var} \left[ \frac{w_t}{N_0} \sum_{j=1}^{N_{t+1}} h_t(\hat{\xi}_t^{(j)}) \bigg| \mathcal{F}_t \right] = \frac{w_t^2}{N_0^2} \sum_{j=1}^{N_{t+1}} \text{Var} \left[ h_t(\hat{\xi}_t^{(j)}) \bigg| \mathcal{F}_t \right]$$

$$= \frac{w_t^2}{N_0^2} \sum_{j=1}^{N_{t+1}} \mathbb{E} \left[ \left| h_t(\hat{\xi}_t^{(j)}) \right|^2 \bigg| \mathcal{F}_t \right] - \frac{w_t^2}{N_0^2} \sum_{j=1}^{N_{t+1}} \left( \mathbb{E} \left[ h_t(\hat{\xi}_t^{(j)}) \bigg| \mathcal{F}_t \right] \right)^2$$

$$= \frac{w_t^2}{N_0} \sum_{j=1}^{N_{t+1}} \mathbb{E} \left[ \left| h_t(\hat{\xi}_t^{(j)}) \right|^2 \bigg| \mathcal{F}_t \right] - \frac{1}{N_0} \sum_{j=1}^{N_{t+1}} \sum_{i=1}^{N_t} w_t^{(i,j)} h_t(\hat{\xi}_t^{(i)})^2$$

$$= \frac{w_t}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} h_t(\hat{\xi}_t^{(i)})^2 - \frac{1}{N_0} h_t^T W_t W_t^T h_t$$

Next, let $W_t^{(1)}, W_t^{(2)}, \ldots, W_t^{(K)}$ be the resampling matrices formed by cyclic permutations of columns $c_{j1}, c_{j2}, \ldots, c_{jk}$. Then, $h_t^T W_t^{(k)} W_t^{(k)^T} h_t = h_t^T W_t W_t^T h_t$ for each $1 \leq k \leq K$. By convexity of $W_t \mapsto h_t^T W_t W_t^T h_t$,

$$h_t^T \left( \frac{1}{K} \sum_{k=1}^{K} W_t^{(k)} \right) \left( \frac{1}{K} \sum_{k=1}^{K} W_t^{(k)} \right)^T h_t$$

$$\leq \frac{1}{K} \sum_{k=1}^{K} h_t^T W_t^{(k)} W_t^{(k)^T} h_t = h_t^T W_t W_t^T h_t$$

The second part of Lemma 3.1 is a useful device for comparing common resampling schemes. In examples below, the lemma is used to analyze efficiency of three common resampling schemes: stratified, multinomial residual, and stratified residual resampling. See also Figure 3, which provides resampling matrices for these three schemes.

Example 3.1 (Stratified resampling). In stratified resampling [21], sample uniform random variables $U_t^{(j)} \sim \text{Unif} \left( \frac{i-1}{N_0}, \frac{i}{N_0} \right)$ for $1 \leq j \leq N_0$ and select particles $\hat{\xi}_t^{(j)} = Q_t \left( U_t^{(j)} \right)$, where

$$Q_t(x) = \xi_t^{(i)}$$

$$\frac{i-1}{N_0} \leq x < \frac{i}{N_0}$$

It is seen in Figure 3 that the resampling matrix for stratified resampling takes a particular form, with nonzero matrix entries forming a path rightwards and downwards. By averaging over all matrix columns, the multinomial resampling matrix is obtained. Thus, by Lemma 3.1, the resampling variance of stratified resampling is always as low or lower than that of multinomial resampling.
Example 3.2. In multinomial residual resampling [2], first select \( \left\lfloor \frac{w(i)}{w_t} \right\rfloor \) copies of each particle \( \xi(i) \). Then, select an additional \( R_t = \sum_{i=1}^{N_t} \left\{ \frac{w(i)}{w_t} \right\} \) particles \( \hat{\xi}(j) \) independently from the distribution

\[
\hat{\xi}(j) \sim \frac{1}{R_t} \sum_{i=1}^{N_t} \left\{ \frac{w(i)}{w_t} \right\} \delta\left(\xi(i)\right)
\]

It is seen in Figure 3 that the resampling matrix for multinomial residual resampling contains a block of columns with just one nonzero matrix entry per column. By averaging over all matrix columns, the multinomial resampling matrix is obtained. Thus, by Lemma 3.1, the resampling variance of stratified resampling is always as low or lower than that of multinomial resampling.

Example 3.3. Stratified residual resampling [1] combines aspects of stratified resampling and multinomial residual resampling. First select \( \left\lfloor \frac{w(i)}{w_t} \right\rfloor \) copies of each particle \( \xi(i) \). Then, for \( 1 \leq j \leq R_t = \sum_{i=1}^{N_t} \left\{ \frac{w(i)}{w_t} \right\} \), sample a uniform random variable \( U(j) \sim \text{Unif}\left(\frac{j-1}{R_t}, \frac{j}{R_t}\right) \) and select the particle \( \hat{\xi}(j) = Q_t\left(U(j)\right) \), where

\[
Q_t(x) = \xi(i), \quad \frac{1}{R_t} \sum_{k=1}^{i-1} \left\{ \frac{w(k)}{w_t} \right\} \leq x < \frac{1}{R_t} \sum_{k=1}^{i} \left\{ \frac{w(k)}{w_t} \right\}
\]

The resampling matrix for stratified residual resampling contains a block of columns where entries for a path rightwards and downwards. By averaging over this block of columns, the multinomial residual matrix is obtained. By Lemma 3.1, the resampling variance of stratified residual resampling is as low or lower than that of multinomial residual resampling.

Fig 3. Examples of resampling matrices \( W_0 \) when \( N_0 = 4, N_1 = 4 \), and particles have weights \( w_0(1) = 3.2, w_0(2) = 2.4, w_0(3) = .8, \) and \( w_0(4) = 1.6 \).

\[
\begin{array}{cccc}
.8 & .8 & .8 & .8 \\
.6 & .6 & .6 & .6 \\
.2 & .2 & .2 & .2 \\
.4 & .4 & .4 & .4 \\
\end{array}
\quad
\begin{array}{cccc}
2 & 1.2 & 1 & 1.6 \\
.8 & .8 & .4 & .4 \\
.2 & .2 & .4 & .4 \\
.4 & .4 & 1 & 1.6 \\
\end{array}
\quad
\begin{array}{cccc}
2 & .6 & .6 & 1.2 \\
2 & .6 & .2 & 1.6 \\
2 & .2 & .4 & .4 \\
.4 & .4 & .4 & .4 \\
\end{array}
\quad
\begin{array}{cccc}
2 & 1.2 & 1 & 1.6 \\
2 & .6 & .6 & 1.2 \\
2 & .6 & .2 & 1.6 \\
2 & .2 & .4 & .4 \\
\end{array}
\]

multinomial  
stratified  
multinomial residual  
stratified residual

Remark 3.1. While Proposition 3.2 requires that functions \( (G_t)_{t \geq 0} \) are bounded, this assumption can be lifted, at the cost of greater complexity. Using methods to be presented in Section 3.4, it can be shown that mutation error converges
in distribution

\[
\frac{1}{\sqrt{N_0}} \left( M_{t+1} - M_t \right) \overset{D}{\rightarrow} N \left( 0, \frac{1}{\sqrt{N}} \right)
\]

whenever the asymptotic variance is finite. The asymptotic distribution does not depend on which resampling scheme is used.

**Remark 3.2.** Similar to the examples above, Douc et al. [6] compared resampling variance between different resampling schemes. But while [6] used explicit resampling variance calculations, the resampling matrix framework provides a quicker route to comparing schemes. In the examples above, it is enough simply to compare columns between resampling matrices and apply Lemma 3.1 to obtain a rigorous error comparison.

### 3.3. Minimizing resampling variance

The goal of SMC is to compute a quantity \( \mathbb{E} \left[ \prod_{s=t+1}^{T} G_s \right] \) with minimal error. Sections 3.1 and 3.2 have demonstrated that the error of an estimate depends critically on the resampling variance. Thus, it is of foremost concern to find resampling schemes that minimize resampling variance.

Theorem 3.1 identifies the minimal variance resampling scheme, a scheme that sorts particles depending on the values \( h_t(\xi_t^{(i)}) \):

**Theorem 3.1.** (a) The following random population scheme minimizes resampling variance \( \hat{V}_t^2 [h_t] \):

1. Add one particle \( \left( w^{(N_t+1)}_t, \xi^{(N_t+1)}_t \right) = (\bar{w}_t, c) \) to the ensemble \( \left( w^{(j)}_t, \xi^{(j)}_t \right)_{1 \leq j \leq N_t} \).
2. Sort the ensemble \( \left( w^{(j)}_t, \xi^{(j)}_t \right)_{1 \leq j \leq N_t+1} \) from highest to lowest by the value of \( h_t(\xi^{(j)}_t) \) so that
   \[
   h_t(\xi^{(1)}_t) \geq h_t(\xi^{(2)}_t) \geq \cdots \geq h_t(\xi^{(N_t)}_t) \geq h_t(\xi^{(N_t+1)}_t)
   \]
3. Apply stratified resampling.

(b) The fixed population scheme that minimizes resampling variance \( \hat{V}_t^2 [h_t] \) is a simpler version of the scheme in part (a). First sort particles from highest to lowest by the value of \( h_t(\xi^{(j)}_t) \) and then apply stratified resampling.

**Proof.** Assume particles have been sorted so that \( h_t(\xi^{(1)}_t) \geq h_t(\xi^{(2)}_t) \geq \cdots \geq h_t(\xi^{(N_t)}_t) \) and consider an arbitrary resampling matrix \( W_t \). By Lemma 3.1, the resampling variance is decreased if \( h^T_t W_t W_t^T h_t \) is increased.
As a first step toward increasing $h_t^T W_t W_t^T h_t$, define $P_t$ and $N_t$ by

$$
\left\{
\begin{array}{l}
p_t^{(i,j)} = w_t^{(i,j)} \mathbf{1} \left\{ h_t \left( \xi_t^{(j)} \right) \geq 0 \right\}, \quad 1 \leq i \leq N_t \\
q_t^{(i,j)} = w_t^{(i,j)} \mathbf{1} \left\{ h_t \left( \xi_t^{(j)} \right) < 0 \right\}, \quad 1 \leq i \leq N_t 
\end{array}
\right.
$$

and

$$
\left\{
\begin{array}{l}
p_t^{(N_t+1,j)} = w_t - \sum_{k=1}^{N_t} p_t^{(k,j)} \\
n_t^{(N_t+1,j)} = w_t - \sum_{k=1}^{N_t} q_t^{(k,j)}
\end{array}
\right.
$$

Then set $S_t = (P_t \ Q_t)$ and observe that $h_t^T W_t W_t^T h_t \leq h_t^T S_t S_t^T h_t$.

Let $c^{(1)}, c^{(2)}, \ldots, c^{(N_t+1)}$ denote the columns of $P_t$, sorted so that $h_t^T c^{(1)} \geq h_t^T c^{(2)} \geq \cdots \geq h_t^T c^{(N_t+1)} \geq 0$. Consider the following algorithm to increase the value of $h_t^T P_t P_t^T h_t$:

1. Call a quadruplet $(i, j, k, \ell)$ a problematic quadruplet if $p_t^{(i,j+k)} > 0$ and $p_t^{(i+k,j)} > 0$ and if $i + k \leq N_t$. Choose a problematic quadruplet with $i$ as small as possible. If there is more than one such quadruplet, choose one with $j$ as small as possible.

2. Set $\alpha = \min \left\{ p_t^{(i,j+k)}, p_t^{(i+k,j)} \right\}$ and update the entries of $P_t$ with

$$
\begin{align*}
p_t^{(i,j)} &= p_t^{(i,j)} + \alpha \\
p_t^{(i,j+k)} &= p_t^{(i,j)} - \alpha \\
p_t^{(i+k,j)} &= p_t^{(i,j+k)} - \alpha \\
p_t^{(i+k,j+k)} &= p_t^{(i,j+k)} + \alpha
\end{align*}
$$

3. If necessary, resort the columns $c^{(j+k)}, c^{(j+k+1)}, \ldots, c^{(N_t+1)}$ to ensure that $h_t^T c^{(j+k)} \geq h_t^T c^{(j+k+1)} \geq \cdots \geq h_t^T c^{(N_t+1)}$.

Note that step 2 of the algorithm increases $h_t^T P_t P_t^T h_t$ or leaves $h_t^T P_t P_t^T h_t$ unchanged, while step 3 ensures that $h_t^T c^{(1)} \geq h_t^T c^{(2)} \geq \cdots \geq h_t^T c^{(N_t+1)} \geq 0$. After repeated applications of the algorithm, all the problematic quadruplets involving column $c^{(1)}$ will eventually be gone and the same too with columns $c^{(2)}$, $c^{(3)}$, etc. Eventually, the algorithm will have no more problematic quadruplets to correct. A similar algorithm can be applied to increase the value of $N_t$. On examination it is seen that the resulting matrix $(Q_t)$ generates the same resampling scheme as described in part (a).

The proof of part (b) is similar. Because fixed population schemes satisfy $\hat{V}_t^2 \left[ h_t \right] = \hat{V}_t^2 \left[ h_t + c \right]$ for $c \in \mathbb{R}$, it can be assumed without loss of generality that $h_t \left( \xi_t^{(i)} \right) > 0$ for all $\xi_t^{(i)} \neq c$. But in this case, the schemes in part (a) and part (b) are identical, and sorted stratified resampling represents the best possible strategy.

The optimal scheme identified in Theorem 3.1 is an example of a sorting scheme. In more general sorting schemes, particles can be sorted using any real-valued coordinate $\theta_t : E_t \rightarrow \mathbb{R}$ and then stratified resampling or stratified residual resampling can be used. Sorting schemes have a long history dating back at least to Madow & Madow [25]. Sorting schemes can lead to a beneficial
stratification effect. Each particle $\hat{\xi}_t^{(j)}$ is drawn from a subset of particles for which $h_t\left(\hat{\xi}_t^{(i)}\right)$ values are similar, thereby reducing resampling variance. Theorem 3.1 indicates that the best possible coordinate for sorting is $\theta_t = h_t$. This is the first known result which proves that sorting particles can produce an optimal resampling scheme.

The optimal scheme in Theorem 3.1 can be difficult to implement exactly, because the function $h_t(x_t) = E \left[ \prod_{s=t+1}^{T-1} G_s f \bigg| X_t = x_t \right]$ can be challenging to compute. However, $h_t$ is not the only coordinate for sorting particles that can lead to an effective resampling scheme. The error formulas of the next section show that effective sorting is possible with a wide range of different coordinates $\theta_t$.

3.4. Asymptotic error

In past work [5, 4, 7], a central tool for for analyzing SMC error has been Central Limit Theorems (CLTs) of the form

$$\sqrt{N_0} \left( \frac{w_{T-1}}{N_0} \sum_{i=1}^{N_T} f \left( \hat{\xi}_{T-1}^{(i)}, \hat{\xi}_{T}^{(i)} \right) - E \left[ \prod_{t=0}^{T-1} G_t f \bigg| X_t = x_t \right] \right) \overset{D}{\rightarrow} N(0, \eta^2)$$

where the quantity $\eta^2$ depends on the particular resampling scheme that is used. CLTs have been proved for multinomial, multinomial residual and Bernoulli resampling [5, 4, 7]. In the present section, new error formulas are presented for stratified resampling and stratified residual resampling. These error formulas are not CLTs; instead they are upper bounds on asymptotic error. Asymptotic error is a new way to measure error that is more general than a CLT and also more flexible for analysis. Before presenting asymptotic error formulas, it is therefore necessary to introduce the key features of asymptotic error and explain how this error measurement tool can be interpreted.

Asymptotic error is a far-reaching generalization of the error rate in a CLT. In a CLT, a sequence of random variables $(Y_n)_{n \geq 1}$ approach a constant $c$ with error measured by an error rate $U_n$.

$$\frac{Y_n - c}{U_n} \overset{D}{\rightarrow} N(0, 1)$$

Thus, a CLT can only be proved when there is very precise knowledge of the error rate $U_n$. In contrast, asymptotic error can be analyzed when knowledge of $U_n$ is less precise and there is only an upper or lower bound on $U_n$. A full definition of asymptotic error is provided below:

**Definition 3.1.** Suppose random variables $(Y_n)_{n \geq 1}$ satisfy

$$\liminf_{n \to \infty} E \left[ \mathbb{1}_{A_n} \left| \frac{Y_n - c}{U_n} \right|^2 \right] \geq 1$$
lemma 3.2. guaranteed by the following lemma:

\[ \lim sup_{n \to \infty} E \left[ \mathbb{I}_{B_n} \left| \frac{Y_n - c}{U_n} \right|^2 \right] \leq 1 \]

for some sequence of sets \((B_n)_{n \geq 1}\) with \( P (B_n) \to 1 \). Then, \( Y_n \) converges to \( c \) with asymptotic error less than or equal to \( U_n \), and write \( |Y_n - c| \gtrsim U_n \).

Suppose random variables \((Y_n)_{n \geq 1}\) satisfy

A CLT can be viewed as a particular example of asymptotic error, as guaranteed by the following lemma:

**Lemma 3.2.** Suppose random variables \((Y_n)_{n \geq 1}\) satisfy \( \frac{Y_n - c}{U_n} \overset{D}{\to} N(0, 1) \) as \( n \to \infty \). Then, \( |Y_n - c| \sim U_n \).

**Proof.** Fatou’s Lemma shows \( \liminf_{n \to \infty} E \left[ \mathbb{I}_{A_n} \left| \frac{Y_n - c}{U_n} \right|^2 \right] \geq 1 \) for all sequences \((A_n)_{n \geq 1}\) with \( P (A_n) \to 1 \). Thus, \( |Y_n - c| \gtrsim U_n \). To show \( |Y_n - c| \lesssim U_n \), construct a sequence \((B_n)_{n \geq 1}\) with the properties \( P (B_n) \to 1 \) and \( \limsup_{n \to \infty} E \left[ \mathbb{I}_{B_n} \left| \frac{Y_n - c}{U_n} \right|^2 \right] \leq 1 \).

1. First let \( L_n \) be the largest number such that

\[ E \left[ \mathbb{I} \left\{ \left| \frac{Y_n - c}{U_n} \right| < L_n \right\} \left| Y_n - c \right|^2 \right] \leq 1 \]

and note that \( L_n \) is well-defined by the Monotone Convergence Theorem. Set \( B_n = \left\{ \left| \frac{Y_n - c}{U_n} \right| < L_n \right\} \). For any \( \epsilon > 0 \), choose \( M > 0 \) large enough that \( P \left\{ |Z| < M \right\} > 1 - \frac{\epsilon}{2} \) where \( Z \sim N(0, 1) \). Since \( \frac{Y_n - c}{U_n} \overset{D}{\to} N(0, 1) \), it follows that \( P \left\{ \left| \frac{Y_n - c}{U_n} \right| < M \right\} > 1 - \epsilon \) for all \( n \) large enough. Since \( x \mapsto \mathbb{I} \{ |x| < M \} |x|^2 \) is bounded and piecewise continuous,

\[ E \left[ \mathbb{I} \left\{ \left| \frac{Y_n - c}{U_n} \right| < M \right\} \left| Y_n - c \right|^2 \right] \to E \left[ \mathbb{I} \{ |Z| < M \} Z^2 \right] < 1 \]

For all \( n \) large enough, it follows that \( M < L_n \), \( \left\{ \left| \frac{Y_n - c}{U_n} \right| < M \right\} \subseteq B_n \), and \( P (B_n) \geq P \left\{ \left| \frac{Y_n - c}{U_n} \right| < M \right\} > 1 - \epsilon \). Since \( \epsilon \) is arbitrary, \( P (B_n) \to 1 \).

Asymptotic error can be compared to mean squared error, which is another common error metric, different from the error rate in the CLT. Both asymptotic error and mean squared error are tools to assess the value of an estimate and to provide confidence intervals around an estimate. By Chebyshev’s inequality,
asymptotic error leads to confidence intervals:

$$|Y_n - c| \lesssim U_n \implies \limsup_{n \to \infty} P \{|Y_n - c| \geq \epsilon U_n\} \leq \frac{1}{\epsilon^2}$$

The chief difference between asymptotic error and mean squared error is robustness to perturbations. Mean squared error $E|Y_n - c|^2$ is quite sensitive to changes in the behavior of $Y_n$ on a set of vanishing probability, but asymptotic error is completely robust to these changes. Thus the confidence intervals derived from asymptotic error bounds can be much tighter than those derived from mean squared error bounds.

The rigorous treatment of asymptotic error leads to new results in SMC analysis, including the first known error formulas for stratified resampling and stratified residual resampling. In the following theorem, these new formulas are presented alongside CLTs for multinomial, multinomial residual, and Bernoulli resampling, which are extended from [5, 4, 7] to have less restrictions on functions $(G_t)_{t \geq 0}$ and $f$:

**Theorem 3.2.** Assume $E \left[ \prod_{s=0}^t G_s \right] < \infty$ for $0 \leq t \leq T - 1$, $E|G_0h_0|^2 < \infty$, and $E \left[ \prod_{s=0}^t G_s |G_{t+1}h_{t+1}|^2 \right]$ for $0 \leq t \leq T - 1$. If multinomial residual or stratified residual resampling is used, assume $E \left[ \prod_{s=0}^{t-1} G_s \{\tilde{G}_t \in \{1, 2, \ldots\}\} \right] = 0$ for $0 \leq t \leq T - 1$ as well. Set $\tilde{G}_t = E \left[ \prod_{s=0}^{t-1} G_s \right] G_t / E \left[ \prod_{s=0}^t G_s \right]$ and set

$$\eta^2 = \text{Var}[G_0h_0] + \sum_{t=0}^{T-1} \eta_t^2 [h_t] + \sum_{t=0}^{T-1} \left( E \left[ \prod_{s=0}^t G_s \right] E \left[ \prod_{s=0}^{t} G_s \text{Var}[G_{t+1}h_{t+1}] X_t \right] \right)$$

where $\eta^2$ depends on a sequence of numbers $(\eta_t [h_t])_{0 \leq t \leq T-1}$.

First assume multinomial resampling, Bernoulli resampling, or multinomial residual resampling is used. Then SMC estimates satisfy the CLT

$$\sqrt{N_0} \left( \frac{\sum_{i=1}^{N_T} f \left( \tilde{z}^{(i)}_{T-1}, \tilde{z}^{(i)}_T \right) - E \left[ \prod_{t=0}^{T-1} G_t f \right]}{N_0} \right) \overset{D}{\to} N(0, \eta^2)$$

where $\tilde{z}^2 [h_t]$ is determined by the resampling scheme:

- **multinomial**
  $$\left( E \left[ \prod_{s=0}^t G_s \right] \right)^2 \min_{c \in \mathbb{R}} E \left[ \prod_{s=0}^t \tilde{G}_s |h_t - c|^2 \right]$$

- **multinomial residual**
  $$\left( E \left[ \prod_{s=0}^t G_s \right] \right)^2 \min_{c \in \mathbb{R}} E \left[ \prod_{s=0}^{t-1} \tilde{G}_s \{\tilde{G}_t\} |h_t - c|^2 \right]$$

- **Bernoulli**
  $$\left( E \left[ \prod_{s=0}^t G_s \right] \right)^2 E \left[ \prod_{s=0}^t \tilde{G}_s \{\tilde{G}_t\} \left( 1 - \{\tilde{G}_t\} \right) h_t^2 \right]$$
Next, assume that at each resampling step particles are sorted by a coordinate \( \theta_t \) and then stratified or stratified residual resampling is used. Then,

\[
\left| \frac{\mathbb{I}_{T-1}}{N_0} \sum_{i=1}^{N_0} f \left( \xi_{T-1}^{(i)}, \xi_{T}^{(i)} \right) - \mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right] \right| \lesssim \frac{\eta_t}{\sqrt{N_0}}
\]

where \( \eta_t^2 [h_t] \) is determined by the resampling scheme:

- **stratified**
  \[
  \left( \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \right)^2 \text{min}_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t} \tilde{G}_s | h_t - p(\theta_t)^2 \right]
  \]

- **stratified residual**
  \[
  \left( \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \right)^2 \text{min}_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t-1} \tilde{G}_s \left\{ \tilde{G}_t \right\} | h_t - p(\theta_t)^2 \right]
  \]

There are two main conclusions that can be drawn from Theorem 3.2 about how best to choose a resampling scheme. The first conclusion is that residual versions of a resampling scheme should be used whenever possible. Error formulas for multinomial and multinomial residual resampling are differentiated by a factor of \( \tilde{G}_t \) for multinomial and a factor of \( \left\{ \tilde{G}_t \right\} \) for multinomial residual resampling. Since \( \left\{ \tilde{G}_t \right\} \) is always as low or lower than \( \tilde{G}_t \), the multinomial residual resampling scheme can lead to reduced SMC error. Similarly, stratified residual resampling has an improved asymptotic error upper bound compared to stratified resampling.

The second conclusion that follows from Theorem 3.2 is that sorting schemes can substantially reduce error, depending on the coordinate \( \theta_t \) used for sorting. Error formulas for multinomial and stratified resampling are distinguished by a factor of \( \text{min}_{c \in \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t} \tilde{G}_s | h_t - c |^2 \right] \) for multinomial and a factor of \( \text{min}_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t} \tilde{G}_s | h_t - p(\theta_t) |^2 \right] \) for stratified resampling. Since \( \text{min}_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t} \tilde{G}_s | h_t - p(\theta_t) |^2 \right] \leq \text{min}_{c \in \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t} \tilde{G}_s | h_t - c |^2 \right] \), asymptotic error for stratified resampling is as low or lower than asymptotic error for multinomial resampling. In the simplest case where \( \theta_t \equiv 0 \), particles are not sorted in any particular order and error reduction may be very mild; on the other hand, as the stratification effect due to sorting by \( \theta_t \) increases, the error contributed at each resampling step approaches zero. Similarly, asymptotic error for stratified residual resampling is as low or lower than asymptotic error for multinomial residual resampling, with a major reduction possible depending on the coordinate \( \theta_t \).

Below, two examples of resampling schemes that use sorting to achieve error reduction are described:

**Example 3.4** (Sorting in \( \mathbb{R}^d \)). When applying SMC to a one-dimensional system, Kitagawa [21] sorted particles \( \xi_t^{(i)} \) by their values in \( \mathbb{R} \) and then applied stratified resampling, leading to a dramatic reduction in resampling variance. Later, Gerber et al. [10] suggested a more general strategy of sorting particles in \( \mathbb{R}^d \) according to a Hilbert curve, a measurable one-to-one mapping from \( \mathbb{R}^d \) into \( \mathbb{R} \). In both cases, Theorem 3.2 gives an upper bound on asymptotic error with
It should be noted however that pre-asymptotic resampling variance for this sorting strategy is difficult to estimate; further research may help elucidate the practical efficiency of Hilbert curve sorting.

Example 3.5 (Binning). In binned resampling [19], the state space is sorted into bins \( B_1, B_2, \ldots, B_K \), and particles \( \xi^{(i)}_t \) are arranged by bin number, from highest to lowest. When stratified resampling is applied, Theorem 3.2 gives an upper bound on asymptotic error with

\[
\hat{\eta}^2_t [h_t] = \left( \mathbb{E} \left[ \prod_{s=0}^t G_s \right] \right)^2 \mathbb{E} \left[ \prod_{s=0}^t \tilde{G}_s \sum_{k=1}^K 1_{B_k} \right] h_t - \frac{\mathbb{E} \left[ \prod_{s=0}^t \tilde{G}_s 1_{B_k} h_t \right]}{\mathbb{E} \left[ \prod_{s=0}^t \tilde{G}_s \right]}^2
\]

As values of \( h_t \) become increasingly similar in each bin \( B_k \), equation (3.5) guarantees that asymptotic error must decrease. In particular, as the diameter of the bins approaches zero in a region that grows to fill the state space \( E_t \), \( \hat{\eta}^2_t [h_t] \) approaches the lowest possible level: \( \hat{\eta}^2_t [h_t] = 0 \).

4. Conclusion

The present work derives a theoretical framework that unifies past SMC scholarship and establishes significant new results. The framework uses a simple parametrization to describe a great variety of resampling schemes. The theoretical framework includes a unified error analysis and asymptotic error formulas with a unified structure that can be used to compare resampling schemes.

The resampling matrix framework combines a fresh look at common resampling schemes with new technical tools to analyze SMC error. Asymptotic error is defined in a new way, as mean squared error outside a set of vanishing probability. This notion of error leads to simple proofs and rigorous comparisons between resampling schemes. Due to this innovation, asymptotic error formulas are now available for stratified resampling and stratified residual resampling, including the full range of unbounded functions \( (G_t)_{t \geq 0} \) and \( f \) used in practical implementations of SMC.

The framework leads to two concrete recommendations for how best to resample:

1. Firstly, practitioners are encouraged to use stratified residual resampling instead of multinomial residual resampling and stratified resampling instead of multinomial resampling in order to reduce resampling variance. Similar recommendations were given in Douc et al. [6], but resampling matrices provide a more intuitive and general explanation for reductions in resampling variance.

2. Secondly, sorting schemes can lead to extremely low asymptotic error rates. These schemes are recommended when there is a coordinate \( \theta_t \) that can be used to sort particles \( \left( \xi^{(i)}_t \right)_{1 \leq i \leq N_t} \) to achieve a beneficial stratification effect in the resampling step.
In summary, the unifying analysis in the current paper shines light on the best ways to resample, providing practical guidance to help SMC users make the most of the powerful and versatile SMC algorithm.

5. Appendix

5.1. Estimates of ratios

SMC is often used to approximate ratios

$$\frac{\sum_{j=1}^{N_T} w_{T-1}^{(j)} f\left(\hat{\zeta}^{(j)}_{T-1}, \hat{\zeta}^{(j)}_{T}\right)}{\sum_{j=1}^{N_T} w_{T-1}^{(j)} \mathbb{1}\left\{\hat{\zeta}^{(j)}_{T-1} \neq c\right\}} \approx \frac{E\left[\prod_{t=0}^{T-1} G_t f\right]}{E\left[\prod_{t=0}^{T-1} G_t\right]}$$

In some cases, the denominator in the SMC estimate may equal zero, and the estimate can be assigned an arbitrary value when this occurs. If $E\left[\prod_{t=0}^{T-1} G_t\right] < \infty$ for $0 \leq t \leq T - 1$ and $E\left[\prod_{t=0}^{T-1} G_t f\right] < \infty$, then Theorem 2.2 guarantees

$$\left\{\frac{\sum_{j=1}^{N_T} w_{T-1}^{(j)} f\left(\hat{\zeta}^{(j)}_{T-1}, \hat{\zeta}^{(j)}_{T}\right)}{\sum_{j=1}^{N_T} w_{T-1}^{(j)} \mathbb{1}\left\{\hat{\zeta}^{(j)}_{T-1} \neq c\right\}}\right\} \to P E\left[\prod_{t=0}^{T-1} G_t f\right] \to E\left[\prod_{t=0}^{T-1} G_t\right]$$

Therefore, SMC estimates of ratios are convergent. While expressions for the bias and variance of these estimates are challenging to derive, asymptotic error for these estimates can be studied with the aid of the following lemma:

**Lemma 5.1.** 1. If $Y_n \overset{D}{=} N(0,1)$ and $W_n \overset{P}{=} c \in \mathbb{R}$, then $Y_n W_n \overset{D}{=} N(0,c^2)$.
2. If $|Y_n| \lesssim 1$ and $W_n \overset{P}{=} c > 0$, then $|Y_n W_n| \lesssim c$.

**Proof.** Part 1 follows from Slutsky’s Theorem. To prove part 2, first construct a sequence of sets $(C_n)_{n \geq 1}$ with the properties $P(C_n) \to 1$ and $\lim sup_{n} \|\mathbb{1}_{C_n} \frac{W_n}{c}\|_{\infty} \leq 1$. Set $E_{m,n} = \left\{\frac{W_n}{c} \leq 1 + \frac{1}{m}\right\}$ for $m, n \geq 1$ and $D_{1,n} = E_{1,n}$ for $n \geq 1$. By the hypothesis in part 2, there exists a number $N(m) \geq 1$ such that $P(E_{m,n}) \geq 1 - \frac{1}{m}$ for $n \geq N(m)$. Accordingly, for $m > 1$ define

$$\begin{cases} D_{m,n} = D_{m-1,n}, & n < N(m) \\ D_{m,n} = E_{m,n}, & n \geq N(m) \end{cases}$$

By this construction, for $m \geq M$ and $n \geq N(M)$, $\|\mathbb{1}_{D_{m,n}} \frac{W_n}{c}\|_{\infty} \leq 1 + \frac{1}{M}$ and $P(D_{m,n}) \geq 1 - \frac{1}{M}$. Setting $C_n = D_{n,n}$ gives the required sequence. Lastly, select $(B_n)_{n \geq 1}$ so that $P(B_n) \to 1$ and $\lim sup_{n} E\left[\mathbb{1}_{B_n} Y_n^2\right] \leq 1$. Then $D_n = B_n \cap C_n$ satisfies $P(D_n) \to 1$ and $\lim sup_{n} E\left[\mathbb{1}_{D_n} \left|\frac{Y_n W_n}{c}\right|^2\right] \leq 1$.  \(\square\)
To apply Lemma 5.1, set \( \hat{f} = f - \frac{E[\Pi_{t=0}^{T-1} G_t f]}{E[\Pi_{t=0}^{T} G_t]} \) and observe

\[
\frac{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} \tilde{f} \left( \hat{\xi}_{T-1}, \hat{\xi}_T \right)}{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} 1 \left\{ \hat{\xi}_{T-1} \neq c \right\}} = \frac{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} \tilde{f} \left( \hat{\xi}_{T-1}, \hat{\xi}_T \right)}{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} 1 \left\{ \hat{\xi}_{T-1} \neq c \right\}} - \frac{E \left[ \Pi_{t=0}^{T-1} G_t f \right]}{E \left[ \Pi_{t=0}^{T-1} G_t \right]}
\]

Since \( \frac{1}{N_T} \sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} 1 \left\{ \hat{\xi}_{T-1} \neq c \right\} \xrightarrow{P} E \left[ \Pi_{t=0}^{T-1} G_t \right] \), the asymptotic error of an SMC estimate is the asymptotic error of \( \frac{1}{N_T} \sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} \tilde{f} \left( \hat{\xi}_{T-1}, \hat{\xi}_T \right) \) scaled by a factor of \( \left( E \left[ \Pi_{t=0}^{T-1} G_t \right] \right)^{-1} \). A corollary of Theorem 3.2 gives precise expressions for asymptotic error:

**Corollary 5.1.** Set \( \tilde{G}_t = E \left[ \Pi_{s=0}^{t-1} G_s | G_t \right] / E \left[ \Pi_{s=0}^{t-1} G_s \right] \) and set

\[
\tilde{h}_t(x_t) = E \left[ \prod_{s=t+1}^{T-1} \tilde{G}_s \left( f - \frac{E \left[ \Pi_{r=0}^{T-1} G_r f \right]}{E \left[ \Pi_{r=0}^{T-1} G_r \right]} \right) | X_t = x_t \right]
\]

Assume that \( E \left[ \Pi_{s=0}^{t} G_s \right] < \infty \) for \( 0 \leq t \leq T - 1 \), \( E \left[ |G\tilde{h}_0|^2 \right] < \infty \), and \( E \left[ \Pi_{s=0}^{t} G_s \right] G_{t+1} \tilde{h}_{t+1}^2 \) for \( 0 \leq t \leq T - 1 \). If multinomial residual or stratified residual resampling is used, also assume \( E \left[ \Pi_{s=0}^{t-1} G_s \left\{ \tilde{G}_t \in \{1, 2, \ldots \} \right\} \right] = 0 \) for \( 0 \leq t \leq T - 1 \). Define

\[
\eta^2 = \text{Var} \left[ \tilde{G}_0 \tilde{h}_0 \right] + \sum_{t=0}^{T-1} \eta_t^2 \left[ \tilde{h}_t \right] + \sum_{t=0}^{T-1} E \left[ \prod_{s=0}^{t} \tilde{G}_s \text{Var} \left[ \tilde{G}_{t+1} \tilde{h}_{t+1}^2 | X_t \right] \right]
\]

where \( \eta^2 \) depends on a sequence of numbers \( \left( \eta_t \left[ \tilde{h}_t \right]^2 \right)_{0 \leq t \leq T - 1} \).

First assume multinomial resampling, Bernoulli resampling, or multinomial residual resampling is used. Then SMC estimates satisfy the CLT

\[
\sqrt{N_0} \left( \frac{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} f \left( \hat{\xi}_{T-1}, \hat{\xi}_T \right)}{\sum_{j=1}^{N_T} \tilde{w}_{T-1}^{(j)} 1 \left\{ \hat{\xi}_{T-1} \neq c \right\}} - \frac{E \left[ \Pi_{t=0}^{T-1} G_t f \right]}{E \left[ \Pi_{t=0}^{T-1} G_t \right]} \right) \xrightarrow{D} N(0, \eta^2)
\]

where \( \eta_t^2 \left[ \tilde{h}_t \right] \) is determined by the resampling scheme:

- **multinomial**: \( E \left[ \Pi_{s=0}^{t} \tilde{G}_s \hat{h}_t^2 \right] \)
- **multinomial residual**: \( \min_{c \in \mathbb{R}} E \left[ \Pi_{s=0}^{t-1} \tilde{G}_s \left\{ \tilde{G}_t \right\} | \hat{h}_t - c \right]^2 \)
- **Bernoulli**: \( E \left[ \Pi_{s=0}^{t-1} \tilde{G}_s \left\{ \tilde{G}_t \right\} \left( 1 - \left\{ \tilde{G}_t \right\} \right) \right] \hat{h}_t^2 \)
Next assume at each resampling step particles are sorted by a coordinate \( \theta_i \) and then stratified or stratified residual resampling is used. Then,

\[
\frac{\sum_{j=1}^{N_T} \tilde{w}_t^{(j)} f \left( \tilde{\xi}_{t-1}^{(j)}, \xi_t^0 \right)}{\sum_{j=1}^{N_T} \tilde{w}_t^{(j)} I \left\{ \xi_t^{(j)} \neq c \right\}} - \frac{\mathbb{E} \left[ \prod_{t=0}^{T-1} G_t f \right]}{\mathbb{E} \left[ \prod_{t=0}^{T-1} G_t \right]} \stackrel{\text{a.s.}}{\sim} \frac{\eta}{\sqrt{N_0}}
\]

where \( \tilde{n}_t^2 \left[ \tilde{h}_t \right] \) is determined by the resampling scheme:

\[
\begin{align*}
\text{stratified} & \quad \min_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t-1} \tilde{G}_s \left| \tilde{h}_t - p \left( \theta_i \right) \right|^2 \right] \\
\text{stratified residual} & \quad \min_{p: \mathbb{R} \to \mathbb{R}} \mathbb{E} \left[ \prod_{s=0}^{t-1} \tilde{G}_s \left\{ \tilde{G}_t \right\} \left| \tilde{h}_t - p \left( \theta_i \right) \right|^2 \right]
\end{align*}
\]

5.2. Proofs for Theorems 2.2, 2.3, and 3.2

To prove Theorem 2.2, first introduce intermediate \( \sigma \)-algebras between \( \mathcal{F}_{t-1} \) and \( \mathcal{F}_t \):

\[
\begin{align*}
\mathcal{F}_{t}^{(0)} & = \mathcal{F}_{t-1} \\
\mathcal{F}_{t}^{(i)} & = \mathcal{F}_{t}^{(i-1)} \vee \sigma \left( \mathcal{F}_{t}^{(i)} \right), \quad t = 0, 1 \leq i \leq N_t - 1 \\
\mathcal{F}_{t}^{(N_t)} & = \mathcal{F}_t \\
\mathcal{F}_{t}^{(i)} & = \mathcal{F}_{t}^{(i-1)} \vee \sigma \left( \mathcal{F}_{t}^{(i)} \right), \quad t > 0, 1 \leq i \leq N_t - 1
\end{align*}
\]

Next, define a martingale \( M_t^{(i)} = \mathbb{E} \left[ \frac{1}{N_0} \sum_{k=1}^{N_T} w_t^{(k)} f \left( \tilde{\xi}_{t-1}^{(k)}, \xi_t^{(k)} \right) \middle| \mathcal{F}_t^{(i)} \right] \). Since pairs \( \left( w_t^{(k)}, \xi_t^{(k)} \right) \) are conditionally independent given \( \mathcal{F}_{t-1} \), it follows

\[
M_t^{(i)} = \frac{1}{N_0} \sum_{k=1}^{i} w_t^{(k)} h_t \left( \xi_t^{(k)} \right) + \frac{1}{N_0} \sum_{k=i+1}^{N_t} \mathbb{E} \left[ w_t^{(k)} h_t \left( \xi_t^{(k)} \right) \middle| \mathcal{F}_{t-1} \right]
\]

(5.1)

The proof of Theorem 2.2 also requires two technical lemmas.

Lemma 5.2. For each \( n \geq 1 \), suppose \( \mathcal{G}_{n0} \subseteq \mathcal{G}_{n1} \subseteq \mathcal{G}_{n2} \subseteq \cdots \subseteq \mathcal{G}_{nk_n} \) is a filtration and \( \left( Y_{nj} \right)_{1 \leq j \leq k_n} \) is a sequence of random variables with \( Y_{nj} \) measurable in \( \mathcal{G}_{nj} \). Suppose

\[
\begin{align*}
&\left\{ \sum_{j=1}^{k_n} \mathbb{E} \left[ \left. \left| Y_{nj} \right| \right| \mathcal{G}_{nj-1} \geq C \right] \right\} \xrightarrow{P} 0, \quad C > 0 \\
&\lim_{\lambda \to \infty} \sup_{n \geq 1} P \left\{ \sum_{j=1}^{k_n} \mathbb{E} \left[ \left. \left| Y_{nj} \right| \right| \mathcal{G}_{nj-1} \geq \lambda \right] \right\} = 0
\end{align*}
\]

Then, \( \sum_{j=1}^{k_n} \left\{ Y_{nj} - \mathbb{E} \left[ Y_{nj} \middle| \mathcal{G}_{nj-1} \right] \right\} \xrightarrow{P} 0 \).

Proof. The lemma can be traced back to the early martingale literature, particularly Hall & Heyde \[14\], pg. 45-47] and McLeish \[26\], pg. 626]. The lemma appears in Douc & Moulines \[7\], who also use the lemma to prove convergence of SMC schemes. \( \square \)
Lemma 5.3. If \( E \left[ \prod_{s=0}^{t} G_s \right] < \infty \) for \( 0 \leq t \leq T-1 \) and \( E \left| \prod_{t=0}^{T-1} G_t f \right| < \infty \), then for each \( 0 \leq t \leq T \) and \( C > 0 \)

\[
E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} \hat{\omega}_{t-1}^{(i)} |f(\xi_t^{(i)})| \mathbf{1} \left\{ \frac{1}{N_0} \hat{\omega}_{t-1}^{(i)} |f(\xi_t^{(i)})| \geq C \right\} \bigg| \mathcal{F}_{t-1} \right] \xrightarrow{P} 0 \quad (5.2)
\]

Proof of Lemma 5.3. Use induction on the time index \( 0 \leq t \leq T \). For the \( t = 0 \) case, the Dominated Convergence Theorem shows

\[
E \left[ \frac{1}{N_0} \sum_{i=1}^{N_0} G_0 (\xi_0^{(i)}) |f(\xi_0^{(i)})| \mathbf{1} \left\{ \frac{1}{N_0} G_0 (\xi_0^{(i)}) |f(\xi_0^{(i)})| \geq C \right\} \right] = E \left[ G_0 |f| \mathbf{1} \left\{ \frac{1}{N_0} G_0 |f| \geq C \right\} \right] \rightarrow 0
\]

Next, assume (5.2) holds for all times \( 0 \leq s \leq t-1 \) and consider a time \( t \geq 1 \). By the induction assumption,

\[
E \left[ \frac{1}{N_0} \sum_{i=1}^{N_{t-1}} \hat{\omega}_{t-1}^{(i)} \mathbf{1} \left\{ \frac{1}{N_0} \hat{\omega}_{t-1}^{(i)} \geq C \right\} \bigg| \mathcal{F}_{t-1} \right] \xrightarrow{P} 0, \quad C > 0
\]

For \( \delta > 0 \), calculate

\[
P \left\{ \max_{1 \leq i \leq N_{t-1}} \frac{\hat{\omega}_{t-1}^{(i)}}{N_0} \geq \delta \right\} = E \left[ P \left\{ \max_{1 \leq i \leq N_{t-1}} \frac{\hat{\omega}_{t-1}^{(i)}}{N_0} \geq \delta \bigg| \mathcal{F}_{t-1} \right\} \right]
\]

\[
\leq E \left[ \min \left\{ 1, \frac{1}{\delta} E \left[ \frac{1}{N_0} \sum_{i=1}^{N_{t-1}} \hat{\omega}_{t-1}^{(i)} \mathbf{1} \left\{ \frac{1}{N_0} \hat{\omega}_{t-1}^{(i)} \geq \delta \right\} \bigg| \mathcal{F}_{t-1} \right] \right\} \right]
\]

Sending \( N_0 \) to infinity, it follows that \( \max_{1 \leq i \leq N_{t-1}} \frac{\hat{\omega}_{t-1}^{(i)}}{N_0} \xrightarrow{P} 0 \), and by Assumption 2.1 \( \max_{1 \leq j \leq N_t} \frac{\hat{\omega}_{t-1}^{(j)}}{N_0} \xrightarrow{P} 0 \).

Now for \( C > 0 \) and \( \delta > 0 \), define \( \nu = |f| \mathbf{1} \{ |f| \geq \frac{C}{\delta} \} \). Calculate

\[
P \left\{ E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} \hat{\omega}_{t-1}^{(i)} |f(\xi_t^{(i)})| \mathbf{1} \left\{ \frac{1}{N_0} \hat{\omega}_{t-1}^{(i)} |f(\xi_t^{(i)})| \geq C \right\} \bigg| \mathcal{F}_t \right] > \epsilon \right\}
\]

\[
\leq P \left\{ \max_{1 \leq j \leq N_t} \frac{\hat{\omega}_{t-1}^{(j)}}{N_0} > \delta \right\} + \frac{1}{\epsilon} E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} \hat{\omega}_{t-1}^{(i)} \nu(\hat{\xi}_{t-1}^{(i)}, \xi_t^{(i)}) \right]
\]

\[
= P \left\{ \max_{1 \leq j \leq N_t} \frac{\hat{\omega}_{t-1}^{(j)}}{N_0} > \delta \right\} + \frac{1}{\epsilon} E \left[ \prod_{s=0}^{t-1} G_s |f| \mathbf{1} \left\{ |f| \geq \frac{C}{\delta} \right\} \right]
\]

Sending \( N_0 \) to infinity and \( \delta \) to zero verifies (5.2) at time \( t \). \( \square \)
Proof of Theorem 2.2. Lemma 5.3 verifies the first condition of Lemma 5.2, namely,
\[ \sum_{t=0}^{T} \sum_{i=1}^{N_t} E \left[ \frac{1}{N_0} w_i^{(i)} \left| h_t \left( \xi_t^{(i)} \right) \right| \mathbb{1} \left\{ \left| \frac{1}{N_0} w_i^{(i)} \left| h_t \left( \xi_t^{(i)} \right) \right| > C \right\} \right| \mathcal{F}_{t}^{(i-1)} \right] P \rightarrow 0 \]

To verify the second condition, observe
\[ P \left\{ \sum_{t=0}^{T} \sum_{i=1}^{N_t} E \left[ \frac{1}{N_0} w_i^{(j)} \left| h_t \left( \xi_t^{(j)} \right) \right| \mathcal{F}_{t}^{(i-1)} \right] > \lambda \right\} \leq \frac{1}{\lambda} \sum_{t=0}^{T} E \left[ \frac{1}{N_0} \sum_{j=1}^{N_t} w_i^{(j)} \left| h_t \left( \xi_t^{(j)} \right) \right| \prod_{t=0}^{T-1} G_t \right] \]

The last quantity tends to zero as \( \lambda \rightarrow \infty \). Apply Lemma 5.2 to conclude. \( \square \)

Proof of Theorem 2.3. The proof uses a standard variance decomposition for martingales:
\[ \text{Var} \left[ M_T \right] = \sum_{t=0}^{T} \sum_{i=1}^{N_t} E \left[ \text{Var} \left[ M_t^{(i)} \right] \mathcal{F}_{t}^{(i-1)} \right] \leq E \left[ \sum_{t=0}^{T} \sum_{i=1}^{N_t} \left\{ \frac{1}{N_0} w_i^{(i)} h_t \left( \xi_t^{(i)} \right) \right\}^2 \right] \]

Since functions \( G_t \) are bounded, weights \( w_i^{(i)} \) are also bounded, with \( w_i^{(i)} \leq \prod_{s=0}^{t} C_s \sup G_s \). Thus, conclude
\[ \text{Var} \left[ M_T \right] \leq \sum_{t=0}^{T} \left( \prod_{s=0}^{t-1} C_s \sup G_s \right) E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} \left| G_t \left( \xi_{t-1}^{(i)}, \xi_t^{(i)} \right) h_t \left( \xi_t^{(i)} \right) \right|^2 \right] \]

The proof of Theorem 3.2 requires a series of lemmas.

Lemma 5.4. For each \( n \geq 1 \), suppose \( \mathcal{G}_{n,0} \subseteq \mathcal{G}_{n1} \subseteq \mathcal{G}_{n2} \subseteq \cdots \subseteq \mathcal{G}_{n,k_n} \) is a filtration and \( S_{n,k_n} = \sum_{j=1}^{k_n} Y_{nj} \) is the sum of martingale differences with \( E \left[ Y_{nj} \right| \mathcal{G}_{n,j-1} \right] = 0 \). Define \( V_{n,k_n}^2 = \sum_{j=1}^{k_n} \text{Var} \left[ Y_{nj} \right| \mathcal{G}_{n,j-1} \right] \).

(a) If \( V_{n,k_n}^2 \) \( \rightarrow 1 \) and if \( \sum_{j=1}^{k_n} E \left[ Y_{nj}^2 \mathbb{1} \left\{ \left| Y_{nj} \right| > C \right\} \right| \mathcal{G}_{n,j-1} \right] \) \( \rightarrow 0 \) for each \( C > 0 \), then \( S_{n,k_n} \) \( \mathcal{N} \rightarrow (0,1) \).

(b) If \( P \left\{ V_{n,k_n}^2 > 1 + \epsilon \right\} \rightarrow 0 \) for all \( \epsilon > 0 \), then \( \left| S_{n,k_n} \right| \leq 1 \).
Proof. Part (a) is the martingale CLT [14, pg.58-59]. To prove part (b), first set \( E_{m,n} = \left\{ \sqrt{2_{n,k_n}} \leq 1 + \frac{1}{m} \right\} \) for \( m, n \geq 1 \) and \( D_{1,n} = E_{1,n} \) for \( n \geq 1 \). By the assumption in part (b), there exists a number \( N(m) \geq 1 \) such that \( P(E_{m,n}) \geq 1 - \frac{1}{m} \) for \( n \geq N(m) \). Accordingly, for \( m \geq 1 \) define

\[
\begin{align*}
&\begin{cases}
D_{m,n} = D_{m-1,n}, & n < N(m) \\
D_{m,n} = E_{m,n}, & n \geq N(m)
\end{cases}
\end{align*}
\]

By this construction, for any \( m \geq M \) and \( n \geq N(M) \), \( E \left[ D_{m,n} S_{n,k_n}^2 \right] \leq 1 + \frac{1}{M} \) and \( P(D_{m,n}) \geq 1 - \frac{1}{M} \). Setting \( B_n = D_{n,n} \) gives \( \limsup_{n \to \infty} E \left[ \mathbb{I}_{B_n} S_{n,k_n}^2 \right] \leq 1 \) and \( P(B_n) \to 1 \). \( \square \)

**Lemma 5.5.** Assume \( E \left[ \prod_{s=0}^t G_s \right] < \infty \) for \( 0 \leq t \leq T - 1 \) and assume \( E \left[ \prod_{s=0}^t G_s | G_t h_t | ^2 \right] < \infty \). Then for each \( 0 \leq t \leq T \) and \( C > 0 \),

\[
E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} w_i^{(i)} h_t \left( \xi_t^{(i)} \right) \right] ^2 \mathbb{I} \left\{ \frac{1}{\sqrt{N_0}} w_i^{(i)} h_t \left( \xi_t^{(i)} \right) \geq C \right\} | \mathcal{F}_{t-1} \} \mathbb{P} \to 0
\]

**Proof.** For the \( t = 0 \) case, use the Dominated Convergence Theorem. For \( 1 \leq t \leq T \) and \( C > 0 \), define \( \nu = |G_t h_t| \mathbb{1} \left\{ |G_t h_t| \geq \frac{C \sqrt{N_0}}{2E[\prod_{s=0}^t G_s]} \right\} \). Calculate

\[
P \left\{ E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} w_i^{(i)} h_t \left( \xi_t^{(i)} \right) \right] ^2 \mathbb{I} \left\{ \frac{1}{\sqrt{N_0}} w_i^{(i)} h_t \left( \xi_t^{(i)} \right) \geq C \right\} | \mathcal{F}_{t-1} \} > \epsilon \right\}
\]

\[
\leq P \left\{ \frac{w_{t-1}}{N_0} > 2 E \left[ \prod_{s=0}^{t-1} G_s \right] \right\} + \frac{2}{\epsilon} E \left[ \prod_{s=0}^{t-1} G_s \right] E \left[ \frac{w_{t-1}}{N_0} \sum_{i=1}^{N_t} \nu \left( \xi_t^{(i)} \right) \right]
\]

\[
= P \left\{ \frac{w_{t-1}}{N_0} > 2 E \left[ \prod_{s=0}^{t-1} G_s \right] \right\} + \frac{2}{\epsilon} E \left[ \prod_{s=0}^{t-1} G_s \right] E \left[ \sum_{s=0}^{t-1} G_s \nu \right]
\]

Since \( \frac{w_{t-1}}{N_0} \mathbb{P} \to E \left[ \prod_{s=0}^{t-1} G_s \right] \) by Theorem 2.2, both terms vanish upon sending \( N_0 \) to infinity. \( \square \)

**Lemma 5.6.** Assume \( E \left[ \prod_{s=0}^t G_s \right] < \infty \) for \( 0 \leq t \leq T - 1 \), \( E |G_0 h_0|^2 < \infty \), and \( E \left[ \prod_{s=0}^t G_s | G_{t+1} h_{t+1} | ^2 \right] \) for \( 0 \leq t \leq T - 1 \). Define

\[
\eta^2 = \text{Var}[G_0 h_0] + \sum_{t=0}^{T-1} \eta_t^2 [h_t] + \sum_{t=0}^{T-1} E \left[ \prod_{s=0}^t G_s \right] E \left[ \prod_{s=0}^t G_s \text{Var}[G_{t+1} h_{t+1}] X_t \right]
\]

where \( \eta^2 \) depends on numbers \( (\eta_t)^2 [h_t] \) for \( 0 \leq t \leq T - 1 \).
(a) If \( N_0 \tilde{V}_T^2 [h_t] \overset{p}{\to} \tilde{\eta}_T^2 [h_t] \) for each \( 0 \leq t \leq T - 1 \), then
\[
\sqrt{N_0} \left( \frac{m_{T-1}^N}{N_0} \sum_{i=1}^{N_T} f \left( \Xi_{T-1}^{(i)}, \tilde{\xi}_T^{(i)} \right) - E \left[ \prod_{t=0}^{T-1} G_t f \right] \right) \overset{p}{\to} N (0, q^2)
\]

(b) If \( \lim_{N_0 \to \infty} P \left\{ N_0 \tilde{V}_T^2 [h_t] \geq \tilde{\eta}_T^2 [h_t] + \epsilon \right\} = 0 \) for each \( 0 \leq t \leq T - 1 \) and \( \epsilon > 0 \),
\[
\sum_{i=1}^{N_T} f \left( \Xi_{T-1}^{(i)}, \tilde{\xi}_T^{(i)} \right) - E \left[ \prod_{t=0}^{T-1} G_t f \right] \geq \frac{\eta}{\sqrt{N_0}}
\]

Proof. The proof uses Lemma 5.4 to analyze asymptotic behavior of the martingale \( \frac{\sqrt{N_0}}{q} M_t^{(i)} \), where \( M_t^{(i)} \) is defined in equation (5.1). First compute the sum of conditional variances

\[
\sum_{i=1}^{N_T} \text{Var} \left[ M_t^{(i)} \mid \mathcal{F}_{t-1} \right] = \frac{\text{Var} \left[ G_0 h_0 \right]}{N_0} + \sum_{i=1}^{T} \text{Var} \left[ M_t \mid \mathcal{F}_{t-1} \right]
\]

\[
= \frac{\text{Var} \left[ G_0 h_0 \right]}{N_0} + \sum_{i=1}^{T} \text{Var} \left[ E \left[ M_t \mid \mathcal{F}_{t-1} \right] \right] + \text{Var} \left[ M_t \mid \mathcal{F}_{t-1} \right]
\]

\[
= \frac{\text{Var} \left[ G_0 h_0 \right]}{N_0} + \sum_{i=1}^{T} \text{Var} \left[ \frac{m_{t-1}^N}{N_0} \sum_{i=1}^{N_t} h_t \left( \xi_t^{(i)} \right) \right] + \text{Var} \left[ G_t h_t \mid X_{t-1} = \xi_t^{(i)} \right]
\]

\[
= \frac{\text{Var} \left[ G_0 h_0 \right]}{N_0} + \sum_{i=1}^{T} \tilde{V}_t^2 [h_t] + \sum_{t=1}^{T} \frac{m_{t-1}^N}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} \text{Var} \left[ G_t h_t \mid X_{t-1} = \xi_t^{(i)} \right]
\]

Theorem 2.2 shows that \( \frac{m_t}{p} \overset{p}{\to} E \left[ \prod_{s=0}^{t-1} G_s \right] \) and

\[
\frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} \text{Var} \left[ G_t h_t \mid X_{t-1} = \xi_t^{(i)} \right] \overset{p}{\to} E \left[ \prod_{s=0}^{t-1} G_s \text{Var} \left[ G_t h_t \mid X_{t-1} = \xi_t^{(i)} \right] \right]
\]

Next, for \( C > 0 \) and \( 0 \leq t \leq T \), a useful inequality of Dvoretzky [9] gives

\[
N_0 \sum_{i=1}^{N_t} E \left[ \left( M_t^{(i)} - M_t^{(i-1)} \right)^2 1 \left\{ \sqrt{N_0} \left| M_t^{(i)} - M_t^{(i-1)} \right| \geq C \right\} \right] \leq 4 E \left[ \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} h_t \left( \xi_t^{(i)} \right)^2 1 \left\{ \left| \frac{1}{\sqrt{N_0}} w_t^{(i)} h_t \left( \xi_t^{(i)} \right) \right| \geq \frac{C}{2} \right\} \right] \]

This last term vanishes upon sending \( N_0 \) to infinity by Lemma 5.5. Thus, the conditions of Lemma 5.4 are satisfied, and parts (a) and (b) follow. \( \square \)
Lemma 5.7. Assume \( \mathbb{E}\left[ \prod_{s=0}^{t} G_s \right] < \infty \) for \( 0 \leq t \leq T - 1 \) and assume \( \mathbb{E}\left[ \prod_{t=0}^{T} G_t \right] < \infty \). Set \( \tilde{G}_T = \mathbb{E}\left[ \prod_{t=0}^{T-1} G_t \right] G_T / \mathbb{E}\left[ \prod_{t=0}^{T} G_t \right] \). Then

\[
\frac{w_{T-1}}{N_0} \sum_{i=1}^{N_T} \left\{ \frac{w_T^{(i)}}{w_T} \right\} \left( 1 - \left\{ \frac{w_T^{(i)}}{w_T} \right\} \right) f \left( \xi_t^{(i)} \right) \rightarrow P \mathbb{E}\left[ \prod_{t=0}^{T-1} G_t \left\{ \tilde{G}_T \right\} \left( 1 - \{ \tilde{G}_T \} \right) f \right] \tag{5.3}
\]

If additionally \( \mathbb{E}\left[ \prod_{t=0}^{T-1} G_t \mathbb{I}\{ \tilde{G}_T \in \{1, 2, \ldots\} \} \right] = 0 \), then

\[
\frac{w_{T-1}}{N_0} \sum_{i=1}^{N_T} \left\{ \frac{w_T^{(i)}}{w_T} \right\} f \left( \xi_t^{(i)} \right) \rightarrow P \mathbb{E}\left[ \prod_{t=0}^{T-1} G_t \left\{ \tilde{G}_T \right\} \nu \right] \tag{5.4}
\]

Proof. For a proof of equation (5.4) and a special case of equation (5.3), see Douc & Moulines [7]. To prove the more general case of (5.3), first define \( L(x) = \{x\} - \{x\}^2 \). By Theorem 2.2,

\[
\frac{w_{T-1}}{N_0} \sum_{i=1}^{N_T} L \left( \tilde{G}_T \left( \xi_t^{(i)}, \xi_t^{(i)} \right) \right) f \left( \xi_t^{(i)} \right) \rightarrow P \mathbb{E}\left[ \prod_{t=0}^{T-1} G_t \left\{ \tilde{G}_T \right\} \left( 1 - \{ \tilde{G}_T \} \right) f \right]
\]

Thus, it suffices to show

\[
\frac{w_{T-1}}{N_0} \sum_{i=1}^{N_T} \left| L \left( \frac{w_T^{(i)}}{w_T} \right) - L \left( \tilde{G}_T \left( \xi_t^{(i)}, \xi_t^{(i)} \right) \right) \right| f \left( \xi_t^{(i)} \right) \rightarrow 0
\]

Since \( x \mapsto L(x) \) has Lipschitz constant 1, for \( \epsilon > 0 \) and \( \delta > 0 \) it follows that

\[
P \left\{ \frac{\hat{w}_{T-1}}{N_0} \sum_{i=1}^{N_t} L \left( \frac{w_T^{(j)}}{w_T} \right) - L \left( \tilde{G}_T \left( \xi_t^{(i)}, \xi_t^{(i)} \right) \right) \left| f \left( \xi_t^{(i)} \right) \right| > \epsilon \right\}
\]

\[
\leq P \left\{ \left| \frac{\hat{w}_{T-1}}{w_T} \mathbb{E}\left[ \prod_{t=0}^{T} G_t \right] \right| - 1 > \delta \right\}
\]

\[
+ \frac{\epsilon}{\epsilon} \mathbb{E}\left[ \frac{\hat{w}_{T-1}}{N_0} \sum_{i=1}^{N_T} \tilde{G}_T \left( \xi_t^{(i)}, \xi_t^{(i)} \right) | f \left( \xi_t^{(i)} \right) | \right]
\]

\[
= P \left\{ \left| \frac{\hat{w}_{T-1}}{w_T} \mathbb{E}\left[ \prod_{t=0}^{T} G_t \right] \right| - 1 > \delta \right\} + \frac{\epsilon}{\epsilon} \mathbb{E}\left[ \prod_{s=0}^{T-1} G_s \tilde{G}_T f \right]
\]

Both terms vanish upon sending \( N_0 \) to infinity and then \( \delta \) to 0. \( \square \)
Lemma 5.8. Assume \(E \left[ \prod_{t=0}^{t} G_s \right] < \infty \) for \(0 \leq t \leq T - 1\) and assume \(E \left[ \prod_{t=0}^{T} G_t f^2 \right] < \infty\). At resampling step \(T\), assume particles are sorted by a coordinate \(\theta_T\) and then stratified or stratified residual resampling is used. Then for any \(\delta : \mathbb{R} \rightarrow \mathbb{R}\) with \(E \left[ \prod_{t=0}^{T} G_t | p (\theta_T)^2 \right] < \infty\),

\[
\limsup_{N_0 \rightarrow \infty} P \left\{ N_0 \hat{V}_T^2 [f] > (1 + \epsilon) N_0 \hat{V}_T^2 [f - p (\theta_T)] + \epsilon \right\} < \epsilon, \quad \epsilon > 0
\]

Proof. Fix \(\delta > 0\) and select \(\eta \in C_c (\mathbb{R})\), which approximates \(p\) so that \(E \left[ \prod_{t=0}^{T} G_t | \eta (\theta_T) - p (\theta_T)|^2 \right] < \delta\). Applying Cauchy’s inequality with \(\epsilon\),

\[
\hat{V}_T^2 [f] \leq \left( 2 + \frac{2}{\epsilon} \right) \left( \hat{V}_T^2 [\eta (\theta_T)] + \hat{V}_T^2 [\eta (\theta_T) - p (\theta_T)] \right) + (1 + \epsilon) \hat{V}_T^2 [f - p (\theta_T)]
\]

To prove the result it suffices to bound \(\hat{V}_T^2 [\eta (\theta_T)]\) and \(\hat{V}_T^2 [\eta (\theta_T) - p (\theta_T)]\). First bound \(\hat{V}_T^2 [\eta (\theta_T)]\). On the event that \(\hat{\omega}_T \leq 2 E \left[ \prod_{t=0}^{T} G_t \right]\), it follows

\[
\hat{V}_T^2 [\eta (\theta_T)] \leq \left( 2 E \left[ \prod_{t=0}^{T} G_t \right] \right)^2 \text{Var} \left[ \frac{1}{N_0} \sum_{j=1}^{N_{T+1}} \eta (\hat{\theta}_T^{(j)}) \right] | \mathcal{F}_T |
\]

where \(\hat{\theta}_T^{(j)}\) denotes \(\theta_T \left( \hat{\xi}_T^{(j)} \right)\). In the resampling step, a series of particles \(\left( \hat{\xi}_T^{(j)} \right)_{1 \leq j \leq J}\) is randomly selected (other particles may be deterministically selected) with \(L^{(0)} \geq \hat{\theta}_T^{(1)} \geq L^{(1)} \geq \hat{\theta}_T^{(2)} \geq \cdots \geq \hat{\theta}_T^{(J)} \geq L^{(J)}\) for some \(\mathcal{F}_T\)-measurable random variables \(L^{(j)}\). Therefore,

\[
\sum_{j=1}^{J} \text{Var} \left[ \eta (\hat{\theta}_T^{(j)}) \right] | \mathcal{F}_T | \leq \frac{1}{4} \sum_{j=1}^{J} \sup_{x \in L^{(j-1)}, L^{(j)}} \eta (x) - \inf_{x \in L^{(j-1)}, L^{(j)}} \eta (x) \leq \sup_{x \in \mathbb{R}} | \eta (x) | V (\eta)
\]

where \(V (\eta)\) is the total variation of \(\eta\).

It remains to bound \(\hat{V}_T^2 [\eta (\theta_T) - p (\theta_T)]\). On the event \(\hat{\omega}_T \leq 2 E \left[ \prod_{t=0}^{T} G_t \right],\)

\[
\hat{V}_T^2 [\eta (\theta_T) - p (\theta_T)] = \sum_{j=1}^{N_{T+1}} \frac{\hat{\omega}_T}{N_0} \text{Var} \left[ \left( \eta (\hat{\theta}_T^{(j)}) - p (\hat{\theta}_T^{(j)}) \right) \right] | \mathcal{F}_T |
\]

\[
\leq 2 E \left[ \prod_{t=0}^{T} G_t \right] E \left[ \frac{\hat{\omega}_T}{N_0} \sum_{j=1}^{N_{T+1}} \left( \eta (\hat{\theta}_T^{(j)}) - p (\hat{\theta}_T^{(j)}) \right)^2 | \mathcal{F}_T \right]
\]
This last term has expectation

\[
2E \left[ \prod_{t=0}^{T} G_t \right] E \left[ \prod_{t=0}^{T} G_t \left| \eta_0 \right| \left( \theta_T - p(\theta_T) \right)^2 \right] < 2\delta E \left[ \prod_{t=0}^{T} G_t \right]
\]

Conclude

\[
\limsup_{N_0 \to \infty} P \left\{ N_0 \hat{V}_t^2 [h_t] > (1 + \epsilon) N_0 \hat{V}_t^2 [h_t - \nu] + \epsilon \right\}
\]

\[
= \limsup_{N_0 \to \infty} P \left\{ \hat{w}_t \leq 2E \left[ \prod_{s=0}^{t} G_s \right], \left( 2 + \frac{2}{\epsilon} \right) N_0 \hat{V}_t^2 [\eta(\theta_T)] \leq \frac{\epsilon}{2} \right\}
\]

\[
+ \limsup_{N_0 \to \infty} P \left\{ \hat{w}_t \leq 2E \left[ \prod_{s=0}^{t} G_s \right], \left( 2 + \frac{2}{\epsilon} \right) N_0 \hat{V}_t^2 [\eta(\theta_T) - p(\theta_T)] > \frac{\epsilon}{2} \right\}
\]

\[
\leq \left( \frac{2}{\epsilon} \right) \left( 2 + \frac{2}{\epsilon} \right) \left( 2\delta E \left[ \prod_{t=0}^{T} G_t \right] \right)
\]

For small enough \( \delta \), this last term is less than \( \epsilon \), proving the result. \( \square \)

Proof of Theorem 3.2. The proof combines Lemma 5.6 with explicit computations of resampling variances \( \hat{V}_t^2 \). For multinomial resampling,

\[
N_0 \hat{V}_t^2 [h_t] = \frac{\hat{w}_t}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} \left| h_t \left( \xi_t^{(i)} \right) \right|^2 - \left| \frac{1}{N_0} \sum_{i=1}^{N_t} w_t^{(i)} h_t \left( \xi_t^{(i)} \right) \right|^2
\]

By Theorem 2.2, therefore,

\[
N_0 \hat{V}_t^2 [h_t] \xrightarrow{P} E \left[ \prod_{s=0}^{t} G_s \right] E \left[ \prod_{s=0}^{t} G_s h_t \right] - \left( E \left[ \prod_{s=0}^{t} G_s h_t \right] \right)^2
\]

\[
= E \left[ \prod_{s=0}^{t} G_s \right] E \left[ \prod_{s=0}^{t} G_s \right] h_t - E \left[ \prod_{s=0}^{t} G_s h_t \right] \frac{E \left[ \prod_{s=0}^{t} G_s h_t \right]}{E \left[ \prod_{s=0}^{t} G_s \right]}
\]

For multinomial residual resampling, \( N_0 \hat{V}_t^2 [h_t] \) takes the form

\[
\frac{\hat{w}_t}{\hat{w}_{t-1}} \left( \frac{\hat{w}_{t-1}}{N_0} \sum_{i=1}^{N_t} \left\{ \frac{w_t^{(i)}}{\hat{w}_t} \right\} \right) \left| h_t \left( \xi_t^{(i)} \right) \right|^2 - \frac{\hat{w}_{t-1}}{N_0} \sum_{i=1}^{N_t} \left\{ \frac{w_t^{(i)}}{\hat{w}_t} \right\} h_t \left( \xi_t^{(i)} \right)^2
\]

By Theorem 2.2 and Lemma 5.7, \( N_0 \hat{V}_t^2 [h_t] \) converges in probability to

\[
\left( \frac{E \left[ \prod_{s=0}^{t} G_s \right]}{E \left[ \prod_{s=0}^{t-1} G_s \right]} \right)^2 E \left[ \prod_{s=0}^{t-1} G_s \left\{ \tilde{G}_t \right\} \right] h_t - \left( E \left[ \prod_{s=0}^{t-1} G_s \left\{ \tilde{G}_t \right\} \right] \right)^2
\]
For Bernoulli resampling, Theorem 2.2 and Lemma 5.7 give

\begin{align*}
N_0 \hat{V}_t^2 [h_t] &= \frac{1}{N_0} \sum_{i=1}^{N_r} \left\{ \frac{w_t^{(i)}}{w_t} \right\} \left( 1 - \left\{ \frac{w_t^{(i)}}{w_t} \right\} \right) \left| h_t \left( s_t^{(i)} \right) \right|^2 \\
&\rightarrow \mathbb{E} \left[ \prod_{s=0}^{t-1} G_s \right] \left( \prod_{s=0}^{t-1} \hat{G}_s \right) \left( 1 - \left\{ \hat{G}_t \right\} \right) h_t^2
\end{align*}

To compute the resampling variance for stratified resampling, consider the function \( p \) that minimizes \( \mathbb{E} \left[ \prod_{s=0}^{t} \hat{G}_s | h_t - p(\theta_t) \right]^2 \). Since this function can be written as an \( L^2 \) projection, it is well-defined. Moreover, by Lemma 3.1, the resampling variance \( N_0 \hat{V}_t^2 [h_t - p(\theta_t)] \) is bounded by the multinomial resampling variance, which converges in probability to

\[ \hat{\eta}^2_t [h_t] = \left( \mathbb{E} \left[ \prod_{s=0}^{t} G_s \right] \right)^2 \mathbb{E} \left[ \prod_{s=0}^{t} \hat{G}_s | h_t - p(\theta_t) \right]^2 \]

Thus, \( \mathbb{P} \left\{ N_0 \hat{V}_t^2 [h_t - p(\theta_t)] > \hat{\eta}^2_t [h_t] + \epsilon \right\} \rightarrow 0 \) for all \( \epsilon > 0 \). By Lemma 5.8, this is enough to guarantee \( \mathbb{P} \left\{ N_0 \hat{V}_t^2 [h_t] > \hat{\eta}^2_t [h_t] + \epsilon \right\} \rightarrow 0 \) for all \( \epsilon > 0 \). The asymptotic variance upper bound for sorted stratified residual resampling is proved similarly.

Acknowledgements

The author would like to thank Jonathan Weare and Omiros Papaspiliopoulos for conversations that helped shape the presentation of results and Alicia Zhao for gracious and patient editorial assistance.

References

[1] Bolic, M., Djuric, P. M., & Hong, S. (2003). New resampling algorithms for particle filters. In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on, volume 2, (pp. II–589). IEEE.

[2] Brindle, A. (1980). Genetic algorithms for function optimization. PhD thesis, University of Alberta.

[3] Chan, H. P., Lai, T. L., et al. (2013). A general theory of particle filters in hidden Markov models and some applications. The Annals of Statistics, 41(6), 2877–2904.

[4] Chopin, N. (2004). Central Limit Theorem for Sequential Monte Carlo methods and its application to Bayesian inference. Annals of statistics, 32(6), 2385–2411.
[5] Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer Science & Business Media.

[6] Douc, R., Cappé, O., & Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, (pp. 64–69). IEEE.

[7] Douc, R. & Moulines, E. (2008). Limit theorems for weighted samples with applications to Sequential Monte Carlo methods. *Annals of Statistics*, 36(5), 2344–2376.

[8] Doucet, A., de Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential Monte Carlo methods in practice*. Springer.

[9] Douc, R. & Moulines, E. (2008). Limit theorems for weighted samples with applications to Sequential Monte Carlo methods. *Annals of Statistics*, 36(5), 2344–2376.

[10] Gerber, M., Chopin, N., & Whiteley, N. (2017). Negative association, ordering and convergence of resampling methods. arXiv preprint arXiv:1707.01845.

[11] Grassberger, P. (1997). Pruned-Enriched Rosenbluth Method: Simulations of $\theta$ polymers of chain length up to 1 000 000. *Physical Review E*, 56(3), 3682.

[12] Grimm, R. & Storer, R. (1971). Monte-Carlo solution of schrödinger’s equation. *Journal of Computational Physics*, 7(1), 134–156.

[13] Hairer, M. & Weare, J. (2014). Improved Diffusion Monte Carlo. *Communications on Pure and Applied Mathematics*, 67(12), 1995–2021.

[14] Hall, P. & Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press.

[15] Hammersley, J. M. & Morton, K. W. (1954). Poor man’s Monte Carlo. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23–38.

[16] Hol, J. D., Schön, T. B., & Gustafsson, F. (2006). On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, (pp. 79–82). IEEE.

[17] Holland, J. H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press.

[18] Hu, X.-L., Schön, T., & Ljung, L. (2008). A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing*, 56(4), 1337–1348.

[19] Huber, G. A. & Kim, S. (1996). Weighted-Ensemble Brownian dynamics simulations for protein association reactions. *Biophysical journal*, 70(1), 97.

[20] Kalos, M. (1962). Monte Carlo calculations of the ground state of three-and four-body nuclei. *Physical Review*, 128(4), 1791.

[21] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1), 1–25.

[22] Li, T., Bolic, M., & Djuric, P. M. (2015). Resampling methods for particle filtering: classification, implementation, and strategies. *IEEE Signal
Processing Magazine, 32(3), 70–86.

[23] Liu, J. S. (2008). Monte Carlo strategies in scientific computing. Springer Science & Business Media.

[24] Liu, J. S., Chen, R., & Wong, W. H. (1998). Rejection control and sequential importance sampling. Journal of the American Statistical Association, 93(443), 1022–1031.

[25] Madow, W. G. & Madow, L. H. (1944). On the theory of systematic sampling, I. The Annals of Mathematical Statistics, 15(1), 1–24.

[26] McLeish, D. L. (1974). Dependent Central Limit Theorems and invariance principles. The Annals of Probability, 2(4), 620–628.

[27] Rosenbluth, M. N. & Rosenbluth, A. W. (1955). Monte Carlo calculation of the average extension of molecular chains. The Journal of Chemical Physics, 23(2), 356–359.

[28] Whiteley, N., Lee, A., Heine, K., et al. (2016). On the role of interaction in Sequential Monte Carlo algorithms. Bernoulli, 22(1), 494–529.