Research statement: Robert J. Webber

My research is dedicated to analyzing and improving Monte Carlo methods. Monte Carlo methods are a collection of computer algorithms that randomly sample from complex models in order to estimate statistics of interest. Monte Carlo methods become crucial when scientific models include thousands or millions of degrees of freedom, so there is no feasible way to assess the complete range of outcomes. While Monte Carlo has proven to be a useful tool in high-dimensional inference, it can also lead to large error bars that limit its precision. My research addresses the question of how to minimize these error bars. It is, at its core, a pursuit of optimization and refinement for a powerful and versatile mathematical tool.

My research is focused on the Monte Carlo methods used in computational chemistry and geophysics. In these disciplines, there is typically a high computational cost to running models forward. For example, molecular dynamics models \[1\] typically use \(10^4 - 10^6\) variables and weather models \[e.g., 2\] use \(10^6 - 10^9\) variables to represent complex systems of interest. Due to the high cost of evolving these models over many small timesteps, it is necessary to develop Monte Carlo approaches that perform well with a small sample size and behave robustly despite possible parameter misspecification. Below, I describe three projects where I established or optimized Monte Carlo methods for this computationally challenging context. Additionally, I discuss future research goals that were inspired by these three projects.

I. A NEW SPLITTING AND KILLING SCHEME FOR TROPICAL CYCLONES

Background. Splitting and killing is a popular approach to rare event sampling with a history dating back to Ulam and Von Neumann in the late 1940s \[3, 4\]. In this approach, which is illustrated in Figure 1, a stochastic model is run forward from a time \(s\) to a later time \(t > s\). Next, the samples \(\xi^i_t\) that have moved closer to the rare event of interest are randomly split into conditionally independent replicates, and the samples \(\xi^i_t\) that have moved farther from the rare event are randomly terminated (“killed”), while preserving the total number of samples \(N\).

![Figure 1: Splitting and killing is used to sample rare, high positions of the x coordinate. White circles indicate that samples are killed while black circles indicate that samples are preserved and possibly split.](image)

For any Markov process \(X_t\), splitting and killing produces unbiased estimates of expectations \(E[f(X_t)]\) using the estimator \(\sum_{i=1}^{N} w^i_t f(\xi^i_t)\). In this estimator, each sample \(\xi^i_t\) is weighted by a positive factor \(w^i_t > 0\), which must be adjusted after each splitting event. In addition to being unbiased, the splitting and killing estimator can achieve high statistical precision. For example, to estimate the probability \(p\) of a rare event with fixed relative accuracy, direct sampling requires \(O(1/p)\) samples. In contrast, splitting and killing schemes can require as few as \(O(\log(1/p))\) samples \[5\], an exponential improvement compared to direct sampling.

Objectives. The main objective of my tropical cyclone research \[6\] was to extend splitting and killing for the first time to study statistics of intense tropical cyclones, particularly the probability of intense Category 5 storms occurring. Tropical cyclones, along with earthquakes, are the costliest natural disasters in many North American and East Asian countries \[7\]. Understanding tropical cyclone statistics is more critical than ever as the frequency and distribution of tropical cyclones potentially shifts due to climate change.

A second objective was to increase the robustness of splitting and killing. Splitting and killing schemes are typically guided by a one-dimensional reaction coordinate \(\theta \in \mathbb{R}\) that indicates the
relative importance of different samples. However, small changes in the reaction coordinate can lead to large changes in statistical efficiency. Increasing robustness is especially important when simulating tropical cyclones, because the cost of running simulations limits the sample size ($10^2$ samples) and limits the potential to tune parameters.

**Results.** In collaboration with my Ph.D. advisor Jonathan Weare and tropical cyclone experts, I developed a new splitting and killing scheme called “quantile diffusion Monte Carlo” (QDMC) and applied it to study statistics of intense tropical cyclones. The major methodological contribution of our work was to increase the robustness of splitting and killing. Previous splitting and killing schemes typically transformed the distribution of samples, relative to direct sampling, according to the likelihood ratio

$$\ell_t(\xi^i_t) \propto \exp\{c\theta(\xi^i_t)\},$$

(1)

where $c > 0$ is an intensity parameter. This common approach can be successful in reducing variance, but the likelihood ratio (1) can also be sensitive to small changes in $c$ and $\theta$. To increase robustness, therefore, QDMC uses a new approach to specify splitting and killing rules. In this approach, the user directly specifies a quantile $q$ of the reaction coordinate $\theta$ to emphasize during simulations. Then, QDMC transforms the samples by the likelihood ratio

$$\ell_t(\xi^i_t) \propto \exp\left\{F^{-1}_N(q) F^{-1}_\theta \circ \theta(\xi^i_t)\right\}.$$

(2)

Here, $F_N$ is the cumulative distribution function for a standard Gaussian, and $F_\theta$ is the cumulative distribution function for the reaction coordinate $\theta$, which is estimated using the samples. The likelihood ratio (2) has the advantage that it is invariant under monotonic transformations of $\theta$, an invariance property that is key to the algorithm’s effectiveness.

In practice, we found that QDMC is more robust than previous splitting and killing schemes and requires minimal tuning in tropical cyclone (TC) simulations. In our implementation, before splitting and killing takes place, the median TC simulation samples the 50th percentile of TC intensity. After the first splitting time, we specified that the median TC simulation sample the 64th percentile of TC intensity. After three additional splitting times, the median TC simulation samples the 84th percentile of TC intensity. Compared to direct sampling, QDMC reduced the variance of TC statistics by factors of at least $2^{10}$. As seen in Figure 2, QDMC produced three intense Category 5 simulations of Hurricane Earl compared to zero Category 5 simulations using direct sampling. Moreover, QDMC simulations were consistent with the “best track” historical record, suggesting that simulated storms followed physically realistic trajectories.

![Earl pressures, direct sampling](image)

**Figure 2:** QDMC (right) samples more intense storms compared to direct sampling (left).

**Future research.** My goal for future research is to address the issue of uncertainty quantification for splitting and killing schemes. In response to the known difficulty of calculating error
bars \[8\]. I aim to develop more accurate variance estimators for splitting and killing estimates. The starting point for my investigations is the second moment estimator

\[ E \left[ \left| \sum_{i=1}^{N} w_i f(\xi_i^t) \right|^2 \right] \approx \sum_{i=1}^{N} \sum_{\text{anc}(\xi_i^t) = i} \left| \sum_{j=1}^{\text{anc}(\xi_i^t)} \xi_j \right|^2, \tag{3} \]

where \( \text{anc}(\xi_i^t) \) denotes the ancestral index of the sample \( \xi_i^t \) traced back across splitting events. The second moment estimator (3) is unbiased in the special case where Bernoulli resampling is used for splitting and killing. However, in more typical splitting and killing schemes, the estimator (3) overestimates the true splitting and killing error, especially when the number of splitting times becomes large. Using numerical experiments and mean-field asymptotic analysis, I aim to identify when and why this estimator overstates the true splitting and killing error. Then, building from this analysis, I intend to identify more accurate estimators of splitting and killing error that can be used to improve uncertainty quantification in the future.

II. Optimized resampling schemes

Background. Resampling schemes are a key component of splitting and killing algorithms, as well as Monte Carlo algorithms for data assimilation and Bayesian inference. In all of these Monte Carlo algorithms, there is an ensemble of samples \( \{\xi_i\}_{1 \leq i \leq N} \), and each sample \( \xi_i \) is assigned an importance weight \( I_i \geq 0 \) with \( \sum_{i=1}^{N} I_i = N \). Then, each sample \( \xi_i \) is replaced by \( N_i \) identical copies, where \( N_i \) is a random number with expected value \( E[N_i] = I_i \). A resampling scheme is any algorithm to determine the random numbers \( \{N_i\}_{1 \leq i \leq N} \).

It is well-known that a poor choice of resampling scheme can lead to excess error in Monte Carlo estimates. In response to this problem, a great number of resampling schemes have been proposed in the Monte Carlo literature \[9\]. However, the relative advantages of different schemes are incompletely understood.

Objective. The main objective of my research project \[10\] was to identify resampling schemes that minimize the resampling variance \( \text{Var} \left[ \sum_{i=1}^{N} N_i f(\xi_i^t) \right] \) where \( f \) is a quantity of interest, which may be known or unknown at the time of resampling.

Results. My research used two main strategies to minimize the resampling variance. As a first strategy, I directly minimized \( \text{Var} \left[ \sum_{i=1}^{N} N_i f(\xi_i^t) \right] \) when the function \( f \) is known:

Theorem. Among a large class of resampling schemes (the “matrix resampling schemes”, see \[10\]), the following scheme minimizes \( \text{Var} \left[ \sum_{i=1}^{N} N_i f(\xi_i^t) \right] \):

1. Relabel the samples \( \{\xi_i\}_{1 \leq i \leq N} \) so that \( f(\xi_1^t) \leq f(\xi_2^t) \leq \cdots \leq f(\xi_N^t) \).

2. Draw independent random variables \( U^j \sim \text{Unif}(0,1) \) for \( 1 \leq j \leq N \). Define \( N_i^j \) to be the number of indices \( 1 \leq j \leq N \) for which

\[ j - U^j \leq \sum_{k=1}^{i-1} I_k^j, \sum_{k=1}^{i} I_k^j \tag{4} \]

The optimal resampling scheme introduced in the theorem sorts the samples to remove as much randomness as possible. Indeed, the advantage of sorting samples to create a beneficial stratification effect has been known for many years in the survey sampling community \[11\], yet it was not fully appreciated in the Monte Carlo community before this research project.

As a second analytic strategy, I derived asymptotic formulas for \( \text{Var} \left[ \sum_{i=1}^{N} N_i f(\xi_i^t) \right] \) when all samples are drawn from a distribution \( \xi_i^t \sim \mu \) and the number of samples \( N \) increases to
infinity. This strategy is particularly suitable for analyzing resampling variance when functions \( f \) are not known in advance of resampling. The strategy revealed that advantageous resampling schemes satisfy two properties:

(a) Each number \( N^i \) takes the value \( \lfloor I^i \rfloor + 1 \) with probability \( I^i - \lfloor I^i \rfloor \) and otherwise takes the value \( \lfloor I^i \rfloor \), where \( \lfloor \cdot \rfloor \) denotes the floor function \( \lfloor x \rfloor = \max \{ z \in \mathbb{Z} : z \leq x \} \).

(b) The sum of the numbers \( N^i \) satisfies \( \sum_{i=1}^{N} N^i = N \) with probability one.

Schemes that satisfy [a] and [b] typically have lower asymptotic variance than schemes that satisfy just property [a] or just property [b].

Future research. In future research, I plan to investigate the open question of whether resampling schemes converge to a continuous-time limit as resampling is applied increasingly frequently. I conjecture that such a limit is possible when appropriate resampling schemes are used. In the continuous-time limit, at random times determined by an exponential clock, one sample is killed and another sample is split in two. I plan to rigorously demonstrate the convergence to this limit using a weak convergence approach. This future research will clarify the uncertain connection between discrete-time Monte Carlo schemes that use resampling and continuous-time interacting particle systems.

### III. Error analysis for spectral estimation

**Background.** In molecular dynamics simulations of biochemical systems, impactful biological events typically occur on a slow timescale (\( \geq 10^{-6} \) s) compared to the fast timestep of numerical integrators (\( 10^{-15} \) s) [13]. Therefore, it is essential to make the most of limited data sets by accurately identifying slow conformation changes in the data. A popular Monte Carlo approach for achieving this goal is spectral estimation.

Spectral estimation aims to compute eigenvalues and eigenvectors of the transition operator \( T_t \), an infinite-dimensional operator defined by \( T_t [f] (x) = \mathbb{E} [f (X_t) | X_0 = x] \). To perform spectral estimation, the user first calculates the time-lagged correlations \( C_{ij} (t) = \mathbb{E} [\phi_i (X_0) \phi_j (X_t)] \) for a set of orthonormal basis functions \( \{ \phi_i \}_{1 \leq i \leq n} \). Next, the user performs an eigendecomposition of the matrix \( C(t) \) at a fixed time lag \( t = \tau \) to estimate eigenvectors and eigenvalues of the transition operator.

In applications, spectral estimation has aided the interpretation of complex data sets. Typically, the leading 2-6 eigenvectors of the transition operator describe the slow conformation

![Figure 3: Error of Monte Carlo estimates using different resampling schemes. Multinomial resampling is a commonly used scheme. Pivotal resampling is an improved scheme satisfying properties (a) and (b). Sorted resampling is the optimal scheme identified in the theorem.](image)
changes in the system, and the associated eigenvalues determine the timescales of the conformation changes. When molecular dynamics trajectories are projected down onto these eigenvectors, the data display clear, interpretable pathways leading to biological functionality [14].

Despite the popularity of spectral estimation in biochemistry, the current mathematical understanding of this approach is limited. There is no published convergence proof for spectral estimation, and no error formulas bound the extent of spectral estimation error. Moreover, because of this incomplete mathematical understanding, it is unclear how best to tune algorithmic parameters to reduce spectral estimation error. In applications, the time lag parameter $\tau$ is seen to play a critical role in the accuracy of spectral estimation [15], yet there is no rigorous standard for how to select $\tau$.

**Objectives.** The first objective of my ongoing research in spectral estimation is to prove convergence of the most popular spectral estimation method in chemistry, called the variational approach in conformation dynamics (VAC, [16]). My second objective is to analyze how best to tune the time lag parameter $\tau$ in order to reduce spectral estimation error.

**Results.** In collaboration with Jonathan Weare and experts in biochemistry, I proved the convergence of VAC estimates for eigenvectors of the transition operator. VAC estimates are guaranteed to converge as the sampling becomes increasingly accurate and as the projection of eigenvectors onto the basis set approaches the true eigenvectors. This result establishes the validity of VAC and provides simple bounds to help gauge error.

To provide a more nuanced understanding of VAC error, I derived condition numbers for the VAC problem with respect to small sampling errors and small errors in the basis set. These condition numbers depend explicitly on time lag $\tau$ through the VAC eigenvalues $\lambda_i = \lambda_i(\tau)$. At short time lags $\tau$, the VAC problem is especially sensitive to small errors in the basis set. At long time lags $\tau$, the VAC problem is especially sensitive to small errors in the sampling. As seen in Figure 4, it is advantageous to select an intermediate time lag parameter, optimizing the conditioning of the VAC problem and thereby reducing error.

![Figure 4: Estimation error increases at short time lags and long time lags. Plot shows error in calculating the second eigenvector of the time-reversible system $dX = -X dt + \sqrt{2} dW$.](image)

Currently, we are building software that automatically calculates condition numbers for the VAC problem using data from simulations. This software will enable users select an appropriate time lag $\tau$ where eigenvector estimates are likely to be most accurate.

**Future research.** The main limitation of the VAC algorithm is the assumption that the Markov process $X_t$ is time-reversible. My goal in future work is therefore to remove the time-reversibility assumption and analyze the performance of more general spectral estimation schemes. In chemistry and geophysics, non-reversible diffusion processes are commonly encountered, so I plan to establish convergence results and error estimates for this important class of processes. Through this analysis, moreover, I plan to investigate strategies for manipulating the time lag to reduce spectral estimation error. Whereas VAC uses a single time lag to compute properties of the transition operator, there is an opportunity to aggregate information at multiple time lags, thereby reducing error through an averaging effect. This research is of direct benefit to scientists that are using spectral estimation to identify the slow dynamics of complex systems, and it extends the applicability of spectral estimation for the future.
References cited


