Biometrika (2021), **xx**, x, *pp*. 1–14
Printed in Great Britain Manuscript

Stratification and Optimal Resampling for Sequential Monte Carlo

BY YICHAO LI
Center for Statistical Science, Tsinghua University, Beijing, 100084, China
liyichao16@mails.tsinghua.edu.cn

WENSHUO WANG
Department of Statistics, Harvard University, Cambridge, Massachusetts, 02138, U.S.A.
wenshuo_wang@g.harvard.edu

KE DENG
Center for Statistical Science, Tsinghua University, Beijing, 100084, China
kdeng@tsinghua.edu.cn

AND JUN S. LIU
Department of Statistics, Harvard University, Cambridge, Massachusetts, 02138, U.S.A.
jliu@stat.harvard.edu

SUMMARY

Sequential Monte Carlo algorithms have been widely accepted as a powerful computational tool for making inference with dynamical systems. A key step in sequential Monte Carlo is resampling, which plays a role of steering the algorithm towards the future dynamics. Several strategies have been used in practice, including multinomial resampling, residual resampling, optimal resampling, stratified resampling, and optimal transport resampling. In the one-dimensional cases, we show that optimal transport resampling is equivalent to stratified resampling on the sorted particles, and they both minimize the resampling variance as well as the expected squared energy distance between the original and resampled empirical distributions. In general d-dimensional cases, if the particles are first sorted using the Hilbert curve, we show that the variance of stratified resampling is $O(m^{-(1+2/d)})$, an improved rate compared to the previously known best rate $O(m^{-(1+1/d)})$, where m is the number of resampled particles. We show this improved rate is optimal for ordered stratified resampling schemes, as conjectured in Gerber et al. (2019). We also present an almost sure bound on the Wasserstein distance between the original and Hilbert-curve-resampled empirical distributions. In light of these results, we show that, for dimension $d > 1$, the mean square error of sequential quasi-Monte Carlo with n particles can be $O(n^{-1-4/\{d(d+4)\}})$ if Hilbert curve resampling is used and a specific low-discrepancy set is chosen. To our knowledge, this is the first known convergence rate lower than $o(n^{-1})$.

Some key words: Hilbert space-filling curve; Particle filter; Resampling; Sequential Monte Carlo; Sequential quasi-Monte Carlo; Stratification.

1. INTRODUCTION

Sequential Monte Carlo, which dates back to the study of self-avoiding random walks and molecular structural optimizations (Hammersley & Morton, 1954; Rosenbluth & Rosenbluth, 1955; Siepmann & Frenkel, 1992; Grassberger, 1997), has been studied intensively in the past two decades and applied broadly to high-dimensional statistical inference, signal processing, biology and many other areas (Liu & Chen, 1998; Doucet et al., 2001). Through building up the sampling (trial) distribution sequentially, a set of weighted samples can be used to approximate the high-dimensional target distribution. The state-space model is a particularly interesting dynamical system that has been treated with sequential Monte Carlo. The model is defined by Markovian dynamics for a hidden state and an emission distribution that relates the hidden state to noisy observations. The hidden state, for instance, can represent the underlying volatility in an economical time series (Taylor, 2008; Gatheral, 2011), or the location in a terrain navigation problem (Bergman et al., 1999; Bergman, 2001; Gustafsson et al., 2002), or many others. In such models, characterizing the distribution of the hidden state is known as the filtering problem, and within this context, sequential Monte Carlo is also known as the bootstrap filter (Gordon et al., 1993), Monte Carlo filter (Kitagawa, 1996), or particle filter (Del Moral, 1997).

Roughly speaking, sequential Monte Carlo is built based on sequential importance sampling, which recursively simulates a future state, reweights the sample path, and then potentially resamples the paths (Liu & Chen, 1998). In sequential imputation (Kong et al., 1994), which is a form of sequential importance sampling without any resampling, weight degeneracy arises as an inevitable problem. Since the importance weights are updated recursively at each step, stochastically the total weights will concentrate on a very few samples, leading to an exponentially increasing coefficient of variation. One effective strategy to avoid weight degeneracy is to resample from the current samples according to the corresponding weights (Liu & Chen, 1995). Resampling alone does not provide any additional information but only adds noise to the estimate of the current state. A main motivation for resampling is the belief that particles with small weights are unpromising for further development and thus should be discarded so as to reallocate resources to particles with larger weights for exploring regions that may be more promising for the future. Liu & Chen (1995) provided an early attempt at analyzing resampling for statistical models, providing some useful insights but lacking a rigorous theory.

There are various means to resample from a collection of weighted particles. Informally, one would like to minimize the "resampling randomness" over a certain class of valid resampling schemes. This goal is closely related to the balanced sampling design in survey sampling (Tillé, 2006, Chapter 8), which seeks to reduce the sampling variance using auxiliary variables. A naïve way to resample is bootstrap resampling or multinomial resampling (Gordon et al., 1993), where the new particles are sampled from independent and identically distributed multinomial distributions based on the original particle weights. Residual resampling (Liu & Chen, 1998) and stratified resampling (Kitagawa, 1996) are two more popular resampling schemes in practice. These methods have also been studied and used in scientific fields outside of statistics under different names of resampling, such as parent selection for genetic algorithms (Brindle, 1980, Chapter 4.2) and, in physics, stochastic reconfiguration (Gubernatis et al., 2016, Chapter 10.3). Douc & Cappé (2005) compared the above resampling schemes and concluded that residual resampling and stratified resampling always have a smaller conditional variance than multinomial resampling does. For discrete state-spaces, the optimal resampling method (Fearnhead & Clifford, 2003) offers an interesting way of diversified sampling. Besides these traditional resampling schemes, Reich (2013) proposed optimal transport resampling, an approach borrowing ideas from transportation theory. However, to the best of our knowledge, there has been no theo-

retical guarantee for optimal transport resampling aside from its validity. Recently, Gerber et al. (2019) showed that stratified resampling after ordering the particles by the Hilbert space-filling curve has a relatively low conditional variance in some cases, which is also one of our interests in this article.

Sequential quasi-Monte Carlo, introduced in Gerber & Chopin (2015), is a class of algorithms taking advantage of Hilbert curve resampling and quasi-Monte Carlo point sets. By constructing a low-discrepancy set on a product space, sequential quasi-Monte Carlo combines resampling and growth and numerically outperforms regular sequential Monte Carlo significantly. Theoretically, however, the convergence rate in terms of the mean square error has only been shown to be $o(n^{-1})$ for certain low-discrepancy sets. It is naturally believed that the rate could be improved and should depend on the dimension d.

In this paper, we focus on theoretical properties of various resampling schemes and sequential quasi-Monte Carlo. We show that, in the one-dimensional case, optimal transport resampling is equivalent to stratified resampling on the sorted particles, which minimizes the resampling variance as well as the expected squared energy distance between the empirical distributions before and after resampling. In the d-dimensional case, a natural generalization of ordered stratified sampling in one dimension is Hilbert curve resampling (Gerber et al., 2019), which is stratified resampling on particles sorted using the Hilbert space-filling curve. We prove that its resampling variance is of the order $O(m^{-(1+2/d)})$ when $d > 1$, where m is the number of resampled particles. This improves the previous best known rate $O(m^{-(1+1/d)})$. We show that the order cannot be further improved by resorting to a different ordering rule, confirming a conjecture in Gerber et al. (2019). We also derive a bound on the Wasserstein distance between the empirical distributions before and after Hilbert curve resampling. Based on the theoretical results on resampling, we further design a low-discrepancy set for sequential quasi-Monte Carlo and prove that the mean square error under this set is of the order $O(n^{-1-4/\{d(d+4)\}})$ for $d > 1$. This improves the original rate $o(n^{-1})$. We believe this low-discrepancy set captures some key intuitions of quasi-Monte Carlo; the tools, moreover, may be of independent interest for the analysis of other low-discrepancy sets.

2. PRELIMINARIES

2.1. *Notations*

We use superscripts to denote the step or iteration and subscripts for the sample index; the temporal notations are omitted for the sake of clarity whenever there is no confusion. The target distribution is denoted as $\pi(x)$, while $g(x)$ denotes a trial distribution. When written without a subscript, X and W mean $(X_1, X_2, \ldots, X_n)$ and $(W_1, W_2, \ldots, W_n)$ for an appropriate n, and the set of tuples $(X_j, W_j)_{j=1}^n$ refers to a set of weighted samples, where $W_j \geq 0, j = 1, 2, \cdots, n$. Unless stated otherwise, the W_j's are normalized so that $\sum_{j=1}^n W_j = 1$. We use $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_m$ to denote the equally weighted samples after resampling, so that in some sense, $\sum_{i=1}^m m^{-1}\delta_{\tilde{X}_i} \approx \sum_{j=1}^n W_j\delta_{X_j}$, where δ_x denotes the Dirac measure at point x. If $X_j \in \mathcal{X}$ for $j = 1, 2, \ldots, n$, we use $\mathcal{X}^n$ to denote the space in which X lives. We use $Z \sim \text{Multinomial}(1, y, p)$ to mean that $\text{pr}(Z = y_i) = p_i$, where p is a probability vector. We write $m_d(\cdot)$ for the Lebesgue measure in d dimensions. The standard L_2 norm is denoted as $\|\cdot\|$. For a vector a, $\text{diag}(a)$ represents the diagonal matrix with the ith diagonal element being a_i. For a real number u, $\lfloor u \rfloor$ denotes the greatest integer less than or equal to u. The symbol $\overset{\text{i.i.d.}}{\sim}$ denotes sampling independent and identically distributed random variables.

2.2. *Sequential Monte Carlo*

To set up future analyses, we here describe a generic sequential Monte Carlo procedure. Let the target distribution $\pi(x)$ be supported in a T-dimensional space, which can be viewed as a joint distribution of a sequence of variables, say $\pi(x^{(1:T)})$. We can sample sequentially from a sequence of distributions $\{\pi_t(x^{(1:t)})\}_{t=1}^T$, where $\pi_T = \pi$. A generic sequential Monte Carlo algorithm is outlined in Algorithm 1.

Algorithm 1. Sequential importance sampling with resampling.

Input: A sequence of target distributions $\{\pi_t(x^{(1:t)})\}_{t=1}^T$ and a sequence of trial distributions $g_1(x^{(1)})$ and $\{g(x^{(t)} \mid x^{(1:t-1)})\}_{t=2}^T$
Output: weighted particles $(X_i^{(1:T)}, W_i^{(T)})_{1\le i\le n}$
At time $t = 1$,
Draw $X_1^{(1)}, \cdots, X_n^{(1)}$ from $g_1(X^{(1)})$.
Calculate and normalize the importance weight: $W_j^{(1)} \propto \pi_1(X_j^{(1)})/g_1(X_j^{(1)})$.
Resample $\tilde{X}_1^{(1)}, \tilde{X}_2^{(1)}, \cdots, \tilde{X}_n^{(1)}$ from $X_1^{(1)}, \cdots, X_n^{(1)}$ with probabilities $W_1^{(1)}, \cdots, W_n^{(1)}$, and reweight the samples $\tilde{X}_1^{(1)}, \tilde{X}_2^{(1)}, \cdots, \tilde{X}_n^{(1)}$ equally with $1/n$.
Let $X_j^{(1)} = \tilde{X}_j^{(1)}$ for $j = 1, 2, \ldots, n$.
for $t = 2$ **to** T **do**
Draw $X_j^{(t)}$ from $g_t(X^{(t)} \mid X_j^{(1:t-1)})$ for $j = 1, 2, \ldots, n$ conditionally independently.
Calculate and normalize the importance weight:

$$W_j^{(t)} \propto \pi_t\left(X_j^{(1:t)}\right) \Big/ \left\{\pi_{t-1}\left(X_j^{(1:t-1)}\right) g_t\left(X_j^{(t)} \mid X_j^{(1:t-1)}\right)\right\}$$

if $t < T$ **then**
Resample $\tilde{X}_1^{(1:t)}, \tilde{X}_2^{(1:t)}, \cdots, \tilde{X}_n^{(1:t)}$ from $X_1^{(1:t)}, \cdots, X_n^{(1:t)}$ with probabilities $W_1^{(t)}, \cdots, W_n^{(t)}$, and reweight the samples $\tilde{X}_1^{(1:t)}, \tilde{X}_2^{(1:t)}, \cdots, \tilde{X}_n^{(1:t)}$ equally with $1/n$.
Let $X_j^{(1:t)} = \tilde{X}_j^{(1:t)}$.
Return $(X_i^{(1:T)}, W_i^{(T)})_{1\le i\le n}$

In the special case of a state-space model, we have

$$\begin{aligned} Y^{(t)} \mid \left(X^{(1:t)} = x^{(1:t)}, Y^{(1:t-1)}\right) &\sim p_y(\cdot \mid x^{(t)}), \\ X^{(t)} \mid \left(X^{(1:t-1)} = x^{(1:t-1)}, Y^{(1:t-1)}\right) &\sim p_x(\cdot \mid x^{(t-1)}), t = 2, \cdots, T, \end{aligned} \tag{1}$$

where p_x and p_y represent distributions as well as density functions, $X^{(1)}, \cdots, X^{(T)}$ are unobserved hidden states, and $Y^{(1)}, \cdots, Y^{(T)}$ are the observed sequence of variables. The filtering problem focuses on the target distribution

$$\pi_T(x^{(1:T)}) \propto \prod_{t=1}^{T} \left\{p_x(x^{(t)} \mid x^{(t-1)}) p_y(y^{(t)} \mid x^{(t)})\right\}.$$

While implementing Algorithm 1 for such a state-space model, the trial distribution at each step can be naturally (or naïvely) chosen as $g_t(x^{(t)} \mid x^{(t-1)}) = p_x(x^{(t)} \mid x^{(t-1)})$, and thus the corresponding importance weight can be updated as $w^{(t)} \propto w^{(t-1)} p_y(y^{(t)} \mid x^{(t)})$.

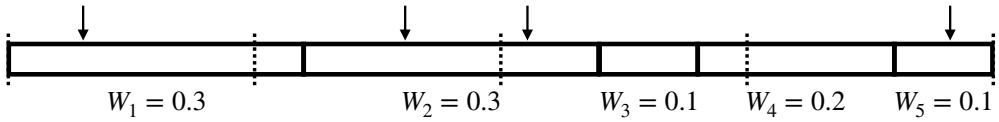


Fig. 1: Illustration of stratified resampling. First line up the weights, then divide the interval into m equal-sized subintervals. Uniformly choose one point from each subinterval and record in which weight's region it lands. In the presented example where $m = 4$, $n = 5$, particles 1 and 5 are resampled once, particle 2 is resampled twice and particles 3 and 4 are discarded.

2.3. *Resampling matrix*

Suppose we have weighted particles $(W_j, X_j)_{j=1}^n$ with weights summing to one. Without loss of generality, we assume that the X_j's are distinct since we can always merge particles with identical values and add up their weights. Consider the family of resampling methods indexed by a matrix $P_{m\times n}$, where the new unweighted particles $(\tilde{X}_i)_{i=1}^m$ are sampled independently from

$$\tilde{X}_i \mid X, W \sim \text{Multinomial}(1, X, (p_{i1}, p_{i2}, \ldots, p_{in})),$$

and P has non-negative entries with $\sum_{i=1}^m p_{ij} = mW_j$ and $\sum_{j=1}^n p_{ij} = 1$. Note that permuting P's rows does not change the resampling scheme. It can be easily verified that such a resampling strategy is unbiased, which means that for any function ϕ we have

$$E\left\{\frac{1}{m}\sum_{i=1}^m \phi(\tilde{X}_i) \mid X, W\right\} = \frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m p_{ij}\phi(X_j) = \sum_{j=1}^n W_j\phi(X_j).$$

We use $\mathcal{P}_{m,W}$ to denote the set of all matrices of this form and the set of all corresponding resampling methods, with a slight abuse of notation. We call this collection of resampling methods matrix resampling methods. The use of resampling matrices appeared at least as early as in Hu et al. (2008), and subsequently in many other works (Reich, 2013; Whiteley et al., 2016; Webber, 2019). Most available resampling methods, as listed below, fit into this framework.

In multinomial resampling, each $\tilde{X}_i$ is an independent and identically distributed sample from the multinomial distribution $\text{Multinomial}(1, X, W)$. This corresponds to $p_{ij} = W_j$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, as shown in Figure 2(a). In stratified resampling, we let $U_i \sim \text{Unif}\,((i-1)/m, i/m]$, independently for $i = 1, \ldots, m$, and let $\tilde{X}_i = X_j$ if $U_i \in \left(\sum_{k=1}^{j-1} W_k, \sum_{k=1}^{j} W_k\right]$. See Figure 1 for an illustration. Stratified resampling corresponds to a staircase matrix; see Figure 2(b) for an example and Definition 1 for a formal definition. In residual resampling, we first make $\lfloor mW_j \rfloor$ copies of X_j for all $j = 1, \ldots, n$; then, apply multinomial or stratified resampling (corresponding to Figure 2(c) and (d), respectively) for drawing the rest $m - \sum_{j=1}^n \lfloor mW_j \rfloor$ particles with $\tilde{W}_j \propto mW_j - \lfloor mW_j \rfloor$.

2.4. *Criteria for choosing resampling schemes*

To choose from the set of valid resampling procedures, we need a measure of goodness of a resampling procedure. Let $\mathbb{P} = \sum_{j=1}^n W_j\delta_{X_j}$ and $\tilde{\mathbb{P}} = \sum_{i=1}^m m^{-1}\delta_{\tilde{X}_i}$. It is natural to favor a stable process, where $\tilde{\mathbb{P}}$ is close to $\mathbb{P}$. Explicitly, we want to minimize $E\{\ell(\mathbb{P}, \tilde{\mathbb{P}}) \mid X, W\}$ for a loss function ℓ. For example, we can pick $\ell(\mathbb{P}, \tilde{\mathbb{P}})$ to be $[E_{\mathbb{P}}\{\phi(X)\} - E_{\tilde{\mathbb{P}}}\{\phi(X)\}]^2$ and use the conditional variance $\text{var}\{m^{-1}\sum_{i=1}^m \phi(\tilde{X}_i) \mid X, W\}$ as a measure of goodness. We can also choose ℓ to be the squared energy distance, which has the advantage of explicit expression and the property that the energy distance is zero if and only if two distributions are the same. The

$$\begin{pmatrix} 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \end{pmatrix} \quad \begin{pmatrix} 1 & & & & \\ 0.2 & 0.8 & & & \\ & 0.4 & 0.4 & 0.2 & \\ & & & 0.6 & 0.4 \end{pmatrix} \quad \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \end{pmatrix} \quad \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ 0.2 & 0.2 & 0.4 & 0.2 & \\ & & & 0.6 & 0.4 \end{pmatrix}$$

(a) Multinomial (b) Stratified (c) Multinomial Residual (d) Stratified Residual

Fig. 2: Examples of resampling matrices with $m = 4$ and $n = 5$, and particle weights $(W_1, W_2, W_3, W_4, W_5) = (0.3, 0.3, 0.1, 0.2, 0.1)$.

energy distance (Rizzo & Székely, 2016) between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as the square root of

$$D^2(\mathbb{P}_1, \mathbb{P}_2) = 2E(\|Y_1 - Y_2\|) - E(\|Y_1 - Y_1'\|) - E(\|Y_2 - Y_2'\|),$$

where Y_1 and Y_1' follow $\mathbb{P}_1$, Y_2 and Y_2' follow $\mathbb{P}_2$, and the four random variables are mutually independent. Another example is the Wasserstein distance, for $p \geq 1$, defined between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ as

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\gamma \in \Gamma(\mathbb{P}_1, \mathbb{P}_2)} \left\{ E_{(Y_1, Y_2) \sim \gamma}(\|Y_1 - Y_2\|^p) \right\}^{1/p},$$

where $\Gamma(\mathbb{P}_1, \mathbb{P}_2)$ is the set of all probability measures that have $\mathbb{P}_1$ and $\mathbb{P}_2$ as their marginal distributions.

In Section 3, we prove that minimizing the conditional variance is equivalent to minimizing the expected squared energy distance in one-dimensional cases, both of which can be achieved by ordered stratified resampling (i.e., stratified resampling on the sorted particles). In Section 4, we give upper bounds for conditional variance and expected Wasserstein distance for ordered stratified resampling, where the particles are sorted according to the Hilbert curve in multiple dimensions.

3. OPTIMAL RESAMPLING IN ONE DIMENSION

A good resampling scheme should ideally incorporate the information of the state values, since the loss function usually depends on them. In this section, we show that, by incorporating the X_j's value information, the stratified resampling method minimizes several objectives proposed in the literature. We consider the case in which the particles take values in a one-dimensional space. One example is the state-space model with one-dimensional hidden states.

We first define the staircase matrix, which is the same as a stratified resampling matrix as we show in Proposition 1. In doing so, we gain a greater insight into why ordering the states before applying stratified resampling can lower the resampling variance.

DEFINITION 1 (STAIRCASE MATRIX). *We call a matrix P staircase matrix if (i) in each row and column of P, non-zero entries are consecutive. In other words, if $p_{ij_1} \neq 0$ and $p_{ij_2} \neq 0$ for $j_1 < j_2$, then for all $j_1 < j < j_2$, $p_{ij} \neq 0$; and (ii) for any quadruplet (i, j, k, l) such that $i < k, j < l$, at least one of p_{il} and p_{kj} is 0.*

A staircase matrix has at most $n + m - 1$ non-negative entries and has a clear spatial structure. The non-negative entries form a path from the top left entry to the bottom right entry, allowing diagonal moves. For example, Figure 2 (b) is a staircase matrix, but the other three are not. Below we show uniqueness of the staircase representation and its relevance to stratified resampling.

LEMMA 1. *For $m, n > 2$ and positive r_i's and c_j's, there can only be one unique m by n staircase matrix with non-negative entries and satisfies $\sum_{j=1}^{n} p_{ij} = r_i$ and $\sum_{i=1}^{m} p_{ij} = c_j$.*

PROPOSITION 1. *Any stratified resampling scheme corresponds to a unique staircase matrix up to row permutations.*

Lemma 1 and Proposition 1 allow us to define the stratified resampling matrix.

DEFINITION 2 (STRATIFIED RESAMPLING MATRIX). *We call a matrix $P_{m,W}^{\mathrm{SR}} \in \mathcal{P}_{m,W}$ the stratified resampling matrix of a set of weighted particles $(X_j, W_j)_{j=1}^{n}$ if $P_{m,W}^{\mathrm{SR}}$ can be converted to a staircase matrix after some row permutation.*

THEOREM 1. *For particles $(X_j, W_j)_{j=1}^{n}$ with $X_1 < X_2 < \cdots X_n$, resampling defined by $P_{m,W}^{\mathrm{SR}}$ minimizes the following objectives:*

(i) The conditional variance $\mathrm{var}_P\left(m^{-1}\sum_{i=1}^{m}\tilde{X}_i \mid X, W\right)$.

(ii) The expected squared energy distance $E_P\left\{D^2\left(\sum_{i=1}^{m} m^{-1}\delta_{\tilde{X}_i}, \sum_{j=1}^{n} W_j\delta_{X_j}\right)\right\}$.

(iii) The earth mover distance $\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\ell(Y_i - X_j)$ where ℓ is a strictly convex function, and $Y_1 < \cdots < Y_m$ is any given sequence of ascending numbers.

Remark 1. If the goal is to estimate $E\{\phi(X)\}$, then ordering the states by function ϕ and then applying stratified resampling gives the minimum variance. This result also appeared in Webber (2019), where it was proved using an optimization argument. Our proof uses a similar idea and directly shows that when the resampling variance is minimized, the resampling matrix must be a staircase matrix and corresponds to ordered stratified resampling. A similar approach is used to prove (iii) as well.

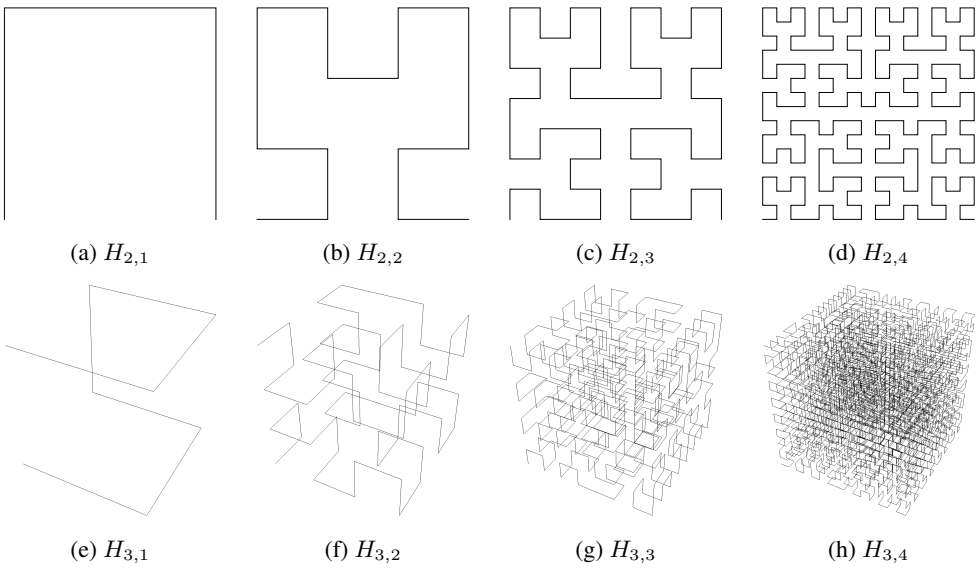
4. ERROR OF ORDERED STRATIFIED RESAMPLING

Intuitively, since the new particles in resampling are sampled independently, in order to minimize a discrepancy measure, we want to make sure that each new particle brings in little randomness. It is easy to see from the staircase structure of a resampling matrix that each $\tilde{X}_i$ takes value in a sequence of consecutive X_j's according to the order that enables the staircase structure. Thus, if these X_j's are ordered in the one-dimensional space and function $\phi(\cdot)$ is Lipschitz, then the values $\phi(\tilde{X}_i)$ are bounded in a small region, which leads to the following result.

THEOREM 2. *Suppose one-dimensional particles $(\tilde{X}_i)_{i=1}^{m}$ are resampled with ordered stratified resampling from $(X_j, W_j)_{j=1}^{n}$, then for any Lipschitz function ϕ with coefficient L_ϕ,*

$$\mathrm{var}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi(\tilde{X}_i) \mid X, W\right\} \le \frac{L_\phi^2}{4m^2}\left(\max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i\right)^2.$$

In multiple dimensions, it has been noticed that the Hilbert space-filling curve (Hilbert, 1935) can help lower the sampling variance (Gerber & Chopin, 2015; He & Owen, 2016; Gerber et al., 2019). In particular, Gerber et al. (2019) used the Hilbert curve in the context of resampling. They showed that the resampling variance for Lipschitz functions with m particles is of order $O(m^{-(1+1/d)})$, where d is the number of dimensions. We improve this bound to $O(m^{-(1+2/d)})$ and show that this new rate is the best for ordered stratified resampling schemes with any ordering, as conjectured in Gerber et al. (2019).



(a) $H_{2,1}$ (b) $H_{2,2}$ (c) $H_{2,3}$ (d) $H_{2,4}$

(e) $H_{3,1}$ (f) $H_{3,2}$ (g) $H_{3,3}$ (h) $H_{3,4}$

Fig. 3: Hilbert curves of the first four orders in two and three dimensions.

A d-dimensional Hilbert curve is a continuous function $H : [0,1] \to [0,1]^d$. Its most important properties relevant to our tasks are as follows:

(i) H is surjective.

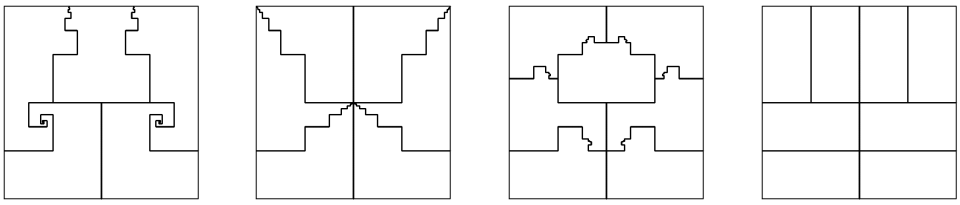
(ii) H is Hölder continuous with exponent $1/d$ (He & Owen, 2016):

$$\|H(x) - H(y)\| \leq 2\sqrt{d+3}|x-y|^{1/d}.$$

(iii) H is measure-preserving. For each Lebesgue measurable $I \subseteq [0,1]$, $m_1(I) = m_d(H(I))$.

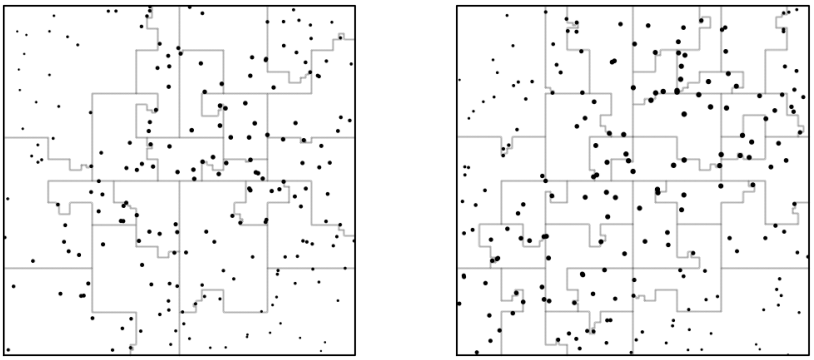
The Hilbert curve can be defined as the limit of a sequence of curves; see Figure 3 for an illustration in two and three dimensions. Many software packages can efficiently convert between x and $H(x)$, such as the Python package hilbertcurve. In practice, the computation cost of this approximation is quite minimal compared to the sampling part. We omit here the rigorous definition of Hilbert curves and refer interested readers to Sagan (2012). For the purpose of resampling, the most relevant property is the Hölder continuity. This ensures that $H(I)$, the image of an interval $I \subseteq [0,1]$, has its diameter bounded above by $2\sqrt{d+3} \cdot m_1(I)^{1/d}$. As an illustration, we plot the images of $H([i/k, (i+1)/k])$ for $i = 0, 1, \ldots, k-1$ and $k = 5, 6, 7, 8$ in Figure 4.

Now we formally introduce the Hilbert curve resampling first proposed in Gerber et al. (2019). Proposition 2 in Gerber et al. (2019) says that there exists a one-to-one Borel measurable function $h : [0,1]^d \to [0,1]$ such that $H(h(x)) = x$ for all $x \in [0,1]^d$. The resampling procedure is to first sort the particles so that $(h(X_j))_{j=1}^n$ is in ascending order, and then apply stratified resampling. Note that in one dimension this reduces to ordered stratified sampling. Following the intuition in the one-dimensional case, each resampled particle is bounded in a small region in $[0,1]^d$ due to the Hölder continuity of H, which limits the variability of $\tilde{X}_i$. See Figure 5 for an illustration. Theorem 3 gives an upper bound on the resampling variance, which is an improved bound compared to the one reported in Theorem 5 of Gerber et al. (2019).



(a) Five parts. (b) Six parts. (c) Seven parts. (d) Eight parts.

Fig. 4: The unit square divided into several parts with equal areas based on the Hilbert curve.



(a) $n = 200$ particles resampled into $m = 20$. (b) $n = 200$ particles resampled into $m = 30$.

Fig. 5: The unit square divided into m parts based on the Hilbert curve and the particle weights. The size of each point represents their particle weight. Each region contains particles with weights summing to one (neighbouring regions divide weights of the particles on the boundary).

THEOREM 3. *Let $\phi : [0,1]^d \to [0,1]$, $d > 1$, be a Lipschitz function with Lipschitz coefficient L_ϕ. If $(X_j)_{j=1}^n$ is sorted in ascending order by the value of $h(X_j)$, then stratified sampling satisfies*

$$\mathrm{var}_{\text{HC-strat}}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi(\tilde{X}_i) \mid X, W\right\} \leq \frac{(d+3)L_\phi^2}{m^{1+2/d}}.$$

Remark 2. The intuition behind Theorems 2 and 3 is the same: in stratified resampling, the variance of each individual resampled particle is controlled because it is sampled from a set of particles spatially close to each other. In fact, one can easily generalize Theorem 3 to the Hölder function case: if $|\phi(x) - \phi(y)| \leq L_\phi \|x - y\|^\beta$, $\beta \in (0, 1]$, then

$$\mathrm{var}_{\text{HC-strat}}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi(\tilde{X}_i) \mid X, W\right\} \leq \frac{(d+3)L_\phi^2}{m^{1+2\beta/d}}.$$

Remark 3. The exponent $1 + 2/d$ in the theorem improves the original rate $1 + 1/d$ in Gerber et al. (2019). Gerber et al. (2019) conjectured that the Hilbert curve is the best choice for ordering the particles. For clarity, we take the Lipschitz coefficient to be 1 and $m = n$. Define the space

of valid probability vectors as

$$\Delta_n = \left\{ (w_1, w_2, \ldots, w_n) \in \mathbb{R}^n : \sum_{j=1}^{n} w_j = 1, w_i \geq 0 \text{ for all } 1 \leq i \leq n \right\}.$$

Theorem 3 implies that

$$\limsup_{n\to\infty} n^{1+\frac{2}{d}} \sup_{X\in[0,1]^{d\times n}} \sup_{W\in\Delta_n} \sup_{\phi\in\Phi_d} \mathrm{var}_{\text{HC-strat}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \phi(\tilde{X}_i) \mid X, W \right\} \leq d+3,$$

where Φ_d denotes the set of 1-Lipschitz functions from $[0,1]^d$ to $[0,1]$, $d > 1$. For other space-filling curves with a different Hölder exponent that may be cheaper to implement, similar results hold with an exponent different from $1 + 2/d$. However, we show in Proposition 2 that no other ordering rule can improve the exponent $1 + 2/d$.

PROPOSITION 2. *Let Φ_d be the set of 1-Lipschitz functions from $[0,1]^d$ to $[0,1]$, $d > 1$. Let $o(x) : [0,1]^d \to [0,1]$ be a one-to-one function. The stratified sampling procedure after ordering particles by o satisfies*

$$\limsup_{n\to\infty} n^{1+\frac{2}{d}} \sup_{X\in[0,1]^{d\times n}} \sup_{W\in\Delta_n} \sup_{\phi\in\Phi_d} \mathrm{var}_{o\text{-strat}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \phi(\tilde{X}_i) \mid X, W \right\} \geq \frac{1}{27d}.$$

Hilbert resampling is also stable in terms of the Wasserstein distance, as stated in Theorem 4. The Wasserstein distance is arguably a more intuitive notion to measure the stability of a resampling algorithm than conditional variance. When $p \leq d$, Theorem 4 is intuitively optimal, since m balls with radius of the order $1/m^{1/d}$ are needed to cover the d-dimensional unit cube.

THEOREM 4. *Under d-dimensional Hilbert curve resampling, $d \geq 1$, the Wasserstein distance W_p between $\tilde{\mathbb{P}} = \sum_{i=1}^{m} m^{-1}\delta_{\tilde{X}_i}$ and $\mathbb{P} = \sum_{j=1}^{n} W_j \delta_{X_j}$ is almost surely upper bounded by $2\sqrt{d+3}\, m^{-\frac{1}{\max(p,d)}}$.*

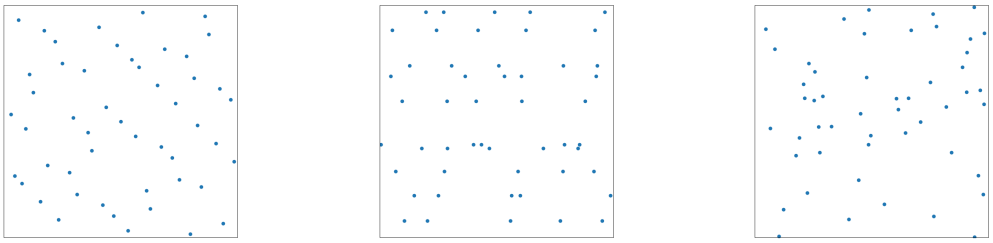
It is worthwhile to point out that in practice, depending on the target quantities of interest, there may exist an effective dimension lower than d. For example, if we only care about functions of the first $\tilde{d}$ coordinates, we should sort the particles using the $\tilde{d}$-dimensional Hilbert curve for the first $\tilde{d}$ coordinates of the particles; if the particles concentrate on a $\tilde{d}$-dimensional subspace, we should project the particles to this subspace and sort the particles using the corresponding $\tilde{d}$-dimensional Hilbert curve.

5. MEAN SQUARE ERROR OF SEQUENTIAL QUASI-MONTE CARLO

5.1. *Sequential quasi-Monte Carlo*

We show here how to utilize the results in previous sections to obtain a new convergence rate for the sequential quasi-Monte Carlo proposed in Gerber & Chopin (2015), which can be structured identically to Algorithm 1 with the same weight computation, but with different resampling and growth steps.

Suppose there exists function $\Gamma_1(\cdot)$ and $\Gamma_t(\cdot,\cdot)$ for $2 \leq t \leq T$ such that $\Gamma_1(V) \sim g_1(\cdot)$ and $\Gamma_t(X, V) \mid X \sim g_t(\cdot \mid X)$, where $V \sim \mathrm{Unif}([0,1]^d)$ is independent of X. Assume at the beginning of step t, we have weighted samples $(X_j^{(1:t-1)}, W_j^{(t-1)})_{j=1}^n$, which have been ordered by



(a) Sobol sequence. (b) Stratified multiple-descendant growth. (c) Independent sampling.

Fig. 6: Comparison of low-discrepancy sets on $[0,1]^2$ ($n = 50$, $k = 10$, $r = 5$).

the Hilbert mapping h so that $h(X_1^{(t-1)}) \leq \cdots \leq h(X_n^{(t-1)})$. Recall that Hilbert-curve stratified sampling can then be implemented by independently sampling $U_i \sim \mathrm{Unif}\{((i-1)/n, i/n]\}$ for $1 \leq i \leq n$ and let $\tilde{X}_i^{(t-1)} = X_{\sigma(U_i,W)}^{(t-1)}$, where $\sigma(U_i, W) = j$ if $\sum_{k=1}^{j-1} W_k < U_i \leq \sum_{k=1}^{j} W_k$. Suppose we have a low-discrepancy set $U^{(t)} = \{(u_i, v_i) : u_i \in [0,1], v_i \in [0,1]^d, 1 \leq i \leq n\}$, labeled in the way such that the $u_{1:n}$ are in the ascending order. Intuitively speaking, a low-discrepancy set is a set that spreads evenly in $[0,1]^{1+d}$; see Gerber & Chopin (2015) for a more detailed discussion. Sequential quasi-Monte Carlo combines resampling and growth by defining $X_j^{(1)} = \Gamma_1(v_j)$ and $X_j^{(t)} = \Gamma_t(X_{\sigma(u_j, W_{1:n}^{(t-1)})}^{(t-1)}, v_j)$, for $2 \leq t \leq T$ and $1 \leq j \leq n$. If the set $U^{(t)}$ contains n independent samples from $\mathrm{Unif}([0,1]^{1+d})$, then we recover Algorithm 1 with Hilbert resampling. It was shown in Gerber & Chopin (2015) that some choice of $U^{(t)}$ (e.g., the nested scrambled Sobol sequence) can achieve a mean square error of order $o(n^{-1})$. Next, we will show that a specifically chosen set can achieve $O(n^{-1-4/\{d(d+4)\}})$.

5.2. *Stratified multiple-descendant growth*

The intuition behind sequential quasi-Monte Carlo is that the consecutive resampled particles $(X_{\sigma(u_j, W_{1:n}^{(t)})}^{(t)})_{j=a}^{b}$ are close in space due to the Hölder continuity of the Hilbert curve, so if $v_{a:b}$ are more spread out, the space can be probed more consistently by stratified growth. The main difficulty of quantifying the convergence rate of sequential quasi-Monte Carlo lies in the deterministic or semi-deterministic nature of the set $U^{(t)}$. We exploit this intuition and construct a specific set that enables a more careful convergence analysis.

Let $n = sr$, and let $U_k \sim \mathrm{Unif}\{((k-1)/s, k/s]\}$ be independent for $1 \leq k \leq s$. Let $V_{(k-1)s+\ell} = H(\tilde{V}_{k\ell})$, where H is the d-dimensional Hilbert curve and $\tilde{V}_{k\ell} \sim \mathrm{Unif}\{((\ell-1)/r, \ell/r]\}$, independently for $1 \leq k \leq s$, $1 \leq \ell \leq r$. We define $U_{\mathrm{SMG}}^{(t)} = \{(U_{\lfloor i/r \rfloor + 1}, V_i) : 1 \leq i \leq n\}$. Here, SMG stands for stratified multiple-descendant growth, because we essentially resample s particles, and let each particle have r descendants in a stratified manner. This idea is also closely related to the optimal resampling in the discrete space (Fearnhead & Clifford, 2003). Figure 6 compares the discrepancy set generated by stratified multiple-descendant growth and two other approaches. The next theorem focuses on bounding the mean square error of the SMG estimate of the posterior mean of ϕ in a state-space model.

THEOREM 5. *In a state-space model* (1)*, we let* $g_t(x^{(t)} \mid x^{(t-1)}) = p_x(x^{(t)} \mid x^{(t-1)})$ *and run sequential quasi-Monte Carlo with* $U_{\mathrm{SMG}}^{(t)}$*. Assume that each* $X^{(t)}$ *falls in a compact set, assuming*

to be $\mathcal{X} = [0,1]^d$ *without loss of generality. Suppose* $(X_j^{(t)}, W_j^{(t)})_{1\le j\le n}$ *are the weighted samples at time* t*, where the number of multiple descendants* $r = cn^{2/(d+4)}$ *and particle dimension* $d \ge 2$*. Assume that, for any* t*,*

(i) $a(v) = \pi_{t-1}(X)^{-1} g_t(v \mid X)^{-1} \pi_t((X,v))$, $b(v) = \pi_{t-1}((X,v))^{-1} \pi_t((X,v,u))$, $c(v) = \Gamma_t(X,v)$*, and* $\Gamma_1(v)$ *are bounded in* $[-M,M]$ *and L-Lipschitz.*
(ii) $\pi_{t-1}((X,v))^{-1} \int_{\mathcal{X}} \pi_t((X,v,u))\,du$ *is lower bounded by* $\underline{e} > 0$*.*

Then, for any L-Lipschitz ϕ *bounded in* $[-M,M]$*,*

$$\left| E\left\{ \frac{\sum_{j=1}^n W_j^{(t)} \phi(X_j^{(t)})}{\sum_{j=1}^n W_j^{(t)}} \right\} - \int \pi_t(x^{(1:t)}) \phi(x^{(t)}) dx^{(1:t)} \right| = O(n^{-\frac{1}{2} - \frac{2}{d(d+4)}}),$$

$$\operatorname{var}\left\{ \frac{\sum_{j=1}^n W_j^{(t)} \phi(X_j^{(t)})}{\sum_{j=1}^n W_j^{(t)}} \right\} = O(n^{-1-\frac{4}{d(d+4)}})$$

for all t*, where the constants in* O *depend only on* M*,* L*,* $\underline{e}$ *and* t*.*

Remark 4. There are different ways to map generally supported random vectors into $[0,1]^d$. Here we recommend the inverse transformation method proposed in Gerber & Chopin (2015). One significant advantage is that the spatial structure of the particles is preserved to a large extent by this method.

In dimension $d = 2$, our simulations in a stochastic volatility model seem to suggest that the rate is rather tight. The results are shown in Figure 7 and the model details are included in Appendix B. We can see that the empirical slope gets closer to the slope $-4/3$ given by Theorem 5 as n gets larger.

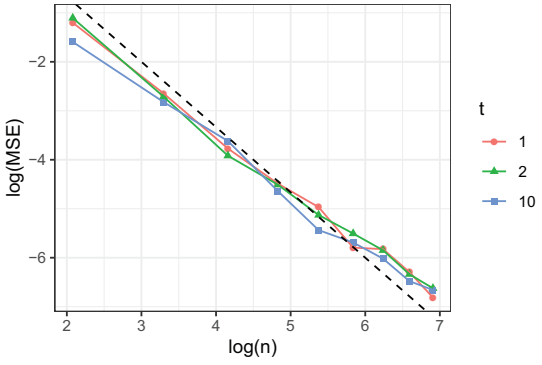


Fig. 7: The m versus the number of particles in logarithmic scales. The dashed line is a reference line with a slope of $-4/3$, the rate shown by our theory.

6. DISCUSSION

We have discussed how one may improve the performance of sequential Monte Carlo and sequential quasi-Monte Carlo via stratified sampling and multi-descendent growth. The matrix resampling framework in Section 2.3 can be generalized to allow resampled particles to carry

unequal weights, such as in the optimal resampling of Fearnhead & Clifford (2003). Let $q_{1:m}$ satisfy $q_i \geq 0$ and $\sum_{i=1}^m q_i = 1$. We can resample according to a matrix $P = (p_{ij})_{m\times n}$ with non-negative entries, where $\sum_{j=1}^n p_{ij} = 1$ and $\sum_{i=1}^m q_i p_{ij} = W_j$, by conditionally independent sampling:

$$X_i^* \mid X, W \sim \text{Multinomial}(1, X, (p_{i1}, p_{i2}, \ldots, p_{in})), i = 1, 2, \ldots, m,$$

and then assigning X_i^* the weight q_i. We focused on the case with $q_i = 1/m$ in this article, but by choosing unequal q_i's, one may further reduce the resampling variance at the cost of less balanced weights. It is unclear what an optimal trade-off might be.

When the resampled particles are not mutually independent conditional on the original particles, the resampling method cannot be represented by a resampling matrix. Systematic resampling (Carpenter et al., 1999) is such an example. All criteria mentioned in Section 2.4 are still well-defined for non-matrix resampling, but techniques developed here are not directly applicable. It is of interest to investigate if the Hilbert curve can still be utilized effectively in a broader class of resampling methods beyond the independent ones.

While our theoretical results are on sequential quasi-Monte Carlo with multiple-descendent growth, we believe that there are better choices of low-discrepancy sets. In fact, the Sobol sequence may be such an example based on our preliminary simulations, and it was conjectured in Gerber & Chopin (2015) that the optimal convergence rate of sequential quasi-Monte Carlo can reach $O(n^{-1-2/d})$. It is of interest to see if the tools developed here can guide the choice of low-discrepancy sets or be generalized to analyze convergence rates of other commonly used low-discrepancy sets.

ACKNOWLEDGEMENT

Y. L. and W. W. contributed equally and are listed in alphabetical order. Y. L. is supported by China Scholarship Council. The research is partly supported by the Natural Science Foundation of China (Grant 11401338 and 11931001, KD PI), Beijing Academy of Artificial Intelligence (Supporting Grant, KD PI), and the National Science Foundation of USA (DMS-1712714 and DMS-1903139, JSL PI). The authors thank Pierre Jacob for useful discussions on particle filters and optimal transport, and the associate editor and reviewers for their valuable suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes all technical details.

REFERENCES

BERGMAN, N. (2001). Posterior Cramér-Rao bounds for sequential estimation. In *Sequential Monte Carlo methods in practice*. Springer, pp. 321–338.

BERGMAN, N., LJUNG, L. & GUSTAFSSON, F. (1999). Terrain navigation using Bayesian statistics. *IEEE Control Systems Magazine* **19**, 33–40.

BRINDLE, A. (1980). Genetic algorithms for function optimization .

CARPENTER, J., CLIFFORD, P. & FEARNHEAD, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation* **146**, 2–7.

DEL MORAL, P. (1997). Nonlinear filtering: Interacting particle resolution. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* **325**, 653–658.

DOUC, R. & CAPPÉ, O. (2005). Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005*. IEEE.

DOUCET, A., DE FREITAS, N. & GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*. Springer, pp. 3–14.

FEARNHEAD, P. & CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 887–899.

GATHERAL, J. (2011). *The volatility surface: a practitioner's guide*, vol. 357. John Wiley & Sons.

GERBER, M. & CHOPIN, N. (2015). Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 509–579.

GERBER, M., CHOPIN, N., WHITELEY, N. et al. (2019). Negative association, ordering and convergence of resampling methods. *The Annals of Statistics* **47**, 2236–2260.

GORDON, N. J., SALMOND, D. J. & SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, vol. 140. IET.

GRASSBERGER, P. (1997). Pruned-enriched Rosenbluth method: Simulations of θ polymers of chain length up to 1 000 000. *Physical Review E* **56**, 3682.

GUBERNATIS, J., KAWASHIMA, N. & WERNER, P. (2016). *Quantum Monte Carlo Methods*. Cambridge University Press.

GUSTAFSSON, F., GUNNARSSON, F., BERGMAN, N., FORSSELL, U., JANSSON, J., KARLSSON, R. & NORDLUND, P.-J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on signal processing* **50**, 425–437.

HAMMERSLEY, J. M. & MORTON, K. W. (1954). Poor man's Monte Carlo. *Journal of the Royal Statistical Society: Series B (Methodological)* **16**, 23–38.

HE, Z. & OWEN, A. B. (2016). Extensible grids: uniform sampling on a space filling curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 917–931.

HILBERT, D. (1935). Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis• Grundlagen der Mathematik• Physik Verschiedenes*. Springer, pp. 1–2.

HU, X.-L., SCHON, T. B. & LJUNG, L. (2008). A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing* **56**, 1337–1348.

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

KONG, A., LIU, J. S. & WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association* **89**, 278–288.

LIU, J. S. & CHEN, R. (1995). Blind deconvolution via sequential imputations. *Journal of the american statistical association* **90**, 567–576.

LIU, J. S. & CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association* **93**, 1032–1044.

REICH, S. (2013). A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing* **35**, A2013–A2024.

RIZZO, M. L. & SZÉKELY, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational statistics* **8**, 27–38.

ROSENBLUTH, M. N. & ROSENBLUTH, A. W. (1955). Monte Carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics* **23**, 356–359.

SAGAN, H. (2012). *Space-filling curves*. Springer Science & Business Media.

SIEPMANN, J. I. & FRENKEL, D. (1992). Configurational bias Monte Carlo: a new sampling scheme for flexible chains. *Molecular Physics* **75**, 59–70.

TAYLOR, S. J. (2008). *Modelling financial time series*. world scientific.

TILLÉ, Y. (2006). *Sampling algorithms*. Springer.

WEBBER, R. J. (2019). Unifying sequential Monte Carlo with resampling matrices. *arXiv preprint arXiv:1903.12583* .

WHITELEY, N., LEE, A., HEINE, K. et al. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22**, 494–529.