Stratification and Optimal Resampling for Sequential Monte Carlo

Yichao Li*[1], Wenshuo Wang*[2], Ke Deng[1], and Jun S Liu[2]

[1]Center for Statistical Science, Tsinghua University, Beijing 100084, China
[2]Department of Statistics, Harvard University, Cambridge, MA 02138

September 11, 2020

Abstract

Sequential Monte Carlo (SMC), also known as particle filters, has been widely accepted as a powerful computational tool for making inference with dynamical systems. A key step in SMC is resampling, which plays the role of steering the algorithm towards the future dynamics. Several strategies have been proposed and used in practice, including multinomial resampling, residual resampling (Liu and Chen 1998), optimal resampling (Fearnhead and Clifford 2003), stratified resampling (Kitagawa 1996), and optimal transport resampling (Reich 2013). We show that, in the one dimensional case, 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 the multidimensional case, the variance of stratified resampling after sorting particles using Hilbert curve (Gerber et al. 2019) in $\mathbb{R}^d$ is $O(m^{-(1+2/d)})$, an improved rate compared to the original $O(m^{-(1+1/d)})$, where m is the number of particles. This improved rate is the lowest 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 theoretical results, we propose the stratified multiple-descendant growth (SMG) algorithm, which allows us to explore the sample space more efficiently compared to the standard i.i.d. multiple-descendant sampling-resampling approach as measured by the Wasserstein metric. Numerical evidence is provided to demonstrate the effectiveness of our proposed method.

Keywords. Hilbert space-filling curve, particle filter, resampling, sequential Monte Carlo (SMC), stratification

1 Introduction

Sequential Monte Carlo has been studied intensively in the past two decades and applied broadly to high-dimensional statistical inference, signal processing, biology and many other fields (Liu and 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, or at least a certain aspect of it. The state-space model is a particularly interesting dynamic system that have been treated with sequential Monte Carlo. The model is governed by the hidden Markovian state equation and the noisy observation equation. The hidden state, for instance, can represent

*These authors contributed equally and are listed in alphabetical order.

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 sequential Monte Carlo is more commonly known as the particle filter in this context (Gordon et al. 1993).

Roughly speaking, sequential Monte Carlo is built based on sequential importance sampling, which recursively simulates a future state and reweighs the sampling path, with additional resampling steps (Liu and Chen 1998). In a vanilla sequential importance sampling procedure, such as sequential imputation (Kong et al. 1994), weight degeneracy arises as an inevitable problem. Since the importance weights are updated recursively at each step, stochastically most of the total weights will concentrate on a very few samples, leading to exponentially increasing variance (Kong et al. 1994). One effective strategy to avoid weight degeneracy is to resample from the current samples according to the corresponding weights. Resampling alone does not provide any information for estimation at the current step, but only introduces additional randomness. The main intuition behind resampling is that particles with small weights are deemed less hopeful and thus discarded so as to save resources in order to explore regions that may be more promising for the future (Liu and Chen 1995). Incidentally, in the bootstrap filter of Gordon et al. (1993), every forward simulation step is followed immediately with a resampling step without investigating its advantages and disadvantages. Liu and Chen (1995) provided a first attempt at analyzing resampling (termed as rejuvenation in that article), providing some useful insights, but was short of a rigorous theory.

Each iteration of sequential Monte Carlo consists of two steps: forward-sampling (or more intuitively, growth) and resampling. In the resampling step, we rejuvenate all the weights where samples with higher weights are more likely to be retained. In the growth step, we generate samples from the trial distribution and calculate the corresponding weight for each sample. Intuitively, the trial distribution should be as close to the target distribution as possible so as to explore the relevant part of the sample space.

There are various means to resample from a collection of weighted particles. The naïvest way to resample is called bootstrap resampling or multinomial resampling (Gordon et al. 1993), where the new particles are sampled from independent and identically distributed (i.i.d.) multinomial distributions based on the original particle weights. Residual resampling (Liu and Chen 1998) and stratified resampling (Kitagawa 1996) are two more popular resampling schemes in practice. Douc and 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 and 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, there has been no theoretical guarantee for the optimal transport resampling (aside from its validity), to the best of our knowledge. 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 and 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 sequential Monte Carlo significantly. Theoretically, however, the convergence rate in terms of the mean squared 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.

We focus on theoretical properties of various resampling schemes and sequential quasi-Monte

Carlo in this paper. We show that, in one dimension cases, 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 d dimensions, 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 particles. This improves the original 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 squared 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 and the tools can be modified to analyze other low-discrepancy sets as well.

The rest of the article is organized as follows. We provide some preliminaries, including relevant notations, definitions, and formulations, in Section 2. In Section 3, we prove the equivalence of several aforementioned resampling approaches in the one dimensional case. In Section 4, we give upper bounds for the resampling error of Hilbert curve resampling in terms of both variance and Wasserstein distance. In Section 5, we focus on exploring sequential quasi-Monte Carlo and derive a better convergence rate based on the theoretical results in Section 4. We wrap up the paper in Section 6 with some important open problems. All proofs are deferred to the supplement.

2 Preliminaries

2.1 Notations

We use superscript to denote the temporal notation (i.e., the step or iteration) and subscript 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 the trial distribution in the sense of importance sampling, which is constructed in a forward sampling (growth) fashion in sequential Monte Carlo. 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 weighed 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 $\mathbb{P}(Z = y_i) = p_i$, 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 $\{\pi_t(x^{(1:t)})\}_{t=1}^T$
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:
        $$W_j^{(t)} \propto \frac{\pi_t\left(X_j^{(1:t)}\right)}{\pi_{t-1}\left(X_j^{(1:t-1)}\right) g_t\left(X_j^{(t)} \mid X_j^{(1:t-1)}\right)}$$
    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 $X_j^{(1:t)} = \tilde{X}_j^{(1:t)}$.
    end
end
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 in 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 $w^{(t)} \propto w^{(t-1)} p_y(y^{(t)} \mid x^{(t)})$.

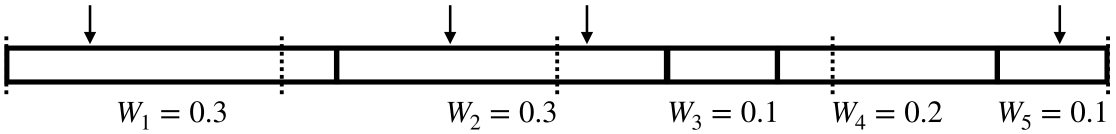


Figure 1: Illustration of stratified resampling. First line up the weights, then divide the interval into m equal parts, uniformly choose one point from each subinterval and record in which weight's region it lands. In the presented example where $m = 4$, $n = 5$, particles 1 and 5 are resampled once, 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

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

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 permutating 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 ϕ we have

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

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 slight abuse of notation. We call this collection of resampling methods matrix resampling methods. 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 Multinomial$(1, X, W)$. This corresponds to $p_{ij} = W_j$ for $i = 1, \dots, m$, $j = 1, \dots, n$, as shown in Figure 2(a). In stratified resampling, we let $U_i \sim \text{Unif}\left((i-1)/m, i/m\right]$, independently for $i = 1, \dots, m$, 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 $j = 1, \dots, n$; 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 $\mathbb{E}[\ell(\mathbb{P}, \tilde{\mathbb{P}}) \mid X, W]$ for a loss function ℓ. For example, we can pick $\ell(\mathbb{P}, \tilde{\mathbb{P}})$ to be $(\mathbb{E}_{\mathbb{P}}[\phi(X)] - \mathbb{E}_{\tilde{\mathbb{P}}}[\phi(X)])^2$ 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 energy distance between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as the square root of

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

$$\begin{pmatrix} 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.2 & 0.1 \end{pmatrix}$$

(a) Multinomial Resampling

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

(b) Stratified Resampling

$$\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \end{pmatrix}$$

(c) Multinomial Residual Resampling

$$\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ 0.2 & 0.2 & 0.4 & 0.2 & \\ & & & 0.6 & 0.4 \end{pmatrix}$$

(d) Stratified Residual Resampling

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

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, defined between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ as

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

where $\Gamma(\mathbb{P}_1, \mathbb{P}_2)$ denotes 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 X_j's, 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. Note that in this section, we consider the case where the particles take values in a one dimensional space. For example, resampling in a state-space model where the hidden state at each step is one-dimensional. In this case, we can focus on the last dimension of each particle, since the other components will not affect the future.

To study the stratified resampling matrix, we first define the staircase matrix. This will help with understanding 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 the following conditions are satisfied:*

(1) 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 similarly for the columns.

(2) 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 (allowing diagonal moves) from the top left entry to the bottom right entry.

Lemma 1. *For $m, n > 2$, there can only be one unique m by n staircase matrix that has non-negative entries and satisfies:*

$$\sum_{j=1}^{n} p_{ij} = r_i > 0 \text{ and } \sum_{i=1}^{m} p_{ij} = c_j > 0$$

By Lemma 1, the staircase resampling matrix is unique given the weights for each particles. Then we can define a stratified resampling matrix.

Definition 2 (Stratified resampling matrix)**.** *We call a matrix $P^{SR}_{m,W} \in \mathcal{P}_{m,W}$ the stratified resampling matrix of a set of weighted particles $(X_j, W_j)_{j=1}^n$ if $P^{SR}_{m,W}$ 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 $P^{SR}_{m,W}$ minimizes the following objectives:*

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

(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 $\mathbb{E}[\phi(X)]$, then ordering the states by function ϕ and then applying stratified resampling gives the minimum variance. This result is noted in Webber (2019), although it appears that only a proof that ordered stratified sampling gives a local maximum is provided. We build upon their idea and offer a detailed proof that it is indeed the global maximum.*

4 Error of ordered stratified resampling

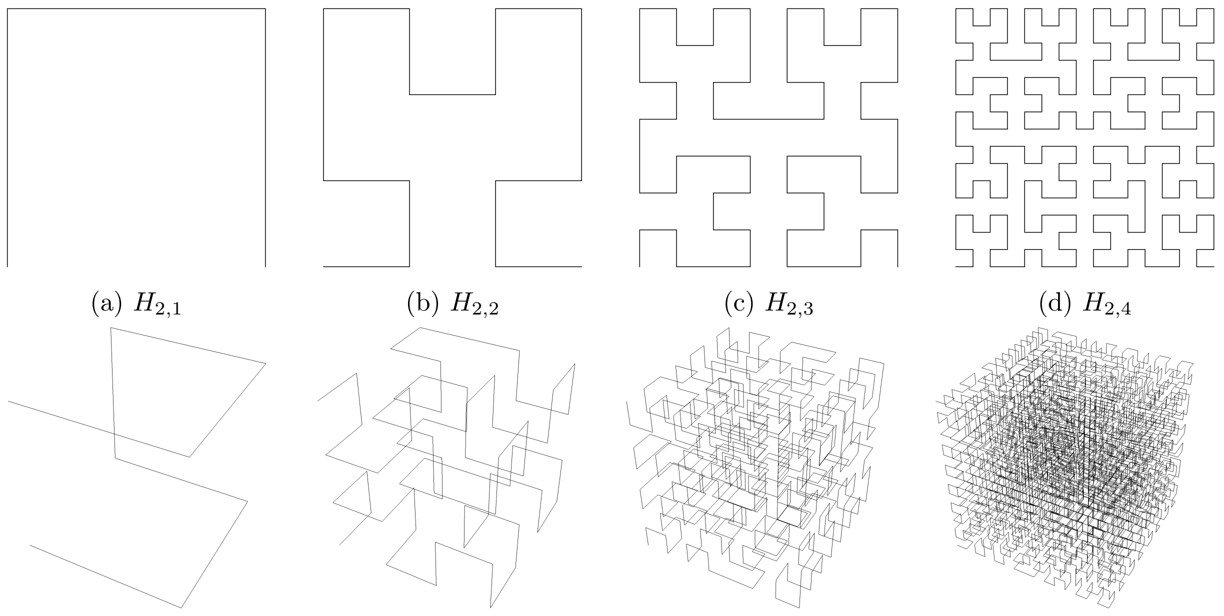
In this section, we analyze the error induced by ordered stratified resampling.

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

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

We here provide some intuition behind ordered stratified sampling. Since the new particles are sampled independently, we only need to make sure that each new particle brings in little randomness. It is easy to see from the staircase structure of the resampling matrix that each X_i^* takes value in a sequence of consecutive X_j's. Since the original particles have been ordered, this sequence of X_j's are close to each other in the space. Together with the fact that ϕ is Lipschitz, we see that for each i, $\phi(\tilde{X}_i)$ is bounded in a small region.

In multiple dimensions, it has been noticed that the Hilbert space-filling curve (Hilbert 1935) can help lower the sampling variance (Gerber and Chopin 2015; He and Owen 2016; 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) $H_{3,4}$

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

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 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.

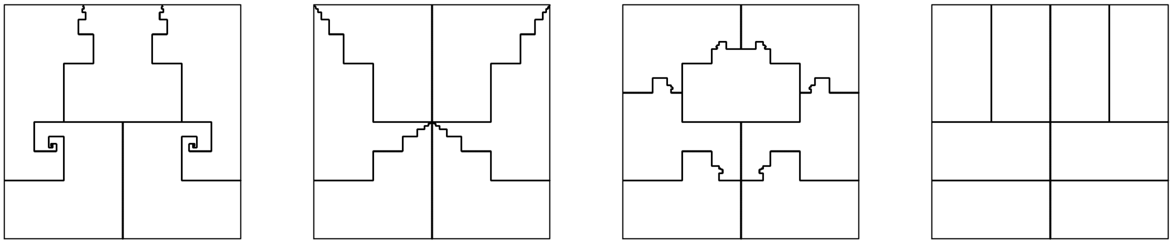
(ii) H is Hölder continuous with exponent $1/d$ (He and Owen 2016):

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

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

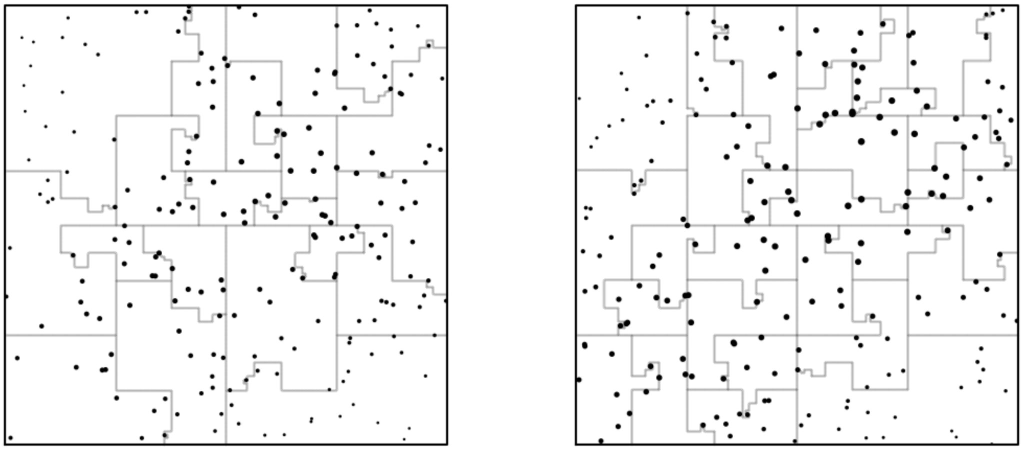
The Hilbert curve is defined as the limit of a sequence of curves; see Figure 3 for an illustration in two and three dimensions. Many software packages can efficiently convert between x and $H(x)$ (e.g., the Python package hilbertcurve). We omit here the rigorous definition of Hilbert curves and refer 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



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

Figure 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 $m = 30$.

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

the one-dimensional case, each new 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 in Gerber et al. (2019).

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 an ascending order by the value of $h(X_j)$, then stratified sampling satisifies*

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

Remark 2. *The exponent $1+2/d$ in the theorem improves the original rate $1+1/d$ shown in Gerber et al. (2019). It is conjectured in Gerber et al. (2019) 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 vector 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 $[0,1]$, $d > 1$. (ref1 minor-com12)

Moreover, we show in Proposition 1 that no other ordering rule can improve the exponent $1+2/d$.

Proposition 1. *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 space.

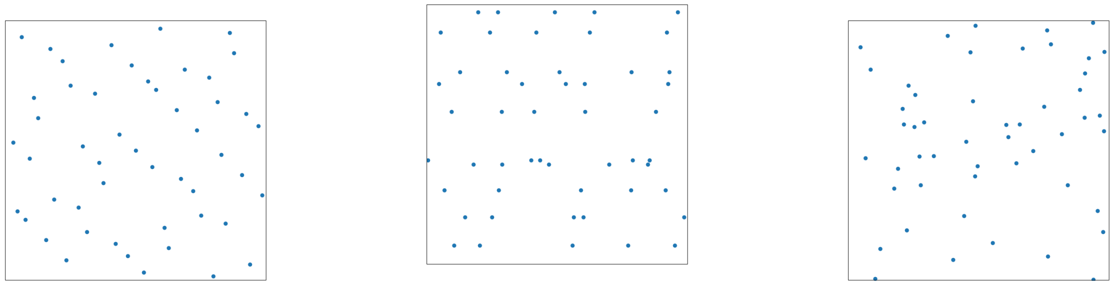
Theorem 4. *Under d-dimensional Hilbert curve resampling, $d \geq 1$, the Wasserstein distance W_p between $\tilde{\mathbb{P}} = \sum_{i=1}^{m} m^{-1}\delta_{\tilde{X}_i}$ and $\mathbb{P} = \sum_{j=1}^{n} W_j \delta_{X_j}$ is almost surely upper bounded by $2\sqrt{d+3}m^{-\frac{1}{\max(p,d)}}$.*

5 Mean square error of sequential quasi-Monte Carlo

5.1 Sequential quasi-Monte Carlo

In this section, we discuss how to utilize the previous results to obtain a new convergence rate for the sequential quasi-Monte Carlo proposed in Gerber and Chopin (2015), which can be structured identically as 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 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 that $u_{1:n}$ are in ascending order. Intuitively speaking, a low-discrepancy set is a set that spreads evenly



(a) Sobol' sequence.

(b) Stratified multiple-descendant growth.

(c) Independent sampling.

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

in $[0,1]^{1+d}$; see Gerber and Chopin (2015) for a more detailed discussion. Sequential quasi-Monte Carlo combines resampling and growth by defining

$$X_j^{(t)} = \begin{cases} \Gamma_1(v_j), & t = 1, \\ \Gamma_t(X^{(t-1)}_{\sigma(u_j, W^{(t-1)}_{1:n})}, v_j), & 2 \le t \le T, \end{cases}, \quad 1 \le j \le n.$$

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 and 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 $(X^{(t)}_{\sigma(u_j, W^{(t)}_{1:n})})_{j=a}^{b}$ 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 at the deterministic or semi-deterministic nature of the set $U^{(t)}$. We exploit this intuition and construct a specific set that enables more careful convergence analysis.

Let $n = sr$, and let $U_k \sim \mathrm{Unif}[(k-1)/s, k/s]$ be independent for $1 \le k \le s$. 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 $1 \le k \le s$, $1 \le \ell \le r$. We define $U^{(t)}_{\mathrm{SMG}} = \{(U_{\lfloor i/r \rfloor + 1}, V_i) : 1 \le i \le n\}$. 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 sampling in the discrete space (Fearnhead and Clifford 2003). Figure 6 compares the discrepancy set generated by stratified multiple-descendant growth and two other approaches.

We focus on the mean square error of the estimation of any proper function ϕ in a state-space model, which is defined as

$$\mathrm{MSE}_t(\phi) = \mathbb{E}\left[\frac{\sum_{j=1}^n W_j^{(t)} \phi(X_j^{(t)})}{\sum_{j=1}^n W_j^{(t)}} - \int \pi_t(x^{(1:t)}) \phi(x^{(t)}) dx^{(1:t)}\right]^2.$$

The mean square error can be decomposed into the squared bias and variance. The following theorem gives a bound for each one, respectively.

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* $U^{(t)}_{SMG}$*. Assume that each* $X^{(t)}$ *falls in a compact set, assuming to be* $\mathcal{X} = [0,1]^d$ *without loss of generality. Suppose* $(X^{(t)}_j, W^{(t)}_j)_{1\le j\le n}$ *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* $[-M,M]$*,*

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

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

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

6 Discussion

We have discussed how one might improve the performance of SMC and SQMC 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 and Clifford (2003)). Let $q_{1:m}$ satisfy $q_i \ge 0$ 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 independent from each other conditional on the original particles, the resampling method cannot be represented as a resampling matrix. Systematic resampling (Carpenter et al. 1999) is such an example. All criteria mentioned in Section 2.4 are also well-defined for non-matrix resampling. It would be interesting to study a broader class of resampling methods including some non-matrix resampling schemes.

Finally, it will be interesting to see if the tools in this paper can guide the choice of low-discrepancy sets or be generalized to analyze other existing low-discrepancy sets more commonly used in practice and show they achieve the same or better convergence rates. In fact, it was conjectured that the optimal convergence rate can reach $O(n^{-1-2/d})$ (Gerber and Chopin 2015).

Acknowledgements

Y. L. is supported by China Scholarship Council. The research is partly supported by the National Natural Science Foundation of China (Grant 11401338, DK PI), Beijing Academy of Artificial Intelligence (Supporting Grant, DK PI), and the National Science Foundation of USA (DMS-1712714 and DMS-1903139, LJ PI). The authors would like to thank Pierre Jacob for useful discussions on particle filters and optimal transport.

References

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

Bergman, N., Ljung, L., and Gustafsson, F. (1999). Terrain navigation using Bayesian statistics. *IEEE Control Systems Magazine*, 19(3):33–40.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7.

Douc, R. and 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.*, pages 64–69. IEEE.

Doucet, A., De Freitas, N., Gordon, N., and others, a. (2001). Sequential Monte Carlo methods in practice.

Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899.

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

Gerber, M. and Chopin, N. (2015). Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579.

Gerber, M., Chopin, N., Whiteley, N., et al. (2019). Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET.

Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., and Nordlund, P.-J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on signal processing*, 50(2):425–437.

He, Z. and Owen, A. B. (2016). Extensible grids: uniform sampling on a space filling curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):917–931.

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

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.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288.

Liu, J. S. and Chen, R. (1995). Blind deconvolution via sequential imputations. *Journal of the american statistical association*, 90(430):567–576.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044.

Popoviciu, T. (1935). Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145.

Reich, S. (2013). A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024.

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

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

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

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

A Proofs

Proof of Lemma 1. Suppose $P = (p_{ij})_{m\times n}$ and $Q = (q_{ij})_{m\times n}$ are both eligible staircase matrices. If $p_{11} \neq q_{11}$, without loss of generality, assume $p_{11} < q_{11}$, then $\sum_{j=2}^{n} p_{1j} = r_1 - p_{11} > r_1 - q_{11} \geq 0$. By condition (1) in the definition of staircase matrix, $p_{12} > 0$. This actually implies that $p_{i1} = 0$ for all $i > 1$. However, $p_{11} = \sum_{i=1}^{m} p_{i1} = \sum_{i=1}^{m} q_{i1} \geq q_{11} > p_{11}$, which is a contradiction.

Then consider p_{12} and q_{12}, suppose $0 \leq p_{12} < q_{12}$, then $\sum_{j=3}^{n} = p_{1j} = r_1 - p_{11} - p_{12} > r_1 - q_{11} - q_{12} \geq 0$. By condition (1) in the definition of staircase matrix, $p_{13} > 0$. This implies that $p_{i2} = 0$ for all $i > 1$. Similarly, $p_{12} = \sum_{i=1}^{m} p_{i2} = \sum_{i=1}^{m} q_{i1} \geq q_{12} > p_{12}$, which is a contradiction. Similarly, we can prove that $p_{1j} = q_{1j}$ for each $j = 1, 2, \cdots, n$. By induction, $P = Q$. □

Proof of Theorem 1. First, we prove the following lemma.

Lemma 2. *Suppose P is an m by n matrix with $m, n > 2$, $\sum_{i=1}^{n} p_{ij} > 0$ for all j, and $\sum_{j=1}^{n} p_{ij} > 0$ for all i, then in Definition 1,* (2) *implies* (1).

Proof of Lemma 2. We consider the rows, and the same proof applies to the columns. Suppose $p_{ij_1} \neq 0$ and $p_{ij_2} \neq 0$, $j_1 < j_2$, for j such that $j_1 < j < j_2$, if $p_{ij} = 0$, because $\sum_{s=1} p_{sj} > 0$, there is a k such that $p_{kj} > 0$. If $k < i$, then (k, j_1, i, j) is an ineligible quadruplet that contradicts (2). If $k > i$, then (i, j, k, j_2) is an ineligible quadruplet that contradicts (2). □

(i) Conditional variance.

Suppose P maximizes $t(P) = X^\top P^\top P X$ and $\sum_{j=1}^{n} p_{ij} X_j$ is ascending with respect to i (note that permutation of rows in P doesn't change the value of $X^\top P^\top P X$). Consider a quadruplet (i, j, k, l) such that $i < k$ and $j < l$. If $p_{il} > 0$ and $p_{kj} > 0$, set $\alpha = \min\{p_{il}, p_{kj}\} > 0$, then update the entries of P as:

$$p_{ij} \leftarrow p_{ij} + \alpha \qquad p_{il} \leftarrow p_{il} - \alpha$$
$$p_{kj} \leftarrow p_{kj} - \alpha \qquad p_{kl} \leftarrow p_{kl} + \alpha$$

We name the updated weight matrix as P', then

$$\begin{aligned} t(P') - t(P) &= (\sum_{s=1}^{n} X_s p_{is} + \alpha(X_j - X_l))^2 + (\sum_{s=1}^{n} X_s p_{ks} + \alpha(-X_j + X_l))^2 - \sum_{s=1}^{n}(X_s p_{is})^2 - \sum_{s=1}^{n}(X_s p_{ks})^2 \\ &= 2\alpha^2 (X_j - X_l)^2 + 2\alpha(X_j - X_l)(\sum_{s=1}^{n} X_s p_{is} - \sum_{s=1}^{n} X_s p_{ks}) > 0, \end{aligned}$$

since $X_j < X_l$ and $\sum_{s=1}^{n} X_s p_{is} \leq \sum_{s=1}^{n} X_s p_{ks}$. This would contradict the fact that P maximizes $t(P)$. Hence, by Lemma 2, P is a staircase matrix.

(ii) Expected squared energy distance.

Note that the squared energy distance admits an explicit expression in one dimension. By some algebra, we find that Lemma 3 enables us to convert the problem of minimizing expected squared energy distance to a simpler problem.

Lemma 3. *In the setting of Theorem 1, the solution to the following optimization problems minimizes the expected squared energy distance:*

$$\underset{P \in \mathcal{P}_{m,W}}{\arg\max} \sum_{k=1}^{n-1} \left[(X_{k+1} - X_k) \sum_{i=1}^{m} \left(\sum_{j=1}^{k} p_{ij} \right)^2 \right].$$

Back to the proof the theorem, let $P_{m,W}^{\mathrm{SR}} = (p_{ij}^*)$ be the ordered stratified resampling matrix. We will prove that for any k and any $P = (p_{ij}) \in \mathcal{P}_{m,W}$,

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}^*\right)^2 \geq \sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}\right)^2 .$$

The result then follows from Lemma 3. Since $\sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}\right) = m\sum_{j=1}^{k} W_j$ and $0 \leq \sum_{j=1}^{k} p_{ij} \leq 1$, the sum of squares attains its maximum when $[m\sum_{j=1}^{k} W_j]$ of them are 1, one of them is $m\sum_{j=1}^{k} W_j - [m\sum_{j=1}^{k} W_j]$, and the rest are 0. It can be easily checked that (p_{ij}^*) satisfies this condition and thus solves the optimization problem.

Proof of Lemma 3. Let $d(\mathbb{P}, \tilde{\mathbb{P}}) = \int_{-\infty}^{\infty} (F_{\mathbb{P}}(x) - F_{\tilde{\mathbb{P}}}(x))^2\, dx$, which is equal to half the squared energy distance (Székely 2003)

$$\mathbb{E}|X - Y| - \frac{\mathbb{E}|X - X'| + \mathbb{E}|Y - Y'|}{2},$$

with X, X', Y, Y' independent, X, X' coming from $\mathbb{P}$ and Y, Y' coming from $\tilde{\mathbb{P}}$. Since the X_j's are ordered as $X_1 < X_2 < \cdots < X_n$, we have

$$\begin{aligned}\mathbb{E}[d(\mathbb{P}, \tilde{\mathbb{P}}) \mid X, W] &= \int_{-\infty}^{\infty} \mathbb{E}[(F_{\mathbb{P}}(x) - F_{\tilde{\mathbb{P}}}(x))^2 \mid X, W]\, dx \\ &= \int_{-\infty}^{\infty} (\mathbb{E}[F_{\tilde{\mathbb{P}}}(x)^2 \mid X, W] - F_{\mathbb{P}}(x)^2)\, dx\,.\end{aligned} \tag{2}$$

Note that

$$\begin{aligned}\mathbb{E}[F_{\tilde{\mathbb{P}}}(x)^2 \mid X, W] &= \frac{1}{m^2}\mathbb{E}[(\#\{i : \tilde{X}_i \leq x\})^2 \mid X, W] \\ &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{k} p_{ij} + \frac{1}{m^2}\sum_{i\neq l}\left(\sum_{j=1}^{k} p_{ij}\right)\left(\sum_{j=1}^{k} p_{lj}\right) \\ &= \underbrace{\frac{1}{m}\sum_{j=1}^{k} W_j}_{\text{constant}} + \frac{1}{m^2}\sum_{i\neq l}\left(\sum_{j=1}^{k} p_{ij}\right)\left(\sum_{j=1}^{k} p_{lj}\right), X_k \leq x < X_{k+1}.\end{aligned}$$

Minimizing equation (2) now becomes minimizing

$$\begin{aligned}&\sum_{k=1}^{n-1}(X_{k+1} - X_k)\left[\sum_{i\neq l}\left(\sum_{j=1}^{k} p_{ij}\right)\left(\sum_{j=1}^{k} p_{lj}\right)\right] \\ &= \sum_{k=1}^{n-1}(X_{k+1} - X_k)\left\{\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}\right)\right]^2 - \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}\right)^2\right]\right\} \\ &= \sum_{k=1}^{n-1}(X_{k+1} - X_k)\left\{\left[m\left(\sum_{j=1}^{k} W_j\right)\right]^2 - \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{k} p_{ij}\right)^2\right]\right\},\end{aligned}$$

which, after discarding constants, simplifies to maximizing

$$\sum_{k=1}^{n-1}(X_{k+1}-X_k)\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{k}p_{ij}\right)^2\right].$$

□

(iii) Earth mover's distance.

Let $t(P)=\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}\ell(Y_i-X_j)$. Let P be the matrix that minimizes $t(P)$. Consider a quadruplet (i,j,k,l) such that $i<k$ and $j<l$. If $p_{il}>0$ and $p_{kj}>0$, set $\alpha=\min\{p_{il},p_{kj}\}>0$, then update the entries of P as:

$$\begin{aligned} p_{ij} &\leftarrow p_{ij}+\alpha & p_{il} &\leftarrow p_{il}-\alpha \\ p_{kj} &\leftarrow p_{kj}-\alpha & p_{kl} &\leftarrow p_{kl}+\alpha \end{aligned}$$

We name the updated weight matrix as P', then

$$t(P')-t(P)=\alpha(\ell(Y_i-X_j)+\ell(Y_k-X_l)-\ell(Y_i-X_l)-\ell(Y_k-X_j)).$$

Since ℓ is convex and

$$\begin{gathered}(Y_i-X_j)+(Y_k-X_l)=(Y_i-X_l)+(Y_k-X_j)\\ |(Y_i-X_j)-(Y_k-X_l)|<|(Y_i-X_l)-(Y_k-X_j)|\end{gathered}$$

we have

$$\ell(Y_i-X_j)+\ell(Y_k-X_l)<\ell(Y_i-X_l)+\ell(Y_k-X_j)),$$

so $t(P')<t(P)$. This would contradict the fact that P is the minimizer, so such a quadruplet does not exist. By Lemma 2, the solution P is a staircase matrix.

□

Proof of Theorem 2. Without loss of generality, suppose $X_1<X_2<\cdots<X_n$ and P is a staircase weight matrix corresponding to stratified resampling. Each $\tilde{X}_i$ can only take values in $X_{il},X_{il+1},\cdots,X_{ir}$, with

$$X_{il}\le X_{ir} \text{ and } X_{ir}\le X_{i+1,l},$$

for $1\le i\le m-1$.

Hence,

$$\begin{aligned}\operatorname{Var}\left[\frac{1}{m}\sum_{i=1}^{m}\phi(\tilde{X}_i)\mid X,W\right] &= \frac{1}{m^2}\sum_{i=1}^{m}\operatorname{Var}\left[\phi(\tilde{X}_i)\mid X,W\right]\\ &\le \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{4}\max_{x,y\in[X_{ir},X_{il}]}(\phi(x)-\phi(y))^2\\ &\qquad\qquad\text{(Popoviciu's inequality on variances)}\\ &\le \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{4}\max_{x,y\in[X_{ir},X_{il}]}L_\phi^2(x-y)^2=\frac{L_\phi^2}{4m^2}\sum_{i=1}^{m}(X_{ir}-X_{il})^2\\ &\le \frac{L_\phi^2}{4m^2}(X_n-X_1)\sum_{i=1}^{m}(X_{ir}-X_{il})=\frac{L_\phi^2}{4m^2}(X_n-X_1)^2.\end{aligned}$$

The Popoviciu's inequality (Popoviciu 1935) on variances is stated as following:

Lemma 4 (Popoviciu's inequaltity on variances)**.** *Let M and m be the upper bound and lower bound of a random variable X, respectively. Then,*

$$\mathrm{Var}(X) \leq (M-m)^2/4$$

□

Proof of Theorem 3. First note that $H(x)$ is Hölder continuous with exponent $1/d$,

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

With Hilbert curve stratified sampling, $\tilde{X}_i$ can only take values in $X_{il}, X_{il+1}, \cdots, X_{ir}$, with

$$h(X_1) = h(X_{1l}) \leq \cdots \leq h(X_{i-1,r}) \leq h(X_{il}) \leq h(X_{ir}) \leq h(X_{i+1,l}) \leq \cdots \leq h(X_{nr}) = h(X_n).$$

Note that

$$\begin{aligned}
&\mathrm{Var}_P\left[\frac{1}{m}\sum_{i=1}^m \phi(\tilde{X}_i) \mid X\right] = \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}[\phi(\tilde{X}_i) \mid X] = \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}[\phi(H(h(\tilde{X}_i))) \mid X] \\
&\leq \frac{1}{4m^2}\sum_{i=1}^m \left(\max_{x:h(x)\in[h(X_{il}),h(X_{ir})]} \phi(x) - \min_{x:h(x)\in[h(X_{il}),h(X_{ir})]} \phi(x)\right)^2 \text{ (Popoviciu's inequality on variances)} \\
&= \frac{1}{4m^2}\sum_{i=1}^m \left(\max_{y\in[h(X_{il}),h(X_{ir})]} \phi(H(y)) - \min_{y\in[h(X_{il}),h(X_{ir})]} \phi(H(y))\right)^2 \\
&= \frac{1}{4m^2}\sum_{i=1}^m \max_{y_1,y_2\in[h(X_{il}),h(X_{ir})]} \|\phi(H(y_1)) - \phi(H(y_2))\|^2 \\
&\leq \frac{1}{4m^2}\sum_{i=1}^m \max_{y_1,y_2\in[h(X_{il}),h(X_{ir})]} L_\phi^2\|H(y_1) - H(y_2)\|^2 \\
&\leq \frac{L_\phi^2}{4m^2}\sum_{i=1}^m \max_{y_1,y_2\in[h(X_{il}),h(X_{ir})]} 4(d+3)|y_1 - y_2|^{2/d} \\
&= \frac{(d+3)L_\phi^2}{m^2}\sum_{i=1}^m (h(X_{ir}) - h(X_{il}))^{2/d} \\
&\leq \frac{(d+3)L_\phi^2}{m^2}\left[\sum_{i=1}^m ((h(X_{ir}) - h(X_{il}))^{2/d})^{d/2}\right]^{2/d} m^{1-2/d} \text{ (Hölder inequality)} \\
&= \frac{(d+3)L_\phi^2 m^{1-2/d}}{m^2}(h(X_m) - h(X_1))^{2/d} \leq \frac{(d+3)L_\phi^2}{m^{1+2/d}}.
\end{aligned}$$

□

Proof of Proposition 1. We will prove that for all $n = 2^{kd}$, where k is a positive integer and $kd > 3$, there exists $\phi \in \Phi_d$, W and X such that

$$\mathrm{Var}_P(\frac{1}{n}\sum_{i=1}^n \phi(\tilde{X}_i) \mid X, W) \geq \frac{1}{27d}\frac{1}{n^{1+2/d}}.$$

Let

$$\mathcal{L}_k = \left\{0, \frac{1}{2^k}, \cdots, \frac{2^k-1}{2^k}\right\}^d$$

be an equally spaced grid of $[0,1]^d$. Let $X = (X_1, X_2, \cdots, X_{2^{dk}})$ be the sequence of points in $\mathcal{L}_k$ ordered by o. Suppose

$$W = (W_1, \cdots, W_{2^{dk}}) \propto (\underbrace{1, \cdots, 1}_{2^{kd-1}}, \underbrace{2, \cdots, 2}_{2^{kd-1}}).$$

The stratified resampling matrix is

$$P = \mathrm{diag}\{\underbrace{P_1, \cdots, P_1}_{(2^{dk-1}-2)/3}, P_2, \underbrace{P_3, \cdots, P_3}_{(2^{dk-1}-2)/3}\},$$

where

$$P_1 = \begin{pmatrix} 2/3 & 1/3 & \\ & 1/3 & 2/3 \end{pmatrix},$$

$$P_2 = \begin{pmatrix} 2/3 & 1/3 & & \\ & 1/3 & 2/3 & \\ & & 2/3 & 1/3 \\ & & & 1 \end{pmatrix},$$

$$P_3 = \begin{pmatrix} 1 & & \\ 1/3 & 2/3 & \\ & 2/3 & 1/3 \\ & & 1 \end{pmatrix}.$$

Let $\phi_k(X = (x_1, \cdots, x_d)) = x_k$ be the function that returns the kth coordinate, $k = 1, 2, \cdots, d$. It is easy to see that ϕ_k is 1-Lipschitz. We prove a simple lemma below.

Lemma 5. *If Z is a random variable defined by*

$$Z = \begin{cases} x, & \textit{with probability } 1/3, \\ y & \textit{with probability } 2/3, \end{cases}$$

where x and y are distinct points in $\mathcal{L}_k$, then $\mathrm{Var}(\phi_k(Z)) \geq \dfrac{2^{-2k+1}}{9}$ *for at least one $k \in \{1, 2, \ldots, d\}$.*

Proof of Lemma 5. By direct calculation, $\mathrm{Var}(\phi_k(Z)) = \dfrac{2}{9}(x_k - y_k)^2$. Since $x \neq y$, at least one k satisfies $|x_k - y_k| \geq 2^{-k}$. □

Now the resampling variance is

$$\begin{aligned}
&\sum_{k=1}^{d}\frac{1}{m^2}\operatorname{Var}_P\left[\sum_{i=1}^{m}\phi_k(\tilde{X}_i)\mid X,W\right]\\
=&\frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=1}^{d}\operatorname{Var}_P\left[\phi_k(\tilde{X}_i)\mid X,W\right]\\
\geq&\frac{1}{m^2}\sum_{i=1}^{(2^{dk}-4)/3}\sum_{k=1}^{d}\operatorname{Var}_P\left[\phi_k(\tilde{X}_i)\mid X,W\right]\\
\geq&\frac{1}{m^2}\sum_{i=1}^{(2^{dk}-4)/3}\frac{2^{-2k+1}}{9}\\
=&\frac{1}{2^{2dk}}\frac{2^{dk}-4}{3}\frac{2^{-2k+1}}{9}\\
\geq&\frac{1}{2^{2dk}}\frac{2^{dk-1}}{3}\frac{2^{-2k+1}}{9}\ \text{(when } dk\geq 3)\\
=&\frac{1}{27}m^{-1-2/d}.
\end{aligned}$$

Hence, there exists at least one $k\in\{1,2,\cdots,d\}$, such that

$$\frac{1}{m^2}\operatorname{Var}_P\left[\sum_{i=1}^{m}\phi_k(\tilde{X}_i)\mid X,W\right]\geq\frac{1}{27d}\frac{1}{m^{1+2/d}}.$$

□

Proof of Theorem 4. We define a coupling between $Y\sim\mathbb{P}=\sum_{j=1}^{n}W_j\delta_{X_j}$ and $\tilde{Y}\sim\tilde{\mathbb{P}}=\sum_{i=1}^{m}\frac{1}{m}\delta_{\tilde{X}_i}$ by letting $(Y,\tilde{Y})=(X_J,\tilde{X}_I)$, where $P(I=i,J=j)=p_{ij}/m$ and p_{ij} is the (i,j)-entry of the Hilbert curve resampling matrix P. Recall that with Hilbert curve stratified sampling, $\tilde{X}_i$ can only take values in $X_{il},X_{il+1},\cdots,X_{ir}$, with

$$h(X_{il})\leq h(X_{ir})\text{ and }h(X_{ir})\leq h(X_{i+1,l}),$$

for $1\leq i\leq n$.

$$\begin{aligned}
\mathbb{E}[\|Y-\tilde{Y}\|^p]&=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{m}p_{ij}\|\tilde{X}_i-X_j\|^p\\
&\leq\frac{1}{m}\sum_{i=1}^{m}\max_{z,z'\in[h(X_{il}),h(X_{ir})]}\|H(z)-H(z')\|^p\\
&\leq\frac{1}{m}\sum_{i=1}^{m}(2\sqrt{d+3}(h(X_{ir})-h(X_{il}))^{1/d})^p\\
&\leq\begin{cases}2^p(d+3)^{p/2}m^{-p/d}, & \text{if } p\leq d,\\ \dfrac{2^p(d+3)^{p/2}}{m}, & \text{if } p>d.\end{cases}
\end{aligned}$$

Thus,

$$W_p(\mathbb{P}^*,\mathbb{P})\leq\frac{2\sqrt{d+3}}{m^{1/\max(p,d)}},\quad a.s.$$

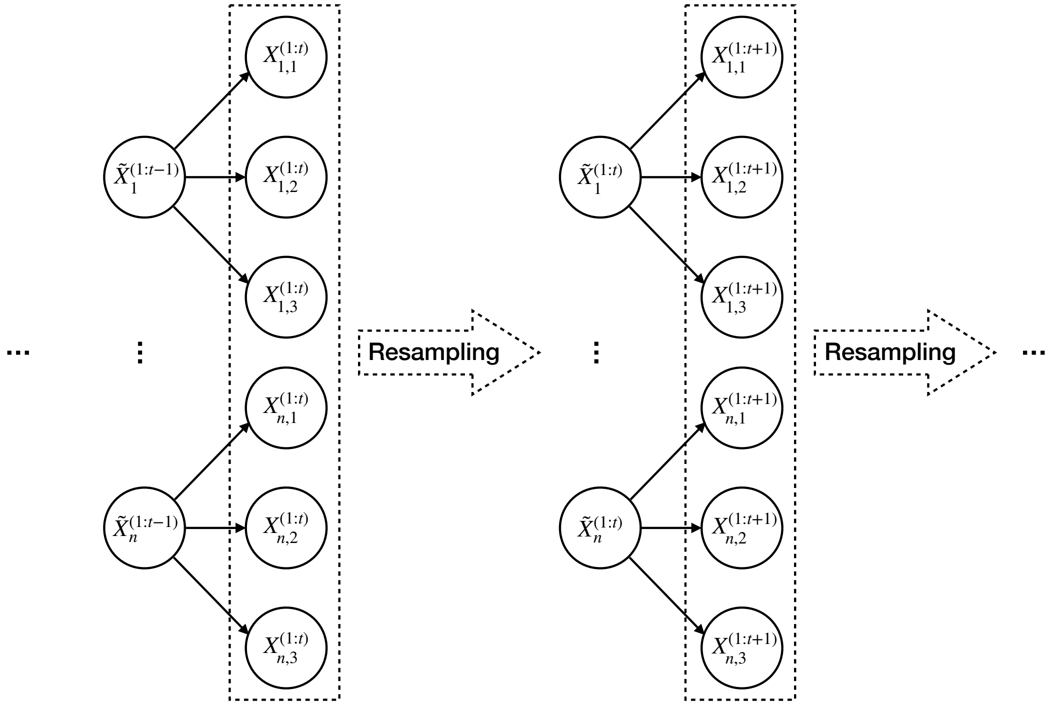


Figure 7: Illustration of multiple-descendant growth.

□

Proof of Theorem 5. We introduce some notations for presentational convenience to reflect the multiple-descendant nature. Let $\tilde{X}_k^{(t-1)} = X^{(t-1)}_{\sigma(U_k, W^{(t-1)}_{1:n})}$ corresponds to the resampled particle (recall the definition of U_k for $U^{(t)}_{\mathrm{SMG}}$) for $t \geq 2$, and let

$$X^{(t)}_{k\ell} = X^{(t)}_{k(s-1)+\ell} = \begin{cases} \Gamma_1(v_{k(s-1)+\ell}), & t = 1, \\ \Gamma_t(\tilde{X}_k^{(t-1)}, v_{k(s-1)+\ell}), & 2 \leq t \leq T, \end{cases}$$

be the ℓth descendant of $\tilde{X}_k^{(t-1)}$. Similarly, $W^{(t)}_{k\ell} = W^{(t)}_{k(s-1)+\ell}$. See Figure 7 for an illustration.

We then introduce two lemmas.

Lemma 6. *Under the assumptions of Theorem 5,*

$$\mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W^{(t)}_{k\ell}\phi(X^{(t)}_{k\ell}) \mid X^{(1:t-1)}, W^{(t-1)}\right] = O(n^{-1-\frac{4}{d(d+4)}}).$$

Lemma 7. *Under the assumptions of Theorem 5,*

$$\mathrm{Var}\left[\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W^{(t)}_{k\ell}\phi(X^{(t)}_{k\ell})}{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W^{(t)}_{k\ell}} \mid X^{(1:t-1)}, W^{(t-1)}\right] = O(n^{-1-\frac{4}{d(d+4)}}).$$

We prove by induction. For $t = 1$,

$$\begin{aligned}
&\left(E\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}\phi(X_{ik}^{(1)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}} - \int_{\mathcal{X}} \pi_1(x^{(1)})\phi(x^{(1)})dx^{(1)}\right]\right)^2 \\
&= \left(E\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}\phi(X_{ik}^{(1)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}} - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}\phi(X_{ik}^{(1)})}{sr}\right]\right)^2 \\
&= \left(E\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}\phi(X_{ik}^{(1)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}\left(1 - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}{sr}\right)\right]\right)^2 \\
&\le E\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}\phi(X_{ik}^{(1)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}\left(1 - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}{sr}\right)\right]^2 \\
&\le M^2\mathbb{E}\left[\left(1 - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}{sr}\right)^2\right] \\
&= M^2 \operatorname{Var}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{ik}^{(1)}}{sr}\right),
\end{aligned} \tag{3}$$

since $n^{-1}\mathbb{E}\left(\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(1)}\right) = 1$. The variance is $O(n^{-1-4/[d(d+4)]})$ by the same analysis as part A in the proof of Lemma 6.

Now suppose we have proved the cases from 1 to $t-1$. Note that in state-space models, $h_t(X_{k\ell}^{(t-1)}, x) = \pi_t\left(\left(\tilde{X}_{k\ell}^{(1:t-1)}, x\right)\right)/\pi_{t-1}\left(X_{k\ell}^{(1:t-1)}\right)$ is only a function of x and $X_{k\ell}^{(t-1)}$.

$$\begin{aligned}
&\left(\mathbb{E}\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr} - \int \pi_t(x^{(1:t)})\phi(x^{(t)})dx^{(1:t)}\right]\right)^2 \\
&= \left(\frac{1}{s}\mathbb{E}\sum_{k=1}^{s}\mathbb{E}\left[\int_{\mathcal{X}} \frac{\pi_t\left(\left(\tilde{X}_k^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_k^{(1:t-1)}\right)}\phi(x)\,dx \mid X_{k\ell}^{(1:t-1)}, W_{k\ell}^{(t-1)}\right]\right. \\
&\quad\left. - \int \frac{\pi_t(x^{(1:t)})}{\pi_{t-1}(x^{(1:t-1)})}\pi_{t-1}(x^{(1:t-1)})\phi(x^{(t)})dx^{(1:t)}\right)^2 \\
&= \left(\mathbb{E}\sum_{k=1}^{s}\sum_{\ell=1}^{r}\frac{W_{k\ell}^{(t-1)}}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t-1)}}\int_{\mathcal{X}} h_t(X_{k\ell}^{(t-1)}, x)\phi(x)\,dx - \int \pi_{t-1}(x^{(1:t-1)})h_t(x^{(t-1)}, x^{(t)})\phi(x^{(t)})dx^{(1:t)}\right)^2 \\
&= \left(\mathbb{E}\left[\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t-1)}\tilde{\phi}_t(X_{k\ell}^{(t-1)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t-1)}}\right] - \int \pi_{t-1}(x^{(1:t-1)})\tilde{\phi}_t(x^{(t-1)})dx^{(1:t-1)}\right)^2 = O(n^{-1-\frac{4}{d(d+4)}}),
\end{aligned} \tag{4}$$

by induction hypothesis, where

$$\tilde{\phi}_t(x) = \int_{\mathcal{X}} h_t(x, u)\phi(u)\,du$$

is bounded and Lipschitz since ϕ is bounded and $h_t(\cdot, u)$ is bounded and uniformly Lipschitz by assumption.

Now we analyze the difference bewteen normalized estimate and unnormalized estimate.

$$
\begin{aligned}
&\left(E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}} - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr}\right]\right)^2 \\
&= \left(E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}\left(1 - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right)\right]\right)^2 \\
&\le E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}\left(1 - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right)\right]^2 \qquad (5)\\
&\le M^2\mathbb{E}\left[\left(1 - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right)^2\right] \\
&= M^2\left(\mathbb{E}\left[1 - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right]\right)^2 + M^2\,\mathrm{Var}\left(\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right).
\end{aligned}
$$

The first term is $O(n^{-1-4/[d(d+4)]})$ by the same deduction as (4). For the second term,

$$
\begin{aligned}
\mathrm{Var}\left(\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr}\right) &= \mathbb{E}\left[\mathrm{Var}\left(\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr} \mid X^{(1:t-1)}, W^{(1:t-1)}\right)\right] \\
&\quad + \mathrm{Var}\left(\mathbb{E}\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}}{sr} \mid X^{(1:t-1)}, W^{(1:t-1)}\right]\right) \\
&= O(n^{-1-\frac{4}{d(d+4)}}) \text{ (Lemma 6)} \\
&\quad + \mathrm{Var}\left(\frac{\sum_{i=1}^n\sum_{k=1}^r W_{ik}^{(t-1)}\tilde{1}_t(X_{ik}^{(t-1)})}{\sum_{i=1}^n\sum_{k=1}^r W_{ik}^{(t-1)}}\right),
\end{aligned}
$$

where

$$
\tilde{1}_t(x) = \int_{\mathcal{X}} h_t(x,u)\,du\,.
$$

Since $h_t(\cdot,u)$ is bounded and uniformly Lipschitz, $\tilde{1}_t$ is bounded and Lipschitz. By induction hypothesis, the variance term is $O(n^{-1-4/[d(d+4)]})$. Now we have

$$
\begin{aligned}
&\left(E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}} - \int \pi_t(x^{(1:t)})\phi(x^{(t)})dx^{(1:t)}\right]\right)^2 \\
&= \left(E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}} - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr}\right]\right. \\
&\quad \left. + \left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr} - \int \pi_t(x^{(1:t)})\phi(x^{(t)})dx^{(1:t)}\right]\right)^2 \\
&\le 2\left(E\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}} - \frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr}\right]\right)^2 \\
&\quad + 2\left(\mathbb{E}\left[\frac{\sum_{k=1}^s\sum_{\ell=1}^r W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr} - \int \pi_t(x^{(1:t)})\phi(x^{(t)})dx^{(1:t)}\right]\right)^2 = O(n^{-1-\frac{4}{d(d+4)}}),
\end{aligned}
$$

by (4) and (5). This completes the induction hypothesis for the bias at step t,

For the variance at step t,

$$\operatorname{Var}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}\right) = \mathbb{E}\left[\operatorname{Var}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} \mid X^{(1:t-1)}, W^{(1:t-1)}\right)\right] + \operatorname{Var}\left(\mathbb{E}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} \mid X^{(1:t-1)}, W^{(1:t-1)}\right]\right).$$

The first term is $O(n^{-1-\frac{4}{d(d+4)}})$ by Lemma 7. For the second term,

$$\begin{aligned}
&\operatorname{Var}\left(\mathbb{E}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} \mid X^{(1:t-1)}, W^{(1:t-1)}\right]\right)\\
&\leq 2\operatorname{Var}\left(\mathbb{E}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr} \mid X^{(1:t-1)}, W^{(1:t-1)}\right]\right)\\
&\quad + 2\operatorname{Var}\left(\mathbb{E}\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr} \mid X^{(1:t-1)}, W^{(1:t-1)}\right]\right)\\
&\leq 2\mathbb{E}\left[\left(\frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} - \frac{\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{sr}\right)^{2}\right]\\
&\quad + 2\operatorname{Var}\left(\frac{\sum_{i=1}^{n}\sum_{k=1}^{r} W_{ik}^{(t-1)}\tilde{\phi}_t(X_{ik}^{(t-1)})}{\sum_{i=1}^{n}\sum_{k=1}^{r} W_{ik}^{(t-1)}}\right)\\
&= O(n^{-1-\frac{4}{d(d+4)}}) \text{ (derivation in (5))}\\
&\quad + O(n^{-1-\frac{4}{d(d+4)}}) \text{ (induction hypothesis).}
\end{aligned}$$

This proves the induction hypothesis for the variance at step t. □

Proof of Lemma 6. We omit the superscript $(t-1)$ and $(1:t-1)$. We can decompose the conditional variance into two parts:

$$\operatorname{Var}\left[\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid X, W\right] = \underbrace{E\left[\operatorname{Var}\left(\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid \tilde{X}_i, \tilde{W}_i\right) \mid X, W\right]}_{A} + \underbrace{\operatorname{Var}\left[E\left(\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid \tilde{X}_i, \tilde{W}_i\right) \mid X, W\right]}_{B}.$$

To make the computation easy to read, we first analyze A and B separately.

Let $l_k(x) = \dfrac{\pi_t\left((\tilde{X}_k, x)\right)\phi(x)}{\pi_{t-1}\left(\tilde{X}_k\right) g\left(x \mid \tilde{X}_k\right)}$, which is Lipschitz by assumption. Suppose the Lipschitz constant is L_k, which, for example, can be $2ML$.

$$
\begin{aligned}
A &= E\left[\operatorname{Var}\left(\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid \tilde{X}_i, \tilde{W}_i\right) \mid X, W\right] \\
&= \frac{1}{n^2}\sum_{k=1}^{s}\sum_{\ell=1}^{r} E\left[\operatorname{Var}\left(W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid \tilde{X}_i, \tilde{W}_i\right) \mid X, W\right] \\
&\le \frac{1}{4n^2}\sum_{k=1}^{s}\sum_{\ell=1}^{r} E\left[\max_{x,y\in\Gamma_t(\tilde{X}_k, H([(\ell-1)/r, \ell/r]))} (l_k(x) - l_k(y))^2 \mid X, W\right] \text{ (Popoviciu's inequality on variance)} \\
&\le \frac{1}{4n^2}\sum_{k=1}^{s}\sum_{\ell=1}^{r} E\left[\max_{x,y\in\Gamma_t(\tilde{X}_k, H([(\ell-1)/r, \ell/r]))} L_k^2 ||x-y||^2 \mid X, W\right] \\
&\le \frac{1}{4n^2}\sum_{k=1}^{s}\sum_{\ell=1}^{r} E\left[\max_{x,y\in H([(\ell-1)/r, \ell/r])} L^2 L_k^2 ||x-y||^2 \mid X, W\right] \\
&\le \frac{1}{4n^2}\sum_{k=1}^{s}\sum_{\ell=1}^{r} E\left[4(d+3)L^2 L_k^2 (1/r)^{2/d} \mid X, W\right] \text{ (Hölder continuity)} \\
&= \frac{d+3}{n^2}\sum_{k=1}^{s} L^2 L_k^2 r^{1-2/d} = \frac{(d+3)L^2 L_k^2 r^{-2/d}}{n} = O(n^{-1-\frac{4}{d(d+4)}}).
\end{aligned}
$$

For part B, we have

$$
\begin{aligned}
B &= \operatorname{Var}\left[E\left(\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)}) \mid \tilde{X}_k, \tilde{W}_k\right) \mid X, W\right] \\
&= \operatorname{Var}\left[\frac{r}{n}\sum_{k=1}^{s}\int_{\mathcal{X}} \frac{\pi_t\left(\left(\tilde{X}_k^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_k^{(1:t-1)}\right)}\phi(x)dx \mid X, W\right]
\end{aligned}
$$

Let

$$
f_k(x) = \int_{\mathcal{X}} \frac{\pi_t\left(\left(\tilde{X}_k^{(1:t-2)}, x, u\right)\right)}{\pi_{t-1}\left(\tilde{X}_k^{(1:t-2)}, x\right)}\phi(u)du,
$$

which is Lipschitz by assumption. Suppose the Lipschitz constant is L_B for all k, which, for example, can be $2ML$. Then by Theorem 3 (Theorem 3 requires the functions to be the same, but actually it can be seen that the proof still applies as long as all the functions have the same Lipschitz constant),

$$
B \le \frac{(d+3)L_B^2}{(n/r)^{1+2/d}} = \frac{(d+3)L_B^2 r^{1+2/d}}{n^{1+2/d}} = O(n^{-1-\frac{4}{d(d+4)}}).
$$

Combining A and B proves the claim. □

Proof of Lemma 7. Let $e(X, W) = \mathbb{E}[\frac{1}{n} \sum_{k=1}^{s} \sum_{\ell=1}^{r} W_{k\ell}^{(t)} \mid X, W]$.

$$\operatorname{Var}\left[\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}} \mid X, W\right] \leq 2\operatorname{Var}\left[\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{e(X,W)} \mid X, W\right]$$
$$+ 2\operatorname{Var}\left[\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}\left(1 - \frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}{e(X,W)}\right) \mid X, W\right].$$

On the right hand side, the first term is $O(n^{-1-4/[d(d+4)]})$ by Lemma 6; for the second term,

$$\operatorname{Var}\left[\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}\left(1 - \frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}{e(X,W)}\right) \mid X, W\right]$$
$$\leq \mathbb{E}\left[\left(\frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}\phi(X_{k\ell}^{(t)})}{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}\right)^2\left(1 - \frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}{e(X,W)}\right)^2 \mid X, W\right]$$
$$\leq M^2\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)}}{e(X,W)}\right)^2 \mid X, W\right]$$
$$= \frac{C_\phi^2}{e(X,W)^2}\operatorname{Var}\left[\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)} \mid X, W\right]$$
$$\leq \frac{C_\phi^2}{\underline{e}^2}\operatorname{Var}\left[\frac{1}{n}\sum_{k=1}^{s}\sum_{\ell=1}^{r} W_{k\ell}^{(t)} \mid X, W\right] = O(n^{-1-\frac{4}{d(d+4)}})$$

by taking ϕ to be the constant function 1 in Lemma 6. □