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

April 7, 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 (SMC) (Doucet et al. 2001; Liu and Chen 1998) has been studied intensively in the past two decades and applied broadly to high-dimensional statistical inference, signal processing, biology and many other fields. 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 one particularly interesting dynamic system that have been treated with SMC. The state-space model consists of the hidden Markovian state equation and the noisy observation equation. The hidden state, for instance, can

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

be interpreted as 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. For the state-space model, characterizing the distribution of the hidden state is well known as the filtering problem; thus, SMC is more commonly known as the particle filter in this context (Gordon et al. 1993).

Roughly speaking, SMC is built based on sequential importance sampling (SIS), which recursively simulates a future state and reweighs the sampling path, with additional resampling steps (Liu and Chen 1998). In a vanilla SIS 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 SMC can be decomposed into two steps: forward-sampling (or more intuitively, growth) and resampling. 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. Since resampling can also reduce the number of particles, which decreases the estimation accuracy of the algorithm at the current step, it is important to have a growth step that can produce "diverse" samples so as to explore the space more fully (Fearnhead and Clifford 2003). This idea will be discussed in more detail in Section 5. In the resampling step, we rejuvenate all the weights where samples with higher weights are more likely to be retained. Various approaches have been proposed for both growth and resampling (Doucet et al. 2001; Lin et al. 2013).

One of the earliest approaches to improving the SIS method designed for simulating chain polymers was proposed in Wall and Erpenbeck (1959), known as the enrichment method. In this method, at each stage we "amplify" each currently "alive" partial polymer chain (some simulated partial chains are "dead" due to the encountering of a conflict) by making r exact copies, and at the next stage we grow each of the enriched copies by adding to it sequentially s more monomers according to the growth rule. Grassberger (1997) improved this method by choosing r adaptively according to the weights and pruning away some low-weight partial polymers probabilistically. The roles of resampling in a SMC framework were first discussed in Liu and Chen (1995), and the connection between resampling and aforementioned pruning and enrichment methods was brought up in Liu et al. (2001). Fearnhead and Clifford (2003) developed an "optimal resampling" method for state-space models when the state space is discrete, where each sample can have multiple descendants—one for each possible value in the state space—to explore the whole space thoroughly. Another direction to improve SMC performance from the growth step is to propose better trial distributions, such as employing the auxiliary particle filter (Pitt and Shephard 1999) and look-ahead strategies (Lin et al. 2013).

There are various means to resample from a collection of weighted particles. The naïvest way to resample is called bootstrap resampling or multinomial resampling (Tibshirani and Efron 1993), where the new particles are sampled from independent and identically distributed (i.i.d.) multi-

nomial 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. Although there has been no theoretical guarantee for the optimal transport resampling (aside from its validity), to the best of our knowledge, sometimes it works very well in practice. 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.

We study both the growth and the resampling steps of SMC in this paper, and our main contributions are:

1. We prove that in one dimension, 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. The equivalences require surprisingly different techniques to prove.

2. 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 show 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.

3. We introduce the stratified multiple-descendant growth (SMG) method for constructing a trial distribution in forward sampling. SMG makes use of the Hilbert space-filling curve to probe the space more consistently in terms of the Wasserstein metric, which we show in simulations can greatly improve the performance of SMC, especially if combined with Hilbert curve resampling.

The rest of the paper is organized as follows. Some relevant notations, definitions, and formulations are introduced 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 describe the SMG algorithm in detail and explain why it enables one to better explore the space. Numerical studies are carried out in Section 6. Section 7 wraps up the paper with some important open problems. With few exceptions, proofs are deferred to the appendix.

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 SMC. 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 $W_j \geq 0, j = 1, 2, \cdots, n$, and 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,

$$\sum_{i=1}^m \frac{1}{m} \delta_{\tilde{X}_i} \approx \sum_{j=1}^n W_j \delta_{X_j},$$

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.

2.2 Sequential Monte Carlo

To set up future analyses, we here describe a generic SMC 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 $\pi_T = \pi$.

We can decompose $\pi(x)$ as

$$\pi(x^{(1:T)}) = \pi_1(x^{(1)}) \frac{\pi_2(x^{(1:2)})}{\pi_1(x^{(1)})} \cdots \frac{\pi_T(x^{(1:T)})}{\pi_{T-1}(x^{(1:T-1)})}.$$

Suppose the trial sampling distribution is constructed as

$$g(x^{(1:T)}) = g_1(x^{(1)}) g_2(x^{(2)} \mid x^{(1)}) \cdots g_T(x^{(T)} \mid x^{(1:T-1)}),$$

which may be selected as a Markov sequence in some problems for computational convenience. Given the target and trial distributions, the importance weight is

$$w(x^{(1:T)}) = \frac{\pi_T(x^{(1:T)})}{g_1(x^{(1)}) g_2(x^{(2)} \mid x^{(1)}) \cdots g_T(x^{(T)} \mid x^{(1:T-1)})}.$$

While sampling sequentially, the importance weight can be updated recursively:

$$w^{(t)}(x^{(1:t)}) = w^{(t-1)}(x^{(1:t-1)}) \frac{\pi_t(x^{(1:t)})}{\pi_{t-1}(x^{(1:t-1)}) g_t(x^{(t)} \mid x^{(1:t-1)})}.$$

Combined with resampling mentioned in Section 1, a generic SMC algorithm, also known as SISR, is outlined in Algorithm 1.

In a state-space model, we have

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

Algorithm 1: Sequential importance sampling with resampling (SISR).

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:

$$W_j^{(1)} \propto \frac{\pi_1(X_j^{(1)})}{g_1(X_j^{(1)})}.$$

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 $(X_i^{(1:T)}, W_i^{(T)})_{1\le i\le n}$

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 SISR 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 $w^{(t)} \propto w^{(t-1)} p_y(y^{(t)} \mid x^{(t)})$.

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 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, which also appears in Reich (2013) and Webber (2019). Most available resampling methods, as listed below, fit into this framework.

- **Multinomial resampling**. Each $\tilde{X}_i$ is an i.i.d. sample from the multinomial distribution 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).

- **Stratified resampling**. Let $U_i \sim \text{Unif}\left(\frac{i-1}{m}, \frac{i}{m}\right]$, independently for $i = 1, \ldots, m$, and sample

$$\tilde{X}_i = X_j \;\text{ if }\; U_i \in \left(\sum_{k=1}^{j-1} W_k, \sum_{k=1}^{j} W_k \right].$$

  See Figure 1 for an illustration of stratified resampling. Stratified resampling corresponds to a staircase matrix; see Figure 2(b) for an example and Definition 1 for a formal definition.

- **Residual resampling**. 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 $\tilde{W}_j \propto mW_j - \lfloor mW_j \rfloor$.

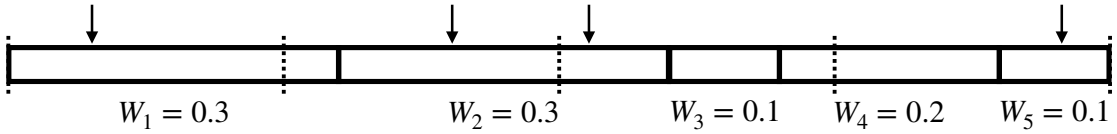


Figure 1: Illustration of stratified resampling. First line up the weights, then divide the interval into m equal parts, uniformly choose one point from each subinterval and record in which weight's region it lands. In the presented example where $m = 4$, $n = 5$, particles 1 and 5 are resampled once,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

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

(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 $(W_1, W_2, W_3, W_4, W_5) = (0.3, 0.3, 0.1, 0.2, 0.1)$.

2.4 Criteria for choosing resampling schemes

To choose from the set of valid resampling procedures, we need some 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. As examples, several possible loss functions are given below.

Conditional variance. By picking $\ell(\mathbb{P}, \tilde{\mathbb{P}})$ to be $(\mathbb{E}_{\mathbb{P}}[\phi(X)] - \mathbb{E}_{\tilde{\mathbb{P}}}[\phi(X)])^2$, we use the conditional variance $\mathrm{Var}[m^{-1} \sum_{i=1}^{m} \phi(\tilde{X}_i) \mid X, W]$ as a measure of goodness. It is straightforward to verify that

$$\mathrm{Var}_P \left[\frac{1}{m} \sum_{i=1}^{m} \phi(\tilde{X}_i) \mid X, W \right] = \frac{1}{m^2} \phi^\top \left(m \cdot \mathrm{diag}\{W_1, W_2, \cdots, W_n\} - P^\top P \right) \phi \tag{1}$$

where $\phi = (\phi(X_1), \phi(X_2), \cdots, \phi(X_n))^\top$ and the subscript P means resampling according to matrix P.

Energy distance. We can 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

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

where Y_1, Y_1' follow $\mathbb{P}_1$, Y_2, Y_2' follow $\mathbb{P}_2$ and the four random variables are independent.

Wasserstein distance. The Wasserstein distance between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined 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 admit $\mathbb{P}_1$ and $\mathbb{P}_2$ as its marginal distributions.

In Section 3, we prove that minimizing the conditional variance and expected squared energy distance are equivalent in the one dimensional case, both of which can be reached 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 the Hilbert curve in multiple dimensions.

2.5 Weighted resampling

We briefly pause to discuss a more flexible resampling objective and strategy as suggested in Liu et al. (2001). Suppose at time $t-1$ we have a set of weighted particles $(X_j, W_j)_{j=1}^n$, which can be treated as a discrete representation of π_{t-1}. We can generate another discrete representation as follows:

- Pick a probability vector $(a_1, a_2, \ldots, a_n)$. For i in $1, \ldots, m$, let $\tilde{X}_i$ be independently sampled from
$$\tilde{X}_i \mid X, W \sim \text{Multinomial}(1, X, (a_1, a_2, \ldots, a_n)),$$
Assign the new weight associated with this particle as $\tilde{W}_i = W_j / a_j$.
- Return the new representation $(\tilde{X}_i, \tilde{W}_i)_{i=1}^m$.

By letting $a_j = W_j$, this weighted resampling is exactly the resampling in Algorithm 1. In this case, all the weights can be fully rejuvenated while bringing a non-negligible variance. In another case, by choosing different a_j, for example, $a_j \propto \sqrt{W_j}$, one can reduce the resampling variance at the cost of less balanced weights. It is unclear how to find an optimal trade-off between reducing randomness brought by resampling and balancing the particle weights.

As a heuristic analysis, we consider a one-step variance for SMC in the following setting. Consider a special sub-class of the resampling weights: $a_j \propto W_j^\gamma$, where $\gamma \in \mathbb{R}$. Suppose we resample $(X_j^{(1:t-1)}, W_j^{(t-1)})_{j=1}^n$ into $(\tilde{X}_i^{(1:t-1)}, \tilde{W}_i^{(t-1)})_{i=1}^m$ with $a_j \propto W_j^\gamma$, and $(X_i^{(1:t)}, W_i^{(t)})_{i=1}^m$ are the weighted particles after an SMC growth step with trial distribution g. For an estimand function ϕ, we can obtain (see Appendix C for derivation)

$$\text{Var}\left[\sum_{i=1}^m W_i^{(t)} \phi(X_i^{(t)}) \mid X, W \right] = m \sum_{j=1}^n W_j^{2-\gamma} C_j \sum_{j=1}^n W_j^\gamma + \text{constant},$$

where both

$$C_j = \int \frac{\pi_t \left((X_j, x) \right)^2 \phi(x)^2}{\pi_{t-1} \left(X_j \right)^2 g \left(x \mid X_j \right)} dx$$

and the constant term do not depend on γ. We have omitted the superscript for step $t-1$ to ease the notations.

Note that if C_j is roughly "independent" of W_j, so $\sum_{j=1}^n W_j^{2-\gamma} C_j \approx \sum_{j=1}^n W_j^{2-\gamma} (\sum_{j=1}^n C_j / n)$, then the best γ is 1 by the Cauchy–Schwarz inequality, which corresponds to equal weights after resampling. If C_j is roughly "positively correlated" with W_j (e.g., $C_j \approx W_j^\alpha$, $\alpha > 0$), the optimal γ should be larger than 1, and vice versa. Since we generally know very little about C_j, $\gamma = 1$ is a reasonable choice.

3 Optimal resampling in one dimension

A good resampling scheme should naturally 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.

3.1 Stratified resampling matrix

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.

$$\begin{array}{ccccc} & \vdots & & \vdots & \\ \cdots & p_{ij} & \cdots & p_{il} & \cdots \\ & \vdots & & \vdots & \\ \cdots & p_{kj} & \cdots & p_{kl} & \cdots \\ & \vdots & & \vdots & \end{array}$$

It is not hard to see that the matrix of stratified resampling is a staircase matrix up to row permutation. A staircase matrix has at most $n + m - 1$ non-negative entries and has a clear spatial structure which looks like the following:

$$\begin{pmatrix} * & * & * & & \\ & & * & * & \\ & & & * & \\ & & & * & * \end{pmatrix}$$

The non-negative entries form a path (allowing diagonal moves) from the top left entry to the bottom right entry.

Lemma 1. *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)*.*

Lemma 2. *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 2, 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 $P^{SR}_{m,W}$ the stratified resampling matrix of a set of weighted particles $(X_j, W_j)_{j=1}^n$ if the following conditions are satisfied:*

(1) $P^{SR}_{m,W} \in \mathcal{P}_{m,W}$.

(2) $P^{SR}_{m,W}$ can be converted to a staircase matrix after some row permutation.

3.2 Minimizing resampling variance

If the goal is to estimate $\mathbb{E}[\phi(X)]$, then ordering the states by the function ϕ and then applying stratified resampling gives the minimum variance. This result is noted in Webber (2019), although it seems 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. Following equation (1), we see minimizing the conditional variance is equivalent to maximizing

$$\sum_{i=1}^{m}(\sum_{j=1}^{n} p_{ij}\phi(X_j))^2 = \|P\phi\|_2^2 = \phi^\top P^\top P\phi, \quad \text{with } \phi = (\phi(X_1), \phi(X_2), \ldots, \phi(X_n))^\top.$$

Theorem 1. *Let ϕ be an n-dimensional vector with distinct elements in increasing order, then the stratified resampling matrix $P^{SR}_{m,W}$ solves*

$$\arg\max_{P\in\mathcal{P}_{m,W}} \phi^\top P^\top P\phi \tag{2}$$

and thus minimizes the resampling variance.

Remark. The problem of finding the solution of (2) is a concave minimization problem, which is very expensive to solve in practice. Knowing that the solution is the ordered stratified resampling matrix drastically simplifies the optimization task.

3.3 Minimizing expected squared energy distance

Interestingly, we observe the following result regarding the energy distance. For simplicity of notations, we assume $X_1 < X_2 < \cdots X_n$, without loss of generality.

Theorem 2. *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 expected squared energy distance among resampling methods in $\mathcal{P}_{m,W}$.*

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 2, the solution to the following optimization problems minimizes the expected squared energy distance:*

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

We now provide a very short and succinct proof of Theorem 2.

Proof of Theorem 2. Let $P^{\mathrm{SR}}_{m,W} = (p^*_{ij})$ 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. □

3.4 Minimizing the earth mover's distance

When $m = n$, i.e., the number of particles remains the same after resampling, a view from coupling can be adopted. A matrix $P \in \mathcal{P}_{n,W}$ defines a "coupling" between $\sum_{j=1}^{n} W_j \delta_{X_j}$ and $\sum_{j=1}^{n} n^{-1} \delta_{X_j}$ (i.e., a joint distribution that retains the two given marginals): $n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \delta_{(X_i, X_j)}$. Perhaps due to this fact, Reich (2013) proposes a resampling method based on optimal transport. With our notations, the optimization problem is equivalent to

$$\operatorname*{arg\,min}_{P \in \mathcal{P}_{n,W}} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \ell(X_i, X_j) = \mathbb{E} \left[\sum_{j=1}^{n} \ell(X_j, X^*_j) \mid X, W \right]$$

Here, the loss function is usually taken to be the squared Euclidean distance. We consider a more general case where ℓ is any strictly convex function and m and n are not necessarily equal.

Theorem 3. *Let ϕ be an n-dimensional vector with distinct elements in increasing order and ψ be an m-dimensional vector with distinct elements in increasing order. Then, the stratified resampling matrix $P^{SR}_{m,W}$ solves*

$$\operatorname*{arg\,min}_{P \in \mathcal{P}_{m,W}} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \ell(\psi_i - \phi_j),$$

where ℓ is a strictly convex function.

Remark 1. It is well known that optimal transport in one dimension has explicit solutions of the above form. For example, Theorem 2.18 in Villani (2008) proves the case of quadratic loss, in which case we obtain that ordered stratified resampling matrix minimizes $\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} (\psi_i - \phi_j)^2$ for all choices of m and ψ. This means that in the original definition of optimal transport resampling, the choice of pairing X_j and X^*_j for calculating loss is not essential in one dimension.

Remark 2. The computation cost of an exact algorithm to optimize the earth mover's distance is of order $O(n^3 \log n)$. Some approximation approaches give a relaxed solution with computational complexity of order $O(n^2)$ (Cuturi 2013; Benamou et al. 2015). Despite the significant reduction of computation cost, $O(n^2)$ is still prohibitively expensive for large n, especially for long sequences. Theorem 3 shows that if the dimension at each step is one, optimal transport resampling can be solved in an order of $O(n \log n)$, which is merely the order for sorting the particles.

4 Error of ordered stratified resampling

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

4.1 One-dimensional case

Theorem 4. *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 include the proof of this theorem in the main text because it is succinct and provides some intuition on the role played by stratification.

Proof of Theorem 4. Without loss of generality, suppose $X_1 < X_2 < \cdots < X_n$ and P is a staircase weight matrix corresponding to stratified resampling. Each X_i^* can only take values in $X_{il}, X_{il+1}, \cdots, X_{ir}$, with

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

Hence,

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

□

Remark. 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 the next section, we see that this intuition regarding spacial ordering of the X_j's also provides useful suggestions for conducting resampling in multiple-dimensional cases.

4.2 Multidimensional case

There are many cases where the particles to be resampled are multidimensional. One example is the state-space model where the hidden state at each step is multidimensional (e.g., the tracking problem in Section 6.2). Moving away from state-space models, we might want to resample the whole path because we care about not only the marginal posterior distributions, but the joint distribution as well (e.g., see Section 6.1).

4.2.1 Hilbert curve and its properties

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). 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).

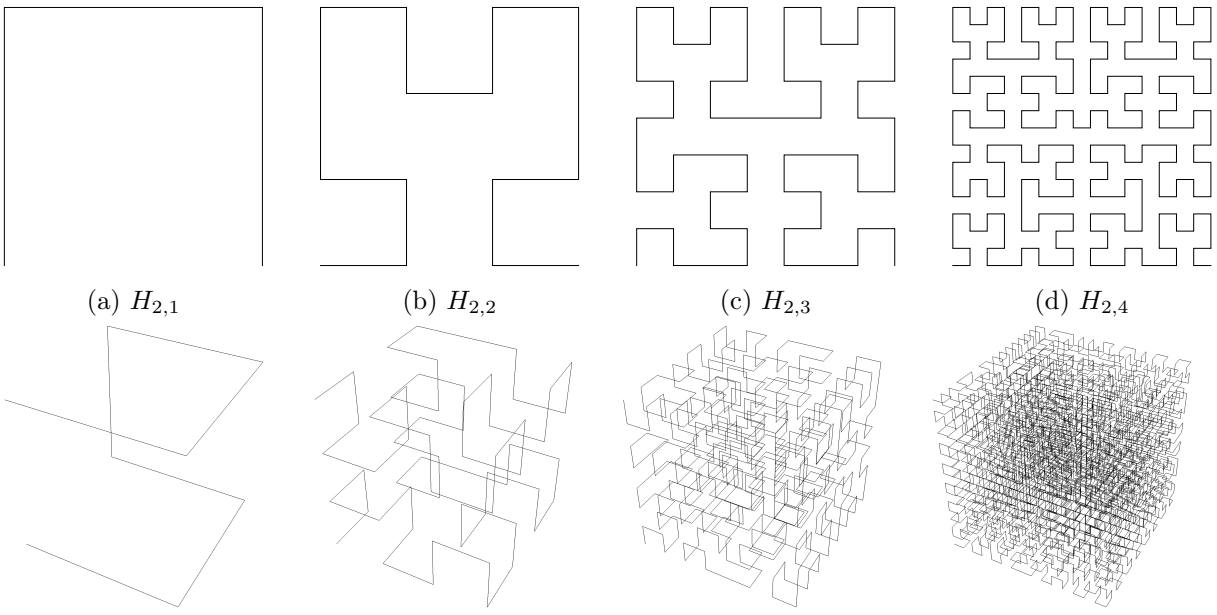
A d-dimensional Hilbert curve is a continuous function $H : [0, 1] \to [0, 1]^d$. Its most important properties relevant to our tasks are as follows:

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

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

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

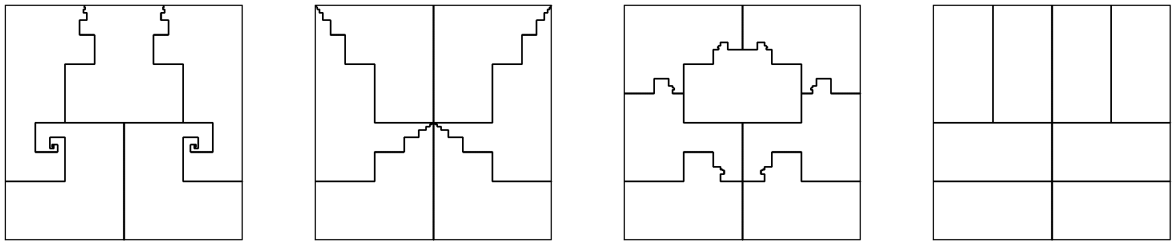
The Hilbert curve is defined as the limit of a sequence of curves; see Figure 3 for an illustration in two and three dimensions. Many software packages can efficiently convert between x and $H(x)$ (e.g., the Python package hilbertcurve). We omit here the rigorous definition of Hilbert curves and refer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



(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.

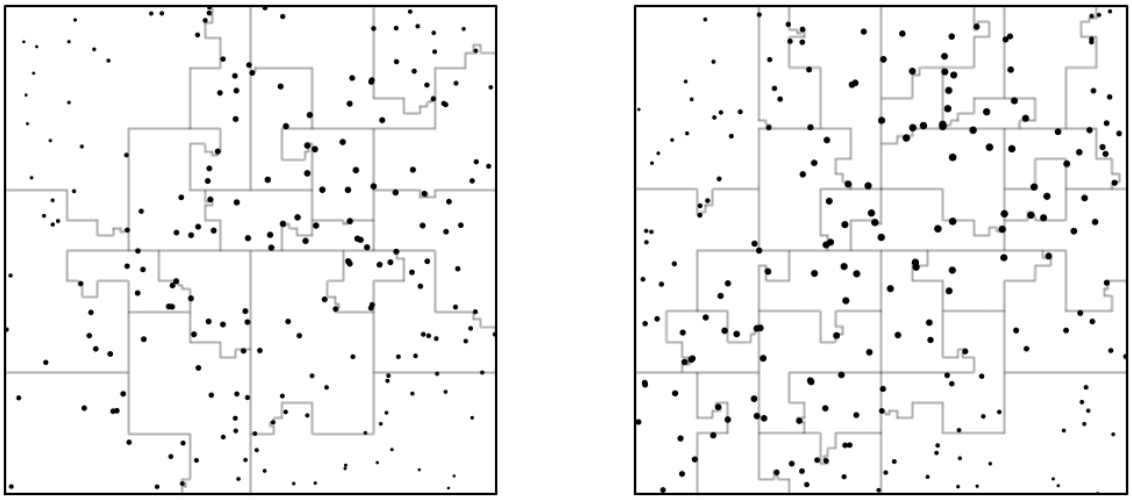


(a) Five parts. (b) Six parts. (c) Seven parts. (d) Eight parts.

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

4.2.2 Hilbert curve resampling

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 simply sorting the particles so $(h(X_j))_{j=1}^n$ is in ascending order, and then applying stratified resampling. Note that in one dimension this reduces to ordered stratified sampling. Following the intuition in 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 5 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).



(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).

Theorem 5. *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 1. The exponent $1+2/d$ improves the original rate $1+1/d$ shown in Gerber et al. (2019).

Remark 2. 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 5 implies that

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

We show in Proposition 1 that no other ordering rule will 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 6. 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 6 is intuitively optimal, since m balls with radius of the order $1/m^{1/d}$ are needed to cover the space.

Theorem 6. *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 Multiple-descendant growth

In this section, we discuss the multiple-descendant growth. Let r be a positive integer. At each step after resampling, conditional on the unweighted particles at the previous generation $\tilde{X}_{1:n}^{(1:t-1)}$, we sample for each particle $X_i^{(1:t-1)}$ independently r descendants:

$$X_{ij}^{(t)} \mid \tilde{X}_i^{(1:t-1)} \sim g(\cdot \mid \tilde{X}_i^{(1:t-1)}),\ j = 1, \ldots, r;\quad i = 1, 2, \ldots, n. \tag{3}$$

The multiple-descendant growth can be thought of as making r copies of each particle and each of them having an independent descendant. To probe the space more stably, instead of picking the

same g for each j, we can do this in a stratified manner. In the discrete case, this is similar to take each possible value exactly one time (Fearnhead and Clifford 2003).

In the interest of clarity, we now suppose the support of each $g(\cdot \mid \tilde{X}_i^{(1:t-1)})$ is $\mathcal{X} = [0,1]^d$ and is absolutely continuous with respect to the Lebesgue measure. This constraint is not essential and SMG can be similarly defined when g has unbounded support. Pick $0 = s_{i,0} \le s_{i,1} \le \cdots \le s_{i,r-1} \le s_{i,r} = 1$ for $i = 1, 2, \ldots, n$ such that

$$\int_{H([s_{i,j-1}, s_{i,j}])} g(x \mid \tilde{X}_i^{(1:t-1)})\, dx = \frac{1}{r}; \quad j = 1, 2, \ldots, r; \quad i = 1, 2, \ldots, n. \tag{4}$$

Now we introduce stratified multiple-descendant growth (SMG).

Definition 3 (SMG)**.** *Conditional on the unweighted particles $\tilde{X}_{1:n}^{(1:t-1)}$, independently sample*

$$\bar{X}_{ij}^{(t)} \mid \tilde{X}_{1:n}^{(1:t-1)} \sim g(\cdot \mid \tilde{X}_i^{(1:t-1)}) \frac{\mathbb{I}(H([s_{i,j-1}, s_{i,j}]))}{1/r}; \quad j = 1, \ldots, r; \quad i = 1, \ldots, n,$$

where the $s_{i,j}$'s satisfy (4) *and $\mathbb{I}(A)$ means the indicator function on A.*

Here we use $\bar{X}$ to distinguish SMG particles from the vanilla multiple-descendent growth. Theorem 7 shows that SMG is unbiased and more stable that i.i.d. multiple-descendant growth.

Theorem 7. *Following Definition 3, define the unnormalized weights as*

$$\bar{W}_{ij}^{(t)} = \mathbb{I}\left(H([s_{i,j-1}, s_{i,j}])\right) \frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, \bar{X}_{ij}^{(t)}\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right) g\left(\bar{X}_{ij}^{(t)} \mid \tilde{X}_i^{(1:t-1)}\right)}$$

for $j = 1, \ldots, r, i = 1, \ldots, n$. We have,

(1) For any Borel-measurable function h on $\mathcal{X} = [0,1]^d$ and $1 \le i \le n$,

$$\frac{1}{r}\mathbb{E}\left(\sum_{j=1}^{r} \bar{W}_{ij}^{(t)} h(\bar{X}_{ij}^{(t)}) \mid \tilde{X}_i^{(1:t-1)}\right) = \int_{\mathcal{X}} \frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)} h(x)\, dx\,.$$

(2) Almost surely, SMG satisfies

$$W_p\left(\frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{r} \delta_{\bar{X}_{ij}^{(t)}}, \frac{1}{n}\sum_{i=1}^{n} g\left(\cdot \mid \tilde{X}_i^{(1:t-1)}\right)\right) \le \frac{2\sqrt{d+3}}{r^{1/\max(p,d)}}.$$

(3) With the same trial distribution g, suppose $(X_{ij}^{(1:t)}, W_{ij}^{(t)})_{1 \le i \le n, 1 \le j \le r}$ are the weighted samples through the vanilla multiple-descendant growth defined in (3), where the weights are defined as

$$W_{ij}^{(t)} = \frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, X_{ij}^{(t)}\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right) g\left(X_{ij}^{(t)} \mid \tilde{X}_i^{(1:t-1)}\right)}$$

without normalization. Then for a Borel-measureable function h on $\mathcal{X} = [0,1]^d$ we have

$$\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j=1}^{r} W_{ij}^{(t)} h(X_{ij}^{(t)}) \mid \tilde{X}_{1:n}^{(1:t-1)}\right) \ge \mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j=1}^{r} \bar{W}_{ij}^{(t)} h(\bar{X}_{ij}^{(t)}) \mid \tilde{X}_{1:n}^{(1:t-1)}\right),$$

as long as both sides are well-defined.

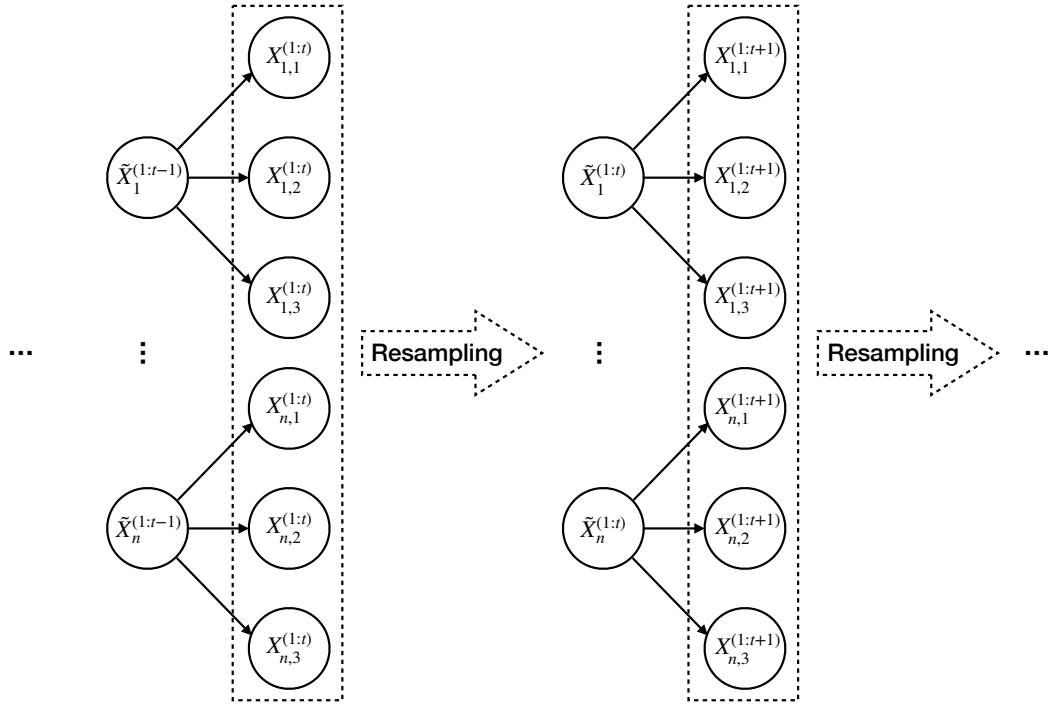


Figure 6: Illustration of multiple-descendant growth.

In the general case where π_t is supported on $\mathbb{R}^d$, we can choose g to be a multivariate Gaussian distribution as in Algorithm 2. The multivariate Gaussian distribution is a particularly convenient choice, because we can easily construct it from the uniform distribution on $[0,1]^d$ via a one-to-one transformation.

Algorithm 2: Gaussian stratified multiple-descendant growth (GSMG).

Input: π_{t-1}, π_t, $\tilde{X}_{1:n}^{(1:t-1)}$, Gaussian parameters μ, Σ, number of descendants r

Output: weighted particles $(\bar{X}_{ij}^{(t)}, \bar{W}_{ij}^{(t)})_{1\le i\le n, 1\le j\le r}$

for $i = 1$ **to** n **do**

Independently sample $U_{i1}, U_{i2}, \ldots, U_{ir}$ with $U_{ij} \sim \mathrm{Unif}[(j-1)/r, j/r]$.

Set $\bar{X}_{ij}^{(t)} = \mu + \Sigma^{1/2}(\Phi^{-1}(u_{ij}^1), \Phi^{-1}(u_{ij}^2), \ldots, \Phi^{-1}(u_{ij}^d))^\top$, where Φ is the cumulative distribution function of the standard Gaussian distribution and $(u_{ij}^1, u_{ij}^2, \ldots, u_{ij}^d) = H(U_{ij})$.

Calculate and normalize weights:

$$\bar{W}_{ij}^{(t)} \propto \frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, \bar{X}_{ij}^{(t)}\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)\varphi\left(\bar{X}_{ij}^{(t)}; \mu, \Sigma\right)},$$

where $\varphi(x; \mu, \Sigma)$ denotes the probability density function of $\mathcal{N}(\mu, \Sigma)$.

end

Return $(\bar{X}_{ij}^{(t)}, \bar{W}_{ij}^{(t)})_{1\le i\le n, 1\le j\le r}$

6 Numerical studies

In this section, we report some simulation results to demonstrate our proposed method. Further simulation details can be found in Appendix B. We report the average runtimes, while the experiments were carried out with parallelization on different machines. Source code can be found at `https://github.com/junliulab/smg` with notebook tutorials at `http://wenshuow.github.io/smg`.

6.1 Sampling from a high-dimensional distribution

In this section, we consider sampling from a T-dimensional distribution. We include a detailed version of the algorithm here, since it is a bit different from the state-space models. We take our target distribution to be the mixture $f_T = 0.5\mathcal{N}(3\mathbf{1}_T, I_T) + 0.5\mathcal{N}(-3\mathbf{1}_T, I_T)$, where $\mathbf{1}_p$ is a p-dimensional vector with all 1's. Let $f_t = 0.5\mathcal{N}(3\mathbf{1}_t, I_t) + 0.5\mathcal{N}(-3\mathbf{1}_t, I_t)$.

1. Let $t = 1$ and sample $X_1^{(1)}, \ldots, X_n^{(1)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 3^2)$; let $W_i^{(1)} = f_1(X_i^{(1)})/\varphi_3(X_i^{(1)})$, where φ_3 is the density of $\mathcal{N}(0, 3^2)$, and then normalize the weights. Resample $(X_i^{(1)}, W_i^{(1)})_{i=1}^n$ via Hilbert curve resampling and denote the unweighted particles again as $X_{1:n}^{(1)}$.

2. From $t = 2$ to $t = T$, for each $X_i^{(1:t-1)}$, sample r $\mathcal{N}(0, 3^2)$ random variables $X_{i1}^{(t)}, \ldots, X_{ir}^{(t)}$ either independently or by stratification. Let $W_{ij}^{(t)} = f_t(X_i^{(1:t-1)}, X_{ij}^{(t)})/(f_{t-1}(X_i^{(1:t-1)})\varphi_3(X_{ij}^{(t)}))$ and then normalize the weights. If $t < T$, resample $\left((X_i^{(1:t-1)}, X_{ij}^{(t)}), W_{ij}^{(t)}\right)_{1\le i\le n, 1\le j\le r}$ to n particles via Hilbert curve resampling and denote the unweighted particles again as $X_{1:n}^{(1:t)}$. If $t = T$, return $\left((X_i^{(1:t-1)}, X_{ij}^{(t)}), W_{ij}^{(t)}\right)_{1\le i\le n, 1\le j\le r}$.

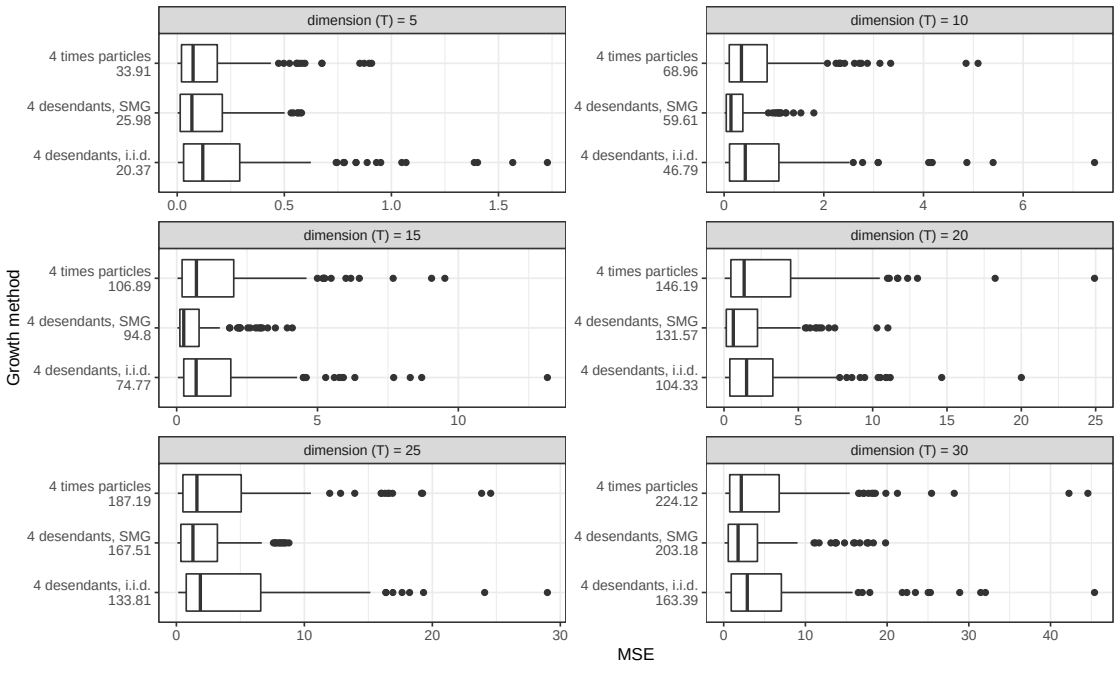


Figure 7: Sampling from a high-dimensional target distribution with $1,000$ particles and 4 descendants or $4,000$ particles and 1 descendant. Numbers below resampling methods indicate average runtimes measured in seconds.

Results are summarized in Figure 7. It can be seen that SMG outperforms i.i.d. multiple-descendant growth, which is close to the results of one-descendant growth with r times the amount of particles when T is moderately large.

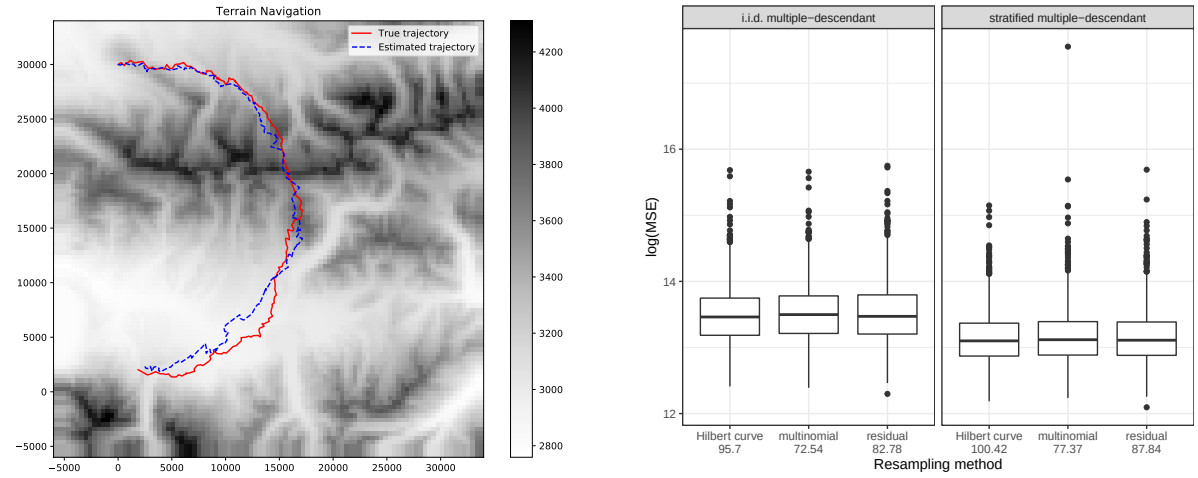
6.2 Terrain navigation

Terrain navigation is a problem with many applications. In a typical terrain navigation problem, an airplane is flying over terrain with known structure and it uses the elevation to estimate its current position.

We conduct simulations of the terrain navigation problem mentioned in Section 1. We take the model and data (a topographical map of a Colorado region) from Givens and Hoeting (2013). Let $X^{(t)} \in \mathbb{R}^2$ be the true location of the plane at time t, $h(z)$ be the true elevation at location z, $Y^{(t)}$ be the observed elevation at time t, $d^{(t)}$ be the measured shift by the inertial navigation system. Let $X^{(0)}$ be known, and the hidden Markov model is defined through

$$X^{(t)} = X^{(t-1)} + d^{(t)} + \epsilon^{(t)},$$
$$X^{(t)} = h(X^{(t)}) + \delta^{(t)},$$

where $\epsilon^{(t)}$ and $\delta^{(t)}$ are independent Gaussian error processes representing the error in drift and elevation measurement, respectively. We are interested in $\pi_t(X^{(t)} \mid Y^{(1:t)})$. Figure 8(a) illustrates one simulation example, where SMC was run with 100-particle multinomial resampling and i.i.d. 4-descendant growth. Figure 8(b) contains box plots of log(MSE) for six different SMC procedures over 960 independent runs; 500 particles were used with 4-descendant growth. Hilbert curve resampling outperforms residual and multinomial resampling, and SMG further lowers the MSE drastically compared to i.i.d. stratified multiple-descendant growth. MSE is defined as the empirical version of $(2T)^{-1}\sum_{t=1}^{T}\mathbb{E}[\|\hat{X}^{(t)} - X^{(t)}\|^2]$, where $\hat{X}^{(t)}$ is the weighted average of the particles at step t.



(a) (b)

Figure 8: (a) One simulation example. The background represents ground elevations indicated by the color bar; (b) simulation results. Numbers below resampling methods are average runtimes in seconds.

6.3 Stochastic volatility

We consider a multivariate stochastic volatility (MSV) model (Harvey et al. 1994). Let Σ be a $p \times p$ matrix with $\Sigma_{ij} = \rho^{|i-j|}$, $X^{(1)} \sim \mathcal{N}(0, \Sigma)$ and

$$X^{(t)} \mid X^{(1:t-1)} \sim \mathcal{N}\left(\alpha X^{(t-1)}, \Sigma\right),$$
$$Y^{(t)} \mid X^{(1:t)}, Y^{(1:t-1)} \sim \mathcal{N}\left(0, \beta^2 \operatorname{diag}(\exp(X^{(t)}))\right),$$

where $\exp(X^{(t)})$ represents the elementwise application of the exponential function to the vector $X^{(t)}$. If $p = 1$, this reduces to the standard SV model. In this section, the MSE is taken with respect to the oracle posterior mean $\mathbb{E}[X^{(t)} \mid Y^{(1:t)}]$ (obtained by running SMC with $3,000$ particles and 4-descendant SMG) instead of the true value $X^{(t)}$, since otherwise the difference between the posterior mean and the true value would dominate the error.

In Figures 9 and 10, "stratified" refers to stratified resampling without ordering the particles. In Figure 10, we compare the performances of five different resampling methods. Hilbert curve resampling consistently performs the best out of all five, and it only takes a little more time than multinomial resampling, residual resampling and stratified resampling. Note that in one dimension, Hilbert curve resampling reduces to ordered stratified resampling, which is the same as optimal transport resampling. They do not perfectly coincide in our simulations because we set the tolerance level so that the time cost is not too large. The results are from $1,600$ independent runs with 100 particles without multiple-descendant growth, since the original optimal transport resampling does not naturally generalize to allow multiple descendants.

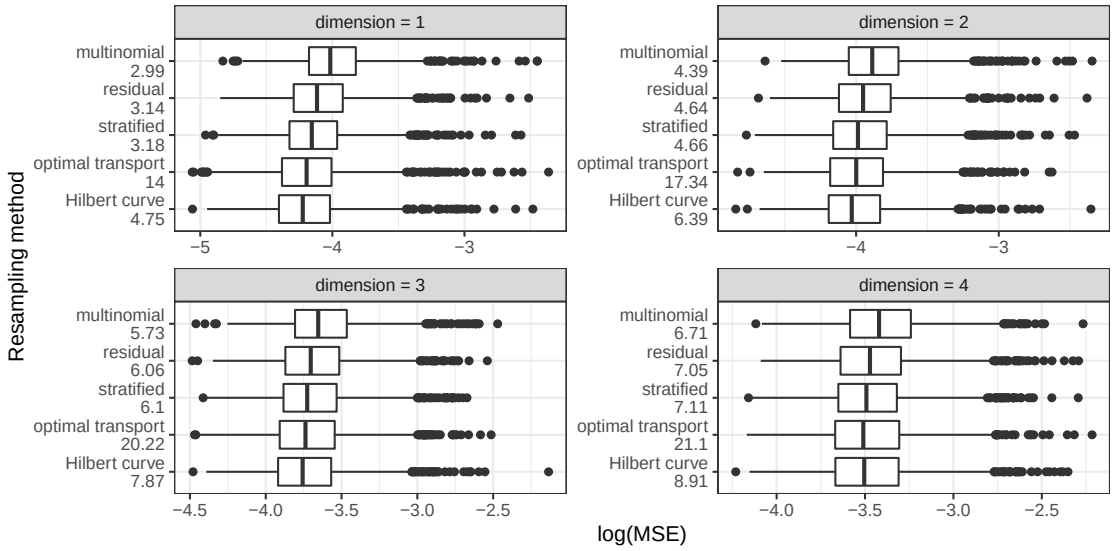


Figure 9: Stochastic volatility simulation results with one descendant. Numbers below resampling methods indicate average time per run measured in seconds.

Figure 10 summaries results from $1,600$ independent runs with 100 particles. We see that Hilbert curve resampling performs the best and the MSE decreases as the number of descendants grows. SMG has lower MSE than i.i.d. multiple-descendant growth, while the gap is especially significant when the number of dimensions is small.

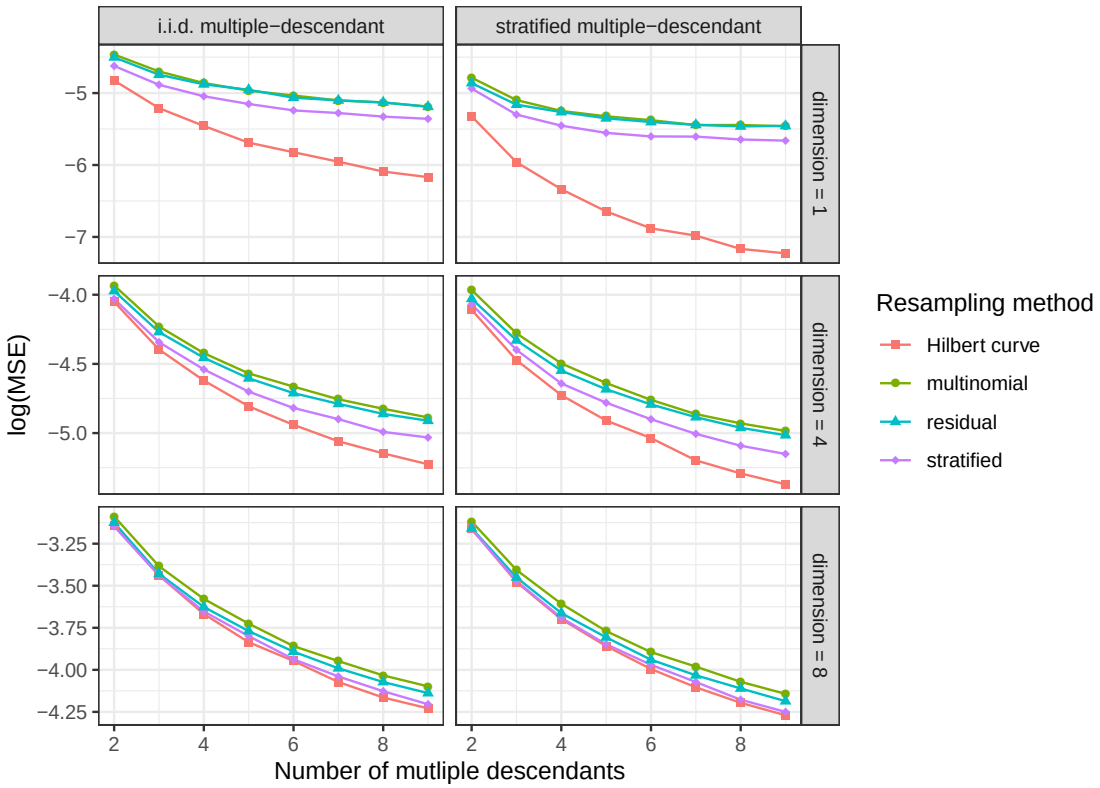


Figure 10: Stochastic volatility simulation results with multiple-descendant growth. Average run times similar across different methods, with Hilbert curve taking slightly longer. Details are reported in Appendix B.

6.4 Weighted Resampling

This section includes simulation results with the SV model, where we implement the weighted resampling idea discussed in Section 2.5. It can be seen that $\gamma = 1$ is a reasonable choice, while sometimes it seems other values of γ might give better results in terms of MSE.

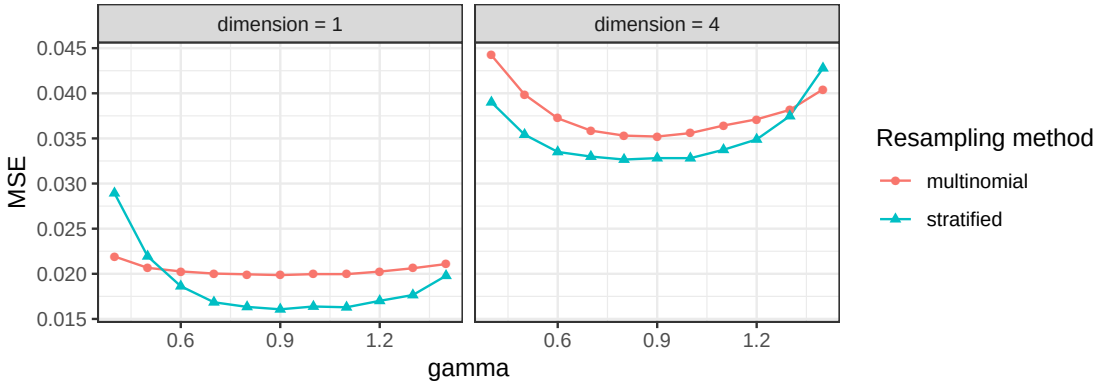


Figure 11: Stochastic volatility simulation results with one descendant and weighted resampling as discussed in Section 2.5. MSE is estimated with $1,600$ independent runs.

7 Discussion

This paper discussed how stratification can help improve the performance of SMC in both resampling and growth steps. For the resampling step, we proved some optimality results for ordered stratified sampling (in multiple dimensions, ordering is given by the Hilbert space-filling curve). For the growth step, we proposed a way to improve the multiple-descendant growth by stratifying the space with the Hilbert curve and discussed its theoretical properties. We provided numerical evidence to support our method. We wish to conclude by highlighting several important unsolved problems.

- **Generalized matrix resampling.** We can generalize the matrix resampling framework in Section 2.3 to allow resampled particles to carry unequal weights (e.g., the optimal resampling introduced in Fearnhead and 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 independently 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 focus on the case with $q_i = 1/m$ in this article. By choosing unequal q_i's, one can further reduce the resampling variance at the cost of less balanced weights. It is unclear what an optimal trade-off might be.

- **Non-matrix resampling.** Not all resampling methods can be represented as a resampling matrix. Systematic resampling (Carpenter et al. 1999) is such an example, since conditional on the original particles the resampled particles are not independent from each other. 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 that includes some non-matrix resampling schemes.

- **What are "correct" objectives of resampling?** This article studies how to minimize the "additional" randomness brought in by resampling to equal weights. As discussed in Section 2.5, we may resample to just obtain a less variable weight distribution. We have discussed several criteria as measurements of randomness, but which measurement of randomness is more "appropriate"? This is related to questions regarding why and how resampling helps and whether there are other criteria that may help better steer the resampling. Answering these questions in a rigorous manner may help us design better objectives for resampling and find alternative ways to tackle the weight degeneracy problem.

- **How to integrate growth and resampling?** Most of the theoretical results in this paper are within either the resampling step or the growth step. How to take into account the fact that the two steps are intertwined is an important open problem.

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

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

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.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

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.

Givens, G. H. and Hoeting, J. A. (2013). *Computational Statistics*. John Wiley & Sons, Inc.

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.

Grassberger, P. (1997). Pruned-enriched Rosenbluth method: Simulations of θ polymers of chain length up to 1 000 000. *Physical Review E*, 56(3):3682.

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.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

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.

Lin, M., Chen, R., and Liu, J. S. (2013). Lookahead strategies for sequential monte carlo. *Statistical Science*, 28(1):69–94.

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.

Liu, J. S., Chen, R., and Logvinenko, T. (2001). A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo methods in practice*, pages 225–246. Springer.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599.

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.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wall, F. and Erpenbeck, J. (1959). New method for the statistical computation of polymer dimensions. *The Journal of Chemical Physics*, 30(3):634–637.

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

A Proofs

Proof of Lemma 1. 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). □

Proof of Lemma 2. 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. Suppose P maximizes $t(P) = \phi^\top P^\top P\phi$ and $\sum_{j=1}^{n} p_{ij}\phi_j$ is ascending with respect to i (note that permutation of rows in P doesn't change the value of $\phi^\top P^\top P\phi$). 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

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

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

Proof of Theorem 3. Let $t(P) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij}\ell(\psi_i - \phi_j)$. 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

$$t(P') - t(P) = \alpha(\ell(\psi_i - \phi_j) + \ell(\psi_k - \phi_l) - \ell(\psi_i - \phi_l) - \ell(\psi_k - \phi_j)).$$

Since ℓ is convex and

$$\begin{aligned} (\psi_i - \phi_j) + (\psi_k - \phi_l) &= (\psi_i - \phi_l) + (\psi_k - \phi_j) \\ |(\psi_i - \phi_j) - (\psi_k - \phi_l)| &< |(\psi_i - \phi_l) - (\psi_k - \phi_j)| \end{aligned}$$

we have

$$\ell(\psi_i - \phi_j) + \ell(\psi_k - \phi_l) < \ell(\psi_i - \phi_l) + \ell(\psi_k - \phi_j)),$$

so $t(P') < t(P)$. This would contradict the fact that P is the minimizer, so such a quadruplet does not exist. By Lemma 1, the solution P is a staircase matrix. □

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

$$\begin{aligned} \mathbb{E}[d(\mathbb{P}, \tilde{\mathbb{P}}) \mid X, W] &= \int_{-\infty}^{\infty} \mathbb{E}[(F_{\mathbb{P}}(x) - F_{\tilde{\mathbb{P}}}(x))^2 \mid X, W] \, dx \\ &= \int_{-\infty}^{\infty} (\mathbb{E}[F_{\tilde{\mathbb{P}}}(x)^2 \mid X, W] - F_{\mathbb{P}}(x)^2) \, dx \, . \end{aligned} \tag{5}$$

Note that

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

Minimizing equation (5) 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

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

□

Proof of Theorem 5. 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 4. *If Z is a random variable defined by*

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

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 4. 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

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

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 6. 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

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

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

Thus,

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

□

Proof of Theorem 7. (1) For any square-integrable function h on $\mathcal{X} = [0, 1]^d$ and $1 \leq i \leq n$,

$$\begin{aligned}
&\frac{1}{r} \mathbb{E} \left(\sum_{j=1}^{r} \bar{W}_{ij}^{(t)} h(\bar{X}_{ij}) \mid \tilde{X}_{1:n}^{(1:t-1)} \right) \\
=&\sum_{j=1}^{r} \int_{H([s_{i,j-1}, s_{i,j}])} \frac{\pi_t \left(\left(\tilde{X}_i^{(1:t-1)}, x \right) \right)}{\pi_{t-1} \left(\tilde{X}_i^{(1:t-1)} \right)} h(x) dx \\
=&\int_{\mathcal{X}} \frac{\pi_t \left(\left(\tilde{X}_i^{(1:t-1)}, x \right) \right)}{\pi_{t-1} \left(\tilde{X}_i^{(1:t-1)} \right)} h(x) dx.
\end{aligned}$$

(2) Let $I \sim \text{Unif}(\{1, 2, \ldots, n\})$ and $J \sim \text{Unif}(\{1, 2, \ldots, r\})$ be independent. Let $Y = \bar{X}_{IJ}^{(t)}$ and $Y' \mid I, J \sim g(\cdot \mid \tilde{X}_I^{(1:t-1)}) \frac{\mathbb{I}(H([s_{I,J-1}, s_{I,J}]))}{1/r}$. It is easy to see that this defines a coupling between the

two distributions at hand.

$$\begin{aligned}
\mathbb{E}[\|Y - Y'\|^p] &= \sum_{i=1}^{n}\sum_{j=1}^{r}\frac{1}{nr}\int_{H([s_{i,j-1},s_{i,j}])} r\|\bar{X}_{ij} - x\|^p g(x \mid \tilde{X}_i^{(1:t-1)})\,dx \\
&\le \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}\int_{H([s_{i,j-1},s_{i,j}])} r(2\sqrt{d+3}|s_{i,j} - s_{i,j-1}|^{1/d})^p g(x \mid \tilde{X}_i^{(1:t-1)})\,dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r}\sum_{j=1}^{r}(2\sqrt{d+3}|s_{i,j} - s_{i,j-1}|^{1/d})^p \\
&= \frac{2^p(d+3)^{p/2}}{n}\sum_{i=1}^{n}\frac{1}{r}\sum_{j=1}^{r}|s_{i,j} - s_{i,j-1}|^{p/d} \\
&\le \begin{cases} 2^p(d+3)^{p/2}r^{-p/d}, & \text{if } p \le d, \\ \dfrac{2^p(d+3)^{p/2}}{r}, & \text{if } p > d. \end{cases}
\end{aligned}$$

Thus,

$$W_p\left(\frac{1}{rn}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{X_{ij}^{(t)}}, \frac{1}{n}\sum_{i=1}^{n} g(\cdot \mid \tilde{X}_i^{(1:t-1)})\right) \le \frac{2\sqrt{d+3}}{r^{1/\max(p,d)}}, \quad a.s.$$

(3) It suffices to show

$$\sum_{j=1}^{r}\mathrm{Var}\left(W_{ij}^{(t)}h(X_{ij}^{(t)}) \mid \tilde{X}_{1:n}^{(1:t-1)}\right) \ge \sum_{j=1}^{r}\mathrm{Var}\left(\bar{W}_{ij}^{(t)}h(\bar{X}_{ij}^{(t)}) \mid \tilde{X}_{1:n}^{(1:t-1)}\right)$$

for any $i = 1, 2, \cdots, n$. Actually,

$$\begin{aligned}
&\frac{1}{r}\,(\text{left hand side} - \text{right hand side}) \\
&= \int_{\mathcal{X}}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)^2 h(x)^2}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)^2 g\left(x|\tilde{X}_{1:n}^{(1:t-1)}\right)}dx - \sum_{j=1}^{r}\int_{H([s_{i,j-1},s_{i,j}])}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)^2 h(x)^2}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)^2 g\left(x|\tilde{X}_{1:n}^{(1:t-1)}\right)}dx \\
&\quad + r\sum_{j=1}^{r}\left(\int_{H([s_{i,j-1},s_{i,j}])}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)}h(x)dx\right)^2 - \left(\int_{\mathcal{X}}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)}h(x)dx\right)^2 \\
&= r\sum_{j=1}^{r}\left(\int_{H([s_{i,j-1},s_{i,j}])}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)}h(x)dx\right)^2 - \left(\int_{\mathcal{X}}\frac{\pi_t\left(\left(\tilde{X}_i^{(1:t-1)}, x\right)\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right)}h(x)dx\right)^2 \\
&\ge 0.
\end{aligned}$$

The last inequality holds by the Cauchy–Schwarz inequality. □

B Simulation details

B.1 Terrain navigation

As it is shown in Section 6.2, the model is defined through

$$X^{(t)} = X^{(t-1)} + d^{(t)} + \epsilon^{(t)}$$
$$Y^{(t)} = h(X^{(t)}) + \delta^{(t)}$$

where $X^{(t)} = (x^{(1t)}, x^{(2t)})$ is the unobserved hidden location, and $Y^{(t)}$ is the observed one-dimensional elevation.

In the simulation study, the drift vectors are set as $d^{(t)} = a^{(t)} - a^{(t-1)}$ for $t = 1, 2, \cdots, 200$, where

$$a^{(t)} = (15000 \sin(\pi t/200, 15000 \cos(\pi t/200))$$

for $t = 0, 1, 2, \cdots, 200$.

The elevator function $h(\cdot)$ is constructed by linear interpolation from *colorado.dat* available at `https://www.stat.colostate.edu/computationalstatistics/`.

The random error in location $\epsilon^{(t)} = (R^{(t)})^\top Z^{(t)}$, where

$$R^{(t)} = \frac{1}{\|X^{(t)}\|_2} \begin{pmatrix} -x^{(1t)} & -x^{(2t)} \\ -x^{(2t)} & x^{(1t)} \end{pmatrix}$$

and

$$Z^{(t)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, q^2 \begin{pmatrix} 1 & 0 \\ 0 & k^2 \end{pmatrix}\right).$$

Here we take $q = 200$ and $k = 1/2$. The measurement errors are $\delta^{(t)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 40^2)$.

B.2 Stochastic volatility

As shown in Section 6.3, the multidimensional stochastic volatility model is

$$X^{(t)} \mid X^{(1:t-1)} \sim \mathcal{N}\left(\alpha X^{(t-1)}, \Sigma\right),$$
$$Y^{(t)} \mid X^{(1:t)}, Y^{(1:t-1)} \sim \mathcal{N}\left(0, \beta^2 \operatorname{diag}(\exp(X^{(t)}))\right),$$

for $t = 1, 2, \cdots, T$. Here we set $\alpha = 0.7$, $\beta = 0.8$, $T = 80$, and $X^{(0)} \sim \mathcal{N}(0, \Sigma)$, where Σ is a $p \times p$ matrix with $\Sigma_{ij} = 0.8^{|i-j|}$. Figure 12 presents extended simulation results. Figure 13 reports the average runtime of each setting.

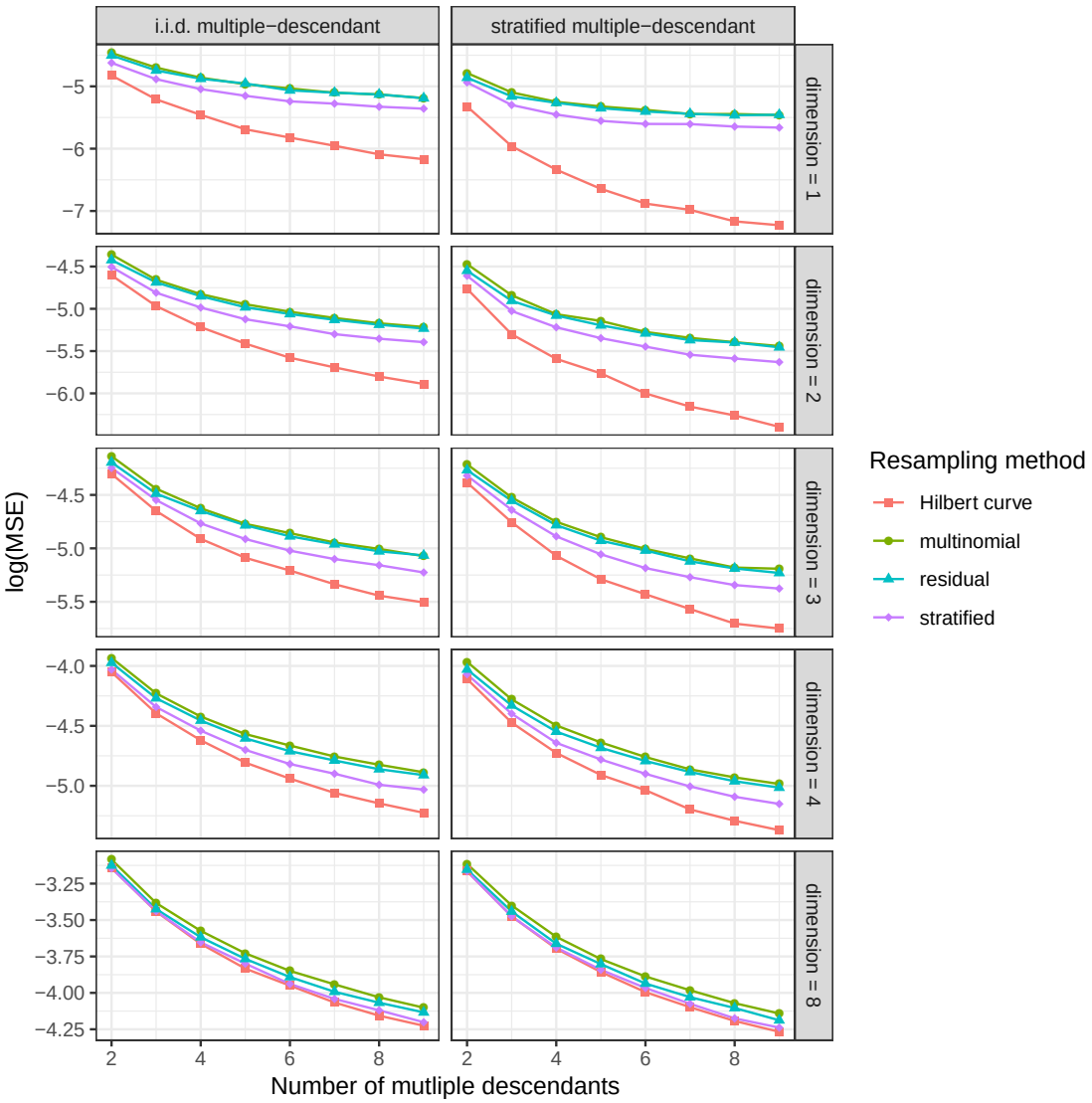


Figure 12: Stochastic volatility extended simulation results.

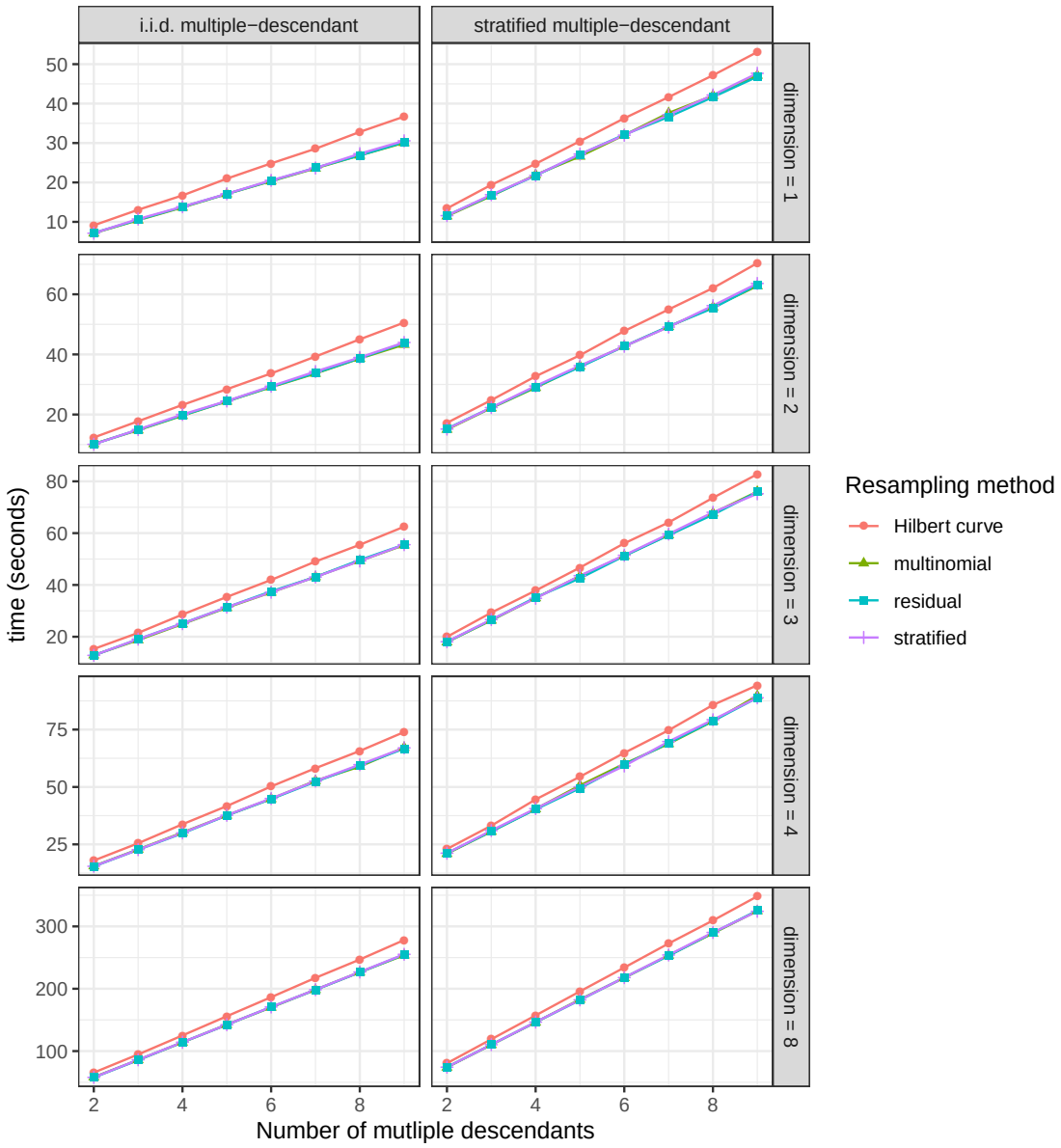


Figure 13: Stochastic volatility extended simulation results (average runtime measured in seconds).

C Weighted resampling

The particles are updated as follows. Sample $\tilde{X}_i^{(t)} \mid \tilde{X}_i^{(1:t-1)} \sim g(\cdot \mid \tilde{X}_i^{(1:t-1)})$ and let

$$\tilde{W}_i^{(t)} = \tilde{W}_i^{(t-1)} \frac{\pi_t\left(\tilde{X}_i^{(1:t)}\right)}{\pi_{t-1}\left(\tilde{X}_i^{(1:t-1)}\right) g\left(\tilde{X}_i^{(t)} \mid \tilde{X}_i^{(1:t-1)}\right)}.$$

We are interested in how $\mathrm{Var}\left[\sum_{i=1}^m \tilde{W}_i^{(t)} \phi(\tilde{X}_i^{(t)}) \mid X^{(1:t-1)}, W^{(t-1)}\right]$ depends on γ.

In the following derivation, we omit the superscripts for time $t-1$ for convenience of notation.

Note that that $a_j = W_j^\gamma / \sum_{k=1}^n W_k^\gamma$.

$$
\begin{aligned}
&\operatorname{Var}\left[\sum_{i=1}^m \tilde{W}_i^{(t)} \phi(\tilde{X}_i^{(t)}) \mid X, W\right] \\
&= m \operatorname{Var}\left[\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid X, W\right] \\
&= m\mathbb{E}\left[\operatorname{Var}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right] + m \operatorname{Var}\left[\mathbb{E}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right] \\
&= m\mathbb{E}\left[\mathbb{E}\left((\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}))^2 \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right] - m\mathbb{E}\left[\left\{\mathbb{E}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right)\right\}^2 \mid X, W\right] \\
&\quad + m\mathbb{E}\left[\left\{\mathbb{E}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right)\right\}^2 \mid X, W\right] - m\left\{\mathbb{E}\left[\mathbb{E}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right]\right\}^2 \\
&= m\mathbb{E}\left[\mathbb{E}\left((\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}))^2 \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right] - m\left\{\mathbb{E}\left[\mathbb{E}\left(\tilde{W}_1^{(t)} \phi(\tilde{X}_1^{(t)}) \mid \tilde{X}_1, \tilde{W}_1\right) \mid X, W\right]\right\}^2 \\
&= m\mathbb{E}\left[\int \frac{\tilde{W}_1^2 \pi_t\left(\left(\tilde{X}_1, x\right)\right)^2 \phi(x)^2}{\pi_{t-1}\left(\tilde{X}_1\right)^2 g\left(x \mid \tilde{X}_1\right)} dx \mid X, W\right] - m\left\{\sum_{j=1}^n \int \frac{W_j \pi_t\left(\left(X_j, x\right)\right) \phi(x)}{\pi_{t-1}(X_j)} dx\right\}^2 \\
&= m\sum_{j=1}^n W_j^{2-\gamma} \int \frac{\pi_t\left(\left(X_j, x\right)\right)^2 \phi(x)^2}{\pi_{t-1}\left(X_j\right)^2 g\left(x \mid X_j\right)} dx \sum_{j=1}^n W_j^\gamma - m\left\{\sum_{j=1}^n \int \frac{W_j \pi_t\left(\left(X_j, x\right)\right) \phi(x)}{\pi_{t-1}(X_j)} dx\right\}^2,
\end{aligned}
$$

where $X = (X_j)_{j=1}^n$ and $W = (W_j)_{j=1}^n$.

By letting $C_j = \int \frac{\pi_t\left(\left(X_j, x\right)\right)^2 \phi(x)^2}{\pi_{t-1}\left(X_j\right)^2 g\left(x \mid X_j\right)} dx$, we have

$$
\operatorname{Var}\left[\sum_{i=1}^m \tilde{W}_i^{(t)} \phi(\tilde{X}_i^{(t)}) \mid X, W\right] = m\sum_{j=1}^n W_j^{2-\gamma} C_j \sum_{j=1}^n W_j^\gamma - m\left\{\sum_{j=1}^n \int \frac{W_j \pi_t\left(\left(X_j, x\right)\right) \phi(x)}{\pi_{t-1}(X_j)} dx\right\}^2.
$$