Free-support Wasserstein Resampling in Sequential Monte Carlo

Mengyu Li[1], Cheng Meng[2*], Jun S. Liu[1*]

[1]Department of Statistics and Data Science, Tsinghua University
[2]Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China

November 11, 2026

Abstract

In modern sequential Monte Carlo (SMC) applications, such as diffusion model samplers, propagating a particle is expensive, so only a few particles can be afforded. In this regime, each resampling step decides which regions of the state space survive, so a better resampling scheme is worth extra computation. However, classical resampling methods only replicate existing particles and suffer from sample impoverishment, especially when the particle size is small relative to the dimension. To create new particles, we relax the unbiasedness of resampling and control the mean squared error via a Wasserstein criterion. Specifically, we propose Wasserstein-Dirichlet resampling (WDR), which constructs an equal-weight measure that optimally approximates the weighted particle distribution by solving a free-support Wasserstein barycenter problem. We then randomize the resulting optimal coupling by Dirichlet mixing. Theoretically, we provide non-asymptotic mean squared error bounds that quantify the bias-variance trade-off and show that the rate depends on the intrinsic dimension. A truncated variant preserves high-weight particles while applying WDR to the remaining mass, targeting mild weight degeneracy. Experiments on state-space models and conditional image generation demonstrate that WDR consistently improves estimation accuracy, stability, and sample diversity. The improvements persist at matched wall-clock cost when the propagation cost dominates.

Keywords: particle filter, state-space model, optimal transport, diffusion model, conditional image generation

*Corresponding author: Cheng Meng (chengmeng@ruc.edu.cn), Jun S. Liu (junsliu@tsinghua.edu.cn)

1 Introduction

Resampling is a fundamental operation that transforms a weighted sample $\{(X_i, W_i)\}_{i=1}^n$ with $\sum_{i=1}^n W_i = 1$ into an equally weighted collection $\{\tilde{X}_j\}_{j=1}^m$. The goal is to ensure that the two discrete distributions are close in some sense, i.e.,

$$\sum_{i=1}^{n} W_i \delta_{X_i} \approx \sum_{j=1}^{m} m^{-1} \delta_{\tilde{X}_j},$$

where δ_x denotes the Dirac measure at x. Resampling plays an important role in a wide range of statistical applications, such as survey sampling (Rao & Wu 1988), causal inference (Morgan & Rubin 2012), and missing data imputation (Rubin 1996). This work focuses on resampling within sequential Monte Carlo (SMC) methods, also known as particle filters, which have emerged as a powerful computational framework for Bayesian inference in dynamical systems (Liu & Chen 1998).

1.1 Background and problem formulation

Sequential Monte Carlo algorithms recursively approximate a sequence of target distributions by maintaining a particle system and updating it through propagation, weighting, and resampling (Doucet & Johansen 2011). Resampling mitigates the progressive concentration of importance weights, known as weight degeneracy (Liu & Chen 1995). Without resampling, the weights become increasingly concentrated on a few particles as iterations proceed, rendering most particles ineffective for downstream estimation and wasting computational resources.

In modern SMC applications, propagating each particle can require substantial time or memory. For example, in diffusion model samplers, every propagation step requires a neural network forward pass for each particle (Wu et al. 2023); in scientific filtering problems, the latent dynamics must be numerically integrated between observation times (van Leeuwen et al. 2019). The number of particles m that can be used is thus severely limited, while

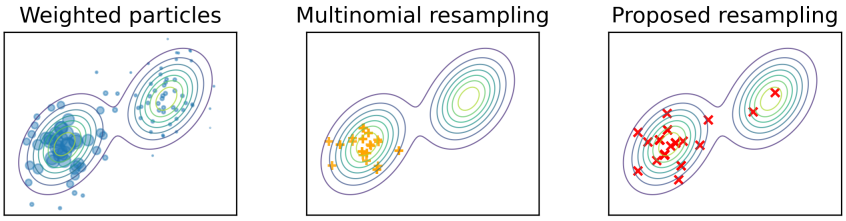


Figure 1: **Left:** weighted particles ($n = 100$) before resampling; marker size is proportional to importance weight, and contour lines depict the target density. **Middle:** equally weighted particles ($m = 30$) produced by multinomial resampling happen to be all located around the dominant high-weight region. **Right:** particles ($m = 30$) produced by the proposed free-support resampling are allowed to relocate, yielding a more spatially dispersed set.

the number required to avoid weight collapse grows exponentially with the effective dimension (Snyder et al. 2008, Bengtsson et al. 2008). With so few particles, it is critical to handle resampling properly.

Various resampling schemes have been developed in the SMC literature. The most straightforward is multinomial resampling (Gordon et al. 1993), which draws particle indices independently from the categorical distribution with probabilities $(W_i)_{i=1}^n$. To reduce the variance introduced by resampling, refined schemes such as residual (Liu & Chen 1998), stratified (Kitagawa 1996), and systematic resampling (Carpenter et al. 1999) have been proposed. These methods are simple to implement, unbiased for arbitrary test functions, and consequently remain widely used in practice.

However, these classical schemes only select among existing particles, i.e., they replicate or discard elements from $\{X_i\}_{i=1}^n$ but never relocate them or create new support points. This limited flexibility leads to sample impoverishment: when a small number of particles carry most of the weight, traditional resampling replicates these particles many times while discarding all others. The resulting particle system concentrates in a narrow region of the state space and fails to represent the global shape of the target distribution; see Fig. 1 (left and middle panels) for a two-dimensional illustration.

To mitigate impoverishment, classical resampling can be augmented with additional ran-

domization, such as adding jitter to resampled particles (Gordon et al. 1993) or performing resample-move MCMC rejuvenation steps (Gilks & Berzuini 2001). While such mechanisms can improve diversity, they do so heuristically and locally. The resampling step itself still depends solely on the weights $\{W_i\}_{i=1}^n$ while leaving particle locations unchanged; any subsequent jitter or MCMC move is applied post-resampling and lacks an explicit global criterion for comparing the weighted empirical measure and its equal-weight approximation within the underlying geometry.

1.2 Related work

The aforementioned limitations have motivated a more recent line of work that directly incorporates geometric information into the resampling step. Existing spatially aware methods fall into three classes: sorting-based, clustering-based, and optimal transport resampling.

Sorting-based methods first order the particles and then apply stratified or systematic resampling to this sequence. In multiple dimensions, the ordering is typically induced by a Hilbert space-filling curve mapping that approximately preserves locality (Gerber & Chopin 2015). Li et al. (2022) established an equivalence between sorted stratified resampling and an optimal transport formulation in one dimension. In higher dimensions, however, the performance depends on the quality of embedding, and the projection may introduce artifacts in highly anisotropic or multimodal settings.

Clustering-based resampling partitions particles into groups using algorithms such as k-means and performs resampling within each cluster (Yang et al. 2019). This prevents clusters with small total weight from being eliminated entirely and encourages local diversity. However, such procedures typically lack theoretical guarantees and may be sensitive to clustering hyperparameters.

The third direction incorporates optimal transport (OT) theory into resampling (Villani 2009, Peyré & Cuturi 2019, Zhang et al. 2021). Reich (2013) proposed the ensemble transform particle filter (ETPF), which computes an OT plan between the weighted empirical measure and an equal-weight empirical measure supported on the original particles, and then uses the associated barycentric map to transform the particles deterministically. Subsequent work developed second-order accurate ETPF (Acevedo et al. 2017) with deterministic corrections to match covariance for bias reduction, and entropy-regularized variants (Kang & Park 2018, Corenflos et al. 2021) to reduce the computational cost. While ETPF preserves geometry better than classical schemes, it optimizes transport only over couplings with fixed supports rather than optimizing the locations of resampled particles. Consequently, the resulting equal-weight empirical measure is not optimal in the Wasserstein sense. Moreover, because both marginals are supported on the same particle set, ETPF is naturally formulated only for $m = n$ and does not accommodate multi-descendant SMC frameworks (Fearnhead & Clifford 2003, Li et al. 2022), where m and n may differ significantly. Finally, this line of work lacks non-asymptotic theoretical results.

1.3 Contributions

Our main idea is to relax the exact unbiasedness imposed by most classical resampling methods and instead control the overall mean squared error (MSE). To this end, we first consider resampling from the point of view of a free-support Wasserstein barycenter problem: given a weighted empirical measure $\mu_n = \sum_{i=1}^n W_i \delta_{X_i}$, we seek an equal-weight measure $\nu_m = m^{-1} \sum_{j=1}^m \delta_{\tilde{X}_j}$ that minimizes the Wasserstein distance $W(\mu_n, \nu_m)$. Unlike fixed-support OT schemes such as ETPF, which constrain both marginals μ_n and ν_m to the original particle support and optimize only over the coupling $\Pi \in \mathbb{R}_+^{n \times m}$, our formulation optimizes jointly over $\Pi \in \mathbb{R}_+^{n \times m}$ and the new support locations $\{\tilde{X}_j\}_{j=1}^m$. Consequently, the resulting equal-weight empirical measure attains the minimum Wasserstein distance to

μ_n among all m-atom equal-weight measures, distributing particles more evenly across the support (see Fig. 1 for illustration).

We next turn this deterministic free-support approximation into a randomized resampling operator. A purely deterministic Wasserstein approximation may underestimate posterior spread, so we propose a Dirichlet mixing mechanism applied to the columns of the optimal coupling. In contrast to a global Dirichlet reweighting on the original support (Rubin 1981, Lyddon et al. 2019), our construction uses the optimal transport plan to build many local Dirichlet mixtures, one per output particle. The resulting method, termed Wasserstein-Dirichlet resampling (WDR), produces a family of randomized resampling schemes interpolating between a deterministic transform and fully randomized ancestor selection, providing a tunable bias-variance trade-off.

Third, we propose some variants and algorithms for SMC implementation. For mild weight-degeneracy regimes, we propose a truncated variant, T-WDR, which preserves a high-weight subset of particles and applies WDR to the remaining mass, thereby allocating more resampling budget to rejuvenate the lower-weight region. Algorithmically, we develop an efficient block coordinate descent method to solve the underlying optimization problem.

Theoretically, we derive non-asymptotic error bounds for the WDR estimator. Under mild regularity conditions, WDR achieves a rate of $\mathcal{O}(m^{-1/s})$ for Lipschitz functions, where s is the intrinsic dimension of the underlying distribution and can be substantially smaller than the ambient dimension d. We also establish the convergence of the proposed block coordinate descent algorithm.

Empirically, we evaluate WDR on linear Gaussian state-space models simulated across a range of dimensions, including a spatial model that requires a block particle filter. Since all of these admit an exact Kalman filter, the Kalman solution provides a clean reference against which the effectiveness of each method can be assessed. Additional nonlinear studies

are reported in the Supplementary Materials. WDR consistently improves filtering accuracy and stability at matched particle sizes, and remains advantageous at matched wall-clock time in propagation-dominating regimes. We further integrate WDR into the twisted diffusion sampler (TDS) (Wu et al. 2023), a state-of-the-art SMC framework for diffusion models. In conditional image generation tasks, WDR consistently produces more diverse and higher-quality samples than baseline resampling strategies.

The rest of the article is organized as follows. Section 2 provides preliminary introductions to the general SMC framework and the Wasserstein barycenter; Section 3 describes our proposed WDR method; Section 4 studies theoretical properties of the proposed method; Section 5 details our simulation evaluations; Section 6 applies WDR to conditional image generation; and Section 7 concludes with a short discussion.

2 Preliminaries

Notations. Superscripts (t) and $[r]$ index timesteps and iterations, respectively; subscripts index particles. Temporal superscripts are omitted when clear from context. We use n for the number of weighted particles and m for the resampled size. For vectors, $\|\cdot\|$ denotes the Euclidean norm. For a measurable space $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ denotes the set of Borel probability measures on $\mathcal{X}$, and $\delta_x \in \mathcal{P}(\mathcal{X})$ denotes the Dirac measure at $x \in \mathcal{X}$. We write 1_m for the m-vector of ones, and $\Delta^{n-1} := \{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\}$ for the probability simplex. For two nonnegative sequences $(x_n)_n$ and $(y_n)_n$, we write $x_n = \mathcal{O}(y_n)$ if there exists a constant $c > 0$ such that $x_n \leq c y_n$, and $x_n = \tilde{\mathcal{O}}(y_n)$ if there exist constants $c, c' > 0$ such that $x_n \leq c' y_n (\log(n))^c$. Sections numbered with the prefix S refer to the Supplementary Materials.

2.1 Sequential Monte Carlo

Consider a state-space model observed at T timesteps, with a latent process $(X^{(t)})_{0\le t\le T}$ taking values in $\mathcal{X}$ and observations $(Y^{(t)})_{1\le t\le T}$ taking values in $\mathcal{Y}$. The latent process evolves from an initial probability density $\mu_0 \in \mathcal{P}(\mathcal{X})$ according to Markov transition kernels $f_t(x \mid x')$, $1 \le t \le T$, where $f_t(x \mid x')$ denotes the probability density associated with moving from x' to x. Conditional on $(X^{(t)})_{0\le t\le T}$, the observations $(Y^{(t)})_{1\le t\le T}$ are statistically independent and their marginal densities are given by $g_t(y \mid x)$. That is,

$$\begin{aligned} X^{(0)} &\sim \mu_0, \\ X^{(t)} \mid (X^{(t-1)} = x^{(t-1)}) &\sim f_t(\cdot \mid x^{(t-1)}), \quad 1 \le t \le T, \\ Y^{(t)} \mid (X^{(t)} = x^{(t)}) &\sim g_t(\cdot \mid x^{(t)}), \quad 1 \le t \le T. \end{aligned} \tag{1}$$

Given observations $y^{(1:t)} = (y^{(1)}, \dots, y^{(t)})$, the filtering problem primarily focuses on the target distribution $\pi_t(x) = \mathbb{P}(X^{(t)} \in \mathrm{d}x \mid Y^{(1:t)} = y^{(1:t)})$ and expectations $\pi_t(\varphi) = \mathbb{E}[\varphi(X^{(t)}) \mid Y^{(1:t)} = y^{(1:t)}]$ for some suitable test function φ.

Sequential Monte Carlo methods approximate π_t by a random empirical measure

$$\pi_n^{(t)} = \sum_{i=1}^{n} W_i^{(t)} \, \delta_{X_i^{(t)}},$$

where $\{X_i^{(t)}\}_{i=1}^n$ are particles and $\{W_i^{(t)}\}_{i=1}^n$ are normalized importance weights. Given a particle set $\{(\tilde{X}_i^{(t-1)}, \tilde{W}_i^{(t-1)})\}_{i=1}^m$, an SMC update at time t proceeds as follows (Doucet & Johansen 2011):

1. **Propagation.** For each ancestor particle $i = 1, \dots, m$, simulate $K_t \ge 1$ descendant particle(s)

$$X_{i,k}^{(t)} \sim q_t(\cdot \mid \tilde{X}_i^{(t-1)}), \quad k = 1, \dots, K_t,$$

where q_t is a proposal kernel; for example, $q_t = f_t$ in the bootstrap filter. The total number of candidate particles at time t is mK_t.

2. **Weighting.** Update the importance weights according to the observation at time t:

$$W_{i,k}^{(t)} = \frac{\bar{W}_{i,k}^{(t)}}{\sum_{i=1}^{m}\sum_{k=1}^{K_t}\bar{W}_{i,k}^{(t)}}, \quad \text{where} \quad \bar{W}_{i,k}^{(t)} \propto W_{i,k}^{(t-1)}\,\frac{g_t(Y^{(t)} \mid X_{i,k}^{(t)})\, f_t(X_{i,k}^{(t)} \mid \tilde{X}_{i,k}^{(t-1)})}{q_t(X_{i,k}^{(t)} \mid \tilde{X}_{i,k}^{(t-1)})}.$$

   For notational convenience, we relabel the subscripts $\{(i,k)\}_{i=1,\ldots,m,\,k=1,\ldots,K_t}$ as $\{i\}_{i=1,\ldots,n}$ with $n = mK_t$.

3. **Resampling.** From $\{(X_i^{(t)}, W_i^{(t)})\}_{i=1}^n$, generate a set $\{\tilde{X}_i^{(t)}\}_{i=1}^m$ with equal weight, $\tilde{W}_i^{(t)} = 1/m$, using some resampling scheme. We denote this operation by

$$\{\tilde{X}_i^{(t)}\}_{i=1}^m \sim \mathcal{R}_m\Big(\{(X_i^{(t)}, W_i^{(t)})\}_{i=1}^n\Big),$$

   where $\mathcal{R}_m$ represents a resampling operator. We keep m fixed over time to ensure that the particle size does not grow uncontrollably over time.

In practice, the multi-descendant proposal and resampling steps can be adapted based on the coefficient of variation (CV) of the weights. The CV of a weight vector $W = (W_1, \ldots, W_n)$ is defined as the ratio of its standard deviation to its mean, i.e., $\mathrm{cv}^2(W) = n^{-1}\sum_{i=1}^n (W_i - \bar{W})^2/\bar{W}^2$, which is equal to $n\sum_{i=1}^n W_i^2 - 1$ if W is normalized. In the survey literature, researchers often compute the effective sample size (ESS) as $\mathrm{ess}(W) = n/(1+\mathrm{cv}^2(W))$ for inference with weighted samples, which was also suggested by Kong et al. (1994) for the SMC framework (Liu & Chen 1995, 1998). As such, CV can be used to provide a general measure of the degeneracy of a particle set, decoupled from the sample size n.

At timestep t, we compute $\mathrm{cv}(W^{(t-1)})$ and compare it with a predefined threshold $c_{\text{thres}} \geq 0$. If $\mathrm{cv}(W^{(t-1)}) > c_{\text{thres}}$, we regard the particle system as strongly degenerate and activate the multi-descendant proposal with $K_t = K > 1$; otherwise we set $K_t = 1$ to save computational cost. After the weighting step at time t, we compute $\mathrm{cv}(W^{(t)})$ and perform resampling if either $K_t > 1$ or $\mathrm{cv}(W^{(t)}) > c_{\text{thres}}$; otherwise we skip resampling and carry the weighted particles forward to the next timestep.

In what follows, we regard resampling as a stand-alone transformation that takes a weighted

empirical measure μ_n and returns a random equal-weight measure ν_m. We say that a resampling operator $\mathcal{R}_m$ is *unbiased* if

$$\mathbb{E}\Big[\nu_m(\varphi) \mid \mu_n\Big] = \mu_n(\varphi), \quad \forall\, \varphi \in \mathcal{F},$$

where $\mathcal{F}$ denotes the set of all finite functions with respect to μ_n. Classical resampling schemes (e.g., multinomial, residual, stratified) are unbiased, while the method developed in this study trades off unbiasedness for improving geometric properties quantified via Wasserstein distance.

2.2 Wasserstein distance and barycenter

For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, a coupling (or transport plan) between μ and ν is a probability measure γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, i.e.,

$$\gamma(A \times \mathbb{R}^d) = \mu(A), \quad \gamma(\mathbb{R}^d \times B) = \nu(B), \quad \forall\, A, B \in \mathcal{B}(\mathbb{R}^d),$$

where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-algebra. The set of all couplings between μ and ν is denoted by $\Gamma(\mu, \nu)$. For $p \geq 1$, the p-Wasserstein distance (Villani 2009, Peyré & Cuturi 2019) between μ and ν is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}.$$

When μ and ν are discrete measures, the OT problem reduces to a finite-dimensional linear program. For weighted empirical measures $\mu_n = \sum_{i=1}^n W_i \, \delta_{X_i}$ and $\nu_m = \sum_{j=1}^m V_j \, \delta_{Y_j}$ with weights $W \in \Delta^{n-1}$ and $V \in \Delta^{m-1}$, any coupling $\gamma \in \Gamma(\mu_n, \nu_m)$ can be represented by a nonnegative matrix $\Pi \in \mathcal{U}(W, V)$, where

$$\mathcal{U}(W, V) = \left\{ \Pi \in \mathbb{R}_+^{n \times m} : \sum_{j=1}^m \Pi_{ij} = W_i, \quad \sum_{i=1}^n \Pi_{ij} = V_j \right\}.$$

The p-Wasserstein distance is then given by

$$W_p^p(\mu_n, \nu_m) = \min_{\Pi \in \mathcal{U}(W,V)} \sum_{i=1}^n \sum_{j=1}^m \Pi_{ij} \, \|X_i - Y_j\|^p.$$

Wasserstein barycenters. Given a collection of probability measures $\mu^{(1)}, \ldots, \mu^{(L)}$ on $\mathbb{R}^d$ and nonnegative weights $\lambda_1, \ldots, \lambda_L$ with $\sum_{l=1}^{L} \lambda_l = 1$, a Wasserstein barycenter of $\{\mu^{(l)}\}_{l=1}^{L}$ is the minimizer

$$\nu^\star = \arg\min_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{l=1}^{L} \lambda_l \, W_p^p(\nu, \mu^{(l)}).$$

This construction generalizes the notion of an average (or Fréchet mean) to the Wasserstein space. In the *free-support* setting, ν is allowed to be any probability measure in $\mathcal{P}(\mathbb{R}^d)$, so both its support and weights are determined by the optimization problem (Claici et al. 2018, Cuturi & Doucet 2014). In the *fixed-support* setting, both ν and the input measures $\mu^{(l)}$ are discrete measures supported on a prescribed common support set. In this case, the support locations are fixed and only the weights on that common support are optimized (Benamou et al. 2015, Lin et al. 2020). Our resampling formulation corresponds to a constrained free-support barycenter in which ν is required to be an equal-weight discrete measure with flexible support locations.

3 Method

3.1 Free-support Wasserstein resampling with Dirichlet mixing

Recall that $\mu_n = \sum_{i=1}^{n} W_i \delta_{X_i}$ is a weighted empirical measure obtained after the weighting step of SMC, and $\nu_m = m^{-1} \sum_{i=1}^{m} \delta_{\tilde{X}_i}$ is the equal-weight measure returned by a resampling operator. For a test function $\varphi : \mathcal{X} \to \mathbb{R}$, the resampling estimator of $\mu_n(\varphi) = \sum_{i=1}^{n} W_i \varphi(X_i)$ is $\nu_m(\varphi) = m^{-1} \sum_{i=1}^{m} \varphi(\tilde{X}_i)$. Classical schemes are designed to satisfy $\mathbb{E}[\nu_m(\varphi) \mid \mu_n] = \mu_n(\varphi)$ while reducing the conditional variance $\mathrm{Var}\{\nu_m(\varphi) \mid \mu_n\}$ (Liu & Chen 1995, Kitagawa 1996, Li et al. 2022).

However, exact unbiasedness for all test functions constrains resampled particles to the original support. If new locations were created with positive probability, one could always construct a function φ separating the new support from the old, yielding $\mathbb{E}[\nu_m(\varphi) \mid \mu_n] \neq$

$\mu_n(\varphi)$. To enable free supports, we relax exact unbiasedness and instead control the conditional MSE $\mathbb{E}[(\nu_m(\varphi) - \mu_n(\varphi))^2 \mid \mu_n]$. For Lipschitz test functions, this MSE is controlled by the p-Wasserstein distance between μ_n and ν_m, as established in Proposition 1.

Proposition 1. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz, i.e. $|\varphi(x) - \varphi(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$. Then, for any $p \geq 1$,*

$$\mathbb{E}\Big[(\nu(\varphi) - \mu(\varphi))^2 \mid \mu\Big] \;\leq\; L^2\, \mathbb{E}\Big[W_p^2(\mu, \nu) \mid \mu\Big]. \tag{2}$$

Inequality (2) motivates bringing ν_m as close as possible to μ_n in the W_p distance. Throughout the remainder of the paper, we focus on $p = 2$ and write $W(\mu, \nu) = W_2(\mu, \nu)$.

Free-support Wasserstein approximation. Define the set of equal-weight discrete measures with m atoms as $\mathcal{A}_m = \{m^{-1} \sum_{j=1}^m \delta_{Y_j} : Y_j \in \mathbb{R}^d\}$. Motivated by Proposition 1, our target resampled measure is the best W_2-approximation of μ_n in $\mathcal{A}_m$:

$$\nu_m^\star = \arg\min_{\nu_m \in \mathcal{A}_m} W^2(\mu_n, \nu_m). \tag{3}$$

This is a constrained free-support Wasserstein barycenter of the single input measure μ_n, restricted to $\mathcal{A}_m$. Using the discrete formulation in Section 2, problem (3) is equivalent to

$$\min_{Y \in (\mathbb{R}^d)^m} \; \min_{\Pi \in \mathcal{U}(W, m^{-1}1_m)} \sum_{i=1}^n \sum_{j=1}^m \Pi_{ij} \, \|X_i - Y_j\|^2, \tag{4}$$

where $W = (W_1, \ldots, W_n)$ are the input weights and $\mathcal{U}(W, m^{-1}1_m)$ is the associated transportation polytope. The minimizer $(\Pi^\star, Y^\star)$ of (4) yields $\nu_m^\star = m^{-1} \sum_{j=1}^m \delta_{Y_j^\star}$, which is the closest distribution to μ_n in the W_2 metric within the class $\mathcal{A}_m$.

Furthermore, $Y^\star$ can be expressed in terms of the columns of $\Pi^\star$. For each $j = 1, \ldots, m$, define $p_i^{(j)} := m\, \Pi_{ij}^\star$, $i = 1, \ldots, n$, so that $\sum_i p_i^{(j)} = 1$ and $p^{(j)} = (p_1^{(j)}, \ldots, p_n^{(j)})$ is a probability vector. Then,

$$Y_j^\star = \frac{\sum_{i=1}^n \Pi_{ij}^\star X_i}{\sum_{i=1}^n \Pi_{ij}^\star} = \sum_{i=1}^n p_i^{(j)} X_i, \quad j = 1, \ldots, m. \tag{5}$$

That is, each $Y_j^\star$ is the barycenter of the input particles under the column weights $p^{(j)}$.

The transformation $\mu_n \mapsto \nu_m^\star$ enjoys optimal geometric properties in the Wasserstein sense, but it is fully deterministic given μ_n and tends to underestimate particle spread (Acevedo et al. 2017). To address this limitation, we introduce controlled stochasticity to $\nu_m^\star$ via a Dirichlet mixing mechanism.

Dirichlet mixing randomization. Given $(\Pi^\star, Y^\star)$, we now define the Wasserstein-Dirichlet resampling (WDR) operator. For $j = 1, \ldots, m$, recall the probability vector $p^{(j)} \in \Delta^{n-1}$ and the barycentric projection $Y_j^\star = \sum_{i=1}^n p_i^{(j)} X_i$. Given a concentration parameter $\tau \geq 0$, we define the Dirichlet-mixed resampled particles as

$$Z_j(\tau) := \sum_{i=1}^n V_i^{(j)} X_i, \quad \text{where} \quad V^{(j)} \sim \text{Dirichlet}(\tau p^{(j)}), \quad j = 1, \ldots, m, \tag{6}$$

independently across j conditional on $(\Pi^\star, X)$. Collecting $\tilde{X}_j := Z_j(\tau)$, $j = 1, \ldots, m$, and assigning equal weights $1/m$ yields the equal-weight empirical measure ν_m^{WDR}.

This construction interpolates continuously between deterministic and fully randomized behavior. As $\tau \to \infty$, the Dirichlet weights $V^{(j)}$ concentrate around their mean $p^{(j)}$, and $Z_j(\tau) \to Y_j^\star$ in probability; in this limit WDR recovers the deterministic map $\mu_n \mapsto \nu_m^\star$ defined by (3). At the other extreme, as $\tau \to 0^+$, the Dirichlet distribution with parameter $\tau p^{(j)}$ concentrates on the vertices of the probability simplex. In the limit case of $\tau = 0$, we obtain $I_j \sim \text{Categorical}(p^{(j)})$ and $Z_j(0) = X_{I_j}$, i.e., randomized resampling based on the optimal plan $\Pi^\star$, which is unbiased. For $d = 1$, the optimal plan $\Pi^\star$ corresponds to a monotone transport map and $Z_j(0)$ reduces to sorted stratified resampling (Li et al. 2022).

For any fixed $\tau > 0$, the weights $V^{(j)}$ have full support on the simplex, and $Z_j(\tau)$ lies in the convex hull of the particles $\{X_i : p_i^{(j)} > 0\}$. The parameter τ thus trades off Monte Carlo variability and shrinkage toward the barycentric projections $Y_j^\star$.

Remark 1. The discrete OT plan $\Pi^\star$ is sparse with at most $n+m-1$ nonzero entries (Gerber & Maggioni 2017), so each column $p^{(j)}$ involves only $\mathcal{O}(1)$ input particles on average. The Dirichlet mixing in (6) is therefore inherently local: each resampled particle combines only

a small number of nearby ancestors rather than aggregating mass from the entire set. This locality preserves local structures and avoids excessive collapse of particles into a single location.

Remark 2. The Dirichlet mixing (6) is related in spirit to the Bayesian bootstrap (Rubin 1981, Lyddon et al. 2019), which also uses Dirichlet weights to randomize an empirical measure. In the Bayesian bootstrap, one draws $V \sim \text{Dirichlet}(1_n)$ on the simplex Δ^{n-1} and forms $\sum_{i=1}^n V_i \delta_{X_i}$ as a random empirical measure. The key difference is that, instead of employing a global Dirichlet weighting on the original support, WDR uses the optimal transport plan to construct many local Dirichlet mixtures, one for each output particle.

Truncated Wasserstein-Dirichlet resampling. In multi-descendant propagation, each ancestor generates multiple candidate descendants that probe the future independently. Hence, when weight degeneracy is mild, an additional resampling step to the entire population can waste budget by allocating multiple equal-weight output particles to essentially the same high-weight region. We instead keep high-weight particles fixed and devote more budget to rejuvenate medium to low-weight regions; related thresholding ideas have appeared in, e.g., Liu et al. (1998) and Fearnhead & Clifford (2003). This leads to truncated WDR (T-WDR), a variant of WDR that preserves high-weight particles as anchors and reallocates only the residual mass, thereby retaining the dominant structure while providing sufficient opportunity for tail rejuvenation. T-WDR shares the same objective (3) as WDR, and differs only in the feasible class $\mathcal{A}_m$. Given a threshold $\kappa > 0$, define $S_\kappa = \{i : W_i \geq \kappa\}$, $r_\kappa = \sum_{i \notin S_\kappa} W_i$, and $m_{\text{tail}} = m - |S_\kappa|$. We then restrict the optimization to $\mathcal{A}_m^{\text{T-WDR}} = \{\sum_{i \in S_\kappa} W_i \delta_{X_i} + r_\kappa m_{\text{tail}}^{-1} \sum_{j=1}^{m_{\text{tail}}} \delta_{Y_j} : Y_j \in \mathbb{R}^d\}$. Under a continuity condition, the threshold κ is characterized as the solution to $m = \sum_{j=1}^n \min(W_j/\kappa, 1)$. Further details are provided in Section S1. We recommend T-WDR as an efficient surrogate for WDR when weight degeneracy is mild; under severe degeneracy, full WDR is generally preferable.

3.2 Block coordinate descent algorithm

In this section, we develop a numerical framework for efficient computation of the problem (4). Although many numerical solvers for Wasserstein barycenters have been proposed, most focus on fixed-support formulations or regularized variants (Benamou et al. 2015, Lin et al. 2020). For the free-support setting, existing algorithms are either not directly compatible with the equal-weight constraint in resampling (Cuturi & Doucet 2014), or are general-purpose solvers (Claici et al. 2018) that can be prohibitively slow in our regime because the barycenter subproblem must be solved repeatedly within SMC.

Therefore, we propose an efficient algorithm tailored to the structure of (4). The objective in (4) is convex in Π for fixed Y and convex in Y for fixed Π, yielding a biconvex structure. We exploit this with a block coordinate descent (BCD) procedure (Bertsekas 2016) that alternates between updating the transport plan and the support locations. In particular, at iteration r, the BCD scheme performs the following two steps.

1. **Transport step.** Given $Y^{[r]}$, compute the optimal transport plan

$$\Pi^{[r+1]} = \arg\min_{\Pi \in \mathcal{U}(W, m^{-1}1_m)} \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{ij} \|X_i - Y_j^{[r]}\|^2. \tag{7}$$

   This is a standard discrete OT problem with fixed supports X and $Y^{[r]}$.

2. **Support step.** Given $\Pi^{[r+1]}$, update the support locations $Y^{[r+1]}$ by minimizing the objective with respect to Y:

$$Y^{[r+1]} = \arg\min_{Y \in (\mathbb{R}^d)^m} \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{ij}^{[r+1]} \|X_i - Y_j\|^2.$$

   This separates across columns and admits the closed-form update as in (5), that is,

$$Y_j^{[r+1]} = \frac{\sum_{i=1}^{n} \Pi_{ij}^{[r+1]} X_i}{\sum_{i=1}^{n} \Pi_{ij}^{[r+1]}}, \qquad j = 1, \ldots, m, \tag{8}$$

   so each $Y_j^{[r+1]}$ is the barycenter of X under the jth column of $\Pi^{[r+1]}$.

Algorithm 1 Wasserstein-Dirichlet resampling (WDR)

Input: weighted particles $\{(X_i, W_i)\}_{i=1}^n$, target size m, concentration parameter $\tau \geq 0$

Compute $(\Pi^\star, Y^\star)$ by solving (4) via block coordinate descent:

  Initialize $Y^{[0]} = (Y_1^{[0]}, \ldots, Y_m^{[0]})$, set $r \leftarrow 0$

  repeat

    Update $\Pi^{[r+1]}$ by solving (7) with $Y^{[r]}$

    Update $Y^{[r+1]}$ using (8) with $\Pi^{[r+1]}$

    $r \leftarrow r + 1$

  until convergence criterion is met

  $(\Pi^\star, Y^\star) \leftarrow (\Pi^{[r]}, Y^{[r]})$

for $j = 1, \ldots, m$ **do**

  Set $p^{(j)} \leftarrow m\,\Pi^\star_{\cdot j}$

  Draw $V^{(j)} \sim \mathrm{Dirichlet}(\tau p^{(j)})$ and set $\tilde{X}_j \leftarrow \sum_{i=1}^n V_i^{(j)} X_i$

end for

Output: equal-weight particles $\{\tilde{X}_j\}_{j=1}^m$

Each iteration decreases the objective in (4), and the sequence $\{(\Pi^{[r]}, Y^{[r]})\}$ has convergence guarantees (see Section 4.2). In practice, we terminate the iterations when the relative change in Y falls below a small tolerance, or after a fixed maximum number of iterations. In our experiments, a small number of iterations (typically fewer than ten) is sufficient to reach numerical convergence.

Equipped with the BCD solver, the complete Wasserstein-Dirichlet resampling procedure is summarized in Algorithm 1. In the SMC context of Section 2, we use Algorithm 1 as a drop-in replacement for the generic resampling operator $\mathcal{R}_m$, with m chosen according to the desired number of descendants.

In Algorithm 1, we initialize $Y^{[0]}$ by subsampling m particles at random. Regarding the subproblems, when n is moderate, we solve the transport subproblem (7) exactly using the network flow method (Flamary et al. 2021). For larger n, we use the sparsified approximation solver proposed by Li et al. (2023), which achieves near-linear time complexity. Due to the sparsity of the OT plan Π with $\tilde{\mathcal{O}}(n)$ nonzero entries, the barycentric update (8) can

also be implemented in $\tilde{\mathcal{O}}(n)$ time. Thus, each BCD iteration has near-linear complexity $\tilde{\mathcal{O}}(n)$. Since the BCD iterations converge within a small number of iterations (as shown in Section S3.1), the overall cost of WDR scales approximately linearly in n, comparable to that of classical resampling methods, while exploiting richer geometric structure.

4 Theoretical results

4.1 Statistical properties

We first quantify the approximation error of the optimal equal-weight atomic approximation to a given measure, then analyze the bias-variance trade-off of the Wasserstein-Dirichlet resampling.

Lemma 1. *Let $s \in \mathbb{N}$ be the intrinsic dimension of the support of μ in the following sense: after a rigid change of coordinates in $\mathbb{R}^d$, the support of μ is contained in an axis-aligned hyperrectangle $Q_s = \prod_{k=1}^{s}[a_k, b_k] \times \{0\}^{d-s} \subset \mathbb{R}^d$ with side lengths $\ell_k := b_k - a_k > 0$ for $k = 1, \ldots, s$. Let $D := (\sum_{k=1}^{s} \ell_k^2)^{1/2}$ be the Euclidean diameter of the active block $\prod_{k=1}^{s}[a_k, b_k]$. Then for any $m \in \mathbb{N}$ and any $p < s$, there exists an equal-weight atomic measure $\nu_m := m^{-1} \sum_{j=1}^{m} \delta_{y_j}$ with $y_1, \ldots, y_m \in Q_s$ such that*

$$W_p(\mu, \nu_m) \leq C_{s,p}\, D\, m^{-1/s},$$

where $C_{s,p}$ depends only on s and p.

This result extends standard quantization rates (Gray & Neuhoff 1998, Kloeckner 2012) to the constrained setting of equal-weight discrete approximation, preserving the polynomial decay in the intrinsic dimension s. While Lemma 1 assumes an s-dimensional axis-aligned support for brevity, the convergence rate $W_p(\mu, \nu_m) \lesssim m^{-1/s}$ can extend to more general structures, such as s-dimensional manifolds (Weed & Bach 2019, Hundrieser et al. 2024), up to a change in the constant factor.

This viewpoint is motivated by the fact that, in many filtering and generative problems, the target distribution is far from isotropic in the ambient space. Strong dependence or latent geometric structure might cause the posterior to concentrate near a lower-dimensional set or along a small number of effective directions. For example, although image generation problems are formulated in high-dimensional pixel space, natural images are commonly believed to lie on a much lower-dimensional manifold. In such cases, geometry-aware resampling is expected to be advantageous. We now state a non-asymptotic bound for the Wasserstein-Dirichlet resampling estimator of L-Lipschitz functions.

Theorem 1. *Let $\varphi : [0,1]^d \to \mathbb{R}$ be L-Lipschitz and let $\mu_n(\varphi) := \sum_{i=1}^n W_i \varphi(X_i)$ be the target function. Define the Wasserstein-Dirichlet resampling estimator $\nu_m^{\mathrm{WDR}}(\varphi) := m^{-1} \sum_{j=1}^m \varphi(Z_j(\tau))$, where $\{Z_j(\tau)\}_{j=1}^m$ are the resampled particles constructed in Section 3 with concentration parameter $\tau \geq 0$. Then, under the assumption of Lemma 1,*

$$\mathbb{E}\left[\left(\nu_m^{\mathrm{WDR}}(\varphi) - \mu_n(\varphi)\right)^2 \mid \mu_n\right] \leq L^2 C_s^2 \left[\frac{\tau}{\tau+1}\, m^{-2/s} + \frac{1}{\tau+1}\, m^{-1-2/s}\right], \tag{9}$$

where s is the intrinsic dimension under the structural condition of Lemma 1, and the constant C_s depends only on s.

The right-hand side of (9) arises from a bias-variance decomposition of MSE. The first term $\mathcal{O}(m^{-2/s})$, weighted by $\tau/(\tau+1)$, corresponds to the squared bias from the free-support Wasserstein approximation. As $\tau \to \infty$, WDR becomes the deterministic barycentric transform, so only this bias term remains. The second term $\mathcal{O}(m^{-1-2/s})$, weighted by $1/(\tau+1)$, corresponds to the conditional variance from Dirichlet mixing. When $\tau = 0$, WDR reduces to ancestor sampling according to the optimal transport plan, and then the bias term vanishes, leaving only the variance term. Especially, for linear test functions φ, the bias term vanishes, so the MSE reduces to the variance term $\mathcal{O}(m^{-1-2/s})$ for all $\tau \geq 0$.

It is also worth noting that the variance rate $\mathcal{O}(m^{-1-2/s})$ matches that of Hilbert curve resampling (Li et al. 2022), which samples ancestors according to a transport plan constructed via Hilbert curve projections. In this sense, WDR generalizes the Hilbert curve resampling

by providing a continuous interpolation between deterministic and randomized schemes, thereby balancing estimation bias and variance.

4.2 Algorithmic convergence

Let $\{(\Pi^{[r]}, Y^{[r]})\}_{r\geq 0}$ be the sequence produced by the BCD part of Algorithm 1, i.e., alternating exact minimization over Π and Y. The next result establishes a well-behaved convergence guarantee.

Theorem 2. *Denote $F(\Pi, Y)$ as the objective function in (4). Recall that, for each $r \geq 0$,*

$$\Pi^{[r+1]} \in \arg\min_{\Pi \in \mathcal{U}(W, m^{-1}\mathbf{1}_m)} F(\Pi, Y^{[r]}), \quad Y^{[r+1]} = \arg\min_{Y \in (\mathbb{R}^d)^m} F(\Pi^{[r+1]}, Y).$$

Then, every limit point of $\{(\Pi^{[r]}, Y^{[r]})\}$ is a stationary point of the problem (4).

The monotone decrease property used in the convergence analysis is established in Lemma S3 of the Supplementary Materials, while Theorem 2 guarantees the stationarity of all limit points. Since F is nonconvex, the theorem does not assert global optimality without additional regularity. Empirically, the BCD solver is numerically stable for the resampling problem and shows little sensitivity to initialization, as shown in Section S3.1.

5 Simulation study

This section evaluates the proposed resampling scheme on linear Gaussian state-space models. Section 5.1 compares accuracy and stability across dimensions and condition numbers at matched particle sizes, and then identifies the propagation cost above which WDR is preferred at matched wall-clock time. Section 5.2 scales the comparison to a high-dimensional spatial model, which requires a block particle filter and has a high propagation cost. Two nonlinear benchmarks, the Lorenz-63 system and a stochastic volatility model, are provided in Sections S3.2 and S3.4.

We compare the proposed (T-)WDR with classical multinomial, residual, and stratified resampling, and with the OT-based Hilbert curve resampling (Li et al. 2022). Another OT-based approach, ETPF (Reich 2013), is not included in the comparison because it is incompatible with the multi-descendant SMC framework. We set the Dirichlet concentration parameter τ to 30 as the default. For the transport subproblem in Algorithm 1, we solve the OT problem exactly using a network flow solver (Flamary et al. 2021) when $m \leq 200$, covering the Lorenz-63 study; while for larger m, such as those considered in this section, we use a sparsified proximal point method (Xie et al. 2020, Li et al. 2023) to approximate the transport plan.

Evaluation metrics. The primary goal is to estimate the posterior mean trajectory $I^\star = (I^{\star(1)}, \ldots, I^{\star(T)})$, where $I^{\star(t)}$ denotes the oracle value of $\mathbb{E}[X^{(t)} \mid Y^{(1:t)}]$ under the state-space model (1). For the linear Gaussian models, the oracle is available in closed form from the Kalman filter.

We consider three metrics to evaluate each resampling method. At time t, a given method produces a weighted particle approximation $\{(X_i^{(t)}, W_i^{(t)})\}_i$ to the filtering distribution. We compute the weighted particle mean $\hat{I}^{(t)} := \sum_i W_i^{(t)} X_i^{(t)}$ and evaluate its discrepancy from the oracle posterior mean $I^{\star(t)}$ using $\mathrm{RMSE}_t = \|\hat{I}^{(t)} - I^{\star(t)}\|/\sqrt{d}$, and then report the time-averaged RMSE over the T timesteps, $\mathrm{RMSE} = T^{-1} \sum_{t=1}^T \mathrm{RMSE}_t$.

We also evaluate the filtering distribution using the energy score (ES), a strictly proper scoring rule (Acevedo et al. 2017). At time t, given the weighted particle approximation $\sum_i W_i^{(t)} \delta_{X_i^{(t)}}$ and the oracle posterior mean $I^{\star(t)}$, the energy score is defined as

$$\mathrm{ES}_t = \sum_{j=1}^n W_j^{(t)} \|X_j^{(t)} - I^{\star(t)}\| - \frac{1}{2} \sum_{j=1}^n \sum_{l=1}^n W_j^{(t)} W_l^{(t)} \|X_j^{(t)} - X_l^{(t)}\|. \tag{10}$$

We then compute its time-averaged version $\mathrm{ES} = T^{-1} \sum_{t=1}^T \mathrm{ES}_t$. The first term in (10) measures the proximity to the oracle state, and the second term accounts for the ensemble spread. Thus, ES penalizes under-dispersion caused by the particle impoverishment. For the

linear Gaussian models, we additionally compare the particle approximation with the exact filtering distribution using the energy distance (ED); details are provided in Section S3.7. For all the three metrics, lower values indicate better performance.

5.1 Linear Gaussian state-space model

We consider a linear Gaussian state-space model, which can be viewed as discrete observations at unit time intervals on the latent continuous-time Ornstein–Uhlenbeck (OU) diffusion, $\mathrm{d}X_u = BX_u\,\mathrm{d}u + C\,\mathrm{d}W_u$, with W_u being a standard d-dimensional Brownian motion:

$$X^{(t)} = AX^{(t-1)} + \xi^{(t)}, \quad \xi^{(t)} \sim \mathcal{N}(0, \Sigma),$$
$$Y^{(t)} = X^{(t)} + \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim \mathcal{N}(0, I_d), \quad t = 1, 2, \ldots, T.$$

The process is initialized at $X^{(0)} \sim \mathcal{N}(0, I_d)$, where $A = e^B$ and $\Sigma = \int_0^1 e^{Bs} CC^\top e^{B^\top s}\,\mathrm{d}s$. We take $A = 0.95\, I_d$ and $\Sigma = 0.5\, Q\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\, Q^\top$, where Q is a fixed random orthogonal matrix and $\lambda_j = \kappa^{-(j-1)/(d-1)}$, so the eigenvalues decrease geometrically from 1 to $1/\kappa$, where κ is called the *condition number* of Σ. To investigate the effects of dimensionality and anisotropy, we consider $d \in \{20, 50, 100\}$ and $\kappa \in \{1, 10, 50\}$, with particle sizes $m \in \{2^8, 2^9, \ldots, 2^{12}\}$. The number of particles needed to avoid weight collapse grows exponentially with the effective dimension (Section 1.1), so even the largest of these particle sizes is small at these dimensions. Every setting thus lies in the small-particle regime that motivates this work.

In each of the twenty independent replications, we simulate a trajectory of $T = 1000$ timesteps and run a bootstrap particle filter with the adaptive multi-descendant proposal and adaptive resampling of Section 2.1, setting $K = 3$ and $c_{\text{thres}} = 2$. We compare WDR with the baselines listed above, excluding T-WDR because the higher-dimensional settings exhibit severe weight degeneracy, for which it is not designed.

5.1.1 Filtering results

Figure 2 reports the time-averaged RMSE and energy score against the Kalman reference $I^{\star(t)}$, and shows that WDR performs the best across all (d, κ), while the four baselines are nearly indistinguishable. For both metrics, the gap widens with κ, because classical schemes only replicate existing particles and thus suffer from weight degeneracy and sample impoverishment in ill-conditioned settings, whereas WDR's OT-informed coupling creates new support points that preserve particle dispersion and posterior coverage. This is consistent with the intrinsic dimension dependence in Theorem 1: stronger anisotropy concentrates the posterior along fewer effective directions, bringing it closer to a lower-dimensional structure, where geometry-aware resampling is beneficial. RMSE depends on the particle distribution only through its mean, where WDR's margin over the baselines becomes smaller as m grows when $d = 20$, although its curve remains lowest throughout. The energy score instead depends on the whole particle distribution, which is what WDR's free-support placement improves, and there its advantage persists at the largest particle size. The energy distance results in Section S3.7 yield rankings consistent with those based on RMSE and energy score.

Per-time trajectories further confirm that WDR's advantage holds over the whole time horizon for all the considered particle sizes; see Section S3.3.1 for details.

5.1.2 Stability analysis

Besides accuracy, it is also important to assess the stability of resampling methods across repeated runs. To this end, we fix one dataset from the linear Gaussian model with $(d, \kappa) = (50, 50)$ and rerun the particle filter twenty times with different random seeds. We consider three stability metrics, moving from the particle level to the estimator level and then to the evaluation level, all averaged over timesteps: (i) the between-run energy distance between the full particle distributions; (ii) the between-run standard deviation of the

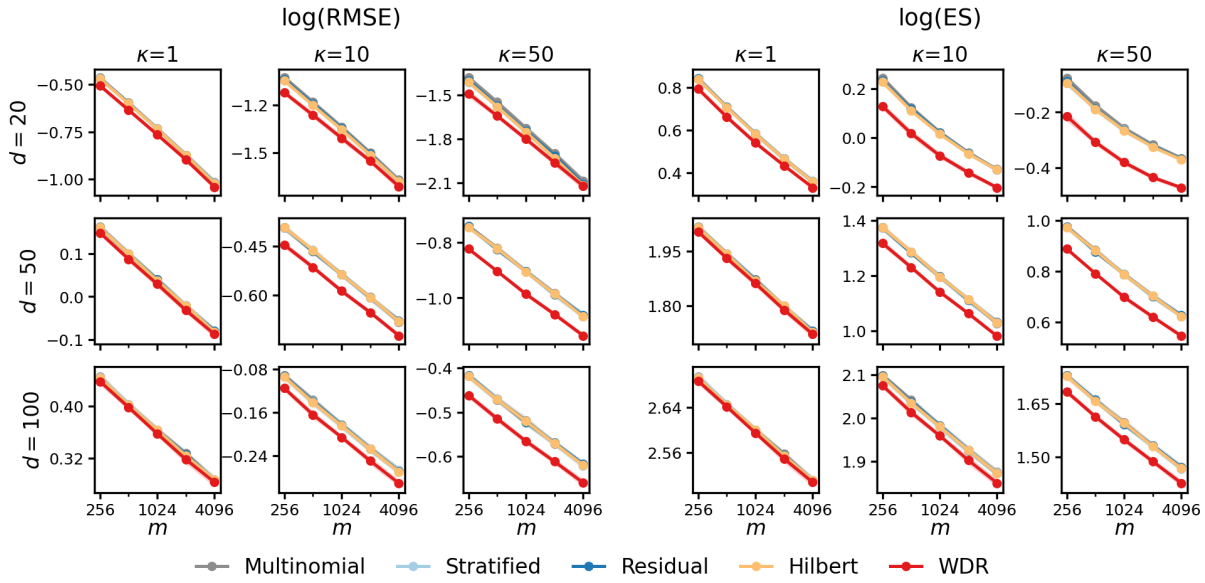


Figure 2: Time-averaged RMSE (left) and energy score (right) versus the particle size m on the log-log scale for the linear Gaussian model. Rows correspond to the dimension $d \in \{20, 50, 100\}$ and columns to the condition number $\kappa \in \{1, 10, 50\}$. Shaded bands represent standard deviations based on twenty replications.

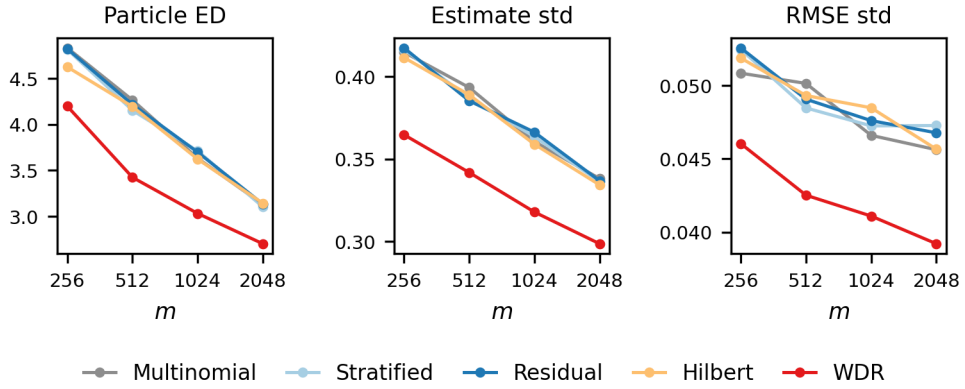


Figure 3: Between-run stability metrics versus particle size m for the linear Gaussian model. From left to right: between-run particle energy distance, between-run standard deviation of the estimate, and between-run standard deviation of RMSE. Lower values are better.

posterior mean estimate; and (iii) the between-run standard deviation of the RMSE relative to the Kalman oracle. Detailed definitions of these metrics are provided in Section S3.3.2. For all three metrics, smaller values indicate greater stability.

Figure 3 plots the stability metrics against particle size m. Although stratified and residual resampling are known to have lower conditional variance than multinomial resampling at a

single step (Chopin & Papaspiliopoulos 2020), the between-run variability accumulates with time and aggregates propagation randomness in the particle filter. Consequently, differences among these classical resampling methods are largely masked in the final run-to-run metrics, and their curves remain close. In contrast, WDR lies below all baseline methods across the full range of m on all three metrics. This can be attributed to the fact that WDR constructs a structured and geometrically coherent transport of the particle cloud, rather than randomly selecting existing particles. This leads to a more stable approximation of the filtering distribution, demonstrating that WDR improves not only estimation accuracy but also between-run stability.

5.1.3 Performance-cost analysis

The above experiments compared methods at equal particle size. Considering that WDR incurs additional resampling overhead, a more practical criterion is performance at equal wall-clock time. We design an experiment that varies the propagation cost to identify when WDR is the preferred choice.

We continue to use the linear Gaussian model with $(d, \kappa) = (50, 50)$, and now make the cost of simulating the latent diffusion explicit. For a general nonlinear diffusion, the transition kernel has no closed form, so each propagation step must integrate the dynamics over many small sub-steps between observations, with cost proportional to the integration resolution. We emulate this cost by simulating the OU dynamics on a grid of γ internal sub-steps per observation interval. Specifically, we set $A = P\,\mathrm{diag}(q_1, \ldots, q_d)\,P^\top$, where P is a fixed random orthogonal matrix and $q_j = 0.95^{\,1+s(2(j-1)/(d-1)-1)}$ with spread $s = 0.5$, and each sub-step applies the transition $A_{\mathrm{sub}} = P\,\mathrm{diag}(q_1^{1/\gamma}, \ldots, q_d^{1/\gamma})\,P^\top$ with a matching sub-step noise covariance Σ_{sub}, so that γ successive sub-transitions under $(A_{\mathrm{sub}}, \Sigma_{\mathrm{sub}})$ exactly reproduce the original one-step transition under (A, Σ). This construction leaves the target filtering distribution and the Kalman oracle unchanged, while making the propagation cost scale

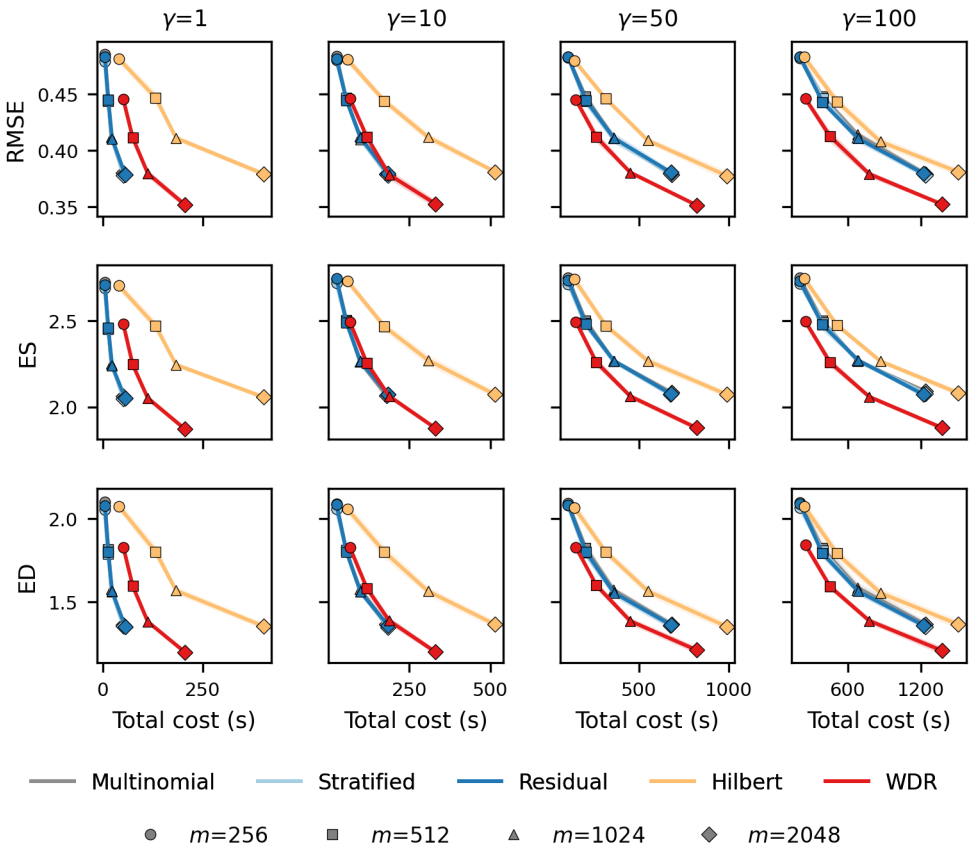


Figure 4: Time-averaged RMSE (top row), energy score (middle row), and energy distance (bottom row) versus total wall-clock cost on the linear Gaussian model with $(d, \kappa) = (50, 50)$. Each panel corresponds to a sub-step multiplier γ. Shaded bands represent standard deviations based on twenty replications.

approximately linearly with γ.

Figure 4 plots the time-averaged RMSE, energy score, and energy distance against the total wall-clock time for each γ. As γ grows, propagation gradually comes to dominate the computational budget, while resampling accounts for an ever-smaller fraction of total runtime. In particular, the ratio of WDR resampling time to propagation time, averaged across m, falls from $5.1\times$ at $\gamma = 1$ to $1.1\times$ at $\gamma = 10$, $0.30\times$ at $\gamma = 50$, and $0.16\times$ at $\gamma = 100$. Consistent with this shift in the computational budget, the practical benefit of WDR becomes increasingly clear as γ rises. For $\gamma \geq 10$, WDR lies on the Pareto frontier for each of the three metrics: no alternative method simultaneously achieves a lower value on the corresponding metric and a lower total runtime.

Translated to absolute time, $\gamma = 10$ corresponds to a per-particle-per-step propagation cost of roughly 0.1 ms, giving a practical threshold above which WDR is the recommended choice. Although this threshold is a representative order-of-magnitude reference in our experiments, the qualitative claim that WDR is preferable when propagation dominates is robust across settings. This regime is directly relevant to modern SMC applications, such as the diffusion model of Section 6, where the propagation cost is far above this threshold. We also conduct sensitivity analyses over the Dirichlet concentration parameter τ (defined in (6)), the multi-descendant factor K, and the CV threshold c_{thres} (see Section 2.1), and observe that WDR performs robustly, with intermediate values of τ performing the best. Details are provided in Section S3.5.

5.2 High-dimensional Gaussian spatial model

To test a problem in which the model itself makes propagation expensive, we consider a linear Gaussian spatial model on a 45×45 lattice with $d = 2025$ grid sites:

$$X^{(t)} = 0.95\, X^{(t-1)} + \xi^{(t)}, \quad \xi^{(t)} \sim \mathcal{N}(0, \Sigma),$$
$$Y^{(t)} = H X^{(t)} + \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim \mathcal{N}(0, I_{d_y}),$$

where each $X^{(t)}$ is d-dimensional; $\Sigma_{ij} = 0.5(1 + \sqrt{3} r_{ij}/\rho) \exp(-\sqrt{3} r_{ij}/\rho)$ is a Matérn covariance with $\rho = 6$ and r_{ij} the distance between sites i and j on the unit-spacing lattice; and $H \in \{0,1\}^{d_y \times d}$ has $d_y = 202$ rows, each selecting one site (only one "1" in each row). Generating the spatially correlated process noise $\xi^{(t)}$ requires a dense $d \times d$ matrix-vector product per particle per step, incurring substantial computational cost when $d = 2025$. At this dimension, the generic bootstrap filter collapses with any resampling scheme, so we employ the block particle filtering strategy (Rebeschini & van Handel 2015), detailed in Section S3.6 of the Supplementary Materials. We simulate a trajectory of $T = 200$ timesteps and compare all the resampling methods discussed in Section 5.1. The oracle reference is again available from the Kalman filter.

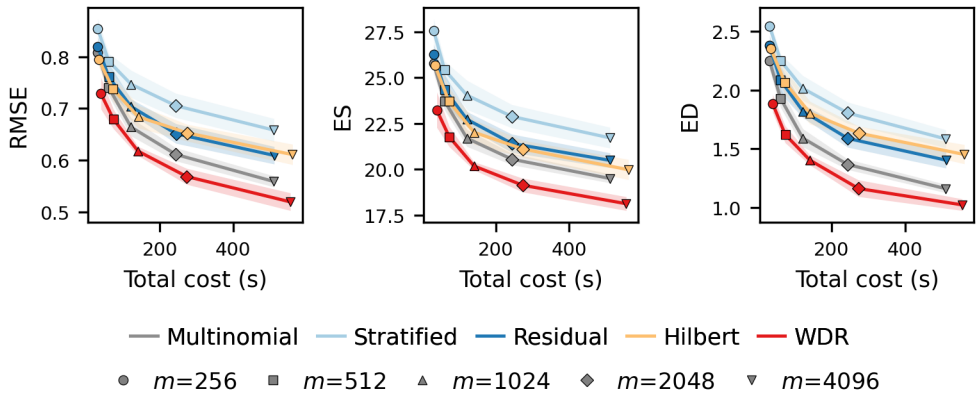


Figure 5: Time-averaged RMSE (left), energy score (middle), and energy distance (right) versus total wall-clock cost on the $d = 2025$ spatial model. Marker shape indicates the particle size m. Shaded bands represent standard deviations based on twenty replications.

Figure 5 reports the evaluation metrics against total wall-clock cost, with the particle size m represented by the marker. WDR outperforms all baselines on all the three metrics at every m, and its advantage persists as m increases. The propagation step costs about 0.6 ms per particle per step, six times the threshold identified in Section 5.1.3, and accounts for over 80% of the per-step cost when WDR is used for resampling. Among the classical schemes, multinomial resampling performs the best, reversing its usual ranking. We attribute this reversal to the i.i.d. output indices generated by multinomial resampling, which preserve the blockwise independence that the block particle filter is based on.

6 Conditional image generation

We next apply Wasserstein-Dirichlet resampling to conditional image generation with diffusion models, using the twisted diffusion sampler (TDS) of Wu et al. (2023) as the baseline SMC framework. Let $x^{(0:T)} = (x^{(0)}, \ldots, x^{(T)})$ denote the latent diffusion trajectory (with t decreasing in the reverse-time sampling procedure) and let y be a conditioning variable, such as a class label. A pretrained diffusion model specifies a joint density $p_\theta(x^{(0:T)})$ over trajectories, and the task is to draw samples from the conditional distribution

$\pi_0(x^{(0)} \mid y)$, which is the marginal of $\pi(x^{(0:T)} \mid y) \propto p_\theta(x^{(0:T)})\, p(y \mid x^{(0)})$, where $p(y \mid x^{(0)})$ is an externally supplied likelihood (e.g., a classifier on clean images).

Twisted diffusion sampler with WDR. TDS views $\pi(x^{(0:T)} \mid y)$ as the extended target of an SMC sampler and runs particles backward in diffusion time. In particular, TDS augments an unconditional score network with a likelihood gradient term as the twisted proposal and corrects this proposal using importance weights, so that the resulting particle approximation converges to $\pi(x^{(0:T)} \mid y)$ as the number of particles grows. Algorithmic details of TDS are provided in Section S4.1. At each reverse timestep, TDS (i) propagates descendants via the twisted kernel, (ii) updates the particle weights by the associated incremental weight, and (iii) resamples particles according to these weights. In the original TDS, systematic sampling (i.e., stratified with a single shared random offset) is used at each resampling step. As shown in Section 5.1, such classical schemes are inefficient in high-dimensional state spaces, whereas Wasserstein-Dirichlet resampling better preserves geometric structure and diversity. We therefore replace the classical resampling in TDS with (T-)WDR and incorporate multi-descendant propagation with $K = 3$.

6.1 Experimental setup

We follow the experimental design of Wu et al. (2023) and consider class-conditional image generation on MNIST and CIFAR-10 datasets, each consisting of ten classes. In both experiments, we use the same pretrained score-based diffusion models and ResNet classifiers as in the original TDS. Experiments were conducted on one NVIDIA H100 GPU (80 GB), using PyTorch 2.0.1 and CUDA 11.8.

For MNIST, the conditioning variable is the digit label, and the goal is to generate 28×28 grayscale images. For CIFAR-10, the conditioning variable is the class label, and the goal is to generate 32×32 RGB images. In all runs, we simulate the reverse diffusion trajectory for $T = 100$ discretization steps and consider particle sizes $m \in \{2^4, 2^5, 2^6, 2^7\}$. Following

Wu et al. (2023), we trigger resampling adaptively when the effective sample size falls below $0.5m$. Considering that the squared Euclidean distances in high-dimensional pixel space are poorly aligned with perceptual similarity, we implement Wasserstein-Dirichlet resampling in a learned low-dimensional representation that better reflects image semantics. Implementation details are provided in Section S4.2.

For each method, we run 100 independent replications per class, with each run using m particles (i.e., m generated images). We evaluate the generated images using four complementary metrics: (i) *Fréchet inception distance (FID)* (Heusel et al. 2017), lower indicates better distributional fidelity; (ii) *classification accuracy* (Wu et al. 2023), higher indicates better label-conditional correctness; (iii) *within-run LPIPS diversity* (Zhang et al. 2018), higher indicates more within-class sample diversity; and (iv) *between-run over within-run LPIPS ratio*, smaller (i.e., closer to one) indicates that different random seeds explore the same correct posterior region rather than collapsing onto different modes. Detailed definitions of these metrics are provided in Section S4.3.

6.2 Experimental results

Figure 6 compares vanilla TDS with its (T-)WDR variants on MNIST and CIFAR-10. To evaluate practical benefit, we plot each performance metric against computational costs directly. As expected, all methods improve as the particle size m increases, reflecting a more accurate approximation of the twisted path measure. Within this overall pattern, TDS-T-WDR performs the best, followed by TDS-WDR, and both consistently outperform vanilla TDS on both datasets.

This result is consistent with the geometric structure of the problem. Although they lie in a high-dimensional pixel space, natural images are concentrated near a much lower-dimensional manifold. Thus, in our conditional generation tasks, weight degeneracy is often mild to moderate when employing the twisted SMC. This is precisely the regime where

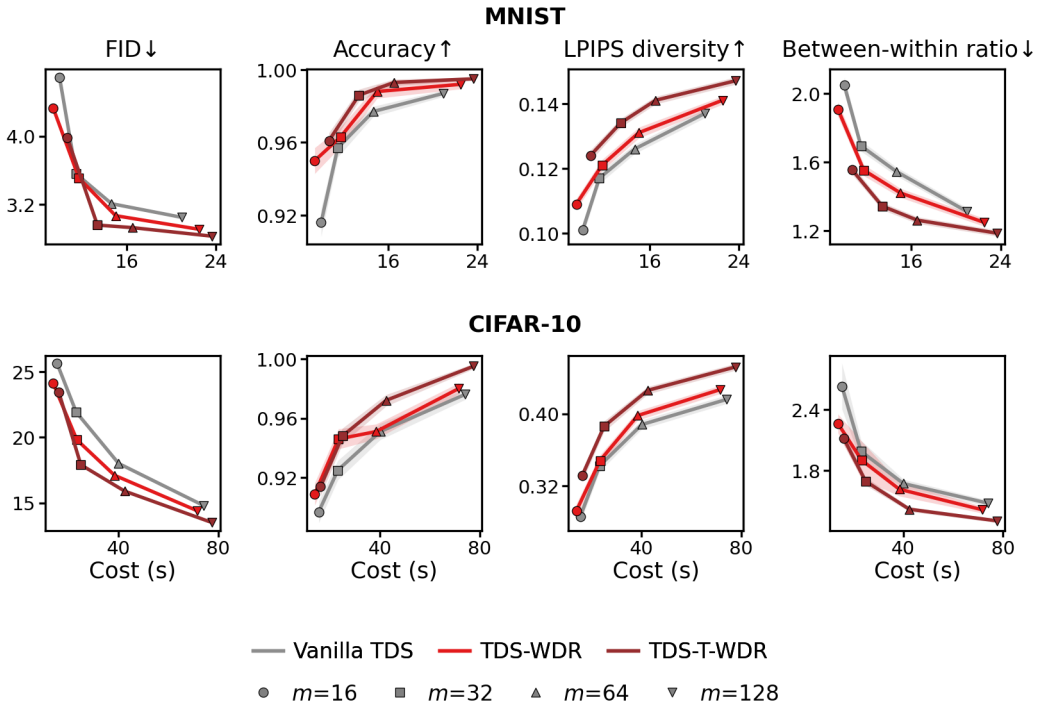


Figure 6: Conditional image generation metrics versus wall-clock time (in seconds). Lower is better for FID and the between-within ratio; higher is better for classification accuracy and within-run LPIPS diversity. Shaded bands represent standard errors based on one hundred replications.

T-WDR is most effective, leading to improved generation qualities.

In terms of computational cost, TDS-WDR introduces negligible additional cost compared to vanilla TDS, whereas TDS-T-WDR incurs a modest overhead, which is consistent with the wall-clock times reported in Section S3.2.2 and can be attributed to more resampling triggers. Overall, (T-)WDR remains computationally competitive across the particle sizes considered. Averaged over m, the per-particle-per-step propagation cost is about 4 ms on MNIST and 8 ms on CIFAR-10, both far above the practical threshold identified in Section 5.1.3. Hence, diffusion-based conditional generation fits the propagation-dominating regime, where the overhead of WDR is negligible, while its performance gains remain practically meaningful.

Figure 7 provides examples of the generated images with $m = 64$. Compared with vanilla TDS, our TDS-T-WDR produces samples that are both more class-consistent and more diverse within each class. For example, the MNIST digits generated by TDS-T-WDR

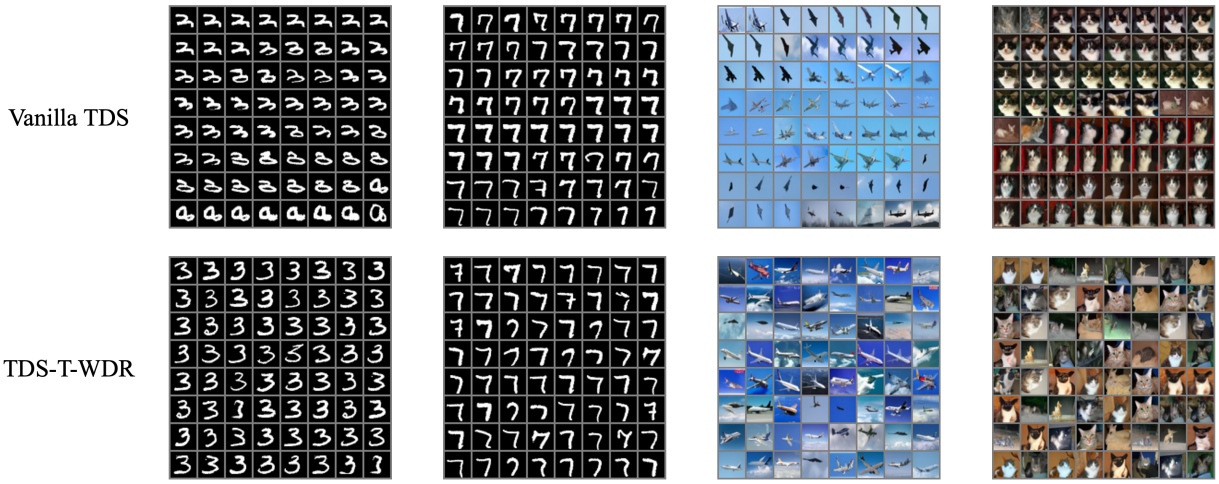


Figure 7: Samples generated by vanilla TDS (top row) versus those by the proposed TDS-T-WDR (bottom row) with $m = 64$. The image batches are randomly selected. From left to right: digit 3 (MNIST), digit 7 (MNIST), airplane (CIFAR-10), and cat (CIFAR-10).

show richer variation in stroke thickness, curvature, and writing style. In contrast, the vanilla TDS occasionally shows low-quality artifacts and exhibits apparent within-class repetition, indicating reduced effective diversity. Taken together, these results demonstrate that incorporating Wasserstein-Dirichlet resampling into TDS improves fidelity, conditional correctness, and sample coverage without compromising practical efficiency.

7 Concluding remarks

We propose a new resampling paradigm, termed WDR, for sequential Monte Carlo from the perspective of free-support Wasserstein barycenters. Unlike conventional replication-only approaches, our scheme employs an optimal transport formulation that jointly optimizes the coupling and support locations, yielding an equal-weight particle system that best approximates the weighted empirical distribution in the Wasserstein sense. A Dirichlet mixing mechanism injects controlled stochasticity, interpolating between a deterministic transform and OT-informed ancestor sampling. Additionally, when weight degeneracy is mild, a truncated version of the method offers further computational efficiency. Algorithmically, we develop a block coordinate descent solver that exploits the problem's biconvex structure

with provable convergence guarantees. Theoretically, we derive a non-asymptotic MSE bound that makes the bias-variance trade-off explicit and reveals that the convergence rate depends only on the intrinsic dimension of the target distribution.

Empirically, WDR consistently improves filtering accuracy and run-to-run stability, particularly in higher-dimensional and anisotropic settings. At per-particle propagation costs above roughly 0.1 ms, the method occupies the Pareto frontier, trading off accuracy and runtime optimally. Moreover, WDR integrates seamlessly into modern SMC-based diffusion samplers, boosting conditional image generation across fidelity, coverage, and diversity metrics—all while remaining computationally competitive.

Several directions are worthy of further investigation. First, it would be valuable to propagate our single-step theoretical guarantee through the full SMC recursion to quantify the overall filtering error. Second, an adaptive selection strategy for the Dirichlet concentration parameter τ would be useful in practice. Third, it is of interest to generalize the truncated WDR to the severely degenerate high-dimensional scenarios. Finally, the proposed resampling paradigm can extend to broader particle filters, such as the SMC sampler (Dai et al. 2022) and particle MCMC (Andrieu et al. 2009), where geometry-aware rejuvenation may help reduce path degeneracy and improve mixing.

Supplementary materials The supplementary materials include: a PDF containing proofs of the theoretical results, additional details, and further numerical results; and a ZIP implementing the proposed methods and reproducing the experiments. The reproducibility materials are also available at https://github.com/Mengyu8042/wdr-repro.

Disclosure statement The authors declare no competing interests.

Data availability statement The datasets used in this study are publicly available at http://yann.lecun.com/exdb/mnist/ and https://www.cs.toronto.edu/~kriz/cifar.html.

Declaration of generative AI use The authors used ChatGPT (OpenAI, GPT-5 and

GPT-6 series) for language polishing. They reviewed and edited the output and take full responsibility for the manuscript.

Acknowledgments The authors thank the Editor, the Associate Editor, and three anonymous referees for their constructive comments that helped improve the work. This work was supported by the National Natural Science Foundation of China (Grant No. 12601554), the China Postdoctoral Science Foundation (Grant No. 2025M783111), and the Shenzhen Science and Technology Program (Grant No. AI2026019).

References

Acevedo, W., de Wiljes, J. & Reich, S. (2017), 'Second-order accurate ensemble transform particle filters', *SIAM Journal on Scientific Computing* **39**(5), A1834–A1850.

Andrieu, C., Doucet, A. & Holenstein, R. (2009), Particle Markov chain Monte Carlo for efficient numerical simulation, *in* 'Monte Carlo and Quasi-Monte Carlo Methods 2008', Springer, pp. 45–60.

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

Bengtsson, T., Bickel, P. & Li, B. (2008), Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, *in* 'Probability and Statistics: Essays in Honor of David A. Freedman', Institute of Mathematical Statistics, pp. 316–334.

Bertsekas, D. P. (2016), *Nonlinear Programming, 3rd Edition*, Athena Scientific.

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

Chopin, N. & Papaspiliopoulos, O. (2020), *An Introduction to Sequential Monte Carlo*, Springer.

Claici, S., Chien, E. & Solomon, J. (2018), Stochastic Wasserstein barycenters, *in* 'International Conference on Machine Learning', PMLR, pp. 999–1008.

Corenflos, A., Thornton, J., Deligiannidis, G. & Doucet, A. (2021), Differentiable particle filtering via entropy-regularized optimal transport, *in* 'International Conference on Machine Learning', PMLR, pp. 2100–2111.

Cuturi, M. & Doucet, A. (2014), Fast computation of Wasserstein barycenters, *in* 'International Conference on Machine Learning', PMLR, pp. 685–693.

Dai, C., Heng, J., Jacob, P. E. & Whiteley, N. (2022), 'An invitation to sequential Monte Carlo samplers', *Journal of the American Statistical Association* **117**(539), 1587–1600.

Doucet, A. & Johansen, A. M. (2011), A tutorial on particle filtering and smoothing: Fifteen years later, *in* 'The Oxford Handbook of Nonlinear Filtering', Oxford University Press, pp. 656–704.

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

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N. et al. (2021), 'POT: Python optimal transport', *Journal of Machine Learning Research* **22**(78), 1–8.

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

Gerber, S. & Maggioni, M. (2017), 'Multiscale strategies for computing optimal transport', *Journal of Machine Learning Research* **18**(72), 1–32.

Gilks, W. R. & Berzuini, C. (2001), 'Following a moving target—Monte Carlo inference for dynamic Bayesian models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(1), 127–146.

Gordon, N. J., Salmond, D. J. & Smith, A. F. M. (1993), 'Novel approach to nonlinear/non-Gaussian Bayesian state estimation', *IEE Proceedings F (Radar and Signal Processing)* **140**(2), 107–113.

Gray, R. M. & Neuhoff, D. L. (1998), 'Quantization', *IEEE Transactions on Information Theory* **44**(6), 2325–2383.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017), GANs trained by a two time-scale update rule converge to a local Nash equilibrium, *in* 'Advances in Neural Information Processing Systems', Vol. 30, pp. 6626–6637.

Hundrieser, S., Staudt, T. & Munk, A. (2024), 'Empirical optimal transport between different measures adapts to lower complexity', *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **60**(2), 824–846.

Kang, C. H. & Park, C. G. (2018), 'Particles resampling scheme using regularized optimal transport for sequential Monte Carlo filters', *International Journal of Adaptive Control and Signal Processing* **32**(10), 1393–1402.

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.

Kloeckner, B. (2012), 'Approximation by finitely supported measures', *ESAIM: Control, Optimisation and Calculus of Variations* **18**(2), 343–359.

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

Li, M., Yu, J., Li, T. & Meng, C. (2023), 'Importance sparsification for Sinkhorn algorithm', *Journal of Machine Learning Research* **24**(247), 1–44.

Li, Y., Wang, W., Deng, K. & Liu, J. S. (2022), 'Stratification and optimal resampling for sequential Monte Carlo', *Biometrika* **109**(1), 181–194.

Lin, T., Ho, N., Chen, X., Cuturi, M. & Jordan, M. I. (2020), Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm, *in* 'Advances in Neural Information Processing Systems', Vol. 33, pp. 5368–5380.

Liu, J. S. & Chen, R. (1995), 'Blind deconvolution via sequential imputations', *Journal of the American Statistical Association* **90**(430), 567–576.

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

Lyddon, S. P., Holmes, C. C. & Walker, S. G. (2019), 'General Bayesian updating and the loss-likelihood bootstrap', *Biometrika* **106**(2), 465–478.

Morgan, K. L. & Rubin, D. B. (2012), 'Rerandomization to improve covariate balance in experiments', *The Annals of Statistics* **40**(2), 1263–1282.

Peyré, G. & Cuturi, M. (2019), 'Computational optimal transport: With applications to data science', *Foundations and Trends® in Machine Learning* **11**(5-6), 355–607.

Rao, J. N. K. & Wu, C. F. J. (1988), 'Resampling inference with complex survey data', *Journal of the American Statistical Association* **83**(401), 231–241.

Rebeschini, P. & van Handel, R. (2015), 'Can local particle filters beat the curse of dimensionality?', *The Annals of Applied Probability* **25**(5), 2809–2866.

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

Rubin, D. B. (1981), 'The Bayesian bootstrap', *The Annals of Statistics* **9**(1), 130–134.

Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American statistical Association* **91**(434), 473–489.

Snyder, C., Bengtsson, T., Bickel, P. & Anderson, J. (2008), 'Obstacles to high-dimensional particle filtering', *Monthly Weather Review* **136**(12), 4629–4640.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R. & Reich, S. (2019), 'Particle filters for high-dimensional geoscience applications: A review', *Quarterly Journal of the Royal Meteorological Society* **145**(723), 2335–2365.

Villani, C. (2009), *Optimal Transport: Old and New*, Vol. 338, Springer.

Weed, J. & Bach, F. (2019), 'Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance', *Bernoulli* **25**(4A), 2620–2648.

Wu, L., Trippe, B. L., Naesseth, C. A., Blei, D. M. & Cunningham, J. P. (2023), Practical and asymptotically exact conditional sampling in diffusion models, *in* 'Advances in Neural Information Processing Systems', Vol. 36, pp. 31372–31403.

Xie, Y., Wang, X., Wang, R. & Zha, H. (2020), A fast proximal point method for computing exact Wasserstein distance, *in* 'Uncertainty in Artificial Intelligence', PMLR, pp. 433–453.

Yang, W., Song, L., Tee, C. A. T., Zheng, Y. & Liu, Y. (2019), 'Unsupervised learning grouping-based resampling for particle filters', *IEEE Access* **7**, 127265–127275.

Zhang, J., Zhong, W. & Ma, P. (2021), A review on modern computational optimal transport methods with applications in biomedical research, *in* 'Modern Statistical Methods for Health Research', Springer, pp. 279–300.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. (2018), The unreasonable effectiveness of deep features as a perceptual metric, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 586–595.