Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

An Invitation to Sequential Monte Carlo Samplers

Chenguang Dai, Jeremy Heng, Pierre E. Jacob & Nick Whiteley

To cite this article: Chenguang Dai, Jeremy Heng, Pierre E. Jacob & Nick Whiteley (2022): An Invitation to Sequential Monte Carlo Samplers, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2087659

To link to this article: https://doi.org/10.1080/01621459.2022.2087659

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Reviews
https://doi.org/10.1080/01621459.2022.2087659

Taylor & Francis
Taylor & Francis Group

An Invitation to Sequential Monte Carlo Samplers

Chenguang Dai[a], Jeremy Heng[b], Pierre E. Jacob[b], and Nick Whiteley[c]

[a]Citadel Securities, Chicago, Illinois, USA; [b]ESSEC Business School, Cergy, France; [c]School of Mathematics, University of Bristol, Bristol, UK

ABSTRACT

Statisticians often use Monte Carlo methods to approximate probability distributions, primarily with Markov chain Monte Carlo and importance sampling. Sequential Monte Carlo samplers are a class of algorithms that combine both techniques to approximate distributions of interest and their normalizing constants. These samplers originate from particle filtering for state space models and have become general and scalable sampling techniques. This article describes sequential Monte Carlo samplers and their possible implementations, arguing that they remain under-used in statistics, despite their ability to perform sequential inference and to leverage parallel processing resources among other potential benefits. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2020
Accepted June 2022

KEYWORDS
Interacting particle systems; Monte Carlo methods; Normalizing constant; Sequential inference

1. Introduction

Motivation. The numerical approximation of probability distributions is ubiquitous in statistics, whether it concerns the distribution of a test statistic, or conditional distributions of parameters and latent variables given observed variables. Monte Carlo methods, originally developed in physics to compute specific expectations of interest, have become fundamental in data analysis. The complexity of distributions encountered in the practice of statistics has increased, and the study of Monte Carlo methods to approximate them has become more formal. Meanwhile, the context surrounding the development of numerical methods has changed, with computation becoming more parallel, users becoming accustomed to stochastic as opposed to deterministic calculations, increasing interest in quantifying Monte Carlo errors, and a wider availability of modifiable software packages.

In this fast-changing landscape, it can be difficult for the nonspecialist to keep track of important developments in Monte Carlo methods. This article shines some light on *Sequential Monte Carlo samplers*, imported from the signal processing literature to statistics in the early 2000s (Chopin 2002; Del Moral, Doucet, and Jasra 2006). These algorithms provide a generic approach to sample from probability distributions, can scale well with the dimension of the state space, provide estimates of the associated normalizing constants, are well-adapted to sequential settings and are largely amenable to parallel computing.

We first recall Markov chain Monte Carlo (MCMC, Robert and Casella 2011; Dunson and Johndrow 2020) and importance sampling (IS, Owen 2013, chap. 9) techniques and highlight some of their limitations. Throughout the article we consider the task of sampling from a target distribution $\pi(dx) = \gamma(x)dx/Z$ defined on a measurable space $(\mathsf{X}, \mathscr{X})$, with unnormalized density $x \mapsto \gamma(x)$ and unknown normalizing constant $Z = \int_{\mathsf{X}} \gamma(x)dx$, that we might also want to approximate.

MCMC. Given a π-invariant Markov transition kernel M, an MCMC method generates a Markov chain $(x_t)_{t\geq 0}$ by sampling x_0 from an initial distribution π_0 and iteratively sampling the next state x_t given x_{t-1} from $M(x_{t-1}, \cdot)$. After discarding an initial portion of the trajectory as "burn-in," the subsequent T states form an empirical approximation $T^{-1}\sum_{t=1}^{T}\delta_{x_t}(\cdot)$ of the target distribution π converging as $T \to \infty$ (Nummelin 2002). Although immensely successful (Diaconis 2009), MCMC methods suffer from serious limitations: their iterative nature prevents straightforward parallelization; tuning the Markov transition kernels to achieve a satisfactory efficiency might be time-consuming; and the estimation of the normalizing constant Z from MCMC runs alone is difficult.

Motivated by these considerations, a number of more elaborate algorithms, or "meta-algorithms," emerged in the 1990s, such as parallel tempering (Geyer 1991) where intermediate distributions of varying complexity are introduced between π_0 and π, and Markov chains targeting these distributions regularly exchange their states. The introduction of a path of distributions also appeared in techniques specifically designed to estimate Z, known as bridge sampling and path sampling (Chen and Shao 1997; Gelman and Meng 1998).

Importance sampling. IS is a method to approximate π and Z using samples from another distribution π_0. Denoting N independent samples from π_0 by $(x^n)_{n\in[N]}$, IS assigns weights computed as $w^n = \gamma(x^n)/\pi_0(x^n)$ for each $n \in [N] = \{1, \ldots, N\}$. This provides the estimator $Z^N = N^{-1}\sum_{n=1}^{N} w^n$ and the weighted empirical measure $(NZ^N)^{-1}\sum_{n=1}^{N} w^n\delta_{x^n}(\cdot)$ that converge to Z and π as $N \to \infty$ under assumptions on the ratio of densities $\pi(x)/\pi_0(x)$ (Owen 2013). The method is amenable to parallel computation but its naive implementation can fail spectacularly when π_0 and π are

CONTACT Pierre E. Jacob pierre.jacob@essec.edu ESSEC Business School, 3 avenue Bernard Hirsch, 95000 Cergy, France.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

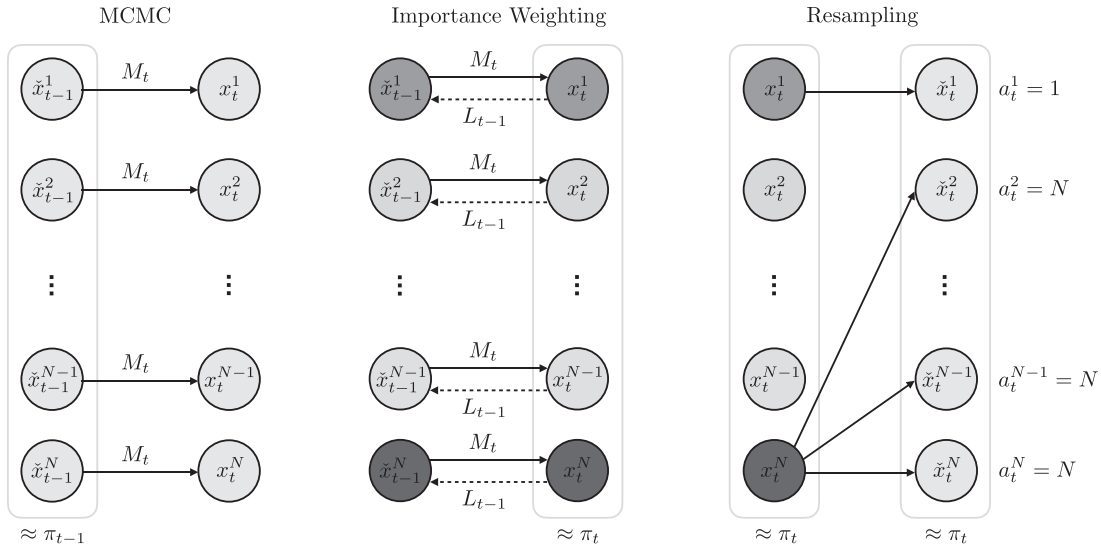


Figure 1. The three steps of SMCS. Particles $(\check{x}_{t-1}^n)_{n\in[N]}$ approximating π_{t-1} evolve separately according to an MCMC kernel M_t (*left*). To approximate π_t, each pair $(\check{x}_{t-1}^n, x_t^n)$ is assigned an importance weight $w_t(\check{x}_{t-1}^n, x_t^n)$, which depends on a backward kernel L_{t-1} (*middle*). The size of weights is depicted by the shade of gray. Equally weighted particles are then obtained by sampling from the particles according to their weights (*right*).

far apart (Agapiou et al. 2017; Chatterjee and Diaconis 2018). This motivated the introduction of "bridging" distributions between π_0 and π and the combination of these intermediate distributions with MCMC moves was proposed in Annealed Importance Sampling (Neal 2001), see also Jarzynski (1997). These works are predecessors to sequential Monte Carlo samplers.

Particle filters. State space models are a flexible way of analyzing time series, where each observation is associated with a latent variable, and these latent variables form a Markov chain. In that context, importance sampling forms the basis of particle filters (Gordon, Salmond, and Smith 1993; Kong, Liu, and Wong 1994), developed for sequential state inference where the target π_t at step t is the distribution of the tth latent variable given observations available up to time t. Particle filters employ Monte Carlo principles that generalize Kalman filters to nonlinear, non-Gaussian state space models. Many variants have been put forward to deal with state space models of various complexities (Cappé, Godsill, and Moulines 2007; Chopin and Papaspiliopoulos 2020), for example with the addition of MCMC moves at each step (Gilks and Berzuini 2001).

SMCS. By the late 1990s, various Monte Carlo algorithms incorporated ideas from both importance sampling and MCMC, and the introduction of intermediate distributions was already familiar. There were concurrent efforts to generalize particle filters to a much wider class of problems arising in statistics, under the name of Iterated Batch Importance Sampling (Chopin 2002) and SMC samplers (SMCS, Del Moral, Doucet, and Jasra 2006). To avoid confusion with particle filters, we refer to SMC samplers as SMCS and we view them as a generic alternative to MCMC and IS for the approximation of π and Z. SMCS share the general structure of particle filters and we elaborate further on the links in the supplementary materials.

SMCS are designed to approximate an arbitrary sequence of target distributions of fixed dimension, that recovers the desired distribution π as its last element, and estimates of Z are obtained as by-products. SMCS involve sequences of bridging distributions or "paths," as in path sampling and parallel tempering. Specifically, SMCS generate a system of N samples, termed particles, that evolve through MCMC moves and importance weights, and interact through resampling steps. A schematic description of SMCS is provided in Figure 1. The benefits of SMCS include the estimation of Z, the ability to adaptively tune MCMC kernels, a large amenability to parallel computing, and improved accuracy relative to plain or annealed IS.

Objectives and outline. We aim to provide an accessible guide with useful references on SMCS for statisticians. We do not assume familiarity with particle filters. Our presentation is self-contained, complementing chap. 17 of Chopin and Papaspiliopoulos (2020) with updated pointers to some of the most significant advances. Section 2 describes SMCS and possible instantiations. As SMCS involve IS steps, and since the performance of one step of IS tends to deteriorate with the dimension, we provide in Section 3 a clarification of the role of bridging distributions. In Section 4 we present methodological consequences of the fact that SMCS generate interacting particle systems (Del Moral 2004, 2013), and not trajectories of Markov chains as in MCMC. We underline how parallel computers can be employed and how the approximation error can be quantified. In Section 5 we illustrate how SMCS translate into algorithms of practical importance in statistics using simple examples, and Section 6 concludes.

2. What are Sequential Monte Carlo Samplers?

We follow Del Moral, Doucet, and Jasra (2006), and describe the required specification of paths, of forward and backward Markov transition kernels and of various algorithmic components. As we will see, many of these choices can be implicitly or adaptively made.

2.1. Algorithmic Description

A generic, nonadaptive sequential Monte Carlo sampler is described in Algorithm 1. As input, a sequence of $T+1$ distributions $\pi_0, \ldots, \pi_T$ is introduced, with each $\pi_t(dx) = \gamma_t(x)dx/Z_t$ defined on the same space $(\mathsf{X}, \mathscr{X})$, where $\gamma_t(x)$ denotes an unnormalized density, and $Z_t = \int_{\mathsf{X}} \gamma_t(x)dx$ a normalizing constant (with $Z_0 = 1$). We assume that we can sample from π_0 to initialize the algorithm (see Step 1(a)) and that the terminal distribution π_T is precisely the target distribution π.

Next we introduce two sequences of Markov kernels. In the sequence $(M_t)_{t\in[T]}$ of "forward" kernels, each M_t is designed to leave π_t invariant, at least approximately. The forward kernel M_t is used to sample x_t given x_{t-1} at the tth step of the algorithm (see Step 2(b) and left of Figure 1). At this stage, one could imagine an importance sampling step with the proposed x_t to approximate π_t. However, for most choices of M_t, the marginal density of the proposed state $q_t(x_t) = \int_{\mathsf{X}} \pi_{t-1}(dx_{t-1})M_t(x_{t-1}, x_t)$ would be intractable, and importance weights could not be computed. To circumvent this issue, the key idea is to define an importance sampling step on the space of (x_{t-1}, x_t), by introducing the sequence $(L_{t-1})_{t\in[T]}$ of "backward" kernels (Neal 2001) (see Step 2(c) and middle of Figure 1). Defining the proposal $\pi_{t-1}(dx_{t-1})M_t(x_{t-1}, dx_t)$ and the target $\pi_t(dx_t)L_{t-1}(x_t, dx_{t-1})$ on the joint space, the weight function

$$w_t(x_{t-1}, x_t) = \frac{\gamma_t(x_t)L_{t-1}(x_t, x_{t-1})}{\gamma_{t-1}(x_{t-1})M_t(x_{t-1}, x_t)} \tag{2.1}$$

can be made tractable even if the marginal distribution of x_t is intractable and even if M_t has an intractable transition density, by choosing L_{t-1} appropriately. Indeed the backward kernel L_{t-1} is introduced specifically so that the importance weight in (2.1) achieves a practical compromise between ease of numerical evaluation and variance. The joint target $\pi_t(dx_t)L_{t-1}(x_t, dx_{t-1})$ admits π_t as marginal on x_t, for any choice of backward kernel L_{t-1}, therefore, (2.1) provides a valid importance weight. As described in Del Moral, Doucet, and Jasra (2006, sec. 3.3), given π_{t-1}, π_t, and M_t, one should ideally select L_{t-1} to keep the variance of $w_t(x_{t-1}, x_t)$ under the joint proposal $\pi_{t-1}(dx_{t-1})M_t(x_{t-1}, dx_t)$ small. Minimal variance is attained by $L_{t-1}(x_t, x_{t-1}) = \pi_{t-1}(x_{t-1})M_t(x_{t-1}, x_t)/q_t(x_t)$, the backward transition under the joint proposal, that reduces to having $w_t(x_{t-1}, x_t) = \gamma_t(x_t)/q_t(x_t)$, which, again, would be typically intractable. In other words tractability of importance sampling through operating on the joint space of (x_{t-1}, x_t) comes with an increase of variance.

Overall, the algorithm propagates N particles $x_t^{1:N} = (x_t^1, \ldots, x_t^N)$ using the forward kernels (M_t), and assigns to the particles some weights that depend on (π_t), (M_t) and (L_{t-1}). The above reasoning could be carried out recursively and the final weights would be obtained as the product over time of the weights computed at each time step (Neal 2001). One might then worry about the variance of the weights after T steps. The resampling step (see Step 2(a) and right of Figure 1) helps to mitigate this issue. According to the magnitude of their weights, some particles are discarded and others duplicated, maintaining a fixed population size of N. Resampling involves a distribution $r(\cdot|w^{1:N})$ of ancestor indices $a^{1:N}$ in $[N]^N$, parametrized by a vector $w^{1:N}$ of probabilities. The simplest resampling scheme is called multinomial resampling (Gordon, Salmond, and Smith 1993), where $a^{1:N}$ are independent Categorical variables on $[N]$ with probabilities $w^{1:N}$. At step t of the algorithm, N particles are obtained by propagating particles with indices $(a_{t-1}^n)_{n\in[N]}$ generated from $r(\cdot|w_{t-1}^{1:N})$. Gerber, Chopin, and Whiteley (2019) provide recent discussions on resampling schemes, Whiteley, Lee, and Heine (2016) study the stabilizing effect of resampling, and Gerber and Chopin (2015) consider the use of quasi-random numbers. One could resample a different number of particles N_t at each step t, and the optimal allocation of these numbers is discussed in Lee and Whiteley (2018). Resampling is the key difference between Annealed Importance Sampling (Neal 2001) and SMCS, and we recover the former by omitting the resampling steps, that is, $\check{x}_{t-1}^n = x_{t-1}^n$ in Step 2(a). Resampling makes the SMCS output non-differentiable with respect to the input, which has motivated works such as Corenflos et al. (2021).

The output of SMCS includes weighted particles $(w_t^n, x_t^n)_{n\in[N]}$ approximating π_t, in the sense that $\pi_t^N(\varphi) = \sum_{n\in[N]} w_t^n \varphi(x_t^n)$ converges to $\pi_t(\varphi) = \int_{\mathsf{X}} \varphi(x_t)\pi_t(dx_t)$, for suitable $\varphi : \mathsf{X} \to \mathbb{R}$, as $N \to \infty$. Another output of the algorithm is an unbiased and consistent normalizing constant estimator Z_t^N, computed using the unnormalized weights.

Algorithm 1 Sequential Monte Carlo sampler

Input: sequence of distributions (π_t), forward Markov kernels (M_t), backward Markov kernels (L_t), resampling distribution $r(\cdot|w^{1:N})$ on $[N]^N$ where $w^{1:N}$ is an N-vector of probabilities.

1. Initialization.
   (a) Sample particle x_0^n from $\pi_0(\cdot)$ for $n \in [N]$ independently.
   (b) Set $w_0^n = N^{-1}$ for $n \in [N]$.
2. For $t \in [T]$, iterate the following steps.
   (a) Sample ancestor indices $(a_{t-1}^n)_{n\in[N]}$ from $r(\cdot|w_{t-1}^{1:N})$, and define $\check{x}_{t-1}^n = x_{t-1}^{a_{t-1}^n}$ for $n \in [N]$.
   (b) Sample particle $x_t^n \sim M_t(\check{x}_{t-1}^n, \cdot)$ for $n \in [N]$.
   (c) Compute weights $w_t(\check{x}_{t-1}^n, x_t^n) = \frac{\gamma_t(x_t^n)L_{t-1}(x_t^n, \check{x}_{t-1}^n)}{\gamma_{t-1}(\check{x}_{t-1}^n)M_t(\check{x}_{t-1}^n, x_t^n)}$ for $n \in [N]$,
   and set $w_t^n \propto w_t(\check{x}_{t-1}^n, x_t^n)$ such that $\sum_{n\in[N]} w_t^n = 1$.

Output: weighted particles $(w_t^n, x_t^n)_{n\in[N]}$ approximating π_t, and estimator $Z_t^N = \prod_{s=1}^t N^{-1} \sum_{n\in[N]} w_s(\check{x}_{s-1}^n, x_s^n)$ of Z_t for $t \in [T]$.

2.2. Paths of Distributions

The initial distribution $\pi_0(dx) = \gamma_0(x)dx/Z_0$ and target distribution $\pi(dx) = \gamma(x)dx/Z$ are considered inputs of the problem. We consider the choice of a path $\pi_t(dx) = \gamma_t(x)dx/Z_t$ for $t \in [T]$, where the number of distributions T can be user-specified or determined adaptively (see Section 2.4).

Geometric path. A popular choice is the geometric path

$$\gamma_t(x) = \gamma_0(x)^{1-\lambda_t}\gamma(x)^{\lambda_t}, \tag{2.2}$$

defined by a sequence $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = 1$, often referred to as inverse temperatures (Kirkpatrick, Jr., Gelatt, C. D., and Vecchi 1983). The unnormalized density $\gamma_t(x)$ and its gradient can be evaluated pointwise if it is possible to evaluate $\gamma_0(x)$ and $\gamma(x)$ and their gradients. If π_0 is a proper prior and π the posterior distribution, then the geometric path corresponds to raising the likelihood to a power. There could be statistical reasons to consider the resulting "tempered" posteriors, for example, see Holmes and Walker (2017) and references therein. A run of SMCS using (2.2) provides approximations of tempered posteriors over a desired sequence of exponents. Geometric paths can be generalized into q-paths, where the geometric mean in (2.2) is replaced by a power mean (Whitfield, Bu, and Straub 2002); see Syed et al. (2021) in the context of parallel tempering.

Path of partial posteriors. In the Bayesian setting, with a prior distribution $\pi_0(dx) = p(dx)$ and the target distribution $\pi(dx) = p(dx|y_{1:T})$ defined as the posterior based on observations $y_{1:T} = (y_1, \ldots, y_T)$, Chopin (2002) proposed SMCS, then termed Iterated Batch Importance Sampling, applied to the sequence of "partial posteriors" $\pi_t(dx) = p(dx|y_{1:t})$ for $t \in [T]$. The procedure provides a richer analysis compared to the approximation of $p(dx|y_{1:T})$ alone, helping one to visualize how the posterior distribution and the marginal likelihood evolve as data points are assimilated. Concepts such as Bayesian updating, sequential analysis and coherency are often presented as central in Bayesian theory (e.g., sec. 2.4.4 of Bernardo and Smith (2009) or Hooten, Johnson, and Brost (2019)), and yet it is common in applied Bayesian analysis to examine only the posterior distribution given all the data. SMCS on the path of partial posteriors enable the consideration of the dynamics of Bayesian inference, as will be illustrated in Section 5. This path is also key to the assessment of sequential predictive performance, such as with the Hyvärinen score (Dawid and Musio 2015; Shao et al. 2019), and plays a special role in Bayesian sequential experimental design (Drovandi, McGree, and Pettitt 2014; Cuturi et al. 2020). The path of partial posteriors is particularly appealing for time series, and when combined with particle filters for nonlinear state space models, the technique is known as SMC^2 (Chopin, Jacob, and Papaspiliopoulos 2013; Fulop and Li 2013).

Improved performance can be obtained by introducing a geometric path between successive partial posteriors. In the presence of improper priors, the initial distribution cannot be set as the prior. As an alternative, one can start the algorithm using a geometric path between some proper distribution and the posterior distribution given enough observations for it to be proper.

Path of truncated distributions. In rare event estimation, the task is to approximate the probability of a set $A \in \mathscr{X}$ under a distribution $\mu(dx) = \mu(x)dx$ defined on $(\mathsf{X}, \mathscr{X})$. Following Cérou et al. (2012), we consider sets $A = \{x \in \mathsf{X} : \Phi(x) \geq \ell\}$ for a function $\Phi : \mathsf{X} \to \mathbb{R}$ and a level $\ell \in \mathbb{R}$. We can define

$$\gamma_t(x) = \mu(x)\mathbb{I}_{A(\ell_t)}(x), \tag{2.3}$$

where $-\infty = \ell_0 < \ell_1 < \cdots < \ell_T = \ell$ is a sequence of levels, and $\mathbb{I}_{A(l)}(x)$ denotes the indicator function on the set $A(l) = \{x \in \mathsf{X} : \Phi(x) \geq l\}$. This defines a path of distributions that gradually truncates $\pi_0(dx) = \mu(dx)$ to $\pi(dx) = \mu(dx)\mathbb{I}_A(x)/Z$, whose normalizing constant $Z = \mu(A)$ is the probability of interest.

In the Bayesian setup where π_0 and π denote a (proper) prior and posterior, respectively, nested sampling (Skilling 2006) represents the marginal likelihood as $Z = \int_0^\infty \pi_0(A(l))dl$ with $A(l)$ defined by level sets of the likelihood function $\Phi(x) = \gamma(x)/\gamma_0(x)$. This identity is leveraged by Salomone et al. (2018) to apply SMCS with the path (2.3). Approximate Bayesian computation provides another setting where a sequence of truncated distributions, indexed by a "tolerance" parameter, can be estimated by SMCS (Sisson, Fan, and Tanaka 2007; Del Moral, Doucet, and Jasra 2012).

Path of least coding effort. In anticipation of the choice of forward Markov kernels (M_t), we might want to introduce a path of distributions such that the associated Markov kernels are readily available. One might already have an MCMC algorithm that targets π, for example a Gibbs sampler that exploits specific aspects of π. To reduce implementation effort, we can then introduce a path (π_t) designed so that only slight modifications to that MCMC algorithm are required. For example, the implementation of Langevin or Hamiltonian Monte Carlo to target any distribution on the geometric path requires minimal modifications relative to the original target π.

Another example can be found in Rischard, Jacob, and Pillai (2018), in the context of logistic regression. Assuming a Normal prior on the regression coefficients, the Pólya–Gamma Gibbs (PGG) sampler of Polson, Scott, and Windle (2013) can be employed to target the posterior distribution, for any matrix of covariates x and binary outcome vector y. We can introduce a path of posterior distributions π_t corresponding to the use of scaled covariates $\lambda_t x$ instead of x, with $\lambda_t \in [0, 1]$. The appeal is that the same implementation of PGG, given inputs $\lambda_t x$ and y, provides a forward kernel M_t for each distribution π_t. A similar approach was considered for probit regressions in Del Moral, Doucet, and Jasra (2007).

There is much freedom in the choice of paths, so that various settings and goals can be accommodated. Sequences of distributions can be further generalized to sets of distributions indexed by trees, with applications to Bayesian hierarchical models in Lindsten et al. (2017).

2.3. Forward and Backward Markov Kernels

Given a path (π_t), the SMCS user must select forward and backward kernels, (M_t) and (L_t). In view of Algorithm 1, one must be able to sample from $M_t(x_{t-1}, \cdot)$ and to evaluate $w_t(x_{t-1}, x_t)$ in (2.1). We would set $M_t(x_{t-1}, dx_t) = \pi_t(dx_t)$ if perfect samples from π_t could be obtained, and by defining $L_{t-1}(x_t, dx_{t-1}) = \pi_{t-1}(dx_{t-1})$ the weight would simplify to Z_t/Z_{t-1}, leading to an estimator of Z_t with zero variance. This section presents more practical choices.

Exact MCMC moves. We can exploit the vast literature on MCMC to design M_t as a π_t-invariant kernel. Although such choices typically do not admit tractable transition densities,

the weight in (2.1) can be tractable if the backward kernel L_{t-1} is chosen judiciously. Following Jarzynski (1997) and Neal (2001), L_{t-1} can be selected as the time reversal of M_t, that is, $L_{t-1}(x_t, x_{t-1}) = \pi_t(x_{t-1})M_t(x_{t-1}, x_t)/\pi_t(x_t)$, leading to the weight $\gamma_t(x_{t-1})/\gamma_{t-1}(x_{t-1})$. When the distributions π_{t-1} and π_t are close, the time reversal provides an approximation of the backward transition $L_{t-1}(x_t, x_{t-1}) = \pi_{t-1}(x_{t-1})M_t(x_{t-1}, x_t)/q_t(x_t)$ yielding minimal variance. Del Moral, Doucet, and Jasra (2006, sec. 3.3) provide more discussions on the choice of (L_t) given (M_t).

SMCS can also accommodate kernels M_t that are not π_t-invariant, while preserving consistency of SMC estimates. We provide an example that shows how to remove time-discretization biases without resorting to Metropolis–Rosenbluth–Teller–Hastings corrections.

Unadjusted Langevin moves. For problems on $\mathsf{X} = \mathbb{R}^d$, we can select forward kernels based on the unadjusted Langevin algorithm (ULA, Grenander and Miller 1994):

$$M_t(x_{t-1}, dx_t) = \mathcal{N}\left(x_t; x_{t-1} + \frac{\varepsilon}{2}\Omega\nabla \log \pi_t(x_{t-1}), \varepsilon\Omega\right) dx_t, \tag{2.4}$$

where $z \mapsto \mathcal{N}(z; \mu, \Sigma)$ denotes the density of a Normal distribution with mean vector μ and covariance matrix Σ, $\varepsilon > 0$ denotes a step size, and $\Omega \in \mathbb{R}^{d\times d}$ is a positive definite preconditioning matrix. In general the ULA transition does not leave π_t invariant for any $\varepsilon > 0$. When an acceptance correction step is added to enforce π_t-invariance, the resulting method is known as the Metropolis-adjusted Langevin algorithm (MALA). In SMCS one can account for the time-discretization using importance sampling instead. The reversibility of the underlying continuous-time Langevin diffusion suggests the choice $L_{t-1}(x_t, dx_{t-1}) = M_t(x_t, dx_{t-1})$ for sufficiently small ε (Nilmeier et al. 2011). With these choices, the weight (2.1) is tractable, and approaches $\gamma_t(x_{t-1})/\gamma_{t-1}(x_{t-1})$ as $\varepsilon \to 0$. The tractability of ULA kernels as an alternative to MALA kernels within SMCS was exploited in the controlled sequential Monte Carlo approach (Heng et al. 2020), which optimizes over the path of distributions (π_t) and forward kernels (M_t) to improve performance, and in the Schrödinger bridge sampler (Bernton et al. 2019) that fixes (π_t) and optimizes over (M_t) and (L_{t-1}) for similar purposes. As an alternative, the supplementary materials describe the use of unadjusted Hamiltonian Monte Carlo (HMC) moves within SMCS. One can also design the forward kernel M_t as a deterministic map that transports π_{t-1} to π_t, and choose the backward kernel L_{t-1} as the inverse map (Vaikuntanathan and Jarzynski 2008; Everitt et al. 2020; Arbel, Matthews, and Doucet 2021; Heng, Doucet, and Pokern 2021; Matthews et al. 2022).

Tuning parameters. Having chosen Markov kernels, there might be some algorithmic parameters to tune. Firstly, it is often worthwhile to iterate the chosen Markov kernel more than once at each step of SMCS. For π_t-invariant forward kernels with time reversals as backward kernels, iterating the forward kernel can be done without modification of the weights. For unadjusted kernels additional care might be required. When we iterate MCMC moves, or when we perform moves that involve intermediate steps such as HMC, it can be advantageous to exploit all intermediate samples (Dau and Chopin 2022); see also the supplementary material.

Each Markov kernel may further depend on parameters such as step sizes, or preconditioning matrices. An appealing specificity of SMCS, relative to classical MCMC, is that approximations of the previous and current bridging distributions are available and can be used to inform the choice of parameters. For example one can select Ω as the estimated covariance of bridging distributions for random walk or MALA moves (Chopin 2002). Fearnhead and Taylor (2013) offer a generic recipe to automate such tuning procedures. Kostov (2006) and Buchholz, Chopin, and Jacob (2020) consider specifically HMC kernels, and Schäfer and Chopin (2013) and South, Pettitt, and Drovandi (2019) discuss strategies to adapt independent proposals on discrete and continuous spaces. While these adaptation rules will not affect consistency properties of SMCS as $N \to \infty$ (Beskos et al. 2016), they may not preserve the unbiasedness property of normalizing constant estimators. When this unbiasedness matters, for example in particle MCMC methods as in Section 4.2, one can run an adaptive SMCS, record the obtained tuning parameters and run a second, nonadaptive sampler, for an approximate 2-fold increase in computing cost.

2.4. Progressing Through a Path of Distributions

The user also needs to address the choice of the number of distributions T and of the particular distributions π_t along a given path. In the case of a geometric path (2.2), one needs a specific choice of inverse temperatures $(\lambda_t)_{t\in[T]}$. We could simply prespecify T and select $\lambda_t = (t/T)^p$ for $t \in [T]$ and some exponent $p > 0$, informed by preliminary runs. The following describes a common procedure that specifies T and $(\lambda_t)_{t\in[T]}$ adaptively. For clarity, we consider only the setting where the forward kernel M_t is π_t-invariant and the backward kernel L_{t-1} is its time reversal, leading to the weight

$$w_t(x_{t-1}) = \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})} = \frac{\gamma(x_{t-1})^{\lambda_t-\lambda_{t-1}}}{\gamma_0(x_{t-1})}. \tag{2.5}$$

As particle weights do not depend on their states at step t in this setting, one can perform weighting (Step 2(c)) and resampling (Step 2(a)) before applying Markov moves (Step 2(b)) to promote sample diversity in Algorithm 1. Equation (2.5) can be seen as an importance weight targeting π_t using proposed samples from π_{t-1}. Suppose π_{t-1} has been determined by some $\lambda_{t-1} \in [0, 1)$ and we seek $\lambda_t \in (\lambda_{t-1}, 1]$ so that π_t can be well-approximated by π_{t-1} through importance sampling. We can control the performance by keeping the χ^2-divergence small (Agapiou et al. 2017), where

$$\begin{aligned}\chi^2(\pi_t|\pi_{t-1}) &= \int_{\mathsf{X}} \left(\frac{\pi_t(x)}{\pi_{t-1}(x)} - 1\right)^2 \pi_{t-1}(dx) \\ &= \frac{\int_{\mathsf{X}} w_t(x)^2\pi_{t-1}(dx)}{\left(\int_{\mathsf{X}} w_t(x)\pi_{t-1}(dx)\right)^2} - 1.\end{aligned} \tag{2.6}$$

Instead of fixing $\chi^2(\pi_t|\pi_{t-1})$ to a desired level, it is more convenient to work with $\varrho_t(\lambda_t) = (1+\chi^2(\pi_t|\pi_{t-1}))^{-1}$ as this quantity

takes values in $[0,1]$. Given samples $(x_{t-1}^n)_{n\in[N]}$ approximating π_{t-1}, an approximation of $\varrho_t(\lambda_t)$ is given by $\hat{\varrho}_t(\lambda_t) = \mathrm{ESS}_t(\lambda_t)/N$, where

$$\mathrm{ESS}_t(\lambda_t) = \frac{\left(\sum_{n=1}^N w_t(x_{t-1}^n)\right)^2}{\sum_{n=1}^N w_t(x_{t-1}^n)^2} = \frac{\left(\sum_{n=1}^N (\gamma/\gamma_0)(x_{t-1}^n)^{\lambda_t-\lambda_{t-1}}\right)^2}{\sum_{n=1}^N (\gamma/\gamma_0)(x_{t-1}^n)^{2(\lambda_t-\lambda_{t-1})}}. \tag{2.7}$$

This effective sample size (ESS) (Kong, Liu, and Wong 1994) takes values in $[1,N]$, achieving the lower bound when one sample holds all the weight and the upper bound when all samples have equal weights.

If $\hat{\varrho}_t(1)$ is greater than a prespecified threshold $\kappa \in (0,1)$, we set $\lambda_t = 1$. Otherwise, we can solve for $\lambda_t \in (\lambda_{t-1}, 1)$ such that $\hat{\varrho}_t(\lambda_t)$ is equal to κ (Jasra et al. 2011). As this enforces the χ^2-divergence between successive distributions to be approximately $\delta = \kappa^{-1} - 1$ in the large N regime, higher thresholds lead to more bridging distributions T. The search for λ_t can be implemented using the bisection method on the interval $[\lambda_{t-1}, 1]$ as the function $\hat{\varrho}_t(\lambda_t)$ is strictly decreasing (Beskos et al. 2016, Lemma 3.1). The cost of this procedure is negligible as evaluations of (2.7) are inexpensive once $(\gamma/\gamma_0)(x_{t-1}^n)$ have been pre-computed. Note that when the path is fixed, adaptively resampling whenever the ESS falls below a threshold does not alter unbiasedness (Whiteley, Lee, and Heine 2016).

Modifications and alternatives to the ESS criterion are proposed in Cornebise, Moulines, and Olsson (2008), Zhou, Johansen, and Aston (2016), and Huggins and Roy (2019), and can be used to select intermediate distributions. An alternative method to determine (λ_t) adaptively is described in Nguyen et al. (2015). In any case, starting with a fixed or an adaptive schedule, we can re-run SMCS with additional intermediate steps to improve performance without necessarily increasing the number of particles, as mentioned in Section 4.2 and the supplementary materials.

3. Effect of Bridging Distributions

3.1. The Curse of Dimension for Importance Sampling

The cost of IS estimators is related to the discrepancy between the proposal and target distributions, as measured by the χ^2 or KL divergence (Agapiou et al. 2017; Chatterjee and Diaconis 2018). For example, the number of samples needed for an estimator of Z to achieve a given variance is proportional to that χ^2-divergence. As the dimension $d \in \mathbb{N}$ of X grows, the χ^2 and KL divergences between π_0 and π would often increase exponentially with d. Since each step of SMCS involves importance sampling, concerns about their performance in high dimension are understandable. Remarkably, SMCS can deliver reliable estimates for problems in high dimension; see, for example, the applications to inverse problems in Kantas, Beskos, and Jasra (2014), and to spatio-temporal models in Naesseth, Lindsten, and Schön (2015). We propose simple elements to explain this operational success.

3.2. Variance of the Normalizing Constant Estimator

We focus on the geometric path (2.2), forward π_t-invariant kernels (M_t), and backward kernels (L_t) taken as their time reversals. We consider the variance of the normalizing constant estimator

$$Z_T^N = \prod_{t=1}^T \frac{1}{N}\sum_{n=1}^N w_t(X_{t-1}^n), \tag{3.1}$$

produced by the modification of Algorithm 1 described in Section 2.4 to promote sample diversity. Cérou, Del Moral, and Guyader (2011) established a formula for this estimator's nonasymptotic variance in the setting of Feynman–Kac formulas. Our first step is to make a simplifying assumption that allows us to capture some of the essence of Cérou, Del Moral, and Guyader (2011) with only simple calculations.

Assumption 3.1. For all $t \in [T]$, the forward kernel is perfect: $M_t(x_{t-1}, dx_t) = \pi_t(dx_t)$.

Our priority here is exposition rather than realism, but in practice if M_t is taken as multiple iterations of an ergodic kernel targeting π_t, then Assumption 3.1 essentially holds if the number of iterates is large enough. Using the unbiased property of Z_T^N and the identity $Z = \prod_{t=1}^T Z_t/Z_{t-1} = \prod_{t=1}^T \{\int_{\mathsf{X}} w_t(x_{t-1})\pi_{t-1}(dx_{t-1})\}$, a calculation shows that

$$\mathrm{var}\left[\frac{Z_T^N}{Z}\right] = \prod_{t=1}^T \left[1 + \frac{\chi^2(\pi_t|\pi_{t-1})}{N}\right] - 1. \tag{3.2}$$

Hence, χ^2-divergences between consecutive distributions appear in the variance of Z_T^N/Z.

3.3. Scaling the Number of Bridging Distributions with Dimension

Next we introduce a sequence of sampling problems, indexed by $d \in \mathbb{N}$. We will specify the inverse temperatures (λ_t) in a way that possibly depends on d. The next assumption captures the idealized performance of the adaptive procedure of Section 2.4 as $N \to \infty$, and serves to dispense with technical subtleties that would arise if the number of distributions T was random.

Assumption 3.2. For all $t \in [T-1]$, π_{t-1} and π_t satisfy $\chi^2(\pi_t|\pi_{t-1}) \le \delta$ for some prespecified $\delta > 0$ which is independent of d, and such that $\chi^2(\pi|\pi_0) > \delta$.

Assumption 3.3. There exists $\alpha > 0$ such that $T = O(d^\alpha)$ as $d \to \infty$.

The above assumption postulates how the number of bridging distributions T scales with dimension d. The fact that Assumptions 3.2–3.3 can hold simultaneously will be illustrated through examples. Since the χ^2-divergence between successive distributions is fixed as $\delta = \kappa^{-1} - 1$ under Assumption 3.2, the relative variance in (3.2) is equal to $(1+\delta/N)^T - 1$. As $d \to \infty$ and hence $T \to \infty$, to ensure stability of the

estimator (3.1), we can choose the number of particles N such that $N = O(T)$ to keep the relative variance of a constant order (since $\lim_{N\to\infty}(1+\delta/N)^N = \exp(\delta)$). Therefore, the overall cost of this idealized SMCS measured in terms of density evaluations would be $O(T^2) = O(d^{2\alpha})$, that is polynomial in d, under Assumption 3.3. We next consider Assumptions 3.2–3.3 on a Normal example.

Example 3.1. Set $\pi_0(dx) = \mathcal{N}(x;\mu_0,\Sigma)dx$ and $\pi(dx) = \mathcal{N}(x;\mu,\Sigma)dx$ for some $\mu_0,\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$. Each distribution along the geometric path (2.2) is Normal $\pi_t(dx) = \mathcal{N}(x;\mu_t,\Sigma)dx$ with mean vector $\mu_t = \mu_0 + \lambda_t(\mu-\mu_0)$ for $t \in [T]$. The χ^2-divergence between successive distributions can be computed in closed-form: $\chi^2(\pi_t|\pi_{t-1}) = \exp((\lambda_t-\lambda_{t-1})^2|\mu-\mu_0|^2_{\Sigma^{-1}}) - 1$, where $|\mu-\mu_0|_{\Sigma^{-1}} = \sqrt{(\mu-\mu_0)^\top\Sigma^{-1}(\mu-\mu_0)}$. Under the specification

$$T = \lceil |\mu-\mu_0|_{\Sigma^{-1}}/\sqrt{\log(1+\delta)}\rceil,$$
$$\lambda_t = t\sqrt{\log(1+\delta)}/|\mu-\mu_0|_{\Sigma^{-1}} \text{ for } t \in [T-1], \quad (3.3)$$

where $\lceil\cdot\rceil$ denotes the ceiling function, Assumption 3.2 is satisfied. Using the bound $|\mu-\mu_0|_{\Sigma^{-1}} \le \Lambda_{\min}(\Sigma)^{-1/2}|\mu-\mu_0|$, where $\Lambda_{\min}(\Sigma)$ denotes the minimum eigenvalue of Σ, it follows from (3.3) that $T = O(\sqrt{d})$ if $\Lambda_{\min}(\Sigma)$ is uniformly bounded away from zero and $|\mu-\mu_0|$ is $O(\sqrt{d})$, both as $d\to\infty$. In this situation, Assumption 3.3 holds with $\alpha = 1/2$.

To address less specific examples on $\mathsf{X} = \mathbb{R}^d$, we consider assumptions along the path $\pi(\lambda,dx) = \gamma(\lambda,x)dx/Z(\lambda)$ for $\lambda \in [0,1]$, where $\gamma(\lambda,x) = \gamma_0(x)^{1-\lambda}\gamma(x)^\lambda$ and $Z(\lambda) = \int_{\mathsf{X}}\gamma(\lambda,x)dx$. The densities $\gamma_0(x)$ and $\gamma(x)$ are assumed to be continuously differentiable. We will write the expectation of $\varphi : \mathbb{R}^d \to \mathbb{R}$ with respect to $\pi(\lambda,dx)$ as $\pi(\lambda,\varphi) = \int_{\mathsf{X}}\varphi(x)\pi(\lambda,dx)$ and $\ell(x) = \log(\gamma(x)/\gamma_0(x))$.

Assumption 3.4. There exist constants $C,\zeta > 0$ and a function $\beta : [0,1] \to \mathbb{R}_+$ with $\inf_{\lambda\in[0,1]}\beta(\lambda) > 0$ such that for each $\lambda \in [0,1]$, the distribution $\pi(\lambda,dx)$ along the geometric path satisfies:

(i) a Poincaré inequality with constant $\beta(\lambda)$, that is, for all differentiable $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have $\pi(\lambda,\varphi^2) - \pi(\lambda,\varphi)^2 \le \beta(\lambda)^{-1}\pi(\lambda,|\nabla\varphi|^2)$;
(ii) a bound on the maximum log-likelihood, $\sup_{x\in\mathsf{X}}\ell(x) \le Cd^\zeta$;
(iii) a bound on the expected log-likelihood, $\pi(\lambda,\ell) \ge -Cd^\zeta$;
(iv) a bound on the expected squared norm of the log-likelihood, $\pi(\lambda,|\nabla\ell|^2) \le Cd^{2\zeta}$.

The Poincaré inequality is an isoperimetric condition with rich implications, such as the exponential convergence of certain MCMC algorithms (Vempala and Wibisono 2019) which can be used to generalize the discussion in Section 3.2 by relaxing the assumption of perfectly mixing kernels (Schweizer 2012a). We refer to references in Vempala and Wibisono (2019, pp. 7 and 16) for conditions to verify a Poincaré inequality and we recall that it is implied by strong log-concavity of the distribution in question. Under Assumption 3.4, we can verify Assumptions 3.2 and 3.3. At step $t \in [T]$, the χ^2-divergence of $\pi_{t-1}(dx) = \pi(\lambda_{t-1},dx)$ from $\pi_t(dx) = \pi(\lambda_t,dx)$ for $0 \le \lambda_{t-1} < \lambda_t \le 1$ can be bounded,

$$\chi^2(\pi_t|\pi_{t-1}) \le \beta(\lambda_{t-1})^{-1}(\lambda_t-\lambda_{t-1})^2 \int_{\mathsf{X}}\frac{\pi_t(x)}{\pi_{t-1}(x)}|\nabla\ell(x)|^2\pi_t(dx). \quad (3.4)$$

This follows from Assumption 3.4(i) for the distribution $\pi(\lambda_{t-1},dx)$ and the function $\varphi(x) = \pi_t(x)/\pi_{t-1}(x)$. To upper bound the ratio of densities in (3.4), we consider

$$\begin{aligned}\log\pi_t(x) - \log\pi_{t-1}(x) &= (\lambda_t-\lambda_{t-1})\ell(x) \\ &\quad - (\log Z_t - \log Z_{t-1}) \quad (3.5)\\ &= (\lambda_t-\lambda_{t-1})(\ell(x) - \pi(\lambda_t^*,\ell))\end{aligned}$$

which holds for some $\lambda_t^* \in (\lambda_{t-1},\lambda_t)$ using the mean value theorem. Hence, using Assumption 3.4(ii)–(iii), we have

$$\sup_{x\in\mathsf{X}}\frac{\pi_t(x)}{\pi_{t-1}(x)} \le \exp(2C(\lambda_t-\lambda_{t-1})d^\zeta). \quad (3.6)$$

Applying this upper bound in (3.4), Assumption 3.4(iv) and the lower bound $\underline{\beta} = \inf_{\lambda\in[0,1]}\beta(\lambda)$, we obtain

$$\chi^2(\pi_t|\pi_{t-1}) \le \underline{\beta}^{-1}(\lambda_t-\lambda_{t-1})^2\exp(2C(\lambda_t-\lambda_{t-1})d^\zeta)Cd^{2\zeta}. \quad (3.7)$$

If we construct a sequence (λ_t) with increment $\lambda_t - \lambda_{t-1} = cd^{-\zeta}$, the constant $c > 0$ can be chosen so that Assumption 3.2 holds, and Assumption 3.3 is satisfied since $T = O(d^\zeta)$.

Formal studies on SMCS in high dimension include that of Beskos, Crisan, and Jasra (2014), which provides stability results in settings where the target π can be factorized into independent components, and discusses the behavior of the required number of bridging steps and of the effective sample size. Relevant discussions can also be found in sec. 6 of Schweizer (2012a), where solid reasons are given to support a polynomial dimension dependence; see also Brosse, Durmus, and Moulines (2018). The supplementary materials contain numerical illustrations of the performance of SMCS in increasing dimensions.

4. Parallel Execution and Confidence Intervals

Having specified the ingredients of SMCS, the user has more than one way of running these algorithms, which leads to different perspectives on the use of parallel processors and on the quantification of errors.

4.1. Interacting Particle Systems

SMCS are instances of interacting particle systems, or equivalently Monte Carlo approximations of Feynman–Kac models. This view has proven fruitful and allows the application of readily-available results (Del Moral 2004, 2013). In particular the estimators $\pi_t^N(\varphi)$, for a function φ, and Z_t^N satisfy central limit theorems:

$$\sqrt{N}(\pi_t^N(\varphi) - \pi_t(\varphi)) \xrightarrow{d.} \mathcal{N}(0,v_t(\varphi)), \quad (4.1)$$
$$\sqrt{N}(Z_t^N/Z_t - 1) \xrightarrow{d.} \mathcal{N}(0,v_t^\star), \quad (4.2)$$

for each t as $N\to\infty$, where $\xrightarrow{d.}$ denotes convergence in distribution, and $v_t(\varphi), v_t^\star > 0$ denote asymptotic variances.

Therefore, valid confidence intervals can be derived from consistent estimators of the asymptotic variances. Such estimators were derived in Chan and Lai (2013), Lee and Whiteley (2018), and Du and Guyader (2021), and address a long-standing gap on the quantification of errors in SMC.

We present a result from Lee and Whiteley (2018) that is valid when multinomial resampling is employed. We introduce the "lineage" of the nth particle at step t:

$$b^n_{t,t} = n, \quad \text{and} \quad b^n_{s-1,t} = a^{b^n_{s,t}}_{s-1} \quad \text{for } 1 \le s \le t, \tag{4.3}$$

where the ancestor indices (a^n_t) are defined when resampling in Step 2(a) of Algorithm 1. Since only the offspring of the particles indexed by $b^{1:N}_{0,t}$ survive at time t, we will refer to such indices as "roots." For a function φ, consider the quantity

$$V^N_t(\varphi) = \pi^N_t(\varphi)^2 - \left(\frac{N}{N-1}\right)^{t+1} \frac{1}{N^2} \sum_{n,m:\, b^n_{0,t} \neq b^m_{0,t}} \varphi(x^n_t)\varphi(x^m_t), \tag{4.4}$$

which can be computed as a by-product of SMCS. Theorem 1 of Lee and Whiteley (2018) states the convergence in probability of $N \cdot V^N_t(\varphi - \pi^N_t(\varphi))$ to $v_t(\varphi)$, and of $N \cdot V^N_t(1)$ to $v^\star_t$, as $N \to \infty$. We can directly write

$$V^N_t(1) = 1 - \left(\frac{N}{N-1}\right)^{t+1} + \left(\frac{N}{N-1}\right)^{t+1} \times \frac{1}{N^2} \sum_{n \in [N]} |\{m\colon b^m_{0,t} = b^n_{0,t}\}|^2. \tag{4.5}$$

The right-hand side features the cardinality of the set of siblings of particle n, that is, the particles that have the same ancestor at time zero. If all particles were siblings, the sum would be of order N^2 and thus the estimated variance would be away from zero, but if all particles have a small number of siblings, the variance estimator is of order N^{-1}.

The $N \to \infty$ regime underpinning the above results is compatible with parallel computing since the propagation of particles and the calculation of weights can be distributed across processors. The resampling step, on the other hand, induces interactions and thus synchronization and communication (Jun, Wang, and Bouchard-Côté 2012; Murray, Lee, and Jacob 2016), so that SMCS are not fully parallelizable. Another limitation is that a direct implementation requires N particles in memory, which can be limiting in certain settings; Jun and Bouchard-Côté (2014) propose a memory-efficient implementation. Finally SMCS in the large N regime are not "anytime": they run for T steps before returning their output and then stop. In contrast, MCMC methods are easier to interrupt and resume. These shortcomings have motivated variants where individual particles can be added sequentially (Brockwell, Del Moral, and Doucet 2010; Paige et al. 2014; Finke, Doucet, and Johansen 2020).

4.2. Independent Systems of Fixed Xize

Consider R independent SMCS runs, with a fixed number of particles N. The runs can be executed on parallel machines, and denote by $(\pi^{N,r})_{r\in[R]}$ and by $(Z^{N,r})_{r\in[R]}$ the resulting particle approximations of π and Z, respectively. We can obtain consistent approximations of π and Z as $R \to \infty$, even though N is fixed (e.g., Whiteley, Lee, and Heine 2016; Rainforth et al. 2016). Since the normalizing constant estimator is unbiased, we can directly average $(Z^{N,r})_{r\in[R]}$ to obtain a consistent estimator of Z as $R \to \infty$. Estimating expectations under π seems more involved as the estimator $\pi^N_t(\varphi)$ is itself biased when N is fixed. We describe how to correct this in the framework of Andrieu, Doucet, and Holenstein (2010).

We select a particle among the N available ones at the terminal step of Algorithm 1, by sampling $k \in [N]$ with probabilities $w^{1:N}_T$, and returning x^k_T. The distribution of all random variables generated by the procedure is

$$q^N(k, \bar{x}, \bar{a}) = \left\{\prod_{n\in[N]} \pi_0(x^n_0)\right\} \prod_{t=1}^{T} \left\{ r(a^{1:N}_{t-1} | w^{1:N}_{t-1}) \prod_{n\in[N]} M_t(x^{a^n_{t-1}}_{t-1}, x^n_t) \right\} w^k_T, \tag{4.6}$$

where $\bar{x} = (x^n_t)_{n\in[N]}$ for $0 \le t \le T$ and $\bar{a} = (a^n_t)_{n\in[N]}$ for $0 \le t \le T-1$. Next we define

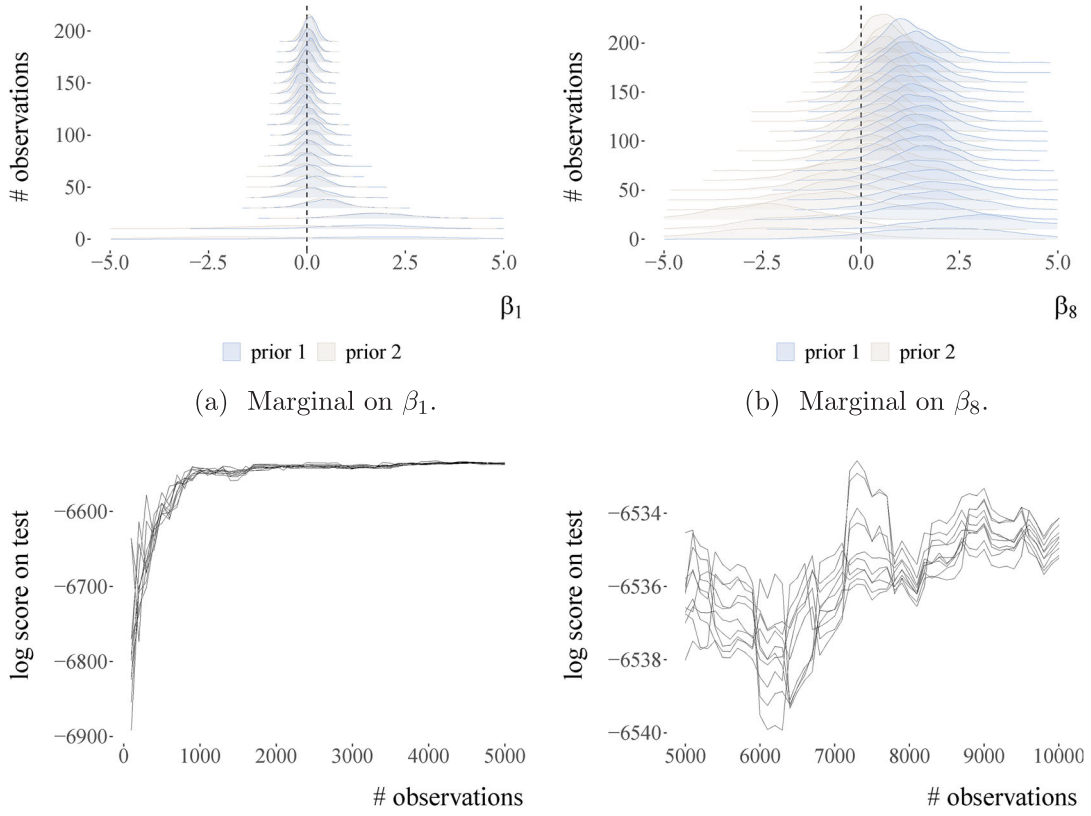
$$\bar{\pi}^N(k, \bar{x}, \bar{a}) = \frac{Z^N_T}{Z_T} q^N(k, \bar{x}, \bar{a}). \tag{4.7}$$

Under a mild assumption on the resampling scheme, Andrieu, Doucet, and Holenstein (2010) observe that (4.7) defines a valid probability distribution, and that its marginal distribution in x^k_T is the target π. Therefore, we can use q^N as a proposal and $\bar{\pi}^N$ as a target in an importance sampling argument, and the corresponding unnormalized weight is Z^N_T. For a function φ, a self-normalized importance sampling estimator after Rao-Blackwellizing the index k is thus,

$$\bar{\pi}^R(\varphi) = \frac{\sum_{r\in[R]} Z^{N,r} \pi^{N,r}(\varphi)}{\sum_{r'\in[R]} Z^{N,r'}}, \tag{4.8}$$

which approximates $\pi(\varphi)$ as $R \to \infty$, for any fixed N. Its asymptotic variance can be estimated consistently as $R \to \infty$ to construct confidence intervals (eq. (9.8) in Owen 2013). There are practical benefits of the large R asymptotics over the large N asymptotics: independent SMCS can be run on parallel machines without communication; results can be refined with more independent runs without hitting memory limits; the procedure is simple to implement, to interrupt and to resume. More advanced schemes where "islands" of particles are allowed to communicate have been studied in for example, Vergé et al. (2015) and Whiteley, Lee, and Heine (2016).

Equation (4.7) suggests the use of SMCS as an independent proposal in a Metropolis–Rosenbluth–Teller–Hastings algorithm (Andrieu, Doucet, and Holenstein 2010). Despite the iterative nature of MCMC, most of the computation here lies in the generation of the independent proposals, which can be fully parallelized. The approach lends itself to generic convergence diagnostics for MCMC (Brooks et al. 2011), and other tools developed for MCMC estimators, including variance reduction (Dellaportas and Kontoyiannis 2012) and debiasing techniques



(a) Marginal on β_1. (b) Marginal on β_8.

(c) Prediction of first 5000 observations. (d) Prediction of next 5000 observations.

Figure 2. Logistic regression with forest cover type data. Evolution of the posterior distribution of β_1 (*top-left*) and β_8 (*top-right*) as more data are assimilated, with initialization from the priors $\mathcal{N}(2, 3)$ (*blue*) and $\mathcal{N}(-2, 3)$ (*beige*). Performance of the posterior predictive distribution on a test dataset as the first 5000 (*bottom-left*) and next 5000 (*bottom-right*) observations are assimilated, estimated using independent runs of SMCS.

(Middleton et al. 2019). The latter reference describes a generic strategy that delivers unbiased estimators of $\pi(\varphi)$ using only standard SMCS runs, thus, bypassing the design of algorithm-specific couplings as in Jacob, O'Leary, and Atchadé (2020).

The performance of (4.8) clearly depends on the performance of each run of SMCS. If N is fixed to a low value, the performance can still be satisfactory provided that the other algorithmic ingredients are well-chosen. The supplementary materials present numerical experiments where the variance of $\log Z^N$ is seen to be stable in problems of increasing dimension d using a fixed value of N, thanks to an adequate scaling of the number of intermediate steps.

5. Illustrations

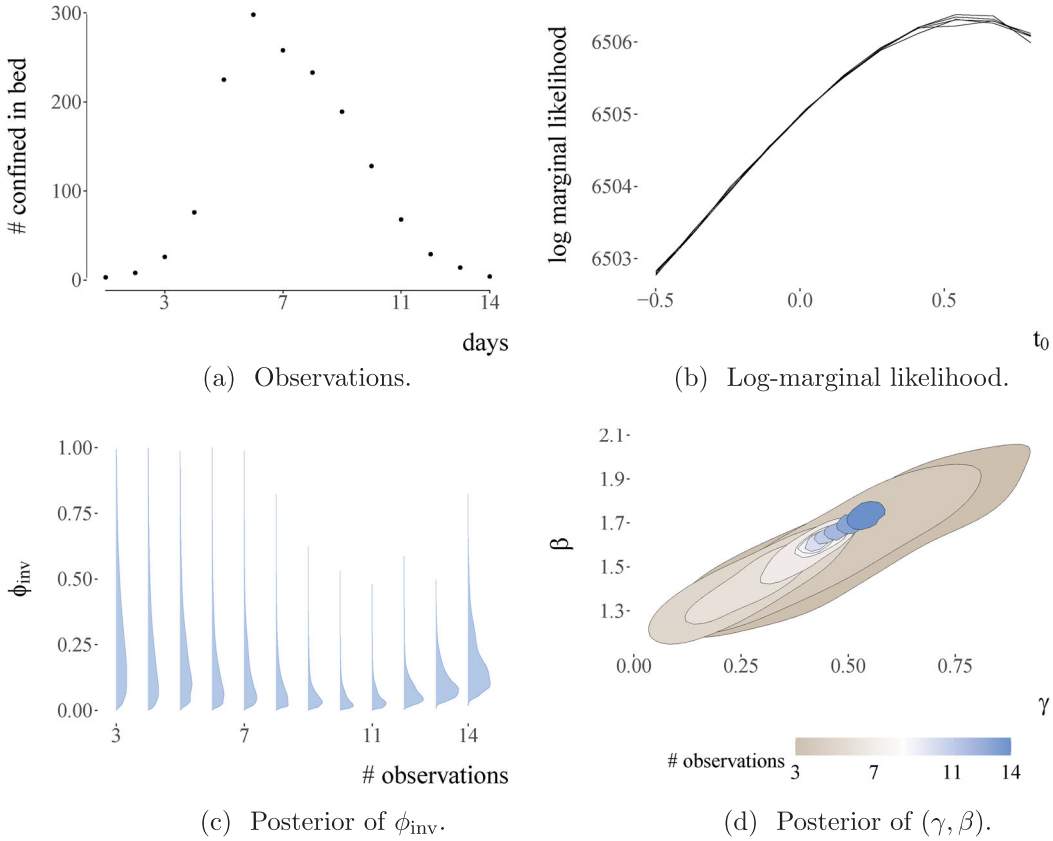
We illustrate some appealing properties of SMCS compared to MCMC methods in two simple examples; all details and the implementation are described in supplementary materials. More challenging problems have been tackled with SMCS, for example, in Bayesian nonparametrics (Griffin 2017), phylogenetic inference (Wang, Bouchard-Côté, and Doucet 2015), fiducial inference (Cisewski and Hannig 2012), financial econometrics (Fulop and Li 2013; Fulop et al. 2021), large-scale graphical models (Naesseth, Lindsten, and Schön 2014), partial differential equations (Beskos et al. 2017) and experimental design (Drovandi, McGree, and Pettitt 2014; Cuturi et al. 2020).

5.1. Logistic Regression

We consider a logistic regression $y = (y_1, \ldots, y_m) \in \{0, 1\}^m$ on covariates $x = (x_1, \ldots, x_m) \in \mathbb{R}^{m \times d}$. Under the model, y_i is a Bernoulli variable with success probability $(1 + \exp(-x_i^\top \beta))^{-1}$ where $\beta \in \mathbb{R}^d$ denote the regression coefficients. We use the "forest cover type" data (Blackard 2000), processed as in Collobert, Bengio, and Bengio (2002).[1] The data contain cartographic information (relating to altitude, slope, azimuth etc) for $30m$ by $30m$ cells in northern Colorado, along with the type of cover (originally spruce/fir, lodgepole pine, Ponderosa pine, cottonwood/willow, spruce/fir and aspen or Douglas-fir, and in Collobert, Bengio, and Bengio (2002) this was simplified to lodgepole pine versus the other categories combined). With a logistic regression, we predict the cover type using cartographic variables. There are $d = 11$ regression coefficients including the intercept, and the prior is Normal(0, 10) on each coefficient unless specified otherwise.

We illustrate the sequential aspect of Bayesian updating with SMCS using the path of partial posteriors; other paths are considered in the supplementary materials. Figure 2(a)–(b) show a phenomenon called "merging" whereby posteriors resulting from different priors eventually coincide as more observations are introduced. We observe that certain components of the posterior "merge" faster than others. Similar figures could help to visualize the Bernstein-von Mises phenomenon, whereby the

[1] *https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html*.



(a) Observations. (b) Log-marginal likelihood.

(c) Posterior of ϕ_{inv}. (d) Posterior of (γ, β).

Figure 3. SIR model with boarding school data. Observations of daily counts (*top-left*). Log-marginal likelihood of the initial time t_0 at which the first individual is assumed to be infected (*top-right*). Evolution of the marginal posterior distribution of ϕ_{inv} (*bottom-left*) and of (γ, β) (*bottom-right*) as more data are assimilated.

posterior distribution becomes closer to a Normal distribution as more data get assimilated.

Sequential inference allows us to monitor not only the evolution of beliefs but also measures of predictive performance. For example, Figure 2(c)–(d) show the logarithmic score associated with the posterior predictive distribution as the number of observations increases, on a test dataset. Predictive performance increases significantly as we start to assimilate data. After a certain point the predictive performance seems to stagnate. Indeed, under model misspecification, there is no guarantee that the posterior predictive performance would improve with more data. The ability to monitor performance can be helpful when deciding whether the model under consideration is able to benefit from the inclusion of more data. The arbitrariness of the ordering of the observations in the setting of regression can be addressed by averaging over orderings, as described in the supplementary materials, where it is also shown how Bayesian asymptotics provide efficient strategies for initializing SMCS.

5.2. Susceptible-Infected-Recovered model

Another setting where sequential inference is particularly relevant is the modeling of disease outbreaks. Parameter calibration involves blending prior knowledge with data arriving regularly, typically daily or weekly. We consider a simple deterministic Susceptible-Infected-Recovered (SIR) model (Bacaër 2012). Inference for such models can be done with MCMC (Grinsztajn et al. 2021). We consider an example from that article, using the classical boarding school data of daily counts of pupils confined to bed during an influenza outbreak, shown in Figure 3(a). The model is described by the differential equations

$$\frac{dS}{dt} = -\beta SI/n, \quad \frac{dI}{dt} = \beta SI/n - \gamma I, \quad \frac{dR}{dt} = \gamma I, \qquad (5.1)$$

where $n = 763$ is the total number of school children, S, I and R represent the number of susceptible, infected and recovered children, respectively, and $\gamma, \beta > 0$ are parameters to be inferred. We assume an initial condition of $(S, I, R) = (n - 1, 1, 0)$ at time $t_0 = 0$, that is, with an infected individual. The observations, which begin at time $t = 1$, are assumed to be noisy measurements of the number I of infected children that day. The observation noise is modeled as a Negative Binomial distribution parameterized by $\phi_{\text{inv}} > 0$. Priors on $\gamma, \beta, \phi_{\text{inv}}$ are taken as $\mathcal{N}(0.4, 0.5^2)$, $\mathcal{N}(2, 1)$ (truncated to $\mathbb{R}_+$) and an exponential distribution with rate 5, respectively. The Stan implementation in Grinsztajn et al. (2021) provides a function that evaluates the posterior log-density, which we use in an adaptive SMCS for the path of partial posteriors.

The bottom row of Figure 3 displays the time evolution of the posterior distribution of parameters. These types of visualization could be used, for example, to study how many observations are necessary to obtain a desired precision on the parameter estimates. Lastly, we consider a simple procedure to infer the initial time t_0 at which the first individual is assumed to be infected. Figure 3(b) plots the marginal likelihood of t_0, which is the normalizing constant of the corresponding posterior distribution, obtained here by running SMCS 10 times independently for different values of t_0; we observe a peak around $t_0 = 0.5$.

6. Discussion

Consider a standard MCMC setting, where we run R chains independently for T steps each, possibly after some early adaptive phase and discarding the first samples as burn-in. From this familiar situation, we might want to: (a) parallelize computation across chains, ideally with a large R and a small T; (b) allow the R chains to communicate in order to accelerate their exploration of the state space; (c) use different Markov kernels depending on the marginal distribution of the chains at the current iteration; (d) estimate the normalizing constant; (e) approximate not a single but multiple, related target distributions. There are various ways of addressing any of these points, but SMCS provide a unified and principled strategy to address them all. We can add to this list: for example there are documented advantages of particle methods over Markov chains for multimodal target distributions (Schweizer 2012b; Paulin, Jasra, and Thiery 2019), and unbiased estimators of normalizing constants lead to useful evidence lower bounds for variational inference (Naesseth et al. 2018). The most important message is perhaps that SMCS provide a viable alternative to MCMC with distinct advantages that can help statisticians.

We take a cautious view regarding performance comparisons between SMCS and MCMC algorithms. There are many tuning choices involved in both families of algorithms, thus, one should not expect to draw fully general conclusions about one algorithm being superior to another. Comparisons can be informative in specific cases (e.g., Matthews et al. 2022). More often comparisons are made between SMCS and annealed importance sampling (e.g., Heng, Doucet, and Pokern 2021) or between variants of SMCS (e.g., Salomone et al. 2018). In principle, any efficient MCMC algorithm could also be used as an ingredient in SMCS, but the choice of paths might not be obvious. For example when the target distribution is supported on a manifold (see e.g., Diaconis, Holmes, and Shahshahani 2013), it might be difficult to define a suitable initial distribution for SMCS, while any point on the manifold provides a valid start for MCMC. Also, as of today the literature on convergence diagnostics is much more developed for MCMC methods (Roy 2020) than for SMCS. Methods discussed in Section 4.1 to construct valid confidence intervals for SMCS estimators are only recent, while the construction of confidence intervals using independent SMCS runs as in Section 4.2 seems to be rarely employed.

Why are sequential Monte Carlo samplers not used more often? The flexibility in the choice of paths and Markov kernels may appear overwhelming to new users. Despite useful efforts to automatize the design of SMCS, for example , using stochastic optimization (Fearnhead and Taylor 2013), or Generative Adversarial Networks (Kempinska and Shawe-Taylor 2017), there remains a number of tuning choices to be addressed in any specific case, which may be a barrier even to computationally-minded statisticians. In addition, software implementations of SMCS exist (Wood, Meent, and Mansinghka 2014; Salvatier, Wiecki, and Fonnesbeck 2016; Murray and Schön 2018) but are not as widely used as MCMC software such as Stan (Carpenter et al. 2017) and do not benefit from a comparable community support. We have described reasons for SMCS to be implemented more often by statisticians in the future.

Supplementary Materials

Supplementary material contains additional information on the link between particle filters and SMCS, unadjusted Hamiltonian Monte Carlo moves, experiments on Normal distributions in increasing dimensions, implementation details for the experiments of Section 5, and additional paths of distributions for the logistic regression example (pdf file).

R-package An R package contains code to reproduce the figures (zip file), also available at *https://github.com/pierrejacob/smcsamplers*.

Tutorials R Markdown files presenting various implementations of SMCS, how to add intermediate distributions and how to remove the bias from adaptive versions of SMCS (zip file).

Acknowledgments

The authors thank Elena Bortolato, Nicolas Chopin, Arnaud Doucet, Sam Power, Rob Salomone, Leah South, Ian and Elizabeth Taylor for useful discussions.

Funding

This work was funded by CY Initiative of Excellence (grant "Investissements d'Avenir" ANR-16-IDEX-0008). Pierre E. Jacob gratefully acknowledges support by the National Science Foundation through grants DMS-1712872 and DMS-1844695.

ORCID

Jeremy Heng https://orcid.org/0000-0003-4959-6856
Pierre E. Jacob https://orcid.org/0000-0002-3126-6966
Nick Whiteley https://orcid.org/0000-0002-8337-626X

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017), "Importance Sampling: Intrinsic Dimension and Computational Cost," *Statistical Science*, 32, 405–431. [2,5,6]

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 72, 269–342. [8]

Arbel, M., Matthews, A., and Doucet, A. (2021), "Annealed Flow Transport Monte Carlo," in *International Conference on Machine Learning*, PMLR, pp. 318–330. [5]

Bacaër, N. (2012), "The Model of Kermack and McKendrick for the Plague Epidemic in Bombay and the Type Reproduction Number with Seasonality," *Journal of Mathematical Biology*, 64, 403–422. [10]

Bernardo, J. M., and Smith, A. F. (2009), *Bayesian Theory* (Vol. 405), Hoboken, NJ: Wiley. [4]

Bernton, E., Heng, J., Doucet, A., and Jacob, P. E. (2019), "Schrödinger Bridge Samplers," arXiv preprint arXiv:1912.13170. [5]

Beskos, A., Crisan, D., and Jasra, A. (2014), "On the Stability of Sequential Monte Carlo Methods in High Dimensions," *The Annals of Applied Probability*, 24, 1396–1445. [7]

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016), "On the Convergence of Adaptive Sequential Monte Carlo Methods," *Annals of Applied Probability*, 26, 1111–1146. [5,6]

Beskos, A., Jasra, A., Law, K., Tempone, R., and Zhou, Y. (2017), "Multilevel Sequential Monte Carlo Samplers," *Stochastic Processes and their Applications*, 127, 1417–1440. [9]

Blackard, J. A. (2000), "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types," PhD thesis, Department of Forest Sciences, Colorado State University. [9]

Brockwell, A., Del Moral, P., and Doucet, A. (2010), "Sequentially Interacting Markov Chain Monte Carlo Methods," *The Annals of Statistics*, 38, 3387–3411. [8]

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011), *Handbook of Markov chain Monte Carlo*, Boca Raton, FL: CRC Press. [8]

Brosse, N., Durmus, A., and Moulines, É. (2018), "Normalizing Constants of Log-Concave Densities," *Electronic Journal of Statistics*, 12, 851–889. [7]

Buchholz, A., Chopin, N., and Jacob, P. E. (2020), "Adaptive Tuning of Hamiltonian Monte Carlo within Sequential Monte Carlo," *Bayesian Analysis* (to appear). [5]

Cappé, O., Godsill, S. J., and Moulines, E. (2007), "An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo," *Proceedings of the IEEE*, 95, 899–924. [2]

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. [11]

Cérou, F., Del Moral, P., Furon, T., and Guyader, A. (2012), "Sequential Monte Carlo for Rare Event Estimation," *Statistics and Computing*, 22, 795–808. [4]

Cérou, F., Del Moral, P., and Guyader, A. (2011), "A Nonasymptotic Theorem for Unnormalized Feynman–Kac Particle Models," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* (Vol. 47), Institut Henri Poincaré, pp. 629–649. [6]

Chan, H. P., and Lai, T. L. (2013), "A General Theory of Particle Filters in Hidden Markov Models and Some Applications," *The Annals of Statistics*, 41, 2877–2904. [8]

Chatterjee, S., and Diaconis, P. (2018), "The Sample Size Required in Importance Sampling," *The Annals of Applied Probability*, 28, 1099–1135. [2,6]

Chen, M.-H. and Shao, Q.-M. (1997), "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants," *The Annals of Statistics*, 25, 1563–1594. [1]

Chopin, N. (2002), "A Sequential Particle Filter Method for Static Models," *Biometrika*, 89, 539–552. [1,2,4,5]

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013), "SMC^2: An Efficient Algorithm for Sequential Analysis of State Space Models," *Journal of the Royal Statistical Society*, Series B, 75, 397–426. [4]

Chopin, N., and Papaspiliopoulos, O. (2020), *An Introduction to Sequential Monte Carlo* (Vol. 4), Cham: Springer. [2]

Cisewski, J., and Hannig, J. (2012), "Generalized Fiducial Inference for Normal Linear Mixed Models," *The Annals of Statistics*, 40, 2102–2127. [9]

Collobert, R., Bengio, S., and Bengio, Y. (2002), "A Parallel Mixture of SVMs for Very Large Scale Problems," in *Advances in Neural Information Processing Systems*, pp. 633–640. [9]

Corenflos, A., Thornton, J., Deligiannidis, G., and Doucet, A. (2021), "Differentiable Particle Filtering via Entropy-Regularized Optimal Transport," in *International Conference on Machine Learning*, PMLR, pp. 2100–2111. [3]

Cornebise, J., Moulines, É., and Olsson, J. (2008), "Adaptive Methods for Sequential Importance Sampling with Application to State Space Models," *Statistics and Computing*, 18, 461–480. [6]

Cuturi, M., Teboul, O., Berthet, Q., Doucet, A., and Vert, J.-P. (2020), "Noisy Adaptive Group Testing Using Bayesian Sequential Experimental Design," arXiv preprint arXiv:2004.12508. [4,9]

Dau, H.-D., and Chopin, N. (2022), "Waste-Free Sequential Monte Carlo," *Journal of the Royal Statistical Society*, Series B, 84, 114–148. [5]

Dawid, A. P., and Musio, M. (2015), "Bayesian Model Selection based on Proper Scoring Rules," *Bayesian Analysis*, 10, 479–499. [4]

Del Moral, P. (2004), *Feynman–Kac Formulae*, New York: Springer. [2,7]

——— (2013), *Mean Field Simulation for Monte Carlo Integration*, Boca Raton, FL: CRC Press. [2,7]

Del Moral, P., Doucet, A., and Jasra, A. (2006), "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 68, 411–436. [1,2,3,5]

——— (2007), "Sequential Monte Carlo for Bayesian Computation," in *Bayesian Statistics* (Vol. 8), eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, pp. 1–34, Oxford: Oxford University Press. [4]

——— (2012), "An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation," *Statistics and Computing*, 22, 1009–1020. [4]

Dellaportas, P., and Kontoyiannis, I. (2012), "Control Variates for Estimation Based on Reversible Markov Chain Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 74, 133–161. [8]

Diaconis, P. (2009), "The Markov Chain Monte Carlo Revolution," *Bulletin of the American Mathematical Society*, 46, 179–205. [1]

Diaconis, P., Holmes, S., and Shahshahani, M. (2013), "Sampling from a Manifold," in *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, Institute of Mathematical Statistics, pp. 102–125. [11]

Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2014), "A Sequential Monte Carlo Algorithm to Incorporate Model Uncertainty in Bayesian Sequential Design," *Journal of Computational and Graphical Statistics*, 23, 3–24. [4,9]

Du, Q., and Guyader, A. (2021), "Variance Estimation in Adaptive Sequential Monte Carlo," *The Annals of Applied Probability*, 31, 1021–1060. [8]

Dunson, D. B., and Johndrow, J. (2020), "The Hastings Algorithm at Fifty," *Biometrika*, 107, 1–23. [1]

Everitt, R. G., Culliford, R., Medina-Aguayo, F., and Wilson, D. J. (2020), "Sequential Monte Carlo with Transformations," *Statistics and Computing*, 30, 663–676. [5]

Fearnhead, P., and Taylor, B. M. (2013), "An Adaptive Sequential Monte Carlo Sampler," *Bayesian Analysis*, 8, 411–438. [5,11]

Finke, A., Doucet, A., and Johansen, A. M. (2020), "Limit Theorems for Sequential MCMC Methods," *Advances in Applied Probability*, 52, 377–403. [8]

Fulop, A., Heng, J., Li, J., and Liu, H. (2021), "Bayesian Estimation of Long-Run Risk Models Using Sequential Monte Carlo," *Journal of Econometrics*, 228, 62–84. [9]

Fulop, A., and Li, J. (2013), "Efficient Learning via Simulation: A Marginalized Resample-Move Approach," *Journal of Econometrics*, 176, 146–161. [4,9]

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185. [1]

Gerber, M., and Chopin, N. (2015), "Sequential Quasi Monte Carlo," *Journal of the Royal Statistical Society*, Series B, 77, 509–579. [3]

Gerber, M., Chopin, N., and Whiteley, N. (2019), "Negative Association, Ordering and Convergence of Resampling Methods," *The Annals of Statistics*, 47, 2236–2260. [3]

Geyer, C. J., (1991), *Markov Chain Monte Carlo Maximum Likelihood*. Interface Foundation of North America. Retrieved from the University of Minnesota Digital Conservancy, *https://hdl.handle.net/11299/58440*. [1]

Gilks, W. R. and Berzuini, C. (2001), "Following a Moving Target—Monte Carlo Inference for Dynamic Bayesian Models," *Journal of the Royal Statistical Society*, Series B, 63, 127–146. [2]

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), "Novel Approach to Nonlinear/non-Gaussian Bayesian State Estimation," in *IEE Proceedings F (Radar and Signal Processing)* (Vol. 140(2)), IET, pp. 107–113. [2,3]

Grenander, U., and Miller, M. I. (1994), "Representations of Knowledge in Complex Systems," *Journal of the Royal Statistical Society*, Series B, 56, 549–581. [5]

Griffin, J. E. (2017), "Sequential Monte Carlo Methods for Mixtures with Normalized Random Measures with Independent Increments Priors," *Statistics and Computing*, 27, 131–145. [9]

Grinsztajn, L., Semenova, E., Margossian, C. C., and Riou, J. (2021), "Bayesian Workflow for Disease Transmission Modeling in Stan," *Statistics in Medicine*, 40, 6209–6234. Available at *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9164* [10]

Heng, J., Bishop, A. N., Deligiannidis, G., and Doucet, A. (2020), "Controlled Sequential Monte Carlo," *The Annals of Statistics*, 48, 2904–2929. [5]

Heng, J., Doucet, A., and Pokern, Y. (2021), "Gibbs Flow for Approximate Transport with Applications to Bayesian Computation," *Journal of the Royal Statistical Society*, Series B, 83, 156–187. [5,11]

Holmes, C. C., and Walker, S. G. (2017), "Assigning a Value to a Power Likelihood in a General Bayesian Model," *Biometrika*, 104, 497–503. [4]

Hooten, M. B., Johnson, D. S., and Brost, B. M. (2019), "Making Recursive Bayesian Inference Accessible," *The American Statistician*, 75, 185–194. [4]

Huggins, J. H., and Roy, D. M. (2019), "Sequential Monte Carlo as Approximate Sampling: Bounds, Adaptive Resampling via ∞-ESS, and an Application to Particle Gibbs," *Bernoulli*, 25, 584–622. [6]

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020), "Unbiased Markov chain Monte Carlo Methods with Couplings," *Journal of the Royal Statistical Society*, Series B, 82, 543–600. [9]

Jarzynski, C., (1997), "Nonequilibrium Equality for Free Energy Differences," *Physical Review Letters*, 78, 2690–2693. [2,5]

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011), "Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo," *Scandinavian Journal of Statistics*, 38, 1–22. [6]

Jun, S.-H., and Bouchard-Côté, A. (2014), "Memory (and Time) Efficient Sequential Monte Carlo," in *International Conference on Machine Learning*, pp. 514–522. [8]

Jun, S.-H., Wang, L., and Bouchard-Côté, A. (2012), "Entangled Monte Carlo," in *Advances in Neural Information Processing Systems*, pp. 2726–2734. [8]

Kantas, N., Beskos, A., and Jasra, A. (2014), "Sequential Monte Carlo Methods for High-Dimensional Inverse Problems: A Case Study for the Navier–Stokes Equations," *SIAM/ASA Journal on Uncertainty Quantification*, 2, 464–489. [6]

Kempinska, K., and Shawe-Taylor, J. (2017), "Adversarial Sequential Monte Carlo," in *Bayesian Deep Learning* (NeurIPS Workshop). [11]

Kirkpatrick, S., Jr., Gelatt, C. D., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680. [4]

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288. [2,6]

Kostov, S. (2006), "Hamiltonian Sequential Monte Carlo and Normalizing Constants," PhD thesis, University of Bristol. [5]

Lee, A., and Whiteley, N. (2018), "Variance Estimation in the Particle Filter," *Biometrika*, 105, 609–625. [3,8]

Lindsten, F., Johansen, A. M., Naesseth, C. A., Kirkpatrick, B., Schön, T. B., Aston, J. A. D., and Bouchard-Côté, A. (2017), "Divide-and-Conquer with Sequential Monte Carlo," *Journal of Computational and Graphical Statistics*, 26, 445–458. [4]

Matthews, A. G., Arbel, M., Rezende, D. J., and Doucet, A. (2022), "Continual Repeated Annealed Flow Transport Monte Carlo," arXiv preprint arXiv:2201.13117. [5,11]

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019), "Unbiased Smoothing Using Particle Independent Metropolis-Hastings," in *Proceedings of Machine Learning Research* (Vol. 89), eds. K. Chaudhuri and M. Sugiyama, PMLR, pp. 2378–2387. [9]

Murray, L., Lee, A., and Jacob, P. (2016), "Parallel Resampling in the Particle Filter," *Journal of Computational and Graphical Statistics*, 25, 789–805. [8]

Murray, L. M., and Schön, T. B. (2018), "Automated Learning with a Probabilistic Programming Language: Birch," *Annual Reviews in Control*, 46, 29–43. [11]

Naesseth, C. A., Lindsten, F., and Schön, T. B. (2014), "Sequential Monte Carlo for Graphical Models," in *Advances in Neural Information Processing Systems*, pp. 1862–1870. [9]

——— (2015), "Nested Sequential Monte Carlo Methods," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning* (Vol. 37), pp. 1292–1301. [6]

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. (2018), "Variational Sequential Monte Carlo," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Vol. 84 of Proceedings of Machine Learning Research, eds. A. Storkey and F. Perez-Cruz, PMLR, pp. 968–977. Available at *https://proceedings.mlr.press/v84/naesseth18a.html* [11]

Neal, R. M. (2001), "Annealed Importance Sampling," *Statistics and Computing*, 11, 125–139. [2,3,5]

Nguyen, T. L. T., Septier, F., Peters, G. W., and Delignon, Y. (2015), "Efficient Sequential Monte Carlo Samplers for Bayesian Inference," *IEEE Transactions on Signal Processing*, 64, 1305–1319. [6]

Nilmeier, J. P., Crooks, G. E., Minh, D. D., and Chodera, J. D. (2011), "Nonequilibrium Candidate Monte Carlo is an Efficient Tool for Equilibrium Simulation," *Proceedings of the National Academy of Sciences*, 108, E1009–E1018. [5]

Nummelin, E. (2002), "MC's for MCMC'ists," *International Statistical Review*, 70, 215–240. [1]

Owen, A. B. (2013), *Monte Carlo theory, Methods and Examples*. Available at: *https://artowen.su.domains/mc/*. [1,8]

Paige, B., Wood, F., Doucet, A., and Teh, Y. W. (2014), "Asynchronous Anytime Sequential Monte Carlo," in *Advances in Neural Information Processing Systems*, pp. 3410–3418. [8]

Paulin, D., Jasra, A., and Thiery, A. (2019), "Error Bounds for Sequential Monte Carlo Samplers for Multimodal Distributions," *Bernoulli*, 25, 310–340. [11]

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models using Pólya–Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349. [4]

Rainforth, T., Naesseth, C., Lindsten, F., Paige, B., Vandemeent, J.-W., Doucet, A., and Wood, F. (2016), "Interacting Particle Markov Chain Monte Carlo„," *International Conference on Machine Learning*, pp. 2616–2625. [8]

Rischard, M., Jacob, P. E., and Pillai, N. (2018), "Unbiased Estimation of Log Normalizing Constants with Applications to Bayesian Cross-validation," arXiv preprint arXiv:1810.01382. [4]

Robert, C., and Casella, G. (2011), "A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data," *Statistical Science*, 26, 102–115. [1]

Roy, V. (2020), "Convergence Diagnostics for Markov Chain Monte Carlo," *Annual Review of Statistics and Its Application*, 7, 387–412. [11]

Salomone, R., South, L. F., Drovandi, C. C., and Kroese, D. P. (2018), "Unbiased and Consistent Nested Sampling via Sequential Monte Carlo," arXiv preprint arXiv:1805.03924. [4,11]

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016), "Probabilistic Programming in Python Using PyMC3," *PeerJ Computer Science*, 2, e55. [11]

Schäfer, C., and Chopin, N., (2013), "Sequential Monte Carlo on Large Binary Sampling Spaces," *Statistics and Computing*, 23, 163–184. [5]

Schweizer, N. (2012a), "Non-asymptotic Error Bounds for Sequential MCMC and Stability of Feynman–Kac Propagators," arXiv preprint arXiv:1204.2382. [7]

——— (2012b), "Non-asymptotic Error Bounds for Sequential MCMC Methods in Multimodal Settings," arXiv preprint arXiv:1205.6733. [11]

Shao, S., Jacob, P. E., Ding, J., and Tarokh, V. (2019), "Bayesian Model Comparison with the Hyvärinen Score: Computation and Consistency," *Journal of the American Statistical Association*, 114, 1–24. [4]

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007), "Sequential Monte Carlo Without Likelihoods," *Proceedings of the National Academy of Sciences*, 104, 1760–1765. [4]

Skilling, J. (2006), "Nested Sampling for General Bayesian Computation," *Bayesian Analysis*, 1, 833–859. [4]

South, L. F., Pettitt, A. N., and Drovandi, C. C. (2019), "Sequential Monte Carlo Samplers with Independent Markov Chain Monte Carlo Proposals," *Bayesian Analysis*, 14, 773–796. [5]

Syed, S., Romaniello, V., Campbell, T., and Bouchard-Côté, A. (2021), "Parallel Tempering on Optimized Paths," in *International Conference on Machine Learning*, PMLR, pp. 10033–10042. [4]

Vaikuntanathan, S., and Jarzynski, C. (2008), "Escorted Free Energy Simulations: Improving Convergence by Reducing Dissipation," *Physical Review Letters*, 100, 190601. [5]

Vempala, S. S., and Wibisono, A. (2019), "Rapid Convergence of the Unadjusted Langevin Algorithm: Log-Sobolev Suffices," in *Advances in Neural Information Processing Systems*, pp. 8092–8104. [7]

Vergé, C., Dubarry, C., Del Moral, P., and Moulines, E. (2015), "On Parallel Implementation of Sequential Monte Carlo Methods: The Island Particle Model," *Statistics and Computing*, 25, 243–260. [8]

Wang, L., Bouchard-Côté, A., and Doucet, A. (2015), "Bayesian Phylogenetic Inference using a Combinatorial Sequential Monte Carlo Method," *Journal of the American Statistical Association*, 110, 1362–1374. [9]

Whiteley, N., Lee, A., and Heine, K. (2016), "On the Role of Interaction in Sequential Monte Carlo Algorithms," *Bernoulli*, 22, 494–529. [3,6,8]

Whitfield, T., Bu, L., and Straub, J. (2002), "Generalized Parallel Sampling," *Physica A: Statistical Mechanics and its Applications*, 305, 157–171. [4]

Wood, F., Meent, J. W., and Mansinghka, V. (2014), "A New Approach to Probabilistic Programming Inference," in *Artificial Intelligence and Statistics*, pp. 1024–1032. [11]

Zhou, Y., Johansen, A. M., and Aston, J. A. (2016), "Toward Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach," *Journal of Computational and Graphical Statistics*, 25, 701–726. [6]