Sequential Monte Carlo Methods

Ajay Jasra

National University of Singapore

KAUST, October 14th 2014

Introduction

A Review of some Monte Carlo methods

- Monte Carlo
- Importance Sampling
- Markov chain Monte Carlo

Sequential Monte Carlo Methods

- Sequential Importance Sampling
- Resampling $+$ Weight Degeneracy
- Path Degeneracy

SMC Samplers

- Algorithm
- Remarks

Particle Markov chain Monte Carlo

Introduction

- These talks will give a basic introduction to sequential Monte Carlo methods.
- These talks will seek to introduce SMC methods for a wide variety of applications.
- It will also provide some details on the theory and implementation of the methodology.

The structure is as follows:

1. Introduction: motivations from Bayesian statistics and standard Monte Carlo.
2. Sequential importance sampling/resampling. This includes the weight and path degeneracy problems.
3. Advanced SMC methods. SMC samplers and particle Markov chain Monte Carlo.
4. Theory and Application. The probabilistic theory and implementation for real problems.

Motivations from Bayesian statistics: Static Inference

- Consider observed data $y_1, \dots, y_n$. Suppose that given a parameter vector $\theta \in \mathbb{R}^d$ the data are i.i.d. with conditional density f

  $$L(y_{1:n}; \theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

  with $y_{1:n} = (y_1, \dots, y_n)$ and $L(\cdot)$ the likelihood function.
- In Bayesian statistical inference, one places a probability density π on θ, prior to seeing the data.

- Then statistical inference is based upon the posterior distribution

$$\pi(\theta|y_{1:n})\frac{L(y_{1:n};\theta)\pi(\theta)}{\int_{\mathbb{R}^d} L(y_{1:n};\theta)\pi(\theta)d\theta} \qquad \theta \in \mathbb{R}^d.$$

- Then, for example, one may interested in the posterior mean of θ, or more generally the posterior expectation function of $\pi-$integrable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$

$$\int_{\mathbb{R}^d} \varphi(\theta)\pi(\theta|y_{1:n})d\theta.$$

- Thus, one cannot typically apply Bayesian statistical models, without having access to a class of numerical methods.

- In most real statistical applications, one can seldom calculate the posterior density, because the marginal likelihood is unknown:

$$\int_{\mathbb{R}^d} L(y_{1:n}; \theta)\pi(\theta)d\theta.$$

- This is because the integral is often in very high dimension. In such scenarios standard deterministic numerical integration is so inaccurate as to be practically useless.

Motivations from Bayesian statistics: Mixture Models

- Consider data $y_1, \dots, y_n$ which are i.i.d. with density

$$f(y_i|\theta) = \sum_{j=1}^{k} \omega_j \phi(y_i; \mu_j, \lambda_j^{-1})$$

$\theta = (\omega_{1:k-1}, \mu_{1:k}, \lambda_{1:k})$, with k known, ϕ the normal density.

- Suppose that

$$\begin{aligned} \mu_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\xi, \kappa) \\ \lambda_j &\overset{\text{i.i.d.}}{\sim} \mathcal{G}a(\alpha, \beta) \\ \omega_{1:k-1} &\sim \mathcal{D}(\delta) \end{aligned}$$

with $\mathcal{N}$ the normal distribution, $\mathcal{G}a$ the Gamma distribution and $\mathcal{D}$ the Dirichlet distribution.

- Thus the posterior distribution is of the form given above.
- As a result, one cannot perform statistical inference without resorting to some numerical method.
- It is remarked that there are direct simulation methods for such models (Fearnhead, P. & Meligkotsidou, 2007; Mukhopadhyay, S. & Bhattacharya, 2011), but unless $k \leq 3$ and n 'not too large' they are too computationally slow to be used.

Motivations from Bayesian statistics: Filtering

- A HMM is a pair of discrete-time processes, $\{X_k\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$.
- The hidden process, $\{X_k\}_{k\geq 0}$, is a Markov chain.
- The observed process $\{Y_k\}_{k\geq 0}$ takes values in $\mathbb{R}^m$.
- Given X_k the Y_k are independent of $Y_0, \dots, Y_{k-1}; X_0, \dots, X_{k-1}$.
- Many real applications: Bioinformatics, econometrics and finance.

Example

- For example, with $k \geq 1$, $X_0 = 0$

  $$\begin{aligned} Y_k &= X_k + \sigma_1 \epsilon_k \\ X_k &= X_{k-1} + \sigma_2 \nu_k \end{aligned}$$

  where ϵ_k, ν_k are i.i.d. standard normals.
- Here, $\theta = (\sigma_1, \sigma_2)$ are assumed known.
- The likelihood is Gaussian and the state-process under-goes a Gaussian Markov transition.

- In the notation to follow:

$$g_\theta(y|x) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_1^2}(y-x)^2\}$$

and

$$q_\theta(x,x') = \frac{1}{\sigma_2\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_2^2}(x'-x)^2\}.$$

- The objective is to compute $\pi_n(x_n|y_{1:n})$ (supressing θ) which can be written:

$$\begin{aligned}\pi_n(x_n|y_{0:n}) &= \int_{\mathbb{R}^{d_x}} \frac{g_\theta(y_n|x_n)q_\theta(x_{n-1},x_n)\pi_{n-1}(x_{n-1}|y_{0:n-1})}{p(y_n|y_{0:n-1})}dx_{n-1}\\ &= \frac{g_n(y_n,x_n)}{p(y_n|y_{1:n-1})}\pi_n(x_n|y_{0:n-1}).\end{aligned}$$

  This is called the filtering distribution.
- That is, given the posterior density at time $n-1$ we perform a prediction step (via $\pi_n(x_n|y_{0:n-1})$) and update to obtain the posterior density at time n.

- Unless the model is linear and Gaussian (as in our example) or the space of the hidden state is finite, one typically cannot compute this distribution (the Kalman filter).
- Moreover, one may also want to calculate the static parameters θ which is a very difficult problem.
- We will discuss methodology which will enable us to achieve the first goal. We will also explain why static parameter estimation is very difficult.

Problem Setting

- Due to the above discussion, our interest (at one level) is in the context where one considers estimation of integrals w.r.t. a probability π

$$\mathbb{E}_\pi[h(X)] = \int h(x)\pi(x)dx.$$

- It is assumed that the density is known point-wise up-to a constant.
- Interested in $-\infty < \mathbb{E}_\pi[h(X)] < \infty$, $X \in \mathbb{R}^d$, for many h.

Monte Carlo

- Often, d is so high that deterministic numerical integration is inaccurate.
- As a result, one often resorts to stochastic numerical methods.
- These are techniques for which the estimate should change (even if the change is small) every time one obtains an answer.

- If it is possible to sample i.i.d. samples $X_1, \ldots, X_N$ from π then one can use

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i)$$

  which converges almost surely via the strong law of large numbers.
- In addition, the method is supposedly dimension independent.
- The problem is that, in a wide variety of problems, one cannot sample from π.

Unbiasedness + Variance

- Before we continue, we remark that:

$$\mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} h(X_i)] = \int h(x)\pi(x)dx.$$

  so the estimate is unbiased.

- In addition, to measure the accuracy of the procedure, one can use the variance:

$$\mathbb{V}\mathrm{ar}[\frac{1}{N}\sum_{i=1}^{N} h(X_i)] = \frac{1}{N}\mathbb{V}\mathrm{ar}_{\pi}[h(X)].$$

  In some scenarios, the variance could be infinite.

$\mathbb{L}_p-$Bound

- It is possible to show (using the Marcinkiewicz-Zygmund inequality e.g. see Cappé et al. (2005) pp. 292) that (for bounded measurable functions, although this can be expanded) for any $p \geq 1$ there exist a $B_p < \infty$

$$\mathbb{E}[|\frac{1}{N}\sum_{i=1}^{N} h(X_i) - \int h(x)\pi(x)dx|^p]^{1/p} \leq \frac{B_p}{\sqrt{N}}.$$

  That is, the rate of convergence is $\mathcal{O}(N^{-1/2})$.
- This bound provides an *finite sample* rate of convergence. In some scenarios ($p = 2!$) the exact moment is known.

Central Limit Theorem

- In addition, one can also establish that for any function that is square integrable w.r.t. π we have (e.g. Shiryaev(1996))

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{h(X_i)-\mathbb{E}_{\pi}[h(X)]\}\right) \Rightarrow \mathcal{N}(0,\mathbb{V}\mathrm{ar}_{\pi}[h(X)]).$$

- The asymptotic variance gives an idea on the accuracy of the approximation, although, here, it is related to the finite sample one. For the methods to be described there will be a clear distinction and the central limit theorem will provide valuable information about accuracy of the approximation adopted.

Importance Sampling

- We consider the scenarios where
  - One cannot sample from π.
  - $\mathbb{V}\mathrm{ar}_{\pi}[h(X)]$ is 'large'.

  In either scenario one cannot or would not want to use Monte Carlo methods.
- We introduce a simple method which can help to alleviate this problem.
- It is based upon a change of measure.

- Let $q(x)$ be any probability density such that if $q(x) > 0$ it implies that $\pi(x) > 0$. Then one has

  $$\mathbb{E}_\pi[h(X)] = \mathbb{E}_q[h(X)w(X)]$$

  where $w(x) = \pi(x)/q(x)$ is the importance weight or Radon-Nikodym derivative.
- Then one can just sample $X_1, \dots, X_N$ i.i.d. from q and use the estimate

  $$\frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i)$$

  which is asymptotically consistent, via the strong law of large numbers.

Unbiasedness + Variance

- As for Monte Carlo, the estimate is unbiased.
- In addition

$$\mathbb{V}\mathrm{ar}[\frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i)] = \frac{1}{N}\mathbb{V}\mathrm{ar}_q[h(X)w(X)].$$

Here it is clear that some q's are better than others. In general, one wants to choose the importance distribution which minimizes the variance.

Optimal Importance Distribution

- It can be shown that the importance distribution that minimizes the variance is (e.g. Rubinstein (1981))

$$q(x) = \frac{|h(x)|\pi(x)}{\int |h(x)|\pi(x)dx}.$$

  Clearly, one cannot evaluate this density, so many importance sampling schemes are based upon approximating this distribution.

- In general, one wants q to follow the shape of $h(x)\pi(x)$. If there are many h one would like the variance of the weights to be, as much as possible, 'low'.

$\mathbb{L}_p-$Bound and Central Limit Theorem

- We can also show, if $w(x)$ is bounded that

$$\mathbb{E}[|\frac{1}{N}\sum_{i=1}^{N} w(X_i)h(X_i) - \int h(x)\pi(x)dx|^p]^{1/p} \leq \frac{B_p}{\sqrt{N}}.$$

- The CLT holds

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{w(X_i)h(X_i) - \mathbb{E}_\pi[h(X)]\}\right) \Rightarrow \mathcal{N}(0, \mathbb{V}\mathrm{ar}_q[h(X)w(X)]).$$

Self-Normalized Estimates

- Finally, we note that if $\pi(x)$ is only known up-to a constant, then we can use the estimate

$$\sum_{i=1}^{N} \overline{w}(X_i)h(X_i)$$

where

$$\overline{w}(X_i) = \frac{w(x)}{\sum_{j=1}^{N} w(X_j)}$$

which does not require the evaluation of any unknown constants of π or q.

- In addition, via the SLLN, this estimate is asymptotically consistent. However, for any finite N it is biased; one of the prices to pay for not evaluating the constants in π and q.

Problems in High Dimensions

- At this stage, it may seem that one can at least solve the problem of (static) Bayesian inference.
- However, in high dimensional situations, e.g. $d \geq 1000$, importance sampling typically collapses as the variance of the importance weights can explode.
- It has been shown (Bickel et al. 2008) that as the dimension goes to infinity, then one needs to increase the number of samples at an exponential rate (for some stability properties) in the dimension.
- This can be too expensive for many practicioners. We will introduce a method, later, whose cost is (at most) cubic in the dimension (Beskos et al. 2014).

MCMC

- A standard technique in statistics/physics is to construct an ergodic Markov chain of stationary distribution π.
- Simulate a Markov chain from an ergodic Markov kernel K and use the estimate:

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i)$$

where $X_1, \dots, X_N$ is the simulated chain.

- Under rather mild conditions, the above quantity also converges almost surely to $\mathbb{E}_\pi[h(X)]$.

Some Markov chain Theory

- We consider an $E-$valued Markov chain ($E = \mathbb{R}^d$) with associated $\sigma-$algebra $\mathcal{B}(E)$.
- Let $K(x, y)$ be a non-negative function on $E \times E$ such that:
  1. For any $x \in E$, $\int_E K(x, y)dy = 1$
  2. For any set A, the function $\int_A K(x, y)dy$ is *measurable*.

  Then K is a **transition kernel**.
- Let $\pi : E \to \mathbb{R}_+$ be such that
  1. $\int_E \pi(x)dx = 1$
  2. $\pi K = \pi$.

  Then π is the stationary (or invariant) density of the chain.

- The idea of a transition kernel, is to replace the concept of the transition matrix. We have

$$\mathbb{P}(X_n \in A | X_{n-1} = x) = \int_A K(x, y) dy.$$

  This is a conditional density which now tells us how the Markov chain evolves over time.

- The $n-$step transition kernel is

$$\int_A K^n(x,y)dy = \int_A \left\{ \int_{E^{n-1}} K(x,x_1)\times\cdots\times K(x_{n-1},y)dx_{1:n-1} \right\} dy$$

  and the Chapman-Kolmogorov:

$$\int_A K^{m+n}(x,y)dy = \int_A \left\{ \int K^m(x,u)K^n(u,y)du \right\} dy.$$

- The marginal density, at time n of the Markov chain is, given an initial density function $\mu^{(0)}$ of X_0:

$$\mu^{(n)}(y) = \int_E \mu^{(n-1)}(x)K(x,y)dx$$

  we use the short-hand $\mu^{(n)} = \mu^{(n-1)}K$.

- Consider a Markov chain $\{X_n\}$ and a probability density λ. The chain is said to be $\lambda-$irreducible if, for any $A \in \mathcal{B}(\mathbb{R}^d)$ with

$$\int_A \lambda(x)dx > 0$$

  and any $x \in E$, there exist a $0 < n_0 < \infty$ such that $\int_A K^{n_0}(x,y)dy > 0$.

- A Markov chain is aperiodic, if there does not exist a disjoint partition $B_1, \dots, B_n$ of $\mathbb{R}^d$, with $n \geq 2$ such that for all $i = 1, \dots, n$, $x \in B_i$

$$\int_{B_{i+1}} K(x, y) dy = 1$$

with the convention $B_{n+1} = B_1$.

- Let λ and η be any two probability densities, then the total variation distance between them is

$$\|\lambda - \eta\|_{TV} := \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| \int_A [\lambda(x) - \eta(x)] dx \right|$$

the distance is based upon the set for which the two probabilities most disagree.

Convergence in Total Variation

- The distance is on $[0, 1]$ (check this) and provides a way to measure the distance between two probability densities. We then have the following result.
- Suppose that K is a $\pi-$irreducible, aperiodic Markov kernel of stationary distribution π. Then for any $\mu^{(0)}$ we have that

$$\lim_{n\to+\infty} \|\mu^{(n)} - \pi\|_{TV} = 0.$$

- Remark that there is a lot of theory on Markov chains (some is below). For a complete introduction see Meyn & Tweedie (2009) and Roberts & Rosenthal (2004).

Strong Law of Large Numbers

- For any Harris positive Markov chain (invariant π) one has

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \rightarrow_{\mathbb{P}} \int_E h(x)\pi(x)dx$$

  i.e. via the SLLN for Markov chains.

- Harris positive, is a Markov chain that is $\pi-$irreducible with invariant π and is Harris recurrent: for every $x \in E$ $\mathbb{P}_x(\eta_A = \infty) = 1$, with η_A the no. of visits to A and A any set with positive $\pi-$measure.

$\mathbb{L}_p-$Bound

- It is also (under some conditions which are not defined here; see Meyn & Tweedie (2009)) possible to establish an $\mathbb{L}_p-$Bound, for example, via the Poisson equation (see Gylnn & Meyn (1996) and below).
- If a $\pi-$invariant Markov chain is geometrically ergodic (possibly sub-geometric) then

$$\mathbb{E}[|\frac{1}{N}\sum_{i=1}^{N} h(X_i) - \int h(x)\pi(x)dx|^p]^{1/p} \leq \frac{B_p}{\sqrt{N}}.$$

Central Limit Theorem

- There is also a CLT, indeed a functional CLT, which makes for a very elegant proof based upon the Poisson equation.
- A function $\widehat{h}$ is said to be a solution to Poisson's equation if

$$h(x) - \pi(h) = \widehat{h}(x) - K(\widehat{h})(x).$$

- We consider a Harris positive Markov chain (invariant π) such that a solution to Poisson exists and $\pi(\widehat{h}^2) < \infty$. Suppose further that

$$\gamma_h^2 = \pi(\widehat{h}^2 - K(\widehat{h})^2) > 0.$$

- Then we have that the CLT holds:

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{h(X_i) - \mathbb{E}_\pi[h(X)]\}\right) \Rightarrow \mathcal{N}(0, \gamma_h^2).$$

- Here, we observe that the asymptotic variance has a very different form to what we have seen thus far. On in-depth analysis, one can find that if the Markov chain mixes quickly, then the variance is 'small'.
- In general, it can be very difficult to compute the variance, even numerically. See e.g. Andrieu & Thoms (2008) and the references therein.

Metropolis-Hastings

- A chain can be set-up, which admits π as its invariant measure, e.g. the Metropolis-Hastings (M-H) kernel (Metropolis et al. 1953; Hastings, 1970).
- This is of the form:

$$\begin{aligned} K(x, dx') &= \left\{1 \wedge \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\right\} q(x,x')dx' \\ &\quad + \delta_x(dx')[1 - \int_{\mathbb{R}^d} \left\{1 \wedge \frac{\pi(u)q(u,x)}{\pi(x)q(x,u)}\right\} q(x,u)du] \end{aligned}$$

- where q is the proposal density.
- Under easily verifiable conditions, the chain is also ergodic.

Optimal Scaling

- In 1997, Gareth Roberts (Roberts et al. 1997) and co-authors proved a stunning result about the Metropolis-Hastings algorithm.
- They proved, when having i.i.d. targets with normal random walk proposals and taking d to infinity that the (say) first component of the Markov chain converges to a Langevin diffusion. This is when the proposal variance is scaled by $1/d$.
- More interestingly, they show that using the limiting diffusion that the optimal, in some sense, scaling induces an acceptance rate of around 0.234; one should aim to scale the proposal so that this rate is achieved.

Metropolis-within-Gibbs

- In general, the MCMC method will not work well, when proposing samples on the whole space (at least if d is big).
- One way to alleviate this problem is to adopt the so-called Metropolis-within-Gibbs (MWG) algorithm. In this scenario one samples from the full conditional densities:

$$\pi(x_i|x_{-i})$$

  with $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ using a Metropolis-Hastings step.
- We remark that the full conditionals can have any dimension less than d.

- A basic MWG algorithm. Given $x_{1:d} \in \mathbb{R}^d$, sample

  $$X_1'|x_{-1}$$

  from any Markov kernel of invariant distribution $\pi(x_1|x_{-1})$ and update the state.
- Every subsequent step for $n = 2, \dots, d$ does the same for the full conditional density $\pi(x_n|x_{-n})$.

- In ‘complex’ problems standard MCMC often does not work well.
- For example, when π is multi-modal, or there are high correlations between sub-blocks of X.
- That is, methods which rely upon mixtures and compositions of M-H kernels (e.g. operating on sub-blocks of π) may converge very slowly.
- In addition, for any real simulation, the samples may fail to traverse the entire support of π.

- This has lead to a vast literature on how to improve over MCMC.
- Methods based upon
  - multiple chains
  - non-linear chains
  - adaptive chains

  have appeared.
- Recently, methods which use MCMC in SMC (and vice-versa) have also appeared and we will, ultimately describe this.

SMC Methods

- SMC methods simulate from a sequence of related distributions of increasing dimension, known point-wise up-to a normalizing constant.
- This technique can be used to sample from a single complex distribution.
- The ideas of importance sampling and resampling are combined.
- The method samples $N > 1$ samples in parallel.

Sequential Importance Sampling

- Denote the sequence of densities $\widetilde{\pi}_1, \dots, \widetilde{\pi}_p$.
- At time step 1, simulate N i.i.d. samples (particles) from a distribution q.
- Calculate the weight

$$W_1^i \propto \frac{\widetilde{\pi}_1(x_1^i)}{q(x_1^i)}.$$

- Can then approximate $\mathbb{E}_{\widetilde{\pi}_1}[h_1(X_1)]$ by

$$\sum_{i=1}^N \left\{ \frac{W_1^i}{\sum_{j=1}^N W_1^j} \right\} h_1(X_1^i).$$

- Use samples at time 1 to help simulate from $\widetilde{\pi_2}$.
- For $i = 1, \dots, N$, to sample $X_2^i | x_1^i \sim K_2(x_1^i, \cdot)$ and calculate

$$W_2^i \propto W_1^i \left[\frac{\widetilde{\pi}_2(x_{1:2}^i)}{\widetilde{\pi}_1(x_1^i) K_2(x_1^i, x_2^i)} \right].$$

- More generally, at time n the weight is

$$W_n^i \propto W_{n-1}^i \left[\frac{\widetilde{\pi}_n(x_{1:n}^i)}{\widetilde{\pi}_{n-1}(x_{1:n-1}^i) K_n(x_{n-1}^i, x_n^i)} \right]$$

and we continue until time p.

$\mathbb{L}_p$—Bound

- We can also show that (under some conditions)

$$\mathbb{E}[|\frac{1}{N}\sum_{i=1}^{N}\overline{W}_n^i h_n(X_n^i). - \int h_n(x_{1:n})\widetilde{\pi}_n(x_{1:n})dx|^p]^{1/p} \leq \frac{B_{p,n}}{\sqrt{N}}.$$

- However, under realistic conditions, $B_{p,n}$ will typically explode with the time parameter. This can mean that the error of the algorithm will increase for any fixed N (indeed this is the case) as n grows.

Resampling + Weight Degeneracy

- In most scenarios, the variance of the weights increase with the time parameter (e.g. Kong et al. (1994)).
- One way to deal with this problem is to resample the particles.
- This uses some stochastic rule to sample the N samples with replacement, according to the weights from the current particle cloud.
- The weights are then reset to one.
- This will mean that SMC is only useful for estimating the marginals of $\widetilde{\pi}_n$ on a fixed dimensional space (we discuss this below).

- Typically, one should not resample at every time step. This can be very inefficient.
- Often use a measure of the quality of the particle approximation.
- Resample when this measure falls below (or moves beyond) some threshold.
- One often used measure is the effective sample size (ESS):

$$\frac{[\sum_{i=1}^{N} W_n^{(i)}]^2}{\sum_{i=1}^{N}(W_n^{(i)})^2}.$$

- This is a number between 1 and N which indicates (roughly) how many useful samples the algorithm has.

Theory of SMC

- The analysis of SMC methods is non-trivial and very different from that of MCMC.
- One can represent the law of the algorithm via a Markov chain, but due to the resampling step, one must think very deeply about how to prove $\mathbb{L}_p-$Bounds and CLTs.
- Most of the theory comes from the pioneering work of Pierre Del Moral; see Del Moral (2004). Many other authors have contributed, to a lesser extent, including: Laurent Miclo, Dan Crisan, Eric Moulines, Francois Le Gland, Randl Douc, Christophe Andrieu, Nicolas Chopin, Arnaud Doucet, Sylvain Rubenthaler, Paul Fearnhead, Andreas Erbele, Hans Kunsch.

- More recently, younger authors: Alex Beskos, Sumeet Singh, Hock Peng Chan, Jimmy Olsson, Omiros Papaspiliopoulos, Adam Johansen (and even myself...) have contributed and it is still possible to do original research in this field.
- Del Moral's work has proved $\mathbb{L}_p-$Bounds and CLTs, large deviation principles, propagation of chaos etc etc for SMC methods.
- Much of the work is thinking about SMC algorithms as approximations of Feynman-Kac formulae.
- These can be thought of as 'importance sampling' identities.
- The results below hold when one resamples at deterministic times, or stochastic times (see Del Moral et al. (2011)).

$\mathbb{L}_p-$Bound

- Denote by $\widetilde{\pi}_n^N$ the approximation of the density $\widetilde{\pi}_n$, $n \geq k$, $h_k \in \mathcal{B}_b(E_{[n-k,n]})$ then one has (under some conditions; see Del Moral (2004))

$$\mathbb{E}\left[\left((\widetilde{\pi}_n^N - \widetilde{\pi}_n)(h_k)\right)^p\right]^{1/p} \leq \frac{B_p}{N}$$

- That is, for a *fixed* computational complexity, in the number of particles, it is *guaranteed* that the marginal can be approximated for a fixed lag; the bound will go to zero, uniformly in time, as $N \to \infty$.
- This is not necessarily the case without resampling.

Central Limit Theorem

- Again, under some conditions, one can prove that for $h_n \in \mathcal{B}_b(E_n)$

$$\sqrt{N}\big(\widetilde{\pi}_n^N - \widetilde{\pi}_n)(h_k)\big) \Rightarrow \mathcal{N}(0, \sigma_n^2).$$

- The asymptotic variance has a very complicated expression and one can see Del Moral (2004) or Chopin (2004).

Resampling

- The basic idea of resampling can be characterized as possible.
- At time n of the algorithm we sample, with replacement, N particles from the current set of particles according to some stochastic rule such that:

$$\mathbb{E}[N_n^i | X_{0:n}^{(i)}] = N w_n^{(i)} \tag{1}$$

where N_n^i is the number of replicates of the i^{th} particle at time n and the argument of w_n is omitted.

Resampling Methods: Multinomial Resampling

- This is the most basic approach, as used in the bootstrap filter of Gordon et al. (1993).
- The procedure is as follows: Resample with replacement N particles, with probabilities proportional to their weights.
- This approach is termed multinomial resampling as, given $X_{1:n}^{(1:N)}$, $(N_n^1, \dots, N_n^N) \sim \mathcal{M}n_N(N, w)$, $w = (\overline{W}_n^{(1)}, \dots, \overline{W}_n^{(N)})$. Here N_n^i are the replicates of the i^{th} particle at time n.

- For $i \in \mathbb{T}_N$ compute the normalized weights:

$$\overline{W}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}.$$

- For $i \in \mathbb{T}_N$ sample $\widehat{X}_{0:n}^{(i)}$ according to the distribution:

$$\widehat{\pi}_n^N(d\widehat{x}_{1:n}) = \sum_{i=1}^N \overline{W}_n^{(i)} \delta_{x_{1:n}^{(i)}}(d\widehat{x}_{1:n}).$$

Resampling Methods: Residual Resampling

- One of the drawbacks of the multinomial scheme is the unnecessarily high variance that it introduces into the algorithm.
- An attempt to deal with this problem is residual resampling (or stochastic remainder resampling) (Baker, 1985), rediscovered by Liu & Chen (1998).

- For $i \in \mathbb{T}_N$ compute the normalized weights.
- For $i \in \mathbb{T}_N$ compute

$$\bar{w}_n^{(i)} = \frac{N\overline{W}_n^{(i)} - \lfloor N\overline{W}_n^{(i)} \rfloor}{N - \sum_{j=1}^N \lfloor N\overline{W}_n^{(i)} \rfloor}.$$

- For $i \in \mathbb{T}_N$ set

$$N_n^i = \lfloor N\bar{w}_n^{(i)} \rfloor + \bar{N}_n^i$$

where $(\bar{N}_n^1, \dots, \bar{N}_n^N) \sim \mathcal{M}n_N(N - \sum_{j=1}^N \lfloor N\bar{w}_n^{(j)} \rfloor, \bar{w})$, $\bar{w} = (\bar{w}_n^{(1)}, \dots, \bar{w}_n^{(N)})$.

Resampling Methods: Stratified Resampling

- This method, introduced by Kitagawa(1996) (see also Fearnhead (1998)) is based upon ideas from survey sampling. To ease the presentation, introduce the following mapping $\mathcal{I}_{\bar{w}_n,N} : [0,1] \to \mathbb{T}_N$ defined as

$$\mathcal{I}_{w_n,N}(u) = i \quad \text{if} \quad u \in \Big(\sum_{j=1}^{i-1} \overline{W}_n^{(i)}, \sum_{j=1}^{i} \overline{W}_n^{(i)}\Big]. \qquad (2)$$

- For $i \in \mathbb{T}_N$ compute the normalized weights.
- For $i \in \mathbb{T}_N$ sample $U_i \sim \mathcal{U}_{[(i-1)/N, i/N]}$.
- For $i \in \mathbb{T}_N$ set $\widehat{X}_{0:n}^{(i)} = x_{0:n}^{\mathcal{I}_{w_n,N}(U_i)}$.

Resampling Methods: Systematic Resampling

- A refinement of the stratified resampling method is systematic resampling (Whitely, 1994), rediscovered by Carpenter et al. (1999). The approach attempts to reduce the randomness in resampling by using only a single uniform random variable
- For $i \in \mathbb{T}_N$ compute the normalized weights.
- Sample $U \sim \mathcal{U}_{[0,1]}$.
- For $i \in \mathbb{T}_N$ set $U_i = (i-1)/N + U$.
- For $i \in \mathbb{T}_N$ set $\widehat{X}_{1:n}^{(i)} = x_{1:n}^{\mathcal{I}_{w_n,N}(U_i)}$.

- In Crisan et al. (1999) several resampling schemes are presented which lead to a random number of particles at each time step. Essentially, they provide some resampling steps that satisfy (1) and the extra condition:

$$\mathbb{E}[|\sum_{i=1}^{N_n} N_n^i f(X_n^{(i)}) - N_n \sum_{i=1}^{N_n} w_n^{(i)} f(X_n^{(i)})|^2 | \mathcal{F}_{n-1}] \leq C N_n \|f\|^2$$

  where the spaces are homogeneous in time $E_n = E_{n-1}$, $f \in \mathcal{C}_b(E)$, N_n is the number of particles at time N and $\mathcal{F}_{n-1}$ is the canonical filtration generated by the process at time $n-1$.

- This procedure, whilst theoretically interesting and advantageous (over multinomial) has some drawbacks. This is because the particle system can both die out, or explode.

- In Fearnhead & Clifford (2003), the following scheme is proposed. Assume that $M \leq N$ particles are to be produced, and given the normalized weights, we may:
    - Calculate the unique solution to $\sum_{i=1}^{N}\{1 \wedge \frac{\bar{w}_n^{(i)}}{\alpha}\} = M$.
    - For $i \in \mathbb{T}_N$ retail $x_{0:n}^{(i)}$ if $\bar{w}_n^{(i)} > \alpha$. Let the number of particles be retained be M'.
    - Adopt any resampling method to resample $M - M'$ particles from the remaining $N - M'$ particles.

  Fearnhead & Clifford (2003) establish that when the systematic resampling method is used in the final step, that this procedure is optimal for the class of problems that they investigate.

- In Chopin (2007) a Rao-Blackwellized scheme is used for the problem of dynamic change-point detection in time-series models. The model is a hidden Markov model with a state

  $$x_n = (\theta_n, i_n) \in E_n = \mathbb{R} \times \{1, i_{n-1} + 1\}.$$

  The Markov prior density is then

  $$p(x_n|x_{n-1}) = \left\{ \begin{array}{lll} (\theta_{n-1}, i_{n-1} + 1) & \mathrm{wp} & \pi_1 \\ (\xi, 1) & \mathrm{wp} & 1 - \pi_1 \end{array} \right.$$

  where π_1 is a known prior probability and ξ is drawn from some known density π_ξ.

- Then, to reduce the variance of the weights, both possibilities are realized; that is, N ξ's are drawn, and two particles are produced with weights corresponding to a kernel with Dirac on the choice $i_n = i_{n-1} + 1$ or $i_n = 1$. Given the $2N$ particles, N new particles are drawn.

Discussion on Resampling

- A clear drawback of the resampling approach is the fact that particles with a very high weight are likely to be resampled many times. As a result, there will be a clear loss in particle diversity. That is, the estimates of functionals of interest, post resampling, may be very poor - this is verified, theoretically in Chopin (2004).
- Indeed, resampling can only be thought of as a technique that will improve estimation in the *future*.

- In terms of computational cost, the standard SIR method, that is with multinomial resampling, is $\mathcal{O}(N \log N)$. However, as noted in Doucet et al. (2000), this operation can be reduced in computational complexity to $\mathcal{O}(N)$.
- Another question, is whether we should resample the particles at all? As noted above, the resampling scheme can introduce extra variance into the estimation procedure and simultaneously, make the algorithm difficult to parallelize, so despite the fact that the variance of the weights increases, should a resampling technique be adopted?

- As noted by Godsill & Clapp (2001) and in implemented in Del Moral et al. (2007), if the discrepancy between consecutive densities is high, we might try to introduce some sort of interpolating distributions between them. This may reduce the need to resample the particles.
- There are, however, at least two counter arguments to not resampling the particles:
  - There is theoretical evidence to the contrary.
  - The introduction of interpolating densities can be very computationally expensive, and not appropriate for real on-line inference problems.

  We have covered the first point. For the second point, it may be necessary to introduce, a large number of intermediate distributions, which may lead to the algorithm being too expensive; see Bickel et al (2008).

- Due to the above points, it is not expected that resampling at every step of the algorithm is required; a sensible criterion is required to decide when to resample. One aspect of IS methods, that we have stressed in general, is the fact that the variance of the importance weights provides a sensible, if not perfect, measure of the performance of the algorithm.
- Typically, the criterion of the effective sample size is used. One approach is to resample the particles when this quantity drops below some pre-specified value, for example $N/2$.
- There is also a rejection method, which at least when the maximum of the weights is known, is preferable to the dynamic approach.

Path Degeneracy

- We discuss a rather important point associated to SMC algorithms. SMC methods, for most applications, should only be used to approximate expectations on the marginal of $\widetilde{\pi}_n$ or, up to some small fixed lag $k \geq 1$, $\widetilde{\pi}(x_{n-k:n})$.
- This fact has an important implication in reference to the static parameter estimation problem.
- Due to the weight degeneracy problem, it is often not possible to accurately approximate the joint distribution $\pi_n(x_{0:n})$ for large n.

- The path degeneracy problem is a product of the weight degeneracy problem.
- Since it is necessary to resample the particles, looking backward in time, many of the particles will be exactly the same. Therefore, the approximation of the joint distribution is in terms of a large number of similar paths; the method cannot be expected to work well.

- This is illustrated in Figure 1. The diagram shows the path of 5 particles for 4 time steps. The size of the circles represent the weight of the particle prior to resampling and the arrows denote the parentage post resampling. It is assumed that resampling occurs at every time-step.

- In Figure 1 the number of unique particles for an SMC algorithm that resamples at every step, falls when looking backwards in time. For extremely efficient algorithms, we can expect that resampling does not occur too regularly; for example 1-2 times in 100 time steps. In this extreme case we have a diverse collection of particles which can approximate the joint quite accurately.

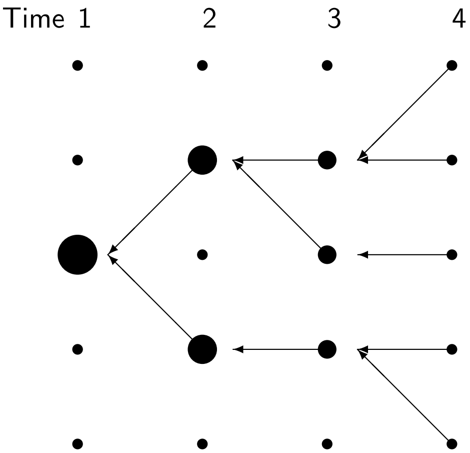

Figure : Path Degeneracy.

- The path degeneracy problem means that estimating static parameters, online, for hidden Markov models and using a Bayesian approach can be very problematic.
- For example, one has

$$\pi(\theta, x_{1:n}|y_{1:n}) \propto \prod_{i=1}^{n} g_\theta(y_i|x_i) q_\theta(x_{i-1}, x_i)\pi(\theta).$$

- Now, it could be that one uses SMC, with an MCMC step after resampling (we discuss this next). However, due to the path degeneracy problem, this technique is *destined to fail*. This is because the posterior distribution depends upon the path of the hidden Markov chain.
- There has been, recently, an elegant solution to this problem (Chopin, et al. 2013), but the computational complexity increases with the time parameter.

SMC Samplers

- We now describe an SMC technique that is designed to simulate from a sequence of probability densities on a *common state-space*.
- This will be (for us), be useful when it is of interest to sample from a single probability π, which is 'complex'.
- The approach here is to introduce a sequence of densities. The sequence starts at a very simple distribution and then moves towards π with related distributions interpolating between π and this initial distribution.
- The method is termed SMC samplers. It has been developed by: Jarzynski (1997); Neal (2001); Gilks & Berzuini (2001); Chopin (2002); Del Moral et al. (2006). It has also been rediscovered by e.g. Botev & Kroese (2008).

- As an example, consider:

$$\pi_n(x) \propto \pi(x)^{\phi_n} \qquad x \in \mathbb{R}^d$$

  with $0 < \phi_1 < \cdots < \phi_p = 1$.
- The idea is to start with a very simple density $\phi_1 \approx 0$ and then move gradually towards π.
- When $\phi_1 \approx 0$ the target density is 'flat' and should be easy to sample from. Then, by appropriately constructing the densities (such that they are not too far apart) it is possible to use the SMC algorithm to interpolate between π_1 and π.
- This idea has been successfully used in many articles.

- Before we discuss the specifics, we just make some remarks.
- The algorithm can be used in many different contexts, such as for rare events estimation (e.g. Cerou et al.(2011)), maximum likelihood estimation (Johansen et al. 2008), as well as approximate Bayesian computation (e.g. Del Moral at el. (2008)).
- There has also been a great deal of interest in the theoretical analysis including: Beskos et al. (2014); Del Moral & Doucet (2003); Erbele & Marinelli (2010); Jasra & Doucet (2008) and Whiteley (2011).

- Recall that SMC methods sample from a sequence of densities of increasing dimension.
- Our sequence of densities are on a common space.
- Consider the following idea. Perform IS w.r.t. π_1 via proposal Υ. Then to move to the next density, use a Markov kernel K_2 say.
- In this scenario, the importance weight is

$$\frac{\pi_2(x_2)}{\int \Upsilon(x_1)K_2(x_1,x_2)dx_1}.$$

- In most scenarios of interest, one cannot compute this importance weight. It is possible to use an $\mathcal{O}(N^2)$ algorithm to remove the integral (even if K_2 is a M-H kernel - see Del Moral et al. (2008)), but this is too costly and in other scenarios (e.g. compositions) one cannot compute the kernel.

- It turns out, that one approach (Jarzynski, 1997; Neal 2001; Del Moral et al. 2006) to circumvent this problem is to introduce a sequence of densities:

$$\widetilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{j=2}^{n} L_j(x_j, x_{j-1})$$

  and use SMC methods on this sequence.
- The $\{L_j\}$ are a sequence of backward Markov kernels and up-to some minimal technical requirements are essentially arbitrary.
- It turns out that one can characterize an optimal (in terms of minimizing the variance of the importance weights) backward kernel; see Del Moral et al. (2006).

- The algorithm is thus nothing more than SIS. The incremental weights are of the form

$$\frac{\pi_n(x_n)L_n(x_n, x_{n-1})}{\pi_{n-1}(x_{n-1})K_n(x_{n-1}, x_n)}.$$

- Throughout this talk we will assume that K_n is an MCMC kernel of invariant distribution π_n
- In this scenario, it turns out that a sensible (and close to optimal in some sense) backward kernel is

$$\frac{\pi_n(x_n)K_n(x_n, x_{n-1})}{\pi_n(x_{n-1})}.$$

- In our example the importance weight can be constructed as

$$W_n^i \propto W_{n-1}^i \frac{\pi_n(x_{n-1}^i)^{\phi_n}}{\pi_{n-1}(x_{n-1}^i)^{\phi_{n-1}}}$$

- This algorithm can work very well in practice; see e.g. (Neal, 2001; Chopin 2002, Del Moral et al. 2006).

Remarks

- It should be first noted that this algorithm can be made *stable* in very high dimensions, when the number of particles is kept *fixed* (see Beskos et al. (2014)).
- In more details, in N are the number of particles and d is the dimension of the random variable, the algorithm will not collapse w.r.t. the ESS, if one has a computational budget of Nd^2.
- From a practical perspective, in our example, it may be difficult to set the $\{\phi_n\}$. However, there is an approach which allows on to do this on the fly (see Jasra et al. (2011) and also Schäfer & Chopin (2011).

- It should also be noted that one can adapt the MCMC kernels on the fly too. In addition, this will not affect the theoretical correctness of the algorithm (contrary to MCMC, where proving ergodicity is rather complex - see Andrieu & Moulines (2006)).

Particle Markov chain Monte Carlo (PMCMC)

- We now present the final algorithm of this talk. For SMC samplers we put MCMC within SMC. PMCMC, uses SMC within MCMC (indeed, although we do not discuss it, there are methods which put SMC within MCMC and this is within SMC - SMC^2 (Chopin et al. 2013)).
- The idea is to use an SMC algorithm as a proposal for an M-H step. This algorithm was developed by Andrieu et al. (2010), although there are some connections to algorithms from molecular simulation e.g. Siepmann & Frenkel (1992).

- We will begin by presenting the simplest generic algorithm found in Andrieu et al. (2010), namely the particle independent Metropolis-algorithm (PIMH). In this case θ and p (the time step of the algorithm) are fixed and PIMH is designed to sample from a pre-specified target distribution $\overline{\pi}_p$ as for SIS. This algorithm proceeds as on the next page.
- For now, suppose the target density is $\overline{\pi}_p(u_{1:p})$, $u_i \in \mathbb{R}$.
- Some notations:

$$\begin{aligned}\psi(\bar{u}_{1:p}, \bar{\mathbf{a}}_{1:p-1}) \quad = \quad & \left(\prod_{i=1}^{N} \overline{M}_1(u_1^{(i)})\right) \prod_{n=2}^{p} \left(\prod_{i=1}^{N} \bar{W}_{n-1}^{(a_{n-1}^i)} \overline{M}_n(u_n^{(i)} \right. \\ & \left. |u_{n-1}^{(a_{n-1}^i)}, \dots, u_1^{(a_1^i)})\right), \qquad (3)\end{aligned}$$

is the SMC algorithm.

- Suppose one resamples, multinomially, at every iteration, except when $n = p$. Denote the resampled index of the ancestor of particle i at time n by $a_n^i \in \mathbb{T}_N$; this is a random variable chosen with probability $\bar{W}_{n-1}^{(a_{n-1}^i)}$.
- An approximation of the normalizing constant of $\overline{\pi}_p$ is

$$\widehat{Z}_p = \prod_{n=1}^{p} \left\{ \frac{1}{N} \sum_{j=1}^{N} W_n^{(j)} \right\}. \tag{4}$$

- 0. Sample $\bar{u}_1, \dots, \bar{u}_p, \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_{p-1}$ from (3). Sample $k \in \mathbb{T}_N$ from $\bar{W}_p^k$ and set this as a new state. Store $\widehat{Z}(0), k(0), \bar{\mathcal{X}}_{1:p}(0), \bar{\mathbf{a}}_{1:p-1}(0)$ (see eq. (4)). Set $i = 1$
- 1. Propose a new $\bar{u}_1', \dots, \bar{u}_p', \bar{\mathbf{a}}_1', \dots, \bar{\mathbf{a}}_{p-1}'$ and k' as in step 0. Accept or reject this as the new state of the chain with probability

  $$1 \wedge \frac{\widehat{Z}'}{\widehat{Z}(i-1)}.$$

  If we accept, store $\left(\widehat{Z}(i), k(i), \bar{\mathcal{X}}_{1:p}(i), \bar{\mathbf{a}}_{1:p-1}(i)\right) = \left(\widehat{Z}', k', \bar{\mathcal{X}}_{1:p}'(i), \bar{\mathbf{a}}_{1:p-1}'\right)$. Set $i = i + 1$.

- In Andrieu et al. (2010) it is shown that the invariant density of the Markov kernel above is exactly

$$\overline{\pi}_p^N(k,\bar{u}_{1:p},\bar{\mathbf{a}}_{1:p-1}) = \frac{\overline{\pi}_p(u_{1:p}^{(k)})}{N^p}\frac{\psi(\bar{u}_{1:p},\bar{\mathbf{a}}_{1:p-1})}{\overline{M}_1(u_1^{(b_1^k)})\prod_{n=2}^{p}\left\{\bar{W}_{n-1}^{(b_{n-1}^k)}\right.}$$

$$\times\frac{1}{\left.\overline{M}_n(u_n^{(b_n^k)}|u_{n-1}^{(b_{n-1}^k)},\ldots,u_1^{(b_1^k)})\right\}}$$

where ψ is as in (3) and as before we have $b_p^k = k$ and $b_n^k = a_n^{b_{n+1}^k}$ for every $k \in \mathbb{T}_N$ and $n \in \mathbb{T}_{p-1}$. The target density of interest, $\overline{\pi}_p$, is the marginal, conditional on k and $\bar{\mathbf{a}}_{1:p-1}$.

- It is also possible to introduce an unknown parameter θ and perform a M-H update; this is called particle marginal M-H.
- The elegant idea of Andrieu and Doucet is to introduce an extended target density which facilitates the use of a particle filter as a proposal and provides an unbiased estimate of the original target density (also related to the exact simulations of diffusion process - see (Beskos et al. 2006)).

- Andrieu, C. & Moulines É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.*, **16**, 1462–1505.
- Andrieu, C. & Thoms J. (2008). A tutorial on adaptive MCMC. *Statist. Comp.*, **18**, 343–373.
- Andrieu, C., Doucet, A. & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. Ser. B*, **72**, 269–342.
- Baker, O. (1985). Adaptive selection methods for genetic algorithms. In *Proc. Intl. Conf. on Genetic Algorithms and their Appl.*, (Ed. J. Greffenstette), 101–111, Erlbaum: Mahwah, NJ.

- Beskos, A., Crisan, D. & Jasra, A. (2014). On the stability of a class of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, **24**, 1396–1445.

- Beskos, A., Papaspiliopoulos, O., Roberts, G. O. & Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc. Ser. B*, **68**, 333–382.

- Bickel, P., Li, B. & Bengtsson, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*, B. Clarke & S. Ghosal, Eds, 318–329, IMS.

- Botev, Z., & Kroese, D. P. (2008). An efficient algorithm for rare event probability estimation, combinatorial optimization and counting. *Methodol. Computat. Appl. Probab.*, **10**, 471–505.

Cappé, O., Ryden, T, & Moulines, É. (2005). *Inference in Hidden Markov Models*. Springer: New York.

Carpenter, S., Clifford, P. & Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEE Proc. Radar Sonar Navigation*, **146**, 2–7.

Cérou, F., Del Moral, P., Furon, T. & Guyader, A. (2011). Sequential Monte Carlo for Rare event estimation. *Statist. Comp.*, (to appear).

Chopin, N. (2002). A sequential particle filter for static models. *Biometrika*, **89**, 539–552.

Chopin, N. (2004). Central limit theorem and its application to Bayesian inference. *Ann Statist.*, **32**, 2385–2411.

Chopin, N. (2007). Dynamic detection of changepoints in long time series. *Ann. Inst. Statist. Math.*, **59**, 349–366.

Chopin, N., Jacob, P. & Papaspiliopoulos, O. (2013). SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *J. R. Statist. Soc. Ser. B.*

Crisan, D., Del Moral P. & Lyons, T. (1999). Discrete filtering using branching and interacting particle systems. *Markov Processes and Relat. Fields*, **5**, 293–318.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

- Del Moral, P. & Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. In *Séminaire de Probabilités*, **37** (eds J. Azema, M. Emery, M. Ledoux and M. Yor), pp. 415446. Berlin: Springer.

- Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.

- Del Moral, P., Doucet, A. & Jasra, A. (2008). A note on the use of Metropolis-Hastings kernels in importance sampling. Unpublished note.

- Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo for Bayesian computation (with discussion). *Bayesian Statistics 8*, Ed. Bayarri, S., Berger, J.

O., Bernardo, J. M., Dawid, A. P., Heckerman, D. Smith, A. F. M. and West, M. 115-149, OUP: Oxford.

- Del Moral, P., Doucet, A. & Jasra, A. (2008). An adaptive sequential Monte Carlo method for approximate Bayesian computation. Technical Report, Imperial College London.
- Del Moral, P., Doucet, A. & Jasra, A. (2011). On adaptive resampling procedures for sequential Monte Carlo methods. *Bernoulli*, (to appear).
- Doucet, A., Godsill, S. & Andrieu, C (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.*, **10**, 197–208.

- Erbele, A. & Marinelli, C. (2010). Quantitative approximations of evolving probability measures. Technical Report, Bonn.

- Fearnhead, P. (1998). *Sequential Monte Carlo Methods in Filter Theory*, D.Phil Thesis, University of Oxford.

- Fearnhead, P. & Clifford, P. (2003). Online inference for well-log data. *J. R. Statist. Soc. Ser B*, **65**, 887-899.

- Fearnhead, P. & Meligkotsidou, L. (2007). Filtering methods for mixture models. *J. Comp. Graph. Statist.*, **16**, 586–607.

- Gilks, W. R. & Berzuini, C. (2001). Following a moving target: Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. Ser. B*, **63**, 127–146.

- Glynn, P. W. & Meyn S. P. (1996). A Lyapunov bound for solutions of the Poisson equation. *Ann. Prob.*, **24**, 916–931.

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

- Godsill, S. J. & Clapp, T. (2001). Improvement strategies for Monte Carlo particle filters. In *Sequential Monte Carlo Methods in Practice* (eds A. Doucet, N. DeFreitsa & N. Gordon). Springer: New York.

- Hastings, W. K. (1970). Monte Carlo sampling using Markov chains and their applications. *Biometrika*, **57**, 97–109.

- Jarzynski, C., 1997. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693.

- JASRA, A. & Doucet, A. (2008). Stability of sequential Monte Carlo samplers via the Foster-Lyapunov condition. *Statist. Probab. Lett.*, **78**, 3062–3069.

- JASRA, A., STEPHENS, D. A., DOUCET, A. & TSAGARIS, T. (2011). Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Statist.*, **38**, 1–22.

- JOHANSEN, A., DOUCET, A. & DAVY, M. (2008). Particle methods for maximum likelihood parameter estimation in latent variable models. *Statist. Comp.*, **18**, 47–57.

- KITAGAWA, G. (1996) Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models, *J. Comp. Graph. Statist.*, **5**, 1–25.

- KONG, A., LIU, J. S. & WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.*, **89**, 278–288.

- LIU, J. S. & CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, **93**, 1032–1044.

- MUKHOPADHYAY, S. & BHATTACHARYA , S. (2011). Perfect simulation for mixtures with known and unknown number of components. Technical Report, Indian Statistical Institute.

- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

- MEYN, S. & TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*. 2nd edition, Cambridge: CUP.
- NEAL, R. M. (2001). Annealed importance sampling. *Statist. Comp.*, **11**, 125–139.
- ROBERTS, G. O., GELMAN, A. & GILKS W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.
- ROBERTS, G.O., & ROSENTHAL, J, S. (2004). General state space Markov chains and MCMC algorithms. *Prob. Surveys*, **1**, 20–71.
- RUBINSTEIN, R. (1981). *Simulation and the Monte Carlo Method*, Wiley: New York.

Schäfer, C. & Chopin, N. (2010). Adaptive Monte Carlo on binary sampling spaces. Technical Report, ENSAE.

Siepmann, J. I. & Frenkel, D. (1992). Configurational-bias Monte Carlo: a new sampling scheme for flexible chains. *Molec. Phys.*, **75**, 59-70.

Shiryaev, A. (1996). *Probability*, Springer: New York.

Whitely, D. (1994). A genetic algorithms tutorial. *Statist. Comp.*, **4**, 65–84.

Whiteley, N. W. (2011).Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions. Technical Report, University of Bristol.