Sequential Monte Carlo Samplers

Arnaud Doucet
Departments of Statistics & Computer Science
University of British Columbia

Objectives

- Let $\{\pi_n\}_{n\geq 1}$ be a *sequence of probability distributions* defined on E such that each $\pi_n(x)$ is known *up to a normalizing constant*, i.e.

$$\pi_n(x) = \underbrace{Z_n^{-1}}_{\text{unknown}} . \underbrace{\gamma_n(x)}_{\text{known}}.$$

- Estimate expectations $\int \varphi(x)\,\pi_n(dx)$ and/or normalizing constants Z_n *sequentially*; i.e. first π_1 and Z_1 then π_2 and Z_2 and so on.
- *Objectives*: Develop efficient Monte Carlo methods to perform numerically these calculations.

Examples

- *Sequential Bayesian Inference:* $\pi_n(x) = p(x|y_{1:n})$.
- *Global optimization:* $\pi_n(x) \propto [\pi(x)]^{\eta_n}$ with $\{\eta_n\}$ increasing sequence such that $\eta_n \longrightarrow \infty$.
- *Sampling from a fixed target* π : $\pi_n(x) \propto [\mu_1(x)]^{\eta_n} [\pi(x)]^{1-\eta_n}$ where μ_1 easy to sample and $\eta_1 = 1$, $\eta_n < \eta_{n-1}$ and $\eta_P = 0$.
- *Rare event simulation* $\pi(A) \ll 1$: $\pi_n(x) \propto \pi(x) 1_{E_n}(x)$ with Z_1 known, $E_1 = E$, $E_n \subset E_{n-1}$ and $E_P = A$ then $Z_P = \pi(A)$.

- Run a Markov chain Monte Carlo (e.g. Metropolis-Hastings) algorithm to sample from each target distribution π_n; i.e. build a Markov kernel $K_n(x, x')$ such that

$$\pi_n(x') = \int_E \pi_n(x) K_n(x, x') dx$$

and simulate a Markov chain $\left\{X_n^{(i)}\right\}$: $X_n^{(1)} \sim \mu_n$ and $X_n^{(i)} \sim K_n\left(X_n^{(i-1)}, \cdot\right)$.

- Under weak assumptions, namely irreducibility & aperiodicity

$$\lim_{i \to \infty} \left\| \mathcal{L}\left(X_n^{(i)}\right) - \pi_n \right\| \to 0,$$

i.e. $X_n^{(i)}$ is asymptotically distributed according to π_n and

$$\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} \varphi\left(X_n^{(i)}\right) = \int \varphi(x) \pi_n(x) dx.$$

Limitations

- Convergence to π_n can be extremely slow and is difficult to diagnose.
- Does not give an estimate of Z_n with 'good' properties.
- If π_{n-1} and π_n are 'close', then it should be possible to devise a cleverer strategy.
- A non-iterative alternative to MCMC is Importance Sampling.

Importance Sampling

- Let the *target distribution* be $\pi_n(x) = Z_n^{-1}\gamma_n(x)$ and μ_n be a so-called *importance distribution* then

$$\begin{aligned}\pi_n(x) &= \frac{w_n(x)\,\mu_n(x)}{\int w_n(x)\,\mu_n(x)\,dx} \text{ where } w_n(x) = \frac{\gamma_n(x)}{\mu_n(x)},\\ Z_n &= \int w_n(x)\,\mu_n(x)\,dx\end{aligned}$$

- By sampling N i.i.d. particles $X_n^{(i)} \sim \mu_n$ then $\widehat{\mu}_n(dx) = \frac{1}{N}\sum_{i=1}^N \delta_{X_n^{(i)}}(dx)$ and

$$\begin{aligned}\widehat{\pi}_n(dx) &= \sum_{i=1}^N W_n^{(i)}\delta_{X_n^{(i)}}(dx) \text{ where } W_n^{(i)} \propto w_n\left(X_n^{(i)}\right),\\ \widehat{Z}_n &= \frac{1}{N}\sum_{i=1}^N w_n\left(X_n^{(i)}\right).\end{aligned}$$

Limitations

- Importance Sampling (IS) is a straightforward method to use if μ_n is easy to sample.
- Under weak assumptions, we can obtain asymptotically consistent estimates of $\int \varphi(x)\,\widehat{\pi}_n(dx)$ and $\widehat{Z}_n$... so why do people use MCMC in 99.99% of cases???
- For the estimates to have reasonable variances, we need to select very carefully the importance distribution.
- To compute $\int \varphi(x)\,\pi_n(dx)$ by IS, the optimal distribution depends on φ but in statistics we often simply want μ_n as "close" to π_n as possible.
- For problems routinely addressed in statistics, this is very difficult.

Iterative Importance Sampling

- "Philosophy": Start by doing simple things before trying to do complex things; same idea used in simulated annealing, simulated tempering etc.
- Develop a sequential/iterative IS strategy where we start by approximating a simple target distribution π_1. Then targets evolve over time and *we build the importance distribution sequentially*.
- At time n, we use μ_{n-1} to build μ_n.
- This approach makes sense if the sequence $\{\pi_n\}$ is not arbitrary; i.e. π_{n-1} somewhat close to π_n.

Initialization

- At time 1, sample N ($N \gg 1$) particles $X_1^{(i)} \sim \mu_1$ to obtain the following IS estimates

$$\begin{aligned} \widehat{\pi}_1(dx) &= \sum_{i=1}^{N} W_1^{(i)} \delta_{X_1^{(i)}}(dx) \\ \text{where } W_1^{(i)} &\propto w_1\left(X_1^{(i)}\right), \ \sum_{i=1}^{N} W_1^{(i)} = 1, \\ \widehat{Z}_1 &= \frac{1}{N} \sum_{i=1}^{N} w_1\left(X_1^{(i)}\right) \end{aligned}$$

- *Remark*: Estimates have reasonable variance only if discrepancy between π_1 and μ_1 small; hence the need to start with easy to sample or approximate π_1.

Moving Particles Forward

- At time $n-1$, one has N particles $\left\{X_{n-1}^{(i)}, W_{n-1}^{(i)}\right\}$

$$X_{n-1}^{(i)} \sim \mu_{n-1} \text{ and } W_{n-1}^{(i)} \propto \frac{\pi_{n-1}\left(X_{n-1}^{(i)}\right)}{\mu_{n-1}\left(X_{n-1}^{(i)}\right)}.$$

- Move the particles according to transition kernel

$$X_n^{(i)} \sim K_n\left(X_{n-1}^{(i)}, \cdot\right) \Rightarrow \mu_n\left(x'\right) = \int \mu_{n-1}\left(x\right) K_n\left(x, x'\right) dx$$

- Optimal transition kernel $K_n\left(x, x'\right) = \pi_n\left(x'\right)$ cannot be used so we need alternatives.

Kernel Selection

- $K_n(x, x') = K_n(x')$ with

  - simple parametric form (e.g. Gaussian, multinomial etc.);
  - semi-parametric based on $\widehat{\mu}_{n-1}(dx)$, complexity $O\left(N^2\right)$.

- $K_n(x, x')$ MCMC kernel of invariant distribution π_n.

  - burn-in correction by importance sampling.
  - scaling of proposal can depend on $\left\{X_{n-1}^{(i)}\right\}$ (Crisan & D., 2000 Chopin, 2002)

- $K_n(x, x')$ approximation of a Gibbs sampler of invariant distribution π_n.

Iterative Importance Sampling

Initialization; $n = 1$.

For $i = 1, ..., N$, sample $X_1^{(i)} \sim \mu_1(\cdot)$ and set

$$w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{\mu_1\left(X_1^{(i)}\right)}, \; W_1^{(i)} \propto w_1\left(X_1^{(i)}\right).$$

At time n; $n \geq 1$.

For $i = 1, ..., N$, sample $X_n^{(i)} \sim K_n\left(X_{n-1}^{(i)}, \cdot\right)$ and set

$$\begin{aligned} w_n\left(X_n^{(i)}\right) &= \frac{\gamma_n\left(X_n^{(i)}\right)}{\mu_n\left(X_n^{(i)}\right)}, \; W_n^{(i)} \propto w_n\left(X_n^{(i)}\right) \\ \text{where } \mu_n(x_n) &= \int \mu_{n-1}(dx_{n-1}) K_n(x_{n-1}, x_n). \end{aligned}$$

Limitations

- In most cases, we *cannot* compute the marginal importance distribution

$$\begin{aligned} \mu_n(x_n) &= \int \mu_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n) \, dx_{n-1} \\ &= \int \mu_1(x_1) \prod_{k=2}^{n} K_k(x_{k-1}, x_k) \, dx_{1:n-1}. \end{aligned}$$

- Hence we cannot use Importance Sampling.

A Potential Solution?

- Monte Carlo approximation

$$\widetilde{\mu}_n(x_n) = \int \widehat{\mu}_{n-1}(dx_{n-1})\, K_n(x_{n-1}, x_n) = \frac{1}{N}\sum_{i=1}^{N} K_n\left(X_{n-1}^{(i)}, x_n\right).$$

$\rightsquigarrow$ Computationally intensive $O\left(N^2\right)$.
$\rightsquigarrow$ Impossible if $K_n(x, x')$ cannot be evaluated pointwise;
e.g. Metropolis-Hastings kernel where

$$K_n(x, x') = \alpha(x, x')\, q(x, x') + \underbrace{\left(1 - \int \alpha(x, u)\, q(x, u)\, du\right)}_{\text{unknown}} \delta_x(x')$$

Importance Sampling on an Extended Space

- *Problem summary*: It is impossible to compute pointwise $\mu_n(x_n)$ hence $\gamma_n(x_n)/\mu_n(x_n)$ except when $n=1$.
- *Solution*: Perform importance sampling on extended space.
- At time 2,

$$\frac{\pi_2(x_2)}{\mu_2(x_2)} = \frac{\pi_2(x_2)}{\int \mu_1(dx_1) K_2(x_1, x_2)} \text{ cannot be evaluated}$$

but alternative weights can be defined

$$\frac{\text{new joint target distribution}}{\text{joint importance distribution}} = \frac{\pi_2(x_2) L_1(x_2, x_1)}{\mu_1(x_1) K_2(x_1, x_2)}$$

where $L_1(x_2, x_1)$ is an *arbitrary* (backward) Markov kernel.

- "Proof" of validity:

$$\int \pi_2(x_2) L_1(x_2, x_1) dx_1 = \pi_2(x_2) \underbrace{\int L_1(x_2, x_1) dx_1}_{=1! \text{ whatever being } L_1} = \pi_2(x_2)$$

- Similarly at time n,

$$Z_n^{-1} w_n(x_n) = \frac{\pi_n(x_n)}{\mu_n(x_n)} \text{ IMPOSSIBLE so USE } Z_n^{-1} w_n(x_{1:n}) = \frac{\widetilde{\pi}_n(x_{1:n})}{\mu_n(x_{1:n})}$$

where $\{\widetilde{\pi}_n\}$ is defined using an *sequence of arbitrary backwards* Markov kernels $\{L_n\}$

$$\text{Artificial target:} \qquad \widetilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{k=1}^{n-1} L_k(x_{k+1}, x_k),$$

$$\text{Importance distribution:} \qquad \mu_n(x_{1:n}) = \mu_1(x_1) \prod_{k=2}^{n} K_k(x_{k-1}, x_k).$$

- "Proof" of validity

$$\int \widetilde{\pi}_n(x_{1:n})\, dx_{1:n-1} = \pi_n(x_n) \underbrace{\int \prod_{k=1}^{n-1} L_k(x_{k+1}, x_k)\, dx_{1:n-1}}_{=1! \text{ whatever being } \{L_k\}} = \pi_n(x_n).$$

- **No free lunch**: By extending the integration space, the variance of the importance weights can only increase.
- The optimal kernel $\{L_{n-1}\}$ is the one bringing us back to the case where there is no space extension; i.e.

$$L_{n-1}^{\text{opt}}\left(x_n, x_{n-1}\right) = \frac{\mu_{n-1}\left(x_{n-1}\right) K_n\left(x_{n-1}, x_n\right)}{\mu_n\left(x_n\right)}$$

- The result follows straightforwardly from the forward-backward formula for Markov processes

$$\mu_n\left(x_{1:n}\right) = \mu_1\left(x_1\right) \prod_{k=2}^{n} K_k\left(x_{k-1}, x_k\right) = \mu_n\left(x_n\right) \prod_{k=2}^{n} L_{k-1}^{\text{opt}}\left(x_k, x_{k-1}\right)$$

- L_{n-1}^{opt} cannot typically be computed (though there are important exceptions) but can be properly approximated in numerous cases (see later). *Even if an approximation is used, the estimates are still asymptotically consistent.*

- We are back to "standard" SMC methods where one is interested in sampling from a sequence of (artificial) distributions $\{\widetilde{\pi}_n\}$ whose dimension is increasing over time.
- **Key difference**: Given $\{K_n\}$, $\{\widetilde{\pi}_n\}$ has been constructed in a "clever" way such that

$$\int \widetilde{\pi}_n\left(x_{1:n}\right) dx_{1:n-1} = \pi_n\left(x_n\right)$$

  whereas usually the sequence of targets $\{\widetilde{\pi}_n\}$ is fixed and $\{K_n\}$ is designed accordingly.
- Because we cannot use $\left\{L_n^{\text{opt}}\right\}$ at each time step, the variance of the weights typically increases over time and it is necessary to resample.

Sequential Monte Carlo Samplers

Initialization; $n = 1$.

For $i = 1, ..., N$, sample $X_1^{(i)} \sim \mu_1(\cdot)$ and set

$$W_1^{(i)} \propto \frac{\pi_1\left(X_1^{(i)}\right)}{\mu_1\left(X_1^{(i)}\right)}.$$

Resample $\left\{W_1^{(i)}, X_1^{(i)}\right\}$ to obtain N new particles $\left\{N^{-1}, X_1^{(i)}\right\}$.

At time n; $n > 1$.

For $i = 1, ..., N$, sample $X_n^{(i)} \sim K_n\left(X_{n-1}^{(i)}, \cdot\right)$ and set

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{\pi_n\left(X_n^{(i)}\right) L_{n-1}\left(X_n^{(i)}, X_{n-1}^{(i)}\right)}{\pi_{n-1}\left(X_{n-1}^{(i)}\right) K_n\left(X_{n-1}^{(i)}, X_n^{(i)}\right)}.$$

Resample $\left\{W_n^{(i)}, X_n^{(i)}\right\}$ to obtain N new particles $\left\{N^{-1}, X_n^{(i)}\right\}$.

- Monte Carlo approximation

$$\widehat{\pi}_n(x) = \sum_{i=1}^{N} W_n^{(i)} \delta_{X_n^{(i)}}(x).$$

- Ratio of normalizing constants

$$\begin{aligned}
\frac{Z_n}{Z_{n-1}} &= \frac{\int \gamma_n(x_n)\, dx_n}{\int \gamma_{n-1}(x_{n-1})\, dx_{n-1}} \\
&= \int \frac{\gamma_n(x_n) L_{n-1}(x_n, x_{n-1})}{\gamma_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)} \pi_{n-1}(dx_{n-1}) K_n(x_{n-1}, dx_n) \\
\Rightarrow \quad \widehat{\frac{Z_n}{Z_{n-1}}} &= \sum_{i=1}^{N} W_{n-1}^{(i)} \frac{\gamma_n\left(X_n^{(i)}\right) L_{n-1}\left(X_n^{(i)}, X_{n-1}^{(i)}\right)}{\gamma_{n-1}\left(X_{n-1}^{(i)}\right) K_n\left(X_{n-1}^{(i)}, X_n^{(i)}\right)}.
\end{aligned}$$

- Like in MCMC, in practice one typically wants to use a mixture of moves

$$K_n\left(x, x'\right) = \sum_{m=1}^{M} \alpha_{n,m}\left(x\right) K_{n,m}\left(x, x'\right)$$

  where $\alpha_{n,m}\left(x\right) > 0$, $\sum_{m=1}^{M} \alpha_{n,m}\left(x\right) = 1$ and $\left\{K_{n,m}\right\}$ is a collection of transition kernels.
- Importance weight can be computed using standard formula but can be too computationally intensive if M is large.
- L_{n-1}^{opt} can be difficult to approximate if M is large.

- Alternative importance sampling on joint space (e.g. Auxiliary Particle Filters by Pitt & Shephard) by introducing explicitly a discrete latent variable M_n

$$\Pr\left(M_n = m\right) = \alpha_{n,m}\left(x\right)$$

and performing importance sampling on the extended space.

- The resulting incremental importance weight becomes

$$\frac{\pi_n\left(x'\right)\beta_{n-1,m}\left(x'\right)L_{n-1,m}\left(x',x\right)}{\pi_{n-1}\left(x\right)\alpha_{n,m}\left(x\right)K_{n,m}\left(x,x'\right)} \text{ instead of } \frac{\pi_n\left(x'\right)L_{n-1}\left(x',x\right)}{\pi_{n-1}\left(x\right)K_n\left(x,x'\right)}$$

where $L_{n-1}\left(x',x\right)$ is the artificial backward Markov kernel

$$L_{n-1}\left(x',x\right) = \sum_{m=1}^{M}\beta_{n-1,m}\left(x'\right)L_{n-1,m}\left(x',x\right)$$

- Optimal choice for $\left\{\beta_{n-1,m}, L_{n-1,m}\right\}$ follows straightforwardly.

- Convergence results follow from general results on Feynman-Kac formula (see Del Moral, 2004).
- When no resampling is performed, one has

$$\sqrt{N}\left(E_{\widehat{\pi}_n}\left[\varphi\right]-E_{\pi_n}\left[\varphi\right]\right)\Rightarrow\mathcal{N}\left(0,\int\frac{\widetilde{\pi}_n^2(x_{1:n})}{\mu_n(x_{1:n})}\left(\varphi\left(x_n\right)-E_{\pi_n}\left(\varphi\right)\right)^2dx_{1:n}\right)$$

- When multinomial resampling is used at each iteration, one has

$$\sqrt{N}\left(E_{\widehat{\pi}_n}\left[\varphi\right]-E_{\pi_n}\left[\varphi\right]\right)\Rightarrow\mathcal{N}\left(0,\sigma_{SMC,n}^2\left(\varphi\right)\right),$$

$$\begin{aligned}\sigma_{SMC,n}^2\left(\varphi\right)&=\int\frac{\widetilde{\pi}_n^2(x_1)}{\mu_1(x_1)}\left(\int\varphi\left(x_n\right)\widetilde{\pi}_n\left(x_n\middle|x_1\right)dx_n-E_{\pi_n}\left(\varphi\right)\right)^2dx_1\\&+\sum_{k=2}^{n-1}\int\frac{\left(\widetilde{\pi}_n(x_k)L_{k-1}(x_k,x_{k-1})\right)^2}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)}\left(\int\varphi\left(x_n\right)\widetilde{\pi}_n\left(x_n\middle|x_k\right)dx_n-E_{\pi_n}\left(\varphi\right)\right)^2dx_{k-1:k}\\&+\int\frac{\left(\pi_n(x_n)L_{n-1}(x_n,x_{n-1})\right)^2}{\pi_{n-1}(x_{n-1})K_n(x_{n-1},x_n)}\left(\varphi\left(x_n\right)-E_{\pi_n}\left(\varphi\right)\right)^2dx_{n-1:n}.\end{aligned}$$

- Under mixing assumptions, $\sigma_{SMC,n}\left(\varphi\right)$ upper bounded over time.

From MCMC to SMC Samplers

- **First step**: Build a sequence of distributions $\{\pi_n\}$ going from π_1 easy to sample/approximate to $\pi_P = \pi$; e.g. $\pi(x) \propto [\mu_1(x)]^{\eta_n} [\pi(x)]^{1-\eta_n}$ where μ_1 easy to sample and $\eta_1 = 1$, $\eta_n < \eta_{n-1}$ with $\eta_P = 0$.
- **Second step**: Introduce a sequence of transition kernels $\{K_n\}$; e.g. K_n MCMC sampler of invariant distribution π_n.
- **Third step**: Introduce a sequence of backward kernels $\{L_n\}$ equal/approximating L_n^{opt}; e.g.

$$\begin{aligned} L_{n-1}(x_n, x_{n-1}) &= \frac{\pi_n(x_{n-1}) K_n(x_{n-1}, x_n)}{\pi_n(x_n)}, \\ \alpha_n(x_{n-1}, x_n) &= \frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})} \end{aligned}$$

Bayesian Analysis of Finite Mixture Models

- Model

$$Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{4} \omega_k \mathcal{N}\left(\mu_k, \lambda_k\right).$$

- Standard conjugate priors on $\theta = (\omega_{1:4}, \mu_{1:4}, \lambda_{1:4})$, no identifiability constraint

$$\mu_k \sim \mathcal{N}(\xi, \kappa^{-1}), \lambda_k \sim \mathcal{G}a(\nu, \chi),\ \omega_{1:4} \sim \mathcal{D}(\rho).$$

- The posterior is a mixture of $4! = 24$ components

- T =100 data with $M = 4$, with $\mu = (-3, 0, 3, 6)$,
  $\lambda = (0.55, 0.55, 0.55, 0.55)$; components "far" from each other.
- We build the sequence of P distributions

$$\pi_n(\theta) \propto l(y_{1:T}; \theta)^{\phi_n} f(\theta)$$

  where $\phi_1 = 0 < \phi_2 < ... < \phi_P = 1$.
- MCMC sampler to sample from π_n
  - Update $\mu_{1:4}$ via a MH kernel with additive normal random walk.
  - Update $\lambda_{1:4}$ via a MH kernel with multiplicative log-normal random walk.
  - Update $\omega_{1:4}$ via a MH kernel with additive normal random walk on the logit scale.

- K_P admits as invariant distribution $\pi_P = \pi$. Very long runs of MCMC get trapped in one of the 4!=24 modes of the distributions.
- We select simply here for $L_{n-1}(\theta_n, \theta_{n-1})$ the reversal kernel

$$L_{n-1}(\theta_n, \theta_{n-1}) = \frac{\pi_n(\theta_{n-1}) K_n(\theta_{n-1}, \theta_n)}{\pi_n(\theta_n)}.$$

- We ran SMC samplers with MCMC kernels for P =50, 100, 200 and 500 time steps with 1 and 10 MCMC iterations per time step.

Sampler Details	Component			
	1	2	3	4
SMC (100 steps, 1 iteration)	0.68	0.91	2.02	2.14
SMC (100 steps, 10 iterations)	1.34	1.44	1.44	1.54
SMC (200 steps, 1 iteration)	1.11	1.29	1.39	1.98
SMC (200 steps, 10 iterations)	1.34	1.37	1.53	1.53
SMC (500 steps, 1 iteration)	0.98	1.38	1.54	1.87
SMC (500 steps, 10 iterations)	1.40	1.44	1.42	1.50

- With reasonable number of intermediate distributions and $N = 1000$, SMC manage to provide reasonable estimates of conditional expectations
- For a fixed computational complexity, it outperforms very significantly the associated homogeneous MCMC trapped in a mode.
- Local MCMC kernels can be combined efficiently through SMC to explore the space in a simple way.

Discussion

- SMC methods are a flexible alternative to MCMC and can address more general problems.
- They are not a black-box and careful design is required.
- Adaptive strategies can easily be implemented.