Sequentially Interacting Markov Chain Monte Carlo

Arnaud Doucet

University of British Columbia
Dept. of Statistics & Dept. of Computer Science

arnaud@stat.ubc.ca

With Anthony E. Brockwell
Dept. of Statistics, Carnegie Mellon Univ., USA.

1.1– Outline of the talk

- Introduction to MCMC in 3 slides
- Objectives
- Sequentially Interacting MCMC
- Applications
- Extensions

1.2– Introduction

- Let $\widetilde{\pi}(dx) = \widetilde{\pi}(x)\,dx$ be a probability measure on E such that

$$\widetilde{\pi}(x) = \underbrace{\widetilde{Z}^{-1}}_{\text{unknown}} \, . \, \underbrace{\widetilde{\gamma}(x)}_{\text{known}}.$$

- *Objectives*: Estimate $\int_E \varphi(x)\,\widetilde{\pi}(dx)$ and/or $\widetilde{Z} = \int_E \widetilde{\gamma}(x)\,dx$.

- *Application*: Bayesian statistics where the target distribution is a posterior distribution

$$\widetilde{\pi}(x) := p(x|\, y) = \frac{l(x;y)\,p(x)}{\int_E l(x;y)\,p(x)\,dx}.$$

1.3– Introduction to Monte Carlo Integration

- To approximate $\int \varphi(x)\,\widetilde{\pi}(dx)$, the Monte Carlo method consists of sampling $N >> 1$ i.i.d. random variables $X^{(i)} \sim \widetilde{\pi}$ and build the empirical measure

$$\widetilde{\pi}^{N}(dx) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(dx)$$

- We estimate $\int \varphi(x)\,\widetilde{\pi}(dx)$ through $\int \varphi(x)\,\widetilde{\pi}^{N}(dx)$; we have

$$\begin{aligned} \mathbb{E}\left[\int \varphi(x)\,\widetilde{\pi}^{N}(dx)\right] &= \int \varphi(x)\,\widetilde{\pi}(dx), \\ var\left[\int \varphi(x)\,\widetilde{\pi}^{N}(dx)\right] &= \frac{\int \varphi^{2}(x)\,\widetilde{\pi}(dx) - \left(\int \varphi(x)\,\widetilde{\pi}(dx)\right)^{2}}{N} \end{aligned}$$

- *Problem*: How to sample from $\widetilde{\pi}$? Standard methods rely on Markov Chain Monte Carlo (MCMC).

1.4– Introduction to (Metropolis)-Hastings Algorithm

- We build a Markov chain $X^{(i)}$ such that $\left\|\mathcal{L}\left(X^{(i)}\right)-\widetilde{\pi}\right\|_{TV}\rightarrow 0.$

- Select a proposal dist. $q\left(x\right)$ such that $q\left(x\right)>0\Rightarrow\widetilde{\pi}\left(x\right)\geq 0,$ set $X^{(1)}$and run the following algorithm.

<u>At iteration i; $i\geq 2$</u>.

Sample $X^{*}\sim q\left(\cdot\right)$.

With probability

$$\alpha\left(X^{(i-1)},X^{*}\right)=1\wedge\frac{\widetilde{\pi}\left(X^{*}\right)}{\widetilde{\pi}\left(X^{(i-1)}\right)}\frac{q\left(X^{(i-1)}\right)}{q\left(X^{*}\right)}$$

set $X^{(i)}=X^{*}$, otherwise set $X^{(i)}=X^{(i-1)}$.

- Uniform ergodicity if for any $x\in E$ $\widetilde{\pi}\left(x\right)/q\left(x\right)<C$; the closer q is from π, the better it works.

1.5– Objectives

- Let $\{\pi_n\}$ $(n = 1, ..., M)$ be a *sequence of probability distributions*, where π_n is defined on $E_n = E^n$ such that $\pi_n(dx_{1:n}) = \pi_n(x_{1:n})\,dx_{1:n}$.

- Each $\pi_n(x_{1:n})$ is known *up to a normalizing constant*, i.e.

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

- *Objectives*: Estimate $\int \varphi_n(x_{1:n})\,\pi_n(dx_{1:n})$ and/or Z_n.

1.6– Motivating example: Optimal Filtering

- $\{X_n\}_{n\geq 1}$ is an unobserved Markov process

$$X_1 \sim \mu \text{ and } \; X_n|\left(X_{n-1} = x_{n-1}\right) \sim f\left(\cdot|\, x_{n-1}\right).$$

- $\{Y_n\}_{n\geq 1}$ is the observation process

$$Y_n|\left(X_n = x_n\right) \sim g\left(\cdot|\, x_n\right).$$

- *Example*:

$$\begin{aligned} X_n &= \varphi\left(X_{n-1}, V_n\right) \text{ where } V_n \overset{\text{i.i.d.}}{\sim} p_V\left(\cdot\right), \\ Y_n &= \Psi\left(X_n, W_n\right) \text{ where } W_n \overset{\text{i.i.d.}}{\sim} p_W\left(\cdot\right). \end{aligned}$$

- *Applications*: Time series, Econometrics, Tracking, Robotics etc.

1.6– Motivating example: Optimal Filtering

- **Optimal Filtering**: Estimate the sequence of posterior distributions $\{p(x_{1:n}|\, y_{1:n})\}$ where $x_{1:n} = (x_1, x_2, \ldots, x_n)$ and $y_{1:n} = (y_1, y_2, \ldots, y_n)$. We have

$$p(x_{1:n}|\, y_{1:n}) \propto \mu(x_1) \prod_{k=2}^{n} f(x_k|\, x_{k-1}) \prod_{k=1}^{n} g(y_k|\, x_k).$$

- **Marginal Likelihood**: Given a model and M observations, compute the marginal likelihood

$$p(y_{1:M}) = \prod_{k=1}^{M} \int g(y_k|\, x_k)\, f(x_k|\, x_{k-1})\, p(x_{k-1}|\, y_{1:k-1})\, dx_{k-1:k}.$$

- **Goodness of Fit**: Compute the residuals

$$\Pr(Y_n \leq y_n|\, y_{1:n-1}) = \int \Pr(Y_n \leq y_n|\, x_n)\, f(x_n|\, x_{n-1})\, p(x_{n-1}|\, y_{1:n-1})\, dx_{n-1:n}.$$

1.7– Motivating example: Sampling from a fixed target distribution

- We want to sample from $\widetilde{\pi}(x) \propto \gamma(x)$ and $\widetilde{Z} = \int \gamma(x)\, dx$.

- Build a path of M distributions so that $\widetilde{\pi}_1(x)$ is easy to sample and $\widetilde{\pi}_M(x)$ is the target: Similar idea in annealing/tempering; e.g. if $\widetilde{\pi}(x) \propto l(x;y)\, p(x)$ then $\widetilde{\pi}_n(x) \propto l(x;y)^{\eta_n}\, p(x)$

- Finally construct

$$\pi_n(x_{1:n}) = \widetilde{\pi}_n(x_n) \prod_{k=1}^{n-1} b_k(x_{k+1}, x_k)$$

where the backward kernels $\{b_k\}$ are selected as a function of the forward kernels used to move from $\widetilde{\pi}_n(x_n)$ to $\widetilde{\pi}_{n+1}(x_{n+1})$.

1.8– More Examples

Any problem which can be rewritten as a Feynman-Kac formula.

- Maximizing marginal distributions.

- Counting the number of self-avoiding random walks (polymers, proteins).

- Estimating the largest eigenvalue and associated eigenmeasure of positive operators (known as quantum Monte Carlo in physics).

- Computing the optimal controller for some nonlinear diffusions.

- Computing the probability of rare events.

1.9– Standard Approaches

- Use M independent MCMC algorithms to sample from each distribution π_n.

$\Rightarrow$ Very computationally intensive and does not use the fact that π_{n-1} and π_n are usually "close".
$\Rightarrow$ No "natural" estimates of $\{Z_n\}$.

- Trans-dimensional MCMC (Green, Biometrika, 1995) cannot be used as $\{Z_n/Z_1\}$ unknown.

- Importance sampling (Durbin & Koopman, JRSS B, 2000): "*In my opinion their approach is a simple, quick and dirty way of deriving a numerical approximation*" (J.Q. Smith, first discussant); inefficient in high dim.

- Standard methods to solve this problem: Sequential Monte Carlo.

1.10– Sequential Monte Carlo Methods

- SMC methods are a combination of importance sampling/resampling algorithms where a collection of N interacting particles approximate the distribution of interest.

- *Advantages*

  - On-line method, can be used for large datasets.

- *Drawbacks*

  - Estimates cannot be improved iteratively.
  - Can be difficult to code for non-specialists

2.1– Objectives

- Develop an alternative MCMC-like algorithm to sample from $\{\pi_n\}$ and compute $\{Z_n\}$.

- Iterative algorithm, trivial to code when one knows MCMC.

- Computationally much cheaper than running M independent MCMC chains.

- Assume for the time being you are able to sample from $\pi_{n-1}(x_{1:n-1})$. You know that $\pi_n(x_{1:n-1}) \approx \pi_{n-1}(x_{1:n-1})$ so it makes sense to use it as a proposal distribution in a Metropolis-Hastings algorithm.

t iteration i; $i \geq 2$.

Sample $X^*_{1:n-1} \sim \pi_{n-1}(\cdot)$ and $X^*_n \sim q_n\left(X^*_{1:n-1}, \cdot\right)$.

With probability

$$\alpha_n\left(X^{(i-1)}_{1:n}, X^*_{1:n}\right) = 1 \wedge \frac{\pi_n\left(X^*_{1:n}\right)}{\pi_n\left(X^{(i-1)}_{1:n}\right)} \frac{\pi_{n-1}\left(X^{(i-1)}_{1:n-1}\right) q_n\left(X^{(i-1)}_{1:n-1}, X^{(i-1)}_n\right)}{\pi_{n-1}\left(X^*_{1:n-1}\right) q_n\left(X^*_{1:n-1}, X^*_n\right)}$$

set $X^{(i)}_{1:n} = X^*_{1:n}$, otherwise set $X^{(i)}_{1:n} = X^{(i-1)}_{1:n}$.

2.2– Basic Idea

- If

$$\frac{\gamma_n\left(x_{1:n}\right)}{\gamma_{n-1}\left(x_{1:n-1}\right)q_n\left(x_{1:n-1},x_n\right)}<M_n<\infty$$

then we have

$$\left\|\mathcal{L}\left(X_{1:n}^{(i)}\right)-\pi_n\right\|_{TV}\leq C_n\alpha_n^i \text{ where } \alpha_n<1.$$

- The Markov chain is *uniformly ergodic* even if π_n is defined on E_n.

- *Problem*: We cannot sample from π_{n-1} in practice!

2.3– Towards Sequentially Interacting MCMC

• We sample from π_1 using a standard MH algorithm and obtained at iteration i the following approximation

$$\widehat{\pi}_1^{(i)}(dx_1) = \frac{1}{i}\sum_{k=1}^{i} \delta_{X_{1,1}^{(k)}}(dx_1).$$

To sample from $\pi_2(dx_{1:2})$, we propose the following algorithm running in parallel.

t iteration i; $i \geq 2$.

Sample $X_1^{*(i)} \sim \widehat{\pi}_1^{(i)}(\cdot)$ and $X_2^{*(i)} \sim q_2\left(X_1^{*(i)}, \cdot\right)$.
With probability

$$\alpha_2\left(X_{2,1:2}^{(i-1)}, X_{1:2}^{*(i)}\right) = 1 \wedge \frac{\pi_2\left(X_{1:2}^{*(i)}\right)}{\pi_2\left(X_{2,1:2}^{(i-1)}\right)} \frac{\pi_1\left(X_{2,1}^{(i-1)}\right) q_2\left(X_{2,1}^{(i-1)}, X_{2,2}^{(i-1)}\right)}{\pi_1\left(X_1^{*(i)}\right) q_2\left(X_1^{*(i)}, X_2^{*(i)}\right)}$$

set $X_{1:2}^{(i)} = X_{1:2}^{*}$, otherwise set $X_{1:2}^{(i)} = X_{1:2}^{(i-1)}$.

2.3– Towards Sequentially Interacting MCMC

- Assume that at iteration i, you have the approximation generated by another MCMC algorithm

$$\widehat{\pi}_{n-1}^{(i)}\left(dx_{1:n-1}\right)=\frac{1}{i}\sum_{k=1}^{i}\delta_{X_{n-1,1:n-1}^{(k)}}\left(dx_{1:n-1}\right)$$

then we approximate the Metropolis-Hastings algorithm
to sample from π_n by the following algorithm.

At iteration i; $i \geq 2$.

Sample $X_{1:n-1}^{*(i)} \sim \widehat{\pi}_{n-1}^{(i)}\left(\cdot\right)$ and $X_{n}^{*(i)} \sim q_n\left(X_{1:n-1}^{*(i)},\cdot\right)$.
With probability

$$\alpha_n\left(X_{n,1:n}^{(i-1)},X_{1:n}^{*(i)}\right)=1\wedge\frac{\pi_n\left(X_{1:n}^{*(i)}\right)}{\pi_n\left(X_{n,1:n}^{(i-1)}\right)}\frac{\pi_{n-1}\left(X_{n,1:n-1}^{(i-1)}\right)q_n\left(X_{n,1:n-1}^{(i-1)},X_{n,n}^{(i-1)}\right)}{\pi_{n-1}\left(X_{1:n-1}^{*(i)}\right)q_n\left(X_{1:n-1}^{*(i)},X_{n}^{*(i)}\right)}$$

set $X_{n,1:n}^{(i)}=X_{1:n}^{*}$, otherwise set $X_{n,1:n}^{(i)}=X_{n,1:n}^{(i-1)}$.

2.4– Sequentially Interacting MCMC

t iteration $i; i \geq 2$

At time $n = 1$

Use MH step of target $\pi_1(x_1)$ with proposal $q_1(x_1)$

to sample $X_1^{(i)}$ and update your estimate $\widehat{\pi}_1^{(i)}(x_1)$ of $\pi_1(x_1)$.

At time $n = 2, ..., M$

Use MH step of target $\pi_n(x_{1:n})$ with proposal $\widehat{\pi}_{n-1}^{(i)}(x_{1:n-1})\, q_n(x_{1:n-1}, x_n)$

to obtain $X_{n,1:n}^{(i)}$ and update your estimate $\widehat{\pi}_n^{(i)}(x_{1:n})$ of $\pi_n(x_{1:n})$.

- At iteration i, we have the approximation for all $n = 1, ..., M$

$$\pi_n^{(i)}\left(dx_{1:n}\right) = \frac{1}{i}\sum_{k=1}^{i}\delta_{X_{n,1:n}^{(k)}}\left(dx_{1:n}\right).$$

- The ratio of normalizing constants can be approximated through

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \frac{1}{i}\sum_{k=1}^{i}\frac{\gamma_n\left(X_{1:n}^{*(k)}\right)}{\gamma_{n-1}\left(X_{1:n-1}^{*(k)}\right)q_n\left(X_{1:n-1}^{*(k)}, X_n^{*(k)}\right)}.$$

2.6– Extensions

- MH step coupled with Accept-Reject can be used to improve performance (Tierney, 1994).

- An auxiliary variable version (Pitt & Shephard, JASA, 1999) of SIMCMC can be derived but it is too computationally intensive.

- Rao-Blackwellisation versions can easily be derived.

3.1– SIMCMC for Optimal Filtering

- Assume you want to sample from $\{p(x_n | y_{1:n})\}$.

- <u>**At iteration** i; $i \geq 2$</u>

  <u>**At time** $n = 1$</u>. Sample $X_1^* \sim \mu$. With proba. $1 \wedge \frac{g(y_1|X_1^*)}{g\left(y_1|X_1^{(i-1)}\right)}$,

  set $X_1^{(i)} = X_1^*$ otherwise $X_1^{(i)} = X_1^{(i-1)}$.

  <u>**At time** $n = 2, ..., M$</u>. Sample $X_{n-1}^* \sim \widehat{p}^{(i)}(x_{n-1}| y_{1:n-1})$ and

  $X_n^* \sim f\left(\cdot| X_{n-1}^*\right)$. With proba. $1 \wedge \frac{g(y_n|X_n^*)}{g\left(y_n|X_n^{(i-1)}\right)}$,

  set $X_n^{(i)} = X_n^*$ otherwise $X_n^{(i)} = X_n^{(i-1)}$.

3.2– Simulations for Optimal Filter

- Linear Gaussian model

$$\begin{aligned} X_n &= \phi X_{n-1} + \sigma_v V_n, \\ Y_n &= X_n + \sigma_w W_n. \end{aligned}$$

- We use SIMCMC with

Prior proposal: $q\left(x_{1:n-1}, x_n\right) = f\left(\left. x_n \right| x_{n-1}\right).$

Optimal Proposal $q\left(x_{1:n-1}, x_n\right) = \frac{g(\left. y_n \right| x_n) f(\left. x_n \right| x_{n-1})}{\int g(\left. y_n \right| x_n) f(\left. x_n \right| x_{n-1}) dx_n}.$

- We compare SIMCMC to Kalman and SMC.

3.2– Simulations for Optimal Filter

- For $N = 5000$ over 10 runs of $M = 100$ observations, results between SIMCMC, SMC and Kalman are virtually identical
in terms of $E\left[X_n \middle| y_{1:n}\right]$ and $\log p\left(y_{1:P}\right)$.

- For lower values of N, the optimal proposal yields significantly better results when σ_v / σ_w is large.

- SIMCMC and SMC performs similarly although SMC yields better estimates for small N.

3.2– Simulations for Optimal Filter

- Switching state-space model

$$\begin{aligned} X_n &= A(I_n) X_{n-1} + B(I_n) V_n, \\ Y_n &= C(I_n) X_n + D(I_n) W_n \end{aligned}$$

where $\{I_n\}$ is an unobserved binary discrete-time Markov chain.

- Optimal filter is a mixture of 2^n Kalman filters at time n.

- We can use SIMCMC to sample from $p(\left. i_{1:n}, x_{1:n} \right| y_{1:n})$ and $p(\left. i_{1:n} \right| y_{1:n})$

(Rao-blackwellisation through Kalman filter).

3.2– Simulations for Optimal Filter

- We use a prior proposal in both cases.

- For $N = 5000$ over 10 runs of $M = 100$ observations, results between SIMCMC and SMC are virtually identical
in terms of $E\left[\left. X_n \right| y_{1:n}\right]$ and $\log p\left(y_{1:P}\right)$.

- For lower values of N, Rao-Blackwellisation significantly improves results and estimates stabilize around $N = 1000$.

3.2– Simulations for Optimal Filter

- Stochastic volatility

$$
\begin{aligned}
X_n &= \phi X_{n-1} + \sigma_v V_n, \\
Y_n &= \beta \exp\left(X_n/2\right) W_n.
\end{aligned}
$$

- We use both the prior distribution and an approximation of the optimal.
- Once more, SIMCMC and SMC yields similar results.

3.3– Bayesian Analysis of Finite Mixture of Gaussians

- Model

$$Y_i \sim \sum_{k=1}^{L} \pi_k \mathcal{N}\left(\mu_k, \sigma_k^2\right).$$

- Standard conjugate priors on $\theta = \left(\pi_k, \mu_k, \sigma_k^2\right)$, no identifiability constraint, posterior is a mixture of $L!$ components.

- Simulations with $L = 4$, components “far” from each other.

- MCMC algorithm sampling directly from $p\left(\theta | y_{1:T}\right)$ get trapped in one mode.

3.4– Algorithm Settings

- To sample $p(\theta|\, y_{1:T})$, set $\pi_n(\theta) \propto l(y_{1:T};\theta)^{\eta_n}\, p(\theta)$
where $n \in \{1, \ldots, M\}$, $N = 5000$, $\eta_1 = 0$, $\eta_n > \eta_{n-1}$ and $\eta_M = 1$.

- q_n is an MCMC kernel of invariant distribution π_n (Thanks to Ajay Jasra).

- Over 10 runs with $M = 800$, SIMCMC discovers the 4! modes.

- Moreover, $\widehat{E}[\mu_1|\, y_{1:T}] \simeq \widehat{E}[\mu_2|\, y_{1:T}] \simeq \widehat{E}[\mu_3|\, y_{1:T}] \simeq \widehat{E}[\mu_4|\, y_{1:T}]$ as expected.

4.1– Discussion

- SIMCMC samplers are an iterative alternative to SMC.
- Can be used on all problems addressed through SMC.
- All your SMC knowledge can be reused straightaway.
- Nice convergence properties inherited from the “ideal” algorithm.

5.1– Extensions: The Bigger Picture

- *MCMC*: Build a Markov transition $K : E \rightarrow \mathcal{P}(E)$ such that

$$\pi = \pi K$$

and the fixed point is approximated through

$$\mu_{n+1} = \mu_n K \rightarrow \pi$$

- *Nonlinear MCMC*: Build a nonlinear Markov transition $K : \mathcal{P}(E) \times E \rightarrow \mathcal{P}(E)$ (e.g. McKean-Vlasov) such that

$$\pi = \pi K_\pi$$

and the fixed point is approximated through

$$\mu_{n+1} = \mu_n K_{\mu_n} \rightarrow \pi.$$

- Nonlinear MCMC can be implemented through particles or self-interacting Markov chains (Del Moral & Doucet, 2003; Andrieu et al., 2006).

6.1– Our papers on the subject

- C. Andrieu et al., Nonlinear MCMC via Self-Interacting Approximation, TR Dept. Math., Bristol Univ., 2006.

- A.E. Brockwell & A.D., Sequentially Interacting MCMC for Bayesian Computation, TR Dept. Stats, CMU, 2006.

- P. Del Moral & A.D., On a Class of Genealogical and Interacting Metropolis Models, Sém. Proba. XXXVII, *Lecture Notes in Mathematics*, Springer-Verlag Berlin, 2003

- P. Del Moral, A.D. & A. Jasra, Sequential Monte Carlo Samplers, *J. Royal Statist. Soc.* B, vol. 68, no. 3, pp. 411-436, 2006.

- P. Del Moral, A.D. & A. Jasra, Sequential Monte Carlo for Bayesian Computation, in *Bayesian Statistics 8*, OUP, 2006.