Advanced Simulation - Lecture 15

Patrick Rebeschini

March 5th, 2018

Outline

- Particle filters and likelihood estimation.

- Pseudo-marginal MCMC.

- A theoretical framework around particle methods.

Sequential Monte Carlo: algorithm

- *At time* $t = 1$
  - Sample $X_1^i \sim q_1(\cdot)$.
  - Compute the weights

$$w_1^i = \frac{\mu(X_1^i) g(y_1 \mid X_1^i)}{q_1\left(X_1^i\right)}.$$

- *At time* $t \geq 2$
  - Resample $(w_{t-1}^i, X_{1:t-1}^i) \to \left(N^{-1}, \overline{X}_{1:t-1}^i\right)$.
  - Sample $X_t^i \sim q_{t|t-1}(\cdot \mid \bar{X}_{t-1}^i)$, $X_{1:t}^i := \left(\bar{X}_{1:t-1}^i, X_t^i\right)$.
  - Compute the weights

$$w_t^i = \omega_t^i = \frac{f\left(X_t^i \mid X_{t-1}^i\right) g\left(y_t \mid X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}.$$

Likelihood estimation

- At time 1,

$$p^N(y_1) = \frac{1}{N}\sum_{i=1}^{N} w_1^i$$
$$\xrightarrow[N\to\infty]{a.s.} \int \frac{\mu(x_1)g(y_1 \mid x_1)}{q_1(x_1)} q_1(x_1)\, dx_1 = p(y_1).$$

- At time t,

$$p^N(y_t \mid y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^{N} w_t^i$$
$$\xrightarrow[N\to\infty]{a.s.} \int w(x_{t-1}, x_t) q_{t|t-1}(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1:t}$$
$$= p(y_t \mid y_{1:t-1}).$$

where
$w(x_{t-1}, x_t) = (f(x_t \mid x_{t-1}) g(y_t \mid x_t))/(q_{t|t-1}(x_t \mid x_{t-1}))$.

Likelihood estimation

- This leads to the estimator

$$p^N(y_{1:t}) = p^N(y_1) \prod_{s=2}^{t} p^N(y_s \mid y_{1:s-1})$$
$$= \prod_{s=1}^{t} \frac{1}{N} \sum_{i=1}^{N} w_s^i \xrightarrow[N\to\infty]{a.s.} p(y_{1:t}).$$

- Surprisingly (?), this estimator is unbiased:

$$\mathbb{E}\left[p^N(y_{1:t})\right] = p(y_{1:t}),$$

whereas for any $t \geq 2$,

$$\mathbb{E}\left[p^N(y_t \mid y_{1:t-1})\right] \neq p(y_t \mid y_{1:t-1}).$$

- Typical particle estimates have a bias of order $\mathcal{O}(1/N)$; the likelihood estimator $p^N(y_{1:t})$ is an exception.

Sequential Monte Carlo: example

- Model equations:

$$\forall t \geq 1 \quad X_t = \phi X_{t-1} + \sigma_V V_t,$$
$$\forall t \geq 1 \quad Y_t = X_t + \sigma_V W_t,$$

with $X_0 \sim \mathcal{N}\left(0, \sigma_V^2\right)$, $V_t, W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\sigma_V = 1$, $\sigma_W = 1$.

- Synthetic data is generated using $\phi^\star = 0.95$, and we estimate the likelihood for a range of values of ϕ.

Sequential Monte Carlo: example

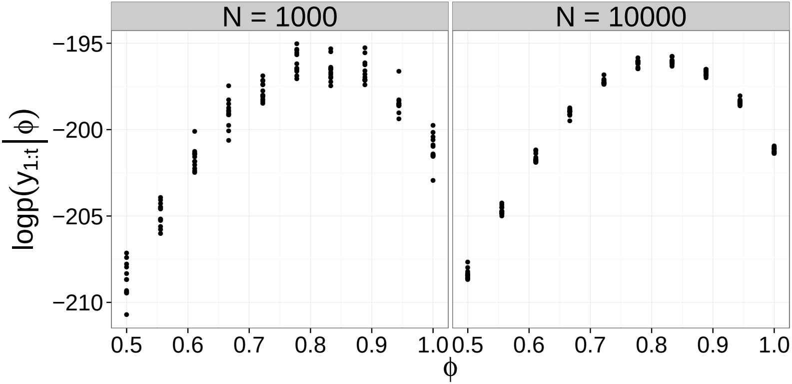

Figure: Log-likelihood estimates $\log p^N(y_{1:t} \mid \phi)$ as a function of ϕ. 12 independent replicates for each value of ϕ.

Likelihood estimation: theory

Consider the estimator of the marginal likelihood

$$p^N(y_{1:t}) = \prod_{s=1}^{t} \frac{1}{N} \sum_{i=1}^{N} w_s^i.$$

- Unbiasedness

$$\mathbb{E}\left[p^N(y_{1:t})\right] = p(y_{1:t}).$$

- Non-asymptotic relative variance

$$\mathbb{E}\left(\left(\frac{p^N\left(y_{1:t}\right)}{p(y_{1:t})} - 1\right)^2\right) \leq \frac{B_3 t}{N}.$$

- Choose $N = \mathcal{O}(t)$ to control the relative variance.

Metropolis–Hastings algorithm

- Target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$.
- Proposal distribution: for any $x, x' \in \mathbb{X}$, we have $q(x'|\, x) \geq 0$ and $\int_{\mathbb{X}} q(x'|\, x)\, dx' = 1$.
- Starting with $X^{(1)}$, for $t = 2, 3, ...$

1. Sample $X^{\star} \sim q\left(\cdot|\, X^{(t-1)}\right)$.
2. Compute

$$\alpha\left(X^{\star}|\, X^{(t-1)}\right) = \min\left(1, \frac{\pi\left(X^{\star}\right) q\left(X^{(t-1)}\middle|\, X^{\star}\right)}{\pi\left(X^{(t-1)}\right) q\left(X^{\star}|\, X^{(t-1)}\right)}\right).$$

3. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X^{\star}|\, X^{(t-1)}\right)$, set $X^{(t)} = X^{\star}$, otherwise set $X^{(t)} = X^{(t-1)}$.

Pseudo-marginal Metropolis–Hastings

- We need to be able to compute point-wise evaluations of $\tilde{\pi}(x) \propto \pi(x)$.

- What if we cannot evaluate these?

- In the setting of hidden Markov models, particle filters provide point-wise unbiased estimates of $\tilde{\pi}(x)$.

- What if we use these estimates instead of $\tilde{\pi}(x)$?

Pseudo-marginal Metropolis–Hastings algorithm

- Starting with $X^{(1)}$, and $Z^{(1)}$ such that $\mathbb{E}(Z^{(1)}) = \tilde{\pi}(X^{(1)})$, for $t = 2, 3, ...$

1. Sample $X^\star \sim q\left(\cdot \mid X^{(t-1)}\right)$.
2. Estimate $\tilde{\pi}(X^\star)$ by $Z^\star$, such that $\mathbb{E}(Z^\star) = \tilde{\pi}(X^\star)$.
3. Compute

$$\alpha\left(X^\star \mid X^{(t-1)}\right) = \min\left(1, \frac{Z^\star\, q\left(X^{(t-1)} \middle| X^\star\right)}{Z^{(t-1)}\, q\left(X^\star \mid X^{(t-1)}\right)}\right).$$

4. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X^\star \mid X^{(t-1)}\right)$, set $(X^{(t)}, Z^{(t)}) = (X^\star, Z^\star)$, otherwise set $(X^{(t)}, Z^{(t)}) = (X^{(t-1)}, Z^{(t-1)})$.

Pseudo-marginal Metropolis–Hastings algorithm

- For any x, denote by Z_x an unbiased estimator of $\tilde{\pi}(x)$, with distribution $g(\cdot \mid x) \equiv g_x$.

- If $\mathbb{V}_{g(\cdot|x)}(Z_x/\tilde{\pi}(x)) << 1$, then the algorithm $\approx$ original Metropolis–Hastings.

- Thus the generated chain $(X^{(t)})_{t\geq 1}$ goes to $\approx \pi$.

- In fact, the limiting law of $(X^{(t)})_{t\geq 0}$ is exactly $\pi\ldots$!

Pseudo-marginal Metropolis–Hastings algorithm

- Introduce an extended target distribution with pdf

$$\bar{\pi}(x, z) \propto z \times g_x(z).$$

- Introduce a proposal kernel $\bar{q}\left((x, z), d(x^\star, z^\star)\right)$ with density

$$\bar{q}\left((x, z), (x^\star, z^\star)\right) = q(x, x^\star) g_{x^\star}(z^\star).$$

- Then the Metropolis–Hastings acceptance ratio would be

$$\begin{aligned} &\min\left(1, \frac{\bar{\pi}(x^\star, z^\star)}{\bar{\pi}(x, z)} \frac{\bar{q}\left((x^\star, z^\star), (x, z)\right)}{\bar{q}\left((x, z), (x^\star, z^\star)\right)}\right) \\ &= \min\left(1, \frac{z^\star}{z} \frac{q(x^\star, x)}{q(x, x^\star)}\right). \end{aligned}$$

This is the algorithm described before.

Pseudo-marginal Metropolis–Hastings algorithm

- The described is a standard Metropolis–Hastings targeting $\bar{\pi}$. What is the distribution of X if (X, Z) follows $\bar{\pi}$?
- By integrating Z out,

$$\begin{aligned}\bar{\pi}_X(x) &\propto \int z g_x(z) dz \\ &= \mathbb{E}_{g_x}[Z_x] \\ &= \tilde{\pi}(x)\end{aligned}$$

  thus the marginal of $\bar{\pi}$ is π.
- Thus if the Markov chain $(X^{(t)}, Z^{(t)})_{t \geq 0}$ converges to $\bar{\pi}$, then the first component $(X^{(t)})$ converges to the first marginal of $\bar{\pi}$, which is π.
- Therefore pseudo-marginal Metropolis–Hastings is *exact*.

Particle Metropolis–Hastings algorithm

- To infer the parameters of a hidden Markov models, one can perform a Metropolis–Hastings algorithm on the parameter space.
- For each proposed parameter $\theta^\star$, run a particle filter to obtain an unbiased estimator $p^N(y_{1:t} \mid \theta^\star)$ of the likelihood $p(y_{1:t} \mid \theta^\star)$.
- Plug these estimators inside the Metropolis–Hastings ratio.
- Produce a chain $(\theta^{(t)})$ targeting the correct posterior distribution.

Numerical experiment

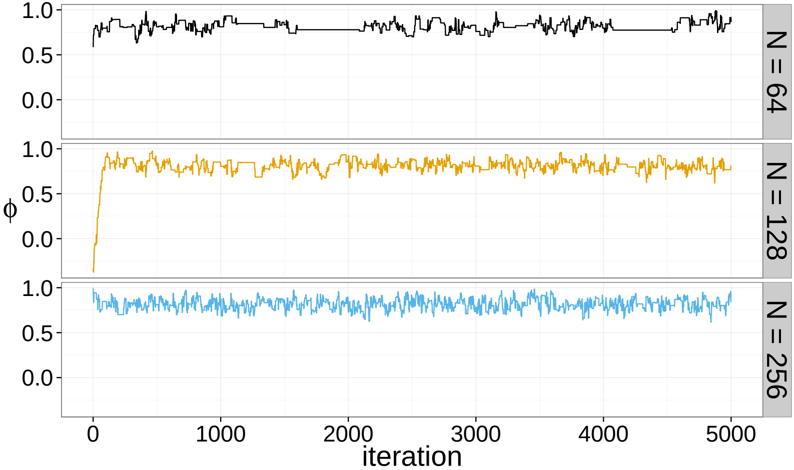

Figure: Trace plot of PMMH chains, for various values of the number of particles N in the particle filter.

Numerical experiment

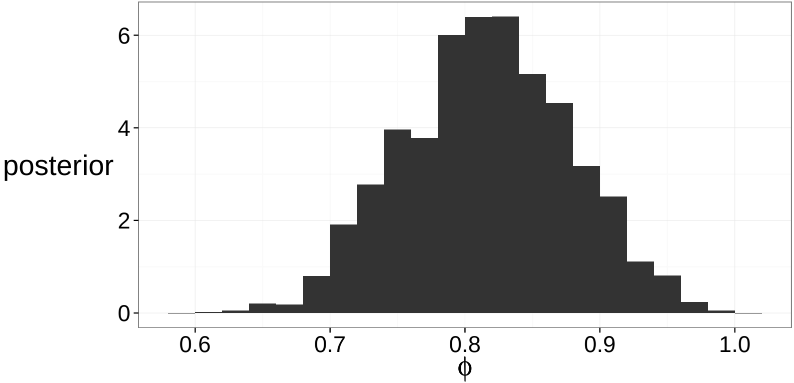

Figure: Histogram of the chain produced with $N = 256$ particles and $T = 5000$ iterations.

Numerical experiment

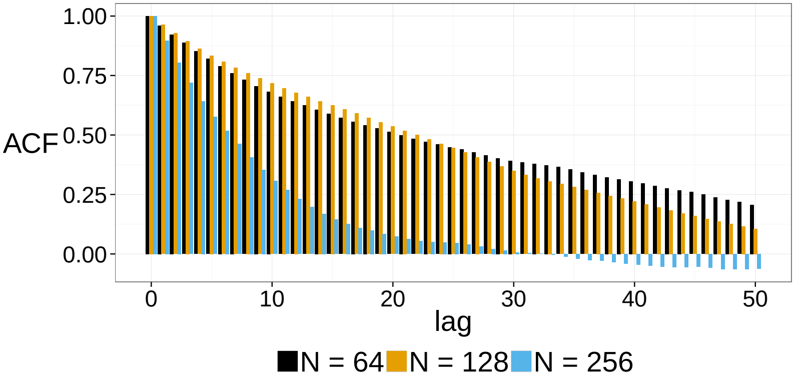

Figure: Autocorrelogram for various values of the number of particles N.

Theoretical framework for particle methods

- A Markov chain (X_n) with initial distribution η_0 and transition kernel M_n at time n.

- A sequence of “potential functions” $G_n : \mathbb{X} \to \mathbb{R}_+$.

- Filtering: $M_n(x, dy) = f(x_n \mid x_{n-1})$, $G_n(x_n) = g(y_n \mid x_n)$.

- Other application: Markov chain in a tube.

Theoretical framework for particle methods

- Sequence of unnormalized measures:

$$\gamma_n(f) = \mathbb{E}\left[f(X_n) \prod_{0 \le k < n} G_k(X_k)\right].$$

- Introduce non-negative kernels:

$$Q_n(x, dy) = G_{n-1}(x) M_n(x, dy)$$

and semi group defined by

$$Q_{p,n} = Q_{p+1} \circ \ldots \circ Q_n$$

such that

$$\gamma_n(f) = \eta_0 Q_{0,n}(f).$$

- Filtering: $\gamma_n(1) = p(y_{0:n-1})$.

Theoretical framework for particle methods

- Normalize γ_n to obtain

$$\eta_n(f) = \gamma_n(f)/\gamma_n(1).$$

- Equivalently

$$\eta_{n+1} = \Phi_n(\eta_n) = \Psi_{G_n}(\eta_n)M_{n+1},$$

where

$$\forall \mu \in \mathcal{P}(E) \quad \Psi_G(\mu)(dx) = \frac{G(x)\mu(dx)}{\int G(x)\mu(dx)} = \frac{G(x)\mu(dx)}{\mu(G)}.$$

- Filtering: η_n corresponds to $p(x_n \mid y_{0:n-1})$.

Theoretical framework for particle methods

- Filtering distributions evolve through:

$$\eta_{n-1} \xrightarrow[reweighting]{} \Psi_{G_{n-1}}(\eta_{n-1}) \xrightarrow[transition]{} \Psi_{G_{n-1}}(\eta_{n-1})M_n.$$

- Particles evolve through the same mechanism:

$$\eta_{n-1}^N \xrightarrow[reweighting]{} \Psi_{G_{n-1}}(\eta_{n-1}^N) \xrightarrow[transition]{} \Psi_{G_{n-1}}(\eta_{n-1}^N)M_n$$

plus a [re]sampling mechanism

$$\Psi_{G_{n-1}}(\eta_{n-1}^N)M_n \xrightarrow[sampling]{} \eta_n^N.$$

- Thus the study of the mechanism itself, i.e.

$$\eta_{n+1} = \Phi_n(\eta_n) = \Psi_{G_n}(\eta_n)M_{n+1},$$

informs about the behaviour of the particles as $n \to \infty$.