Advanced Simulation - Lecture 16

Patrick Rebeschini

March 7th, 2018

Outline

- Particle methods for static problems.
- Evidence estimation.
- The End!

Sequence of posterior distributions

- Consider the problem of estimating

$$\int \varphi(x)\pi(x)dx$$

for test function $\varphi : \mathbb{X} \to \mathbb{R}$ and target distribution π on $\mathbb{X}$.

- Introduce intermediate distributions

$$\pi_0, \pi_1, \ldots, \pi_T$$

such that:

  - π_0 is easy to sample from,
  - π_t and π_{t+1} are not too different,
  - $\pi_T = \pi$.

Sequence of posterior distributions

- Case where $\pi(\theta) \propto p(\theta)p(y_{1:T} \mid \theta)$.
- Introduce the partial posterior $\pi_t(\theta) \propto p(\theta)p(y_{1:t} \mid \theta)$.
- We have $\pi_0(\theta) = p(\theta)$, the prior distribution.
- We have $\pi_T(\theta) = \pi(\theta) = p(\theta \mid y_{1:T})$, the full posterior distribution.
- We can expect $\pi_t(\theta)$ to be similar to $\pi_{t+1}(\theta)$, at least when t is large, when the data is i.i.d.

Sequence of posterior distributions

- General target distribution $\pi(x)$.
- Introduce a simple parametric distribution q.
- We can introduce

$$\pi_t(x) \propto \pi(x)^{\gamma_t} q(x)^{1-\gamma_t}$$

  where $0 = \gamma_0 < \gamma_1 < \ldots < \gamma_T = 1$.

- Then $\pi_0 = q$ and $\pi_T = \pi$.
- If γ_t is close to γ_{t+1}, then π_t is close to π_{t+1}.

Sequential Importance Sampling: algorithm

- *At time* $t = 0$
  - Sample $X^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$.
  - Compute the weights

$$\forall i \in \{1, \ldots, N\} \quad w_0^i = \frac{1}{N}.$$

- *At time* $t \geq 1$
  - Compute the weights

$$\begin{aligned} w_t^i &= w_{t-1}^i \times \omega_t^i \\ &= w_{t-1}^i \times \frac{\pi_t(X^i)}{\pi_{t-1}(X^i)}. \end{aligned}$$

At all times, $(w_t^i, X^i)_{i=1}^N$ approximates π_t.

Sequential Importance Sampling: diagnostics

- As already seen for IS, we can compute the effective sample size

$$\mathrm{ESS}_t = \frac{\left(\sum_{i=1}^N w_t^i\right)^2}{\left(\sum_{i=1}^N (w_t^i)^2\right)} = \frac{1}{\sum_{i=1}^n \left(W_t^i\right)^2}.$$

- $\mathrm{ESS}_t = N$ if $W_t^i = N^{-1}$ for all i.

- If there exists i such that $W_t^i \approx 1$, and for $j \neq i$, $W_t^j \approx 0$, then $\mathrm{ESS}_t \approx 1$.

- As a rule of thumb, the higher the ESS the better our approximation.

Toy model

- Data $y_t \sim \mathcal{N}(\mu, 1)$, 1000 points generated with $\mu^\star = 2$.
- Prior $\mu \sim \mathcal{N}(0, 1)$
- Sequence of partial posterior $\pi_t(\mu) \propto p(\mu) \prod_{s=1}^{t} p(y_s \mid \mu)$.
- Incremental weights:

$$\frac{\pi_t(\mu)}{\pi_{t-1}(\mu)} \propto p(y_t \mid \mu)$$

- We can look at the evolution of the effective sample size with t.

Toy model

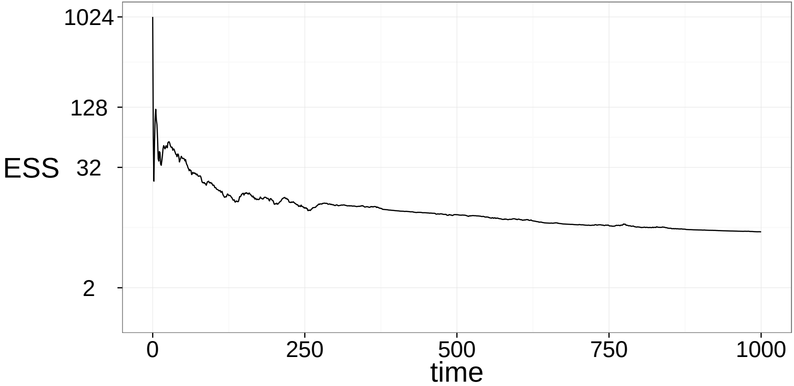

Figure: Effective sample size against "time", using sequential importance sampling.

Sequential Importance Sampling with Resampling

- *At time* $t = 0$
  - Sample $X_0^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$.
  - Compute the weights

$$\forall i \in \{1, \ldots, N\} \quad w_0^i = \frac{1}{N}.$$

- *At time* $t \geq 1$
  - Resample $(w_{t-1}^i, X_{t-1}^i) \rightarrow \left(N^{-1}, \overline{X}_{t-1}^i\right)$.
  - Define $X_t^i = \bar{X}_{t-1}^i$.
  - Compute the weights

$$w_t^i = \omega_t^i = \frac{\pi_t(X_t^i)}{\pi_{t-1}(X_t^i)}.$$

Problem: there are less and less unique values in $(X_t^i)_{i=1}^N$.

Toy model

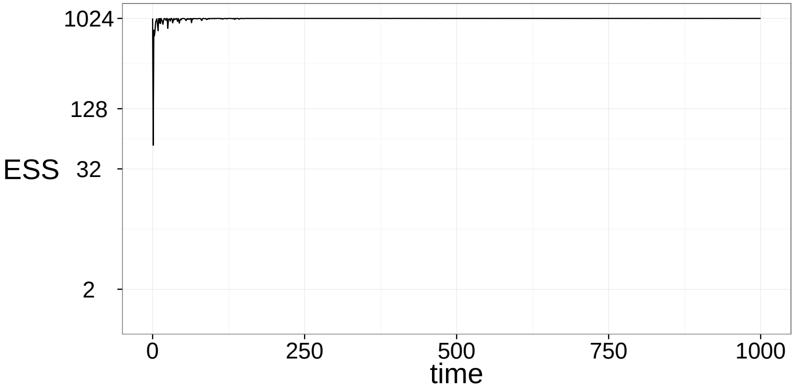

Figure: Effective sample size against “time”, using sequential importance sampling with resampling.

Toy model

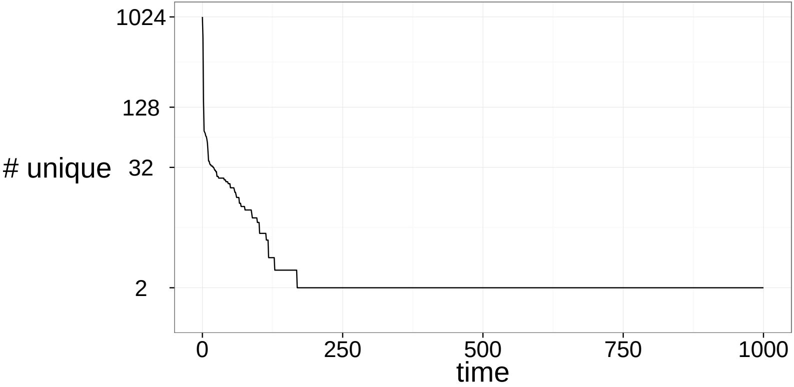

Figure: Number of unique values against "time", using sequential importance sampling with resampling.

Move steps

- Consider particles $(N^{-1}, \bar{X}_t^i)_{i=1}^N$, approximating π_t.
- The ESS is maximum ($=N$), but multiple values within $(\bar{X}_t^i)_{i=1}^N$ are identical.
- How to diversify the particles while still approximating π_t?
- Apply a Markov kernel K_t to each $\bar{X}_t^i$:

$$\forall i \in \{1, \ldots, N\} \quad X_t^i \sim K_t(\bar{X}_t^i, dx)$$

- Can we find K_t such that (N^{-1}, X_t^i) still approximates π_t? e.g.

$$\frac{1}{N} \sum_{i=1}^{N} \varphi(X_t^i) \xrightarrow[N\to\infty]{a.s.} \int \varphi(x)\pi_t(x)dx.$$

Move steps

- Assume that K_t leaves π_t invariant:

$$\int \pi_t(x) K_t(x, y) dx = \pi_t(y).$$

- If $\bar{X}_t^i \sim \pi_t$, then $X_t^i \sim K_t(\bar{X}_t^i, dy)$ also follows π_t.

- Also, if $\bar{X} \sim q$ and $w(x) = \pi(x)/q(x)$, such that

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

  then, for $X \sim K(\bar{X}, dx)$,

$$\mathbb{E}_{qK}[w(\bar{X})\varphi(X)] = \mathbb{E}_\pi[\varphi(X)].$$

Move steps

Result: if $(w^i, \bar{X}^i)$ approximates π, then we can apply a π-invariant kernel to each $\bar{X}^i$ and not change the weights.

- Draw $X^i \sim K(\bar{X}^i, dx)$ for all $i \in \{1, \ldots, N\}$.
- Keep the weights unchanged.
- (w^i, X^i) still approximates π.

Sequential Monte Carlo Sampler: algorithm

- *At time* $t = 0$
  - Sample $X_0^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$.
  - Compute the weights

$$\forall i \in \{1, \ldots, N\} \quad w_0^i = \frac{1}{N}.$$

- *At time* $t \geq 1$
  - Resample $(w_{t-1}^i, X_{t-1}^i) \to \left(N^{-1}, \overline{X}_{t-1}^i\right)$.
  - Draw

$$X_t^i \sim K_{t-1}(\bar{X}_{t-1}^i, dx).$$

  - Compute the weights

$$w_t^i = \omega_t^i = \frac{\pi_t(X_t^i)}{\pi_{t-1}(X_t^i)}.$$

At all times, $(w_t^i, X_t^i)_{i=1}^N$ approximates π_t.

Adaptive resampling

- Problem: if K_t is a Metropolis-Hastings with invariant distribution π_t, computational cost typically linear in t.
- Thus applying K_t at each step $t \in \{1, \ldots, T\}$ yields an overall cost in
$$\sum_{t=1}^{T} \mathcal{O}(t) = \mathcal{O}(T^2)$$
- We can save time by performing the resample-move step only when necessary.
- Use the Effective Sample Size to know whether to trigger a resample-move step.

Sequential Monte Carlo Sampler: algorithm

- *At time* $t = 0$
  - Sample $X_0^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$.
  - Compute the weights

$$\forall i \in \{1, \ldots, N\} \quad w_0^i = \frac{1}{N}$$

- *At time* $t \geq 1$
  - If ESS $< \tilde{N}$, then:
    - Resample $\left(w_{t-1}^i, X_{t-1}^i\right) \to \left(N^{-1}, \overline{X}_{t-1}^i\right)$.
    - Draw

$$X_t^i \sim K_{t-1}(\bar{X}_{t-1}^i, dx).$$

  - Compute the weights

$$w_t^i = w_{t-1}^i \times \frac{\pi_t(X_t^i)}{\pi_{t-1}(X_t^i)}.$$

At all times, $(w_t^i, X_t^i)_{i=1}^N$ approximates π_t.

Toy model

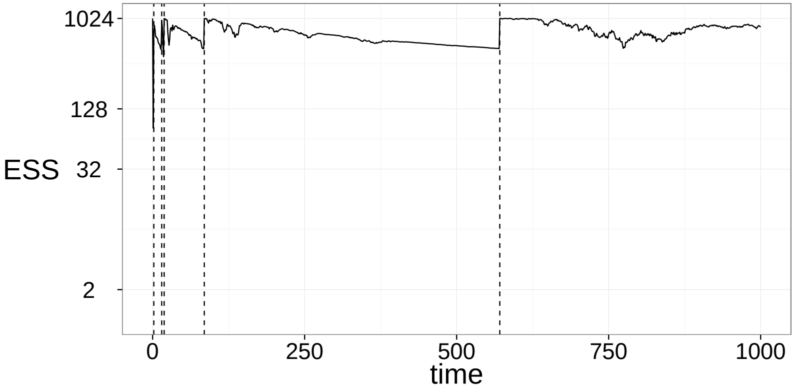

Figure: Effective sample size against "time", using sequential Monte Carlo sampling. Dashed lines indicate resampling times.

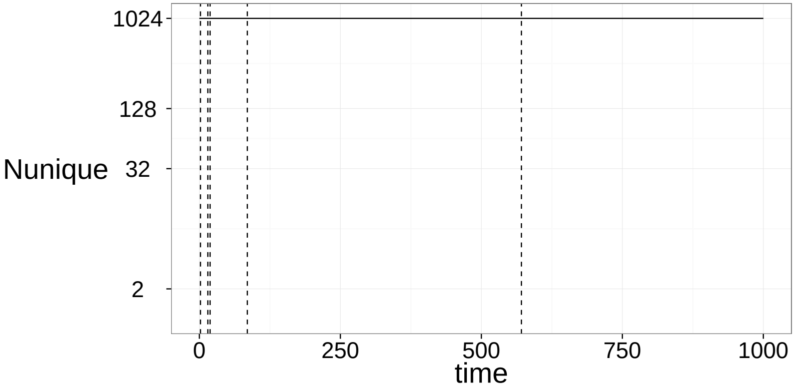

Figure: Number of unique values against “time”, using sequential Monte Carlo sampling. Dashed lines indicate resampling times.

Toy model

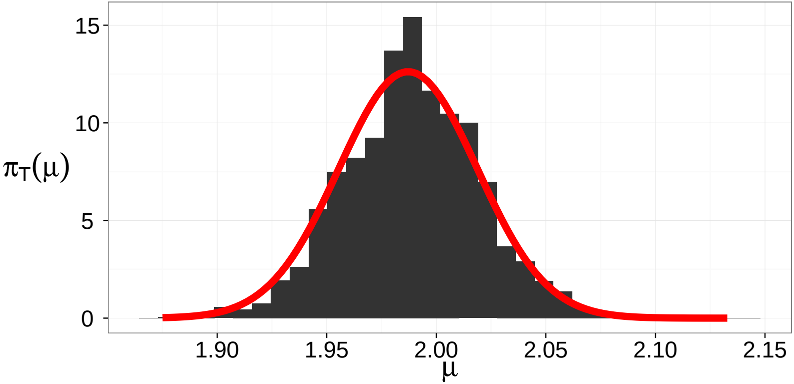

Figure: Posterior approximation (in black), and true posterior (in red), after 1000 observations.

Evidence estimation

- Assume $(w_t^i, X_t^i)_{i=1}^N$ approximates the posterior distribution $\pi_t(\theta) \propto p(\theta)p(y_{1:t} \mid \theta)$ at time t.
- Then the following estimator

$$\frac{\sum_{i=1}^N w_t^i \, p(y_{t+1} \mid X_t^i)}{\sum_{i=1}^N w_t^i}$$

converges to

$$\begin{aligned}
&\int p(y_{t+1} \mid \theta)\pi_t(\theta)d\theta \\
&= \int p(y_{t+1} \mid \theta)\frac{p(\theta)p(y_{1:t} \mid \theta)}{p(y_{1:t})}d\theta \\
&= \frac{1}{p(y_{1:t})}\int p(y_{1:t+1} \mid \theta)p(\theta)d\theta = p(y_{t+1} \mid y_{1:t}).
\end{aligned}$$

- Similar to the likelihood estimator in particle filters.

Toy model

- We compare the SMC estimator with the estimator obtained by importance sampling from the prior distribution:

$$\forall i \in \{1, \dots, N\} \quad X^i \sim p(\theta),$$
$$\forall i \in \{1, \dots, N\} \quad L_t^i = p(y_{1:t} \mid X^i),$$
$$p^N(y_{1:t}) = \frac{1}{N} \sum_{i=1}^{N} L_t^i.$$

- We plot the following relative error

$$r_t^N = \frac{|p^N(y_{1:t}) - p(y_{1:t})|}{|p(y_{1:t})|}.$$

Toy model

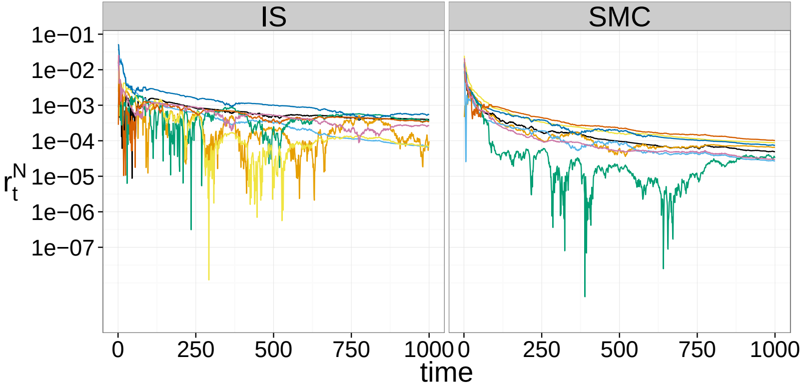

Figure: Relative error using SMC samplers and importance sampling from the prior; 10 independent experiments.

The End

Inversion, Transformation, Composition, Accept-reject, Importance Sampling, Metropolis–Hastings, Gibbs sampling, Adaptive Multiple Importance Sampling, Reversible Jump, Slice Sampling, Sequential Importance Sampling, Particle filter, Pseudo-marginal Metropolis–Hastings, Sequential Monte Carlo Sampler...

What about *Nested Sampling, Path Sampling, Hamiltonian MCMC, Pinball Sampler, Sequential Quasi-Monte Carlo, Multiple-Try Metropolis–Hastings, Coupling From the Past, Multilevel Splitting, Wang–Landau algorithm, Free Energy MCMC, Stochastic Gradient Langevin Dynamics, Firefly MCMC, Configurational Bias Monte Carlo...?*

Good luck for the exams!