Advanced Simulation - Lecture 14

Patrick Rebeschini

February 28th, 2018

Outline

- Sequential Monte Carlo.
- Path degeneracy.
- Likelihood estimation.
- Selected theoretical results.

Hidden Markov Models

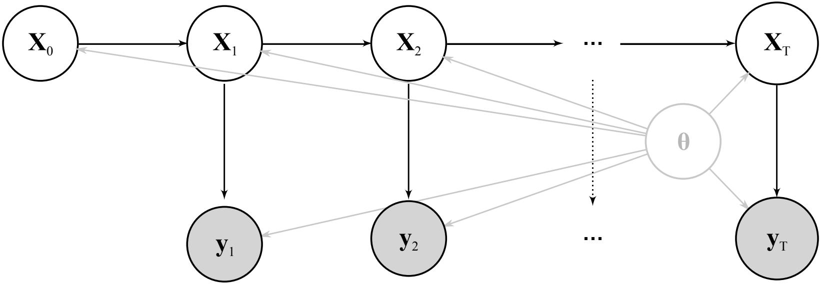

Figure: Graph representation of a general HMM.

(X_t): initial distribution μ_θ, transition f_θ.
(Y_t) given (X_t): measurement g_θ.
Prior on the parameter $\theta \in \Theta$.

Inference in HMMs, Cappé, Moulines, Ryden, 2015.

Sequential Importance Sampling: 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* $t \geq 2$
  - Sample $X_t^i \sim q_{t|t-1}(\cdot | \, X_{t-1}^i)$, $X_{1:t}^i := \left(X_{1:t-1}^i, X_t^i\right)$.
  - Compute the weights

$$\begin{aligned} w_t^i &= w_{t-1}^i \times \omega_t^i \\ &= w_{t-1}^i \times \frac{f\left(X_t^i \middle| \, X_{t-1}^i\right) g\left(y_t \middle| \, X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}. \end{aligned}$$

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 $X_t^i \sim q_{t|t-1}(\cdot | \, \bar{X}_{t-1}^i)$, $X_{1:t}^i := \left(\bar{X}_{1:t-1}^i, X_t^i\right)$.
  - Compute the weights

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

Sequential Monte Carlo: output

- Particle approximation of filtering $p(x_t \mid y_{1:t}, \theta)$:

$$\frac{1}{\sum_{j=1}^{N} w_t^j} \sum_{i=1}^{N} w_t^i \delta_{X_t^i}(dx_t),$$

or, after resampling,

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{\bar{X}_t^i}(dx_t).$$

- Particle approximation of path filtering $p(x_{1:t} \mid y_{1:t}, \theta)$:

$$\frac{1}{\sum_{j=1}^{N} w_t^j} \sum_{i=1}^{N} w_t^i \delta_{X_{1:t}^i}(dx_{1:t}),$$

or, similarly, the one after resampling.

Sequential Monte Carlo: complexity

- Propagating and weighting the particles is $\mathcal{O}(N)$.
- Each particle can be propagated and weighted in parallel.
- Multinomial resampling is $\mathcal{O}(N)$ if the uniforms are generated in sorted order.
- Resampling cannot be completely parallel, since it creates correlation between the particles.
- The memory cost is $\mathcal{O}(N)$ if only the latest particles are stored.
- The memory cost is at most $\mathcal{O}(Nt)$ if the paths are stored; efficient implementations reduce this to $\mathcal{O}(t + N \log N)$.

Sequential Monte Carlo: example

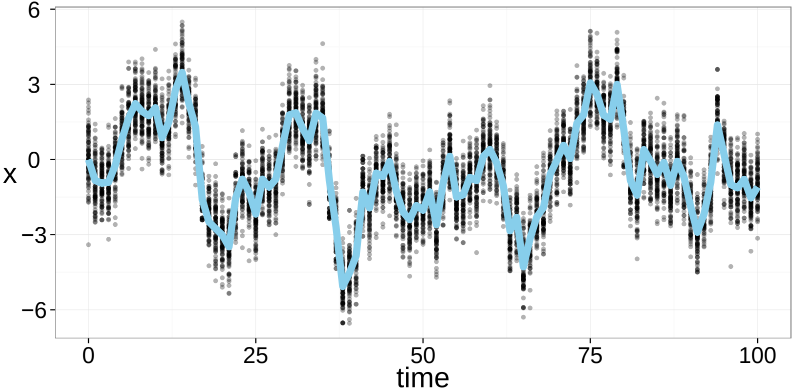

Figure: Support of the approximation $(\bar{X}_t^i)_{i=1}^N$ of $p(x_t \mid y_{1:t})$, over time. The blue curve shows the expectation $\mathbb{E}(x_t \mid y_{1:t})$ at all times t.

Sequential Monte Carlo: example

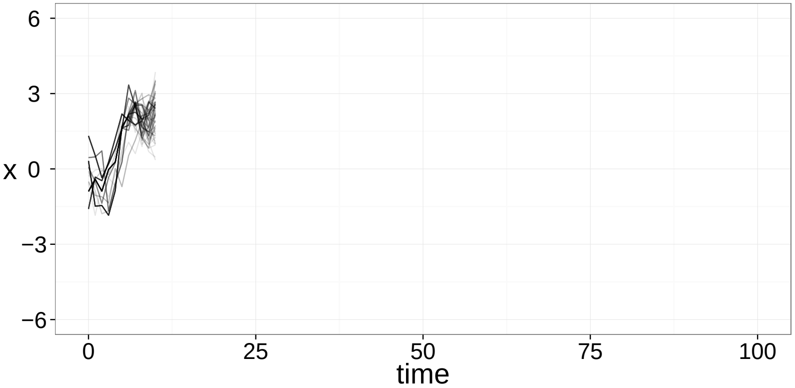

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 10$.

Sequential Monte Carlo: example

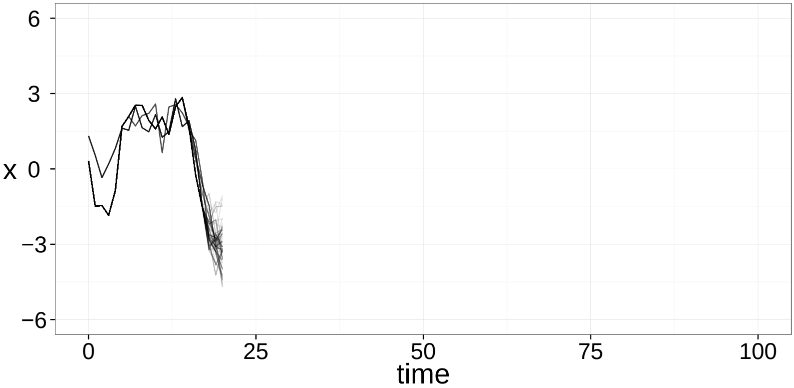

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 20$.

Sequential Monte Carlo: example

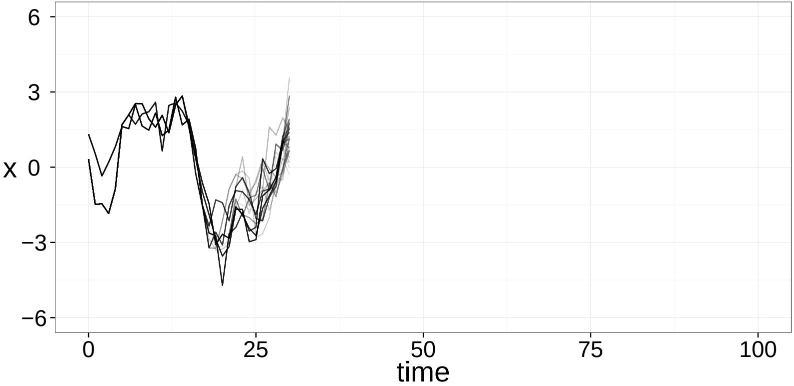

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 30$.

Sequential Monte Carlo: example

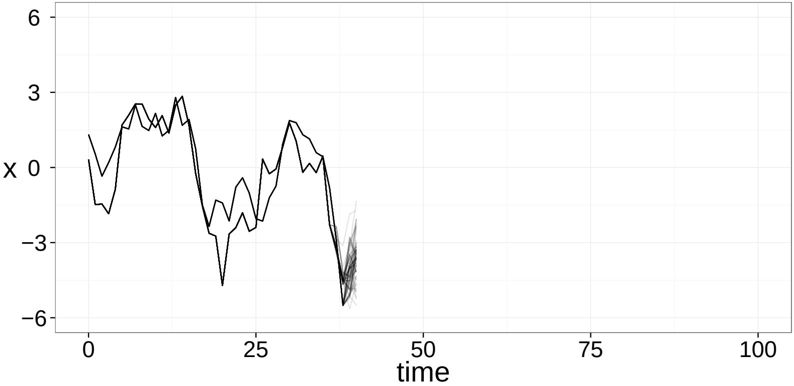

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 40$.

Sequential Monte Carlo: example

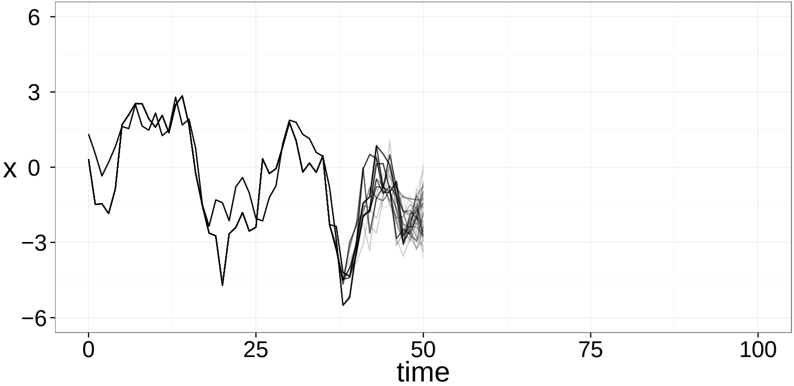

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 50$.

Sequential Monte Carlo: example

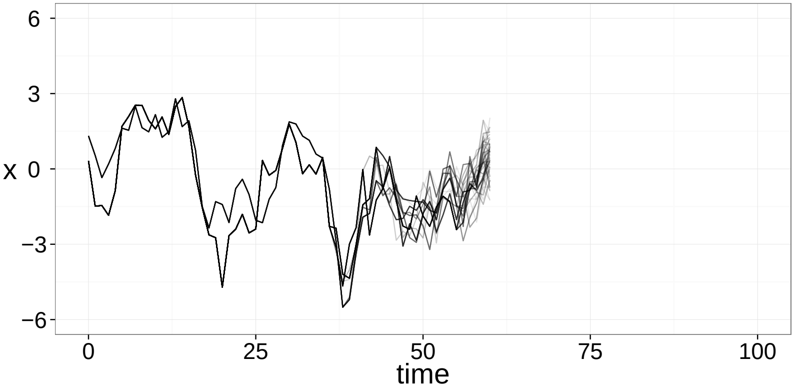

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 60$.

Sequential Monte Carlo: example

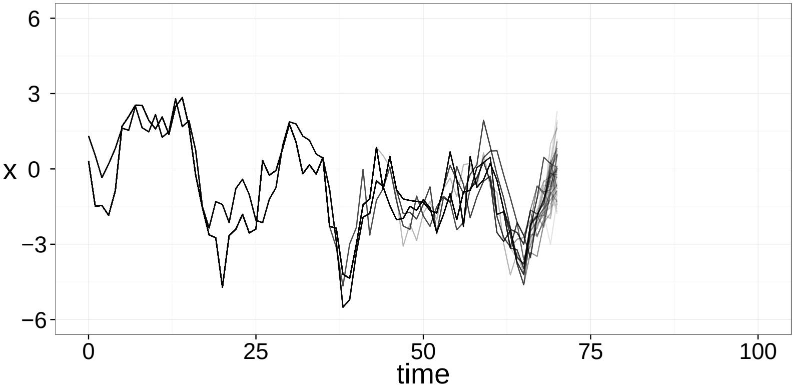

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 70$.

Sequential Monte Carlo: example

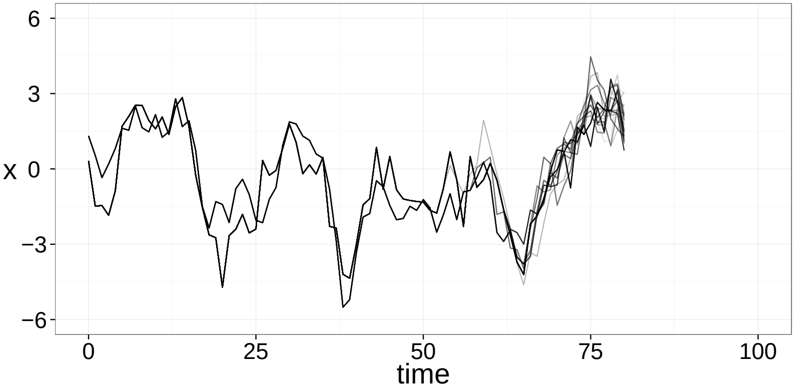

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 80$.

Sequential Monte Carlo: example

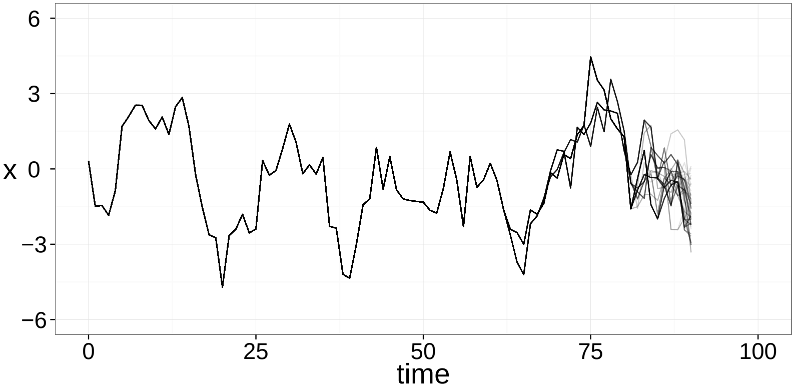

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 90$.

Sequential Monte Carlo: example

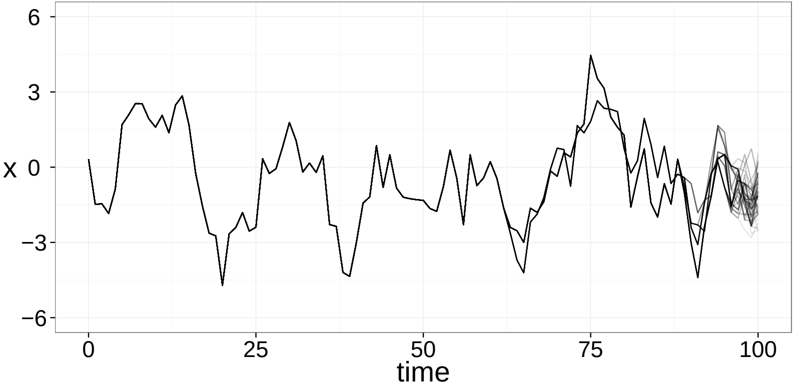

Figure: Trajectories $\bar{X}^i_{1:t}$, at time $t = 100$.

Path degeneracy

- Particle filters approximate well $p(x_t \mid y_{1:t})$ but not $p(x_s \mid y_{1:t})$ for $s << t$.
- Specific particle methods have been developped for this task: fixed lag smoother, forward filtering backward smoothing, etc.
- The simplest is the fixed lag smoother: $p(x_s \mid y_{1:t})$ is approximated by the particle approximation of $p(x_s \mid y_{1:(s+\Delta)\wedge t})$ for a small integer Δ.
- Fixed-lag smoothing introduces a bias but reduces the variance.

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

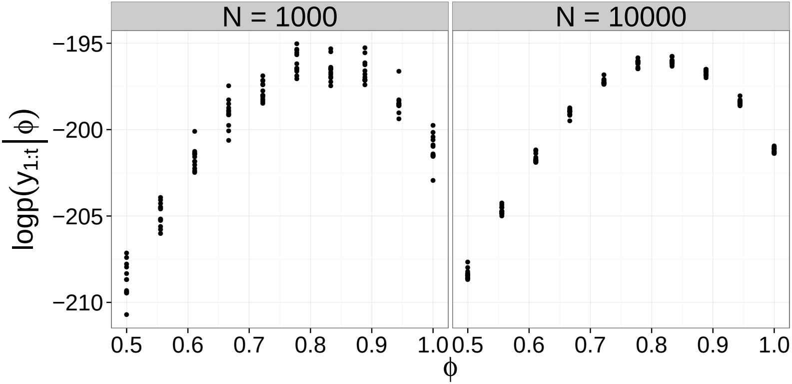

Figure: Log-likelihood estimates $\log p^N(y_{1:t} \mid \phi)$ as a function of ϕ. 12 independent replicates for each value of ϕ.

Selected theoretical results

- Particle filters have been theoretically studied in the past 20 years.
- Convergence results include Central Limit Theorems and non-asymptotic results.
- They provide guidelines to select the number of particles as a function of T, the size of the data, and other algorithmic parameters.
- Consistency as $N \to \infty$ is simple to prove, as each step (propagation, weighting, resampling) is itself consistent.

Selected theoretical results

Consider $I(\varphi_t) = \int \varphi_t(x_{1:t}) p(x_{1:t} \mid y_{1:t}) dx_{1:t}$.

- L_p-bound on the path space:

$$\mathbb{E}\left[\left|I^N(\varphi_t) - I(\varphi_t)\right|^p\right]^{1/p} \leq \frac{B(t)c(p)\left|\left|\varphi_t\right|\right|_\infty}{\sqrt{N}},$$

- Central limit theorem on the path space.

$$\sqrt{N}\left(I^N(\varphi_t) - I(\varphi_t)\right) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right),$$

- As expected, $B(t)$ and σ_t^2 typically grow exponentially fast with t. This is the path degeneracy problem.

Selected theoretical results

Consider instead $I(\varphi_t) = \int \varphi_t(x_t) p(x_t \mid y_{1:t}) dx_t$.

- L_p-bound:

$$\mathbb{E}\left[\left|I^N(\varphi_t) - I(\varphi_t)\right|^p\right]^{1/p} \leq \frac{B_1 c(p) \left|\left|\varphi_t\right|\right|_\infty}{\sqrt{N}}$$

$$\sqrt{N}\left(I^N(\varphi_t) - I(\varphi_t)\right) \xrightarrow[N \to \infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right),$$

- For the filtering estimates, the error is independent of the time t: $\sigma_t^2 < \sigma_{\max}^2$ for all t, and B_1 independent of t.

- Particle filters are fully *online*.

Selected theoretical results

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

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

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