SQMC (Sequential quasi-Monte Carlo)

nicolas.chopin@ensae.fr
(based on a previous PG course with O. Papaspiliopoulos)

Outline

Particle filtering (a.k.a. Sequential Monte Carlo) is a set of *Monte Carlo* techniques for sequential inference in state-space models. The error rate of PF is therefore $O_P(N^{-1/2})$.

Outline

Particle filtering (a.k.a. Sequential Monte Carlo) is a set of *Monte Carlo* techniques for sequential inference in state-space models. The error rate of PF is therefore $O_P(N^{-1/2})$.

Quasi Monte Carlo (QMC) is a substitute for standard Monte Carlo (MC), which typically converges at the faster rate $O(N^{-1+\epsilon})$. However, standard QMC is usually defined for IID problems.

Outline

Particle filtering (a.k.a. Sequential Monte Carlo) is a set of *Monte Carlo* techniques for sequential inference in state-space models. The error rate of PF is therefore $O_P(N^{-1/2})$.

Quasi Monte Carlo (QMC) is a substitute for standard Monte Carlo (MC), which typically converges at the faster rate $O(N^{-1+\epsilon})$. However, standard QMC is usually defined for IID problems.

We derive a QMC version of PF, which we call SQMC (Sequential Quasi Monte Carlo).

Consider the standard MC approximation

$$\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n) \approx \int_{[0,1]^d}\varphi(u)\mathrm{d}u$$

where the N vectors U^n are IID variables simulated from $\mathcal{U}\left([0,1]^d\right)$.

QMC basics

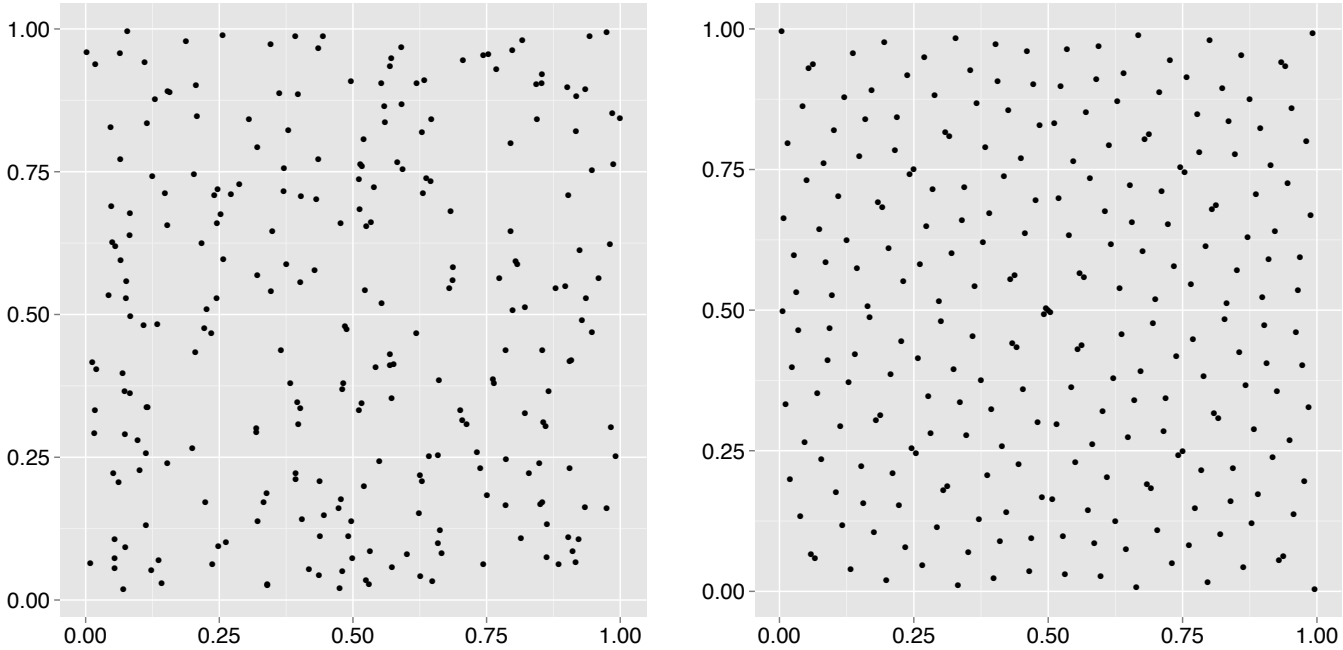
Consider the standard MC approximation

$$\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n) \approx \int_{[0,1]^d}\varphi(u)\mathrm{d}u$$

where the N vectors U^n are IID variables simulated from $\mathcal{U}\left([0,1]^d\right)$.

QMC replaces $U^{1:N}$ by a set of N points that are more evenly distributed on the hyper-cube $[0,1]^d$. This idea is formalised through the notion of *discrepancy*.

QMC vs MC in one plot



QMC versus MC: $N = 256$ points sampled independently and uniformly in $[0,1]^2$ (left); QMC sequence (Sobol) in $[0,1]^2$ of the same length (right)

Discrepancy

Koksma–Hlawka inequality:

$$\left| \frac{1}{N} \sum_{n=1}^{N} \varphi(u^n) - \int_{[0,1]^d} \varphi(u) \, \mathrm{d}u \right| \leq V(\varphi) D^\star(u^{1:N})$$

where $V(\varphi)$ depends only on φ, and the star discrepancy is defined as:

$$D^\star(u^{1:N}) = \sup_{[\mathbf{0},\boldsymbol{b}]} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(U^n \in [\mathbf{0}, \boldsymbol{b}]\right) - \prod_{i=1}^{d} b_i \right|.$$

Discrepancy

Koksma–Hlawka inequality:

$$\left| \frac{1}{N} \sum_{n=1}^{N} \varphi(u^n) - \int_{[0,1]^d} \varphi(u) \, \mathrm{d}u \right| \leq V(\varphi) D^\star(u^{1:N})$$

where $V(\varphi)$ depends only on φ, and the star discrepancy is defined as:

$$D^\star(u^{1:N}) = \sup_{[\mathbf{0},\boldsymbol{b}]} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(U^n \in [\mathbf{0}, \boldsymbol{b}]\right) - \prod_{i=1}^{d} b_i \right|.$$

There are various ways to construct point sets $P_N = \left\{ U^{1:N} \right\}$ so that $D^\star(u^{1:N}) = O(N^{-1+\epsilon})$.

Examples: Van der Corput, Halton

As a simple example of a low-discrepancy sequence in dimension one, $d = 1$, consider

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8} \ldots$$

or more generally,

$$\frac{1}{p}, \ldots, \frac{p-1}{p}, \frac{1}{p^2}, \ldots .$$

Examples: Van der Corput, Halton

As a simple example of a low-discrepancy sequence in dimension one, $d = 1$, consider

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8} \ldots$$

or more generally,

$$\frac{1}{p}, \ldots, \frac{p-1}{p}, \frac{1}{p^2}, \ldots .$$

In dimension $d > 1$, a Halton sequence consists of a Van der Corput sequence for each component, with a different p for each component (the first d prime numbers).

RQMC (randomised QMC)

RQMC randomises QMC so that each $U^n \sim \mathcal{U}\left([0,1]^d\right)$ marginally. In this way

$$\mathbb{E}\left\{\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n)\right\} = \int_{[0,1]^d}\varphi(u)\,\mathrm{d}u$$

and one may evaluate the MSE through independent runs.

RQMC (randomised QMC)

RQMC randomises QMC so that each $U^n \sim \mathcal{U}\left([0,1]^d\right)$ marginally. In this way

$$\mathbb{E}\left\{\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n)\right\} = \int_{[0,1]^d}\varphi(u)\,\mathrm{d}u$$

and one may evaluate the MSE through independent runs.

A simple way to generate a RQMC sequence is to take $U^n = W + V^n \equiv 1$, where $W \sim U([0,1]^d)$ and $V^{1:N}$ is a QMC point set.

RQMC (randomised QMC)

RQMC randomises QMC so that each $U^n \sim \mathcal{U}\left([0,1]^d\right)$ marginally. In this way

$$\mathbb{E}\left\{\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n)\right\} = \int_{[0,1]^d}\varphi(u)\,\mathrm{d}u$$

and one may evaluate the MSE through independent runs.

A simple way to generate a RQMC sequence is to take $U^n = W + V^n \equiv 1$, where $W \sim U([0,1]^d)$ and $V^{1:N}$ is a QMC point set.

Owen (1995, 1997a, 1997b, 1998) developed RQMC strategies such that (for a certain class of smooth functions φ):

$$\mathrm{Var}\left\{\frac{1}{N}\sum_{n=1}^{N}\varphi(U^n)\right\} = O(N^{-3+\epsilon})$$

Particle Filtering: Hidden Markov models

Consider an unobserved Markov chain (X_t), $X_0 \sim m_0(\mathrm{d}x_0)$ and

$$X_t | X_{t-1} = x_{t-1} \sim M_t(x_{t-1}, \mathrm{d}x_t)$$

taking values in $\mathcal{X} \subset \mathbb{R}^d$, and an observed process (Y_t),

$$Y_t | X_t \sim g(y_t | x_t).$$

Particle Filtering: Hidden Markov models

Consider an unobserved Markov chain (X_t), $X_0 \sim m_0(\mathrm{d}x_0)$ and

$$X_t | X_{t-1} = x_{t-1} \sim M_t(x_{t-1}, \mathrm{d}x_t)$$

taking values in $\mathcal{X} \subset \mathbb{R}^d$, and an observed process (Y_t),

$$Y_t | X_t \sim g(y_t | x_t).$$

Sequential analysis of HMMs amounts to recover quantities such as $p(x_t | y_{0:t})$ (filtering), $p(x_{t+1} | y_{0:t})$ (prediction), $p(y_{0:t})$ (marginal likelihood), etc., recursively in time. Many applications in engineering (tracking), finance (stochastic volatility), epidemiology, ecology, neurosciences, etc.

Feynman-Kac formalism

Taking $G_t(x_{t-1}, x_t) := g_t(y_t|x_t)$, we see that sequential analysis of a HMM may be cast into a Feynman-Kac model. In particular, *filtering* amounts to computing

$$\mathbb{Q}_t(\varphi) = \frac{1}{Z_t}\mathbb{E}\left[\varphi(X_t)G_0(X_0)\prod_{s=1}^{t} G_s(X_{s-1}, X_s)\right],$$

$$\text{with } Z_t = \mathbb{E}\left[G_0(X_0)\prod_{s=1}^{t} G_s(X_{s-1}, X_s)\right]$$

and expectations are wrt the law of the Markov chain (X_t).

Feynman-Kac formalism

Taking $G_t(x_{t-1}, x_t) := g_t(y_t|x_t)$, we see that sequential analysis of a HMM may be cast into a Feynman-Kac model. In particular, *filtering* amounts to computing

$$\mathbb{Q}_t(\varphi) = \frac{1}{Z_t}\mathbb{E}\left[\varphi(X_t)G_0(X_0)\prod_{s=1}^{t} G_s(X_{s-1}, X_s)\right],$$

$$\text{with } Z_t = \mathbb{E}\left[G_0(X_0)\prod_{s=1}^{t} G_s(X_{s-1}, X_s)\right]$$

and expectations are wrt the law of the Markov chain (X_t).

Note: FK formalism has other applications that sequential analysis of HMM. In addition, for a given HMM, there is a more than one way to define a Feynmann-Kac formulation of that model.

Particle filtering: the algorithm

Operations must be be performed for all $n \in 1:N$.
At time 0,

(a) Generate $X_0^n \sim M_0(\mathrm{d}x_0)$.

(b) Compute $W_0^n = G_0(X_0^n) / \sum_{m=1}^N G_0(X_0^m)$.

Recursively, for time $t = 1:T$,

(a) Generate $A_{t-1}^n \sim \mathcal{M}(W_{t-1}^{1:N})$.

(b) Generate $X_t^n \sim M_t(X_{t-1}^{A_{t-1}^n}, \mathrm{d}x_t)$.

(c) Compute

$$W_t^n = G_t(X_{t-1}^{A_{t-1}^n}, X_t^n) / \sum_{m=1}^N G_t(X_{t-1}^{A_{t-1}^m}, X_t^m)$$

Formalisation

We can formalise the succession of Steps (a), (b) and (c) at iteration t as an importance sampling step from random probability measure

$$\sum_{n=1}^{N} W_{t-1}^n \delta_{X_{t-1}^n}(\mathrm{d}\widetilde{x}_{t-1}) M_t(\widetilde{x}_{t-1}, \mathrm{d}x_t) \qquad (0.1)$$

to

$$\{\text{same thing}\} \times G_t(\widetilde{x}_{t-1}, x_t).$$

Formalisation

We can formalise the succession of Steps (a), (b) and (c) at iteration t as an importance sampling step from random probability measure

$$\sum_{n=1}^{N} W_{t-1}^{n} \delta_{X_{t-1}^{n}}(\mathrm{d}\widetilde{x}_{t-1}) M_t(\widetilde{x}_{t-1}, \mathrm{d}x_t) \qquad (0.1)$$

to

$$\{\text{same thing}\} \times G_t(\widetilde{x}_{t-1}, x_t).$$

Idea: use QMC instead of MC to sample N points from (0.1); i.e. rewrite sampling from (0.1) this as a function of uniform variables, and use low-discrepancy sequences instead.

Intermediate step

More precisely, we are going to write the simulation from

$$\sum_{n=1}^{N} W_{t-1}^n \delta_{X_{t-1}^n}(\mathrm{d}\widetilde{x}_{t-1}) M_t(\widetilde{x}_{t-1}, \mathrm{d}x_t)$$

as a function of $U_t^n = (u_t^n, V_t^n)$, $u_t^n \in [0,1]$, $V_t^n \in [0,1]^d$, such that:

1. We will use the scalar u_t^n to choose the ancestor $\widetilde{X}_{t-1}$.
2. We will use V_t^n to generate X_t^n as

$$X_t^n = \Gamma_t(\widetilde{X}_{t-1}, V_t^n)$$

where Γ_t is a deterministic function such that, for $V_t^n \sim \mathcal{U}[0,1]^d$, $\Gamma_t(\widetilde{X}_{t-1}, V_t^n) \sim M_t(\widetilde{X}_{t-1}, \mathrm{d}x_t)$.

Intermediate step

More precisely, we are going to write the simulation from

$$\sum_{n=1}^{N} W_{t-1}^n \delta_{X_{t-1}^n}(\mathrm{d}\widetilde{x}_{t-1}) M_t(\widetilde{x}_{t-1}, \mathrm{d}x_t)$$

as a function of $U_t^n = (u_t^n, V_t^n)$, $u_t^n \in [0,1]$, $V_t^n \in [0,1]^d$, such that:

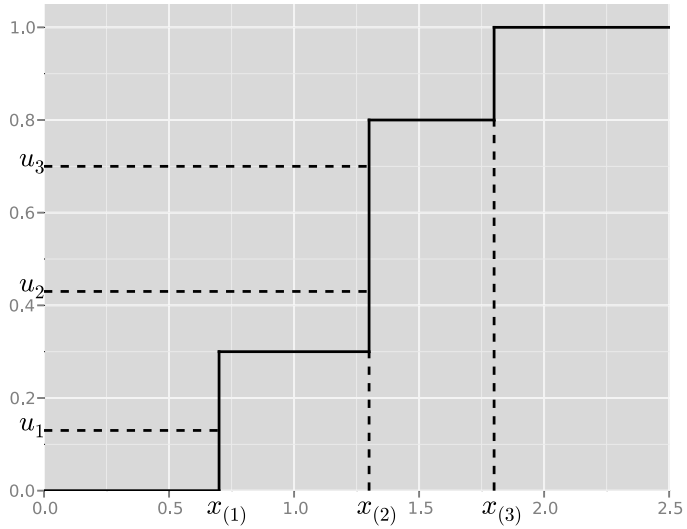
1. We will use the scalar u_t^n to choose the ancestor $\widetilde{X}_{t-1}$.
2. We will use V_t^n to generate X_t^n as

$$X_t^n = \Gamma_t(\widetilde{X}_{t-1}, V_t^n)$$

where Γ_t is a deterministic function such that, for $V_t^n \sim \mathcal{U}[0,1]^d$, $\Gamma_t(\widetilde{X}_{t-1}, V_t^n) \sim M_t(\widetilde{X}_{t-1}, \mathrm{d}x_t)$.

The main problem is point 1.

Case $d = 1$



Simply use the inverse transform method: $\tilde{X}_{t-1}^n = \hat{F}^{-1}(u_t^n)$, where $\hat{F}$ is the empirical cdf of

$$\sum_{n=1}^{N} W_{t-1}^n \delta_{X_{t-1}^n}(\mathrm{d}\widetilde{X}_{t-1}).$$

From $d = 1$ to $d > 1$

When $d > 1$, we cannot use the inverse CDF method to sample from the empirical distribution

$$\sum_{n=1}^{N} W_{t-1}^{n} \delta_{X_{t-1}^{n}}(\mathrm{d}\widetilde{x}_{t-1}).$$

Idea: we "project" the X_{t-1}^{n}'s into $[0,1]$ through the (generalised) inverse of the *Hilbert curve*, which is a fractal, space-filling curve $H : [0,1] \to [0,1]^{d}$.

From $d = 1$ to $d > 1$

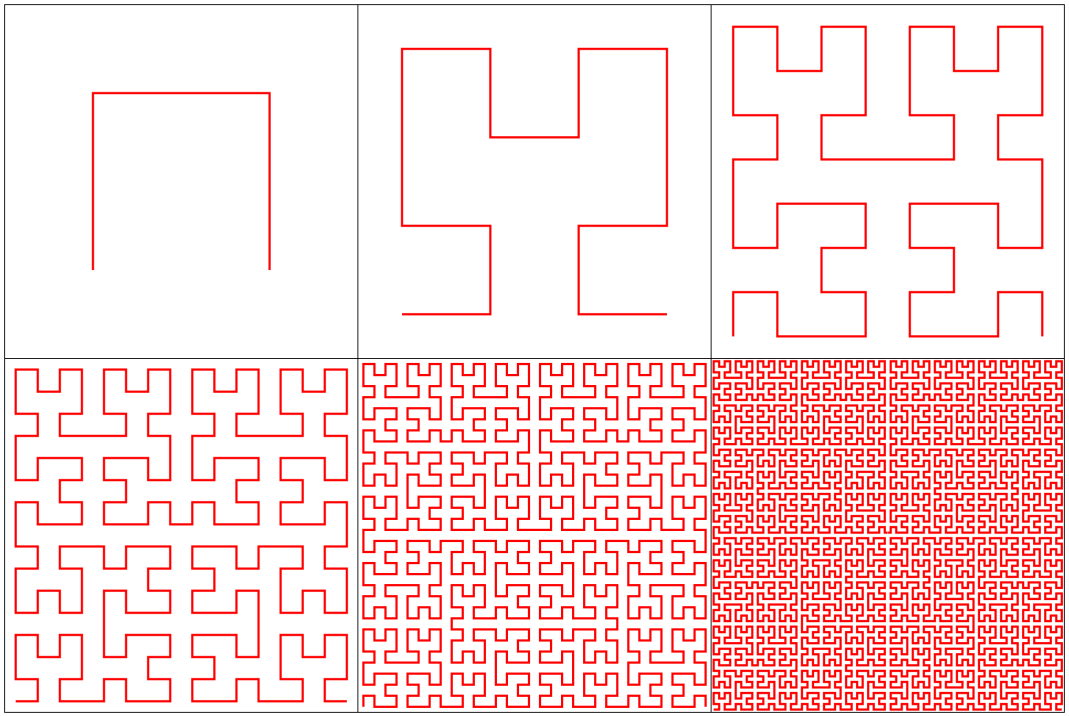
When $d > 1$, we cannot use the inverse CDF method to sample from the empirical distribution

$$\sum_{n=1}^{N} W_{t-1}^n \delta_{X_{t-1}^n}(\mathrm{d}\widetilde{x}_{t-1}).$$

Idea: we "project" the X_{t-1}^n's into $[0,1]$ through the (generalised) inverse of the *Hilbert curve*, which is a fractal, space-filling curve $H : [0,1] \to [0,1]^d$.

More precisely, we transform $\mathcal{X}$ into $[0,1]^d$ through some function ψ, then we transform $[0,1]^d$ into $[0,1]$ through $h = H^{-1}$.

Hilbert curve



The Hilbert curve is the limit of this sequence. Note the locality property of the Hilbert curve: if two points are close in $[0,1]$, then the the corresponding transformed points remains close in $[0,1]^d$. (Source for the plot: Wikipedia)

SQMC Algorithm

At time 0,

(a) Generate a QMC point set $U_0^{1:N}$ in $[0,1]^d$, and compute $X_0^n = \Gamma_0(U_0^n)$. (e.g. $\Gamma_0 = F_{m_0}^{-1}$)

(b) Compute $W_0^n = G_0(X_0^n) / \sum_{m=1}^N G_0(X_0^m)$.

Recursively, for time $t = 1 : T$,

(a) Generate a QMC point set $U_t^{1:N}$ in $[0,1]^{d+1}$; let $U_t^n = (u_t^n, V_t^n)$.

(b) Hilbert sort: find permutation σ such that $\circ\psi(X_{t-1}^{\sigma(1)}) \leq \ldots \leq \circ\psi(X_{t-1}^{\sigma(N)})$.

(c) Generate $a_{t-1}^{1:N}$ using inverse CDF Algorithm, with inputs $\mathrm{sort}(u_t^{1:N})$ and $W_{t-1}^{\sigma(1:N)}$, and compute $X_t^n = \Gamma_t(X_{t-1}^{\sigma(a_{t-1}^n)}, V_t^{\sigma(n)})$. (e.g. $\Gamma_t = F_{M_t}^{-1}$)

(e) Compute $W_t^n = G_t(X_{t-1}^{\sigma(a_{t-1}^n)}, X_t^n) / \sum_{m=1}^N G_t(X_{t-1}^{\sigma(a_{t-1}^m)}, X_t^m)$.

Some remarks

- Because two sort operations are performed, the complexity of SQMC is $O(N \log N)$. (Compare with $O(N)$ for SMC.)

Some remarks

- Because two sort operations are performed, the complexity of SQMC is $O(N \log N)$. (Compare with $O(N)$ for SMC.)
- The main requirement to implement SQMC is that one may simulate from Markov kernel $M_t(x_{t-1}, \mathrm{d}x_t)$ by computing $X_t = \Gamma_t(X_{t-1}, _t)$, where $_t \sim \mathcal{U}[0,1]^d$, for some deterministic function Γ_t (e.g. multivariate inverse CDF).

Some remarks

- Because two sort operations are performed, the complexity of SQMC is $O(N \log N)$. (Compare with $O(N)$ for SMC.)
- The main requirement to implement SQMC is that one may simulate from Markov kernel $M_t(x_{t-1}, \mathrm{d}x_t)$ by computing $X_t = \Gamma_t(X_{t-1}, _t)$, where $_t \sim \mathcal{U}[0,1]^d$, for some deterministic function Γ_t (e.g. multivariate inverse CDF).
- The dimension of the point sets $_t^{1:N}$ is $1 + d$: first component is for selecting the parent particle, the d remaining components is for sampling X_t^n given $X_{t-1}^{a_{t-1}^n}$.

Extensions

- If we use RQMC (randomised QMC) point sets $\overset{1:N}{t}$, then SQMC generates an unbiased estimate of the marginal likelihood Z_t.

Extensions

- If we use RQMC (randomised QMC) point sets $\overset{1:N}{t}$, then SQMC generates an unbiased estimate of the marginal likelihood Z_t.
- This means we can use SQMC within the *PMCMC* framework. (More precisely, we can run e.g. a PMMH algorithm, where the likelihood of the data is computed via SQMC instead of SMC.)

Extensions

- If we use RQMC (randomised QMC) point sets $\frac{1:N}{t}$, then SQMC generates an unbiased estimate of the marginal likelihood Z_t.
- This means we can use SQMC within the *PMCMC* framework. (More precisely, we can run e.g. a PMMH algorithm, where the likelihood of the data is computed via SQMC instead of SMC.)
- We can also adapt quite easily the different particle smoothing algorithms: forward smoothing, backward smoothing, two-filter smoothing.

Extensions

- If we use RQMC (randomised QMC) point sets $\frac{1:N}{t}$, then SQMC generates an unbiased estimate of the marginal likelihood Z_t.
- This means we can use SQMC within the *PMCMC* framework. (More precisely, we can run e.g. a PMMH algorithm, where the likelihood of the data is computed via SQMC instead of SMC.)
- We can also adapt quite easily the different particle smoothing algorithms: forward smoothing, backward smoothing, two-filter smoothing.

Main results

We were able to establish the following types of results: *consistency*

$$\mathbb{Q}_t^N(\varphi) - \mathbb{Q}_t(\varphi) \to 0, \quad \text{as } N \to +\infty$$

for certain functions φ, and *rate of convergence*

$$\mathrm{MSE}\left[\mathbb{Q}_t^N(\varphi)\right] = (N^{-1})$$

(under technical conditions, and for certain types of RQMC point sets).
Theory is non-standard and borrows heavily from QMC concepts.

Some concepts used in the proofs

Let $\mathcal{X} = [0,1]^d$. Consistency results are expressed in terms of the star norm

$$\|\mathbb{Q}_t^N - \mathbb{Q}_t\|_\star = \sup_{[\mathbf{0},\boldsymbol{b}]\subset[0,1)^d} \left| \left(\mathbb{Q}_t^N - \mathbb{Q}_t \right)(B) \right| \to 0.$$

This implies consistency for bounded functions φ,
$\mathbb{Q}_t^N(\varphi) - \mathbb{Q}_t(\varphi) \to 0$.
The Hilbert curve conserves discrepancy:

$$\|\pi^N - \pi\|_\star \to 0 \quad \Rightarrow \quad \|\pi_h^N - \pi_h\|_\star \to 0$$

where $\pi \in \mathcal{P}([0,1]^d)$, $h : [0,1]^d \to [0,1]$ is the (pseudo-)inverse of the Hilbert curve, and π_h is the image of π through π.

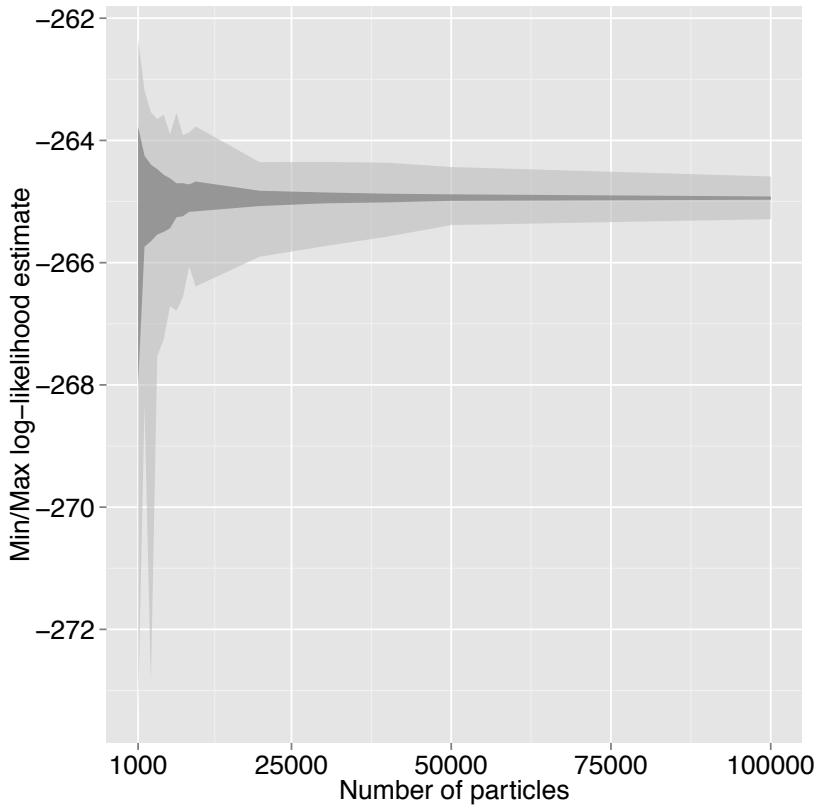
Examples: Kitagawa ($d = 1$)

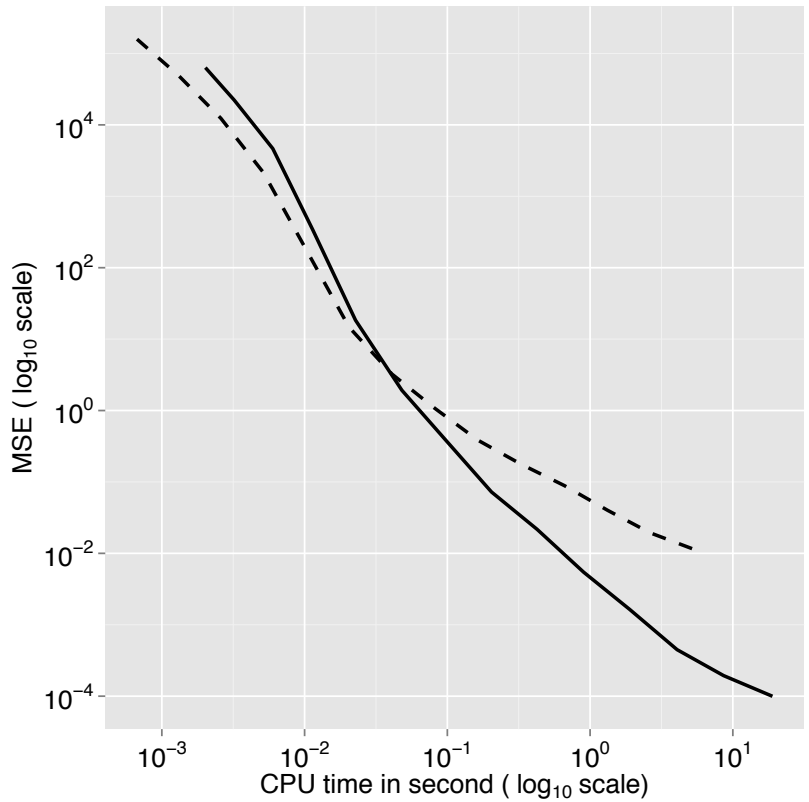
Well known toy example (Kitagawa, 1998):

$$\begin{cases} y_t = \frac{x_t^2}{a} + \epsilon_t \\ x_t = b_1 x_{t-1} + b_2 \frac{x_{t-1}}{1+x_{t-1}^2} + b_3 \cos(b_4 t) + \sigma \nu_t \end{cases}$$

No paramater estimation (parameters are set to their true value).
We compare SQMC with SMC (based on systematic resampling) both in terms of N, and in terms of CPU time.

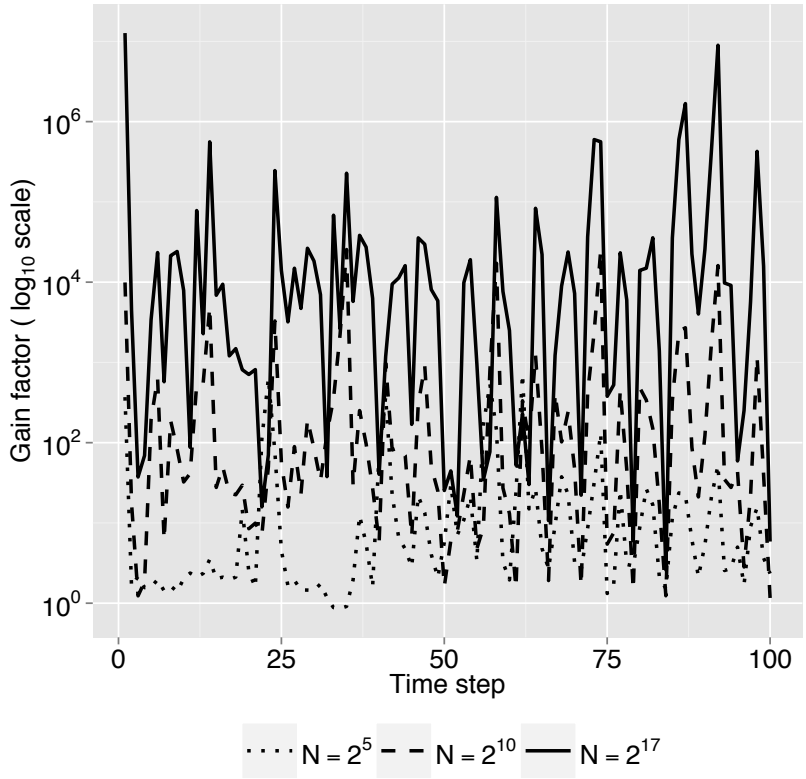
Examples: Kitagawa ($d = 1$)

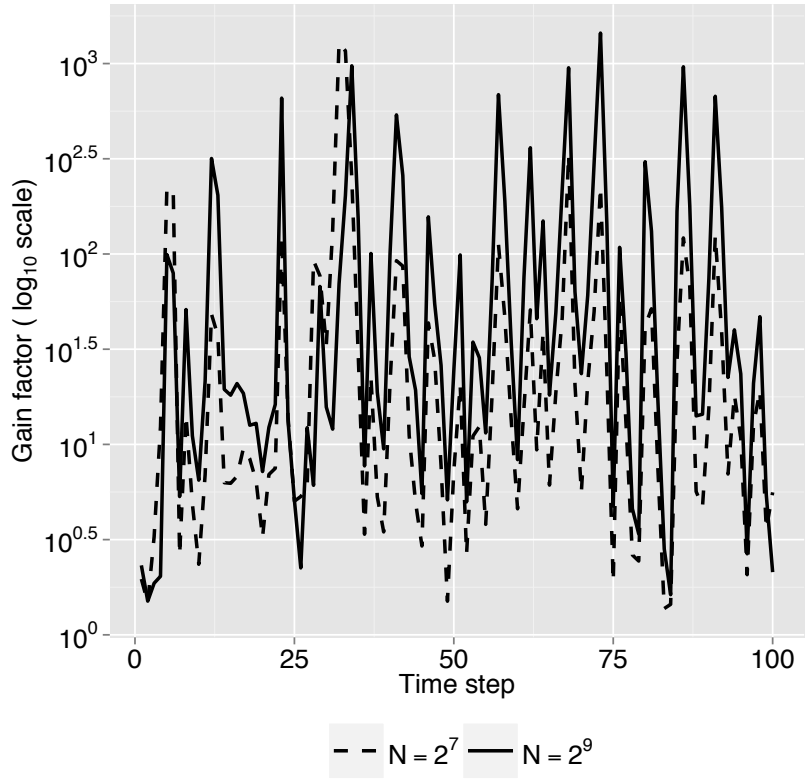




Log-likelihood evaluation (based on $T = 100$ data point and 500 independent SMC and SQMC runs).

Examples: Kitagawa ($d = 1$)





Filtering: computing $\mathbb{E}(X_t|_{0:t})$ at each iteration t. Gain factor is MSE(SMC)/MSE(SQMC).

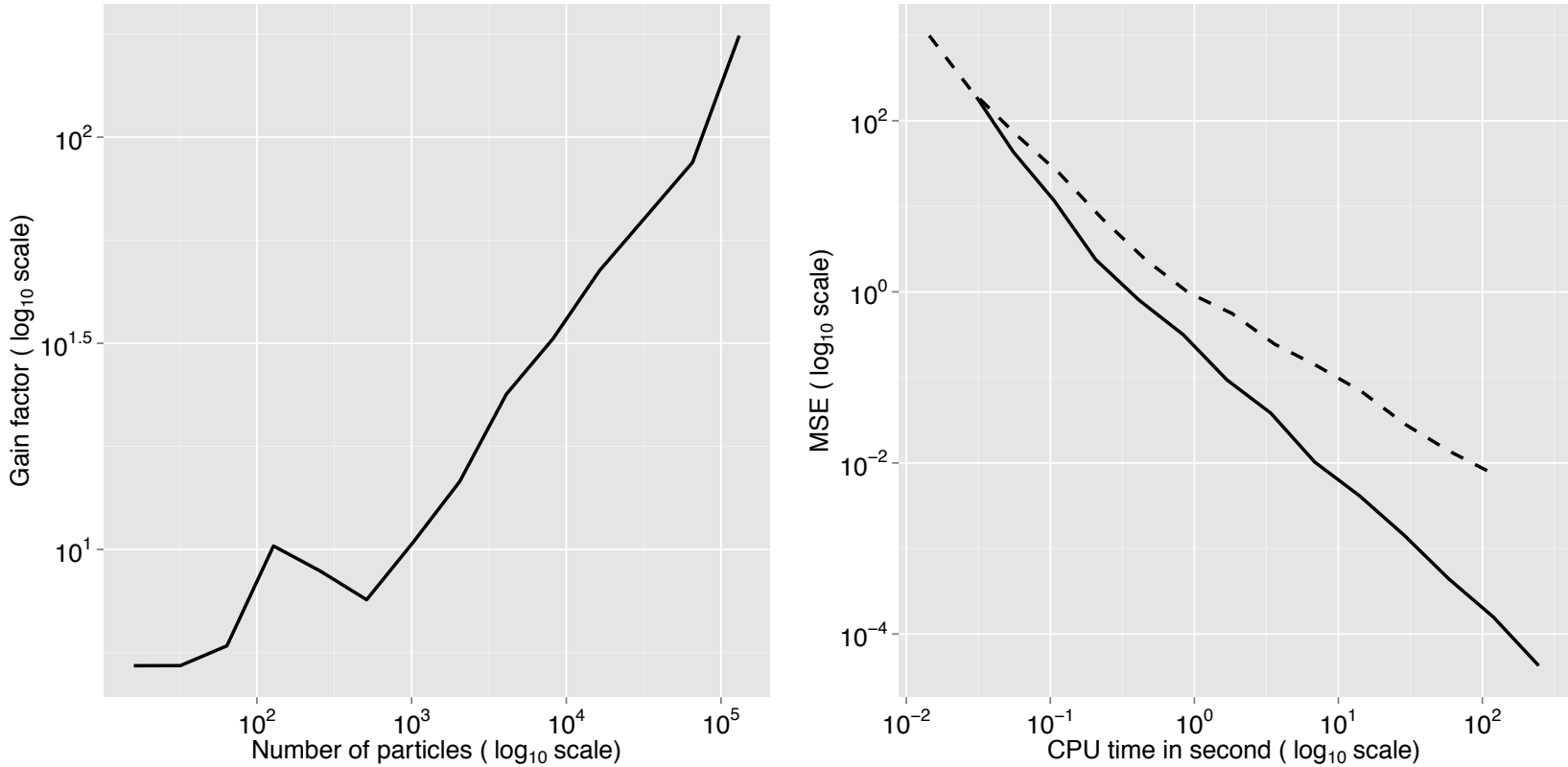
Examples: Multivariate Stochastic Volatility

Model is

$$\begin{cases} {}_t = S_t^{\frac{1}{2}} \epsilon_t \\ X_t = \mu + \Phi(X_{t-1} - \mu) + \Psi^{\frac{1}{2}} \nu_t \end{cases}$$

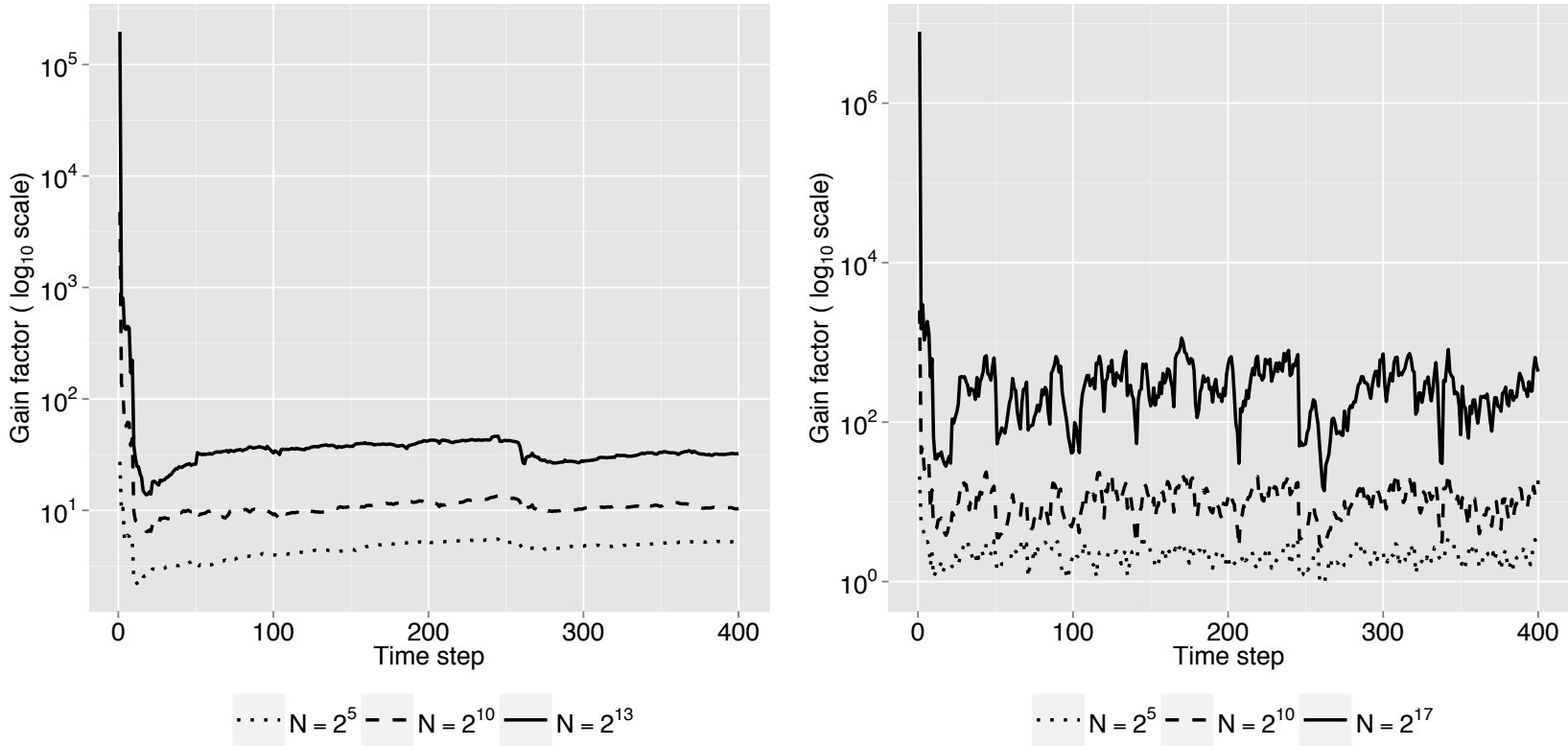
with possibly correlated noise terms: $(\epsilon_t, \nu_t) \sim N_{2d}(\mathbf{0}, \boldsymbol{C})$.
We shall focus on $d = 2$ and $d = 4$.

Examples: Multivariate Stochastic Volatility ($d = 2$)



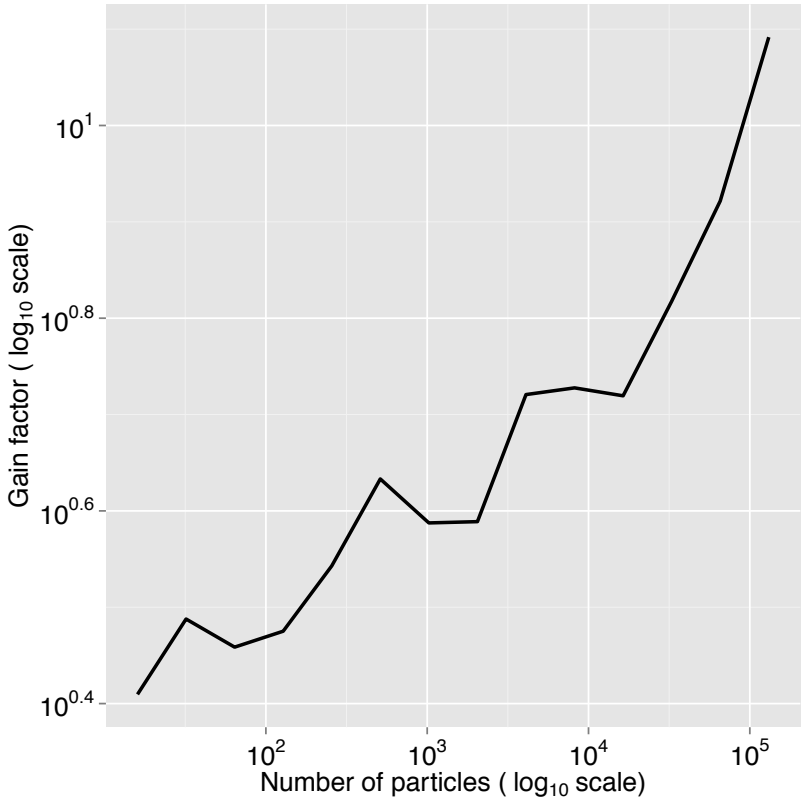
Log-likelihood evaluation (based on $T = 400$ data points and 200 independent runs).

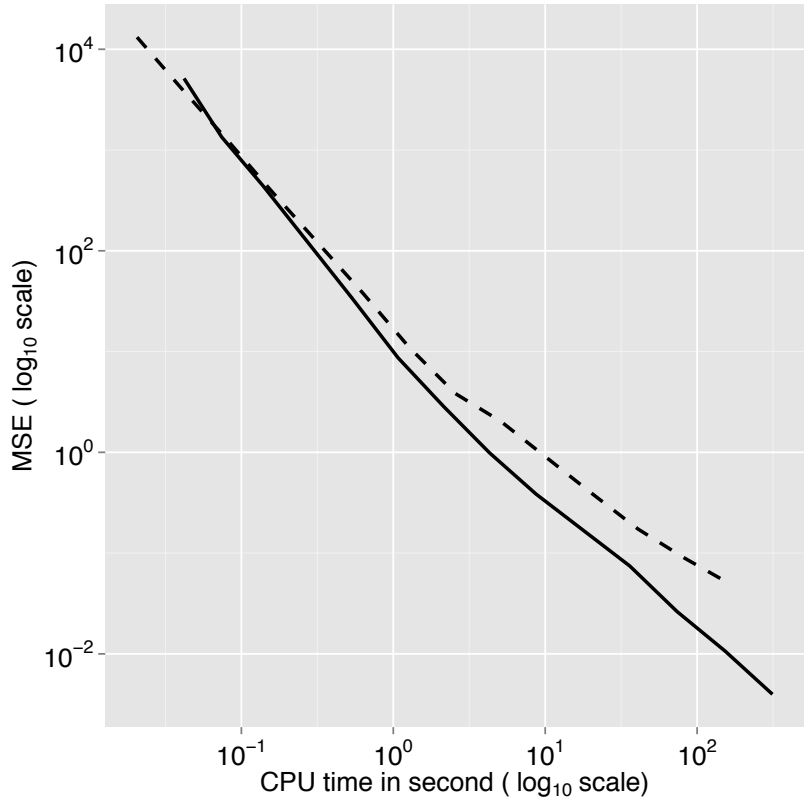
Examples: Multivariate Stochastic Volatility ($d = 2$)



Log-likelihood evaluation (left) and filtering (right) as a function of t.

Examples: Multivariate Stochastic Volatility ($d = 4$)





Log-likelihood estimation (based on $T = 400$ data points and 200 independent runs)

Conclusion

- Only requirement to replace SMC with SQMC is that the simulation of $X_t^n | X_{t-1}^n$ may be written as a $X_t^n = \Gamma_t(X_{t-1}^n, \epsilon_t^n)$ where $\epsilon_t^n \sim U[0,1]^d$.
- We observe *very impressive* gains in performance (even for small N or $d = 6$).
- Supporting theory.

Further work

- Adaptive resampling (triggers resampling steps when weight degeneracy is too high).
- Adapt SQMC to situations where sampling from $M_t(X_{t-1}^n, \mathrm{d}x_t)$ involves some accept/reject mechanism.
- Adapt SQMC to situations where sampling from $M_t(X_{t-1}^n, \mathrm{d}x_t)$ is a Metropolis step. In this way, we could develop SQMC counterparts of *SMC samplers* (Del Moral et al, 2006).
- SQMC^2 (QMC version of SMC^2, C. et al, 2013)?

Further work

- Adaptive resampling (triggers resampling steps when weight degeneracy is too high).
- Adapt SQMC to situations where sampling from $M_t(X_{t-1}^n, \mathrm{d}x_t)$ involves some accept/reject mechanism.
- Adapt SQMC to situations where sampling from $M_t(X_{t-1}^n, \mathrm{d}x_t)$ is a Metropolis step. In this way, we could develop SQMC counterparts of *SMC samplers* (Del Moral et al, 2006).
- SQMC^2 (QMC version of SMC^2, C. et al, 2013)?

Paper on Arxiv, will be published soon as a read paper in JRSSB.

Particles as auxiliary variables: PMCMC and related algorithms

nicolas.chopin@ensae.fr
(based on a previous PG course with O. Papaspiliopoulos)

Particles as auxiliary variables: PMCMC and related algorithms

nicolas.chopin@ensae.fr
(based on a previous PG course with O. Papaspiliopoulos)

Outline

Tractable models

For a standard Bayesian model, defined by (a) prior $p(\theta)$, and (b) likelihood $p(y|\theta)$, a standard approach is to use the Metropolis-Hastings algorithm to sample from the posterior

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

Metropolis-Hastings

From current point θ_m

1. Sample $\theta_\star \sim H(\theta_m, \mathrm{d}\theta_\star)$
2. With probability $1 \wedge r$, take $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$, where

$$r = \frac{p(\theta_\star)p(y|\theta_\star)h(\theta_m|\theta_\star)}{p(\theta_m)p(y|\theta_m)h(\theta_\star|\theta_m)}$$

This generates a Markov chain which leaves $p(\theta|y)$ invariant.

Metropolis Proposal

Note that proposal kernel $H(\theta_m, \mathrm{d}\theta_\star)$ (to simulate proposed value $\theta^\star$, conditional on current value θ_m). Popular choices are:

- random walk proposal: $h(\theta^\star|\theta_m) = N(\theta^\star; \theta_m, \Sigma)$; usual recommendation is to take $\Sigma \approx c_d \Sigma_{\mathrm{post}}$, with $c_d = 2.38^2/d$.
- independent proposal: $h(\theta^\star|\theta_m) = h(\theta^\star)$.
- Langevin proposals.

Intractable models

This generic approach cannot be applied in the following situations:

1. The likelihood is $p(y|\theta) = h_\theta(y)/Z(\theta)$, where $Z(\theta)$ is an intractable normalising constant; e.g. log-linear models, network models, Ising models.
2. The likelihood $p(y|\theta)$ is an intractable integral

$$p(y|\theta) = \int_{\mathcal{X}} p(y, x|\theta)\, dx.$$

3. The likelihood is even more complicated, because it corresponds to some scientific model involving some complicate *generative* process (scientific models, "likelihood-free inference", ABC).

Example of likelihoods as intractable integrals

When $p(y|\theta) = \int p(y, x|\theta)\, dx$.

- phylogenetic trees (Beaumont, 2003);
- state-space models (see later);
- other models with latent variables.

We will focus on this case, but certain ideas may also be applied to the two other cases.

Outline

General framework

Consider posterior

$$\pi(\theta, x) \propto p(\theta)p(x|\theta)p(y|x, \theta)$$

where typically x is of much larger dimension than θ.
One potential approach to sample from the posterior is *Gibbs sampling*: iteratively sample $\theta|x, y$, then $x|\theta, y$. However, there are many cases where Gibbs is either difficult to implement, or quite inefficient.
Instead, we would like to sample *marginally* from

$$\pi(\theta) \propto p(\theta)p(y|\theta), \quad p(y|\theta) = \int_{\mathcal{X}} p(x, y|\theta)\, dx$$

but again $p(y|\theta)$ is intractable...

Importance sampling

I cannot compute $p(y|\theta)$, but I can compute an *unbiased* estimator of this quantity:

$$\hat{p}(y|\theta) = \frac{1}{N}\sum_{n=1}^{N} \frac{p(y, x^n|\theta)}{q(x^n)}, \quad x^{1:N} \overset{iid}{\sim} q(x)$$

using *importance sampling*.

The pseudo-marginal approach

GIMH (Beaumont, 2003)

From current point θ_m

1. Sample $\theta_\star \sim H(\theta_m, \mathrm{d}\theta_\star)$
2. With prob. $1 \wedge r$, take $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$, with

$$r = \frac{p(\theta_\star)\hat{p}(y|\theta_\star)h(\theta_m|\theta_\star)}{p(\theta_m)\hat{p}(y|\theta_m)h(\theta_\star|\theta_m)}$$

Note that $\hat{p}(y|\theta_\star)$ is based on independent samples generated at iteration m.
Question: Is GIMH a *non-standard* HM sampler w.r.t. *standard* target $\pi(\theta)$?

Validity of GIMH

Property 1

The following function

$$\bar{\pi}(\theta, x^{1:N}) = \prod_{n=1}^{N} q(x^n) \frac{p(\theta)\hat{p}(y|\theta)}{p(y)}$$

is a joint PDF, whose θ-marginal is $\pi(\theta) \propto p(\theta)p(y|\theta)$.

Proof: Direct consequence of unbiasedness; fix θ then

$$\int \prod_{n=1}^{N} q(x^n) p(\theta)\hat{p}(y|\theta)\, dx^{1:N} = p(\theta)\mathbb{E}\left[\hat{p}(y|\theta)\right] = p(\theta)p(y|\theta)$$

GIMH as a Metropolis sampler

Property 2

GIMH is a Metropolis sampler with respect to joint distribution $\bar{\pi}(\theta, x^{1:N})$. The proposal density is $h(\theta_\star|\theta_m)\prod_{n=1}^N q(x_\star^n)$.

Proof: current point is $(\theta_m, x_m^{1:N})$, proposed point is $(\theta_\star, x_\star^{1:N})$ and HM ratio is

$$r = \frac{\cancel{\prod_{n=1}^N q(x_\star^n)}\, p(\theta_\star)\hat{p}(y|\theta_\star)h(\theta_m|\theta_\star)\cancel{\prod_{n=1}^N q(x_m^n)}}{\cancel{\prod_{n=1}^N q(x_m^n)}\, p(\theta_m)\hat{p}(y|\theta_m)h(\theta_\star|\theta_m)\cancel{\prod_{n=1}^N q(x_\star^n)}}$$

Thus, GIMH is a *standard* Metropolis sampler w.r.t. *non-standard* (extended) target $\bar{\pi}(\theta, x^{1:N})$.

There is more to life than this

Property 3

Extend $\bar{\pi}(\theta, x^{1:N})$ with $k|\theta, x^{1:N} \propto \pi(\theta, x^k)/q(x^k)$, then,

- the marginal dist. of (θ, x^k) is $\pi(\theta, x)$.
- Conditional on (θ, x^k), $x_n \sim q$ for $n \neq k$, independently.

Proof: let

$$\bar{\pi}(\theta, x^{1:N}, k) = \left\{ \prod_{n=1}^{N} q(x^n) \right\} \frac{\pi(\theta, x^k)}{q(x^k)} = \left\{ \prod_{n \neq k} q(x^n) \right\} \pi(\theta, x^k)$$

then clearly the sum w.r.t. k gives $\bar{\pi}(\theta, x^{1:N})$, while the above properties hold.

We can do Gibbs!

One consequence of Property 3 is that we gain the ability to perform *Gibbs*, in order to regenerate the $N-1$ non-selected points x^n, $n \neq k$. More precisely:

1. Sample $k \sim \pi(k|\theta, x^{1:N}) \propto \pi(\theta, x^k)/q(x^k)$
2. regenerate $x^n \sim q$, for all $n \neq k$.

Could be useful for instance to avoid "getting stuck", because say the current value $\hat{\pi}(\theta)$ is too high.

Main lessons

- We can replace an intractable quantity by an unbiased estimate, *without introducing any approximation*.
- In fact, we can do more: with Proposition 3, we have obtained that
  1. it is possible to sample from $\pi(\theta, x)$ jointly;
  2. it is possible to do a Gibbs step where the $N-1$ x^n, $n \neq k$ are regenerated (useful when GIMH "get stucks"?)
- but careful, it is possible to get it wrong...

Unbiasedness without an auxiliary variable representation

This time, consider instead a target $\pi(\theta)$ (no x), involving an intractable *denominator*; an important application is Bayesian inference on likelihoods with intractable normalising constants:

$$\pi(\theta) \propto p(\theta)p(y|\theta) = p(\theta)\frac{h_\theta(y)}{Z(\theta)}$$

Liang & Lin (2010)'s sampler

From current point θ_m

1. Sample $\theta_\star \sim H(\theta^m, \mathrm{d}\theta_\star)$
2. With prob. $1 \wedge r$, take $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$, with

$$r = \left(\frac{\widehat{Z(\theta_m)}}{Z(\theta_\star)}\right)\frac{p(\theta_\star)h_{\theta_\star}(y)h(\theta^m|\theta_\star)}{p(\theta_m)h_{\theta_m}(y)h(\theta_\star|\theta^m)}.$$

Russian roulette

See the Russian roulette paper of Girolami et al (2013, arxiv) for a valid algorithm for this type of problem. Basically they compute an unbiased estimator of $Z(\theta)^{-1}$ at every iteration.

Note the connection with Bernoulli factories: from unbiased estimates $\hat{Z}_i(\theta)$ of $Z(\theta)$, how do you obtain an unbiased estimate of $\varphi(Z(\theta))$? here $\varphi(z) = 1/z$.

Outline

PMCMC: introduction

PMCMC (Andrieu et al., 2010) is akin to GIMH, except a more complex proposal mechanism is used: a PF (particle filter).
The same remarks will apply:

- Unbiasedness (of the likelihood estimated provided by the PF) is only an intermediate result for establishing the validity of the whole approach.
- Unbiasedness is not enough to give you intuition on the validity of e.g. Particle Gibbs.

Objective

Objectives

Sample from

$$p(\mathrm{d}\theta, \mathrm{d}x_{0:T}|y_{0:T})$$

for a given state-space model.

Why are these models difficult?

Because the likelihood is intractable

$$p_T^\theta(y_{0:T}) = \int \prod_{t=0}^{T} f_t^\theta(y_t|x_t) \prod_{t=1}^{T} p_t^\theta(x_t|x_{t-1}) p_0^\theta(x_0)$$

Feynman-Kac formalism

Taking $\{M_t^\theta, G_t^\theta\}_{t\geq 0}$ so that

- $M_t^\theta(x_{t-1}, \mathrm{d}x_t)$ is a Markov kernel (for fixed θ), with density $m_t^\theta(x_t|x_{t-1})$
- and

$$G_t^\theta(x_{t-1}, x_t) = \frac{f_t^\theta(y_t|x_t)p_t^\theta(x_t|x_{t-1})}{m_t^\theta(x_t|x_{t-1})}$$

we obtain the Feynman-Kac representation associated to a guided PF that approximates the filtering distribution at every time t.

If we take $m_t^\theta(x_t|x_{t-1}) = p_t^\theta(x_t|x_{t-1})$, we recover the bootstrap filter (which does not require to be able to evaluate $p_t^\theta(x_t|x_{t-1})$ pointwise).

Particle filters: pseudo-code

All operations to be performed for all $n \in 1:N$.
At time 0:

(a) Generate $X_0^n \sim M_0^\theta(\mathrm{d}x_0)$.

(b) Compute $w_0^n = G_0^\theta(X_0^n)$, $W_0^n = w_0^n / \sum_{m=1}^N w_0^m$, and $L_0^N = N^{-1} \sum_{n=1}^N w_0^n$.

Recursively, for $t = 1, \ldots, T$:

(a) Generate ancestor variables $A_t^n \in 1:N$ independently from $\mathcal{M}(W_{t-1}^{1:N})$.

(b) Generate $X_t^n \sim M_t^\theta(X_{t-1}^{A_t^n}, \mathrm{d}x_t)$.

(c) Compute $w_t^n = G_t^\theta(x_{t-1}, x_t)$, $W_t^n = w_t^n / \sum_{m=1}^N w_t^m$, and $L_t^N(\theta) = L_{t-1}^N(\theta) \times \{N^{-1} \sum_{n=1}^N w_t^n\}$.

Unbiased likelihood estimator

A by-product of PF output is that

$$L_T^N(\theta) = \left(\frac{1}{N} \sum_{n=1}^{N} G_0^\theta(X_0^n) \right) \prod_{t=1}^{T} \left(\frac{1}{N} \sum_{n=1}^{N} G_t^\theta(x_{t-1}, x_t) \right)$$

is an *unbiased* estimator of the likelihood $L_T(\theta) = p(y_{0:T}|\theta)$.

(Not trivial, see e.g Proposition 7.4.1 in Pierre Del Moral's book.)

PMCMC

Breakthrough paper of Andrieu et al. (2011), based on the unbiasedness of the PF estimate of the likelihood.

Marginal PMCMC

From current point θ_m (and current PF estimate $L_T^N(\theta_m)$):

1. Sample $\theta_\star \sim H(\theta_m, \mathrm{d}\theta_\star)$
2. Run a PF so as to obtain $L_T^N(\theta_\star)$, an unbiased estimate of $L_T(\theta_\star) = p(y_{0:T}|\theta_\star)$.
3. With probability $1 \wedge r$, set $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$ with

$$r = \frac{p(\theta_\star) L_T^N(\theta_\star) h(\theta_m|\theta_\star)}{p(\theta_m) L_T^N(\theta_m) h(\theta_\star|\theta_m)}$$

Validity

Property 1

Let $\psi_{T,\theta}(\mathrm{d}x_{0:T}^{1:N}, \mathrm{d}a_{1:T}^{1:N})$ be the joint dist' of all the the rv's generated by a PF (for fixed θ), then

$$\pi_T(\mathrm{d}\theta, \mathrm{d}x_{0:T}^{1:N}, \mathrm{d}a_{1:T}^{1:N}) = \frac{p(\mathrm{d}\theta)}{p(y_{0:T})} \psi_{T,\theta}(\mathrm{d}x_{0:T}^{1:N}, \mathrm{d}a_{1:T}^{1:N}) L_T^N(\theta)$$

is a joint pdf, such that the θ-marginal is $p(\theta|y_{0:T})\mathrm{d}\theta$.

Proof: fix θ, and integrate wrt the other variables:

$$\begin{aligned} \int \pi_T(\cdot) &= \frac{p(\theta)}{p(y_{0:T})} \mathbb{E}\left[L_T^N(\theta)\right] \mathrm{d}\theta \\ &= \frac{p(\theta)p(y_{0:T}|\theta)}{p(y_{0:T})} \mathrm{d}\theta = p(\theta|y_{0:T})\mathrm{d}\theta \end{aligned}$$

More direct proof for $T = 1$

$$\psi_{1,\theta}(\mathrm{d}x_{0:1}^{1:N}, \mathrm{d}a_1^{1:N}) = \prod_{n=1}^{N} M_0^\theta(\mathrm{d}x_0^n) \left\{ \prod_{n=1}^{N} M_1^\theta(x_0^{a_1^n}, \mathrm{d}x_1^n) W_{0,\theta}^{a_1^n} \mathrm{d}a_1^n \right\}$$

with $W_{0,\theta}^n = G_0^\theta(x_0^n) / \sum_{m=1}^N G_0^\theta(x_0^m)$. So

$$\pi_1(\cdot) = \frac{p(\theta)}{p(y_{0:t})} \psi_{1,\theta}(\cdot) \left\{ \frac{1}{N} \sum_{n=1}^{N} G_0^\theta(x_0^n) \right\} \left\{ \frac{1}{N} \sum_{n=1}^{N} G_1^\theta(x_0^{a_1^n}, x_1^n) \right\}$$

$$= \frac{p(\theta)}{N^2 p(y_{0:t})} \sum_{n=1}^{N} G_1^\theta(x_0^{a_1^n}, x_1^n) M_1^\theta(x_0^{a_1^n}, x_1^n) \frac{G_0^\theta(x_0^{a_1^n})}{\sum_{m=1}^N G_0^\theta(x_0^m)} \left\{ \sum_{m=1}^{N} G_0^\theta(x_0^m) \right\}$$

$$\times M_0^\theta(\mathrm{d}x_0^{a_1^n}) \left\{ \prod_{i \neq a_1^n} M_0^\theta(\mathrm{d}x_0^i) \right\} \left\{ \prod_{i \neq n} M_1^\theta(x_0^{a_1^i}, \mathrm{d}x_1^i) W_1^{a_1^i} \mathrm{d}a_1^i \right\}$$

Interpretation

$$\pi_1(\mathrm{d}\theta, \mathrm{d}x_{0:1}^{1:N}, \mathrm{d}a_1^{1:N}) = \frac{1}{N} \times \left[\frac{1}{N} \sum_{n=1}^{N} p(\mathrm{d}\theta, \mathrm{d}x_0^{a_1^n}, \mathrm{d}x_1^n | y_{0:1}) \right.$$
$$\left. \prod_{i \neq a_1^n} M_0^\theta(\mathrm{d}x_0^i) \left\{ \prod_{i \neq n} M_1^\theta(x_0^{a_1^i}, \mathrm{d}x_1^i) W_0^{a_1^i} \right\} \right]$$

which is a mixture distribution, with probability $1/N$ that path n follows $p(\mathrm{d}\theta, \mathrm{d}x_{0:1}|y_{0:1})$, A_1^n is Uniform in $1:N$, and other paths follows a conditional SMC distribution (the distribution of a particle filter conditional on one trajectory being fixed). From this calculation, one easily deduce the unbiasedness property (directly!) but also properties similar to those of the GIMH.

Additional properties (similar to GIMH)

Property 2

Marginal PMCMC is a Metropolis sampler with invariant distribution π_T, and proposal distribution $h(\theta_\star|\theta)\mathrm{d}\theta_\star\psi_{T,\theta_\star}(\cdot)$. (In particular, it leaves invariant the posterior $p(\mathrm{d}\theta|y_{0:T})$.)

Proof: write the MH ratio, same type of cancellations as for GIMH.

Additional properties (similar to GIMH)

Property 3

If we extend π_T by adding component $k \in 1:N$ with conditional probability $\propto W_T^k$, then the joint pdf $\pi_T(\mathrm{d}\theta, \mathrm{d}x_{0:T}^{1:N}, \mathrm{d}a_{1:T-1}^{1:N}, \mathrm{d}k)$ is such that

(a) $(\theta, X_{0:T}^\star) \sim p(\mathrm{d}\theta, \mathrm{d}x_{0:T}|y_{0:T})$ marginally; and

(b) Given $(\theta, X_{0:T}^\star)$, the $N-1$ remaining trajectories follow the conditional SMC distribution.

where $X_{0:T}^\star$ is the $k-$th *complete* trajectory: $X_t^\star = X_t^{B_t}$ for all t, with $B_T = k$, $B_{T-1} = A_T^k$, ... $B_0 = A_1^{B_1}$.

Outline

Don't listen to Jeff!

Proposal: Gaussian random walk, variance Σ.
Naive approach:

- Fix N
- target acceptance rate 0.234

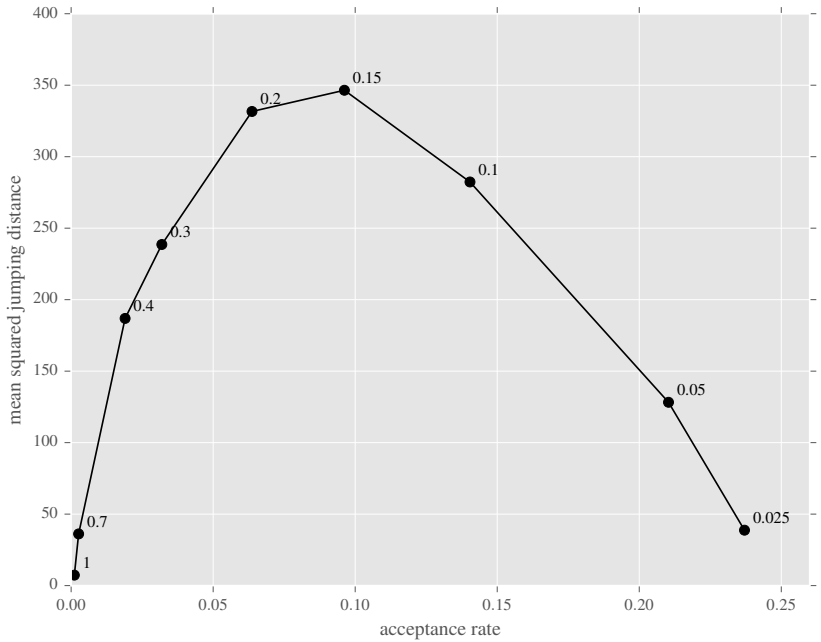


Figure: Acceptance rate vs N, when $\Sigma = \tau I_3$, and τ varies, PMMH for a toy linear Gaussian model

Recommended approach

- Through pilot runs, try to find N such that variance of log-likelihood estimate is $<< 1$;
- Then calibrate in order to minimise the SJD (squared jumping distance) or some other criterion;
- "Best" acceptance rate will be $<< 0.234$.
- Adaptative MCMC is kind of dangerous in this context; consider SMC^2 instead.

Also: state-space model likelihoods are nasty

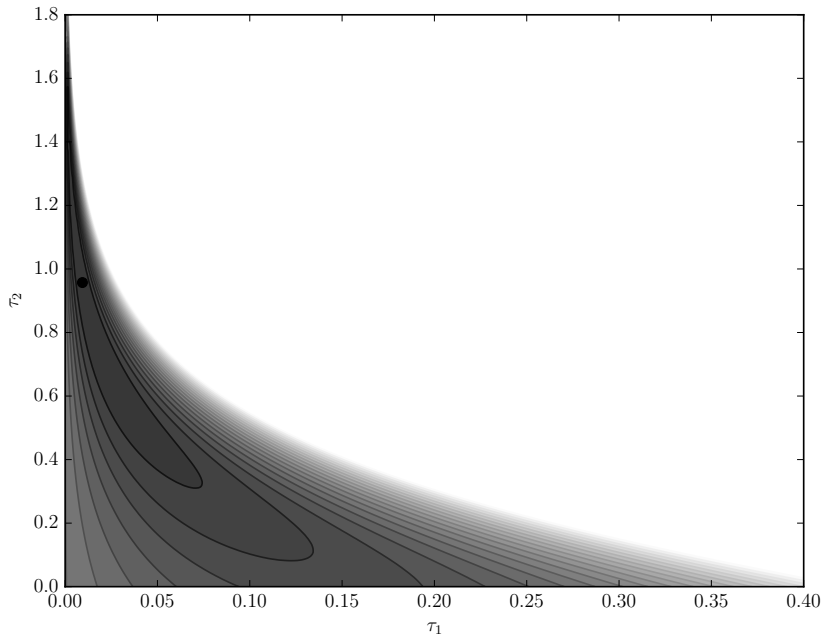


Figure: Log-likelihood contour for nutria data and Ricker state-space model (third parameter is fixed).

Outline

CSMC

- The formalisation of PMCMC offers the possibility to regenerate the $N-1$ trajectories that have not been selected; this is essentially a Gibbs step, conditional on θ, and the selected trajectory $X^{\star}_{0:T}$.
- This CSMC step cannot be analysed with the same tools as marginal PMCMC, as in Andrieu and Vihola (2012).

From now on, we drop θ from the notations.

Algorithmic description ($T = 1$)

Assume selected trajectory is $X_{0:1}^{\star} = (X_0^1, X_1^1)$; i.e. $k = 1$, $A_1^k = 1$.
At time $t = 0$:

(a) sample $X_0^n \sim M_0(\mathrm{d}x_0)$ for $n \in 2 : N$.

(b) Compute weights $w_0^n = G_0(X_0^n)$ and normalise, $W_0^n = w_0^n / \sum_{m=1}^N w_0^m$.

At time $t = 1$:

(a) Sample $A_1^{2:N}$ $\mathcal{M}(W_0^{1:N})$.

(b) Sample $X_1^n \sim M_1(X_1^{A_0^n}, \mathrm{d}x_1)$ for $n \in 2 : N$.

(c) Compute weights $w_1^n = G_1(X_0^{A_1^n}, X_1^n)$ and normalise, $W_1^n = w_1^n / \sum_{m=1}^N w_1^m$.

(d) select new trajectory k with probability W_1^k.

then return $\tilde{X}_{0:1}^{\star} = (X_0^{A_1^k}, X_1^k)$.

Some remarks

- One may show that the CSMC update does not depend on the labels of the frozen trajectory. This is why we set these arbitrarily to $(1, \ldots, 1)$. Formally, this means that the CSMC kernel is such that $K^N_{\mathrm{CSMC}} : \mathcal{X}^T \to \mathcal{P}(\mathcal{X}^T)$.
- This remains true for other resampling schemes (than multinomial); see next two* slides for an example

Properties of the CSMC kernel

Theorem

Under appropriate conditions, one has, for any $\varepsilon > 0$,

$$\left| K^N_{\mathrm{CSMC}}(\varphi)(x_{0:T}) - K^N_{\mathrm{CSMC}}(\varphi)(x'_{0:T}) \right| \leq \varepsilon$$

for N large enough, and $\varphi : \mathcal{X}^T \to [-1, 1]$.

This implies uniform ergodicity. Proof based on a coupling construction.

Assumptions

- G_t is upper bounded, $G_t(x_t) \leq g_t$.
- We have

$$\int M_0(dx_0) G_0(x_0) \geq \frac{1}{g_0}, \quad \int M_t(x_{t-1}, dx_t) G_t(x_t) \geq \frac{1}{g_t}$$

But no assumptions on the kernels M_t.

Backward sampling

Nick Whiteley (in his RSS discussion of PMCMC) suggested to add an extra *backward* step to CSMC, where one tries to modify (recursively, backward in time) the ancestry of the selected trajectory.
In our $T = 1$ example, and for multinomial resampling, this amounts to draw A_1^k from

$$\mathbb{P}(A_1^k = a | k, x_{0:1}^{1:N}) \propto W_0^a m_1(x_1^k | x_0^a)$$

where $m_1(x_1^k | x_0^a)$ is the PDF at point x_1^k of $M_1(x_0^a, dx_1)$, then return $x_{0:1}^\star = (x_0^a, x_1^k)$.

BS for other resampling schemes

More generally, BS amounts to draw a_1^k from

$$P(a_1^k = a|k, x_{1:2}^{1:N}) \propto \rho_1(W_1^{1:N}; a_1^k = a|a_1^{-k}) m_2(x_1^a, x_2^k)$$

where a_1^{-k} is $a_1^{1:N}$ minus a_1^k.

So we need to be able the conditional probability $\rho_1(W_1^{1:N}; a_1^k = a|a_1^{-k})$ for alternative resampling schemes.

Why BS would bring an improvement?

C. and Singh (2014) prove that CSMC+BS dominates CSMC in efficiency ordering (i.e. asymptotic variance). To do so, they prove that these two kernels are reversible; see Tierney (1998), Mira & Geyer (1999).

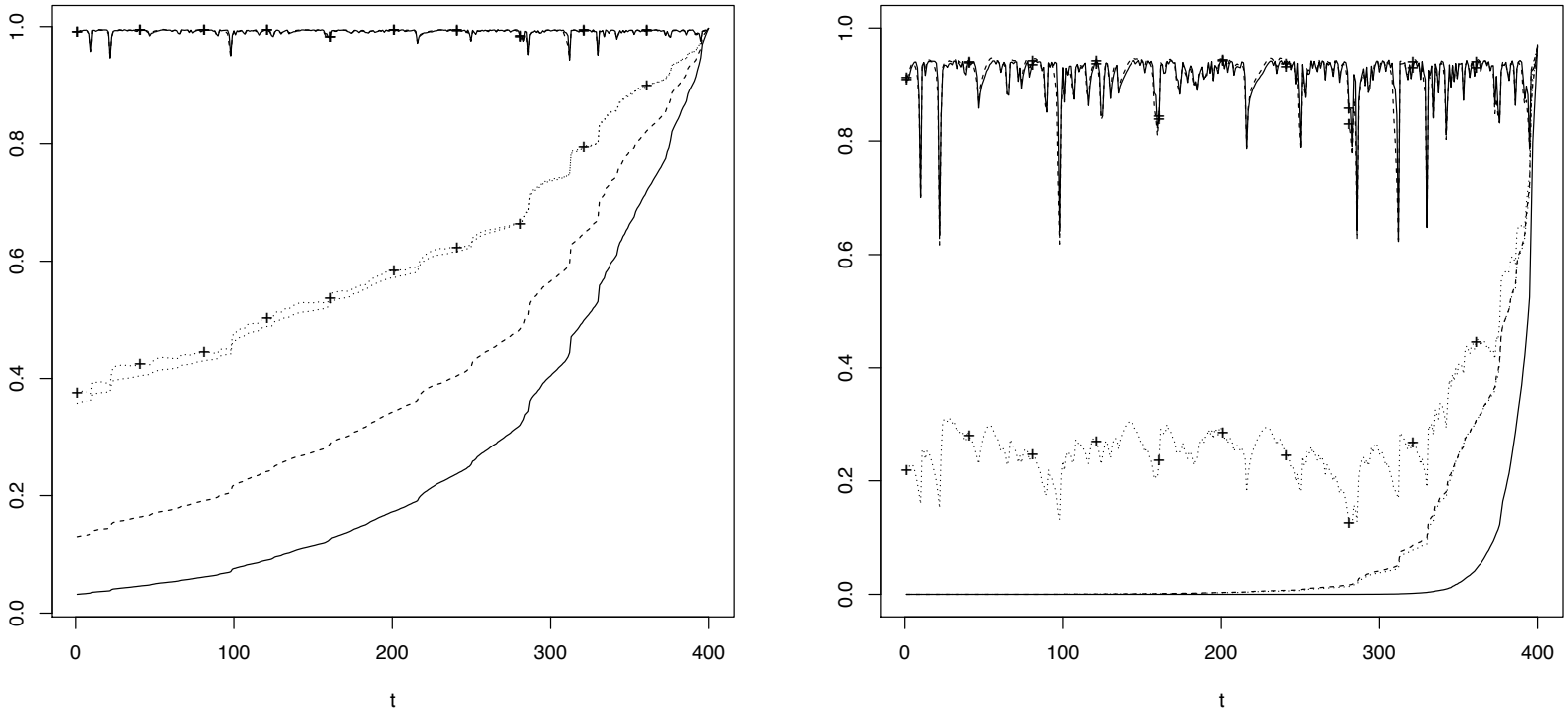
Simulations

See the plots in next slide, based on the following simple state-space model, with $\theta = (\mu, \phi, \sigma)$:

$$x_t - \mu = \phi(x_{t-1} - \mu) + \sigma\epsilon_t, \quad \epsilon_t \sim N(0,1)$$

$$y_t | x_t \sim \text{Poisson}(e^{x_t})$$

Update rate of X_t



Left: $N = 200$, right: $N = 20$. Solid line: multinomial, Dashed line: residual; Dotted line: Systematic. Crosses mean BS has been used.

Conclusion

- When the backward step is possible, it should be implemented, because it improves mixing dramatically. In that case, multinomial resampling is good enough.
- When the backward step cannot be implemented, switching to systematic resampling helps.

But what's the point of PG?

It's a bit the same discussion as marginal Metropolis (in θ-space) versus Gibbs:

- Gibbs does not work so well when they are strong correlations (here between θ and $X^{\star}_{0:T}$);
- Metropolis requires a good proposal to work well.

In some cases, combining the two is helpful: in this way, the CSMC update will refresh the particle system, which may help to get "unstuck".