Sequential Monte Carlo opens up for nonlinear system identification

Thomas Schön

Seminar at the Division of Automatic Control, Linköping University, Oct. 23, 2014.
Joint work with

Local team:

- Fredrik Lindsten
- Johan Dahlin
- Johan Wågberg
- Christian A. Naesseth
- Andreas Svensson
- Liang Dai
Nonlinear system identification

A state space model (SSM) consists of a Markov process \( \{x_t\}_{t \geq 1} \) that is indirectly observed via a measurement process \( \{y_t\}_{t \geq 1} \),

\[
\begin{align*}
  x_{t+1} | x_t & \sim f_{\theta,t}(x_{t+1} | x_t, u_t), \\
  y_t | x_t & \sim g_{\theta,t}(y_t | x_t, u_t), \\
  x_1 & \sim \mu_{\theta}(x_1), \\
  (\theta & \sim \pi(\theta)).
\end{align*}
\]

We observe

\[
y_{1:T} \triangleq \{y_1, \ldots, y_T\}, \quad \text{and possibly } u_{1:T} \triangleq \{u_1, \ldots, u_T\}.
\]

(Leaving the latent variables \( x_{1:T} \) unobserved).

**Identification problem:** Find \( \theta \) based on \( y_{1:T} \) (and \( u_{1:T} \)).
Two commonly used problem formulations

**Maximum likelihood (ML) formulation** – model the unknown parameters as a deterministic variable and solve

\[ \hat{\theta}_{\text{ML}} = \arg \max_{\theta \in \Theta} p_{\theta}(y_{1:T}). \]

**Bayesian formulation** – model the unknown parameters as a random variable \( \theta \sim \pi(\theta) \) and compute

\[ p(\theta \mid y_{1:T}) = \frac{p_{\theta}(y_{1:T})\pi(\theta)}{p(y_{1:T})}, \]

where \( p_{\theta}(y_{1:T}) = p(y_{1:T} \mid \theta) \).
Central object – the likelihood

The central object in both formulations is the likelihood

\[ p_\theta(y_{1:T}) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{1:t-1}). \]

The likelihood is computed by marginalizing the joint density \( p_\theta(x_{1:T}, y_{1:T}) \) w.r.t. the state sequence \( x_{1:T} \)

\[ p_\theta(y_{1:T}) = \int p_\theta(x_{1:T}, y_{1:T}) dx_{1:T} = \prod_{t=1}^{T} \int g_\theta(y_t \mid x_t)p_\theta(x_t \mid y_{1:t-1}) dx_t. \]

**Key challenge:** How to deal with the latent states.

**Our solution:** Sequential Monte Carlo (SMC) including particle filters/smoothers.
One realisation from \( x[k + 1] = 0.8x[k] + v[k] \) where \( v[k] \sim \mathcal{N}(0, 1) \). Initialise in \( x[0] = -40 \).

This will eventually generate samples from the following stationary distribution:

\[
\pi^s(x) = \mathcal{N} \left( x \left| 0, \frac{1}{1 - 0.8^2} \right. \right)
\]

as \( t \to \infty \).
The true stationary distribution is showed in black and the empirical histogram obtained by simulating the Markov chain $x[k + 1] = 0.8x[k] + v[k]$ is plotted in gray.

The initial 1000 samples are discarded (burn-in).
In the example, the Markov chain was fully specified and the stationary distribution could be expressed in closed form.

Not possible in the situations we are interested in, but we can (since 2010) find a Markov chain that has the target distribution (e.g. $p(\theta | y_{1:T})$) as its stationary distribution.

Two constructive ways of doing this are:

1. Metropolis Hastings (MH) algorithm
2. Gibbs sampling

Markov chain Monte Carlo (MCMC) methods allow us to generate samples from a target distribution by simulating a Markov chain which has the target distribution as its stationary distribution.
Outline

1. Problem formulation
2. Micro – MCMC
3. Sketching identification strategies for nonlinear SSMs
   a. Marginalization
   b. Data augmentation
4. Sequential Monte Carlo (SMC)
5. Using SMC as a proposal mechanism within MCMC
6. A nontrivial example
7. The nonlinear SSM is just a special case...
Identification strategies

The two identification strategies we are concerned with are:

- **Marginalization** Deal with the states by marginalizing them out.
- **Data augmentation** Deal with the states by treating them as auxiliary variables to be estimated along with the parameters.

<table>
<thead>
<tr>
<th></th>
<th>Marginalization</th>
<th>Data augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Direct optimization</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Metropolis Hastings</td>
<td>Gibbs sampling</td>
</tr>
</tbody>
</table>
Identification strategy – marginalization

Deal with the states by marginalizing them out.

1. **Direct optimization** work directly with the optimization problem

\[
\hat{\theta}_{\text{ML}} = \arg \max_{\theta \in \Theta} \prod_{t=1}^{T} \int g_{\theta}(y_{t} | x_{t}) p_{\theta}(x_{t} | y_{1:t-1}) dx_{t}.
\]

Cannot be solved in closed form, use iterative numerical methods

\[
\theta_{k+1} = \theta_{k} + \alpha_{k}s_{k}.
\]

The search direction is typically computed according to

\[
s_{k} = H_{k}g_{k}, \quad g_{k} = \nabla_{\theta} p_{\theta}(y_{1:T}) \bigg|_{\theta=\theta_{k}}.
\]

**SMC** used to approximate the cost function and its derivative(s).
Identification strategy – marginalization

2. **Metropolis Hastings (MH)** is an MCMC method that produce a sequence of random variables \( \{\theta[m]\}_{m \geq 1} \) by iterating

1. Propose a new sample \( \theta' \)

\[
\theta' \sim q(\cdot | \theta[m]).
\]

2. Accept the new sample with probability

\[
\alpha = \min \left( 1, \frac{p_{\theta'}(y_1:T)\pi(\theta')}{p_{\theta[m]}(y_1:T)\pi(\theta[m])} \frac{q(\theta[m] | \theta')}{q(\theta' | \theta[m])} \right)
\]

The above procedure results in a Markov chain \( \{\theta[m]\}_{m \geq 1} \) with \( p(\theta | y_T) \) as its stationary distribution!

- **SMC** used to approximate the likelihood \( p_{\theta}(y_1:T) \) in the acceptance probability.
Identification strategy – data augmentation

Deal with the states by treating them as auxiliary variables to be estimated along with the parameters.

Intuitively: Alternate between updating $\theta$ and $x_{1:T}$.

1. **Expectation Maximization (EM)**

**(E)** Compute a conditional expectation

$$Q(\theta, \theta[k]) \triangleq \int \log p_{\theta}(x_{1:T}, y_{1:T}) p_{\theta[k]}(x_{1:T} | y_{1:T}) \, dx_{1:T}.$$ 

**(M)** Maximize $Q(\theta, \theta[k])$ w.r.t. $\theta$

$$\theta[k + 1] = \arg \max_{\theta} Q(\theta, \theta[k]).$$

**SMC** is used to approximate the JSD $p_{\theta[k]}(x_{1:T} | y_{1:T})$. 
2. **Gibbs sampling** aim at compute \( p(\theta, x_{1:T} \mid y_{1:T}) \).

Gibbs sampling (blocked) for SSMs amounts to iterating

- Draw \( \theta[m] \sim p(\theta \mid x_{1:T}[m-1], y_{1:T}) \),
- Draw \( x_{1:T}[m] \sim p(x_{1:T} \mid \theta[m], y_{1:T}) \).

The above procedure results in a Markov chain,

\[
\{ \theta[m], x_{1:T}[m] \}_{m \geq 1}
\]

with \( p(\theta, x_{1:T} \mid y_T) \) as its stationary distribution!

**SMC** is used to generate a state sequence \( x_{1:T}[m] \) from \( p(x_{1:T} \mid \theta[m], y_{1:T}) \).
Outline

1. Problem formulation
2. Micro – MCMC
3. Sketching identification strategies for nonlinear SSMs
   a. Marginalization
   b. Data augmentation
4. **Sequential Monte Carlo (SMC)**
5. Using SMC as a proposal mechanism within MCMC
6. A nontrivial example
7. The nonlinear SSM is just a special case...
Sequential Monte Carlo (SMC)

The particle filter provides an approximation \( p(x_{1:t} | y_{1:t}) \), when the state evolves according to an SSM,

\[
\begin{align*}
    x_{t+1} | x_t & \sim f_\theta(x_{t+1} | x_t), \\
    y_t | x_t & \sim g_\theta(y_t | x_t), \\
    x_1 & \sim \mu_\theta(x_1).
\end{align*}
\]

The particle filter maintains an empirical distribution made up \( N \) samples (particles) and corresponding weights

\[
\tilde{p}(x_{1:t} | y_{1:t}) = \sum_{i=1}^{N} w_t^i \delta_{x_{1:t}^i} (x_{1:t}).
\]

“Think of each particle as one simulation of the system state. Only keep the good ones.”
Particle filter

SMC = resampling + sequential importance sampling

1. Resampling: \( \mathbb{P}(a^i_t = j) = \bar{w}^j_{t-1} / \sum_l \bar{w}^l_{t-1} \).

2. Propagation: \( x^i_t \sim f_\theta(x_t | x^{a^i_t}_{1:t-1}) \) and \( x^i_{1:t} = \{x^{a^i_t}_{1:t-1}, x^i_t\} \).

3. Weighting: \( \bar{w}^i_t = W_t(x^i_t) = g_\theta(y_t | x_t) \).

The **ancestor indices** \( \{a^i_t\}_{i=1}^N \) are very **useful** auxiliary variables! They make the stochasticity of the resampling step explicit.
Sequential Monte Carlo (SMC)

Let

\[ x_t \triangleq \{ x^1_t, \ldots, x^N_t \}, \quad a_t \triangleq \{ a^1_t, \ldots, a^N_t \} \]

denote all particles and ancestor indices generated at time \( t \).

The SMC algorithm generates a single realization of a collection of random variables

\[ \{ x_{1:T}, a_{2:T} \} \in X^{NT} \times \{ 1, \ldots, N \}^{N(T-1)} \]

distributed according to

\[
\psi(x_{1:T}, a_{2:T}) \triangleq \prod_{i=1}^{N} q_1(x^i_1) \prod_{t=2}^{T} \prod_{i=1}^{N} M_t(a^i_t, x^i_t),
\]

where

\[
M_t(a_t, x_t) = \frac{\bar{w}^{a_t}_{t-1}}{\sum_l \bar{w}^l_{t-1}} f_t(x_t \mid x^a_{1:t-1}).
\]
The particle system degenerates (illustration)

Clearly motivates the need for particle smoothers.

Self-contained introduction to particle smoothing using BS and AS

Using SMC within MCMC (PMCMC)

Particle MCMC (PMCMC) is a systematic way of combining SMC and MCMC.

**Intuitively:** SMC is used as a high-dimensional proposal mechanism on the space of state trajectories $X^T$.

**A bit more precise:** Construct a Markov chain with $p(\theta | y_{1:T})$ as its stationary distribution.

Pioneered by the work

Reminder – identification strategies

<table>
<thead>
<tr>
<th></th>
<th>Marginalization</th>
<th>Data augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Bayesian</td>
<td>Direct optimization</td>
<td>Expectation</td>
</tr>
<tr>
<td></td>
<td>Metropolis Hastings</td>
<td>Metropolis Hastings</td>
</tr>
</tbody>
</table>

Iterating the two steps below will result in a Markov chain \( \{\theta[m]\}_{m \geq 1} \) with \( p(\theta | y_T) \) as its stationary distribution.

1. Propose a new sample \( \theta' \) according to \( \theta' \sim q(\cdot | \theta[m]) \).
2. Accept the new sample with probability

\[
\alpha = \min \left( 1, \frac{p_{\theta'}(y_{1:T}) \pi(\theta')}{p_{\theta[m]}(y_{1:T}) \pi(\theta[m])} \frac{q(\theta[m] | \theta')}{q(\theta' | \theta[m])} \right)
\]
Fact (non-trivial): SMC produce an unbiased estimate of the likelihood!

\[ \hat{p}_\theta(y_{1:T}) = \hat{p}_\theta(y_1) \prod_{t=2}^{T} \hat{p}_\theta(y_t \mid y_{1:t-1}) = \prod_{t=1}^{T} \left( \frac{1}{N} \sum_{i=1}^{N} \bar{w}_t^i \right). \]

Intuitive idea: What about using this estimate within MH?!
The extended target distribution

Introduce an auxiliary variable

\[ u = (x_{1:T}, a_{2:T}), \quad u \sim \psi(u | \theta). \]

Note that,

\[ p(\theta, u | y_{1:T}) = \frac{p_{\theta,u}(y_{1:T})\psi(u | \theta)p(\theta)}{p(y_{1:T})} = \frac{p_{\theta,u}(y_{1:T})\psi(u | \theta)p(\theta | y_{1:T})}{p(y_{1:T} | \theta)}. \]

**Non-trivial construction:** Consider the following extended target distribution

\[ \phi(\theta, u) = \frac{\hat{p}_{\theta,u}(y_{1:T})\psi(u | \theta)p(\theta | y_{1:T})}{p_{\theta}(y_{1:T})}, \]

defined on \( \Theta \times X^{NT} \times \{1, \ldots, N\}^{N(T-1)}. \)
Marginalization

Marginalize (recall strategy) out the auxiliary variables \( u \)

\[
\int \phi(\theta, u) du = \frac{p(\theta | y_{1:T})}{p_\theta(y_{1:T})} \int \hat{p}_{\theta,u}(y_{1:T}) \psi(u | \theta) du.
\]

What can we do about the integral?

SMC produce an unbiased estimate of \( \hat{p}_{\theta,u}(y_{1:T}) \)

\[
E_u | \theta [\hat{p}_{\theta,u}(y_{1:T})] = \int \hat{p}_{\theta,u}(y_{1:T}) \psi(u | \theta) du = p_\theta(y_{1:T}),
\]

**Result:** \( p(\theta | y_{1:T}) \) is recovered **exactly** as the marginal of the extended target distribution \( \phi(\theta, u) \), despite the fact that we employ an SMC **approximation** of the likelihood using a finite number of particles \( N \).
Particle Metropolis Hastings (PMH)

Based on the current sample \((\theta[m], u[m])\) a new sample \((\theta', u')\) is proposed according to

\[
\theta' \sim q(\cdot \mid \theta[m], u[m]), \quad u' \sim \psi(\cdot \mid \theta').
\]

The probability of accepting this sample is given by

\[
\alpha = \min \left( 1, \frac{\hat{p}_{\theta'[m],u'[m]}(y_{1:T})p(\theta')}{\hat{p}_{\theta[m],u[m]}(y_{1:T})p(\theta[m])} \frac{q(\theta[m] \mid \theta', u')}{q(\theta' \mid \theta[m], u[m])} \right).
\]

Note: Very importantly, \(\alpha\) does not require evaluation of \(\psi(u \mid \theta')\)!

Originally appeared in (different derivation)


and further studied in,

Johan Dahlin, Fredrik Lindsten and Thomas B. Schön, Particle Metropolis Hastings using gradient and Hessian information, Statistics and Computing, 2014. (accepted for publication)
Example – semiparametric Wiener model

Parametric LGSS and a nonparametric static nonlinearity:

\[
\begin{align*}
\begin{bmatrix} x_{t+1} \\ z_t \\ y_t \\ e_t \\ v_t \\
\end{bmatrix} &= 
\begin{bmatrix} A & B \\ \Gamma \\
\end{bmatrix} 
\begin{bmatrix} x_t \\ u_t \\
\end{bmatrix} + 
\begin{bmatrix} v_t \\ e_t \\
\end{bmatrix}, \\
\end{align*}
\]

\( v_t \sim \mathcal{N}(0, Q), \) \\
\( e_t \sim \mathcal{N}(0, R). \)
Example – semiparametric Wiener model

“Parameters”: \( \theta = \{A, B, Q, g(\cdot), r\} \).

**Bayesian model** specified by priors

- Conjugate priors for \( \Gamma = [A \ B] \), \( Q \) and \( r \),
  - \( p(\Gamma, Q) = \text{Matrix-normal inverse-Wishart} \)
  - \( p(r) = \text{inverse-Wishart} \)
- Gaussian process prior on \( g(\cdot) \),
  \[
g(\cdot) \sim \mathcal{GP}(z, k(z, z')) \]

**Inference** using PGAS with \( N = 15 \) particles. \( T = 1000 \) measurements. We ran 15 000 MCMC iterations and discarded 5 000 as burn-in.
**Example – semiparametric Wiener model**

Show movie

Bode diagram of the 4th-order linear system. Estimated mean (dashed black), true (solid black) and 99% credibility intervals (blue).

Static nonlinearity (non-monotonic), estimated mean (dashed black), true (black) and the 99% credibility intervals (blue).

The nonlinear SSM is just a special case...

A **graphical model** is a probabilistic model where a graph \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \) represents the conditional independency structure between random variables,

1. a set of **vertices** \( \mathcal{V} \) (nodes) represents the random variables
2. a set of **edges** \( \mathcal{E} \) containing elements \((i, j) \in \mathcal{E}\) connecting a pair of nodes \((i, j) \in \mathcal{V}\)

\[
p(x_{0:T}, y_{1:T}) = p(x_0) \prod_{t=1}^{N} p(x_t \mid x_{t-1}) \prod_{t=1}^{N} p(y_t \mid x_t).
\]
The nonlinear SSM is just a special case...

SMC samplers are used to approximate a sequence of probability distributions on a sequence of probability spaces.

Constructing an artificial sequence of intermediate target distributions for an SMC sampler is a powerful (and quite possibly underutilized) idea.


Conclusion

1. Overview of identification strategies for nonlinear SSMs.
2. Focused on **marginalization** today, where we made use of the unbiased likelihood estimate \( \hat{p}_\theta(y_{1:T}) \) from SMC within MH.
3. Powerful tools useful also outside the class of nonlinear SSMs.

**A lot of interesting research that remains to be done!!**

Information about a PhD course (*Computational learning in dynamical systems*) on the topic is available via

user.it.uu.se/~thosc112/CIDS.html

Manuscript is also available (ask me for a draft if you want)

References to some of our work

Self-contained introduction to particle smoothing using BS and AS

ML identification of nonlinear SSMs

PMCMC for Bayesian identification of nonlinear SSMs (and more)

SMC methods for graphical models

Seminar: http://www.newton.ac.uk/seminar/20140425104011151