Sequential Monte Carlo methods for graphical models

“Standard SMC samplers using a non-standard construction of the intermediate target distributions”

Thomas Schöhn
Division of Systems and Control
Department of Information Technology
Uppsala University, Sweden

Ongoing joint work with: Christian A. Naesseth (Linköping University) and Fredrik Lindsten (University of Cambridge).

Thomas Schöhn (user.it.uu.se/~thosc112), Sequential Monte Carlo methods for graphical models
A **graphical model** is a probabilistic model where a graph \( G = (\mathcal{V}, \mathcal{E}) \) 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 \((i, j) \in \mathcal{V}\)

\[
p(x_0:T, y_{1:T}) = p(x_0) \prod_{t=1}^{N} p(x_t \mid x_{t-1}) \prod_{t=1}^{N} p(y_t \mid x_t).
\]
For an undirected graphical model (Markov random field), the joint PDF over all the involved random variables is

\[ p(X_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C), \]

where \( \mathcal{C} \) is the set of cliques in \( \mathcal{G} \).

**Undirected graph**

**Factor graph** making interactions explicit.
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.

**Key idea:** Perform and make use of a **sequential decomposition** of the graphical model.

Using this SMC sampler within a particle MCMC sampler allows us to construct high-dimensional MCMC kernels for graphical models.
1. Example – from information theory
2. SMC for general graphical models
3. Particle MCMC
   a) Micro module – the idea underlying PGAS
   b) Partial blocking
4. Example – Markov random field
5. Conclusions

“Standard SMC samplers using a non-standard construction of the intermediate target distributions.”
Consider a 2D binary-input channel that is *constrained* in the sense that no two horizontally or vertically adjacent variables may be both be equal to 1.

```
... ... ... ... ...
... 0 1 0 ...
... 0 0 1 ...
... 0 1 0 ...
... ... ... ...
```

Such channels are of interest in magnetic and optical storage.

The channel can be described by a square lattice *undirected graphical model*. 
The variables are binary $x_{\ell,j} \in \{0, 1\}$ and the interactions are pair-wise between adjacent variables. Factors:

$$
\psi(x_{\ell,j}, x_{m,n}) = \begin{cases} 
0, & x_{\ell,j} = x_{m,n} = 1 \\
1, & \text{otherwise}
\end{cases}
$$
The resulting joint PDF is given by

\[ p(X_V) = \frac{1}{Z} \prod_{(j,mn) \in E} \psi(x_{\ell,j}, x_{m,n}). \]

For a channel of dimension \( M \times M \) we can write the finite-size noiseless capacity as

\[ C_M = \frac{1}{M^2} \log_2 Z. \]

Unfortunately calculating \( Z \) exactly for these types of models is computationally prohibitive, since the complexity is exponential in the number of variables \( M^2 \).

The example is borrowed from:


Thomas Schön (user.it.uu.se/~thosc112), *Sequential Monte Carlo methods for graphical models*  
Rewrite the GM as a high-dimensional undirected chain by introducing a new set of variables $x_k$.

\[
\psi(x_k) = \prod_{j=1}^{M-1} \psi(x_{j+1,k}, x_{j,k}),
\]

\[
\psi(x_k, x_{k-1}) = \prod_{j=1}^{M} \psi(x_{j,k}, x_{j,k-1}).
\]
The undirected chain

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \rightarrow x_6 \]

results in the following joint PDF

\[
p(X_V) = \frac{1}{Z} \prod_{k=1}^{M} \psi(x_k) \prod_{k=2}^{M} \psi(x_k, x_{k-1}),
\]

which provides a very natural sequence of target distributions for an SMC sampler! Indeed, SMC methods are designed to sample sequentially from a sequence of target distributions \( \gamma_k(x_{1:k}) \),

\[
\gamma_1(x_1) = \psi(x_1),
\]

\[
\gamma_k(x_{1:k}) = \gamma_{k-1}(x_{1:k-1}) \psi(x_k) \psi(x_k, x_{k-1}).
\]
2D channel capacity – 60 × 60 example


For the 2D channel we (Christian) have implement a rather efficient **fully adapted** SMC sampler. We also use FFBS for the “columns”.

This was just a special case, the important question is, can we do this for a general graphical model?! **Yes!**
Constructing an artificial sequence of intermediate target distributions for an SMC sampler is a powerful (and quite possibly underutilized) idea. For some applications, see e.g.,


**Key idea:** Perform and make use of a sequential decomposition of the graphical model.

Defines a sequence of intermediate (auxiliary) target distributions defined on an increasing sequence of probability spaces.

Target this sequence using SMC.
The joint PDF of the set of random variables indexed by $\mathcal{V}$,

$$X_{\mathcal{V}} \triangleq \{x_1, \ldots, x_{|\mathcal{V}|}\}$$

$$p(X_{\mathcal{V}}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C).$$

Sequential decomposition of the above factor graph (the target distributions are built up by adding factors at each iteration),

$$\gamma_1(X_{\mathcal{L}_1}) \quad \gamma_2(X_{\mathcal{L}_2})$$
Sequential decomposition of GMs – equations

Let \( \{ \psi_k \}_{k=1}^{K} \) be a sequence of factors,

\[
\psi_k(X_{\mathcal{I}_k}) = \prod_{C \in \mathcal{C}_k} \psi_C(X_C),
\]

where \( \mathcal{I}_k \subseteq \{1, \ldots, |\mathcal{V}|\} \) is the set of indices in the domain of \( \psi_k \).

The sequential decomposition is based on these factors,

\[
\tilde{\gamma}_k(X_{\mathcal{L}_k}) \triangleq \prod_{\ell=1}^{k} \psi_{\ell}(X_{\mathcal{I}_{\ell}}),
\]

where \( \mathcal{L}_k \triangleq \bigcup_{\ell=1}^{k} \mathcal{I}_\ell \).

By construction, \( \mathcal{L}_K = \mathcal{V} \) and the joint PDF \( p(X_{\mathcal{L}_K}) \propto \tilde{\gamma}_K(X_{\mathcal{L}_K}) \).
Algorithm SMC sampler for graphical models

1. **Initialize** \((k = 1)\): Draw \(X^i_{L_1} \sim r_1(\cdot)\) and set \(w^i_1 = W_1(X^i_{L_1})\).

2. **For** \(k = 2\) **to** \(K\) **do:**
   
   (a) Draw \(a^i_k \sim \text{Cat}(\{w^j_{k-1}\}_{j=1}^N)\).
   
   (b) Draw \(\xi^i_k \sim r_k(\cdot|X^{a^i_k}_{L_{k-1}})\) and set \(X^i_{L_k} = X^{a^i_k}_{L_{k-1}} \cup \xi^i_k\).

   (c) Set \(w^i_k = W_k(X^i_{L_k})\).

Also provides an estimate of the **partition function**!

Again, this is a standard SMC sampler on a non-standard construction of the intermediate target distributions.
Problems with SMC, it is not enough since:

1. It does not solve the parameter learning problem.
2. The quality of the marginals 

\[ p(X_{\mathcal{L}_k}) = \int \tilde{\gamma}_K(X_{\mathcal{L}_K}) dX_{\mathcal{L}_K \setminus \mathcal{L}_k} \]

deteriorates for \( k \ll K \) (particle degeneracy).

(One) solution: Use particle Gibbs with ancestor sampling (PGAS). Allows us to construct high-dimensional MCMC kernels for graphical models.

This allows us to:

1. Simulate, jointly, blocks of variables using an MCMC scheme.
2. Opens up for learning unknown parameters of the model.
Let $x'_{1:K} = (x'_1, \ldots, x'_K)$ be a fixed reference trajectory.

- At each $k$, sample only $N - 1$ particles in the standard way.
- Set the $N^{th}$ particle deterministically: $x^N_k = x'_k$.

CPF causes us to degenerate to the something that is very similar to the reference trajectory, resulting in slow mixing.
CPF vs. CPF-AS – motivation

**Problem with the CPF:** It suffers from slow mixing due to particle degeneracy.

**Solution:** Change the reference trajectory! The resulting algorithm will still degenerate, but it will degenerate to something very different, resulting in better mixing.

This can be achieved by backward simulation, but using *ancestor sampling* we can do this without an explicit backwards pass.

**Implication:** Ancestor sampling opens up for inference in a wider class of models, e.g. non-Markovian SSMs, GMs and BNP models.

Ancestor sampling is conceptually similar to backward simulation, but instead of separate forward and backward sweeps, we achieve the same effect in a **single forward sweep**.
Let $x'_{1:K} = (x'_1, \ldots, x'_K)$ be a fixed *reference trajectory*.

- At each step $k$, sample only $N - 1$ particles in the standard way.
- Set the $N^{th}$ particle deterministically: $x^N_k = x'_k$.
- Generate an artificial history for $x^N_k$ by ancestor sampling.

CPF-AS causes us to degenerate to something that is very different from the reference trajectory, resulting in better mixing.
1. Run CPF-AS(\(x'_{1:K}\)) targeting \(\gamma(x_{1:K} \mid \theta)\).
2. Sample \(x^*_{1:K}\) with \(P(x^*_{1:K} = x^i_{1:K}) \propto w^i_K\).

- Maps \(x'_{1:K}\) stochastically into \(x^*_{1:K}\)
- Implicitly defines an ergodic Markov kernel \((P^N_\theta)\) referred to as the PGAS (particle Gibbs with ancestor sampling) kernel.

**Theorem**

*For any number of particles \(N \geq 2\) and \(\theta \in \Theta\), the PGAS kernel \(P^N_\theta\) is uniformly ergodic and it leaves \(\gamma(x_{1:K} \mid \theta)\) invariant.*
Two extremes of how to sample the variables:

1. Simulate all the latent variables $X_{L_K}$ jointly.
2. Simulate one variable $x_j$ at a time.

With PGAS we can create algorithms that sits in between these two extremes by simulating blocks of variables jointly (partial blocking).

Simulate all the latent variables $X_{L_K}$ jointly.  
Partial blocking via PGAS.  
Simulate one variable $x_j$ at a time.
Let \( \{ \mathcal{V}^m, m \in \{1, \ldots, M\} \} \) be a partition of \( \mathcal{V} \).

We could then (ideally) construct a Gibbs sampler simulating from the conditional distributions

\[
p(X_{\mathcal{V}^m} \mid X_{\mathcal{V} \setminus \mathcal{V}^m}) \propto \prod_{C \in \mathcal{C}^m} \psi_C(X_C), \quad \text{for } m = 1, \ldots, M.
\]

where \( \mathcal{C}^m = \{ C \in \mathcal{C} : C \cap \mathcal{V}^m \neq \emptyset \} \).

\underline{Problem:} It is (in general) impossible to sample from these conditionals.

\underline{(One) solution:} Use PGAS.
Consider a standard square lattice Gaussian MRF of size $10 \times 10$,

$$p(X_\mathcal{V}, Y_\mathcal{V}) \propto \prod_{i \in \mathcal{V}} e^{\frac{1}{2\sigma_i^2} (x_i - y_i)^2} \prod_{(i,j) \in \mathcal{E}} e^{\frac{1}{2\sigma_{ij}^2} (x_i - x_j)^2}$$

with latent variables $X_\mathcal{V} = \{x_1, \ldots, x_{100}\}$ and measurements $Y_\mathcal{V} = \{y_1, \ldots, y_{100}\}$ (simulated with $\sigma_i = 1$ and $\sigma_{ij} = 0.1$).

**Goal:** Compute the posterior distribution $p(X_\mathcal{V} \mid Y_\mathcal{V})$.

We run four MCMC samplers:

1. Standard one-at-a-time Gibbs
2. Tree sampler (Hamze & de Freitas, 2004)
3. PGAS – fully blocked ($N = 50$)
4. PGAS – partially blocked ($N = 50$)
The arrows show the order in which the factors are added.

The two block structures used by the tree sampler and PGAS with partial blocking.

Thomas Schön (user.it.uu.se/~thosc112), *Sequential Monte Carlo methods for graphical models*
The one-step-at-a-time Gibbs sampler is struggling due to the strong interactions.
The tree sampler implements an “ideal” partially blocked Gibbs sampler.
PGAS with partial blocking is an approximation of the tree sampler. Already for relatively few particles we obtain a performance similar to the “ideal” tree sampler.
The fully blocked PGAS performs best, which is not surprising, since it samples all the (dependent) latent variables jointly.

The downside of PGAS is that it is computationally more expensive.
Conclusions

- We have derived SMC-based inference methods for graphical models of arbitrary topologies with discrete or continuous random variables.

- **Key insight:** We exploit a sequential decomposition of the graphical model.

  "Standard SMC samplers using a non-standard construction of the intermediate target distributions"

- Examples involving:
  1. estimating the partition function
  2. inferring the latent variables.

- If you have interesting and challenging problems involving graphical models, let us know!
SMC and PMCMC methods for graphical models


Self-contained introduction to BS and AS (not limited to SSMs)


Particle Gibbs with ancestor sampling (PGAS)


Thank you!!