Ancestor sampling in state space models, graphical models and beyond

“Ancestor sampling is a way of exploiting backward simulation ideas without needing an explicit backward pass.”

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

Joint work with (alphabetical order): Michael I. Jordan (UC Berkeley), Fredrik Lindsten (University of Cambridge) and Christian A. Naesseth (Linköping University).

Thomas Schön (user.it.uu.se/thosc112), Ancestor sampling in state space models, graphical models and beyond
1. Bayesian learning
   a) Problem formulation
   b) Gibbs sampling

2. Sequential Monte Carlo (SMC)

4. Particle Gibbs with ancestor sampling (PGAS)

5. SMC and PGAS for graphical models

The sequential Monte Carlo samplers are fundamental to both the ML and the Bayesian approaches.
Consider a Bayesian SSM (\( \theta \) is now a r.v. with a prior density \( p(\theta) \))

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

**Learning problem:** Compute the posterior \( p(\theta, x_{1:T} \mid y_{1:T}) \), or one of its marginals.

**Key challenge:** There is no closed form expression available.
Gibbs sampler for SSMs (I/II)

**Aim:** Compute \( p(\theta, x_{1:T} \mid y_{1:T}) \).

**MCMC:** 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!
**Aim:** Compute $p(\theta, x_{1:T} \mid y_{1:T})$.

**MCMC:** Gibbs sampling (blocked) for SSMs amounts to iterating

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

**Problem:** $p(x_{1:T} \mid \theta[m], y_{1:T})$ not available!

**Idea:** Approximate $p(x_{1:T} \mid \theta[m], y_{1:T})$ using a sequential Monte Carlo method!
Approximate a **sequence** of probability distributions on a sequence of probability spaces of **increasing dimension**.

Let \( \{ \gamma_{\theta,t}(x_{1:t}) \}_{t \geq 1} \) be a sequence of unnormalized densities and

\[
\tilde{\gamma}_{\theta,t}(x_{1:t}) = \frac{\gamma_{\theta,t}(x_{1:t})}{Z_{\theta,t}}
\]

**Ex.** (SSM)

\[
\tilde{\gamma}_{\theta,t}(x_{1:t}) = p_{\theta}(x_{1:t} \mid y_{1:t}), \quad \gamma_{\theta,t}(x_{1:t}) = p_{\theta}(x_{1:t}, y_{1:t}),
\]

\[
Z_{\theta,t} = p_{\theta}(y_{1:t}).
\]
Consider a toy 1D localization problem.

**Dynamic model:**

\[ x_{t+1} = x_t + u_t + v_t, \]

where \( x_t \) denotes position, \( u_t \) denotes velocity (known), \( v_t \sim \mathcal{N}(0, 5) \) denotes an unknown disturbance.

**Measurements:**

\[ y_t = h(x_t) + e_t. \]

where \( h(\cdot) \) denotes the world model (here the terrain height) and \( e_t \sim \mathcal{N}(0, 1) \) denotes an unknown disturbance.

The same idea has been used for the Swedish fighter JAS 39 Gripen. Details are available in,

Highlights two key capabilities of SMC:

1. Automatically handles an unknown and dynamically changing number of hypotheses.

2. Work with nonlinear/non-Gaussian models.

\[ p(x_t \mid y_{1:t}) \approx \sum_{i=1}^{N} w_t^i \delta_{x_t^i}(x_t) \]
SMC = resampling + sequential importance sampling

1. **Resampling:** \( P(a_t^i = j) = \frac{w_{t-1}^j}{\sum_l w_{t-1}^l} \).

2. **Propagation:** \( x_t^i \sim r_{\theta,t}(x_t | x_{a_t^i}^{1:t-1}) \) and \( x_{1:t}^i = \{ x_{a_t^i}^{1:t-1}, x_t^i \} \).

3. **Weighting:** \( w_t^i = W_{\theta,t}(x_{1:t}^i) = \frac{\gamma_{\theta,t}(x_{1:t}^i)}{\gamma_{\theta,t-1}(x_{1:t-1}^i) r_{\theta,t}(x_t^i | x_{1:t-1}^i)} \).

The result is a new weighted set of particles \( \{ x_{1:t}^i, w_t^i \}_{i=1}^N \).
1. Bayesian learning  
   a) Problem formulation  
   b) Gibbs sampling

2. Sequential Monte Carlo (SMC)

3. Particle Gibbs with ancestor sampling (PGAS)

4. SMC and PGAS for graphical models

The sequential Monte Carlo samplers are fundamental to both the ML and the Bayesian approaches.
Aim: Compute $p(\theta, x_{1:T} \mid y_{1:T})$.

MCMC: Gibbs sampling (blocked) for SSMs amounts to iterating
- Draw $\theta[m] \sim p(\theta \mid x_{1:T}[m-1], y_{1:T})$; \textbf{OK!}
- Draw $x_{1:T}[m] \sim p(x_{1:T} \mid \theta[m], y_{1:T})$. \textbf{Hard!}

Problem: $p(x_{1:T} \mid \theta[m], y_{1:T})$ not available!

Idea: Approximate $p(x_{1:T} \mid \theta[m], y_{1:T})$ using a sequential Monte Carlo method!
Sampling based on SMC

With \( \mathbb{P}(x^*_{1:T} = x^i_{1:T}) \propto w^i_T \) we get, \( x^*_{1:T} \xrightarrow{\text{approx.}} p(x_{1:T} | \theta, y_{1:T}) \).
Problems and a solution

Problems with this approach,

• Based on a PF $\Rightarrow$ approximate sample.
• Does not leave $p(x_{1:T} \mid \theta, y_{1:T})$ invariant!
• Relies on large $N$ to be successful.
• A lot of wasted computations.

To get around these problems,

Use a conditional particle filter (CPF). One pre-specified reference trajectory is retained throughout the sampler.


Thomas Schöen (user.it.uu.se/ thosc112), Ancestor sampling in state space models, graphical models and beyond Seminar at the Department of Statistics, University of Oxford, UK, February 21, 2014.
The idea underlying PMCMC is to make use of a certain SMC sampler to construct a Markov kernel leaving the joint smoothing distribution $p(x_{1:T} \mid \theta, y_{1:T})$ invariant.

This Markov kernel is then used in a standard MCMC algorithm (e.g. Gibbs, results in the Particle Gibbs (PG)).

For a self-contained introduction (focused on BS and AS),

Three SMC samplers leaving $p(x_{1:T} \mid \theta, y_{1:T})$ invariant:

1. **Conditional particle filter (CPF)**


2. **CPF with backward simulation (CPF-BS)**


3. **CPF with ancestor sampling (CPF-AS)**

   Fredrik Lindsten, Michael I. Jordan and Thomas B. Schön, *Particle Gibbs with ancestor sampling*, *Journal of Machine Learning Research (JMLR)*, 2014. (accepted for publication)


Thomas Schön (user.it.uu.se/ thosc112), *Ancestor sampling in state space models, graphical models and beyond*
Let \( x'_{1:T} = (x'_1, \ldots, x'_T) \) be a fixed reference trajectory.

- At each time \( t \), sample only \( N - 1 \) particles in the standard way.
- Set the \( N^{th} \) particle deterministically: \( x^N_t = x'_t \).

CPF causes us to degenerate to the something that is very similar to the reference trajectory, resulting in slow mixing.
BS is problematic for models with more intricate dependencies.

**Reason:** Requires complete trajectories of the latent variable in the backward sweep.

**Solution:** Modify the computation to achieve the same effect as BS, but **without** an explicit backwards sweep.

**Implication:** Ancestor sampling opens up for inference in a wider class of models, e.g. non-Markovian SSMs, PGMs 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 $x'_{1:T} = (x'_1, \ldots, x'_T)$ be a fixed reference trajectory.

- At each time $t$, sample only $N - 1$ particles in the standard way.
- Set the $N^{th}$ particle deterministically: $x^N_t = x'_t$.
- Generate an artificial history for $x^N_t$ by ancestor sampling.

CPF-AS causes us to degenerate to something that is very different from the reference trajectory, resulting in better mixing.
The PGAS Markov kernel

1. Run CPF-AS($x'_{1:T}$) targeting $p(x_{1:T} | \theta, y_{1:T})$.
2. Sample $x^*_{1:T}$ with $\mathbb{P}(x^*_{1:T} = x^i_{1:T}) \propto w^i_T$.

- Maps $x'_{1:T}$ stochastically into $x^*_{1:T}$
- 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 $N \geq 1$ and $\theta \in \Theta$, the PGAS kernel $P^N_\theta$ leaves $p(x_{1:T} | \theta, y_{1:T})$ invariant,*

\[
p(dx^*_{1:T} | \theta, y_{1:T}) = \int P^N_\theta(x'_{1:T}, dx^*_{1:T}) p(dx'_{1:T} | \theta, y_{1:T})
\]
Bayesian learning: Gibbs + CPF-AS = PGAS

Algorithm Particle Gibbs with ancestor sampling (PGAS)

1. **Initialize:** Set $\{\theta[0], x_{1:T}[0]\}$ arbitrarily.
2. **For $m \geq 1$, iterate:**
   (a) Draw $\theta[m] \sim p(\theta \mid x_{1:T}[m - 1], y_{1:T})$.
   (b) Run CPF-AS($x_{1:T}[m - 1]$), targeting $p(x_{1:T} \mid \theta[m], y_{1:T})$.
   (c) Sample with $P(x_{1:T}[m] = x_{1:T}^i) \propto w^i_T$. 
Consider the stochastic volatility model,

\[ x_{t+1} = 0.9x_t + w_t, \quad w_t \sim \mathcal{N}(0, \theta), \]
\[ y_t = e_t \exp \left( \frac{1}{2} x_t \right), \quad e_t \sim \mathcal{N}(0, 1). \]

Let us study the ACF for the estimation error, \( \hat{\theta} - E[\theta | y_{1:T}] \).
Plots of the update rate of $x_t$ versus $t$, i.e. the proportion of iterations where $x_t$ changes value. This provides another comparison of the mixing.
Constructing an artificial sequence of intermediate (auxiliary) target distributions for an SMC sampler is a powerful (and quite possibly underutilized) idea. For some applications, see


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

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

\[
\psi_k(X_{I_k}) = \prod_{C \in C_k} \psi_C(X_C),
\]

where \( I_k \subseteq \{1, \ldots, |\mathcal{V}|\} \).

The \textbf{sequential decomposition} is based on these factors,

\[
\gamma_k(X_{\mathcal{L}_k}) \triangleq \prod_{\ell=1}^{k} \psi_\ell(X_{I_\ell}),
\]

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

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

1. **Initialize** \((k = 1)\): Draw \(X_{L_1}^i \sim r_1(\cdot)\) and set \(w_1^i = W_1(X_{L_1}^i)\).

2. **For** \(k = 2\) **to** \(K\) **do**: 
   
   (a) Draw \(a_k^i \sim \text{Cat}(\{w_{k-1}^j\}_{j=1}^N)\).
   
   (b) Draw \(\xi_k^i \sim r_k(\cdot|X_{L_{k-1}}^{a_k^i})\) and set \(X_{L_k}^i = X_{L_{k-1}}^{a_k^i} \cup \xi_k^i\).
   
   (c) Set \(w_k^i = W_k(X_{L_k}^i)\).

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

**Problem:** SMC is not enough since:

1. It does not solve the parameter learning problem.
2. The quality of the marginals \(p(X_{L_k})\) deteriorates for \(k \ll K\).

**Solution:** Use PGAS.
Consider a standard squared lattice Gaussian MRF of size $10 \times 10$,

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

Four MCMC samplers:

1. PGAS – fully blocked
2. PGAS – partially blocked
3. Standard one-at-a-time Gibbs
4. Tree sampler (Hamze & de Freitas, 2004)

The arrows show the order in which the factors are added.
Example – Gaussian MRF (II/II)

Thomas Schön (user.it.uu.se/thosc112), *Ancestor sampling in state space models, graphical models and beyond*

We have introduced several SMC-based inference methods for PGMs of arbitrary topologies with discrete or continuous variables. The sequential decomposition is not unique and its form will affect

• accuracy
• computational efficiency
• simplicity of implementation

Details and a loopy, non-Gaussian and non-discrete PGM example,


Thomas Schön (user.it.uu.se/ thosc112), Ancestor sampling in state space models, graphical models and beyond Seminar at the Department of Statistics, University of Oxford, UK, February 21, 2014.
Conclusions

- Think of the PGAS kernel as a component that can be used in different inference algorithms.
- Not at all limited to SSMs. Particularly useful for models with more complex dependencies, such as
  - Non-Markovian models
  - Bayesian nonparametric models
  - Probabilistic graphical models
- PGAS is built upon two main ideas
  1. Conditioning the underlying SMC sampler on a reference trajectory ensures the correct stationary distribution for any \( N \).
  2. Ancestor sampling causes degeneration to different trajectories, drastically improving the mixing of the sampler.

There is a lot of interesting research that remains to be done!!
Some references

Novel introduction of PMCMC (given us lots of inspiration)

Forthcoming book
Thomas B. Schön and Fredrik Lindsten, *Learning of dynamical systems – Particle filters and Markov chain methods*, 2014 (or 2015...).

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

PGAS
Fredrik Lindsten, Michael I. Jordan and Thomas B. Schön, *Particle Gibbs with ancestor sampling*, *Journal of Machine Learning Research (JMLR)*, 2014. (accepted for publication)

SMC methods for graphical models

Some MATLAB code is available from the web-site.

Thomas Schön (user.it.uu.se/ thosc112), *Ancestor sampling in state space models, graphical models and beyond*