UPPSALA UNIVERSITET

Sequential Monte Carlo methods and their use in graphical models

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

UPPSALA UNIVERSITET

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

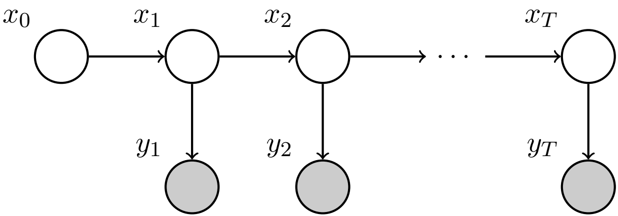
Email: thomas.schon@it.uu.se,
www: user.it.uu.se/~thosc112

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

Background – graphical models (I/II)

A **probabilistic graphical model** (PGM) 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 $(i, j) \in \mathcal{V} \times \mathcal{V}$



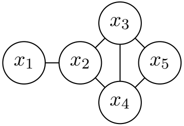
$$p(x_{0:T}, y_{1:T}) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1}) \prod_{t=1}^{T} p(y_t \mid x_t).$$

Background – graphical models (II/II)

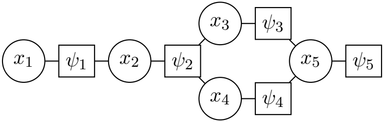
For an undirected graphical model (Markov random field), the joint PDF over all the involved random variables is

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

where $\mathcal{C}$ is the set of cliques in $\mathcal{G}$, and $Z = \int \prod_{C \in \mathcal{C}} \psi_C(X_C) dX_{\mathcal{V}}$.



Undirected graph



Example of a **factor graph** making interactions explicit,
$p(x_{1:5}) = \frac{1}{Z} \prod_{i=1}^{5} \psi_i(\cdot)$.

Background – sequential Monte Carlo

Approximate a **sequence** of probability distributions on a sequence of probability spaces of **increasing dimension**.

Let $\{\gamma_k(x_{1:k})\}_{k\geq 1}$ be a sequence of unnormalised densities and

$$\bar{\gamma}_k(x_{1:k}) = \frac{\gamma_k(x_{1:k})}{Z_k}$$

Approximates

$$\bar{\gamma}_k(x_{1:k}) \approx \sum_{i=1}^{N} \frac{w_k^i}{\sum_{l=1}^{N} w_k^l} \delta_{x_{1:k}^i}(x_{1:k}).$$

Ex. (state space model (SSM))

$$\bar{\gamma}_k(x_{1:k}) = p(x_{1:k} \mid y_{1:k}), \qquad \gamma_k(x_{1:k}) = p(x_{1:k}, y_{1:k}),$$

$$Z_k = p(y_{1:k}).$$

Sequential Monte Carlo – particle filter

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

$$x_{t+1} \mid x_t \sim f_\theta(x_{t+1} \mid x_t),$$
$$y_t \mid x_t \sim g_\theta(y_t \mid x_t),$$
$$x_1 \sim \mu_\theta(x_1).$$

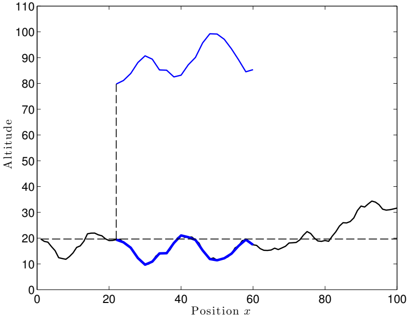
The particle filter maintains an empirical distribution made up of N samples (particles) $\{x_{1:t}^i\}_{i=1}^N$ and corresponding weights $\{w_{1:t}^i\}_{i=1}^N$

$$\widehat{p}(x_{1:t} \mid y_{1:t}) = \sum_{i=1}^{N} w_t^i \delta_{x_{1:t}^i}(x_{1:t}).$$

"Think of each particle as one simulation of the system state. Keep the ones that best explains the measurements."

The particle filter – toy problem

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:

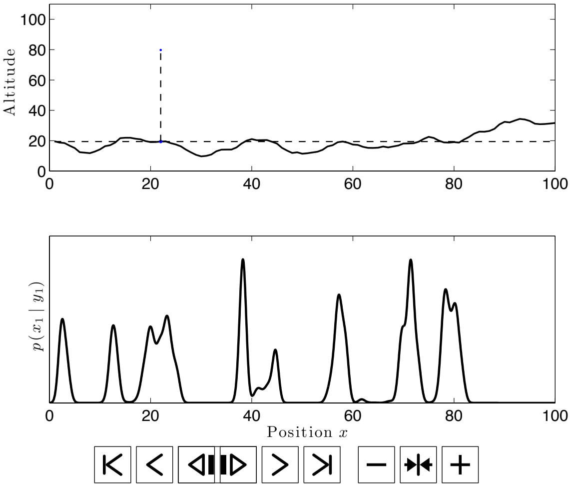
$$y_t = h(x_t) + e_t.$$

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,

Thomas Schön, Fredrik Gustafsson, and Per-Johan Nordlund. **Marginalized particle filters for mixed linear/nonlinear state-space models**. *IEEE Transactions on Signal Processing*, 53(7):2279-2289, July 2005.

The particle filter – toy problem

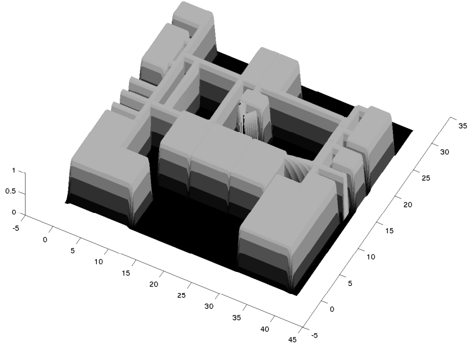


Highlights two **key capabilities** of the PF:

1. Automatically handles an unknown and dynamically changing number of hypotheses.
2. Work with nonlinear/non-Gaussian models.

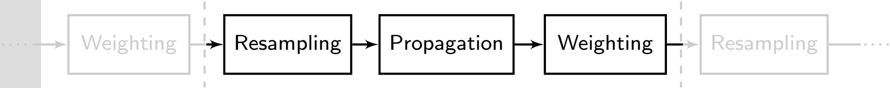
Application example – indoor localization

Aim: Compute the position of a person moving around indoors using sensors (inertial, magnetometer and radio) located in an ID badge and a map.



Show movie

Sequential Monte Carlo – particle filter



SMC = resampling + sequential importance sampling

Given, $\{x^i_{1:t-1}, w^i_{t-1}\}_{i=1}^N$, repeat for $i = 1, \ldots, N$:

1. **Resampling:** $\mathbb{P}\left(\check{x}^i_{1:t-1} = x^j_{1:t-1}\right) = w^j_{t-1} / \sum_l w^l_{t-1}$.
2. **Propagation:** $x^i_t \sim f_\theta(x_t \mid \check{x}^i_{1:t-1})$ and $x^i_{1:t} = \{\check{x}^i_{1:t-1}, x^i_t\}$.
3. **Weighting:** $w^i_t = W_t(x^i_t) = g_\theta(y_t \mid x_t)$.

(a hopefully) intuitive preview

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

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

Key idea: Perform and make use of various **decompositions** of graphical models to design SMC inference methods.

Outline

1. Background – graphical models
2. Background – sequential Monte Carlo
3. **Example – from information theory**
4. SMC for general graphical models
5. Particle MCMC (very brief)
6. Example – Markov random field
7. Conclusions

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

Information theory – 2D channel capacity

Example borrowed from:

M. Molkaraie and H.-A. Loeliger, **Monte Carlo algorithms for the partition function and information rates of two-dimensional channels**, *IEEE Transactions on Information Theory*, 59(1): 495–503, 2013.

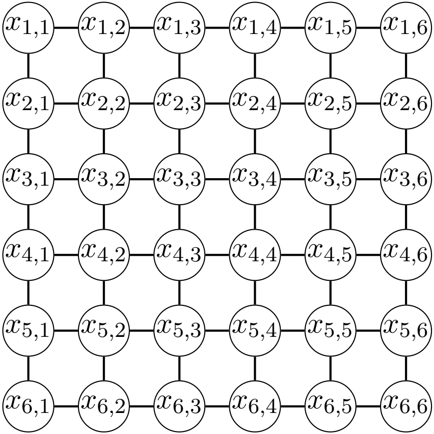
2D binary-input channel with the **constraint** that no two horizontally or vertically adjacent variables may be both be equal to 1.

$$\begin{matrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & \cdots \\ \cdots & 0 & 1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{matrix}$$

Of interest in magnetic and optical storage solutions.

The channel can be described by a square lattice **undirected graphical model**.

2D channel capacity – graphical model



The variables are binary $x_{\ell,j} \in \{0, 1\}$ and the interactions are pair-wise between adjacent variables.

$$\text{Factors: } \psi(x_{\ell,j}, x_{m,n}) = \begin{cases} 0, & x_{\ell,j} = x_{m,n} = 1 \\ 1, & \text{otherwise} \end{cases}$$

2D channel capacity – graphical model

The resulting joint PDF is given by

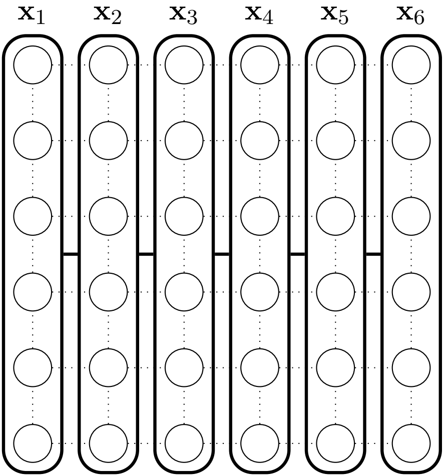
$$p(X_{\mathcal{V}}) = \frac{1}{Z} \prod_{(\ell j, mn) \in \mathcal{E}} \psi(x_{\ell,j}, x_{m,n}).$$

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

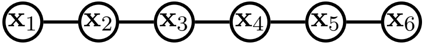
$$C_M = \frac{1}{M^2} \log_2 Z.$$

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

2D channel capacity – undirected chain



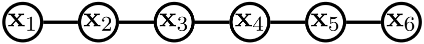
Rewrite the PGM as a high-dimensional **undirected chain** by introducing a new set of variables $\mathbf{x}_k$.



$$\boldsymbol{\phi}(\mathbf{x}_k) = \prod_{j=1}^{M-1} \psi(x_{j+1,k}, x_{j,k}),$$

$$\boldsymbol{\psi}(\mathbf{x}_k, \mathbf{x}_{k-1}) = \prod_{j=1}^{M} \psi(x_{j,k}, x_{j,k-1}).$$

2D channel capacity – SMC algorithm



The **undirected chain** results in the following joint PDF

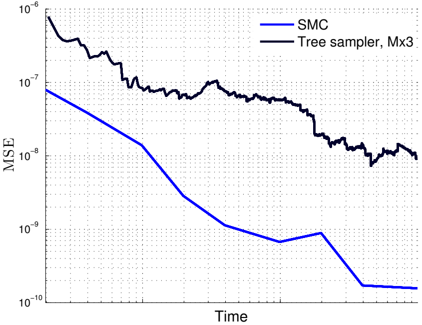
$$p(X_{\mathcal{V}}) = \frac{1}{Z} \prod_{k=1}^{M} \phi(\mathbf{x}_k) \prod_{k=2}^{M} \psi(\mathbf{x}_{k-1}, \mathbf{x}_k).$$

Provides a **natural sequence of target distributions** for SMC!

Sequential decomposition:

$$\begin{aligned} \gamma_1(\mathbf{x}_1) &= \phi(\mathbf{x}_1), \\ \gamma_k(\mathbf{x}_{1:k}) &= \gamma_{k-1}(\mathbf{x}_{1:k-1}) \phi(\mathbf{x}_k) \psi(\mathbf{x}_{k-1}, \mathbf{x}_k). \end{aligned}$$

2D channel capacity – 60×60 example



Our SMC sampler compared to the **tree sampler** by

F. Hamze and N. de Freitas, **From fields to trees**, *In Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*, Banff, Canada, July, 2004.

implemented according to

M. Molkaraie and H.-A. Loeliger, **Monte Carlo algorithms for the partition function and information rates of two-dimensional channels**, *IEEE Transactions on Information Theory*, 59(1): 495–503, 2013.

For the 2D channel: **fully adapted** SMC sampler. To sample exactly the $\mathbf{x}_k$'s we use a forward/backward algorithm.

This was just a special case, the important question is, can we do this for a general probabilistic graphical model?! **Yes!**

Using “standard” SMC for PGMs – the idea

Key idea:

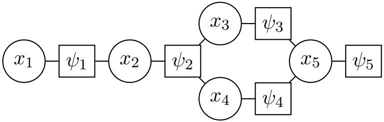
- Perform a **sequential decomposition** of the graphical model.
- Each **subgraph** induces an artificial target distribution.
- Apply SMC to the sequence of artificial target distributions.

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

Sequential decomposition of PGMs – pictures

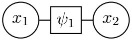
The joint PDF of the set of random variables indexed by $\mathcal{V}$, $X_{\mathcal{V}} \triangleq \{x_1, \ldots, x_{|\mathcal{V}|}\}$

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

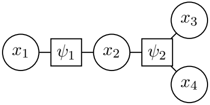


Example of a sequential decomposition of the above factor graph (the target distributions are built up by adding factors at each iteration),

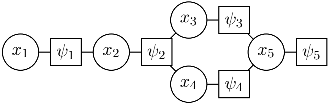
$\gamma_1(X_{\mathcal{L}_1})$



$\gamma_2(X_{\mathcal{L}_2})$



$\gamma_3(X_{\mathcal{L}_3}) \propto p(X_{\mathcal{V}})$



Sequential decomp. of PGMs – 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 ψ_k.

The **sequential decomposition** is based on these factors,

$$\gamma_k(X_{\mathcal{L}_k}) \triangleq \prod_{\ell=1}^{k} \psi_\ell(X_{\mathcal{I}_\ell}),$$

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

By construction, $\mathcal{L}_K = \mathcal{V}$ and the joint PDF $p(X_{\mathcal{L}_K}) \propto \gamma_K(X_{\mathcal{L}_K})$.

SMC sampler for graphical models

Algorithm SMC sampler for graphical models

1. **Initialize ($k = 1$):** Draw $X^i_{\mathcal{L}_1} \sim r_1(\cdot)$ and set $w^i_1 = W_1(X^i_{\mathcal{L}_1})$.
2. **For** $k = 2$ **to** K **do:**
   (a) Draw $a^i_k \sim \mathcal{C}(\{w^j_{k-1}\}^N_{j=1})$.
   (b) Draw $\xi^i_k \sim r_k(\cdot | X^{a^i_k}_{\mathcal{L}_{k-1}})$ and set $X^i_{\mathcal{L}_k} = X^{a^i_k}_{\mathcal{L}_{k-1}} \cup \xi^i_k$.
   (c) Set $w^i_k = W_k(X^i_{\mathcal{L}_k})$.

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

A few examples where the partition function is interesting:

1. Likelihood-based learning of parameters in the PGM.
2. Capacity calculations of a channel (information theory).
3. Free energy of a system of objects (statistical mechanics).

UPPSALA UNIVERSITET

Outline

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

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 (SSM special case): Construct a Markov chain with $p(\theta \mid y_{1:T})$ (or $p(\theta, x_{1:T} \mid y_{1:T})$) as its stationary distribution.

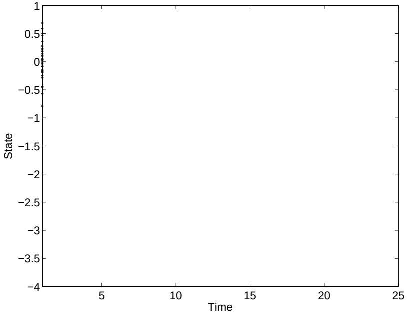
Pioneered by the work

Christophe Andrieu, Arnaud Doucet and Roman Holenstein, **Particle Markov chain Monte Carlo methods**, *Journal of the Royal Statistical Society: Series B*, 72:269-342, 2010.

Particle MCMC deals with SMC problems

Problems with SMC, it is not enough since:

1. It does not solve the parameter learning problem.
2. The quality of the marginals $p(X_{\mathcal{L}_k}) = \int \tilde{\gamma}_K(X_{\mathcal{L}_K}) \mathrm{d}X_{\mathcal{L}_K \setminus \mathcal{L}_k}$ deteriorates for $k \ll K$ (particle degeneracy).



UPPSALA UNIVERSITET

Particle MCMC deals with SMC problems

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

Fredrik Lindsten, Michael I. Jordan and Thomas B. Schön. **Particle Gibbs with ancestor sampling**. *Journal of Machine Learning Research (JMLR)*, 15:2145-2184, June 2014.

We do not need the details, but if you want a talk on it, see

`www.newton.ac.uk/seminar/20140425104011151`

This allows us to:

1. Simulate, jointly, blocks of variables using an MCMC scheme.
2. Opens up for learning unknown parameters of the model.

Partial blocking

Two extremes of how to sample the variables:

1. Simulate all the latent variables $X_{\mathcal{L}_K}$ 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 $X_{\mathcal{L}_K}$ jointly.	**Partial blocking via PGAS.**	Simulate one variable x_j at a time.

Example – Gaussian MRF

Consider a standard square lattice Gaussian MRF of size 10×10,

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

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 $\sigma_{ij} = 0.1$).

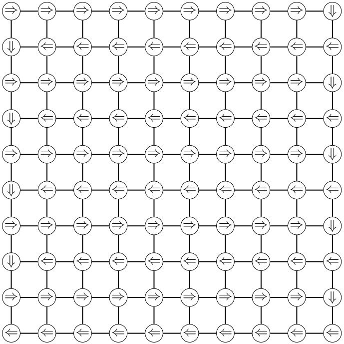
Goal: Compute the posterior distribution $p(X_{\mathcal{V}} \mid Y_{\mathcal{V}})$.

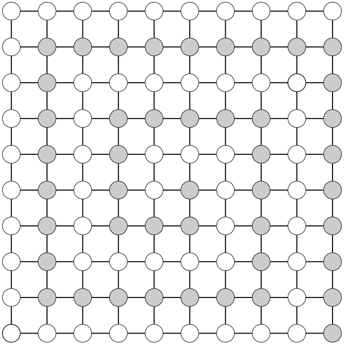
We run four MCMC samplers:

1. Standard one-at-a-time Gibbs
2. Tree sampler (Hamze & de Freitas, 2004)
3. PGAS – fully blocked ($N = 50$)
4. PGAS – partially blocked ($N = 50$)

UPPSALA UNIVERSITET

Example – Gaussian MRF

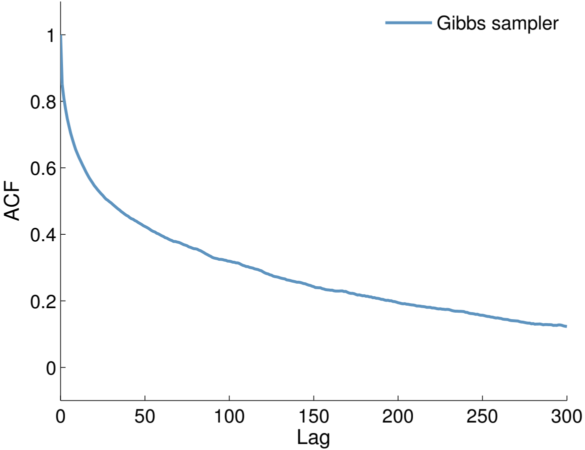




The arrows show the order in which the factors are added.

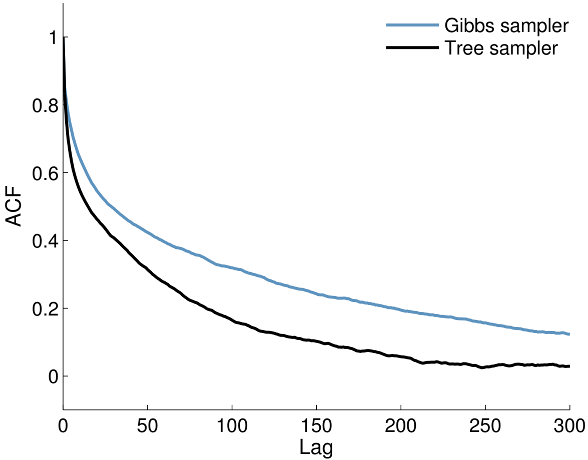
The two block structures used by the tree sampler and PGAS with partial blocking.

Example – Gaussian MRF



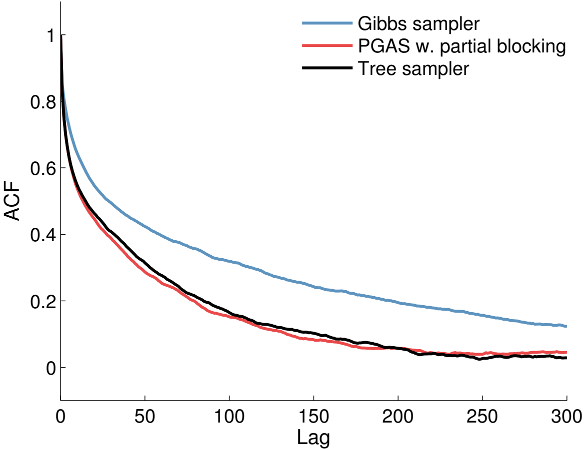
The one-step-at -a-time Gibbs sampler is struggling due to the strong interactions.

Example – Gaussian MRF



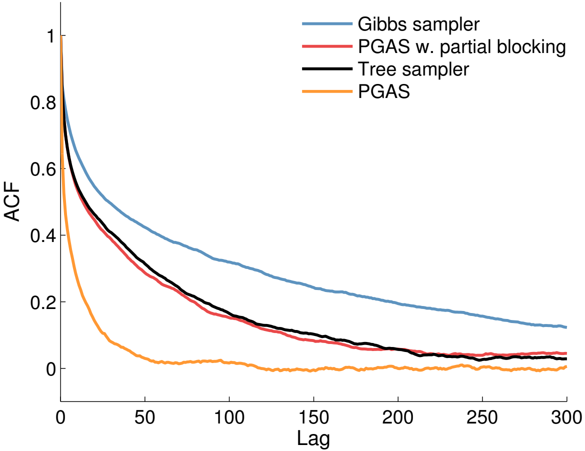
The tree sampler implements an “ideal” partially blocked Gibbs sampler.

Example – Gaussian MRF



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.

Example – Gaussian MRF



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.

For more challenging examples, see our papers.

Conclusions

- Derived SMC-based inference methods for PGMs of arbitrary topologies with discrete or continuous random variables.
- **Key insight:** We exploit a sequential decomposition of the graphical model.
- Using the SMC sampler as a proposal within MCMC provides highly useful constructions.

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)

Thomas B. Schön and Fredrik Lindsten. **Learning of dynamical systems – Particle filters and Markov chain methods**, 2014.

References to some of our work

SMC methods for graphical models

Christian A. Naesseth, Fredrik Lindsten and Thomas B. Schön, **Sequential Monte Carlo methods for graphical models**. *Advances in Neural Information Processing Systems (NIPS)* 27, Montreal, Canada, December, 2014.

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

Fredrik Lindsten, Adam M. Johansen, Christian A. Naesseth, Bonnie Kirkpatrick, Thomas B. Schön, John Aston and Alexandre Bouchard-Côté. **Divide-and-Conquer with Sequential Monte Carlo**. *arXiv:1406.4993*, June 2014.

Information theory example

Christian A. Naesseth, Fredrik Lindsten and Thomas B. Schön, **Capacity estimation of two-dimensional channels using Sequential Monte Carlo**. *Proceedings of the 2014 IEEE Information Theory Workshop (ITW)*, November, 2014.

PMCMC methods

Fredrik Lindsten, Michael I. Jordan and Thomas B. Schön. **Particle Gibbs with ancestor sampling**. *Journal of Machine Learning Research (JMLR)*, 15:2145-2184, June 2014.

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

Self-contained introduction to particle smoothing using BS and AS

Fredrik Lindsten and Thomas B. Schön, **Backward simulation methods for Monte Carlo statistical inference**, *Foundations and Trends in Machine Learning*, 6(1):1-143, 2013.