Gibbs sampling for state space models
Blocking, stability, and particle MCMC

Fredrik Lindsten
Uppsala University
March 2, 2016
Outline

1. Background – Particle Gibbs

2. Blocking strategies and stability

3. Particle Gibbs with Ancestor Sampling
Inference in state-space models

Consider a nonlinear discrete-time state-space model,

\[ X_t \mid X_{t-1} \sim m_\theta(X_{t-1}, \cdot), \]
\[ Y_t \mid X_t \sim g_\theta(X_t, \cdot), \]

and \( X_1 \sim \mu. \)

We observe \( Y_{1:T} = y_{1:T} := (y_1, \ldots, y_T) \) and wish to estimate \( \theta \) and/or \( X_{1:T}. \)
**Gibbs sampler for SSMs**

Let

\[ \phi_{T,\theta}(dx_{1:T}) = p(x_{1:T} \mid \theta, y_{1:T})dx_{1:T}, \]

denote the *joint smoothing distribution*.

---

**MCMC:** Gibbs sampling for state-space models. Iterate,

- Draw \( \theta[k] \sim p(\theta \mid X_{1:T}[k - 1], y_{1:T}) \); \hspace{1cm} OK!
- Draw \( X_{1:T}[k] \sim \phi_{T,\theta[k]}(\cdot) \). \hspace{1cm} Hard!

---

**One-at-a-time:** \( X_t[k] \sim p(x_t \mid \theta[k], X_{t-1}[k], X_{t+1}[k - 1], y_t) \)

**Particle Gibbs:** Approximate \( \phi_{T,\theta}(dx_{1:T}) \) using a particle filter.
The particle filter

The particle filter approximates $\phi_{t,\theta}(dx_{1:t})$, $t = 1, \ldots, T$ by

$$
\hat{\phi}_{t,\theta}^N(dx_{1:t}) := \sum_{i=1}^{N} \frac{\omega_t^i}{\sum_{\ell} \omega_t^\ell} \delta_{X_{1:t}^i}(dx_{1:t}).
$$

- **Resampling:** $\{X_{1:t-1}^i, \omega_{t-1}^i\}_{i=1}^{N} \rightarrow \{\tilde{X}_{1:t-1}^i, 1/N\}_{i=1}^{N}$.
- **Propagation:** $X_t^i \sim q_{t,\theta}(\tilde{X}_{t-1}^i, \cdot)$ and $X_{1:t}^i = (\tilde{X}_{1:t-1}^i, X_t^i)$.
- **Weighting:** $\omega_t^i = W_{t,\theta}(\tilde{X}_{t-1}^i, X_t^i)$.

$$
\Rightarrow \{X_{1:t}^i, \omega_t^i\}_{i=1}^{N}
$$
MCMC using particle filters

In MCMC we need a Markov kernel with invariant distribution $\phi_T$. (From now on we drop $\theta$ from the notation.)

Conditional particle filter (CPF)
Let $x_{1:T}' = (x_1', \ldots, x_T')$ be a given reference trajectory.

- Sample only $N - 1$ particles in the standard way.
- Set the $N$th particle deterministically:
  $$X_t^N = x_t' \text{ and } A_t^N = N.$$
Consider the procedure:

1. Run CPF($N, x'_{1:T}$) targeting $\phi_T(dx_{1:T})$,
2. Sample $X^*_{1:T}$ with $\mathbb{P}(X^*_{1:T} = X^i_{1:T} \mid \mathcal{F}_T^N) \propto \omega^i_T$. 
The PG Markov kernel (I/II)

Consider the procedure:

1. Run CPF($N, x'_{1:T}$) targeting $\phi_T(dx_{1:T})$,
2. Sample $X^*_{1:T}$ with $\mathbb{P}(X^*_{1:T} = X^i_{1:T} \mid \mathcal{F}_T^N) \propto \omega^i_T$. 

```
Time
State
5 10 15 20 25 30 35 40 45 50
-3
-2
-1
0
1
2
3
```

fredrik.lindsten@it.uu.se
The PG Markov kernel (II/II)

This procedure:

- Maps $x_{1:T}'$ stochastically into $X_{1:T}^*$. 
- Implicitly defines a Markov kernel $P_N$ on $(X^T, X^T)$ (the PG kernel),

$$P_N(x_{1:T}', A) = \mathbb{E}[\mathbbm{1}_A(X_{1:T}^*)]$$

- $P_N$ is $\phi_T$-invariant for any number of particles $N \geq 1$.

What about ergodicity?
Minorisation

Theorem

The PG kernel is minorised by $\phi_T$:

$$P_N(x'_{1:T}, A) \geq (1 - \varepsilon_{T,N})\phi_T(A).$$

If we further assume that the SSM is strongly mixing, then

$$1 - \varepsilon_{T,N} = \left(1 - \frac{1}{c(N-1)+1}\right)^T$$

for $c \in (0, 1]$ (depending on mixing).


- Implies uniform ergodicity: $\|\mu P_N^k(\cdot) - \phi_T(\cdot)\|_{TV} \leq \varepsilon_{T,N}^k$.
- Stable as $T \to \infty$ if $N \sim \gamma T$,

$$\limsup_{T \to \infty} \varepsilon_{T,\gamma T} = 1 - \exp \left(-\frac{1}{\gamma c}\right) < 1.$$
Gibbs sampling for state space models

Alternative Gibbs sampling strategies:

**Particle Gibbs:** \( X^\star_{1:T} \sim P_N(x_{1:T}, \cdot) \).
- Samples \( X_{1:T} \) in one “block”.
- Requires \( N \propto T \) as \( T \to \infty \) for stability (strong mixing)

\[ \Rightarrow O(T^2) \text{ computational cost!} \]

**One-at-a-time:** \( X^*_t \sim p(x_t \mid x_{-t}, y_t), \ t = 1, \ldots, T \).
- Slow mixing/convergence speed!
- Stable as \( T \to \infty \)?
Outline

1. Background – Particle Gibbs

2. Blocking strategies and stability

3. Particle Gibbs with Ancestor Sampling
Blocking strategy

Intermediate strategy — *blocked Particle Gibbs*:

\[ P_N^J(x_{J+}, dx^*_J) \]  
PG kernel for \[ p(x_J \mid x_{\partial J}, y_J). \]

**Trade off:**

1. Mixing of *ideal blocked Gibbs sampler* \( \uparrow \) as \( |J| \uparrow \)  
   (how fast? stable?)

2. “Mixing of \( P_N^J \) = \( 1 - \frac{1}{c(N-1)+1} \)^{|J|}, i.e., \( \downarrow \) as \( |J| \uparrow \)

\( \partial J = \{ t \in J^c : t + 1 \in J \text{ or } t - 1 \in J \} \) ("boundary points for block \( J' \))

\( J^+ = J \cup \partial J \)
Stability of ideal blocked Gibbs sampler

**Theorem**

Let \( J = \{ J_1, \ldots, J_m \} \) be a cover of \( \{1, \ldots, T\} \) and let \( \mathcal{P} = P^{J_1} \cdots P^{J_m} \) be the ideal Gibbs kernel for one complete sweep. Let all blocks have common size \( L \) and common overlap \( p \). Then

\[
|\mu \mathcal{P}^k(f) - \phi_T(f)| \leq 2\lambda^{k-1} \sum_{t=1}^{T} \text{osc}_t(f),
\]

where \( \lambda = \alpha^{p+1} + \alpha^{L-p} \) and \( \alpha \in [0, 1) \) is a constant depending on the mixing coefficients of the model (assuming strong mixing).

**Def:**

\[
\text{osc}_t(f) = \sup_{x, z \in X^T} |f(x) - f(z)| \quad \text{for } x - t = z - t
\]
Stability of ideal blocked Gibbs sampler

- If $L \geq 2p + 1$ the odd/even blocks can be updated in parallel!
- To control the rate $\lambda = \alpha^{p+1} + \alpha^{L-p}$ we need to increase both block size $L$ and overlap $p$!

  ex) with $L = 2p + 1$ ($\lesssim 50\%$ overlapping blocks) we get

  $$\lambda < 1 \quad \text{if} \quad L > \frac{\log 4}{\log \alpha - 1} - 1.$$ 

- For left-to-right and parallel blocking the rate is $\sim \lambda^2$. 
Stability of blocked Particle Gibbs sampler

The *blocked Particle Gibbs sampler* $\mathcal{P}_N$ can be seen as a perturbation of the ideal blocked Gibbs sampler $\mathcal{P}$.

**Theorem**

$$|\mu \mathcal{P}_N^k(f) - \phi_T(f)| \leq 2\lambda_N^{k-1} \sum_{t=1}^{T} \text{osc}_t(f)$$

$$\lambda_N = \lambda + \text{const.} \times \epsilon_{L,N}, \quad \epsilon_{L,N} = 1 - \left(1 - \frac{1}{c(N-1)+1}\right)^L.$$

- $\lambda \rightarrow 0$ with increasing block size $L$ and overlap $p$.
- $\epsilon_{L,N} \downarrow$ as $N \uparrow$; $\epsilon_{L,N} \uparrow$ as $L \uparrow$.
- Bound independent of $T$ for marginals of $\phi_T(dx_{1:T})$.
- $\|\mu \mathcal{P}_N^k(\cdot) - \phi_T(\cdot)\|_{TV} \leq 2T\lambda_N^{k-1}$. 

fredrik.lindsten@it.uu.se

Uppsala University
Outline

1. Background – Particle Gibbs

2. Blocking strategies and stability


3. Particle Gibbs with Ancestor Sampling

The PGAS Markov kernel

**Standard Particle Gibbs:**
- At time $t$: set $X^N_t = x'_t$ and $A^N_t = N$.

**Particle Gibbs with “Ancestor Sampling” (PGAS):**
- At time $t$: set $X^N_t = x'_t$ and sample $A^N_t \in \{1, \ldots, N\}$ with

  $$
  \mathbb{P}(A^N_t = j \mid \mathcal{F}^N_{t-1}) = \frac{\omega^j_{t-1} m(X^j_{t-1}, x'_t)}{\sum_l \omega^l_{t-1} m(X^l_{t-1}, x'_t)}.
  $$

- Can be understood as an application of Bayes’ theorem:

  $$
  \mathbb{P}({X^j_{t-1} \text{ is the ancestor of } x'_t}) \propto \omega^j_{t-1} m(X^j_{t-1}, x'_t).
  $$

  "prior" "likelihood"
PGAS vs. PG

PGAS

PG
Plots of the update rate of $X_t$ versus $t$, i.e. the proportion of iterations where $X_t$ changes value. (Simulated data from a simple stochastic volatility model.)
Coloured regions = intervals between coalescence points.

PGAS ⇔ stochastic and adaptive blocking?

If yes, then ∃\(N_0\) such that PGAS is stable as \(T \to \infty\) for \(N \geq N_0\)?
Summary

- Particle Gibbs – mimics sampling from $\phi_{T,\theta}(dx_{1:T})$ (e.g., in a Gibbs sampler or stochastic approximation method).
- **Uniformly ergodic** under weak conditions.
  - *Strong mixing conditions:* stable if $N = \gamma T$.
  - *(Weaker) Moment conditions:* stable if $N = T^{1/\gamma}$.
- **Blocking** $\Rightarrow$ stable as $T \to \infty$ for constant $N$.
  - Set block size $L$ and overlap $p$ s.t. ideal sampler is stable.
  - Set $N$ large enough to obtain a stable Particle Gibbs sampler (depends only on $L$, not $T$).
  - Opens up for parallelisation!
  - Requires evaluation of $m_{\theta}(x_{t-1}, x_{t})$!
- **Ancestor sampling** $\Rightarrow$ much improved empirical performance
  - Can AS be viewed as adaptive and stochastic blocking?
  - Stable as $T \to \infty$ for fixed $N$?
  - Requires evaluation of $m(x_{t-1}, x_{t})$!
Wasserstein estimates

**Def:** For $f : X^T \mapsto \mathbb{R}$, the oscillation in the $i$-th coordinate is

$$osc_i(f) = \sup_{x,z \in X^T, x-i = z-i} |f(x) - f(z)|$$

**Def:** $W$ is a *Wasserstein matrix* for Markov kernel $P$ if

$$osc_i(Pf) \leq \sum_{j=1}^{T} W_{ij} osc_j(f).$$
Wasserstein matrix for blocked Gibbs sampler

Under strong mixing

\[
W^J = \begin{bmatrix}
1 & & & & & \\
& 1 & & & & \\
& & \alpha & & & \\
& & 0 & \cdots & 0 & \alpha^{|J|} \\
& & \alpha^2 & 0 & \cdots & 0 & \alpha^{|J|-1} \\
& \ddots & \ddots & \ddots & \ddots & \ddots & \\
& & \alpha^{|J|} & 0 & \cdots & 0 & \alpha \\
& & & \alpha_{|J|} & 1 & \\
& & & & & & 1
\end{bmatrix},
\]

is a Wasserstein Matrix for the \textit{ideal} Gibbs kernel updating block \( J \),

\[
P^J(x_{1:T}, dx_{1:T}^*) : \left\{ \begin{aligned}
    X^*_J &\sim p(x_J \mid x_{\partial J}, y_J) dx_J, \\
    X^*_Jc &= x_{Jc}
\end{aligned} \right.
\]

where \( \alpha \in [0, 1) \) is a constant depending on the mixing coefficients.
Stability of blocked Gibbs sampler

**Theorem**

Let \( \mathcal{J} = \{J_1, \ldots, J_m\} \) be a cover of \( \{1, \ldots, T\} \) and let \( \mathcal{P} = P^{J_1} \cdots P^{J_m} \) be the Gibbs kernel for one complete sweep. Let \( \partial = \bigcup_{J \in \mathcal{J}} \partial J \). Then, if

\[
\sup_{i \in J \cap \partial} \sum_{j=1}^{T} W_{ij}^J \leq \lambda < 1 \quad \forall J \in \mathcal{J}, \quad (\star)
\]

it follows that

\[
|\mu \mathcal{P}^k(f) - \phi_T(f)| \leq 2\lambda^{k-1} \sum_{i=1}^{T} \text{osc}_i(f).
\]

- With \( \lesssim 50\% \) overlapping equally sized blocks, (\( \star \)) is satisfied if the block size satisfies \( |J| > \frac{\log 4}{\log \alpha^{-1}} - 1 \).
- For left-to-right and parallel blocking the rate is \( \sim \lambda^2 \).