Monte Carlo methods for inference in dynamical systems

Fredrik Lindsten

Uppsala University

March 30, 2016
Consider a nonlinear, discrete-time state-space model,

\[ x_{t+1} \mid x_t \sim f_{\theta}(x_{t+1} \mid x_t), \]
\[ y_t \mid x_t \sim g_{\theta}(y_t \mid x_t), \]

and \( x_1 \sim \mu(x_1) \).

We observe \( y_{1:T} := (y_1, \ldots, y_T) \) and wish to estimate \( \theta \).

**Bayesian model:** \( \theta \) is a random variable with prior density \( p(\theta) \).

**Aim:** Compute \( p(\theta \mid y_{1:T}) \).
Gibbs sampler for SSMs

**Aim:** Find \( p(\theta \mid y_{1:T}) \).

Alternate between updating \( \theta \) and updating \( x_{1:T} \).

⇒ Gibbs sampling
Markov kernel

Let \( \{X_k\}_{k \geq 0} \) be a Markov chain on a general state space. The Markov transition kernel is given by,

\[
P(x_{k-1}, A) = \mathbb{P}(X_k \in A \mid X_{k-1} = x_{k-1}).
\]

Stationary distribution

If \( \pi \) is a probability distribution such that,

\[
\pi(A) = \int P(x, A)\pi(x) \, dx,
\]

we say that \( \pi \) is a stationary distribution for \( P \).
Crash course in MCMC (II/IV)

**Ergodic theorem**

If \( \{X_k\}_{k \geq 0} \) is ergodic with stationary distribution \( \pi \), then

\[
\mathbb{E}_\pi[g(X)] = \int g(x) \pi(x) \, dx = \lim_{m \to \infty} \frac{1}{m} \sum_{k=1}^{m} g(X_k)
\]

- average over space
- average over "time"

**Markov chain Monte Carlo:** We want to compute \( \mathbb{E}_\pi[g(X)] \)

1. Construct an ergodic Markov kernel \( P \) with stationary distribution \( \pi \).
2. Simulate \( \{X_k\}_{k=1}^{m} \) according to \( P \) for large \( m \).
3. Approximate \( \mathbb{E}_\pi[g(X)] \approx \frac{1}{m} \sum_{k=1}^{m} g(X_k) \)
Two common methods for constructing $P$:

1. Metropolis-Hastings
2. Gibbs sampling

Update components $X_j$ of the vector $X = (X_1, \ldots, X_d)$ in turn,

$$X_j \sim \pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d), \quad j = 1, \ldots, d.$$
Sample from, $\pi(x, y) = \mathcal{N} \left( \begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \right)$.

Gibbs sampler

- Draw $x' \sim \pi(x \mid y)$;
- Draw $y' \sim \pi(y \mid x')$. 

![Graph showing the distribution and the sampling process](image)
Gibbs sampler for SSMs, cont’d

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

Alternate between updating \( \theta \) and updating \( x_{1:T} \).

**MCMC:** Gibbs sampling for state-space models. Iterate,
- Draw \( \theta[k] \sim p(\theta \mid x_{1:T}[k-1], y_{1:T}) \);
- Draw \( x_{1:T}[k] \sim p(x_{1:T} \mid \theta[k], y_{1:T}) \).

The above procedure results in a Markov chain,
\[
\{\theta[k], x_{1:T}[k]\}_{k \geq 1}
\]
with stationary distribution \( p(\theta, x_{1:T} \mid y_{1:T}) \).
Gibbs sampling for linear system identification.

\[ x_{t+1} = Ax_t + v_t, \quad v_t \sim \mathcal{N}(0, Q), \]
\[ y_t = Cx_t + e_t, \quad e_t \sim \mathcal{N}(0, R), \]

with \( \theta = (A, C, Q, R) \).

**Gibbs sampling:** Iterate,

- Draw \( \theta' \sim p(\theta \mid x_{1:T}, y_{1:T}) \);
- Draw \( x'_{1:T} \sim p(x_{1:T} \mid \theta', y_{1:T}) \).
  - Run a Kalman filter for \( t = 1, \ldots, T \).
  - Simulate backward, \( x'_T, x'_{T-1}, \ldots, x'_1 \).
Gibbs sampler for general SSM?

What about the general nonlinear/non-Gaussian case?

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

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

**Idea:** Approximate $p(x_{1:T} \mid \theta, y_{1:T})$ using a particle filter.
The particle filter

Consider state inference for a fixed value of $\theta$.

The particle filter approximates $p(x_t \mid \theta, y_{1:t})$, $t = 1, \ldots, T$ by

$$\hat{p}^N(x_t \mid \theta, y_{1:t}) := \sum_{i=1}^{N} \frac{w_t^i}{\sum_{\ell} w_t^\ell} \delta_{x_t^i}(x_t).$$

$t = 1$, importance sampling from prior: For $i = 1, \ldots, N$,

- Sample $x_1^i \sim \mu(x_1)$;
- Compute weights, $w_1^i = g_{\theta}(y_1 \mid x_1^i)$. 
The particle filter

\[ t > 1: \]

- **Resampling:** \( \{x^i_{t-1}, w^i_{t-1}\}^N_{i=1} \rightarrow \{\tilde{x}^i_{t-1}, 1/N\}^N_{i=1} \).

- **Propagation:** \( x^i_t \sim f_\theta(x_t \mid \tilde{x}^i_{t-1}) \).

- **Weighting:** \( w^i_t = g_\theta(y_t \mid x^i_t) \).

\[ \Rightarrow \{x^i_t, w^i_t\}^N_{i=1} \]

N.B. This is the simplest PF and many extensions and improvements are available.
The particle filter

Algorithm Bootstrap particle filter (all steps for $i = 1, \ldots, N$)

1. Initialize ($t = 1$):
   (a) Draw $x_1^i \sim \mu(x_1)$.
   (b) Set $w_1^i = g_\theta(y_1 \mid x_1^i)$.

2. for $t = 2, 3, \ldots$:
   (a) Draw $a_t^i \sim \text{Discrete}(\{w_{t-1}^j\}_{j=1}^N)$.
   (b) Draw $x_t^i \sim f_\theta(x_t \mid x_{t-1}^{a_t^i})$.
   (c) Set $w_t^i = g_\theta(y_t \mid x_t^i)$.

Theoretical justification:

- Strongly consistent, central limit theorem at rate $\sqrt{N}$.
- Stable as $t \to \infty$ under mixing assumptions.
- Unbiased estimate of normalizing constant $p(y_{1:t} \mid \theta)$.
- ...
PF illustration
Filtering or smoothing?

- Possible to reconstruct *state trajectories* by tracing the genealogy of the particles,

\[ \left( \ldots, x_{T-1}^{a_i}, x_T^{a_i}, x_T^i \right) =: x_{1:T}^i \]

- At \( t = T \) the particle filter provides an approximation of the *joint smoothing distribution*, \( p(x_{1:T} \mid \theta, y_{1:T}) :\)

\[ \hat{p}^N(x_{1:T} \mid \theta, y_{1:T}) := \sum_{i=1}^{N} \frac{w_t^i}{\sum_{\ell} w_{T}^\ell} \delta_{x_{1:T}^i}(x_{1:T}). \]
Path degeneracy
Path degeneracy
Sampling based on the PF

Recall Gibbs sampler:

Want to sample from $p(x_{1:T} \mid \theta, y_{1:T})$ for a fixed value of $\theta$.

With $\mathbb{P}(x_{1:T}^* = x_{1:T}^i) \propto w_T^i$, we get $x_{1:T}^* \approx p(x_{1:T} \mid \theta, y_{1:T})$. 

[Graph showing a time series plot with state on the y-axis and time on the x-axis, illustrating the sampling process.]
Particle MCMC idea

Problems with this approach,

- Based on a PF $\Rightarrow$ approximate sample.
- $N \to \infty$ required to get a correct MCMC.
- A lot of wasted computations.

Analyze PF + MCMC together $\Rightarrow$ Particle MCMC

Particle Gibbs sampling
Let $x_{1:T}' = (x_1', \ldots, x_T')$ be a fixed reference trajectory.

- Sample only $N-1$ particles in the standard way.
- Set the $N$th particle deterministically: $x_t^N = x_t'$ and $a_t^N = N$.

Particle Gibbs
Particle Gibbs

- Sampling $x_{1:T}^*$ with $\mathbb{P}(x_{1:T}^* = x_{1:T}^i) \propto w_T^i$ stochastically “maps” $x'_{1:T}$ into $x_{1:T}^*$.

- Implicitly defines a Markov kernel (the PG kernel) on $X^T$, the space of state trajectories.
Theorem (Stationary distribution)

*The joint smoothing distribution* \( p(x_{1:T} \mid \theta, y_{1:T}) \) *is a stationary distribution of the PG kernel for any* \( N \geq 1 \).


Theorem (Uniform ergodicity)

*The PG kernel is uniformly ergodic under weak conditions.*

*Additionally, under strong mixing conditions the convergence rate does not degenerate as* \( T \to \infty \) *provided that* \( N \propto T \).

ex) Particle Gibbs

Stochastic volatility model,

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

Consider the ACF of \( \theta[k] - \mathbb{E}[\theta | y_{1:T}] \).
Ancestor Sampling

Particle Gibbs:

Let $x'_{1:T} = (x'_1, \ldots, x'_T)$ be a fixed reference trajectory.

- Sample only $N - 1$ particles in the standard way.
- Set the $N$th particle deterministically: $x^N_t = x'_t$.
- Set $a^N_t = N$.
- Sample $a^N_t \in \{1, \ldots, N\}$ with

$$
P(a^N_t = j) \propto w^j_t f_\theta(x'_t | x^j_{t-1}).$$

---

Particle Gibbs with Ancestor Sampling
Particle Gibbs with Ancestor Sampling
PGAS vs. PG

PGAS

PG
ex) Stochastic volatility model, cont’d

Stochastic volatility model,

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

Consider the ACF of \( \theta[k] - \mathbb{E}[\theta \mid y_{1:T}] \).
ex) Wiener system identification

\[ y_t = G(u_t) + h(\cdot) + e_t \]

**Semi-parametric model:** State-space model for \( G \), Gaussian process model for \( h(\cdot) \).

---


fredrik.lindsten@it.uu.se
ex) Gaussian process state-space models

A GP-SSM is a flexible nonparametric dynamical systems model,

\[ f(\cdot) \sim \mathcal{GP}(m_\theta(x_t), k_\theta(x_t, x'_t)), \]
\[ x_{t+1} \mid f, x_t \sim \mathcal{N}(f(x_t), Q), \]
\[ y_t \mid x_t \sim g_\theta(y_t \mid x_t). \]

Idea: Marginalize out \( f(\cdot) \) and infer \( x_{1:T} \) (and \( \theta \)) directly.

Marginalization of \( f \) introduces non-Markovian dependencies in \( \{x_t\} \).

**ex) Epidemiological forecasting**

**Figure:** Disease activity over an eight year period. First half used for inference, second half for one-month predictions.

Inference (and prediction) for SIR model with environmental noise and seasonal fluctuations

\[
S_{t+dt} = S_t + \mu P dt - \mu S_t dt - (1 + F v_t) \beta_t S_t P^{-1} I_t dt,
\]

\[
I_{t+dt} = I_t - (\gamma + \mu) I_t dt + (1 + F v_t) \beta_t S_t P^{-1} I_t dt,
\]

\[
R_{t+dt} = R_t + \gamma I_t dt - \mu R_t dt,
\]

Fredrik Lindsten

Uppsala University
Applications of PGAS

- Oceanography [Marcos et al., JGR, 2015]
- Econometrics [Nonejad, Economics Letters, 2015]
- Synaptic plasticity models [Linderman et al., NIPS, 2014]
- Probabilistic programming [Meent et al., AISTATS, 2015]
- Power disaggregation [Valera et al., NIPS, 2015]
- ...
Ongoing: Blocking

- Update *blocks of states* using PG(AS).
- Stable as $T \to \infty$ using fixed $N$.
- Opens up for parallelisation.
- Possible to do apply blocking also over components of the state vector (spatial blocking).

Ongoing: Spatio-temporal models

Spatio-temporal model:

- State $x_t$ spatially structured
- High dimensionality prevents conventional methods!
- Using multiple Nested Particle Filters we can handle dimensions $\sim 100 - 1000$.

Figure: Drought probability in Sahel region in 1989.

Ongoing: Probabilistic graphical models

- Solving static problems using sequential methods.
- Divide-and-Conquer PF – extending particle filtering from chains to trees
- In particular: hierarchical Bayesian models
- Applications in genetics, linguistics, social science, ...
A framework for studying synaptic plasticity with neural spike train data.

M. Marcos, F. M. Calafat, A. Berihuete, and S. Dangendorf.
Long-term variations in global sea level extremes.

N. Nonejad.
Flexible model comparison of unobserved components models using
particle Gibbs with ancestor sampling.

I. Valera, F. Ruiz, L. Svensson, and F. Perez-Cruz.
Infinite factorial dynamical model.

J.-W. van de Meent, Y. Hongseok, V. Mansinghka, and F. Wood.
Particle Gibbs with ancestor sampling for probabilistic programs.
In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.