Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 8
Markov chain Monte Carlo I
February 13, 2020

Plan of today's lecture

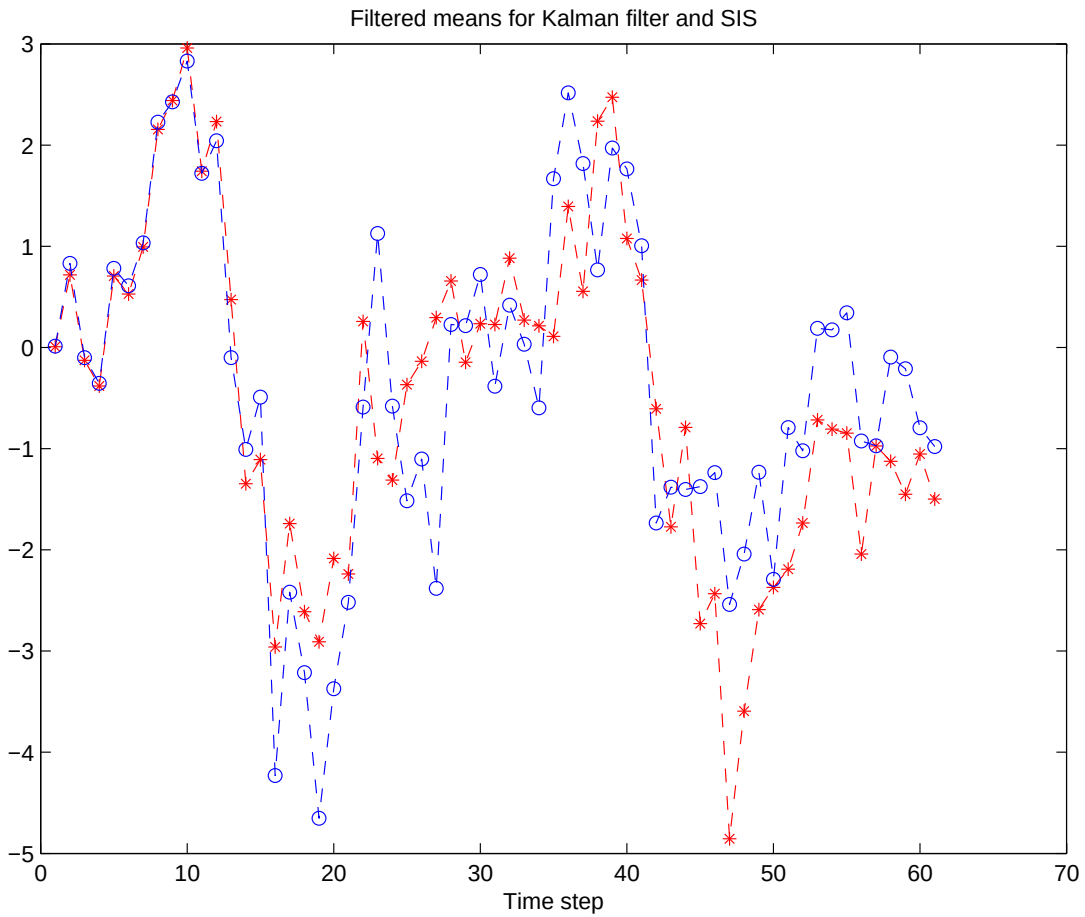
1. Last time: sequential Monte Carlo methods

2. Markov chain Monte Carlo (MCMC, Ch. 7)
   - Overview of MCMC
   - More on Markov chains
   - What's next?

We are here $\longrightarrow \bullet$

Example: Linear/Gaussian HMM, SIS implementation

Comparison of SIS ($\circ$) with exact values ($*$) provided by the Kalman filter (possible only for linear/Gaussian models):



Multinomial resampling

A simple—but revolutionary!—idea: duplicate/kill particles with large/small weights! (Gordon *et al.*, 1993)

The most natural approach to such selection is to simply draw, with replacement, new particles $\tilde{X}_1^{0:n}, \tilde{X}_2^{0:n}, \ldots, \tilde{X}_N^{0:n}$ among the SIS particles $X_1^{0:n}, X_2^{0:n}, \ldots, X_N^{0:n}$ with probabilities given by the normalized importance weights.

Formally, this amounts to set, for $i = 1, 2, \ldots, N$,

$$\tilde{X}_i^{0:n} = X_j^{0:n} \text{ with probability } \frac{\omega_n^j}{\sum_{\ell=1}^N \omega_n^\ell}.$$

Multinomial resampling (cont.)

After this, the resampled particles $(\tilde{X}_i^{0:n})$ are assigned equal weights $\tilde{\omega}_n^i = 1$, say, and we replace

$$\sum_{i=1}^N \frac{\omega_n^i}{\sum_{\ell=1}^N \omega_n^\ell} \phi(X_i^{0:n}) \quad \text{by} \quad \frac{1}{N} \sum_{i=1}^N \phi(\tilde{X}_i^{0:n}).$$

Multinomial resampling does not add bias to the estimator:

Theorem

For all $N \geq 1$ and $n \geq 0$,

$$\mathbb{E}\left(\frac{1}{N} \sum_{i=1}^N \phi(\tilde{X}_i^{0:n}) \right) = \mathbb{E}\left(\sum_{i=1}^N \frac{\omega_n^i}{\sum_{\ell=1}^N \omega_n^\ell} \phi(X_i^{0:n}) \right).$$

The operation adds however some variance due to additional randomness.

Sequential importance sampling with resampling (SISR)

After selection, we proceed with standard SIS and move the selected particles $(\tilde{X}_i^{0:n})$ according to $g_{n+1}(x_{n+1}|x_{0:n})$.

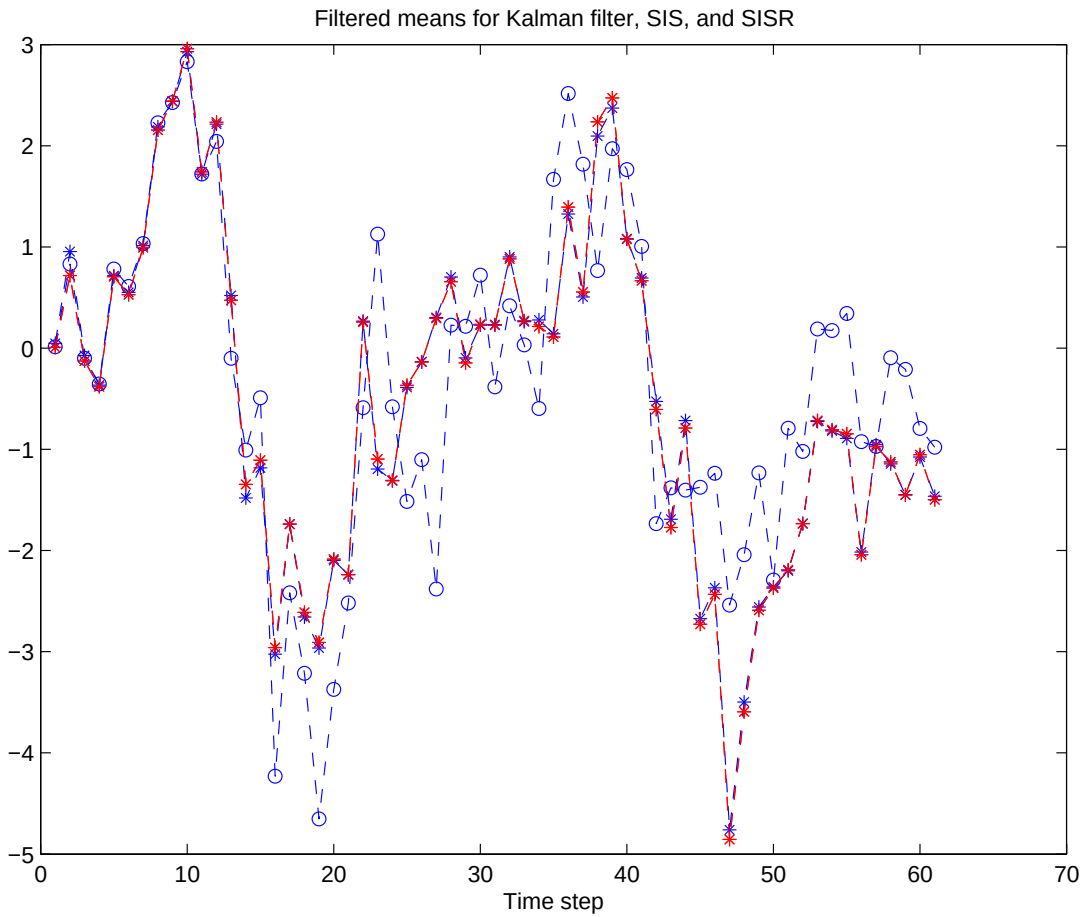
The full scheme goes as follows. Given $(X_i^{0:n}, \omega_n^i)$,

- (*selection*) draw, with replacement, $(\tilde{X}_i^{0:n})$ among $(X_i^{0:n})$ according to probabilities $(\omega_n^i / \sum_{\ell=1}^N \omega_n^\ell)$
- (*mutation*) draw, for all i, $X_i^{n+1} \sim g_{n+1}(x_{n+1}|\tilde{X}_i^{0:n})$,
- set, for all i, $X_i^{0:n+1} = (\tilde{X}_i^{0:n}, X_i^{n+1})$, and
- set, for all i,

$$\omega_{n+1}^i = \frac{z_{n+1}(X_i^{0:n+1})}{z_n(\tilde{X}_i^{0:n}) g_{n+1}(X_i^{n+1}|\tilde{X}_i^{0:n})}.$$

Example: Linear/Gaussian HMM, SIS implementation

Comparison of SIS ($\circ$) and SISR ($*$, blue) with exact values ($*$, red) provided by the Kalman filter:



We are here $\longrightarrow \bullet$

We are here $\longrightarrow$ •

Markov Chain Monte Carlo (MCMC)

- Basic idea: To sample from a density f we construct a Markov chain having f as stationary distribution. A law of large numbers for Markov chains guarantees convergence.
- If f is complicated and/or defined on a space of high dimension this is often easier than transformation methods and rejection sampling.
- The samples will however not be independent.
- MCMC is currently the most common method for sampling from complicated and/or high dimensional distributions.
- Dates back to the 1950's with two key papers being
  - *Equations of state calculations by fast computing machines* (Metropolis *et al.*, 1953) and
  - *Monte Carlo sampling methods using Markov chains and their applications* (Hastings, 1970).

We are here $\longrightarrow \bullet$

More on Markov chains

A Markov chain on $\mathsf{X} \subseteq \mathbb{R}^d$ is a family of random variables (= stochastic process) $(X_k)_{k \geq 0}$ taking values in X such that

$$\mathbb{P}(X_{k+1} \in \mathsf{A} | X_0, X_1, \ldots, X_k) = \mathbb{P}(X_{k+1} \in \mathsf{A} | X_k).$$

The density q of the distribution of X_{k+1} given $X_k = x$ is called the transition density of (X_k). Consequently,

$$\mathbb{P}(X_{k+1} \in \mathsf{A} | X_k = x_k) = \int_{\mathsf{A}} q(x_{k+1} | x_k) \, \mathrm{d}x_{k+1}.$$

As a first example we considered an AR(1) process:

$$X_0 = 0, \quad X_{k+1} = \alpha X_k + \epsilon_{k+1},$$

where α is a constant and (ϵ_k) are i.i.d. noise variables.

Markov chains (cont.)

The following theorem provides the joint density $f(x_0, x_1, \ldots, x_n)$ of $X_0, X_1, \ldots, X_n$.

Theorem

Let (X_k) be Markov with initial distribution χ. Then for $n > 0$,

$$f(x_0, x_1, \ldots, x_n) = \chi(x_0) \prod_{k=0}^{n-1} q(x_{k+1}|x_k).$$

Corollary (Chapman-Kolmogorov equation)

Let (X_k) be Markov. Then for $n > 1$,

$$f(x_n|x_0) = \int \cdots \int \left(\prod_{k=0}^{n-1} q(x_{k+1}|x_k) \right) dx_1 \cdots dx_{n-1}.$$

Stationary Markov chains

A distribution $\pi(x)$ is said to be stationary if

$$\int q(x|z)\pi(z)\mathrm{d}z = \pi(x). \quad \text{(Global balance)}$$

For a stationary distribution π it holds that

$$\begin{aligned}
&\chi = \pi \\
&\Rightarrow f(x_1) = \int q(x_1|x_0)\chi(x_0)\,\mathrm{d}x_0 = \int q(x_1|x_0)\pi(x_0)\,\mathrm{d}x_0 = \pi(x_1) \\
&\Rightarrow f(x_2) = \int q(x_2|x_1)f(x_1)\,\mathrm{d}x_1 = \int q(x_2|x_1)\pi(x_1)\,\mathrm{d}x_1 = \pi(x_2) \\
&\Rightarrow \ldots \Rightarrow f(x_n) = \pi(x_n), \quad \forall n.
\end{aligned}$$

Thus, if the chain starts in the stationary distribution, it will always stay in the stationary distribution. In this case we call also the chain stationary.

Local balance

Let (X_k) have transition density q and let $\lambda(x)$ be a distribution satisfying the local balance condition

$$\lambda(x)q(z|x) = \lambda(z)q(x|z), \quad \forall x, z \in \mathsf{X}.$$

Interpretation:

$$\text{"flow" from state } x \to z = \text{"flow" from state } z \to x.$$

Theorem

Assume that λ satisfies local balance. Then λ is a stationary distribution.

The converse is not true.

Ergodic Markov chains

A Markov chain (X_n) with stationary distribution π is called ergodic if for all initial distributions χ,

$$\sup_{A \subseteq X} |\mathbb{P}(X_n \in A) - \pi(A)| \to 0, \quad \text{as} \quad n \to \infty.$$

Theorem (Geometric ergodicity)

Assume that there exists a density $\mu(x)$ and a constant $\epsilon > 0$ such that for all $x, z \in X$,

$$q(z|x) \geq \epsilon \mu(z). \quad (*)$$

Then the chain (X_n) is geometrically ergodic, i.e. there is $\rho < 1$ such that for all χ,

$$\sup_{A \subseteq X} |\mathbb{P}(X_n \in A) - \pi(A)| \leq \rho^n.$$

Geometrically ezrgodic Markov chains

In other words, geometric ergodicity means that the chain forgets its initial distribution geometrically fast. Two remarks:

- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set). It can however be weakened considerably to hold also for non-compact state spaces.
- Geometric ergodicity implies in general that

$$|\mathbb{C}(\phi(X_m), \phi(X_n))| \leq C\tilde{\rho}^{|n-m|}$$

for some $\tilde{\rho} < 1$ and some constant $C > 0$ depending on ϕ.

A coupling-based proof of geometric ergodicity

Define the transition density

$$\tilde{q}(x_{k+1}|x_k) = \frac{q(x_{k+1}|x_k) - \epsilon\mu(x_{k+1})}{1-\epsilon} \quad (\geq 0 \text{ by } (*))$$

and let χ and χ' be two initial distributions. Define two new Markov chains (X_k) and (X'_k) as follows:

- Draw $X_0 \sim \chi$ and $X'_0 \sim \chi'$.
- given X_k and X'_k, toss an ϵ-coin. If
  1. head (w. pr. ϵ), draw $X_{k+1} \sim \mu(x_{k+1})$ and set $X'_{k+1} = X_{k+1}$ ($\Rightarrow$ *coupling*).
  2. tail (w. pr. $1-\epsilon$), draw $X_{k+1} \sim \tilde{q}(x_{k+1}|X_k)$. In addition, draw independently $X'_{k+1} \sim \tilde{q}(x_{k+1}|X'_k)$; however, if the chains have coupled earlier, keep $X'_{k+1} = X_{k+1}$.

Example: a chain on a discrete set

Let $\mathsf{X} = \{1, 2, 3\}$ and

$$\begin{pmatrix} q(1|1) = 0.4 & q(2|1) = 0.4 & q(3|1) = 0.2 \\ q(1|2) = 0 & q(2|2) = 0.7 & q(3|2) = 0.3 \\ q(1|3) = 0 & q(2|3) = 0.1 & q(3|3) = 0.9 \end{pmatrix}.$$

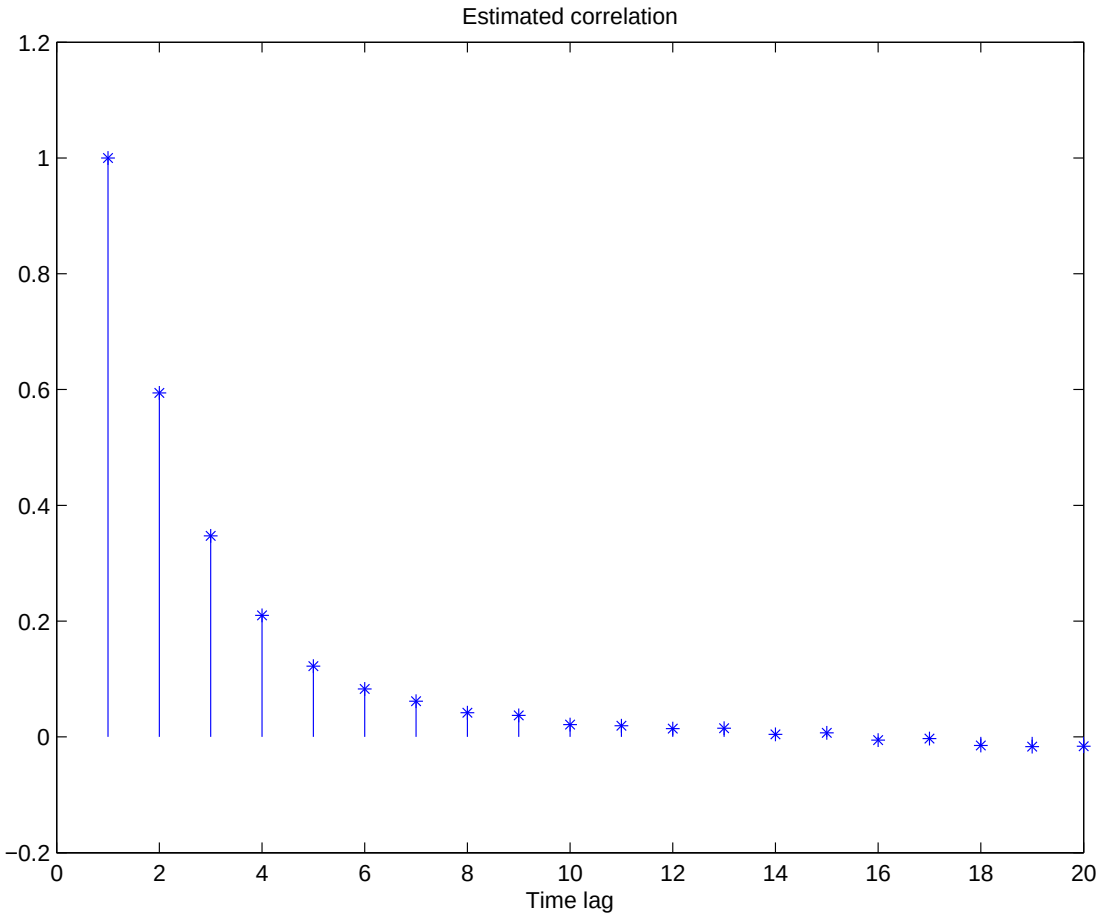
This chain has $\pi = (0, 0.25, 0.75)$ as stationary distribution (check global balance). Moreover, the chain satisfies $(*)$ with

$$\epsilon = 0.3 \quad \text{and} \quad \mu = (0, 1/3, 2/3).$$

It is thus geometrically ergodic.

Example: a chain on a discrete set

Estimated correlation obtained by simulating the chain 1000 time steps:



A law of large numbers for Markov chains

In the case when (X_k) is geometrically ergodic, the states are only weakly dependent. There is thus, just like in the case of independent variables, a law of large numbers:

Theorem (Law of large numbers for Markov chains)

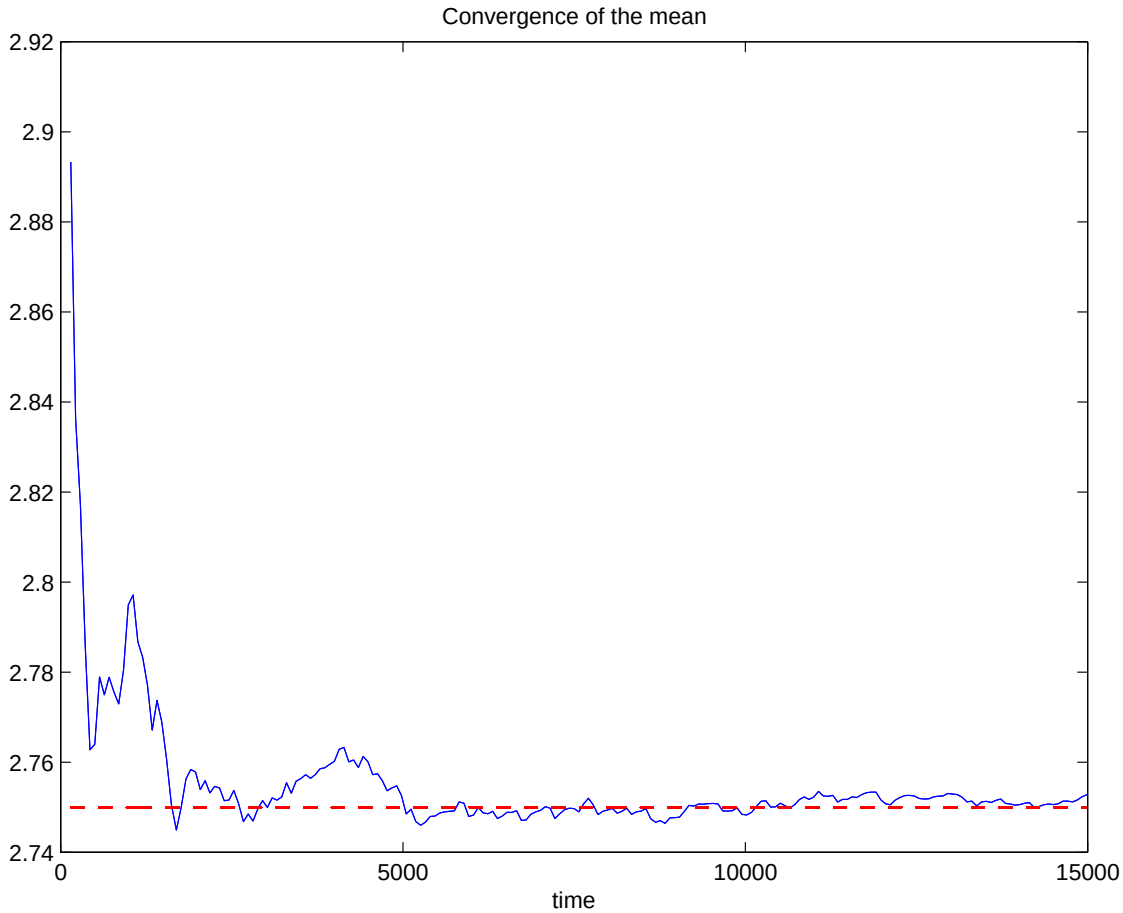
Let (X_n) be a geometrically ergodic Markov chain with stationary distribution π. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\phi(X_k) - \int \phi(x)\pi(x)\,dx\right| \geq \epsilon\right) \to 0 \quad \textit{as} \quad n \to \infty.$$

This means that $\frac{1}{n}\sum_{k=1}^{n}\phi(X_k)$ converges in probability to the mean of $\phi(X_k)$ under π.

Example: a chain on a discrete set reconsidered

Plot of means $\frac{1}{n}\sum_{k=1}^{n} X_k$ with increasing n. Here the mean of the stationary distribution is $1 \cdot 0 + 2 \cdot 0.25 + 3 \cdot 0.75 = 2.75$ (red line).



We are here $\longrightarrow \bullet$

Next week

Now we have gained enough understanding of Markov chains to be able to understand MCMC in some detail.

Thus, next week we will deal with the main objective of MCMC, namely how to, given a density f, construct a Markov chain (X_k) having f as stationary distribution.

Focus will be set on

- the Metropolis-Hastings algorithm and
- the Gibbs sampler.

We will also work out a full example of an implementation.

See you!