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

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 9
Markov chain Monte Carlo II
February 19, 2019

Plan of today's lecture

We are here $\longrightarrow \bullet$

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 high dimensional this is often easier than transformation methods and rejection sampling.
- The price is it that samples will be statistically dependent.
- 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).

Last time: Stationary Markov chains

We called a distribution $\pi(x)$ 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

$$\chi = \pi \Rightarrow f(x_n) = \pi(x_n), \quad \forall n.$$

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:

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

Then the following holds.

Theorem

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

The converse is not true in general.

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 $x, z \in X$,

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

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

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

Example: a chain on a discrete set

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

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

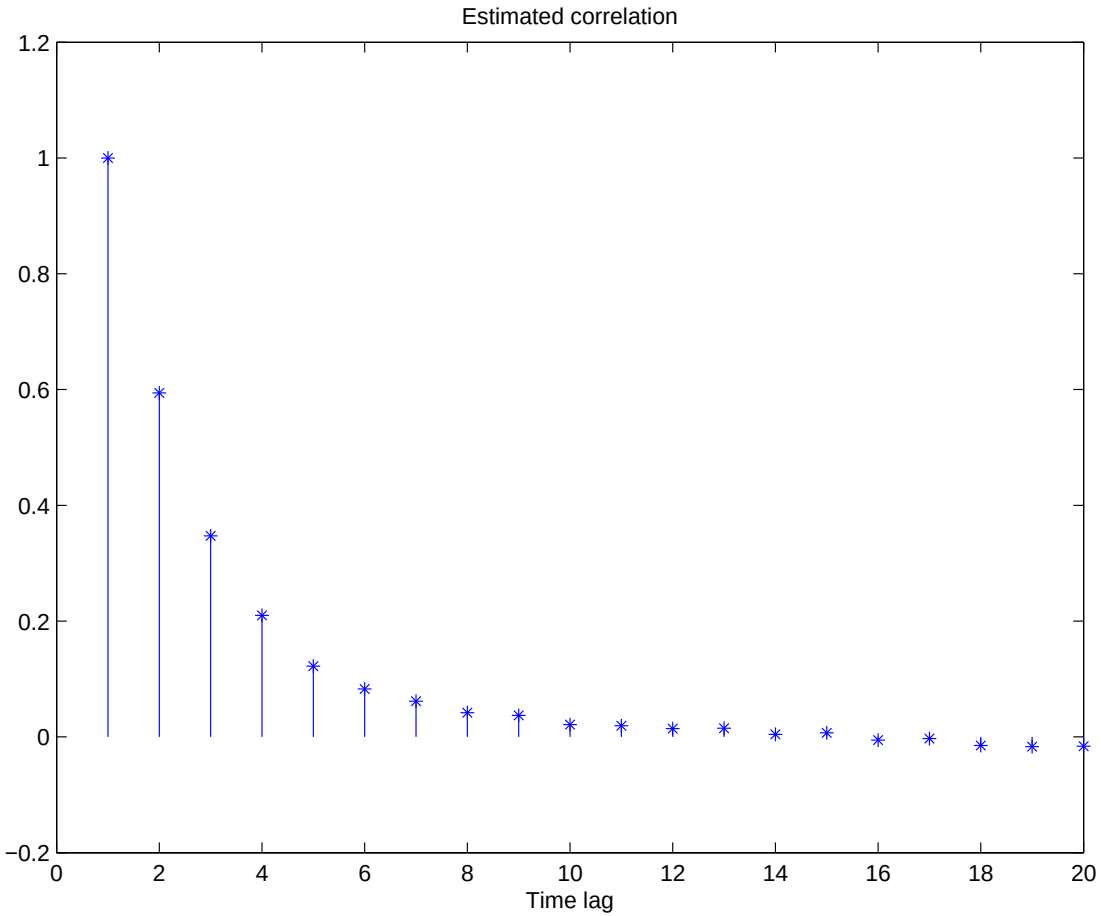
This chain has $\pi = (0, 0.25, 0.75)$ as stationary distribution (check global balance). Moreover, the chain satisfies $(*)$ with

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

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_n) is geometrically ergodic, the states are only weakly dependent. For such Markov chains there is, just like in the case of independent variables, a law of large numbers (LLN):

Theorem (Law of large numbers for Markov chains)

Let (X_n) be a stationary Markov chain. Denote by π the stationary distribution. In addition, assume that

$$\sum_{k=1}^{\infty} |\mathbb{C}\left(\phi(X_1), \phi(X_k)\right)| < \infty. \qquad (**)$$

Then for all $\epsilon > 0$,

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

Proof of LLN (helping Lemma)

Lemma (Chebyshev's inequality)

Let $Z \geq 0$ be a r.v. with $\mathbb{E}Z^2 < \infty$ then

$$\mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}[Z^2]}{\epsilon^2}.$$

Proof.

$$\begin{aligned}\mathbb{P}(Z > \epsilon) &= \mathbb{E}[I(Z > \epsilon)] \leq \mathbb{E}\left[\left(\frac{Z}{\epsilon}\right)^2 I(Z > \epsilon)\right] \\ &\leq \mathbb{E}\left[\left(\frac{Z}{\epsilon}\right)^2\right] = \frac{\mathbb{E}[Z^2]}{\epsilon^2}\end{aligned}$$

□

A law of large numbers for Markov chains (cont.)

In the theorem above,

- the condition $(**)$ is often satisfied for geometrically ergodic Markov chains for which the dependence between states decreases geometrically fast.
- the condition $(**)$ can be weakened considerably (it is however needed for the corresponding CLT; see next lecture).
- the convergence type is called convergence in probability and one typically writes

$$\frac{1}{n}\sum_{k=1}^{n}\phi(X_k) \xrightarrow{\mathbb{P}} \int_{\mathsf{X}} \phi(x)\pi(x)\,\mathrm{d}x \qquad \text{as } n \to \infty.$$

It is possible to extend the LLN to stronger types of convergence, such as convergence a.s.

We are here $\longrightarrow \bullet$

The principle of MCMC

The LLN for Markov chains makes it possible to estimate expectations

$$\tau = \mathbb{E}(\phi(X)) = \int_{\mathsf{X}} \phi(x) f(x) \, \mathrm{d}x$$

by simulating, N steps, a Markov chain (X_k) with stationary distribution f and letting

$$\tau_N = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \to \tau \quad \text{as} \quad N \to \infty.$$

This is the principle of all MCMC methods.

To make this idea practicable requires simulation schemes that guarantee

- that simulating the chain (X_k) is an easily implementable process
- that the stationary distribution of (X_k) indeed coincides with the desired distribution f.
- that the chain (X_k) converges towards f irrespectively of the initial value X_1.

We will now discuss two major classes of such algorithms, namely the Metropolis-Hastings algorithm and the Gibbs sampler.

The Metropolis-Hastings (MH) algorithm

In the following we assume that we can simulate from a transition density $r(z|x)$, referred to as the proposal kernel, on X.

The MH algorithm simulates a sequence of values (X_k), forming a Markov chain on X, through the following mechanism: given X_k,

- generate $X^* \sim r(z|X_k)$ and
- set

$$X_{k+1} = \begin{cases} X^* & \text{w. pr. } \alpha(X_k, X^*) \stackrel{\text{def}}{=} 1 \wedge \dfrac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}, \\ X_k & \text{otherwise.} \end{cases}$$

Here we used the notation $a \wedge b \stackrel{\text{def}}{=} \min\{a, b\}$. The scheme is initialized by drawing X_1 from some initial distribution χ.

The MH algorithm: Pseudo-code

draw $X_1 \sim \chi$
for $k = 1 \rightarrow (N-1)$ **do**
  draw $X^* \sim r(z|X_k)$
  set $\alpha(X_k, X^*) \leftarrow 1 \wedge \frac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}$
  draw $U \sim \mathcal{U}(0,1)$
  if $U \leq \alpha$ **then**
    $X_{k+1} \leftarrow X^*$
  else
    $X_{k+1} = X_k$
  end if
end for
set $\tau_N \leftarrow \sum_{k=1}^{N} \phi(X_k)/N$
return τ_N

A look at $\alpha(X_k, X^*)$

Recall that

$$\alpha(X_k, X^*) = 1 \wedge \frac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}$$

is the probability of accepting the new state X^* given the old state X_k.

First, ignore the transition kernel r. Then

- the ratio $f(X^*)/f(X_k)$ says: accept (keep) the proposed state X^* if it is "better" than the old state X_k (as measured by f);
- otherwise, if the proposed state is "worse" than the old one, accept it with a probability proportional to how much worse.

A look at $\alpha(X_k, X^*)$

Recall again

$$\alpha(X_k, X^*) = 1 \wedge \frac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}.$$

At the same time we also want to explore the state space, where some states may be easier to reach than others. This is compensated for by the factor $r(X_k|X^*)/r(X^*|X_k)$.

- If it is easy to reach X^* from X_k, the denominator $r(X^*|X_k)$ will reduce the acceptance probability;
- if it is easy to get back to X_k from X^*, the numerator $r(X_k|X^*)$ will increase the acceptance probability.

Convergence of the MH algorithm

The following result is fundamental.

Theorem (Global balance of the MH sampler)

The chain (X_k) generated by the MH sampler has f as stationary distribution.

In addition, one may prove, under weak assumptions, that the MH algorithm is also geometrically ergodic, implying that, as $N \to \infty$,

$$\tau_N = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \to \tau = \int \phi(x) f(x) \, \mathrm{d}x.$$

Given some starting value X_1, there will be, say, B iterations before the distribution of the chain can be considered as "sufficiently close" to the stationary distribution. The values $(X_k)_{k=1}^{B}$ are referred to as burn-in and are typically discarded in the analysis.

different types of proposal kernels

There are a number of different ways of constructing the proposal kernel r. The three main classes are

- independent proposals,
- symmetric proposals, and
- multiplicative proposals.

Independent proposal

Independent proposals are characterized as follows.

- Draw the candidates from $r(z)$, i.e. independently of the current state x.
- The acceptance probability reduces to

$$\alpha(x, z) = 1 \wedge \frac{f(z)r(x)}{f(x)r(z)}.$$

- Here it is required that $\{x : f(x) > 0\} \subseteq \{x : r(x) > 0\}$ to ensure convergence.
- If we take $r(x) = f(x)$, which is of course infeasible in general, the acceptance probability reduces to 1 and we get independent samples from f.

Symmetric proposal

Symmetric proposals are characterized by the following.

- It holds that $r(z|x) = r(x|z)$, $\forall (x, z) \in \mathsf{X}^2$.
- In this case the acceptance probability reduces to

$$\alpha(x, y) = 1 \wedge \frac{f(z)}{f(x)}.$$

- Commonly this corresponds to $X^* = X_k + \epsilon$ (random walk proposal) with, e.g.,
  - $\epsilon \in \mathcal{N}(0, \sigma^2)$ or
  - $\epsilon \in \mathcal{U}(-a, a)$.

Multiplicative proposals

An easy way to obtain an asymmetric proposal where the size of the jump depends on the current state $X_k = x$ is to take

$$X^* = x\epsilon,$$

where ϵ is drawn from some density p.

The proposal kernel now becomes $r(z|x) = p(z/x)/x$ and the acceptance probability becomes

$$\alpha(x, z) = 1 \wedge \frac{f(z)p(x/z)/z}{f(x)p(z/x)/x}.$$

Example: a tricky integral

Since the target density f enters the acceptance probability $\alpha(x, z)$ only via the ratio $f(z)/f(x)$, we only need to know f up to a normalizing constant (cf. rejection sampling). This is one of the main strengths of the MH sampler.

As an example we estimate the variance $\tau = \mathbb{E}(X^2)$ under

$$f(x) = \exp(\cos^2(x))/c, \quad x \in (-\pi/2, \pi/2),$$

where $c > 0$ is unknown, using the MH algorithm.

We propose new candidates according to a simple symmetric random walk initialized in the origin, i.e.

$$r(z|x) = \mathcal{N}(z; x, \sigma^2)$$

and $X_1 = 0$.

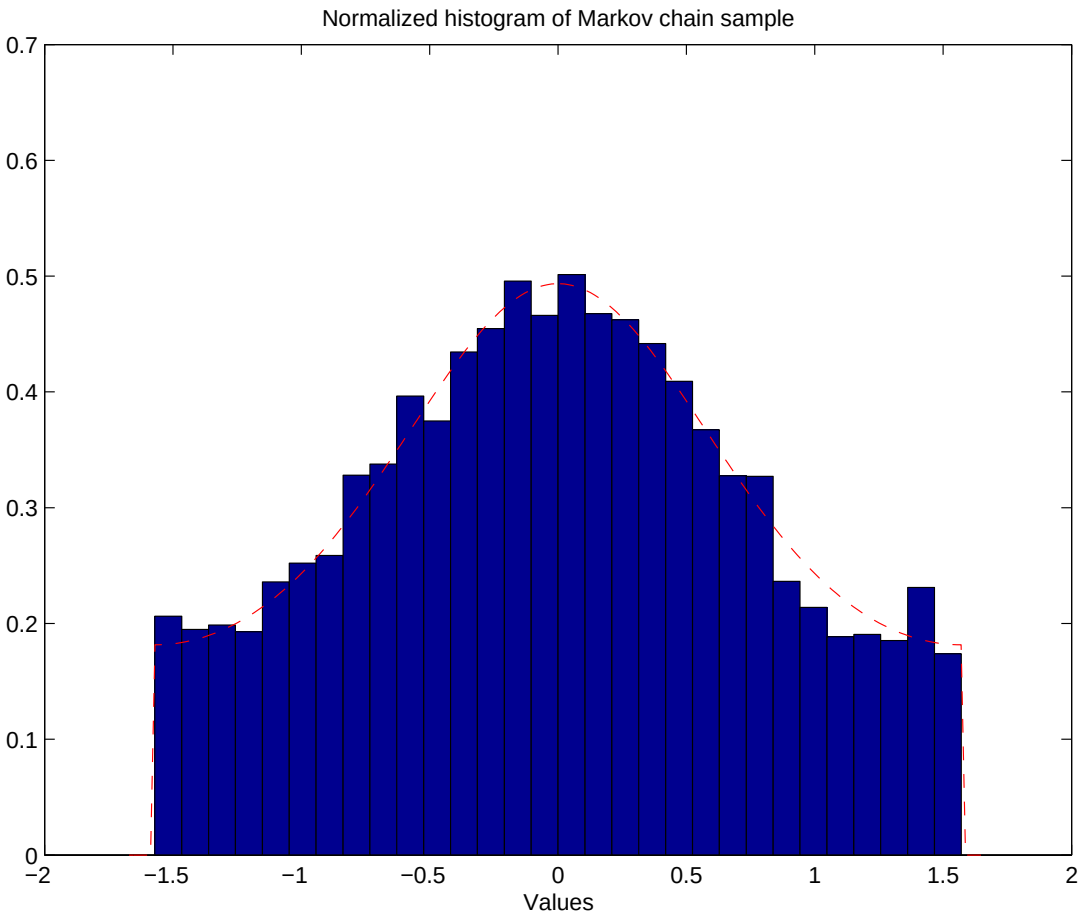
Example: a tricky integral (cont.)

In Matlab:

```
z = @(x) exp(cos(x).^2).*(x > -pi/2).*(x < pi/2);
burn_in = 2000;
M = N + burn_in
X = zeros(1,M);
X(1) = 0;
for k = 1:(M - 1),
    cand = X(k) + randn*sigma;
    alpha = z(cand)/z(X(k));
    if rand <= alpha,
        X(k + 1) = cand;
    else
        X(k + 1) = X(k);
    end
end
tau = mean(X(burn_in:M).^2);
```

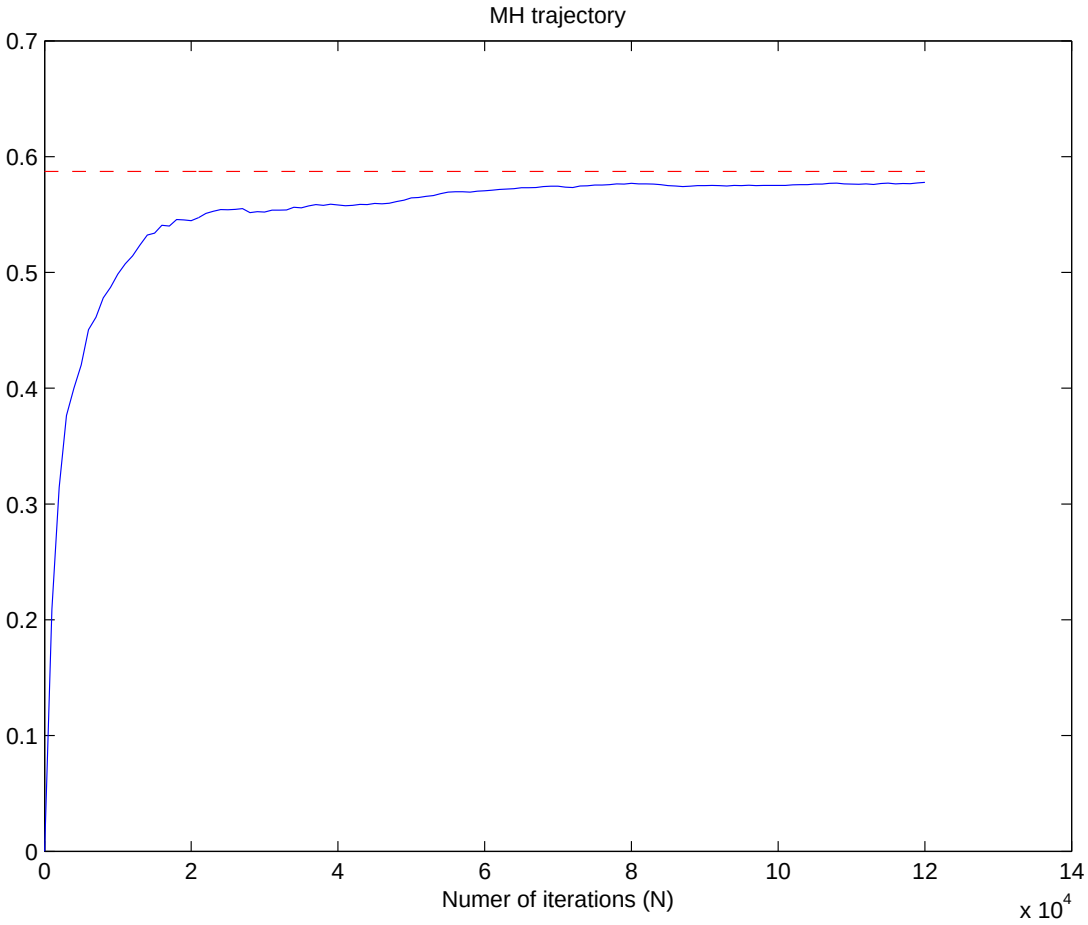
Example: a tricky integral (cont.)

Comparison between the true density and the histogram of X_k, $k = 2001, \ldots, 22000$.



Example:a tricky integral (cont.)

MH output (τ_N) for increasing N (blue) and true value (red):



We are here $\longrightarrow \bullet$

A few comments on HA 1

- Always provide numerical values (not only figures), preferably in a table.
- Focus on describing precisely how you obtained your results rather than on describing the general theory. But be concise!
- Analyze your results.
- A figure caption cannot almost never be too long!
- Check your importance weights properly!

Next time

Next time we will

- prove the global balance theorem above and
- move on to the Gibbs sampler.