# 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 18, 2020

Plan of today's lecture



# 2 The Metropolis-Hastings algorithm (Ch. 5.3)

# 3 Comments on HA 1



# 1 Last time: Introduction to MCMC

# 2 The Metropolis-Hastings algorithm (Ch. 5.3)

# 3 Comments on HA 1

# 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

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

For a stationary distribution  $\pi$  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:

"flow" from state  $x \to z =$  "flow" from state  $z \to x$ .

Then the following holds.

#### Theorem

Assume that  $\lambda$  satisfies local balance. Then  $\lambda$  is a stationary distribution.

The converse is not true in general.

#### Ergodic Markov chains

A Markov chain  $(X_n)$  with stationary distribution  $\pi$  is called ergodic if for all initial distributions  $\chi$ ,

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

### Theorem (Geometric ergodicity)

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

$$q(z|x) \ge \epsilon \mu(z). \tag{(*)}$$

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

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

Example: a chain on a discrete set

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

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

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  $\pi$  the stationary distribution. In addition, assume that

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

Then for all  $\epsilon > 0$ ,

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

Proof of LLN (helping Lemma)

Lemma (Chebyshev's inequality) Let  $Z \geq 0$  be a r.v. with  $\mathbb{E}Z^2 < \infty$  then

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

Proof.

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

# 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

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

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



### Last time: Introduction to MCMC

# 2 The Metropolis-Hastings algorithm (Ch. 5.3)

# 3 Comments on HA 1

# 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  $\left(X_k\right)$  with stationary distribution f and letting

$$au_N = rac{1}{N} \sum_{k=1}^N \phi(X_k) o au$$
 as  $N o \infty.$ 

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

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

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  $\chi$ .

# The MH algorithm: Pseudo-code

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

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

#### Recall that

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

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

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

- 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  $\epsilon$  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  $au = \mathbb{E}(X^2)$  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 = Q(x) \exp(\cos(x) \cdot 2) \cdot (x > -pi/2) \cdot (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  $(\tau_N)$  for increasing N (blue) and true value (red):





### Last time: Introduction to MCMC

# 2 The Metropolis-Hastings algorithm (Ch. 5.3)

# 3 Comments on HA 1

### 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.