Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 9
Markov chain Monte Carlo II
21 April 2017

Plan of today's lecture

Hand in 1

- Some small notes:
  - Make sure to vectorize the code. That is do not use any for loops over the particles!
  - Run the algorithms first on your own simulated data before runing it on the provided data.
- Always provide numerical values (not only figures), preferably in a table.
- Solve the problem!
- 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 be too long!

Outline

1. Last time: Introduction to MCMC
2. The Metropolis-Hastings algorithm (Ch. 5.3)

Last time: 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.

Last time: stationary Markov chains

- We called a distribution π stationary if

$$\int q(x \mid z)\pi(z)\,dz = \pi(x) \quad \text{(global balance)}.$$

- For a stationary distribution π it holds that

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

  (where χ denotes the initial distribution). 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.

Last time: detailed balance

- Let $(X_k)_{k \geq 0}$ have transition density q and let λ be a distribution satisfying the detailed balance condition

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

- Then the following holds true.

Theorem

Assume that λ satisfies detailed balance for q. Then λ is a stationary distribution for q.

The converse is not true in general.

Last time: detailed balance

- Let $(X_k)_{k \geq 0}$ have transition density q and let λ be a distribution satisfying the detailed balance condition

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

- Then the following holds true.

Theorem

Assume that λ satisfies detailed balance for q. Then λ is a stationary distribution for q.

The converse is not true in general.

Last time: ergodic Markov chains

- We introduced the following definitions.

Definition

A Markov chain $(X_n)_{n\geq 0}$ with stationary distribution π is called

(i) ergodic if for all initial distributions χ,

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

(ii) uniformly ergodic if there is $\rho < 1$ such that for all initial distributions χ,

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

Last time: ergodic Markov chains (cont.)

- The following theorem provides geometric ergodicity under the so-called Doeblin condition $(*)$:

Theorem (uniform ergodicity)

Assume that there exists a density μ and a constant $\varepsilon > 0$ such that for all $x, z \in \mathsf{X}$,

$$q(z \mid x) \geq \varepsilon \mu(z). \quad (*)$$

Then the chain $(X_n)_{n \geq 0}$ is uniformly ergodic for

$$\rho = 1 - \varepsilon.$$

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

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

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

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

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

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

Uniformly ergodic Markov chains

- In other words, uniform ergodicity means that the chain forgets its initial distribution geometrically fast.
- The condition $(*)$ is typically satisfied when X is compact (which is e.g. the case when X is finite set); the previous result can however be established under weaker versions of the condition that hold also for non-compact state spaces.
- Uniform ergodicity implies in general that for a large class of objective functions ϕ,

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

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

A coupling-based proof

- Define the transition density

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

  and let χ and χ' be two initial distributions.
- Define two new Markov chains $(X_k)_{k \geq 0}$ and $(X'_k)_{k \geq 0}$ as follows:
  - Draw $X_0 \sim \chi$ and $X'_0 \sim \chi'$.
  - given X_k and X'_k, toss an ε-coin. If
    (i) head (w. pr. ε), draw $X_{k+1} \sim \mu(x_{k+1})$ and set $X'_{k+1} = X_{k+1}$ ($\Rightarrow$ *coupling*).
    (ii) tail (w. pr. $1 - \varepsilon$), draw $X_{k+1} \sim \tilde{q}(x_{k+1} \mid X_k)$. In addition, draw independently $X'_{k+1} \sim \tilde{q}(x_{k+1} \mid X'_k)$; however, if the chains have coupled earlier, keep $X'_{k+1} = X_{k+1}$.

A coupling-based proof

- Define the transition density

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

  and let χ and χ' be two initial distributions.
- Define two new Markov chains $(X_k)_{k \geq 0}$ and $(X'_k)_{k \geq 0}$ as follows:
  - Draw $X_0 \sim \chi$ and $X'_0 \sim \chi'$.
  - given X_k and X'_k, toss an ε-coin. If
    - (i) head (w. pr. ε), draw $X_{k+1} \sim \mu(x_{k+1})$ and set $X'_{k+1} = X_{k+1}$ ($\Rightarrow$ *coupling*).
    - (ii) tail (w. pr. $1 - \varepsilon$), draw $X_{k+1} \sim \tilde{q}(x_{k+1} \mid X_k)$. In addition, draw independently $X'_{k+1} \sim \tilde{q}(x_{k+1} \mid X'_k)$; however, if the chains have coupled earlier, keep $X'_{k+1} = X_{k+1}$.

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

$$\varepsilon = 0.2 \quad \text{and} \quad \mu = (0, 0.5, 0.5).$$

It is thus uniformly ergodic.

Example: a chain on a discrete set (cont.)

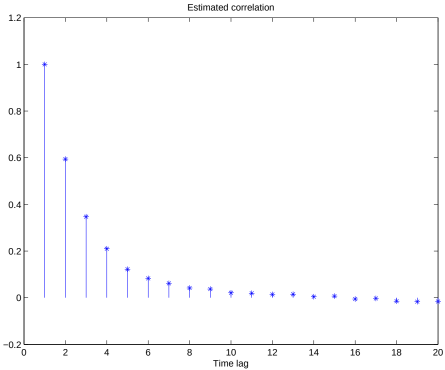

Figure: Estimated correlation obtained by simulating the chain 1000 time steps.

A law of large numbers for Markov chains

- In the case where the states of (X_k) are only weakly dependent there is, just like in the case of independent variables, an LLN:

Theorem (law of large numbers for Markov chains)

Let $(X_n)_{n\geq 0}$ be a stationary Markov chain (with stationary distribution π) and ϕ a function s.t.

$$\mathbb{C}(\phi(X_0), \phi(X_n)) \to 0 \quad \textit{as} \quad n \to \infty.$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} \phi(X_k) \overset{\mathbb{P}}{\to} \int \phi(x)\pi(x)\,dx \quad \textit{as} \quad n \to \infty.$$

Example: a chain on a discrete set reconsidered

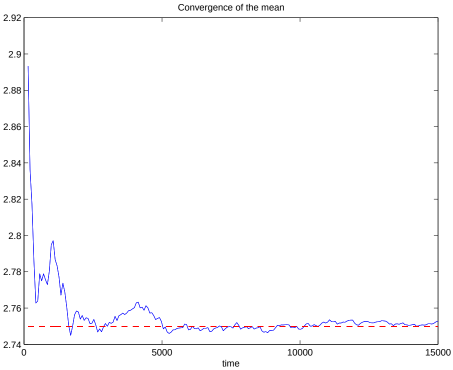

Figure: Plot of means $\frac{1}{n}\sum_{k=1}^{n} X_k$ with increasing n. Here the mean of the stationary distribution is $1 \cdot 0 + 2 \cdot 0.25 + 3 \cdot 0.75 = 2.75$ (red line).

Outline

1 Last time: Introduction to MCMC

2 The Metropolis-Hastings algorithm (Ch. 5.3)

The principle of MCMC

- The LLN for Markov chains makes it possible to estimate expectations

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

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

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

This is the main principle of MCMC methods.

The principle of MCMC (cont.)

- In order for the approach to be practically useful, we require that
  - simulating the chain (X_k) is an easily implementable process.
  - the stationary distribution of (X_k) coincides indeed with the desired distribution f.
  - the chain (X_k) converges to f irrespectively of the initial value X_1.
  - the target density f needs to be known only up to a normalizing constant.
- We will discuss two major classes of such algorithms, namely the Metropolis-Hastings algorithm (today) and the Gibbs sampler (next lecture).

The principle of MCMC (cont.)

- In order for the approach to be practically useful, we require that
  - simulating the chain (X_k) is an easily implementable process.
  - the stationary distribution of (X_k) coincides indeed with the desired distribution f.
  - the chain (X_k) converges to f irrespectively of the initial value X_1.
  - the target density f needs to be known only up to a normalizing constant.
- We will discuss two major classes of such algorithms, namely the Metropolis-Hastings algorithm (today) and the Gibbs sampler (next lecture).

The Metropolis-Hastings (MH) algorithm

- In the following we assume that we are able to simulate from a transition density $r(z \mid x)$, referred to as the proposal kernel, on X.
- The MH algorithm simulates recursively a sequence of draws (X_k), forming a Markov chain on X, through the following mechanism: given X_k,
  - draw $X^* \sim r(z \mid X_k)$ and
  - set $X_{k+1} = \begin{cases} X^* & \text{w. pr. } \alpha(X_k, X^*) \overset{\text{def}}{=} 1 \wedge \dfrac{f(X^*) r(X_k \mid X^*)}{f(X_k) r(X^* \mid X_k)}, \\ X_k & \text{otherwise.} \end{cases}$

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

The Metropolis-Hastings (MH) algorithm

- In the following we assume that we are able to simulate from a transition density $r(z \mid x)$, referred to as the proposal kernel, on X.
- The MH algorithm simulates recursively a sequence of draws (X_k), forming a Markov chain on X, through the following mechanism: given X_k,
  - draw $X^* \sim r(z \mid X_k)$ and
  - set $X_{k+1} = \begin{cases} X^* & \text{w. pr. } \alpha(X_k, X^*) \stackrel{\text{def}}{=} 1 \wedge \dfrac{f(X^*)r(X_k \mid X^*)}{f(X_k)r(X^* \mid X_k)}, \\ X_k & \text{otherwise.} \end{cases}$

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

The MH algorithm: pseudo-code

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

A closer look at α

- Recall that

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

is the probability of accepting the candidate 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 only with a probability proportional to $f(X^*)/f(X_k)$.

A closer look at α

- Recall that

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

  is the probability of accepting the candidate 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 only with a probability proportional to $f(X^*)/f(X_k)$.

A closer look at α (cont.)

- 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 $r(X_k \mid X^*)/r(X^* \mid X_k)$ in the acceptance probability:

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

- Consequently,
  - if it is easy to reach X^* from X_k, the denominator $r(X^* \mid X_k)$ will reduce the acceptance probability;
  - if it is easy to return to X_k from X^*, the numerator $r(X_k \mid X^*)$ will increase the acceptance probability.

A closer look at α (cont.)

- 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 $r(X_k \mid X^*)/r(X^* \mid X_k)$ in the acceptance probability:

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

- Consequently,
  - if it is easy to reach X^* from X_k, the denominator $r(X^* \mid X_k)$ will reduce the acceptance probability;
  - if it is easy to return to X_k from X^*, the numerator $r(X_k \mid X^*)$ will increase the acceptance probability.

A closer look at α (cont.)

- 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 $r(X_k \mid X^*)/r(X^* \mid X_k)$ in the acceptance probability:

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

- Consequently,
  - if it is easy to reach X^* from X_k, the denominator $r(X^* \mid X_k)$ will reduce the acceptance probability;
  - if it is easy to return to X_k from X^*, the numerator $r(X_k \mid X^*)$ will increase the acceptance probability.

Convergence of the MH algorithm

- The following result is fundamental.

Theorem (detailed balance of the MH sampler)

The MH sampler satisfies detailed balance for the target density f.

- Consequently, the following holds true.

Corollary (global balance of the MH sampler)

The Markov chain generated by the MH sampler allows f as a stationary distribution.

Convergence of the MH algorithm

- The following result is fundamental.

Theorem (detailed balance of the MH sampler)

The MH sampler satisfies detailed balance for the target density f.

- Consequently, the following holds true.

Corollary (global balance of the MH sampler)

The Markov chain generated by the MH sampler allows f as a stationary distribution.

Convergence of the MH algorithm (cont.)

- The MH algorithm is in general not uniformly ergodic.
- However, under weak assumptions one may prove that the MH algorithm is geometrically ergodic, i.e., there exist $\rho < 1$ and a function C on X such that for all initial states $\chi = \delta_x$,

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

- Also geometrically ergodic Markov chains satisfy the LLN.
- 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.

Convergence of the MH algorithm (cont.)

- The MH algorithm is in general not uniformly ergodic.
- However, under weak assumptions one may prove that the MH algorithm is geometrically ergodic, i.e., there exist $\rho < 1$ and a function C on X such that for all initial states $\chi = \delta_x$,

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

- Also geometrically ergodic Markov chains satisfy the LLN.
- 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.

Convergence of the MH algorithm (cont.)

- The MH algorithm is in general not uniformly ergodic.
- However, under weak assumptions one may prove that the MH algorithm is geometrically ergodic, i.e., there exist $\rho < 1$ and a function C on X such that for all initial states $\chi = \delta_x$,

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

- Also geometrically ergodic Markov chains satisfy the LLN.
- 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.

Convergence of the MH algorithm (cont.)

- The MH algorithm is in general not uniformly ergodic.
- However, under weak assumptions one may prove that the MH algorithm is geometrically ergodic, i.e., there exist $\rho < 1$ and a function C on X such that for all initial states $\chi = \delta_x$,

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

- Also geometrically ergodic Markov chains satisfy the LLN.
- 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.

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

- Using an independent proposal, candidates are drawn from $r(z)$ 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 practice, the acceptance probability reduces to 1 and we get independent samples from f.

Independent proposal

- Using an independent proposal, candidates are drawn from $r(z)$ 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 practice, the acceptance probability reduces to 1 and we get independent samples from f.

Independent proposal

- Using an independent proposal, candidates are drawn from $r(z)$ 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 practice, the acceptance probability reduces to 1 and we get independent samples from f.

Symmetric proposal

- For a symmetric proposal it holds that $r(z \mid x) = r(x \mid z)$ for all $(x, z) \in \mathsf{X}^2$.
- In this case the acceptance probability simplifies to

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

- Commonly this is obtained by letting $X^* = X_k + \varepsilon$ (random walk proposal) with, e.g.,
  - $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ or
  - $\varepsilon \sim \mathrm{U}(-a, a)$.

Symmetric proposal

- For a symmetric proposal it holds that $r(z \mid x) = r(x \mid z)$ for all $(x, z) \in \mathsf{X}^2$.
- In this case the acceptance probability simplifies to

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

- Commonly this is obtained by letting $X^* = X_k + \varepsilon$ (random walk proposal) with, e.g.,
  - $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ or
  - $\varepsilon \sim \mathsf{U}(-a, a)$.

Symmetric proposal

- For a symmetric proposal it holds that $r(z \mid x) = r(x \mid z)$ for all $(x, z) \in \mathsf{X}^2$.
- In this case the acceptance probability simplifies to

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

- Commonly this is obtained by letting $X^* = X_k + \varepsilon$ (random walk proposal) with, e.g.,
  - $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ or
  - $\varepsilon \sim \mathrm{U}(-a, a)$.

Multiplicative proposals

- An easy way of obtaining an asymmetric proposal where the size of the jump depends on the current state $X_k = x$ is to take
$$X^* = x\varepsilon,$$
where ε is drawn from some density p.
- The proposal kernel now becomes $r(z \mid x) = p(z/x)/x$, yielding the acceptance probability
$$\alpha(x, z) = 1 \wedge \frac{f(z)p(x/z)/z}{f(x)p(z/x)/x}.$$

Normalizing constants

- 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 or self-normalized importance sampling).
- This is one of the main strengths of the MH sampler.

Example: the tricky distribution (again)

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

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

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.,

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

and $X_1 = 0$.

Example: the tricky distribution (again)

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

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

  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.,

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

  and $X_1 = 0$.

Example: the tricky distribution (again) (cont.)

```
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.)

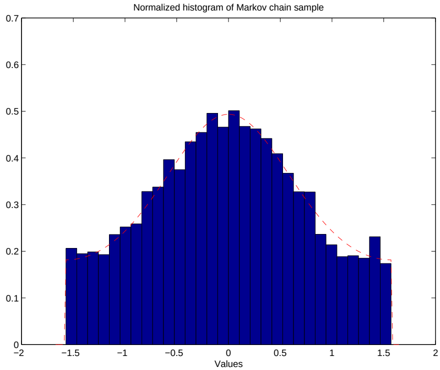

Figure: Comparison between the true density and the histogram of X_k, $k = 2001, \ldots, 22000$.

Example:a tricky integral (cont.)

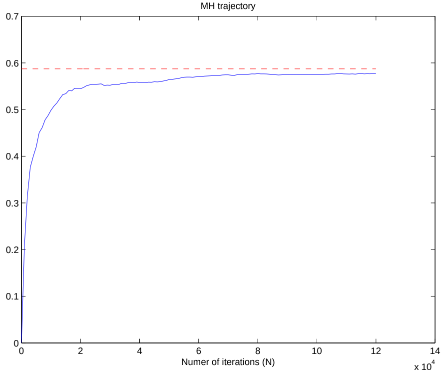

Figure: MH output (τ_N) for increasing N (blue) and true value (red).

Next time

Next time we will

- prove the MH detailed balance theorem above and
- move on to the Gibbs sampler.
- Notice that next week the following change in the regular shcedule:
    - there is a lecture on wednesday
    - the exercise class is on thrusday