

January 3, 2005
EH

**Markov Chains on
general state spaces
fall 2004**

Paper 3

Formal requirements: the paper must be handed in no later than Monday February 14 2004 at noon. The paper must be given to me personally.

The paper can be written in danish or english. It is strongly encouraged that the paper is produced electronically.

It is not prohibited that participants cooperate in the problem solving phase - indeed, it is encouraged. But the final paper must be an individual piece of work.

Ernst Hansen

Markov Chain Monte Carlo

Statement of the problem

Let $(\mathcal{X}, \mathbb{E})$ be measurable space, and let π be a probability measure hereon. This measure is usually given explicitly through some sort of a formula, and so it is “known”. In most cases of interest we have that $\pi = f \cdot \mu$, where μ is a well-known reference measure, and where the density f is explicitly given.

We will refer to π as the **target measure**. The challenge is to compute integrals with respect to π . That is, given a function $g : \mathcal{X} \rightarrow \mathbb{R}$ which is π -integrable (it may for instance be bounded), how do we compute the integral $\int g d\pi$? Just as π is explicitly given, the integrand g will be explicitly given, and we want the integral as a *real number*.

This may not seem like a problem that belongs to probability theory. After all there is a huge mathematical field known as numerical analysis, and a branch of this field is concerned with *numerical integration*. But implicit in the problem is the understanding that \mathcal{X} is high dimensional - say \mathbb{R}^{100} or \mathbb{R}^{1000} - and numerical integration in high dimensions is a very tricky business, where probabilistic methods play a very important role.

Before we describe these probabilistic methods, it is useful to discuss a variation of the problem we have stated. Frequently we must face the complication that the target measure π is not *completely* known. . . It is true that $\pi = f \cdot \mu$, but the density f has the form

$$f(x) = c f_0(x) \tag{1}$$

where f_0 is explicitly given, but where the normalizing constant c is unknown. The problem, formulated with this complication, occurs in numerous statistical examples. In particular, it lies at the heart of Bayesian statistics, where the issue is to make sense of the so-called posterior distribution. One way of ‘making sense’ is to compute integrals. Posterior distributions are inevitably given in terms of unnormalized densities like (1), where f_0 is easily computed, but where the normalizing constant may be very difficult to obtain.

From a certain angle, the complication does not matter much. We observe that

$$\int g(x) d\pi(x) = \frac{\int g(x) f_0(x) d\mu(x)}{\int f_0(x) d\mu(x)},$$

so instead of computing one innocently looking integral, we simply have to compute two. Unfortunately, this may be a lot worse than doubling our problems. The integral in the denominator may be difficult to get right. In many cases g is a 'local' function, perhaps an indicator function, so the numerator integral may only involve integration over a limited part of space. While the denominator integral per definition is global, and involves integration over the entire space.

Monte Carlo methods

The classical idea of *Monte Carlo integration* can be described very easily: If X_1, X_2, \dots are independent \mathcal{X} -valued random variables, each with distribution π , then the strong law of large numbers shows that

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \rightarrow \int g d\pi \quad \text{for } n \rightarrow \infty \quad \text{a.e.}$$

So 'computation' of the integral may be interpreted as the approximation

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \approx \int g d\pi \tag{2}$$

for some large n . Some care has to be taken in order to choose n large enough, but purely deterministic methods of computing the integral will have similar problems - they also rely on replacing the integral with a suitable finite sum.

When Monte Carlo integration can be carried out, it usually performs very well in high dimensional settings - it is not so vulnerable to the so-called *curse of dimensionality* that makes the computing time for deterministic methods explode as the dimension is increased. Magically, it also seems that the complication with unnormalized densities somehow vapourises. At least it is hidden in the way the iid. variables X_1, X_2, \dots are produced.

But unfortunately, simulation of these variables may be impossible. Simulation from an explicitly given density is an easy task in one dimension. But it turns out to be prohibitively difficult in higher dimensions, unless the density has special features that can be exploited.

A strategy for tackling this situation, which has had spectacularly success for the last 10-15 years, is based on the strong law of large numbers for Markov

chains: Construct an update scheme $\phi : \mathcal{X} \times (0, 1) \rightarrow \mathcal{X}$ such that the corresponding Markov kernel P is

- 1) irreducible
- 2) aperiodic
- 3) Harris recurrent

and such that π as an invariant probability measure for P . If we choose X_0 in some suitable way, and generate variables X_1, X_2, \dots from this update scheme, the strong law of large numbers for Markov chains says that

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \rightarrow \int g d\pi \quad \text{for } n \rightarrow \infty \text{ a.e.}$$

irrespective of the distribution of X_0 . And hence we may 'compute' the desired integral by the approximation (2) as before. We stress that we need an update scheme, not just a Markov kernel, because we need to be able to simulate the corresponding Markov chain effectively.

At first it would seem that the Markov Chain Monte Carlo scheme is a joke: If it is too difficult to simulate from π , what hope do we have of construction a Markov kernel P with good ergodic properties **and** with π as invariant distribution, in such a way that we can simulate from P ? Much effort in Markov chain theory is devoted to the understanding of the invariant distribution for a given Markov kernel, and it is usually a very difficult task to make the link between these two concepts. But Markov Chain Monte Carlo somehow turns the tables around, and starts with the initial distribution, and this changes the situation completely. Two general classes of techniques can be used to construct a relevant Markov kernel P surprisingly painless.

The idea of **Hastings modifications** is to take a Markov kernel with reasonable ergodic properties, but with no obvious connection to π , and modify it through an acceptance/rejection step - the details will be explained in great details in the following pages. Proper tuning of the acceptance/rejection step produces a new kernel, that magically has π as invariant distribution. And with a bit of luck, the new kernel inherits (most of) the ergodic properties of the original kernel.

The other idea in the business is generally referred to as **Gibbs sampling**. It is usually quite easy to produce Markov kernels that have π as invariant measure, if we are prepared to accept that these kernels have horrible ergodic

properties - they are typically non-irreducible and whatnot. But if several of these bad kernels can be constructed, they might be combined to a kernel with good ergodic properties while keeping π as invariant measure.

If $\mathcal{X} = \mathbb{R}^2$, Gibbs sampling will typically involve two kernels: Q^1 which constructs updates with fixed x -coordinate, such that we have updates of the form

$$(x, y) \mapsto (x, y')$$

And Q^2 which constructs updates with fixed y -coordinate, such that we get updates of the form

$$(x, y) \mapsto (x', y)$$

In the **random scan** combination of these two kernels the update is done in two steps: an axis is chosen at random, and the corresponding Q^i is used to update along that axis. The **fixed scan** combination, on the other hand, has a full update corresponding to two partial updates: the first partial update is through Q^1 , the second is through Q^2 . The random scan is often easier to analyse, while the fixed scan is somewhat more effective in practise.

Usually Gibbs samplers are very effective, but they can not be made to work in quite as general settings as the Hastings modifications. On the other hand, there are a number of applications, where the two ideas can be used in conjunction, for instance in terms of a fixed scan of several Hastings modified proposal mechanisms.

In this paper we will only pursue a study of Hastings modifications. We will not care about practical matters, such as how long the chain should be simulated, or if it is relevant to include a **burn-in**, that is if the initial (and probably very atypical) observations X_0, X_1, \dots, X_k should be discarded before the timeaverages are taken. We will exclusively be occupied with the demonstration that the acceptance/rejection step in very general situations can be constructed such that Hastings-modified kernels have the desired invariant distributions and have good ergodic properties.

1 Reversibility

A measure λ on $(\mathcal{X} \times \mathcal{X}, \mathbb{E} \otimes \mathbb{E})$ is **reversible** if it is invariant under reflection of the two coordinates. That is, if $\phi(\lambda) = \lambda$, where

$$\phi(x, y) = (y, x) \quad \text{for all } x, y \in \mathcal{X}.$$

PROBLEM 1. Show that λ is reversible if and only if

$$\lambda(A \times B) = \lambda(B \times A) \quad \text{for all } A, B \in \mathbb{E}.$$

PROBLEM 2. Suppose that λ is reversible. Show that

$$\int h(x, y) d\lambda(x, y) = \int h(y, x) d\lambda(x, y)$$

for all non-negative measurable functions h .

We say that a Markov kernel P on \mathcal{X} and a probability measure ν on \mathcal{X} form a **reversible pair** if the integration of P with respect to μ is a reversible measure on $\mathcal{X} \times \mathcal{X}$.

PROBLEM 3. Show that P and ν forms a reversible pair if

$$\int_A P_x(B) d\nu(x) = \int_B P_x(A) d\nu(x) \quad \text{for all } A, B \in \mathbb{E}.$$

PROBLEM 4. Show that if P and ν form a reversible pair, then ν is an invariant distribution for P .

Our strategy in search of good Markov kernels for the Markov Chain Monte Carlo scheme will be to look for Markov kernels P that form a reversible pair with the target distribution π . Lots of kernels will have π as an invariant distribution without forming a reversible pair with it, so in some ways we have made our quest harder with this strategy. But the search will be successful nonetheless.

2 Hastings modifications

Let Q be a Markov kernel on \mathcal{X} , called the **proposal kernel**. Let it have an associated update scheme $\psi : \mathcal{X} \times (0, 1) \rightarrow \mathcal{X}$, which we typically will refer to as the proposal mechanism. Let $\alpha : \mathcal{X} \rightarrow [0, 1]$ be a measurable function, called the **acceptance probability**.

The Hastings modification of Q using α works in the following way: From the initial point x we construct a proposed update y using Q . The true update will be either this proposed update y , or it will be the point x where we started. The choice between these two outcomes is governed by the acceptance probability $\alpha(x, y)$.

Formally, this leads to the update scheme $\phi : \mathcal{X} \times (0, 1)^2 \rightarrow \mathcal{X}$ of the form

$$\phi(x, u_1, u_2) = \begin{cases} y & \text{if } u_2 < \alpha(x, y) \\ x & \text{if } u_2 \geq \alpha(x, y) \end{cases} \quad \text{where } y = \psi(x, u_1),$$

which are supposed to be fed by an initial x and two independent standard uniformly distributed U 's. It is very easy to produce computer code for this Hastings modification step - the hard part is the update scheme for Q .

PROBLEM 5. Show that the Hastings modified kernel P is given as

$$P_x(A) = \int_A \alpha(x, y) dQ_x(y) + (1 - \beta(x)) \epsilon_x(A),$$

where ϵ_x is the one-point measure in x , and where $\beta(x)$ is the **overall acceptance probability** from x , given as

$$\beta(x) = \int \alpha(x, y) dQ_x(y)$$

PROBLEM 6. Show that the Hastings modification of Q using α form a reversible pair with π if

$$\iint h(x, y) \alpha(x, y) dQ_x(y) d\pi(x) = \iint h(y, x) \alpha(y, x) dQ_y(x) d\pi(y) \quad (3)$$

for all non-negative measurable functions h .

Equation (3) is typically referred to as **gross balance**. The challenge is: given the target distribution π and the proposal kernel Q , find an acceptance probability α that satisfies the gross balance condition.

3 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a specific formula for the acceptance probability in a Hastings modification that ensures gross balance. It is applicable in the case where all the measures involved have densities with respect to a common reference measure μ . Assume from now on that

$$\pi = f \cdot \mu, \quad Q_x = g_x \cdot \mu. \quad (4)$$

We suppose that $(x, y) \mapsto g_x(y)$ is measurable.

PROBLEM 7. Let P be the Hastings modification of Q with respect to α . Show that P and π will form a reversible pair if the **detailed balance equation**,

$$\alpha(x, y) f(x) g_x(y) = \alpha(y, x) f(y) g_y(x)$$

is satisfied for $\mu \otimes \mu$ -almost all (x, y) .

PROBLEM 8. Show that the Metropolis-Hastings acceptance probabilities

$$\alpha(x, y) = \min \left\{ \frac{f(y) g_y(x)}{f(x) g_x(y)}, 1 \right\} \quad (5)$$

will make P and π form a reversible pair. How should the formula be interpreted if the numerator and/or denominator of the fraction is zero?

PROBLEM 9. Explain why it is immaterial for the Metropolis-Hastings algorithm whether the target density f is normalised or not.

The user only has to supply an proposal mechanism for the Metropolis-Hastings algorithm. Of course different choices of proposal mechanisms will not work equally well - in some cases there can be huge advantages in choosing the proposal mechanism carefully. But in general, the algorithm is quite robust to the choice of proposals, it is extremely easy to code, and it works like a charm.

If $\mathcal{X} = \mathbb{R}^n$ and $\pi = f \cdot \mu$, a common choice of a proposal mechanism is based on a random walk, where the increments have density with respect to μ . To be specific, let g be a probability density with respect to μ , and define

$$g_x(y) = g(y - x).$$

The proposal kernel Q corresponding to these proposal densities is simply the kernel for the random walk on \mathbb{R}^n with increment distribution $g \cdot \mu$.

PROBLEM 10. Find the acceptance probabilities for a random walk based Metropolis-Hastings algorithm, if the proposal density g is symmetric around 0,

$$g(x) = g(-x) \quad \text{for all } x \in \mathbb{R}^n .$$

The Metropolis-Hastings algorithm requires that both the target and the proposal mechanism has density with respect to a fixed reference measure. Still, the Metropolis-Hastings kernel itself does usually **not** have density with respect to the reference measure, due to the possibility that the moves are rejected. The kernel will usually have a mixed character, with a continuous component (the proposed transitions) and a discrete component (rejected moves). This mixed character makes the kernel quite difficult to handle for people not versed in measure theory.

4 Irreducibility and aperiodicity

In order to obtain rigorous results on the ergodic properties of the Metropolis-Hastings algorithm, we will sharpen the density requirements in (4). We will work under the assumption of strictly positive densities,

$$f(x) > 0, \quad g_x(y) > 0 \quad \text{for all } x, y \in \mathcal{X} . \quad (6)$$

Actually, it is unreasonable restrictive to assume that the proposal densities are strictly positive. Considerably less will do - at the expense of a longer argument. But normal distributions of the relevant dimension are the favourite choice for proposal distributions, and they certainly satisfy the condition. It is harder to relax the assumption that the target density should be positive everywhere. If it is not, the algorithm can have a very problematic behaviour if it is started outside the support of π .

PROBLEM 11. Suppose that π and Q satisfies (6), and let P be the corresponding Metropolis-Hastings kernel. Show that P is irreducible.

PROBLEM 12. Suppose that π and Q satisfies (6), and let P be the corresponding Metropolis-Hastings kernel. Show that P is aperiodic.

We now have a complete proof that Metropolis-Hastings algorithm works, if we can find conditions under which it is Harris recurrent. To aid this discussion, we will examine the notion of harmonic functions.

5 Harmonic functions

In this section we will leave aside the Markov Chain Monte Carlo set-up, and consider a general Markov kernel P . A bounded, measurable function $h : \mathcal{X} \rightarrow \mathbb{R}$ is **harmonic** with respect to P if $\Delta h(x) = 0$ for all x . That is, if

$$\int h(y) dP_x(y) = h(x) \quad \text{for all } x \in \mathcal{X}.$$

PROBLEM 13. Show that if h is bounded and harmonic, and if X_0, X_1, \dots is a P -chain, then $h(X_0), h(X_1), \dots$ is converging almost surely and in L^1 .

Hint: use the martingale convergence theorem.

PROBLEM 14. Suppose that P is Harris recurrent and that h is bounded and harmonic. Show that for any sublevel set of h ,

$$A = \{y \in \mathcal{X} \mid h(y) \leq a\}$$

it holds that if $A \in \mathbb{E}^+$ then $A = \mathcal{X}$.

Hint: Suppose $A \in \mathbb{E}^+$, and consider a fixed $x \in \mathcal{X}$. Choose a P -chain X_0, X_1, \dots with $X_0 = x$. Let Z be the limit of $h(X_n)$. Show that $Z \leq a$ almost surely, and that $EZ = h(x)$.

PROBLEM 15. Show that if P is Harris recurrent and h is bounded and harmonic, then h is constant.

Hint: Use the previous problem on both h and $-h$.

For the remaining problems in this section, it will be useful to recall that if P is recurrent, there is a decomposition of the space

$$\mathcal{X} = H \cup \mathcal{N},$$

such that H is non-empty and absorbing, and such that P restricted to H is Harris recurrent. This is Meyn and Tweedies theorem 9.1.5, where they furthermore have a characterization of H as *maximal Harris* - a notion that need not occupy us, but which essentially makes the decomposition unique.

PROBLEM 16. Show that if P is irreducible and positive with invariant distribution π , then any non-empty absorbing set A must satisfy that $\pi(A) = 1$.

Hint: Show that

$$\int_{A^c} P_x^k(A) d\pi(x) = 0 \quad \text{for all } k \in \mathbb{N}.$$

PROBLEM 17. Show that if P is irreducible and positive with invariant distribution π , and if h is bounded and harmonic, then h is π -almost surely constant.

PROBLEM 18. Show that P is irreducible and positive, and if all bounded, harmonic functions are constant, then P is Harris recurrent.

Hint: Take $A \in \mathbb{E}^+$ and define the function

$$h(x) = P(X_n \in A \text{ i.o.})$$

where X_0, X_1, \dots is a P -Markov chain with $X_0 = x$. Explain that h is harmonic, and show that $h(x) = 1$ for at least some x 'es.

6 Harris recurrence

PROBLEM 19. Suppose that π and Q satisfies (6), and let P be the corresponding Metropolis-Hastings kernel. Show that P is Harris recurrent.

Hint: Let h be bounded and harmonic. By problem 17 we know that

$$h(x) = \int h d\pi \quad \text{for } \pi\text{-almost all } x.$$

Show that harmonicity implies that

$$h(x) = \beta(x) \int h d\pi + (1 - \beta(x)) h(x) \quad \text{for all } x \in \mathcal{X},$$

where $\beta(x)$ is the overall acceptance probability.