

An Introduction
to
Markov Chain Monte Carlo (MCMC)

30 April 2007
Center for Data Analysis & Statistics

LTC Mick Smith
Department of Mathematical Sciences
US Military Academy

Outline

- Examples of Monte Carlo Methods
- Extension to Markov Chain Monte Carlo
- Convergence Theorems
- Example Algorithms
- Two Applications

Reference:

Robert & Casella, *Monte Carlo Statistical Methods*, Springer 1999.

Monte Carlo Integration

Consider the task of evaluating $\int_{\mathbb{R}} h(x) f(x) dx$.

Idea: If f is the probability density function associated with some random variable X , then interpret the integral as being equal to $\mathbb{E}_f[h(X)]$.

How does this help? If we can sample from f , then the **Strong Law of Large Numbers** states that

$$\frac{1}{n} \sum_{j=1}^n h(X_j) \rightarrow \mathbb{E}_f[h(X)] \quad \text{w.p. 1 as } n \rightarrow \infty,$$

where $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} f$.

Implicit here is that Statistics requires integration

Motivation for Importance Sampling

Question: What can block our progress along the route below?

$$\int_{\mathbb{R}} h(x) f(x) dx \approx \frac{1}{n} \sum_{j=1}^n h(X_j).$$

Motivation for Importance Sampling

Question: What can block our progress along the route below?

$$\int_{\mathbb{R}} h(x) f(x) dx \approx \frac{1}{n} \sum_{j=1}^n h(X_j).$$

Answer: Generating $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} f$ might be infeasible.

Importance Sampling

Let f be a density, presumed unworkable for sampling. Let g be a density for another distribution from which we *can* sample. Require that $\text{supp}\{g\} \supset \text{supp}\{f\}$. Then

$$\begin{aligned} \int_{\mathbb{R}} h(x) f(x) dx &= \int_{\mathbb{R}} h(x) \frac{f(x)}{g(x)} g(x) dx \\ &= \mathbb{E}_f[h(X)] \\ &\approx \frac{1}{n} \sum_{j=1}^n h(X_j) \frac{f(X_j)}{g(X_j)}, \end{aligned}$$

where $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} g$, density for the *instrumental distribution*.

An Alternative: Markov Chain Monte Carlo

Definition: A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f .

An Alternative: Markov Chain Monte Carlo

Definition: A *Markov chain Monte Carlo (MCMC) method* for the **simulation** of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f .

An Alternative: Markov Chain Monte Carlo

Definition: A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose **stationary distribution** is f .

An Alternative: Markov Chain Monte Carlo

Definition: A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an **ergodic Markov chain** $(X^{(t)})$ whose stationary distribution is f .

Introduction to Markov Chains

Fix a probability space (Ω, \mathcal{F}, P) on which we define random variables $X_j : \Omega \rightarrow \mathcal{X}$. Here $\mathcal{X} \subset \mathbb{R}$ is called the *state space* and is endowed with the usual Borel σ -algebra so that $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a measurable space.

Let $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$ be such that

- $K(x, \cdot)$ is a probability measure for all $x \in \mathcal{X}$; and
- $K(\cdot, A)$ is measurable for all $A \in \mathcal{B}(\mathcal{X})$.

Then K is a *transition kernel*.

In case \mathcal{X} is *discrete*, then K is a *transition probability matrix*:

$K = [P_{xy}]$ where

$$P_{xy} = P(X_{n+1} = y | X_n = x), \quad x, y \in \mathcal{X}.$$

Markov Chains Defined

Definition: Given a transition kernel K , a sequence $X_0, X_1, \dots, X_n, \dots$ is a *Markov chain*, denoted by (X_n) , if, for any t , the conditional distribution of X_t given $x_{t-1}, x_{t-2}, \dots, x_0$ is the same as the distribution of X_t given x_{t-1} ; that is,

$$\begin{aligned} P(X_{k+1} \in A \mid x_0, x_1, x_2, \dots, x_k) &= P(X_{k+1} \in A \mid x_k) \\ &= \int_A K(x_k, dx). \end{aligned}$$

Let $X_0 \sim \mu_0$. Probabilistic behavior of (X_n) is completely determined by μ_0 and K . In case \mathcal{X} is discrete, $X_n \sim \mu_0 K^n$ obtained by repeated matrix multiplication.

Markov Chains: Invariant Measures

Definition: A σ -finite measure π is *invariant* for the transition kernel $K(\cdot, \cdot)$ (and for the associated chain) if

$$\pi(B) = \int_{\mathcal{X}} K(x, B) \pi(dx), \quad \forall B \in \mathcal{B}(\mathcal{X}).$$

In case $\pi(\mathcal{X}) = 1$, we say that π is a *stationary distribution* since $X_0 \sim \pi$ implies that $X_n \sim \pi$ for every n .

Convergence of the Markov Chain

Let (X_n) be a Markov chain having stationary distribution π and for which $X_n \sim P^n$.

Question 1: What do we mean by $P^n \rightarrow \pi$?

Question 2: What conditions on (X_n) suffice to ensure that $P^n \rightarrow \pi$?

Question 1: Total Variation Norm

Let μ_1 and μ_2 be measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Definition: The *total variation norm* is given by

$$\|\mu_1 - \mu_2\|_{TV} = \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu_1(A) - \mu_2(A)|.$$

We take “ $P^n \rightarrow \pi$ ” to mean that $\|P^n - \pi\|_{TV} \rightarrow 0$ as $n \rightarrow \infty$.

Question 2: Ergodic Markov Chains

Here is an inelegant pseudo-theorem:

Claim: Let (X_n) be irreducible, recurrent, and aperiodic. Then

$$\lim_{n \rightarrow \infty} \left\| \int_{\mathcal{X}} K^n(x, \cdot) \mu(dx) - \pi \right\|_{TV} = 0$$

for every initial distribution μ .

Question 2: Ergodic Markov Chains

Claim: Let (X_n) be **irreducible**, recurrent, and aperiodic. Then

$$\lim_{n \rightarrow \infty} \left\| \int_{\mathcal{X}} K^n(x, \cdot) \mu(dx) - \pi \right\|_{TV} = 0$$

for every initial distribution μ .

Question 2: Ergodic Markov Chains

Claim: Let (X_n) be irreducible, **recurrent**, and aperiodic. Then

$$\lim_{n \rightarrow \infty} \left\| \int_{\mathcal{X}} K^n(x, \cdot) \mu(dx) - \pi \right\|_{TV} = 0$$

for every initial distribution μ .

Question 2: Ergodic Markov Chains

Claim: Let (X_n) be irreducible, recurrent, and **aperiodic**. Then

$$\lim_{n \rightarrow \infty} \left\| \int_{\mathcal{X}} K^n(x, \cdot) \mu(dx) - \pi \right\|_{TV} = 0$$

for every initial distribution μ .

Remind Me: What is MCMC?

Definition: A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f .

Let's proceed toward an example of such a method: the **Metropolis-Hastings algorithm**.

Metropolis-Hastings

Algorithm: Given $x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.

2. Take

$$X^{(t+1)} = \begin{cases} Y_t, & \text{w.p. } \rho(x^{(t)}, Y_t); \\ x^{(t)}, & \text{w.p. } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

$$\rho(x, y) = \min \left\{ \frac{f(y) q(x|y)}{f(x) q(y|x)}, 1 \right\}.$$

The distribution q is called the *instrumental* (or *proposal*) *distribution*.

2 More Definitions: Detailed Balance & Reversibility

Definition: A stationary Markov chain (X_n) is *reversible* if the distribution of X_{n+1} conditionally on $X_{n+2} = x$ is the same as the distribution of X_{n+1} conditionally on $X_n = x$.

Definition: A Markov chain with transition kernel K satisfies the *detailed balance condition* if there exists a function f satisfying

$$K(y, x)f(y) = K(x, y)f(x)$$

for every (x, y) .

Underpinning of Metropolis-Hastings

Theorem: Suppose that a Markov chain with transition kernel K satisfies the detailed balance condition with f a probability density function. Then

1. The density f is the invariant density of the chain.
2. The chain is reversible.

If follows that for every conditional distribution q whose support contains $\text{supp}(f)$, f is a stationary distribution of the chain $(X^{(t)})$ produced by the Metropolis-Hastings Algorithm.

Advance slide & then pause for proofs.

Convergence of the MH Chain

Theorem: Suppose that the Metropolis-Hastings Markov chain $(X^{(t)})$ is f -irreducible.

1. If $h \in L^1(f)$, then

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T h(X^{(t)}) = \int_{\mathcal{X}} h(x) f(x) dx \quad a.e. f.$$

2. If, in addition, $(X^{(t)})$ is aperiodic, then

$$\lim_{n \rightarrow \infty} \left\| \int_{\mathcal{X}} K^n(x, \cdot) \mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution μ , where $K^n(x, \cdot)$ denotes the kernel for n transitions.

Another Popular MCMC Method

Suppose that for some $p > 1$, the random vector $\mathbf{X} = (X_1, \dots, X_p) \in \mathcal{X}^n$ is such that we can sample from the corresponding univariate conditional densities f_1, \dots, f_p :

$$X_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p \sim f_i(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p),$$

for $i = 1, 2, \dots, p$.

The **Gibbs sampler** is the name of the following algorithm that specifies transition from $\mathbf{X}^{(t)}$ to $\mathbf{X}^{(t+1)}$.

The Gibbs Sampler

Algorithm: Given $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_p^{(t)})$, generate

$$1. X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \dots, x_p^{(t)})$$

$$2. X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_{p-1}^{(t+1)})$$

...

$$p. X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \dots, x_p^{(t)})$$

The densities f_1, \dots, f_p are called the *full conditionals*.

Question: Are these enough to specify the distribution of \mathbf{X} ?

Answer: Yes; up to a normalizing constant, the *Hammersley-Clifford Theorem* states that *positivity* is a sufficient condition.

Two Applications of MCMC (Metropolis-Hastings)

1. Sampling from a posterior on the *scene space* in my thesis
2. Sampling from a Pearson Type-III distribution

Target Distribution for Thesis M-H Chain

$$\nu(X | Y_1, Y_2, Y_3, Y_4) \propto L_1(Y_1 | X) L_2(Y_2 | X) L_3(Y_3 | X) L_4(Y_4 | X) \nu_0(X).$$

Basic space is $\mathcal{X} = \bigcup_{n=0}^{\infty} (\mathcal{D} \times \mathcal{A})^n$, where $\mathcal{D} \subset \mathbb{R}^2$ is a battlefield region of interest, $\mathcal{A} = \{\alpha_1, \dots, \alpha_M, \alpha_\emptyset\}$ is a set of M possible target types (α_\emptyset means that no target is present), and n is the number of targets present.

Proposal Distribution for Thesis M-H Chain

$$\begin{aligned} G(y | X^{(t)}) &= \\ &w_D \frac{1}{|\mathcal{N}_D(X^{(t)})|} \mathbf{1}_{\mathcal{N}_D(X^{(t)})(y)} + w_C \frac{1}{|\mathcal{N}_C(X^{(t)})|} \mathbf{1}_{\mathcal{N}_C(X^{(t)})(y)} \\ &+ w_A \frac{1}{|\mathcal{N}_A(X^{(t)})|} \mathbf{1}_{\mathcal{N}_A(X^{(t)})(y)} + w_B \mathbb{P}_{T_{X^{(t)}}}(\tau) \mathbf{1}_{\mathcal{N}_B(X^{(t)})(y)}, \end{aligned}$$

where $\mathbb{P}_{T_{X^{(t)}}}(\cdot)$ is a probability mass function on $(\mathcal{D} \times \mathcal{A}) \setminus T_{X^{(t)}}$.