Introduction to Markov chain Monte Carlo

Adam M. Johansen[1]

February 12, 2018

[1]Based closely on slides produced by Anthony Lee in previous years.

Outline

Outline

Introduction

- This is a module on stochastic simulation.
- Monte Carlo methods are certainly stochastic simulation techniques.
- They are also very important in many modern statistical analyses.
- I will cover "fundamental" theory and methodology for Markov chain Monte Carlo.
  - fundamental here means I cannot even cover 1% of what is interesting.
- There are other methods of stochastic simulation, and also deterministic counterparts to Monte Carlo.
- I hope that after the lectures you will understand why we can use MCMC, and how to construct your own Monte Carlo Markov chains.

Approximating expectations

- Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. We have a target probability measure $\pi : \mathcal{X} \to [0, 1]$ and we would like to approximate the quantity

$$\pi(f) := \int_{\mathsf{X}} f(x)\pi(\mathrm{d}x),$$

  where $f \in L_1(\mathsf{X}, \pi) = \{f : \pi(|f|) < \infty\}$, i.e., expectations w.r.t. π.
- We will assume that one can calculate π's associated density $\pi : \mathsf{X} \to \mathbb{R}_+$ w.r.t. some dominating measure (e.g., Lebesgue or counting).
- A major motivation for this in statistics is to compute posterior expectations in Bayesian inference.

Posterior expectations

- We have
  - a prior probability distribution for an unknown X-valued parameter with probability density function $p : \mathsf{X} \to \mathbb{R}_+$, and
  - a collection of probability distributions with probability density functions $\{g_x; x \in \mathsf{X}\}$ for some observed data $y \in \mathsf{Y}$.
- We can use Bayes' rule to obtain that the conditional distribution of the unknown X-valued parameter is defined by the probability density function

$$\pi(x) = \frac{p(x)g_x(y)}{\int_{\mathsf{X}} p(z)g_z(y)\mathrm{d}z}.$$

- Posterior expectations $\pi(f)$ cannot generally be calculated analytically, and so numerical methods are needed to approximate them.

The Strong Law of Large Numbers

Theorem (Strong Law of Large Numbers)

Assume $(X_n)_{n\geq 1}$ is a sequence of i.i.d. random variables distributed according to μ. Define

$$S_n(f) := \sum_{i=1}^{n} f(X_i),$$

for $f \in L_1(X, \mu)$. Then

$$\lim_{n\to\infty} \frac{1}{n} S_n(f) = \mu(f)$$

almost surely.

Monte Carlo Integration

- We can apply the SLLN with $\mu = \pi$ to use $n^{-1}S_n(f)$ as an estimate of $\pi(f)$, if we can sample according to π.
- There are some ways of doing this in special cases, e.g.,
  - inverse transform,
  - composition,
  - special representations in terms of random variables we can simulate easily.
  - other methods in, e.g., Devroye [1986]
- Most of the time in practical applications, we cannot easily sample according to π.

Radon–Nikodym derivative

- If μ and ν are densities w.r.t. the Lebesgue measure and $\nu(x) > 0 \Rightarrow \mu(x) > 0$ then

$$\int_A \frac{\nu(x)}{\mu(x)} \mu(x) \mathrm{d}x = \int_A \nu(x) \mathrm{d}x = \nu(A),$$

  for an arbitrary measurable A.
- If μ and ν are σ-finite measures on $(\mathsf{X}, \mathcal{X})$ and μ dominates ν ($\nu \ll \mu$)then there is a function f such that

$$\int_A f(x) \mu(\mathrm{d}x) = \nu(A), \qquad A \in \mathcal{X},$$

  and we call it the Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Rejection sampling

Rejection sampling

1. Sample $X \sim \mu$.
2. With prob. $\frac{1}{M}\frac{\pi(X)}{\mu(X)}$ output X, otherwise go back to step 1.

- A general purpose method for sampling from π when we can sample from μ and

$$\sup_x \frac{\pi(x)}{\mu(x)} \leq M < \infty.$$

- Letting $Y = \mathbb{I}\left(U < \frac{1}{M}\frac{\pi(X)}{\mu(X)}\right)$ where U is uniformly distributed on $[0, 1]$ we obtain

$$\begin{aligned} \Pr(X \in A \mid Y = 1) &= \frac{\Pr(X \in A, Y = 1)}{\Pr(Y = 1)} \\ &= \frac{\int_A \frac{1}{M}\frac{\pi(x)}{\mu(x)}\mu(x)\mathrm{d}x}{\int_{\mathsf{X}} \frac{1}{M}\frac{\pi(x)}{\mu(x)}\mu(x)\mathrm{d}x} = \pi(A). \end{aligned}$$

Cost of rejection sampling

- We have

$$\Pr(Y = 1) = \int_{\mathsf{X}} \frac{1}{M} \frac{\pi(x)}{\mu(x)} \mu(x) \mathrm{d}x = \frac{1}{M}.$$

- It follows that the time until acceptance is a geometric random variable with success probability M^{-1}.
- The expected time to obtain a single sample is M.
- In many practical applications M is prohibitively large.
  - Toy example: consider what happens as d increases when $\pi(x) = \prod_{i=1}^{d} p(x_i)$, $\mu(x) = \prod_{i=1}^{d} g(x_i)$ and $\sup_x \frac{p(x)}{g(x)} > 1$.
- Practical intuition: for complicated π we do not usually know how to find a "good" μ.

Importance sampling

- Recall that $(X_n)_{n\geq 1}$ is a sequence of i.i.d. μ-distributed random variables.
- We again appeal to the SLLN, but now assume only that $\pi \ll \mu$ and we define

$$\tilde{f}(x) := f(x)w(x), \quad x \in \mathsf{X},$$

where $f \in L_1(\mathsf{X}, \pi)$ is the function defining the expectation of interest and

$$w(x) := \frac{\pi(x)}{\mu(x)}, \quad x \in \mathsf{X},$$

is the "importance weight" function.

- It follows that

$$\mu(\tilde{f}) = \int_{\mathsf{X}} f(x)\frac{\pi(x)}{\mu(x)}\mu(x)\mathrm{d}x = \int_{\mathsf{X}} f(x)\pi(x)\mathrm{d}x = \pi(f).$$

Cost of importance sampling

- Consider
$$\tilde{f}(x) := f(x)w(x).$$
- Then if $\tilde{f} \in L_2(\mathrm{X}, \mu)$ we have
$$\mathrm{var}(\tilde{f}(X)) = \int_{\mathrm{X}} \tilde{f}(x)^2 \mu(\mathrm{d}x) - \mu(\tilde{f})^2 = \mu(\tilde{f}^2) - \mu(\tilde{f})^2.$$
- One can then obtain
$$\mathrm{var}(n^{-1} S_n(\tilde{f})) = \frac{\mu(\tilde{f}^2) - \mu(\tilde{f})^2}{n}.$$
- Note: it is possible that $f \in L_2(\mathrm{X}, \pi)$ but $\tilde{f} \notin L_2(\mathrm{X}, \mu)$.
  - in practice, one can avoid this by having $\sup_x \pi(x)/\mu(x) < \infty$.
- In many practical situations, the numerator of this expression is prohibitively large.

Self-normalized importance sampling

- In many situations, one can only compute π up to an unknown normalizing constant.
- We define the self-normalized estimate via

$$I_n(f, \pi, \mu) := \frac{S_n(\tilde{f})}{S_n(w)} = \frac{\sum_{i=1}^n f(X_i) w(X_i)}{\sum_{i=1}^n w(X_i)},$$

  and it is clear that one only needs to know π up to an unknown normalizing constant.
- Then

$$\lim_{n \to \infty} I_n(f, \pi, \mu) = \pi(f)$$

  almost surely.
- If $\int_{\mathsf{X}} \left[1 + f(x)^2\right] \frac{\pi(x)}{\mu(x)} \pi(x) \mathrm{d}x < \infty$ then asymptotically the variance of $I_n(f)$ is

$$\frac{1}{n} \int_{\mathsf{X}} \left[f(x) - \pi(f)\right]^2 \frac{\pi(x)}{\mu(x)} \pi(x) \mathrm{d}x.$$

- Note: this expression can be smaller than $\mathrm{var}(n^{-1} S_n(\tilde{f}))$.

Outline

Markov chains and stochastic stability

- If unspecified, the source of a definition or theorem is

  Meyn, S. and Tweedie, R. L. (2009) Markov chains and stochastic stability, 2nd ed.

- This is a great single and reasonably accessible source for a lot of what you might want to know about Markov chains on a general state space.
- There is a free version online at http://probability.ca/MT/

Time homogeneous, discrete time Markov chains

- We will assume now that $\mathcal{X}$ is countably generated, e.g. the Borel σ-algebra on $\mathbb{R}^d$.
- Let $\mathbf{X} := (X_n)_{n\geq 0}$ be a discrete time Markov chain evolving on X with some initial distribution for X_0.
- This means that for $A \in \mathcal{X}$

$$\Pr\left(X_n \in A \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}\right) = \Pr\left(X_n \in A \mid X_{n-1} = x_{n-1}\right),$$

  i.e. $\mathbf{X}$ possesses the Markov property.
- We will restrict our attention to the time-homogeneous case:

$$\Pr\left(X_n \in A \mid X_{n-1} = x\right) = \Pr\left(X_1 \in A \mid X_0 = x\right),$$

  for any $n \in \mathbb{N}$.
- Then $\mathbf{X}$ is described by a single Markov transition kernel $P : \mathrm{X} \times \mathcal{X} \to [0, 1]$ with

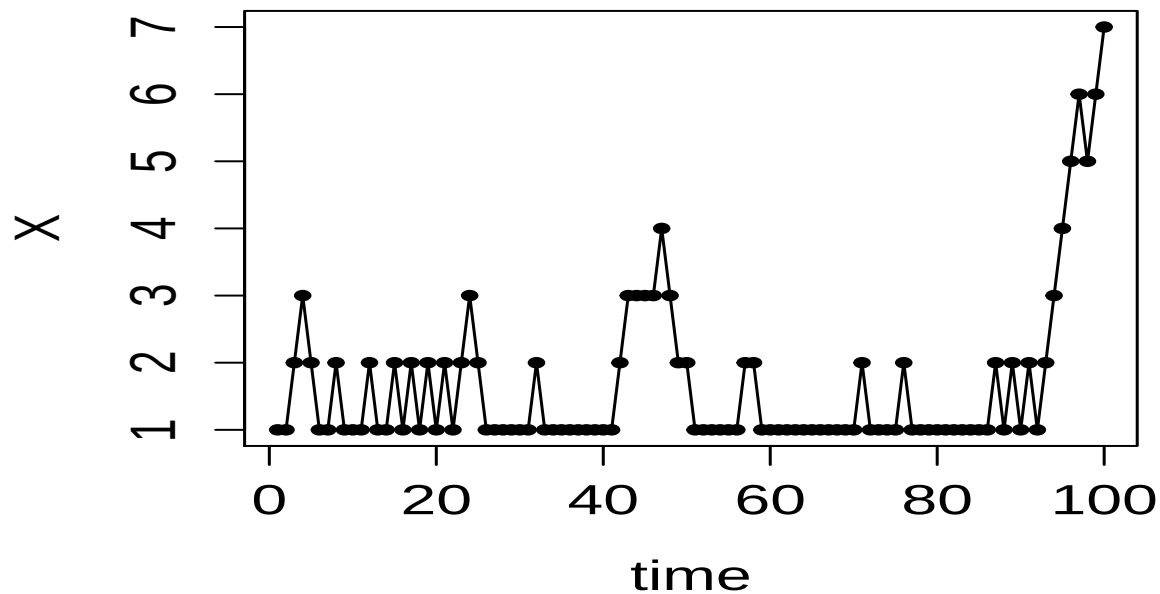
$$\Pr(X_1 \in A \mid X_0 = x) = P(x, A).$$

Example: simple random walk on $\mathbb{N}$

- Let $p, q \in (0, 1]$ and $r \in [0, 1)$ such that $p + q + r = 1$, and

$$P(i,j) := \begin{cases} p & j = i - 1, \\ q & j = i + 1, \\ r & j = i, \\ 0 & \text{otherwise}, \end{cases} \qquad i \geq 2$$

with $P(1,1) = p + r$ and $P(1,2) = q$.



Example: simple random walk on $\mathbb{N}$

- How can we characterize the behaviour of $\mathbf{X}$?
- Does it "escape to infinity"?
- Will it visit every point at least once?
- Will it visit each point infinitely many times?
- Does it have a "stationary distribution"?
- Does it look the same forwards and backwards in time?
- How do the partial sums

$$S_n(f) := \sum_{i=1}^{n} f(X_i)$$

behave?

Example: simple random walk on $\mathbb{N}$

When $q < p$:

- It is recurrent - it almost surely visits every state infinitely often.
- It is therefore not transient.
- It has an invariant probability measure $\mu(x) = \mathrm{Geo}(x; q/p)$.
- It is (time)-reversible - if $X_0 \sim \mu$ then

$$\mathcal{L}(X_0, X_1, \ldots, X_n) = \mathcal{L}(X_n, X_{n-1}, \ldots, X_0).$$

- It is irreducible.
- The proportion of time it spends at each point x converges almost surely to $\mu(x)$.
- It is aperiodic and for each $i \in \mathbb{N}$ (irrespective of x_0),

$$\lim_{n \to \infty} \Pr(X_n = i \mid X_0 = x_0) = \mu(i).$$

- The list could go on...

Example: simple random walk on $\mathbb{N}$

When $q > p$:

- It is transient - the expected number of visits to each state is finite.
- It does not have an invariant probability measure.
- It is not time-reversible (depending on the definition!).
- It is aperiodic and irreducible.

When $q = p$:

- It is recurrent.
- It does not have an invariant probability measure.
- It is not time-reversible (again, depending on the definition).
- It is aperiodic and irreducible.

Our interest is in Markov chains that behave as in the case $q < p$.

Stability properties of Markov chains

- Many of the properties discussed above can be verified in this specific case in a number of different ways.
- We are interested, however, in more general classifications.
- Consider a simple random walk on $\mathbb{R}_+$ with $X_0 = 0$ and

$$X_n = \max\{X_{n-1} + W_n, 0\},$$

where $(W_n)_{n\geq 1}$ is a sequence of i.i.d. random variables with mean β.

- Is $\mathbf{X}$ recurrent or transient? Does it have an invariant (probability) measure?
- Clearly this chain has some differences to the simple random walk on $\mathbb{N}$.
  - e.g., it does not visit an arbitrary $x \in \mathbb{R}_+ \setminus \{0\}$ with positive probability.
- Since most statistical applications involve $\mathsf{X} \subseteq \mathbb{R}^d$ we need to discuss properties of Markov chains on general state spaces.

Why do we care?

Theorem (An Ergodic Theorem (an LLN for Markov chains))

Suppose that $\mathbf{X} = (X_n)_{n \geq 0}$ *is a positive Harris Markov chain with invariant probability measure* π. *Then for any* $f \in L_1(\mathsf{X}, \pi) = \{f : \pi(|f|) < \infty\}$,

$$\lim_{n \to \infty} \frac{1}{n} S_n(f) = \pi(f),$$

almost surely for any initial distribution for X_0.

- We need to understand some of these definitions.

Outline

General state spaces

- When considering a Markov chain $\mathbf{X}$ on a general state space, we must start to think about sets $A \in \mathcal{X}$ rather than points $x \in \mathsf{X}$.
- When statements about the chain $\mathbf{X}$ are made in probability P or expectation E, we can use a subscript x or μ to denote the "initial" or marginal distribution of X_0.
- We define P^n to be the n-step transition kernel by $P^1(x, A) := P(x, A)$ and

$$P^n(x, A) := \int_{\mathsf{X}} P(z, A) P^{n-1}(x, \mathrm{d}z), \quad n \geq 2.$$

- We will use P to denote the linear operator associated with this Markov transition kernel, which acts to the left on measures:

$$\mu P(A) := \int_{\mathsf{X}} \mu(\mathrm{d}x) P(x, A).$$

φ-irreducibility

Definition

X is φ-irreducible if φ is a measure on $\mathcal{X}$ such that whenever $\varphi(A) > 0$ and $x \in \mathsf{X}$, there exists some n possibly depending on both x and A such that $P^n(x, A) > 0$.

- It is important to note that this holds for every $x \in \mathsf{X}$ and is therefore rather strong.
- One can think of **X** having a maximal irreducibility probability measure ψ whenever it is φ-irreducible, such that (MT Proposition 4.2.2)
  1. **X** is ψ-irreducible;
  2. **X** is φ'-irreducible if and only if $\varphi' \ll \psi$.

Recurrent and transient sets

- We define the occupation time of a set $A \subseteq \mathsf{X}$ to be

$$\eta_A := \sum_{n=1}^{\infty} \mathbb{I}\{X_n \in A\}.$$

Definition

The set A is recurrent if $\mathsf{E}_x[\eta_A] = \infty$ for all $x \in A$.

Definition

The set A is uniformly transient if there exists $M < \infty$ such that $\mathsf{E}_x[\eta_A] \leq M$ for all $x \in A$.

Recurrence/transience dichotomy

- When $\mathbf{X}$ is ψ-irreducible, a dichotomy theorem describes whether the chain $\mathbf{X}$ is recurrent or transient.

Theorem (MT Theorem 8.0.1)

Suppose that $\mathbf{X}$ is ψ-irreducible. Then either

1. *every set $A \in \mathcal{X}$ with $\psi(A) > 0$ is recurrent, and we call $\mathbf{X}$ recurrent, or*
2. *there is a countable cover of X with uniformly transient sets, and we call $\mathbf{X}$ transient.*

- You can find alternative definitions of recurrence and transience in MT Appendix A, e.g., $\mathbf{X}$ is recurrent iff

$$\sum_{n \geq 0} P^n(x, A) = \infty, \quad x \in \mathsf{X}, \quad \psi(A) > 0,$$

or statements about $\mathsf{P}_x(\mathbf{X} \text{ visits } A \text{ i.o.}) = 1$.

Harris recurrence

- In order to make statements about $\mathbf{X}$ regardless of the value of X_0 one requires a stronger definition of recurrence.

Definition

A set A is Harris recurrent if $\mathsf{P}_x(\eta_A = \infty) = 1$ for all $x \in A$.

Definition

$\mathbf{X}$ is Harris (recurrent) if it is ψ-irreducible and every set $A \in \mathcal{X}$ such that $\psi(A) > 0$ is Harris recurrent.

- The difference between recurrence and Harris recurrence is the difference between

$$\mathsf{E}_x[\eta_A] = \infty \qquad \text{and} \qquad \mathsf{P}_x(\eta_A = \infty) = 1.$$

Example of non-Harris recurrence

- The difference between recurrence and Harris recurrence is the difference between

$$\mathsf{E}_x[\eta_A] = \infty \qquad \text{and} \qquad \mathsf{P}_x(\eta_A = \infty) = 1.$$

- When $\mathsf{X} = \mathbb{N}$, consider Charlie Geyer's example:

$$P(1,1) = 1,\ P(x, x+1) = 1 - x^{-2},\ P(x,1) = x^{-2}.$$

Then $\psi(\{x\}) > 0 \iff x = 1$, and for all x, $\mathsf{E}_x[\eta_{\{1\}}] = \infty$ since $\mathsf{P}_x(X_1 = 1) > 0$. However,

$$\mathsf{P}_x(X_n = x + n \text{ for all } n) = \prod_{j=x}^{\infty} \left(1 - \frac{1}{j^2}\right) = \frac{x-1}{x} > 0,$$

so $\mathsf{P}_x(\eta_{\{1\}} = \infty) < 1$.

Invariant measures

Definition

A sigma-finite measure μ is an invariant measure for $\mathbf{X}$ if

$$\mu P = \mu.$$

- Our interest in invariant measures is related to viewing a special version of $\mathbf{X}$ as a stationary process.
- Indeed, assume that μ is a probability measure and that $\Pr(X_0 \in A) = \mu(A)$ for all $A \in \mathcal{X}$.
- Then it is not too difficult to see that $\mathbf{X}$ is a stationary process, i.e. the marginal distribution of $(X_n, \ldots, X_{n+k})$ does not change as n varies.
- In general, invariant measures are not necessarily finite.
- When $\mathbf{X}$ is recurrent, the unique (up to constant multiples) invariant measure for $\mathbf{X}$ is equivalent (as a measure) to ψ (MT Theorem 10.4.9)

Positive and null chains

Definition

If $\mathbf{X}$ is ψ-irreducible and admits an invariant probability measure then it is positive. If $\mathbf{X}$ does not admit such a measure then it is null.

Example

Consider $\mathbf{X}$ being a simple random walk on $\mathbb{N}$ as before. If $p > q$, $\mathbf{X}$ is positive (recurrent). If $p = q$ then $\mathbf{X}$ is null (but) recurrent. If $q > p$ then $\mathbf{X}$ is (null and) transient.

The LLN again

Theorem (An Ergodic Theorem for Harris Chains)

Suppose that $\mathbf{X} = (X_n)_{n\geq 0}$ *is a positive Harris Markov chain with invariant probability measure* π*. Then for any* $f \in L_1(\mathsf{X}, \pi) = \{f : \pi(|f|) < \infty\}$,

$$\lim_{n\to\infty} \frac{1}{n} S_n(f) = \pi(f),$$

almost surely for any initial distribution for X_0.

- One can replace Harris recurrence with φ-irreducibility and positivity but then the statement holds only for π-almost all X_0. This is eventually a consequence of Birkhoff's Pointwise Ergodic Theorem.
- Being positive Harris implies that if an LLN holds for f and some initial distribution then it must hold for every initial distribution (MT Proposition 17.1.6).

Null recurrent vs transient: simplified classic example

- Let $(X_n^{(i)})_{n\geq 1}$ be independent, simple random walks on $\mathbb{Z}$: $p = q = \frac{1}{2}$, for each $i \in \{1, \ldots, d\}$.
- We have $\mathsf{P}_0(X_n^{(i)} = 0) = 0$ for odd n, and

$$\mathsf{P}_0(X_{2n}^{(i)} = 0) = \Pr(B_{2n} = n) \sim \frac{1}{\sqrt{\pi n}}$$

where B_{2n} is a $\mathrm{Binomial}(2n, \frac{1}{2})$ r.v.

- Consider the Markov chain $(X_n^{(1)}, \ldots, X_n^{(d)})$ started at $\mathbf{0}$. Then

$$\begin{aligned}
\mathsf{E}_{\mathbf{0}}\left[\eta_{\{\mathbf{0}\}}\right] &= \mathsf{E}_{\mathbf{0}}\left[\sum_{n=1}^{\infty} \mathbb{I}\left(X_n^{(1)} = \cdots = X_n^{(d)} = 0\right)\right] \\
&= \sum_{n=1}^{\infty} \mathsf{P}_{\mathbf{0}}\left(X_{2n}^{(1)} = \cdots = X_{2n}^{(d)} = 0\right) \\
&\sim \sum_{n=1}^{\infty} (\pi n)^{-d/2},
\end{aligned}$$

which is infinite only for $d \in \{1, 2\}$.

Outline

Motivation

- The LLN motivates the following question:

  Can we construct a Harris recurrent or at least φ-irreducible Markov chain with invariant distribution π where all we compute is the density $\pi(x)$ (up to an unknown normalizing constant) for any $x \in \mathsf{X}$?

- If so, then we can produce a realization $\mathbf{X}$ and estimate $\pi(f)$ via $n^{-1}S_n(f)$ where

$$S_n(f) := \sum_{i=1}^{n} f(X_i).$$

- A positive, constructive answer to this question was a pivotal moment in Bayesian statistics, and many other sciences.

Metropolis–Hastings

- There are a large number of ways of constructing such Markov chains, but we will focus on the essentials.
- By far the most commonly used Markov chains in practice are constructed using Metropolis–Hastings Markov transition kernels.
- These owe their development to the seminal papers Metropolis et al. [1953] and Hastings [1970].
- Assume π has a density w.r.t. μ.
- In order to define the Metropolis–Hastings kernel for a particular target π we require only to specify a proposal Markov kernel Q admitting a density q w.r.t. μ, i.e.

$$Q(x, \mathrm{d}z) = q(x, z)\mu(\mathrm{d}z).$$

Metropolis–Hastings

To simulate according to $P_{\mathrm{MH}}(x, \cdot)$:

1. Simulate $Z \sim Q(x, \cdot)$.
2. With prob. $\alpha_{\mathrm{MH}}(x, Z)$ output Z; otherwise, output x, where

$$\alpha_{\mathrm{MH}}(x, z) := 1 \wedge \frac{\pi(z) q(z, x)}{\pi(x) q(x, z)}$$

- Equivalently,

$$P_{\mathrm{MH}}(x, A) := \int_A \alpha_{\mathrm{MH}}(x, z) Q(x, \mathrm{d}z) + r_{\mathrm{MH}}(x) \mathbf{1}_A(x),$$

where

$$r_{\mathrm{MH}}(x) := 1 - \int_{\mathsf{X}} \alpha_{\mathrm{MH}}(x, z) Q(x, \mathrm{d}z).$$

- We need only know the density π up to a normalizing constant.

Metropolis–Hastings validity

- In order to show that P leaves π invariant, we need to check

$$\pi P = \pi$$

  i.e., that

$$\int_{\mathsf{X}} \pi(\mathrm{d}x) P(x, A) = \pi(A), \qquad \forall A \in \mathcal{X}.$$

- Verifying $\pi P = \pi$ is extremely difficult in general.
- Determining the invariant measure of a given Markov kernel is also v. difficult.
- The π-invariance of the Metropolis–Hastings Markov chain is a special case of the π-invariance of π-reversible Markov chains.

Reversible Markov chains

Definition

A π-reversible Markov chain is a stationary Markov chain with invariant probability measure π satisfying

$$\mathrm{P}_\pi(X_0 \in A_0, \ldots, X_n \in A_n) = \mathrm{P}_\pi(X_0 \in A_n, \ldots, X_n \in A_0).$$

- It suffices to check that

$$\mathrm{P}_\pi(X_0 \in A, X_1 \in B) = \mathrm{P}_\pi(X_0 \in B, X_1 \in A),$$

  i.e.

$$\int_A \pi(\mathrm{d}x) P(x, B) = \int_B \pi(\mathrm{d}x) P(x, A).$$

- Moreover, π-invariance is obvious by considering $A = \mathsf{X}$:

$$\int_{\mathsf{X}} \pi(\mathrm{d}x) P(x, B) = \int_B \pi(\mathrm{d}x) P(x, \mathsf{X}) = \pi(B).$$

Reversible Markov chains

- That $\int_A \pi(\mathrm{d}x)P(x,B) = \int_B \pi(\mathrm{d}x)P(x,A)$ implies reversibility is slightly laborious in the general state space context.
- For intuition, consider a discrete state space where the property becomes

$$\mathsf{P}_\pi(X_0 = x_0, \ldots, X_n = x_n) = \mathsf{P}_\pi(X_0 = x_n, \ldots, X_n = x_0),$$

which is indeed implied by $\pi(x)P(x,z) = \pi(z)P(z,x)$ since

$$\begin{aligned}
& \mathsf{P}_\pi(X_0 = x_0, \ldots, X_n = x_n) \\
= \;& \pi(x_0)P(x_0,x_1)\cdots P(x_{n-1},x_n) \\
= \;& P(x_1,x_0)\pi(x_1)P(x_1,x_2)\cdots P(x_{n-1},x_n) \\
= \;& P(x_1,x_0)P(x_2,x_1)\cdots P(x_n,x_{n-1})\pi(x_n) \\
= \;& \mathsf{P}_\pi(X_0 = x_n, \ldots, X_n = x_0).
\end{aligned}$$

Verifying π-reversibility

- When $P(x, A) = \int_A p(x, z)\mu(\mathrm{d}z) + r(x)\mathbf{1}_A(x)$, we can verify reversibility by considering the densities $\pi(x)$ and $p(x, z)$ each w.r.t μ. Indeed if the detailed balance condition

$$\pi(x)p(x, z) = \pi(z)p(z, x), \quad x, z \in \mathsf{X}$$

holds then

$$\begin{aligned}
& \int_A \pi(\mathrm{d}x)P(x, B) \\
= & \int_A \pi(x) \left[\int_B p(x, z)\mu(\mathrm{d}z) + r(x)\mathbf{1}_B(x) \right] \mu(\mathrm{d}x) \\
= & \int_B \pi(z) \left[\int_A p(z, x)\mu(\mathrm{d}x) \right] \mu(\mathrm{d}z) + \int_{A\cap B} \pi(x)r(x)\mu(\mathrm{d}x) \\
= & \int_B \pi(z) \left[\int_A p(z, x)\mu(\mathrm{d}x) + r(z)\mathbf{1}_A(x) \right] \mu(\mathrm{d}z) \\
= & \int_B \pi(\mathrm{d}x)P(x, A).
\end{aligned}$$

Verifying π-reversibility for Metropolis–Hastings

- The key utility of detailed balance is it need only be checked pointwise — no integrals necessary!
- We now verify for P_{MH}:

$$\begin{aligned}
\pi(x)p_{\mathrm{MH}}(x,z) &= \pi(x)q(x,z)\left[1 \wedge \frac{\pi(z)q(z,x)}{\pi(x)q(x,z)}\right] \\
&= [\pi(x)q(x,z) \wedge \pi(z)q(z,x)] \\
&= \pi(z)q(z,x)\left[\frac{\pi(x)q(x,z)}{\pi(z)q(z,x)} \wedge 1\right] \\
&= \pi(z)p_{\mathrm{MH}}(z,x).
\end{aligned}$$

- This is extremely versatile and most Markov chains used in statistics are constructed using reversible Markov transition kernels.

What about Harris recurrence?

- That P_{MH} is π-reversible implies that if it is also π-irreducible then it is positive and has the right invariant probability measure.
- Verifying φ-irreducibility is *typically* very easy.
  - e.g., $\pi(A) > 0$, $A \in \mathcal{X}$ and $q(x, A) > 0$, $x \in \mathsf{X}, A \in \mathcal{X}$.

Theorem (Tierney [1994, Corollary 2], Roberts and Rosenthal [2006, Theorem 8])

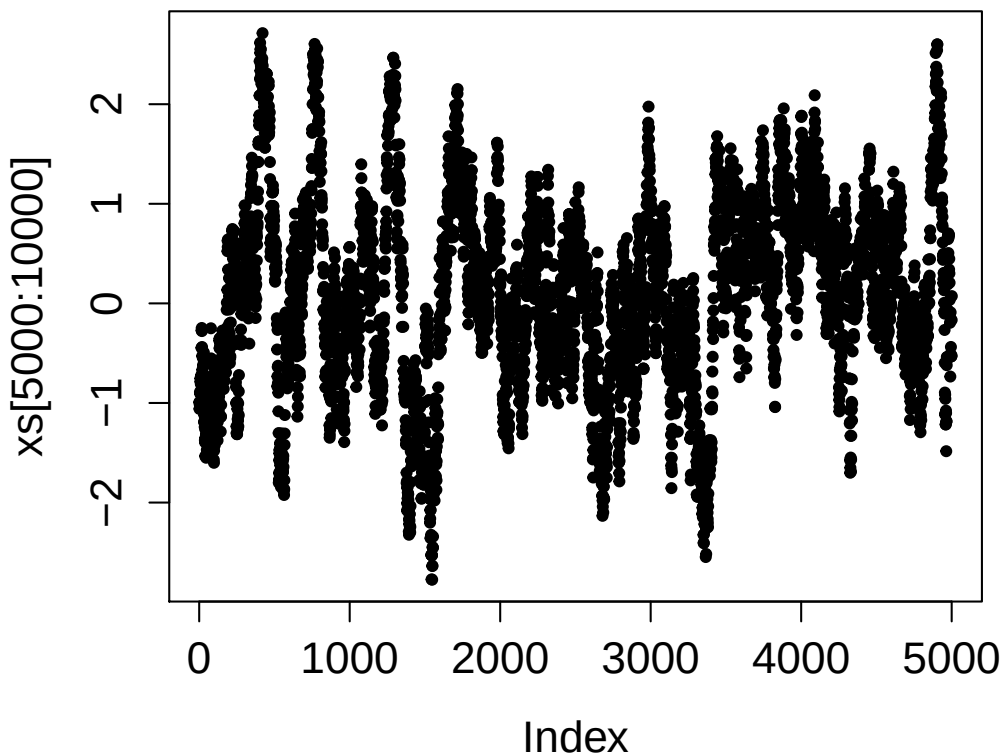
Every π-irreducible, full-dimensional Metropolis–Hastings Markov chain is Harris recurrent.

- That's all you need to know to construct some sophisticated Markov chains!

Random walk Metropolis–Hastings

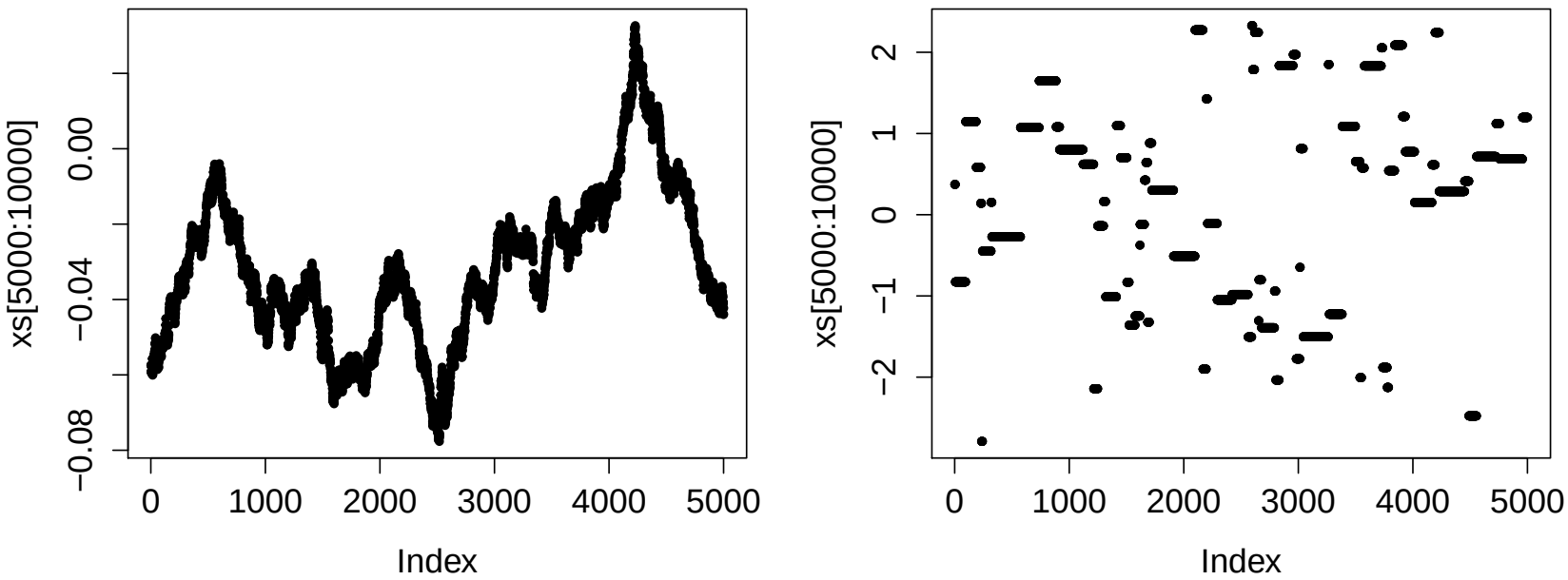
- Let π be given and let Q satisfy $q(x, z) = q(\|z - x\|)$. Then the Metropolis–Hastings acceptance probability is

$$\alpha_{\mathrm{MH}}(x, z) = 1 \wedge \frac{\pi(z)}{\pi(x)}.$$



Random walk Metropolis–Hastings

- Choice of the proposal is important, even though the Markov chain is "valid".



- On the left, the variance of $Q(x, \cdot)$ is too small and on the right it is too large.

Independent Metropolis–Hastings

- One can even choose $Q(x, \cdot) = q(\cdot)$ to be independent of x.
- Then we have

$$\alpha_{\mathrm{MH}}(x, z) = 1 \wedge \frac{\pi(z)q(x)}{\pi(x)q(z)}.$$

- It can be difficult to find a good q in practice, but we will return to this example later.
- As before, it is helpful if

$$\sup_x \frac{\pi(x)}{q(x)} < \infty.$$

Hybrid Markov chains

- We can easily construct π-invariant Markov chains out of different π-invariant Markov transition kernels.
- In practice, such hybrid chains are commonplace.
  - the Gibbs sampler is an example.
- Generally speaking, we will have $(P_s)_{s \in S}$ and we will try to make a mixture or cycle or combination of the two out of them.

Mixtures of Markov kernels

Definition

A Markov kernel P is a mixture of the Markov kernels $(P_s)_{s\in S}$ if

$$P(x, A) = \sum_{s\in S} w(s)P_s(x, A),$$

where w is a p.m.f. (independent of x). Alternatively, $P = \sum_{s\in S} w(s)P_s$.

Fact

A mixture of π-invariant Markov kernels is π-invariant.

Proof.

$$\pi P(A) = \sum_{s\in S} w(s)\pi P_s(A) = \sum_{s\in S} w(s)\pi(A) = \pi(A).$$

□

Cycles of Markov kernels

Definition

A Markov kernel P is a cycle of Markov kernels P_1 and P_2 if

$$P(x, A) = \int_{\mathsf{X}} P_1(x, \mathrm{d}z) P_2(z, A),$$

i.e., $P = P_1 P_2$.

Fact

A cycle of π-invariant Markov kernels is π-invariant.

Proof.

$$\pi P(A) = \pi P_1 P_2(A) = \pi P_2(A) = \pi(A).$$

□

Remarks on hybrid chains

- If P is φ-irreducible then so is a mixture including P with positive probability.
- The same is not necessarily true for cycles, but it is often true in practice.
- A mixture of π-reversible Markov kernels is π-reversible.
- A cycle of π-reversible Markov kernels is generally not π-reversible.
- We will now see a special kind of hybrid Markov chain called the Gibbs sampler.

The Gibbs sampler

- Let $\mathsf{X} = \mathsf{X}_1 \times \cdots \times \mathsf{X}_d$.
- Let $-i$ denote the sequence $(1, \ldots i-1, i+1, \ldots, d)$ with the convention that $(1, 0) = (d+1, d) = ()$ is the empty sequence.
- If $s = (s_1, \ldots, s_j)$ then let $x_s := (x_{s_1}, x_{s_2}, \ldots, x_{s_j})$.
- Assume we can sample from each "full" conditional distribution defined by

$$\pi_{i,x_{-i}}(A) = \Pr(X_i \in A \mid X_{-i} = x_{-i}),$$

  which has a density $\pi_i(\cdot|x_{-i})$ w.r.t. some dominating μ.
- Now define

$$P_i(x, A_1 \times \cdots \times A_d) := \pi_{i,x_{-i}}(A_i) \prod_{j \neq i} \mathbb{I}(x_j \in A_j).$$

- It follows that P_i is in fact a Metropolis–Hastings kernel with acceptance probability 1 since

$$\alpha_{\mathrm{MH}}(x, z) = 1 \wedge \frac{\pi(z_1, \ldots, z_d)\pi_i(x_i|z_{-i})}{\pi(x_1, \ldots, x_d)\pi_i(z_i|x_{-i})} = \frac{\pi(z_{-i})}{\pi(x_{-i})} = 1.$$

The Gibbs sampler

- Gibbs samplers are commonly used to sample from Bayesian hierarchical models.
- Example:

$$\begin{aligned} Y_i \mid \theta_i &\sim F_{\theta_i}, \qquad i \in \{1, \ldots, n\} \\ \theta_i \mid \theta_0 &\sim G_{\theta_0}, \qquad i \in \{1, \ldots, n\} \\ \theta_0 &\sim H. \end{aligned}$$

- By fixing, e.g., $(\theta_1, \ldots, \theta_n)$ one may know the distribution of θ_0 conditional upon $\theta_1, \ldots, \theta_n$ and by fixing θ_0 one may know the distribution of $\theta_1, \ldots, \theta_n$ conditional upon θ_0 and $Y_1, \ldots, Y_n$.
- Originally introduced in statistical physics, then to statistics in Geman and Geman [1984] and popularized in Gelfand and Smith [1990]

Random scan and deterministic scan

- There are two major approaches to constructing a Gibbs sampler.
- Random scan:

$$P(x, A) = \sum_{s \in S} w(s) P_s(x, A),$$

  with $S = \{1, \ldots, d\}$ and usually $w(s) = d^{-1}$ for each $s \in S$.
- Deterministic scan:

$$P = P_{\sigma(1)} \cdots P_{\sigma(d)}$$

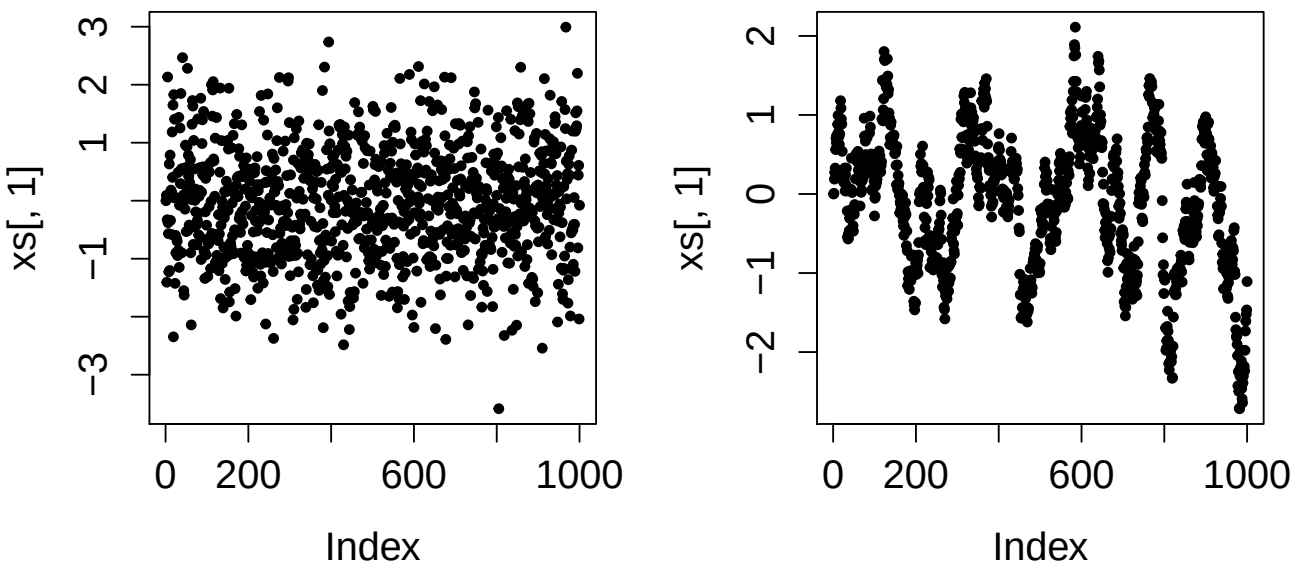
  where σ is some permutation of $(1, \ldots, d)$.
- These are just mixtures or cycles of the constituent kernels.

Gibbs sampler: toy example

- Consider the case $\mathsf{X} = \mathbb{R} \times \mathbb{R}$ and $\pi(x) = \mathcal{N}(x, 0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \rho \in (-1, 1).$$

- Then we have $\pi(x_1|x_2) = \mathcal{N}(x_1; \rho x_2, 1 - \rho^2)$ and $\pi(x_2|x_1) = \mathcal{N}(x_2; \rho x_1, 1 - \rho^2)$.
- Below we have a plot of the first coordinate of $\mathbf{X}$ when $\rho = .1$ (left) and $\rho = .99$ (right).



Gibbs sampler: blocking

- Imagine that $\mathsf{X} = \mathbb{R}^3$ and the correlation between the first two coordinates is large whilst the third is not very correlated.
- Then it makes sense to treat (x_1, x_2) and x_3 as two components in a Gibbs sampler.
- This is called "blocking", as one updates several variables together from their "joint" conditional.

Metropolis-within-Gibbs samplers

- In some cases, only some of the conditional distributions can be sampled from.
- So for any i such that we can't sample from $\pi_{i,x_{-i}}$ we can instead perform a Metropolis–Hastings update that updates only the ith coordinate of x.

Auxiliary variables

- Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space.
- Let $\tilde{\pi}$ be a probability measure on $\mathcal{X} \otimes \mathcal{Y}$ such that $\tilde{\pi}(A, \mathsf{Y}) = \pi(A)$.
- Then it is clear that if we can construct a positive Harris $\tilde{\pi}$-invariant Markov chain $(X_n, Y_n)_{n \geq 1}$, we can use

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

  to estimate $\pi(f)$ — we "discard" the auxiliary variables $(Y_n)_{n \geq 1}$.
- There are a huge number of auxiliary variable methods now.
- I will cover two interesting examples.

Latent variable model

- Consider a target density where

$$\pi(x) \propto p(x) \int_{\mathsf{Y}} g(y) f(x, y) \mathrm{d}y.$$

- For example, y represents a latent variable whose conditional density given x is $f(x, \cdot)$ and $g(y)$ is the conditional density of some observed data given y.
- Assume further that we cannot evaluate the function $x \mapsto \int_{\mathsf{Y}} g(y) f(x, y) \mathrm{d}y$ pointwise.
- We can instead define an extended target density

$$\tilde{\pi}(x, y) \propto p(x) g(y) f(x, y),$$

  and construct a Markov chain with invariant distribution $\tilde{\pi}$.
- More complicated alternatives: pseudo-marginal methods.

Pseudo-marginal methods

- For each $x \in \mathsf{X}$ let $W \sim Q_x$ be a non-negative random variable with $\mathbb{E}_x[W] = \pi(x)$.
- Define

$$\tilde{\pi}(x, w) = \pi(x) \left[Q_x(w) \frac{w}{\pi(x)} \right].$$

  and observe that $\tilde{\pi}(x) = \int_{\mathbb{R}_+} \tilde{\pi}(x, w) \mathrm{d}w = \pi(x)$.
- Metropolis–Hastings for $\tilde{\pi}$: at (x, w) simulate $Z \sim q(x, \cdot)$ and $U \sim Q_Z$ and "accept" with probability

$$\begin{aligned} \alpha(x, w; Z, U) &= 1 \wedge \frac{\tilde{\pi}(Z, U) q(Z, x) Q_x(w)}{\tilde{\pi}(x, w) q(x, Z) Q_Z(U)} \\ &= 1 \wedge \frac{U}{w} \cdot \frac{q(Z, x)}{q(x, Z)}. \end{aligned}$$

- No need to evaluate π exactly!

Hamiltonian Markov chain Monte Carlo (v. briefly)

- This Markov chain is motivated by Hamiltonian dynamics in physics.
- Assume $\mathsf{X} = \mathbb{R}^d$ and π is differentiable.
- We imagine a particle in X evolving in continuous time according to fictitious dynamics according to π and an auxiliary momentum variable p.
- Hamiltonian dynamics are time reversible and measure-preserving:
  - if x is distributed according to $\tilde{\pi}$ and follows these dynamics to produce Z then $Z \sim \tilde{\pi}$.
- The formulation is to use $H = U + V$, where U is the potential energy and V the kinetic energy.
  - U is related to π and V describes the momentum variables.
  - We have $\tilde{\pi}(x, p) \propto \exp(-H(x, p)) = \exp\left(-U(x) - V(p)\right)$.
- In practice, we cannot simulate the system in continuous time so discretization is required.

Simple Hamiltonian Markov chain Monte Carlo

- Define $\tilde{\pi}(x, p) := \pi(x)\mathcal{N}(p; 0, 1)$ and set parameters $h = \frac{1}{L}$, $L \in \mathbb{N}$ and $T \in \mathbb{N}$.
- The following "leapfrog" scheme is an approximation of Hamiltonian dynamics in one dimension.
- At (x, p), sample $P_0 \sim \mathcal{N}(\cdot; 0, 1)$ and set $Z_0 = x$.
- For $l = 0, \ldots, LT - 1$:
  - Set $P_{(l+\frac{1}{2})h} = P_{lh} + \frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}x}\log\pi(Z_{lh})$.
  - Set $Z_{(l+1)h} = Z_{lh} + hP_{(l+\frac{1}{2})h}$.
  - Set $P_{(l+1)h} = P_{(l+\frac{1}{2})h} + \frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}x}\log\pi(Z_{(l+1)h})$.
- Accept $(z, q) := (Z_T, P_T)$ with probability

$$\alpha_{\mathrm{MH}}(x, p; z, q) = 1 \wedge \frac{\tilde{\pi}(z, q)}{\tilde{\pi}(x, p)}.$$

HMC: brief explanation

- We have $U(x) = -\log \pi(x)$ and $V(p) = C(M) + \frac{1}{2} p^T M^{-1} p$:

$$\tilde{\pi}(x, p) = \exp(-U(x) - V(p)) = \pi(x)\mathcal{N}(p; 0, M).$$

- The Hamiltonian dynamics are given by

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial U}{\partial x} = \frac{1}{2}\nabla \log \pi(x), \qquad \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial V}{\partial p} = M^{-1} p.$$

- The h in the algorithm is a discretization step size.
- The deterministic part is "volume preserving" and reversible, the proposal is "symmetric".
- The acceptance probability corrects the time discretization: by discretizing, energy is not preserved.

Multivariate Hamiltonian Markov chain Monte Carlo

- Define $\tilde{\pi}(x, p) := \pi(x)\mathcal{N}(p; 0, M)$ and set parameters $h = \frac{1}{L}$, $L \in \mathbb{N}$ and $T \in \mathbb{N}$.
- At (x, p), sample $P_0 \sim \mathcal{N}(\cdot; 0, M)$ and set $Z_0 = x$.
- For $l = 0, \ldots, LT - 1$:
  - Set $P_{(l+\frac{1}{2})h} = P_{lh} + \frac{h}{2}\nabla \log \pi(Z_{lh})$.
  - Set $Z_{(l+1)h} = Z_{lh} + hM^{-1}P_{(l+\frac{1}{2})h}$.
  - Set $P_{(l+1)h} = P_{(l+\frac{1}{2})h} + \frac{h}{2}\nabla \log \pi(Z_{(l+1)h})$.
- Accept $(z, q) := (Z_T, P_T)$ with probability

$$\alpha_{\mathrm{MH}}(x, p; z, q) = 1 \wedge \frac{\tilde{\pi}(z, q)}{\tilde{\pi}(x, p)}.$$

- M is a "mass matrix". The choice of M, L and T is important.

Outline

Central limit theorems

- Recall that $S_n(f) := \sum_{i=1}^n f(X_i)$, for some $f \in L_1(\mathsf{X}, \pi)$.

Definition

A central limit theorem holds for f if there exists a constant $\sigma^2(f) < \infty$ such that

$$\frac{1}{\sqrt{n}} S_n(\bar{f}) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f))$$

as $n \to \infty$, where $\bar{f} = f - \pi(f)$.

- When a CLT holds for f and a particular chain $\mathbf{X}$ then it is an indication that results can be reliable.
- Perhaps more obvious that if a CLT does not hold, then it is unusual for $n^{-1} S_n(f)$ to be close to $\pi(f)$.

Central limit theorems

- A huge amount of research has gone into characterizations of when a CLT holds.
- In some situations one can verify that it holds!
- We cannot cover even a small fraction of this research.
- Instead, we will look at important classifications of Markov chains for which we can be assured that a CLT holds for all or nearly all reasonable functions f.

Some central limit theorems

Theorem ([Cogburn et al., 1972])

Assume that $\mathbf{X}$ *is positive Harris and uniformly ergodic and that* $\pi(f^2) < \infty$. *Then a CLT holds for* f *and*

$$\sigma^2(f) = \mathsf{E}_\pi\left[\bar{f}(X_0)^2\right] + 2\sum_{k=1}^{\infty} \mathsf{E}_\pi\left[\bar{f}(X_0)\bar{f}(X_k)\right] < \infty.$$

Some central limit theorems

Theorem ([Ibragimov and Linnik, 1971, Chan and Geyer, 1994])

Assume that $\mathbf{X}$ *is positive Harris and geometrically ergodic with invariant probability measure* π*, and that* $\pi(|f|^{2+\delta}) < \infty$ *for some* $\delta > 0$*. Then a CLT holds for* f *and*

$$\sigma^2(f) = \mathsf{E}_\pi\left[\bar{f}(X_0)^2\right] + 2\sum_{k=1}^{\infty} \mathsf{E}_\pi\left[\bar{f}(X_0)\bar{f}(X_k)\right] < \infty.$$

Some central limit theorems

Theorem ([Roberts and Rosenthal, 1997])

Assume that $\mathbf{X}$ *is positive Harris,* π*-reversible and geometrically ergodic, and that* $\pi(f^2) < \infty$*. Then a CLT holds for* f *and*

$$\sigma^2(f) = \mathrm{E}_\pi\left[\bar{f}(X_0)^2\right] + 2\sum_{k=1}^{\infty} \mathrm{E}_\pi\left[\bar{f}(X_0)\bar{f}(X_k)\right] < \infty.$$

Remarks

- There are a number of different CLTs, with different conditions.
- There are also different proof techniques and different expressions for $\sigma^2(f)$.
- It appears from the above that uniform and geometric ergodicity are beneficial properties.
- While true, they are not essential nor necessarily better than non-geometrically ergodic counterparts in specific settings.

Asymptotic variance

- The expression for $\sigma^2(f)$ we have seen is not unusual.
- Imagine $\mathbf{X}$ with initial distribution π and $f \in L_2(\mathsf{X}, \pi)$. Then

$$\begin{aligned}
\mathrm{var}(S_n(f)) &= \mathrm{var}(S_n(f) - \pi(f)) = \mathrm{var}(S_n(\bar{f})) \\
&= \mathsf{E}_\pi[\{\sum_{i=1}^{n} \bar{f}(X_i)\}^2] - n\pi(\bar{f})^2 \\
&= \mathsf{E}_\pi\left[\sum_{i=1}^{n} \bar{f}(X_i)^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \bar{f}(X_i)\bar{f}(X_j)\right] \\
&= n\pi(\bar{f}^2) + 2\sum_{k=1}^{n-1}(n-k)\mathsf{E}_\pi\left[\bar{f}(X_0)\bar{f}(X_j)\right].
\end{aligned}$$

So the variance of $\frac{1}{\sqrt{n}} S_n(f)$ is

$$\mathsf{E}_\pi\left[\bar{f}(X_0)^2\right] + 2\sum_{k=1}^{n-1} \frac{n-k}{n} \mathsf{E}_\pi\left[\bar{f}(X_0)\bar{f}(X_k)\right]$$

whose limit (if it exists) is $\sigma^2(f)$.

Optimality of Metropolis–Hastings

Theorem (Peskun [1973], Tierney [1998])

Let Q be fixed. Amongst reversible Markov kernels P of the form

$$P(x, A) = \int_A Q(x, \mathrm{d}z)\alpha(x, z) + r(x)\mathbf{1}_A(x),$$

where $r(x) = 1 - \int_{\mathsf{X}} \alpha(x, z)Q(x, \mathrm{d}z)$, the one minimizing $\sigma^2(f)$ for all $f \in L_2(\mathsf{X}, \pi)$ is the Metropolis–Hastings kernel.

- This is a statement about the form of $\alpha_{\mathrm{MH}}(x, z)$.
- There are many valid "acceptance probability" functions but they are dominated by α_{MH}.
- Note: this tells us nothing about non-reversible Markov kernels, or about non-asymptotic variance.

Outline

Total variation distance

Definition

The total variation distance between two probability measures μ and ν on $\mathcal{X}$ is

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \in \mathcal{X}} |\mu(A) - \nu(A)|.$$

Ergodic Markov chains

Definition

A Markov chain with invariant probability measure π and Markov transition kernel P is ergodic if

$$\lim_{n \to \infty} \|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} = 0,$$

for any $x \in \mathsf{X}$.

- That is, the probability measure associated with X_n when $X_0 = x$ is converging to π in total variation.
- Note: this is not a universal definition of ergodic.

Note on aperiodicity

- It is important to note that an ergodic Markov chain, as we have defined, cannot be periodic.
- Loosely speaking, there cannot be disjoint sets $D_1, \ldots, D_d$ such

$$\inf_{x \in D_i} P(x, D_{i+1}) = 1, \quad i \in \{1, \ldots, d-1\}$$

  and $\inf_{x \in D_d} P(x, D_1) = 1$, where $d > 1$.
- While clearly this is not obviously an issue for the LLN or even the CLT, we will assume from now on that we are dealing with aperiodic Markov chains.
- In fact, periodic behaviour is exceedingly rare amongst Monte Carlo Markov chains.

Uniform ergodicity

Definition

A Markov chain with invariant probability measure π and Markov transition kernel P is uniformly ergodic if

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq M\rho^n, \qquad x \in \mathsf{X}$$

for some constant M and $\rho < 1$.

- The total variation distance decreases geometrically fast, with ρ governing this rate, and the bound is independent of x.

Geometric ergodicity

Definition

A Markov chain with invariant probability measure π and Markov transition kernel P is geometrically ergodic if

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq M(x)\rho^n, \qquad x \in \mathsf{X}$$

for some function M finite for π-almost all $x \in \mathsf{X}$ and $\rho < 1$.

- The total variation distance decreases geometrically fast, with ρ governing this rate, and the bound is dependent on x.
- For some intuition, recall the simple random walk chain on $\mathbb{N}$.

Verifying uniform ergodicity

- One way to verify uniform ergodicity for an aperiodic, π-irreducible Markov chain is to check that

$$P^m(x, A) \geq \epsilon \nu(A), \quad x \in \mathsf{X}, A \in \mathcal{X},$$

  for some $m \in \mathbb{N}$, $\epsilon > 0$ and probability measure ν.
- This is called a minorization condition.
- In this case it is basically Doeblin's condition and is equivalent to uniform ergodicity.
- Important observation: $P^m(x, \cdot)$ and $P^m(x', \cdot)$ have, loosely speaking, some degree of similarity.

A simple quantitative proof of uniform ergodicity

- We will look at the case where the minorization condition is satisfied for $m = 1$, for simplicity.
- The method of proof is by coupling, due to Doeblin.
- We assume that $P(x, \cdot) \geq \epsilon\nu(\cdot)$ and will show that

$$\|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} \leq (1 - \epsilon)^n.$$

- We define a residual Markov kernel

$$R(x, A) := \frac{P(x, A) - \epsilon\nu(A)}{1 - \epsilon}, \qquad x \in \mathsf{X}, A \in \mathcal{X},$$

and observe that $P(x, \cdot) = \epsilon\nu(\cdot) + (1 - \epsilon)R(x, \cdot)$.

A simple quantitative proof of uniform ergodicity

- Loosely, a coupling between two probability measures μ and ν on $\mathcal{X}$ is a pair of random variables (X, Y) defined on a common probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν.
- The coupling inequality states that for any such construction

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \Pr(X \neq Y).$$

- So we will show an explicit coupling such that

$$\Pr(X_n \neq Y_n) \leq (1 - \epsilon)^n$$

where X_n is distributed according to $P^n(x, \cdot)$ and Y_n is distributed according to π.

A simple quantitative proof of uniform ergodicity

- Let $X_0 = x$ and $Y_0 \sim \pi$.
- Now follow the procedure for each time $n \geq 1$:
  1. If $X_{n-1} = Y_{n-1}$, sample $Z_n \sim P(X_{n-1}, \cdot)$, set $X_n = Y_n = Z_n$.
  2. Otherwise, with probability ϵ, sample $Z_n \sim \nu$ and set $X_n = Y_n = Z_n$.
  3. Otherwise, sample $X_n \sim R(X_{n-1}, \cdot)$ and $Y_n \sim R(Y_{n-1}, \cdot)$ independently.
- We observe that we have not changed the marginal distributions of X_n or Y_n, so $X_n \sim P^n(x, \cdot)$ and $Y_n \sim \pi P^n = \pi$.
- We also observe that

$$\Pr(X_n \neq Y_n) \leq (1 - \epsilon)^n.$$

- Hence, $\|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \Pr(X_n \neq Y_n) \leq (1 - \epsilon)^n$.

Example: independent Metropolis–Hastings

- Recall that $P(x, A) = \int_A q(z)\alpha_{\mathrm{MH}}(x, z)\mathrm{d}z + r(x)\mathbf{1}_A(x)$, where $\alpha_{\mathrm{MH}}(x, z) = 1 \wedge \frac{\pi(z)q(x)}{\pi(x)q(z)}$.
- Now assume that $\sup_x \pi(x)/q(x) = K < \infty$. Then we have

$$\begin{aligned} q(z)\alpha_{\mathrm{MH}}(x, z) &= q(z)\left[1 \wedge \frac{\pi(z)q(x)}{\pi(x)q(z)}\right] \\ &= \pi(z)\left[\frac{q(z)}{\pi(z)} \wedge \frac{q(x)}{\pi(x)}\right] \geq K^{-1}\pi(z) \end{aligned}$$

  and so $P(x, A) \geq K^{-1}\pi(A)$.
- Therefore, $\mathbf{X}$ is uniformly ergodic and

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq (1 - K^{-1})^n.$$

- In fact, if $\sup_x \pi(x)/q(x) = \infty$ then $\mathbf{X}$ is not even geometrically ergodic.

Small sets

- When $\mathbf{X}$ is evolving on a general state space, there is no guarantee that two independent copies of $\mathbf{X}$ will visit a particular state simultaneously.
- The minorization condition allowed us to successfully couple the two Markov chains with probability ϵ at each time.
- Of course, uniform ergodicity and therefore the minorization condition we have seen is very strong in practice.
- This motivates the definition of a small set, which is essentially a set for which points are "similar".

Definition

A set $C \in \mathcal{X}$ is small if

$$P^m(x, A) \geq \epsilon \nu(A), \quad x \in C, A \in \mathcal{X},$$

for some $m \in \mathbb{N}$, $\epsilon > 0$ and probability measure ν.

Verifying geometric ergodicity

- The presence of a small set is only one of two ingredients required for an aperiodic, π-irreducible Markov chain to be geometrically ergodic.
- Intuitively, one can use a coupling argument if both chains are in the small set C.
- We need to ensure that they are both in C simultaneously "often enough".
- A "drift condition" that ensures geometric ergodicity is

$$\int_{\mathsf{X}} V(z)P(x, \mathrm{d}z) \leq \lambda V(x) + b\mathbf{1}_C(x),$$

where $\lambda \in (0, 1)$, $b < \infty$ and $V : \mathsf{X} \to [1, \infty]$ satisfies $V(x) < \infty$ for at least one $x \in \mathsf{X}$.

- This condition guarantees that

$$\sup_{x \in C} \mathsf{E}_x \left[\kappa^{\tau_C}\right] < \infty,$$

for some $\kappa > 1$, where $\tau_A := \inf\{n \geq 1 : X_n \in A\}$.

Example: simple random walk on $\mathbb{N}$

- Here the small set is no problem, we can take $C = \{1\}$ and so $P(x, A) = \nu(A)$ for each $x \in C$ where $\nu(\cdot) = P(x, \cdot)$.
- We take $V(x) = c^x$, $c > 1$ and we have

$$\begin{aligned} \int_{\mathsf{X}} V(z) P(x, \mathrm{d}z) &= rV(x) + pV(x-1) + qV(x+1) \\ &= V(x)(r + p/c + qc). \end{aligned}$$

- If $q < p$ and $c \in (1, \frac{p}{q})$ then $r + \frac{p}{c} + qc < 1$.
- One choice, e.g., is $c = \sqrt{p/q}$, so that one can take $\lambda = r + 2\sqrt{pq}$.
- In Kovchegov [2010], e.g., it is shown that

$$\|P^n(1, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq A \left(\frac{p}{p+r} \right)^n + B \left(r + 2\sqrt{pq} \right)^n,$$

where $A, B \in \mathbb{R}_+$ for a very slight difference of the Markov chain's behaviour at 1.

Remarks

- In practice, small sets are often possible to identify.
- The drift condition is usually harder, but it is still possible in some cases.
- Drift conditions and return times are alternative ways to characterize many of the stability criteria we have talked about.
- For example, **X** is "regular" (and therefore positive) iff

$$\sup_{x \in C_j} \mathsf{E}_x(\tau_A) < \infty, \quad A \in \mathcal{X},\ \psi(A) > 0,\ \mathsf{X} = \cup_j C_j.$$

- Alternatively, **X** is regular iff

$$\int_{\mathsf{X}} V(z) P(x, \mathrm{d}z) \leq V(x) - 1 + b\mathbf{1}_C(x), \quad x \in \mathsf{X}, \quad C \text{ "petite".}$$

Outline

Conclusions

- We have covered a tiny fraction of what is interesting and relevant.
- Hopefully, you have a clear idea of the fundamental theorems underpinning the use of MCMC in statistical computations.
- If you are doing modern Bayesian inference, it is very common to use MCMC.
- Research in this area is extremely varied:
  - theory
  - intuition-based methodology
  - theory-based methodology
  - hybrids of the two
  - applications $\leftrightarrow$ methodology $\leftrightarrow$ theory $\leftrightarrow$ applications.

What we didn't cover

- We have covered a tiny fraction of what is interesting and relevant.
- Markov chains and their use in Monte Carlo are very large research areas.
- Just a few things that we didn't cover are:
  - the splitting construction underpinning many of the results
  - perfect simulation
  - spectral properties of P
  - adaptive Markov chain Monte Carlo
  - optimal scaling
  - subgeometric rates of convergence and corresponding CLTs
  - genuinely non-reversible Markov chains
  - more methodology
  - non-homogeneous Markov chains
  - exact approximations
  - inexact approximations
  - ...

Further reading

- Meyn & Tweedie: Markov chains & stochastic stability (available online to the public)
- Handbook of Markov chain Monte Carlo (available online through proxy).
- Robert & Casella. Monte Carlo Statistical Methods.
- Liu: Monte Carlo Strategies in Scientific Computing.
- Roberts & Rosenthal: General state space Markov chains and MCMC algorithms
- Jones: On the Markov chain central limit theorem.
- Look for Markov chain Monte Carlo papers in Ann. Stat., JRSS B, Biometrika, JASA, JCGS, Stats & Comp.
- It is impossible to be comprehensive!

References I

K. S. Chan and C. J. Geyer. Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1747–1758, 1994.

R. Cogburn et al. The central limit theorem for Markov processes. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab*, volume 2, pages 485–512, 1972.

L. Devroye. *Non-uniform random variate generation*. Springer Verlag, 1986.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

References II

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

I. A. Ibragimov and Y. V. Linnik. *Independent and stationary sequences of random variables*. Wolters-Noordhoff, 1971.

Y. Kovchegov. Orthogonality and probability: mixing times. *Electronic Communications in Probability*, 15:59–67, 2010.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2 edition, 2009.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

References III

G. O. Roberts and J. S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2(2): 13–25, 1997.

G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Annals of Applied Probability*, 16(4):2123–2139, 2006.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1762, 1994.

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8(1):1–9, 1998.