Markov Chain Monte Carlo Methods

Juan F. Rubio-Ramírez

Emory University

“Bayesianism has obviously come a long way. It used to be that you could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogations, and other manifestations of paranoia. Things have changed...”

Peter Clifford (1993)

Our Goal

- We have a distribution:
$$X \sim f(X)$$
such that $f > 0$ and $\int f(x)\,dx < \infty$.
- How do we draw from it?
- We could use Importance Sampling...
- ...but we need to find a good source density.

Transition Kernels I

- The function $P(x, A)$ is a transition kernel for $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ (a Borel σ-field) such that:
  1. For all x, $P(x, \cdot)$ is a probability measure.
  2. For all A, $P(\cdot, A)$ is measurable.
- When $\mathcal{X}$ is discrete, $P_{xy} = P(X_n = y \mid X_{n-1} = x)$.
- When $\mathcal{X}$ is continuous,

$$P(X \in A \mid x) = \int_A P(x, x')\, dx'.$$

Transition Kernels II

- Clearly $P(x, \mathcal{X}) = 1$.
- It may be that $P(x, \{x\}) \neq 0$.
- Examples in economics: capital accumulation, job search, prices in financial markets, . . .

A Particular Transition Kernel

Define:

$$P(x, dy) = p(x, y)\, dy + r(x)\, \delta_{\{x\}}(dy)$$

or

$$P(x, A) = \int_A p(x, y)\, dy + r(x) \int_A \delta_{\{x\}}(dy) = \int_A p(x, y)\, dy + r(x) 1_A(x)$$

1. $p(x, y) \geq 0,\ p(x, x) = 0.$
2. $\delta_{\{x\}}(dy)$ is the Dirac delta.
3. $P(x, x)$, the probability the chain stays at x, is $r(x)$.
4. By construction:

$$r(x) = 1 - \int_{\mathcal{X}} p(x, y)\, dy.$$

Markov Chain

- Given a transition kernel P, a sequence $X_0, X_1, \ldots$ is a Markov Chain if for any k

$$P(X_{k+1} \in A \mid x_0, \ldots, x_k) = P(X_{k+1} \in A \mid x_k) = \int_A P(x_k, dx).$$

- We focus on **time-homogeneous** chains: the distribution of $(X_{t_1}, \ldots, X_{t_k})$ given x_{t_0} is the same as that of $(X_{t_1 - t_0}, \ldots, X_{t_k - t_0})$ given x_0.

Chapman–Kolmogorov Equations

- For a time-homogeneous chain:

$$P^{m+n}(x, A) = \int_{\mathcal{X}} P^{n}(y, A)\, P^{m}(x, dy).$$

- Implies convolution formula: $P^{m+n} = P^{m} \star P^{n}$.
- Discrete case: this is a matrix product. Continuous case: P acts as an operator:

$$Ph(dy) = \int_{\mathcal{X}} P(x, dy) h(x)\, dx.$$

Two Important Questions in Markov Chains

Knowing $P(x, B)$:

- Does there exist a fixed point π_s such that

$$\pi_s(B) = \int_{\mathcal{X}} P(x, B)\pi_s(dx)?$$

- If $P^1(x, dy) = P(x, dy)$, does $P^n(x, B) \to \pi_s(B)$ as $n \to \infty$?
- Reference: Meyn & Tweedie (1993), *Markov Chains and Stochastic Stability*.

MCMC Questions: Knowing π_s

- Does there exist a kernel $P(x, B)$ such that

$$\pi_s(B) = \int_{\mathcal{X}} P(x, B)\pi_s(dx)?$$

- If $P^1(x, dy) = P(x, dy)$, does $P^n(x, B) \to \pi_s(B)$ as $n \to \infty$?
- Reference: Chib & Greenberg (1995), “Understanding the Metropolis–Hastings Algorithm.”

Markov Chain Monte Carlo Methods

- An MCMC method simulates $f(x)$ by producing an ergodic Markov Chain with invariant distribution $f(x)$.
- We seek a chain such that if $X^1, X^2, \ldots, X^t$ are realizations,

$$X^t \to X \sim f(x)$$

as $t \to \infty$.

Turning the Theory Around

- For equilibrium models: we know the kernel (policy functions) $\rightarrow$ find invariant distribution.
- For MCMC: we know the invariant distribution $\rightarrow$ find a kernel that produces it.
- Question: How do we find such a transition kernel?

Roadmap

We search for a transition kernel that:

1. Has stationary distribution $f(x)$.
2. Stays within that stationary distribution (convergence).
3. Converges to it.
4. Obeys a Law of Large Numbers.
5. Admits a Central Limit Theorem.

f as Invariant Density and f-Reversibility

Definition
The density f is stationary for a Markov kernel P if

$$\int_A f(y)\,dy = \int_{\mathcal{X}} f(x)\,P(x,A)\,dx \qquad \forall A.$$

Definition
The Markov kernel P is f-*reversible* if

$$\forall g \qquad \iint g(x,y)\,f(x)\,dx\,P(x,dy) = \iint g(y,x)\,f(y)\,dy\,P(y,dx).$$

Time Reversibility

Definition
The Markov kernel P is *time reversible* if

$$f(x)\,p(x,y) = f(y)\,p(y,x) \quad \text{for a.e. } (x,y).$$

where

$$P(x,dy) = p(x,y)\,dy + r(x)\,\delta_{\{x\}}(dy)$$

Equivalence

Then the following are equivalent:

(i) P is f-reversible

(ii) P is time reversible

Proof.
Expanding both sides of the f-reversibility condition, the singular terms involving δ_x cancel automatically. The remaining absolutely continuous parts imply

$$\iint g(x,y)\, f(x)p(x,y)\, dx\, dy = \iint g(y,x)\, f(y)p(y,x)\, dy\, dx \quad \forall g,$$

which holds if and only if $f(x)p(x,y) = f(y)p(y,x)$ a.e. □

Detailed Balance

Let f be a density and let P be a Markov kernel.

Definition (Detailed balance)
We say that (f, P) satisfies *detailed balance* if for all A, B,

$$\int_B f(x)\, P(x, A)\, dx = \int_A f(y)\, P(y, B)\, dy.$$

Lemma: Detailed Balance ⇒ Stationary and Reversible

Lemma
If (f, P) satisfies detailed balance, then

1. *f is stationary for P, i.e. for all A,*

$$\int_{\mathcal{X}} f(x)\, P(x, A)\, dx = \int_A f(y)\, dy;$$

2. *P is f-reversible, i.e. for all g,*

$$\iint g(x, y)\, f(x)\, dx\, P(x, dy) = \iint g(y, x)\, f(y)\, dy\, P(y, dx).$$

Proof (Sketch)

Proof sketch.

- **Stationarity:** Apply detailed balance with $B = \mathcal{X}$:

$$\int_{\mathcal{X}} f(x)\, P(x, A)\, dx = \int_{A} f(y)\, P(y, \mathcal{X})\, dy = \int_{A} f(y)\, dy,$$

  since $P(y, \mathcal{X}) = 1$.

- *f*-**reversibility:** Detailed balance implies equality of the two measures,

$$\mu_1(dx, dy) := f(x)\, dx\, P(x, dy), \qquad \mu_2(dx, dy) := f(y)\, dy\, P(y, dx),$$

  because $\mu_1(B \times A) = \mu_2(B \times A)$ for all measurable rectangles $B \times A$. Hence $\mu_1 = \mu_2$, and integrating any g yields the stated identity (we can use the monotone class theorem)

Searching for a Transition Kernel $P(x, A)$

- Let $P(x, dy) = p(x, y)dy + r(x)\delta_{\{x\}}(dy)$.
- If $f(x)p(x, y) = f(y)p(y, x)$ (time reversibility), then

$$\int_A f(y)\,dy = \int_{\mathcal{X}} P(x, A)f(x)\,dx.$$

- This means that $f(x)$ is a stationary distribution.
- Time reversibility also called detailed balance.

Proof Sketch

$$\begin{aligned}
\int_{\mathcal{X}} P(x,A)f(x)\,dx &= \int_{\mathcal{X}} \left[\int_A p(x,y)\,dy\right] f(x)\,dx + \int_{\mathcal{X}} r(x)\delta_{\{x\}}(A)f(x)\,dx \\
&= \int_A \left[\int_{\mathcal{X}} p(x,y)f(x)\,dx\right] dy + \int_A r(x)f(x)\,dx \\
&= \int_A \left[\int_{\mathcal{X}} p(y,x)f(y)\,dx\right] dy + \int_A r(x)f(x)\,dx \\
&= \int_A (1 - r(y))f(y)\,dy + \int_A r(x)f(x)\,dx = \int_A f(y)\,dy.
\end{aligned}$$

Remarks

- The condition $f(x)p(x, y) = f(y)p(y, x)$ ensures f is the invariant distribution.
- Time reversibility is the key property for MCMC algorithms.

Convergence

- We proved f is a fixed point of the kernel operator.
- We ask: does $P^m(x, A) \to \pi_s(A)$ as $m \to \infty$?

Sufficient Conditions for Convergence

If the kernel satisfies time reversibility and:

- **Irreducibility:** any x can reach any A with positive probability.
- **Aperiodicity:** chain does not have periodic behavior.
- A irreducible Markov chain is **Harris recurrent** if for any measurable set $\mu(A) > 0$, we have

$$\forall x \in \mathcal{X} \qquad \mathbb{P}_x(\eta_A = \infty) = 1.$$

where

$$\eta_A = \sum_{k=1}^{\infty} 1_A(X_k)$$

A Law of Large Numbers

If $P(x, A)$ is irreducible and aperiodic with invariant distribution π_s:

1. π_s is unique.
2. For all integrable h:

$$\frac{1}{M}\sum_{i=1}^{M} h(x_i) \to \int h(x)\pi_s(dx),$$

i.e.

$$\widehat{h} \to Eh.$$

Building our MCMC

We need a kernel $P(x, A)$ such that:

1. Time reversible ($f(x)$ invariant).
2. Irreducible (convergence + LLN).
3. Aperiodic (convergence + LLN).
4. Harris-recurrent and geometrically ergodic (CLT).

These are sufficient for validity of MCMC.

MCMC and Metropolis–Hastings

- The Metropolis–Hastings algorithm is the canonical MCMC method.
- Gibbs sampler is a special case.
- Many variants exist—be cautious!
- The frontier: Perfect Sampling.

On the Use of MCMC

- Motivation: draw from a posterior distribution.
- But MCMC applies more broadly:
    1. It can sample from any distribution, not necessarily Bayesian posteriors.
    2. It explores distributions and can be used for classical estimation as well.