# 10 Markov Chain Monte Carlo Simulations

This is a very broad area and would actually merit a full main section of its own. Maybe later.

In many practical applications of Markov chains, one is interested not just in sampling according to the stationary distribution $\pi$, but also in computing expected values of various quantities with respect to it:

$$E_\pi[f] = \sum_{\sigma \in S} f(\sigma)\pi_\sigma \quad (\text{also denoted } \langle f \rangle_\pi)$$

E.g. one might want to compute the average magnetisation of a spin glass model at a given inverse temperature $\beta$ (cf. page 62):

$$\langle M \rangle = \sum_{\sigma \in S} M(\sigma) \cdot \underbrace{e^{-\beta H(\sigma)}/Z_\beta}_{\text{Gibbs density}}$$

The task could be approached by producing many independent sample states $\sigma$ according to $\pi$, computing $f(\sigma)$ for each and controlling the estimation error.

However, it is customary to compute the estimates from a single (or a few) long runs of the chain:

$$E_\pi[f] \approx \frac{1}{N} \sum_{k=1}^{N} f(X_k(\omega)), \quad N \text{ large}$$

(More precisely, maybe

$$E_\pi[f] \approx \frac{1}{N - N_0} \sum_{k=N_0+1}^{N} f(X_k(\omega)),$$

where $N_0$ is an initial “burn-in” time to eliminate systematic effects of choice of the initial state.)

For this approach to work properly, the Markov chains must be “path-ergodic” in the sense that the stationary distribution is sampled properly along almost every individual path of the chain.

In fact, if the word was not already so overused, we could define a Markov chain $\mathcal{M} = (X_1, X_2, \ldots)$ to be *ergodic with stationary distribution* $\pi$ if for any initial distribution $\mu$ and for all states $\sigma \in S$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} I_\sigma(X_k) = \pi_\sigma \quad \mu\text{-almost surely,}$$

i.e.

$$\Pr_\mu\left(\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} I_\sigma(X_k(\omega)) \neq \pi_\sigma\right) = 0,$$

where $I_\sigma$ is an indicator function for state $\sigma$:

$$I_\sigma(\xi) = \begin{cases} 1, & \text{if } \xi = \sigma \\ 0, & \text{if } \xi \neq \sigma \end{cases}$$

Luckily, all regular (finite) Markov chains are ergodic also in this strong sense. In fact, even more is true:

**Theorem 10.1 (Ergodic Theorem for Regular Markov Chains)**
*Let $\mathcal{M} = (X_1, X_2, \ldots)$ be a regular Markov chain with state space $S$, and $f : S \to \mathbb{R}$ any function. Then for any initial distribution $\mu$:*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} f(X_k) = E_\pi[f] \quad \mu\text{-almost surely.}$$

We do not have all the tools (or the time) to give a complete proof of Theorem 10.1, but here are the key components:

**Theorem 10.2 (Kolmogorov's Strong Law of Large Numbers)**
*Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables defined on probability space $(\Omega, \mathcal{F}, P)$, and such that $E[|X_k|] = E[|X_1|] < \infty$ for all $k$. Then*

$$\lim_{N\to\infty}\frac{1}{N}(X_1 + \ldots + X_N) = E[X_1] \quad P\text{-almost surely.}$$

**Lemma 10.3 (Regenerative Cycle Lemma / Strong Markov Property)**
*Let $\mathcal{M} = (X_0, X_1, \ldots)$ be a regular finite Markov chain with state space $S$. Fix any state $0 \in S$. Then $0$ is visited on any given sample path of $\mathcal{M}$ infinitely often (almost surely), and denoting $\tau_0, \tau_1, \tau_2, \ldots$ the successive times of visit to $0$, the sample path segments*

$$\{X_{\tau_k}, X_{\tau_k+1}, \ldots, X_{\tau_{k+1}-1}\}, \quad k \geq 0,$$

*are independent and identically distributed.*

*Proof of Theorem 10.1:* Recall that for any $\sigma \in S$:

$$\pi_\sigma = \frac{\rho_\sigma}{\mu_0} = \frac{1}{\mu_0} \cdot E_0 \left[ \sum_{n \geq 1} I_{[X_n = \sigma]} I_{[\tau_1 > n]} \right] = \frac{1}{\mu_0} E_0 \left[ \sum_{n=1}^{\tau_1} I_{[X_n = \sigma]} \right],$$

where $E_0[\cdot] = E[\cdot | X_0 = 0]$, $\tau_1$ is the time of first return to 0, and $\mu_0 = E[\tau_1]$.

Given a sample path starting at state 0, let $\tau_1, \tau_2, \ldots$ be the successive return times to 0, and define

$$U_p = \sum_{n=\tau_p+1}^{\tau_{p+1}} f(X_n).$$

By Lemma 10.3, the $U_p$'s are independent and identically distributed random variables. Assuming $f \geq 0$ we obtain:

$$\begin{aligned} E[U_0] &= E_0 \left[ \sum_{n=1}^{\tau_1} f(X_n) \right] \\ &= E_0 \left[ \sum_{n=1}^{\tau_1} \sum_{\sigma \in S} f(\sigma) I_{[X_n = \sigma]} \right] \\ &= \sum_{\sigma \in S} f(\sigma) E_0 \left[ \sum_{n=1}^{\tau_1} I_{[X_n = \sigma]} \right] \\ &= \mu_0 \sum_{\sigma \in S} f(\sigma) \pi_\sigma = \mu_0 E_\pi[f] \end{aligned}$$

By Theorem 10.2 (Strong Law of Large Numbers), then:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{p=1}^{n} U_p = E[U_0] = \mu_0 E_\pi[f] \quad \eta\text{-almost surely,}$$

i.e.

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=\tau_1+1}^{\tau_{n+1}} f(X_k) = \mu_0 E_\pi[f] \quad \eta\text{-almost surely.} \tag{7}$$

Define then random variables $\nu(n)$ as:

$$\nu(n) = \sum_{k=1}^{n} I_{[X_k = 0]}$$

($\sim$ number of returns to 0 by time $n$). Clearly $\tau_{\nu(n)} \leq n < \tau_{\nu(n)+1}$ for all $n$, so that

$$\frac{1}{\nu(n)} \sum_{k=1}^{\tau_{\nu(n)}} f(X_k) \leq \frac{1}{\nu(n)} \sum_{k=1}^{n} f(X_k) < \frac{1}{\nu(n)} \sum_{k=1}^{\tau_{\nu(n)+1}} f(X_k) \quad \text{almost surely.}$$

Since by Lemma 10.3, $\nu(n) \to \infty$ as $n \to \infty$, we obtain from equation (7):

$$\lim_{n\to\infty} \frac{1}{\nu(n)} \sum_{k=1}^{n} f(X_k) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\tau_{n+1}} f(X_k) = \mu_0 E_\pi[f] \quad \text{almost surely.} \tag{8}$$

However, asymptotically also

$$n \sim \tau_{\nu(n)} = \sum_{i=0}^{\nu(n)-1} (\tau_{i+1} - \tau_i) \quad \text{almost surely,}$$

so by Lemma 10.3 and Theorem 10.2:

$$\frac{n}{\nu(n)} \sim \frac{1}{\nu(n)} \sum_{i=0}^{\nu(n)-1} (\tau_{i+1} - \tau_i) = E[\tau_1] = \mu_0 \quad \text{almost surely.}$$

Thus $\mu_0 \nu(n) \sim n$, and by combining equations (7) and (8):

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) &= \lim_{n\to\infty} \frac{1}{\mu_0 \nu(n)} \sum_{k=1}^{n} f(X_k) \quad \text{almost surely} \\ &= E_\pi[f]. \end{aligned}$$

The case of general $f : S \to \mathbb{R}$ can be handled by treating separately the nonnegative functions

$$f^+ = \max\{f, 0\} \text{ and } f^- = \max\{-f, 0\}$$

and summing up the resulting equalities. □

### Convergence Rates of MCMC Simulation Algorithms

Let $\mathcal{M} = (X_0, X_1, \dots)$ be a regular finite Markov chain with state space $S = \{1, \dots, r\}$, transition probability matrix $P$, and stationary distribution $\pi$. Denote:

$$\Pi = \begin{bmatrix} \pi_1 & \cdots & \pi_r \\ \pi_1 & \cdots & \pi_r \\ & \vdots & \\ \pi_1 & \cdots & \pi_r \end{bmatrix} \quad \text{(i.e. for any distribution } \mu,\ \mu^T \Pi = \pi^T\text{).}$$

The *fundamental matrix* of chain $\mathcal{M}$ is defined as

$$Z = (I - (P - \Pi))^{-1}.$$

**Proposition 10.4** *For a regular chain $\mathcal{M}$, the fundamental matrix $Z$ is well-defined, and*

$$Z = I + \sum_{n \geq 1} (P^n - \Pi).$$

*Proof:* It is easy to verify that for all $k \geq 1$:

$$P\Pi^k = \Pi^k P = \Pi.$$

Thus,

$$\begin{aligned}
(P-\Pi)^n &= \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} P^k \Pi^{n-k} \\
&= P^n + \sum_{k=0}^{n-1} \binom{n}{k} (-1)^{n-k} \Pi \\
&= P^n - \Pi.
\end{aligned}$$

Therefore, with $A = P - \Pi$,

$$(I-A)(I+A+A^2+\ldots+A^{n-1}) = I - A^n = I + P^n - \Pi,$$

and consequently

$$(I-A)(I+\sum_{n\geq 1} A^n) = \lim_{n\to\infty}(I+P^n-\Pi) = I.$$

Hence the matrix $I - A = I - (P - \Pi)$ is invertible, and

$$(I-(P-\Pi))^{-1} = I + \sum_{n\geq 1}(P-\Pi)^n = I + \sum_{n\geq 1}(P^n - \Pi). \;\square$$

The fundamental matrix has many uses (analogous to the fundamental matrix of transient states) in computing expected recurrence times etc.

We, however, quote only the one of main interest to us (and even that without its somewhat technical proof). Given a Markov chain $\mathcal{M}$ with finite state space $S$, and any functions $f, g : S \to \mathbb{R}$, denote:

$$\begin{aligned}
\langle f, g\rangle_\pi &= E_\pi[f(X)g(X)] = \sum_{i\in S} \pi(i) f(i) g(i) \\
\mathrm{Var}_\mu(f) &= E_\mu[(f(X)-\bar{f})^2] = E_\mu[f(X)^2] - (\underbrace{E_\mu[f(X)]}_{\bar{f}})^2
\end{aligned}$$

**Theorem 10.5 (Asymptotic variance of Ergodic Estimates)** *For a regular chain $\mathcal{M}$, and any function $f : S \to \mathbb{R}$,*

$$\lim_{N\to\infty} \frac{1}{N} Var_\mu \left( \sum_{k=1}^{N} f(X_k) \right) = \underbrace{2\langle f, Zf\rangle_\pi - \langle f, (I+\Pi)f\rangle_\pi}_{Denote\ v(f,P,\pi)}$$

*for any initial distribution $\mu$.*

*Proof:* E.g. Brémaud 1999, pages 232-234. □

Since by Theorem 10.1,

$$\tilde{f}_N = \frac{1}{N}\sum_{k=1}^{N} f(X_k) \underset{\text{a.s.}}{\longrightarrow} \bar{f} = E_\pi[f],$$

by Chebyshev's inequality we see that for any $\delta > 0$ and for "large $N$":

$$\Pr(|\tilde{f}_N - \bar{f}| \geq \delta) \leq \frac{1}{\delta^2}\mathrm{Var}(\tilde{f}_N) = \frac{1}{\delta^2 N^2}\mathrm{Var}\left(\sum_{k=1}^{N} f(X_k)\right) \approx \frac{v(f,P,\pi)}{\delta^2 N}$$

independent of the initial distribution $\mu$.

Suppose then that the transition probability matrix $P$ has $r$ distinct eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r > -1$, with associated left and right eigenvectors $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$, respectively (normalized so that $u_i^T v_i = 1 \quad \forall\ i$). Then:[2]

$$P^n = \sum_{i=1}^{r} \lambda_i^n v_i u_i^T = \Pi + \sum_{i=2}^{r} \lambda_i^n v_i u_i^T,$$

and so

$$Z = I + \sum_{n\geq 1}(P^n - \Pi) = I + \sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} v_i u_i^T.$$

Thus

$$\begin{aligned} v(f,P,\pi) &= 2\langle f, Zf\rangle_\pi - \langle f, (I+\Pi)f\rangle_\pi \\ &= 2\langle f, f\rangle_\pi + 2\sum_{i=2}^{r}\frac{\lambda_i}{1-\lambda_i}\langle f, v_i\rangle_\pi (u_i^T f) - \langle f, f\rangle_\pi - \langle f, \Pi f\rangle_\pi \\ &= \underbrace{\langle f, (I-\Pi)f\rangle_\pi}_{\mathrm{Var}_\pi(f(X_0))} + 2\sum_{i=2}^{r}\frac{\lambda_i}{1-\lambda_i}\langle f, v_i\rangle_\pi (f^T u_i). \end{aligned}$$

[2] Cf. page 16. Also left eigenvectors are here represented as column vectors, however.

For a reversible chain ($D^{1/2}PD^{-1/2}$ symmetric), $u_i = Dv_i$ and therefore $f^T u_i = \langle f, v_i\rangle_\pi$. Applying the decomposition $f = \sum_i \langle f, v_i\rangle_\pi v_i$ we obtain in this case

$$v(f,P,\pi) = \sum_{i=2}^{r} \frac{1+\lambda_i}{1-\lambda_i} |\langle f, v_i\rangle_\pi|^2.$$

Let us then consider the task of designing good “Metropolis-like” reversible Markov chains with given stationary distribution $\pi$ and as good convergence rate as possible.

To achieve a given stationary distribution $\pi$, the detailed balance conditions require only that

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{for all states } i, j \in S \tag{9}$$

There are potentially an infinite number of transition matrices $P$ satisfying conditions (9). Let us focus on solutions of the form

$$p_{ij} = q_{ij}\alpha_{ij},$$

where $Q = (q_{ij})$ is an irreducible *candidate-generation matrix*, and $\alpha_{ij} \in (0,1]$ are the *acceptance probabilities* for given tentative state transitions.

W. Hastings (1970) proposed the following general class of acceptance probability matrices guaranteeing the validity of the detailed balance conditions (9):

$$\alpha_{ij} = \frac{s_{ij}}{1+t_{ij}},$$

where

$$t_{ij} = \frac{\pi_i q_{ij}}{\pi_j q_{ji}}.$$

and $s_{ij} = s_{ji}$ are numbers chosen so that $\alpha_{ij} \in (0,1]$, i.e.

$$0 < s_{ij} \leq 1 + \min\{t_{ij}, t_{ji}\} \quad \forall\ i, j. \tag{10}$$

Enforcing equality in condition (10) results in the Metropolis-Hastings algorithm

$$\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

(check this!), whereas always choosing $s_{ij} = 1$ defines the so called *Barker's algorithm*:

$$\alpha_{ij} = \frac{\pi_j q_{ij}}{\pi_j q_{ji} + \pi_i q_{ij}}.$$

Let us then compare the various Hastings-type MCMC algorithms with respect to their asymptotic variance (Theorem 10.5). We quote the following result without proving it:

**Theorem 10.6**
*Let $P = (p_{ij})$ and $P' = (p'_{ij})$ be regular transition matrices over finite state space S, with the same stationary distribution $\pi$. If $p_{ij} \geq p'_{ij}$ for all $i \neq j$, then*

$$v(f,P,\pi) \leq v(f,P',\pi)$$

*holds for all functions $f : S \to \mathbb{R}$.*

*Proof:* E.g. Brémaud page 300. □

**Corollary 10.7**
*For a given candidate-generation matrix Q, the Metropolis-Hastings algorithm has optimal asymptotic variance in the class of Hastings algorithms.*

*Proof:* Since the $\alpha_{ij}$ are probabilities, the upper bound on $s_{ij}$ given in condition (10) cannot be exceeded. The Metropolis-Hastings algorithm matches the upper bound. □

# 11 Genetic Algorithms

Genetic algorithms (GA) are a general purpose “black-box” optimization method (cf. simulated annealing) proposed by J. Holland (1975) and K. DeJong (1975).

The subject has attracted lots of interest recently, but the theory is still incomplete and the empirical results inconclusive. Being general-purpose, parallelizable (?) and incrementally adaptive to changing cost functions (“on-line optimization”) are advantages of genetic algorithms. However, they are typically very slow. (Not competitive for serial optimization of a stable cost function?)

### The Basic Algorithm

We consider the so called “simple genetic algorithm”; also many other variations exist.

Assume we wish to maximize a cost function *c* defined on *n*-bit binary strings:

$$c : \{0,1\}^n \to \mathbb{R}$$

Other types of domains must be encoded into binary strings, which is a nontrivial problem.