Markov Chain Concepts Related to Estimation Algorithms

Thet Paing Soe[1], Ko Ko Oo[2]

Abstract

The purpose of this paper is to give an introduction to some of the theoretical ideas from Markov chain theory. We begin with a brief review of basic concepts and rates of convergence for Markov chain. An account of estimation from ergodic averages follows.

Keywords: Markov chain, Rate of Convergence, Estimation, Batch Means, Window Estimator

Markov Chain

A Markov chain X is a discrete time stochastic process $\{X_0, X_1, \dots \}$ with the property that the distribution of X_t given all previous values of the process, $X_0, X_1, \dots, X_{t-1}$ only depends upon X_{t-1}. Mathematically, we write

$$P\{X_t \in A \mid X_0, X_1, \dots, X_{t-1}\} = P\{X_t \in A \mid X_{t-1}\}$$

for any set A, where $P\{\circ \mid \circ\}$ denotes a conditional probability.

Typically (*but not always*) for Markov chain, the Markov chain takes values in R^n (n-dimensional Euclidean space). However, to illustrate the main ideas, for most of this paper, we shall restrict attention to discrete state-spaces.

Extension to general state-spaces are more technical, but do not require any major new concepts. Therefore, we consider transition probabilities of the form $P_{ij}(t) = P\{X_t = j \mid X_0 = i\}$.

For the distribution of X_t to converge to a stationary distribution, the chain needs to satisfy three important properties. First, it has to be **irreducible**. *That is*, from all starting points, the Markov chain can reach any non-empty set with positive probability, in some number of iterations. This is essentially a probabilistic connectedness condition. Second, the chain needs to be **aperiodic**. This stops the Markov chain from oscillating between different sets of states in a regular periodic movement. Finally, and most importantly, the chain must be positive recurrent. This can be expressed in terms of the existence of a stationary distribution $\pi(\circ)$. There are also various equivalent definitions. These ideas are made precise in the following definition.

Let τ_{ii} be ^{i}the time of the first return to state i, ($\tau_{ii} = \mathit{min}\{t > 0 : X_t = i \mid X_0 = i\}$).

Definition

(i) X is called **irreducible** if for all i, j, there exists a t > 0 such that $P_{ij}(t) > 0$.

(ii) An irreducible chain X is **recurrent** of $P\{\tau_{ii} < \infty\} = 1$ for some (*and hence for all*) i. Otherwise, X is **transient**. Another equivalent condition for recurrence is

$$\sum_{i} P_{ij}(t) = \infty \text{ for all i, j.}$$

[1] Tutor, Department of Mathematics, Hinthada University
[2] Lecturer, Department of Mathematics, Hinthada University

(iii) An irreducible recurrent chain X is called **positive recurrent** if $E[\tau_{ii}] < \infty$ for some (*and hence for all*) i. Otherwise, it is called **null-recurrent**. Another equivalent condition for positive recurrence is the existence of a stationary probability distribution for X, that is there exists $\pi(\circ)$ such that

$$\sum_{i} \pi(i)\, P_{ij}(t) = \pi(j). \qquad (1.1)$$

for all j and $t \geq 0$.

(iv) An irreducible chain X is called **aperiodic** if for some (*and hence for all*) i,

greatest common divisor $\{t > 0 : P_{ii}(t) > 0\} = 1$.

In Markov chain, we already have a target distribution $\pi(\circ)$, so that by (iii) above, X will be positive recurrent if we can demonstrate irreducibility.

In practice, output from Markov chain is summarized in terms of **ergodic average** of the form

$$\overline{f}_N = \tfrac{1}{N}\sum_{i=1}^{N} f(X_t)$$

where $f(\circ)$ is a real valued function. Therefore asymptotic properties of $\overline{f}_N$ are very important.

Theorem If X is positive recurrent and aperiodic then its stationary distribution $\pi(\circ)$ is the unique probability distribution satisfying (1.1). We then sat that X is ergodic and the following consequences hold :

(i) $P_{ij}(t) \rightarrow \pi(j)$ as $t \rightarrow \infty$ for all i, j.

(ii) If $E_\pi[\ |f(X)|\ ] < \infty$, then $P\{\overline{f}_N \rightarrow E_\pi[f(X)]\} = 1$,

where $E_\pi[f(X)] = f(i)\ \pi(i)$, the expectation of f(X) with respect to $\pi(\circ)$.

Part (ii) of this Theorem is clearly very important in practice for Markov chain, although it does not offer any reassurance as to show how long we need to run the Markov chain before its iterations are distributed approximately according to $\pi(\circ)$, and it offers no estimate as to the size of the error of any estimate $\overline{f}_N$.

Most of the Markov chains produced in Markov chain Monte Carlo are reversible, are derived from reversible components, or have reversible versions. A Markov chain is said to be reversible if it is positive recurrent with stationary distribution $\pi(\circ)$, and

$$\pi(i)\, P_{ij} = \pi(j)\, P_{ji.}$$

We shall assume that the Markov chains we consider are reversible unless otherwise stated.

Rate of Convergence

We say that X is geometrically ergodic (*in total variation norm*), if it is ergodic (*positive recurrent and aperiodic*) and there exists $0 \le \lambda < 1$ and a function $V(\circ) > 1$ such that

$$\sum_{j} |P_{ij}(t) - \pi(j)| \le V(i)\, \lambda^{t} \qquad (1.2)$$

for all i. The smallest λ for which there exists a function V satisfying (1.2) is called the **rate of convergence**. We shall denote this by λ^*. (*Formally, we define* λ^* *as inf*{λ : ∃V such that (1.2) holds}.)

Sufficiently regular problems have transition probabilities described by a sequence of eigenvalues $(\lambda_0, \lambda_1, \ldots\}$, where $\lambda_0 = 1$, and corresponding left eigenvalues $\{e_0, e_1, \ldots\}$, that is

$$\sum_{i} e_k(i)\, P_{ij}(t) = \lambda_k\, e_k(j)$$

for all j and for each k, such that

$$P_{ij}(t) = \sum_{k} e_k(i)\, e_k(j)\, \lambda_k^{t}. \qquad (1.3)$$

Here $e_0(\circ) = \pi(\circ)$. In general, the eigenvalues can be complex with modulus bounded by unity. Reversibility of the Markov chain ensures that the eigenvalues and eigenvectors are real. In general, for infinite state-spaces, there are an infinite number of eigenvalues. However, for geometrically ergodic chains, all but the principal eigenvalues, $\lambda_0 = 1$, are uniformly bounded away from ±1. Chains which fail to be geometrically ergodic have an infinite number of eigenvalues in any open interval containing wither –1 or 1.

For large t, the dominant term in (1.3) is $\pi(j) = e_0(j)$. However, the speed at which convergence is achieved depends on the second largest eigenvalue in absolute value, which is just the rate of convergence of the Markov chain.

Therefore, an alternative definition of λ^* is

$$\lambda^* = |\lambda_k|.$$

In practice, it is usually too difficult to obtain useful upper bounds on λ^*.

Estimation

One of the most important consequences of geometric convergence is that it allows the existence of central limit theorems for ergodic averages, *that is*, results of the form

$$\sqrt{N}\,(\bar{f}_N - E_\pi[f(X)]) \to N(0, \sigma^2) \qquad (1.4)$$

for some positive constant σ, as $N \to \infty$, where the convergence is in distribution. Such results are essential in order to put inference from Markov chain output on a sound footing. Even when central limit theorems do exist, algorithms can often be extremely inefficient in case where σ is large in comparison with the variance (*under* π) of f(X).

An extensive treatment of geometric convergence and central limit theorems for Markov chains can be found in Meyn and Tweedie (1993), applications to Metropolis-Hastings algorithms appear in Roberts and Tweedie (1994), and to the Gibbs sampler in Chan

(1993), Schervish and Carlin (1992) and Roberts and Polson (1994). Tierney (1995: this volume) discusses these results further, and Geyer (1992) provides a review of the use of central limit theorems for Markov chain.

We will assume that (1.4) holds for the function f(∘). The following result gives equivalent expressions for σ^2.

Theorem
$$\sigma^2 = \mathrm{Var}_\pi[f(X_0)] + 2\sum_{i=1}^{\infty} \mathrm{Cov}(X_0, X_1) \qquad (1.5)$$
$$= \sum_{j=1}^{\infty} \frac{1+\lambda_j}{1-\lambda_j} a_i \mathrm{Var}_\pi[f(X_0)]$$
$$\leq \frac{1+\lambda^*}{1-\lambda^*}$$

for some nonnegative constant a_i, where X_0 is distributed according to π(∘), and

$$\sum_{i=0}^{\infty} a_i = 1.$$

The ratio

$$eff_{\bar{f}} = \frac{1}{\sigma^2} \mathrm{Var}_\pi[f(X_0)]$$

is measure of efficiency of the Markov chain for estimating $E_\pi[f(X)]$.

To assess the accuracy of our estimate of $E_\pi[f(X)]$, $\bar{f}_N$, it is essential to be able to estimate σ^2. This problem is reviewed extensively in Geyer (1992). We content ourselves here with a short description of two of the simplest and most commonly used methods.

Batch Means

Run the Markov chain for N = mn iterations, where n is assumed sufficiently large that

$$Y_k = \frac{1}{n} \sum_{i=(k-1)n+1}^{kn} f(X_i) \qquad (1.6)$$

are approximately independently $N(E_\pi[f(X)], \sigma^2)$.

Therefore σ^2 can be approximated by

$$\frac{n}{m-1} \sum_{k=1}^{m} (Y_k - \bar{f}_N)^2,$$

or alternatively a t-test can be used to give bounds on the accuracy of $\bar{f}_N$.

Window Estimator

From (1.5), an obvious way to try to approximate σ^2 is to estimate

$\tau_i \cong \mathrm{Cov}_\pi[f(X_0), f(X_i)]$ by the empirical covariance function

$$\tau_i = \frac{1}{n} \sum_{j=1}^{N-i} [f(X_j) - \overline{f}_N]\,[f(X_{j+1}) - \overline{f}_N] \qquad (1.7)$$

Unfortunately, this approach runs into trouble since the estimates become progressively worse as i increases (*there are less and less terms in the average* (1.7)).

In fact, the estimator produced by this approach is not even consistent. Instead, it is necessary to use a truncated window estimator of σ^2,

$$\delta_N^2 = \tau_0 + 2 \sum_{i=1}^{\infty} \omega_N(i)\, \tau_i, \qquad (1.8)$$

where $0 \le \omega_N(i) \le 1$. Typically, the $\omega_N(i)$ are chosen to be unity within a certain range (*which can depend upon N*), and zero outside this range.

Acknowledgments

I would like to express my sincere gratitude to Dr. Tin Htwe, Rector, Hinthada University, and Dr. Theingi Shwe, Pro-Rector, Hinthada University, for their kind permission to submit this paper. Special thanks are also due to Dr. Khin Su Su Shein, Associate professor and Head of the Department of Mathematics, Hinthada University, for her encouragement and comments.

References

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. *Markov Chain Monte Carlo in Practice*, First Edition, Chapman & Hall, UK.

Kandasamy, P., Thilagavathi, K., and Gunavathi, K., 2010. *Probability, Random Variables & Random Processes*, Reprint, S-Chand & Company Limited, Delhi.

Natarajan, A.M. and Tamilarsi, A., 2007. *Probability Random Processes and Queuing Theory*, Reprint, New Age International Publishing Company, New Delhi.