Math 577 Monte Carlo Methods – Spring 2019
Comments on Lectures 10–13

Kevin K. Lin*

February 8–25, 2019

Main topics:

- Ergodic theorem
- Initialization bias
- Estimator variance and the Kubo formula

References: for basic Markov chain results, I like [HPS]. For the error analysis, I am largely following §3 of [S].

Setting and problem. We now step away from any specific MCMC algorithm and look at Markov chains in general. The setting is a Markov chain X_n with state space S, transition matrix P, and initial distribution π^0. We assume that X_n is irreducible and aperiodic. For simplicity, we assume $|S| < \infty$. Let π^* be the (unique) stationary distribution. For most of these notes, we do not need to assume detailed balance, so the results and methods we discuss may cover algorithms other than Metropolis-Hastings.

The basic problem we consider is: suppose we have a function $\varphi : S \to \mathbb{R}$ and we wish to compute the expectation

$$E_{\pi^*}(\varphi) \;=\; \sum_{x\in S} \varphi(x)\pi^*(x) \;=\; \pi^* \cdot \varphi \tag{1}$$

with

$$\pi^* = \begin{pmatrix} \pi^*_1 & \cdots & \pi^*_{|S|} \end{pmatrix} \qquad \text{and} \qquad \varphi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_{|S|} \end{pmatrix}. \tag{2}$$

We use the standard estimator

$$\widehat{\varphi}_N = \frac{1}{N}\sum_{n=1}^{N} \varphi(X_n). \tag{3}$$

*E-mail: klin@math.arizona.edu.

The *ergodic theorem* asserts that $\lim_{N\to\infty} \widehat{\varphi}_N = \pi^* \cdot \varphi$ as $N \to \infty$ with probability 1; this can be viewed as a law of large numbers for Markov chains; see Ch. 2 of [HPS]. This fundamental limit theorem serves as the basis for all MCMC algorithms. However, in practice we can never actually take $N = \infty$, and for finite N the estimator $\widehat{\varphi}_N$ is a random variable. These lectures are concerned with characterizing the errors resulting from finite N.

There are two main sources of errors:

1) *Initialization bias.* Since $\pi^0 \neq \pi^*$ in general, this will affect the accuracy of the estimate $\widehat{\varphi}_N$.

2) *Estimator variance.* Even if $\pi^0 = \pi^*$, the estimator $\widehat{\varphi}_N$ still has a positive variance for finite N.

The standard way to deal with initialization bias is to throw away the first N_{init} samples, or "burn in" the chain. The question is: how do we know we have thrown away enough samples? As for estimator variance, this would be easy to assess if the samples were independent (as we did for simple Monte Carlo integration). But in MCMC, the samples are correlated. As you might expect, this can mean the variance of $\widehat{\varphi}_N$ is larger than simple sample variance might suggest. We thus need to a reliable method for estimating the variance of $\widehat{\varphi}_N$.

This note lays out the relevant theory; the next one will discuss practical issues of how to deal with data.

Initialization bias and exponential autocorrelation time. A natural way to measure initialization bias is to compute $E(\widehat{\varphi}_N)$: *if* $\pi^0 = \pi^*$, then we have

$$\begin{aligned} E(\widehat{\varphi}_N) &= \frac{1}{N}\sum_{n=1}^{N} E(\varphi(X_n)) && \text{(4a)} \\ &= E_{\pi^*}(\varphi) && \text{(4b)} \end{aligned}$$

since the X_n are identically distributed if $X_0 \sim \pi^*$. When $\pi^0 \neq \pi^*$, we *generally* expect $E(\widehat{\varphi}_N) \neq 0$ for finite N (though this may not be the case for special choices of φ).

A key property of a Markov chain that determines how fast initialization bias decays is the *exponential autocorrelation time* τ_{exp}. First, observe that $E(\varphi(X_n)) = \pi^0 \cdot P^n \cdot \varphi$. Recall that we for $P^\perp = P - \mathbb{1} \cdot \pi^*$, we have

$$P^n = \mathbb{1} \cdot \pi^* + (P^\perp)^n. \tag{5}$$

Since for an irreducible aperiodic chain, all eigenvalues λ of $P^\perp$ satisfy $|\lambda| < 1$, we have

$$\begin{aligned} E(\varphi(X_n)) &= \pi^0 \cdot P^n \cdot \varphi && \text{(6a)} \\ &= \pi^* \cdot \varphi + \pi^0 \cdot (P^\perp)^n \cdot \varphi && \text{(6b)} \\ &= \pi^* \cdot \varphi + O(\rho^n) && \text{(6c)} \end{aligned}$$

where $\rho = \max\left\{|\lambda| \;\middle|\; \lambda \in \operatorname{spec}(P^\perp)\right\}$ is the *spectral radius* of $P^\perp$; ρ is also the size of the second largest eigenvalue of P.

From this, we see that the initialization bias decays geometrically, at a rate governed by ρ. If we introduce the exponential autocorrelation time

$$\tau_{\text{exp}} = -\frac{1}{\log \rho}, \tag{7}$$

then the above suggests

$$E(\varphi(X_n)) = E_{\pi^*}(\varphi) + O(e^{-n/\tau_{\text{exp}}}). \tag{8}$$

We do need to be a little more careful: what the above equation really means is

$$\left| E(\varphi(X_n)) - E_{\pi^*}(\varphi) \right| \leqslant K e^{-n/\tau_{\text{exp}}} \tag{9}$$

for some constant $K > 0$ depending on both φ and π^0. In general, K may be quite large, because there may be large transients in the LHS before exponential decay sets in. However, for P that satisfy detailed balance with respect to π^*, we can expand φ in terms of an orthonormal basis[1] of eigenectors η_k of $P^\perp$, so that $\varphi = \sum_k a_k \eta_k$. From this, we get

$$\begin{aligned}
\left| E(\varphi(X_n)) - E_{\pi^*}(\varphi) \right| &= \left| \pi^0 \cdot (P^\perp)^n \cdot \varphi \right| && \text{(10a)} \\
&= \left| \sum_k a_k \lambda_k^n \pi^0 \cdot \eta_k \right| && \text{(10b)} \\
&\leqslant \sqrt{\sum_k |a_k|^2 \, \lambda_k^{2n}} \cdot \sqrt{\sum_k |\pi^0 \cdot \eta_k|^2} \qquad \text{(Caucy-Schwartz)} && \text{(10c)} \\
&\leqslant \rho^n \sqrt{\sum_k |a_k|^2} \cdot \sqrt{\sum_k |\pi^0 \cdot \eta_k|^2} \qquad (\text{since } |\lambda_k| \leqslant \rho) && \text{(10d)} \\
&= \rho^n \sqrt{\operatorname{var}_{\pi^*}(\varphi)} \cdot \sqrt{\operatorname{var}_{\pi^*}(\pi^0/\pi^*)} \qquad \text{(see PS2).} && \text{(10e)}
\end{aligned}$$

In the above, $\|v\|_\pi := \sqrt{\sum_x v(x)^2 \pi(x)}$ is a weighted vector norm, and $\|A\|_\pi = \max\{\|Av\| \mid \|v\|_\pi = 1\}$ is the associated matrix norm. Defining

$$\begin{aligned}
K_0 &= \sqrt{\operatorname{var}_{\pi^*}(\pi^0/\pi^*)} && \text{(11a)} \\
&= \sqrt{\sum_{x \in S} \left| \frac{\pi^0(x)}{\pi^*(x)} - 1 \right|^2 \pi^*(x)}, && \text{(11b)}
\end{aligned}$$

we obtain

$$\left| E(\varphi(X_n)) - E_{\pi^*}(\varphi) \right| \leqslant K_0 \rho^n \sqrt{\operatorname{var}_{\pi^*}(\varphi)}. \tag{12}$$

The constant K_0 depends only on π^0, and measures how far π^0 and π^* are from each other. If $\pi^0 = \pi^*$, then $K_0 = 0$ as one would expect (since there is no initialization bias in this case).

Important note. In class I stated this with $K_0 = 1$. That was a mistake!

[1] Orthonormal with respect to the inner product $\langle u, v \rangle = \sum_x u(x)v(x)\pi^*(x)$. See PS2 Problem 4.

Autocorrelation functions. The reason τ_{exp} is called the exponential autocorrelation time is that it is connected with *autocorrelation functions*[2], defined by

$$C_\varphi(m,n) = E\big((\varphi_m - \mu_m)\cdot(\varphi_n - \mu_n)\big) \tag{13}$$

where $\varphi_n = \varphi(X_n)$ and $\mu_n = E(\varphi_n)$, i.e., it is just the covariance of φ_m and φ_n. When $\pi^0 = \pi^*$, we have $C_\varphi(m,n) = C_\varphi(m-n,0)$ for all m,n, and we just have

$$C_\varphi(n) = E\big((\varphi_n - \mu)\cdot(\varphi_0 - \mu)\big) \tag{14}$$

where $\mu = E_{\pi^*}(\varphi)$.

Notice $C_\varphi(0) = \mathrm{var}_{\pi^*}(\varphi)$. By Cauchy-Schwartz, we have

$$|C_\varphi(m,n)| \leqslant \sqrt{\mathrm{var}(\varphi_m)}\cdot\sqrt{\mathrm{var}(\varphi_n)} \tag{15}$$

and, in the stationary case,

$$|C_\varphi(n)| \leqslant \mathrm{var}_{\pi^*}(\varphi). \tag{16}$$

So we introduce the normalized autocorrelation function $\overline{C}_\varphi(n) = C_\varphi(n)/C_\varphi(0)$ in the stationary case[3].

Observe that

$$C_\varphi(n) = \pi^*\cdot M_\varphi\cdot(P^\perp)^n\cdot\varphi \tag{17}$$

where M_φ is the multiplication operator

$$M_\varphi = \begin{pmatrix} \varphi_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \varphi_{|S|} \end{pmatrix}. \tag{18}$$

Thus, we expect $|C_\varphi(n)| = O(e^{-n/\tau_{\mathrm{exp}}})$, i.e., for an irreducible aperiodic finite-state Markov chain, autocorrelation functions decay exponentially, at the same rate as the initialization bias. But, since the autocorrelation function is defined *in stationarity,* it can be estimated from a single long simulation, rather than many repeated simulations (as is the case for the initialization bias $E(\widehat{\varphi}_N)$).

Estimator variance and Kubo formula. Let us now consider estimator variance, i.e., assuming $\pi^0 = \pi^*$, what is the variance of $\widehat{\varphi}_N$? This is given by the *Kubo formula*

$$\mathrm{var}(\widehat{\varphi}_N) = \frac{\mathrm{var}_{\pi^*}(\varphi)\cdot\tau_{\mathrm{int}}}{N} + O(1/N^2), \tag{19}$$

[2] These are more accurately called autocovariance functions, and are often referred to as such in the statistics literature.

[3] We will not need normalized autocorrelation functions for nonstationary chains.

where

$$\mathrm{var}_{\pi^*}(\varphi) = \sum_{x \in S} (\varphi(x) - E_{\pi^*}(\varphi))^2 \pi^*(x) \tag{20}$$

is the variance of φ with respect to π^*, and τ_{int} is the *integrated autocorrelation time*

$$\tau_{\mathrm{int}} = \sum_{n=-\infty}^{\infty} \overline{C}(n). \tag{21}$$

Notice if the samples X_n were IID, then $C_\varphi(n) = 0$ for all $n \neq 0$ and $\tau_{\mathrm{int}} = 1$. The integrated autocorrelation time measures, in a sense, the number of steps one must take to obtain 1 effectively new sample. Comparing the Kubo formula with the variance of the simple Monte Carlo integrator suggests that we think of the quantity N/τ_{int} as the effective number of samples.

Eq. (19) is fairly straightforward to prove. To simplify notation, let us assume $E_{\pi^*}(\varphi) = 0$. Then

$$\mathrm{var}(\widehat{\varphi}_N) = \mathrm{var}\Big(\frac{1}{N}\sum_{n=1}^{N} \varphi_n\Big) \tag{22a}$$

$$= E\left[\Big(\frac{1}{N}\sum_{n=1}^{N} \varphi_n\Big)^2\right] \tag{22b}$$

$$= \frac{1}{N^2} \sum_{m,n=1}^{N} E\big(\varphi_m \varphi_n\big) \tag{22c}$$

$$= \frac{1}{N^2} \sum_{m,n=1}^{N} C_\varphi(m-n). \tag{22d}$$

Since the summand is a function of $k = m - n$, it makes sense to rewrite the sum in terms of k. A little work then gives

$$\mathrm{var}(\widehat{\varphi}_N) = \frac{1}{N^2} \sum_{k=-(N-1)}^{N-1} (N - |k|) \cdot C_\varphi(k) \tag{23a}$$

$$= \underbrace{\frac{1}{N} \sum_{k=-(N-1)}^{N-1} C_\varphi(k)}_{(I)} - \underbrace{\frac{1}{N^2} \sum_{k=-(N-1)}^{N-1} |k| \cdot C_\varphi(k)}_{(II)}. \tag{23b}$$

Recall that $C_\varphi(n) = O(\rho^n)$ where $0 \leqslant \varphi < 1$ is the spectral radius of $P^\perp$. So, we can write Term (I) as

$$\sum_{k=-(N-1)}^{N-1} C_\varphi(k) = \sum_{k=-\infty}^{\infty} C_\varphi(k) + O(\rho^N). \tag{24}$$

Similarly, we have

$$\sum_{k=-(N-1)}^{N-1} |k| \cdot |C_\varphi(k)| < \infty \tag{25}$$

so Term (II) is $O(1/N^2)$. Putting all this together gives

$$\mathrm{var}(\widehat{\varphi}_N) = \frac{\sum_n C_\varphi(n)}{N} + O(1/N^2). \tag{26}$$

The Kubo formula follows.

What happens in infinite dimensions? The main difference is that even when the chain is irreducible and aperiodic, and possesses a (unique) stationary distribution, correlation functions need not decay exponentially. Indeed, because the spectrum of the transition operator P is now typically infinite, the spectral radius of P may be 1!

It is useful to consider some special cases. For example, often correlation functions exhibit power law decay, i.e., $C_\varphi(n) = O(|n|^{-\beta})$ for some $\beta > 0$. In this case, we can expect $\sum_{|n|\geqslant N} C_\varphi(n) = O(N^{-\beta+1})$. And if $\beta > 2$, then the sum $\sum_n |n| C_\varphi(n)$ converges. Combining these two facts, we see that the Kubo formula will continue to hold so long as $\beta > 2$. But if $\beta \leqslant 2$, then the whole thing falls apart, and a different way of characterizing estimator accuracy is needed.