Convergence of the Metropolis-Hastings Algorithm

Peter Racioppo
UID: 103953689 | pcracioppo@ucla.edu

March 22, 2020

1 Introduction

Markov Chain Monte Carlo algorithms are commonly-used tools for sampling from hard-to-sample probability distributions. Given a target distribution $\pi(x)$, and a starting state X_0, we form a Markov chain of approximate samples $X_0, X_1, ..., X_T$ from $\pi(x)$. In practice, we cannot draw X_0 from the target distribution, and it may be randomly set. However, under mild constraints, the distribution of the Markov chain will approach the target distribution as the number of samples goes to infty. It may be difficult to determine how quickly the Markov chain will converge. Typically, a burn in length *B* is set such that the first *B* samples from the Markov chain are discarded. If we wish, for example, to compute the integral $I = E_\pi[h(x)] = \int_X h(x)\pi(x)$ we can estimate *I* by $I \approx \frac{1}{N}\sum_{i=1}^{N} h(X_{B+i})$ [1].

2 Background

A Markov chain is a stochastic process that "remembers" only the most recent step. Formally, a Markov chain on a state space Ω is a stochastic process $\{X_0, X_1, ..., X_T\}$, with every $X_i \in \Omega$, such that $Pr(X_{t+1} = y|X_t = x, X_{t-1} = x_{t-1}, ..., X_0 = x_0) = Pr(X_{t+1} = y|X_t = x) =: P(x, y)$. We may define *P* as the matrix with (x, y)th entry $P(x, y)$. Then, *P* is a stochastic matrix, meaning that $P(x, y) \geq 0$ and $\sum_{y\in\Omega} P(x, y) = 1, \forall x, y \in \Omega$.

Definition (Reducibility): A Markov chain *P* is said to be reducible if there is a nonzero probability of moving from any state to any other state, that is, if $\forall x, y, \exists t$ s.t. $P_t(x, y) > 0$.

Definition (Periodicity): The period of state *i* is defined as $\gcd\{n > 0 : Pr(X_n = i|X_0 = i\}$. A state with period 1 is called aperiodic. A Markov chain with only aperiodic states is called aperiodic.

Definition (Stationarity): A vector π is called a stationary distribution for a stochastic matrix *P* if $\pi = \pi P$.

Definition (Reversibility): A Markov chain is called reversible if it satisfies the detailed balance condition $\pi(x)P(x, y) = \pi(y)P(y, x)$.

Metropolis-Hastings: The Metropolis-Hastings Algorithm considers base chains $P(x, y)$ and $P(y, x)$. The detailed balance is maintained at stationarity by defining an acceptance probability $A(x, y)$ that limits the number of transitions between states *x* and *y* to $\min(\pi(x)P(x, y), \pi(y)P(y, x))$. Then, a proportion of $\frac{\min(\pi(x)P(x,y),\pi(y)P(y,x))}{\pi(x)P(x,y)}$ $= \min(1, \frac{\pi(y)P(y,x)}{\pi(x)P(x,y)}) := A(x, y)$ are allowed to move from *x* to *y*. The transition probability is $M(x, y) = P(x, y)A(x, y) = P(x, y)\min(1, \frac{\pi(y)P(y,x)}{\pi(x)P(x,y)})$ for $x \neq y$ and $M(x, x) = 1 - \sum_{y\neq x} M(x, y)$. The Metropolis Algorithm is the case when $P(x, y) = P(y, x)$ [1].

3 Convergence of Markov Chains

Proposition 1: If a Markov chain P is reversible with respect to π, then π is a stationary distribution for P.
Proof: $(\pi P)(y) = \sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y)\sum_x P(y,x) = \pi(y)$.

Proposition 2: Any reversible Markov chain P is similar to a symmetric stochastic matrix and hence has only real eigenvalues.
Proof: Left and right multiplying $\pi P = \pi$ by diag$(1/\sqrt{\pi(x)})$, we have that diag$(\sqrt{\pi(x)})P$ diag$((\sqrt{\pi(x)})^{-1}) = I$.

Theorem 1 (Fundamental Theorem of Markov Chains): Any irreducible, aperiodic Markov chain P has a unique stationary distribution π, given by the normalized left eigenvector of P corresponding to eigenvalue 1, and $P^t(x,y) \to \pi(y)$ as $t \to \infty$, $\forall x, y \in \Omega$ [2].
Proof (reversible case): By Prop. 2, P has strictly real eigenvalues. By the Perron Frobenius Theorem, any irreducible, aperiodic, stochastic matrix P has an eigenvalue $\lambda_0 = 1$ with a unique left eigenvector $e_0 > 0$, and all other eigenvalues of P have magnitude less than 1. By reversibility w.r.t π and Prop. 1, $e_0 = \pi$ is the unique stationary distribution. A standard power method analysis then shows that $P^t(x,y) \to \pi(y)$ up to a constant. Since both the right hand side and π must integrate to 1, this constant is 1. **Note:** This theorem also holds for irreversible Markov chains, but the proof is more involved [2].

Proposition 3: Let P be an irreducible, stochastic matrix. For any $0 < \alpha < 1$, the matrix $P' = \alpha P + (1-\alpha)I$ is stochastic, irreducible, and aperiodic, and has the same stationary distribution as P. In practice, the periodicty of a Markov chain can be ignored by instead simulating P', which slows down the dynamics by a factor of $1/\alpha$ [2].

4 Convergence of Metropolis Hastings

An aperiodic, irreducible Markov chain has a unique stationary distribution d (independent of the starting state) by the results of the previous section. Random walks on proper probability distributions are aperiodic, and the extra requirement of irreducibility imposes only mild constraints on proposal densities (it suffices for the proposal distribution to be positive everywhere). It remains only to show that the stationary distribution d of the MCMC Markov chain is the target distribution π.

We will now show that π is stationary with respect to P. Consider the probability of moving between states x_a and x_b over steps i and $i+1$, that is $P(x_i = x_a, x_{i+1} = x_b)$ and suppose, without loss of generality, that $\pi(x_b)P(x_a,x_b) \geq \pi(x_a)P(x_b,x_a)$. We have that $P(x_i = x_a, x_{i+1} = x_b) = \pi(x_a)M(x_a,x_b) = \pi(x_a)P(x_a,x_b)\text{min}(1, \frac{\pi(x_b)P(x_a,x_b)}{\pi(x_a)P(x_b,x_a)}) = \pi(x_a)P(x_a,x_b)$. On the other hand, the probability of moving between states x_b and x_a is $P(x_i = x_b, x_{i+1} = x_a) = \pi(x_b)P(x_b,x_a)\text{min}(1, \frac{\pi(x_a)P(x_b,x_a)}{\pi(x_b)P(x_a,x_b)}) = \pi(x_b)P(x_b,x_a)(1, \frac{\pi(x_a)P(x_b,x_a)}{\pi(x_b)P(x_a,x_b)}) = \pi(x_a)P(x_a,x_b)$.

Thus, the joint density of x_i and x_{i-1} is symmetric, so π satisfies the detailed balance condition, and hence, by Prop. 1, is a stationary distribution for P. Since P's stationary distribution d is unique, we have that $d = \pi$, or in other words, π is the unique stationary distribution of P. It follows from Theorem 1 that $P^t(x, y) \to \pi(y)$ as $t \to \infty$, $\forall x, y \in \Omega$ [3].

5 Data Augmentation (Tanner & Wong)

Tanner and Wong's 1987 paper *The calculation of posterior distributions by data augmentation* proposed a two-step iterative algorithm for the computation of posterior probability distributions. The posterior density can be written as $p(\theta|y) = \int_Z p(\theta|z, y)p(z|y)dz$, where y is the data and z is "latent data" which is added to y to form "augmented data." The authors propose to solve for $p(\theta|y)$ by iteratively solving the integral equation
$g(\theta) = \int K(\theta, \phi)g(\phi)d\phi := Tg(\theta)$. The posterior $p(\theta|y)$ is a solution of this equation, since $p(z|y) = \int_\Theta p(z|\phi, y)p(\phi|y)d\phi$ and therefore $p(\theta|y) = \int K(\theta, \phi)p(\phi|y)d\phi$, where
$K(\theta, \phi) = \int p(\theta|z, y)p(z|\phi, y)$. It follows that the Markov chain $g_{i+1} = Tg_i$ converges to $p(\theta|y)$. In Sec. 6 of the paper, the authors prove convergence results for their two-part algorithm. Under mild conditions on K and g, if f has nonzero positive and negative parts, $||Tf|| < ||f||$ (Lemma 2), and therefore $||g_i - g*||$ is strictly decreasing, where $g* := p(\theta|y)$. The authors also show explicitly that g^* is the only fixed point of the operator T (Theorem 2) and the Markov chain $g_{i+1} = Tg_i$ converges to g^* linearly (Theorem 3) [4].

6 Conclusion

We have seen that for sufficiently large b, the bth MCMC Markov chain sample X_b is arbitrarily close to the target distribution, independently of the selection of the first state X_0. In practice, how quickly the algorithm converges is highly dependent on the proposal distribution P. Rules of thumb exist for tuning the "acceptance rate" and "effective sample size." More sophisticated calibration techniques, including "diminishing adaptation" methods, can improve the rate of convergence of these algorithms [5].

References

[1] Y. Wu. *A Note on Monte Carlo Methods. UCLA Statistics*. STATS 102C, 2017.

[2] A. Sinclair. *CS294 Markov Chain Monte Carlo: Foundations & Applications*.
2009. (*https://people.eecs.berkeley.edu/ sinclair/cs294/n2.pdf*).

[3] Gelman et al. *Bayesian Data Analysis*. (pg. 290-291). Chapman & Hall, 2004.

[4] M. Tanner, W. Wong. *The calculation of posterior distributions by data augmentation*. Journal of the American statistical Association, 82(398):528–540, 1987.

[5] C.P. Robert. *The Metropolis–Hastings algorithm*. Universite Paris-Dauphine, University of Warwick, and CREST. 2016. (*https://arxiv.org/pdf/1504.01896.pdf*)