Stationarity and Convergence of the Metropolis–Hastings Algorithm

STACY D. HILL and
JAMES C. SPALL

INSIGHTS INTO THEORETICAL ASPECTS

PHOTO COURTESY OF JOHNS HOPKINS UNIVERSITY ENGINEERING FOR PROFESSIONALS

Markov chain Monte Carlo (MCMC) is a versatile sampling approach that is useful in a wide range of estimation and simulation applications. Fundamentally, MCMC is a powerful general means of generating random samples from probability distributions from which it is otherwise difficult to draw samples. The MCMC method is named for its reliance on the construction of a Markovian (dependent) sequence of random variables. Under modest conditions, the sequence has a limiting probability distribution that corresponds to the distribution of interest, often called the *target distribution*. The sequence, as will be seen, is easily constructed, and the target distribution may be almost any distribution of interest. These two features of MCMC make it a popular choice for Monte Carlo simulation. The primary forms of MCMC are the Metropolis–Hastings (M-H) algorithm and Gibbs sampling. Although both forms are useful, we focus on M-H, which is more flexible and easier to implement. A general

Digital Object Identifier 10.1109/MCS.2018.2876959
Date of publication: 16 January 2019

1066-033X/19©2019IEEE

discussion of many aspects of MCMC, including several examples, is given in [1]. As discussed in "Summary," the purpose of this article is to provide some of the theoretical support for M-H that was not given in [1], focusing specifically on the stationarity and convergence of the underlying Markov process.

Although the implementation of M-H is easy, developing and understanding the theoretical foundation is not. Furthermore, related results are scattered throughout the mathematics, probability, statistics, and engineering literature, making it challenging to obtain a clear picture of what is known about the theoretical properties of M-H. Nontrivial concepts from measure theory, general Markov processes, and ergodic theory are needed to provide a full justification in the form of convergence of the transition

Summary

Markov chain Monte Carlo (MCMC) is a powerful general means for generating random samples from probability distributions from which it is otherwise difficult to draw samples. MCMC has applications in diverse areas, such as Monte Carlo simulation, Bayesian parameter estimation, system identification, and state estimation in dynamical systems. Although the implementation of MCMC algorithms is often straightforward, developing and understanding their theoretical foundation is not. This article discusses some of the theoretical support underlying the Metropolis–Hastings (M-H) form of MCMC, focusing on the issues of stationarity and convergence. The M-H form is one of the two most popular general forms of MCMC; the other is the related Gibbs sampling. This article synthesizes and interprets several results on the theory of Markov processes scattered throughout the probability, statistics, and engineering literature that have application to MCMC. This article also illustrates the application of M-H to two-system modeling and identification problems.

probabilities for the sequence of generated random samples and convergence in mean square or almost surely (a.s.) for functions of the random samples.

The samples generated in the Markov sequence, which asymptotically represent samples from the target distribution, can be used in diverse areas, such as Monte Carlo simulation [2, Ch. 6], parameter estimation [3], [4], [S1], or state estimation in dynamical systems [1], [5]. The two practical applications in [4] and [S1] of MCMC to parameter estimation are discussed further in "Nonlinear Control-System Identification" and "Pharmacokinetic Model Estimation," respectively. See [6] for another practical application of MCMC in controls that uses MCMC to avoid the well-known degeneracy problem in importance sampling (where excess samples are generated in unimportant regions of the sample space) in the context of a particle filter.

Let $\boldsymbol{X}_k$ be the random vector generated at the kth iteration of M-H. Our discussion focuses on the case in which the target probability distribution is continuous and the state space $\mathcal{X}$ is a subset of $\mathbb{R}^d$, where d is the dimension of $\boldsymbol{X}_k$. The general ideas also apply with discrete probabilities and other more general probability distributions for noncontinuous cases. Some authors reserve the term Markov chain for a process that has a discrete state space. However, following convention in MCMC, we make no distinction here between discrete and nondiscrete state spaces and use the term "Markov chain" whether the state space is discrete or not.

Specifically, we show that the target density $p(\cdot)$ on $\mathcal{X}$ (which is the probability density associated with the target distribution) is a stationary probability density of the Markov chain $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$; that is, if $\boldsymbol{X}_k$ is sampled from

Nonlinear Control-System Identification

The following example illustrates the application of Metropolis–Hastings to nonlinear parametric system identification [4]. The problem considered in [4] is to estimate an unknown parameter $\boldsymbol{\theta}$ in a discrete-time, scalar nonlinear model:

$$y(k) = f((y(k-1),\ldots,y(k-l),u(k-1),\ldots,u(k-l)),\boldsymbol{\theta}) + v(k) \tag{S1}$$

$$= f(\boldsymbol{\Phi}(k),\boldsymbol{\theta}) + v(k), \quad k = 1,\ldots,n, \tag{S2}$$

where $y(k)$ is the system output, $u(k)$ is the system input, $v(k)$ is the measurement noise, and $\boldsymbol{\Phi}(k) = (y(k-1),\ldots,y(k-l), u(k-1),\ldots,u(k-l))$ the regressor, for $k = 1,\ldots,n$. The (real-valued) function $f(\cdot,\cdot)$ is assumed to be known but depends on the unknown parameter vector $\boldsymbol{\theta}$ that lies in a subset S of $\mathbb{R}^q$. The system identification problem is to estimate $\boldsymbol{\theta}$ given the input–output data $\boldsymbol{Z} = \{\{y(k)\}_{k=1}^n, \{u(k)\}_{k=1}^n\}$.

The parameter estimate $\hat{\boldsymbol{\theta}}$ derived in [4] is the minimum variance estimate, which is given by

$$\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\boldsymbol{Z}] = \int_S \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{Z})\, d\boldsymbol{\theta}, \tag{S3}$$

where $p(\boldsymbol{\theta}|\boldsymbol{Z})$, the *posterior density function*, satisfies

$$p(\boldsymbol{\theta}|\boldsymbol{Z}) \propto p(\boldsymbol{Z}|\boldsymbol{\theta}) I_S(\boldsymbol{\theta}), \tag{S4}$$

where the prior probability density function is taken to be the uniform density function on S. The conditional density $p(\boldsymbol{Z}|\boldsymbol{\theta})$ is readily computed for any value of $\boldsymbol{\theta}$ in terms of the $f(\boldsymbol{\Phi}(k),\boldsymbol{\theta})$ and probability densities of the $v(k)$, which are assumed to be of known form [4]. Furthermore, the support set of the density $p(\boldsymbol{\theta}|\boldsymbol{Z})$, for each fixed $\boldsymbol{Z}$, coincides with the region S.

The computation of the estimate $\hat{\boldsymbol{\theta}}$ by Markov chain Monte Carlo (MCMC) reduces to generating, for each fixed $\boldsymbol{Z}$, a convergent Markov chain $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$ from the target density $p(\boldsymbol{\theta}|\boldsymbol{Z})$ defined in (S4), forming a sample average of the $\boldsymbol{\theta}_k$ and then using the fact that the sample averages of the chain converge (a.s.) to $\hat{\boldsymbol{\theta}}$. The proposal density $q(\boldsymbol{w}|\boldsymbol{\theta})$ in [4] is the Gaussian density on $\mathbb{R}^q$ with mean equal to $\boldsymbol{\theta}$ and covariance that is a scalar multiple of the identity matrix. Thus, $q(\boldsymbol{\theta}|\boldsymbol{w}) = q(\boldsymbol{w}|\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ and $\boldsymbol{w}$ in $\mathbb{R}^q$. This identity and (S4) imply that the criterion for accepting the candidate value of $\boldsymbol{w}$ for the next state (given that the current state value is $\boldsymbol{\theta}$) is

$$\rho(\boldsymbol{\theta},\boldsymbol{w}) = \min\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{w})}{p(\boldsymbol{Z}|\boldsymbol{\theta})}, 1\right\}, \quad \boldsymbol{\theta}, w \in \mathbb{R}^q. \tag{S5}$$

As in (1), let $\rho(\boldsymbol{\theta},\boldsymbol{w}) = 1$ when $p(\boldsymbol{Z}|\boldsymbol{\theta}) = 0$. The acceptance probability is easily evaluated for all values of $\boldsymbol{\theta}$ and $\boldsymbol{w}$.

We verify that the conditions for convergence are satisfied for this example. Condition C1 holds as a consequence of the following three assumptions in [4]:

1) S is bounded.
2) The prior probability density function (the uniform density function on S) is positive and constant on S.
3) For each fixed $\boldsymbol{Z}$, the density function $p(\boldsymbol{Z}|\boldsymbol{\theta})$ satisfies $0 < p(\boldsymbol{Z}|\boldsymbol{\theta}) < \infty$ in S. As a consequence, for the target density function $p(\boldsymbol{\theta}|\boldsymbol{Z})$, we also have $0 < p(\boldsymbol{\theta}|\boldsymbol{Z}) < \infty$ in S.

It is easy to see that the Gaussian proposal density satisfies conditions required for C2. The two conditions for convergence are therefore satisfied.

the target distribution, then so is $\boldsymbol{X}_{k+1}$ for all $k \geq 0$. For completeness, this article also discusses the convergence of the distribution for $\boldsymbol{X}_k$ to the target distribution.

The remainder of our discussion is organized as follows. The "Brief Review of the Metropolis–Hastings Algorithm" section summarizes the steps of the M-H algorithm. The "Transition Kernel in the Metropolis–Hastings Algorithm" section provides the analytical form of the transition kernel for the Markov chain generated by the M-H algorithm. The "Stationarity" section shows that the stationary density for the transition kernel is the target density. Finally, the "Convergence" section discusses the convergence of the distribution of the Markov chain to the distribution defined by the target density.

BRIEF REVIEW OF THE METROPOLIS–HASTINGS ALGORITHM

We first summarize the M-H algorithm for generating $\boldsymbol{X}_k$. The approach is based on two stages: 1) the generation of a candidate random value from a proposal distribution (sometimes referred to as a candidate-generating distribution), from which random sampling is easily performed, and 2) the acceptance or rejection of the candidate value according to the Metropolis criterion. The Metropolis criterion accepts the candidate value of $\boldsymbol{w}$ for the next state, given that the current state value is $\boldsymbol{x}$, with probability

$$\rho(\boldsymbol{x},\boldsymbol{w}) = \min\left\{\frac{p(\boldsymbol{w})q(\boldsymbol{x}|\boldsymbol{w})}{p(\boldsymbol{x})q(\boldsymbol{w}|\boldsymbol{x})}, 1\right\}, \quad \boldsymbol{x}, \boldsymbol{w} \in \mathcal{X}, \tag{1}$$

where $q(\cdot|\boldsymbol{x})$ is the proposal density function. (See "Selecting a Proposal Density" for a discussion on selecting a proposal density function.) For each $\boldsymbol{x}$, the probability distribution determined by $q(\cdot|\boldsymbol{x})$ is assumed to be a continuous distribution (this is not a general requirement).

To ensure that this acceptance probability is well defined, as in [7], let $\rho(\boldsymbol{x},\boldsymbol{w}) = 1$ when $p(\boldsymbol{x})q(\boldsymbol{w}|\boldsymbol{x}) = 0$. Given $\boldsymbol{X}_k$, a candidate point $\boldsymbol{W}$ is sampled from $q(\cdot|\boldsymbol{X}_k)$, and, if accepted, $\boldsymbol{X}_{k+1} = \boldsymbol{W}$; otherwise, $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k$. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be dummy variables for $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k+1}$ (so if $\boldsymbol{y}$ is the dummy variable for

$\boldsymbol{X}_{k+1}$, then $\boldsymbol{y}$ will ultimately represent a $\boldsymbol{w}$ or an $\boldsymbol{x}$, depending on whether or not the candidate $\boldsymbol{W}$ is accepted).

The steps of the M-H algorithm are repeated in Algorithm 1 from [1]. The algorithm begins by choosing an initial state value $\boldsymbol{x}_0$ for $\boldsymbol{X}_0$ in step 1 to be some point in the *support set* of $p(\cdot)$, defined as the set S consisting of all points $\boldsymbol{x}$ in $\mathcal{X}$, such that $p(\boldsymbol{x}) > 0$. The convergence results allow for initial state values to be any point in $\mathcal{X}$. Choosing an initial state value in S ensures that the subsequent states $\boldsymbol{X}_k$, $k \geq 1$, will also lie in S (that is, the chain a.s. never leaves S). Step 3 is easily implemented by generating a point U from the uniform distribution $U(0, 1)$ and setting $\boldsymbol{X}_{k+1} = \boldsymbol{W}$ when $U \leq \rho(\boldsymbol{X}_k, \boldsymbol{W})$ and $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k$ otherwise.

The following example illustrates the Metropolis acceptance probability (1) for a bivariate target density function of truncated exponential form. Generating samples from this target density using the M-H algorithm is straightforward, in contrast to generating samples directly from the target density.

Pharmacokinetic Model Estimation

This example considers a problem in biochemical kinetics and pertains to identifying a nonlinear model used in characterizing the rate of enzymatic reactions. The specific application is to model the rate of uptake (as a function concentration) of β-methyl-glucoside in the intestinal tissue of the guinea pig [S1]–[S3]. The identification problem is to estimate unknown parameters in the model for uptake, where the parameter estimates are the posterior mean, given a prior distribution on the parameters. Markov chain Monte Carlo (MCMC) is used to generate samples from a given posterior density function, which are then used to derive an estimate of the posterior mean. In this sense, the parameter estimation problem is similar to that in "Nonlinear Control-System Identification."

Data were collected from an experiment conducted on 50 intestinal tissue samples from each of the eight guinea pigs [S1]. The rate of uptake of β-methyl-glucoside at each of ten concentration levels was measured. The measurements were taken at ten different concentration levels and repeated five times, using a fresh tissue sample for each of the 50 measurements. The five measurements taken at the same concentration level were averaged to provide a total of ten measurements of uptake for each guinea pig.

The uptake z_{ij} for the ith subject at concentration level j is assumed to be [S1]

$$z_{ij} = \log\left[\frac{\psi_{1i} C_{ij}}{\psi_{2i} + C_{ij}} + \psi_{3i} C_{ij}\right] + \varepsilon_{ij} \quad (i = 1, \ldots, 8,\ j = 1, \ldots, 10), \tag{S6}$$

where ε_{ij} (the measurement error) is assumed to be normal with mean zero and variance σ^2, C_{ij} is the jth concentration level of β-methyl-glucoside for the ith subject, ψ_{1i} is the maximal rate of uptake, and ψ_{3i} is a diffusion constant for subject i. The term ψ_{2i} for subject i is the Michaelis affinity constant K_M in modeling biochemical reactions [S4].

In [S3], it is assumed that for each i, $\log\psi_{1i}, \log\psi_{2i}$, and $\log\psi_{3i}$ are jointly normal with mean μ_i and covariance Σ_i, $i = 1, \ldots, 8$. For notational convenience,

$$\boldsymbol{z} = (z_{11}, z_{12}, \ldots, z_{1,10}; z_{21}, z_{22}, \ldots, z_{2,10}; \ldots; z_{81}, z_{82}, \ldots, z_{8,10}),$$

and

$$\boldsymbol{\psi} = (\psi_{11}, \psi_{21}, \psi_{31}; \psi_{12}, \psi_{22}, \psi_{32}; \ldots; \psi_{18}, \psi_{28}, \psi_{38}).$$

Given a prior probability density for $\boldsymbol{\psi}$, it is of interest to generate samples from the posterior probability density function of $\boldsymbol{\psi}$ given the data $\boldsymbol{z}$. Similar to target density in "Nonlinear Control-System Identification," the target density function is the posterior probability density function

$$p(\boldsymbol{\psi}|z) \propto p(z|\boldsymbol{\psi}) p(\boldsymbol{\psi}). \tag{S7}$$

One choice of a proposal density used in [S3] is the normal distribution with mean $\tilde{\boldsymbol{\psi}}$ and covariance matrix equal to a constant times the inverse Fisher information matrix

$$-E\left(\left[\frac{\partial^2 \log p(z|\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right]\Bigg|_{\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}}\right)^{-1}. \tag{S8}$$

In [S3], $\tilde{\boldsymbol{\psi}}$ is the maximum likelihood estimate of $\boldsymbol{\psi}$.

Results of a simulation experiment in which samples were generated from the target posterior density using the described proposal density are reported in [S3]. The results are compared to the Gibbs sampling method for MCMC sampling [1]. The total number of iterations recommended for convergence varied by parameter and experimental subject, for example, from 36,000 for guinea pig 2 to 96,000 for guinea pig 6. The number of recommended iterations, averaged over all parameters and experimental subjects, was approximately 49,000. The study [S3] also reported the empirical acceptance rate, which is defined to be the ratio of the number of accepted proposal samples to the total number of samples generated. A theoretical verification of convergence of the parameter estimates requires an argument similar to that given in "Nonlinear Control-System Identification" and is therefore omitted.

REFERENCES

[S1] S. Johansen, *Functional Relations, Random Coefficients, and Nonlinear Regression with Application to Kinetic Data*, vol. 22. New York: Springer-Verlag, 1984.

[S2] M. J. Linstrom, and D. M. Bates, "Nonlinear mixed effects models for repeated measures data," *Biometrics*, vol. 46, no. 3, pp. 672–687, 1990.

[S3] J. E. Bennett, A. Racine-Poon, and J. Wakefield, "MCMC for nonlinear hierarchical models," in *Markov Chain Monte Carlo In Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. London: Chapman-Hall, 1996, pp. 339–357.

[S4] Wikipedia. (2018). Michaelis–Menten kinetics. [Online]. Available: https://en.wikipedia.org/wiki/Michaelis-Menten_kinetics.

Selecting a Proposal Density

There is wide flexibility in the choice of a proposal density $q(\cdot|\boldsymbol{x})$ to generate candidate values. This sidebar discusses the desired properties of proposal densities. For the Metropolis–Hastings (M-H) algorithm to converge in either of the two forms, (28) and (29), the proposal density must be capable of generating random samples that can reach any subset of S that has nonzero probability in a finite number of steps, starting from any state value in $\mathcal{X}$. More precisely, suppose that z is a point in $\mathcal{X}$ and let $\boldsymbol{Z}_0 = \boldsymbol{z}$. Define $\boldsymbol{Z}_1$ by generating a sample $\boldsymbol{Z}$ from $q(\cdot|\boldsymbol{z})$ and setting $\boldsymbol{Z}_1 = \boldsymbol{Z}$. Having defined $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k$ $(k = 1, 2, 3, \ldots)$, define $\boldsymbol{Z}_{k+1}$ by generating a sample $\boldsymbol{Z}$ according to the density $q(\cdot|\boldsymbol{Z}_k)$, then setting $\boldsymbol{Z}_{k+1} = \boldsymbol{Z}$. This sequence is said to *reach* set A starting from initial state $\boldsymbol{z}$ if $P(\boldsymbol{Z}_k \in A | \boldsymbol{Z}_0 = \boldsymbol{z}) > 0$ for some $k \geq 1$ (depending on $\boldsymbol{z}$ and A). It is necessary that the proposal density generate a sequence $\{\boldsymbol{Z}_k\}$ that is capable of reaching every set A for which $\int_A p(\boldsymbol{y})\,d\boldsymbol{y} > 0$, starting from any initial state $\boldsymbol{z} \in \mathcal{X}$. The proposal density must also be capable of generating an M-H Markov chain $\{\boldsymbol{X}_k\}$ that is irreducible.

The required conditions on the proposal density hold if there is some sufficiently large $\eta > 0$, such that $q(\boldsymbol{y}|\boldsymbol{x}) > 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ with $\|\boldsymbol{y} - \boldsymbol{x}\| < \eta$. The value η is sufficiently large if the proposal density satisfies conditions 1) and 2) in the section "Conditions for Irreducibility."

Another aspect of the proposal density is its acceptance rate, defined in [S5] to be the proportion of proposed moves accepted by the chain in steady state and given by:

$$\int \rho(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})\,d\boldsymbol{x}d\boldsymbol{y}. \tag{S9}$$

The optimal acceptance rate is defined to be the rate that minimizes the asymptotic variance of the estimate $\sum_{K=1}^{n} f(\boldsymbol{X}_k)/n$ of $\int f(\boldsymbol{x})\,d\boldsymbol{x}$. It is shown in [S5], for example, that in a high-dimensional state space, the optimal acceptance rate is approximately 23.4% for Gaussian proposal densities with diagonal covariance matrices, under the assumption that random vectors from the target density have independent identically distributed components.

REFERENCE

[S5] G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various Metropolis-Hastings algorithms," *Statist. Sci*, vol. 16, no. 4, pp. 351–367, 2001. doi: 10.1214/ss/1015346320.

Example 1: Metropolis Criterion Acceptance Probability

Let $\mathcal{X}$ be the open rectangle $(0, B) \times (0, B)$ in $\mathbb{R}^2$, where $B > 0$ is fixed. For each $\boldsymbol{x} = (x_1, x_2)$ in $\mathbb{R}^2$,

$$p(x_1, x_2) = \begin{cases} \beta e^{-x_1 x_2} & \text{if } 0 < x_j < B,\ j = 1, 2, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where β is the normalizing constant. It is easy to see that the support S of $p(x_1, x_2)$ equals the set $\mathcal{X}$. This target probability density function is obtained by taking the product of the appropriate conditional and marginal density functions in [8] (see the discussion in [8, Ex. 2, p. 171] or [9, Ex. 16.5]). Suppose that the proposal density function is bivariate uniform and independent of the current point; that is, $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{y})$ for all $\boldsymbol{y}$ and $\boldsymbol{x}$ in $\mathcal{X}$, and

$$q(\boldsymbol{y}) = \frac{1}{B^2} \boldsymbol{I}_{\mathcal{X}}(\boldsymbol{y}), \quad \boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) \in \mathcal{X}. \tag{3}$$

The Metropolis criterion (1) then becomes

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1, & \text{if } y_1 y_2 < x_1 x_2, \\ e^{-(y_1 y_2 - x_1 x_2)}, & \text{if } y_1 y_2 \geq x_1 x_2, \end{cases} \tag{4}$$

for each $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to $\mathcal{X}$. We suppress the lower bound 0 for $x_1 x_2$ and $y_1 y_2$ in the inequalities in (4) and elsewhere, because both products are nonzero in $\mathcal{X}$. ■

The probabilities of the Markov chain transitioning from one state value to another are determined by steps 2 and 3 of the M-H algorithm. Thus, given $\boldsymbol{X}_k = \boldsymbol{x}$, the probability of transitioning to $\boldsymbol{X}_{k+1} = \boldsymbol{y}$ depends only on the Metropolis acceptance probability $\rho(\cdot, \cdot)$ and the proposal density $q(\cdot|\cdot)$, neither of which depends on the index k. It follows that the conditional probability density for transitioning from one state value to another is not indexed by k. This conditional probability density, which is denoted by $K(\boldsymbol{y}|\boldsymbol{x})$, is called the *transition kernel* and is analogous to the transition matrix $\boldsymbol{P}$ in a discrete Markov chain [2, Sec. 1.12.1]. This kernel, when integrated over subsets of $\mathcal{X}$, yields the one-step transition probabilities for the Markov chain, which are the conditional probabilities of $\boldsymbol{X}_{k+1}$ given $\boldsymbol{X}_k$. Specifically, the transition probability for the set $A \subseteq \mathcal{X}$, given point $x \in \mathcal{X}$,

ALGORITHM 1 The Metropolis–Hastings Algorithm

Step 1. (**Initialization**) Choose the initial state value $\boldsymbol{x}_0$ for $\boldsymbol{X}_0$. Set $k = 0$.

Step 2. Given $\boldsymbol{X}_k$, generate a candidate point $\boldsymbol{W}$ according to the proposal density $q(\cdot|\boldsymbol{X}_k)$.

Step 3. Given $\boldsymbol{W}$, set

$$\boldsymbol{X}_{k+1} = \begin{cases} \boldsymbol{W}, & \text{with probability } \rho(\boldsymbol{X}_k, \boldsymbol{W}), \\ \boldsymbol{X}_k, & \text{with probability } 1 - \rho(\boldsymbol{X}_k, \boldsymbol{W}). \end{cases}$$

Increment counter: $k \to k + 1$.

Step 4. Repeat steps 2 and 3 until some terminal value $\boldsymbol{X}_n$ is available.

The MCMC method is named for its reliance on the construction of a Markovian (dependent) sequence of random variables.

is provided by the integral $\int_A K(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}$. This conditional probability is denoted $P(\boldsymbol{X}_{k+1} \in A|\boldsymbol{X}_k = \boldsymbol{x})$. For each $\boldsymbol{x} \in \mathcal{X}$,

$$P(\boldsymbol{X}_{k+1} \in A|\boldsymbol{X}_k = \boldsymbol{x}) = \int_A K(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}, \quad k = 0, 1, 2, \ldots. \tag{5}$$

We provide the analytical expression for both the kernel $K(\cdot|\cdot)$ and conditional probability $P(\cdot|\cdot)$ in (5) in the "Stationarity" section.

The k-step transition kernel, denoted $K^{(k)}(\cdot|\cdot)$, determines the probability density function (and the associated probability) for transitions across k steps. This kernel is the conditional probability density function for $\boldsymbol{X}_k$ given $\boldsymbol{X}_0$ and is determined by the Chapman–Kolmogorov equations [10, Lemma 6.7],

$$K^{(k+1)}(\boldsymbol{y}|\boldsymbol{x}) = \int K^{(k)}(\boldsymbol{y}|\boldsymbol{v})K(\boldsymbol{v}|\boldsymbol{x})d\boldsymbol{v} = \int K(\boldsymbol{y}|\boldsymbol{v})K^{(k)}(\boldsymbol{v}|\boldsymbol{x})d\boldsymbol{v}, \tag{6}$$

for $k = 1, 2, 3, \ldots$, where $K^{(1)}(\cdot|\cdot)$ is the transition kernel $K(\cdot|\cdot)$ itself. The convention here and throughout is that the domain of integration, when left unspecified, is $\mathcal{X}$ and that all probabilities are well defined (all sets are Borel-measurable subsets of $\mathcal{X}$ and all functions are measurable). The resulting k-step transition probability $P(\boldsymbol{X}_k \in A|\boldsymbol{X}_0 = \boldsymbol{x})$ for set A is then given by $\int_A K^{(k)}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}$. Thus, for each $\boldsymbol{x} \in \mathcal{X}$,

$$P(\boldsymbol{X}_k \in A|\boldsymbol{X}_0 = \boldsymbol{x}) = \int_A K^{(k)}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}. \tag{7}$$

TRANSITION KERNEL IN THE METROPOLIS–HASTINGS ALGORITHM

We now provide the analytical form of $K(\cdot|\cdot)$. From the M-H algorithm, given $\boldsymbol{x}$ and $\boldsymbol{w}$, the outcome $\boldsymbol{y}$ satisfies

$$\boldsymbol{y} = \begin{cases} \boldsymbol{w} & \text{with probability } \rho(\boldsymbol{x}, \boldsymbol{w}), \\ \boldsymbol{x} & \text{with probability } 1 - \rho(\boldsymbol{x}, \boldsymbol{w}). \end{cases} \tag{8}$$

Thus, the (conditional) probability density of $\boldsymbol{y}$ given $\boldsymbol{x}$ and $\boldsymbol{w}$ is a mixture of two point-mass probability density functions: one concentrated at $\boldsymbol{w}$ and weighted by $\rho(\boldsymbol{x}, \boldsymbol{w})$ and the other concentrated at $\boldsymbol{x}$ and weighted by $1 - \rho(\boldsymbol{x}, \boldsymbol{w})$, which is given by

$$\delta(\boldsymbol{y} - \boldsymbol{w})\rho(\boldsymbol{x}, \boldsymbol{w}) + \delta(\boldsymbol{y} - \boldsymbol{x})[1 - \rho(\boldsymbol{x}, \boldsymbol{w})]. \tag{9}$$

The function $\delta(\cdot)$ is the Dirac delta function at $\mathbf{0}$: $\int_A \delta(\boldsymbol{x}')d\boldsymbol{x}' = 1$ if the set $A \subseteq \mathbb{R}^d$ contains $\mathbf{0}$; otherwise, $\int_A \delta(\boldsymbol{x}')d\boldsymbol{x}' = 0$. More precisely, $\delta(\cdot)$ is the point-mass probability measure at $\mathbf{0}$.

The integral of (9) with respect to $q(\cdot|\boldsymbol{x})$ yields the conditional probability density of $\boldsymbol{y}$ given $\boldsymbol{x}$, which is the kernel

$$\begin{aligned} K(\boldsymbol{y}|\boldsymbol{x}) &= \int \{\delta(\boldsymbol{y} - \boldsymbol{w})\rho(\boldsymbol{x}, \boldsymbol{w}) + \delta(\boldsymbol{y} - \boldsymbol{x})(1 - \rho(\boldsymbol{x}, \boldsymbol{w})\} \\ &\quad \times q(\boldsymbol{w}|\boldsymbol{x})d\boldsymbol{w} \\ &= \rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x}) + \delta(\boldsymbol{y} - \boldsymbol{x})(1 - r(\boldsymbol{x})), \end{aligned} \tag{10}$$

where

$$r(\boldsymbol{x}) = \int \rho(\boldsymbol{x}, \boldsymbol{w})q(\boldsymbol{w}|\boldsymbol{x})d\boldsymbol{w}. \tag{11}$$

The first product on the right-hand side of (10) represents the transition probability when candidate values are accepted (that is, $\boldsymbol{y} = \boldsymbol{w}$), and the second product represents the transition probability from $\boldsymbol{x}$ to $\boldsymbol{x}$ given the rejection of candidates $\boldsymbol{w}$. Thus, the transition probability density has both a continuous and a discrete component, and for each set A satisfies

$$\int_A K(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} = \int_A \rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} + I_A(\boldsymbol{x})[1 - r(\boldsymbol{x})], \tag{12}$$

where $I_A(\boldsymbol{x})$ is the indicator function of A given by

$$I_A(\boldsymbol{x}) = \int_A \delta(\boldsymbol{y} - \boldsymbol{x})d\boldsymbol{y} = \begin{cases} 1, & \text{if } \boldsymbol{x} \in A, \\ 0, & \text{if } \boldsymbol{x} \notin A. \end{cases} \tag{13}$$

It can be seen that $K(\cdot|\boldsymbol{x})$ integrates to unity for each $\boldsymbol{x} \in \mathcal{X}$ by taking A in (12) to be $\mathcal{X}$. That is, when $A = \mathcal{X}$, the integral (with respect to $\boldsymbol{y}$) on the right side of (12) becomes $r(\boldsymbol{x})$ and the indicator $I_{\mathcal{X}}(\boldsymbol{x}) = 1$, so that $\int K(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} = 1$, as claimed.

STATIONARITY

Motivation

In this section, we discuss the stationary density for a Markov chain and a property (referred to as irreducibility) that guarantees the uniqueness of a stationary density. We also discuss the implications of this uniqueness. The target density is shown to be a stationary density of the kernel.

The notion of stationarity plays an important role in the convergence of $\boldsymbol{X}_k$. Stationarity is invoked as a condition in proving that the limit distribution of the M-H algorithm is the target density function. The density $p(\cdot)$ is stationary (or invariant) for $K(\cdot|\cdot)$ [10, Def. 6.35], [11, Def. 7.10] if

$$\int_A p(\boldsymbol{y})d\boldsymbol{y} = \int \left[\int_A K(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}\right]p(\boldsymbol{x})d\boldsymbol{x} \qquad \text{for all } A \subseteq \mathcal{X},$$

or, equivalently,

$$\int_A p(\boldsymbol{y})\,d\boldsymbol{y} = \int_A \Big[\int K(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}\Big]d\boldsymbol{y} \quad \text{for all } A \subseteq \mathcal{X}. \quad (14)$$

A stationary density for a kernel is also said to be a stationary density for the associated Markov chain. A necessary and sufficient condition for (14) is that

$$p(\boldsymbol{y}) = \int K(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} \quad \text{a.s. with respect to } p(\cdot) \text{ on } \mathcal{X}, \quad (15)$$

that is, the equality in (15) holds for all $\boldsymbol{y}$ in $\mathcal{X}$, except possibly for a set $E \subseteq \mathcal{X}$ that has probability zero with respect to $p(\cdot)$. Example 2 illustrates the stationarity in (15) using the truncated exponential target density from Example 1.

It is straightforward to show that (15) implies (14). That (15) follows from (14) is a consequence of the Radon–Nikodym theorem (for example, see [12, p. 128]).

It follows from (5) and (15) (and mathematical induction) that if $\boldsymbol{X}_k$ is sampled from $p(\cdot)$, then $\boldsymbol{X}_{k+1}$ will also be distributed according to $p(\cdot)$ for all $k \geq 0$ when $p(\cdot)$ is stationary for $K(\cdot|\cdot)$. In other words, a density that is stationary for $K(\cdot|\cdot)$ is self-replicating.

As will be shown, the target density satisfies (15) and therefore is, by definition, stationary for $K(\cdot|\cdot)$ and self-replicating. If the solution to (15) is unique, then a density of arbitrary other (that is, nontarget) form $p'(\cdot)$ will not be self-replicating. Hence, if $p(x)$ replaces the target $p'(x)$ on the right-hand side of (14), then the resulting density for $\boldsymbol{X}_{k+1}$ on the left side of (15) will not be of the form $p'(\boldsymbol{y})$. That is, for some set $A \subseteq \mathcal{X}$ that has nonzero probability with respect to $p(\cdot)$ and for all $\boldsymbol{y}$ in A,

$$p'(\boldsymbol{y}) \neq \int K(\boldsymbol{y}|\boldsymbol{x})p'(\boldsymbol{x})\,d\boldsymbol{x}. \quad (16)$$

Nonequality on some set of nonzero probability allows for the possibility that equality holds in (16) on some set of probability (strictly) less than one.

The limiting density function of a Markov chain whose k-step transition probabilities (7) converge, independent of the initial state $\boldsymbol{x}_0$, is necessarily a stationary density for $K(\cdot|\cdot)$ and, as a consequence, is also a solution to (15) [11, pp. 133–134]. It follows that, if (15) has a unique solution, then the limiting density of a convergent Markov chain necessarily equals the target density. Thus, the uniqueness of a solution to (15) guarantees that if the MCMC converges in some appropriate sense, then the limiting distribution is the target distribution.

According to [7, Thrm. 1], the target density is the unique stationary density for $K(\cdot|\cdot)$ if the kernel is *irreducible with respect to* $p(\cdot)$ (often simply called *irreducible*). The kernel is irreducible when the state value of the chain will enter any set A that has a nonzero probability (with respect to the target density) in a finite number of steps, starting from any initial state value in $\mathcal{X}$. Specifically (see [10, Def. 6.13] or [7, p. 1711]), the kernel $K(\cdot|\cdot)$ is irreducible with respect to $p(\cdot)$ if, for each $\boldsymbol{x}_0 \in \mathcal{X}$ and each set A with $\int_A p(\boldsymbol{x})\,d\boldsymbol{x} > 0$, there exists a $k > 0$ (depending on $\boldsymbol{x}_0$ and A) such that $P(\boldsymbol{X}_k \in A | \boldsymbol{X}_0 = \boldsymbol{x}_0) > 0$. Thus, an irreducible kernel generates a Markov chain that eventually reaches every set that has a positive $p(\cdot)$ probability, no matter where it starts.

Irreducibility can be formulated in terms of the first entry time of a Markov chain to a set [13, Prop. 4.2.1]. The first entry time of the Markov chain $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ to set the A is $T(A) = \min\{k \geq 1: \boldsymbol{X}_k \in A\}$, where $T(A) = \infty$ if the minimum does not exist. The probability that the Markov chain ever enters A, starting from $\boldsymbol{x}_0$, is given by $P(T(A) < \infty | \boldsymbol{X}_0 = \boldsymbol{x}_0)$. Thus, the kernel is irreducible if $P(T(A) < \infty | \boldsymbol{X}_0 = \boldsymbol{x}_0) > 0$ for every $\boldsymbol{x}_0 \in \mathcal{X}$, whenever $\int_A p(\boldsymbol{x})\,d\boldsymbol{x} > 0$.

When $\boldsymbol{X}_k \in A$, the Markov chain is said to *visit* A at time k, and each index k for which this happens marks a visit to A. Specifically, $T(A)$ is the time of the first visit to A after time $k = 0$. An irreducible kernel is characterized as generating a chain that has some chance of visiting A in finite time after time zero, whenever A has positive probability (with respect to the target density). From [13, Prop. 10.1.1], an irreducible kernel that admits a stationary distribution generates a chain that has some chance of visiting any set A with a positive $p(\cdot)$ probability an infinite number of times [that is, for every integer $n > 0$, there is some $k \geq n$, such that $P(\boldsymbol{X}_k \in A | \boldsymbol{X}_0 = \boldsymbol{x}_0) > 0$] starting from any point $\boldsymbol{x}_0$ in $\mathcal{X}$. Furthermore, the expected total number of visits to any such set A is infinite, and the Markov chain is therefore said to be *recurrent*.

The following is a simple example of a kernel that is not irreducible. Let A_1 and A_2 be two disjoint sets, and suppose that $p_1(\cdot)$ and $p_2(\cdot)$ are probability density functions such that $\int_{A_1} p_1(\boldsymbol{y})\,d\boldsymbol{y} = 1$ and $\int_{A_2} p_2(\boldsymbol{y})\,d\boldsymbol{y} = 1$. Thus, $\int_{A_1} p_2(\boldsymbol{y})\,d\boldsymbol{y} = 0$, and, similarly, $\int_{A2} p_1(\boldsymbol{y})\,d\boldsymbol{y} = 0$. Let $\mathcal{X} = A_1 \cup A_2$, and consider the transition probability density defined by

$$K(\boldsymbol{y}|\boldsymbol{x}) = \begin{cases} p_1(\boldsymbol{y}), & \text{if } \boldsymbol{x} \in A_1, \\ p_2(\boldsymbol{y}), & \text{if } \boldsymbol{x} \in A_2. \end{cases} \quad (17)$$

It is straightforward to check that, for this transition kernel, $P(T(A_i) < \infty | \boldsymbol{X}_0 = \boldsymbol{x}_0) = 0$ if $\boldsymbol{x}_0 \in A_j$ and $i \neq j$. Thus, a Markov chain with the transition kernel defined in (17) can never reach set A_1 starting from points in A_2 and vice versa, and it is therefore not irreducible if $\int_{A_1} p(\boldsymbol{y})\,d\boldsymbol{y} > 0$ or $\int_{A_2} p(\boldsymbol{y})\,d\boldsymbol{y} > 0$.

Conditions for Irreducibility

The conditions presented here for irreducibility cover a broad range of practical applications of MCMC. Let $B(\boldsymbol{x}, R)$ (called the *ball* with *center* $\boldsymbol{x} \in \mathbb{R}^d$) and radius R (where $R > 0$) be the set of all points $\boldsymbol{y} \in \mathbb{R}^d$, such that $\|\boldsymbol{x} - \boldsymbol{y}\| < R$, where $\|\cdot\|$ denotes Euclidean norm in $\mathbb{R}^d$.

The chain, starting from a point $\boldsymbol{x}$, will enter a set $A \subseteq \mathcal{X}$ if there is a (finite) sequence of one-step transitions from $\boldsymbol{x}$ to A. Such one-step transitions are possible if, about each

point in $\mathcal{X}$, there is a sufficiently large ball in which all points may be proposed and accepted.

The following two conditions are sufficient for irreducibility (see the proof of irreducibility in [14, Thm. 2.2]):

1) The proposal density is nonzero in a fixed neighborhood about each point of $\mathcal{X}$. That is, there is an $\eta > 0$ such that

$$q(\boldsymbol{y}|\boldsymbol{x}) > 0 \quad \text{for every } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \text{ such that } \|\boldsymbol{x} - \boldsymbol{y}\| < \eta. \tag{18}$$

2) Any point in S [the support of $p(\cdot)$] can be connected to any subset of S that has a positive $p(\cdot)$ probability using a finite number of balls that have radii less than or equal to $\eta/2$ and overlap in sets of a positive $p(\cdot)$ probability. That is, if $\boldsymbol{x}' \in S$ and A is a subset of S that has positive $p(\cdot)$ probability, there are a finite number of balls, $B_1, B_2, \ldots, B_m$ (each with radius less than or equal to $\eta/2$), such that $\boldsymbol{x}' \in B_1$ and

$$\int_{B_i \cap B_{i+1}} p(\boldsymbol{x})\,d\boldsymbol{x} > 0, \quad i = 1, \ldots, m-2, \int_{B_m \cap A} p(\boldsymbol{x})\,d\boldsymbol{x} > 0. \tag{19}$$

The convergence results for the M-H algorithm allow for the possibility of starting the chain at an initial value outside of the set S. If the current value $\boldsymbol{X}_k$ lies outside of S, then the candidate value is always accepted, because $\rho(\boldsymbol{X}_k, \boldsymbol{W}) = 1$ for all such $\boldsymbol{X}_k$. This probability of acceptance, together with condition 1 (which allows the proposal density to sample candidate values within some fixed positive distance from the current value), implies that if the Markov chain starts outside of the support S, then it can enter S in a finite number of steps. Taking the radii in condition 2 to be less than or equal to $\eta/2$ guarantees that $q(\boldsymbol{y}|\boldsymbol{x}) > 0$, when $\boldsymbol{x} \in B_i \cap B_{i+1}$ and $\boldsymbol{y} \in B_{i+1} \cap B_{i+2}$, or when $\boldsymbol{x} \in B_{m-1} \cap B_m$ and $\boldsymbol{y} \in B_m \cap A$. It follows from (19) that the chain, starting from any point in the support, can reach A in a finite number of steps, provided that A has a positive $p(\cdot)$ probability.

For an independence proposal density, that is, a sampling density that satisfies $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{y})$ for all $\boldsymbol{y}$ and $\boldsymbol{x}$ in $\mathcal{X}$, condition 1 is true when $q(\boldsymbol{y}) > 0$ for all $\boldsymbol{y}$ in $\mathcal{X}$, except possibly for a set of points that has probability zero with respect to $p(\cdot)$. The proof that conditions 1 and 2 imply irreducibility is similar to the proof in [10, Lemma 7.6]. Figure 1 illustrates condition 2 for a hypothetical support region S in $\mathbb{R}^2$. Recall that condition 2 requires that there is a finite number of overlapping balls that connect any two points in the support of the target density.

TARGET DENSITY AS A STATIONARY EQUATION SOLUTION

We show that $p(\cdot)$ satisfies (14) [and equivalently (15)] and is therefore stationary for $K(\cdot|\cdot)$. From (12), for each set A,

$$\int_A \int K(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}\,d\boldsymbol{y} = \int_A \int \rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}\,d\boldsymbol{y} + \int I_A(\boldsymbol{y})[1 - r(\boldsymbol{y})]p(\boldsymbol{y})\,d\boldsymbol{y}. \tag{20}$$

Consider the first expression on the right side of (20). Observe that

$$q(\boldsymbol{x}|\boldsymbol{y})\rho(\boldsymbol{y}, \boldsymbol{x})p(\boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})\rho(\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{x}), \tag{21}$$

which is an easy consequence of the definition of the Metropolis acceptance probability (1). From (21) and Fubini's theorem (which justifies the interchange of the order of integration; for example, see [12, p. 88]), we obtain

$$\int \Big[\int_A \rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x})\,d\boldsymbol{y}\Big]p(\boldsymbol{x})\,d\boldsymbol{x} = \int_A \Big[\int \rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}\Big]d\boldsymbol{y} = \int_A \Big[\int \rho(\boldsymbol{y}, \boldsymbol{x})q(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{x}\Big]d\boldsymbol{y}. \tag{22}$$

Furthermore,

$$\begin{aligned}\int_A \Big[\int \rho(\boldsymbol{y}, \boldsymbol{x})q(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{x}\Big]d\boldsymbol{y} &= \int_A \Big[\int \rho(\boldsymbol{y}, \boldsymbol{x})q(\boldsymbol{x}|\boldsymbol{y})\,d\boldsymbol{x}\Big]p(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int_A r(\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{y} \quad \text{(from (11))} \\ &= \int I_A(\boldsymbol{y})r(\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{y}.\end{aligned}$$

This last equality and (20) yield

$$\begin{aligned}\int_A \Big[\int K(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}\Big]d\boldsymbol{y} &= \int I_A(\boldsymbol{y})r(\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{y} + \int I_A(\boldsymbol{y})[1 - r(\boldsymbol{y})]p(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int I_A(\boldsymbol{y})p(\boldsymbol{y})\,d\boldsymbol{y},\end{aligned}$$

so that

$$\int_A \Big[\int K(\boldsymbol{y}|\boldsymbol{x})\,p(\boldsymbol{x})\,d\boldsymbol{x}\Big]d\boldsymbol{y} = \int_A p(\boldsymbol{y})\,d\boldsymbol{y}.$$

Because the choice of the set A in this last identity is arbitrary, the target density satisfies (14) [and therefore

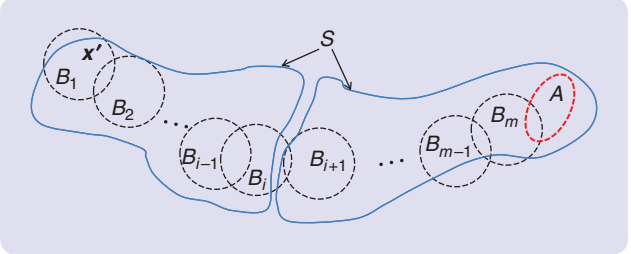

FIGURE 1 A sequence of overlapping balls that connect point $\boldsymbol{x}'$ and set A in disconnected support set S. The proposal density $q(\boldsymbol{y}|\boldsymbol{x}) > 0$ for all $\boldsymbol{y}$ in a neighborhood of radius η about each point $\boldsymbol{x}$ in S. The balls connecting $\boldsymbol{x}'$ and A are of radius less than or equal to $\eta/2$.

(15)], as desired; that is, $p(\cdot)$ is a stationary distribution for M-H.

Example 2: Stationarity of the Target Density

Consider the target and proposal densities in Example 1. It is shown by direct computation that the target density satisfies (15). For this, the terms in the kernel (10) must first be calculated, specifically the product $\rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y})$ and its integral $r(\boldsymbol{x})$ [recall that $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{y})$].

The product term follows immediately from (3) and (4):

$$\rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}) = \begin{cases} \dfrac{1}{B^2}, & \text{if } y_1 y_2 < x_1 x_2, \\ \dfrac{e^{-(y_1 y_2 - x_1 x_2)}}{B^2}, & \text{if } y_1 y_2 \geq x_1 x_2, \end{cases} \tag{23}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to $\mathcal{X}$. As in (4), we have suppressed the lower bound zero for $y_1 y_2$ and $x_1 x_2$ in (23).

Next, consider the term $r(\cdot)$. To evaluate the integral to obtain $r(\cdot)$, it will be convenient to partition $\mathcal{X}$ into three disjoint subsets, denoted $\mathcal{X}_1, \mathcal{X}_2$, and $\mathcal{X}_3$, which are defined in terms of the two inequalities in (23) (see Figure 2). The first two sets of the partition are determined by the inequality $y_1 y_2 < x_1 x_2$. The points $\boldsymbol{y}$ that satisfy this inequality are all (y_1, y_2) in $\mathbb{R}^d$ such that

$$0 < y_1 < B, \text{ and } y_2 < \min\{x_1 x_2 / y_1, B\}, \tag{24}$$

where $0 < x_1 < B$ and $0 < x_2 < B$. Observe that

$$\min\{x_1 x_2 / y_1, B\} = \begin{cases} B, & \text{if } \ 0 < y_1 \leq x_1 x_2 / B, \\ x_1 x_2 / y_1, & \text{if } \ x_1 x_2 / B < y_1. \end{cases} \tag{25}$$

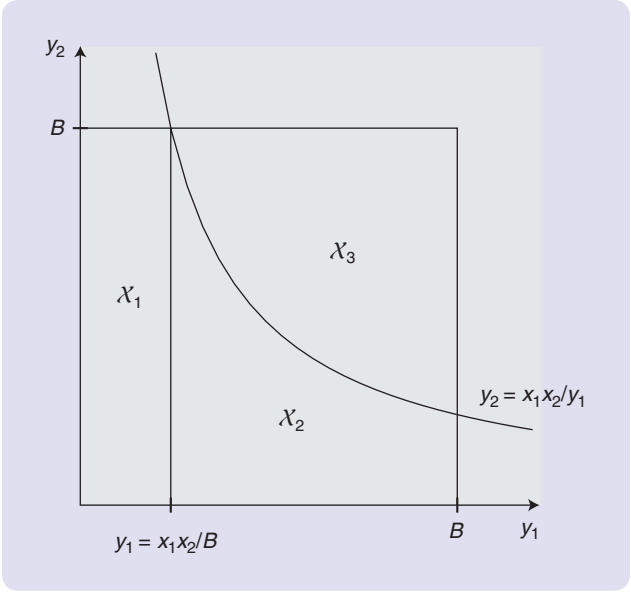

FIGURE 2 The partition of $\mathcal{X}$ into three subsets.

The two intervals for y_1 that define the minimum in (25) split the points in (24) into two regions and provide the first two sets in the partition for $\mathcal{X}$: the rectangle

$$\mathcal{X}_1 = \{\boldsymbol{y} \in \mathcal{X} : 0 < y_1 \leq x_1 x_2 / B, \ \ 0 < y_2 < B\}$$

and the region beneath the curve $y_2 = x_1 x_2 / y_1$

$$\mathcal{X}_2 = \{\boldsymbol{y} \in \mathcal{X} : x_1 x_2 / B < y_1 < B, \ \ 0 < y_2 \leq x_1 x_2 / y_1\}.$$

The remaining points in $\mathcal{X}$ (that is, the points $\boldsymbol{y}$, such that $y_1 y_2 \geq x_1 x_2$ or, equivalently, $y_2 \geq x_1 x_2 / y_1$) are the points in $\mathcal{X}$ that lie on or above the curve $y_2 = x_1 x_2 / y_1$:

$$\mathcal{X}_3 = \{\boldsymbol{y} \in \mathcal{X} : x_1 x_2 / B \leq y_1 < B, \ \ x_1 x_2 / y_1 < y_2 < B\}.$$

Integrating $\rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y})$ with respect to $\boldsymbol{y}$ over $\mathcal{X}$ yields

$$\begin{aligned} r(x) &= \int_{\{y_1 y_2 < x_1 x_2\}} \rho(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{y}) d\boldsymbol{y} + \int_{\{x_1 x_2 \leq y_1 y_2\}} \rho(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{y}) d\boldsymbol{y} \\ &= \int_{\{y_1 y_2 < x_1 x_2\}} \frac{1}{B^2} d\boldsymbol{y} + \int_{\{x_1 x_2 \leq y_1 y_2\}} \frac{e^{-(y_1 y_2 - x_1 x_2)}}{B^2} d\boldsymbol{y} \\ &= \int_{\mathcal{X}_1} \frac{1}{B^2} d\boldsymbol{y} + \int_{\mathcal{X}_2} \frac{1}{B^2} d\boldsymbol{y} + \int_{\mathcal{X}_3} \frac{e^{-(y_1 y_2 - x_1 x_2)}}{B^2} d\boldsymbol{y} \\ &= \int_0^B \left\{ \int_0^{x_1 x_2 / B} \frac{1}{B^2} dy_1 \right\} dy_2 + \int_{x_1 x_2 / B}^B \left\{ \int_0^{x_1 x_2 / y_1} \frac{1}{B^2} dy_2 \right\} dy_1 \\ &\quad + \frac{e^{x_1 x_2}}{B^2} \int_{x_1 x_2 / B}^B \left\{ \int_{x_1 x_2 / y_1}^B e^{-y_1 y_2} dy_2 \right\} dy_1. \end{aligned}$$

Thus,

$$\begin{aligned} r(\boldsymbol{x}) &= \frac{x_1 x_2}{B^2} + \frac{x_1 x_2}{B^2} \log\left(\frac{B^2}{x_1 x_2}\right) \\ &\quad + \frac{e^{x_1 x_2}}{B^2} \left\{ e^{-x_1 x_2} \log\left(\frac{B^2}{x_1 x_2}\right) - \int_{x_1 x_2 / B}^B \frac{e^{-B y_1}}{y_1} dy_1 \right\}, \end{aligned} \tag{26}$$

which completes the calculation for $r(\cdot)$. The remaining integral in (26) is not available in closed form; a closed-form or other solution is not needed to verify (15), which is shown next.

Using (2) and (23),

$$\begin{aligned} \int \rho(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{y}) p(\boldsymbol{x}) d\boldsymbol{x} &= \int_{\{x_1 x_2 < y_1 y_2\}} \frac{\beta e^{-y_1 y_2}}{B^2} d\boldsymbol{x} + \int_{\{y_1 y_2 \leq x_1 x_2\}} \frac{\beta e^{-x_1 x_2}}{B^2} d\boldsymbol{x} \\ &= \beta e^{-y_1 y_2} \left\{ \int_{\{x_1 x_2 < y_1 y_2\}} \frac{1}{B^2} d\boldsymbol{x} \right. \\ &\quad \left. + \int_{\{y_1 y_2 \leq x_1 x_2\}} \frac{e^{-(x_1 x_2 - y_1 y_2)}}{B^2} d\boldsymbol{x} \right\} \\ &= \beta e^{-y_1 y_2} \left\{ \frac{y_1 y_2}{B^2} + \frac{y_1 y_2}{B^2} \log\left(\frac{B^2}{y_1 y_2}\right) \right. \\ &\quad \left. + \frac{e^{y_1 y_2}}{B^2} \left[e^{-y_1 y_2} \log\left(\frac{B^2}{y_1 y_2}\right) - \int_{y_1 y_2 / B}^B \frac{e^{-B x_1}}{x_1} dx_1 \right] \right\} \\ &= \beta e^{-y_1 y_2} r(\boldsymbol{y}), \end{aligned}$$

where the expressions in the third equality use the assumption that the products $x_1 x_2$ and $y_1 y_2$ are both positive. This last identity implies that

A stationary density for a Markov chain is unique if the transition kernel of the chain is irreducible.

$$\begin{aligned}&\int [\rho(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{y}) + \delta(\boldsymbol{y}-\boldsymbol{x})(1-r(\boldsymbol{x}))]p(\boldsymbol{x})d\boldsymbol{x}\\ &= \beta e^{-y_1 y_2} r(\boldsymbol{y}) + (1-r(\boldsymbol{y}))\beta e^{-y_1 y_2}\\ &= \beta e^{-y_1 y_2}\\ &= p(\boldsymbol{y}),\end{aligned}$$

which establishes that the target density (2) satisfies (15) and is, therefore, a stationary density for the M-H kernel.

It is also verified that the two conditions for irreducibility are met for these choice-of-target and proposal density functions. Because $q(\boldsymbol{x})$ is positive everywhere in $\mathcal{X}$, condition 1 is satisfied. Similarly, because $p(\boldsymbol{x})$ is positive everywhere in $\mathcal{X}$, condition 2 is satisfied. It follows, therefore, that (2) is the unique solution to (15).

CONVERGENCE

Recall that a probability density $\tilde{p}(\cdot)$ is stationary for a Markov chain if its probabilities are the limit of the k-step transition probabilities of the chain, independent of the starting point, that is, if $P(\boldsymbol{X}_k \in A | \boldsymbol{X}_0 = \boldsymbol{x}_0) \to \int_A \tilde{p}(\boldsymbol{y})d\boldsymbol{y}$ for each set A independent of $\boldsymbol{x}_0$. Furthermore, a stationary density for a Markov chain is unique if the transition kernel of the chain is irreducible. Hence, if the M-H algorithm converges (in an appropriate sense) and its kernel is irreducible, then the limiting distribution is the target distribution. Thus, when choosing a proposal density for the M-H algorithm, the density must be one for which the associated transition kernel is irreducible (as discussed in the "Conditions for Irreducibility" section). Figure 3 illustrates the relationship among convergence, irreducibility, and stationary densities for Markov chains.

Convergence takes various forms, two of which are discussed in this section: the convergence of probabilities and the convergence of the sample average of a function of random variables. Only the statements of these forms of convergence and conditions under which they hold are given here. Proofs of convergence require measure-theoretic details that are beyond the scope of this article. Numerous references wherein convergence proofs can be found include [7], [10], and [13]–[17].

Consider first the convergence of probabilities. In this form of convergence (for example, [7, Thrm. 1] or [10, Thrm. 7.4 (ii)]), for every $\boldsymbol{x}_0 \in S$ and every $\varepsilon > 0$, there is an integer $k_0 = k_0(\boldsymbol{x}_0, \varepsilon)$ such that for every set $A \subseteq \mathcal{X}$,

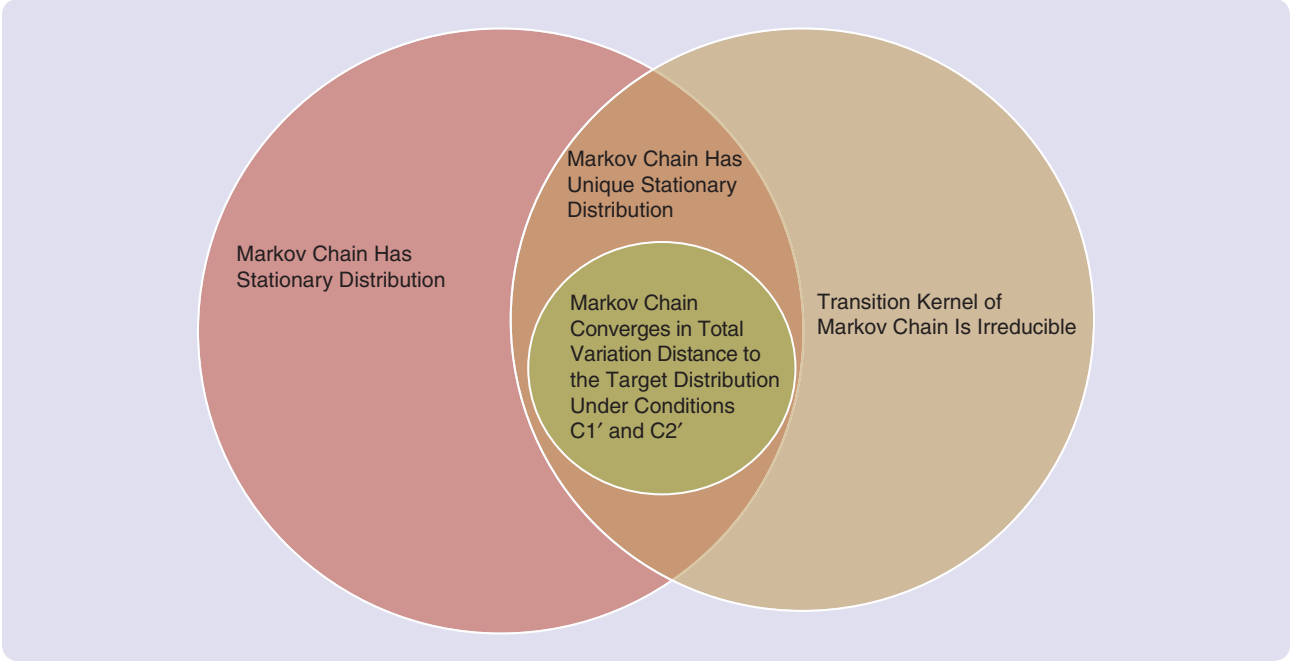

FIGURE 3 An illustration of the relationship among stationarity, convergence, and irreducibility. If the Markov chain is irreducible and has a stationary distribution, then the Markov chain has a unique stationary distribution. The target distribution is a stationary distribution of the Metropolis–Hastings Markov chain. If conditions (C1′) and (C2′) hold, then the k-step transition probabilities converge in total variation distance to the target distribution, where the initial state is any point in S.

$$\left\| P(\boldsymbol{X}_k \in A \,|\, \boldsymbol{X}_0 = \boldsymbol{x}_0) - \int_A p(\boldsymbol{x})\, d\boldsymbol{x} \right\| < \varepsilon \quad \text{whenever } k \geq k_0. \quad (27)$$

An equivalent formulation of (27) is

$$\sup_{A \subseteq \mathcal{X}} \left\{ \left| P(\boldsymbol{X}_k \in A \,|\, \boldsymbol{X}_0 = \boldsymbol{x}_0) - \int_A p(\boldsymbol{x})\, d\boldsymbol{x} \right| \right\} \longrightarrow 0 \quad \text{as } k \to \infty, \quad (28)$$

that is, the conditional probabilities $P(\boldsymbol{X}_k \in A \,|\, \boldsymbol{X}_0 = \boldsymbol{x}_0)$ converge to $\int_A p(\boldsymbol{x})\, d\boldsymbol{x}$ uniformly in A for all $\boldsymbol{x}_0 \in S$. The supremum on the left side of (28) defines the *total variation distance* between the conditional probability distribution of $\boldsymbol{X}_k$ given $\boldsymbol{X}_0 = \boldsymbol{x}_0$ and the target distribution.

The uniform convergence in (28) (that is, convergence in total variation distance) implies convergence in distribution. Convergence in distribution requires only that the distribution function of $\boldsymbol{X}_k$ converge to the distribution function $F(\cdot)$ of the target density pointwise at each point of continuity of F. Thus, the convergence in (28) implies convergence in distribution. To illustrate this, simply take A in (28) to be any set of the form $\{\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 \leq y_1, \ldots, x_d \leq y_d\} \cap \mathcal{X}$, where $\boldsymbol{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$. For such a set,

$$F(\boldsymbol{y}) = \int_{\{x_1 \leq y_1, \ldots, x_d \leq y_d\}} p(\boldsymbol{x})\, d\boldsymbol{x}.$$

According to (28), convergence holds at *all* points of the distribution function F. The converse (that is, that convergence in distribution to the same stationary distribution from all starting points implies convergence in total variation distance) need not be true in general. (See [18] for results on the relationship between convergence in distribution and total variation for Markov chains.)

To state convergence for sample averages, let f be a real-valued (measurable) function, such that the integral $\int |f(\boldsymbol{x})|\, p(\boldsymbol{x})\, d\boldsymbol{x}$ is finite. In this form of convergence (see [7, Thm. 3] or [10, Thm. 7.4 (i)]), for each $\boldsymbol{x}_0 \in S$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(\boldsymbol{X}_k) = \int f(\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} \text{ a.s., given that } \boldsymbol{X}_0 = \boldsymbol{x}_0. \quad (29)$$

Thus, the sample (ergodic) averages of $f(\boldsymbol{X}_k)$ of a chain $\boldsymbol{X}_k$ that starts from state $\boldsymbol{X}_0 = \boldsymbol{x}_0$ converge to $E[f(\boldsymbol{X})]$ a.s., where $\boldsymbol{X}$ is a random vector whose probability density is the target density. The convergence of sample averages of $f(\boldsymbol{X}_k)$ in (29) can also be considered in the mean-square sense. (See the discussion in [1, p. 36] for details on the conditions for mean-square convergence.)

An immediate application of (29) is the use of M-H to compute the expected value

$$\int \boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x}. \quad (30)$$

Two such examples are discussed in "Nonlinear Control-System Identification" and "Pharmacokinetic Model Estimation."

The following two conditions imply that the transition kernel in the M-H algorithm is irreducible (see the "Conditions for Irreducibility" section and "Selecting a Proposal Density") and satisfies additional conditions for the convergence in (28):

C1) For every ball B with center $\boldsymbol{x} \in S$ and arbitrary (positive) radius

$$\inf_{\boldsymbol{x}' \in S \cap B} p(\boldsymbol{x}') > 0 \quad \text{and} \quad \sup_{\boldsymbol{x}' \in S \cap B} p(\boldsymbol{x}') < \infty. \quad (31)$$

C2) There is a $\gamma > 0$ and a $\eta > 0$ such that

$$\text{if } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \quad \text{and} \quad \|\boldsymbol{x} - \boldsymbol{y}\| < \eta \quad \text{then} \quad q(\boldsymbol{y} \,|\, \boldsymbol{x}) \geq \gamma. \quad (32)$$

If, in addition to these two conditions, the integral $\int |f(\boldsymbol{x})|\, p(\boldsymbol{x})\, d\boldsymbol{x}$ is finite, then the convergence in (29) also holds (see [10, Cor. 7.7]). Condition C1 states that the target density is bounded and also bounded away from zero on every bounded (nonempty) subset of S. The condition ensures that the assumptions for the target density in [10, Cor. 7.7] are met, wherein the sets B in C1 are taken to be compact. Condition C2 states that the proposal density is bounded away from zero in some fixed ball about each point of $\mathcal{X}$.

Proofs of the two forms of convergence for general Markov chains can also be found in [15], which proves that the convergence holds from all starting points in $\mathcal{X}$, except for a set of points that has $p(\cdot)$ probability zero. See, in particular, [15, Thrms. 1 and 2]. The two theorems give conditions under which convergence in (28) and (29) holds for almost all starting points $\boldsymbol{x}_0$ in $\mathcal{X}$. For the M-H algorithm, the convergence holds for all starting points $\boldsymbol{x}_0$ in S. For a discussion of the convergence, see [7, Thrms. 1 and 3 and Cor. 2]. It is assumed in [7] that, for all $\boldsymbol{x}_0 \notin S$, $\int_S q(\boldsymbol{y} \,|\, \boldsymbol{x}_0)\, d\boldsymbol{y} = 1$, that is, the support of the proposal density $q(\cdot \,|\, \boldsymbol{x}_0)$ contains S when $\boldsymbol{x}_0$ does not belong to S. This assumption states that, if the M-H algorithm starts outside of S, it reaches S in one step. This restriction on proposal densities implies that the convergence in (28) and (29) holds from all starting points $\boldsymbol{x}_0 \in \mathcal{X}$. See [7] for details. We do not impose this restriction on proposal densities here and, therefore, can only assert that the convergence will hold when the Markov chain starts from points in S.

The proofs of Theorems 1 and 2 in [15] require a form of irreducibility [15, Cons. (1.4) and (1.5)] that is met if the proposal density satisfies conditions 1) and 2) in the section "Conditions for Irreducibility." The assumption in the two theorems that the probability distribution determined by $p(\cdot)$ is stationary for K automatically holds for the M-H algorithm [see the definition in (14) and (15)]. The remaining condition in Theorem 1 of [15]—condition (1.6) there, which is referred to as the *aperiodicity condition*—is satisfied when C1 and C2 hold (compare with the discussion in [15, p. 73]). The remaining condition of Theorem 2 is $\int_S |f(\boldsymbol{x})|\, p(\boldsymbol{x})\, d\boldsymbol{x} < \infty$, which yields the convergence in (29) for almost all starting points $\boldsymbol{x}_0$.

We close this section by giving conditions for the convergence in (28) for the M-H algorithm that are less restrictive than C1 and C2 (and [10, Cor. 7.7]):

C1′) The transition kernel in the M-H algorithm is irreducible.

C2′) $\inf_{x,y\in A}\ \rho(\boldsymbol{x},\boldsymbol{y})\,q(\boldsymbol{x}\,|\,\boldsymbol{y}) > 0$ for some set $A \subseteq \mathcal{X}$ that has positive $p(\cdot)$ probability.

Again, the assumption in Theorems 1 and 2 of [15] that the probability distribution determined by $p(\cdot)$ is stationary for K automatically holds for the M-H algorithm. Condition C1′ and the assumption that A has positive $p(\cdot)$ probability guarantee that the irreducibility conditions (1.4) and (1.5) of [15] are satisfied. Condition C2′ guarantees that condition (1.6) of Theorem 1 of [15] is met, which is the remaining condition of the theorem. Furthermore, if $\int |f(\boldsymbol{x})|p(\boldsymbol{x})d\boldsymbol{x} < \infty$, then Theorem 2 of [15] implies that the convergence in (29) holds for almost all starting points.

CONCLUSION

The usefulness of the M-H form of the MCMC method lies in the ease with which the associated Markov chain can be generated. The justification for the use of M-H relies on the fact that the chain converges in the distributional sense (and other senses) and the limit distribution is the target distribution.

This article interprets and synthesizes several scattered results related to the convergence of M-H. As discussed, the uniqueness of the target density as a stationary density of the chain is key to establishing the target density as the limiting density of the M-H algorithm.

ACKNOWLEDGMENTS

The authors wish to thank Editor-in-Chief Jonathan How and the anonymous reviewers for their comments and suggestions on improving this article. James C. Spall was supported by an APL Sabbatical Professorship at the JHU Whiting School of Engineering.

AUTHOR INFORMATION

Stacy D. Hill (stacy.hill@jhuapl.edu) is a member of the senior professional staff at the Johns Hopkins University (JHU), Applied Physics Laboratory, Laurel, Maryland. He is also a research faculty member in the Applied and Computational Mathematics (ACM) Program of the JHU Whiting School of Engineering Programs for Professionals and serves on the ACM program advisory committee. He has extensive theoretical and practical experience in systems modeling and analysis and has published articles on diverse topics in engineering and statistics, including stochastic simulation optimization, reliability analysis, and parameter estimation. His article "Optimization of Discrete Event Dynamic Systems via Simultaneous Perturbation Stochastic Approximation" (with M.C. Fu) received a Best Paper Award from the Institute of Industrial Engineers and was published in a special issue of *IIE Transactions*. He is a Member of the IEEE. He can be contacted at Johns Hopkins University, Applied Physics Laboratory, 11100 Johns Hopkins Rd., Laurel, MD 20723 USA.

James C. Spall is a member of the principal professional staff at the Johns Hopkins University (JHU), Applied Physics Laboratory, Laurel, Maryland, and is a research professor in the JHU Department of Applied Mathematics and Statistics, Baltimore, Maryland. He is also chair of the Applied and Computational Mathematics Program and cochair of the Data Science Program within the JHU Engineering Programs for Professionals. He has published extensively in the areas of statistics and control and holds two U.S. patents (both licensed) for inventions in control systems. He is the editor and coauthor of the book *Bayesian Analysis of Time Series and Dynamic Models* (Marcel Dekker, 1988) and the author of *Introduction to Stochastic Search and Optimization* (Wiley, 2003). He was one of the inaugural senior editors for *IEEE Transactions on Automatic Control* (2009–2017) and is a contributing editor for the Current Index to Statistics. He was the program chair for the 2007 IEEE Conference on Decision and Control. He is a Fellow of the IEEE.

REFERENCES

[1] J. C. Spall, "Estimation via Markov chain Monte Carlo," *IEEE Control Syst. Mag.*, vol. 23, no. 2, pp. 34–45, 2003.

[2] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 3rd ed. Hoboken, NJ: Wiley, 2017.

[3] B. Ninness and S. Henriksen, "Bayesian system identification via Markov chain Monte Carlo techniques," *Automatica*, vol. 46, no. 1, pp. 40–51, 2010. doi: 10.1016/j.automatica.2009.10.015.

[4] E.-W. Bai, H. Ishii, and R. Tempo, "A Markov chain Monte Carlo approach to nonlinear parametric system identification," *IEEE Trans. Autom. Control*, vol. 60, no. 9, pp. 2542–2546, 2015. doi: 10.1109/TAC.2014.2380997.

[5] B. P. Carlin, N. G. Polson, and D. S. Stoffer, "A Monte Carlo approach to nonnormal and nonlinear state-space modeling," *J. Amer. Statistical Assoc.*, vol. 87, no. 41, pp. 493–500, 1992. doi: 10.1080/01621459.1992.10475231.

[6] W. Cai and J. Wang, "Estimation of battery state-of-charge for electric vehicles using an MCMC-based auxiliary particle filter," *Proc. 2016 American Control Conf.*, pp. 4018–4021.

[7] L. Tierney, "Markov chains for exploring posterior distributions," *Ann. Statist.*, vol. 22, no. 4, pp. 1701–1762, 1994. doi: 10.1214/aos/1176325750.

[8] G. Casella and E. George, "Explaining the Gibbs sampler," *Amer. Statistician*, vol. 46, no. 3, pp. 167–174, 1992. doi: 10.2307/2685208.

[9] J. C. Spall, *Introduction to Stochastic Search and Optimization*. Hoboken, NJ: Wiley, 2003.

[10] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer-Verlag, 2004.

[11] L. Breiman, *Probability*. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

[12] G. G. Roussas, *An Introduction to Measure-Theoretic Probability*, 2nd ed. Oxford, UK: Academic, 2014.

[13] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. New York: Cambridge University Press, 2009.

[14] G. O. Roberts and R. L. Tweedie, "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms," *Biometrika*, vol. 83, no. 1, pp. 95–110, 1996. doi: 10.1093/biomet/83.1.95.

[15] K. B. Athreya, H. Doss, and J. Sethuraman, "On the convergence of the Markov chain simulation method," *Ann. Statist.*, vol. 24, no. 1, pp. 69–100, 1996. doi: 10.1214/aos/1033066200.

[16] S. Asmussen and P. W. Glynn, "A new proof of convergence of MCMC via the Ergodic theorem," *Statist. Probability Lett.*, vol. 81, no. 10, pp. 1482–1485, 2011. doi: 10.1016/j.spl.2011.05.004.

[17] E. Nummelin, *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge, UK: Cambridge Univ. Press, 1984.

[18] R. L. Tweedie, "Modes of convergence of Markov chain transition probabilities," *J. Math. Anal. Appl.*, vol. 60, no. 1, pp. 280–291, 1977. doi: 10.1016/0022-247X(77)90067-1.