A History of the Metropolis-Hastings Algorithm

David B. Hitchcock

University of Florida[1]

August 12, 2003

[1]David B. Hitchcock is a Ph.D. Student, Department of Statistics, University of Florida, Gainesville, FL 32611 (email: dhitchco@stat.ufl.edu). The author thanks the editor, an associate editor, a referee, and George Casella for their helpful comments.

Abstract

The Metropolis-Hastings algorithm is an extremely popular Markov chain Monte Carlo technique among statisticians. This article explores the history of the algorithm, highlighting key personalities and events in its development. We relate reasons for the delay in the acceptance of the algorithm and reasons for its recent popularity.

KEY WORDS: Biography; Markov chain; Monte Carlo method; Simulation; Statistical computing.

1 Introduction

The Metropolis-Hastings (M-H) algorithm, a Markov chain Monte Carlo (MCMC) method, is one of the most popular techniques used by statisticians today. It is primarily used as a way to simulate observations from unwieldy distributions. The algorithm produces a Markov chain whose members' limiting distribution is the *target density* $\pi(x)$. At step j, an observation x_j is generated from an *instrumental density* $q(\cdot|x_i)$ (which is typically easy to simulate from). This candidate observation becomes the next value in the Markov chain with probability

$$\rho = \min\left\{\frac{\pi(x_j)q(x_i|x_j)}{\pi(x_i)q(x_j|x_i)}, 1\right\};$$

with probability $1-\rho$, set $x_j = x_i$, the previous value in the chain (Robert and Casella 1999, p. 233). Under certain conditions, the limiting distribution of the observations in the Markov chain is $\pi(x)$; see Chib and Greenberg (1995) for a detailed introduction.

The M-H algorithm is a relatively old technique, having been introduced in 1953. Yet for decades it languished below the radar, outside the knowledge base of the typical statistician. Consider the statistical landscape just

two decades ago. In the 1982 edition of the *Encyclopedia of Statistical Sciences*, there was no entry for "Metropolis-Hastings algorithm", "Metropolis", "Hastings", or "Markov Chain Monte Carlo" (Kotz and Johnson 1982). Long after the method had originated and even after it had been theoretically validated, this comprehensive reference of all things statistical did not bother to mention it. What was the origin of the method and what factors accounted for its sudden rise to prominence?

2 The Early Days of Monte Carlo Methods

The use of Monte Carlo methods, defined broadly as the field of experiments using random numbers (Hammersley and Handscomb 1964, p. 2), existed well before the twentieth century. In 1777, Georges Louis Leclerc Comte de Buffon established a method for approximating π by repeatedly, randomly throwing a needle onto a grid of parallel lines and tracking how often the needle landed on a line (Liu 2001, p. vii). In the early 20th century, William Gosset ("Student") used simulations with random numbers to help determine the sampling distributions of the correlation coefficient and the t-statistic. But the mathematical branch of Monte Carlo methods really began in earnest

in the 1940s among scientists at the Los Alamos Laboratory in New Mexico, which is where the seeds of the M-H algorithm were sown.

Nicholas C. Metropolis was born in 1915 in Chicago. He attended the University of Chicago, eventually receiving a doctorate in experimental physics there. He researched nuclear reactors with Enrico Fermi and Edward Teller, and, through his work with such noteworthy scientists, he came to the attention of J. Robert Oppenheimer, head of the Manhattan Project — the United States government's plan to build the first atomic bomb. In 1943, at the height of World War II, Oppenheimer recruited Metropolis to Los Alamos to develop mathematical equations to describe the states of physical materials (Ravo 1999). Annoyed by the slow, unwieldy electromechanical calculators they had to use, Metropolis and colleagues Richard Feynman and John von Neumann became interested in the prospect of fast electronic calculators (Santa Fe Institute Bulletin 2000).

After the war, Metropolis went back to the University of Chicago to teach, but in 1948 returned to Los Alamos, where state-sponsored research was burgeoning under America's top scientists. Metropolis led the design of the first programmable super-computer, which he called MANIAC (Mathematical Analyzer, Numerical Integrator and Computer) (Liu 2001, p. vii).

Metropolis chose this name as a satirical poke at the acronyms favored by scientists, but it stuck.

Finally, the computing power was available to drive the development of Monte Carlo (MC) methods, and MC applications soon followed. The motivating example was the random behavior of neutrons in the fissile material in atomic bombs. Two leading mathematicians at Los Alamos, Stanislaw Ulam and John von Neumann, thought of the idea of performing computations via simulation, and Metropolis apparently coined the catchy name "Monte Carlo methods" (Liu 2001, p. viii). Motivated by their physics problems, Metropolis and Ulam (1949) introduced their idea to the statistics community in their paper, "The Monte Carlo Method." They gave an example of estimating the probability of success of a solitaire strategy by undertaking the strategy in many trials and tracking what proportion were successful. Appealing to the theory of probability, they noted, "The estimate will never be confined within given limits with certainty, but only — if the number of trials is great — with great probability." (Metropolis and Ulam 1949, p. 336). They gave another classic example of finding the volume of a 20-dimensional region within a unit cube when the required multiple integrals were intractable. The sensible solution, they explained, might be to subdivide the cube into 10^{20} equally

spaced lattice points, "Take, say 10^4 points at random from this ensemble and ... count how many of the selected points satisfy all the given inequalities" that define the region. Ergodic theorems imply that the resulting estimate should be highly accurate with great probability, they said (Metropolis and Ulam 1949, pp. 336-337).

3 The Birth of an Algorithm

If the 1949 paper introduced the Monte Carlo philosophy, the landmark 1953 paper by Metropolis and the husband and wife teams of Marshall and Arianna Rosenbluth and Edward and Augusta "Mici" Teller took another step forward in providing a specific method in detail. Edward Teller and Marshall Rosenbluth were physicists who had earlier collaborated on research that led to the development of the first hydrogen bomb in 1952. Rosenbluth provided the theoretical calculations to implement Teller's ideas about the hydrogen bomb. They were conducting their research at Los Alamos, which brought them into contact with Metropolis.

The part of the 1953 paper of interest to statisticians is Section II. In this section the authors describe the method of putting N particles at points

on a square to allow the calculation of a $2N$-dimensional integral which is a function of the "energy" E. (Given the computing power of the day, "N may be as [large] as several hundred.") The naive method of Monte Carlo integration would randomly place the N particles in the square, calculate the energy E of this configuration, and weight this configuration by $\exp(-E/kT)$ (what statisticians would later call the objective function). Interestingly, they never define kT in this paper, although we call T the temperature, and k is Boltzmann's constant (Hammersley and Handscomb 1964, p. 117). The problem with this method is that the random choice of a configuration means that "with high probability we choose a configuration where $\exp(-E/kT)$ is very small," so, "instead of choosing configurations randomly, then weighting them with $\exp(-E/kT)$, we choose configurations with a probability $\exp(-E/kT)$ and weight them evenly." (Metropolis et al. 1953, p. 1088)

Metropolis et al. (1953) describe a method which we would today call simulated annealing, in which each particle on the square is moved according to a random (uniform) perturbation, forming a new configuration. This new configuration is accepted if the effected change in energy $\Delta E < 0$ (low energy is good here); or if $\Delta E > 0$, the new configuration is accepted with probability $\exp(-\Delta E/kT)$. Otherwise the previous one is retained.

Having described this simulated annealing method, Metropolis et al. (1953) proceed to the important theoretical result of this paper: that the method is ergodic and the system tends to a distribution $\propto \exp(-E_r/kT)$ for each state r. They note that in their method, the perturbations of the particles are uniform; $P_{rs} = P_{sr}$, where P_{rs} = probability of considering the move from state r to state s *before* accounting for $\exp(-\Delta E/kT)$. Today, we would say that the method uses a symmetric instrumental distribution.

While the paper appeared in a chemical physics journal and was written from a physics viewpoint, it gained some recognition from statisticians. J.M. Hammersley and D.C. Handscomb, in their 1964 book *Monte Carlo Methods,* mention the Metropolis method, but they seemingly fail to grasp its great potential. They list it as a method of solving "problems in equilibrium statistical mechanics," seeing the method only as a way to solve an integral like that in the Metropolis paper, not as a general way to simulate observations from virtually any distribution (Hammersley and Handscomb 1964, pp. 117-121).

In 1965 a mathematical physicist, A.A. Barker, of the University of Adelaide, Australia, published a paper (Barker 1965) with a competing method which employed a slightly different algorithm of sampling from the instrumen-

tal density. Again, the paper was written wholly using physics terminology and focused explicitly on the physics problem of Metropolis et al. (1953).

The Metropolis method was generalized and improved by a professor from the University of Toronto named W. Keith Hastings (1970). Hastings viewed the Metropolis algorithm chiefly as a way to sample from high-dimensional probability distributions, which reflects its primary modern use. Hastings' paper was written in a more statistical style, noting that at its heart the Metropolis method involved the transition matrix of a Markov chain. He presented the target distribution in terms of the invariant distribution $\pi(x)$ of the Markov chain rather than the physically based objective function of Metropolis and Barker.

Hastings' generalization included the possibility of a non-symmetric instrumental distribution. The decision whether to move from x_i to x_j was based on the ratio $[\pi(x_j)q(x_i|x_j)]/[\pi(x_i)q(x_j|x_i)]$ (where $q(x_j|x_i)$ represents the instrumental density from which the candidate x_j was drawn) whereas Metropolis considered only the situation in which $q(x_j|x_i) = q(x_i|x_j)$.

As it turns out, both the Metropolis method and the Barker method arise as special cases of Hastings' generalization. Comparing those two predecessors, Hastings wrote, "Little is known about the relative merits of these two

choices," but suggested that "Metropolis's method may be preferable since it seems to encourage a better sampling of the states." (Hastings 1970, p. 100) Hastings' reasoning was as follows: The Metropolis method would move from x_i to x_j with probability

$$\min\left\{1, \frac{\pi(x_j)q(x_i|x_j)}{\pi(x_i)q(x_j|x_i)}\right\}$$

whereas Barker (1965) would make this move with probability

$$\frac{\pi(x_j)q(x_i|x_j)}{\pi(x_i)q(x_j|x_i) + \pi(x_j)q(x_i|x_j)}.$$

So under a symmetric instrumental distribution, if moving from x_i to x_j resulted in the same value for the target density, Metropolis would make the move with probability 1 but Barker would move with probability 1/2 (1970).

The issue of which method "reigned supreme" was answered by P.H. Peskun, a Ph.D. student of Hastings at the University of Toronto, who devoted his thesis to the topic. In 1973, Peskun (by then at York University in Toronto) published a *Biometrika* paper (Peskun 1973) proving that the general Metropolis-Hastings (M-H) method was optimal. Peskun showed that under the M-H algorithm, the transition matrix led to a sampling method

that was asymptotically as precise as possible. Meanwhile, the Barker algorithm led to a sampling method that was asymptotically no better — and in most cases less precise — than M-H (Peskun 1973).

4 Modern Times

While Peskun's result established the superiority of M-H, it did not lead to statisticians immediately using the method in practice, or even being aware of it. To be sure, those who specialized in the subfield of Monte Carlo methods knew of it, but few practitioners of statistics used the method and the statistical literature contained only passing references to it until the 1990s.

In the early 1990s, the popular MCMC method of Gibbs sampling became known, mainly due to the work of Gelfand and Smith (1990), who built on the seminal paper of Geman and Geman (1984). The Gibbs sampler is a related simulation algorithm that is especially useful for sampling multivariate distributions, particularly when the full univariate conditional densities are known, or are easy to sample from (Robert and Casella 1999, p. 287). Like the Metropolis et al. (1953) paper, Geman and Geman's work, which

introduced the Gibbs sampler, was closely related to optimization problems in statistical physics. Gelfand and Smith (1990) solidified the theory behind Gibbs sampling and some related methods, and, crucially for statistical practitioners, gave examples of common Bayesian analyses which could be greatly enhanced by these methods. Notably, although Gibbs sampling is newer than the M-H algorithm (and is, in fact, a special case of M-H), it came into common use among statisticians slightly earlier than M-H.

In a review (Kass 1997) of the 1996 book *Markov Chain Monte Carlo in Practice*, Robert Kass, a professor at Carnegie-Mellon University, recalled that a 1991 conference at Ohio State University provided the impetus for the introduction of both Gibbs sampling and the M-H algorithm into the statistical mainstream. Kass pointed to a presentation by Luke Tierney of the University of Minnesota which, Kass wrote, "illustrated the use of the Metropolis algorithm to simulate from posterior distributions, pointed out its close relationship to Gibbs sampling, and ... signaled the advent of what has, within statistics, come to be called MCMC." (Kass 1997, p. 1645) The next year, Andrew Gelman (1992) showed that the Gibbs sampler was, formally, a special case of the M-H algorithm. Tierney (1994) would write an influential paper which summarized the history and theory of the M-H

algorithm and showed how it (and the Gibbs sampler) could be employed to solve the problem of working with intractable posterior distributions that often arise in Bayesian inference. Siddhartha Chib and Edward Greenberg (1995) wrote a review article for *The American Statistician* explaining M-H to a wide audience of statistical practitioners. At long last, the M-H algorithm had reached the mainstream.

Kass (1997) pointed out an undeniable factor in the popularization of the M-H algorithm and other MCMC methods: the rise of computing power in the late 1980s and throughout the 1990s. The speed with which computer simulations could be done (not only with giant mainframes like those at Los Alamos, but on everyday personal computers that adorned statisticians' desks) made such algorithmic computations practical. Wrote Kass, "Large numbers of researchers could, in the early 1990s, implement [MCMC] on their desktops for interesting, nontrivial problems." (Kass 1997, p. 1645)

Certainly the M-H algorithm and its MCMC cousins have been a godsend for Bayesians, who, with the ability to simulate from complicated posteriors, were freed from the reliance on conjugate priors. Frequentists, as well, could work with more complicated likelihoods, thanks to these methods.

In closing, let us return to the story of Nicholas Metropolis, who planted

the seeds for this revolution. After writing the 1953 paper, he founded the Institute for Computer Research at Chicago in 1957 but returned in 1964 to the Los Alamos Laboratory, where he would spend the rest of his career (Ravo 1999). Named a senior fellow at the Laboratory in 1980 and given emeritus status in 1987 by the University of California, he enjoyed a storied scientific career, editing many scientific journals and volumes. With Gian Carlo Rota, Metropolis wrote many articles on the dilemmas caused by the rise of computers for the foundations of mathematics — a topic statisticians must face now and in the future (Santa Fe Institute Bulletin 2000). Metropolis died, aged 84, in October 1999 in a nursing home in Los Alamos, New Mexico.

Edward Teller and Marshall Rosenbluth, Metropolis's co-authors, also continued in the scientific limelight after 1953. Teller served as Associate Director of Livermore Laboratory (which he helped establish) in Berkeley, California, from 1954 to 1975 and has been Director Emeritus since 1975. In the 1980s he resurfaced in the public eye as an important proponent of the Strategic Defense Initiative ("Star Wars") and other nuclear defense systems. Marshall Rosenbluth has enjoyed a distinguished career as a physicist that included stints at the General Atomic Corporation, the University of

California-San Diego, and the Institute for Advanced Study in Princeton, New Jersey (*McGraw-Hill Modern Scientists and Engineers* 1980).

References

[1] Barker, A. A. (1965), "Monte Carlo Calculations of the Radial Distribution Functions for a Protonelectron Plasma," *Australian Journal of Physics*, 18, 119-133.

[2] Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327-335.

[3] Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

[4] Gelman, A. (1992), "Iterative and Non-Iterative Simulation Algorithms," *Computing Science and Statistics (Interface Proceedings)*, 24, 433-438.

[5] Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 6, 721-741.

[6] Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Chapman and Hall.

[7] Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.

[8] Kass, R. (1997), "Review of *Markov Chain Monte Carlo in Practice*, by W. R. Gilks, S. Richardson and D. J. Spiegelhalter," *Journal of the American Statistical Association*, 92, 1645-1646.

[9] Kotz, S. and Johnson, N. L., (eds.) (1982), *Encyclopedia of Statistical Sciences*, New York: Wiley.

[10] Liu, J. S. (2001), *Monte Carlo Strategies in Statistical Computing*, New York: Springer.

[11] *McGraw-Hill Modern Scientists and Engineers* (1980), New York: McGraw-Hill.

[12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1092.

[13] Metropolis, N. and Ulam, S. (1949), "The Monte Carlo Method," *Journal of the American Statistical Association*, 44, 335-341.

[14] Peskun, P. H. (1973), "Optimum Monte Carlo Sampling Using Markov Chains," *Biometrika*, 60, 607-612.

[15] Ravo, N. (1999), "Nicholas Metropolis, a Maker of the A-Bomb and Computers," *New York Times* obituary, Oct. 23, 1999.

[16] Robert, C. and Casella, G. (1999), *Monte Carlo Statistical Methods.* New York: Springer.

[17] Santa Fe Institute Bulletin (2000), "News: Nicholas Metropolis, 1915-1999," (Vol. 15, no. 1).

[18] Tanner, M. and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528-550.

[19] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *Annals of Statistics*, 22, 1701-1786.