annalen der physik

www.ann-phys.org

The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo

Michael Betancourt

From its inception in the 1950s to the modern frontiers of applied statistics, Markov chain Monte Carlo has been one of the most ubiquitous and successful methods in statistical computing. The development of the method in that time has been fueled by not only increasingly difficult problems but also novel techniques adopted from physics. Here, the history of Markov chain Monte Carlo is reviewed from its inception with the Metropolis method to the contemporary state-of-the-art in Hamiltonian Monte Carlo, focusing on the evolving interplay between the statistical and physical perspectives of the method.

This particular conceptual emphasis, not to mention the brevity here, requires a necessarily incomplete treatment. A complementary, and entertaining, discussion of the method from the statistical perspective is given in ref. [1]. Similarly, a more thorough but still very readable review of the mathematics behind Markov chain Monte Carlo and its implementations is given in the excellent survey by ref. [2].

I will begin with a discussion of the mathematical relationship between physical and statistical computation before reviewing the historical introduction of Markov chain Monte Carlo and its first implementations. Then I will continue to the subsequent evolution of the method with increasing more sophisticated implementations, ultimately leading to the advent of Hamiltonian Monte Carlo.

1. From Physics to Statistics and Back Again

At the dawn of the twentieth-century, physics became increasingly focused on understanding the equilibrium behavior of thermodynamic systems, especially ensembles of particles. For a system at constant temperature, T, that equilibrium behavior is completely characterized by the *canonical* probability distribution,

$$\pi(q, p) \propto \exp(-H(q, p)/k\,T),$$

where k is the Boltzmann constant, q and p are respectively the positions and momenta of the particles, and $H(q, p) = K(p) + U(q)$ is the *Hamiltonian* defining the physical system. In particular, all physical observables of the equilibrium system become expectations with respect to this canonical distribution,

$$\mathbb{E}[f] = \int \mathrm{d}q\,\mathrm{d}p\,\pi(q, p)\,f(q, p).$$

Often physical observables are independent of the kinematics of the system, reducing the calculation to expectations entirely on the configuration space,

$$\mathbb{E}[f] = \int \mathrm{d}q\,\mathrm{d}p\,\pi(q, p)\,f(q) = \int \mathrm{d}q\,\pi(q)\,f(q),$$

where the equilibrium distribution over configuration space is given by

$$\pi(q) \propto \exp(-U(q)/k\,T).$$

Ultimately equilibrated physical systems are equivalent to probabilistic systems, both specified by a given probability distribution with well-defined characteristics given by expectation values.

By exploiting this equivalence we can facilitate the computation of physical observables. The direct approach for computing observables of equilibrated physical systems, for example, is to simulate the dynamics long enough that time averages over the dynamics converge to the equilibrium observables. Because of this equivalence, however, we can also employ techniques developed in statistics to avoid the costly simulations.

Moreover, the equivalence can be just as useful in statistical problems. Transforming a statistical system into a physical one allows us to utilize the corresponding *pseudodynamics* to improve existing computational methods. The history of Markov chain Monte Carlo, and really much of modern statistical computation, is woven from this interplay between physics and statistics.

2. The Inception of Markov Chain Monte Carlo

A key insight from statistics is that every probability distribution admits a stochastic representation comprised of a *sampling procedure* that generates arbitrarily long sequences of points, $\{\theta_1, \ldots, \theta_N\}$, or *samples*, whose empirical average for any function, f, converges to the corresponding expectation value,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\theta_n) = \mathbb{E}[f].$$

Dr. M. Betancourt
Applied Statistics Center
Columbia University
New York, 10026, New York, United States
E-mail: betanalpha@gmail.com

DOI: 10.1002/andp.201700214

Without knowledge of the sampling procedure itself, the individual points in any such sample appear to be randomly distributed across parameter space, stochastically jumping between neighborhoods of high probability from one iteration to another.

In particular, if we can generate samples from the canonical distribution, or its projection onto configuration space, then we can quickly calculate physical observables without having to resort to expensive simulations. Unfortunately, as one considers increasing complex systems the generation of those samples becomes increasingly challenging.

2.1. "...had to go to Monte Carlo"

Exact sampling procedures generate each point in a sample independently of the others. Generating and publishing tables of exact samples from common probability distributions was a major effort of statistics in the early 20th century, but the work was laborious and limited to very simple distributions. Still the method was a constant curiosity for physicists. Enrico Fermi, for example, would exploit the method to make what seemed like impossibly-quick predictions of experimental outcomes as early as the 1930s.[3] Similarly, Stanislaw Ulam would play with sampling techniques to reason about the outcomes of random games like poker and solitaire.[4]

The computational technologies spurred by World War II and the Manhattan Project, however, soon provided an alternative means of generating samples from physical systems. Soon after the war ended, John von Neumann realized the potential of the Electronic Numerical Integrator and Computer, or ENIAC, for automating the grueling task of generating random samples for estimating physical observables.

Together with Stan Ulam he developed computational methods for generating exact samples from arbitrary univariate distributions[4] while his wife, Klari von Neumann, worked with Nicholas Metropolis to program those algorithms into the ENIAC.[3] At the same time Metropolis and Ulam formalized the initial mathematical foundations of the method.[5] The program quickly became a success for the post-war weapons program and the method was anointed "Monte Carlo" as an homage to the stories Ulam would tell about his uncle who was always asking for money to gamble at the infamous casino.

Not much later the ENIAC was disassembled and moved from the University of Pennsylania to Aberdeen Proving Ground in Maryland, shortly delaying the work of von Neumann, Metrolis, and Ulam. In the meantime, however, Enrico Fermi developed an entirely *analog* version of the method.[3] A small trolley would trace the path of a neutron across a two-dimensional material using tables of random numbers to determine speed, direction, and collision times without the need of expensive computer calculations. The so-called FERMIAC is still around, on display at the Bradbury Science Museum in Los Alamos.[6]

2.2. Markov Chain Reactions

With von Neumann returning to Princeton after the war, Metropolis continued on at Los Alamos, leading the group applying Monte Carlo to study more and more advanced thermodynamic reactions. As the reactions became more complex, however, generating exact samples became infeasible, even with the increasing computational power available from the upgraded ENIAC and its successor, MANIAC.

Along with Arianna Rosenbluth, Marshall Rosenbluth, Edward Teller, and Augusta Teller, Metropolis introduced a new scheme to generate *correlated* samples from the equilibrium distribution.[7] Arianna, an accomplished physicist herself, was responsible for the critical task of implementing the new algorithm on the MANIAC, while Marshall did the bulk of the methodological development.[8] The resulting scheme became known as the Metropolis method, although perhaps a more appropriate name would have been the Rosenbluth method.

Regardless, the scheme introduced a stochastic dynamics that, while unrelated to the true dynamics of the physical system being studied, generates correlated samples from the equilibrium distribution on the configuration space. The artificial pseudodynamics perturb the current state, q, to give a *proposal*, q', with is then accepted with probability

$$a(q, q') = \min\left(1, \frac{\pi(q')}{\pi(q)}\right).$$

If the proposal is rejected then the current state is repeated as a new point in the sample.

Provided that the nature of the perturbations does not depend on the current state, this procedure will generate a *Markov chain* of points in the configuration space that defines a sample from the equilibrium distribution which can be used to estimate observables. The generality of the approach suddenly made sampling methods feasible not only for the thermonuclear problems of immediate interest but also for a variety of other challenging problems in physics and chemistry.

After decades of empirical success in the physical sciences, the Metropolis method was generalized by the statistician W. K. Hastings[9] who realized that the proposal could be given by sampling from an arbitrary distribution, $Q(q \mid q')$, provided that the acceptance probability was modified to

$$a(q, q') = \min\left(1, \frac{Q(q \mid q')\,\pi(q')}{Q(q' \mid q)\,\pi(q)}\right).$$

The resulting *Metropolis–Hastings method* is still to this date one of the most common methods of generating Markov chains for Monte Carlo estimation, a method now known as *Markov chain Monte Carlo*.

The generality of the Metropolis–Hastings method and Markov chain Monte Carlo, however, does not guarantee reasonable practical performance. In particular, if the Markov chain doesn't explore the target distribution well enough and fast enough then we will exhaust our computational resources long before generating a large enough sample to estimate expectations with any reasonable accuracy. For the Metropolis–Hastings method this requires a proposal distribution sufficiently well-suited to the target distribution that each proposal strongly perturbs the initial state without being rejected, ensuring large distances between neighboring points in the subsequent Markov chain.

2.3. The Markov Chain Monte Carlo Revolution in Statistics

Despite Hastings' seminal contribution, Markov chain Monte Carlo techniques were not yet strongly embraced by the statistics community. Although exact sampling methods were finding success in applied problems, for example in refs. [10] and [11], Markov chain Monte Carlo itself was largely considered untrustworthy as the correlations inherent in the samples initially made it difficult to build a rigorous theoretical understanding of when the method would yield well-behaved results. That hesitancy to employ Markov chain Monte Carlo would soon change, however, with the introduction of a particularly useful implementation of the algorithm.

The rapid improvement and proliferation of computing in the 1970s stimulated not only physical scientists but also computer scientists, especially those working on the reconstruction of images corrupted by noise and deterioration. These reconstructions utilized statistical correlations between the black and white pixels in an image, correlations that looks suspiciously like those in spin lattices such as the Ising model that had been extensively studied in the physics literature.

Conveniently, methods for simulating from the equilibrium distribution of these lattices had already been developed in physics. Chief amongst these was *Glauber dynamics*,[12] which updates the orientation of each spin one at a time based on the configuration of its neighbors on the lattice.

In 1984 Stuart and Donald Geman formalized this equivalency and applied Glauber dynamics to the problem of image reconstruction.[13] The resulting *Gibbs sampler* updated each pixel one at a time by sampling from the conditional distribution determined by the configuration of all of the other pixels in the image, generating a Markov chain of correlated pixel configurations distributed according to the desired distribution.

Not long after, Gelfand and Smith introduced the Gibbs sampler to the statistics literature.[14] Having recently addressed the general problem of how to decomposing a multivariate probability distribution into the univariate conditional probability distributions needed to implement a Gibbs sampler, the statistics field was particularly ripe for this implementation of Markov chain Monte Carlo. Gibbs sampling became a rapid success, enabling a flurry of applied statistical analyses that had been up to that point infeasible.

The introduction of the software *Bayesian inference Using Gibbs Sampling*, or BUGS,[15] was especially critical to this revolution. By allowing users to specify a large class of Bayesian posterior distributions bespoke to their analysis and then automating the application of Gibbs sampling, the software facilitated statistical analyses across a diversity of scientific fields. Following its inception in 1989 and first public prototype in 1991, the software quickly became indispensable in many applied domains after its first stable release in 1995.[16]

3. Drifting Towards Scalable Markov Chain Monte Carlo

The inertia of the Markov chain Monte Carlo revolution in statistics carried over through the next few decades, fueling the development of a rigorous understand of the performance and robustness of not only the Gibbs sampler but also Metropolis–Hastings samplers and Markov chain Monte Carlo in general.[17] At the same time practical diagnostics such as the Gelman-Rubin statistic[18] were introduced, testing the consistency of multiple Markov chains to identify the pathologies that had been identified as obstructions to accurate Markov chain Monte Carlo estimation. When carefully employed, these diagnostics could promote the robustness of Markov chain Monte Carlo in practical applications.

The success of these early Markov chain Monte Carlo implementations, however, soon contributed to their own demise. Taking the success for granted, practitioners quickly advanced to increasingly complicated problems that saturated their capabilities.

When targeting distributions with more than ten or so dimensions or non-trivial dependencies amongst the parameters, Metropolis–Hastings implementations with simple proposals and Gibbs samplers become fragile or slow and quickly lose the ability to provide accurate estimators. This diminishing performance motivated the development of the novel Markov chain Monte Carlo implementations needed to tackle the challenging problems to which practical interest had progressed.

3.1. Langevin Monte Carlo

The diminishing performance of the early Metropolis–Hastings implementations is not inherent to the Metropolis–Hastings method itself, but rather a consequence of the simple proposal distributions being used. When applied to complex problems, simple proposals such as the random walk proposal first proposed in ref. [7] either suffer from increasingly high rejection rates or increasingly small jumps.[19] In either case we end up with a Markov chain that moves through parameter space so slowly that we exhaust our computational resources before we can adequately explore the target distribution.

Ultimately the problem is that the random walk proposal generates an isotropic diffusion that in high-dimensions drifts away from the neighborhoods of parameters space relevant to the target distribution. If we let the diffusion drift for an extended time without applying a Metropolis–Hastings correction then we suffer from high rejection rates. At the same time, however, if we apply the correction with high frequency then we interrupt the pseudodynamics and slow the exploration. In other words, the pseudodynamics induced by random walk Metropolis proposals are incompatible with the structure of high-dimensional target distributions.

There are, however, more sophisticated diffusions that can exploit the structure of a given target distribution. In particular, *Langevin diffusions* utilize *differential* information about the target distribution to confine the induced pseudodynamics to the relevant neighborhoods of parameter space. The more directed exploration of this measure-preserving diffusion yields much more efficient exploration than the naive random walk of the original Metropolis algorithm.

One immediate issue with Langevin diffusions, however, is that they cannot be simulated exactly for most problems, and error in the numerical integration of the diffusion will bias the resulting samples. Reference [20] noted that if numerical

integration of the Langevin diffusion is used as the proposal in a Metropolis–Hastings algorithm then the acceptance procedure will compensate for any numerical errors. Provided that the gradients of the target probability density function are available, this approach can drastically improve the performance of Markov chain Monte Carlo.

Reference [20], however, was largely unknown within the statistics literature, and Langevin methods were not much considered until ref. [21] was read before the Royal Statistical Society. The paper introduced a discretized Langevin diffusion without any Metropolis–Hastings correction, which the authors argued was adequate for practical problems. Disagreement by Julian Besag in the formal discussion of the paper motivated Gareth Roberts and Richard Tweedie to undertake a theoretical analysis[22] which demonstrated that the correction was critical to the stability of the algorithm.[23]

This analysis confirmed that by carefully incorporating differential information about the target distribution into the proposal the *Metropolis Adjusted Langevin Algorithm* drastically improved the performance and scalability of Markov chain Monte Carlo. That said, the need to evaluate gradients of the target density proved a substantial burden that limited the adoption of the algorithm both in statistics and science.

3.2. Molecular Dynamics

While Markov chain Monte Carlo was evolving, the continued growth in computing technologies stimulated work on the direct simulation of physical systems.[24–27]

Symplectic integrators[28] allowed larger and more complex physical systems to be simulated with enough accuracy that the resulting dynamical averages were not be too strongly biased from the equilibrium observables. In challenging problems these *molecular dynamics* methods typically yielded estimates much more quickly than competing Markov chain Monte Carlo approaches, although determining whether the bias induced by the numerical integration was indeed small enough for a given application was a persistent challenge.

One particular complication with molecular dynamics, however, is that because the physical dynamics are energy-preserving the dynamical averages yield only *microcanonical* expectations conditioned on the given energy. Quantifying the distribution over the energies themselves requires the difficult computation of the density of states.

Reference [29] introduced a workaround with the introduction of stochastic collisions into the physical dynamics. The modified dynamics would proceed largely as before, evolving along the initial energy level set for a random amount of time before a collision perturbed the momentum and hence the energy of the system. For sufficiently strong perturbations the resulting dynamics explore the entirety of configuration space and in many cases ensure that the resulting dynamical averages converge to the equilibrium observables.

Interestingly, as the time between collisions becomes more frequent these dynamics converge to a Langevin diffusion. The efficacy of the modified dynamics, however, decreases in this limit as the constant collisions limit how much the dynamics explore each energy level set. The efficiency in which the unencumbered dynamics explores those level sets explains why molecular dynamics can often be more efficient than Langevin Monte Carlo methods. In practice, the time between collisions has to be carefully tuned to ensure optimal performance.

3.3. Hybrid Monte Carlo

Although they evolved largely independently, researchers would eventually identify that Markov chain Monte Carlo and molecular dynamics methods could be combined to yield an even more powerful technique.

Motivated by experimental verification of perturbative quantum chromodynamics in the 1970s, numerous groups entered the 1980s in a battle to tackle non-perturbative quantum chromodynamics. Most efforts considered *lattice* methods that discretized the quantum fields on grids where the expectations over the space of field configurations became expectations over the discretized lattice configurations. For bosonic fields the expectations over the lattice configurations were amenable to existing Markov chain Monte Carlo methods, but expectations over the complex-valued fermionic fields were still out of reach.

Initial work relied on *quenching* approximations that simply ignored the fermionic degrees of freedom in quantum chromodynamics. By employing some of the most powerful supercomputers in the world, Gibbs samplers running across the discretized lattice were able to reasonably approximate the resulting bosonic expectations.[30] In this early work the error induced by ignoring the fermionic fields was largely obscured by the error induced from the discretization of the fields themselves. As computation improved and lattices became smaller, however, the fermionic fields would have to be taken into consideration.

The incorporation of fermionic fields was made possible when ref. [31] showed that expectations over fermionic fields could be calculated by augmenting a purely bosonic system with artificial momenta and taking expectations over the corresponding pseudodynamical system. Unfortunately, while expectations over these completely bosonic systems were amenable to Markov chain Monte Carlo, the systems were too complex for existing samplers to be practical, even with the available supercomputing resources.

References [32] and [33] then recognized that certain observables in quantum chromodynamics were equivalent to microcanonical expectations over the pseudodynamical system introduced in ref. [31]. Consequently these observables could be efficiently estimated with a molecular dynamics approach that simulated the pseudodynamics and took dynamical averages. As with molecular dynamics, however, the numerical error induced by the numerical integration of the pseudodynamics was difficult to quantify and limited the utility of the computations.

References [34] and [35] followed with applications of Langevin diffusions to the pseudo-dynamical system, allowing for arbitrary observables to be estimated and not just those equivalent to microcanonical expectations ref. [36] added a Metropolis correction to the Langevin dynamics, independently proposing the Metropolis Adjusted Langevin Algorithm yet again.

The stage was then set for the introduction of a *hybrid* method.[37,38] Similar to ref. [29], this hybrid method simulated the pseudodynamics for an extended time before resampling the momenta and, consequently, the energy, allowing observables to be estimated with a single, prolonged time average.

Finally, in order to correct for the inevitable error arising from the numerical integration of the pseudodynamics ref. [39], considered using the hybrid method as a proposal distribution for the Metropolis method. As with the Metropolis Adjusted Langevin Algorithm, the acceptance procedure rejected simulations that had accumulated too much error, exactly compensating for the inaccuracies of the numerical integrator. The resulting *Hybrid Monte Carlo* algorithm then alternated between two steps – a Metropolis transition driven by the pseudodynamics and a Gibbs transition that sampled new values of the auxiliary momenta parameters.

Before streamlining the name for publication, the authors had originally referred to this final approach as "hybrid-guided" Monte Carlo. Behind the scenes, however, Pendleton and Roweth punned "guided" into "guid", the Scottish slang for "good" appropriate to their positions at the University of Edinburgh. The method became know colloquially as "guid" Monte Carlo for its superior performance compared other algorithms.[40]

4. Going with the Hamiltonian Flow

Hybrid Monte Carlo quickly became the standard for lattice quantum chromodynamics calculations, but that would not be its only application. In an effort to read all papers every written about Markov chain Monte Carlo,[41] Radford Neal came across the method in the review ref. [42] and recognized that the introduction of auxiliary momentum parameters could lift *any* probabilistic system into a pseudodynamical system amenable to Hybrid Monte Carlo

After this realization Neal pioneered its application outside of physics, beginning with his seminal review of Markov chain Monte Carlo[2] and thesis on Bayesian neural networks,[43] and then culminating in a comprehensive review aimed at statisticians.[44]

After his thesis Neal began to appreciate the importance of the pseudo-Hamiltonian system in the construction of the method and transitioned to the revised name *Hamiltonian Monte Carlo*.[41] After many discussions about the method with Neal, David MacKay adopted the new name in his influential textbook[45] which introduced the method to many, including this humble author! The improved naming was facilitated by Neal and MacKay not being aware that Hamiltonian Monte Carlo had already been used to denote unrelated algorithms in lattice quantum chromodynamics.

Eventually Hamiltonian Monte Carlo started to pique the interest of statisticians and practitioners. In particular, the promise of a scalable method and the recognition that *automatic differentiation*[46] could automate the calculation of the necessary derivatives stimulated the development of general-purpose software to facilitate the use of Hamiltonian Monte Carlo in practical applications.

Initial attempts at software implementations, such as that in the Automatic Differentiation Model Builder, or `ADMB`, package popular in ecology,[47] were limited by the need to delicately tune the algorithm to achieve the promised performance. It wasn't until 2011 when Hoffman and Gelman introduced the No-U-Turn sampler capable of automatically tuning Hamiltonian Monte Carlo to achieve high performance in a given problem[48] that the method really began to flourish in practice.

Along with a user-friendly probabilistic programming language[49] and a high-performance automatic differentiation package,[50] the No-U-Turn sampler formed the basis of `Stan`, a multi-environment platform for both specifying statistical models and estimating posterior expectations[51] in the spirit of BUGS. Today `Stan` has revolutionized applied statistics, admitting novel analyses in everything from political science, ecology, and medicine all the way back to astronomy and physics. The dynamic Hamiltonian Monte Carlo methods pioneered in Stan have also been adopted in packages such as `ADMB` and `TMB`,[52] `PyMC3`,[53] and `NONMEM`,[54] with implementations in numerous other packages currently in development.

Parallel to the exploitation of Hamiltonian Monte Carlo, theoretical analysis has illuminated the mathematical foundations of its success. Ultimately the method is driven by a *measure-preserving flow* that rapidly explores a given probability distribution, efficiently aggregating the information needed to construct accurate Markov chain Monte Carlo estimators.[55] The introduction of auxiliary momenta parameters is simply the unique means of lifting the given probabilistic system into a Hamiltonian system naturally equipped with the desired flow. Exploiting this differential geometric understanding has illuminated both how to optimally implement the method in practice and how to diagnose potential biases. A non-technical review of these developments is given in ref. [56].

Identifying measure-preserving flows over more intricate spaces, such as tree spaces, in order to generalize Hamiltonian Monte Carlo is an active field of research.

5. Conclusion

From their inception, the development of Monte Carlo and Markov chain Monte Carlo methods has been constantly fueled by input from physics and related fields. The equivalence between probabilistic systems and equilibrium physical systems provides a bridge that allows the fruitful exchange of ideas and techniques between physics and statistics.

This exchange extends beyond sampling methods. Variational inference,[57] for example has proven successful for computation within certain classes of probability distributions. Similarly, thermodynamic methods[58–61] employ the structure of non-equilibrium systems to facilitate computation in multimodal problems.

Formalizing this connection and identifying unexploited dualities promises to accelerate this exchange between the two fields and push both into new generations of computation.

Acknowledgements

I am extraordinarily thankful to Brian Pendleton, Simon Duane, Andrew Gelman, Tony Kennedy, Radford Neal, Christian Robert, Gareth Roberts,

and Don Rubin for providing critical comments as well as personal recollections of the history presented in this article.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

Hamiltonian Monte Carlo, Markov chain Monte Carlo, molecular dynamics, Monte Carlo

Received: June 5, 2017
Revised: November 29, 2017
Published online: March 23, 2018

[1] C. Robert, G. Casella, *Stat. Sci.* **2011**, *26*, 102.
[2] R. Neal, Probabilistic inference using Markov Chain Monte Carlo methods, Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.
[3] N. Metropolis, *Los Alamos Science* **1987**, *Special Issue*, 125.
[4] R. Eckhardt, *Los Alamos Science* **1987**, *Special Issue*, 131.
[5] N. Metropolis, S. Ulam, *J. Am. Stat. Assoc.* **1949**, *44*, 335.
[6] Brandbury Science Museum, http://www.lanl.gov/museum/, (accessed: April **2017**).
[7] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087.
[8] J. Gubernatis, *Phys. Plasmas* **2005**, *12*, 057303.
[9] W. K. Hastings, *Biometrika* **1970**, *57*, 97.
[10] D. B. Rubin, *J. Educ. Behav. Stat.* **1981**, *6*, 377.
[11] A. P. Dempster, M. R. Selwyn, B. J. Weeks, *J. Am. Stat. Assoc.* **1983**, *78*, 221.
[12] R. J. Glauber, *J. Math. Phys.* **1963**, *4*, 294.
[13] S. Geman, D. Geman, *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6* **1984**, *6*, 721.
[14] A. E. Gelfand, A. F. Smith, *J. Am. Stat. Assoc.* **1990**, *85*, 398.
[15] D. J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, *Stat. Comput.* **2000**, *10*, 325.
[16] D. Lunn, D. Spiegelhalter, A. Thomas, N. Best, *Stat. Med.* **2009**, *28*, 3049.
[17] S. P. Meyn, R. L. Tweedie, *Markov Chains and Stochastic Stability*, Cambridge University Press, New York **2009**.
[18] A. Gelman, D. B. Rubin, *Stat. Sci.* **1992**, *7*, 457.
[19] G. O. Roberts, A. Gelman, W. R. Gilks, *Ann. Appl. Probab.* **1997**, *7*, 110.
[20] P. Rossky, J. Doll, H. Friedman, *J. Chem. Phys.* **1978**, *69*, 4628.
[21] U. Grenander, M. I. Miller, *J. R. Stat. Soc., Ser. B: Methodol.* **1994**, *56*, 549.
[22] G. Roberts, Personal Communication, June **2017**.
[23] G. O. Roberts, R. L. Tweedie, *Bernoulli* **1996**, *2*, 341.
[24] E. Fermi, J. Pasta, S. Ulam, *Los Alamos Report LA-1940* **1955**, *978*.
[25] T. Dauxois, *Phys. Today* **2008**, *61*, 55.
[26] B. J. Alder, T. E. Wainwright, *J. Chem. Phys.* **1959**, *31*, 459.
[27] A. Rahman, *Phys. Rev.* **1964**, *136*, A405.
[28] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, New York **2006**.
[29] H. C. Andersen, *J. Chem. Phys.* **1980**, *72*, 2384.
[30] M. Creutz, L. Jacobs, C. Rebbi, *Phys. Rev. D* **1979**, *20*, 1915.
[31] D. Weingarten, D. Petcher, *Phys. Lett. B* **1981**, *99*, 333.
[32] D. J. E. Callaway, A. Rahman, *Phys. Rev. Lett.* **1982**, *49*, 613.
[33] J. Polonyi, H. Wyld, *Phys. Rev. Lett.* **1983**, *51*, 2257.
[34] G. Batrouni, G. Katz, A. Kronfeld, G. Lepage, B. Svetitsky, K. Wilson, *Phys. Rev. D* **1985**, *32*, 2736.
[35] A. Horowitz, *Phys. Lett. B* **1985**, *156*, 89.
[36] R. T. Scalettar, D. J. Scalapino, R. L. Sugar, *Phys. Rev. B* **1986**, *34*, 7911.
[37] S. Duane, *Nucl. Phys. B* **1985**, *257*, 652.
[38] S. Duane, J. Kogut, *Nucl. Phys. B* **1986**, *275*, 398.
[39] S. Duane, A. Kennedy, B. J. Pendleton, D. Roweth, *Phys. Lett. B* **1987**, *195*, 216.
[40] B. Pendleton, Personal Communication, May **2017**.
[41] R. M. Neal, Personal Communication, November **2017**.
[42] D. Toussaint, *Comput. Phys. Commun.* **1989**, *56*, 69.
[43] R. M. Neal, Ph.D. thesis, University of Toronto, **1995**.
[44] R. Neal, in: *Handbook of Markov Chain Monte Carlo*, (Ed. S. Brooks, A. Gelman, G. L. Jones, X. L. Meng), CRC Press, New York **2011**.
[45] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, New York **2003**.
[46] H. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris, *Automatic Differentiation: Applications, Theory, and Implementations*, Springer, New York **2006**.
[47] D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, J. Sibert, *Optim. Methods Software* **2012**, *27*, 233.
[48] M. D. Hoffman, A. Gelman, *J. Mach. Learn. Res.* **2014**, *15*, 1593.
[49] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, A. Riddell, *J. Stat. Software* **2017**, *76*.
[50] B. Carpenter, M. D. Hoffman, M. Brubaker, D. Lee, P. Li, M. Betancourt, *ArXiv e-prints* **2015**, *1509.07164*.
[51] Stan Development Team, Stan: A C++ library for probability and sampling, version 2.15.0, http://mc-stan.org/, **2017**.
[52] C. Monnahan, adnuts: An r package for nuts sampling using tmb and admb, https://github.com/colemonnahan/adnuts, November **2017**.
[53] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, *PeerJ Comput. Sci.* **2016**, *2*, e55.
[54] R. J. Bauer, *NONMEM Users Guide: Introduction to NONMEM 7.4.1*, 7.4.1 edition, ICON Plc, Gaithersburg, MD, USA **2017**.
[55] M. Betancourt, S. Byrne, S. Livingstone, M. Girolami, *Bernoulli* **2017**, *23*, 2257.
[56] M. Betancourt, ArXiv e-prints **2017**, 1701.02434.
[57] M. J. Wainwright, M. I. Jordan, *Found. Trends. Mach. Learn.* **2008**, *1*, 1.
[58] V. Černøy, *J. Optim. Theory Appl.* **1985**, *45*, 41.
[59] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, *Science* **1983**, *220*, 671.
[60] E. Marinari, G. Parisi, *Europhys. Lett.* **1992**, *19*, 451.
[61] M. Betancourt, *ArXiv e-prints* **2014**, *1405.3489*.

