Exact Convergence Analysis of the Independent Metropolis-Hastings Algorithms

Guanyang Wang (Rutgers Stats)

Joint Statistical Meetings (JSM) 2021

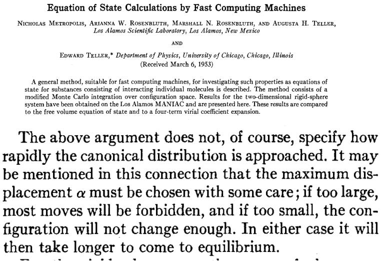
Outline

MCMC: History

- Invented by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in 1953.
- Generalized by Hastings in 1970's.
- Popularized by Gelfand and Smith in 1990 (Gibbs sampler)
- Provides a general (and incredibly popular) approach to simulate from the posterior distribution
- Helped turning Bayesian methods into practically useful tool

MCMC: History

- The authors in the 1953 paper wrote: ‘The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.’



Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

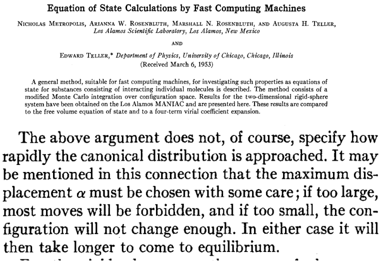
EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*
(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement α must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.

MCMC: History

- The authors in the 1953 paper wrote: 'The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.'
- In 2021, getting quantitative convergence rates is still the central problem in MCMC theory.



Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement α must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.

MCMC: The Metropolis-Hastings algorithm

- Target: Sample from a distribution $\pi(x)$. We know π up to a normalizing constant.
- Key idea: Construct a Markov chain $\{x_1, x_2, x_3, \cdots\}$ with stationary distribution π.
- Given input x_0, transition kernel $q(\cdot, \cdot)$, the algorithm is implemented as follows:

Algorithm 1 Metropolis-Hastings MCMC

for $t = 0, 1, \cdots, T-1$ **do**
  Set $x = x_t$
  Propose $x' \sim q(x, \cdot)$
  Compute $a = \frac{q(x', x)\pi(x')}{q(x, x')\pi(x)}$
  Draw $r \sim \text{Uniform}[0, 1]$
  If $(r < a)$ **then** set $x_{t+1} = x'$
  Else $x_{t+1} = x$
end for

MCMC: The Metropolis-Hastings algorithm

- The sample space can be discrete or continuous, low or high dimensional.
- The correctness is relatively easy. The effectiveness heavily depends on the proposal q.
- Popular choices include $q(x, y) := q(y)$ (independent MH), $q(x, y) = q(\|y - x\|)$ (random-walk MH), gradient-based MH such as MALA/HMC.

Algorithm 2 Metropolis-Hastings MCMC

for $t = 0, 1, \cdots, T-1$ **do**
  Set $x = x_t$
  Propose $x' \sim q(x, \cdot)$
  Compute $a = \frac{q(x',x)\pi(x')}{q(x,x')\pi(x)}$
  Draw $r \sim \text{Uniform}[0, 1]$
  If $(r < a)$ **then** set $x_{t+1} = x'$
  Else $x_{t+1} = x$
end for

Outline

Markov Chain Monte Carlo in 1953

MCMC convergence theory

The IMH algorithm

MCMC convergence: uniform and geometric ergodicity

Definition (Total variation distance)

Let μ, ν be two probability measures on a sigma-algebra $\mathcal{F}$ of subsets of a probability space Ω, the total variation distance between μ and ν is defined as:

$$\|\mu - \nu\|_{\mathsf{TV}} = \max_{A \subset \Omega, A \in \mathcal{F}} |\mu(A) - \nu(A)|$$

A Markov chain P with stationary distribution π is called:

- uniformly ergodic, if

$$\sup_{x \in \mathcal{X}} \|P^n(x, \cdot) - \pi\|_{\mathsf{TV}} \leq C r^n$$

- geometrically ergodic, if

$$\|P^n(x, \cdot) - \pi\|_{\mathsf{TV}} \leq C(x) r^n$$

for $C, C(x) > 0$ and $0 < r < 1$.

MCMC convergence: uniform and geometric ergodicity

- Despite numerous progresses have been made, sharp bounds for for practical MCMC algorithms are very rare.
- Existing techniques mostly rely on the ‘drift-and-minorization’ framework, which often gives conservative bounds.

MCMC convergence: uniform and geometric ergodicity

- Despite numerous progresses have been made, sharp bounds for for practical MCMC algorithms are very rare.
- Existing techniques mostly rely on the ‘drift-and-minorization’ framework, which often gives conservative bounds.

We therefore ask the following two questions:

- (Q1) *How to get sharp convergence rate r of the inequality*

$$\|P^n(x,\cdot) - \pi\|_{TV} \le C(x) r^n?$$

- (Q2) *Does every point x have the same convergence rate?*

Why should we care about Q1 and Q2?

- Q1 seems to be the most natural question after establishing the geometric ergodicity.

Why should we care about Q1 and Q2?

- Q1 seems to be the most natural question after establishing the geometric ergodicity.
- Q2, the convergence speed analysis for different initializations, may be an interesting, important but overlooked question from both a mathematical and an algorithmic point of view.
  - Mathematically, natural extension of Q1.
  - Algorithmically, suppose there exists a Markov chain, such that the convergence rate at one point (say x_1) equals 0.001, while the convergence rate at another point (say x_2) equals 0.999. Then the bound given by Q1 would be practically useless when one starts the chain at x_1.
  - A lot more to be done. Perhaps the only existing work (I am aware of) is [6] by Lubetzky and Sly (2020, PTRF) in the context of Ising Models
- We give complete answers to Q1 and Q2 for independent MH (IMH) algorithms.

Outline

The IMH algorithm

- Short recap: The IMH algorithm is the MH algorithm with $q(x, y) := q(y)$ (the proposed position is independent of the current position).
- The acceptance ratio is of the form $\frac{w(x')}{w(x)}$ given a proposed move $x \to x'$, where $w(x) := \frac{\pi(x)}{q(x)}$.
- The IMH algorithm is commonly used. Some modern variants and applications of IMH algorithm include the Adaptive IMH [4] and the Particle IMH [1], [8]. IMH algorithms are routinely used as a component of auxiliary Monte Carlo methods, such as the Pseudo-marginal Monte Carlo sampler [2].

The IMH algorithm, existing results

- When $\mathcal{X}$ is discrete and finite:
  - Liu [5] calculates all the eigenvalues and the eigenvectors of the IMH transition matrix.
- When $\mathcal{X}$ is continuous ($\mathbb{R}$ or $\mathbb{R}^d$):
  - Mengersen and Tweedie [7] proves: For the IMH algorithm

    $$\text{Uniform Ergodicity} \Leftrightarrow \text{Geometric Ergodicity} \Leftrightarrow w^\star < \infty.$$

    where $w^\star := \sup_{x \in \mathcal{X}} w(x)$.
  - Mengersen and Tweedie [7], Smith and Tierney [11] proves

    $$\|P^n(x, \cdot) - \pi\|_{\mathsf{TV}} \leq (1 - \frac{1}{w^\star})^n,$$

    for every x, given $w^\star < \infty$.
  - Smith and Tierney [11] derives the formula for the n-step transition probability.

Our contribution

- (Q1) Exact convergence rate analysis for the IMH algorithm:
  - General state spaces:

$$\sup_{x \in \mathsf{supp}(\pi)} \|P^n(x,\cdot) - \pi\|_{\mathsf{TV}} = (1 - \frac{1}{w^\star})^n. \tag{1}$$

    It is worth mentioning that formula (1) completely characterizes the worst-case convergence speed for the IMH chain.
  - Discrete state spaces:

$$c_1(1 - \frac{1}{w^\star})^n \leq \sup_{x \in \mathsf{supp}(\pi)} \|P^n(x,\cdot) - \pi\|_{\mathsf{TV}} \leq c_2(1 - \frac{1}{w^\star})^n, \tag{2}$$

    where $0 < c_1 \leq c_2 \leq 1$ are two computable constants.
- (Q2) Convergence rate analysis with different initializations:
  - For both cases, we prove that $P^n(x,\cdot)$ converges to π at the same rate $(1 - \frac{1}{w^\star})$ for all $x \in \mathcal{X}$ under certain conditions.

Some proof ideas: lower bound

The $(1 - \frac{1}{w^\star})^n$ lower bound relies on the following lemma:

Lemma

Let $R(x)$ be the rejection probability at x of the IMH chain on general state space. Then

$$\|P^n(x,\cdot) - \pi\|_{\mathsf{TV}} \geq (R(x))^n. \tag{3}$$

- If $x^\star = \operatorname{argmax} w(x)$, then straightforward calculation yields $R(x^\star) = 1 - 1/w^\star$.
- Otherwise, there is a sequence x_n with $R(x_n) \to 1 - 1/w^\star$.

Hint:

$$\|\mu - \nu\|_{\mathsf{TV}} = \max_{A \subset \Omega, A \in \mathcal{F}} |\mu(A) - \nu(A)|$$

Some proof ideas: different intializations

- Step 1 – Measure transformation: Let $\tilde{\Pi}$ and $\tilde{Q}$ be two cumulative distribution functions (CDFs) on $\mathbb{R}$ defined by:

$$\tilde{\Pi}(s) := \pi(C(s)) = \int_{y \in C(s)} \pi(y) dy$$
$$\tilde{Q}(s) := q(C(s)) = \int_{y \in C(s)} p(y) dy,$$

  where $C(s) := \{x \in \mathcal{X} : w(x) \leq s\}$. Essentially, we are reparameterizing the measure π and p according to w.

Some proof ideas: different intializations

- Step 2 – Exact transition probability: The n-step transition kernel for the IMH chain is given by [11]:

$$P^n(x, dy) = T_n(\max\{w(x), w(y)\})\pi(y)dy + R^n(w(x))\delta_x(dy),$$

where $T_n : \mathbb{R}^+ \to \mathbb{R}$ is defined by:

$$T_n(w) = \int_w^\infty \frac{n\lambda^{n-1}(v)}{v^2} dv.$$

and

$$\lambda(s) = \int_{v \le s} (1 - \frac{v}{s})\tilde{Q}(dv) = \tilde{Q}(s) - \frac{\tilde{\Pi}(s)}{s}.$$

- Step 3 – Our result follows from a careful estimate of the $P^n(x, dy)$ formula.

Connections with rejection sampling and CFTP

- Comparison with rejection sampling: The assumption $w^\star < \infty$ implies we can do rejection sampling.
  - Let h be the function of interest, with $\mathbb{E}_\pi(h) = \mu$ and $\mathsf{Var}_\pi(h) = 1$.
  - The asymptotic variance of the rejection sampling estimator $\sigma^2(\hat{h}_{\mathsf{REJ}})$ is at least $w^\star$
  - The asymptotic variance of the IMH estimator $\sigma^2(\hat{h}_{\mathsf{IMH}})$ between 1 and $2w^\star - 1$.
  - Conclusion: The IMH estimator is always less effective than the $i.i.d.$ samples from π, but may be preferable to the rejection estimator, as it does not require the knowledge of $w^\star$.

Connections with coupling from the past (CFTP)

- The CFTP algorithm is first proposed by Propp and Wilson [10] and has then been a very active area for more than twenty years.
- The condition $w^\star < \infty$ allows one to apply the CFTP technique to draw samples from π directly.
- The CFTP algorithm for IMH is described in page 493 of Murdoch and Green [9] and in page 303 of Corcoran and Tweedie [3], using slightly different languages.
- Idea: We can define an order among all the states according to the value of w function. Given any proposed move for all the chains, if the chain at $x^\star$ agrees to move, all the paths merge into one simultaneously.

Thanks!

- Check out the paper on arxiv https://arxiv.org/abs/2008.02455. "Exact Convergence Rate Analysis of the Independent Metropolis-Hastings Algorithms"

References I

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein.
Particle Markov chain Monte Carlo methods.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(3):269–342, 2010.

Christophe Andrieu and Gareth O Roberts.
The pseudo-marginal approach for efficient Monte Carlo computations.
The Annals of Statistics, 37(2):697–725, 2009.

Jem N Corcoran and Richard L Tweedie.
Perfect sampling from independent Metropolis-Hastings chains.
Journal of statistical planning and inference, 104(2):297–314, 2002.

Lars Holden, Ragnar Hauge, and Marit Holden.
Adaptive independent Metropolis–Hastings.
The Annals of Applied Probability, 19(1):395–413, 2009.

Jun S Liu.
Metropolized independent sampling with comparisons to rejection sampling and importance sampling.
Statistics and Computing, 6(2):113–119, 1996.

Eyal Lubetzky and Allan Sly.
Fast initial conditions for glauber dynamics.
Probability Theory and Related Fields, pages 1–21, 2020.

Kerrie L Mengersen and Richard L Tweedie.
Rates of convergence of the Hastings and Metropolis algorithms.
The Annals of Statistics, 24(1):101–121, 1996.

References II

Lawrece Middleton, George Deligiannidis, Arnaud Doucet, and Pierre E Jacob.
Unbiased smoothing using particle independent Metropolis-Hastings.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2378–2387. PMLR, 2019.

Duncan J Murdoch and Peter J Green.
Exact sampling from a continuous state space.
Scandinavian Journal of Statistics, 25(3):483–502, 1998.

James Gary Propp and David Bruce Wilson.
Exact sampling with coupled markov chains and applications to statistical mechanics.
Random Structures & Algorithms, 9(1-2):223–252, 1996.

Richard L Smith and Luke Tierney.
Exact transition probabilities for the independence Metropolis sampler.
1996.