Markov Chain Monte Carlo and Couplings Part I

Pierre E. Jacob

Optimization-Conscious Econometrics Summer School

June 3-6th, 2024

Outline

Outline

Outline

Goals

Obtain a sample X from a probability distribution π.

Compute expectations $\mathbb{E}[h(X)]$ with respect to π:

$$\mathbb{E}[h(X)] = \int h(x)\pi(x)dx = \pi(h),$$

where h is called a *test function*.

Sampling, numerical integration, Monte Carlo...

Importance in econometrics & statistics

Many point estimators are defined as extrema (*M-estimators*):

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} m(\theta, y_i),$$

thus they naturally relate to the field of optimization.

The objective function $m(\theta, y_i)$ could involve an integral over latent variables, e.g.

$$m(\theta, y_i) = -\log p(y_i|\theta) = -\log \left(\int p(y, u|\theta) du \right).$$

That integral could be intractable.

Digression on the EM algorithm

Expectation Maximization (EM), where integration and optimization are intertwined.

- Compute distribution $q^{(t)}(u) = p(u|y, \theta^{(t)})$.
- Maximize $\mathbb{E}_{q^{(t)}}[\log p(y, U|\theta)]$ to obtain $\theta^{(t+1)}$.

Can be seen as coordinate-wise optimization of

$$\mathcal{F}(q, \theta) = \mathbb{E}_q[\log p(y, U|\theta)] - \mathbb{E}_q[\log q(U)].$$

Neal & Hinton, 1998, *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*,
Wei & Tanner, 1990, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms.*

Importance in econometrics & statistics

Some estimators are directly defined as *expectations*, e.g. posterior means in Bayesian inference:

$$\hat{\theta} = \int \theta \, d\text{posterior}(\theta).$$

Chernozhukov & Hong, 2003, *A MCMC approach to classical estimation*,
Gallant, Hong, Leung & Li, 2021, *Constrained estimation using penalization and MCMC*.

Importance in econometrics & statistics

Hypothesis testing, the p-value is

$$\mathbb{P}(\text{test statistic} > \text{observed test statistic})$$

under the null hypothesis. The distribution of the test statistic might not be χ^2 (believe it or not!), and numerical integration techniques can be necessary.

Besag, 2001, *Markov Chain Monte Carlo for Statistical Inference*, Barber & Janson, 2022, *Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling*.

Importance in econometrics & statistics

In Bayesian model comparison the key quantity is the evidence or marginal likelihood

$$Z = \int \text{likelihood}(\theta) d\text{prior}(\theta).$$

It is often approximated cheaply by the Bayesian Information Criterion (BIC), because the integral is typically intractable.

Monte Carlo methods can be used to obtain consistent approximations.

Geweke, 2007, *Bayesian model comparison and validation.*

Importance in causal inference

- Finding Directed Acyclic Graphs compatible with the data.

  Friedman & Koller, 2003, *Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks.*

  Kuipers & Moffa, 2017, *Partition MCMC for Inference on Acyclic Digraphs.*

  Agrawal, Broderick & Uhler, 2018, *Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models.*

- In the potential outcome framework, success of Bayesian nonparametric methods in causal inference competitions.

  Oganisian & Roy, 2020, *A Practical Introduction to Bayesian Estimation of Causal Effects: Parametric and Nonparametric Approaches.*

  Linero & Antonelli, 2021, *The how and why of Bayesian nonparametric causal inference.*

Monte Carlo

- Originates from physics in the 1940s, and an active research topic in physics, applied maths, statistics, econometrics, etc.

  Brooks, Gelman, Jones, & Meng, 2011, *Handbook of Markov Chain Monte Carlo*.

- Often state-of-the-art for numerical integration e.g.

  Novak, 2016, *Some results on the complexity of numerical integration*.

- Strong links between sampling and optimization,

  Chewi, 2023, *Log-Concave Sampling*.

Monte Carlo

Monte Carlo methods are algorithms: with input, output, algorithmic/tuning parameters, running cost, memory cost, etc.

Inverse CDF transform, rejection sampling, importance sampling...

Hard to come up with a "global" approximation of the target π.

$\rightarrow$ Markov chain Monte Carlo (MCMC) to the rescue.

MCMC

Start with a sample X_0, from $\pi_0 \neq \pi$.

Apply random perturbations to X_0 to obtain X_1.

Apply random perturbations to X_1 to obtain X_2.

...

Obtain X_T, approximately from π if T is large enough.

Note that this generates a discrete-time Markov chain. Some algorithms operate on continuous time, e.g. piecewise deterministic MCMC.
Vanetti, Bouchard-Côté, Deligiannidis, & Doucet, 2017,
Piecewise-deterministic Markov chain Monte Carlo.

A few long runs, or many short runs?

Long runs: $\int h(x)d\pi(x)$ can be approximated by $T^{-1}\sum_{t=0}^{T-1} h(X_t)$ if T is large enough.

Parallel runs: we can also run R chains for T steps and $\int h(x)d\pi(x)$ can be approximated by $R^{-1}\sum_{r=1}^{R} h(X_T^{(r)})$.

Second approach is risky as finite time bias decreases with T and not with R. Variance decreases with either T or R.

Rosenthal, 2000, *Parallel computing and Monte Carlo algorithms.*

Outline

Let's try

Target distribution:

$$\pi(x) \propto \exp\left(-0.01(2+\cos(x))x^2\right).$$

It is indeed a probability density function on $\mathbb{R}$, once normalized.

Let's start from $X_0 = 0$. How do we construct $(X_t)_{t\geq 1}$?

Metropolis–Rosenbluth–Teller–Hastings transition

With Markov chain at state X_t,

1. propose $X^\star \sim q(X_t, \cdot)$,
2. sample $U \sim \text{Uniform}(0, 1)$,
3. if
$$U \leq \frac{\pi(X^\star) q(X^\star, X_t)}{\pi(X_t) q(X_t, X^\star)},$$
set $X_{t+1} = X^\star$, otherwise set $X_{t+1} = X_t$.

Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, 1970.

Note: everyone calls this the Metropolis algorithm, or Metropolis–Hastings.

Dr. Arianna Wright Rosenbluth (1927-2020)

Dr. Arianna Wright Rosenbluth in 2013. She helped create what has become one of the most important algorithms of all time.
via Rosenbluth family

From `https://www.nytimes.com/2021/02/09/science/arianna-wright-dead.html`, by Katie Hafner.

The germinal paper

Equation of State Calculations by Fast Computing Machines (1953) by Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller

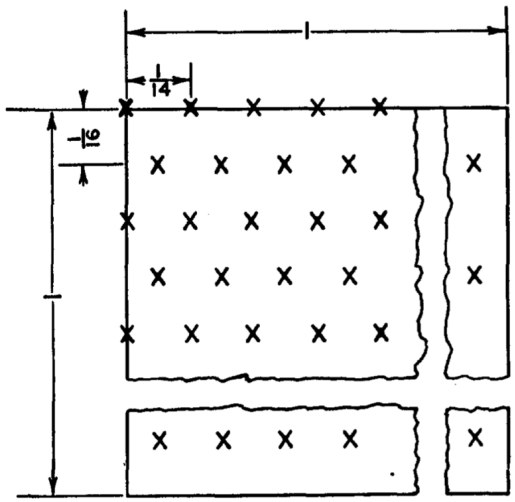


FIG. 2. Initial trigonal lattice.

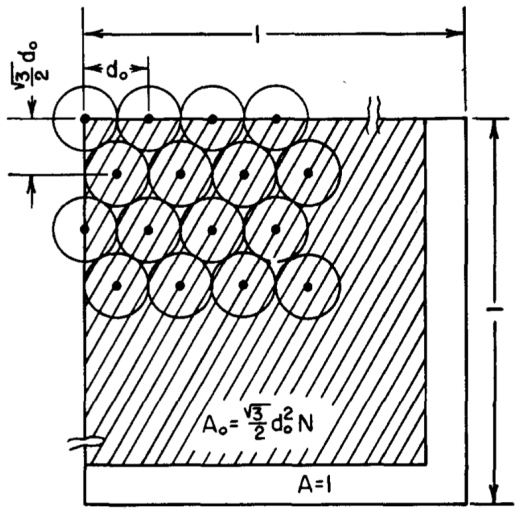


FIG. 3. The close-packed arrangement for determining A_0.

$r_{ij}=d_0$ and $\sum_j F_{ij}$ is given by Eq. (8), so we have

$$\langle \sum_i \mathbf{X}_i^{(\mathrm{int})} \cdot \mathbf{r}_i \rangle_{\mathrm{Av}} = -(Nm\bar{v}^2/2)\pi d_0^2 \bar{n}. \qquad (9)$$

The unit cell is a parallelogram with interior angle 60°, side d_0, and altitude $3^{\frac{1}{2}}d_0/2$ in the close-packed system.

Every configuration reached by proceeding according to the method of the preceding section was analyzed in

Diaconis, Lebeau & Michel, 2011, *Geometric analysis for the Metropolis algorithm on Lipschitz domains*

Fixed point iteration and detailed balance

MCMC = fixed point iteration on the space of measures.

Let π_t be the marginal distribution of X_t at time t, and $P(x, \cdot)$ be the conditional distribution of X_{t+1} given $X_t = x$.

We have $\pi_{t+1} = \pi_t P$ i.e.

$$\pi_{t+1}(dx') = (\pi_t P)(dx') = \int \pi_t(dx) P(x, dx').$$

P is π-invariant if $\pi = \pi P$.

Fixed point iteration and detailed balance

MRTH transition kernel:

$$P_{\text{MRTH}}(x, dx') = r(x)\delta_x(dx') + q(x, dx')\alpha(x, x'),$$

where $r(x)$ is $\ldots$ in terms of q and $\alpha\ldots$

Check detailed balance:

$$\pi(x)P_{\text{MRTH}}(x, x') = \pi(x')P_{\text{MRTH}}(x', x) \quad \forall x, x'.$$

Detailed balance implies π-invariance.

The "right" acceptance probability?

Why should the acceptance probability in MRTH be equal to

$$\alpha(X_t, X^\star) = \min\left(1, \frac{\pi(X^\star)q(X^\star, X_t)}{\pi(X_t)q(X_t, X^\star)}\right)$$

...?

It could be different (Barker 1965). Peskun (1973) shows that the above formula leads to the maximal acceptance rate, for a given proposal q.

In turn this leads to a minimal asymptotic variance for the estimator $T^{-1}\sum_{t=0}^{T-1} h(X_t)$, for all test functions h.

MRTH as a modified random walk

The MRTH transition kernel can be seen as the projection of the proposal kernel q onto the space of π-invariant Markov kernels, under the distance

$$d(P_1, P_2) = \int_{\{x \neq x'\}} \pi(dx)|P_1(x, x') - P_2(x, x')|dx'.$$

Thus it is the smallest (in that sense) modification of q such that the chain (X_t) converges to π.

Other distances could lead to other acceptance probabilities.

Billera & Diaconis, 2001, *A geometric interpretation of the Metropolis–Hastings algorithm.*
Chewi, 2023, *Log-Concave Sampling*, Section 7.2.

Gibbs sampling

Update components of the state, conditional on the others.
Note: MRTH 1953 paper introduced a Gibbs sampler.

Order of updates: random or systematic.

Example: slice sampling.

Exercise: sample from the distribution with density $\pi(x) = \exp(-\sqrt{x})/2$, for $x \geq 0$.

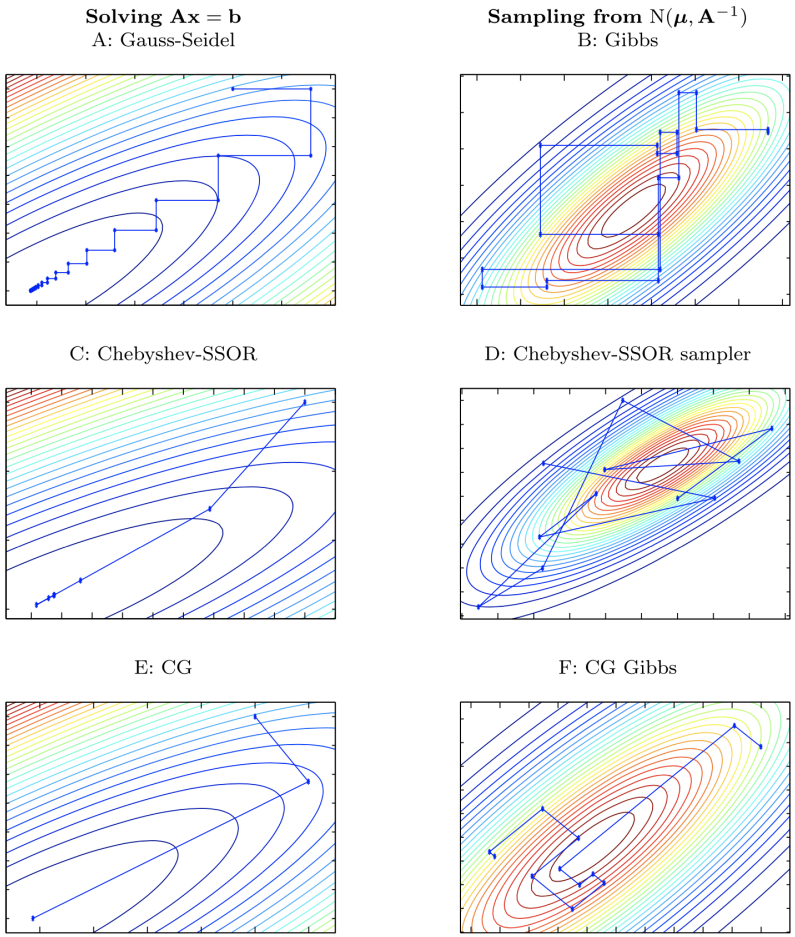
Exercise: probit regression: $\mathbb{P}(Y = 1|X) = \Phi(X\beta)$, where Φ is Normal$(0, 1)$ cumulative distribution function (CDF), $\beta \sim \text{Normal}(\mu, \Sigma)$.

Albert & Chib, 1993, *Bayesian Analysis of Binary and Polychotomous Response Data*.

Gibbs sampling

Gibbs sampling for Normals $\leftrightarrow$ numerical linear algebra.

Fox & Parker, 2017, *Accelerated Gibbs sampling of Normal distributions using matrix splittings and polynomials.*



Langevin Monte Carlo

Langevin diffusion in continuous time:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t,$$

where U is the potential function, dB_t is the increment of the Brownian motion.

Convergence to equilibrium distribution

$$\pi(dx) = \frac{\exp(-U(x))dx}{Z},$$

under e.g. strong convexity of U, and $\mathbb{E}[|X_0|^2] < \infty$.

Simple proof by coupling, as we'll see tomorrow!

Langevin Monte Carlo

Thus if we want to obtain samples from π, we can define

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t,$$

and obtain samples X_T for T large enough.

Discretize time (forward explicit "Euler" scheme):

$$X^\epsilon_{t+1} = X^\epsilon_t + \epsilon \nabla \log \pi(X^\epsilon_t) + \sqrt{2\epsilon} Z_{t+1}, \quad Z_{t+1} \sim \text{Normal}(0, 1).$$

This is the "unadjusted Langevin algorithm" (ULA), dating back to Ermak, 1975 or Parisi, 1981.

Due to the discretization error, X^ϵ_t does not converge to π but to an approximation of it (consistent as $\epsilon \to 0$).

Langevin Monte Carlo

Links between
ULA: $X_{t+1}^{\epsilon} = X_t^{\epsilon} + \epsilon \nabla \log \pi(X_t^{\epsilon}) + \sqrt{2\epsilon} Z_{t+1}$.

Gradient descent: $X_{t+1} = X_t + \epsilon_t \nabla \log \pi(X_t)$.

In fact the Langevin diffusion is gradient descent of the entropy functional on the space of measures metrized by W_2.

Jordan, Kinderlehrer & Otto, 1998, *The variational formulation of the Fokker–Planck equation*,

Explosion of papers on the convergence of these algorithms, well reviewed in Chewi, 2023, Log-Concave Sampling.

Langevin Monte Carlo

We can use "ULA" as a proposal distribution:

$$X^\star \sim \text{Normal}(X_t + \epsilon \nabla \log \pi(X_t), 2\epsilon I),$$

and accept/reject using the MRTH step.

This is the Metropolis-adjusted Langevin algorithm (MALA), just MRTH with a particular choice of proposal distribution.

Rossky, Doll & Friedman, 1978, *Brownian dynamics as smart Monte Carlo simulation.*
Besag, 1994, *Comment on Grenander & Miller's "Representations of Knowledge in Complex Systems".*
Dwivedi, Chen, Wainwright, Yu, 2018, *Log-concave sampling: Metropolis–Hastings algorithms are fast!.*

Langevin as limit of MRTH with small steps

Curiously, MRTH with component-wise random walk updates in random order,

$$X_i^\star \sim \text{Normal}(X_{t,i}, \sigma^2),$$

without gradients employed explicitly in the proposal, also converges weakly to the Langevin diffusion as $\sigma^2 \to 0$.

Gelfand & Mitter, 1991, *Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions.*

Hamiltonian Monte Carlo

Extended distribution $\bar{\pi}(q, p) = \pi(q) \cdot \text{Normal}(p; 0, I)$.

Minus log: potential energy $U(q) = -\log \pi(q)$, kinetic energy $K(p) = \frac{1}{2}|p|^2$, and total energy $E(q, p) = U(q) + \frac{1}{2}|p|^2$.

Hamiltonian dynamics for $(q(s), p(s))$, where $s \geq 0$:

$$\frac{d}{ds} q(s) = \nabla_p E(q(s), p(s)),$$
$$\frac{d}{ds} p(s) = -\nabla_q E(q(s), p(s)).$$

The total energy $E(q, p)$ is constant along trajectories.

Solving Hamiltonian dynamics exactly is not feasible, but we can discretize it and add an accept/reject step to generate a π-invariant chain.

Hamiltonian Monte Carlo

From a current state X_t, set $q_0 = X_t$.

Sample $p_0 \sim \text{Normal}(0, I)$.

With stepsize η, number of "leap-frog" steps L, for $\ell = 1, \ldots, L$,

$$p_{\ell+1/2} = p_\ell - \frac{\eta}{2}\nabla U(q_\ell),$$
$$q_{\ell+1} = q_\ell + \eta p_{\ell+1/2},$$
$$p_{\ell+1} = p_{\ell+1/2} - \frac{\eta}{2}\nabla U(q_{\ell+1}).$$

This is a "Verlet" integrator $\neq$ Euler integrator.

Accept or reject q_L with probability:

$$\alpha = \min(1, \exp(-E(q_L, p_L) + E(q_0, p_0))) = \min(1, \frac{\bar{\pi}(q_L, p_L)}{\bar{\pi}(q_0, p_0)}).$$

If $L = 1$, we recognize MALA with $\epsilon = \eta^2/2$.

Hamiltonian Monte Carlo

Appeal of HMC: can do large moves without sacrificing the acceptance probability.
Neal, 2011, *MCMC using Hamiltonian dynamics*.
Dates back to Duane, Kennedy, Pendleton, Roweth, 1987.

In many software packages. Whenever one can do ULA/MALA one can do HMC, and it is generally believed to be a good idea.
Chen & Gatmiry, 2023, *When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?*

Relates to optimization methods with momentum, e.g.
Maddison, Paulin, Teh, O'Donoghue & Doucet, 2018, *Hamiltonian Descent Methods*.

Visualizations

The Markov Chain Monte Carlo Interactive Gallery:
`https://chi-feng.github.io/mcmc-demo/`

HarlMCMC shake:
`https://www.youtube.com/watch?v=Vv3f0QNWvWQ`

Statistical Rethinking:
`https://www.youtube.com/watch?v=rZk2FqX2XnY`

Outline

Outstanding questions

- Stopping criteria. Seems much harder than in optimization. Number of iterations reported in the literature spans many orders of magnitude (dozens, millions, trillions).

- Theoretical results provide insight on the performance of MCMC but rarely precise guidelines for practitioners.

- Since MCMC methods are iterative, they are not obvious to parallelize. Since $\sim$20 years computing power increases mostly through parallelization.

Coming soon. . .

- Total variation distance, Wasserstein distance, and probability couplings.
- Examples of theoretical results about MCMC methods: convergence of marginals and central limit theorems.
- Theoretical devices: couplings of Markov chains, and the Poisson equation.
- And after that: concrete methods derived from them.

Outline

Outline

Outline