Optimization Methods III. The Markov Chain Monte Carlo.

Anatoli Iambartsev
IME-USP

[D] A Brief Introduction to Markov Chains.

“Let $\mathcal{X}$ be a finite set. A *Markov chain* is defined by a matrix $K = (K(x,y), x, y \in \mathcal{X})$ with $K(x,y) \geq 0$ and $\sum_y K(x,y) = 1$ for each x. Thus each row is a probability measure so K can direct a kind of random walk: from x, choose y with probability $K(x,y)$, and so on. We refer to the outcomes

$$X_0 = x, X_1 = y, X_2 = z, \ldots$$

as a run of the chain starting at x. From the definitions

$$\mathbb{P}(X_1 = y \mid X_0 = x) = K(x,y),\ \mathbb{P}(X_2 = z, X_1 = y \mid X_0 = x) = K(x,y)K(y,z)$$

From this,

$$\mathbb{P}(X_2 = z \mid X_0 = x) = \sum_{y \in \mathcal{X}} K(x,y)K(y,z) = K^2(x,z).$$

The n-th power of the matrix has x, y entry $K^n(x,y) = \mathbb{P}(X_n = y \mid X_0 = x)$.”

[D] A Brief Introduction to Markov Chains.

"All of the Markov chains considered here have *stationary distributions* $\pi = (\pi(x), x \in \mathcal{X}), \sum_x \pi(x) = 1$ with

$$\sum_x \pi(x)K(x,y) = \pi(y). \tag{1}$$

Thus π is a left eigenvector of K with eigenvalue 1. The probabilistic interpretation of (1) is "pick x from π and take a step from $K(x,y)$; the chance of being at y is $\pi(y)$." Thus π is stationary for the evolution."

[D] A Brief Introduction to Markov Chains.

“The fundamental theorem of Markov chains (a simple corollary of the Perron-Frobenius theorem) says, under a simple connectedness condition, π is unique and high powers of K converge to the rank one matrix with all rows equal to π.

Theorem 1. Let $\mathcal{X}$ be a finite set and $K(x,y)$ a Markov chain indexed by $\mathcal{X}$. If there is n_0 so that $K^n(x,y) \geq 0$ for all $n > n_0$, then K has a unique stationary distribution π and, as $n \to \infty$,

$$K^n(x,y) \to \pi(y), \quad \text{for each } x,y \in \mathcal{X}.$$

The probabilistic content of the theorem is that from any starting state x, the n-th step of a run of the Markov chain has chance close to $\pi(y)$ of being at y if n is large. In computational settings, $|X|$ is large, it is easy to move from x to y according to $K(x,y)$ and it is hard to sample from π directly.”

[D] A Brief Introduction to Markov Chains.

Markov chain theory:

Given K find their invariant measure π

Markov Chain Monte Carlo theory:

Given π find Markov chain K with invariant measure is π

[D] A Brief Introduction to Markov Chains.

Theorem 2. Let K be irreducible and aperiodic Markov chain. Let the measure π satisfy

$$\pi(x)K(x,y) = \pi(y)K(y,x), \text{ for any } x \neq y \in \mathcal{X}.$$

Then, the chain is called *reversible* and the measure π is invariant.

[D] A Brief Introduction to Markov Chains. Convergence.

“A basic problem of Markov chain theory concerns the rate of convergence in $K^n(x,y) \to \pi(y)$. How long must the chain be run to be suitably close to π? It is customary to measure distances between two probabilities by total variation distance

$$\|K_x^n - \pi\|_{TV} = \frac{1}{2}\sum_y |K^n(x,y) - \pi(y)| = \max_{A\subseteq\mathcal{X}} |K^n(x,A) - \pi(A)|.$$

This yields the math problem: Given K, π, x and $\varepsilon > 0$, how large n so

$$\|K_x^n - \pi\|_{TV} < \varepsilon?$$

”

[D] A Brief Introduction to Markov Chains. Convergence.

“Suppose that the Markov chain is reversible: $\pi(x)K(x,y) = \pi(y)K(y,x)$. Let $L^2(\pi)$ be $\{g : \mathcal{X} \to \mathbb{R}\}$ with inner product $\langle g,h\rangle = \sum_x g(x)h(x)\pi(x)$. Then K operates on L^2 by $Kg(x) = \sum_y g(y)K(x,y)$. Reversibility implies $\langle Kg,h\rangle = \langle g,Kh\rangle$, so K is self-adjoint. Now, the spectral theorem says there is an orthonormal basis of eigenvectors ψ_i and eigenvalues β_i (so $K\psi_i = \beta_i\psi_i$) for $0 \leq i \leq |\mathcal{X}|-1$ and $1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{|\mathcal{X}|-1} \geq -1$. By elementary manipulations

$$K(x,y) = \pi(y)\sum_{i=0}^{|\mathcal{X}|-1} \beta_i\psi_i(x)\psi_i(y),\; K^n(x,y) = \pi(y)\sum_{i=0}^{|\mathcal{X}|-1} \beta_i^n\psi_i(x)\psi_i(y).$$

”

[D] A Brief Introduction to Markov Chains. Convergence.

"Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
4\|K_x^n - \pi\|_{TV}^2 &= \sum_y |K^n(x,y) - \pi(y)| \\
&= \sum_y \frac{|K^n(x,y) - \pi(y)|}{\sqrt{\pi(y)}} \sqrt{\pi(y)} \\
&\leq \sum_y \frac{(K^n(x,y) - \pi(y))^2}{\pi(y)} = \sum_{i=1}^{|\mathcal{X}|-1} \beta_i^{2n} \psi_i^2(x).
\end{aligned}$$

This bound is the basic eigenvalue bound used to get rates of convergence for the examples presented here." Observe that the maximal eigenvalue $\beta_0 \equiv 1$ is missing in the last sum, and the eigenvector corresponding to β_0 can be chosen as identical $\psi_0(\cdot) \equiv 1$, i.e. $\psi_0 = (1, 1, \ldots, 1)$.

[D] General state space.

“If (X,B) is a measurable space, a Markov kernel $K(x, dy)$ is a probability measure $K(x, \cdot)$ for each x. Iterates of the kernel are given by, e.g.,

$$K^2(x, A) = \int K(z, A)K(x, dz).$$

A *stationary distribution* is a probability $\pi(dx)$ satisfying

$$\pi(A) = \int K(x, A)\pi(dx)$$

under simple conditions $K^n(x, A) \to \pi(A)$ and exactly the same problems arise.

[D] Metropolis Algorithm.

Let $\mathcal{X}$ be a finite state space and $\pi(x)$ a probability on $\mathcal{X}$ (perhaps specified only up to an unknown normalizing constant). Let $J(x,y)$ be a Markov matrix on $\mathcal{X}$ with $J(x,y) > 0 \leftrightarrow J(y,x) > 0$. At the start, J is unrelated to π. The Metropolis algorithm changes J to a new Markov matrix $K(x,y)$ with stationary distribution π. It is given by a simple recipe:

$$K(x,y) = \begin{cases} J(x,y), & \text{if } x \neq y, A(x,y) \geq 1 \\ J(x,y)A(x,y), & \text{if } x \neq y, A(x,y) < 1 \\ J(x,y) + \sum_{z:A(x,z)<1} J(x,z)(1 - A(x,z)), & \text{if } x = y, \end{cases}$$

where A is the acceptance ratio $A(x,y) = \pi(y)J(y,x)/\pi(x)J(x,y)$.

[D] Metropolis Algorithm.

$$K(x,y) = \begin{cases} J(x,y), & \text{if } x \neq y, A(x,y) \geq 1 \\ J(x,y)A(x,y), & \text{if } x \neq y, A(x,y) < 1 \\ J(x,y) + \sum_{z:A(x,z)<1} J(x,z)(1 - A(x,z)), & \text{if } x = y, \end{cases}$$

the acceptance ratio $A(x,y) = \pi(y)J(y,x)/\pi(x)J(x,y)$.

The formula has a simple interpretation: from x, choose y with probability $J(x,y)$; if $A(x,y) \geq 1$, move to y; if $A(x,y) < 1$, flip a coin with this success probability and move to y if success occurs; in other cases, stay at x. Note that the normalizing constant cancels out in all calculations. The new chain satisfies

$$\pi(x)K(x,y) = \pi(y)K(y,x).$$

[D] Metropolis Algorithm.

Thus

$$\sum_x \pi(x)K(x,y) = \sum_x \pi(y)K(y,x) = \pi(y).$$

so that π is a left eigenvector with eigenvalue 1. If the chain K is connected, Theorem 1 is in force:

$$K^n(x,y) \to \pi(y), \text{ as } n \to \infty.$$

After many steps of the chain, the chance of being at y is approximately $\pi(y)$, no matter what the starting state $\mathcal{X}$.

Metropolis Algorithm.

Summarizing in algorithmic form initialized with the (arbitrary) value $x^{(0)}$: given $x^{(t)}$

1. Generate $Y_t \sim J(x^{(t)}, \cdot)$.

2. Take

$$x^{(t+1)} = \begin{cases} Y_t \text{ with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} \text{ with probability } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

   where

$$\rho(x, y) = \min\{A(x, y), 1\} = \min\left\{\frac{\pi(y)}{\pi(x)} \frac{J(y, x)}{J(x, y)}, 1\right\}.$$

[P] Metropolis Algorithm. Remark.

A very important feature of the Metropolis chain is that it only depends on the ratios $\pi(x)/\pi(y)$. Frequently $\pi(x)$ is only be explicitly known up to a normalizing constant. The optimization chains described below are examples of this type. The normalizing constant is not needed to run the Metropolis chain.

Metropolis Algorithm. Example: symmetric group.

Let $\mathcal{X} = S_n$, the symmetric group on n letters. Define a probability measure on S_n by

$$\pi(\sigma) = \frac{1}{Z}\theta^{d(\sigma,\sigma_0)}, \ 0 < \theta \leq 1.$$

where $d(\sigma, \sigma_0)$ is a metric on symmetric group defined as a minimum number of transpositions required to bring σ to σ_0. A transposition is a permutation which exchanges two elements and keeps all others fixed; for example (13) is a transposition. Every permutation can be written as a product of transpositions; for instance, the permutation $g = (125)(34)$ from above can be written as $g = (12)(25)(34)$.

The normalizing constant can be calculated explicitly

$$Z = \sum_{\sigma} \theta^{d(\sigma,\sigma_0)} = \prod_{i=1}^{n} \Big(1 + \theta(i-1)\Big).$$

Note that if $\theta = 1$, $\pi(\sigma)$ is the uniform distribution on S_n. For $\theta < 1, \pi(\sigma)$ is largest at σ_0 and falls off from its maximum as σ moves away from σ_0. It serves as a natural non-uniform distribution on S_n, peaked at a point.

Metropolis Algorithm. Example: symmetric group.

How can samples be drawn from $\pi(\sigma), \sigma \in S_n$?

Metropolis Algorithm. Example: symmetric group.

One route is to use the Metropolis algorithm, based on random transpositions. Thus, from σ, choose a transposition (i,j) uniformly at random and consider $(i,j)\sigma = \sigma^*$. If $d(\sigma^*,\sigma_0) \leq d(\sigma,\sigma_0)$ the chain moves to σ^*. If $d(\sigma^*,\sigma_0) > d(\sigma,\sigma_0)$, flip a θ-coin. If this comes up heads, move to σ^*; else stay at σ. In symbols,

$$K(\sigma,\sigma^*) = \begin{cases} 1/\binom{n}{2}, & \text{if } \sigma^* = (i,j)\sigma,\ d(\sigma^*,\sigma_0) < d(\sigma,\sigma_0) \\ \theta/\binom{n}{2}, & \text{if } \sigma^* = (i,j)\sigma,\ d(\sigma^*,\sigma_0) > d(\sigma,\sigma_0) \\ c(1-\theta/\binom{n}{2}), & \text{if } \sigma^* = \sigma,\ \text{with} \\ & c = \#\{(i,j) : d((i,j)\sigma,\sigma_0) > d(\sigma,\sigma_0) \\ 0, & \text{otherwise.} \end{cases}$$

Observe that this Markov chain is "easy to run". The Metropolis construction guarantees that:

$$\pi(\sigma)K(\sigma,\sigma^*) = \pi(\sigma^*)K(\sigma^*,\sigma)$$

so that the chain has stationary distribution π.

Metropolis Algorithm. Example: symmetric group.

When $n = 3$ and $\sigma_0 = id$, the transition matrix is

	id	(12)	(13)	(23)	(123)	(132)
id	$1-\theta$	$\frac{\theta}{3}$	$\frac{\theta}{3}$	$\frac{\theta}{3}$	0	0
(12)	$\frac{1}{3}$	$\frac{2}{3}(1-\theta)$	0	0	$\frac{\theta}{3}$	$\frac{\theta}{3}$
(13)	$\frac{1}{3}$	0	$\frac{2}{3}(1-\theta)$	0	$\frac{\theta}{3}$	$\frac{\theta}{3}$
(23)	0	0	$\frac{2}{3}(1-\theta)$	$\frac{\theta}{3}$	$\frac{\theta}{3}$	
(123)	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
(132)	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0

Remember that permutation, say (12) we read as: $1 \to 2$ and $2 \to 1$.

The stationary distribution is the left eigenvector proportional to $(1, \theta, \theta, \theta, \theta^2, \theta^2)$.

Metropolis Algorithm. Example: Ising Model.

The Ising model was invented by Wilhelm Lenz and developed by his student Ernst Ising. It is used for modeling ferromagnetic and anti-ferromagnetic materials. The model represents a lattice occupied by atoms which can each have dipole magnetic moments (called spins). The model predicts a second order phase transition occurring at the Curie temperature for dimensions higher than 1. Phase transition is identified from ensemble properties and compared with the theoretical model which has been solved exactly for zero external field.

Metropolis Algorithm. Example: Ising Model.

Each atom $i \in \Lambda$ can adopt two states, corresponding to $\sigma_i \in \{-1, 1\}$, where σ_v represents the spin and the spin interactions are dependent on the coupling parameter J_{ij}. The lattice model has periodic boundary conditions and extends infinitely. The Hamiltonian is defined as below: let $\sigma \in \{-1, 1\}^\Lambda$

$$H(\sigma) = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i,$$

where J_{ij} is coupling parameter between neighbors atoms i and j; h is external field strength.

Metropolis Algorithm. Example: Ising Model.

The probability measure on $\{-1,1\}^{\Lambda}$ is defined as follows: let $\sigma \in \{-1,1\}^{\Lambda}$, then

$$\pi_{\beta,\Lambda}(\sigma) = \frac{1}{Z_{\beta,\Lambda}} e^{-\beta H(\sigma)},$$

where β is inverse temperature, and $Z_{\beta,\Lambda}$ the normalized constant which called *partition function*.

$$Z_{\beta,\Lambda} = \sum_{\sigma} e^{-\beta H(\sigma)}.$$

The problem: how to sample configurations σ from distribution $\pi_{\beta,\Lambda}$.

Metropolis Algorithm. Example: Ising Model.

Metropolis Algorithm:

1. Initialize the system randomly with spins, at a given temperature (fixed β);

2. set the value of the external field (in most cases $h = 0$);

3. make a random flip in the spin of some atom;

4. compute the energy change ΔH arising from this, due to only the neighboring atoms;

5. ensure that the periodic boundary conditions are in place to take care of edge effects;

6. if $\Delta H < 0$, accept this configuration and continue this process;

7. If $\Delta H > 0$, accept this configuration with a probability of $p = e^{-\beta \Delta H}$, else retain the old configuration.

References.

[D] Diaconis, Persi. *The Markov Chain Monte Carlo Revolution*.

[P] Levin D.A., Peres Y., Wilmer E.L. *Markov Chains and Mixing Times.* 2007.

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series "Use R!". Springer

Further reading.

Chib S., Greenberg, E. *Understanding the Metropolis-Hasting Algorithm.* American Stat. Ass., 1995, Vol. 49, No 4.

Biller L.J., Diaconis P. *A Geometric Interpretation of the Metropolis-Hasting Algorithm.* Statistical Science, Vol. 16, No. 4, pp. 335–339

Diaconis, P., Saloff-Coste, L. *What Do We Know about the Metropolis Algorithm?* J. of Computer and System Sciences, **57**, 20–36, 1998