MCMC: The Metropolis Algorithm

Andrei Nenciu

18 Feb. 2020

The Oxford University Invariant Society

Why MCMC?

from *SIAM News*, Volume 33, Number 4

The Best of the 20th Century: Editors Name Top 10 Algorithms

Why MCMC?

- physical sciences
- engineering
- computational biology
- computer graphics
- machine learning ...

Table of contents

Motivation: Bayesian Statistics

- Let $X \sim \mathrm{Bernoulli}(\theta)$, for some unknown $\theta \in [0, 1]$.
- Draw n i.i.d. samples $X_1, X_2, ..., X_m \sim X$.
- What can we say about θ?

The Bayesian Approach

1. Treat θ like a random variable over $\Theta := \mathbb{R}$ and assign it a distribution $P(\theta)$, called the prior.
2. Use the data $X_1, X_2, \ldots, X_m$ and Bayes' rule to update the prior:

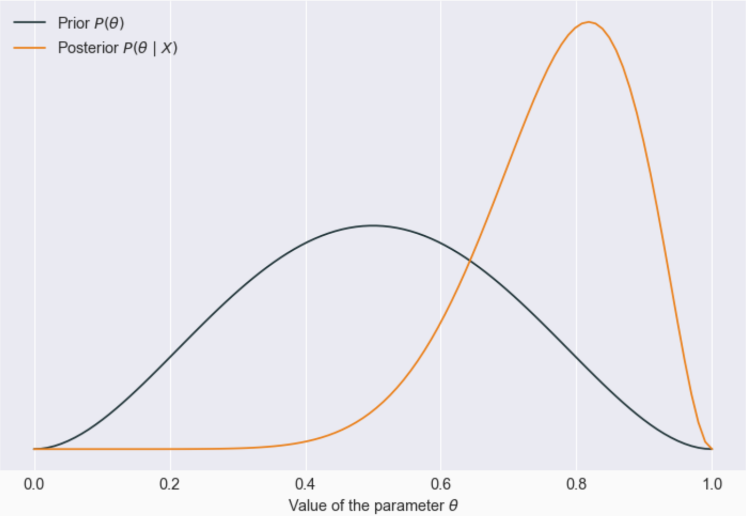
   $$\mu(\theta) := P(\theta \mid X) \propto P(\theta)P(X \mid \theta).$$

   The resulting distribution $\mu(\theta)$ is called the posterior distribution.

The Bayesian Approach

1. Treat θ like a random variable over $\Theta := \mathbb{R}$ and assign it a distribution $P(\theta)$, called the prior.
2. Use the data $X_1, X_2, \ldots, X_m$ and Bayes' rule to update the prior:
$$\mu(\theta) := P(\theta \mid X) \propto P(\theta)P(X \mid \theta).$$
The resulting distribution $\mu(\theta)$ is called the posterior distribution.

Example of Prior and Posterior



- To compute average of μ:

$$\int_{\Theta} \theta \mu(\theta) d\theta$$

- To compute the variance of μ, also need:

$$\int_{\Theta} \theta^2 \mu(\theta) d\theta$$

- In general:

$$I[f] := \mathbb{E}_{\theta \sim \mu}[f(\theta)]$$

  for a suitable $f : \Theta \to \Theta$.

- To compute average of μ:

$$\int_\Theta \theta\mu(\theta)d\theta$$

- To compute the variance of μ, also need:

$$\int_\Theta \theta^2\mu(\theta)d\theta$$

- In general:

$$I[f] := \mathbb{E}_{\theta\sim\mu}[f(\theta)]$$

for a suitable $f : \Theta \to \Theta$.

- To compute average of μ:

$$\int_\Theta \theta \mu(\theta) d\theta$$

- To compute the variance of μ, also need:

$$\int_\Theta \theta^2 \mu(\theta) d\theta$$

- In general:

$$I[f] := \mathbb{E}_{\theta \sim \mu}[f(\theta)]$$

for a suitable $f : \Theta \to \Theta$.

Reality Check

- Real-life Bayesian models can have tens, hundreds or even thousands of parameters, so $\Theta \sim \mathbb{R}^d$ for large d. In general, the integral:

$$I[f] := \mathbb{E}_{\theta\sim\mu}[f(\theta)]$$

  will be intractable.
- Numerical methods?

Reality Check

- Real-life Bayesian models can have tens, hundreds or even thousands of parameters, so $\Theta \sim \mathbb{R}^d$ for large d. In general, the integral:

$$I[f] := \mathbb{E}_{\theta \sim \mu}[f(\theta)]$$

  will be intractable.
- Numerical methods?

The Monte Carlo Method

1. Simulate n samples $\theta_1, \theta_2, ..., \theta_n \sim \mu$.
2. Approximate $I[f] := \mathbb{E}_{\theta\sim\mu}[f(\theta)]$ by the empirical estimate:

$$I_n[f] = \frac{1}{n}\sum_{i=1}^{N} f(\theta_i).$$

3. By the Strong Law of Large Numbers, almost surely:

$$I_n[f] \to I[f].$$

4. Under mild assumptions, the Central Limit Theorem gives:

$$I[f] - I_n[f] \propto n^{-\frac{1}{2}}$$

The Monte Carlo Method

1. Simulate n samples $\theta_1, \theta_2, ..., \theta_n \sim \mu$.
2. Approximate $I[f] := \mathbb{E}_{\theta \sim \mu}[f(\theta)]$ by the empirical estimate:

$$I_n[f] = \frac{1}{n} \sum_{i=1}^{N} f(\theta_i).$$

3. By the Strong Law of Large Numbers, almost surely:

$$I_n[f] \to I[f].$$

4. Under mild assumptions, the Central Limit Theorem gives:

$$I[f] - I_n[f] \propto n^{-\frac{1}{2}}$$

The Monte Carlo Method

1. Simulate n samples $\theta_1, \theta_2, ..., \theta_n \sim \mu$.
2. Approximate $I[f] := \mathbb{E}_{\theta \sim \mu}[f(\theta)]$ by the empirical estimate:

$$I_n[f] = \frac{1}{n} \sum_{i=1}^{N} f(\theta_i).$$

3. By the Strong Law of Large Numbers, almost surely:

$$I_n[f] \to I[f].$$

4. Under mild assumptions, the Central Limit Theorem gives:

$$I[f] - I_n[f] \propto n^{-\frac{1}{2}}$$

The Monte Carlo Method

1. Simulate n samples $\theta_1, \theta_2, ..., \theta_n \sim \mu$.
2. Approximate $I[f] := \mathbb{E}_{\theta\sim\mu}[f(\theta)]$ by the empirical estimate:

$$I_n[f] = \frac{1}{n}\sum_{i=1}^{N} f(\theta_i).$$

3. By the Strong Law of Large Numbers, almost surely:

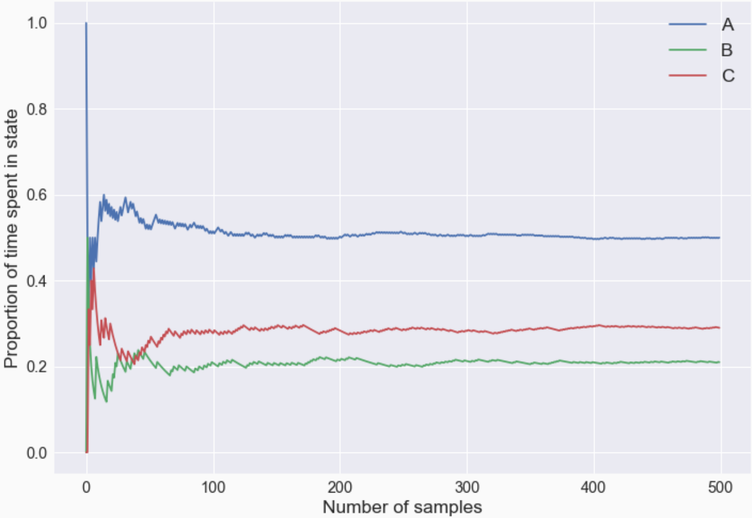
$$I_n[f] \to I[f].$$

4. Under mild assumptions, the Central Limit Theorem gives:

$$I[f] - I_n[f] \propto n^{-\frac{1}{2}}$$

Markov Chains

Proportion of Time Spent in Each State



As $n \to \infty$, the set $\{\theta_1, \theta_2, ..., \theta_n\}$ looks like a set of samples from the limiting distribution P of the Markov chain $(\{A, B, C\}, K)$:

$$P(A) = 0.5$$

$$P(B) = 0.2$$

$$P(C) = 0.3$$

.

Let μ be a distribution on $\Theta = R^d$. We want to sample from μ.

1. Construct a Markov Chain (Θ, K) with μ as its limiting distribution.
2. Run the Markov chain for n steps, to obtain $\{\theta_1, \theta_2, ..., \theta_n\}$.
3. Use the empirical estimate, as in the standard Monte Carlo setup:

$$I_n[f] = \frac{1}{n} \sum_{i=1}^{N} f(\theta_i).$$

Let μ be a distribution on $\Theta = R^d$. We want to sample from μ.

1. Construct a Markov Chain (Θ, K) with μ as its limiting distribution.
2. Run the Markov chain for n steps, to obtain $\{\theta_1, \theta_2, ..., \theta_n\}$.
3. Use the empirical estimate, as in the standard Monte Carlo setup:

$$I_n[f] = \frac{1}{n} \sum_{i=1}^{N} f(\theta_i).$$

The Markov Chain Monte Carlo Method

Let μ be a distribution on $\Theta = R^d$. We want to sample from μ.

1. Construct a Markov Chain (Θ, K) with μ as its limiting distribution.
2. Run the Markov chain for n steps, to obtain $\{\theta_1, \theta_2, ..., \theta_n\}$.
3. Use the empirical estimate, as in the standard Monte Carlo setup:

$$I_n[f] = \frac{1}{n} \sum_{i=1}^{N} f(\theta_i).$$

The Metropolis Algorithm

- Given a distribution μ be a distribution on $\Theta = R^d$, we want to construct a Markov Chain on Θ with invariant distribution μ.
- Need a transition rule (i.e. a kernel K).

The Metropolis Kernel (Metropolis et al., aprox. 1950)

Say we are currently at $\theta_n \in \Theta$.

- Sample a proposal step: $\theta_p \sim N(\theta_n, \sigma I)$.
- Compute the acceptance probability:

$$p_{acc} = min\left(1, \frac{\mu(\theta_p)}{\mu(\theta_n)}\right)$$

- Accept the proposal θ_p with probability p_{acc}:

$$\theta_{n+1} = \begin{cases} \theta_p, & \text{with probability } p_{acc} \\ \theta_n, & \text{with probability } 1 - p_{acc} \end{cases}$$

The Metropolis Kernel (Metropolis et al., aprox. 1950)

Say we are currently at $\theta_n \in \Theta$.

- Sample a proposal step: $\theta_p \sim N(\theta_n, \sigma I)$.
- Compute the acceptance probability:

$$p_{acc} = min\left(1, \frac{\mu(\theta_p)}{\mu(\theta_n)}\right)$$

- Accept the proposal θ_p with probability p_{acc}:

$$\theta_{n+1} = \begin{cases} \theta_p, & \text{with probability } p_{acc} \\ \theta_n, & \text{with probability } 1 - p_{acc} \end{cases}$$

The Metropolis Kernel (Metropolis et al., aprox. 1950)

Say we are currently at $\theta_n \in \Theta$.

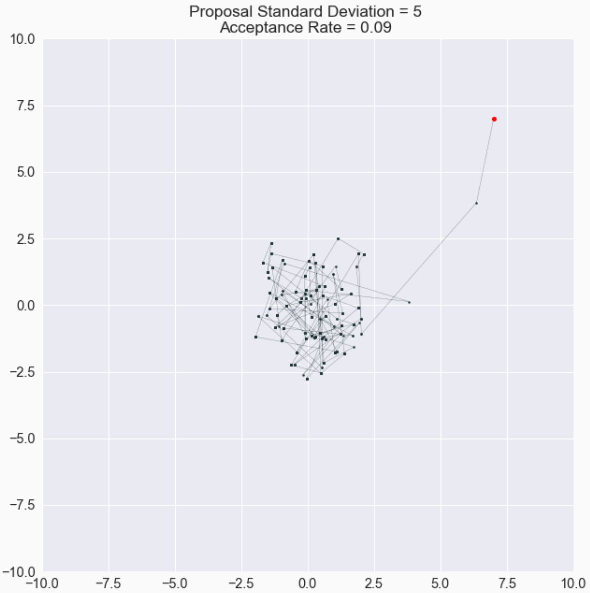
- Sample a proposal step: $\theta_p \sim N(\theta_n, \sigma I)$.
- Compute the acceptance probability:

$$p_{acc} = min\left(1, \frac{\mu(\theta_p)}{\mu(\theta_n)}\right)$$

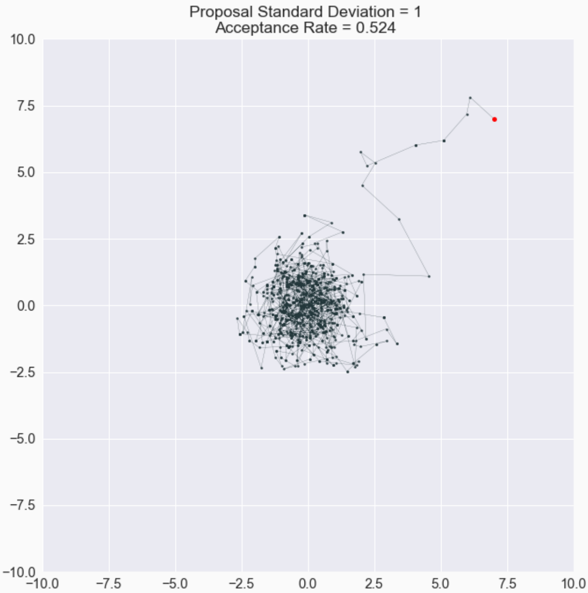
- Accept the proposal θ_p with probability p_{acc}:

$$\theta_{n+1} = \begin{cases} \theta_p, & \text{with probability } p_{acc} \\ \theta_n, & \text{with probability } 1 - p_{acc} \end{cases}$$

Metropolis for $\mu = N(0, I_2)$



Metropolis for $\mu = N(0, I_2)$



Metropolis for $\mu = N(0, I_2)$