Introduction to Markov chain Monte Carlo

Michael Choi

The Chinese University of Hong Kong, Shenzhen
Institute for Data and Decision Analytics (iDDA)

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

数据运筹科学研究院
Institute for Data and Decision Analytics

May 2019

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where

$$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where
  $$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$
- **Difficulty**: At times it is impossible to apply classical Monte Carlo methods, since π is often of the form
  $$\pi(x) = \frac{e^{-\beta H(x)}}{Z},$$
  where Z is a normalization constant that cannot be computed.

Introduction

- Let π be a discrete or continuous distribution.
  Goal: Sample from π or estimate $\pi(f)$, where
  $$\pi(f) = \sum_x f(x)\pi(x), \quad \text{or} \quad \pi(f) = \int f(x)\pi(dx).$$
- **Difficulty**: At times it is impossible to apply classical Monte Carlo methods, since π is often of the form
  $$\pi(x) = \frac{e^{-\beta H(x)}}{Z},$$
  where Z is a normalization constant that cannot be computed.
- **Idea of Markov chain Monte Carlo (MCMC)**: Construct a Markov chain that converges to π, which only depends on the ratio
  $$\frac{\pi(y)}{\pi(x)}.$$
  Thus there is no need to know Z.

Motivation from Bayesian statistics

- Suppose that we have a statistical model on the parameter θ, and we observe data $\mathbf{x} = (x_i)_{i=1}^n$ generated from this model.
  Likelihood function of $\mathbf{x}$ given θ: $L(\theta|\mathbf{x})$.
  Prior distribution of θ: $f(\theta)$.

Motivation from Bayesian statistics

- Suppose that we have a statistical model on the parameter θ, and we observe data $\mathbf{x} = (x_i)_{i=1}^n$ generated from this model.
  Likelihood function of $\mathbf{x}$ given θ: $L(\theta|\mathbf{x})$.
  Prior distribution of θ: $f(\theta)$.
- By Bayes theorem, the **posterior distribution** of θ given $\mathbf{x}$ is
  $$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})f(\theta)}{\int L(\theta|\mathbf{x})f(\theta)\,d\theta},$$
  where the integral is often impossible to calculate.

Motivation from Bayesian statistics

- Suppose that we have a statistical model on the parameter θ, and we observe data $\mathbf{x} = (x_i)_{i=1}^n$ generated from this model.
  Likelihood function of $\mathbf{x}$ given θ: $L(\theta|\mathbf{x})$.
  Prior distribution of θ: $f(\theta)$.
- By Bayes theorem, the **posterior distribution** of θ given $\mathbf{x}$ is
  $$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})f(\theta)}{\int L(\theta|\mathbf{x})f(\theta)\,d\theta},$$
  where the integral is often impossible to calculate.
- To conduct Bayesian inference, we need to sample from $\pi(\theta|\mathbf{x})$ or estimate $\pi(f)$ (e.g. the posterior mean). MCMC is thus very useful in the Bayesian statistics community.

The Metropolis-Hastings algorithm

- Two ingredients:
  (i). Target distribution: π
  (ii). Proposal chain with transition matrix $Q = (Q(x,y))_{x,y}$.

The Metropolis-Hastings algorithm

Algorithm 1: The Metropolis-Hastings algorithm

Input: Proposal chain Q, target distribution π

1 Given X_n, generate $Y_n \sim Q(X_n, \cdot)$

2 Take

$$X_{n+1} = \begin{cases} Y_n, & \text{with probability } \alpha(X_n, Y_n), \\ X_n, & \text{with probability } 1 - \alpha(X_n, Y_n), \end{cases}$$

where

$$\alpha(x, y) := \min \left\{ \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1 \right\}$$

is known as the acceptance probability.

The Metropolis-Hastings algorithm

Definition

The Metropolis-Hastings algorithm, with proposal chain Q and target distribution π, is a Markov chain $X = (X_n)_{n \geq 1}$ with transition matrix

$$P(x,y) = \begin{cases} \alpha(x,y)Q(x,y), & \text{for } x \neq y, \\ 1 - \sum_{y;\ y \neq x} P(x,y), & \text{for } x = y. \end{cases}$$

The Metropolis-Hastings (MH) algorithm

Theorem

Given target distribution π and proposal chain Q, the Metropolis-Hastings chain is

- ***reversible***, *that is, for all x, y,*

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

- *(Ergodic theorem of MH) If P is irreducible, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \pi(f).$$

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x,y) = Q(y,x)$, and so the acceptance probability is

$$\alpha(x,y) = \min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x,y) = Q(y,x)$, and so the acceptance probability is

$$\alpha(x,y) = \min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

- **Random walk MH**: We take a random walk proposal chain with $Q(x,y) = Q(y-x)$. E.g., $Q(x,\cdot)$ is the probability density function of $N(x,\sigma^2)$.

The Metropolis-Hastings algorithm

- Different choices of Q give rise to different MH algorithms
- **Symmetric MH**: We take a symmetric proposal chain with $Q(x,y) = Q(y,x)$, and so the acceptance probability is

$$\alpha(x,y) = \min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

- **Random walk MH**: We take a random walk proposal chain with $Q(x,y) = Q(y-x)$. E.g., $Q(x,\cdot)$ is the probability density function of $N(x,\sigma^2)$.
- **Independence sampler**: Here we take $Q(x,y) = q(y)$, where $q(y)$ is a probability distribution. In words, $Q(x,y)$ does not depend on x.

Example 1: logistic regression

- We observe $(x_i, y_i)_{i=1}^n$ according to the model

$$Y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

Example 1: logistic regression

- We observe $(x_i, y_i)_{i=1}^n$ according to the model

$$Y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

- The likelihood function is

$$L(\alpha, \beta | \mathbf{x}, \mathbf{y}) \propto \prod_{i=1}^n \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right)^{y_i} \left(\frac{1}{1+e^{\alpha+\beta x_i}} \right)^{1-y_i},$$

and prior distribution

$$\pi_\alpha(\alpha|\hat{b})\pi_\beta(\beta) = \frac{1}{\hat{b}} e^\alpha e^{-e^\alpha/\hat{b}},$$

i.e. exponential prior on $\log \alpha$ and a flat prior on β. $\hat{b}$ is chosen such that $\mathbb{E}(\alpha) = \hat{\alpha}$, where $\hat{\alpha}$ is the MLE of α.

Example 1: logistic regression

- We observe $(x_i, y_i)_{i=1}^n$ according to the model

$$Y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

- The likelihood function is

$$L(\alpha, \beta|\mathbf{x}, \mathbf{y}) \propto \prod_{i=1}^n \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\alpha+\beta x_i}}\right)^{1-y_i},$$

  and prior distribution

$$\pi_\alpha(\alpha|\hat{b})\pi_\beta(\beta) = \frac{1}{\hat{b}} e^\alpha e^{-e^\alpha/\hat{b}},$$

  i.e. exponential prior on $\log \alpha$ and a flat prior on β. $\hat{b}$ is chosen such that $\mathbb{E}(\alpha) = \hat{\alpha}$, where $\hat{\alpha}$ is the MLE of α.
- Goal: sample from the posterior of (α, β) using the MH algorithm

Example 1: logistic regression

- **Choosing a good Q to accelerate convergence**: Let $\hat{\alpha}$ and $\hat{\beta}$ be the MLE of α and β respectively, and $\widehat{\sigma^2_{\hat{\beta}}}$ be the variance of $\hat{\beta}$.

Example 1: logistic regression

- **Choosing a good Q to accelerate convergence**: Let $\hat{\alpha}$ and $\hat{\beta}$ be the MLE of α and β respectively, and $\widehat{\sigma^2_{\hat{\beta}}}$ be the variance of $\hat{\beta}$.
- We take an independent MH with proposal chain

$$f(\alpha, \beta) = \pi_\alpha(\alpha|\hat{b})\phi(\beta),$$

where $\phi(\beta)$ is the pdf of normal distribution with mean $\hat{\beta}$ and variance $\widehat{\sigma^2_{\hat{\beta}}}$.

Example 1: logistic regression

Algorithm 2: Independent MH on logistic regression

1 Given (α_n, β_n), generate $(\alpha^{'}, \beta^{'}) \sim f(\alpha, \beta)$, that is, generate $\log \alpha^{'}$ following exponential distribution with parameter $\hat{b}$, and $\beta^{'} \sim N(\hat{\beta}, \widehat{\sigma^2_{\hat{\beta}}})$.

2 Accept $(\alpha^{'}, \beta^{'})$ with probability

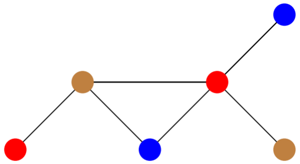
$$\min\left\{\frac{L(\alpha^{'}, \beta^{'}|\mathbf{x}, \mathbf{y})\phi(\beta_n)}{L(\alpha_n, \beta_n|\mathbf{x}, \mathbf{y})\phi(\beta^{'})}, 1\right\}$$

Example 2: graph colouring

- Let $G = (V, E)$ be an undirected graph without self-loop on the vertex set V and edge set E. We want to colour each vertex with one of the q colours such that a vertex's colour differs from that of all its neighbours.

Example 2: graph colouring

- Let $G = (V, E)$ be an undirected graph without self-loop on the vertex set V and edge set E. We want to colour each vertex with one of the q colours such that a vertex's colour differs from that of all its neighbours.



Example 2: graph colouring

- Let S be the set of possible colour configurations on G, and $x = (x_v, v \in V) \in S$ is a particular colour configuration. A **proper q-colouring** of G is any configuration x such that for all $v, w \in V$, if $(v, w) \in E$, then $x_v \neq x_w$.

Example 2: graph colouring

- Let S be the set of possible colour configurations on G, and $x = (x_v, v \in V) \in S$ is a particular colour configuration. A **proper q-colouring** of G is any configuration x such that for all $v, w \in V$, if $(v, w) \in E$, then $x_v \neq x_w$.
- **Goal:** Sample uniformly among the proper q-colourings of G. In other words, we would like to sample from

$$\pi(x) = \frac{\mathbf{1}_{\{\text{x is a proper q-colouring}\}}}{Z}, \quad x \in S,$$

where Z is the number of proper q-colourings of G.

Example 2: graph colouring

- Let S be the set of possible colour configurations on G, and $x = (x_v, v \in V) \in S$ is a particular colour configuration. A **proper q-colouring** of G is any configuration x such that for all $v, w \in V$, if $(v, w) \in E$, then $x_v \neq x_w$.
- **Goal:** Sample uniformly among the proper q-colourings of G. In other words, we would like to sample from

$$\pi(x) = \frac{\mathbf{1}_{\{\text{x is a proper q-colouring}\}}}{Z}, \quad x \in S,$$

  where Z is the number of proper q-colourings of G.
- Computing Z is non-trivial. Using Metropolis-Hastings, we can still sample π without computing Z.

Example 2: graph colouring

Algorithm 3: MH on graph colouring

1 Given a proper q-colouring x

2 Select a vertex $v \in V$ uniformly at random

3 Select a colour $c \in \{1, 2, \dots, q\}$ uniformly at random

4 If c is an allowed colour at v, then recolour v, i.e. set $x_v = c$; do nothing otherwise

5 Repeat step 2 - 4

Example 3: Ising model

- Let $G = (V, E)$ be an undirected graph without self-loop on the vertex set $V = \{1, 2, \dots, N\}$ and edge set E. Variables $\sigma_v \in \{-1, 1\}$ are attached to the vertices $v \in V$. These variables are called spins. The state space is made up of spin assignments $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_N) \in \{-1, 1\}^N$.

Example 3: Ising model

- Let $G = (V, E)$ be an undirected graph without self-loop on the vertex set $V = \{1, 2, \ldots, N\}$ and edge set E. Variables $\sigma_v \in \{-1, 1\}$ are attached to the vertices $v \in V$. These variables are called spins. The state space is made up of spin assignments $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N) \in \{-1, 1\}^N$.
- We would like to sample from the **Gibbs** distribution:

$$\pi(\sigma) = \frac{1}{Z} \exp\left(\sum_{(v,w) \in E} \beta J_{vw} \sigma_v \sigma_w \right),$$

where $\beta > 0$ is the inverse temperature, $J_{vw} \in \mathbb{R}$ is the interaction strength and Z is the normalization constant

$$Z = \sum_{\sigma \in \{-1,1\}^N} \exp\left(\sum_{(v,w) \in E} \beta J_{vw} \sigma_v \sigma_w \right).$$

Example 3: Ising model

Algorithm 4: MH on Ising model

1 Given an initial spin assignment σ

2 Select a vertex $v \in V$ uniformly at random

3 Consider the spin assignment $\sigma^{(v)}$ where the initial spin σ_v is flipped, i.e. $\sigma_v^{(v)} = -\sigma_v$.

4 Accept $\sigma^{(v)}$ with probability

$$\min\left\{\frac{\pi(\sigma^{(v)})}{\pi(\sigma)}, 1\right\} = \min\left\{e^{-\beta 2\sigma_v \sum_w J_{vw}\sigma_w}, 1\right\}$$

; do nothing otherwise.

5 Repeat step 2 - 4

MH as L^1 minimizer

Theorem (Billera and Diaconis '01, Choi and Huang '19)

Given a target distribution π and proposal chain Q on a finite state space, let P be the transition matrix of MH. Then

$$d_\pi(Q, P) = \inf_{K \in R(\pi)} d_\pi(Q, K),$$

where $R(\pi)$ is the set of reversible transition matrix with respect to π, and

$$d_\pi(Q, K) = \sum_x \sum_{y \neq x} \pi(x) |Q(x, y) - K(x, y)|.$$

In words, P minimizes the distance d_π between Q and $R(\pi)$.

The scaling limit of MH is the Langevin diffusion

- Suppose that $U : \mathbb{R}^d \to \mathbb{R}$, and U is continuously differentiable with Lipschitz continuous gradient.

The scaling limit of MH is the Langevin diffusion

- Suppose that $U : \mathbb{R}^d \to \mathbb{R}$, and U is continuously differentiable with Lipschitz continuous gradient.
- Target distribution: Gibbs distribution with density

$$\pi(\mathbf{x}) = \frac{e^{-U(\mathbf{x})/T}}{\int e^{-U(\mathbf{x})/T}\, dx}$$

Proposal chain: Gaussian proposal with $Q_\epsilon(\mathbf{x}, \mathbf{y})$ being the pdf of $N(\mathbf{x}, \epsilon I)$.

The scaling limit of MH is the Langevin diffusion

Theorem (Gelfand and Mitter '91)

Given target distribution π and proposal chain Q_ϵ, let $(X_n^\epsilon)_{n\geq 0}$ be the MH chain. Then

$$X^\epsilon_{\lfloor t/\epsilon \rfloor} \Rightarrow X_t,$$

where $(X_t)_{t\geq 0}$ is a rescaled version of the Langevin diffusion described by the SDE

$$dX_t = -\nabla U(X_t)/2T dt + dW_t,$$

where $(W_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion. In words, the scaled MH chain converges weakly in the Skorokhod topology to a rescaled Langevin diffusion.

Simulated annealing

- **Goal**: Find the global minimizers of a target function U.

Simulated annealing

- **Goal**: Find the global minimizers of a target function U.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to π_∞, which is supported on the set of global minima of U.

Simulated annealing

- **Goal**: Find the global minimizers of a target function U.
- **Idea of simulated annealing**: Construct a **non-homogeneous** Metropolis-Hastings Markov chain that converges to π_∞, which is supported on the set of global minima of U.
- Target distribution: Gibbs distribution $\pi_{T(t)}$ with temperature $T(t)$ that depends on time t

$$\pi_{T(t)}(x) = \frac{e^{-U(x)/T(t)}}{Z_{T(t)}},$$

$$Z_{T(t)} = \sum_x e^{-U(x)/T(t)}.$$

Proposal chain Q: symmetric

Simulated annealing

- The temperature cools down $T(t) \to 0$ as $t \to \infty$, and we expect the Markov chain get "frozen" at the set of global minima U_{min}:

$$\pi_\infty(x) := \lim_{t\to\infty} \pi_{T(t)}(x) = \begin{cases} \dfrac{1}{|U_{min}|}, & \text{for } x \in U_{min}, \\ 0, & \text{for } x \notin U_{min}. \end{cases}$$

$$U_{min} := \{x;\ U(x) \leq U(y) \text{ for all } y\}.$$

Simulated annealing

Algorithm 5: Simulated annealing

Input: Symmetric proposal chain Q, target distribution $\pi_{T(t)}$, temperature schedule $T(t)$

1 Given X_t, generate $Y_t \sim Q(X_t, \cdot)$

2 Take

$$X_{t+1} = \begin{cases} Y_t, & \text{with probability } \alpha_t(X_t, Y_t), \\ X_t, & \text{with probability } 1 - \alpha_t(X_t, Y_t), \end{cases}$$

where

$$\alpha_t(x, y) := \min\left\{\frac{\pi_{T(t)}(y)Q(y,x)}{\pi_{T(t)}(x)Q(x,y)}, 1\right\} = \min\left\{e^{\frac{U(x)-U(y)}{T(t)}}, 1\right\}$$

is the acceptance probability.

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge
- $T(t)$ cannot converge to zero too fast: we can prove that with positive probability the Markov chain may get stuck at local minimum.

Optimal cooling schedule

- The temperature schedule $T(t)$ cannot be too slow: it may take too long for the Markov chain to converge
- $T(t)$ cannot converge to zero too fast: we can prove that with positive probability the Markov chain may get stuck at local minimum.

Theorem (Hajek '88, Holley and Stroock '88)

The Markov chain generated by simulated annealing converges to π_∞ if and only if for any $\epsilon > 0$,

$$T(t) = \frac{c + \epsilon}{\ln(t+1)},$$

where c is known as the optimal hill-climbing constant that depends on the target function U and proposal chain Q.

Other MCMC algorithms

- Glauber dynamics/heat bath algorithm/Gibbs sampler
- Perfect simulation/Coupling from the past
- Hamilitonian Monte Carlo
- Metropolis adjusted Langevin algorithm (MALA)

References

- Roberts, G. O., & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. Probability surveys, 1, 20-71.
- Robert, C., & Casella, G. (2013). Monte Carlo statistical methods. Springer Science & Business Media.
- Hajek, B. (1988). Cooling schedules for optimal annealing. Mathematics of Operations Research, 13(2), 311-329.
- Holley, R., & Stroock, D. (1988). Simulated annealing via Sobolev inequalities. Communications in Mathematical Physics, 115(4), 553-569.
- Billera, L. J., & Diaconis, P. (2001). A geometric interpretation of the Metropolis-Hastings algorithm. Statistical Science, 335-339.
- Gelfand, S. B., & Mitter, S. K. (1991). Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions. Journal of Optimization Theory and Applications, 68(3), 483-498.

Thank you! Question(s)?