Math 459: Lecture 15

Todd Kuffner

Last time

- Pseudo-random number generators.
- Generating random samples using inverse transform sampling.
- Basics of Markov chains.

This week: Metropolis-Hastings, Gibbs sampling, reversible jump, convergence diagnostics

Reminder of Motivation

Except for conjugate Bayesian analysis, computations involving the posterior are often analytically intractable. Imagine if, *even though we don't know the posterior* $p(\theta|y)$, we could still generate an aribitrarily large number of samples $\theta_1^*, \ldots, \theta_m^*$ from $p(\theta|y)$. Then

post. mean $\widehat{E(\theta|y)} = \bar{\theta}^* = m^{-1} \sum_{j=1}^m \theta_j^*$

post. variance $\widehat{\text{Var}(\theta|y)} = (m-1)^{-1} \sum_{j=1}^m (\theta_j^* - \bar{\theta}^*)^2$

density function use histogram or other density estimate from samples

credible intervals compute quantiles of the samples, i.e. to find an estimate of the the 0.05th quantile of the posterior density, find the value q that solves

$$m^{-1} \sum_{j=1}^m I(\theta_j^* \leq q) = 0.05$$

with $I(A)$ the indicator function for the event A.

These are Monte Carlo estimates of the true posterior quantities.

Generating i.i.d. samples

Suppose the target $f(x)$ cannot be sampled from directly.

von Neumann's idea: rejection sampling

Suppose there exists some other density $g(x)$ with the same support as $f(x)$ which is easy and fast to sample from, and there exists some constant M such that $f(x) \leq Mg(x)$ for all x. Then $0 \leq f(x)/Mg(x) \leq 1$ for all x.

1. Generate $Y \sim g(y)$ and generate $U \sim \text{Unif}(0, 1)$.
2. Set $X = Y$ (accept) if

$$U \leq \frac{f(Y)}{Mg(Y)},$$

otherwise go back to step 1 (reject).

Comments

- Both $f(Y)$ and $g(Y)$ are random variables; thus the *ratio* $f(Y)/Mg(Y)$ is a random variable.
- Expected number of iterations of steps 1 and 2 to obtain a draw is M^{-1} $\Rightarrow$ algorithm optimized by setting

$$M = \sup_x \frac{f(x)}{g(x)}$$

- von Neumann showed that if the probability of acceptance is

$$P(U \leq \frac{f(Y)}{Mg(Y)} | Y \leq y) = \frac{f(y)}{Mg(y)}$$

then the algorithm produces random draws from f

More Comments

Rejection sampling also called *accept-reject method*.

Metropolis et al. (1953) allowed for serial dependence in the sequence of sampled values by combining von Neumann's (1951) rejection sampling idea with the theory of Markov chains.

(Review)

A Markov chain $\{X^{(t)}\}$ is a sequence of dependent random variables

$$X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(t)}, \ldots$$

such that the probability distribution of $X^{(t)}$ *given the past variables* depends only on $X^{(t-1)}$.

- this conditional probability distribution is the **transition or Markov kernel** K:

$$X^{(t+1)} | X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(t)} \sim K(X^{(t)}, X^{(t+1)}).$$

Example (Simple Random Walk)

$$X^{(t+1)} = X^{(t)} + \epsilon_t,$$

with $\epsilon_t \sim \mathcal{N}(0, 1)$ independently of $X^{(t)}$.
$\Rightarrow K(X^{(t)}, X^{(t+1)}) \equiv \mathcal{N}(X^{(t)}, 1)$

Stationarity

We often will work with Markov chains which exhibit a very strong property called **stationarity**. This means that there exists a probability distribution f such that if $X^{(t)} \sim f$, then $X^{(t+1)} \sim f$.

- this means the transition kernel and stationary distribution must satisfy

$$\int_{\mathcal{X}} K(x, y) f(x) dx = f(y).$$

- the existence of a stationary distribution imposes a constraint on the kernel called irreducibility

Definition

A Markov chain is said to be **irreducible** if, regardless of the starting value $X^{(0)}$, there is a positive probability to eventually reach any part of the state space.

Recurrence and Ergodicity

Definition

An (irreducible) Markov chain is said to be **recurrent** if the chain returns to any arbitrary part of the state space infinitely many times.

For recurrent chains, the stationary distribution is also a limiting distribution.

- this means the limiting distribution of $X^{(t)}$ is f for almost all initial values $X^{(0)}$
- this property is termed ergodicity

Magical result: if a kernel K produces an *ergodic* Markov chain with stationary distribution f, then generating a chain from this kernel K will eventually produce simulations from f

LLN for Markov Chains

Recall that we justified Monte Carlo methods by saying that a suitable LLN assures that the Monte Carlo estimate (the average of the simulated values) converges to the true mean.

For integrable functions h, the average

$$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \to E_f[h(X)]$$

In this setting, such a result is often called the **Ergodic Theorem**.

Implication: we can learn about the target f with *arbitrary accuracy* by waiting for stationarity to kick in and then taking a large enough sample

Answer to Question

Does recurrence imply irreducibility or vice versa?

- two states i and j are said to *communicate* if one is reachable from the other $i \leftrightarrow j$: an irreducible Markov chain is one for which all states communicate with each other
- let τ_{ii} be the number of time periods (steps of the chain) until the chain returns to state i after starting at state i
- if $P(\tau_{ii} < \infty) = 1$, state i is **recurrent**, and if $P(\tau_{ii} < \infty) < 1$, the state is transient

For an irreducible Markov chain, either all states are recurrent, or all states are transient.

If all states of an irreducible Markov chain are recurrent, then the Markov chain is said to be recurrent. An irreducible Markov chain is *always recurrent* if the state space is finite.

Positive recurrence

A recurrent state i is said to be **positive recurrent** if the expected value of the amount of time to return to state i, given that the chain started in state i, is finite:

$$E(\tau_{ii}) < \infty$$

Otherwise the recurrent state i is called null recurrent.

When all states in an irreducible Markov chain are positive recurrent, the chain is called positive recurrent.

Departure from i.i.d. samples

We like to assume that we can generate independent and identically distributed samples from a density f of interest.

Problem: standard methods for generating i.i.d. samples (inverse transform sampling, importance sampling, etc.) require that we know the target f

Advantages of using Markov chains: (a sequence of *dependent* variables)

- convergence properties of Markov chains can be exploited to make things easier
- minimal requirements on f
- allows for decompositions of high-dimensional sampling problem into sequence of smaller problems

Basic idea of MCMC

Given a target density f:

- build a Markov kernel $K(x, y) = \Pr(X^{(t)} = y | X^{(t-1)} = x)$ with stationary distribution π the same as the target f
- generate a Markov chain $\{X^{(t)}\}$ using the kernel K *such that* the limiting distribution of $\{X^{(t)}\}$ is f and integrals can be approximated using the Ergodic Theorem, i.e. for all integrable functions h, $T^{-1} \sum_{t=1}^{T} h(X^{(t)}) \to E_f[h(X)]$.

To sample from f using MCMC, we need a kernel K with stationary distribution π equal to f.

To ensure this, sufficient to show that K satisfies the **detailed balance** property

$$f(x)K(x,y) = f(y)K(y,x),$$

for all x, y in the state space of the Markov chain.

- this implies

$$\sum_y f(y)K(y,x) = \sum_y f(x)K(x,y) = f(x)\sum_y K(x,y) = f(x)$$

  which shows that $fK = f$
- also need $\{X^{(t)}\}$ to be ergodic so that the stationary distribution π is unique

Potentially hard part: constructing the kernel K that is associated with an *arbitrary* density f

But there *are* methods, which are valid *for any* density f, to derive such kernels

Motivating the Metropolis-Hastings Algorithm

Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953, *J. Chem. Physics*) sought to simulate random draws from the posterior $p(\theta|y)$ by constructing a Markov chain possessing these attributes:

- should have the same state space as θ
- should be easy to simulate from
- stationary distribution should be the posterior $p(\theta|y)$
- should be able to ignore the normalizing constant in $p(\theta|y)$ to implement the algorithm, which means $p(\theta|y)$ should appear only through ratios of the form

$$\frac{p(\theta|y)}{p(\theta'|y)}$$

so that the normalizing constant cancels

Idea of Metropolis-Hastings

The Metropolis-Hastings algorithm generates correlated variables from a Markov chain.

- the target density f is associated with a 'working' conditional density, $q(y|x)$, that can be easily simulated in practice
- $q(\cdot|x)$ is arbitrary except that it must satisfy:
  1. the ratio $f(y)/q(y|x)$ is known up to a constant *independent of* x
  2. $q(\cdot|x)$ is disperse enough to explore the entire support of f

  Note that the Markov kernel q is not the Markov kernel K of the algorithm.

Magic of M-H: for every given q, we can construct a M-H kernel K *such that* f is its stationary distribution.

Generic Metropolis-Hastings Algorithm

Given an initial value $x^{(t)}$ and conditional density $q(y|x^{(t)})$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Set

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

where

$$\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

- q is the **proposal** or **candidate** distribution
- $\rho(x, y)$ is the acceptance probability; Hastings (1970, *Biometrika*) proved this choice gives you the correct stationary distribution for the chain

Comments

- closely related to rejection sampling, but the proposal distribution changes over time (e.g. when you reject you don't go back to the beginning, you stay where you are and iterate again)
- the acceptance rate is the average of the acceptance probability over iterations:

$$\bar{\rho} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \rho(X^{(t)}, Y_t) = \int \rho(x,y) f(x) q(y|x) dy dx.$$

- the algorithm is independent of normalizing constants
- in the symmetric case, i.e. when $q(x|y) = q(y|x)$, the acceptance probability reduced to a ratio independent of q!

Detailed Balance

The Markov chain $\{X^{(t)}\}$ satisfies the *detailed balance* condition if there exists a function f (e.g. a density) such that

$$f(x)K(y|x) = f(y)K(x|y).$$

- to see this, integrate both sides with respect to x
- LHS: unconditional probability of moving from x to y when x is generated from f
- RHS: unconditional probability of moving from y to$x$, when y is also generated from f

Also called time-reversibility.

Independent Metropolis-Hastings

The generic M-H algorithm allows a proposal q which depends only on the current state of the chain.

- if we require q to be independent of the current state, i.e. $q(y|x) = g(y)$, we have a special case

Given initial state $x^{(t)}$

1. Generate $Y_t \sim g(y)$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min\left\{\frac{f(Y_t)g(x^{(t)})}{f(x^{(t)})g(Y_t)}, 1\right\} \\ x^{(t)} & \text{otherwise} \end{cases}.$$

This is like a generalization of rejection sampling.

Comparison of rejection with M-H

- rejection sample is i.i.d.; M-H sample is not–even though Y_t's generated independently–because acceptance probability of Y_t depends on $X^{(t)}$
- M-H involves repeated occurrences of the same value since rejection of Y_t results in repetition of $X^{(t)}$ at time $t+1$
- rejection sampling requires calculation of the upper bound $\sup_x f(x)/g(x) \leq M$, which is <u>not</u> required for M-H

Selecting a proposal

Generic approach: random walk Metropolis-Hastings

- simulate Y_t according to
$$Y_t = X^{(t)} + \varepsilon_t$$
where $\varepsilon_t \sim g$ is random, independent of $X^{(t)}$ (e.g. uniform or normal)
- for instance $g \sim \mathcal{N}(0, \tau^2)$, so that $Y_t \sim \mathcal{N}(X^{(t)}, \tau^2)$
- the proposal $q(y|x)$ is of the form $g(y - x)$
- the Markov chain associated with q is a **random walk** when g is symmetric around zero
- However the M-H Markov chain is not a random walk, due to the acceptance step

RWMH

Given initial $x^{(t)}$

1. Generate $Y_t \sim g(y - x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min\{1, f(Y_t)/f(x^{(t)})\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

The acceptance probability does not depend on g, i.e. for a given pair $(x^{(t)}, y_t)$, the probability of acceptance is the same whether g is any symmetric density (student-t, normal, uniform, Cauchy, etc.).

- However, different choices of g will result in different ranges for Y_t and hence different *acceptance rates*

Comments

RWMH is very appealing intuitively, but some drawbacks:

- requires very long chains to fully explore the density if it has some very low probability regions and many modes
- because of the symmetric qualities, half the time is spent revisiting regions it has already explored

A (theoretically) good alternative:

- Langevin algorithm (Roberts & Rosenthal, 1998), favors moves toward higher values of the target by including gradient in proposal

$$Y_t = X^{(t)} + \frac{\sigma^2}{2}\nabla \log f(X^{(t)}) + \sigma\epsilon_t, \quad \epsilon_t \sim g(\epsilon)$$

  and σ is a scale factor of the proposal.
- requires a lot of tuning in practice

Example: 2-D RWMH

To generate samples from the posterior of $\theta = (\theta_1, \theta_2)^T$, we would need a way of randomly sampling from a bivariate proposal distribution.

- could assume independence of parameters in the posterior and then sample each from the same univariate proposal density (not a good idea)
- could specify a bivariate normal density with some covariance matrix (requires tuning)
- fortunately, it is possible to generate multivariate normal samples using univariate normal random samplers

Gibbs Sampling

Suppose the parameter vector θ can be divided into d subvectors $\theta = (\theta_1, \ldots, \theta_d)^T$.

- an iteration of the Gibbs sampler draws values of each subvector, conditional on the values of all the other subvectors, i.e. there are d steps in iteration t
- this is possible when we can explicitly write down the conditional posterior distribution of each subvector, i.e.

$$p(\theta_j | \theta_{(-j)}^{t-1}, y)$$

with $\theta_{(-j)}^{t-1}$ all the components except for θ_j, at their current values:

$$\theta_{(-j)}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})^T$$

which are the iteration t values for those subvectors *already* updated and the iteration $(t-1)$ values for those subvectors not yet updated

- most often applied when the the conditional distributions are conjugate distributions which are easy to simulate from

Example from BDA

Consider a single observation (y_1, y_2) from bivariate normal population with mean $\theta = (\theta_1, \theta_2)^T$ and known covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

- use uniform prior on θ, then posterior is

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \Big| y \sim \mathcal{N}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

We could directly sample from joint posterior here, but to illustrate Gibbs we don't.

- conditional posterior distributions are given by

$$\theta_1 | \theta_2, y \sim \mathcal{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2),$$

$$\theta_2 | \theta_1, y \sim \mathcal{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2).$$

Choose some initial values (e.g. the MLEs), then alternate sampling from these two normal distributions.

Key Questions

1. What is a good initial value?
2. How long until convergence to stationary distribution?
3. How do we assess convergence? (convergence diagnostics)
4. After convergence to stationary distribution, how many samples to take? (controlling *finite simulation* error)

Next time

Reversible Jump

Plus: convergence diagnostics and implementation in `R`