Markov Chains on Continuous State Space

1 Markov Chains Monte Carlo

1. Consider a **discrete** time Markov chain $\{X_i, i = 1, 2, ...\}$ that takes values on a continuous state space $\mathcal{S}$.

   Examples of continuous $\mathcal{S}$ are:

   - $\mathbf{R}^d$, d dimensional space of real numbers
   - a d-dimensional unit simplex, a subset of $\mathbf{R}^d$
   - the Mandelbrot Set
   - the Brownian motion
   - set of all color configurations of an image
   - set of all geographical topologies
   - any uncountably infinite set

   Since $P_{xy} = 0, \forall x, y \in \mathcal{S}$ We have to modify the definition of transition matrix P_{xy} to **transition kernel** $P(x, A)$ where for all $x \in \mathcal{S}$ and $A \subset \mathcal{S}$. Likewise, the concepts of irreducibility, recurrence, and positive recurrence, have to be re-define on a continuous state space.

2. The **transition kernel** $P(x, .)$

   For all $x \in \mathcal{S}$, $P(x, A)$ is defined as the probability of reaching the measurable set A from state x.

3. The **transition kernel density** $p(x, .)$

   For all $x \in \mathcal{S}$, $p(x, y)$ is defined as a non-negative function such that

   $$P(x, A) = \int_{y \in A} p(x, y) dy.$$

   Note that for a given x, $p(x, .)$ ***is a pdf***, and

   $$P(x, \mathcal{S}) = \int_{y \in \mathcal{S}} p(x, y) dy = 1.$$

4. *ϕ-Irreducible:* For each measurable set A with $\phi(A) > 0$, for each x there exists n such that $P^n(x, A) > 0$.

5. *Recurrent:* For every measurable set B with $\phi(B) > 0$,

   $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) > 0$$

   for all X_0, and

   $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) = 1$$

   for ϕ almost-every X_0.

6. *Harris recurrence* means that

   $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) = 1$$

   for all X_0.

7. A probability distribution π is called the *stationary distribution* for the Markov chain with transition density p (or transition kernel P) if

   $$\pi(y) = \int p(x, y)\pi(x)dx,$$

   for all $y \in \mathcal{S}$, or equivalently,

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

   for all measurable sets $A \subset \mathcal{S}$, in which case we may write $\pi P = \pi$.

8. *Periodic:* There exists $n \geq 2$ and a sequence of nonempty, disjoint measurable sets $A_1, A_2, \ldots, A_n$ such that $\forall x \in A_j$ $(j < n)$, $P(x, A_{j+1}) = 1$, $\forall x \in A_j$, $P(x, A_1) = 1$ $\forall x \in A_n$.

9. *Aperiodic:* If the transition kernel has density $q(.|x)$, a sufficient condition for aperiodicity is that $q(.|x)$ is positive in a neighborhood of x, since the chain can them remain in this neighborhood for an arbitrary number of times before visiting other set A.

10. **Theorem:** Let P be a Markov transition kernel, with invariant distribution π, and suppose that P is π irreducible.

    Then π is the unique (invariant distribution of P).

    If P is aperiodic, then for π almost-every x

    $$||P^n(x, A) - \pi(A)||_{TV} \rightarrow 0$$

    where the distance between two probability measures π_1 and π_2 is measured using the *Total Variation Norm:* $||\pi_1 - \pi_2||_{TV} = sup_A|\pi_1(A) - \pi_2(A)|$.

    If P is Harris recurrent, then the convergence occurs for all x.

11. In words, we can start at essentially any X, run the chain for a long time, and the final draw has a distribution that is approximately π. The "long time" is called the *burn in*, and it is generally impossible to know how long this needs to be in order to get a good approximation.

12. We often don't need independence for the most common application: estimating the mean of a function $E_\pi f$ with respect to the invariant distribution. This is true as long as the chain is *uniformly ergodic*:

$$\sup_x \|P^n(x,\cdot) - \pi\|_{TV} \leq Mr^n$$

for some $M > 0$ and $r < 1$. In this case there is a central limit theorem: let $\bar{f}_n = \sum_{j=1}^n f(X_j)/n$ denote the sample mean. Then $\sqrt{n}(\bar{f}_n - E_\pi f)$ converges to a normal distribution with mean 0.

2 Metropolis-Hastings Algorithm

One of the most popular MCMC technique used in approximate sampling from complicated distributions is the Metropolis-Hastings algorithm. The setup is as earlier: we are interested in generating samples of a random variable X distributed according to the density $f(x)$.

Let $q_x(.) = q(x,.) = q(.|x)$ be a pdf such that:

1. It is easy to sample from $q(.|x)$ for all x.

2. The support of q contains the support of f.

3. The functional form of $q(y|x)$ is known or $q(y|x)$ is symmetric in y and x. As shown later, it is not necessary to know the normalizing constant in $q(y|x)$ as long as it does not depend upon x.

Given $f(x)$ and a choice of $q(y|x)$, that satisfies the above mentioned properties, the Metropolis- Hastings algorithm can be stated as follows:

(Metropolis-Hastings Algorithm) Choose an initial condition X_0 in the support of $f(x)$. The Markov chain $X_1, X_2, ..., X_n$ is constructed iteratively according to the steps:

1. Generate a candidate $Y \sim q(y|X_t)$.

2. Update the state to X_{t+1} according to

$$X_{t+1} = \begin{cases} Y & \text{w.p. } \alpha(X_t, Y) \\ X_t & \text{w.p. } 1 - \alpha(X_t, Y) \end{cases}$$

where

$$\alpha(x,y) = \min\{\frac{f(y)}{f(x)}\frac{q(y,x)}{q(x,y)}, 1\}$$

$q(x,y)$ is called the proposal density and $\alpha(x,y)$ is called the acceptance ratio/probability.

Consider first the case where the ratio

$$\frac{f(y)}{f(x)}\frac{q(y,x)}{q(x,y)}$$

values more than one and hence the acceptance probability takes the value 1.0. In this case, we set $X_{t+1} = Y$ with probability one.

In case this ratio goes below one, we set X_{t+1} to Y with acceptance probability $\alpha(X_t, Y)$.

Note that the normalizing constants in the two densities f and q cancel out and hence are not explicitly needed. However, if the normalizing constant for $q(y|x)$ depends upon x, then it does not cancel out and is needed in the expression for $\alpha(x, y)$.

In the algorithm, one generates samples from the proposal q at every step **independently** but the elements of the chain are **not independent** of each other. In fact, many times it is possible to have X_t and X_{t+1} be identical. There are some special cases that occur often in practice:

1. When $q(y|x)$ is symmetric in the two arguments, the expression for the acceptance ratio simplifies to

$$\alpha(x, y) = min\{\frac{f(y)}{f(x)}, 1\}$$

where

$$\frac{f(y)}{f(x)}$$

is often called the **likelihood ratio**. This is the case Metropolis and his colleagues considered in their 1953 paper: "Equation of state calculation by fast computing machines," Journal of Chemical Physics, 21: 1087-1092, 1953. This is the paper that introduces the world to the "Metropolis Algorithm," the basic MCMC algorithm.

2. Independent M-H: In cases where the proposal density is independent of the current state, i.e. $q(y|x) = q(y)$, then the algorithm is called Independent Metropolis-Hastings algorithm. In this case, the acceptance ratio becomes:

$$\alpha(x, y) = \min\{\frac{f(y)}{f(x)}\frac{q(x)}{q(y)}, 1\}$$

3. Random Walk M-H: In some applications, it is useful to generate proposals using a random walk. That is, the proposal is obtained using the equation: $Y = X_t + \epsilon$, where ϵ has density $g(\epsilon)$ that is symmetric and unimodal at zero. The proposal density is symmetric: $q(y|x) = g(y - x)$, and the algorithm simplifies.

For a broad set of proposal densities, it can be shown that the M-H algorithm generates a Markov chain that asymptotically becomes stationary and generates samples from the target density $f(x)$. To derive the proof, we need to define the **detailed balance** or **reversibility** condition.

A homogenous Markov chain with transition kernel density function $p(x, y)$ is said to satisfy the **detailed balance** or **reversibility** condition if there exists a function f such that:

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

for every pair x and y.

This condition can be loosely understood as:

- *the process moves from x to y* equals to *the moves from y to x*, or, in other words,
- the probabilistic transition mechanism has no direction of time, they are similar stochastically when the chain runs backward in time.
- this is the sufficient condition for a chain to achieve **stationarity**

Theorem: Suppose that a Markov chain with the transition function p satisfies the detailed balance condition with f being a pdf. Then, f is a stationary probability density of that chain.
Proof: For all $y \in \mathcal{S}$, we integrate both sides of the detailed balance condition,

$$\int_{\mathcal{S}} f(x)p(x,y)dx = \int_{\mathcal{S}} f(y)p(y,x)dx = f(y)\int_{\mathcal{S}} p(y,x)dx = f(y)$$

Thus, f is a stationary probability density.

Theorem: For any proposal density q that satisfies the three conditions listed earlier, f is a stationary probability density of the Markov chain produced by M-H algorithm.
Proof: We just need to that the transition function of the M-H Markov chain satisfies the detailed balance condition with the given $f(x)$. The transition kernel pdf can be represented as

$$p(x,y) = q(x,y)\alpha(x,y) + r(x)\delta_x(y)$$

where δ_x is the point mass at x, i.e.,

$$\delta_x(y) = \begin{cases} 0 & \text{if } y \neq x \\ 1 & \text{if } y = x \end{cases}$$

and

$$r(x) = 1 - \int q(x,y)\alpha(x,y)dy$$

is the marginal probability of remaining at x.

We would like to show that the **detailed balance** condition

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

is satisfied.

When $x = y$, then the **detailed balance** condition

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

is trivially true.

When $x \neq y$, the transition kernel pdf becomes

$$p(x,y) = q(x,y)\alpha(x,y)$$

or equivalently

$$p(y,x) = q(y,x)\alpha(y,x)$$

- Case 1: If $\alpha(x, y) \leq 1$, then $\alpha(y, x) = 1$, and

  $$\alpha(x, y) = \frac{f(y)}{f(x)} \frac{q(y, x)}{q(x, y)}$$

  Thus,

  $$\begin{aligned} f(x)p(x, y) &= f(x)q(x, y)\alpha(x, y) \\ &= f(x)q(x, y)\frac{f(y)}{f(x)} \frac{q(y, x)}{q(x, y)} \\ &= f(y)q(y, x) \\ &= f(y)q(y, x)\alpha(y, x) \\ &= f(y)p(y, x) \end{aligned}$$

  Hence the detailed balance condition is satisfied.

- Case 2: If $\alpha(y, x) \leq 1$, then $\alpha(x, y) = 1$, and

  $$\alpha(y, x) = \frac{f(x)}{f(y)} \frac{q(x, y)}{q(y, x)}$$

  Thus,

  $$\begin{aligned} f(y)p(y, x) &= f(y)q(y, x)\alpha(y, x) \\ &= f(y)q(y, x)\frac{f(x)}{f(y)} \frac{q(x, y)}{q(y, x)} \\ &= f(x)q(x, y) \\ &= f(x)q(x, y)\alpha(x, y) \\ &= f(x)p(x, y) \end{aligned}$$

  Hence the detailed balance condition is satisfied.

Some intuition for case 1:

When $\alpha(x, y) < 1$, then $f(y)q(y, x) < f(x)q(x, y)$ the process moves from x to y too often, and from y to x too rarely. To correct this condition for *stationarity*, we reduce the number of moves from x to y by introducing a probability $\alpha(x, y) < 1$ that it moves to y, and a positive probability of $1 - \alpha(x, y)$ that it stays at x.

Careful reader will motice that case 2 is just the dual image (change x to y, and y to x) of case 1. Moreover, $\alpha(y, x) = 1$ is equivalent to $\alpha(x, y) = 1$, and is equivalent to $q(x, y) = p(x, y)$, and the detail balance condition is obviously true under this special case.

3 The Gibbs sampler

A commonly encountered version of MCMC is the Gibbs sampler. The Gibbs sampler is essentially a variation of Metropolis-Hastings. We consider $X_{t+1} = (X_{t+1}^{(1)}, X_{t+1}^{(2)}, \cdots, X_{t+1}^{(p)})$ consists of p components, and each component is sampled in turn from its conditional distribution given the values of all the other components. Typically the component-wise transitions are applied *systematically in cycle* as follows:

$$\begin{aligned}
X_{t+1}^{(1)} &\sim f(X^{(1)}|X_t^{(2)}, X_t^{(3)}, \cdots, X_t^{(p)}) \\
X_{t+1}^{(2)} &\sim f(X^{(2)}|X_{t+1}^{(1)}, X_t^{(3)}, \cdots, X_t^{(p)}) \\
&\vdots \\
X_{t+1}^{(p)} &\sim f(X^{(p)}|X_{t+1}^{(1)}, X_{t+1}^{(2)}, \cdots, X_{t+1}^{(p-1)})
\end{aligned}$$

Note that the new value for each component is used *immediately* as soon as it is available.

A reasonable alternative to *systematically in cycle* is to cycle through a random permutation of the components, either a fixed permutation, or a new permutation for each cycle.

Whether Gibbs sampling can be applied depends heavily on whether it is possible to easily sample from the conditional distributions of the components. In the best case, the conditional distribution has some parametric form from which we know how to sample.

For some highly structured problems such as Markov random fields, the components may be too dependent for Gibbs sampling to be efficient. However, in such cases the structure often permits update proposals from the conditional distributions for larger **component blocks** instead of just **individual components**.

Whether Gibbs sampling can be applied depends heavily on whether it is possible to easily sample from the conditional distributions of the components. **In the best case, the conditional distribution has some *parametric form* from which we know how to sample.**

For some highly structured problems such as Markov random fields, the components may be too dependent for Gibbs sampling to be efficient. However, in such cases the structure often permits update proposals from the conditional distributions for larger **component blocks** instead of just **individual components**.

Hammersley-Clifford Theorem:

The joint density can be constructively derived from the *conditional* densities.

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

so

$$\frac{f(y)}{f(x)} = \frac{f(y|x)}{f(x|y)}$$

and

$$\int \frac{f(y)}{f(x)} dy = \frac{1}{f(x)} = \int \frac{f(y|x)}{f(x|y)} dy$$

Thus,

$$f(x, y) = f(y|x)f(x) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dy}.$$

So, what is the joint density $f(x, y)$ when $f(x|y) \propto ye^{-xy}$ and $f(y|x) \propto xe^{-xy}$?

Note that: a joint density **cannot** be derived from the *marginal* densities. Lots of counter-examples...

Examples:

- Uniform sampling inside the unit sphere in the R^p space

  $$\{X \in R^p : ||X|| \leq 1\}$$

  The volume of the unit sphere S in R^p is

  $$S = \frac{\pi^{k/2}}{\Gamma(k/2 + 1)}$$

  The smallest cube C which contains S in p-dimensional space is

  $$C = \{X = (x_1, \cdots, x_p) \in R^p : ||x_i|| \leq 1, \forall i\}$$

  with volume 2^p.

  The ratio of the volume of S to C is

  $$\frac{\pi^{p/2}}{\Gamma(p/2 + 1)2^p}$$

  which goes to 0 for large p.

  For p from 1 to 10, the ratio in percentage (%) is

```
100.0000000  78.5398163  52.3598776  30.8425138  16.4493407
  8.0745512   3.6912234   1.5854344   0.6442400   0.2490395
```

- One example in Casella, G. and George, E. (1992). Explaining the Gibbs Sampler. The American Statistician 46, 167 - 174 is the following:

  For

  $$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1 - y)^{n-x+\beta-1},$$

  where $x = 0, 1, \cdots, n$ and $0 \leq y \leq 1$, we have $f(x|y)$ is Binomial(n, y) and $f(y|x)$ is Beta$(x + \alpha, n - x + \beta)$.

- Bivariate Normal
- Multivariate Normal. Let

  $$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

  where $X \in R^p$, $X_1 \in R^{p_1}$ and $X_2 \in R^{p_2}$ with $p = p_1 + p_2$.

  When $|\Sigma_{22}| > 0$, the conditional distribution of X_1 given $X_2 = x_2$ is

  $$N_{p_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

- $$f(x,y) \propto \exp(-|x| - |y|)$$

- $$f(x,y) \propto (1 + x^2 + y^2 + 0.9xy)^{-6} + (1 + 9(x-2)^2 + 4y^2)^{-6}$$