31/03/2022 LECTURE 16: NOTES MADE ON THE VISUALISER.

3.3.1 Useful Results for Markov chains

A Markov chain is a discrete time [as simulation is discrete, consider time to be discrete] stochastic process $\{X_0, X_1, \dots\}$ with the property that the distribution of X_t given all previous values of the process only depends on X_{t-1}

$$\mathbb{P}(X_t \in A \mid X_0 = x_0, X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t \in A \mid X_{t-1} = x_{t-1})$$

for any set A.

The probability of a transition or jump from x_{t-1} at time $t-1$ to X_t at time t is given by the transition kernel

- If X_t is discrete then the transition kernel is a matrix with elements

$$P_{ij} = \mathbb{P}(X_t = j \mid X_{t-1} = i)$$

(moving from state i to j)

(time homogeneous: probabilities do not depend on t).

- If X_t is continuous then

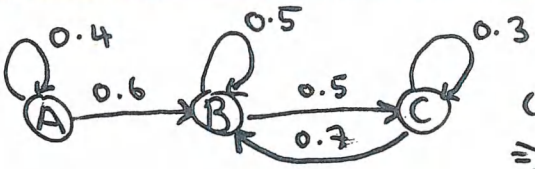
$$\mathbb{P}(X_t \in A \mid X_{t-1} = x_{t-1}) = \int_A q(x \mid x_{t-1})\, dx$$

where $q(x \mid x_{t-1})$, the transition kernel, is often written $q(x_{t-1}, x)$ to represent moving from x_{t-1} to x [NB. It's not the joint distribution!]

We now discuss useful definitions and results for Markov chains. Assume state space is finite (and therefore discrete) [MCMC methods are applicable to countability infinite and continuous state spaces]

Aim: sample from a distribution (typically the posterior) using a Markov chain.

Irreducible



Once left state A, can't get back to it. ⇒ Not irreducible.

Definition

A Markov chain is said to be **irreducible** *if for every i, j there exists k such that*

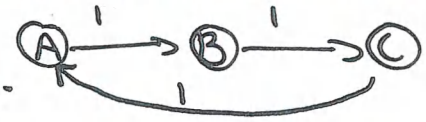
$$P(X_{t+k} = j \mid X_t = i) \quad > \quad 0$$

that is, all states can be reached from any other state in a finite number of moves.

We will want our Markov chain to be irreducible

Periodic/Aperiodic

e.g.



Each state has period 3

Definition

A state i is said to be **periodic** *with period d_i if starting from state i the chain returns to it within a fixed number of steps d_i or a multiple of d_i.*

$$d_i \quad = \quad gcd\{t : P(X_t = i \mid X_0 = i) > 0\}$$

where gcd is the greatest common divisor. If $d_i = 1$ then the state i is said to be **aperiodic**.

Irreducible Markov chains have the property that all states have the same period. A Markov chain is called **aperiodic** if some (and hence all) states are aperiodic.

Aperiodic means there's no pattern to when we return to a state. We will want this.

Recurrent/Positive Recurrent

Definition

Let τ_{ii} be the time of the first return to state i

$$\tau_{ii} = \min\{t > 0 : X_t = i \,|\, X_0 = i\}$$

A state i is **recurrent** *if $P(\tau_{ii} < \infty) = 1$ and* **positive recurrent** *if $E(\tau_{ii}) < \infty$.*

Thus, a state i is recurrent if the chain will return to state i with probability 1 and positive recurrent if, with probability 1, it will return in a finite time. An irreducible Markov chain is positive recurrent if some (and hence all) states i are positive recurrent.

Ergodic

Definition

A state is said to be **ergodic** *if it is aperiodic and positive recurrent. A Markov chain is ergodic if all of its states are ergodic.*

Stationary

Definition

A distribution π is said to be a **stationary distribution** *of a Markov chain with transition probabilities $P_{ij} = P(X_t = j \mid X_{t-1} = i)$, if*

$$\sum_{i \in S} \pi_i P_{ij} = \pi_j \quad \forall j \in S$$

where S denotes the state space.

In matrix notation if P is the matrix of transition probabilities and π the vector with ith entry π_i then the stationary distribution satisfies

$$\pi = \pi P$$

$\pi = (\pi_i : i \in S, \ \sum_{i \in S} \pi_i = 1, \ \pi_i \geq 0)$ row vector.

$$\mathbb{P}(X_t = j) = \sum_{i \in S} \mathbb{P}(X_t = j \mid X_{t-1} = i)\, \mathbb{P}(X_{t-1} = i)$$

$$= \sum_{i \in S} P_{ij}\, \mathbb{P}(X_{t-1} = i)$$

Suppose we are stationary at time $t-1$ then $\mathbb{P}(X_{t-1} = i) = \pi_i$ and

so $\mathbb{P}(X_t = j) = \sum_{i \in S} P_{ij} \pi_i = \pi_j$

i.e. X_{t-1} and X_t have the same distribution which is π.

i.e. Once we reach the stationary distribution, all future observations are from this distribution.

Note: ideas generalise to the continuous case,

e.g. stationary distribution π satisfies

$$\pi(x) = \int_{x_{t-1}} q(x \mid x_{t-1})\, \pi(x_{t-1})\, dx_{t-1}$$

$$f_{X_t}(x) = \int_{x_{t-1}} q(x \mid x_{t-1}) f_{X_{t-1}}(x_{t-1})\, dx_{t-1} = \pi(x)$$

if $f_{X_{t-1}}(x_{t-1}) = \pi(x_{t-1})$.

Existence and uniqueness

Theorem

Each irreducible and aperiodic Markov chain has a unique stationary distribution π.

Convergence

Theorem

Let X_t be an irreducible and aperiodic Markov chain with stationary distribution π and arbitrary initial value $X_0 = x_0$. Then

$$P(X_t = x \mid X_0 = x_0) \;\rightarrow\; \pi(x)$$

as $t \rightarrow \infty$.

- When does it converge?
- How does the choice of $X_0 = x_0$ affect the rate of convergence?

Ergodic

Theorem

Let X_t be an ergodic Markov chain with limiting distribution π. If $E\{g(X) \mid X \sim \pi(x)\} < \infty$ then the sample mean converges to the expectation of $g(X)$ under π,

$$P\left\{\frac{1}{N}\sum_{i=1}^{N} g(X_i) \to E\{g(X) \mid X \sim \pi(x)\}\right\} = 1.$$

Note: includes all X_t, including observations prior to convergence.
Better estimate if disregard early observations prior to convergence.

Consequence of these theorems

If we can construct an ergodic Markov chain θ_t which has the posterior distribution $f(\theta \mid x)$ as the stationary distribution $\pi(\theta)$ then, starting from an initial point θ_0, if we run the Markov chain for long enough, we will sample from the posterior.

- *for large t, $\theta_t \sim \pi(\theta) = f(\theta \mid x)$*
- *for each $s > t$, $\theta_s \sim \pi(\theta) = f(\theta \mid x)$*
- *the ergodic averages converge to the desired expectations under the target distribution.*

3.3.2 THE METROPOLIS-HASTINGS ALGORITHM.

The Metropolis-Hastings algorithm provides a way to sample from

$$f(\theta \mid x) = c\, g(\theta) \qquad g(\theta) \propto f(x \mid \theta) f(\theta)$$

without requiring us to know c.