Markov Chains

Markov Chain Monte Carlo (MCMC) methods are a broad class of stochastic sampling algorithms. Their common feature is that they aim to construct a *Markov chain* that has the target distribution as its *stationary distribution*. If one can construct such a chain and evolve it for a large enough number of iterations then samples from P may be obtained by taking widely separated points from the chain. In practice, determining when the chain has reached equilibrium (after an initial transient phase usually called the "burnin") and then drawing independent samples from the subsequent correlated chain are important parts of implementing any MCMC algorithm. Many algorithms exist for constructing MCMC chains with the desired stationary distribution: e.g. Gibbs sampling, Metropolis-Hasting sampling, Hamiltonian Monte Carlo, slice sampling, affine invariant ensemble sampling, and so on. We will discuss some of these methods here..

As we will see, the defining property of a *Markov chain* is that the transition probability ρ to the next point in the chain depends only on the current point; it does not depend on the previous history of the chain. If the chain satisfies some technical conditions described below, then after evolving for many iterations the distribution of points in the chain will approach a *stationary distribution* π. By carefully choosing ρ it is possible to make the stationary distribution equal any desired target distribution, $\pi = P$.

Definition. A *Markov chain* is a ordered sequence (or *chain*) of random points $x_0, x_1, x_2, \ldots$ in the sample space (i.e. $x_i \in \mathcal{X}$) that satisfies the *Markov property*,

$$P(x_{i+1}|x_0, x_2, \ldots, x_i) = P(x_{i+1}|x_i). \tag{1}$$

I.e. the distribution of each point x_i depends only on the previous point x_{i-1} in the chain.

A Markov chain is specified by the *transition probabilities*, $P(x_{i+1}|x_i)$[1].

Definition. A Markov chain is said to be *time-homogeneous* if the transition probabilities do not depend on the chain position; i.e. if $P(x_{i+1}|x_i) = P(x_1|x_0)$ for all i.

For time-homogeneous Markov chains the transition probabilities are described by a single time-independent transition probability PDF; $\rho(x', x) \equiv P(x'|x)$.

We use the following notation for MCMC chains. Points in the chain are denoted $x_i \in \mathcal{X}$, with $i = 0, 1, 2, \ldots$. When $d = \dim(\mathcal{X}) > 1$, the components of points are denoted

[1] Strictly speaking, a full specification should also include the initial point x_0. Typically this is set by drawing randomly $x_0 \sim \alpha$ from some *initialisation distribution* α on $\mathcal{X}$.

$x = (x^0, x^1, \ldots, x^{d-1})$. Subscripts label chain position and superscripts label vector components.

We are interested in the long-term behaviour of the chain. Given a time-homogeneous Markov chain with initial position x_0 the distribution of the next point in the chain is, by definition, given by the transition probability; i.e. $P(x_1|x_0) = \rho(x_1, x_0)$. The distribution of the third point in the chain can be obtained as follows.

$$P(x_2|x_0) = \int \mathrm{d}x_1 \; P(x_2|x_1, x_0)P(x_1|x_0) \qquad (\textit{law of total probability}) \tag{2}$$

$$= \int \mathrm{d}x_1 \; P(x_2|x_1)P(x_1|x_0) \qquad (\textit{Markov property}) \tag{3}$$

$$= \int \mathrm{d}x_1 \; \rho(x_2, x_1)\rho(x_1, x_0) \qquad (\textit{def. of transition prob}) \tag{4}$$

It follows by induction that for a given starting position the distribution of the i^{th} point is

$$P(x_i|x_0) = \int \mathrm{d}x_{i-1} \int \mathrm{d}x_{i-2} \; \ldots \int \mathrm{d}x_1 \; \rho(x_i, x_{i-1})\rho(x_{i-1}, x_{i-2}) \ldots \rho(x_1, x_0). \tag{5}$$

Note, nearby points in the chain are *not* independent.

We want our Markov chains to be able to move all over the space $\mathcal{X}$.

Definition. A Markov chain is *irreducible* if for any starting point $x_0 \in \mathcal{X}$ for any (measurable) region $A \subset \mathcal{X}$ there exists an integer $n \geq 1$ such that $\int_A \mathrm{d}x \; P(x_n|x_0) > 0$.

Irreducibility will be an importance property for our intended applications. It ensures that (given enough iterations) our chains have a *go anywhere* tendency that means they can fully explore the sample space $\mathcal{X}$. This property is sometimes called *ergodicity*.

We consider only *irreducible* and *time-homogeneous* Markov chains here. Again, we are interested in the long-term behaviour of the chain. Together, the irreducibility and time-homogeneity ensures that the chain never stops moving, e.g. by converging to a point. However, the distribution of points in the chain might still converge to something.

Definition. A time-homogeneous Markov chain is said to approach a *limiting distribution* λ on the space $\mathcal{X}$ if

$$\lim_{n\to\infty} P(x_n|x_0) = \lambda(x_n). \tag{6}$$

A limiting distribution doesn't have to exist, but if it does exist then it will be unique.

If we start the chain in the limiting distribution then it will remain there. To see that this is the case, let $x_0 \sim \lambda$ and consider the distribution of the next point in the chain;

$$P(x_1) = \int \mathrm{d}x_0 \; \rho(x_1, x_0)\lambda(x_0) = \lambda(x_1), \tag{7}$$

where the final equality follows from from Eq. 5 and the definition of λ.

The limiting distribution is an example of a *stationary distribution*.

Definition. Given a Markov chain with transition probabilities $\rho(x', x)$, a distribution π on $\mathcal{X}$ is said to be a *stationary distribution* of the Markov chain if

$$\pi(x') = \int \mathrm{d}x \; \pi(x)\rho(x', x). \tag{8}$$

Note that a limiting distribution is necessarily stationary, but a stationary distribution is not necessarily a limiting distribution.

In order to be useful for stochastic sampling, we now need a way not only of determining the stationary distribution, but also of ensuring that it matches out target distribution; i.e. $\pi = P$. The easiest way do this is to impose the stricter condition of *detailed balance*.

Definition. Consider a time-homogeneous Markov chain with transition probabilities $\rho(x', x)$. The chain is said to satisfy *detailed balance* with respect to π if

$$\pi(x)\rho(x', x) = \pi(x')\rho(x, x'). \tag{9}$$

In order to see the significance of the *detailed balance equations* in Eq. 9, consider two regions $A, B \subset \mathcal{X}$ (see Fig. 1). Suppose a chain at position $x \sim \pi$ moves at the next iteration to x'. Consider the probability that it moves from A to B,

$$\mathrm{Prob}(x \in A \textbf{ and } x' \in B) = \mathrm{Prob}(x \in A)\mathrm{Prob}(x' \in B|x) \tag{10}$$

$$= \int_{x \in A} \mathrm{d}x \; \pi(x) \int_{x' \in B} \mathrm{d}x' \; \rho(x', x). \tag{11}$$

The integrand on the right-hand side of Eq. 11 is the left-hand side of Eq. 9. Similarly, integrating the right-hand side of the detailed balance equation gives the probability that a chain initially in B and modes to A. The detailed balance equation ensure these probabilities are equal for any regions A and B. Detailed balance ensures that there is no net flux of probability anywhere in the space, when the chain is in the stationary distribution.

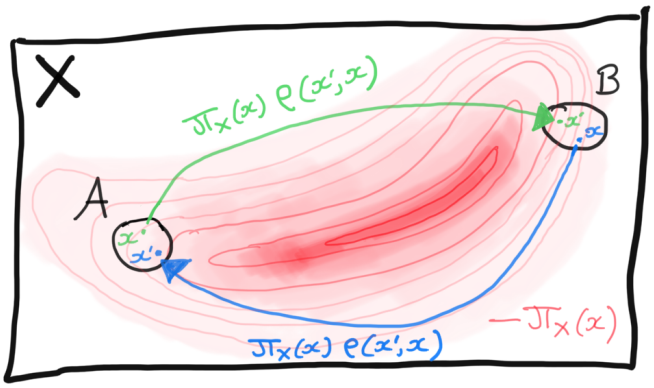

Figure 1: Illustration of detailed balance. A Markov chain moving in a 2D space $\mathcal{X}$ has a stationary distribution π (shown in red). If the chain satisfies detailed balance then there is no net flow of probability anywhere in $\mathcal{X}$. In particular, the flow from any region A into any region B (green arrow) is balanced by the flow from B into A (blue arrow).

Lemma 0.0.1. *If a Markov chain satisfies detailed balance with respect to π then π is a stationary distribution of the Markov chain.*

Proof. Integrate Eq. 9 w.r.t. x'. On LHS, use fact that π is normalised to get Eq. 8. □

We assume without proof if a time-homogeneous, irreducible Markov chain satisfies detailed balance then π is the *unique* stationary distribution and is also a limiting distribution.

Note that detailed balance is a sufficient, but not a necessary, condition for the existence of a stationary distribution, π. In other words, the detailed balance condition is a stricter than we really need. However, the detailed balance condition is convenient because it gives an easy way of finding the stationary distribution. The (local) detailed balance conditions in Eq. 9 are usually easier to check than the (global) stationarity conditions in Eq. 8.

For understanding the following sections, the most important conclusion to take away from this brief discussion of Markov chains is...

If we can design transition probabilities $\rho(x', x)$ for a time-homogeneous, irreducible Markov chain s.t. they satisfy detailed balance with $\pi = P$, then after evolving for enough iterations the distribution of the chain will approach P.