19 - MCMC

MCMC, or Markov Chain Monte Carlo, is a technique for sampling points from a distribution. Say you have a probability distribution $p(\mathbf{x})$; you might want to find a set of points $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ which are distributed according to $p(\mathbf{x})$. This is useful in a whole range of machine learning type problems where you want to find the expectation value of some function

$$I = \int p(\mathbf{x})A(\mathbf{x})dV = \langle A(\mathbf{x})\rangle_{\mathbf{x}} \tag{1}$$

Sampling from the function allows you to approximate

$$I \approx \frac{1}{n}\sum_i A(\mathbf{x}_i \tag{2}$$

For simple distributions and small numbers of dimensions this can be done analytically, for example if we are in one-dimension and $p(x)$ is the normal distibution:

$$p(x) = \mathcal{N}(0,1) \tag{3}$$

then there is a simple algorithm, the Box-Muller algorithm, for sampling points points on this distribution. One oddity of this algorithm is it gives you pairs of values: if u_1 and u_2 are both uniformly distributed on $(0,1)$ then

$$\begin{aligned} z_1 &= \sqrt{-\log_e u_1}\sin 2\pi u_2 \\ z_2 &= \sqrt{-\log_e u_1}\cos 2\pi u_2 \end{aligned} \tag{4}$$

are independent values from the distribution. We won't discuss in detail how this is derived, basically there is a change of variable to make the distribution uniform and this change of variable requires the integral of the distribution, the reason there are two values drawn rather than one is because $\exp(-x^2 - y^2)$ can be integrated when $\exp -x^2$ cannot.

Either way, the main point is that working out an explicit algorithm like this is not typically possible even if $P(\mathbf{x})$ is known explicitely and, of course, $P(\mathbf{x})$ is not always known explicitely. If the number of dimensions is low there is a procedure known as *rejection sampling* that solves this problem, MCMC is there for when the dimension is high.

The idea behind MCMC is to construct a Markov chain; that is a set of states with probabilities for a transition from state to state. Imagine you have states A, B and so on, so that if you are in state A you have probability p_{AB} of going from that state to state B; for a Markov process it is required that this probability does not depend on what transitions have been made, just on the fact you are at A now. The transition probabilities are actually conditional probabilities,

$$p_{AB} = p(B|A) \tag{5}$$

Over time, once the system has reached equilibrium about which more will be said later, it wanders around between the states. Under certain circumstances there is a 'stationary distribution', that is different states have unique probabilities $p(A)$, in other words, there is a probability $p(A)$ of getting the state A and

In Markov Chain Monte Carlo the 'states' are different values of $\mathbf{x}$ and transition probabilities are chosen so that at equilibrium $p(\mathbf{x})$ is precisely the distribution you are trying to sample from. This means that simulating the Markov Chain will give you a collection of points sampled from the distribution; obviously the individual points are not independent but the collection as a whole will have the right distribution, in other words, if $p(\mathbf{x}_2|\mathbf{x}_1)$ is high then $\mathbf{x}_1$ is likely to be followed by $\mathbf{x}_2$ in the list, but ignoring the order of the $\mathbf{x}_i$ in the list, overall the probability of occurances of $\mathbf{x}_1$ and $\mathbf{x}_2$ in the list is correct.

This will all be clearer when we have looked at how to do all this in detail; in fact there are lots of different ways to do MCMC, the algorithm we will focus on here is the Metropolis-Hastings algorithm. In the Metropolis-Hastings algorithm a new point is 'proposed' based on the old one by drawing from a proposal distribution $g(\mathbf{x}|\mathbf{x}')$

This starts with the condition that there is a stationary distribution, we rush over that above, but basically for a Markov chain to have a stationary distribution it must satisfy a condition called detailed balance:

$$p(\mathbf{x})p(\mathbf{x}|\mathbf{x}') = p(\mathbf{x}')p(\mathbf{x}'|\mathbf{x}) \tag{6}$$

There is a second condition for the stationary distribution to be unique, but we will ignore that here. Now