Markov Chain Monte Carlo

- Dimensionality of space to be sampled is a problem for rejection & importance sampling
- MCMC scales better with dimensionality
- Rooted in stat. physics

- Similar to rejection & importance sampling we sample from a proposal dist, however
  a) keep track of current state $z^{(\tau)}$
  b) proposal depends on $z^{(\tau)}$

The sequence of samples $z^{(1)}, z^{(2)}, \ldots$ form a Markov chain and are the samples from $\tilde{p}(z)$

Basic Metropolis algorithm (powerful!)

Assume we want to sample from $p(z) = \tilde{p}(z)/Z_p$ an unnormalized dist. of interest. for which $\tilde{p}(z)$ can be computed easily.

Choose a symmetric proposal dist, usually Normal centred at current sample s.t. $q(z_A | z_B) = q(z_B | z_A)$

Initialize $z^{(\tau)}$

Repeat:

Propose $z^* \sim q(z^* | z^{(\tau)})$

Accept z^* w.p.

$$A(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)}{\tilde{p}(z^{(\tau)})}\right)$$

If z^* is accepted set $z^{\tau+1} = z^*$ otherwise discard z^* and set $z^{\tau+1} = z^{\tau}$

Increment τ

Note: samples are replicated

Properties of Metropolis alg.:

1) Samples not independent, samples highly correlated

↑ can be fixed by subsampling

2)

Understanding MCMC:

Theory of Markov chains useful

A 1st order Markov chain is one in which, for $m \in \{1, \ldots, M-1\}$ and for a sequence of R.V.'s $z^{(1)}, \ldots, z^{(M)}$ the following cond. indep. property holds:

$$p(z^{(m+1)} \mid z^{(1)}, \ldots, z^{(m)}) = p(z^{(m+1)} \mid z^{(m)})$$

(Remember chain G.M.)

Such a Markov chain can be specified by the initial dist $p(z^{(0)})$ and transition prob's

$$T_m(z^{(m)}, z^{(m+1)}) \equiv p(z^{(m+1)} \mid z^{(m)})$$

Note transposition of order.

Def: A Markov chain is homogeneous if

$$T_1 = T_2 = \cdots T_m \equiv T$$

the transition functions are the same for all m.

The marginal prob. of a particular var. can be expressed in terms of the marginal prob. of the variable earlier in the chain:

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)} | z^{(m)}) p(z^{(m)})$$

Important: def: A dist. is said to be invariant or stationary w.r.t. to a Markov chain if the transition function of that M.C. leaves that distribution unchanged.

Looking forward: The dist. we are interested in sampling from will be set up as the invariant dist of a Markov chain and that chain will be simulated with a single "particle"/sample(s) long run occupancy in subsets of the parameter space being the "sample" from the distribution.

The dist. $p^*(z)$ is ~~then~~ invariant dist. of the Markov chain with transition function $T(z', z)$ if

$$p^*(z) = \sum_{z'} T(z', z)\, p^*(z')$$

Some transition functions are trivial -- these are not of interest.

Whatever transition function we define/choose can be demonstrated to ~~have~~ leave $p^*(z)$ invariant if it satisfies detailed balance w.r.t. $p^*(z)$

Detailed balance:

$$p^*(z)\, T(z, z') = p^*(z')\, T(z', z)$$

If a transition function satisfies detailed balance w.r.t. a particular dist. then that dist. will be invariant under T. This can be seen by

$$\sum_{z'} p^*(z') T(z', z) = \sum_{z'} p^*(z) T(z, z') \leftarrow \text{detailed balance}$$

$$= p^*(z) \sum_{z'} p(z', z) \leftarrow \text{const \& def of } T(\,)$$

$$= p^*(z)$$

Goal: use Markov chains to sample from a given dist. This can be done if we set up a Markov chain s.t. the desired dist. is invariant.

To accomplish this the Markov Chain must also be ergodic, i.e. we must require that for $m \to \infty$ the dist. $p(z^{(m)})$ converges to the required invariant dist $p^*(z)$ regardless of starting dist $p(z^{(0)})$. This is called ergodicity and the invariant dist is called the equilibrium dist.

Note: an ergodic Markov chain can have only one equilibrium dist.

Let's show that a ~~ergodic~~ homogeneous M.C. will be ergotic in most of the situations we will encounter.

From Neal:

<u>Fundamental Theorem</u>: If a homogeneous Markov chain on a <u>finite</u> state space with transition probs $T(x,x')$ has π as an invariant dist and

$$\nu = \min_{x} \min_{x' : \pi(x')>0} T(x,x')/\pi(x') > 0$$

then the Markov chain is ergodic, i.e. regardless of the initial probs, $p_0(x)$

$$\lim_{n\to\infty} p_n(x) = \pi(x) \qquad (a)$$

for all x. A bound on the rate of convergence is given by

$$|\pi(x) - p_n(x)| \le (1-\nu)^n \qquad (b)$$

Further, if $a(x)$ is a real-valued function of the state, then the expectation of a w.r.t. the dist. p_n, written $E_n[a]$ converges to its expectation w.r.t. π, written $\langle a \rangle$, with

$$|\langle a \rangle - E_n[a]| \le (1-\nu)^n \max_{x,x'} |a(x) - a(x')|$$

<u>Proof</u>: (Synopsis) The proof consists of showing that at all times n the distribution can be expressed as a "mixture" of the invariant distribution and another arbitrary distribution. The invariant distribution's weight/proportion in the mixture will approach 1 as $n\to\infty$

Specifically the proof demonstrates that at all times n $p_n(x)$ can be written as

$$p_n(x) = \left[1-(1-\nu)^n\right]\pi(x) + (1-\nu)^n r_n(x) \quad (1)$$

where $r_n(x)$ is a prob. dist. Note $v \leq 1$ by condition since $T(x,x') > v\pi(x')\ \forall x'$

The proof is by induction. The base case, $n=0$ can be satisfied by setting $r_0(x) = p_0(x)$. Assume (1) holds for $n = \bar{n}$ then

$$P_{\bar{n}+1}(x) = \sum_{\tilde{x}} P_{\bar{n}}(\tilde{x}) T(\tilde{x}, x)$$

← def of trans. function

$$= \left[1-(1-v)^{\bar{n}}\right] \sum_{\tilde{x}} \pi(\tilde{x}) T(\tilde{x}, x) + (1-v)^{\bar{n}} \sum_{\tilde{x}} r_n(\tilde{x}) T(\tilde{x}, x)$$

π is invariant dist of T

$$= \left[1-(1-v)^{\bar{n}}\right] \pi(x) + (1-v)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x}) \left[T(\tilde{x}, x) - v\pi(x) + v\pi(x)\right]$$

adding and subtracting same thing

$$= \left[1-(1-v)^{\bar{n}}\right] \pi(x) + (1-v)^{\bar{n}} v\, \pi(x)$$

← $r_{\bar{n}}(\tilde{x})$ a dist.

$$+ (1-v)^{\bar{n}} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x}) \left[T(\tilde{x}, x) - v\pi(x)\right]$$

look 2 pgs back

$$= \left[1-(1-v)^{\bar{n}+1}\right] \pi(x) + (1-v)^{\bar{n}+1} \sum_{\tilde{x}} r_{\bar{n}}(\tilde{x}) \frac{T(\tilde{x}, x) - v\pi(x)}{1-v}$$

$$= \left[1-(1-v)^{\bar{n}+1}\right] \pi(x) + (1-v)^{\bar{n}+1} r_{\bar{n}+1}(x)$$

where $r_{n+1}(x) = \sum_{\tilde{x}} r_n(\tilde{x}) \left[T(\tilde{x}, x) - v\pi(x)\right] / (1-v)$

By condition $T(\tilde{x}, x) - v\pi(x) > 0$ so $r_{n+1}(x) \geq 0$.

$$\sum_x r_{n+1}(x) = \sum_{\tilde{x}} r_n(\tilde{x}) \left[\sum_x T(\tilde{x}, x) - v \sum_x \pi(x)\right] / (1-v)$$

$$= \sum_{\tilde{x}} r_n(\tilde{x}) \left[1 - v\right] / (1-v) = 1$$

so $r_{n+1}(x)$ is proper dist.

So we establish

$$p_{\bar{n}+1}(x) = \left[1-(1-r)^{\bar{n}+1}\right]\pi(x) + (1-r)^{\bar{n}+1} r_{\bar{n}+1}(x)$$

from $p_{\bar{n}}(x) = \cdots$

and thus by induction this is true for $p_n(x)$ $\forall n$.

Using (1) we can show that (a) holds by

$$|\pi(x) - p_n(x)| = \left|\pi(x) - [1-(1-r)^n]\pi(x) - (1-r)^n r_n(x)\right|$$

$$= \left|(1-r)^n \pi(x) - (1-r)^n r_n(x)\right|$$

$$= (1-r)^n |\pi(x) - r_n(x)|$$

$$\leq (1-r)^n$$

Also (b) holds by

$$|\langle a \rangle - \mathbb{E}_n[a]| = \left|\sum_{\tilde{x}} a(\tilde{x})\pi(\tilde{x}) - \sum_{\tilde{x}} a(\tilde{x}) p_n(\tilde{x})\right|$$

$$= \left|\sum_{\tilde{x}} a(\tilde{x})\pi(\tilde{x}) - \sum_{\tilde{x}} a(\tilde{x})\left(\left[1-(1-r)^n\right]\pi(\tilde{x}) + (1-r)^n r_n(\tilde{x})\right)\right|$$

$$= \left|\sum_{\tilde{x}} a(\tilde{x})\left((1-r)^n \pi(\tilde{x}) + (1-r)^n r_n(\tilde{x})\right)\right|$$

$$= (1-r)^n \left|\sum_{\tilde{x}} a(\tilde{x})\pi(\tilde{x}) - \sum_{\tilde{x}} a(\tilde{x}) r_n(\tilde{x})\right|$$

$$\leq (1-r)^n \max_{x,x'} |a(x) - a(x')|$$

What does this mean? If the condition holds then the Markov chain is Ergodic and has a single equilibrium dist. Furthermore, if we run the chain long enough then regardless of initial dist. $p_n(x) \to \pi(x)$ at some rate which is a function the "reachability" of some part of the space to be sampled.

<u>Alternative view of Markov Chains of finite state spaces</u>:

The probabilities at time n can be interpretted as a row vector $\vec{p}_n$ and homogenous transition probabilities as a matrix T (a stochastic matrix whose elements are all pos., rows sum to one)

A homo. M.C. can then be written as

$$\vec{p}_{n+1} = \vec{p}_n T$$

and

$$\vec{p}_n = p_0 T^n$$

Clearly $\vec{\pi}$ is an invariant dist if

$\vec{\pi} = \vec{\pi} T$, i.e. $\vec{\pi}$ is an eigenvector of T associated with eigenvalue $\lambda = 1$

If we write $\vec{p}_0 = \vec{\pi} + a_2 \vec{v}_2 + a_3 \vec{v}_3 + \ldots$

where $\vec{v}_2, \vec{v}_3, \ldots$ are eigenvectors of T with eigenvalues < 1 then the distribution over states after the n-th step of the Markov chain will be:

$$\bar{p}_n = p_0 T^n = a_2 v_2 T^n + a_3 v_3 T^n + \dots + \pi T^n$$

$$= \lambda_2^n a_2 v_2 + \lambda_3^n a_3 v_3 + \dots + \pi$$

i.e. $p_n \rightarrow \pi$ with rate given by the size of the second largest eigenvalue.

Back to Detailed Balance

If the transition probs obey detailed balance then the dist of interest will be invariant under that Markov chain (pg 29.) A Markov chain that satisfies detailed balance is called <u>reversible</u>.

Goal: use Markov chains to sample from a distribution

need invariance & ergodicity

homogeneous Markov chain → ergodic (proof from Neal)

s.t. restrictions on transition probs → invariant dist.

IMPORTANT:

Transitions can be constructed by either "mixing" transitions or chaining transitions

$$T(z', z) = \sum_{k=1}^{K} \alpha_k B_k(z', z) \qquad \alpha_k \geq 0 \quad \sum \alpha_k = 1$$

i.e. can mix Gibbs & MH moves

$$T(z', z) = \sum_{z_1} \dots \sum_{z_{n-1}} B_1(z', z_1) \dots B_{k-1}(z_{k-2}, z_{k-1}) B_k(z_{k-1}, z)$$

MH on conditionals

mixture a chain

Invariance holds for both if
the individual transitions all hold a distribution
invariant.

Detailed balance holds for the mixture
transition if all of the transitions in the sum
preserve detailed balance. The chain proposal does
not, but symmetrizing fixes this $(B_1, B_2 \ldots B_K, B_K, \ldots, B_1)$

Metropolis Hastings

Proposal dist not necessarily symmetric

At step τ, current state of M.C. is
$z^{(\tau)}$. Sample $z^* \sim q_k(z \mid z^{(\tau)})$, accept
w.p. $A_k(z^*, z^{(\tau)})$ where

$$A_k(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)\, q_k(z^{(\tau)} \mid z^*)}{\tilde{p}(z^{(\tau)})\, q_k(z^* \mid z^{(\tau)})}\right)$$

For symmetric proposal

$$q(z^{(\tau)} \mid z^*) = q(z^* \mid z^{(\tau)})$$

~~and~~ the resulting algorithm is called the
Metropolis algorithm. (obviously stuff cancels)

We can show that $p(z)$ is an invariant
dist of the M.C. defined by the MH alg.
by showing that detailed balance holds.

How do we choose a proposal distribution?

- Art, commonly Gaussian centered at current state
- small variance of proposal $\Leftrightarrow$ frequent acceptance
- big variance " " " low acceptance

Aim for 20-40% acceptance.

Gibbs sampling (Geman Bro's)

Consider $p(\vec{z}) = p(z_1, \ldots, z_M)$. Let's say we want to sample from it. Further, let's say we can sample from

$$p(z_i \mid \vec{z}_{\setminus i})$$

the cond. dist. of z_i given all values other than i. This can be done using Rejection sampling, IS, ARS, or slice sampling. Often, in conjugate models, this conditional has a known analytic form.

Algorithm

1) Initialize $z_i \ \forall i$
2) For $\tau = 1, \ldots T$
   - sample $z_1^{(\tau+1)} \sim p(z_1 \mid z_2^{(\tau)}, z_3^{(\tau)} \ldots)$
   - sample $z_2^{(\tau+1)} \sim p(z_2 \mid z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots)$

   ⋮

Gibbs can be interpreted as a version of MH where the acceptance prob = 1.

$$A_k(z^*, z^{(\tau)}) = \frac{p(z^*)\, q_k(z^{(\tau)} \mid z^*)}{p(z^{(\tau)})\, q_k(z^{(\tau)} \mid z^*)}$$

$$= \cancel{\frac{p(z_k^* \mid \vec{z}^*_{\setminus k})\, p(z_k \mid \vec{z}^*_{\setminus k})\, p(\vec{z}^*_{\setminus k})}{p(z_k^{(\tau)})}}$$

$$= \frac{\overbrace{p(z_k^* \mid \vec{z}_{\setminus k})\, p(\vec{z}^*_{\setminus k})}^{\text{Factorization of joint}}\, p(z_k^{(\tau)} \mid z^*_{\setminus k})}{p(z_k^{(\tau)} \mid z^{(\tau)}_{\setminus k})\, p(\vec{z}^{(\tau)}_{\setminus k})\, p(z_k^* \mid z_k^{(\tau)})}$$

but note $z^*_{\setminus k} = \vec{z}^{(\tau)}_{\setminus k}$ so

$$= 1$$