Monte-Carlo and Metropolis Hastings

This is a powerful algorithm to estimate target quantities of the form

$$E_\pi(g) = \int g(x)\pi(x)\,dx$$

which are drawn from a distribution (defined be the pdf $\pi(\cdot)$) which is difficult to sample from. However, this target distribution must be easy to evaluate at a given point up to a unknown constant. This flexibility related to the unknown constant is much appreciated since intractabile normalizing constants are quite common.

Monte-Carlo Metropolis Hastings makes use of the rejection sampling (Acceptance/Rejection Method However, it additionally constructs a Markov chain which move towards the target density. For high dimensions, this is advantageous over simple rejection sampling. For low dimensions where simple rejection sampling is efficient, the latter is the better method, since, contrary to the Markov chain sampling, independent samples are obtained.

Let us look at rejection sampling. Given $M\lambda(x) \geq \pi(x)$, one first obtains sample x from a density $x \sim \lambda(x)$ different from the target $(\pi(x))$, where M is a scalar. Then one obtains a sample u from a uniform distribution $U[0,1]$ and accepts the realization x from $\lambda(x)$ if

$$u \leq \frac{\pi(x)}{M\lambda(x)}$$

The **Probability of Accepting the Sample from** $\lambda(x)$ is

$$\Pr\left(u \leq \frac{\pi(x)}{M\lambda(x)}\middle| x\right)$$

This probability can be calculated as:

$$\frac{\pi(x)}{M\lambda(x)} \int \lambda(x)dx = \frac{1}{M}$$

In practice, $M\lambda(x)$ has to be a nice cover of $\pi(x)$, but its choice is difficult in high dimensions.

Let us proof that this way of sampling indeed enables sampling from π.

Recall that, if A is the event 'the sample from λ is accepted', then

$$\Pr(A) = M^{-1}$$

so that

$$\Pr(x|A) = \frac{\Pr(x \cap A)}{\Pr(A)} = M \Pr(x \cap A)$$

The infinitesimal probability of generating and accepting x using rejection sampling is

$$\Pr(x \cap A) = \lambda(x)dx \; \Pr\left(U \leq \frac{\pi(x)}{M\lambda(x)}\right) = \frac{\lambda(x)dx\,\pi(x)}{M\lambda(x)}$$

Hence, we can conclude that

$$\Pr(x|A) = M\frac{\lambda(x)dx\,\pi(x)}{M\lambda(x)} = \pi(x)dx$$

This concludes the Proof of correctness.

In Markov Chain Metropolis Hastings the target is

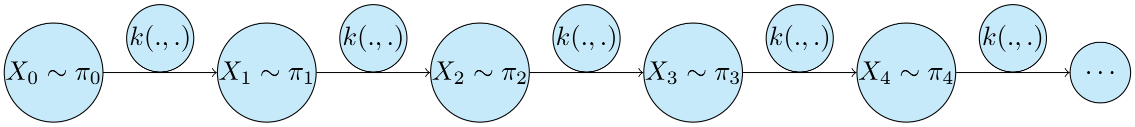
$$E_\pi(g) = \int g(x)\pi(x)\,dx$$

which can be estimated as

$$E_\pi(g) \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

where $x_1, x_2, \ldots, x_n$ are non-independent realizations from π. Consider the transition kernel of a stationary Markov chain. Assume X_{t-1} has probability density π_{t-1}. If π_t is the probability density of X_t, one has:

$$\pi_t(a) = \int k(a,x)\pi_{t-1}(x)\,dx$$



π is an invariant probability density if

$$\pi(a) = \int k(a,x)\pi(x)\,dx$$

Let p be an invariant density for the chain.
The chain is **irreducible** if, for any x and $A \in \mathcal{B}$ with $p(A) > 0$,
there exists $t > 0$ such that

$$\Pr(X_t \in A \mid X_0 = x) > 0.$$

Let $\{X_t\}$ be an irreducible Markov chain having p as an invariant density.
One has:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n} g(X_t) = E_\pi(g)$$

for any initial state (except for a set of probability zero).

The Metropolis Hasting Algorithm constructs a Markov chain whose invariant distribution is the target distribution. By transitioning from sample to sample according to the transition kernel, it ensures that after many iterations the samples are obtained from the invariant target distribution. A limitation of this approach is that the samples are generally not independent.
Consider the **current chain state** $X_t = x$. The algorithm proceeds as follows

- propose a new sample c, where $q(\cdot \mid x)$ is the proposal density of the chain.
$$c \sim q(\cdot \mid x)$$
- with a certain probability $a(c, x)$, we accept the candidate c, i.e., $X_{t+1} = c$.
- otherwise, $X_{t+1} = x$.

If the acceptance probability is given by:

$$\alpha(c, x) = \min\left(1, \frac{\pi(c)q(x \mid c)}{\pi(x)q(c \mid x)}\right),$$

then π becomes the invariant density of the generated Markov chain.
The detailed balance condition, which must be fulfilled for the algorithm to work, is

$$\pi(X_t)q(X_{t+1} \mid X_t)\alpha(X_{t+1}, X_t) = \pi(X_{t+1})q(X_t \mid X_{t+1})\alpha(X_t, X_{t+1})$$

Let us examine if it does hold. Consider Group 1:

$$\frac{\pi(X_{t+1})q(X_t \mid X_{t+1})}{\pi(X_t)q(X_{t+1} \mid X_t)} < 1$$

This implies:

$$\alpha(X_{t+1}, X_t) = \frac{\pi(X_{t+1})q(X_t \mid X_{t+1})}{\pi(X_t)q(X_{t+1} \mid X_t)}$$

and

$$\alpha(X_t, X_{t+1}) = 1$$

Consider furthermore Group 2:

$$\frac{\pi(X_t)q(X_{t+1} \mid X_t)}{\pi(X_{t+1})q(X_t \mid X_{t+1})} < 1$$

This implies

$$\alpha(X_t, X_{t+1}) = \frac{\pi(X_t)q(X_{t+1} \mid X_t)}{\pi(X_{t+1})q(X_t \mid X_{t+1})}$$

and

$$\alpha(X_{t+1}, X_t) = 1$$

Thus the balance equation follows for both groups.

The kernel of the chain is

$$k(X_{t+1} \mid X_t) = q(X_{t+1} \mid X_t)\alpha(X_{t+1}, X_t) + \delta(X_{t+1} = X_t)\left(1 - \int q(c \mid X_t)\alpha(c, X_t)\,dc\right)$$

which together with the balance equation implies

$$\pi(X_t)k(X_{t+1} \mid X_t) = \pi(X_{t+1})k(X_t \mid X_{t+1})$$

and

$$\int \pi(X_t)k(X_{t+1} \mid X_t)\,dX_t = \pi(X_{t+1}) \int k(X_t \mid X_{t+1})\,dX_t$$

Thus:

$$\int \pi(X_t)k(X_{t+1} \mid X_t)\,dX_t = \pi(X_{t+1})$$

since

$$\int k(X_t \mid X_{t+1})\,dX_t = 1$$

π is indeed the invariant density.

With this a proof has been provided showing that the kernel defined via rejection sampling does indeed converge to the target density.

Some preliminaries:

- we must be able to evaluate the target density π apart from a normalization factor.
- theoretically, the algorithm works for any $q(\cdot \mid \cdot)$ (if the chain is irreducible), but in practice, the choice of q is crucial.

Often, it is useful to adopt random-walk proposals:

$$q(c \mid x) = f(c - x) = q(x \mid c)$$

where

$$c = x_t + \epsilon$$

and

$$\epsilon \sim N(0, \Sigma)$$

which implies

$$q(c \mid x) = N(x, \Sigma)$$

S provides information on how to move locally around the current point.

$$\alpha(c, x) = \min\left(1, \frac{\pi(c)}{\pi(x)}\right)$$

The acceptance probability becomes:

$$\alpha(c, x)$$