Posterior Sampling & MCMC via Metropolis-Hastings

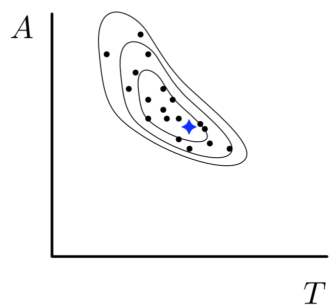
Notes for the Astrostatistics Summer School, India, July 2010
Tom Loredo <loredo@astro.cornell.edu>

Posterior Sampling & MCMC

Posterior Sampling

Recall the Monte Carlo algorithm for finding credible regions:

1. Create a RNG that can sample $\mathcal{P}$ from $p(\mathcal{P}|D_{\rm obs})$
2. Draw N samples; record $\mathcal{P}_i$ and $q_i = \pi(\mathcal{P}_i)\mathcal{L}(\mu_i)$
3. Sort the samples by the q_i values
4. An HPD region of probability P is the $\mathcal{P}$ region spanned by the $100P\%$ of samples with highest q_i



This approach is called *posterior sampling*.

Building a posterior sampler (step 1) is *hard*!

Posterior Sampling & MCMC

1. Posterior sampling
2. Accept-reject algorithm
3. Markov chains
4. Metropolis-Hastings algorithm

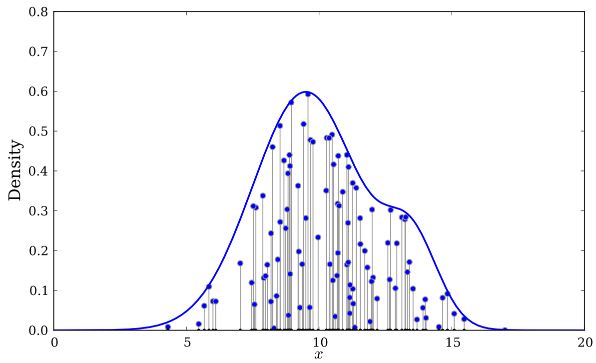
Basic Accept-Reject Algorithm

Goal: Given $q(\mathcal{P}) \equiv \pi(\mathcal{P})\mathcal{L}(\mathcal{P})$, build a RNG that draws samples from the probability density function (*pdf*)

$$f(\mathcal{P}) = \frac{q(\mathcal{P})}{Z} \quad \text{with} \quad Z = \int d\mathcal{P}\, q(\mathcal{P})$$

The probability for a region under the *pdf* is the *area (volume) under the curve (surface).*

$\rightarrow$ Sample points uniformly in volume under q; their $\mathcal{P}$ values will be draws from $f(\mathcal{P})$.



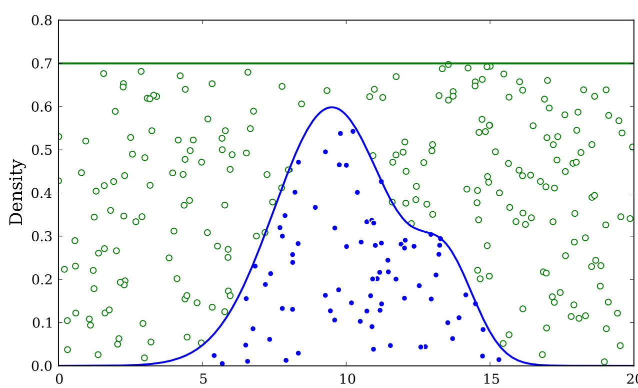
The fraction of samples with $\mathcal{P}$ ("x" in the fig) in a bin of size $\delta\mathcal{P}$ is the fractional area of the bin.

How can we generate points uniformly under the *pdf*?

Suppose $q(\mathcal{P})$ has compact support: it is nonzero in a finite contiguous region of volume V.

Generate *candidate* points uniformly in a rectangle enclosing $q(\mathcal{P})$.

Keep the points that end up under q.



Basic accept-reject algorithm

1. Find an upper bound Q for $q(\mathcal{P})$
2. Draw a candidate parameter value $\mathcal{P}'$ from the uniform distribution in V
3. Draw a uniform random number, u
4. If the ordinate $uQ < q(\mathcal{P}')$, record $\mathcal{P}'$ as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of areas (volumes), $Z/(QV)$.

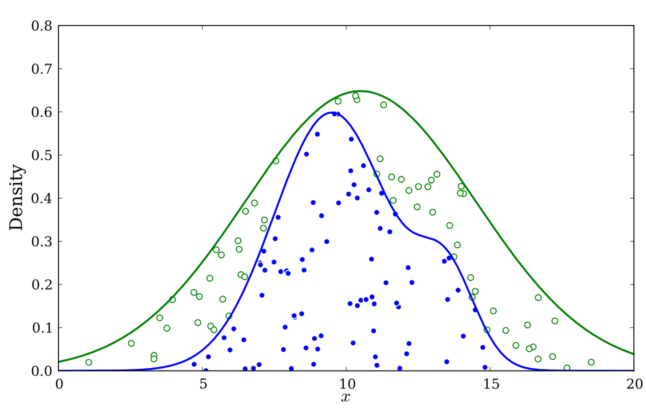
Two issues

- Increasing efficiency
- Handling distributions with infinite support

Envelope Functions

Suppose there is a *pdf* $h(\mathcal{P})$ that we know how to sample from and that roughly resembles $q(\mathcal{P})$:

- Multiply h by a constant C so $Ch(\mathcal{P}) \geq q(\mathcal{P})$
- Points with coordinates $\mathcal{P}' \sim h$ and ordinate $uCh(\mathcal{P}')$ will be distributed uniformly under $Ch(\mathcal{P})$
- Replace the hyperrectangle in the basic algorithm with the region under $Ch(\mathcal{P})$



Accept-Reject Algorithm

1. Choose an envelope function $h(\mathcal{P})$ and a constant C so it bounds q
2. Draw a candidate parameter value $\mathcal{P}' \sim h$
3. Draw a uniform random number, u
4. If $q(\mathcal{P}') < Ch(\mathcal{P}')$, record $\mathcal{P}'$ as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of volumes, Z/C.

In problems of realistic complexity, the efficiency is intolerably low for parameter spaces of more than a few dimensions.

Key idea: *Propose candidates that may be accepted or rejected*

Posterior Sampling & MCMC

Markov Chain Monte Carlo

Accept/Reject aims to produce *independent* samples—each new $\mathcal{P}$ is chosen irrespective of previous draws.

To enable exploration of complex *pdf*s, let's introduce *dependence*: Choose new $\mathcal{P}$ points in a way that

- Tends to *move toward* regions with higher probability than current
- Tends to *avoid* lower probability regions

The simplest possibility is a *Markov chain*:

$$p(\text{next location}|\text{current and previous locations}) \\ = p(\text{next location}|\text{current location})$$

A Markov chain "has no memory."

Equilibrium Distributions

Start with some (possibly random) point $\mathcal{P}_0$; produce a sequence of points labeled in order by a "time" index, $\mathcal{P}_t$.

Ideally we'd like to have $p(\mathcal{P}_t) = q(\mathcal{P}_t)/Z$ for each t. Can we do this with a Markov chain?

To simplify discussion, discretize parameter space into a countable number of *states*, which we'll label by x or y (i.e., cell numbers). If $\mathcal{P}_t$ is in cell x, we say state $S_t = x$.

Focus on *homogeneous Markov chains*:

$$p(S_t = y|S_{t-1} = x) = T(y|x), \quad \text{transition probability (matrix)}$$

Note that $T(y|x)$ is a probability distribution over y, and does not depend on t.

Aside: There is no standard notation for any of this—including the order of arguments in T!

What is the probability for being in state y at time t?

$$\begin{aligned} p(S_t = y) &= p(\text{stay at } y) + p(\text{move to } y) - p(\text{move from } y) \\ &= p(S_{t-1} = y) \\ &\quad + \sum_{x \neq y} p(S_{t-1} = x) T(y|x) - \sum_{x \neq y} p(S_{t-1} = y) T(x|y) \\ &= p(S_{t-1} = y) \\ &\quad + \sum_{x \neq y} \left[p(S_{t-1} = x) T(y|x) - p(S_{t-1} = y) T(x|y) \right] \end{aligned}$$

If the sum vanishes, then there is an *equilibrium distribution*:

$$p(S_t = y) = p(S_{t-1} = y) \equiv p_{\rm eq}(y)$$

If we *start* in a state drawn from $p_{\rm eq}$, every subsequent sample will be a (dependent) draw from $p_{\rm eq}$.

Reversibility/Detailed Ballance

A sufficient (but not necessary!) condition for there to be an equilibrium distribution is for *each* term of the sum to vanish:

$$\begin{aligned} p_{\rm eq}(x) T(y|x) &= p_{\rm eq}(y) T(x|y) \qquad \textit{or} \\ \frac{T(y|x)}{T(x|y)} &= \frac{p_{\rm eq}(y)}{p_{\rm eq}(x)} \end{aligned}$$

This is called the *detailed balance* or *reversibility* condition.

If we set $p_{\rm eq} = q/Z$, and we build a reversible transition distribution for this choice, then *the equilibrim distribution will be the posterior distribution*.

Convergence

Problem: What about $p(S_0 = x)$?

If we start the chain with a draw from the posterior, every subsequent draw will be from the posterior. But we can't do this!

Convergence

If the chain produced by $T(y|x)$ satisifies two conditions:

- It is *irreducible*: From any x, we can reach any y with finite probability in a finite $\#$ of steps
- It is *aperiodic*: The transitions never get trapped in cycles

then $p(S_t = s) \rightarrow p_{\rm eq}(x)$.

Early samples will show evidence of whatever procedure was used to generate the starting point $\rightarrow$ discard samples in an initial "burn-in" period.

Posterior Sampling & MCMC

Designing Reversible Transitions

Set $p_{\rm eq}(x) = q(x)/Z$; how can we build a $T(y|x)$ with this as its EQ dist'n?

Steal an idea from accept/reject: Start with a proposal or candidate distribution, $k(y|x)$. Devise an accept/reject criterion that leads to a reversible $T(y|x)$ for q/Z.

Using any $k(y|x)$ will not guarantee reversibility. E.g., from a particular x, the transition rate to a particular y may be too large:

$$q(x)k(y|x) > q(y)k(x|y) \qquad \textit{Note: Z dropped out!}$$

When this is true, we should use rejections to reduce the rate to y.

Acceptance probability: Accept y with probability $\alpha(y|x)$; reject it with probability $1 - \alpha(y|x)$ and stay at x:

$$T(y|x) = k(y|x)\alpha(y|x) + [1 - \alpha(y|x)]\delta_{y,x}$$

The detailed balance condition is a requirement for $y \neq x$ transitions, for which $\delta_{y,x} = 0$; it gives a condition for α:

$$q(x)k(y|x)\alpha(y|x) = q(y)k(x|y)\alpha(x|y)$$

Suppose $q(x)k(y|x) > q(y)k(x|y)$; then we want to suppress $x \rightarrow y$ transitions, but we want to maximize $y \rightarrow x$ transitions. So we should set $\alpha(x|y) = 1$, and the condition becomes:

$$\alpha(y|x) = \frac{q(y)k(x|y)}{q(x)k(y|x)}$$

If instead $q(x)k(y|x) < q(y)k(x|y)$, the situation is reversed: we want $\alpha(y|x) = 1$, and $\alpha(x|y)$ should suppress $y \rightarrow x$ transitions.

We can summarize the two cases as:

$$\alpha(y|x) = \begin{cases} \frac{q(y)k(x|y)}{q(x)k(y|x)} & \text{if } q(y)k(x|y) < q(x)k(y|x) \\ 1 & \text{otherwise} \end{cases}$$

or equivalently:

$$\alpha(y|x) = \min\left[\frac{q(y)k(x|y)}{q(x)k(y|x)}, 1\right]$$

Metropolis-Hastings algorithm

Given a target quasi-distribution $q(x)$ (it need not be normalized):

1. Specify a proposal distribution $k(y|x)$ (make sure it is irreducible and aperiodic).
2. Choose a starting point x; set $t = 0$ and $S_t = x$
3. Increment t
4. Propose a new state $y \sim k(y|x)$
5. If $q(x)k(y|x) < q(y)k(x|y)$, set $S_t = y$; goto (3)
6. Draw a uniform random number u
7. If $u < \frac{q(y)k(x|y)}{q(x)k(y|x)}$, set $S_t = y$; else set $S_t = x$; goto (3)