Overview & MCMC Convergence

Note Title 6/8/2006

Overview

(1) Sampling basic (unstructured) distributions $\pi(x)$

Main Methods:
- Rejection Sampling.
- Importance Sampling.

sample from $g(x)$, weight by $\frac{\pi(x)}{g(x)}$

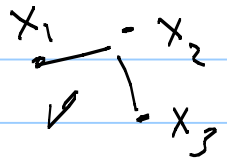
Both special cases of Weighted Importance Sampling
- eg. Rejection Control.

Variance of Sampler determines sampling efficiency (want small variance)

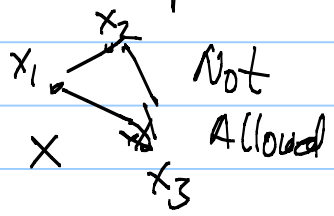
If normalization factor of distribution is unknown, then normalize by the weights.

Rao-Blackwellization → do as much as possible analytically (to decrease variance)

(2) Sampling with Structured Distribution where structure is a tree or polytree.

$\underline{X} \rightarrow$ 

No closed loops



Exploit the linear structure of the graph — ordering of nodes.

Two types of algorithm:

(1) Dynamic Programming

(2) Bayes-Kalman Filtering (e.g. particle filtering)

particle filtering is a big success story of Monte Carlo sampling in the last 5-10 years

Before, people used Kalman filters (propagate the means and covariances) and tricks when the distributions were not Gaussian (Kalman only works for Kalman).

(3) Structured Distributions of Any Distribution

MCMC — Repeated sampling from a transition kernel $K(x|y)$.

obey detailed balance

eg. $K(x|y)p(y) = K(y|x)p(x)$

$p(x)$ - target distribution

(Some people advocate transition kernels which do not obey detailed balance, but these are non-standard)

Best MCMC → Metropolis-Hastings

$K(x|y)$ divided into two steps

proposal $T(x|y)$ & acceptance $\min\left\{1, \frac{p(x)T(y|x)}{p(y)T(x|y)}\right\}$

(Automatically satisfies detailed balance.)

Variants of Metropolis-Hastings

Multiple-Try Metropolis-Hastings.

Hybrid Monte Carlo

Simulated Annealing.

Alternative MCMC

Gibbs Sampling. $P(x, y)$

Sample x from $P(x|y)$

y from $P(y|x)$

Obeys Detailed Balance.

(Gibbs can be thought of as a proposal distribution for Metropolis-Hastings where the proposals are always accepted - due to detailed balance)

Gibbs Sampling can be extended to

Data Augmentation — MCMC's alternative to the EM algorithm.

Swendsen-Wang.

Note: extensions of Swendsen-Wang involve Metropolis-Hastings.

More Alternatives

Genetic Algorithms – not an MCMC, but it sets the stage for

Evolutionary MCMC – a population of Markov-Chains that interact with each other (e.g. by competing for selection, or by cross-over).

Deterministic approximations

Belief Propagation and Mean Field Theory – Deterministic, and can get stuck in bad solutions.

But can work very fast and reliably. Need to be evaluated by Computer simulations (very hard to prove useful theoretical results).

Convergence of MCMC.

Claim exponentially fast convergence. Proof for transition kernels obeying detailed balance (results hold for other kernels, but are harder to prove.)

The proof proceeds in a series of steps.

Subclaim: if $K(x|y)$ obeys detailed balance, then $Q(x,y) = p(y)^{1/2} K(x|y) p(x)^{-1/2}$ is symmetric.

This can be checked as follows.

If $Q(x,y) = Q(y,x)$, then

$$p(y)^{1/2} K(x|y) p(x)^{-1/2} = p(x)^{1/2} K(y|x) p(y)^{-1/2}$$

$$\Rightarrow K(x|y) p(y) = K(y|x) p(x) \quad \text{detailed balance.}$$

(converse follows directly).

Treat $Q(x,y)$ as a symmetric matrix and exploit standard linear algebra results.

Linear Algebra

(1) $Q(x,y)$ has real eigenvalues λ^μ and eigenvectors $e^\mu(x)$

s.t. $\sum_y Q(x,y)\, e^\mu(y) = \lambda^\mu e^\mu(x)$

(2) These eigenvectors are orthogonal

$$\sum_x e^\mu(x) e^\nu(x) = 0, \text{ if } \mu \neq \nu$$
$$= 1, \text{ if } \mu = \nu.$$

(3) We can express:

$$Q(x,y) = \sum_\mu \lambda^\mu e^\mu(x) e^\mu(y)$$

(4) which implies that

$$Q^M(x,y) = \sum_\mu \{\lambda^\mu\}^M e^\mu(x) e^\mu(y)$$

$Q^M(x,y)$ is the matrix product of $Q(\cdot,\cdot)$ with itself M times.

Hence, we can write

$$K^M(y|x) = \sum_\mu (\lambda_\mu)^M P(x)^{-1/2} P(y)^{1/2} e^\mu(x) e^\mu(y).$$

Re-express as

$$K^M(y|x) = \sum_\mu (\lambda_\mu)^M u^\mu(x) v^\mu(y)$$

with $u^\mu(x) = e^\mu(x) P(x)^{-1/2}$

$v^\mu(x) = e^\mu(x) P(x)^{1/2}$

It can be checked that

$$\sum_x v^\mu(x) K(y|x) = \lambda^\mu v^\mu(y)$$

$$\sum_y u^\mu(y) K(y|x) = \lambda^\mu u^\mu(x)$$

u & v are left and right eigenvectors of $K(y|x)$.

They come in pairs $u^\mu(x)$ & $v^\mu(x)$ with same eigenvalue λ^μ (easy to check!)

(If $K(y|x)$ is symmetric, then $u^\mu(x) = v^\mu(x)$ and we are back to standard linear algebra.)

Also $\sum_x v^\mu(x) u^\nu(x) = \begin{cases} 1, & \text{if } \mu = \nu \\ 0, & \text{if } \mu \neq \nu \end{cases}$

This implies that any function (vector) can be expanded as

$$f(x) = \sum_{\mu} \left\{ \sum_{y} f(y)\, u^{\mu}(y) \right\} v^{\mu}(x)$$

$$f(x) = \sum_{\mu} \left\{ \sum_{y} f(y)\, v^{\mu}(y) \right\} u^{\mu}(x)$$

Now, the first eigenvalue λ^1, and its left and right eigenvectors $u^1(x)$ & $v^1(x)$ can be computed to be

$$v^1(x) = P(x) \quad \leftarrow \text{target distribution.}, \quad u^1(x) = 1, \quad \lambda^1 = 1.$$

(This follows directly from the detailed balance equations).

The other eigenvalues, $\lambda^i : i = 2, \ldots, N$ obeys $|\lambda^i| < 1$ (except for pathological cases).

This follows from the condition that

$$\sum_{y} K(y|x) = 1, \text{ for all } x.$$

Now suppose we start the MCMC with distribution $P_0(x)$ (any distribution).

We can express

$$P_0(x) = \sum_\mu \left(\sum_y P_0(y) u^\mu(y) \right) v^\mu(x)$$

$$P_0(x) = P(x) + \sum_{\mu=2}^{N} \left(\sum_y P_0(y) u^\mu(y) \right) v^\mu(x)$$

$P(x)$: target distribution

(Because $u^1(y) = 1$, hence $\sum_y P_0(y) u^1(y) = 1$, normalization
and $v^1(x) = P(x)$)

Now

$$K^M(y|x) P_0(x) = \sum_{\mu=1}^{N} \alpha^\mu \{\lambda^\mu\}^M v^\mu(x)$$

$$= P(x) + \sum_{\mu=2}^{N} \alpha^\mu \{\lambda^\mu\}^M v^\mu(x)$$

($\lambda^1 = 1$, so $\{\lambda^1\}^M = 1$)

decays exponentially fast, because $|\lambda^\mu| < 1$ for $\mu = 2 \to N$.

Hence $K^M(y|x) P_0(x) \to P(x)$ exponentially fast independent of the starting condition $P_0(x)$.

∴ MCMC converges exponentially fast.

Practical Problem

It is nice to know that MCMC convergences exponentially fast with fall off $|\lambda_2|^M$, where λ_2 is the second largest (modulus) eigenvalue of

$$Q(x,y) = p(y)^{1/2} K(x|y)\, p(x)^{-1/2}.$$

Problem: It is impossible to calculate λ_2, except for very simple transition kernels (ones you could never want to use).

Some very clever people have put bounds on the magnitude of $|\lambda_2|$. But these bounds are not tight, which means convergence in practice is a lot faster than the theory says.

So you just have to run MCMC on a computer and see how fast it converges (heuristics (like autocorrelation) can help).