Notes for CHRISTOPHE ANDRIEU et al. "An Introduction to MCMC for Machine Learning"

Malte Brinch
University of Copenhagen
Septemper 2019

1. Introduction

This section is history skip it.

2. MCMC motivation

MCMC is important for integration and optimisation problems in large dimensional spaces. some examples are:

- Bayesian inference and learning. For some given unknown parameters and data the normalization, marginalisation and expectation (value) can be found.
- Statistical mechanics. compute the partition function for a system with a given number of states and a Hamiltonian
- Optimization. Chi squared minimization and the like.
- Penalised likelihood model selection. This involves two steps. First, one finds the maximum likelihood (ML) estimates for each model separately. Then one uses a penalisation term (for example MDL, BIC or AIC) to select one of the models. computationally heavy.

2.1. The Monte Carlo principle

The advantage of Monte Carlo integration over deterministic integration arises from the fact that the former positions the integration grid (samples) in regions of high probability.

When the density p(x) has standard form, like a Gaussian, it is straightforward to sample from it using easily available routines. In nonstandard cases we need to introduce more sophisticated techniques based on rejection sampling, importance sampling and MCMC.

2.2. Rejection sampling

it is possible to sample from a distribution p(x) by sampling from another distribution q(x) where $p(x) \leq Mq(x)$ where M<∞. Problem for this method is that for high M the probability of acceptance is very small as it goes as $\frac{1}{M}$. This is a problem in higher dimensions.

2.3. Importance sampling

Here we introduce the importance weight $w(x) = \frac{p(x)}{q(x)}$ such that for an integration problem we have

$$I(f) = \int f(x)w(x)q(x)dx$$

A Monte Carlo estimate of I(f) would then be

$$I_N(f) = \sum_{i=1}^{N} f(x^{(i)})w(x^{(i)})$$

An important criterion for choosing an optimal proposal distribution (q(x)) is to find one that minimises the variance of the estimator $I_N(f)$. We try to look for $q(x) \approx p(x)$. There is high efficiency in looking for regions where |f(x)|p(x) is relatively large.

As the dimension of the x increases, it becomes harder to obtain a suitable q(x) from which to draw samples.

A sensible strategy is to adopt a parameterised q(x,θ) and to adapt θ during the simulation. This is called Adaptive importance sampling.
sampling importance resampling or SIR is used to obtain the constant M.

3. MCMC algorithms

We use MCMC when we cannot draw samples from p(x) directly, but can evaluate p(x) up to a normalising constant. An MCMC chain converges to an invariant distribution p(x) as long as the transition probability matrix has two properties:
1. Irreducibility. For any state of the Markov chain, there is a positive probability of visiting all other states. That is, the matrix T cannot be reduced to separate smaller matrices, which is also the same as stating that the transition graph is connected.
2. Aperiodicity. The chain should not get trapped in cycles
(A sufficient, but not necessary, condition is the reversibility (detailed balance) condition)

MCMC samplers are irreducible and aperiodic Markov chains that have the target distribution as the invariant distribution. One way to design these samplers is to ensure that detailed balance is satisfied. However, it is also important to design samplers that converge quickly.

The paper goes through some of the different MCMC algorithms like the other papers wrote about plus some extras. A lot of math and some pseudo code is provided.