Markov Chain Monte Carlo [1]

Lesson No. 12

Leonardo Alves de Melo - 156188

1 Introduction

Monte Carlo is a really good technic to approximate the posterior when getting the derivative seems to be intractable, but it turns out that for high dimensions it does not work so well. A way to solve that is using Markov Chain Monte Carlo, that the main idea is: given a target density $p^*(x)$ which is a stacionary distribution of a state space X in a Markov Chain. Some advantages of using Markov Chain Monte Carlo is that it is: easy to implement, applicable to various models, fast.

2 Curiosities

Markov Chain Monte Carlo was considered one of the top 10 most importants algorithms of the 20th century by a SIAM News survey, and it is the most popular method for sampling from high-dimensional distributions. It was developed during the second world war by physicists during the creation of the first atomic bomb, and was initialy unnoticed by the statistical community until a published work in 90's, and since then is used in topics like machine learning.

3 Gibbs Sampling

This is one of the most famous Monte Carlo markov Chain algorithms, and its main idea is that for a n-dimensional space, we have to take a joint sample with the n variables, and then itarates the next joint sample using the first one (just like a Markov Chain does). It can be seen in Equation 1, that is called full conditional conditional for variable i. For a graphical model, $p(x)$ can be found in its neighbors in the graph, making Gibbs Sampling a distributed algorithm, but it can be not parallelized.

$$x_i^{s+1} \sim p(x_i|x_2^s, x_3^s, ..., x_n^s) \tag{1}$$

For a case that x_i is a visible value, we do not have to sample it. This algorithm can be applied in image denoising, ilustrated by Figure 1.

Another approach of this method is the Collapsed Gibbs Sampling, where we analytically integrate out some of the unknown quantities, and just sample the rest, resulting in a more efficient algorithm. Given the Theorem 1 named as Rao-Blackwell, we do the Rao-Blackwellisation, which is, sampling z and integrating θ, knowing that θ do not participate on the Markov Chain, we can draw conditionally independent samples $\theta^S \sim p(\theta|z^S, D)$, having lower variance.

Theorem 1 *Let z and θ be dependent random variables, and $f(z,\theta)$ be some scalar function. Then* $var_{z,\theta}[f(z,\theta)] \geq var_z[\mathbb{E}_\theta[f(z,\theta)|z]]$

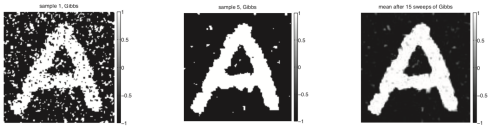

Figure 1. Example of image denoising using Gibbs Sampling

4 Metropolis Hastings

This is a more generalized algorithm used for cases that Gibbs Sampling could not handle, like computing $p(w|D)$ in a Logistic Regression, because the corresponding graphical model has no useful Markov structure.

The main idea is that for each step we move from the first state to the second one with the proposion distribution q by $q(x|x')$, where x is the first state and x' is the second one. The proposion distribution can be chosen by the user, with some conditions. Generaly it is used the Random Walk Metropolis Algorithm, where the proposal is a symmetric Gaussian distribution centered on the current state, given by Equation 2.

$$q(x|x') = \mathcal{N}(x'|x, \Sigma) \tag{2}$$

Another approach similar to importance sampling is called Independence Sampler and is given by Equation 3.

$$q(x|x') = q(x') \tag{3}$$

Now we have to use some formula in a way that it decides if the new step is accepted or not, i.e. the fraction of time spent in each state is proportional to $p^*(x)$. For a simetric proposal, this formula is given by Equation 4, and for a assimetric it is given by Equation 5, where q is used to compensate that some states may be favorized.

$$r = min(1, \frac{p^*(x')}{p^*(x)}) \tag{4}$$

$$r = min(1, \frac{p^*(x')/q(x'|x)}{p^*(x)/q(x|x')}) \tag{5}$$

The Gibbs Sampling is a subset of Metropolis Hastings, where the proposal is in the form of Equation 6.

$$q(x', x) = p(x'_i|x_{-i})\mathbb{I}(x'_{-i} = x_{-i}) \tag{6}$$

The whole steps of this algorithm can be seen in Algorithm 1

Algorithm 1 Metropolis Hastings algorithm

1: Initialize x^0;
2: **for** $s = 1, 2, 3...$ **do**
3: Define $x = x^s$;
4: Sample $x' \sim q(x'|x)$;
5: Compute acceptance probability $\alpha = \frac{\tilde{p}(x')q(x|x')}{\tilde{p}(x)q(x'|x)}$;
6: Compute $r = min(1, \alpha)$;
7: Sample $u \sim U(0, 1)$;
8: Set new Sample to $x^{s+1} = \begin{cases} x', & \text{if } u < r \\ x^s, & \text{otherwise} \end{cases}$

Every proposal distribution has to follow the Equation 7. Some examples of then is: Gaussian proposals, Mixture proposals (convex combination of base proposals), and others.

$$supp(p^*) = \subseteq \cup_x supp(q(\cdot|x)) \tag{7}$$

5 Auxiliary variable Monte Carlo Markov Chain

There is a way to improve the efficiency of sampling, and it is done introducing the auxiliary variables, which reduces the correlation between the original ones. We can get the Equation 8, where x is the original variable, z is the auxiliaries variables and $p(x, z)$ is simpler to sample than $p(x)$.

$$\Sigma_z p(x, z) = p(x) \tag{8}$$

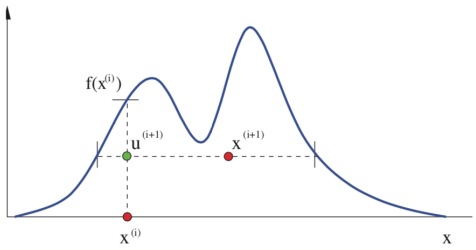

Figure 2. Illustration of the principle behind slice sampling

In the case of Logistic Regression, that its prediction is given by Equation 9, we can define the Equations 10, 11, 12 and 13 to reach posterior of the Equation 14.

$$p(y_i = 1|x_i, w) = sigm(w^T x_i) \tag{9}$$

$$\lambda_i \sim Ga(\nu/2, \nu/2) \tag{10}$$

$$\epsilon_i \sim \mathcal{N}(0, \lambda_i^{-1}) \tag{11}$$

$$z_i \stackrel{\Delta}{=} w^T x_i + \epsilon_i \tag{12}$$

$$y_i = 1|z_i = \mathbb{I}(z_i \geq 0) \tag{13}$$

$$p(\nu|\lambda) \propto p(\nu) \prod_{i=1}^{N} \frac{1}{\Gamma(\nu/2)(\nu/2)^{(\nu/2)}} \lambda_i^{\nu/2-1} e^{-\nu\lambda_i/2} \tag{14}$$

In the case of univariate multimodal distribution $\tilde{p}(x)$, we can introduce an auxiliary variable u to improve the ability to make large moves, just like the Equation 15, where Z_p is the integral of $\tilde{p}(x)$, and the marginal distribution is given by $p(x)$. Now we sample from $p(x)$ using $hatp(x,u)$ and then ignoring u using the conditionals of Equations 16 and 17, and A is given by 18 and is the set of points on or above the chosen height u, slicing just like the Figure 2.

$$\hat{p}(x,u) = \begin{cases} 1/Z_p, & \text{if } 0 \leq u \leq \tilde{p}(x) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$$p(u|x) = U_{[0,\tilde{p}(x)]}(u) \tag{16}$$

$$p(x|u) = U_A(x) \tag{17}$$

$$A = \{x : \tilde{p}(x) \geq u\} \tag{18}$$

6 Approximating the marginal likelihood

Another uses of Monte Carlo Markov Chain is in approximating the marginal likelihood (given by Equation 19), when sometimes is intractable to compute. Some methods to compute it is: The candidate method (Equation 20), Harmonic mean estimate and Annealed importance sampling.

$$p(D|M) = \int p(D|\theta, M) p(\theta|M) d\theta \tag{19}$$

$$p(D|M) = \frac{p(D|\theta, M) p(\theta|M)}{p(\theta|D, M)} \tag{20}$$

7 Conclusion

In this summary we could learn about Monte Carlo Markov Chain methods, and how it is applied in different kinds of context, and different kinds of methods.