CIS 620: Advanced Topics in Machine Learning **September 27, 2018**

Sampling-based Inference II

Lecturer: Shivani Agarwal *Scribe: Rushab Munot*

Disclaimer: These notes are designed to be a supplement to the lecture. They may or may not cover all the material discussed in the lecture (and vice versa).

Outline

- Recap
- Basics of Markov Chains
- Metropolis-Hastings Algorithm
- Gibbs Sampling

1 Recap

The methods that we discussed in the previous lecture and those that we are going to see today are used for tasks like generating samples for a random variable or approximating expectations. In the previous lecture, we looked at sampling methods including the Inverse CDF Method, Rejection Sampling and Importance Sampling. Inverse CDF sampling was used to draw samples from simple distributions for which the inverse of the cdf could be calculated. For more complicated distributions, rejection sampling and importance sampling methods were used. However even these methods had their own limitations. Sampling became more and more (computationally) inefficient with increasing dimensionality of the random variable. For example, in Rejection Sampling, the rejection probability increased exponentially with the number of dimensions. We will quickly go over these sampling techniques before addressing their limitations.

1.1 Inverse CDF Sampling

The task was to generate random samples for a continuous random variable Y assuming we know how to generate random samples for a continuous random variable X. We devised a strictly monotonically increasing function $g : Support(X) \rightarrow Support(Y)$ such that $Y = g(X)$. If we had random samples from X, then applying g over those samples generated random samples for Y. The function g was designed as

$$g^{-1}(y) = F_X^{-1}\big(F_Y(y)\big) \tag{1}$$

If X was distributed according to the standard uniform distribution, then $g^{-1}(y) = F_Y(y) \implies g(x) = F_Y^{-1}(x)$. We saw an example of the exponential distribution for which $g(x) = \frac{-1}{\lambda}\ln(1-x)$ where λ was the parameter of the exponential distribution.

1.2 Rejection Sampling

Rejection sampling is used in case where the distribution of the random variable X is known only up to a constant ($p(x) = \widetilde{p}(x)/Z_p$), implying that the cdf cannot be calculated. We used a proposal distribution q such that $\exists\ k > 0,\ \widetilde{p}(x) \leq\ k \cdot q(x)$, which was easy to sample from. We then used a sample from q, say x_0, to generate a sample u from the uniform distribution over $[0,\ k \cdot q(x_0)]$, which was accepted if it was less than $\widetilde{p}(x_0)$.

Algorithm 1 Rejection Sampling

```
for i = 1, 2,3 ... do
    Sample $x_0 \sim q(x)$
    Sample $u \sim Unif(0,\ kq(x_0))$
    Accept $x_0$ if $u \leq tildep(x_0)$, else reject $x_0$
end for
```

We noted that rejection sampling gets harder as the dimension of x increases, specifically probability of rejection increases exponentially with the dimension of x. We also noted that finding a proposal distribution easy to sample from as well as having a small value of k becomes difficult to visualize with higher dimensional distributions.

1.3 Importance Sampling

This sampling method was used to compute expectations by assigning importance weights to the value of the function on samples generated from a proposal distribution (easy to sample from). Both, the original distribution as well as the proposal distribution could be evaluated up to a normalizing constant. Specifically, if $p(x) = \frac{\widetilde{p}(x)}{Z_p}$ is the original distribution and $q(x) = \frac{\widetilde{q}(x)}{Z_q}$ is the proposal distribution, we used the following equation to approximate the expected value of a function f with respect to X.

$$\begin{aligned} \mathbf{E}[f] &\approx \frac{Z_q}{Z_p} \cdot \frac{1}{n} \sum_{i=1}^{n} f(x^i) \frac{p(x^i)}{q(x^i)} \ \ldots\ x^i \sim q \\ \frac{Z_q}{Z_p} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\widetilde{p}(x^i)}{\widetilde{q}(x^i)} \end{aligned} \tag{2}$$

We also briefly looked at Sampling Importance Resampling, where we re-sampled the the values using importance weights. In this lecture we will look at more sophisticated methods that are free of some of the limitations of the sampling methods above.

2 Introduction

In today's lecture we will focus on Markov Chain Monte Carlo (MCMC) Methods for sampling including Metropolis-Hastings and Gibbs Sampling. MCMC methods are generally used for sampling in higher dimensions. Consecutive samples generated from these methods are correlated and not independent. To reduce correlation, once a sample from the algorithm is accepted, several samples are skipped (usually some constant number) before accepting the next one. For unidimensional or bidimensional sampling, Rejection Sampling, Adaptive Rejection Sampling (for log-concave distributions), Importance Sampling are easier to use and do not suffer from the problem of correlated samples.

In simple terms, a (discrete) Markov Chain is a sequence of random variables, such that the next observation depends only on the current observation and no previous observation. MCMC methods use a conditional propositional distribution, which defines a Markov Chain, to iteratively generate samples from an approximation of the original distribution. This approximate distribution converges to the original distribution as the iterations progress. We will look at two such algorithms - the Metropolis-Hastings algorithm and Gibbs Sampling, which can be looked at as a special case of Metropolis Hastings.

Gibbs Sampling is used in tasks where it is easy to sample from the conditional distribution of one parameter given all other parameters. The variables are sampled cyclically.

These algorithms are used in a wide range of problems, especially in probabilistic models in unsupervised learning settings (which usually have a lot of latent variables). Metropolis Hastings can be used to sample from posterior distributions in probabilistic models, which in many realistic models , are not conjugate to the priors and hence, generally intractable. Gibbs Sampling is used for inference in tasks like Dirichlet Process Mixture Models, Topic Models (example Latent Dirichlet Allocation), probabilistic Matrix Factorization, etc. where local conjugacy is observed.

3 Markov Chains

Definition 3.1. (Markov Chains) A sequence of random variables (or random vectors) $X^0, X^1, X^2, \ldots$ form a Markov Chain if

$$\mathbf{P}\Big(X^{t+1} = x^{t+1} \mid X^0 = x^0, X^1 = x^1, \ldots, X^t = x^t\Big) = \mathbf{P}\Big(X^{t+1} = x^{t+1} \mid X^t = X^t\Big) \tag{3}$$

The above definition is valid when $X^i s$ are discrete. For the continuous cases, probabilities will be replaced by densities.

The discrete case Markov chain can be looked upon as a random walk on a graph with m nodes (states). $P(X^{t+1} = x' \mid X^t = x)$ denotes the probability of going from state x to state x' at the t^{th} step. If this probability is independent of t, we say that the Markov Chain is homogeneous.

Definition 3.2. (Homogeneous Markov Chains) A Markov Chain is is said to be homogeneous if $\forall\ t,\ t'$, we have

$$P(X^{t+1} = x' \mid X^t = x) = P(X^{t'+1} = x' \mid X^{t'} = x) \tag{4}$$

We denote this constant probability (or density in case of continuous random variables) by $M(x, x')$. For the discrete case, M is a matrix of size $m \times m$ called the transition matrix. We will overload M to denote the Markov chain and the transition matrix. For this lecture, We will assume that M is homogeneous until specified otherwise.

Let M be a homogeneous Markov Chain. If a state x is chosen at random from a Multinoulli distribution $\pi_{m\times 1}$, the probability of being at state x' next is $\sum_{x=1}^{m} \pi(x)M(x, x')$ where $\pi(x)$ denotes the x^{th} component of π. Let π^t denote the probability vector of being at each state at time t. Then $\pi^{t+1} = M \cdot \pi^t$. If one were to follow a random walk for t steps, the state he or she would end up in would be distributed according to π^t. It is equivalent to saying that $X^t | X^0 \sim \pi^t$. Further,

$$\pi^{t+1} = M^t \pi_0 \tag{5}$$

If $\pi = M\pi$, then $X^0, X^1, X^2, \ldots$ would be distributed according to π. If such a distribution exists we call it a stationary distribution.

Definition 3.3. (Stationary distribution) A distribution π over m states is said to be stationary with respect to a Markov chain M over those m states if

$$M\pi = \pi \tag{6}$$

M is said to be a stationary Markov chain.

3.1 Properties of Markov Chains

Definition 3.4. (Irreducible Markov Chain) A Markov chain is said to be irreducible if $\forall\ x, x',\ \forall t,\ \exists\ t' > 0$, such that

$$P(X^{t+t'} = x' | X^t = x) > 0$$

Definition 3.5. (Period of a state in a Markov Chain) The period $d(x)$ of state x in a Markov Chain represented by matrix M is defined by

$$d(x) = \gcd\Big(t \in \mathbb{N} \;\Big|\; P(X^t = x \mid X^0 = x) > 0\Big) \tag{7}$$

We now state two important theorems Theorem 3.1 and Theorem 3.2 without proof.

Theorem 3.1. All states in an irreducible Markov chains have the same period. This period is called the period of the Markov chain.

Definition 3.6. (Periodic Markov Chain) An irreducible Markov chain is said to be aperiodic if its period is equal to 1.

Definition 3.7. (Limiting Distribution) π is said to be the limiting distribution of a Markov chain with transition matrix M if $\forall\ x, x'$ we have

$$\pi(x') = \lim_{n\to\infty} \Big(P(X_n = x' \;\Big|\; X^0 = x)\Big) \tag{8}$$

$$\implies \pi = \lim_{t\to\infty} M^t \pi' \text{ for all distributions } \pi' \tag{9}$$

Note that, a limiting distribution, if it exists, is always a stationary distribution.

Theorem 3.2. For an irreducible and aperiodic Markov chain, there exists a unique stationary distribution which is also the limiting distribution.

We will now state a sufficient condition for a distribution π over the states to be a stationary distribution

Theorem 3.3. (Detailed Balance) A distribution π over possible states of the Markov chain with transition matrix M is stationary if it satisfies the detailed balance property.

$$\pi(x)M(x, x') = \pi(x')M(x', x) \tag{10}$$

Proof.

$$\begin{aligned}\sum_x \pi(x)M(x, x') &= \sum_x \pi(x')M(x', x)\\ &= \pi(x') \sum_x M(x', x)\\ &= \pi(x')\end{aligned}$$

□

Detailed balance says that the probability of traversing an edge between two states is the same in both directions. This is a stronger condition than stationarity. A Markov chain is said to be reversible with respect to π if the transition matrix satisfies detailed balance with respect to π.

Theorem 3.4. Let $M_1,\ M_2,\ \ldots,\ M_K$ be K Markov chains, all of which have π as a stationary distribution. Then

1. π is a stationary distribution of all convex combinations of $M_1, M_2, \ldots, M_K$ given by

$$M = \alpha_1 M_1 + \alpha_2 M_2 + \cdots + \alpha_K M_K \ \ \ldots\ \alpha_i \geq 0,\ \sum_{i=1}^{K} \alpha_i = 1 \tag{11}$$

   This property also holds for detailed balance.

2. π is a stationary distribution of the composite Markov chain M given by

$$M = M_1 M_2 \ldots M_K \tag{12}$$

   This does not hold for detailed balance.

4 Metropolis Hastings

Like the sampling algorithms of the previous lecture, we will use a proposal distribution which is easy to sample from and use it to generate samples from the original distribution. However, the proposal distribution for Metropolis Hastings is a conditional distribution, with the distribution for next sample being dependent on the current sample. Let us formally give the setup required for Metropolis-Hastings algorithm.

As before, the task is to generate sample from $p(x) = \frac{\widetilde{p}(x)}{Z_p}$, where Z_p is not known and $\widetilde{p}(x)$ is easy to calculate for all x in the support. A distribution $q(x'|x)$ is used as the proposal distribution. The Metropolis Hastings algorithm constructs a Markov Chain with stationary distribution equal to $p(x)$ [2].

Samples generated by Metropolis Hastings are not independent. To reduce correlation, a certain number of samples are skipped before the next sample is outputted. Though there is still some correlation, this method works well in practice.

Metropolis gave a the Metropolis' algorithm in 1953, which requires a symmetric proposal distribution, that is $q(x'|x) = q(x|x')$. It is obvious that all samples generated by q cannot be accepted as samples from p. As in Rejection Sampling, an acceptance probability for the generated sample is required. Given a random sample x' from the proposal distribution $q(x'|x)$, acceptance probability must be higher when $\widetilde{p}(x') > \widetilde{p}(x)$. The acceptance probability given by Metropolis was $A(x, x') = \min\left(1,\ \frac{\widetilde{p}(x')}{\widetilde{p}(x)}\right)$ Hastings extended this algorithm in 1970 to include asymmetrical distributions and modified the acceptance probability to $A(x, x') = \min\left(1,\ \frac{\widetilde{p}(x')}{\widetilde{p}(x)} \cdot \frac{q(x|x')}{q(x'|x)}\right)$. Theorem 4.1 shows that p is indeed the stationary distribution of the resulting Markov chain.

Since p is the limiting distribution of the Markov chain generated by Metropolis Hastings, initial samples generated from the algorithm are not indicative of p. In practice, a *burn-in* period is set, before which all samples are discarded.

Algorithm 2 The Metropolis Hastings Algorithm

```
Initialize $x^0$
for t = 0, 1, 2, ... do
    Sample $x' \sim q(x' \mid x^t)$
    Sample $u \sim \text{Unif}(0,1)$
    $A(x^t, x') = \min\left(1,\ \frac{\widetilde{p}(x')}{\widetilde{p}(x)} \cdot \frac{q(x^t|x')}{q(x'|x^t)}\right)$
    if $u < A(x^t, x')$ then
        $x^{t+1} \leftarrow x'$
    else
        $x^{t+1} \leftarrow x^t$
    end if
end for
```

Theorem 4.1. p is the stationary distribution of the Markov chain resulting from the Metropolis Hastings algorithm.

$$p(x)M(x,x') = p(x)q(x'|x)A(x,x') + p(x)\mathbf{1}(x'=x)\int_u q(u|x)(1-A(x,u))du \tag{13}$$

$$= \min\Big(p(x)q(x'|x), p(x')q(x'|x)\Big) + p(x)\mathbf{1}(x'=x)\int_u q(u|x)(1-A(x,u))du \tag{14}$$

Since the above expression is symmetric in x and x', $p(x)M(x,x') = p(x')M(x',x)$ and thus p satisfies detailed balance impyling that p is the stationary distribution.

$$\tag{15}$$

5 The Gibbs Sampling Algorithm

In Gibbs Sampling, we want to sample from a distribution $p(x) = p(x_1, x_2, \ldots, x_m)$, given that we can easily sample from $p(x_i|x_{\setminus i})$, where $x_{\setminus i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$.

In many problems involving Bayesian inference, the prior and likelihood are not conjugates of each other. In such cases, the posterior often does not have a closed form solution. However, the conditional likelihood, $p(\theta_i|\theta_{\setminus i})$, is conjugate to the prior, $p(\theta)$, where θ is the set of parameters. Such a model is known as a locally conjugate model. Many important models including Dirichlet Process Mixture Models, Latent Dirichlet Allocation, Bayesian PCA, etc. have local conjugacy and make use of Gibbs sampling (or variants) for inference. Gibbs sampling is also used for inference in matrix factorization based recommender system models.

The Gibbs Sampling algorithm updates each variable, by sampling a value from its conditional distribution. This new value is then used to sample other variables. The variable to be updated is either chosen cyclically or randomly. The sampling procedure is given in Algorithm. 5

Define m distributions, $q_i(x'|x) = p(x'_i|x_{\setminus i})$ for $i \in [m]$. Each of these define a Markov Chain in the Metropolis Hastings Algorithm. The Markov chain when Gibbs sampling is viewed as Metropolis Hastings

Algorithm 3 The Gibbs Sampling Algorithm

```
Initialize $x^0$
for t do = 0, 1, 2, . . .
    $x_0^{t+1} \sim p(x_0|x^t_{\backslash 0})$
    $x_1^{t+1} \sim p(x_1|x^t_{\backslash 1})$
    ⋮
    $x_m^{t+1} \sim p(x_m|x^t_{\backslash m})$
end for
```

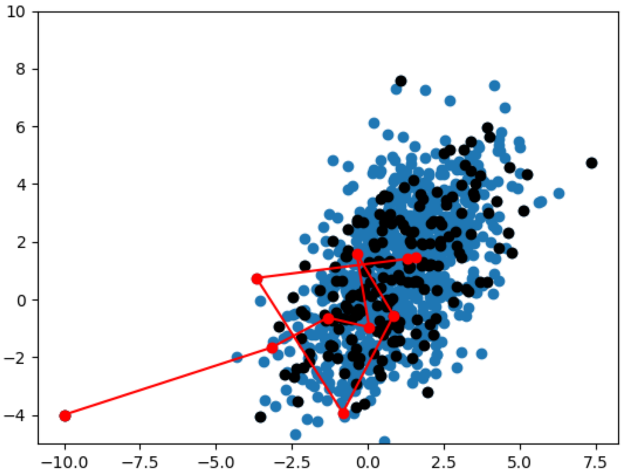

Figure 1: **Illustration of Gibbs Sampling** for a Gaussian Distribution $\mathcal{N}\left(\begin{bmatrix}1 & 1\end{bmatrix}, \begin{bmatrix}2 & 1\\ \frac{3}{2} & \frac{5}{2}\end{bmatrix}\right)$. The red portion is the burn-in period. The black points are accepted samples (i.e. after skipping a constant(5) number of points). The blue points denote all samples produced by the algorithm

is the composite Markov chain $M = M_1 M_2 \dots M_m$. Acceptance probability for all m Markov chains is 1.

$$A_i(x^t, x') = \min\Big(1,\ \frac{p(x')}{p(x)}\frac{q_i(x^t|x')}{q_i(x'|x^t)}\Big) \tag{16}$$

$$= \min\Big(1,\ \frac{p(x'_i|x^t_{\backslash i})p(x^t_{\backslash i})}{p(x_i|x^t_{\backslash i})p(x^t_{\backslash i})}\frac{p(x_i|x^t_{\backslash i})}{p(x'_i|x^t_{\backslash i})}\Big) \tag{17}$$

$$= min(1, 1) = 1 \tag{18}$$

6 Simulated Annealing

Algorithms like Metropolis Hastings can be used to approximate averages and expectations but not to compute quantities like the mode. In this section, we describe a method called *simulated annealing*[1] that approximates the global maximum of the density function.

This method generates samples from a non-homogeneous Markov chain, by slightly modifying the Metropolis Hastings algorithm. The stationary distribution at time t is p^{1/T_t}, where T_t is a decreasing function or cooling schedule of t such that $\lim_{t\to\infty} T_t = 0$. The acceptance probability is defined as follows

$$A(x^t, x') = \frac{\widetilde{p}^{1/T_t}(x')}{\widetilde{p}^{1/T_t}(x)} \cdot \frac{q(x^t|x')}{q(x'|x^t)} \tag{19}$$

The intuition behind doing so is that the limiting density $\lim_{u\to\infty} \frac{p^u x}{\int_x' p^u(x')dx'}$ is concentrated on the global maximum(s).

References

[1] Crishtophe Andrieu Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning https://doi.org/10.1023/A:1020281327116*, (50: 5), 2003.

[2] Cristopher Bishop. Pattern Recognition and Machine Learning. *Springer-Verlag, Berlin, Heidelberg*, 2006.

[3] Jane Lee, William Brown, and Shivani Agarwal. CIS620 - Lecture 7 - Basic Sampling Methods. *University of Pennsylvania*, September 25, 2018.