Sampling

Tuesday, November 6, 2018 00:20

For the glory of God

The importance of Sampling

· It is clear intuitively that we should get some samples from the interesting or important region.

- For example, we may want to sample points where the Gaussian distribution has very high probability.

· Then, how would you sample the points ? Because the only thing that we know is the range such as $x = [-10, 10]$

⇒ Basically, we do this by sampling methods such as inverse transform sampling from a distribution.

· This could be better than just randomly sampling.

· In this note, we will dive into a couple of sampling techniques.

Inverse transform sampling

a) What is inverse transform sampling ?

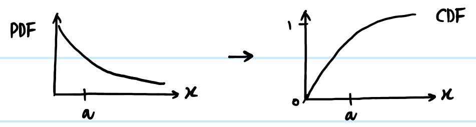
· Inverse transform sampling is a basic method for generating sample numbers at random from any probability distribution given its cumulative distribution function.

b) How does it work ?

· As we discussed, CDF must be given to use the inverse transform sampling.

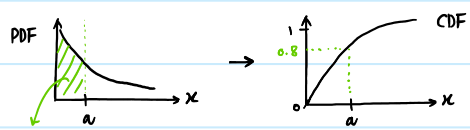
· Let us take an exponential function to explain a procedure ;

1) Calculate the CDF from PDF



$PDF(x) = e^{-x}$ $\quad$ $CDF(x) = \int_0^x e^{-x}\,dx = 1 - e^{-x}$

2) Let's say that you want to know a ; then,



; once we know CDF, we can find a because we know the function.

e.g. $x = 0.4$ when $y = 0.8$

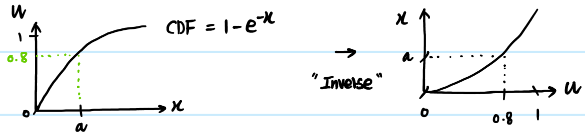
Let's say $\int_0^a PDF = 0.8$

∴ If we can calculate the inverse CDF, then we can obtain a random sample from the distribution.

↓

we will get what it is ; however, you can easily think about it as $y = f(x) \leftrightarrow x = f(y)$

3) Calculate the inverse CDF for inverse transform sampling.



; Let $u = 1-e^{-x}$

$\Leftrightarrow e^{-x} = 1-u \Leftrightarrow -x = \log(1-u) \quad \therefore x = -\log(1-u)$

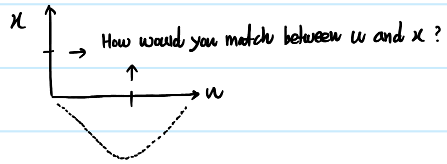
4) Here, we need to keep in mind that the inverse transform sampling takes uniform distribution (u) between 0 and 1.

What is the significance of uniform distribution ?

: If it's not uniform, it would be hard to say that $x = f(u)$

↳ This is because uniform distribution has always same probability regardless of x.

e.g. If $u \sim$ Gaussian distribution



Rejection Sampling

a) What is the Rejection sampling ?

- Basically, it is a basic technique used to generate observations from a distribution. (It's inefficient especially for multi-dimensional distributions)
- Suppose that we want to sample from a distribution $p(x)$ that is difficult or impossible to sample from directly.

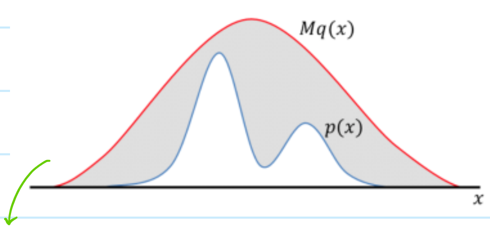
  ↳ Instead, let's say that we have a simpler distribution $q(x)$ from which sampling is easy.
- The idea behind rejection sampling is to sample from $q(x)$ and apply some rejection / acceptance criterion such that the samples that are accepted are distributed according to $p(x)$.

  ↳ we will get there soon !

b) How does it work ?

- To begin with, let's assume that we know the probability density function (PDF) of $p(x)$; but it's hard to sample directly from the distribution.

It can be either uniform, normal, and so forth.



; where $q(x)$ = proposal distribution where it's easy to sample

$p(x)$ = target probability density function where it's difficult to sample

M = constant used for rejection sampling

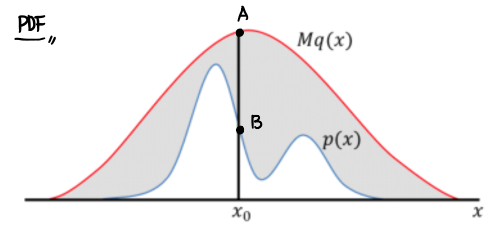
Note that ;

- $q(x)$ must cover / envelope $p(x)$ distribution. (This is why $Mq(x)$ is often called the envelope distribution)

- This is generally done by choosing a constant $M > 1$ such that $Mq(x) > f(x)$ for all x.

· Let us walk through the procedure ;

- Generate a sample x from the proposal distribution $q(x)$



; Let's say we want to know X_0 where $P(X)$ has B.

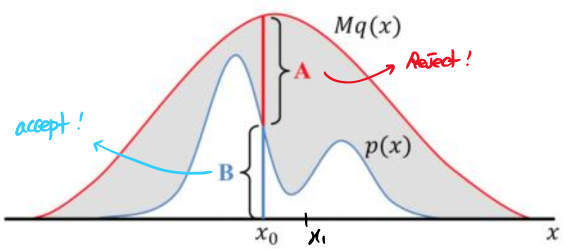
- Generate a $[0,1)$ uniform random number (between 0 and $Mq(X_0)$)

- Check whether or not $u < P(X)/Mq(X)$

↳ A common criterion for accepting samples is based on the ratio of the two probabilities.

- if $u < \frac{P(X)}{Mq(X)}$, accept the point as a sample
- if $u > P(X)/Mq(X)$, reject them

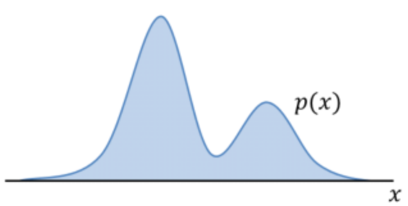
; This represents
- if the ratio is close to one, $P(X)$ must have a large amount of probability mass around x
- if the ratio is small, $P(X)$ has low probability mass around x.



; To be more specific,

1) Sample the X_0 from the proposal distribution
2) Generate u randomly between 0 and $Mq(x)$
3) Check the criterion
   - Accept the X_0 if $u < \frac{P(x)}{Mq(x)}$
   - Reject the X_0 as a sample for $p(x)$ if $u < \frac{P(x)}{Mq(x)}$
4) Try with the next sample point (X_1)

- By repeating the process for all x, the rejection sampling would provide the distribution by the sample points ;



; This distribution is generated by samples calculated by rejection sampling.

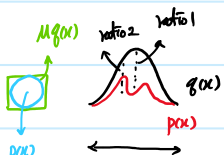
c) Acceptance probability

· The acceptance probability is defined as following ;

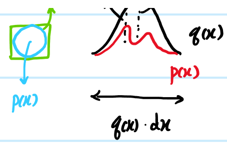
$$p(\text{accept}) = \int \left(\frac{P(X)}{Mq(x)}\right) q(x)\,dx = \frac{1}{M}\int p(x)\,dx$$

↓

You can think about the derivation of Monte-Carlo Simulation.



You can think about the derivation of Monte-Carlo Simulation.



· It seems that ;

- The acceptance probability is inverse proportional to the constant M.

- This means that we may need to define the M as small as possible in order to maximize the acceptance probability.

· If you get really high rejection rate,

- You may need to change either M or the proposal distribution.

Gibbs sampling

a) What is the Gibbs sampling ?

· Gibbs sampling was proposed in the early 1990s.

· The Gibbs sampling might argue with the idea of Metropolis Hastings algorithm at that time ;

- The Gibbs sampling proposed to use only a target probability if it is given in order to sample points.

- The MH algorithm handles both a target and proposal distributions ; whereas, the Gibbs sampling only uses a target.

∴ For this reason, the Gibbs sampling is considered as a special case of Metropolis - Hastings.

· Gibbs sampling is a MCMC algorithm that repeatedly samples from conditional distribution of one variable of the target distribution P, given all of the other variables. → For more information, take a look the hand-written note

b) Why Gibbs sampling ?

· Even though the MH algorithm works well to sample points from a target distribution,

- The MH algorithm requires a proposal distribution.

- The MH algorithm might not work well for high-dimensional cases.

· The Gibbs sampling is very attractive because it could sample / handle the high-dimensional cases.

The main idea is to break the problem of sampling from the high-dimensional joint probability into a series of samples from low-dimensional conditional distributions.

· While the MH algorithm either accepts or rejects the point based on criteria, the Gibbs sampling always accepts the point as one of sample points.

- For this reason, the acceptance probability for Gibbs sampling is always equal to one.

· Even though we accepts all the time, it's okay because we are going to burn the points from the beginning.

c) How does it work ?

↳ And then, we'd consider the characteristic of Markov chain !

· In order to understand a procedure, let us take an example of Gibbs sampling with three variables. $P(z_1, z_2, z_3)$

· To begin with, we need to make sure that the distribution has to be a full conditional joint probability.

- The full conditional usually arises in the context of MCMC or Gibbs sampling.

- Essentially, a conditional in Bayesian analysis is generally the distribution of parameter $\theta = (\theta_1, \theta_2, \cdots, \theta_K)$ given the data $y = (y_1, y_2, \cdots y_n)$ as following ;

$$P(\theta_1, \theta_2, \cdots \theta_K \mid y_1, y_2, \cdots y_n)$$

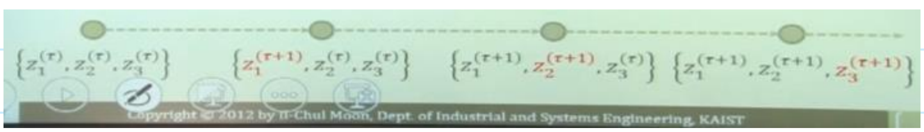
- However, when we sample for particular parameters in the Gibbs sampling, we consider the distribution as follows ;

$$P(\theta_j \mid \theta_1, \theta_2, \cdots \theta_K, y_1, y_2, \cdots y_n)$$; This is called as full conditional distribution of θ_j

· Anyhow, the procedure is as follows ;

1) Given full joint probability : $P(z_1, z_2, z_3)$
2) Sample $z_1 \sim P(z_1 \mid z_2^{t}, z_3^{t}) \Rightarrow$ obtain a value of z_1^{t+1} ; Note that t would be very large
3) Sample $z_2 \sim P(z_2 \mid z_1^{t+1}, z_3^{t}) \Rightarrow$ obtain a value of z_2^{t+1}
4) Sample $z_3 \sim P(z_3 \mid z_1^{t+1}, z_2^{t+1}) \Rightarrow$ obtain a value of z_3^{t+1}

· Therefore, the Markov chain may look like ;



d) Example and demo of Gibbs sampling

1) Example

· Let's say that there is a distribution $P(z_1, z_2, z_3)$ over three variables.

· Suppose that we want to sample one point from the distribution using the Gibbs sampling.

· The first step is to select a point randomly such as $z^0 = (z_1^0, z_2^0, z_3^0)$

· Next, starting from the initial point, we are going to sample a new point $z' = (z_1', z_2', z_3')$

· How so ? Let's take a look the steps as below ;

- Replace z_1^0 by new value z_1' obtained by sampling from $P(z_1' \mid z_2^0, z_3^0)$
- Replace z_2^0 by new value z_2' obtained by sampling from $P(z_2' \mid z_1', z_3^0)$
- Replace z_3^0 by new value z_3' obtained by sampling from $P(z_3' \mid z_1', z_2')$

· Finally, we could obtain $z' = (z_1', z_2', z_3')$; which is considered as one of sample points.

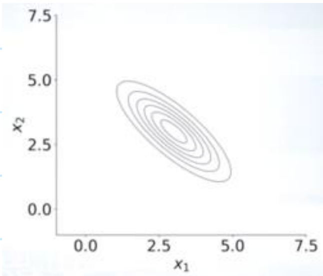
· We are going to repeat the process until the Markov chain is converged to the stationary status.

↳ Then, we can burn the sample points from the beginning such as z^0, which is not a sample point anymore.

↓ It's possible based on the characteristic of Markov chain that the event is only dependent on the previous one.

2) Demo (from coursera)

· Let us take an example with two-dimensional Gaussian distribution as below ;

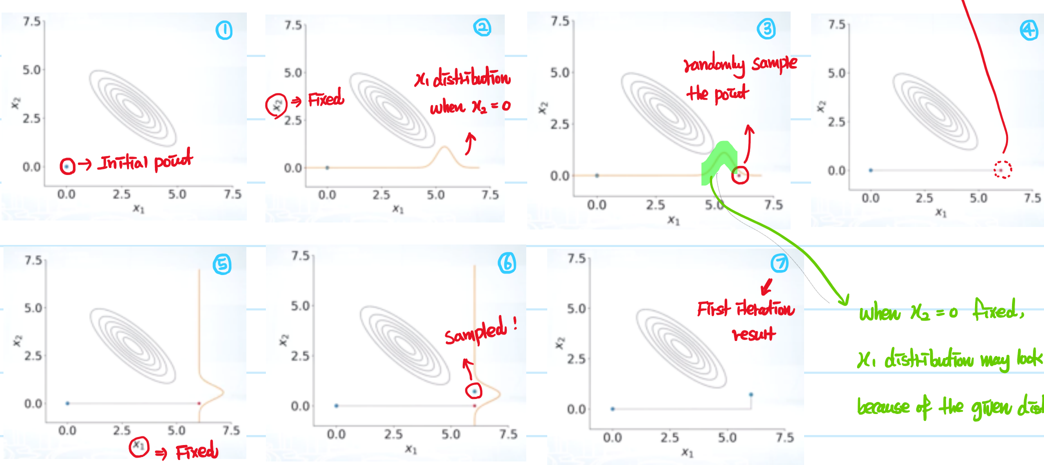


; Even if this is 2D, note that it's possible for Gibbs to handle higher dim.

please note that ;

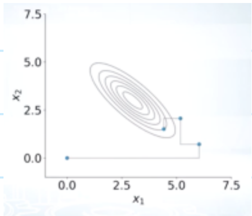
the point is sampled by considering $p(x_1^1 | x_2^0)$

· Let's say that our initial point is $(x_1, x_2) = (0, 0)$. Let's walk through how to sample the first point.

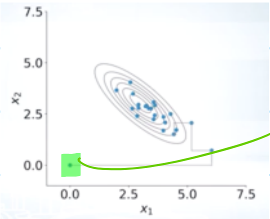


when $x_2 = 0$ fixed, x_1 distribution may look like that because of the given distribution

· Repeat the process, then we will get ;



→ keep going



we may want to burn it out !