Sampling Methods

Nipun Batra

September 12, 2023

IIT Gandhinagar

Topics

Rejection Sampling

1. Markov Chains
2. Importance Sampling
3. Gibbs Sampling
4. Markov Chain Monte Carlo

Rejection Sampling

- Let $p(x)$ be the target distribution from which we want to sample.

Rejection Sampling

- Let $p(x)$ be the target distribution from which we want to sample.
- Let $q(x)$ be a proposal distribution from which we can sample.

Rejection Sampling

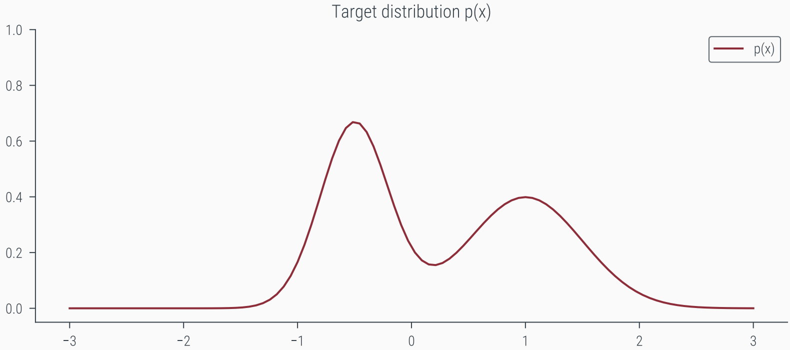
- Let $p(x)$ be the target distribution from which we want to sample.
- Let $q(x)$ be a proposal distribution from which we can sample.
- Let M be a constant such that $M \geq \frac{p(x)}{q(x)} \forall x$.

Rejection Sampling

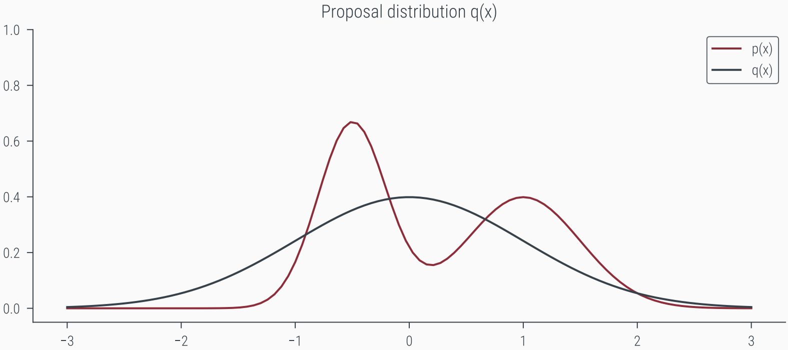
- Let $p(x)$ be the target distribution from which we want to sample.
- Let $q(x)$ be a proposal distribution from which we can sample.
- Let M be a constant such that $M \geq \frac{p(x)}{q(x)} \forall x$.
- Then, we can sample from $p(x)$ by sampling from $q(x)$ and accepting the sample with probability $\frac{p(x)}{Mq(x)}$.

Notebook: `rejection-sampling.ipynb`

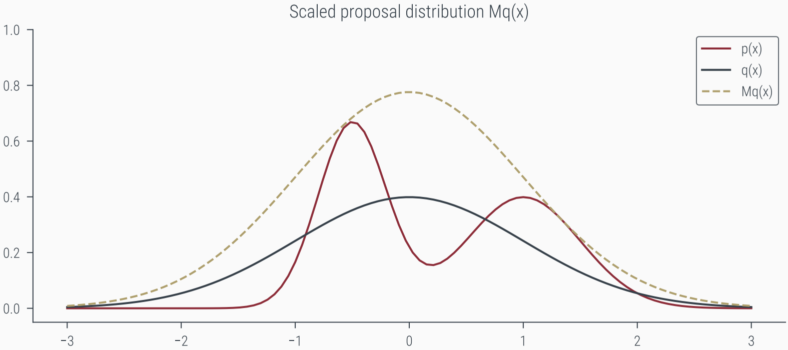
Rejection Sampling



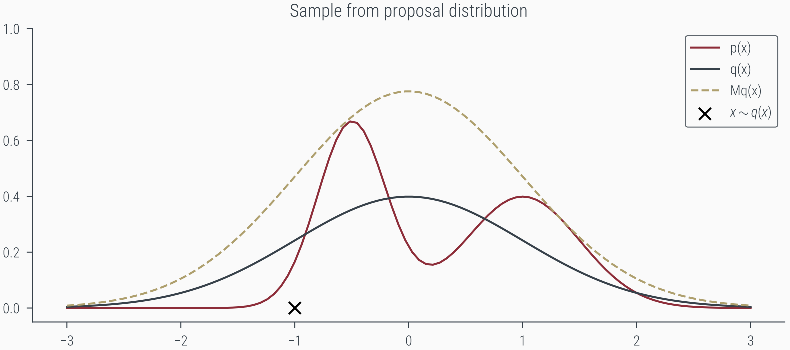
Rejection Sampling



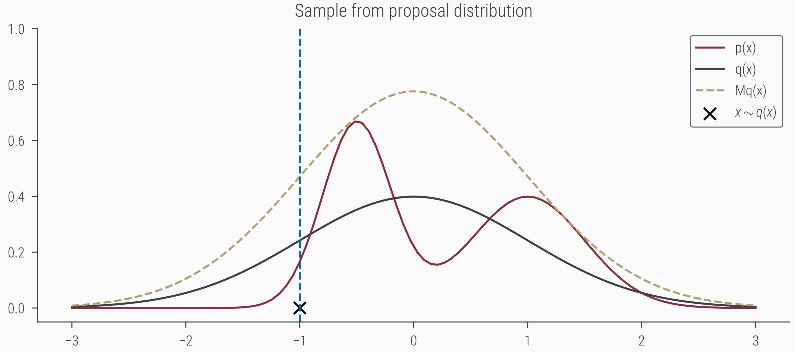
Rejection Sampling



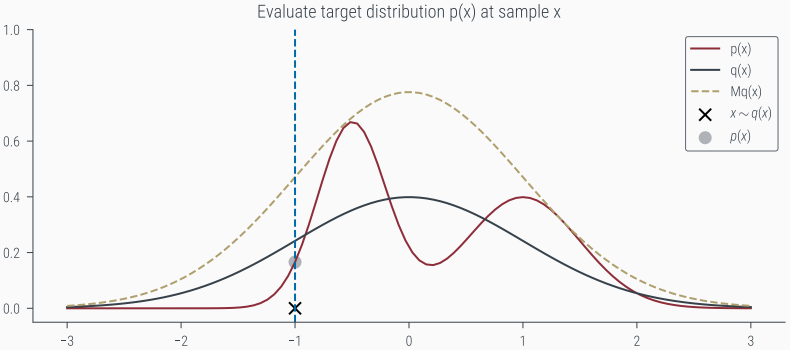
Rejection Sampling



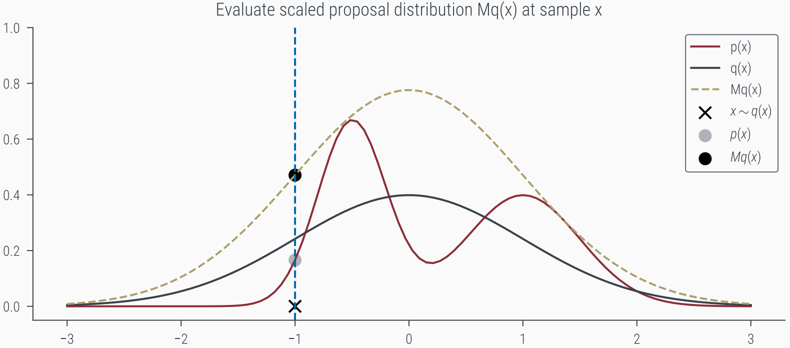
Rejection Sampling



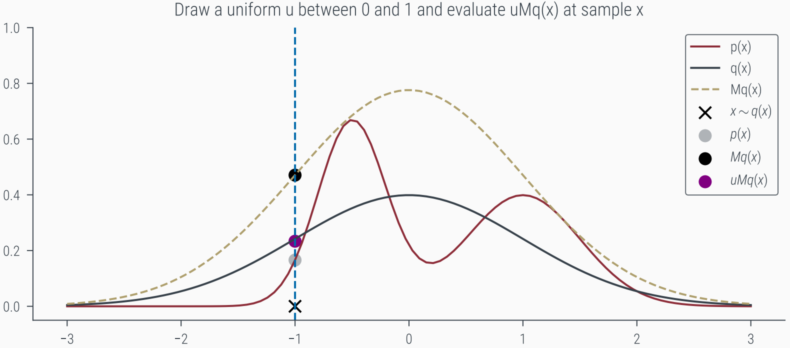
Rejection Sampling



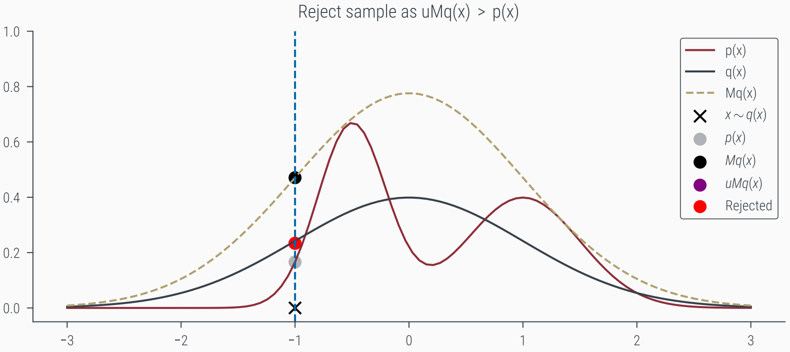
Rejection Sampling



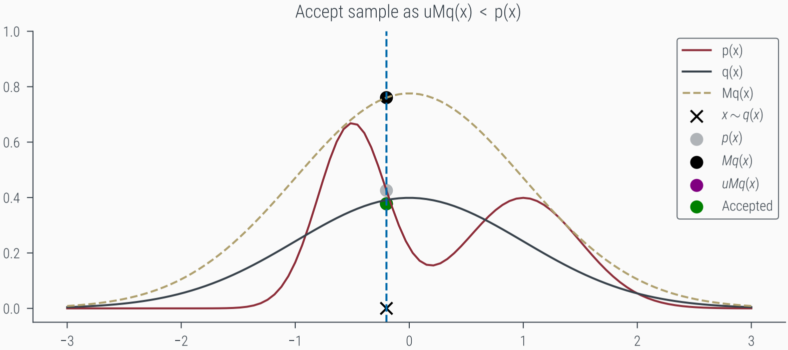
Rejection Sampling



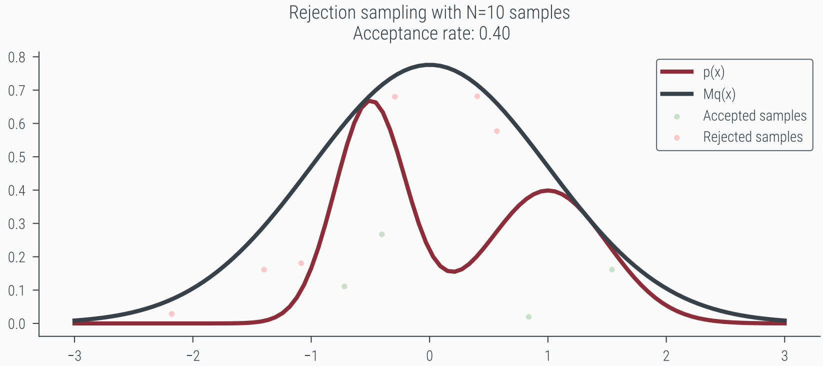
Rejection Sampling (Rejected Sample)



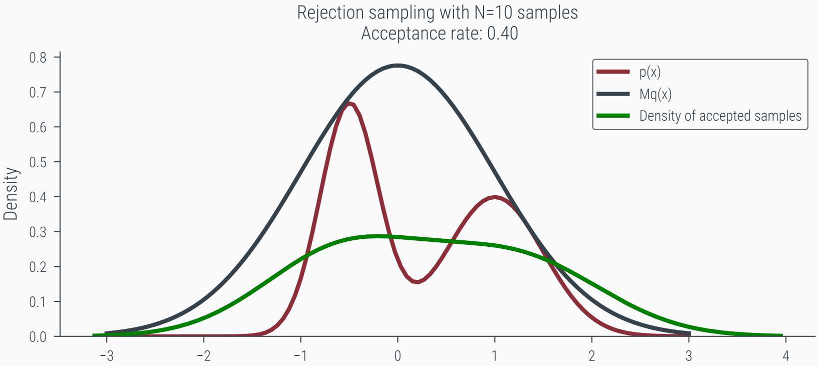
Rejection Sampling (Accepted Sample)



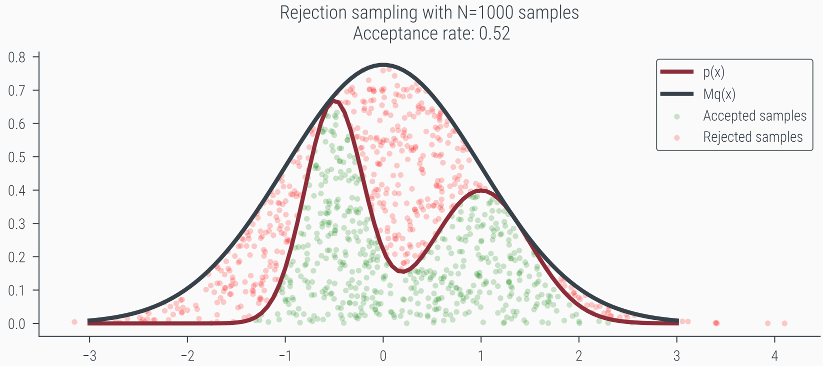
Rejection Sampling (10 samples)



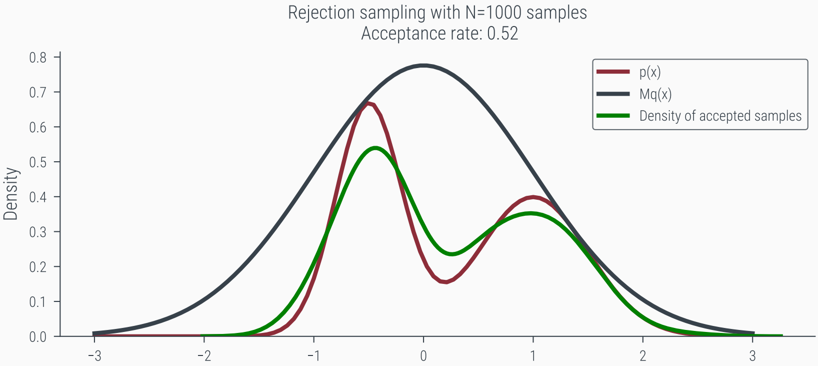
Rejection Sampling (10 samples) (KDE)



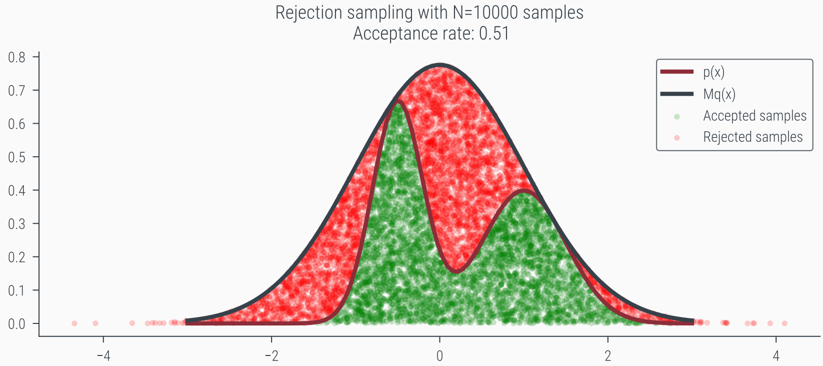
Rejection Sampling (1000 samples)



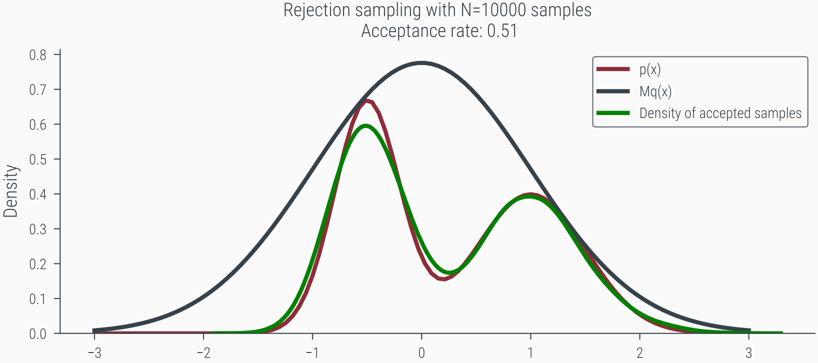
Rejection Sampling (1000 samples) (KDE)



Rejection Sampling (10000 samples)



Rejection Sampling (10000 samples) (KDE)



Proof of Rejection Sampling

Acceptance Probability $\alpha(x)$

$$\alpha(x) = \frac{p(x)}{Mq(x)} \tag{1}$$

Bayes Rule for Acceptance

$$P(Sample|Accept) = \frac{P(Accept|Sample)P(Sample)}{P(Accept)} \tag{2}$$

P(Sample)

We draw samples from $q(x)$, so $P(Sample) = q(x)$.

Proof of Rejection Sampling

Further, $P(Accept|Sample) = \alpha(x) = \frac{p(x)}{Mq(x)}$.

Finally, $P(Accept) = \int P(Accept|Sample)P(Sample)dSample = \int \alpha(x)q(x)dx = \frac{1}{M}\int p(x)dx = \frac{1}{M}$.

P(Accept)

$$P(Accept) = \frac{1}{M} \tag{3}$$

Thus, $P(Sample|Accept) = \frac{p(x)}{Mq(x)} \times \frac{q(x)}{1/M} = p(x)$.

Thus, we have shown that the samples we accept are distributed according to $p(x)$.

Rejection Sampling Completed Example

Note: Figures not on github.

Challenges with Rejection Sampling

- Rejection sampling is inefficient when the target distribution is very different from the proposal distribution.
- In this case, we will reject a lot of samples.
- This is a problem when sampling from high-dimensional distributions.
- Acceptance probability $\alpha(x)$ is very low.

Markov Chains

https://nipunbatra.github.io/hmm/

Global Optimization

Notebook: mcmc=optimization.ipynb

Importance Sampling

General Form

In rejection sampling, we saw that due to less acceptance probability, a lot of samples were wasted leading to more time and higher complexity to approximate a distribution.

Computing $p(x), q(x)$ thus seems wasteful. Let us rewrite the equation as:

$$\phi = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx$$
$$\sim \frac{1}{N}\sum_{i=1}^{N} f(x_i)\frac{p(x_i)}{q(x_i)} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)w_i$$

Here, $x_i \sim q(x)$. w_i is known as the importance(weight) of sample i.

However the normalization constant Z is generally not known to us. Thus writing:

$$p(x) = \frac{\tilde{p}(x)}{Z} \tag{4}$$

Now inserting this in earlier equations, we get:

$$\begin{aligned}\phi &= \frac{1}{Z}\int f(x)\tilde{p}(x)dx = \frac{1}{Z}\int f(x)\frac{\tilde{p}(x)}{q(x)}q(x)dx \\ &\sim \frac{1}{NZ}\sum_{i=1}^{N} f(x_i)\frac{\tilde{p}(x_i)}{q(x_i)} = \frac{1}{NZ}\sum_{i=1}^{N} f(x_i)w_i\end{aligned}$$

We know that:

$$\begin{aligned}Z &= \int_{\infty}^{\infty} \tilde{p}(x)dx = \int_{\infty}^{\infty} \frac{\tilde{p}(x)}{q(x)}q(x)dx \\ &= \frac{1}{N}\sum_{i=1}^{N} w_i\end{aligned}$$

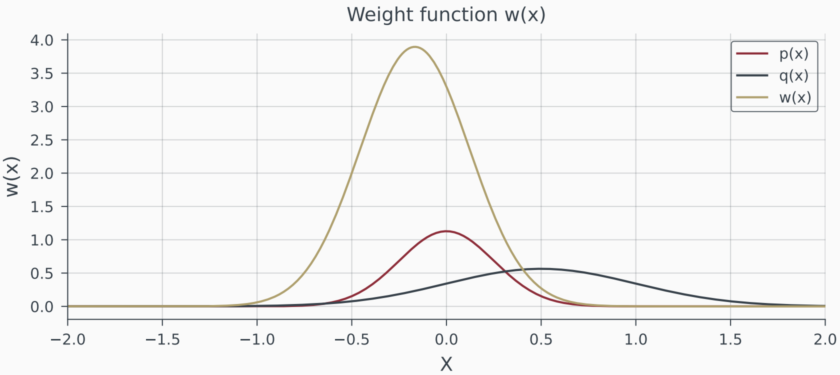
Substuting this value of Z in the equation above, we get:

$$\begin{aligned}\phi &= \frac{1}{N}\sum_{i=1}^{N} f(x_i)w_i = \frac{\sum_{i=1}^{N} f(x_i)w_i}{\sum_{i=1}^{N} w_i} \\ &= \sum_{i=1}^{N} f(x_i)W_i\end{aligned}$$

Here $W_i = \frac{w_i}{\sum_{i=1}^{N} w_i}$ are the normalized weights.

Limitations

- Recall that Var $\hat{\phi} = \frac{var(f)}{N}$. Importance sampling replaces $var(f)$ with $var(f\frac{p}{q})$. At positions where $p >>> q$, the weight can tend to ∞!



Gibbs Sampling

General Form

Suppose we wish to sample $\theta_1, \theta_2 \sim p(\theta_1, \theta_2)$, but cannot use:

- direct simulation
- accept-reject method
- Metropolis-Hasting

But we can sample using the conditionals i.e.:

- $p(\theta_1|\theta_2)$ and
- $p(\theta_2|\theta_1)$,

then we can use Gibbs sampling.

Suppose $\theta_1, \theta_2 \sim p(\theta_1, \theta_2)$ and we can sample from $p(\theta_1, \theta_2)$. We begin with an initial value (θ_1^0, θ_2^0), the workflow for Gibbs algorithm is:

1. sample $\theta_1^j \sim p(\theta_1|\theta_2^{j-1})$ and then
2. sample $\theta_2^j \sim p(\theta_2|\theta_1^j)$.

One thing to note here is that the sequence in which the theta's are sampled are not independent!

Bivariate Normal Example

Suppose

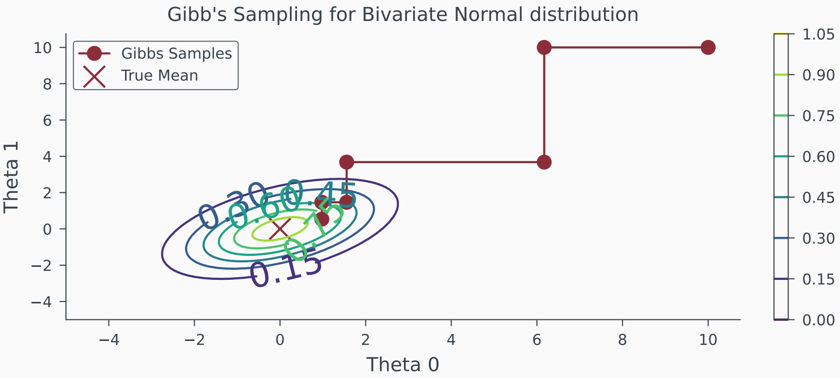
$\theta \sim N_2(0, \Sigma)$ and $\Sigma = \begin{matrix} 1 & \rho \\ \rho & 1 \end{matrix}$

Then, we have:

$\theta_1 | \theta_2 \sim N(\rho\theta_2, [1 - \rho^2])$

$\theta_2 | \theta_1 \sim N(\rho\theta_1, [1 - \rho^2])$ are the conditional distributions. The Gibbs sampling proceeds as follows:

Iteration	Sample θ_1	Sample θ_2
1	$\theta_1 \sim N(\rho\theta_2^0, [1 - \rho^2])$	$\theta_2 \sim N(\rho\theta_1^1, [1 - \rho^2])$
	.	
	.	
k	$\theta_1 \sim N(\rho\theta_2^{k-1}, [1 - \rho^2])$	$\theta_2 \sim N(\rho\theta_1^k, [1 - \rho^2])$


Gibb's Sampling for Bivariate Normal distribution
Gibbs Samples
True Mean
Theta 1
Theta 0
10
8
6
4
2
0
−2
−4
10
8
6
4
2
0
−2
−4
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00

Multivariate case

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$, the Gibbs workflow is as follows:
$\theta_1^j = p(\theta_1|\theta_2^{j-1}, \ldots, \theta_K^{j-1})$
$\theta_2^j = p(\theta_2|\theta_1^j, \theta_3^{j-1}, \ldots, \theta_K^{j-1})$

.

.

$\theta_k^j = p(\theta_k|\theta_1^j, \ldots, \theta_{k-1}^j, \theta_{k+1}^{j-1}, \ldots, \theta_K^{j-1})$

.

.

$\theta_K^j = p(\theta_K|\theta_1^j, \ldots, \theta_{K-1}^j)$
The distributions above are call the full conditional distributions.

Advantages

Gibbs sampling can be used to draw samples from $p(\theta)$ when:

- Other methods don't work quite well in higher dimensions.
- Draw samples from the full conditional distributions is easy, $p(\theta_k|\theta_{-k})$.

Markov Chain Monte Carlo

Limitations of basic sampling methods

- *Transformation based methods*: Usually limited to drawing from standard distributions.
- *Rejection and Importance sampling*: Require selection of good proposal distirbutions.

In high dimensions, usually most of the density $p(x)$ is concentrated within a tiny subspace of x. Moreover, those subspaces are difficult to be known a priori.

A solution to these are MCMC methods.

Markov Chain

- **Markov Chain**: A joint distribution $p(X)$ over a sequence of random variables $X = \{X_1, X_2, \ldots, X_n\}$ is said to have the Markov property if

$$p(X_i|X_1, \ldots, X_{i-1}) = p(X_i|X_{i-1})$$

  The sequence is then called a Markov chain.
- The idea is that the estimates contain information about the shape of the target distribution p.

Metropolis Hastings

- The basic idea is propose to move to a new state x_{i+1} from the current state x_i with probability $q(x_{i+1}|x_i)$, where q is called the proposal distribution and our target density of interest is $p(=\frac{1}{Z}\tilde{p})$.
- The new state is accepted with probability $\alpha(x_i, x_{i+1})$.
  - If $p(x_{i+1}|x_i) = p(x_i|x_{i+1})$, then $\alpha(x_i, x_{i+1}) = \min(1, \frac{p(x_{i+1})}{p(x_i)})$.
  - If $p(x_{i+1}|x_i) \neq p(x_i|x_{i+1})$, then $\alpha(x_i, x_{i+1}) = \min(1, \frac{p(x_{i+1})q(x_i|x_{i+1})}{p(x_i)q(x_{i+1}|x_i)}) = \min(1, \frac{\tilde{p}(x_{i+1})q(x_i|x_{i+1})}{\tilde{p}(x_i)q(x_{i+1}|x_i)})$
- Evaluating α, we only need to know the target distribution up to a constant of proportionality or without normalization constant.

Algorithm: Metropolis Hastings

1. Initialize x_0.
2. for $i = 1, \ldots, N$ do:
3. Sample $x^* \sim q(x^*|x_{i-1})$.
4. Compute $\alpha = \min(1, \frac{\tilde{p}(x^*)q(x_{i-1}|x^*)}{\tilde{p}(x_{i-1})q(x^*|x_{i-1})})$
5. Sample $u \sim \mathcal{U}(0,1)$
6. if $u \leq \alpha$:

   $x_i = x^*$

   else:

   $x_i = x_{i-1}$

Pop Quiz

How do we choose the initial state x_0?

Pop Quiz

How do we choose the initial state x_0?

1. Start the Markov Chain at an initial x_0.
2. Using the proposal $q(x|x_i)$, run the chain long enough, say N_1 steps.
3. Discard the first $N_1 - 1$ samples (called 'burn-in' samples).
4. Treat x_{N_1} as first sample from $p(x)$.

MCMC demo

https://chi-feng.github.io/mcmc-demo/app.html