Alexander V. Mantzaris (2008/2009)

Basic Introduction into-

MCMC (Markov Chain Monte Carlo)

*Considered to be one of the top ten most important algorithms ever

To see how great MCMC is, we will look at the motivation for sampling first, and then the methods used before its introduction

Motivation of sampling

- Many functions, equations, and distributions cannot be integrated analytically. For example:

$$e^{x^2}$$

Even such a simple function cannot be integrated, without numerical methods.

In the field of probability, integrals/summations (continuous/discrete respectively), are vital for calculating the expectation or expected values of distributions.

And this is essentially what marginalisation is all about as well.

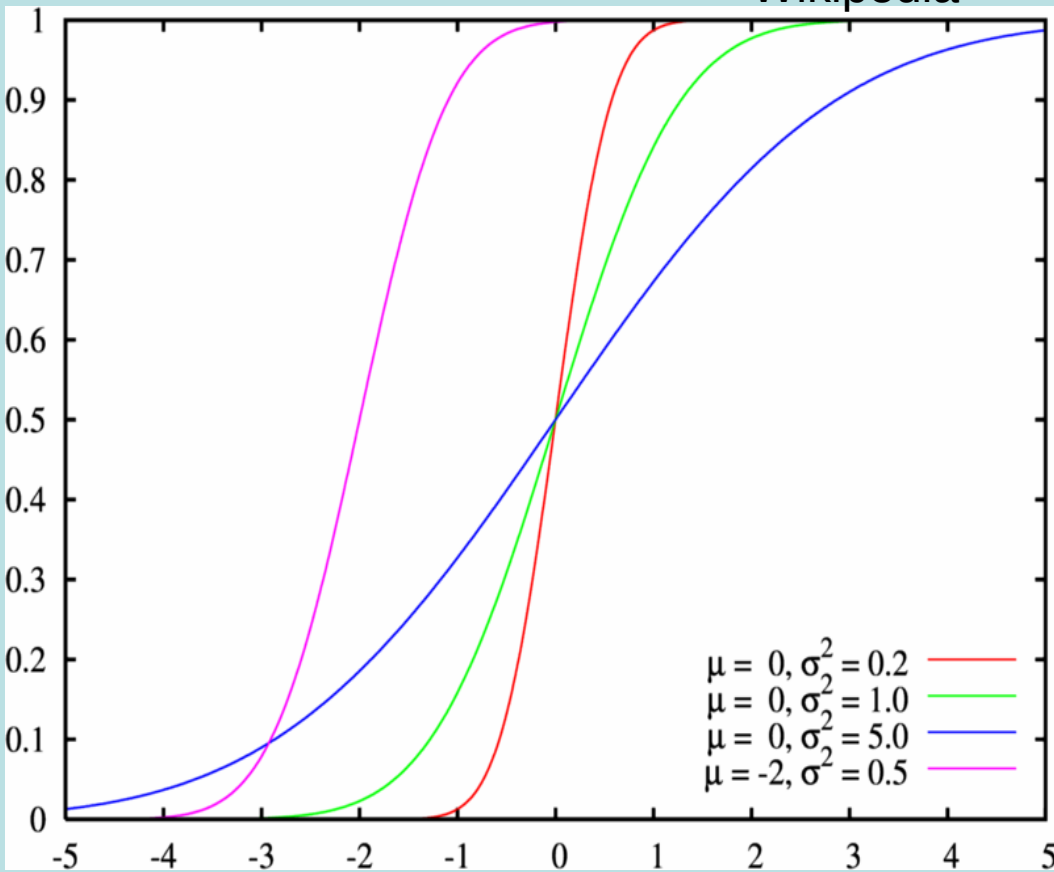
Motivation of sampling

- Drawing statistically consistent samples from a distribution- By drawing samples consistent to your distribution, you are effectively creating a simulation. Which can be useful for analyzing how fast the means, variances, etc, progress, and if the distribution's samples are required as input into another distribution.
- We sample when the Cumulative Distribution Function (CDF) cannot be found analytically. Sampling from the CDF is how Matlab comes up with random samples for distributions like the Gaussian… Next slide

Motivation of sampling

- With normalised distributions the integral=1, and sampling random numbers in [0,1] to then map them from the CDF y-axis to the CDF x-axis, gives us the point where a sample is chosen

Wikipedia



On the original distribution, where the height is larger the slope on the CDF is higher exposing it to greater chances of a random number getting chosen on the y-axis

Monte Carlo

- The most widely used place of Monte Carlo sampling is in Monte Carlo Integration
- Uses randomness to come up with a random variable estimates, similar to the gambling process in casinos, where the name derives.
- Defined domain along which we sample, uniformly (non-uniform is a prior harder to incorporate here), and since each point was taken with equal probability, the integral (expectation for random variables), is simply the average of the samples.
- Each sample is independently sampled

N random samples along the domain

$$x_1, x_2, \ldots, x_N$$

The average of the independent samples

$$E(f;N) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Monte Carlo

From the law of large numbers, as N increases the confidence in the estimate provide increases. And various measures exist to indicate the degree of convergence towards the true underlying values.

Numerical methods that have deterministic policies uniformly over the domain provide more certainty for distributions that have erratic changes in values. But in high dimensional space this causes a problem.

Monte Carlo Integration has convergence of: $1/\sqrt{N}$

Pros

-Simple and easy to implement

-Exploration does not get ‘stuck’ on local optima so easily as with dependent sampling

Cons

-Independent sampling may draw samples that ‘miss’ the most interesting parts of the domain

-Convergence tests are not as strong as those with dependent samples

-Does not extract statistically consistent samples

Importance Sampling

- Importance sampling exists mainly to address one of the greatest problems with Monte Carlo sampling in that many samples are redundant by falling into regions containing very little value. Which does have a purpose in its own right to bring down the global value, but we don’t need tons of estimates to do that, if we know that the surrounding areas of what is of interest are values close to zero we can easily in one step recalculate the average value of high density regions over a wider domain.
- High density regions are more important to sample.

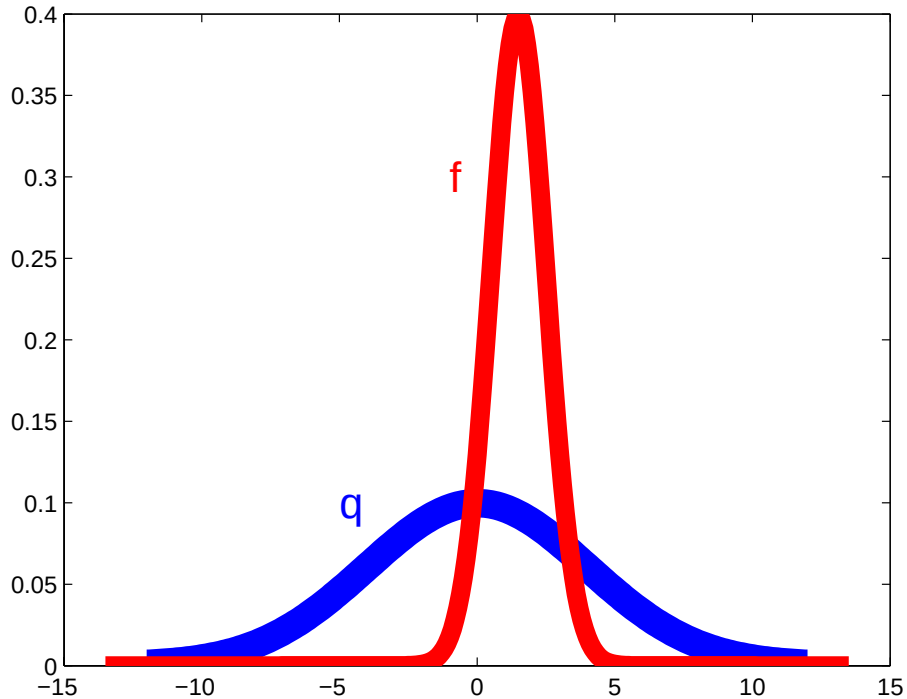
Importance Sampling

- This method focuses samples drawn according to a distribution we can sample from rather than a uniform distribution, to not exclude certain areas but drawn less samples from certain area

The distribution we wish to sample is, **f**.

The distribution we will use to draw samples from is **q**.

The distribution, **q**, is chosen to be one which can be sampled from, eg. Gaussian distribution. Its mean is placed to focus on the high density and variance to spread over the most important areas for **f**



Importance Sampling

- We use **q** to generate a series of independent samples $x_1, x_2, \ldots, x_N$

- Sample points are proportional to q(x) and importance weights are used to correct for the differences in q(x) and f(x): $w(x) = \frac{f(x)}{q(x)}$

In the case where q and f are unequal, w(x), acts as a correcting term to the number of times f(x) is sampled at that point.

Importance Sampling

- We use the weights to recalibrate, eg. If w(x)=0.5 then that means that q(x) is twice f(x) and will be generating twice as many samples as desired from this point, so multiplying all the samples of f(x) with w(x) corrects for the number of samples.
- To get the Monte Carlo expectation estimate from this method of sampling, we take the sum of the calibrated samples values divided by the sum of the weights as a normalisation for the weights to get the expected value

$$E(f) = \frac{\sum_{i=1}^{N} f(x_i) w(x_i)}{\sum_{k=1}^{N} w(x_k)}$$

Importance Sampling

Pros

-With good information on where the regions of high density for f(x) lie, q(x) can be placed suitably for effective sampling

-It can eliminate a large amount of the redundant samples and converge quickly

Cons

-Placing a suitable q(x) may not be possible and especially if the distribution, f(x) is multimodal

-Does not generate statistically consistent samples

Importance weights deviating strongly from 1 indicate that q(x) is not optimal for sampling f(x)

Rejection Sampling

As with Importance sampling a new distribution q(x) is used to sample from f(x).

With Importance sampling each and every sample drawn from q(x) was used to calculate the expected value of f(x) with the weights, but here a fraction of the samples will be rejected/discarded.

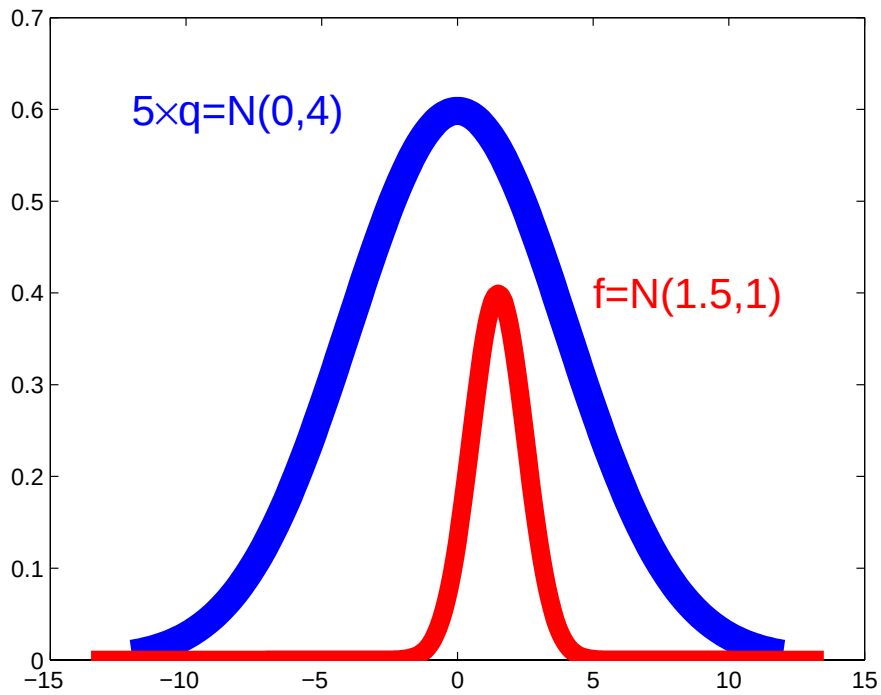
!!The rejection rate is proportional to the to how much larger q(x) is than f(x)

Rejection Sampling

- The reason q(x) as a probabilistic distribution is greater that f(x) in general is because we need to multiply q(x) by a constant k so, k*q(x)>f(x)

As samples are drawn from q(x), they are accepted with the frequency:

$$Accept = \frac{f(x_i)}{k \bullet q(x_i)}$$



Rejection Sampling

- Renormalisation is not required because the rejection rate incorporates it and the expectation is computed directly as an average

Pros

-Can generate statistically consistent samples according to the target distribution f(x)

-A good choice in q(x) and k can lead to fast convergence

Cons

-Many samples are discarded wasting computing time

-The choice of the constant k may not always be possible in highly peaked distributions. And the acceptence probability may too sensitive to k

Some Markov Chains Properties (Good to know)

Homogeneous: A Markov Chain is called homogeneous if the transition probabilities do not change in the progression of state transitions.

Ergodicity: As the number of iterations on the Markov Chain approach infinity

A distribution independent of the initial distribution is *invariant* to further simulation can be called the equilibrium distribution. An ergodic Markov chain can have only one equilibrium distribution.

Most Important: **Detailed Balance** (property of being reversible)

$$P(x, x')\pi(x) = P(x', x)\pi(x')$$

Markov Chain Monte Carlo

- This approach combines all the desired features in one method that is simple
- It generates statistically consistent samples from the target distribution without the help of another distribution and rejecting samples
- The expectation is calculated with samples drawn more proportionally from higher density regions without another distribution and the problems arising from poor weights is avoided too.
- The samples drawn are dependent and follow a Markovian framework which allows more robust convergence diagnostics to be used autocorrelation function, Gelman-Rubin statistics, etc.

MCMC

- How can we sample the target distribution f(x) in a way to obtain all of these benefits?
- If there was a proposal mechanism on the distribution that would operate as a valid Markov Chain, all of the properties mentioned arise.

$$A_{\min}\left(\frac{f(x')}{f(x)},1\right)$$

Between the comma separated values the smallest is chose as the probability to transition from point, x to point x’

MCMC

- This transition acceptance probability satisfies detailed balance!

$$\pi(x)A_{\min}\left(\frac{f(x')}{f(x)},1\right)=\pi(x')A_{\min}\left(\frac{f(x)}{f(x')},1\right)$$

We will change the notation here for a moment so that the concept at hand is evident; since f(x) is a probability distribution substitute it with p(x), and because the value of pi(x) is a free parameter that changes in the simulation, convergence occurs when it takes the value in proportion of the iteration as with the density at that point

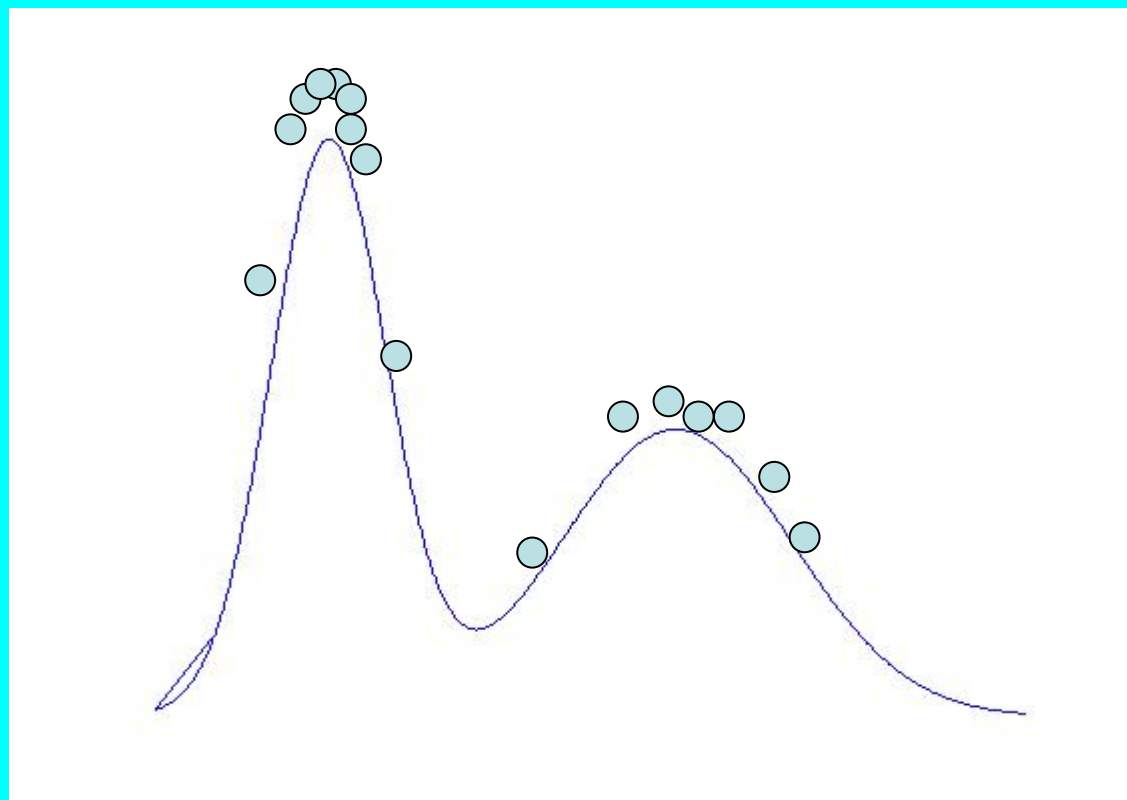
$$\left(p(x'),\,p(x)\right)=\left(p(x),\,p(x')\right)$$

MCMC

- *A point to be made is that before we speak of proposal mechanisms, this is the most simple form that hold when the proposal distribution from an x to x', is symmetric.*
- Starting from a randomly chosen initial value in the domain, as the new values are accepted the subsequent samples become independent of the initial point

MCMC

- Clusters towards the high density areas but can explore multimodal distributions given enough transitions; the low density regions between the local optima make it more unlikely to happen unfortunately



MCMC

- The percentage of accepted moves is important. Conventionally accepted percentages range from 30% to 70% and indicate that the chain has good mixing.
- Less than 30% shows that the sampler is not exploring much of the space and that few points will represent much of the simulation
- More than 70% can imply that the proposals were made too close to the current value so that the ratio of densities in the acceptance function is close to 1

MCMC

- To remove the bias from the initially chosen starting point, we consider a certain number of the first iterations to be discarded as the **Burnin** stage
- This number of iterations can be found in many ways, but the simplest is to use the autocorrelation function to see after how many samples does the correlation with the first samples decay to zero
- In the sample phase it is valid to tune parameters of the sampler (proposal distribution) since these samples will be discarded

MCMC

- Choosing a uniform interval whose mid point is set to the current value is one symmetric proposal mechanism
- The size of the interval is adapted (altered) during the burnin phase to a value that gives an acceptance percentage desired
- For distributions with boundaries eg. [0,inf) and the current value being 0.01 it is possible that the next proposed value falls on -0.49; then the value is simply *reflected* back onto the positive section. **Reflection** is used very frequently
- MCMC might not be suitable for parameter estimation in multimodal distributions

Hastings Ratio

- The proposal distribution may not be symmetric. The probability of being on a point x, and transitioning to x' may not be the same as the probability of being on point x' and transitioning back to x
- On a chess board if you have a king moving randomly (random walk), the boarder squares are visited less often. This is because those squares are not surrounded by as many squares as the ones in the center are. Eg. The corner square transitions into a non-border sq with p=1/3, where as the reverse move occurs with p=1/8.
- The same happens when proposing values from distributions like the dirichlet. Given the parameters and conditioning on the current value, the proposal probablities are not uniform

Hastings Ratio

- This ratio is used a factor for correcting for the bias in sampling of the proposal distribution, but it also has an active role. It can speed up convergence if proposals are made towards peak densities.
- The increases in the target distribution densities will likely outweigh the Hastings ratio leading to accepted transitions. Since the stationary distribution of the Markov chain has not been changed, a good proposal distribution may be more useful than a uniform proposal.

$$\frac{p(x, x')}{p(x', x)}$$

Metropolis-Hasting MCMC

$$A_{\min}\left(\frac{p(x,x')}{p(x',x)}\bullet\frac{p(x')}{p(x)},1\right)$$

Model Selection

$$P(M_k \mid D) = \frac{P(M_k)P(D \mid M_k)}{\sum M_i}$$

Since the denominator is too large to compute, we use MCMC to produce/generate statistically valid samples which we average over for the posterior distribution

This holds for other parameters which we assumed have analytically been integrated out here, eg. Conjugate priors

Bayes Factors

$$k = \frac{P(D \mid M_1)}{P(D \mid M_2)} = \frac{\int P(\theta \mid M_1) P(D \mid \theta, M_1) d\theta}{\int P(\theta \mid M_2) P(D \mid \theta, M_2) d\theta}$$

The higher the value of k, the stronger the evidence for supporting model 1.

There are motivations to make this equivalent to many frequentist approaches, likelihood ration tests, t-tests, etc

Simulated Annealing

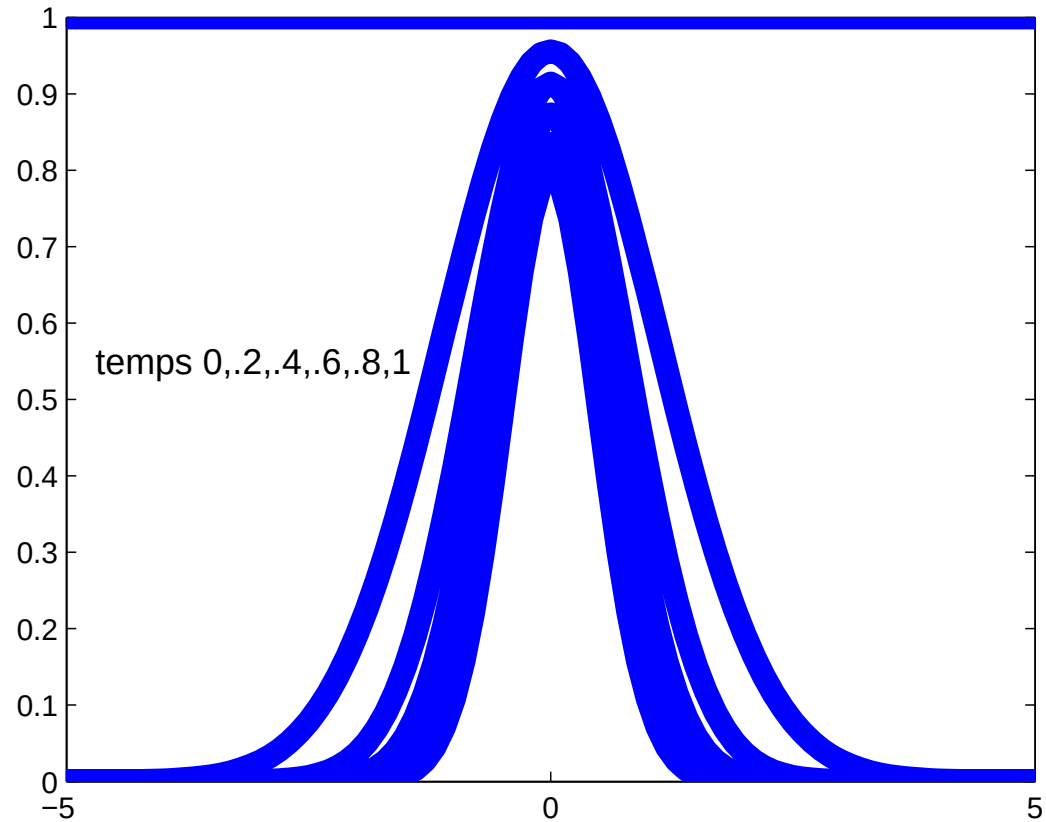
- Any number raised to the power of 0 is 1
- Any number raised to the power of 1 is itself
- Statistical physicists have been using models which interpret exponents on distributions for thermodynamics as temperatures
- Power 0 is hot, and power 1 is cold
- At power 0 the distribution is flat sitting at value 1
- At power 1 the distribution has the original form

Simulated Annealing

- The temperature schedule is the progression of the exponents values 0,0.2,0.4…1
- The schedule does not have to be linear, some distributions will develop difficult local optima to traverse within a certain range of temperatures
- This is very helpful to use in the burnin period for very peaky distributions, and multimodal distributions to have samples begin on high density regions

Simulated Annealing

Gaussian distribution with mean 0 and var 0.5 and it is evident more steps in the interval [0,0.2] were needed



Correspondence

- If you have any questions I would like to use my personal email address:

  dog_of_thunder

  @hotmail.com