MTMD03 Monte Carlo Techniques and Particle Filters

Chapter 1 - Introduction to Monte Carlo Sampling and Integration

Dr Sarah Dance

Spring 2011

Contents

1 What is a Monte Carlo method?

Monte Carlo methods

- Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to compute their results.
- Monte Carlo methods are often used in simulating physical and mathematical systems.
- Because of their reliance on repeated computation of random or pseudo-random numbers, these methods are most suited to calculation by a computer
- They tend to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm.

Monte Carlo methods and data assimilation

- The class of Monte Carlo methods is rather general
- In data assimilation we concentrate on Monte Carlo methods to sample from *prediction, filtering* and *smoothing* probability densities (pdfs) and estimate their associated features.

- e.g., the expectation

$$I\left[f\right]=\int f(\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_k)d\mathbf{x}_k$$

- We will begin by considering how we can approximate integrals such as this one using statistical sampling.

2 A reminder of probabilistic notation

A reminder of some probability theory

Recall that the *probability measure* or *probability distribution* of a random variable X in an arbitrary domain Ω is a measure function P such that

$$P(D)=Pr\{X\in D\},$$

for any measureable set $D\subset\Omega$. In particular $P(\Omega)=1$.
The corresponding *probability density function* or *pdf* is defined as the *Radon-Nikodym* derivative

$$p(x)=\frac{dP}{d\mu}(x),$$

which is simply the function p that satisfies

$$P(D)=\int_D p(x)d\mu(x).$$

Note that p depends on the measure μ used to define it.

Example - uniform distribution

For many of the distributions we deal with in this module we will not need to use the full power of Lebesgue measure.

For example, the uniform distribution for the interval $[a,b]$ has pdf

$$p(x)=\left\{\begin{array}{ll}\frac{1}{b-a} & \text{for } a\le x\le b,\\ 0 & \text{for } x<a \text{ or } x>b.\end{array}\right.,$$

and cumulative distribution function

$$P(x)=Pr(X\le x)=\left\{\begin{array}{ll}0 & \text{for } x<a\\ \frac{x-a}{b-a} & \text{for } a\le x\le b\\ 1 & \text{for } x\ge b\end{array}\right..$$

So in this case $p(x)=\frac{dP}{dx}(x)$.

Expectation and variance

The *expected value*, or *expectation* of a random variable $Y=f(X)$ is defined as

$$E\left[Y\right]=\int_\Omega f(x)p(x)d\mu(x),$$

while its *variance* is

$$V[Y]=E\left[(Y-E(Y))^2\right].$$

Note that we often write the square root of the variance as the standard deviation $\sigma[Y]=\sqrt{V[Y]}$. We will assume that the expectation and variance exist (i.e., the integrals are finite).

Exercise

Find the mean value and variance for the uniform distribution (take f(X)=X).

3 Monte-Carlo integration

Monte Carlo quadrature formula

For a random variable with a density $p(x)$, the expectation of a function $f(x)$ is

$$E[f] = I[f] = \int f(x)p(x)dx.$$

For a sequence of IID (independent, identically distributed) random variables $\{X_n\}_{n=1}^{N} \sim p$ then an empirical approximation to the integral is

$$I_N[f] = \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

Monte Carlo convergence

By the Strong Law of Large Numbers, the Monte Carlo approximation converges almost surely,

$$I_N[f] \overset{a.s.}{\to} I[f]$$

in other words

$$Pr\left(\lim_{N\to\infty} I_N[f] = I[f]\right) = 1.$$

Unbiased approximation

It is straightforward to show that the approximation is unbiased, i.e.

$$E\left[I_N[f]\right] = I[f]$$

for any N (see problem sheet).

N.B. Note that this statement is true for an average over many realizations - we would not expect it to hold for an individual realization.

Convergence rates

Define the Monte Carlo integration error as

$$\varepsilon_N[f] = I[f] - I_N[f],$$

so that the bias is $E\left[\varepsilon_N[f]\right]$ and the RMSE (root mean square error) is

$$E\left[[\varepsilon_N[f]^2]\right]^{1/2} = \sigma\left[\varepsilon_N[f]\right].$$

Let $\xi_i = I[f] - f(X_i)$.

Exercise: Show that

$$\begin{aligned} E[\xi_i] &= 0 \\ E[\xi_i^2] &= V[f] \\ E[\xi_i\xi_j] &= 0 \text{ if } i \neq j. \end{aligned}$$

Now consider the sum,

$$\frac{1}{N}\sum_{n=1}^{N} \xi_i = \varepsilon_N.$$

Its variance is

$$\begin{aligned}
E[\varepsilon_N^2] &= E\left[\frac{1}{N^2}\left(\sum_{n=1}^{N}\xi_n\right)^2\right] \\
&= \frac{1}{N^2}\left\{E\left[\sum_{n=1}^{N}\xi_n^2\right] + E\left[\sum_{n=1}^{N}\sum_{m\neq n}\xi_n\xi_m\right]\right\} \\
&= \frac{1}{N^2}\left\{\sum_{n=1}^{N}V[f] + 0\right\} \\
&= \frac{1}{N}V[f].
\end{aligned}$$

Therefore

$$\sigma\left[\varepsilon_N[f]\right] = \sigma[f]N^{-1/2}$$

which shows that the RMSE is of size $O(N^{-1/2})$. For large N we can use the Central Limit Theorem (CLT) to obtain a bound on the error as

$$\lim_{N\to\infty} Pr\left\{a < \frac{\sqrt{N}}{\sigma[f]}\varepsilon_N < b\right\} = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}dx$$

where the RHS is the cumulative standard normal distribution ($N(0,1)$).

This result does not provide an absolute upper bound on the error; rather it says that the error is of a certain size with some probability.

Comparison of Monte Carlo integration with Grid based methods

Q. How can a random array be better than a grid for integrating a function?

- The number of points required for grid based quadrature depends on the dimension, whereas Monte Carlo convergence rate is $O(N^{-1/2})$ regardless of the dimension or the smoothness of the integrand.
- Monte Carlo is simple - only two basic operations are needed: sampling and point evaluation
- Monte Carlo is easy to set up in any geometry, whereas designing a grid for some domains can be very complex

4 Random variable generation

4.1 Pseudo-random numbers

Pseudo random numbers

- The "random" numbers generated by a computer are not random
- Instead they are a deterministic sequence that has many of the properties of random number sequences
- A pseudo random number generator can be started from an arbitrary starting state, often called a *seed state*
- It will always produce the same sequence thereafter when initialized from that state.
- This is useful for carrying out controlled experiments.
- The maximum length of the sequence before it begins to repeat is known as the *period.*

Pseudo-Random number generators

Pseudo-random number generation is a well developed subject, but it is worth being aware that some of the earlier algorithms may have problems, such as

- Shorter than expected periods for some seed states
- Lack of uniformity of distribution for large amounts of generated numbers
- Correlation of successive values

Mersenne Twister

- In matlab (versions 7.4 and later) the default algorithm for generation of uniform [0,1] samples is the *Mersenne Twister* by Nishimura and Matsumoto
- This method generates double-precision values in the closed interval $[2^{-53}, 1-2^{-53}]$, with a period of $(2^{19937}-1)/2$ and is very efficient.

4.2 Generating samples from other distributions

Non-uniform variables

Standard pseudo-random number generators usually only generate samples from the uniform [0,1] distribution.

We will now consider several ways to generate samples from non-uniform distributions.

Transformation method

The *transformation method* is a method for producing a general random variable X through transformation of a uniform random variable.

Let X be a random variable, and let $F(x)$ be the distribution function for X so that

$$Pr\{X \leq x\} = F(x).$$

Assume that $F(x)$ is a strictly increasing function and that F^{-1} exists.

Define $U = F(X)$, so U is a random variable with values in $[0,1]$.

We find that the distribution function for U satisfies

$$\begin{aligned} Pr\{U \leq u\} &= Pr\{F(X) \leq u\} \\ &= Pr\{X \leq F^{-1}(u)\} \\ &= F(F^{-1}(u)) = u \text{ for } 0 \leq u \leq 1. \end{aligned}$$

Hence U has the uniform distribution on $[0,1]$

Thus to generate a sample, X

- Generate a sample $U \sim$ unif$[0,1]$
- Set $X = F^{-1}(U)$

This formulation can be convenient since it is explicit, but it may not be easy to compute F^{-1} in closed form.

Exercise

The negative exponential cumulative distribution function is given by

$$F(x) = 1 - e^{-\lambda x},$$

where $\lambda > 0$ is the rate parameter for the distribution. Find $F^{-1}(U)$.

Box-Muller method for Gaussian variables

For the Gaussian variable, special transformations are a useful alternative to the transformation method because here

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Thus $F(x)$ is not of closed form. Hence computing F^{-1} is not straightforward.

Suppose (X, Y) are a pair of Gaussian random variables.

Consider a change from Cartesian to polar coordinates, i.e.,

$$(x, y) = (r\cos\theta, r\sin\theta)$$

so that

$$dxdy = rdrd\theta.$$

Thus the corresponding transformation of the cdf is

$$\begin{aligned} F(x, y) &= \frac{1}{2\pi} \int_{-\infty}^{x} \int_{-\infty}^{y} e^{-(s^2+t^2)/2} dsdt \\ &= \frac{1}{2\pi} \int_{\rho=0}^{r} \int_{\phi=0}^{\theta} e^{-\rho^2/2} \rho d\rho d\phi \end{aligned}$$

We can write this as a product of two cdfs:

$$H(\theta) = \frac{1}{2\pi} \int_{\phi=0}^{\theta} d\phi = \frac{\theta}{2\pi}$$

from which it is clear that $U = \Theta/2\pi \sim$ unif$[0, 1]$.

$$G(r) = \int_{\rho=0}^{r} e^{-\rho^2/2} \rho d\rho = 1 - e^{r^2/2},$$

and r can therefore be easily sampled using the transformation method.

Exercise

Show that if $V \sim$ unif$[0, 1]$ then

$$R = \sqrt{-2\log(1 - V)}.$$

Thus, normal variables X, Y can be obtained from uniform variables U, V as

$$X = R\cos\Theta,\ Y = R\sin\Theta,$$

where

$$R = \sqrt{-2\log(1 - V)},\ \Theta = 2\pi U.$$

Accept-reject

- Another way of generating samples from a given density is based on a probabilistic approach.
- The accept-reject method shows the power and flexibility of stochastic methods
- MCMC methods are based on similar ideas so you will see more of this later in the module

We introduce the method by way of an example...

Example

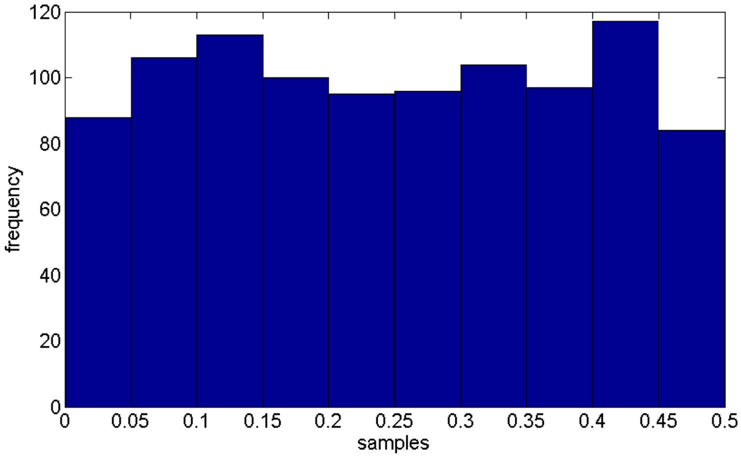
Idea: Generate $X \sim$ unif$[0, \frac{1}{2}]$ using $U \sim$ unif$[0, 1]$

Try the algorithm:

Step 1 Generate $U \sim$ unif$[0, 1]$

Step 2 If $U < \frac{1}{2}$ then set $X = U$, else go to Step 1.

Results for N=1000



The method seems to behave in the right way. How can we show this is in fact correct?

General Accept-Reject Method

Suppose we wish to sample $Y \sim p(x)$, and we know how to sample $U \sim$ unif$[0, 1]$ and $X \sim q(x)$ such that

$$p(x) \leq cq(x),$$

for all x for some positive constant c.
Algorithm:

Step 1 Generate $X \sim q(x),\ U \sim$ unif$[0, 1]$

Step 2 If $U \leq \frac{p(X)}{cq(X)}$ then set $Y = X$, else go to Step 1.

Proof

We will show that the conditional distribution

$$Pr\left\{X < y | U \leq \frac{p(X)}{cq(X)}\right\} = Pr\{Y \leq y\}.$$

Consider the LHS. By definition

$$\frac{Pr\left\{X < y, U \leq \frac{p(X)}{cq(X)}\right\}}{Pr\left\{U \leq \frac{p(X)}{cq(X)}\right\}} = \frac{\int_{x=-\infty}^{y} \int_{u=0}^{p(x)/cq(x)} q(x) du dx}{\int_{x=-\infty}^{\infty} \int_{u=0}^{p(x)/cq(x)} q(x) du dx}.$$

Carrying out the u-integration first, we find that this expression simplifies to

$$\frac{\frac{1}{c}\int_{-\infty}^{y} p(x)dx}{\frac{1}{c}\int_{-\infty}^{\infty} p(x)dx} = \int_{-\infty}^{y} p(x)dx = Pr\{Y \leq y\},$$

as required. □

Unconditional acceptance probability

The *unconditional acceptance probability* is the probability that a sample is accepted at any given iteration:

$$Pr\left\{U \leq \frac{p(X)}{cq(X)}\right\} = \int_{x=-\infty}^{\infty} \int_{u=0}^{p(x)/cq(x)} q(x)dudx = \frac{1}{c}.$$

If c is low, fewer samples are rejected, and the required number of samples for the target distribution is obtained more quickly.

Because c must be no less than the maximum of $p(x)/q(x)$, the unconditional acceptance probability is higher the less that ratio varies. Rejection sampling can lead to a lot of unwanted samples being taken if the function being sampled is highly concentrated in a certain region, for example a function that has a spike at some location.

5 Controlling the Monte-Carlo variance

Variance reduction

In Monte Carlo integration, we have seen that the error ε and the number of samples N are related by

$$\begin{aligned} \varepsilon &= O\left(\sigma N^{-1/2}\right) \\ N &= O\left(\frac{\sigma^2}{\varepsilon^2}\right). \end{aligned}$$

The computational time required for the method is proportional to the size of $N = O\left(\frac{\sigma^2}{\varepsilon^2}\right)$ so that the computational time grows rapidly as the desired accuracy is tightened.

One way to accelerate the method is to reduce the variance σ^2 by transforming the integrand. In this section we describe a few methods for variance reduction. (There are more, see Caflisch, 1998, section 4).

5.1 Antithetic Variables

Antithetic variables

Consider the variance for the sum of two variables

$$\begin{aligned} V[Y_1 + Y_2] &= E\left[(Y_1 + Y_2 - E[Y_1 + Y_2])^2\right] \\ &= E\left[(\{Y_1 - E[Y_1]\} + \{Y_2 - E[Y_2]\})^2\right] \\ &= V[Y_1] + V[Y_2] + 2E\left[(Y_1 - E[Y_1])(Y_2 - E[Y_2])\right]. \end{aligned}$$

If Y_1 and Y_2 are iid. then

$$V[Y_1 + Y_2] = 2V[Y_1].$$

With antithetic variables we pick Y_1 and Y_2 (not iid) such that $V[Y_1] = V[Y_2]$ but that the covariance $\text{cov}(Y_1, Y_2) = E\left[(Y_1 - E[Y_1])(Y_2 - E[Y_2])\right] < 0$, resulting in a reduction in variance.

Example

Suppose we wish to estimate the integral

$$I[f] = \int_0^1 e^x dx = e - 1.$$

We first observe that

$$I[f] = E[e^U]$$

with $U \sim \text{unif}[0, 1]$, so the standard Monte Carlo quadrature estimate would be

$$I_N[f] = \frac{1}{N} \sum_{n=1}^{N} e^U.$$

This has variance

$$V[I_N[f]] = V[e^U]$$

To apply antithetic variables we take $U_m \sim \text{unif}[0,1]$ and $V_m = 1 - U_m$.
Exercise: show that $V_m \sim \text{unif}[0,1]$.

We take our new Monte Carlo estimate of the integral as

$$I'_N[f] = \frac{1}{2N} \sum_{n=1}^{N} \left\{ e^{U_n} + e^{V_n} \right\}.$$

The covariance

$$\begin{aligned} cov\left[e^U, e^V\right] &= E\left[\left(e^U - I[f]\right)\left(e^V - I[f]\right)\right] \\ &= E\left[e^U e^{1-U} - I[f]^2\right] \\ &= e - I[f]^2 = -(e-1)^2 < 0. \end{aligned}$$

Thus

$$V[I'_N[f]] < V[I_N[f]].$$

Remarks

The choice of relationship between the initial variables and the extra variables will depend on the underlying distribution. For example

- U and $1 - U$ are a good choice for the standard uniform distribution
- X and $-X$ are a good choice for the standard normal distribution

It may be more efficient (especially in higher dimensions) to explore space using simplexes rather than just pairs of variables (and this may also help to reduce errors in the higher moments)e.g. the unscented transform.

5.2 Stratified sampling

Stratified sampling

Stratification combines the benefits of a grid with those of random variables.

Example - simplest case - regular grid with uniform density in 1D.

Split the integration domain $\Omega = [0,1]$ into M equally sized pieces Ω_k, so

$$\Omega_k = \left[\frac{k-1}{M}, \frac{k}{M}\right],$$

and $|\Omega_k| = 1/M$.(sketch) For each k sample $N_k = N/M$ points $\{X_{i_k}\}$ uniformly distributed in Ω_k. Then the stratified quadrature formula is

$$I_N = \frac{1}{N} \sum_{k=1}^{M} \sum_{i_k=1}^{N_k} f(X_{i_k}),$$

i.e., just the sum of the quadratures over each subset. Define the averages over each k such that

$$\overline{f_k} = \frac{1}{|\Omega_k|} \int_{\Omega_k} f(x)dx.$$

Then the Monte Carlo quadrature error for this stratified sum is

$$\varepsilon \simeq N^{-1/2}\sigma_s$$

where

$$\sigma_S^2 = \sum_{k=1}^{M} \int_{\Omega_k} \left(f(x) - \overline{f_k}\right)^2 dx$$

In this example, stratified MC quadrature always beats unstratified MC, since

$$\sigma_s \leq \sigma.$$

A proof of this result is straightforward (see problem sheet).

Stratified sampling general framework

Split the integration region into M pieces Ω_k such that

$$\Omega = \cup_{k=1}^{M} \Omega_k.$$

Take N_k samples in each piece such that

$$\sum_{k=1}^{M} N_k = N.$$

Choose samples $X_{i_k} \in \Omega_k$ such that $X_{i_k} \sim p_k(x)$ where

$$p_k(x) = p(x)/\overline{p_k}, \text{ and } \overline{p_k} = \int_{\Omega_k} p(x)dx.$$

The stratified quadrature formula is given by

$$I_N[f] = \sum_{k=1}^{M} \frac{\overline{p_k}}{N_k} \sum_{i_k=1}^{N_k} f(X_{i_k}).$$

Stratification always lowers the integration error if the distribution of points is *balanced.* The balance condition is that

$$\overline{p_k}/N_k = 1/N,$$

i.e., the number of points in a subset is proportional to its weighted size $\overline{p_k}$. Then the Monte Carlo quadrature error for this stratified sum is

$$\varepsilon \simeq N^{-1/2}\sigma_s$$

where

$$\sigma_S^2 = \sum_{k=1}^{M} \int_{\Omega_k} \left(f(x) - \overline{f_k}\right)^2 p(x)dx$$

and

$$\overline{f_k} = \frac{1}{\overline{p_k}} \int_{\Omega_k} f(x)p(x)dx.$$

It can be easily shown that

$$\sigma_s \leq \sigma,$$

so stratification always lowers the integration error.

5.3 Importance sampling

Importance sampling

Importance sampling will be a key component of the particle filter later in the module.

The idea is to rewrite a simple integral by introducing a new density function $\pi(x)$,

$$I[f] = \int f(x)dx = \int \frac{f(x)}{\pi(x)}\pi(x)dx.$$

Note that π must have the same (or larger) support as f.

Now think of this as an expectation with respect to the density π so that the Monte Carlo estimate is

$$I_N[f] = \frac{1}{N}\sum_{n-1}^{N}\frac{f(X_n)}{\pi(X_n)},\ X_n \sim \pi.$$

The resulting error is

$$\varepsilon_N[f] = I[f] - I_N[f] \simeq \sigma_p N^{-1/2},$$

where

$$\sigma_p^2 = V_\pi\left[\frac{f}{\pi}\right].$$

So importance sampling will reduce the quadrature error if

$$\sigma_p \ll \sigma$$

Using the definition of variance it is clear that the theoretical best case is when

$$\sigma_p^2 = 0\ \Rightarrow \frac{f}{\pi} = I[f],\ \text{or}\ \pi = \frac{f}{I[f]}.$$

Of course it is not practical to choose π in this way since the formula involves the very integral $I[f]$ that we are trying to calculate. However, it does give some insight into what importance sampling does!

- $\pi = f/I[f]$ is simply the normalized density defined by f (Sketch).
- Before applying IS we can think of the law of X as being the uniform distribution.
- In IS, the law of X is redistributed so that regions with high values of f are sampled more frequently than regions with low values of f.
- Hence the name “Importance sampling”.

6 Further reading

Further reading

Basic probability

Feller (1971) An introduction to probablility theory and its applications, Wiley (main library call number 519.2-FEL)

Grimmett and Stirzaker(1992) Probability and random processes. Oxford: Clarendon Press (main library call number 519.2-GRI)

Monte Carlo integration, sampling and variance reduction

R.E. Caflisch (1998) Acta Numerica, pp 1-49.
(particularly sections 1-4)

Pseudo-Random number generators

For more information on the Mersenne Twister see `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`

The unscented transform used in ensemble data assimilation

Wang et al (2004) Which is better, an ensemble of positive negative pairs or a centered simplex ensemble?, Mon. Weather Rev. 132, 1590-1605. (see especially appendix A).