SUMMER SCHOOL IN STATISTICS 2014
VIETNAM

Introduction to particle methods
and
Importance Sampling and Splitting

Agnès LAGNOUX

lagnoux@univ-tlse2.fr
http://perso.math.univ-toulouse.fr/lagnoux/enseignements/

INSTITUT
de MATHEMATIQUES
de TOULOUSE

Introduction

What is a rare event ?

When does it occur ? In which domains ?

- biology
- reliability
- telecommunications
- aeronautics...

How to study the probability of such events ?

- statistical analysis based on extreme value distributions but needs a long observation period ;
- modeling based on Monte Carlo simulations.

Introduction

Nevertheless, crude simulation is impracticable for estimating such small probabilities : to estimate probabilities of order 10^{-10} with acceptable confidence would require the simulation of at least 10^{12} events (which corresponds to the occurrence of only one hundred rare events).

Introduction

Nevertheless, crude simulation is impracticable for estimating such small probabilities : to estimate probabilities of order 10^{-10} with acceptable confidence would require the simulation of at least 10^{12} events (which corresponds to the occurrence of only one hundred rare events).

To overcome these limits, fast simulation techniques are applied :

- Importance Sampling (IS) that relies on a change of probability of the underlying process and requires a deep knoweledge of the process studied ;
- Importance Splitting (ISp) that relies on a partitioning of the state space ;
  Systems of particles introduced by P.Del Moral
  RESTART introduced by J. and M. Villén Altamirano
  complex but rely on a common simple basis ISp.

OUTLINE OF LECTURE I

1. Introduction
   - Description of the Monte Carlo method
   - Limits and convergence
2. Importance Sampling (IS)
3. Other methods to reduce the variance
   - Control variables
   - Antithetic variables
   - Method of stratification
   - Average value or conditioning

INTRODUCTION

Introduction - Monte Carlo method

Assume we want to calculate

$$I = \int_{[0,1]^d} g(u_1, \ldots, u_d) du_1 \ldots du_d.$$

Remark that

if we set $X = g(U_1, \ldots, U_d)$ where $U_1, \ldots, U_d$ are uniform iid rv on $[0, 1]$, we get

$$I = \mathbb{E}(X) = \mathbb{E}(g(U_1, \ldots, U_d)).$$

Introduction - Monte Carlo method

For the simulation, assume that $(U_i,\ i \geqslant 1)$ is a sequence of iid uniformly distributed rvs over $[0,1]$ and set

$$X_1 = g(U_1, \ldots, U_d),$$
$$X_2 = g(U_{d+1}, \ldots, U_{2d}) \ \ldots$$

Then the sequence $(X_i,\ i \geqslant 1)$ is a sequence of iid rvs under the distribution X and a good approximation of I is given by

$$\frac{1}{n}(X_1 + \ldots + X_n).$$

This quantity is called the empirical mean of the sample.

We remark that this method is easy to program. It is also notable that it does not depend on the regularity of f, which can be simply measurable.

Introduction - Monte Carlo method

More generally, we often want to evaluate an integral in $\mathbb{R}^d$ of the form

$$I = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

where $f(x)$ is positive and sums to one (i.e. $\int f(x)dx = 1$). Then I can be written in the form $\mathbb{E}(g(X))$ with X a rv valued in $\mathbb{R}^d$ having probability density function f with respect to the Lebesgue measure.

We can therefore approximate I by

$$\widehat{I}_n := \frac{1}{n}(g(X_1) + \ldots + g(X_n)),$$

if $(X_i,\ i \geqslant 1)$ is sampled from the distribution $f(x)dx$.

Introduction - Monte Carlo method

One can easily check that

Proposition

$\widehat{I}_n$ is an unbiased estimator of I (which means that $\mathbb{E}(\widehat{I}_n) = I$).

Probabilistic questions

$\rightarrow$ How and when does this method converge ?

$\rightarrow$ What can we say about the precision of the approximation i.e. what is the rate of convergence ?

Introduction - Limits and convergence

The answers of the previous questions are given by two of the most important probabilistic theorems.

Theorem (Strong Law of Large Numbers)

Let $(X_i,\ i \geqslant 1)$ be a sequence of iid rvs distributed as a rv X. We assume that $\mathbb{E}(|X|) < +\infty$. Then for almost every ω

$$\mathbb{E}(X) = \lim_{n \to +\infty} \frac{1}{n}(X_1 + \ldots + X_n).$$

This theorem then states that the empirical mean is a "good" approximation of I in the case where the function g is integrable :

$$\widehat{I}_n = \frac{1}{n}(g(X_1) + \ldots + g(X_n)) \underset{n \to \infty}{\longrightarrow} \mathbb{E}(g(X)) = \int_{\mathbb{R}^d} f(x_1, \ldots, x_d) g(x_1, \ldots, x_d) dx,$$

where $x = (x_1 \ldots x_d)$.

Introduction - Limits and convergence

The (random) error committed is given by

$$\epsilon_n = \mathbb{E}(g(X)) - \frac{1}{n}\left[g(X_1) + \ldots + g(X_n)\right] = I - \widehat{I}_n.$$

We want to evaluate this error.

The Central Limit Theorem gives a quantity that is asymptotically equal (in distribution) to the random error ϵ_n but which is also random (standard Gaussian distributed).

Introduction - Limits and convergence

Theorem (Central Limit Theorem)

Let $(X_i,\ i \geqslant 1)$ be a sequence of iid rvs distributed as a rv X. We assume that $\mathbb{E}(X^2) < +\infty$ and denote by σ^2 the variance of X. Then

$$\frac{\sqrt{n}}{\sigma}\epsilon_n \xrightarrow{(d)} G,$$

while n goes to infinity and G being a rv with a standard Gaussian distribution.

This means that if h is a bounded Borel function $\mathbb{E}\left(h\left(\frac{\sqrt{n}}{\sigma}\epsilon_n\right)\right)$ converges to $\mathbb{E}(h(G))$.

Introduction - Confidence intervals

The Central Limit Theorem never allows us to bound the random error ϵ_n since the support of a Gaussian is equal to the whole of $\mathbb{R}$.

Nevertheless it leads to a description of the error of the Monte Carlo method by giving the standard deviation $\frac{\sigma}{\sqrt{n}}$ of ϵ_n or by giving a $(1-\alpha)$%-confidence interval (CI) for the result. That means that the result is found with $(1-\alpha)$% chance in the given interval (and with α% chance of being outside). Indeed, we can deduce from the previous theorem that for all $c_1 < c_2$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\sigma}{\sqrt{n}}c_1 \leqslant \epsilon_n \leqslant \frac{\sigma}{\sqrt{n}}c_2\right) = \int_{c_1}^{c_2} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

In practical applications, we then approximate ϵ_n by a centered Gaussian distribution with variance $\frac{\sigma^2}{n}$.

Introduction - Confidence intervals

In our context, we then derive the following asymptotic approximation for $\mathbb{E}(g(X))$:

$$\mathbb{P}\Big(\frac{\sqrt{n}}{\sigma}|\epsilon_n| \leqslant z_\alpha\Big) = \mathbb{P}\Big(\frac{\sqrt{n}}{\sigma}|I - \widehat{I}_n| \leqslant z_\alpha\Big) \approx \mathbb{P}(|G| \leqslant z_\alpha) = 1 - \alpha,$$

where z_α is the $(1-\alpha)$-quantile of the absolute value of a standard Gaussian distribution.
In other words,

$$\mathbb{P}\Big(\mathbb{E}(g(X)) \in \Big[\widehat{I}_n - \frac{\sigma}{\sqrt{n}}z_\alpha, \widehat{I}_n + \frac{\sigma}{\sqrt{n}}z_\alpha\Big]\Big) \approx 1 - \alpha$$

or else the $(1-\alpha)$%-CI is given by

$$\Big[\mathbb{E}(g(X)) - \frac{\sigma}{\sqrt{n}}z_\alpha, \mathbb{E}(g(X)) + \frac{\sigma}{\sqrt{n}}z_\alpha\Big].$$

Introduction - Estimate of the variance

The result above shows that it is important to know the order of the size of the variance σ of the rv used in the Monte Carlo technique. It is easy to estimate this variance.

The variance σ_n^2 of the estimator $\widehat{I}_n$ based on a n-sample $(X_i, i \geqslant 1)$ is given by

$$\frac{1}{n-1}\sum_{i=1}^{n}(g(X_i) - \widehat{I}_n)^2.$$

Remark : The reason for dividing by $n-1$ instead of n is to have an unbiased estimator (which means that $\mathbb{E}(\sigma_n^2) = \sigma^2$). Although from the practical point of view it is not relevant since n will be usually large enough.

σ_n^2 is called the empirical variance of the sample.

Introduction - Estimate of the variance

We can therefore obtain a $(1-\alpha)$%-CI by replacing σ by σ_n in the CI given by the Central Limit Theorem :

$$\left[\widehat{I}_n - \frac{\sigma_n}{\sqrt{n}} z_\alpha, \widehat{I}_n + \frac{\sigma_n}{\sqrt{n}} z_\alpha\right].$$

We therefore see that with no extra calculation (just by evaluating σ_n on the sample already taken) we could give a dependable estimate of the approximation error of I by $\widehat{I}_n$. It is one of the greatest advantages of the Monte Carlo method to give a realistic estimate of the error at a minimum cost.

Introduction - CI for rare event probabilities

If the quantity of interest I is a rare event probability (say less than 10^{-9}), one might be cautious studying CI for ϵ_n. In that case, it is more correct to study the relative random error instead of ϵ_n itself and by the way

$$\mathbb{P}\Big(\frac{\sqrt{n}}{\sigma}\frac{|I-\widehat{I}_n|}{I}\leqslant z_\alpha\Big) \text{ instead of } \mathbb{P}\Big(\frac{\sqrt{n}}{\sigma}|I-\widehat{I}_n|\leqslant z_\alpha\Big).$$

We are then led to a $(1-\alpha)$%-CI of the kind

$$\left[\mathbb{E}(g(X))-\frac{\sigma}{\sqrt{n}}z_\alpha I, \mathbb{E}(g(X))+\frac{\sigma}{\sqrt{n}}z_\alpha I\right],$$

which means that

$$\mathbb{P}\left(\mathbb{E}(g(X))\in\left[\widehat{I}_n-\frac{\sigma}{\sqrt{n}}z_\alpha I, \widehat{I}_n+\frac{\sigma}{\sqrt{n}}z_\alpha I\right]\right)\approx 1-\alpha.$$

An example in telecommunications

In telecommunication, the loss probability of a packet of information is less than 10^{-9}. In other words, one must simulate a billion of information packets by loss packet. To achieve a good approximation one needs the simulation of at least 100 billions of packets that would take hundred of days.

1. Determine the size n needed to achieve a fixed RE in function of z_α.
2. Application : $I = 10^{-9}$, $RE = 10\%$ and $\alpha = 5\%$.

Second example

Assume that we want to calculate $E := \mathbb{E}(e^{\beta G})$ where G is a standard Gaussian rv.

1. Prove that clearly $E = e^{\beta^2/2}$.
2. To apply a Monte Carlo method, we consider the rv $X = e^{\beta G}$. Compute its variance.
3. Determine the size n needed to achieve a fixed RE.
4. Application : $t = 10\%$, $\beta = 5$ and $\alpha = 5\%$.

Second example

The following tabular contains the results of a simulation based on 10^5 trials in the case $\beta = 5$:

	exact value	:	268337
n=100000	estimated 95% CI	:	[-467647,2176181]
	estimated value	:	854267

This approximation is really far to be precise ! But importantly the calculated CI contains the exact value.

Second example

The following tabular contains the results of a simulation based on 10^5 trials in the case $\beta = 5$:

	exact value	:	268337
n=100000	estimated 95% CI	:	[-467647,2176181]
	estimated value	:	854267

This approximation is really far to be precise ! But importantly the calculated CI contains the exact value.

This is the reassuring aspect of the Monte Carlo method : the approximation may be mediocre but we are well aware of it. This example shows the limit of the Monte Carlo method when the variance of the rv used is large.

IMPORTANCE SAMPLING

Importance Sampling (IS)

Assume that we want to calculate $I = \mathbb{E}(g(X))$ where the distribution of X is given by $f(x)dx$. The quantity that we want to estimate is then

$$I = \mathbb{E}(g(X)) = \int g(x)f(x)dx.$$

In that view, we introduce a new function $\widetilde{f}$ such that $\widetilde{f} > 0$ and $\int \widetilde{f}(x)dx = 1$. Obviously, the quantity to estimate can be written as

$$\mathbb{E}(g(X)) = \int \frac{g(x)f(x)}{\widetilde{f}(x)}\widetilde{f}(x)dx = \mathbb{E}\left[\frac{g(Y)f(Y)}{\widetilde{f}(Y)}\right],$$

if Y follows the distribution $\widetilde{f}(x)dx$.

Importance Sampling (IS)

This means that we therefore have another method of estimating $\mathbb{E}(g(X))$ based on a n-sample $Y_1, Y_2, \ldots, Y_n$ distributed as Y, the approximation being

$$\widehat{I}_n := \frac{1}{n} \sum_{i=1}^{n} \frac{g(Y_i) f(Y_i)}{\widetilde{f}(Y_i)}.$$

This procedure will be efficient if the rv Z defined by $Z = \frac{g(Y) f(Y)}{\widetilde{f}(Y)}$ has a smaller variance than that of $g(X)$. Easily we have

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{g(Y) f(Y)}{\widetilde{f}(Y)} \right) = \int \frac{g(x)^2 f(x)^2}{\widetilde{f}(x)} dx - \mathbb{E}(g(X))^2.$$

The quantity $L(x) := \frac{f(x)}{\widetilde{f}(x)}$ is called the likelihood ratio.

Importance Sampling (IS) - Optimal change

Note that if $g > 0$, the function

$$\widetilde{f}(x) = \frac{g(x)f(x)}{\mathbb{E}(g(X))}$$

cancels the variance ! This means that there is an optimal change of measure leading to a zero-variance estimator. The simulation becomes a kind of "pseudo-simulation" leading to the exact value in only one sample (unbiased estimator with variance equal to zero).

Unfortunately, this result is not tractable since this optimal function $\widetilde{f}$ depends on $\mathbb{E}(g(X))$ the quantity to evaluate !

Importance Sampling (IS) - Optimal change

Nevertheless, this observation leads to two remarks :

1. there is an optimal change of measure which suggests that there are other good and even very good changes of measures
2. it allows us to justify the following heuristic : in practice, we choose $\widetilde{f}$ as close as possible as $|gf|$ then we proceed to a normalization to recover a probability density function.

Importance Sampling (IS)

Remark : In order to avoid the calculation of the normalizing constant, we use the following estimate

$$\widetilde{I}_n = \frac{\sum_{i=1}^n g(Y_i) f(Y_i)/\widetilde{f}(Y_i)}{\sum_{i=1}^n f(Y_i)/\widetilde{f}(Y_i)} = \sum_{i=1}^n g(Y_i)\omega_i$$

where $Y_1, \ldots, Y_n$ are iid rvs with common distribution $\widetilde{f}(x)dx$ and the importance weights $\omega_1, \ldots, \omega_n$ are given by

$$\omega_i = \frac{f(Y_i)/\widetilde{f}(Y_i)}{\sum_{i=1}^n f(Y_i)/\widetilde{f}(Y_i)}.$$

For a fixed n, $\widetilde{I}_n$ is biased but it is asymptotically unbiased. $\widetilde{I}_n$ is nothing more than the function $g(x)$ integrated with respect to the empirical measure

$$\sum_{i=1}^n \omega_i \delta_{Y_i}(dy)$$

where δ_a is the Dirac measure at a.

IS - Rare event framework

In the rare event setting, the quantity of interest writes

$$\Gamma = \mathbb{P}(X \in A) = \mathbb{E}(\mathbb{1}_A(X)) = \int_{-\infty}^{+\infty} \mathbb{1}_A(x) f(x) dx$$

and can be estimated trough the Monte Carlo method using a n-sample $X_1, \ldots, X_n$ iid with common density $f(x)dx$. It yields the following unbiased estimator

$$\widehat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i).$$

As showed previously, the necessary sampling size to achieve a RE less than t with probability $1 - \alpha$ should be

$$n = \left(\frac{z_\alpha \sigma_n}{t\Gamma} \right)^2,$$

i.e. proportional to the inverse of the square root of the rare event probability Γ.

IS - Rare event framework

Applying IS methodology, we rewrite Γ in the form

$$\Gamma = \int_{-\infty}^{+\infty} \mathbb{1}_A(x) \frac{f(x)}{\widetilde{f}(x)} \widetilde{f}(x) dx = \mathbb{E}(\mathbb{1}_A(Y) L(Y))$$

Now with a n-sample $Y_1, \dots Y_n$ distributed following $\widetilde{f}(x)dx$, an unbiased estimate of Γ is given by

$$\widehat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(Y_i) L(Y_i).$$

To find a good change of measure, one needs to have a good knowledge of the system under study. Moreover, it is possible that IS does not lead to an improvement and even with a bad change of measure, the result can be worse !

OTHER METHODS TO REDUCE THE VARIANCE

Other methods - Control variables

The principle is the same as the one of IS : we want to evaluate $\mathbb{E}(g(X))$ that we write in the form

$$\mathbb{E}(g(X)) = \mathbb{E}(g(X) - h(X)) + \mathbb{E}(g(X))$$

where $\mathbb{E}(g(X))$ can be calculated explicitly and $\mathrm{Var}(g(X) - h(X))$ is much more smaller than $\mathrm{Var}(g(X))$.

We then use a Monte Carlo method to evaluate $\mathbb{E}(g(X) - h(X))$ and a direct evaluation for $\mathbb{E}(h(X))$.

Other methods - Antithetic variables

If X is uniformly distributed over $[0,1]$ and since $x \mapsto 1-x$ leaves the measure dx invariant, we also have

$$I = \int_0^1 g(x)dx = \mathbb{E}(g(X)) = \frac{1}{2}\int_0^1 (g(x)+g(1-x))dx$$

We can therefore apply the Monte Carlo technique to estimate I by

$$\widehat{I}_{2n} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(g(X_i)+g(1-X_i)) = \frac{1}{2n}\sum_{i=1}^{n} g(X_i)+g(1-X_i),$$

where $(X_i,\ i \geqslant 1)$ iid rvs uniformly distributed over $[0,1]$.

If the function g is continuous and monotone, the quality of the approximation is improved with respect to a direct Monte Carlo method based on $2n$ realizations of the rv X.

Other methods - Method of stratification

Assume that we want to calculate

$$I = \mathbb{E}(g(X)) = \int_{\mathbb{R}^d} g(x) f(x) dx$$

where X is a rv valued in $\mathbb{R}^d$ following the distribution $f(x)dx$.

We take a partition $(D_i, 1 \leqslant i \leqslant m)$ of $\mathbb{R}^d$ and we decompose the integral in the following way

$$I = \sum_{i=1}^{m} \mathbb{E}(\mathbb{1}_{D_i}(X) g(X)) = \sum_{i=1}^{m} \mathbb{E}(g(X)|X \in D_i) \mathbb{P}(X \in D_i)$$

When we know the numbers $p_i := \mathbb{P}(X \in D_i)$, we can use a Monte Carlo method to estimate the integrals $I_i := \mathbb{E}(g(X)|X \in D_i)$ by distributing optimally the n realizations.

Other methods - Method of stratification

Assume that we approximate the integral I_i by $\widehat{I}_i$ based on n_i independent trials. The variance of the approximation error is then given by σ_i^2/n_i, if we denote $\sigma_i^2 := \mathrm{Var}\,(g(X)|X \in D_i)$. We then approximate I by

$$\widehat{I} := \sum_{i=1}^{m} p_i \widehat{I}_i.$$

Since the samples used to obtain the estimates $\widehat{I}_i$ are assumed to be independent, the variance of the estimate $\widehat{I}$ is obviously given by

$$\sum_{i=1}^{m} p_i^2 \frac{\sigma_i^2}{n_i}.$$

It is then natural to minimize this error for a fixed number of trials $n = \sum_{i=1}^{m} n_i$.

Other methods - Method of stratification

We can check that the n_i's minimizing $\mathrm{Var}\left(\widehat{I}\right)$ are given by

$$n_i = n\frac{p_i\sigma_i}{\sum_{i=1}^m p_i\sigma_i}.$$

The minimum of the variance of $\widehat{I}$ then becomes

$$\frac{1}{n}\left(\sum_{i=1}^m p_i\sigma_i\right)^2$$

which is less than the variance obtained with n random trials by the classical Monte Carlo method. In fact, the variance becomes

$$\begin{aligned}
\mathrm{Var}(g(X)) &= \sum_{i=1}^m p_i\mathbb{E}(g(X)^2|X\in D_i) - \left(\sum_{i=1}^m p_i\mathbb{E}(g(X)|X\in D_i)\right)^2 \\
&= \sum_{i=1}^m p_i\mathrm{Var}(g(X)|X\in D_i) + \sum_{i=1}^m p_i\mathbb{E}(g(X)|X\in D_i)^2 \\
&\quad - \left(\sum_{i=1}^m p_i\mathbb{E}(g(X)|X\in D_i)\right)^2
\end{aligned}$$

by using the definition of the conditional variance.

Other methods - Method of stratification

We then use twice the convexity inequality for x^2

$$\left(\sum_{i=1}^{m} p_i a_i\right)^2 \leqslant \sum_{i=1}^{m} p_i a_i^2$$

if $\sum_{i=1}^{m} p_i = 1$ to show that

$$\mathrm{Var}(g(X)) \geqslant \sum_{i=1}^{m} p_i \mathrm{Var}(g(X)|X \in D_i) \geqslant \left(\sum_{i=1}^{m} p_i \sigma_i\right)^2,$$

which proves that provided we have an optimal strategy of trials, we can obtain by stratification, an approximation with lower variance.

Other methods - Method of stratification

Remark

Unfortunately, note that it is possible to obtain an approximation with greater variance than the initial estimate if the assignment of the points is arbitrary. Despite this, there exist other strategies to choose the points on domains that reduce the variance. For example, if we assigns a number of points proportional to the probability of the domain : $n_i = np_i$, we then obtain an approximation with variance equals to

$$\frac{1}{n}\sum_{i=1}^{m} p_i\sigma_i^2.$$

Now we see that $\sum_{i=1}^{m} p_i\sigma_i^2$ is a bound for $\mathrm{Var}(g(X))$. This allocation strategy is sometimes useful when we explicitly know the probabilities p_i.

Other methods - Average value or conditioning

Assume we want to calculate

$$\mathbb{E}(g(X,Y)) = \int g(x,y)f(x,y)dxdy$$

where $f(x,y)dxdy$ is the distribution of the pair (X,Y). Let

$$h(x) = \frac{1}{\int f(x,y)dy} \int g(x,y)f(x,y)dy$$

then $\mathbb{E}(g(X,Y)) = \mathbb{E}(h(X))$. Indeed, the distribution of X is given by $m(x)dx := \left(\int f(x,y)dy\right) dx$ and thus

$$\mathbb{E}(h(X)) = \int h(X)m(x)dx = \int dx \int g(x,y)f(x,y)dy = \mathbb{E}(g(X,Y)).$$

Other methods - Average value or conditioning

We can recover that result noting that

$$\mathbb{E}(g(X,Y)|X) = h(X).$$

This interpretation as a conditional expectation allows us to prove that the variance of $h(X)$ is lower than that of $g(X,Y)$.

If we can not calculate directly $h(x)$, we use a Monte Carlo technique for $h(X)$.

Thank you for your attention

Cám o'n ban dã quan tâm cúa ban

See you tomorrow

Ban vào ngày mai