Monte Carlo integration and variance reduction

Marcin Chrząszcz
mchrzasz@cern.ch

University of Zurich[UZH]

Monte Carlo methods,
3 March, 2016

Monte Carlo and integration

$\hookrightarrow$ **All MC calculations are equivalent to preforming an integration.**

$\rightrightarrows$ Assumptions: r_i random numbers from $\mathcal{U}(0,1)$. The MC result:

$$F = F(r_1, r_2, ...r_n)$$

is unbias estimator of an integral:

$$I = \int_0^1 ... \int_0^1 F(x_1, x_2, ..., x_n) dx_1, dx_2..., dx_n$$

aka the expected value of the I integral is:

$$E(F) = I.$$

$\Rrightarrow$ This mathematical identity is the most useful property of the MC methods. It is a link between mathematical analysis and statistic world. Now we can use the best of the both world!

If we want to calculate the integral in different range then $(0,1)$ we just scale the the previous result:

$$\frac{1}{N} \sum_{i=1}^{N} f(x_i) \xrightarrow{N \to \infty} E(f) = \frac{1}{b-a} \int_a^b f(x) dx$$

Uncertainty from Monte Carlo methods

$\Rightarrow$ In practice we do not have $N \to \infty$ so we will never know the exact result of an integral :(

$\longmapsto$ Let's use the statistical world and estimate the uncertainty of an integral in this case :)

$\rightarrowtail$ A variance of a MC integral:

$$V(\hat{I}) = \frac{1}{n}\Big\{E(f^2) - E^2(f)\Big\} = \frac{1}{n}\Big\{\frac{1}{b-a}\int_a^b f^2(x)dx - I^2\Big\}$$

$\hookrightarrow$ To calculate $V(\hat{I})$ one needs to know the value of I!

$\Rightarrow$ In practice $V(\hat{I})$ is calculated via estimator:

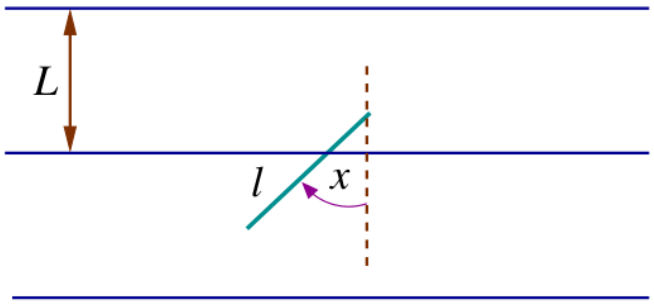
$$\hat{V}(\hat{I}) = \frac{1}{n}\hat{V}(f), \qquad \hat{V}(f) = \frac{1}{n-1}\sum_{i=1}^{n}\Big[f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\Big]^2.$$

$\Rightarrow$ MC estimator of standard deviation: $\hat{\sigma} = \sqrt{\hat{V}(\hat{I})}$

Buffon needle - π number calculus

$\Rightarrow$ Buffon needle (Buffon 1777, Laplace 1886): We are throwing a needle (of length l) on to a surface covered with parallel lines width distance L. If a thrown needle touches a line we count a hit, else miss. Knowing the number of hits and misses one can calculate the π number.

Experiment:



n - number of hits
N number of hits and misses, aka number of tries.

Theory:

$\Rightarrow$ x - angle between needle and horizontal line, $x \in \mathcal{U}(0, \pi)$. $\Rightarrow$ the probability density function (p.d.f.) for x:

$$\rho(x) = \frac{1}{\pi}$$

$\Rightarrow p(x)$ probability to hit a line for a given x value:

$$p(x) = \frac{l}{L}|\cos x|$$

$\Rightarrow$ Total hit probability:

$$P = E[p(x)] = \int_0^\pi p(x)\rho(x)dx = \frac{2l}{\pi L}$$

Now one can calculate $\hat{P}$ from MC : $\hat{P} = \frac{n}{N} \xrightarrow{N\to\infty} P = \frac{2l}{\pi L} \Rightarrow \hat{\pi} = \frac{2Nl}{nL}$

Buffon needle - Simplest Carlo method

Monte Carlo type "hit or miss"

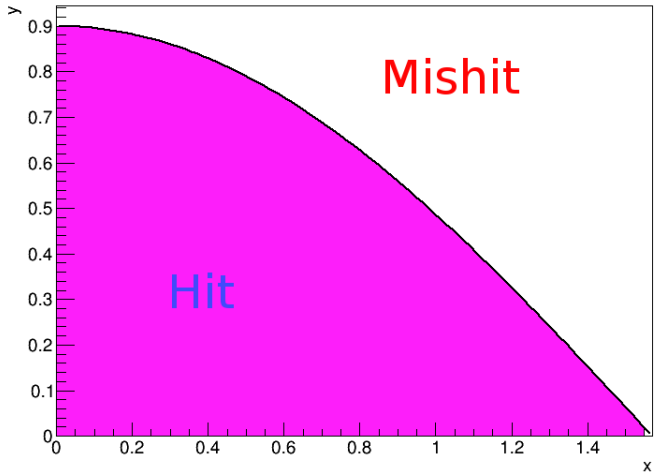
Let's use the summery of $p(x)$ function and stake $0 < x < \frac{\pi}{2}$.

$\Rightarrow$ Algorithm:

Generate 2 dim. distribution:

$$(x, y) : \mathcal{U}(0, \frac{\pi}{2}) \times \mathcal{U}(0, 1) \text{ and}$$

$$y \begin{cases} \leqslant p(x): & \text{hit,} \\ > p(x): & \text{miss.} \end{cases}$$



Let's define weight function: $w(x, y) = \Theta(p(x) - y)$,
where $\Theta(x)$ is the step function.

$\rightarrowtail$ p.d.f.: $\varrho(x, y) = \rho(x)g(y) = \frac{2}{\pi} \cdot 1$

$\Rightarrow$ Integrated probability:

$$P = E(w) = \int w(x, y)\varrho(x, y)dxdy = \frac{2l}{\pi L} \xleftarrow{N \to \infty} \hat{P} = \frac{1}{N} \sum_{i=1}^{N} w(x_i, y_i) = \frac{n}{N}$$

Standard deviation for $\hat{P}$: $\hat{\sigma} = \frac{1}{\sqrt{N-1}} \sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)}$

Heads or tails MC

⇛ MC estimator of an integral that is based on counting the numbers of positive trials compared to the failed ones is called ”hit or miss”

⇛ The probability is described by the Bernoulli distribution:

$$\mathcal{P}(n) = \binom{N}{n} P^n (1-P)^{N-n},$$

where P is the probability of success, N is the number of trials and n is the number of successes.

⇛ The following are true:

$$E(n) = NP,$$

$$V(n) = NP(1-P),$$

⇛ Translating this into probability basis:

$$E(\hat{P}) = P, \quad V(\hat{P}) = \frac{P(1-P)}{N}.$$

⇛ E2.1 prove the above.

Buffon needle

$\Rightarrow$ Lets make this toy experiment and calculate the π number.
$\hookrightarrow$ We can simulate the central position (y) of an needle between $(-L, L)$ from $\mathcal{U}(-L, L)$.

Symmetry:

Please note the symmetry of the problem, if the position of the needle would be $> L$ then we can shift the needle by any number of L's.

$\hookrightarrow$ New we simulate the angle (ϕ) with a flat distribution from $(0, \pi)$. $\hookrightarrow$ The maximum and minimum y position of the needle are:

$$y_{\max} = y + |\cos\phi| l$$
$$y_{\min} = y - |\cos\phi| l$$

$\hookrightarrow$ Now we check if the needle touches any of the lines: $y = L$, $y = 0$ or $y = -L$. If yes we count the events.

N	$\hat{\pi}$	$\hat{\pi} - \pi$	$\sigma(\hat{\pi})$
10000	3.12317	−0.01842	0.03047
100000	3.14707	0.00547	0.00979
1000000	3.13682	−0.00477	0.00307
10000000	3.14096	−0.00063	0.00097

$\Rightarrow$ E2.2 Write the program that calculates the π number using the Buffon needle.

Crude Monte Carlo method of integration

⇛ Crude Monte Carlo method of integration is based on Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \xrightarrow{N\to\infty} \frac{1}{b-a}\int_a^b f(x)dx = E(f)$$

⇛ The standard deviation can be calculated:

$$\sigma = \frac{1}{\sqrt{N}}\sqrt{\left[E(f^2) - E^2(f)\right]}$$

⇛ From LNT we have:

$$P = \int w(x)\rho(x)dx = \int_0^{\pi/2} (\frac{l}{L}\cos x)\frac{2}{\pi}dx = \frac{2l}{\pi L} \xrightarrow{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} w(x_i)$$

⇛ Important comparison between "Hit and mishit" and Crude MC methods. One can analytically calculate:

$$\hat{\sigma}^{\text{Crude}} < \hat{\sigma}^{\text{Hit and mishit}}$$

⇛ Crude MC is **always** better then "Hit and mishit" method. We will prove this on an example (can be proven analytically as well).

Crude vs "hit or miss"

⇛ The Crude MC is never worse then the "hit or miss" method.

⇛ Prove: Let's assume we calculate an integral:

$$I = \int_0^1 f(x)dx, \text{ and } 0 \leqslant f(x) \leqslant 1 \; \forall x \in (0,1)$$

⇛ The variation for the "hit-or-miss"(HM) method: $V(\hat{I}_{HM}) = \frac{1}{N}(I - I^2)$ ⇛ The variation for the crude method: $V(\hat{I}_{Crude}) = \frac{1}{N}[\int_0^1 f^2(x)dx - I^2]$ ⇛ Now the difference:

$$V(\hat{I}_{HM}) - V(\hat{I}_{Crude}) = \frac{1}{N}[I - \int_0^1 f^2(x)dx] = \frac{1}{N}\int_0^1 f(x)[1 - f(x)]dx \geqslant 0 \text{ q.e.d}$$

⇛ E2.3 Calculate the following integrals with uncertainties using "hit or miss" and crude methods:

$$\int_0^1 dx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\int_{x^2+y^2 \leqslant 1} \frac{1}{4}\sqrt{1 - (x^2 + y^2)}dxdy$$

Generalization to multi-dimension case, Crude method

⇛ Let $x = (x_1, x_2, ..., x_n)$- vector in the n-dim vector space $\mathcal{R}^n$.
$\Omega \subset \mathcal{R}^n$ - some subspace in the n-dim space.
$V \equiv (\Omega)$ - volume of the Ω subspaces.

$$I = \int_\Omega f(x)dx = V \int_\Omega f(x)dx/V = V \int_\Omega f(x)dp(x) \equiv VJ = VE(f),$$

where the MC estimator:

$$\hat{J} = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}),\ x \in \mathcal{U}(\Omega)$$

⇛ The standard deviation:

$$\hat{\sigma}(\hat{J}) \frac{1}{\sqrt{N(N-1)}} \sqrt{\sum_{i=1}^{N} f^2(x^{(i)}) - \frac{1}{N}[\sum_{i=1}^{N} f(x^{(i)})]^2}$$

⇛ In the end we get:

$$\hat{I} = V\hat{J},\ \ \hat{\sigma}(\hat{I}) = V\sigma(\hat{J})$$

Generalization to multi-dimension case, "Hit-or-miss"

$\Rightarrow$ Let $x = (x_1, x_2, ..., x_n)$- vector in the n-dim vector space $\mathcal{R}^n$.
$\Omega \subset \mathcal{R}^n$ - some subspace in the n-dim space.
$V \equiv (\Omega)$ - volume of the Ω subspaces.

$$I = \int_\Omega dx \int_0^{f_{max}} dy \Theta(f(x) - y) = V f_{max} \int_\Omega \frac{dx}{V} \int_0^{f_{max}} \frac{dy}{f_{max}} \Theta(f(x) - y)$$

where $(x, y) \in \mathcal{U}(\Omega \times [0, f_{max}]$. $\Rightarrow$ Now we define K:

$$K = \int_\Omega \frac{dx}{V} \int_0^{f_{max}} \frac{dy}{f_{max}} \Theta(f(x) - y) = E(\Theta)$$

$\Rightarrow$ We generator: $(x, y) \in \mathcal{U}(\Omega \times [0, f_{max}])$ and check:

$$y = \begin{cases} \leqslant f(x) \text{hit, weight=1} \\ > f(x) \text{hit, weight=0} \end{cases}$$

$\Rightarrow$ In the end:

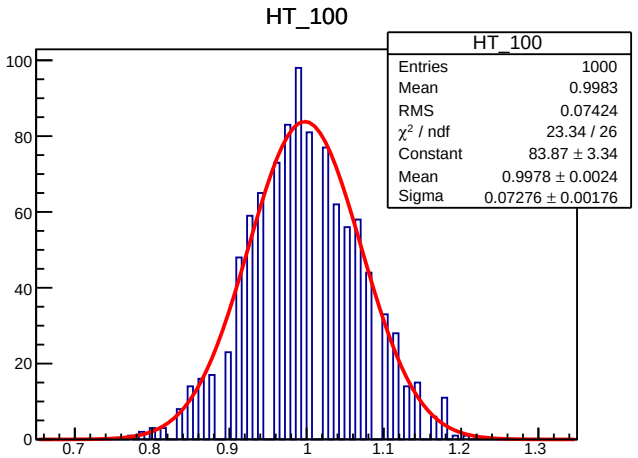
$$\hat{K} = \frac{n}{N}, \quad \hat{\sigma}(\hat{K}) = \frac{1}{\sqrt{N-1}} \sqrt{\hat{K}(1 - \hat{K})}$$

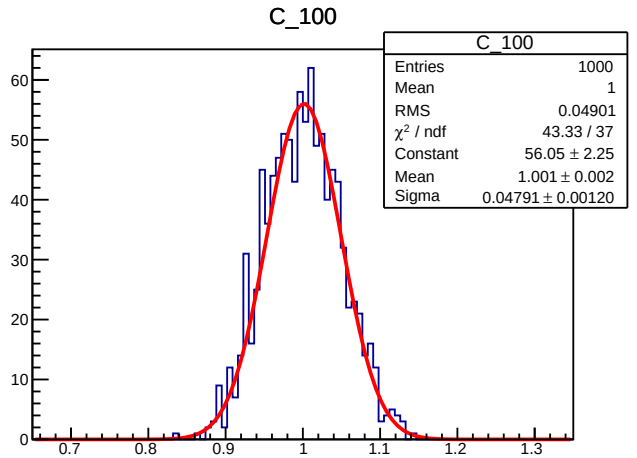
$$\hat{I} = f_{max} V \hat{K}, \quad \hat{\sigma}(\hat{I}) = f_{max} V \hat{\sigma}(\hat{K})$$

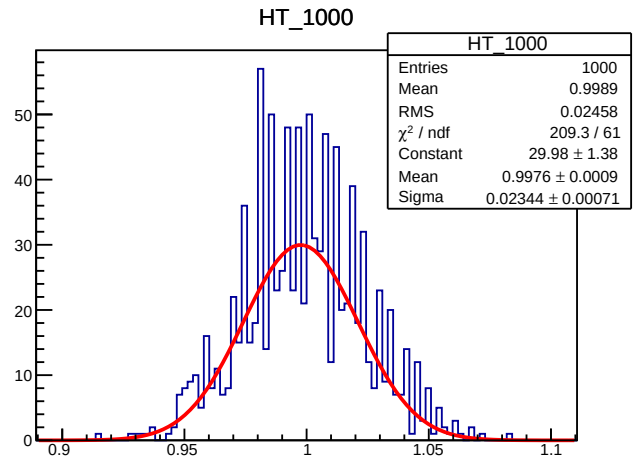
Crude MC vs "Hit and misss"

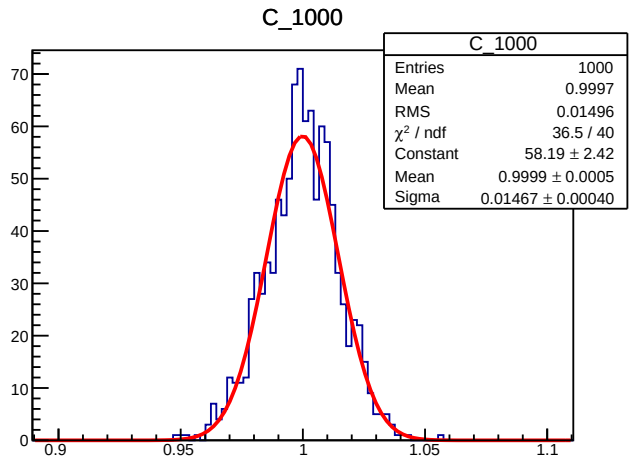
$\Rrightarrow$ We can repeat a toy MC studies as we did in the Euler needle case.

$\hookrightarrow$ In this example we want to calculate $\int_0^{\pi/2} \cos x dx$





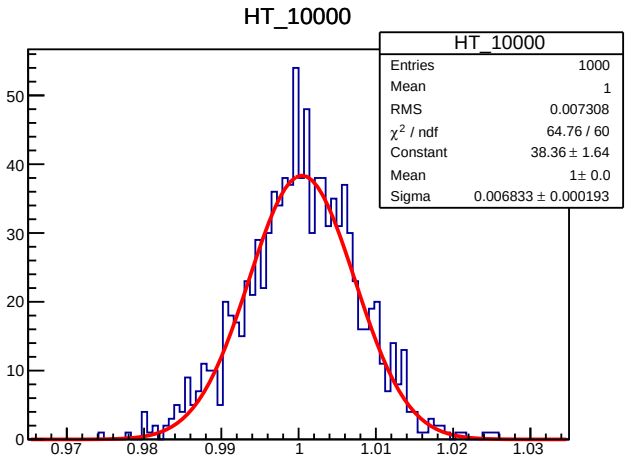


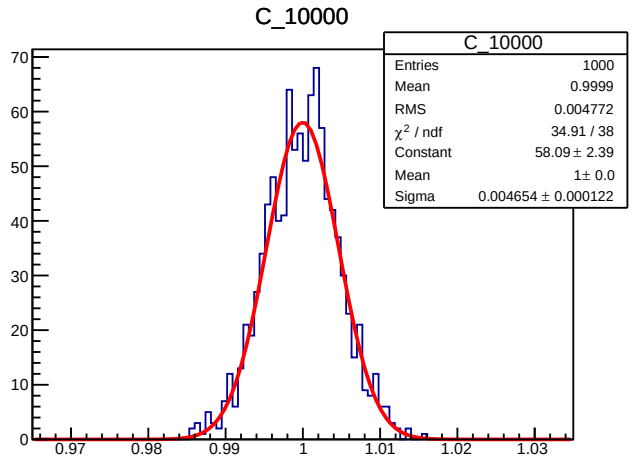


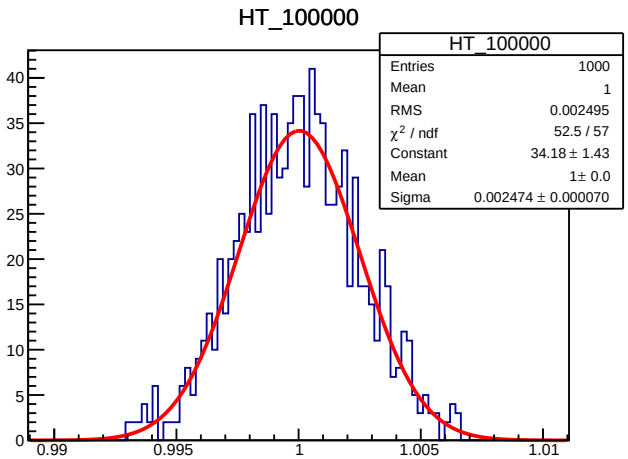
Crude MC vs "Hit and misss"

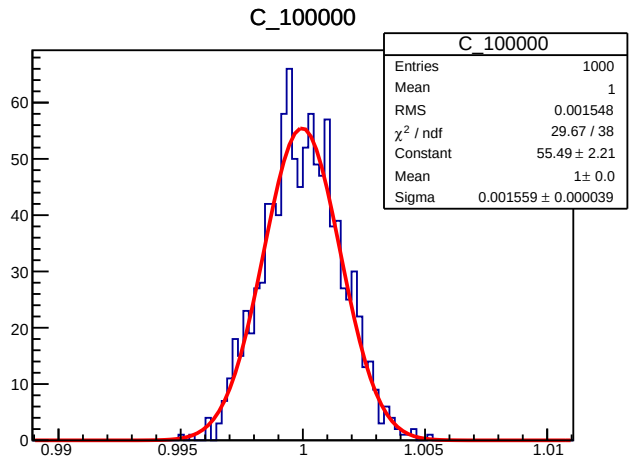
$\Rightarrow$ We can repeat a toy MC studies as we did in the Euler needle case.

$\hookrightarrow$ In this example we want to calculate $\int_0^{\pi/2} \cos x dx$





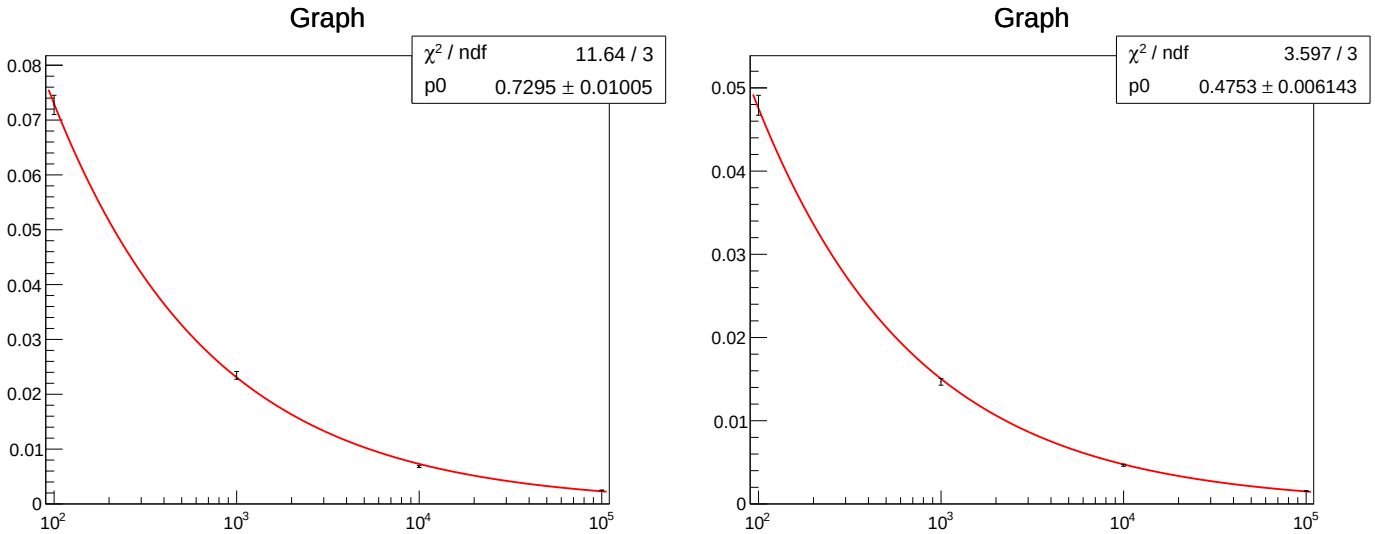




Crude MC vs "Hit and misss"

$\Rightarrow$ We can repeat a toy MC studies as we did in the Euler needle case.

$\hookrightarrow$ In this example we want to calculate $\int_0^{\pi/2} \cos x dx$



$\Rightarrow$ One clearly sees that both methods follow $1/\sqrt{N}$ dependence and that the Crude MC is always better then the "Hit and mishit".

$\Rightarrow$ Please note that for the "Hit and mishit" we are suing 2 times more random numbers than for the Crude method so in terms of timing the Crude MC is also much faster.

Classical methods of variance reduction

$\Rightarrow$ In Monte Carlo methods the statistical uncertainty is defined as:

$$\sigma = \frac{1}{\sqrt{N}} \sqrt{V(f)}$$

$\Rightarrow$ Obvious conclusion:

- To reduce the uncertainty one needs to increase N.
  $\rightrightarrows$ Slow convergence. In order to reduce the error by factor of 10 one needs to simulate factor of 100 more points!

$\Rightarrow$ How ever the other handle ($V(f)$) can be changed! $\longrightarrow$ Lot's of theoretical effort goes into reducing this factor.

$\Rightarrow$ We will discuss four classical methods of variance reduction:

1. Stratified sampling.
2. Importance sampling.
3. Control variates.
4. Antithetic variates.

Stratified sampling

$\Rrightarrow$ The most intuitive method of variance reduction. The idea behind it is to divide the function in different ranges and to use the Riemann integral property:

$$I = \int_0^1 f(u)du = \int_0^a f(u)du + \int_a^1 f(u)du,\ 0 < a < 1.$$

$\Rrightarrow$ The reason for this method is that in smaller ranges the integration function is more flat. And it's trivial to see that the more flatter you get the smaller uncertainty.
$\Rightarrow$ A constant function would have zero uncertainty!

General schematic:

Let's take our integration domain and divide it in smaller domains. In the j^{th} domain with the volume w_j we simulate n_j points from uniform distribution. We sum the function values in each of the simulated points for each of the domain. Finally we sum them with weights proportional to w_i and anti-proportional to n_i.

Stratified sampling - mathematical details

Let's define our integrals and domains:

$$I = \int_{\Omega} f(x)dx, \quad \Omega = \bigcup_{i=1}^{k} w_i$$

The integral over j^{th} domain:

$$I_j = \int_{w_j} f(x)dx, \quad \Rightarrow I = \sum_{j=1}^{k} I_i$$

$\rightrightarrows$ p_j uniform distribution in the w_j domain: $dp_j = \frac{dx}{w_j}$.

$\rightrightarrows$ The integral is calculated based on crude MC method. The estimator is equal:

$$\hat{I}_j = \frac{w_j}{n_j} \sum_{i=1}^{n_j} f(x_j^i)$$

Now the total integral is just a sum:

$$\hat{I} = \sum_{j=1}^{k} \hat{I}_j = \sum_{j=1}^{k} \frac{w_j}{n_j} \sum_{i=1}^{n_j} f(x_j^{(i)}),$$

Variance: $V(\hat{I}) = \sum_{j=1}^{k} \frac{w_j^2}{n_j} V_j(f)$, and it's estimator: $\hat{V}(\hat{I}) = \sum_{j=1}^{k} \frac{w_j^2}{n_j} \hat{V}_j(f)$

Stratified sampling in practice

One can show that splitting the integration region Ω into equal regions will not increase the variance!

$\Rightarrow$ For example in case of two sub samples:

$$V(I_{\text{crude}}) - V(I_{\text{SS}}) = \frac{1}{N}\left[\int_{\omega_1} f(x)dx - \int_{\omega_2} f(x)dx\right]^{-2} \geqslant 0$$

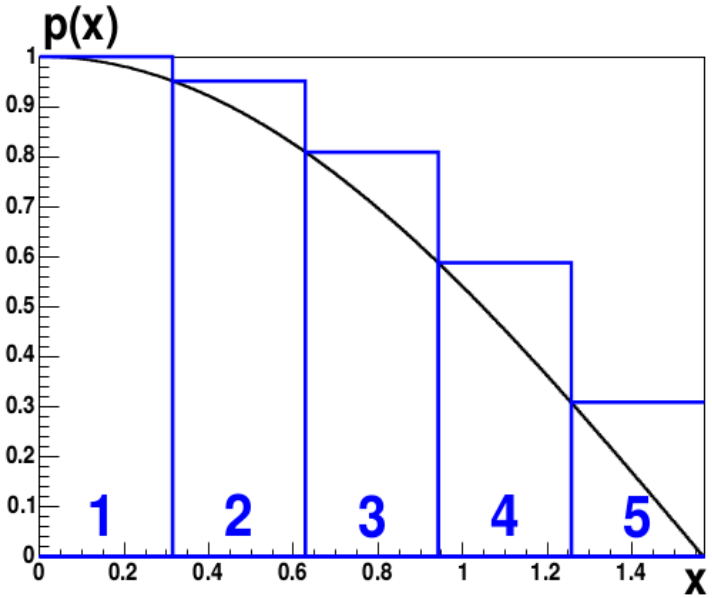
A2.1 Prove the above.

Practical advise:

If we know very little about the integrating function the equal splitting of the Ω space is the best option!

Stratified sampling for the Buffon needle

⇛ Lets apply our Stratified sampling to my favourite :) Buffon needle with 5 samples.



⇛ We have $\omega_i = \Omega/5 = \frac{\pi}{10}$ and $n_i = \frac{N}{5}$.

⇛ The integral estimator:

$$\hat{P} = \frac{1}{\Omega}\sum_{j=1}^{5}\sum_{i=1}^{N/5} p(x_i^j) = \frac{1}{N}\sum_{i=1}^{N} p(x_i)$$

⇛ The standard deviation (for $l = L$):

$$\sigma(\hat{\pi})_{SS} = \frac{0.34}{\sqrt{N}} < \sigma(\hat{\pi})_{Crude} = \frac{1.52}{\sqrt{N}}$$

⇛ In the following example we generated a constant number of events ($N/5$) for each subsample independently of their impact on the integral.

⇛ We can improve this by generating events in each of the sub sample accordingly to the area of the blue rectangle.

⇛ E2.4 Using the Stratified Sampling please calculate the integrals from E2.3 by dividing the are into 5 samples. Compute the errors and compare them to the ones obtained from the Crude method.

Importance sampling

$\Rrightarrow$ If the function is changing rapidly in its domain one needs to use a more elegant method: make the function more stable.

$\rightrightarrows$ The solution is from first course of mathematical analysis: change the integration variable :)

$$f(x)dx \longrightarrow \frac{f(x)}{g(x)}dG(x), \text{ where } g(x) = \frac{dG(x)}{dx}$$

Schematic:

- Generate the distribution from $G(x)$ instead of $\mathcal{U}$.
- For each generate point calculate the weight: $w(x) = \frac{f(x)}{g(x)}$.
- We calculate the expected value $\hat{E}(w)$ and its variance $\hat{V}_G(w)$ for the whole sample.

- If $g(x)$ is choose correctly the resulting variance can be much smaller.
- There are some mathematical requirements:
  - $g(x)$ needs to be non-negative and analytically integrable on its domain.
  - $G(x)$ invertible or there should be a direct generator of g distribution

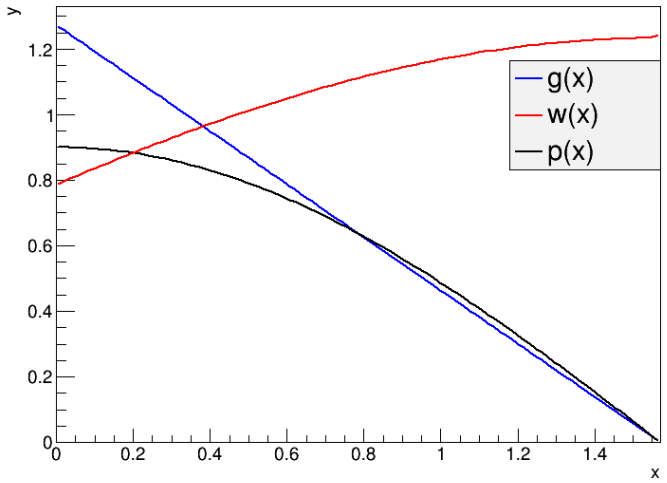
Importance sampling - Example

$\Rightarrow$ Let's take our good old π determination example.

$\Rightarrow$ Let's take here for simplicity: $L = l$.

- Let's take a trivial linear weight function: $g(x) = \frac{4}{\pi}(1 - \frac{2}{\pi}x)$
- It's invertible analytically: $G(x) = \frac{4}{\pi}x(1 - \frac{x}{\pi})$
- The weight function:

$$w(x) = \frac{p(x)}{g(x)} = \frac{\pi}{4}\frac{\cos x}{1 - 2x/\pi}$$



- Now the new standard deviation is smaller:

$$\sigma_\pi^{\mathrm{IS}} \simeq \frac{0.41}{\sqrt{N}} < \sigma_\pi \simeq \frac{1.52}{\sqrt{N}}$$

- Importance sampling has advantages:
  - Big improvements of variance reduction.
  - The only method that can cope with singularities.

$\Rightarrow$ Calculate the first function from E2.3 using the importance sampling. As a weight function $g(x)$ take a linear function.

Control variates

$\Rightarrow$ Control variates uses an other nice property of Riemann integral:

$$\int f(x)dx = \int [f(x) - g(x)]dx + \int g(x)dx$$

- $g(x)$ needs to be analytically integrable.
- The uncertainty comes only from the integral: $\int [f(x) - g(x)]dx$.
- Obviously: $V(f \to g) \xrightarrow{f \to g} 0$

$\Rightarrow$ Advantages:

- Quite stable, immune to the singularities.
- $g(x)$ doesn't need to be invertible analytically.

$\Rightarrow$ Disadvantage:

- Useful only if you know $\int g(x)dx$

Antithetic variates

$\Rrightarrow$ In MC methods usually one uses the independent random variables. The Antithetic variates method on purpose uses a set of correlated variables (negative correlation is the important property):

- Let f and $f\prime$ will be functions of x on the same domain.
- The variance: $V(f + f\prime) = V(f) + V(f\prime) + 2Cov(f, f\prime)$.
- If $Cov(f, f\prime) < 0$ then you can reduce the variance.

$\Rrightarrow$ Advantages:

- If you can pick up f and $f\prime$ so that they have negative correlation one can significantly reduce the variance!

$\Rrightarrow$ Disadvantages:

- There are no general methods to produce such a negative correlations.
- Hard to generalize this for multidimensional case.
- You can't generate events from $f(x)$ with this method.

Wrap up

$\Rightarrow$ To sum up:

- We discussed basic mathematical properties of MC methods.
- We shown that besides the stochastic nature of MC they can be used to determine totally non stochastic quantities.
- We demonstrated there is a perfect isomorphism between MC method and integration.
- We learned how co calculate integrals and estimate the uncertainties.
- Finally we discussed several classical methods of variance reduction.