Rare event simulation

Josselin Garnier

ETICS 2020

Rare event simulation: methods and simulations

- Estimation of the probability of a rare event (such as the failure of a complex system).
- Standard methods (quadrature, Monte Carlo, reliability).
- Advanced Monte Carlo methods (different variance reduction techniques: importance sampling, control variates, with adaptive versions).
- Interacting particle systems (IPS with mutation-selection-resampling, multilevel splitting).

Uncertainty propagation

- Context: numerical code (black box) or experiment

$$Y = f(\boldsymbol{X})$$

with

$Y =$ scalar output

$\boldsymbol{X} =$ random input parameters, with known distribution (with pdf $p(\boldsymbol{x})$)

$f =$ deterministic function $\mathbb{R}^d \to \mathbb{R}$ (computationally expensive).

- Goal: estimation of a quantity of the form

$$\mathbb{E}[g(Y)]$$

with an "error bar" and the minimal number of simulations.

Examples (for a real-valued output Y):

- $g(y) = y \to$ mean of Y, $\mathbb{E}[Y]$
- $g(y) = y^2 \to$ variance of Y, $\mathrm{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$
- $g(y) = \mathbf{1}_{[y_s,\infty)}(y) \to$ probability $\mathbb{P}(Y \geq y_s)$.

Analytic method

- The quantity to be estimated is a d-dimensional integral:

$$I = \mathbb{E}[g(Y)] = \mathbb{E}[h(\boldsymbol{X})] = \int_{\mathbb{R}^d} h(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{X}$ and $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$.

- In simple cases (when the pdf p and the function h have explicit expressions), one can sometimes evaluate the integral exactly (exceptional situation).

Quadrature method

- The quantity to be estimated is a d-dimensional integral:

$$I = \mathbb{E}[g(Y)] = \mathbb{E}[h(\boldsymbol{X})] = \int_{\mathbb{R}^d} h(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{X}$ and $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$.

- If $p(\boldsymbol{x}) = \prod_{i=1}^{d} p_0(x_i)$, then it is possible to apply Gaussian quadrature with a tensorized grid with n^d points:

$$\hat{I} = \sum_{j_1=1}^{n} \cdots \sum_{j_d=1}^{n} \rho_{j_1} \cdots \rho_{j_d} h(\xi_{j_1}, \ldots, \xi_{j_d})$$

with the weights $(\rho_j)_{j=1,\ldots,n}$ and the points $(\xi_j)_{j=1,\ldots,n}$ associated to the quadrature with weighting function p_0.

- There exist quadrature methods with sparse grids (cf Smolyak).
- Quadrature methods are efficient when:

- the function $\boldsymbol{x} \to h(\boldsymbol{x})$ is smooth (and not only f),
- the dimension d is "small" (even with sparse grids).

They require many calls.

Monte Carlo method

- Principle: replace the statistical expectation

$$I = \mathbb{E}[g(Y)] = \mathbb{E}[h(\boldsymbol{X})]$$

by an empirical mean.

- There are different probabilistic representations, that give different simulation methods.

Monte Carlo method

For a given y_{s} we want to estimate

$$p_{\mathrm{s}} = \mathbb{P}(f(\boldsymbol{X}) \geq y_{\mathrm{s}})$$

The quantity of interest is an expectation:

$$p_{\mathrm{s}} = \mathbb{E}\left[\mathbf{1}_{[y_{\mathrm{s}},\infty)}(f(\boldsymbol{X}))\right]$$

- Monte Carlo method:

1) Let $(\boldsymbol{X}^{(k)})_{k=1}^{n}$ be a n-sample of $\boldsymbol{X}$.
2) Compute

$$Z^{(k)} = \mathbf{1}_{[y_{\mathrm{s}},\infty)}(f(\boldsymbol{X}^{(k)}))$$

3) Define the empirical estimator of p_{s}:

$$\hat{P}_n := \frac{1}{n}\sum_{k=1}^{n} Z^{(k)}$$

- Empirical estimator of $p_{\rm s}$:

$$\hat{P}_n := \frac{1}{n}\sum_{k=1}^{n} Z^{(k)}$$

- The estimator is unbiased:

$$\mathbb{E}\left[\hat{P}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} Z^{(k)}\right] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Z^{(k)}] = \mathbb{E}[Z^{(1)}] = p_{\rm s}$$

- The law of large numbers shows that the estimator is convergent:

$$\hat{P}_n = \frac{1}{n}\sum_{k=1}^{n} Z^{(k)} \overset{n\to\infty}{\longrightarrow} \mathbb{E}[Z^{(1)}] = p_{\rm s}$$

because the $Z^{(k)}$'s are independent and identically distributed (i.i.d.).

- Empirical estimator of $p_{\rm s}$:

$$\hat{P}_n := \frac{1}{n}\sum_{k=1}^{n} Z^{(k)}$$

- Mean square error:

$$\begin{aligned}\mathbb{E}\left[(\hat{P}_n - p_{\rm s})^2\right] &= \mathrm{Var}(\hat{P}_n) = \frac{1}{n}\mathrm{Var}(Z^{(1)}) \\ &= \frac{1}{n}(p_{\rm s} - p_{\rm s}^2)\end{aligned}$$

- The relative error is therefore:

$$\mathrm{Error} = \frac{\sqrt{\mathrm{Var}(\hat{P}_n)}}{\mathbb{E}[\hat{P}_n]} = \frac{\sqrt{\mathrm{Var}(\hat{P}_n)}}{p_{\rm s}} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{p_{\rm s}} - 1} \overset{p_{\rm s}\ll 1}{\simeq} \frac{1}{\sqrt{np_{\rm s}}}$$

$\hookrightarrow$ If $p_{\rm s} \ll 1$, then we need $np_{\rm s} > 1$ so that the relative error is smaller than 1 (not surprising) !

• *Question*: The estimator $\hat{P}_n$ gives an approximate value of $p_{\rm s}$, all the better as n is larger. How to quantify the error ?

• *Answer*: We build a confidence interval at the level 0.95, i.e. an empirical interval $[\hat{a}_n, \hat{b}_n]$ such that

$$\mathbb{P}\Big(p_{\rm s} \in [\hat{a}_n, \hat{b}_n]\Big) \geq 0.95$$

Construction based on Central Limit Theorem:

$$\sqrt{n}\big(\hat{P}_n - p_{\rm s}\big) = \sqrt{n}\Big(\frac{1}{n}\sum_{k=1}^{n} Z^{(k)} - p_{\rm s}\Big) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, p_{\rm s} - p_{\rm s}^2) \text{ in distribution}$$

Therefore

$$\mathbb{P}\left(\Big|\hat{P}_n - p_{\rm s}\Big| < c\frac{\sqrt{p_{\rm s} - p_{\rm s}^2}}{\sqrt{n}}\right) \overset{n\to\infty}{\longrightarrow} \frac{2}{\sqrt{2\pi}}\int_0^c e^{-x^2/2}dx$$

$$\mathbb{P}\left(p_{\rm s} \in \left[\hat{P}_n - 1.96 \frac{\sqrt{p_{\rm s} - p_{\rm s}^2}}{\sqrt{n}}, \hat{P}_n + 1.96 \frac{\sqrt{p_{\rm s} - p_{\rm s}^2}}{\sqrt{n}} \right] \right) \simeq 0.95$$

The unknown parameter $p_{\rm s}$ is still in the bounds of the interval !
Two solutions:
- $p_{\rm s} \in [0,1]$, therefore $\sqrt{p_{\rm s} - p_{\rm s}^2} < 1/2$ and

$$\mathbb{P}\left(p_{\rm s} \in \left[\hat{P}_n - 0.98 \frac{1}{\sqrt{n}}, \hat{P}_n + 0.98 \frac{1}{\sqrt{n}} \right] \right) \geq 0.95$$

- asymptotically, we can replace $p_{\rm s}$ in the bounds by $\hat{P}_n$ (OK if $np_{\rm s} > 10$ and $n(1-p_{\rm s}) > 10$):

$$\mathbb{P}\left(p_{\rm s} \in \left[\hat{P}_n - 1.96 \frac{\sqrt{\hat{P}_n - \hat{P}_n^2}}{\sqrt{n}}, \hat{P}_n + 1.96 \frac{\sqrt{\hat{P}_n - \hat{P}_n^2}}{\sqrt{n}} \right] \right) \simeq 0.95$$

[Proof: consistency of $\hat{P}_n$ and Slutsky's theorem].

Monte Carlo estimation: general model

- Black box model (numerical code)

$$Y = f(\boldsymbol{X})$$

We want to estimate

$$I = \mathbb{E}[g(Y)]$$

for some function $g : \mathbb{R} \to \mathbb{R}$.
For instance $g(y) = \mathbf{1}_{[y_s,\infty)}(y)$.

- Empirical estimator:

$$\widehat{I}_n = \frac{1}{n}\sum_{k=1}^{n} g(f(\boldsymbol{X}^{(k)}))$$

where $(\boldsymbol{X}^{(k)})_{k=1}^n$ is a n-sample of $\boldsymbol{X}$.
This is the empirical mean of a sequence of i.i.d. random variables.

- The estimator $\widehat{I}_n$ is unbiased: $\mathbb{E}[\widehat{I}_n] = I$.
- The law of large numbers gives the convergence of the estimator:

$$\widehat{I}_n \overset{n\to\infty}{\longrightarrow} I \quad \text{with probability 1}$$

- Error:

$$\mathrm{Var}(\widehat{I}_n) = \frac{1}{n}\mathrm{Var}(g(Y))$$

Proof: the variance of a sum of i.i.d. random variables is the sum of the variances.

- Asymptotic confidence interval:

$$\mathbb{P}\left(I \in \left[\widehat{I}_n - 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}, \widehat{I}_n + 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right]\right) \simeq 0.95$$

where

$$\hat{\sigma}_n = \left(\frac{1}{n}\sum_{k=1}^{n} g(f(\boldsymbol{X}^{(k)}))^2 - \widehat{I}_n^2\right)^{1/2}$$

- Advantages of the MC method:

1) no regularity condition for f, g (condition: $\mathbb{E}[g(f(\boldsymbol{X}))^2] < \infty$).
2) convergence rate $1/\sqrt{n}$ in any dimension.
3) can be applied for any quantity that can be expressed as an expectation.
4) embarrassingly parallel.

- One needs to simulate many samples of $\boldsymbol{X}$ and to call many times f.

Uncertainty propagation by metamodels

The complex code/experiment f is replaced by a metamodel (reduced model) f_{r} and one of the previous techniques is applied with f_{r} (analytic, quadrature, Monte Carlo).

$\rightarrow$ It is possible to call many times the metamodel.

$\rightarrow$ The choice of the metamodel is critical.

$\rightarrow$ The error control is not simple.

Taylor expansions

• The output $Y = f(\boldsymbol{X})$ is approximated by a Taylor series expansion $Y_{\mathrm{r}} = f_{\mathrm{r}}(\boldsymbol{X})$.

• Example (sandwich method):

- We want to estimate $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ for $Y = f(\boldsymbol{X})$ given $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ and $\mathbf{C} = \mathrm{Cov}(\boldsymbol{X})$.

- We approximate $Y = f(\boldsymbol{X})$ by $Y_{\mathrm{r}} = f_{\mathrm{r}}(\boldsymbol{X}) = f(\boldsymbol{\mu}) + \nabla f(\boldsymbol{\mu}) \cdot (\boldsymbol{X} - \boldsymbol{\mu})$.

We find:

$$\mathbb{E}[Y] \simeq \mathbb{E}[Y_{\mathrm{r}}] = f(\boldsymbol{\mu}), \qquad \mathrm{Var}(Y) \simeq \mathrm{Var}(Y_{\mathrm{r}}) = \nabla f(\boldsymbol{\mu})^T \mathbf{C} \nabla f(\boldsymbol{\mu})$$

We just need to compute $f(\boldsymbol{\mu})$ and $\nabla f(\boldsymbol{\mu})$ (evaluation of the gradient by finite differences, about $d + 1$ calls to f, or adjoint method).

• Rapid, analytic, allows to evaluate approximately central trends of the output (mean, variance).

• Suitable for small variations of the input parameters and a smooth model (that can be linearized).

Reliability methods

- We consider the output $Y = f(\boldsymbol{X})$, with $\boldsymbol{X} = (X_i)_{i=1}^d$ a random vector with known pdf p.
- We want to evaluate

$$p_{\mathrm{s}} = \mathbb{P}(Y \geq y_{\mathrm{s}}) = \mathbb{P}(\boldsymbol{X} \in F) = \int_{\mathbb{R}^d} \mathbf{1}_F(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \int_F p(\boldsymbol{x}) d\boldsymbol{x},$$

where

$$F = \{\boldsymbol{x} \in \mathbb{R}^d,\, f(\boldsymbol{x}) \geq y_{\mathrm{s}}\}$$

is called the failure domain.

- FORM-SORM method (First or Second-Order Reliability Method):

Let us assume that the X_i's are *i.i.d. with distribution* $\mathcal{N}(0,1)$ (see the slides on *isoprobabilist transform*):

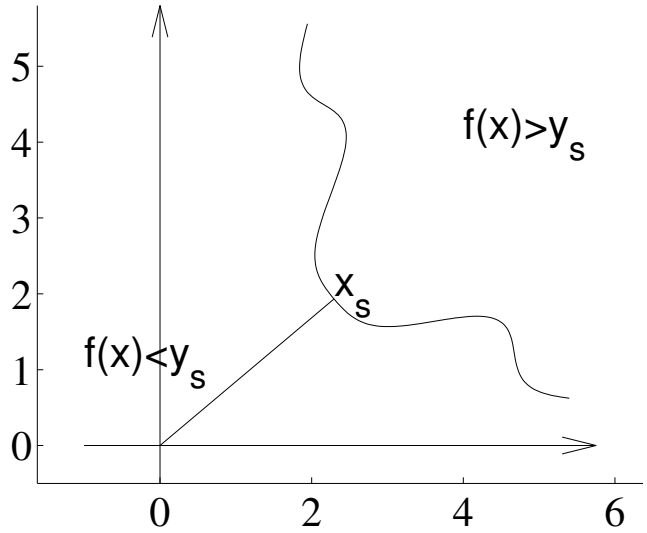
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{\|\boldsymbol{x}\|^2}{2}\Big)$$

[Cf. O. Ditlevsen et H. O. Madsen, Structural Reliability Methods, Wiley, 1996.]

$$p_{\mathrm{s}} = \int_F p(\boldsymbol{x}) d\boldsymbol{x}, \qquad F = \{\boldsymbol{x} \in \mathbb{R}^d,\, f(\boldsymbol{x}) \geq y_{\mathrm{s}}\}$$

• we find by constrained optimization the design point $\boldsymbol{x}_s$, i.e. the point with the smallest norm s.t. $f(\boldsymbol{x}_s) = y_{\mathrm{s}}$ (assuming it exists and is unique).

$$\boldsymbol{x}_s = \underset{\boldsymbol{x} \in F}{\operatorname{argmin}} \ \|\boldsymbol{x}\|^2$$



• the failure domain F is approximated by a half-space with smooth and simple boundary $\hat{F}$ going through $\boldsymbol{x}_s$, which makes it possible to compute

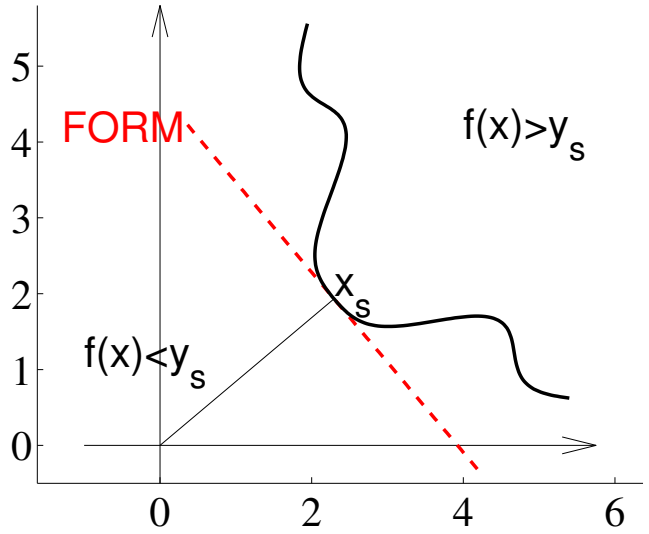
$$\hat{p}_s = \int_{\hat{F}} p(\boldsymbol{x}) d\boldsymbol{x}$$

$$\hat{p}_s = \int_{\hat{F}} p(\boldsymbol{x}) d\boldsymbol{x}$$

• The half-space $\hat{F}$ is determined by a hyperplane by the FORM method; it goes through the point $\boldsymbol{x}_s$ and is orthogonal to the vector $\boldsymbol{x}_s$. We get:

$$\hat{p}_s = \Phi(-\|\boldsymbol{x}_s\|),$$

where Φ is the cdf of the distribution $\mathcal{N}(0,1)$.



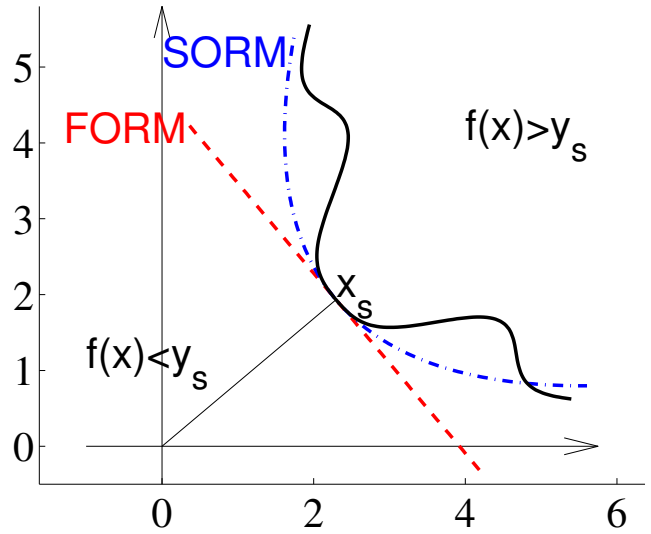
Proof: Introduce an orthonormal basis of $\mathbb{R}^d$ whose first vector is $\boldsymbol{x}_s/\|\boldsymbol{x}_s\|$. Carry out the change of variable $\boldsymbol{x}' = \mathbf{O}\boldsymbol{x}$, where $\mathbf{O}$ is the orthogonal matrix of basis change. □

$$\hat{p}_s = \int_{\hat{F}} p(\boldsymbol{x}) d\boldsymbol{x}$$

• The half-space $\hat{F}$ is determined by a quadratic surface by the SORM method. We get (Breitung formula):

$$\hat{p}_s \simeq \Phi(-\|\boldsymbol{x}_s\|) \prod_{i=1}^{d-1} \frac{1}{\sqrt{1+\|\boldsymbol{x}_s\|\kappa_i}} \Big(1 + \underset{\|\boldsymbol{x}_s\|\to\infty}{o}(1)\Big),$$

where the κ_i are the curvatures of the failure surface at $\boldsymbol{x}_s$ (computed from the gradient and Hessian of f at $\boldsymbol{x}_s$).



Isoprobabilist transform

- Let the distribution of $\boldsymbol{X} = (X_i)_{i=1}^d$ be given.

How do we transform the problem into the "standard" form where X_i are i.i.d. with distribution $\mathcal{N}(0,1)$?

- Assume that $(X_i)_{i=1}^d$ are *independent* with continuous cdf $(F_i)_{i=1}^d$.
- Let

$$\phi(\boldsymbol{x}) = \big(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d))\big),$$

where Φ is the cdf of the distribution $\mathcal{N}(0,1)$.

- The vector

$$\tilde{\boldsymbol{X}} = \phi(\boldsymbol{X})$$

has independent coordinates which satisfy

$$\mathbb{P}(\tilde{X}_i \leq x) = \mathbb{P}\big(\Phi^{-1}(F_i(X_i)) \leq x\big) = \mathbb{P}\big(F_i(X_i) \leq \Phi(x)\big) = \Phi(x),$$

because $F_i(X_i)$ has distribution $\mathcal{U}(0,1)$.

$\rightarrow$ The coordinates of $\tilde{\boldsymbol{X}}$ are i.i.d. with distribution $\mathcal{N}(0,1)$.

$\rightarrow$ $p = \mathbb{P}(f(\boldsymbol{X}) \geq y_{\mathrm{s}}) = \mathbb{P}(\tilde{f}(\tilde{\boldsymbol{X}}) \geq y_{\mathrm{s}})$ with $\tilde{f} = f \circ \phi^{-1}$.

• The Rosenblatt transform: If $\boldsymbol{X}$ is a random vector with arbitrary pdf, then there exists a function $\phi : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\tilde{\boldsymbol{X}} = \phi(\boldsymbol{X})$$

has distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

• In two steps:

$$\phi = \phi^{(2)} \circ \phi^{(1)}$$

where $\phi^{(1)}$ and $\phi^{(2)}$ are the functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ defined by:
1) $\phi_i^{(1)}(\boldsymbol{x}) = F_{i|1,\ldots,i-1}(x_i|x_1,\ldots,x_{i-1})$, $i = 1,\ldots,d$,
2) $\phi^{(2)}(\boldsymbol{z}) = (\Phi^{-1}(z_1),\ldots,\Phi^{-1}(z_d))$.
Here the function $F_{i|1,\ldots,i-1}(x_i|x_1,\ldots,x_{i-1})$ is the cdf of the variable X_i given $\{X_1 = x_1,\ldots,X_{i-1} = x_{i-1}\}$.

• The coordinates of $\phi^{(1)}(\boldsymbol{X})$ are i.i.d. with distribution $\mathcal{U}(0,1)$.
• The coordinates of $\phi^{(2)} \circ \phi^{(1)}(\boldsymbol{X})$ are i.i.d. with distribution $\mathcal{N}(0,1)$.
$\hookrightarrow$ Not easy to implement and to interpret (e.g., dependence w.r.t. order).

Proof.
Let $\boldsymbol{Z} = \phi^{(1)}(\boldsymbol{X})$. We want to show that $\boldsymbol{Z}$ has the distribution $\mathcal{U}([0,1]^d)$.
Let g be a test function ($d = 2$).

$$\mathbb{E}[g(\boldsymbol{Z})] = \mathbb{E}[g(\phi^{(1)}(\boldsymbol{X}))] = \iint_{\mathbb{R}^2} g(\phi^{(1)}(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$$

We make the change of variable $\boldsymbol{z} = \phi^{(1)}(\boldsymbol{x}) = (F_1(x_1), F_{2|1}(x_2|x_1))$, whose Jacobian is:

$$\mathbf{J} = \begin{pmatrix} \partial_{x_1} F_1(x_1) & \partial_{x_2} F_1(x_1) \\ \partial_{x_1} F_{2|1}(x_2|x_1) & \partial_{x_2} F_{2|1}(x_2|x_1) \end{pmatrix} = \begin{pmatrix} p_1(x_1) & 0 \\ * & p_{2|1}(x_2|x_1) \end{pmatrix}$$

and the determinant of the Jacobian is:

$$\mathrm{Det}(\mathbf{J}) = p_1(x_1)p_{2|1}(x_2|x_1) = p(x_1, x_2)$$

Therefore

$$\mathbb{E}[g(\boldsymbol{Z})] = \iint_{[0,1]^2} g(\boldsymbol{z})d\boldsymbol{z}$$

Variance reduction techniques

Goal: reduce the variance of the Monte Carlo estimator:

$$\mathbb{E}\left[(\hat{I}_n - I)^2\right] = \frac{1}{n}\mathrm{Var}(h(\boldsymbol{X}))$$

where $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$, $I = \mathbb{E}[h(\boldsymbol{X})]$, $\widehat{I}_n = \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)})$.

- The methods

- Importance sampling
- Control variates
- Stratification

reduce the constant without changing $1/n$, stay close to the Monte Carlo method (parallelizable).

- The methods

- Quasi-Monte Carlo

aim at changing $1/n$.

- The methods

- Interacting particle systems

are different from Monte Carlo (sequential).

Importance sampling

- The goal is to estimate $I = \mathbb{E}[h(\boldsymbol{X})]$ for $\boldsymbol{X}$ a random vector and $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ a deterministic function.
- Observation: the representation of I as an expectation is not unique:

$$I = \mathbb{E}_p[h(\boldsymbol{X})] = \int h(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int \frac{h(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}_q\left[\frac{h(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]$$

The choice of the pdf q depends on the user.

- Idea: when we know that $h(\boldsymbol{X})$ is sensitive to certain values of $\boldsymbol{X}$, instead of sampling $\boldsymbol{X}^{(k)}$ with the original pdf $p(\boldsymbol{x})$ of $\boldsymbol{X}$, a biased pdf $q(\boldsymbol{x})$ is used that makes more likely the "important" realizations.
- Using the representation

$$I = \mathbb{E}_p[h(\boldsymbol{X})] = \mathbb{E}_q\left[h(\boldsymbol{X})\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]$$

we can propose the estimator:

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)})\frac{p(\boldsymbol{X}^{(k)})}{q(\boldsymbol{X}^{(k)})}, \qquad \boldsymbol{X}^{(k)} \text{ i.i.d. with pdf } q.$$

- The estimator is unbiased (provided $\mathrm{supp}(p) \subset \mathrm{supp}(q)$):

$$\begin{aligned}\mathbb{E}_q[\hat{I}_n] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_q\left[h(\boldsymbol{X}^{(k)})\frac{p}{q}(\boldsymbol{X}^{(k)})\right] = \mathbb{E}_q\left[h(\boldsymbol{X})\frac{p}{q}(\boldsymbol{X})\right] \\ &= \int h(\boldsymbol{x})\frac{p}{q}(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = \int h(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}_p\left[h(\boldsymbol{X})\right] = I\end{aligned}$$

- The estimator is convergent:

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)})\frac{p(\boldsymbol{X}^{(k)})}{q(\boldsymbol{X}^{(k)})} \overset{n\to\infty}{\longrightarrow} \mathbb{E}_q\left[h(\boldsymbol{X})\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] = \mathbb{E}_p\left[h(\boldsymbol{X})\right] = I$$

- The variance of the estimator is:

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\mathrm{Var}_q\left(h(\boldsymbol{X})\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right) = \frac{1}{n}\left(\mathbb{E}_p\left[h(\boldsymbol{X})^2\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] - \mathbb{E}_p\left[h(\boldsymbol{X})\right]^2\right)$$

By a judicious choice of q the variance can be dramatically reduced.

- Critical points: it is necessary to know the likelihood ratio $\dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ and to know how to simulate $\boldsymbol{X}$ with the pdf q.

- The estimator is asymptotically normal

$$\sqrt{n}\big(\hat{I}_n - I\big) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}\Big(0, \mathrm{Var}_q\Big(h(\boldsymbol{X})\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\Big)\Big)$$

- It is (in theory) possible to construct asymptotic confidence intervals.

The empirical estimator of the asymptotic variance is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n} h^2(\boldsymbol{X}^{(k)})\frac{p^2(\boldsymbol{X}^{(k)})}{q^2(\boldsymbol{X}^{(k)})} - \hat{I}_n^2, \qquad \boldsymbol{X}^{(k)} \text{ i.i.d. with pdf } q.$$

The estimator $\hat{\sigma}_n^2$ is consistent (with standard moment conditions) but it may be strongly fluctuating.

$\hookrightarrow$ the construction of confidence interval is not easy.

- **Optimal importance sampling.**

The best biased distribution is the one that minimizes $\mathrm{Var}(\hat{I}_n)$.
$\hookrightarrow$ solution of the minimization problem: find the pdf $q(\boldsymbol{x})$ minimizing

$$\mathbb{E}_p\left[h(\boldsymbol{X})^2 \frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] = \int h(\boldsymbol{x})^2 \frac{p^2(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

Solution (when h is nonnegative-valued):

$$q_{\mathrm{opt}}(\boldsymbol{x}) = \frac{h(\boldsymbol{x})p(\boldsymbol{x})}{\int h(\boldsymbol{x}')p(\boldsymbol{x}')d\boldsymbol{x}'}$$

We then find

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\left(\mathbb{E}_p\left[h(\boldsymbol{X})^2 \frac{p(\boldsymbol{X})}{q_{\mathrm{opt}}(\boldsymbol{X})}\right] - \mathbb{E}_p\left[h(\boldsymbol{X})\right]^2\right) = 0 \ !$$

Pratically: the denominator of $q_{\mathrm{opt}}(\boldsymbol{x})$ is the desired quantity $\mathbb{E}[h(\boldsymbol{X})]$, which is unknown; we do not know how to sample q_{opt}.
Therefore the optimal IS method cannot be implemented.
But it is the principle for an adaptive method.

- Example: We want to estimate

$$I = \mathbb{P}(X \geq 4) = \mathbb{E}[h(X)]$$

with $X \sim \mathcal{N}(0,1)$ and $h(x) = \mathbf{1}_{[4,\infty)}(x)$.

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{1}_{[4,\infty)}(x) e^{-\frac{x^2}{2}} dx = \Phi(-4) = \frac{1}{2}\mathrm{erfc}\left(\frac{4}{\sqrt{2}}\right) \simeq 3.17\,10^{-5}$$

Monte Carlo: Let $(X^{(k)})_{k=1}^n$ be i.i.d. with the original distribution $\mathcal{N}(0,1)$.

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X^{(k)} \geq 4}$$

We have $\mathrm{Var}(\hat{I}_n) = \frac{1}{n} 3.17\,10^{-5}$.

Importance Sampling: Let $(X^{(k)})_{k=1}^n$ be i.i.d. with the distribution $\mathcal{N}(4,1)$.

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X^{(k)} \geq 4} \frac{e^{-\frac{(X^{(k)})^2}{2}}}{e^{-\frac{(X^{(k)}-4)^2}{2}}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X^{(k)} \geq 4} e^{-4X^{(k)}+8}$$

We have $\mathrm{Var}(\hat{I}_n) = \frac{1}{n} 5.53\,10^{-8}$.

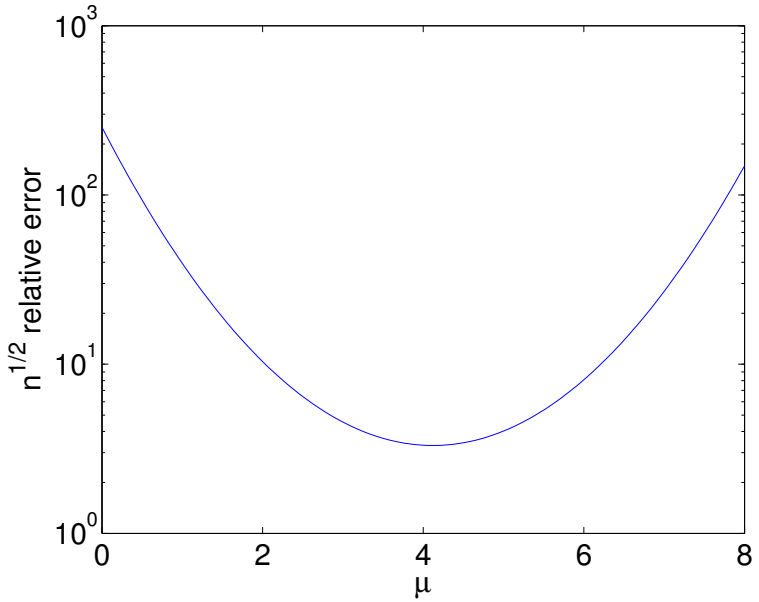
IS needs 600 times less simulations then MC to reach the same precision !

Warning: we should not bias too much.

Importance Sampling: Let $(X^{(k)})_{k=1}^n$ be i.i.d. with the distribution $\mathcal{N}(\mu, 1)$.

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X^{(k)}\geq 4} \frac{e^{-\frac{(X^{(k)})^2}{2}}}{e^{-\frac{(X^{(k)}-\mu)^2}{2}}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X^{(k)}\geq 4} e^{-\mu X^{(k)} + \frac{\mu^2}{2}}$$

$\hookrightarrow \mathrm{Var}(\hat{I}_n) = \frac{1}{n}\frac{e^{\mu^2}}{2}\mathrm{erfc}\left(\frac{4+\mu}{\sqrt{2}}\right) - \frac{1}{n}I^2$, which gives the normalized relative error $\sqrt{n}\mathbb{E}[(\hat{I}_n - I)^2]^{1/2}/I$:



If the bias is large, the fluctuations of the likelihood ratios become large.

- Example: we want to estimate

$$I = \mathbb{E}[h(X)]$$

with $X \sim \mathcal{N}(0,1)$ and $h(x) = \exp(x)$.

$$I = \frac{1}{\sqrt{2\pi}} \int e^x e^{-\frac{x^2}{2}} dx = e^{\frac{1}{2}}$$

The large values of X are important.
Importance Sampling: With a sample $(X^{(k)})_{k=1,\ldots,n}$ with the distribution $\mathcal{N}(\mu,1)$, $\mu > 0$.

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} h(X^{(k)}) \frac{e^{-\frac{[X^{(k)}]^2}{2}}}{e^{-\frac{[X^{(k)}-\mu]^2}{2}}} = \frac{1}{N} \sum_{k=1}^{n} h(X^{(k)}) e^{-\mu X^{(k)} + \frac{\mu^2}{2}}$$

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n} \left(e^{\mu^2 - 2\mu + 2} - e^1 \right)$$

Monte Carlo $\mu = 0$: $\mathrm{Var}(\hat{I}_n) = \frac{1}{n} \left(e^2 - e^1 \right)$
Optimal importance sampling $\mu = 1$: $\mathrm{Var}(\hat{I}_n) = 0$.

• Gaussian random walk $X_p = X_{p-1} + \theta_p$, $X_0 = 0$, where θ_p is a sequence of i.i.d. Gaussian random variables $\mathcal{N}(0,1)$.

$$p_{\rm s} = \mathbb{P}(X_M \geq y_{\rm s}) = \frac{1}{\sqrt{2\pi M}} \int_{y_{\rm s}}^{\infty} \exp\big(-\frac{x^2}{2M}\big) dx$$

MC: sample n trajectories $(X_i^{(k)})_{i=0}^M$, $k = 1, \ldots, n$, and estimate

$$\hat{P}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_M^{(k)} \geq y_{\rm s}}, \qquad \hat{e}_n = \frac{1}{\sqrt{n}} \left(\frac{1}{\hat{P}_n} - 1 \right)^{1/2}$$

IS: sample n trajectories $(X_p^{(k)})_{p=0}^M$, $k = 1, \ldots, n$, with biased distribution $\theta_p \sim \mathcal{N}(a/M, 1)$ and estimate:

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_M^{(k)} \geq y_{\rm s}} \exp\left(\frac{y_{\rm s}^2}{2M} - \frac{y_{\rm s}}{M} X_M^{(k)} \right)$$

$$\hat{e}_n = \frac{1}{\sqrt{n}} \left\{ \frac{\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_M^{(k)} \geq y_{\rm s}} \exp\left(\frac{y_{\rm s}^2}{M} - \frac{2y_{\rm s}}{M} X_M^{(k)} \right)}{\hat{I}_n^2} - 1 \right\}^{1/2}$$

- Adaptive importance sampling: choice of a family of biased distributions

$$\{q_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q\}$$

- Affine transform

Change of mean $q_{\boldsymbol{\theta}_\mu}(\boldsymbol{x}) = p(\boldsymbol{x} - \boldsymbol{\theta}_\mu)$

Change of mean and variance $q_{\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma}(\boldsymbol{x}) = |\det \boldsymbol{\theta}_\sigma|^{-1} p\big(\boldsymbol{\theta}_\sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta}_\mu)\big)$

Easy to implement $\tilde{\boldsymbol{X}} = \boldsymbol{\theta}_\sigma \boldsymbol{X} + \boldsymbol{\theta}_\mu \sim q_{\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma}$ iff $\boldsymbol{X} \sim p$.

Often used with Gaussian distributions.

- Adaptive importance sampling: choice of $\boldsymbol{\theta}$.
- Variance minimization.

Principle: minimization of the variance of the estimator:

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_p\left[h(\boldsymbol{X})^2 \frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X})\right]$$

Empirically

$$\hat{\boldsymbol{\theta}}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)})^2 \frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}^{(k)}), \qquad \boldsymbol{X}^{(k)} \text{ i.i.d. with pdf } p$$

Studied with Gaussian distribution and change of mean and variance.
With additional hypotheses, $\hat{\boldsymbol{\theta}}_n \to \hat{\boldsymbol{\theta}}^*$ a.s., with central limit theorem.
But: the likelihood ratio is strongly fluctuating.

[B. Jourdain and J. Lelong. Robust adaptive importance sampling for normal random vectors. Ann. Appl. Probab. **19**, 1687-1718, 2009.]

- Adaptive importance sampling: choice of $\boldsymbol{\theta}$.
- Cross entropy

Let q^* be the optimal pdf $q^*(\boldsymbol{x}) = \frac{h(\boldsymbol{x})}{\mathbb{E}_p[h(\boldsymbol{X}))]} p(\boldsymbol{x})$ (here $h \geq 0$).

Principle: minimization of the Kullback-Leibleir distance between q^* and $q_{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}} D(q^*, q_{\boldsymbol{\theta}}), \qquad D(q^*, q_{\boldsymbol{\theta}}) = \mathbb{E}_{q^*}\Big[\log \frac{q^*}{q_{\boldsymbol{\theta}}}(\boldsymbol{X})\Big]$$

In fact

$$\operatorname*{argmin}_{\boldsymbol{\theta}} D(q^*, q_{\boldsymbol{\theta}}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_p\Big[h(\boldsymbol{X}) \log \frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X})\Big]$$

Empirically

$$\hat{\boldsymbol{\theta}}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) \log \frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}^{(k)}), \qquad \boldsymbol{X}^{(k)} \text{ i.i.d. with pdf } p$$

With additional hypotheses, $\hat{\boldsymbol{\theta}}_n \to \hat{\boldsymbol{\theta}}^*$ a.s., with central limit theorem.

The log-likelihood ratio is less fluctuating than the likelihood ratio used in the variance minimization method.

- Adaptive importance sampling: choice of $\boldsymbol{\theta}$.
- Cross entropy

For any $\boldsymbol{\theta}_0$,

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \; D(q^*, q_{\boldsymbol{\theta}}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \; \mathbb{E}_{q_{\boldsymbol{\theta}_0}} \left[h(\boldsymbol{X}) \log \left(\frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}) \right) \frac{p}{q_{\boldsymbol{\theta}_0}}(\boldsymbol{X}) \right]$$

Empirically, if $\boldsymbol{X}^{(k)}$ i.i.d. with pdf $q_{\boldsymbol{\theta}_0}$, then

$$\hat{\boldsymbol{\theta}}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}} \; \frac{1}{n} \sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) \log \left(\frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}^{(k)}) \right) \frac{p}{q_{\boldsymbol{\theta}_0}}(\boldsymbol{X}^{(k)})$$

- Adaptive importance sampling: choice of $\boldsymbol{\theta}$.
- Recursive cross entropy

Principle: go progressively towards $\boldsymbol{\theta}^*$.

Algorithm:

Step 0: Set $\boldsymbol{\theta}_0$ (e.g. $\boldsymbol{\theta}_0 = \mathbf{0}$, $q_{\boldsymbol{\theta}_0} = p$).

Step $j \geq 1$: Generate $\boldsymbol{X}^{(k)}$ i.i.d. with pdf $q_{\boldsymbol{\theta}_{j-1}}$

$$\boldsymbol{\theta}_j \in \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N_{j-1}} \sum_{k=1}^{N_{j-1}} h(\boldsymbol{X}^{(k)}) \log \left(\frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}^{(k)}) \right) \frac{p}{q_{\boldsymbol{\theta}_{j-1}}}(\boldsymbol{X}^{(k)}),$$

with N_{j-1} large enough, until a stopping criterium is met, e.g. $|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}| \leq \delta$.

Problem: if $h(\boldsymbol{x}) = \mathbf{1}_{f(\boldsymbol{x}) \geq y_{\mathrm{s}}}$, then one needs realizations that achieve $f(\boldsymbol{X}^{(k)}) \geq y_{\mathrm{s}}$ during the first step $(\boldsymbol{\theta}_0, N_0)$.

- Adaptive importance sampling: choice of $\boldsymbol{\theta}$.
- Adaptive cross entropy

Assume $h(\boldsymbol{x}) = \mathbf{1}_{f(\boldsymbol{x}) \geq y_{\mathrm{s}}}$.
Principle: go progressively towards $\boldsymbol{\theta}^*$ by "increasing" y_{s}.
Algorithm:
Step 0: Set $\boldsymbol{\theta}_0$ (e.g. $\boldsymbol{\theta}_0 = \mathbf{0}$, $q_{\boldsymbol{\theta}_0} = p$), y_0, and α (e.g. $\alpha = 0.1$).
Step $j \geq 1$: Generate $\boldsymbol{X}^{(k)}$ i.i.d. with pdf $q_{\boldsymbol{\theta}_{j-1}}$

$$\boldsymbol{\theta}_j \in \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N_{j-1}} \sum_{k=1}^{N_{j-1}} \mathbf{1}_{f(\boldsymbol{X}^{(k)}) \geq y_{j-1}} \log \left(\frac{p}{q_{\boldsymbol{\theta}}}(\boldsymbol{X}^{(k)}) \right) \frac{p}{q_{\boldsymbol{\theta}_{j-1}}}(\boldsymbol{X}^{(k)}),$$

$$y_j = (1-\alpha)-\text{empirical quantile of } (f(\boldsymbol{X}^{(k)}))_{k=1}^{N_{j-1}}$$

with N_{j-1} large enough, until $y_j \geq y_{\mathrm{s}}$.
Estimation of I with or without recycling.

[P.T. De Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein. A tutorial on the cross-entropy method, Annals of Operations Research **134**, 19-67, 2005.]

Control variates

- The goal is to estimate $I = \mathbb{E}[h(\boldsymbol{X})]$ for $\boldsymbol{X}$ a random vector and $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ a deterministic function.
- Assume that we have a reduced model $f_{\mathrm{r}}(\boldsymbol{X})$.

- Importance sampling method:

First we evaluate (we approximate) the optimal density $q_{\mathrm{opt,r}}(\boldsymbol{x}) = \frac{g(f_{\mathrm{r}}(\boldsymbol{x}))p(\boldsymbol{x})}{I_{\mathrm{r}}}$, with $I_{\mathrm{r}} = \int g(f_{\mathrm{r}}(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$.
Second we use it as a biased density for estimating I.
Dangerous, use conservative version.

Control variates

- The goal is to estimate $I = \mathbb{E}[h(\boldsymbol{X})]$ for $\boldsymbol{X}$ a random vector and $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ a deterministic function.
- Assume that we have a reduced model $f_{\mathrm{r}}(\boldsymbol{X})$.

- Control variates method:

We denote $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$, $h_{\mathrm{r}}(\boldsymbol{x}) = g(f_{\mathrm{r}}(\boldsymbol{x}))$.
We assume that we know $I_{\mathrm{r}} = \mathbb{E}[h_{\mathrm{r}}(\boldsymbol{X})]$.
By considering the representation

$$I = \mathbb{E}[h(\boldsymbol{X})] = I_{\mathrm{r}} + \mathbb{E}[h(\boldsymbol{X}) - h_{\mathrm{r}}(\boldsymbol{X})]$$

we propose the estimator:

$$\hat{I}_n = I_{\mathrm{r}} + \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) - h_{\mathrm{r}}(\boldsymbol{X}^{(k)}),$$

where $(\boldsymbol{X}^{(k)})_{k=1}^{n}$ is a n-sample (with the pdf p).

- Estimator:

$$\hat{I}_n = I_r + \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) - h_r(\boldsymbol{X}^{(k)})$$

- The estimator is unbiased:

$$\begin{aligned}\mathbb{E}\big[\widehat{I_n}\big] &= I_r + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[h(\boldsymbol{X}^{(k)}) - h_r(\boldsymbol{X}^{(k)})\big] = I_r + \mathbb{E}[h(\boldsymbol{X})] - \mathbb{E}[h_r(\boldsymbol{X})] \\ &= I_r + \mathbb{E}[h(\boldsymbol{X})] - I_r = I\end{aligned}$$

- The estimator is convergent:

$$\widehat{I_n} \overset{n\to\infty}{\longrightarrow} I_r + \mathbb{E}\big[h(\boldsymbol{X}) - h_r(\boldsymbol{X})\big] = I$$

- The variance of the estimator is:

$$\mathrm{Var}(\widehat{I_n}) = \frac{1}{n}\mathrm{Var}\big[h(\boldsymbol{X}) - h_r(\boldsymbol{X})\big]$$

$\hookrightarrow$ The use of a reduced model can reduce the variance.

- Example: we want to estimate

$$I = \mathbb{E}[h(X)]$$

with $X \sim \mathcal{U}(0,1)$, $h(x) = \exp(x)$.
Result: $I = e - 1 \simeq 1.72$.
Monte Carlo.

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \exp[X^{(k)}]$$

Variance of the MC estimator$= \frac{1}{n}(2e-1) \simeq \frac{1}{n} 4.44$.
Control Variates. Reduced model: $h_{\mathrm{r}}(x) = 1 + x$ (here $I_{\mathrm{r}} = \frac{3}{2}$). CV estimator:

$$\hat{I}_n = I_{\mathrm{r}} + \frac{1}{n} \sum_{k=1}^{n} \left\{ \exp[X^{(k)}] - 1 - X^{(k)} \right\}$$

Variance of the CV estimator $= \frac{1}{n}(3e - \frac{e^2}{2} - \frac{53}{12}) \simeq \frac{1}{n} 0.044$.
The CV method needs 100 times less simulations to reach the same precision as MC !

- Application: estimation of

$$I = \mathbb{E}[g(f(\boldsymbol{X}))]$$

We have a reduced model f_{r} of the full code f. The ratio between the computational cost of one call of f and one call of f_{r} is $q > 1$.
Estimator

$$\hat{I}_n = \frac{1}{n_{\mathrm{r}}} \sum_{k=1}^{n_{\mathrm{r}}} h_{\mathrm{r}}(\tilde{\boldsymbol{X}}^{(k)}) + \frac{1}{n} \sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) - h_{\mathrm{r}}(\boldsymbol{X}^{(k)})$$

with $n_{\mathrm{r}} > n$, $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$, $h_{\mathrm{r}}(\boldsymbol{x}) = g(f_{\mathrm{r}}(\boldsymbol{x}))$.
Allocation between calls to the complete code and calls to the reduced model can be optimized with the contraint $n_{\mathrm{r}}/q + n(1 + 1/q) = n_{\mathrm{tot}}$:

$$\frac{n}{n_{\mathrm{tot}}} = \frac{q}{1+q} \frac{1}{1 + \frac{1}{\sqrt{1+q}} \frac{\sqrt{\mathrm{Var}(h_{\mathrm{r}}(\boldsymbol{X}))}}{\sqrt{\mathrm{Var}((h-h_{\mathrm{r}})(\boldsymbol{X}))}}}$$

Classical trade-off between approximation error and estimation error.
Used when $f(\boldsymbol{X})$ is the solution of an ODE or PDE with fine grid, while $f_{\mathrm{r}}(\boldsymbol{X})$ is the solution obtained with a coarse grid (MultiLevel Monte Carlo).

- Optimal control variate

By considering the representation (for a fixed λ)

$$I = \mathbb{E}[h(\boldsymbol{X})] = \lambda I_{\mathrm{r}} + \mathbb{E}[h(\boldsymbol{X}) - \lambda h_{\mathrm{r}}(\boldsymbol{X})]$$

we propose the estimator:

$$\hat{I}_n = \lambda I_{\mathrm{r}} + \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) - \lambda h_{\mathrm{r}}(\boldsymbol{X}^{(k)}),$$

where $(\boldsymbol{X}^{(k)})_{k=1}^{n}$ is a n-sample (with the pdf p).
The estimator is unbiased and consistent for any λ. The variance is

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\mathrm{Var}\big(h(\boldsymbol{X}) - \lambda h_{\mathrm{r}}(\boldsymbol{X})\big)$$

The λ that minimizes the variance is

$$\lambda = \frac{\mathrm{Cov}\big(h(\boldsymbol{X}), h_{\mathrm{r}}(\boldsymbol{X})\big)}{\mathrm{Var}\big(h_{\mathrm{r}}(\boldsymbol{X})\big)}$$

and then

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\mathrm{Var}\big(h(\boldsymbol{X})\big)\big(1-\rho^2\big), \qquad \rho = \mathrm{Corr}\big(h(\boldsymbol{X}), h_{\mathrm{r}}(\boldsymbol{X})\big)$$

- Empirical optimal control variate

$$\hat{I}_n = \hat{\lambda}_n I_{\mathrm{r}} + \frac{1}{n}\sum_{k=1}^{n} h(\boldsymbol{X}^{(k)}) - \hat{\lambda}_n h_{\mathrm{r}}(\boldsymbol{X}^{(k)}),$$

with

$$\hat{\lambda}_n = \frac{\sum_{k=1}^{n}\left(h(\boldsymbol{X}^{(k)}) - \frac{1}{n}\sum_{k'=1}^{n} h(\boldsymbol{X}^{(k')})\right)\left(h_{\mathrm{r}}(\boldsymbol{X}^{(k)}) - \frac{1}{n}\sum_{k'=1}^{n} h_{\mathrm{r}}(\boldsymbol{X}^{(k')})\right)}{\sum_{k=1}^{n}\left(h_{\mathrm{r}}(\boldsymbol{X}^{(k)}) - \frac{1}{n}\sum_{k'=1}^{n} h_{\mathrm{r}}(\boldsymbol{X}^{(k')})\right)^2}$$

The estimator is slightly biased ($O(1/n)$).
We have the (optimal) asymptotic normality result

$$\sqrt{n}\left(\hat{I}_n - I\right) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}\left(0, \mathrm{Var}\left(h(\boldsymbol{X})\right)\left(1-\rho^2\right)\right)$$

Stratification

Principle: The sample $(\boldsymbol{X}^{(k)})_{k=1}^n$ is enforced to obey prescribed proportions in some "strata".
Method used in polls (representative sample).
Here: we want to estimate $\mathbb{E}[h(\boldsymbol{X})]$, $\boldsymbol{X}$ with values in D.

- Two ingredients:

i) A partition of the state space $D = \bigcup_{i=1}^m D_i$. We know $p_i = \mathbb{P}(\boldsymbol{X} \in D_i)$.
ii) Total probability formula:

$$I = \mathbb{E}[h(\boldsymbol{X})] = \sum_{i=1}^m \underbrace{\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]}_{J_i} \quad \underbrace{\mathbb{P}(\boldsymbol{X} \in D_i)}_{p_i}$$

- Estimation:

1) For all $i = 1, \ldots, m$, J_i is estimated by MC with n_i simulations:

$$\widehat{J}_{i,n_i} = \frac{1}{n_i} \sum_{k=1}^{n_i} h(\boldsymbol{X}^{(i,k)}), \qquad \boldsymbol{X}^{(i,k)} \sim \mathcal{L}(\boldsymbol{X}|\boldsymbol{X} \in D_i) \text{ ind.}$$

2) The estimator is $\widehat{I}_n = \sum_{i=1}^m \widehat{J}_{i,n_i} p_i$

$$\widehat{I}_n = \sum_{i=1}^{m} p_i \widehat{J}_{i,n_i}, \qquad \widehat{J}_{i,n_i} = \frac{1}{n_i} \sum_{k=1}^{n_i} h(\boldsymbol{X}^{(i,k)}), \quad \boldsymbol{X}^{(i,k)} \sim \mathcal{L}(\boldsymbol{X} | \boldsymbol{X} \in D_i)$$

The total number of simulations is $n = \sum_{i=1}^{m} n_i$.

- The estimator is unbiased, convergent and its variance is

$$\mathrm{Var}\big(\widehat{I}_n\big)_S = \sum_{i=1}^{m} p_i^2 \mathrm{Var}\big(\widehat{J}_{i,n_i}\big) = \sum_{i=1}^{m} p_i^2 \frac{\sigma_i^2}{n_i}, \text{ with } \sigma_i^2 = \mathrm{Var}\big(h(\boldsymbol{X}) | \boldsymbol{X} \in D_i\big)$$

The user is free to choose the n_i (with the constraint $\sum_{i=1}^{m} n_i = n$).

- **Proportional stratification**: $n_i = p_i n$.

$$\widehat{I}_n = \sum_{i=1}^{m} \frac{p_i}{n_i} \sum_{k=1}^{n_i} h(\boldsymbol{X}^{(i,k)}) = \frac{1}{n} \sum_{i=1}^{m} \sum_{k=1}^{n_i} h(\boldsymbol{X}^{(i,k)}), \quad \boldsymbol{X}^{(i,k)} \sim \mathcal{L}(\boldsymbol{X} | \boldsymbol{X} \in D_i)$$

Then

$$\mathrm{Var}\big(\widehat{I}_n\big)_{SP} = \frac{1}{n} \sum_{i=1}^{m} p_i \sigma_i^2$$

$$\mathrm{Var}\big(\widehat{I}_n\big)_{MC} = \frac{1}{n} \mathrm{Var}\big(h(\boldsymbol{X})\big) \geq \frac{1}{n} \sum_{i=1}^{m} p_i \sigma_i^2 = \mathrm{Var}\big(\widehat{I}_n\big)_{SP}$$

Proof: We have

$$\begin{aligned}
\mathbb{E}[h(\boldsymbol{X})]^2 &= \Big(\sum_{i=1}^m p_i \mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]\Big)^2 \\
&\leq \Big(\sum_{i=1}^m p_i\Big)\Big(\sum_{i=1}^m p_i \mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]^2\Big) \\
&= \sum_{i=1}^m p_i \mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]^2
\end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{i=1}^m p_i \sigma_i^2 &= \sum_{i=1}^m p_i \Big(\mathbb{E}[h(\boldsymbol{X})^2|\boldsymbol{X} \in D_i] - \mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]^2\Big) \\
&= \mathbb{E}[h(\boldsymbol{X})^2] - \sum_{i=1}^m p_i \mathbb{E}[h(\boldsymbol{X})|\boldsymbol{X} \in D_i]^2 \\
&\leq \mathbb{E}[h(\boldsymbol{X})^2] - \mathbb{E}[h(\boldsymbol{X})]^2 = \mathrm{Var}(h(\boldsymbol{X}))
\end{aligned}$$

However, the proportional allocation is not optimal !

- The optimal allocation is the one that minimizes the variance

$\mathrm{Var}(\widehat{I}_n)_S = \sum_{i=1}^m p_i^2 \frac{\sigma_i^2}{n_i}$.

It is the solution of the minimization problem: find $(n_i)_{i=1,\dots,m}$ minimizing

$$\sum_{i=1}^m p_i^2 \frac{\sigma_i^2}{n_i} \text{ with the constraint } \sum_{i=1}^m n_i = n$$

Solution (optimal allocation, obtained with Lagrange multiplier method):

$$n_i = n \frac{p_i \sigma_i}{\sum_{l=1}^m p_l \sigma_l}$$

The minimal variance is

$$\mathrm{Var}\left(\widehat{I}_n\right)_{SO} = \frac{1}{n} \left(\sum_{i=1}^m p_i \sigma_i \right)^2,$$

We have:

$$\mathrm{Var}\left(\widehat{I}_n\right)_{SO} \leq \mathrm{Var}\left(\widehat{I}_n\right)_{SP} \leq \mathrm{Var}\left(\widehat{I}_n\right)_{MC}$$

Practically: the σ_i's are unknown, therefore the optimal allocation is unknown (principle of an adaptive method).

- Example: we want to estimate

$$\mathbb{E}[h(X)]$$

with $X \sim \mathcal{U}(-1,1)$, $h(x) = \exp(x)$.
Result: $\mathbb{E}[\exp(X)] = \sinh(1) \simeq 1.18$.
MC. With a sample $X^{(1)}$, ..., $X^{(n)}$ with the distribution $\mathcal{U}(-1,1)$

$$\widehat{I}_n = \frac{1}{n}\sum_{k=1}^{n} \exp[X^{(k)}]$$

Variance of the estimator $= \frac{1}{n}(\frac{1}{2} - \frac{e^{-2}}{2}) \simeq \frac{1}{n}0.43$.
Proportional stratification. With a sample
- $X^{(1)}$, ..., $X^{(n/2)}$ with the distribution $\mathcal{U}(-1,0)$,
- $X^{(n/2+1)}$, ..., $X^{(n)}$ with the distribution $\mathcal{U}(0,1)$.

$$\widehat{I}_n = \frac{1}{n}\sum_{k=1}^{n/2} \exp[X^{(k)}] + \frac{1}{n}\sum_{k=n/2+1}^{n} \exp[X^{(k)}] = \frac{1}{n}\sum_{k=1}^{n} \exp[X^{(k)}]$$

Variance of the PS estimator $\simeq \frac{1}{n}0.14$.
Here PS needs 3 times less simulations to reach the same precision as MC.

Nonproportional stratification. With a sample
- $X^{(1)}$, ..., $X^{(n/4)}$ with the distribution $\mathcal{U}(-1,0)$,
- $X^{(n/4+1)}$, ..., $X^{(n)}$ with the distribution $\mathcal{U}(0,1)$.

$$\widehat{I_n} = \frac{2}{n}\sum_{k=1}^{n/4} \exp[X^{(k)}] + \frac{1}{2n}\sum_{k=n/4+1}^{n} \exp[X^{(k)}]$$

Variance of the estimator $\simeq \frac{1}{n}0.048$.
The stratification method needs 9 times less simulations to reach the same precision as MC.
Similar to importance sampling with a stepwise constant biased pdf.

Antithetic variables

- We want to compute

$$I = \int_{[0,1]^d} h(\boldsymbol{x}) d\boldsymbol{x}$$

MC with a n-sample $(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)})$ with the distribution $\mathcal{U}([0,1]^d)$:

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} h(\boldsymbol{X}^{(k)})$$

$$\mathbb{E}\big[(\hat{I}_n - I)^2\big] = \frac{1}{n} \mathrm{Var}(h(\boldsymbol{X})) = \frac{1}{n} \Big(\int_{[0,1]^d} h^2(\boldsymbol{x}) d\boldsymbol{x} - I^2 \Big)$$

- We consider the representations

$$I = \int_{[0,1]^d} h(1 - \boldsymbol{x}) d\boldsymbol{x} \text{ and } I = \int_{[0,1]^d} \frac{h(\boldsymbol{x}) + h(1 - \boldsymbol{x})}{2} d\boldsymbol{x}$$

MC with a $n/2$-sample $(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n/2)})$ distributed as $\mathcal{U}([0,1]^d)$:

$$\tilde{I}_n = \frac{1}{n} \sum_{k=1}^{n/2} h(\boldsymbol{X}^{(k)}) + h(1 - \boldsymbol{X}^{(k)})$$

- MC estimator with the sample $(\tilde{\boldsymbol{X}}^{(1)}, \ldots, \tilde{\boldsymbol{X}}^{(n)}) := (\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n/2)}, 1 - \boldsymbol{X}^{(1)}, \ldots, 1 - \boldsymbol{X}^{(n/2)})$ that is not i.i.d.:

$$\tilde{I}_n = \frac{1}{n} \sum_{k=1}^{n} h(\tilde{\boldsymbol{X}}^{(k)})$$

The function f is called n times.

- Error:

$$\begin{aligned} \mathbb{E}\big[(\tilde{I}_n - I)^2\big] &= \frac{1}{n}\Big(\mathrm{Var}(h(\boldsymbol{X})) + \mathrm{Cov}(h(\boldsymbol{X}), h(1-\boldsymbol{X}))\Big) \\ &= \frac{1}{n}\Big(\int_{[0,1]^d} h^2(\boldsymbol{x}) + h(\boldsymbol{x})h(1-\boldsymbol{x})d\boldsymbol{x} - 2I^2\Big) \end{aligned}$$

The variance is reduced if $\mathrm{Cov}(h(\boldsymbol{X}), h(1-\boldsymbol{X})) < 0$.
Sufficient condition: h is monotoneous.
Proof: If $\boldsymbol{X}, \boldsymbol{X}'$ i.i.d.

$$\begin{aligned} [h(\boldsymbol{X}) - h(\boldsymbol{X}')][-h(1-\boldsymbol{X}) + h(1-\boldsymbol{X}')] &\geq 0 \text{ a.s.} \\ -2\mathbb{E}\big[h(\boldsymbol{X})h(1-\boldsymbol{X})\big] + 2\mathbb{E}\big[h(\boldsymbol{X})\big]^2 &\geq 0 \end{aligned}$$

- Example:

$$I = \int_0^1 \frac{1}{1+x} dx$$

Result: $I = \ln 2$.

Monte Carlo:

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{1 + X^{(k)}}$$

$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\left(\int_0^1 (1+x)^{-2} dx - \ln 2^2 \right) = \frac{1}{n}\left(\frac{1}{2} - \ln 2^2\right) \simeq \frac{1}{n} 1.95\, 10^{-2}$.

Antithetic variables:

$$\tilde{I}_n = \frac{1}{n} \sum_{k=1}^{n/2} \frac{1}{1 + X^{(k)}} + \frac{1}{2 - X^{(k)}}$$

$\mathrm{Var}(\tilde{I}_n) = \frac{2}{n}\left(\int_0^1 \left(\frac{1}{2(1+x)} + \frac{1}{2(2-x)}\right)^2 dx - \ln 2^2 \right) \simeq \frac{1}{n} 1.2\, 10^{-3}$.

The AV method requires 15 times less simulations than MC.

• More generally: one needs to find a pair $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ such that $h(\boldsymbol{X})$ and $h(\tilde{\boldsymbol{X}})$ have the same distribution and $\mathrm{Cov}(h(\boldsymbol{X}), h(\tilde{\boldsymbol{X}})) < 0$.

• Monte Carlo with an i.i.d. sample $((\boldsymbol{X}^{(1)}, \tilde{\boldsymbol{X}}^{(1)}), \ldots, (\boldsymbol{X}^{(n/2)}, \tilde{\boldsymbol{X}}^{(n/2)}))$:

$$\tilde{I}_n = \frac{1}{n} \sum_{k=1}^{n/2} h(\boldsymbol{X}^{(k)}) + h(\tilde{\boldsymbol{X}}^{(k)})$$

$$\mathbb{E}\big[(\tilde{I}_n - I)^2\big] = \frac{1}{n}\Big(\mathrm{Var}(h(\boldsymbol{X})) + \mathrm{Cov}(h(\boldsymbol{X}), h(\tilde{\boldsymbol{X}}))\Big)$$

• Recent application: computation of effective tensors in stochastic homogenization (the effective tensor is the expectation of a functional of the solution of an elliptic PDE with random coefficients; antithetic pairs of the realizations of the composite medium are sampled; gain by a factor 3; in fact, better results with control variates; cf C. Le Bris, F. Legoll).

• Not very useful for the estimation of probabilities of rare events.