Modeling Extreme Events from Computer Simulations
Part III

Emmanuel Vazquez

SUPELEC, Gif-sur-Yvette, France

Summer School CEA-EDF-INRIA, 2011

Outline of Part III

1. Estimating a probability of failure by Monte Carlo

Monte Carlo integration with importance sampling

Recall the importance-sampling estimator of a probability of failure

- Assume given a probability $\mathrm{P}_{\mathbb{X}}$ with density p, on a factor space $\mathbb{X}$, a function $f : \mathbb{X} \to \mathbb{R}$, and a threshold $u \in \mathbb{R}$
- Choose an instrumental distribution $\mathrm{Q}_{\mathbb{X}}$ on $\mathbb{X}$, with density q
- Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathrm{Q}_{\mathbb{X}}$
- Then

$$\alpha_n = \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1}_{f(X_i)>u}$$

  with $w_i = \frac{p(X_i)}{q(X_i)}$, $i = 1 \ldots n$, is an unbiased estimator of $\alpha = \int \mathbb{1}_{f>u}$.
- The random variable $Z_i = \mathbb{1}_{f(X_i)>u}$ has a Bernoulli distribution $\mathrm{B}(\tilde{\alpha})$, with $\tilde{\alpha} = \mathrm{Q}_{\mathbb{X}}(f > u)$
- Thus,

$$\mathrm{var}(\alpha_n) = \frac{\sum_{i=1}^{n} w_i^2}{n^2} \tilde{\alpha}(1 - \tilde{\alpha})$$

  which is minimum if $q = q^\star$, with

$$q^\star(x) = \frac{\mathbb{1}_{f(x)>u}\, p(x)}{\alpha}$$

Monte Carlo integration with importance sampling

- What makes the problem of computing α difficult?
- The difficulty is to choose $Q_{\mathbb{X}}$ in such a way that there is a high proportion of points $X_1, \ldots, X_n$ in the domain of failure $\Gamma = \{x; f(x) > u\}$
- Γ small, unknown set ➟ it is not possible to find a good instrumental density $Q_{\mathbb{X}}$ before any evaluation is made
- This observation being made, the idea is then to consider an adaptive strategy: after having made some evaluations of f, and if f is reasonably smooth, we may have an idea of regions of $\mathbb{X}$ that are interesting to explore in order to find Γ
- Two main routes have been proposed in the literature
  - sequential importance sampling
  - control variate sampling

2. Sequential Monte Carlo for estimating a probability of failure

Some reminders about MCMC

MCMC: given a probability $\mathrm{P}_{\mathbb{X}}$ on a measurable space $(\mathbb{X}, \mathcal{X})$, construct a Markov chain $(X_n)_{n\in\mathbb{N}}$ in such a way that p is an invariant density of the chain

Markov transitions

- A Markov transition (or Markov kernel) on $\mathbb{X}$ is a set of probability distributions

$$\{K(x,\cdot)\,; x \in \mathbb{X}\}$$

such that for any measurable subset $A \in \mathcal{X}$, $x \mapsto K(x, A)$ is measurable application.
➠ $K(x, A)$ is "the probability to go to A starting from x".

- Given a kernel K, we can define two integral operations
  1. If $f : \mathbb{X} \to \mathbb{R}$ is a measurable and bounded function, define $Kf : \mathbb{X} \to \mathbb{R}$ by

$$(Kf)(x) = \int f(y)K(x, \mathrm{d}y)\,, \quad x \in \mathbb{X}$$

  2. If μ is a probability on $(\mathbb{X}, \mathcal{X})$, define a measure μK by

$$(\mu K)(A) = \int K(y, A)\mathrm{d}\mu(y)\,, \quad A \in \mathcal{X}$$

- Given two kernels K_1 and K_2, define a composite kernel K_1K_2 by

$$(K_1K_2)(x, A) = (K_1(x,\cdot)K_2)(A) = \int K_1(x, \mathrm{d}y)K_2(y, A)\,, \quad (x, A) \in \mathbb{X} \times \mathcal{X}$$

$(K_1K_2)(x, A)$ is the probability to from x to A using a first transition K_1 and a second transition K_2

- Given a kernel K, the iterated kernel K^n, $n \geq 1$, is defined by induction using the composition rule

Some reminders about MCMC

Markov chains

A random process $(X_n)_{n\in\mathbb{N}}$ is a Markov chain if there exists a sequence of Markov transitions $(K_n)_{n\geq 1}$ such that for all measurable and bounded function $f : \mathbb{X} \to \mathbb{R}$

$$\mathrm{E}[f(X_{n+1}) \mid X_n, \ldots, X_0] = (K_{n+1}f)(X_n) \quad \text{a.s}$$

(X_n) is said to be stationary or homogeneous if for all n, $K_n = K$ for some K

Invariant measures

Let (X_n) be a homogeneous Markov Chain. A probability measure π is an invariant measure of (X_n) if

$$\pi K = \pi$$

Foundation of MCMC for the estimation of a probability of failure

Let (X_n) be a π-invariant Markov Chain. Under certain conditions, given $\phi \in L^1$,

$$\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) \to \int_{\mathbb{X}} \phi \mathrm{d}\pi \quad \text{a.s}$$

In particular, if $\mathrm{P}_{\mathbb{X}}$ is invariant for (X_n),

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)>u} \to \alpha^u(f) \quad \text{a.s}$$

Some reminders about MCMC

Given a probability distribution $\mathrm{P}_{\mathbb{X}}$, how to construct a kernel K such that (X_n) is $\mathrm{P}_{\mathbb{X}}$-invariant?

(NB: of course we can choose $K(x, \cdot) := \mathrm{P}_{\mathbb{X}}$)

Metropolis-Hastings algorithm

- Given a probability distribution $\mathrm{P}_{\mathbb{X}}$, with density p, the Metropolis-Hastings algorithm makes it possible to construct a $\mathrm{P}_{\mathbb{X}}$-invariant Markov chain (X_n)
- Consider a kernel Q such that $\forall x \in \mathbb{X}$, $Q(x, \cdot)$ has a density $q(\cdot \mid x)$
- Given $X_n = x$

  1. Generate $Y \sim \mathrm{Q}(x, \cdot)$
  2. Take

$$X_{n+1} = \begin{cases} Y & \text{with probability } \rho(x, Y) \\ x & \text{with probability } 1 - \rho(x, Y) \end{cases}$$

  with

$$\rho(x, y) = \min\left\{ \frac{p(y)}{p(x)} \frac{q(x \mid y)}{q(y \mid x)}, 1 \right\}$$

- The transition kernel of the chain (X_n) is

$$K(x, \mathrm{d}z) = \rho(x, z) Q(x, \mathrm{d}z) + (1 - r(x)) \delta_x(\mathrm{d}z)$$

  with $r(x) = \int \rho(x, y) Q(x, \mathrm{d}y)$
- We have $\mathrm{P}_{\mathbb{X}} = \mathrm{P}_{\mathbb{X}} K$

Sequential Monte Carlo

- With MCMC methods, it is often difficult to assess when the Markov chain has reached its stationary regime
- Moreover, MCMC methods are designed to sample from a fixed distribution π
- Sequential Monte Carlo methods address these limitations by running several Markov chains in parallel: $X^{(i)} = (X_n^{(i)})$, $i = 1, \ldots, N$
- At time n, each chain $X^{(i)}$ is given a weight $w_n^{(i)}$, which is determined so that, for a distribution π_n, and a function $\phi \in L^1(\pi_n)$, we have

$$\sum_{i=1}^{N} w_n^{(i)} \phi(X_n^{(i)}) \to_N \int \phi \mathrm{d}\pi_n \quad \text{a.s.}$$

- A pair $(w_n^{(i)}, X_n^{(i)})$ is called a particle
- Depending on the application, several methods have been proposed in the literature to determine the weights and the transition kernels
- Here, we shall focus only on the problem of the estimation of a probability of failure

Estimation of a probability of failure by subset simulation

- The idea of subset simulation is the following:
  ➠ when u is a high threshold, it may be difficult to deal with the problem of estimating α^u, but we can try to decompose the problem into a series of easier problems
- If u is not too high, then we can get a good approximation of α^u with only a few evaluations of f
- Consider a finite sequence of increasing thresholds

$$-\infty = u_0 < u_1 < u_2 \cdots < u_S = u$$

and define the corresponding sequence of nested subsets

$$\Gamma_k = \{x \in \mathbb{X}; f(x) > u_k\}, \quad k = 0, \ldots, S$$

- We can write

$$\begin{aligned}
\alpha^u &= \mathrm{P}_{\mathbb{X}}(\Gamma) = \mathrm{P}_{\mathbb{X}}\left(\bigcap_{k=1}^{S} \Gamma_k\right) = \mathrm{P}_{\mathbb{X}}\left(\Gamma_S \,\middle|\, \bigcap_{k=1}^{S-1} \Gamma_k\right) \mathrm{P}_{\mathbb{X}}\left(\bigcap_{k=1}^{S-1} \Gamma_k\right) \\
&= \mathrm{P}_{\mathbb{X}}\left(\Gamma_S \mid \Gamma_{S-1}\right) \mathrm{P}_{\mathbb{X}}\left(\bigcap_{k=1}^{S-1} \Gamma_k\right) \\
&= \prod_{k=1}^{S} \mathrm{P}_{\mathbb{X}}(\Gamma_k \mid \Gamma_{k-1})
\end{aligned}$$

Estimation of a probability of failure by subset simulation

- Thus, α^u can be computed as a product of the probabilities $\mathrm{P}_{\mathbb{X}}(\Gamma_{k+1} \mid \Gamma_k)$
- How to compute/estimate a probability $\mathrm{P}_{\mathbb{X}}(\Gamma_k \mid \Gamma_{k-1})$?
- For $k \geq 0$, denote by μ_k the normalized restriction of $\mathrm{P}_{\mathbb{X}}$ to the domain Γ_k

$$\mu_k(\mathrm{d}x) = \frac{1}{\mathrm{P}_{\mathbb{X}}(\Gamma_k)} \mathbb{1}_{\Gamma_k}(x) \mathrm{P}_{\mathbb{X}}(\mathrm{d}x)$$

- In particular, we have

$$\mu_0 = \mathrm{P}_{\mathbb{X}}$$

and

$$\mu_S(\mathrm{d}x) = \frac{1}{\alpha} \mathbb{1}_{\Gamma}(x) \mathrm{P}_{\mathbb{X}}(\mathrm{d}x),$$

which is the optimal instrumental distribution for estimating α(!)

- We have $\mathrm{P}_{\mathbb{X}}(\Gamma_{k+1} \mid \Gamma_k) = \mu_k(\Gamma_{k+1})$
- Thus, to estimate $\mathrm{P}_{\mathbb{X}}(\Gamma_{k+1} \mid \Gamma_k)$, we could use a MC approach, using μ_k as the sampling distribution (provided μ_k is known, or at least, we know how to sample from μ_k)
- If $\mu_k(\Gamma_{k+1})$ is not too small, we could get a good MC estimate of this probability with a moderate effort

Estimation of a probability of failure by subset simulation

- An example:
  - Recall that the number of MC evaluations needed to estimate α with a given standard deviation $\delta\alpha$ is approximately $1/(\delta^2\alpha)$
  - Suppose $\alpha \approx 10^{-4}$
  - Setting $\delta = 0.1$, we need approximately 10^6 evaluations to estimate α by a simple MC approach
  - Now, suppose that the thresholds u_k, $i = 1, \ldots, S-1$, are chosen in such a way that $\mu_k(\Gamma_{k+1}) \approx 0.1$. We need approximately 1000 evaluations to estimate $\mu_k(\Gamma_{k+1})$. Since $10^{-4} = (0.1)^4$, we need, in principle, a total of $4 \times 1000 = 4000$ evaluations to estimate α by subset sampling
- So, the questions that need to be addressed are:
  1. How to sample from μ_k?
  2. How to choose the u_ks so that $\mu_k(\Gamma_{k+1})$ is not too small?

Subset sampling algorithms

- Several versions of the subset sampling algorithm have been proposed
- The most popular version of subset sampling algorithm is that proposed by Au and Beck (2001)
- Here, we shall present the recent version of Cérou et al. (2011)
- We begin with a fixed-threshold algorithm

Fixed-threshold algorithm

Assume given a set of thresholds

$$-\infty = u_0 < u_1 < u_2 \cdots < u_S = u$$

and a transition kernel K which is $\mathrm{P}_{\mathbb{X}}$-invariant (a MH kernel will do, for instance)

1. **Initialization.** Generate an N-sample $X_0^{(j)} \overset{\text{i.i.d}}{\sim} \mu_0 = \mathrm{P}_{\mathbb{X}}$, $1 \leq j \leq n$
2. **For** $k = 0$ **to** $S - 1$**.**
   2.1 Let $I_{k+1} = \{j \colon X_k^{(j)} \in \Gamma_{k+1}\}$. Set $\widehat{\alpha}_{k+1} = \frac{\#I_k}{N}$
   2.2 For $j = 1$ to N,
   - take
   $$Y = \begin{cases} X_k^{(j)} & \text{if } j \in I_{k+1} \\ X_k^{(l)} & \text{with } l \text{ randomly chosen in } I_{k+1} \text{ if } j \notin I_{k+1} \end{cases}$$
   - then, generate
   $$Z \sim K(Y, \cdot)$$
   - Take
   $$X_{k+1}^{(j)} = \begin{cases} Z & \text{if } Z \in \Gamma_{k+1} \\ Y & \text{if not} \end{cases}$$
3. **Set** $\widehat{\alpha} = \prod_{k=1}^{S} \widehat{\alpha}_k$

Fixed-threshold algorithm

- Cérou et al. (2011) show that a simplified version of the fixed-threshold algorithm produced an unbiased estimator $\widehat{\alpha}$ of α^u
- Cérou et al. (2011) also show that the variance of $\widehat{\alpha}$ is minimized if the thresholds are set in such a way that
$$\mu_0(\Gamma_1) = \ldots = \mu_{S-1}(\Gamma_S) = p_0 = \alpha^{1/S}$$
$\rightarrow$ this is a difficult issue in practice
- Instead of determining the thresholds u_k, we can try to prescribe a number N_0 of particles that will be kept at each stage k. Then, at stage k, the threshold u_k is defined implicitly by the $(N - N_0)$th-order statistic of the N-sample $f(X_k^{(j)})$, $j = 1, \ldots, N$.

Adaptive algorithm

Assume given a transition kernel K. Prescribe $N_0 < N$ a fixed number of "succeeding particles".

1. **Initialization.** Generate an N-sample $X_0^{(j)} \overset{\text{i.i.d}}{\sim} \mu_0 = \mathrm{P}_{\mathbb{X}}$, $1 \leq j \leq n$.
   Set
   $$u_1 = [f(X_0^{(j)})]_{(N-N_0)}$$
   where $[f(X_0^{(j)})]_{(N-N_0)}$ stands for the $(N - N_0)$-th order statistic of the N-sample $f(X_0^{(j)})$, $j = 1, \ldots, N$
2. **Set** $k = 1$. **While** $u_k < u$.
   2.1 Starting from an N_0-sample with distribution μ_k, draw an i.i.d. N-sample $X_k^{(j)}$, $1 \leq j \leq N$ with the same distribution μ_k
   2.2 Set
   $$u_{k+1} = [f(X_k^{(j)})]_{(N-N_0)}$$
   2.3 Set $k = k + 1$
3. **Let** N_u be the number of particles such that $f(X_{k-1}^{(j)}) > u$. Set $\widehat{\alpha} = \frac{N_u}{N} \left(\frac{N_0}{N}\right)^{k-1}$

Adaptive algorithm

- Step 2.1 of the adaptive algorithm is obviously the main difficulty of the algorithm $\rightarrow$ use Step 2.2 of the fixed-threshold algorithm
- For this, we need a $\mathrm{P}_{\mathbb{X}}$-invariant kernel K with good mixing properties $\rightarrow$ in practice, K should be a parametrized kernel, whose parameter is tuned to keep the acceptance rate in a reasonable range
- Cérou et al. provides an analysis of the properties of the adaptive algorithm. They show that $\widehat{\alpha}$ has a bias that decreases at rate $1/N$, but the mean square error is actually smaller than that of the fixed-threshold algorithm
- In applications, subset sampling algorithms perform very well $\rightarrow$ more expensive that geometrical methods, but much cheaper than simple MC

Some references

- Au S.-K. and Beck J. (2001), "Estimation of small failure probabilities in high dimensions by subset simulation", in Probabilistic Engineering Mechanics
- Cérou F., Del Moral P., Furon T. and Guyader A. (2011), "Sequential Monte Carlo for Rare Event Estimation", Technical Report

Control-variate sampling

- The idea of control-variate sampling for the estimation of a probability of failure is to make use of a cheap approximation g of f so that the random variable $Z = f(X)$ can be predicted by $W = g(X)$.
- W will be our control variate
- Control variates method are standard variance reduction techniques used in Monte Carlo methods
- Let $\alpha_n(f)$ be the MC estimator of $\alpha(f)$:

$$\alpha_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)>u}\,, \quad X_i \overset{\text{i.i.d.}}{\sim} \mathrm{P}_{\mathbb{X}}$$

- Consider the estimator

$$\tilde{\alpha}_n = \alpha_n(f) + \gamma[\alpha_n(g) - \alpha(g)], \quad \gamma \in \mathbb{R}$$

- We have

$$\mathrm{E}(\tilde{\alpha}_n) = \alpha(f)$$

and

$$\mathrm{var}(\tilde{\alpha}_n) = \mathrm{var}(\alpha_n(f)) + \gamma^2\,\mathrm{var}(\alpha_n(g)) + 2\gamma\,\mathrm{cov}(\alpha_n(f), \alpha_n(g))$$

Control-variate sampling

- For the optimal choice
$$\gamma^\star = -\frac{\mathrm{cov}(\alpha_n(f), \alpha_n(g))}{\mathrm{var}(\alpha_n(g))}$$
we have
$$\mathrm{var}(\tilde{\alpha}_n) = (1-\rho^2)\,\mathrm{var}(\alpha_n(f))$$
where ρ is the correlation coefficient between $\alpha_n(f)$ and $\alpha_n(g)$
- If g is close to f, we expect ρ to be high, so the control variate estimator $\tilde{\alpha}_n$ will have a small variance wrt α_n
- How to choose g?
- An idea is to use the framework of FORM: in the standardized Gaussian space, compute a first-order approximation of f at the design point $x^\star$ ➟ this yields an affine approximation g of f, for which we can compute $\alpha(g)$ exactly, using the formula $\alpha(g) = \Phi(-\beta)$
- How to compute/estimate the optimal $\gamma^\star$?
- An idea is to use the approximation

$$\begin{aligned}
\mathrm{cov}(\alpha_n(f), \alpha_n(g)) &= \mathrm{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{1}_{f(X_i)>u}\mathbb{1}_{g(X_j)>u}\right) - \alpha(f)\alpha(g) \\
&= \frac{1}{n}\mathrm{E}\left(\mathbb{1}_{f(X_i)>u}\mathbb{1}_{g(X_i)>u}\right) + \frac{n-1}{n}\alpha(f)\alpha(g) - \alpha(f)\alpha(g) \\
&\approx \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)>u}\mathbb{1}_{g(X_i)>u} - \frac{1}{n}\alpha_n(f)\alpha_n(g)
\end{aligned}$$

Some references

- Cannamela C., Garnier J., Iooss B. (2008), Controlled stratification for quantile estimation, Annals of Applied Statistics
- Hesterberg T. C. and Nelson B. L. (1998). Control variates for probability and quantile estimation. Management Science