ENSIIE. Simulation methods: MC integration and Important Sampling

Abass SAGNA,
abass.sagna@ensiie.fr

Maître de Conférences à l'ENSIIE,
Laboratoire de Mathématiques et Modélisation d'Evry
Université d'Evry Val-d'Essonne, UMR CNRS 8071

http://www.math-evry.cnrs.fr/members/asagna/

February 21, 2019

Plan

1. Monte Carlo integration
   - The principle
   - Properties of the sample mean
   - Convergence rate - confidence interval

2. Variance reduction techniques
   - A toy example with a discrete r.v.
   - The general case: Important sampling

Plan

1. Monte Carlo integration
   - The principle
   - Properties of the sample mean
   - Convergence rate - confidence interval
2. Variance reduction techniques
   - A toy example with a discrete r.v.
   - The general case: Important sampling

Plan

$\rightsquigarrow$ In usual practical problems in statistics, we are led to compute

$$\mathbb{E}[g(X)], \quad X \text{ r.v. valued in } \mathbb{R}^d, \qquad f : \mathbb{R}^d \to \mathbb{R}. \tag{1}$$

$\rightsquigarrow$ When the distribution of X is known, we sometimes may have an analytical expression for (1). For example,

- if $g : \mathbb{R} \mapsto \mathbb{R}$, with $g(x) = x^n$, $n \in \mathbb{N}$, and if $X \in \mathcal{N}(0;1)$, then
$$\mathbb{E}[g(X)] = \mathbb{E}(X^n) = \frac{(2n)!}{2^n n!}.$$
- If $X = (X_1, X_2)$, $X_i \sim \text{Exp}(\lambda_i)$, $i = 1,2$, with $X_1 \perp\!\!\!\perp X_2$, and if $g : \mathbb{R}^2 \mapsto \mathbb{R}$, with $g(x_1, x_2) = x_1 + x_2$, for any $x = (x_1, x_2) \in \mathbb{R}^2$, then
$$\mathbb{E}[g(X)] = \mathbb{E}(X_1 + X_2) = 1/\lambda_1 + 1/\lambda_2.$$

$\rightsquigarrow$ In general (1) has no analytical solution. We can use numerical integration methods to approximate it, keeping in mind that X has pdf f,

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

$\rightsquigarrow$ The Monte Carlo (MC) method is an alternative to these methods.

$\rightsquigarrow$ The numerical integration approximation methods depend on the dimension d of X and the quality of the approximation deface when d increases.

$\rightsquigarrow$ The MC integration method does not depend on d. This makes it the widely used numerical approximation method in high dimension.

$\rightsquigarrow$ The MC method may be used once we may simulate a sample from X even if the density of X is unknown or does not exist.

$\rightsquigarrow$ In general, the MC can be used in the following situations.

- *The Law of X is known but $\mathbb{E}[g(X)]$ has no analytical solution*. Example: $\mathbb{E}[g(X)]$, $X \sim \mathcal{N}(0;1)$ and $g(x) = \exp(x)$.
- *The Law of X is not explicit but may be simulated*. Example: $\mathbb{E}(X_n)$, where X_n is obtained from the recursion (for $X_0 = 0$ and $(Z_k) \perp\!\!\!\perp X_0$):

$$X_{k+1} = \mu_k X_k + \sigma_k Z_{k+1}, \; k = 0, \ldots, n-1, \quad Z_k \sim \mathcal{N}(0,1).$$

Now, how to approximate $\mathbb{E}[g(X)]$ by MC when a sample $X_1, \ldots, X_N$ (a sequence of iid r.v.) from X of size N is available.

$\triangleright$ *X is a discrete r.v. valued in $\{x_1, \ldots, x_n, \ldots\}, x_i \in \mathbb{R}^d$.* We have

$$\mathbb{E}[g(X)] = \sum_{i=1}^{+\infty} g(x_i)\mathbb{P}(X = x_i).$$

When N is large enough, $\mathbb{P}(X = x_i) \approx f_N(x_i)$, where $\forall i \in \{1, \ldots, N\}$,

$$f_N(x_i) = \mathrm{card}(\{\ell \in \{1, \ldots, N\} : \; X_\ell = x_i\})/N := n_i/N.$$

Then,

$$\begin{aligned}\mathbb{E}[g(X)] \approx M_N(g(X)) \;\; &:= \;\; \sum_{i=1}^{+\infty} g(x_i) f_N(x_i) \\ &= \;\; \frac{1}{N}\sum_{i=1}^{+\infty} n_i g(x_i) \; = \; \frac{1}{N}\sum_{k=1}^{N} g(X_k)\end{aligned}$$

▷ *X is a continuous r.v.*. In this case:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

Let $(X_1(\omega), \ldots, X_N(\omega)) = (x_1, \ldots, x_N)$ and suppose that $x_1 \leq x_2 \leq \ldots \leq x_N$. Set (with $x_{0-} = -\infty, x_{N+} = +\infty$)

$$x_{k-} = \frac{x_i + x_{k-1}}{2}, \quad x_{k+} = \frac{x_k + x_{k+1}}{2}, k = 1, \ldots, N.$$

We have

$$\begin{aligned} \mathbb{E}[g(X)] &= \int_{-\infty}^{+\infty} g(x) \mathbb{P}_X(dx) \\ &= \sum_{k=1}^{N} \int_{x_{k-}}^{x_{k+}} g(x) \mathbb{P}_X(dx) \\ &\approx \sum_{k=1}^{N} g(x_k) \mathbb{P}(X \in]x_{k-}, x_{k+}]). \end{aligned}$$

Since $]x_{k-}, x_{k+}]$ only contains x_k, we have $\mathbb{P}(X \in]x_{k-}, x_{k+}]) \approx f_N(x_k) = n_k/N$. It follows that

$$\mathbb{E}[g(X)] \approx \sum_{k=1}^{N} g(x_k) f_N(x_k) = \frac{1}{N} \sum_{k=1}^{N} g(X_k) = M_N(g(X)).$$

$\rightsquigarrow$ The same arguments apply to r.v. valued in $\mathbb{R}^d$.

$\rightsquigarrow$ So, in general, when X is a r.v. valued in $\mathbb{R}^d$ and $g : \mathbb{R}^d \mapsto \mathbb{R}$, then, $\mathbb{E}[g(X)]$ may be approximated by

$$M_N(g(X)) = \frac{1}{N} \sum_{k=1}^{N} g(X_k),$$

for a sample $X_1, \ldots, X_N$ from X of size N.

Plan

1. Monte Carlo integration
   - The principle
   - Properties of the sample mean
   - Convergence rate - confidence interval

2. Variance reduction techniques
   - A toy example with a discrete r.v.
   - The general case: Important sampling

$\rightsquigarrow$ Owing to the forgoing, for any r.v. X valued in $\mathbb{R}^d$ and any Borel function $g : \mathbb{R}^d \mapsto \mathbb{R}$ we can approximate $\mathbb{E}[g(X)]$ by $M_N(g(X))$.

What are the properties of $M_N(g(X))$?

Proposition. Let $X_1, \ldots, X_N$. We have the following results:

1. $M_N(g(X))$ is an unbiased and consistent estimator of $\mathbb{E}[g(X)]$.
2. If $\text{Var}(g(X))$ exists, the mean square error of $M_N(g(X))$ is

$$\mathbb{E}\left[M_N(g(X)) - \mathbb{E}(g(X))\right]^2 = \text{Var}(M_N(g(X)) = \frac{\text{Var}(g(X))}{N}.$$

Remark. In general, $\text{Var}(g(X))$ is unknown and can be estimated by the (unbiased) sample variance (show that $\mathbb{E}S^2_{N,g(X)} = \text{Var}(g(X))$)

$$S^2_{N,g(X)} = \frac{1}{N-1}\sum_{k=1}^{N}\left(g(X_k) - M_N(g(X))\right)^2.$$

1. For the first statement it is clear that $\mathbb{E}M_N(g(X)) = \mathbb{E}(g(X))$, so that the estimator is unbiased. The consistency follows from the Law of Large Numbers which states that: if $X_1, \ldots, X_N$ is a sequence of iid r.v. then

- the sample mean $\bar{X}_N = (X_1 + \ldots + X_N)/N$ converges in probability towards $\mathbb{E}X$: for any $\varepsilon > 0$, $\lim_{N\to+\infty} \mathbb{P}(|\bar{X}_N - \mathbb{E}X| > \varepsilon) = 0$.
- If in addition $\mathbb{E}|X| < +\infty$, then $\bar{X}_N$ converges almost surely towards $\mathbb{E}X$: $\mathbb{P}(\{\omega \in \Omega,\ \bar{X}_N(\omega) \not\to \mathbb{E}X\}) = 0$.

2. Since the X_k's are independent we have

$$\begin{aligned}
\mathrm{Var}(M_N(g(X)) &= \mathrm{Var}\big[g(X_1) + \ldots + g(X_N)\big]/N^2 \\
&= \mathrm{Var}(g(X_1)) + \ldots + \mathrm{Var}(g(X_N))/N^2 \\
&= \big[N \times \mathrm{Var}(g(X_1))\big]/N^2 \\
&= \mathrm{Var}(g(X))/N.
\end{aligned}$$

Plan

⇝ The variance of $\mathrm{Var}(g(X))$ is useful to deduce the convergence rate of $M_N(g(X))$ toward $\mathbb{E}g(X)$: the sample size N that we need to achieve a given level of accuracy.

⇝ In fact, it follows from Chebechev's inequality that

$$\mathbb{P}\Big(\big|M_N(g(X)) - \mathbb{E}g(X)\big| > \frac{1}{\sqrt{N}}\Big) \leq N\,\mathrm{Var}(M_N(g(X)) = \mathrm{Var}(g(X))$$

⇝ From the Central Limit Theorem, we have:

$$\sqrt{N}\,\frac{M_N(g(X)) - \mathbb{E}g(X)}{\sqrt{\mathrm{Var}(g(X))}} \xrightarrow{d} \mathcal{N}(0,1).$$

Then, for large N (Φ is the cdf of the $\mathcal{N}(0,1)$),

$$\mathbb{P}\left(\big|M_N(g(X)) - \mathbb{E}g(X)\big| \geq c\sqrt{\frac{\mathrm{Var}(g(X))}{N}}\right) \approx 2(1-\Phi(c)) \qquad (1)$$

$\rightsquigarrow$ We can use (1) to give a $(1-\alpha)100\%$ confidence interval for $\mathbb{E}g(X)$. In fact, choosing $c = c_\alpha$ s.t. $2(1-\Phi(c_\alpha)) = \alpha$ we get the CI

$$\left(M_N(g(X)) - c_\alpha\sqrt{\frac{\mathrm{Var}(g(X))}{N}}, M_N(g(X)) + c_\alpha\sqrt{\frac{\mathrm{Var}(g(X))}{N}}\right)$$
$$\approx \quad \left(M_N(g(X)) - c_\alpha\frac{S_{N,g(X)}}{\sqrt{N}}, M_N(g(X)) + c_\alpha\frac{S_{N,g(X)}}{\sqrt{N}}\right).$$

$\rightsquigarrow$ As a consequence, if we want to be $(1-\alpha)100\%$ confident that is $M_N(g(X))$ is within ε of the true value of $\mathbb{E}g(X)$ we may increase the sample size N until

$$c_\alpha S_{N,g(X)}/N < \varepsilon.$$

Plan

Plan

$\rightsquigarrow$ Let X be a r.v. taking values $\{-1, 0, 1\}$ with: $\mathbb{P}(X = -1) = 1/3$, $\mathbb{P}(X = 0) = 1/6$, $\mathbb{P}(X = 1) = 1/2$: $\mathbb{E}(X) = 1/6$ and $\mathrm{Var}(X) = 29/36$.

$\rightsquigarrow$ We can use the sample $\bar{X}_N = (X_1 + \ldots + X_N)/N$, where the X_k's are iid r.v. with the same distribution as X, to estimate $\mathbb{E}(X)$.

$\rightsquigarrow$ Our aim: find another estimator of $\mathbb{E}(X)$ with smaller variance, means,

- we find Y such that

$$\mathbb{E}(X) = \mathbb{E}(Y) \quad \text{and} \quad \mathrm{Var}(Y) < \mathrm{Var}(X)$$

- and use the sample mean $\bar{Y}_N$ to estimate $\mathbb{E}(X)$.

$\rightsquigarrow$ Remark that 0 is closer to $\mathbb{E}(X) = 1/6$ than 1 which, in turn, is closer to $\mathbb{E}(X)$ than -1.

$\rightsquigarrow$ To keep the same expectation and reduce the variance we define a r.v. Y which puts the most weighting on 0 or the lowest weighting on -1.

$\rightsquigarrow$ Two examples of such r.v: Let Y_1 be s.t. $\mathbb{P}(Y_1 = 0) = 1/2$ and find the other weights to put for -1 and 1.

$\rightsquigarrow$ Let $p_{-1} = \mathbb{P}(Y_1 = -1)$, $p_0 = \mathbb{P}(Y_1 = 0)$ and $p_1 = \mathbb{P}(Y_1 = 1)$. We want $\mathbb{E}(Y_1) = -p_{-1} + p_1 = 1/6$.

$\rightsquigarrow$ Since $p_{-1} + p_0 + p_1 = 1$. We get $p_{-1} = 1/6$ and $p_1 = 1/3$. Then $\text{Var}(Y_1) = 17/36 < \text{Var}(X)$.

$\rightsquigarrow$ Choosing Y_2 s.t. $\mathbb{P}(Y_2 = 0) = 4/5$ we get $\mathbb{P}(Y_2 = -1) = 1/60$, $\mathbb{P}(Y_2 = 1) = 11/60$ and $\text{Var}(Y_2) = 8/36$, so that

$$\mathbb{E}(Y_2) = \mathbb{E}(Y_1) = \mathbb{E}(X) \text{ and } \text{Var}(Y_2) < \text{Var}(Y_1) < \text{Var}(X).$$

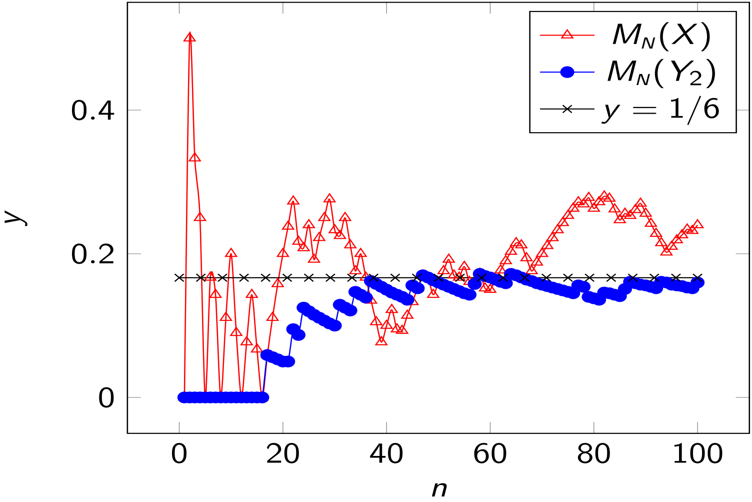

Figure: Abscissa: size N of the sample. Ordinate: $\bar{X}_N = M_N(X)$, $M_N(Y_2)$ and the line $y = 1/6 = \mathbb{E}(X) = \mathbb{E}(Y_2)$.

Plan

1. Monte Carlo integration
   - The principle
   - Properties of the sample mean
   - Convergence rate - confidence interval
2. Variance reduction techniques
   - A toy example with a discrete r.v.
   - The general case: Important sampling

$\rightsquigarrow$ Let X be a r.v. with density f: $X \sim f$ and g a Borel function on $\mathbb{R}^d$.

$\rightsquigarrow$ Let Z be another r.v. with density h s.t. $\forall\, x \in \mathbb{R}^d$, $h(x) = 0$ only if $f(x)g(x) = 0$. Then, setting $\psi(z) = g(z)\frac{f(z)}{h(z)}$,

$$\mathbb{E}g(X) = \int_{\mathbb{R}^d} g(x)f(x)dx = \int_{\mathbb{R}^d} g(x)\frac{f(x)}{h(x)}h(x)dx = \mathbb{E}\left(\psi(Z)\right).$$

$\rightsquigarrow$ $\hat{\theta}_N = (g(X_1) + \ldots + g(X_N))/N$ is the MC estimator of $\mathbb{E}g(X)$,

$\rightsquigarrow$ The estimator $\hat{\theta}_N^{IS} = (\psi(Z_1) + \ldots + \psi(Z_N))/N$ is its IS estimator, where (X_i) and (Z_i) are samples of size N of X and Z, resp.

$\rightsquigarrow$ Recall that

$$\mathbb{E}\left[\hat{\theta}_N - \mathbb{E}(g(X))\right]^2 = \mathrm{Var}(M_N(g(X)) = \frac{\mathrm{Var}(g(X))}{N}.$$

- We have $\mathbb{E}\left[\hat{\theta}_N^{IS} - \mathbb{E}(g(X))\right]^2 = \mathrm{Var}(M_N(\psi(Z)) = \frac{\mathrm{Var}(\psi(Z))}{N}$.
- The estimator $\hat{\theta}_N^{IS}$ is preferable to $\hat{\theta}_N$ if $\mathrm{Var}(\psi(Z)) < \mathrm{Var}(g(Z)$.
- Since $\mathbb{E}\,g(X) = \mathbb{E}\,\psi(Z)$, $\hat{\theta}_N^{IS}$ is preferable to $\hat{\theta}_N$ if $\mathbb{E}\psi^2(Z) < \mathbb{E}g^2(X)$.
- Now,

$$
\begin{aligned}
\mathbb{E}\psi^2(Z) &= \int_{\mathbb{R}^d} g^2(x)\frac{f^2(x)}{h(x)}dx = \int_{\mathbb{R}^d} g^2(x)f(x)\frac{f(x)}{h(x)}dx \\
\text{and} \quad \mathbb{E}g^2(X) &= \int_{\mathbb{R}^d} g^2(x)f(x)dx
\end{aligned}
$$

- Then, $\hat{\theta}_N^{IS}$ is preferable to $\hat{\theta}_N$ if $\frac{f(x)}{h(x)}$ is small where $g^2(x)f(x)$ is large.
- h is chosen to satisfy the previous property.
- h may be a family of density and we can choose the parameter which minimise $\mathbb{E}(\psi^2(Z))$.

⇝ Suppose $h(\mu, z)$ is a family of density depending on $\mu \in A \subset \mathbb{R}^d$ s.t.

$$\mathbb{E}\psi^2(Z^\mu) = \mathbb{E}K(\mu, \xi), \text{where } \xi \text{ is another r.v.}$$

⇝ Our aim is

1. to find μ^* that minimizes $Q(\mu) = \mathbb{E}K(\mu, \xi)$: $Q(\mu^*) = \min_{\mu \in A} Q(\mu)$,
2. to use $\hat{\theta}_N^{IS} = (\psi(Z_1^{\mu^*}) + \ldots + \psi(Z_N^{\mu^*}))/N$ as the IS estimator of $\mathbb{E}g(X)$

⇝ We can use the Robbins-Monro algorithm to approximate μ^*.

⇝ It searches μ^* s.t. (with a formal interchange of the gradient and the expectation) $\nabla Q(\mu^*) = \mathbb{E}\nabla K(\mu^*, \xi) = 0$.

⇝ It states (under some assumptions on $h(\mu, z)$) that the following sequence (μ_n) converges a.s. towards μ^* (where $(\xi_n) \overset{iid}{\sim} \xi$):

$$\mu_{n+1} = \mu_n - \gamma_{n+1}\nabla K(\mu_n, \xi_{n+1})$$

where (γ_n) decreases to 0 and $\sum_{n\geq 1} \gamma_n = +\infty$, i.e., $\gamma_n = 1/n$.

Important Sampling for the Normal distribution. We want to estimate $\mathbb{E}(X\mathbb{1}_{\{X>c\}})$ where $X \sim \mathcal{N}(0,\sigma^2)$ and $c > 3\sigma$.

1. The MC estimator $\hat{\theta}_N$ is poor because very few of the sample values will exceed c.
2. We use IS to have more sample values of Z that exceed c.
3. We suppose $Z \sim \mathcal{N}(\mu,\sigma^2)$ with density

$$h_\mu(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

$\rightsquigarrow$ If f denotes the density of X,

$$\frac{f(x)}{h_\mu(x)} = \exp\Big(\frac{\mu(\mu-2x)}{2\sigma^2}\Big).$$

$\rightsquigarrow$ We want to compute $\mathbb{E}(g(X))$, with $g(x) = x\mathbb{1}_{\{x>c\}}$.

$\rightsquigarrow$ We compare the MC estimator $\hat{\theta}_N$ and the IS estimator

$$\hat{\theta}_N^{IS} = (\psi(Z_1^{\mu^*}) + \ldots + \psi(Z_N^{\mu^*}))/N, \quad \psi(x) = g(x)\frac{f(x)}{h_{\mu^*}(x)},$$

where μ^* is s.t. $Q(\mu^\star) = \min_{\mu \in A} Q(\mu)$, with $A = [0, 6]$ and

$$\begin{aligned} Q(\mu) = \mathbb{E}(\psi^2(Z^\mu)) &= \mathbb{E}(\psi^2(\mu + \sigma\xi)) \\ &= \mathbb{E}\Big(g^2(\mu + \sigma\xi)\exp\Big(\frac{2\mu(\mu - 2(\mu + \sigma\xi))}{2\sigma^2}\Big)\Big) \\ &:= \mathbb{E}K(\mu, \xi), \qquad \xi \sim \mathcal{N}(0,1). \end{aligned}$$

$\rightsquigarrow$ We approximate μ^* from the sequence (μ_n) defined by $(\mu_0 \in A)$

$$\mu_{n+1} = \mu_n - \gamma_{n+1}K'(\mu_n, \xi_{n+1}), \quad \xi_n \overset{iid}{\sim} \mathcal{N}(0,1).$$

$\rightsquigarrow$ Make an application with $N = 10^6$, $c = 3$ and $\sigma = 1$.