UNIVERSITÉ PARIS DAUPHINE

Méthodes de Monte Carlo

Julien STOEHR, stoehr@ceremade.dauphine.fr

Département MIDO

Master 1 Mathématiques

2022–2023

Table des matières

1 Introduction

1.1 Principe de la méthode

Les méthodes de Monte Carlo permettent d'estimer des quantités en utilisant la simulation de variables aléatoires. Les problèmes pouvant être rencontrés comprennent le calcul d'intégrales, les problèmes d'optimisation et la résolution de systèmes linéaires. La simplicité, la flexibilité et l'efficacité pour les problèmes en grande dimension de la méthode en font un outil intéressant, pouvant servir d'alternative ou de référence pour d'autres méthodes numériques.

Supposons que l'on souhaite connaître la valeur d'une certaine quantité δ. La première étape de la méthode consiste à écrire le problème sous la forme d'une espérance. Soient une variable aléatoire $\mathbf{X} = (X_1, \dots, X_d)$ de loi ν sur $\mathbb{R}^d$ (on abrègera cela par $\mathbf{X} \sim \nu$) et une fonction $h : \mathbb{R}^d \to \mathbb{R}$. Le problème traité par les méthodes de Monte Carlo est l'estimation de

$$\delta = \mathbb{E}_\nu[h(\mathbf{X})] = \int_{\mathbb{R}^d} h(\mathbf{x})\nu(\mathrm{d}\mathbf{x}). \tag{1.1}$$

La solution standard à ce problème est de simuler une suite $(\mathbf{X}_n)_{n\geq 1} = (X_{1,n}, \dots, X_{d,n})_{n\geq 1}$ de variables aléatoires indépendantes identiquement distribuées (*i.i.d.*) suivant la loi ν, puis d'estimer l'espérance $\mathbb{E}_\nu[h(\mathbf{X})]$ par la moyenne empirique, *i.e.*,

$$\overline{h}_n = \frac{1}{n}\sum_{k=1}^{n} h(\mathbf{X}_k) \tag{1.2}$$

Remarque. Lorsqu'il n'y a pas d'ambiguïté sur la loi de $\mathbf{X}$, on omettra ν dans les notations.

Exemple 1.1. Calcul d'une intégrale

Soit $h : [a,b]^d \to \mathbb{R}$. On cherche à calculer

$$\delta = \int_{[a,b]^d} h(x_1, \dots, x_d)\mathrm{d}x_1 \dots \mathrm{d}x_d.$$

On peut réécrire $\mathscr{I}$ sous la forme

$$\delta = (b-a)^d \int_{\mathbb{R}^d} h(x_1, \dots, x_d)\frac{1}{(b-a)^d}\mathrm{d}x_1 \dots \mathrm{d}x_d.$$

Si l'on pose $\mathbf{X} = (X_1, \dots, X_d)$ un d-uplet de variables *i.i.d.* suivant la uniforme sur $[a,b]$, on a alors

$$\delta = (b-a)^d \mathbb{E}[h(\mathbf{X})].$$

De façon générale, si $\delta = \int_{R^d} h(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$, avec f un densité de probabilité sur $\mathbb{R}^d$ et h une fonction borélienne, alors on peut écrire, sous les hypothèses d'existence de δ, $\delta = \mathbb{E}\big[g(\mathbf{X})\big]$ et l'estimer.

1.2 Validité et comportement de la méthode

1.2.1 Convergence de la méthode

La convergence de la méthode est assurée par la loi des grands nombres, sous l'hypothèse que h est intégrable par rapport à la mesure ν, *i.e.*, $\mathbb{E}_\nu[|h(\mathbf{X})|]$ existe.

Théorème 1.1. Loi faible et loi forte des grands nombres

Soit $(\mathbf{X}_n)_{n\geq 1}$ une suite de variables aléatoires réelles *i.i.d.* de loi ν telle que $h(\mathbf{X}_1)$ soit ν-intégrable. Alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left[\left|\overline{h}_n - \mathbb{E}[h(\mathbf{X}_1)]\right| \leq \varepsilon\right] \underset{n\to+\infty}{\longrightarrow} 1 \quad \text{(loi faible)}, \quad \mathbb{P}\left[\left|\overline{h}_n - \mathbb{E}[h(\mathbf{X}_1)]\right| \underset{n\to+\infty}{\longrightarrow} 0\right] = 1 \quad \text{(loi forte)}.$$

La loi des grands nombres renseigne donc de deux façons sur l'erreur que l'on commet en estimant δ par la moyenne empirique $\overline{h}_n$. Selon la version faible, il y a une probabilité nulle de commettre une erreur plus grande que ε, tandis que la version forte dit qu'une fois que l'erreur d'estimation est inférieure à ε, alors cette erreur ne peut que décroître.

La loi des grands nombres n'apporte en revanche pas d'informations pratiques sur l'erreur commise. Ainsi, elle ne permet pas de choisir une taille d'échantillon n qui assure que la méthode produira une erreur aussi petite que l'on souhaite, ni de dire si pour un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de taille n fixée, l'erreur sera arbitrairement petite ou non.

1.2.2 Vitesse de convergence et estimation de l'erreur Monte Carlo

Il est possible de préciser davantage le comportement de la méthode lorsque h est de carré intégrable par rapport à la mesure ν, *i.e.*, $\mathbb{E}_\nu\big[h(\mathbf{X})^2\big] < \infty$. $\overline{h}_n$ étant un estimateur sans biais de $\delta = \mathbb{E}_\nu[h(\mathbf{X})]$, *i.e.* $\mathbb{E}_\nu\left[\overline{h}_n\right] = \delta$, son erreur quadratique moyenne est

$$\mathbb{E}\left[\left\{\overline{h}_n - \mathbb{E}[h(\mathbf{X})]\right\}^2\right] = \frac{\sigma^2}{n} =, \quad \text{où} \quad \sigma^2 = \mathbb{V}\mathrm{ar}[h(\mathbf{X})]. \tag{1.3}$$

Autrement dit, la vitesse de convergence de la méthode de Monte Carlo classique est en $\mathcal{O}(n^{-1/2})$. On peut d'ores et déjà remarquer que cette vitesse de convergence ne permet pas d'estimer δ avec une relativement grande précision, chaque chiffre significatif supplémentaire, nécessitant un coût de simulation 100 fois supérieur.

Lorsque la variance de $h(\mathbf{X})$ est finie, le théorème Central-Limite permet d'établir que $\overline{h}_n - \mathbb{E}_\nu[h(\mathbf{X})]$ suit asymptotiquement une loi normale.

Théorème 1.2. Théorème central limite

Soit $(\mathbf{X}_n)_{n\geq 1}$ une suite de variables aléatoires réelles *i.i.d.* de loi ν telle que $h(\mathbf{X}_1)^2$ soit ν-

intégrable. Alors

$$\sqrt{n}\Big(\overline{h}_n - \mathbb{E}[h(\mathbf{X}_1)]\Big) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\big(0,\sigma^2\big) \quad \Leftrightarrow \quad \forall z \in \mathbb{R}, \quad \mathbb{P}\left[\sqrt{\frac{n}{\sigma^2}}\Big(\overline{h}_n - \mathbb{E}[h(\mathbf{X}_1)]\Big) \le z\right] \xrightarrow[n\to+\infty]{} \Phi(z),$$

où Φ est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

L'erreur commise par la méthode de Monte Carlo est aléatoire. Elle ne peut être bornée mais elle peut être quantifiée via un intervalle de confiance. Sous les hypothèse du théorème central limite, pour tout q réel,

$$\mathbb{P}\left[\sqrt{\frac{n}{\sigma^2}}\Big|\overline{h}_n - \mathbb{E}[h(\mathbf{X})]\Big| \le q\right] \xrightarrow[n\to+\infty]{} \int_{-q}^{q} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t = 2\Phi(q) - 1.$$

On en déduit l'intervalle de confiance bilatérale symétrique au niveau de confiance asymptotique $1-\alpha$

$$\mathrm{IC}_{1-\alpha} = \left[\overline{h}_n - q_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{h}_n + q_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] \quad \text{avec} \quad \Phi(q_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Remarque. Le niveau de confiance usuel est $1-\alpha = 0.95$. Alors $q_{1-\alpha/2} = \Phi^{-1}(0.975) \approx 1.96$.

L'erreur quadratique moyenne et l'intervalle de confiance asymptotique dépendent tous deux de la variance σ^2, inconnue en pratique. Lorsque la variance est finie, il est possible d'utiliser l'échantillon $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ pour obtenir un estimateur de σ^2, noté $\widehat{\sigma}^2_n$ dans la suite (*c.f.*, Exemple 1.2).

Exemple 1.2. Estimateurs de la variance

1. $\widehat{\sigma}^2_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2 = \frac{1}{n}\sum_{k=1}^{n} Y_k^2 - (\overline{Y}_n)^2.$
2. $\widehat{\sigma}^2_{n-1} = \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2 = \frac{1}{n-1}\sum_{k=1}^{n} Y_k^2 - \frac{n}{n-1}(\overline{Y}_n)^2.$
3. $\widehat{\sigma}^2_{2n} = \dfrac{1}{2n}\displaystyle\sum_{k=1}^{n}(Y_{2k-1} - Y_{2k})^2.$

En substituant σ^2 par $\widehat{\sigma}^2_n$, on obtient une approximation de l'erreur quadratique moyenne mais également de l'intervalle de confiance. En effet, on rappelle le théorème de Slutsky.

Théorème 1.3. Théorème de Slutsky

Soient $(Y_n)_{n\ge1}$ et $(Z_n)_{n\ge1}$ deux suites de variables aléatoires. S'il existe une variable aléatoire Y telle que $(Y_n)_{n\ge1}$ converge en loi vers Y, et une constante c telle que $(Z_n)_{n\ge1}$ converge en probabilité vers c, alors $(Y_n, Z_n)_{n\ge1}$ converge en loi vers (Y,c). En particulier, $Z_n Y_n \xrightarrow[n\to+\infty]{\mathscr{L}} cY$.

On en déduit que la convergence en loi du théorème 1.2 est préservée lorsque la variance est remplacée par un estimateur asymptotique sans biais.

Lemme 1.1

Sous les hypothèses du théorème 1.2, on a

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\left(\overline{h}_n - \mathbb{E}[h(\mathbf{X}_1)]\right) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1),$$

et l'intervalle de confiance bilatéral symétrique au niveau de confiance asymptotique $1-\alpha$ est

$$\left[\overline{h}_n - q_{1-1/\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}, \overline{h}_n + q_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right].$$

1.2.3 Monte Carlo *v.s.* Méthodes déterministes

La vitesse de convergence de la méthode de Monte Carlo est lente comparée à d'autres méthodes numériques.

Exemple 1.3. Intégration numérique

Soit $h : [0,1] \mapsto \mathbb{R}$. Supposons que l'on souhaite calculer

$$\delta = \int_0^1 h(x)\mathrm{d}x.$$

Par exemple, si la fonction h est de classe $\mathscr{C}^2$ sur $[0,1]$, on obtient par la méthode des trapèzes une erreur en $\mathscr{O}(n^{-2})$. Autrement dit, pour que l'on obtienne par la méthode de Monte calcul la même précision que la méthode des trapèzes avec n points, il faudra de l'ordre de n^4 tirages. Pour des fonctions davantage régulières, la différence est encore plus marquée. Si la fonction est de classe $\mathscr{C}^4$ sur $[0,1]$, la méthode de Simpson fournit une erreur en $\mathscr{O}(n^{-4})$, ce qui équivaudrait à n^8 tirages pour la méthode de Monte Carlo!

Contrairement à ses compétiteurs, la méthode de Monte Carlo a néanmoins l'avantage de ne pas dépendre de la régularité de la fonction h et peut donc s'adapter sans problèmes à des problèmes non réguliers. Une autre caractéristique intéressante des méthodes de Monte Carlo est que l'erreur quadratique moyenne ne dépend pas de la dimension d de l'espace d'états de $\mathbf{X}$. Ceci n'est pas le cas pour les méthodes numériques classiques.

Exemple 1.4. Comparaison avec des méthodes déterministes

Reprenons l'exemple 1.1.

1. Si la fonction est de classe $\mathscr{C}^4$ sur $[0,1]^d$, en utilisant la méthode de Simpson, on obtient une erreur en $\mathscr{O}(n^{-4/d})$. La méthode devient donc inefficace dès lors que d devient trop grand.
2. On peut également approcher δ à l'aide sommes de Riemann

$$\delta = \lim_{n\to+\infty} \frac{1}{n^d}\sum_{k_1=1}^{n}\cdots\sum_{k_d=1}^{n} h\left(\frac{k_1}{n},\ldots,\frac{k_d}{n}\right).$$

Supposons que h soit ℓ-lipschitzienne. Alors,

$$\left|\delta - \frac{1}{n^d}\sum_{k_1=1}^{n}\cdots\sum_{k_d=1}^{n} h\left(\frac{k_1}{n},\ldots,\frac{k_d}{n}\right)\right| \leq \ell\frac{\sqrt{d}}{n}.$$

Le coût de la méthode est $T = n^d$ pour une précision en $T^{-1/d}$, alors que la méthode de

Monte Carlo (sous les hypothèses du Théorème central limite) aura une précision de $T^{-1/2}$. Les méthodes de Monte Carlo sont donc avantageuse à partir de $d = 3$.

En résumé, si les méthodes déterministes sont efficaces pour les problèmes réguliers de très petites dimensions, les méthodes de Monte Carlo les surpassent et sont très compétitives pour les problèmes non-réguliers en grande dimension.

1.3 Conclusion

Les méthodes de Monte Carlo nécessitent uniquement de savoir formuler le problème sous la forme d'une espérance par rapport à une loi ν. Le choix de ν est guidée par le fait que l'on sache simuler suivant cette mesure (*c.f.*, Chapitre 2) mais également par le fait que l'on puisse contrôler la variance de la méthode. En effet, l'équation (1.3) montre que les deux seuls facteurs influant sur les performances de la méthode son n et σ^2 (ou $\widehat{\sigma}_n^2$). On peut donc réduire l'erreur en augmentant n au prix d'un coup de simulation (temps de calcul) plus important ou en réduisant σ^2. Par exemple, si l'on reformule le problème par rapport à une mesure η telle que

$$\delta = \mathbb{E}_\nu[h(\mathbf{X})] = \mathbb{E}_\eta\big[g(\mathbf{Y})\big] \quad \text{et} \quad \mathbb{V}\text{ar}_\eta\big[g(\mathbf{Y})\big] = \frac{1}{2}\mathbb{V}\text{ar}_\nu[h(\mathbf{X})], \quad \text{avec } g \text{ une fonction mesurable de } \mathbb{R}^d,$$

l'estimateur basée sur η présentera la même précision que l'estimateur basée sur ν en utilisant deux fois moins de tirages (*c.f.*, Chapitre 3).

Ce chapitre apporte une réponse (partielle) à comment simule-t-on suivant une loi ν ? Il présentera notamment les principes généraux de la méthode d'inversion et de la méthode du rejet, ainsi que quelques cas particuliers utiles.

2.1 Simulation suivant la loi uniforme

Initialement, le générateur de nombres aléatoires suivant loi uniforme sur $[0,1]$ est le seul dont on dispose. Disposer d'un tel générateur n'est pas trivial : un ordinateur ne disposant d'aucun composant aléatoire. Un générateur de nombres aléatoires est donc un programme déterministe qui produit une suite de valeurs "suffisamment" désordonnées pour ressembler à un échantillon aléatoire (*c.f.*, [1] pour la définition suivante).

Définition 2.1. Générateur pseudo-aléatoire

Un générateur de nombres pseudo-aléatoire est un algorithme qui, à partir d'une valeur initiale u_0, appelée graine (*seed*), et une transformation D, produit une suite $(u_n)_{n\geq 1} = \{D^n(u_o)\}$ dans $[0,1]$. Pour tout n, les valeurs $u_1,\dots,u_n$ reproduisent le comportement d'un échantillon *i.i.d.* de loi uniforme vis à vis des tests usuels.

La loi uniforme joue un rôle central aussi bien dans la simulations de lois usuelles que de lois plus complexes. Si des générateurs sont généralement disponibles pour les lois usuelles, il est important de comprendre le rôle de ces processus de génération qui permettent dans certains cas de réduire la variance des méthodes de Monte Carlo.

2.2 Méthode de la fonction inverse

Cette méthode permet, étant donné un générateur aléatoire uniforme, de simuler suivant la loi d'une variable aléatoire réelle lorsque l'on est capable de calculer l'inverse généralisé de sa fonction de répartition.

Définition 2.2. Inverse généralisé

Soit X une variable aléatoire réelle de fonction de répartition F. On appelle inverse généralisé (ou fonction quantile) de F, noté $F^{\leftarrow}$, la fonction définie pour tout $u \in [0,1]$ par

$$F^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}.$$

Remarque. L'inverse généralisé ne correspond à l'inverse (au sens bijection) que lorsque F est continue et strictement croissante.

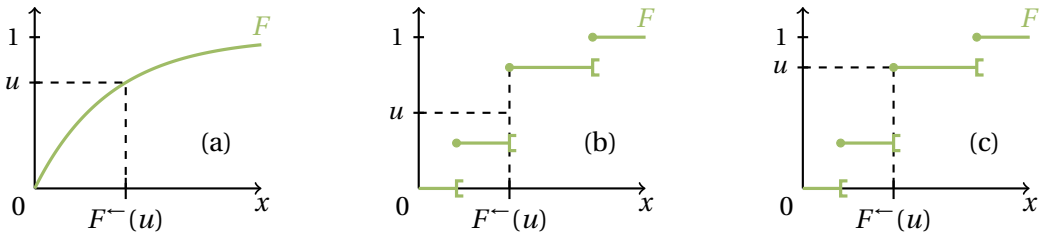

FIGURE 2.1 – Lecture graphique de $F^{\leftarrow}(u)$, *i.e.*, quantile de la loi d'ordre u. (a) L'équation $F(x) = u$ a une unique solution. (b) L'équation $F(x) = u$ n'a pas de solutions. (c) L'équation $F(x) = u$ a une infinité de solutions.

Par construction, l'inverse généralisé est une fonction croissante : pour $0 \le u \le v \le 1$, F étant croissante, on a

$$\{x \in \mathbb{R} : F(x) \ge v \ge u\} \subseteq \{x \in \mathbb{R} : F(x) \ge u\} \Rightarrow F^{\leftarrow}(u) \le F^{\leftarrow}(v).$$

L'inverse généralisé est donc en particulier mesurable et sa composition avec une variable aléatoire aléatoire demeure donc une variable aléatoire.

Lemme 2.1. Méthode de la fonction inverse ou méthode d'inversion

Soit X une variable aléatoire réelle de fonction de répartition F et $U \sim \mathcal{U}([0,1])$. Alors $F^{\leftarrow}(U)$ est une variable aléatoire suivant la loi de X.

Exemple 2.1. Loi discrète

Soit X une variable aléatoire discrète de support $\Omega = \{x_k \in \mathbb{R},\ k \in \mathbb{N}^*\}$. Notons, pour $k \ge 1$,

$$p_k = \mathbb{P}[X = x_k], \quad s_0 = 0 \quad \text{et} \quad s_k = \sum_{i=1}^{k} p_k.$$

Pour tout $u \in [0,1]$, l'inverse généralisé est donné par

$$F^{\leftarrow}(u) = \inf\left\{x \in \mathbb{R} : \sum_{k=1}^{+\infty} p_k \mathbb{1}_{\{x_k \le x\}} \ge u\right\} = \{x_k : s_{k-1} < u \le s_k\}.$$

Alors pour $U \sim \mathcal{U}([0,1])$, la variable aléatoire Z définie ci-dessous suit la même loi que X

$$Z = \begin{cases} x_1 & \text{si } u \in [0, s_1], \\ x_k & \text{si } u \in]s_{k-1}, s_k],\ k \ge 2. \end{cases}$$

En pratique, il suffit donc de trouver, pour une réalisation u suivant la loi $\mathcal{U}([0,1])$, l'unique indice k tel que $s_{k-1} < u \le s_k$. Ceci est facile à mettre en œuvre lorsque Ω est fini mais peut être plus complexe lorsque Ω est infini dénombrable (*e.g.*, loi de Poisson).

Exemple 2.2. Loi exponentielle

Soit $X \sim \mathcal{E}(\lambda)$ une variable aléatoire de loi exponentielle de paramètre $\lambda > 0$.

Rappel. La loi exponentielle a pour densité $f(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x \ge 0\}}$ et pour fonction de répartition $F(x) = \left(1 - e^{-\lambda x}\right) \mathbb{1}_{\{x \ge 0\}}$.

F est bijective (la bijection réciproque et l'inverse généralise coïncident) et pour tout $u \in [0,1]$,

$$F^{\leftarrow}(u) = \frac{-1}{\lambda}\ln(1-u).$$

Ainsi pour $U \sim \mathscr{U}([0,1])$, $F^{\leftarrow}(U) = -\lambda^{-1}\ln(1-U) \sim -\lambda^{-1}\ln(U) \sim \mathscr{E}(\lambda)$, car $1-U$ et U ont même loi sur $[0,1]$.

Exemple 2.3. Loi de Weibull

La loi de Weibull est utilisée dans différents domaines (ingénierie, théorie des extrêmes, assurances, hydrologie,...). On dit qu'une variable aléatoire X suit une loi de Weibull de paramètre $\lambda, k \in \mathbb{R}_+^*$ lorsque sa fonction de répartition est donnée, pour tout réel $x \geq 0$, par

$$F(x) = 1 - \exp\left\{-(x/\lambda)^k\right\}.$$

Comme pour la loi exponentielle, la fonction de répartition est bijective et

$$F^{\leftarrow}(u) = \lambda\{-\ln(1-u)\}^{1/k}.$$

Ainsi pour $U \sim \mathscr{U}([0,1])$, $\lambda\{-\ln(U)\}^{1/k}$ suit une loi de Weibull de paramètre $\lambda, k \in \mathbb{R}_+^*$.

La méthode de la fonction inverse est la méthode standard de simulation de variables aléatoires, mais elle est également utile pour des méthodes de réduction de variance comme la variable antithétique ou la stratification (*c.f.*, Chapitre 3). Bien que facile à mettre en œuvre, elle n'est néanmoins exacte qu'à condition de connaître l'expression explicite de $F^{\leftarrow}$ (exponentielle, double exponentielle, Weibull, Cauchy...), ce qui n'est pas toujours évident. Un contre-exemple classique à cette méthode est la loi normale $\mathscr{N}(0,1)$: il n'existe pas de formulation simple de sa fonction de répartition Φ ou de son inverse Φ^{-1}. Il faut alors utiliser un algorithme d'approximation : il existe des polynômes donnant une approximation de Φ et Φ^{-1}. La méthode de la fonction inverse peut s'appliquer à ces polynômes pour simuler une loi normale moyennant une approximation raisonnable. C'est la méthode utilisée par défaut dans R.

2.3 Autres exemples de transformations

Lorsque $F^{\leftarrow}$ n'a pas d'expression explicite ou lorsque l'on cherche des méthodes de simulation alternatives plus rapide, on peut essayer de trouver des transformations entre la loi ν et des lois dont on dispose déjà de générateurs.

2.3.1 Un premier exemple : somme de variables aléatoires

Simuler une variable aléatoire X de loi binomiale $\mathscr{B}(n,p)$ avec la méthode de la fonction inverse présente l'inconvénient de devoir calculer la fonction de répartition, et donc des coefficients binomiaux et des puissances de p et $1-p$. À la place, on peut utiliser le résultat suivant

Lemme 2.2

La somme de n variables aléatoires *i.i.d.* de Bernoulli de paramètre p suit une loi $\mathscr{B}(n,p)$.

Pour générer des variables de Bernoulli, on peut utiliser la méthode de la fonction inverse (*c.f.*, Exemple 2.1). Soient $U_1,\ldots,U_n$ des variables aléatoires *i.i.d.* suivant la loi $\mathscr{U}([0,1])$. Alors $\mathbb{1}_{\{U_i \leq p\}}$, $i = 1,\ldots,n$,

suit une loi de Bernoulli de paramètre p et

$$\sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq p\}} \sim \mathscr{B}(n,p).$$

Ce type d'approche est également utilisé pour simuler la loi de Poisson de paramètre λ à partir de la somme de variables aléatoires *i.i.d.* suivant la loi exponentielle de même paramètre λ.

$$T = \sup\left\{n \in \mathbb{N}^* : X_1 + \ldots + X_n \leq 1\right\}, \quad X_i \sim \mathscr{E}(\lambda),\ i = 1,\ldots,n.$$

2.3.2 Algorithmes de Box-Muller

L'algorithme de Box-Muller est l'un des exemples les plus connus de transformation. Il permet de générer des variables aléatoires gaussiennes $\mathscr{N}(0,1)$ à partir de variables aléatoires uniformes sans avoir recours à l'évaluation et à l'inversion de sa fonction de répartition Φ.

Proposition 2.1. Algorithmes de Box-Muller

Version n°1. Soient U_1 et U_2 des variables *i.i.d.* de loi $\mathscr{U}([0,1])$. Alors les variables X_1 et X_2 définies ci-dessous sont *i.i.d.* de loi $\mathscr{N}(0,1)$:

$$X_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \quad \text{et} \quad X_2 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2).$$

Version n°2. Soit (U_1,U_2) une variable aléatoire de loi uniforme sur le disque unité $\mathscr{D}_1 = \{(u_1,u_2) \in \mathbb{R}^2 : u_1^2 + u_2^2 \leq 1\}$. Alors les variables X_1 et X_2 définies ci-dessous sont *i.i.d.* de loi $\mathscr{N}(0,1)$:

$$X_1 = U_1\sqrt{\frac{-2\ln\left(U_1^2+U_2^2\right)}{U_1^2+U_2^2}} \quad \text{et} \quad X_2 = U_2\sqrt{\frac{-2\ln\left(U_1^2+U_2^2\right)}{U_1^2+U_2^2}}.$$

2.3.3 Simulation d'un vecteur gaussien

Un générateur de la loi $\mathscr{N}(0,1)$ permet également de simuler un vecteur aléatoire $\mathbf{X} = (X_1,\ldots,X_d)$ de $\mathbb{R}^d$ de loi normale multivariée $\mathscr{N}(\boldsymbol{\mu},\Sigma)$.

Matrice de covariance Σ diagonale. Il suffit de générer d variables X_i, $i \in [\![1,d]\!]$, indépendantes de lois normale respectives $\mathscr{N}(\mu_i,\Sigma_{ii})$. Pour ce faire, il suffit de simuler $Z_1,\ldots,Z_d$ variables aléatoires *i.i.d.* suivant la loi $\mathscr{N}(0,1)$ et d'utiliser la transformation $X_i = \mu_i + \sqrt{\Sigma_{ii}}Z_i$.

Matrice de covariance Σ non diagonale. La méthode précédente n'est pas envisageable. On utilise alors la décomposition de Cholesky de Σ.

Proposition 2.2

Soit $\mathbf{Z}$ un vecteur aléatoire de loi $\mathscr{N}(\mathbf{0}_d,\mathbf{I}_d)$. Étant donnée la décomposition de Cholesky de Σ, *i.e.*, $\Sigma = LL^T$ avec L une matrice triangulaire inférieure, la variable définie par

$$\mathbf{X} = \mu + L\mathbf{Z}, \quad \boldsymbol{\mu} \in \mathbb{R}^d,$$

suit une loi normale multivariée $\mathscr{N}(\boldsymbol{\mu},\Sigma)$.

Remarques. **1.** $\mathbf{Z}$ peut être simulé en utilisant l'algorithme de Box-Muller (les composantes sont

i.i.d. suivant $\mathcal{N}(0,1)$).

2. La fonction `chol` de R permet d'obtenir la décomposition de Cholesky d'une matrice symétrique définie-positive.

3. La décomposition de Cholesky s'applique uniquement si la matrice est symétrique définie positive. Si la loi est dégénérée, la matrice de covariance est simplement semi-définie positive. La matrice L est alors obtenue par décomposition spectrale, *i.e.*, on diagonalise Σ dans une base orthogonale : $\Sigma = QDQ^T$, puis on pose $L = OD^{1/2}$.

Exemple 2.4. En dimension 2

Soit $\mathbf{Z} = (Z_1, Z_2) \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$, $\rho \in [-1,1]$ une constante. En posant,

$$X_1 = \mu_1 + \sigma_1 Z_1 \quad \text{et} \quad X_2 = \mu_2 + \sigma_2 \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2 \right),$$

on a

$$\mathbf{X} = (X_1, X_2) \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right\}. \tag{2.1}$$

2.3.4 Simulation du mouvement Brownien

La simulation du mouvement brownien est un autre exemple classique basé sur la génération de variables aléatoires gaussiennes. On ne s'intéresse pas ici à la construction du mouvement brownien mais simplement à sa simulation.

Définition 2.3

Un processus stochastique $\mathbf{W} = (W_t)_{t \in \mathbb{R}_+}$ est un mouvement brownien standard si, et seulement si,

(1) $\mathbf{W}$ est issu de 0, *i.e.*, $W_0 = 0$ *p.s.*,

(2) $\mathbf{W}$ est à trajectoires continues,

(3) $\mathbf{W}$ est à accroissements indépendants, *i.e.*, pour toutes séquences $0 = t_0 < t_1 < \ldots < t_n$, les variables $\left(W_{t_2} - W_{t_1}, \ldots, W_{t_n} - W_{t_{n-1}}\right)$ sont indépendantes,

(4) $\mathbf{W}$ est à accroissements stationnaires, *i.e.*, pour tous $s < t$ la valeur de $W_t - W_s$ ne dépend que de la valeur de $t - s$ et $W_t - W_s \sim \mathcal{N}(0, t-s)$.

La propriété centrale dans la simulation du mouvement brownien est celle des accroissements indépendants et stationnaires.

Proposition 2.3. *Forward simulation*

Soit $(t_n)_{n\in\mathbb{N}}$ une suite de $\mathbb{R}_+$ et $(Z_n)_{n\in\mathbb{N}}$ une suite de variables *i.i.d.* de loi $\mathcal{N}(0,1)$. On définit le processus $\mathbf{W} = (W_t)_{t\in\mathbb{R}_+}$ par

$$W_0 = 0 \quad \text{et} \quad W_{t_{n+1}} = W_{t_n} + \sqrt{t_{n+1} - t_n} Z_n.$$

Alors $\mathbf{W}$ est une réalisation de trajectoires du mouvement brownien aux instants $(t_n)_{n\in\mathbb{N}}$.

Exemple 2.5. Modèle de Black-Scholes

En finance, un exemple classique d'utilisation du mouvement brownien $(W_t)_{t\in\mathbb{R}_+}$ est le calcul de la valeur d'une option avec le modèle de Black-Scholes, *i.e.*, le processus $(S_t)_{t\in\mathbb{R}_+}$ défini par

$$S_t = S_0 \exp\left(\frac{r-\sigma^2}{2}t + \sigma W_t\right), \quad \text{avec} \quad \begin{cases} r & \text{le taux d'intérêt sans risque,} \\ \sigma & \text{la volatilité du prix de l'option.} \end{cases}$$

2.4 Méthodes d'acceptation-rejet

La méthode d'acceptation rejet est une méthode générale de simulation de variables aléatoires qui ne nécessite de connaître la densité cible f qu'à une constante près. Le principe est le suivant : on génère des réalisations suivant une densité alternative g, appelée densité instrumentale, qui domine f et pour laquelle on sait facilement simuler (*e.g.*, lois uniforme, exponentielle, normale...) puis on introduit un biais (à l'aide d'une procédure acceptant certaines réalisations de g et rejetant les autres) qui permet d'obtenir un échantillon suivant f.

Proposition 2.4. Méthode d'acceptation-rejet

Soit $\mathbf{X} = (X_1, \ldots, X_d)$ une variable aléatoire de densité f de $\mathbb{R}^d$ et g une densité de $\mathbb{R}^d$ telle qu'il existe une constante $M \geq 1$ satisfaisant

$$f(\mathbf{x}) \leq Mg(\mathbf{x}), \quad \text{pour tout } \mathbf{x} \text{ dans } \mathbb{R}^d.$$

Soient $(U_n)_{n\geq 1}$ une suite de variables *i.i.d.* de loi $\mathscr{U}([0,1])$ et $(\mathbf{Y}_n)_{n\geq 1} = (Y_{1,n}, \ldots, Y_{d,n})_{n\geq 1}$ une suite de variables *i.i.d.* de loi de densité g telles que ces deux suites soient indépendantes. Alors, pour T défini par

$$T := \inf\{n \geq 1 : U_n \leq \alpha(\mathbf{Y}_n)\}, \quad \text{où} \quad \alpha(\mathbf{Y}_n) := \frac{f(\mathbf{Y}_n)}{Mg(\mathbf{Y}_n)},$$

$\mathbf{Y}_T$ suit la loi de densité f. Autrement dit pour simuler $\mathbf{X} \sim f$, il suffit de simuler

$$\mathbf{Y} \sim g \quad \text{et} \quad U \sim \mathscr{U}([0,1]) \quad \text{jusqu'à ce que} \quad uMg(\mathbf{y}) < f(\mathbf{y}).$$

Remarques.

1. $\alpha(\mathbf{Y}_n)$ est bien défini pour presque tout ω, car $\mathbb{P}\left[\mathbf{Y}_n \in \{\mathbf{x} : g(\mathbf{x}) = 0\}\right] = 0$.
2. T est presque sûrement fini et l'algorithme ne peut donc pas tourner indéfiniment. En effet, du fait de l'indépendance, pour tout $n \in \mathbb{N}^*$,
$$\mathbb{P}[T = +\infty] \leq \left(1 - \frac{1}{M}\right)^n \underset{n\to+\infty}{\longrightarrow} 0.$$
3. f et g étant des densités, on a nécessairement $M \geq 1$. En effet,
$$f(\mathbf{x}) \leq Mg(\mathbf{x}) \quad \Rightarrow \quad \int_{\mathbb{R}^d} f(\mathbf{x})\mathrm{d}\mathbf{x} \leq M \int_{\mathbb{R}^d} g(\mathbf{x})\mathrm{d}\mathbf{x} \quad \Rightarrow \quad 1 \leq M.$$
4. Ce résultat est valable pour une densité f par rapport à une mesure ν quelconque. Pour une variable aléatoire discrète sur une ensemble fini ou dénombrable, ν est la mesure de comptage.

Une conséquence de la proposition 2.4 est que la probabilité d'acceptation est exactement $1/M$. Au-

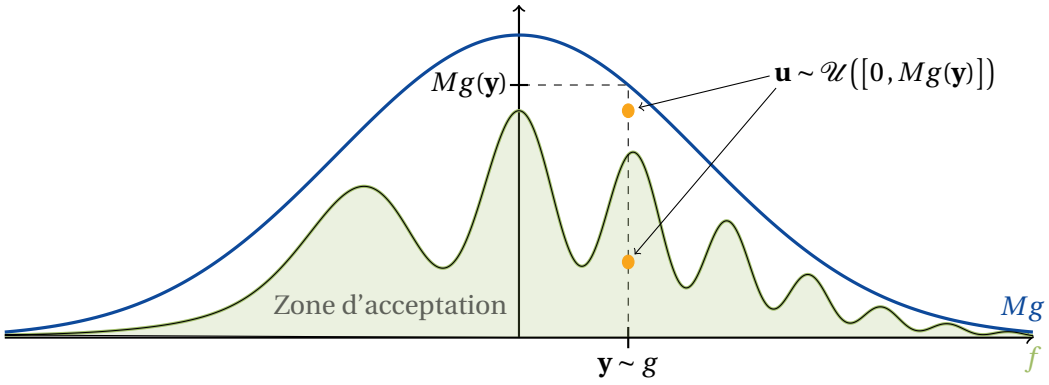

FIGURE 2.2 – Illustration de l'algorithme d'acceptation-rejet

trement dit, T suit une loi géométrique de paramètre $1/M$ et le nombre d'essais moyen jusqu'à ce qu'une variable soit acceptée est M. Ce résultat permet de choisir entre différentes densités instrumentales $g_1, \ldots, g_k$ au travers de leurs bornes respectives $M_1, \ldots, M_k$. Ainsi une façon d'optimiser la méthode est de trouver la plus petite borne M possible, bien que cette technique ait certaines limitations.

2.5 Conclusion

La plupart des langages comme R ou Python proposent des générateurs aléatoires pour les lois de probabilité usuelles. Certains des résultats vu dans ce chapitre peuvent donc paraître anecdotiques. Ils permettent néanmoins de réfléchir aux processus de génération utilisés et à leur utilisation dans des contextes où des générateurs ne sont pas disponibles. Par ailleurs ce chapitre ne fait pas un état de l'art exhaustif des méthodes existantes. Certaines ont été omises comme le rapport d'uniformes et les méthodes de Monte Carlo par chaînes de Markov (MCMC) qui nécessitent des notions du cours de processus discrets.

Dans ce chapitre, on se concentre sur le problème d'estimation de $\delta = \mathbb{E}_\nu[h(\mathbf{X})]$ avec h une fonction mesurable telle que h est de carré intégrable par rapport à la mesure ν (h est donc intégrable par rapport à la mesure ν et δ est bien définie).

On a vu au Chapitre 1, que la méthode de Monte Carlo classique présente une erreur de l'ordre de $\sigma/\sqrt{n}$. On peut diminuer l'erreur en augmentant n, au prix d'un coût de calcul plus important, ou en diminuant σ^2. Ce chapitre décrit les méthodes dites de réduction de variance, ayant pour but de construire un estimateur sans biais $\widehat{\delta}_n$ de variance inférieure à la méthode de Monte Carlo classique, *i.e.*, un estimateur tel que

$$\mathbb{E}\left[\widehat{\delta}_n\right] = \delta \quad \text{et} \quad \mathbb{V}\text{ar}\left[\widehat{\delta}_n\right] < \mathbb{V}\text{ar}\left[\overline{h}_n\right].$$

Critères de comparaison

Le point central de ce chapitre sera donc la comparaison en terme de variance entre un estimateur sans biais $\widehat{\delta}_n$ et $\overline{h}_n$. Lorsque l'estimateur s'écrit $\widehat{\delta}_n = n^{-1}\sum_{k=1}^n Y_k$, avec $(Y_n)_{n\geq 1}$ une suite de variables aléatoires réelles *i.i.d.* de carré intégrable, cela revient, de façon équivalente, à vérifier si

$$\mathbb{E}[Y_1] = \delta \quad \text{et} \quad \sigma_1^2 = \mathbb{V}\text{ar}[Y_1] < \sigma^2 = \mathbb{V}\text{ar}[h(\mathbf{X})].$$

Limiter la comparaison des méthodes à la simple comparaison des variances σ^2 et σ_1^2 n'est néanmoins pas nécessairement judicieux si la méthode alternative n'est pas de complexité comparable (*e.g.* coût de calcul et utilisation de la mémoire plus important) à la méthode classique.

Supposons que le coût de la méthode classique pour simuler un n-échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ et évaluer n fois h est Cn. Pour obtenir une erreur quadratique moyenne ε^2, il faut prendre $n = \sigma^2/\varepsilon^2$ et le coût de la méthode est alors $C\sigma^2/\varepsilon^2$. Si le coût de calcul de $\widehat{\delta}_n$ est $C_1 n$, alors pour obtenir la même précision ε^2, le coût de la méthode est $C_1\sigma_1^2/\varepsilon^2$. On en déduit l'efficacité relative de $\widehat{\delta}_n$ par rapport $\overline{h}_n$

$$R(\overline{h}_n, \widehat{\delta}_n) = \frac{C\sigma^2}{C_1\sigma_1^2}.$$

Autrement dit, pour obtenir une erreur quadratique moyenne donnée, le calcul de $\overline{h}_n$ prend $R(\overline{h}_n, \widehat{\delta}_n)$ fois le temps de calcul de $\widehat{\delta}_n$. Ainsi, une méthode de réduction de variance est d'autant plus efficace que $R(\overline{h}_n, \widehat{\delta}_n)$ est grand devant 1.

Il n'est pas toujours facile de quantifier l'effet du rapport des coûts C/C_1, celui-ci pouvant dépendre du langage, de la façon de programmer les méthodes, des algorithmes de simulation utilisés, ... Dans le cours, on se concentrera sur la comparaison des variances. Ce point sera abordé dans les séances de travaux pratiques.

3.1 Échantillonnage préférentiel (*Importance Sampling*)

L'échantillonnage préférentiel est un principe général, particulièrement adapté au cas où la fonction h est presque nulle (ou nulle) en dehors d'un domaine $\mathscr{D}$ qui est tel que $\mathbb{P}[\mathbf{X} \in \mathscr{D}]$ soit proche de 0. Cela se produit par exemple lorsque le volume de $\mathscr{D}$ est petit ou lorsque $\mathscr{D}$ est défini par les queues de distributions de ν. Il est alors difficile d'obtenir des réalisations de $\mathbf{X}$ dans $\mathscr{D}$ et donc d'estimer δ.

La méthode repose sur l'idée que pour calculer δ, au lieu d'échantillonner suivant ν, il est plus intéressant d'échantillonner suivant une loi dont les réalisations seront dans $\mathscr{D}$ avec forte probabilité, puis de pondérer cet échantillon pour corriger le changement de loi d'échantillonnage.

Exemple 3.1. Probabilité d'événement rare

On souhaite calculer $\delta = \mathbb{P}[X > 4.5]$ pour $X \sim \mathcal{N}(0,1)$. Étant donnée $(X_n)_{n\geq 1}$ une suite de $n = 10000$ variables aléatoires *i.i.d.* suivant la loi $\mathcal{N}(0,1)$, on a

$$\overline{h}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > 4.5\}} = 0.$$

La simulation « naïve » s'avère très inefficace car il s'agit d'un événement rare. La solution est de forcer la réalisation de cet événement. Soit $Z \sim \tau\mathscr{E}(4.5,1)$ une variable aléatoire suivant la loi exponentielle de paramètre 1 et translatée de 4.5, *i.e.* de densité donnée pour tout réel z par

$$g(z) = e^{-z+4.5}\mathbb{1}_{\{z\geq 4.5\}}.$$

On peut écrire

$$\mathbb{P}[X > 4.5] = \int_{4.5}^{+\infty} \phi(x)\mathrm{d}x = \int_{4.5}^{+\infty} \frac{\phi(x)}{g(x)} g(x)\mathrm{d}x.$$

Ainsi, pour une suite $(Z_n)_{n\geq 1}$ de $n = 10000$ variables aléatoires *i.i.d.* suivant la loi de Z, on a l'estimation

$$\widehat{\delta}_n = \frac{1}{n}\sum_{k=1}^{n} \frac{\phi(Z_k)}{g(Z_k)} \mathbb{1}_{\{Z_k > 4.5\}} = 3.36\,10^{-6}.$$

Définition 3.1. Estimateur d'échantillonnage préférentiel

Soit g une densité telle que $\mathrm{supp}(hf) = \{\mathbf{x} \in \mathbb{R}^d : h(\mathbf{x})f(\mathbf{x}) > 0\} \subseteq \mathrm{supp}(g) = \{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) > 0\}$. Alors on peut écrire (pour $Z \sim g$)

$$\mathbb{E}_f[h(\mathbf{X})] = \mathbb{E}_g\left[\frac{h(\mathbf{Z})f(\mathbf{Z})}{g(\mathbf{Z})}\right]. \tag{3.1}$$

Étant donnée $(\mathbf{Z}_n)_{n\geq 1}$ une suite de variables aléatoires *i.i.d.* suivant la densité g, on définit l'estimateur d'échantillonnage préférentiel par

$$\widehat{\delta}_n(g) = \frac{1}{n}\sum_{k=1}^{n} \frac{f(\mathbf{Z}_k)}{g(\mathbf{Z}_k)} h(\mathbf{Z}_k) = \frac{1}{n}\sum_{k=1}^{n} w_k h(\mathbf{Z}_k).$$

La densité g est appelée loi instrumentale (ou loi d'importance) et le rapport $w_k = f(\mathbf{Z}_k)/g(\mathbf{Z}_k)$ est appelé poids d'importance.

Remarque. Le coût de calcul de $\widehat{\delta}_n(g)$ peut être différent de celui de l'estimateur classique : le coût

de simulation suivant g pouvant être différent du coût de simulation suivant f. Cela influe donc sur le rapport C/C_1 de la fonction $R(\overline{h}_n, \widehat{\delta}_n(g))$.

Biais de l'estimateur. Sous l'hypothèse $\text{supp}(hf) \subseteq \text{supp}(g)$, on obtient directement que l'estimateur est sans biais en utilisant l'équation (3.1).

Convergence de l'estimateur. Les variables aléatoires $(Y_n)_{n\geq 1} = \left(h(\mathbf{Z}_n) f(\mathbf{Z}_n)/g(\mathbf{Z}_n)\right)_{n\geq 1}$ sont *i.i.d.* et d'espérance finie sous g. La loi forte des grands nombre donne

$$\widehat{\delta}_n(g) \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \mathbb{E}[Y_1] = \mathbb{E}_g\left[\frac{h(\mathbf{Z})f(\mathbf{Z})}{g(\mathbf{Z})}\right] = \mathbb{E}_f[h(\mathbf{X})]. \tag{3.2}$$

Variance de l'estimateur. Les variables aléatoires $(\mathbf{Z}_n)_{n\geq 1}$ étant *i.i.d.* suivant la densité g, la variance de l'estimateur s'écrit

$$\mathbb{V}\text{ar}_g\left[\widehat{\delta}_n(g)\right] = \frac{1}{n}\left\{\mathbb{E}_g\left[\frac{h^2(\mathbf{Z})f^2(\mathbf{Z})}{g^2(\mathbf{Z})}\right] - \delta^2\right\} = \frac{1}{n}\int_{\mathscr{D}} \frac{\{h(\mathbf{x})f(\mathbf{x}) - \delta g(\mathbf{x})\}^2}{g(\mathbf{x})}\mathrm{d}\mathbf{x}. \tag{3.3}$$

En pratique, un estimateur naturel de $\sigma_1^2 = \mathbb{V}\text{ar}[Y_1]$ est

$$\widehat{\sigma}_1^2(g) = \frac{1}{n-1}\sum_{k=1}^{n}\left\{w_k h(\mathbf{Z}_k) - \widehat{\delta}_n(g)\right\}^2.$$

Choix de la densité instrumentale

La méthode d'échantillonnage préférentiel repose sur le fait qu'il n'y a pas unicité de la loi par rapport à laquelle on intègre pour calculer δ. Cela laisse beaucoup de liberté quant au choix de g parmi les lois faciles à simuler. L'égalité (3.3) nous renseigne sur le choix de bonnes lois instrumentales.

1. La première égalité indique que g est d'autant meilleure que $\mathbb{E}_g\left[h^2(\mathbf{X})f^2(\mathbf{X})/g^2(\mathbf{X})\right]$ est petite (borne supérieure pour la variance). Mais elle illustre également un inconvénient majeur de l'estimateur d'échantillonnage préférentiel : sa variance n'est pas toujours finie. Lorsque $\mathbb{V}\text{ar}_f[h(\mathbf{X})] < +\infty$, une condition suffisante pour garantir que $\widehat{\delta}_n(g)$ soit de variance finie est de choisir une densité g telle que f/g soit bornée ((*i.e.*, les distributions g ayant des queues de distribution plus légères que celles de f sont à proscrire). Si ce rapport n'est pas borné, les poids d'importance $f(\mathbf{Z}_k)/g(\mathbf{Z}_k)$, $k = 1,\ldots,n$, sont très variables et seuls certains termes dans la définition de $\widehat{\delta}_n(g)$ ont un poids significatif. Autrement dit, on observe des variations brusques de la valeur estimée d'une itération à l'autre.
2. La seconde égalité dans (3.3) indique qu'en prenant g proportionnel à hf on obtient un estimateur de variance presque nulle, voire nulle si la fonction h est de signe constant. Plus précisément, la loi instrumentale optimale en terme de variance est donnée par le résultat suivant.

Proposition 3.1

L'estimateur d'échantillonnage préférentiel de variance minimale, $\widehat{\delta}_n(g^\star)$, est obtenu pour la densité instrumentale

$$g^\star(\mathbf{z}) = \frac{|h(\mathbf{z})|f(\mathbf{z})}{\int_{\text{supp}(f)}|h(\mathbf{x})|f(\mathbf{x})\mathrm{d}\mathbf{x}}, \quad \mathbf{z}\in\mathbb{R}^d.$$

Ce résultat n'a qu'un intérêt limité en pratique. Lorsque $h > 0$, la densité instrumentale optimale est $g^\star = hf/\mathbb{E}_f[h(\mathbf{X})]$ et $\mathbb{V}\text{ar}_{g^\star}\left[\widehat{\delta}_n(g^\star)\right] = 0$. Néanmoins, cela requiert de connaître $\delta = \mathbb{E}_f[h(\mathbf{X})]$ qui est

justement la quantité d'intérêt. La proposition 3.1 fournit en revanche une stratégie pour choisir g : un candidat pertinent est tel que $|h|f/g$ soit quasi constant et de variance finie.

Estimateur auto-normalisé

Il est possible que les densités f et/ou g ne soient connues qu'à une constante de normalisation près, *i.e.*, $f = c_f \tilde{f}$ et $g = c_g \tilde{g}$. Dans ce cas, on travaille avec une version de l'estimateur dite auto-normalisée.

Définition 3.2

Sous les même hypothèses que la définition 3.1 et étant donnée $(\mathbf{Z}_n)_{n\geq 1}$ une suite de variables aléatoires *i.i.d.* suivant g, l'estimateur auto-normalisée est défini par

$$\widehat{\delta}_n(\tilde{g}) = \frac{\sum_{k=1}^n \tilde{w}_k h(\mathbf{Z}_k)}{\sum_{k=1}^n \tilde{w}_k}, \quad \text{avec} \quad \tilde{w}_k = \frac{\tilde{f}(\mathbf{Z}_k)}{\tilde{g}(\mathbf{Z}_k)}, \quad k = 1, \ldots, n, \tag{3.4}$$

$$= \frac{\sum_{k=1}^n w_k h(\mathbf{Z}_k)}{\sum_{k=1}^n w_k}. \tag{3.5}$$

Proposition 3.2

L'estimateur (3.4) converge presque sûrement vers δ.

À la différence de l'estimateur d'échantillonnage préférentiel (3.2), l'estimateur auto-normalisé est biaisé. Il présente néanmoins l'avantage d'être exact lorsque la fonction h est constante (ce qui n'est pas le cas pour (3.2)). Certes, il n'y a pas d'intérêt à calculer des intégrales pour des fonctions h constantes, en revanche cette propriété permet d'éviter des comportements atypiques.

Exemple 3.2

Supposons que l'on souhaite estimer $\delta = \mathbb{E}_f[h(\mathbf{X}) + C]$ avec C une constante. Soit $(\mathbf{Z}_n)_{n\geq 1}$ une suite de variables aléatoires *i.i.d.* suivant g. Alors

$$\frac{1}{n}\sum_{k=1}^n w(\mathbf{Z}_k)\{h(\mathbf{Z}_k) + C\} = \frac{1}{n}\sum_{k=1}^n w_k h(\mathbf{Z}_k) + \frac{C}{n}\sum_{k=1}^n w_k \neq C + \frac{1}{n}\sum_{k=1}^n w_k h(\mathbf{Z}_k).$$

Autrement dit, l'estimateur d'échantillonnage préférentiel de δ n'est pas égal à l'estimateur d'échantillonneur préférentiel de $\mathbb{E}_f[h(\mathbf{X})]$ auquel on ajoute C. L'estimation ne préserve pas les propriétés de l'espérance. En revanche, si on considère la version auto-normalisée la propriété attendue est bien vérifiée :

$$\frac{\sum_{k=1}^n w_k\{h(\mathbf{Z}_k) + C\}}{\sum_{k=1}^n w_k} = C + \frac{\sum_{k=1}^n w_k h(\mathbf{Z}_k)}{\sum_{k=1}^n w_k}.$$

Diagnostique et taille efficace d'échantillon (*effective sample size*)

S'il est possible de comparer les performances en terme de variance de l'échantillonnage préférentiel avec la méthode classique, cette comparaison est fortement sensible à la qualité d'estimation de la variance et donc au choix de g. Il peut être préférable de mener d'autres diagnostiques en amont, notamment pour s'assurer que le choix de g est judicieux. Par exemple, lorsque l'on sait calculer explicitement les poids d'importances (f et g sont connues explicitement et pas à une constante près), on peut utiliser que $\mathbb{E}_g\left[f(\mathbf{Z})/g(\mathbf{Z})\right] = 1$ et donc si la moyenne empirique des poids $\overline{w}_n$ diffère beaucoup de 1, le choix de g n'est pas bon.

On utilise généralement la taille efficace d'échantillon pour juger du choix de g :

$$ESS = \frac{\left(\sum_{k=1}^n w_k\right)^2}{\sum_{k=1}^n w_k^2} = n\frac{\overline{w}_n^2}{\overline{w_n^2}}.$$

ESS représente le nombre d'observations apportant effectivement de l'information. Ainsi si la taille efficace est très petite devant n, on considèrera que l'estimation n'est pas fiable. Il existe de nombreuses définitions de la taille efficace d'un échantillon dans la littérature. Il n'est pas possible d'en faire une description exhaustive dans ce cours. On retiendra néanmoins, qu'il ne s'agit pas d'un outil de diagnostique idéal, il permet au mieux d'identifier un problème concernant les poids d'importance mais jamais de valider la méthode.

3.2 Variables antithétiques

La méthode de la variable antithétique a pour objectif de réduire l'erreur d'estimation en utilisant des symétries du problème. Elle s'applique lorsque la loi ν présente des propriétés d'invariances (*e.g.*, symétrie ou rotation). Autrement dit, on suppose qu'il existe une transformation mesurable $A : \mathbb{R}^d \to \mathbb{R}^d$ telle que $A(\mathbf{X}) \sim \nu$.

Exemple 3.3

1. Si $U \sim \mathcal{U}([a,b])$, alors $b + a - U \sim \mathcal{U}([a,b])$.
2. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $2\mu - X \sim \mathcal{N}(\mu, \sigma^2)$.

Définition 3.3. Méthode de la variable antithétique

La variable aléatoire $A(\mathbf{X})$ est appelée variable antithétique de $\mathbf{X}$ et l'estimateur de la variable antithétique est défini par

$$\widehat{\delta}_n = \frac{1}{n}\sum_{k=1}^n \frac{h(\mathbf{X}_k) + h \circ A(\mathbf{X}_k)}{2} \tag{3.6}$$

où $(\mathbf{X}_n)_{n\geq 1}$ est une suite de variables aléatoires *i.i.d.* suivant la loi de $\mathbf{X}$.

Biais de l'estimateur. Les variables aléatoires $\mathbf{X}$ et $A(\mathbf{X})$ étant identiquement distribuées, on a $\mathbb{E}[h(\mathbf{X})] = \mathbb{E}[h \circ A(\mathbf{X})]$. On obtient donc que l'estimateur (3.6) est sans biais, *i.e.*, $\mathbb{E}\left[\widehat{\delta}_n\right] = \mathbb{E}[h(\mathbf{X})]$.

Convergence de l'estimateur. La loi forte des grands nombres pour les suites de variables aléatoires *i.i.d.* $(h(\mathbf{X}_n))_{n\in\mathbb{N}}$ et $(h \circ A(\mathbf{X}_n))_{n\in\mathbb{N}}$ (et le théorème de continuité) donne

$$\left.\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{k=1}^n h(\mathbf{X}_k) \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[h(\mathbf{X})] \\ \dfrac{1}{n}\displaystyle\sum_{k=1}^n h \circ A(\mathbf{X}_k) \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[h \circ A(\mathbf{X})] = \mathbb{E}[h(\mathbf{X})] \end{array}\right\} \quad \Rightarrow \quad \widehat{\delta}_n \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[h(\mathbf{X})].$$

Intervalle de confiance. Les variables aléatoires $(Y_n)_{n\geq 1} = (0.5\{h(\mathbf{X}_n) + h \circ A(\mathbf{X}_n)\})_{n\geq 1}$ sont *i.i.d.* et de variance finie (h étant de carré intégrable par rapport à ν). Le théorème centrale limite donne

$$\sqrt{n}\left(\widehat{\delta}_n - \mathbb{E}[h(\mathbf{X})]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_1^2\right), \quad \text{avec} \quad \sigma_1^2 = \mathbb{V}\mathrm{ar}[0.5\{h(\mathbf{X}) + h \circ A(\mathbf{X})\}].$$

On en déduit l'intervalle de confiance au niveau de confiance $1-\alpha$,

$$\mathrm{IC}_{1-\alpha} = \left[\widehat{\delta}_n - q_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n}}, \widehat{\delta}_n + q_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n}}\right] = \left[\widehat{\delta}_n - q_{1-\alpha/2}\sqrt{\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\right]}, \widehat{\delta}_n + q_{1-\alpha/2}\sqrt{\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\right]}\right],$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Dans la pratique, on peut estimer la variance σ_1^2 via la variance empirique

$$\widehat{\sigma}_1^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(0.5\left\{h(\mathbf{X}_k) + h\circ A(\mathbf{X}_k)\right\} - \widehat{\delta}_n\right)^2.$$

Performances de l'estimateur

Les variables $(\mathbf{X}_n)_{n\geq 1}$ étant *i.i.d.* suivant ν, ajouté au fait que $\mathbf{X}$ et $A(\mathbf{X})$ sont de même loi, on a

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\right] &= \frac{1}{4n}\left(\mathbb{V}\mathrm{ar}[h(\mathbf{X})] + \mathbb{V}\mathrm{ar}[h\circ A(\mathbf{X})] + 2\mathrm{Cov}[h(\mathbf{X}), h\circ A(\mathbf{X})]\right)\\ &= \frac{1}{2n}\mathbb{V}\mathrm{ar}[h(\mathbf{X})] + \frac{1}{2n}\mathrm{Cov}[h(\mathbf{X}), h\circ A(\mathbf{X})] = \frac{1}{2n}\mathbb{V}\mathrm{ar}[h(\mathbf{X})](1+\rho),\end{aligned}$$

avec $\rho = \mathrm{Cov}[h(\mathbf{X}), h\circ A(\mathbf{X})]/\mathbb{V}\mathrm{ar}[h(\mathbf{X})]$.

Cette expression de la variance justifie *a posteriori* pourquoi dans la définition de l'estimateur (3.6), on prend le second échantillon identiquement distribué selon la loi de $\mathbf{X}$, mais pas indépendant de $(\mathbf{X}_n)_{n\geq 1}$. Si l'on n'utilise pas la suite $(\mathbf{X}_n)_{n\geq 1}$ pour construire la suite de réalisations correspondant à la variable antithétique $A(\mathbf{X})$, mais une suite de variables aléatoires *i.i.d.* $\left(\tilde{\mathbf{X}}_n\right)_{n\geq 1}$, alors $\rho = 0$. La méthode n'est rien d'autre que la méthode de Monte Carlo classique pour un échantillon de taille $2n$.

Le cas favorable pour la méthode de la variable antithétique est celui de la corrélation négative. Dans le cas particulier où $\rho = -1$, on peut même constater que l'on a un estimateur de variance nulle.

Proposition 3.3

On a la condition nécessaire et suffisante suivante

$$\mathrm{Cov}[h(\mathbf{X}), h\circ A(\mathbf{X})] < 0 \iff \mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\right] < \frac{1}{2}\mathbb{V}\mathrm{ar}\left[\overline{h}_n\right] = \mathbb{V}\mathrm{ar}\left[\overline{h}_{2n}\right].$$

Ce résultat nous dit que sous la condition $\rho < 0$, la variance de l'estimateur est au moins divisée par rapport à la méthode de Monte Carlo classique qui utilise le même nombre de simulations et fait au moins aussi bien que la méthode de Monte Carlo pour laquelle on utilise deux fois plus de simulation et on évalue le même nombre de fois h.

On peut comparer l'efficacité relative de $\widehat{\delta}_n$ par rapport à $\overline{h}_n$ ou $\overline{h}_{2n}$ en prenant en compte le coût C_ν de simulation suivant ν et le coût C_h d'évaluation de h :

$$R(\overline{h}_n, \widehat{\delta}_n) = R(\overline{h}_{2n}, \widehat{\delta}_n) = \frac{C_\nu + C_h}{C_\nu + 2C_h}\frac{2}{1+\rho} := \frac{C}{C_1}\frac{2}{1+\rho}, \quad \text{avec} \quad -1 \leq \rho \leq 1.$$

Cas n°1 : $C_\nu \gg C_h$**.** Si le coût d'estimation est dominé par le coût de simulation suivant ν, alors $C_1 = C$ et le gain $R(\overline{h}_n, \widehat{\delta}_n) \geq 1$. $\widehat{\delta}_n$ fait au moins aussi bien que $\overline{h}_n$ ou $\overline{h}_{2n}$. Si de plus $\rho < 0$, $\widehat{\delta}_n$ est au moins deux fois plus efficace que $\overline{h}_n$ ou $\overline{h}_{2n}$.

Cas n°2 : $C_\nu \ll C_h$**.** Si le coût d'estimation est dominé par le coût d'évaluation de h, alors $C_1 = 2C$

et $R(\overline{h}_n, \widehat{\delta}_n) \geq 0.5$. Ainsi si $0 \leq \rho \leq 1$, $\widehat{\delta}_n$ est moins efficace que $\overline{h}_n$ ou $\overline{h}_{2n}$. Dans le cas limite $\rho = 1$, la méthode de la variable antithétique double la variance de la méthode classique pour un même nombre d'évaluation de h (ou de façon équivalente $\widehat{\delta}_n$ atteint la même variance que $\overline{h}_n$ en deux fois plus de temps). En revanche, lorsque $\rho < 0$, $\widehat{\delta}_n$ fait au moins aussi bien que $\overline{h}_n$ ou $\overline{h}_{2n}$. Donc même si on utilisait deux fois plus de simulations avec la méthode classique, la méthode de la variable antithétique resterait meilleure.

Cas n°3 : $C_v \approx C_h$**.** On a alors $C_1 = 3C/2$ et $R(\overline{h}_n, \widehat{\delta}_n) \geq 1$ dès lors que $\rho \leq 1/3$. Il suffit donc là encore $\rho \leq 0$ pour que $\widehat{\delta}_n$) soit plus efficace que $\overline{h}_n$ ou $\overline{h}_{2n}$

Sous certaines conditions, on est assuré que la covariance associée à la méthode est négative.

Proposition 3.4

Soient $\mathbf{X}$ une variable aléatoire de $\mathbb{R}^d$ de loi ν. Si Φ et Ψ sont des fonctions mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$ vérifiant :

(i) $\Phi(\mathbf{X})$ et $\Psi(\mathbf{X})$ ont même loi et $\mathbb{E}_\nu\left[\Phi^2(\mathbf{X})\right] < +\infty$;

(ii) Φ et Ψ sont respectivement croissante et décroissante en chacune de leurs coordonnées,

alors $\mathrm{Cov}[\Phi(\mathbf{X}), \Psi(\mathbf{X})] \leq 0$.

Proposition 3.5

Si A est une transformation de $\mathbb{R}^d$ décroissante en chacune de ces coordonnées laissant la loi ν invariante et $h : \mathbb{R}^d \to \mathbb{R}$ est monotone en chacune de ses coordonnées. Alors $\mathrm{Cov}[h(X), h \circ A(X)] \leq 0$, avec égalité uniquement si h est presque sûrement constante.

Exemple 3.4. Cas gaussien et retour sur le modèle de Black-Scholes

Lorsque $X \sim \mathcal{N}(\mu, \sigma^2)$, on a la variable antithétique $A(X) = 2\mu - X$. La transformation A est donc décroissante et la Proposition 3.5 s'applique. Un exemple d'application est le calcul d'une option d'achat. On se place dans le modèle de Black-Scholes

$$S_t = S_0 \exp\left(\frac{r - \sigma^2}{2} t + \sigma W_t\right), \quad \text{avec} \quad \begin{cases} r & \text{le taux d'intérêt sans risque,} \\ \sigma & \text{la volatilité du prix de l'option,} \end{cases}$$

avec $(W_t)_{t \in \mathbb{R}_+}$ un mouvement brownien standard. On cherche à calculer

$$\delta = \mathbb{E}\left[\exp(-rT)\,(S_T - K)+\right], \quad K \geq 0.$$

En remarquant que $W_T \sim \mathcal{N}(0, T)$, on peut écrire $\delta = \mathbb{E}[h(X)]$ avec $X \sim \mathcal{N}(0,1)$ et

$$h : x \mapsto \exp(-rT)\left(S_0 \exp\left(\frac{r - \sigma^2}{2} t + \sigma\sqrt{T}x\right) - K\right)^+.$$

La fonction h est strictement croissante. Ainsi d'après la Proposition 3.5 donne

$$\mathrm{Cov}[h(X), h \circ A(X)] < 0,$$

et la méthode de la variable antithétique fournit un estimateur de Monte Carlo plus efficace.

Le résultat sur la monotonie a une utilité théorique pour justifier la réduction de la variance, mais il ne garantit pas nécessairement un gain significatif (*e.g.*, si $\rho < 0$ mais proche de 0) dans la pratique.

La condition est uniquement suffisante. On peut donc *a contrario* obtenir une corrélation négative pour une fonction h qui n'est pas monotone. En pratique, il peut donc, dans certains cas, étudier empiriquement la corrélation même si la condition de monotonie n'est pas vérifiée.

3.3 Variables de contrôle

La méthode de la variable de contrôle consiste à réduire la variance de la méthode Monte Carlo en introduisant un problème Monte Carlo que l'on sait résoudre exactement.

3.3.1 Cas unidimensionnel

Définition 3.4. Méthode de la variable de contrôle

Soit $h_0 : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}[h_0(\mathbf{X})] = m$ soit facile à calculer et $\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})] < +\infty$. Pour tout réel b, étant donné $(\mathbf{X}_n)_{n\geq 1}$ une suite de variables aléatoires *i.i.d.* suivant la loi ν, on définit l'estimateur

$$\widehat{\delta}_n(b) = \frac{1}{n}\sum_{k=1}^{n}\{h(\mathbf{X}_k) - b[h_0(\mathbf{X}_k) - m]\}. \tag{3.7}$$

Biais de l'estimateur. Comme $m = \mathbb{E}[h_0(\mathbf{X})]$, on obtient directement $\mathbb{E}\left[\widehat{\delta}_n\right] = \mathbb{E}[h(\mathbf{X})]$.

Convergence de l'estimateur. La loi forte des grands nombres pour les suites de variables aléatoires *i.i.d.* $(h(\mathbf{X}_n))_{n\geq 1}$ et $(h_0(\mathbf{X}_n))_{n\geq 1}$, donne

$$\left.\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} h(\mathbf{X}_k) \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \mathbb{E}[h(\mathbf{X})] \\ \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} h_0(\mathbf{X}_k) - m \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} 0 \end{array}\right\} \quad \Rightarrow \quad \widehat{\delta}_n(b) \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \mathbb{E}[h(\mathbf{X})], \quad \text{pour tout réel } b.$$

Intervalle de confiance. Les variables aléatoires $(Y_n)_{n\geq 1} = (h(\mathbf{X}_n) - b\{h_0(\mathbf{X}_n) - m\})_{n\geq 1}$ sont *i.i.d.* et de variance finie, notée $\sigma_1^2(b)$. Le théorème centrale limite donne alors, pour tout réel b,

$$\sqrt{n}\left(\widehat{\delta}_n(b) - \mathbb{E}[h(\mathbf{X})]\right) \underset{n\to+\infty}{\overset{\mathscr{L}}{\longrightarrow}} \mathcal{N}\left(0, \sigma^2(b)\right).$$

On en déduit l'intervalle de confiance au niveau de confiance $1-\alpha$

$$\begin{aligned} \mathrm{IC}_{1-\alpha} &= \left[\widehat{\delta}_n - q_{1-\alpha/2}\frac{\sigma_1(b)}{\sqrt{n}}, \widehat{\delta}_n + q_{1-\alpha/2}\frac{\sigma_1(b)}{\sqrt{n}}\right] \\ &= \left[\widehat{\delta}_n - q_{1-\alpha/2}\sqrt{\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b)\right]}, \widehat{\delta}_n + q_{1-\alpha/2}\sqrt{\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b)\right]}\right], \end{aligned}$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Dans la pratique, on estime la variance $\sigma_1^2(b)$ via la variance empirique associée aux réalisations de la variable aléatoire Y :

$$\widehat{\sigma}_1^2(b) = \frac{1}{n-1}\sum_{k=1}^{n}\left\{h(\mathbf{X}_k) - b(h_0(\mathbf{X}_k) - m) - \widehat{\delta}_n(b)\right\}^2.$$

Performances de l'estimateur

Les variables aléatoires $(Y_n)_{n\geq 1}$ étant *i.i.d.*, on obtient que la variance de l'estimateur (3.7) est

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b)\right] &= \frac{1}{n}\sigma^2(b) = \frac{1}{n}\left\{\mathbb{V}\mathrm{ar}[h(\mathbf{X})] + b^2\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})] - 2b\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})]\right\} \\ &= \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right] + \frac{1}{n}\left\{b^2\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})] - 2b\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})]\right\}.\end{aligned}$$

Proposition 3.6

L'estimateur $\widehat{\delta}_n(b)$ est de variance plus faible que l'estimateur classique $\overline{h}_n$ si, et seulement si, on choisit b et h_0 telles que

$$b^2\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})] - 2b\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})] < 0.$$

L'estimateur de variance minimale, $\widehat{\delta}_n(b^\star)$, est obtenu pour

$$b^\star = \underset{b\in\mathbb{R}}{\arg\min}\, \mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b)\right] = \frac{\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})]}{\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})]}. \tag{3.8}$$

On a alors

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b^\star)\right] = \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right]\left(1-\rho(h,h_0)^2\right), \quad \rho(h,h_0) = \frac{\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})]}{\sqrt{\mathbb{V}\mathrm{ar}[h(\mathbf{X})]\mathbb{V}\mathrm{ar}[h_0(\mathbf{X})]}}.$$

Remarque. Comme pour la méthode de la variable antithétique, il est important de réutiliser la suite $(\mathbf{X}_n)_{n\geq 1}$ pour construire les réalisations de la variables $h_0(\mathbf{X})$. Dans le cas contraire, la méthode a la même variance que la méthode classique.

L'estimateur de variance minimale a une variance inférieure à celle de l'estimateur classique $\overline{h}_n$ dès lors que la corrélation $\rho(h,h_0)$ est non nulle (et cela indépendamment du signe de la corrélation). La méthode sera cependant d'autant plus efficace que $\rho(h,h_0)^2$ est proche de 1 et devient même exacte lorsque cette corrélation vaut 1. Le gain potentiel de la méthode dépend également du coût de sa mise en œuvre. Si le coût de la méthode classique est $C = nC_v + nC_h$, la méthode de la variable de contrôle implique un coût supplémentaire du fait de l'évaluation de h_0, à savoir $C_1 = C + nC(h_0)$. L'efficacité relative de $\widehat{\delta}_n(b)$ par rapport à $\overline{h}_n$ est

$$R(\overline{h}_n, \widehat{\delta}_n(b)) = \frac{C}{C+nC(h_0)}\frac{1}{1-\rho(h,h_0)^2}.$$

La variable de contrôle est plus efficace que la méthode classique lorsque $R(\overline{h}_n, \widehat{\delta}_n(b)) > 1$, autrement dit lorsque

$$|\rho(h,h_0)| > \sqrt{\frac{C(h_0)}{C_v + C_h + C(h_0)}}.$$

Comment déterminer le coefficient optimal $b^\star$? Dans la pratique, il y a peu de chances que l'on connaisse $\mathrm{Cov}[h(\mathbf{X}), h_0(\mathbf{X})]$ et donc $b^\star$ (*c.f.*, Exemple 3.3.1). On l'estime donc en utilisant

$$\widehat{b}^\star_\ell = \frac{\sum_{k=1}^{\ell}(h_0(\mathbf{X}_k) - m)\left(h(\mathbf{X}_k) - \overline{h}_\ell\right)}{\sum_{k=1}^{\ell}(h_0(\mathbf{X}_k) - m)^2}. \tag{3.9}$$

Stratégie n°1 : période de chauffe (*burn-in period*). On utilise les ℓ premiers termes (ℓ petits) de la suite $(\mathbf{X}_n)_{n\geq 1}$ pour estimer $b^\star$ et les $n-\ell$ termes restants pour calculer $\widehat{\delta}_{n-\ell}\left(\widehat{b}_\ell^\star\right)$. $\widehat{\delta}_{n-\ell}\left(\widehat{b}_\ell^\star\right)$ est alors sans biais ($\widehat{b}_\ell^\star$ et $(h_0(\mathbf{X}_n)-m)_{n>\ell}$ sont indépendantes). Néanmoins ℓ a vocation a être petit et donc l'estimateur $\widehat{b}_\ell^\star$ est peu précis.

Stratégie n°2. On utilise l'ensemble de l'échantillon de taille n pour estimer $b^\star$ (*i.e.*, $\ell = n$) et $\widehat{\delta}_n\left(\widehat{b}_n^\star\right)$. Cette solution offre une meilleure estimation de $\widehat{b}_\ell^\star$. Néanmoins $\widehat{\delta}_n\left(\widehat{b}_n^\star\right)$ est alors biaisé mais de variance asymptotique $\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\left(b^\star\right)\right]$.

Exemple 3.5. Distribution symétrique

Soit X une variable aléatoire réelle de densité f symétrique par rapport à μ. On souhaite calculer la probabilité d'obtenir une réalisation au delà de μ,

$$\mathbb{P}[X \geq a] = \int_a^{+\infty} f(x)\mathrm{d}x, \quad a > \mu,$$

Un candidat naturel pour appliquer la méthode de la variable de contrôle est $h_0(X) = \mathbb{1}_{\{X\geq\mu\}}$. Dans ce cas $m = \mathbb{P}\left[X \geq \mu\right] = 0.5$. Pour une suite $(X_n)_{n\geq 1}$ de variables aléatoires *i.i.d.* suivant la loi de X, on considère alors l'estimateur

$$\widehat{\delta}_n(b) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{X_k>a\}} - b\left(\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{X_k\geq\mu\}} - \mathbb{P}\left[X\geq\mu\right]\right).$$

On a

$$\mathbb{V}\mathrm{ar}[h_0(X)] = \mathbb{P}\left[X\geq\mu\right]\left(1-\mathbb{P}\left[X\geq\mu\right]\right) \quad \text{et} \quad \mathrm{Cov}[h(X),h_0(X)] = \mathbb{P}[X>a]\left(1-\mathbb{P}\left[X\geq\mu\right]\right).$$

On en déduit que l'estimateur $\widehat{\delta}_n(b)$ est plus efficace que l'estimateur $\overline{h}_n$ lorsque

$$0 < b < 2\frac{\mathrm{Cov}[h(X),h_0(X)]}{\mathbb{V}\mathrm{ar}[h_0(X)]} = 2\frac{\mathbb{P}[X>a]}{\mathbb{P}\left[X\geq\mu\right]} = 4\mathbb{P}[X>a].$$

Or $\mathbb{P}[X>a]$ est justement la quantité que l'on cherche à calculer.

3.3.2 Cas des variables de contrôle multiples

La méthode peut se généraliser à plusieurs variables de contrôle.

Définition 3.5. Variables de contrôle multiples

Soit $h_1,\ldots,h_s$ un ensemble de fonctions de $\mathbb{R}^d$. On pose $\mathbf{Z}_k = (h_1(\mathbf{X}_k),\ldots,h_s(\mathbf{X}_k))$, $k\geq 1$. Pour tout vecteur b de $\mathbb{R}^s$, on définit

$$\widehat{\delta}_n(b) = \frac{1}{n}\sum_{k=1}^n h(\mathbf{X}_k) - < b, Z_k - \mathbb{E}[Z_k] > .$$

Tout comme dans le cas unidimensionnel, l'estimateur $\widehat{\delta}_n(b)$ est sans biais et on peut écrire

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b)\right] = \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right] - \frac{1}{n}\left(2b^T\Sigma_{h(\mathbf{X}),\mathbf{Z}} - b^T\Sigma_{\mathbf{Z}}b\right),$$

avec $\Sigma_{\mathbf{Z}}$ la matrice de variance covariance de $\mathbf{Z} = (h_1(\mathbf{X}),\ldots,h_s(\mathbf{X}))$ et $\Sigma_{h(\mathbf{X}),\mathbf{Z}}$ la matrice $s\times 1$ dont la i-ème composante, $i = 1,\ldots,s$, est donnée par $\mathrm{Cov}[h(\mathbf{X}),h_i(\mathbf{X})]$. L'estimateur de variance minimale

est alors obtenu pour

$$b^\star = \Sigma_{\mathbf{Z}}^{-1}\Sigma_{h(\mathbf{X}),\mathbf{Z}} \quad \text{et} \quad \mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(b^\star)\right] = \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right]\left(1 - \frac{1}{\mathbb{V}\mathrm{ar}\left[h(\mathbf{X})\right]}\Sigma_{h(\mathbf{X}),\mathbf{Z}}^T\Sigma_{\mathbf{Z}}^{-1}\Sigma_{h(\mathbf{X}),\mathbf{Z}}\right).$$

On obtient donc que l'estimateur de variance minimale est plus efficace que l'estimateur $\overline{h}_n$.

3.4 Méthodes de stratification

Soient $\mathbf{Z}$ une variable aléatoire de $\mathbb{R}^d$ et $(D_k : k = 1,\ldots,K)$ une partition de l'ensemble $\mathscr{D}$ de l'ensemble des valeurs prises par $\mathbf{Z}$. Le principe de la méthode de stratification est de combiner l'information obtenue sur chaque élément de la partition en utilisant la relation :

$$\delta = \mathbb{E}_\nu[h(\mathbf{X})] = \sum_{k=1}^{K} \mathbb{P}[\mathbf{Z} \in D_k]\mathbb{E}[h(\mathbf{X}) \mid \mathbf{Z} \in D_k].$$

Le postulat de base de ces méthodes est que pour tout élément D_k, $k = 1,\ldots,K$:

(H3) $\mathbb{P}[\mathbf{Z} \in D_k]$ est connue explicitement et est strictement positif;

(H4) on sait simuler suivant la loi conditionnelle $\mathscr{L}(\mathbf{X} \mid \mathbf{Z} \in D_k)$.

Ces conditions sont naturelles et permettent simplement d'assurer que l'on a un estimateur Monte Carlo de $\mathbb{E}[h(\mathbf{X}) \mid \mathbf{Z} \in D_k]$ et donc de $\mathbb{E}[h(\mathbf{X})]$. En pratique, elles servent de guide à la construction de la partition.

Remarque. Les résultats sont écrits en toute généralité pour une variable aléatoire $\mathbf{Z}$. Ils sont en particulier vrais pour $\mathbf{Z} = \mathbf{X}$ et on pourra considérer que c'est le cas en première lecture.

Exemple 3.6. Simulation de lois conditionnelles $\mathscr{L}(\mathbf{X} \mid \mathbf{Z} \in D_k)$

1. Soit X une variable aléatoire réelle de fonction de répartition F. On considère $Z = X$ et pour $k = 1,\ldots,K$, $D_k =]d_k, d_k + 1]$ avec d_k des constantes réelles. Si $U \sim \mathscr{U}([0,1])$, alors

$$X^{(k)} = F^{\leftarrow}\left[F(d_k) + U\{F(d_{k+1}) - F(d_k)\}\right], k = 1,\ldots,K,$$

suit la loi de $X \mid X \in D_k$ et $Y^{(k)} = h\left(X^{(k)}\right)$ suit la loi de $h(X) \mid X \in D_k$. On applique cette méthode lorsque l'on sait évaluer F et $F^{\leftarrow}$ ou lorsque la partition est définie à l'aide de quantiles (*e.g.*, loi normale).

2. Soit $\mathbf{X}$ une variable aléatoire de $\mathbb{R}^d$ de densité f. On considère $\mathbf{Z} = \mathbf{X}$. Alors, pour $k \in \{1,\ldots,K\}$,

$$f_k(\mathbf{x}) = \frac{f(\mathbf{x})}{\mathbb{P}[\mathbf{X} \in D_k]}\mathbb{1}_{\{\mathbf{x} \in D_k\}}, \quad x \in \mathbb{R}^d,$$

est la densité de la loi conditionnelle $\mathscr{L}(\mathbf{X} \mid \mathbf{X} \in D_k)$. On peut alors appliquer l'algorithme du rejet.

Définition 3.6. Estimateur stratifié

Pour chaque élément D_k de la partition, on considère $\left(\mathbf{X}_n^{(k)}\right)_{n\in\mathbb{N}}$ une suite de variables aléatoires *i.i.d.* suivant la loi conditionnelle $\mathscr{L}(\mathbf{X} \mid \mathbf{Z} \in D_k)$. Alors pour $n = n_1 + \ldots + n_K$, on définit

l'estimateur stratifié par

$$\widehat{\delta}_n(n_1,\dots,n_K) = \sum_{k=1}^{K} \frac{\mathbb{P}[\mathbf{Z}\in D_k]}{n_k} \sum_{i=1}^{n_k} h\left(\mathbf{X}_i^{(k)}\right).$$

On appelle :
- variable de stratification : la variable aléatoire $\mathbf{Z}$;
- strates : les éléments D_k de la partition ;
- allocation : le choix des nombre de tirages $n_1,\dots,n_K$ que l'on fait pour les lois conditionnelles $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}\in D_k)$, $k=1,\dots,K$, sous la contrainte que le nombre total de simulation est $n = n_1 + \dots + n_K$.

Cet estimateur nécessite de choisir des strates, une variable de stratification et de se donner une allocation ; tant de paramètres qui peuvent rendre l'implémentation d'un tel estimateur délicate.

Par soucis de concision, on notera pour chaque strate D_k, $k\in\{1,\dots,K\}$:

$$p_k = \mathbb{P}[\mathbf{Z}\in D_k], \quad q_k = n_k/n, \quad \mu_k = \mathbb{E}[h(\mathbf{X}) \mid \mathbf{Z}\in D_k], \quad \sigma_k^2 = \mathbb{V}\mathrm{ar}[h(\mathbf{X}) \mid \mathbf{Z}\in D_k]$$

Biais de l'estimateur. Pour tout k, on a

$$\mathbb{E}\left[\frac{1}{n_k}\sum_{i=1}^{n_k} h\left(\mathbf{X}_i^{(k)}\right)\right] = \begin{cases} \mu_k & \text{si } n_k > 0, \\ 0 & \text{sinon.} \end{cases} \quad \Rightarrow \quad \mathbb{E}\left[\widehat{\delta}_n(n_1,\dots,n_K)\right] = \sum_{k=1}^{K} p_k\mu_k \mathbb{1}_{\{n_k>0\}}.$$

L'estimateur est donc sans biais si pour tout $k\in\{1,\dots,K\}$, $n_k>0$, *i.e.* pour chaque strate, on prends au moins un tirage suivant la loi conditionnelle.

Convergence de l'estimateur. Pour $k\in\{1,\dots,K\}$, tel que $n_k>0$, la loi forte des grands nombres appliquée aux variables aléatoires *i.i.d.* $\left(\mathbf{X}_n^{(k)}\right)_{n\in\mathbb{N}}$ donne

$$\frac{1}{n_k}\sum_{i=1}^{n_k} h\left(\mathbf{X}_i^{(k)}\right) \underset{n_k\to+\infty}{\overset{p.s.}{\longrightarrow}} \mu_k. \quad \Rightarrow \quad \widehat{\delta}_n(n_1,\dots,n_K) \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \sum_{k=1}^{K} p_k\mu_k\mathbb{1}_{\{n_k>0\}}.$$

L'estimateur converge donc vers $\mathbb{E}[h(\mathbf{X})]$ si pour tout $k\in\{1,\dots,K\}$, $n_k>0$. On supposera cette hypothèse désormais satisfaite.

Intervalle de confiance. On peut écrire

$$\sqrt{n}\left(\widehat{\delta}_n(n_1,\dots,n_K) - \mathbb{E}[h(\mathbf{X})]\right) = \sum_{k=1}^{K} \frac{p_k}{\sqrt{q_k}} \sqrt{n_k}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} h\left(\mathbf{X}_i^{(k)}\right) - \mu_k\right).$$

Or le théorème centrale limite appliqué aux variables aléatoires *i.i.d.* $\left(\mathbf{X}_n^{(k)}\right)_{n\in\mathbb{N}}$, $k=1,\dots,K$ donne

$$\sqrt{n_k}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} h\left(\mathbf{X}_i^{(k)}\right) - \mu_k.\right) \underset{n_k\to+\infty}{\overset{\mathscr{L}}{\longrightarrow}} \mathcal{N}\left(0,\sigma_k^2\right).$$

L'indépendance entre les strates permet alors d'écrire

$$\sqrt{n}\left(\widehat{\delta}_n(n_1,\dots,n_K) - \mathbb{E}[h(\mathbf{X})]\right) \underset{n\to+\infty}{\overset{\mathscr{L}}{\longrightarrow}} \mathcal{N}\left(0,\sigma^2(q_1,\dots,q_K)\right), \quad \text{où} \quad \sigma^2(q_1,\dots,q_K) = \sum_{k=1}^{K} \frac{p_k^2}{q_k}\sigma_k^2.$$

On en déduit l'intervalle de confiance au niveau de confiance $1-\alpha$

$$\mathrm{IC}_{1-\alpha} = \left[\widehat{\delta}_n - q_{1-\alpha/2}\frac{\sigma(q_1,\ldots,q_K)}{\sqrt{n}}, \widehat{\delta}_n + q_{1-\alpha/2}\frac{\sigma(q_1,\ldots,q_K)}{\sqrt{n}}\right],$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Dans la pratique, on estime la variance $\sigma^2(q_1,\ldots,q_K)$ via les variances intra-strate empiriques associées aux réalisations des variables aléatoires $\left(\mathbf{X}_n^{(k)}\right)_{n\in\mathbb{N}}$ dans chacune des strates D_k, $k=1,\ldots,K$:

$$\widehat{\sigma}_n^2(q_1,\ldots,q_K) = \sum_{k=1}^{K}\frac{p_k^2}{q_k}\frac{1}{n_k-1}\sum_{i=1}^{n_k}\left\{h\left(\mathbf{X}_i^{(k)}\right) - \frac{1}{n_k}\sum_{j=1}^{n_k}h\left(\mathbf{X}_j^{(k)}\right)\right\}^2.$$

Il s'agit d'un estimateur convergent de $\sigma(q_1,\ldots,q_K)$ et donc l'intervalle de confiance reste valide en remplaçant $\sigma(q_1,\ldots,q_K)$ par son estimation (théorème de Slutsky).

Performances de l'estimateur

Les variables étant indépendantes (entre les strates et à l'intérieur des strates) et identiquement distribuées à l'intérieur d'une strate, on a

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(n_1,\ldots,n_K)\right] = \sum_{k=1}^{K}\frac{p_k^2}{n_k}\sigma_k^2 = \frac{1}{n}\sigma^2(q_1,\ldots,q_K).$$

La variance de l'estimateur dépend donc du choix des strates et du choix de l'allocation $(n_1,\ldots,n_K)$ ou de façon équivalente de $(q_1,\ldots,q_K)$. On peut chercher l'allocation optimale, *i.e.* l'allocation telle que la variance de l'estimateur soit minimale. On cherche alors

$$\left(q_1^\star,\ldots,q_K^\star\right) = \underset{(q_1,\ldots,q_K)\in]0,1[^K}{\operatorname{argmin}} \sigma^2(q_1,\ldots,q_K), \quad \text{tel que} \quad \sum_{k=1}^{K} q_k = 1.$$

Proposition 3.7. Allocation optimale

(i) L'estimateur de variance minimal, $\widehat{\delta}_n(n_1^\star,\ldots,n_K^\star)$, est obtenu pour l'allocation optimale définie pour $k=1,\ldots,K$ par

$$q_k^\star = \frac{p_k\sigma_k}{\sum_{i=1}^{K}p_i\sigma_i}.$$

On a alors

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(q_1^\star,\ldots,q_K^\star)\right] = \frac{1}{n}\left(\sum_{k=1}^{K}p_k\sigma_k\right)^2.$$

(ii) Sous l'hypothèse d'allocation optimale, l'estimateur stratifié $\widehat{\delta}_n(q_1^\star,\ldots,q_K^\star)$ est plus efficace que l'estimateur de Monte Carlo classique, *i.e.*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(q_1^\star,\ldots,q_K^\star)\right] \leq \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right].$$

Le choix de l'allocation optimale dépend néanmoins des variances intra-strate σ_k et peut donc être incalculable en pratique. Une solution est de remplacer la variance σ_k par un estimateur convergeant évalué sur un premier jeu de simulations indépendant du jeu de simulations utilisé pour calculer l'estimateur stratifié. Il existe également des méthodes plus sophistiquées et performantes basées sur des techniques itératives.

Une alternative est de choisir une allocation tel que l'estimateur stratifié est de variance plus faible

que l'estimateur de Monte Carlo classique sans être de variance minimale.

Définition 3.7. Allocation proportionnelle

L'allocation est dite proportionnelle lorsque l'allocation pour chaque strate D_k, $k = 1,\dots,K$, est proportionnelle à la probabilité d'appartenance à la strate, *i.e.* $q_k = p_k$. Dans ce cas l'estimateur stratifié est noté $\widehat{\delta}_n(p_1,\dots,p_K)$.

Cette allocation permet toujours de construire un estimateur stratifié de variance plus faible que l'estimateur de Monte Carlo classique $\overline{h}_n$.

Proposition 3.8

Sous l'hypothèse d'allocation proportionnelle, l'estimateur stratifié $\widehat{\delta}_n(p_1,\dots,p_K)$ est plus efficace que l'estimateur de Monte Carlo classique, *i.e.*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(p_1,\dots,p_K)\right] = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2 \leq \mathbb{V}\mathrm{ar}\left[\overline{h}_n\right].$$

Interprétation de l'estimateur stratifié à allocation proportionnelle. On peut toujours écrire

$$n\mathbb{V}\mathrm{ar}\left[\overline{h}_n\right] = n\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n(p_1,\dots,p_K)\right] + \sum_{k=1}^{K} p_k\left(\mu_k - \sum_{j=1}^{K} p_j\mu_j\right)^2.$$

Le second terme correspond à la variance inter-strates. Ainsi l'estimateur à allocation proportionnelle élimine la variabilité inter-strates pour ne conserver que la variabilité intra-strates. Il réduit donc d'autant plus la variance que la variance intra-strates est faible. Cela renseigne sur la stratégie à adopter pour choisir les strates, *i.e.* prendre des strates pour lesquelles la quantité d'intérêt varie peu au sein de chaque strate.

Exemple 3.7. Retour sur le modèle de Black-Scholes

Pour le calcul d'une option d'achat (*c.f.* Exemple 3.4), il est possible de calculer un estimateur stratifié de $\mathscr{I}$ en prenant $Z = X$ pour variable de stratification et $D_k =]d_k, d_k+1]$, $k = 1,\dots,K$. La méthode consiste à

(i) choisir une stratification $d_1,\dots,d_K$ et une allocation $n_1,\dots,n_K$;

(ii) à simuler, pour chaque strate $k \in \{1,\dots,K\}$, une suite $\left(X_n^{(k)}\right)_{n\geq 1}$ de variables aléatoires *i.i.d.* suivant la loi conditionnelle $X \mid X \in]d_k, d_k+1]$;

(iii) calculer $\widehat{\delta}_n(n_1,\dots,n_K)$, *i.e.*, la somme pondérée des moyennes,

$$\widehat{\delta}_n(n_1,\dots,n_K) = \sum_{k=1}^{K}\{\Phi(d_{k+1}) - \Phi(d_k)\}\frac{1}{n_k}\sum_{i=1}^{n_k} h\left(X_i^{(k)}\right).$$

Pour l'étape (ii), on peut simuler une suite $\left(U_n^{(k)}\right)_{n\geq 1}$ de variables aléatoires *i.i.d.* suivant la loi $\mathscr{U}([0,1])$ et poser pour tout $i \geq 1$,

$$X_i^{(k)} = \Phi^{-1}\left[\Phi(d_k) + U_i^{(k)}\{\Phi(d_{k+1}) - \Phi(d_k)\}\right].$$

Cela ne fonctionne néanmoins que pour des événement $X \in]d_k, d_k+1]$ où l'on connait précisément $\Phi(d_{k+1})$ et $\Phi(d_k)$. Dans le cas contraire, il est préférable de procéder par rejet.

Conclusion. Bien que l'on puisse toujours définir un estimateur stratifié de variance plus faible que celle de l'estimateur classique (allocation proportionnelle), il existe plusieurs difficultés en pratique.

Notamment le choix de strates et d'une variable de stratification pour lesquelles on connaît explicitement $\mathbb{P}[\mathbf{Z} \in D_k]$. Le choix des strates n'est pas sans conséquence sur la réduction de variance (des techniques adaptatives efficaces existent pour choisir) mais également sur le coût de calcul de l'estimateur : le temps de simulation pouvant dépendre de la strate considérée et l'architecture de l'algorithme est généralement très différente des autres méthodes vues dans ce chapitre. Il est donc important d'étudier l'efficacité relative de l'estimateur stratifié par rapport à la méthode de Monte Carlo classique.

Références utiles

[1] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer Texts in Statistics. Springer-Verlag New York, 2 edition, 2004. doi : 10.1007/978-1-4757-4145-2.

[2] B. Ycart. *Modèles et Algorithmes Markoviens*, volume 39 of *Mathématiques et Applications*. Springer-Verlag Berlin Heidelberg, 2002.