Méthodes de Monte Carlo – Projet

stoehr@ceremade.dauphine.fr

- **À rendre avant le 28 décembre 2020. Chaque jour de retard sera pénalisé d'un point.**
- R **est le seul langage autorisé.** Les questions nécessitant un code R sont indiquées par le symbole ♠.
- Le rapport (**nom du fichier :** `numero_groupe_rapport_noms`)
  - à rendre au format **.pdf** et doit contenir vos réponses et commentaires. Une rédaction soignée est attendue. Il est important de justifier/commenter les résultats théoriques et numériques
  - Pour intégrer tout ou partie de votre code et des sorties dans votre rapport, vous pouvez utiliser les outils dédiés : Notebook, Rmarkdown ou LaTeX+ knitr. En revanche, **il est interdit de copier-coller du code brut dans le corps du texte**.
  - Les graphiques doivent être soigneusement annotés et présentés (titre, couleur, légendes, ...).
- Une version du code pouvant être testée doit être fournie (**même nom de fichier**). Ce code doit
  - s'exécuter sans erreurs et permettre de reproduire l'intégralité des résultats présentés dans le rapport. Vous préciserez la graine utilisée pour les résultats obtenus.
  - être bien commenté. Il est possible qu'une explication orale vous soit demandée.
  - utiliser autant que possible les spécificités du language (bonus pour les codes les plus efficaces).

Exercice 1. Soit f une densité de $\mathbb{R}^2$ définie pour $(x,y) \in \mathbb{R}^2$ par $f(x,y) = a\psi(x,y)$ avec $a \in \mathbb{R}_+^*$ et

$$\psi(x,y) = \left[\left|\sin\left(\frac{2}{\pi}x^2 - \frac{\pi}{4}\right)\right| + 4\cos(x)^2 + y^4\right] e^{-2(x+|y|)} \mathbb{1}_{\{x\in[-\pi/2,\pi/2]\}} \mathbb{1}_{\{y\in[-1,1]\}}.$$

Pour (X,Y) de densité f, l'objectif est d'estimer f_X la densité marginale de X.

Contrainte. Les générateurs `runif` et `rexp` peuvent être utilisés directement. Les autres générateurs de variables aléatoires doivent être démontrés et codés en conséquence.

Simulation suivant la densité f

1. Montrer que pour simuler suivant f, il n'est pas nécessaire de connaître a et il suffit de trouver une constante $m \in \mathbb{R}_+^*$ et une densité g pour laquelle on dispose d'un générateur aléatoire telles que

$$\forall (x,y) \in \mathbb{R}^2, \quad \psi(x,y) \le mg(x,y). \tag{1}$$

Trouver alors m et g qui satisfont (1).

Dans la suite, on désigne par ratio d'acceptation, la fonction définie pour $(x,y) \in \mathrm{supp}(g)$ par

$$\rho(x,y) = \frac{\psi(x,y)}{mg(x,y)}.$$

2. *(♠)* Coder les fonctions `rgen_g(n)` qui simule n réalisations suivant la densité g et `rgen_f` qui retourne n réalisations suivant la densité f ainsi que toutes les valeurs du ratio d'acceptation utilisées pour obtenir ces réalisations.

3. *(♠)* Simuler un échantillon z de taille $n = 10000$ suivant f à l'aide de la fonction `rgen_f`. Auto-évaluer votre solution à l'aide du tableau suivant. n_t désigne le nombre moyen de simulations suivant g pour différents choix de g possibles. n_ℓ est le nombre moyen de passages dans une boucle `for` ou `while` pour le code utilisé pour générer les réalisations de f.

g	$\star$		$\star\star$		$\star\star\star$		$\star\star\star\star$		$\star\star\star\star\star$	
Code	n_t	n_ℓ	n_t	n_ℓ	n_t	n_ℓ	n_t	n_ℓ	n_t	n_ℓ
$\star$		527000		263000		84000		42000		36000
$\star\star$	527000	503	263000	248	84000	75	42000	36	36000	30
$\star\star\star$		3		3		4		4		7

Méthode n°1 – Estimation de a

4. **(a)** Construire un estimateur de a en fonction de ρ, noté $\widehat{b}_n$. Montrer qu'il est biaisé et converge presque sûrement. En déduire un intervalle de confiance asymptotique de a au niveau $1-\alpha$ calculable en pratique.

(b) *(♠)* À l'aide des variables aléatoires simulées question **3.**, évaluer $\widehat{b}_n$ et l'intervalle de confiance au niveau 95%.

(c) *(♠)* Proposer une méthode d'estimation du biais ne nécessitant pas de simulations supplémentaires suivant f ou g.

5. **(a)** Montrer que l'algorithme de simulation suivant f fournit un autre estimateur de a, noté $\widehat{a}_n$, qui converge presque sûrement mais qui est sans biais. En déduire un intervalle de confiance asymptotique de a au niveau $1-\alpha$ calculable en pratique.

(b) *(♠)* À l'aide de l'échantillon z, évaluer $\widehat{a}_n$ et l'intervalle de confiance au niveau 95%.

6. *(♠)* Exprimer le rapport des coûts pour lesquels $\widehat{b}_n$ et $\widehat{a}_n$ atteignent la même précision. Évaluer le à l'aide des résultats précédents. Quel est l'estimateur le plus efficace ?

7. **(a)** Pour $x \in [-\pi/2, \pi/2]$ donner un estimateur $\widehat{f}_{X,n}(x)$ de $f_X(x)$ à l'aide de $\widehat{a}_n$.

(b) *(♠)* Comparer graphiquement la distribution marginale de l'échantillon z à l'estimateur $\widehat{f}_{X,n}(x)$.

Méthode n°2 – Estimateur ponctuel

8. Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ une suite de variables indépendantes suivant la loi jointes $f(x, y)$ et $w(\cdot)$ une densité quelconque telle que $\mathrm{supp}(w) \subseteq \mathrm{supp}(f_X)$. Montrer que

$$\widehat{\omega}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \frac{\psi(x, Y_k)\, w(X_k)}{\psi(X_k, Y_k)} \xrightarrow[n\to+\infty]{p.s.} f_X(x).$$

En déduire un intervalle de confiance asymptotique de $f_X(x)$ au niveau $1-\alpha$ calculable en pratique.

9. Pour quel choix de w obtient-on l'estimateur de variance minimale ? Commenter ce résultat et expliquer comment l'utiliser en pratique.

10. *(♠)* À l'aide de l'échantillon z, évaluer $\widehat{\omega}_n(-1)$ et l'intervalle de confiance au niveau 95%.

11. *(♠)* Exprimer le rapport des coûts pour lesquels $\widehat{\omega}_n(-1)$ et $\widehat{f}_{X,n}(-1)$ atteignent la même erreur quadratique moyenne. Évaluer le à l'aide des résultats précédents. Quel est l'estimateur le plus efficace ?

Exercice 2. Soit $\mathbf{X} = (X_1, X_2, X_3)$ un vecteur aléatoire de $\mathbb{R}^3$ distribué suivant la loi $\mathcal{N}(\mu, \Sigma)$ avec

$$\mu = \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} \quad \text{et} \quad \Sigma = \begin{pmatrix} 0.047 & 0 & 0.0117 \\ 0 & 0.047 & 0 \\ 0.0117 & 0 & 0.047 \end{pmatrix}.$$

On s'intéresse à

$$\delta = \mathbb{E}\left[\min\left(3, \frac{1}{3}\sum_{k=1}^{3} e^{-X_k}\right)\right].$$

Contrainte. Le générateur `rnorm` peut être utilisé directement. Les autres générateurs de variables aléatoires doivent être codés en conséquence.

1. *(♠)* Écrire une fonction `rmvnorm(`n`, mu, sigma)` qui permet de générer n réalisation de la loi normale multivariée de moyenne `mu` et de matrice de variance-covariance `sigma`. Simuler à l'aide de cette fonction un échantillon $\mathbf{x}$ de taille $n = 10000$ suivant la loi de $\mathbf{X}$.

2. **(a)** Étant donné une ensemble de variables aléatoires $\mathbf{X}_i = (X_{1,i}, X_{2,i}, X_{3,i})$, $i = 1, \dots, n$, *i.i.d.* suivant la loi de $\mathbf{X}$, donner l'expression de l'estimateur de Monte Carlo de δ, noté $\overline{\delta}_n$.

(b) *(♠)* Pour l'échantillon $\mathbf{x}$, évaluer $\overline{\delta}_n$ et l'erreur quadratique moyenne associée.

3. **(a)** Montrer qu'il existe une transformation mesurable A qui laisse la loi $\mathcal{N}(\mu, \Sigma)$ invariante et telle que pour l'estimateur de δ par la méthode de la variable antithétique, noté $\widehat{\delta}_n$, $\mathbb{V}\mathrm{ar}\left[\widehat{\delta}_n\right] \leq \mathbb{V}\mathrm{ar}\left[\overline{\delta}_n\right]/2$. Exprimer le facteur de réduction de variance théorique, noté R_1, de $\widehat{\delta}_n$ par rapport à $\overline{\delta}_n$.

(b) *(♠)* Pour l'échantillon $\mathbf{x}$, évaluer $\widehat{\delta}_n$, l'erreur quadratique moyenne associée et R_1. Qu'en concluez vous ?

4. **(a)** *(♠)* En utilisant des moments d'ordre 1 et/ou d'ordre 2 associés à la loi de $\mathbf{X}$, trouver une fonction h_0, telle que la corrélation entre $h_0(\mathbf{X})$ et $\min(3, \sum_{k=1}^{3} e^{-X_k}/3)$ soit supérieure à 0.5. En déduire, pour $b \in \mathbb{R}$, l'expression de l'estimateur par la méthode de la variable de contrôle simple, noté $\widehat{\delta}_n(b)$.

(b) *(♠)* Pour l'échantillon $\mathbf{x}$ et une valeur de b judicieusement choisie, évaluer $\widehat{\delta}_n(b)$ et l'erreur quadratique moyenne associée. Discuter le résultat obtenu en fonction du nombre global de simulations effectuées et du nombre de simulations utilisées pour le calcul de $\widehat{\delta}_n(b)$.

Exercice 3. On suppose Y est distribué suivant la loi géométrique $\mathcal{G}(p)$, *i.e.*, pour $k \in \mathbb{N}^*$, $\mathbb{P}[Y = k] = p(1-p)^{k-1}$. Pour $(X_n)_{n\geq 1}$ de variables aléatoires *i.i.d.* suivant la loi gamma $\Gamma(m, \theta)$, on s'intéresse à

$$\delta = \mathbb{E}[S], \quad \text{avec} \quad S = \sum_{i=1}^{Y} \log(X_i + 1).$$

On prendra $p = 0.2$, $m = 2$ et $\theta = 2$.

1. *(♠)* Pour $n = 10000$ tirages, donner une estimation de δ par la méthode de Monte Carlo classique et de l'erreur quadratique moyenne associée.

2. **(a)** Proposer un ensemble de strates $D_1, \dots, D_L$, $L \in \mathbb{N}^*$. En déduire un estimateur de δ par la méthode de stratification avec allocation proportionnelle $(n_1, \dots, n_L)$. On le notera $\widehat{\delta}_n(n_1, \dots, n_L)$.

(b) *($\spadesuit$)* Évaluer $\widehat{\delta}_n(n_1,\ldots,n_L)$ pour $n = 10000$ tirages et $L = 15$ strates. Donner l'erreur quadratique moyenne associée. Quelle est l'efficacité relative $\widehat{\delta}_n(n_1,\ldots,n_L)$ par rapport à la méthode de Monte Carlo classique ? Discuter de façon concise les résultats obtenus.

Auto-évaluation du code. Évaluer votre code à l'aide des critères suivants :

- Nombre de déclarations du type « `c()` »

<table>
<tr><th>Code</th><th>Ex. n°1</th><th>Ex. n°2</th><th>Ex. n°3</th></tr>
<tr><td>$\star$</td><td>≥ 3</td><td rowspan="2">≥ 1</td><td rowspan="2">≥ 1</td></tr>
<tr><td>$\star\star$</td><td rowspan="2">≤ 2</td></tr>
<tr><td>$\star\star\star$</td><td>0</td><td>0</td></tr>
</table>

- Nombre de boucles `for` ou `while` utilisées

Code	Ex. n°1	Ex. n°2	Ex. n°3
$\star$	≥ 3	≥ 2	≥ 4
$\star\star$	2	1	≤ 3
$\star\star\star$	1	0	0

- Nombre de boucles conditionnelles `if` utilisées

<table>
<tr><th>Code</th><th>Ex. n°1</th><th>Ex. n°2</th><th>Ex. n°3</th></tr>
<tr><td>$\star$</td><td>≥ 1</td><td>≥ 1</td><td>≥ 1</td></tr>
<tr><td>$\star\star$</td><td rowspan="2">0</td><td rowspan="2">0</td><td rowspan="2">0</td></tr>
<tr><td>$\star\star\star$</td></tr>
</table>

- Nombre d'appel à `Vectorize` (`sapply`)

<table>
<tr><th>Code</th><th>Ex. n°1</th><th>Ex. n°2</th><th>Ex. n°3</th></tr>
<tr><td>$\star$</td><td>≥ 1 (0 ou ≥ 2)</td><td>≥ 1 (≥ 1)</td><td>≥ 1 (0)</td></tr>
<tr><td>$\star\star$</td><td rowspan="2">0 (1)</td><td rowspan="2">0 (0)</td><td>0 (3)</td></tr>
<tr><td>$\star\star\star$</td><td>0 (4)</td></tr>
</table>