IFT3655 - Devoir 7

Mathias La Rochelle

Le samedi 16 novembre 2024

Question 1

On a trois variables aléatoires indépendantes X_1, X_2, et X_3, et on veut estimer $p = \mathbb{P}[Y > b]$ où $Y = X_1 + X_2 + X_3$ et b est une constante. On sait que p est très petit et on veut l'estimer par simulation en utilisant l'importance sampling (IS). Pour ramener la variance à 0 dans ce cas-ci, il faudrait générer le vecteur (X_1, X_2, X_3) selon sa loi conditionnelle à $Y > b$. L'estimateur serait alors égal au rapport de vraisemblance L, qui serait une constante. Mais cela est difficile et compliqué à implanter, car sous cette loi conditionnelle, les variables X_1, X_2, X_3 ne sont plus indépendantes.

Une stratégie heuristique plus facile à implanter pourrait être la suivante. Supposons que X_j a la densité π_j pour chaque j. On va remplacer π_j par une autre densité g_j pour $j = 1$ et 2 , on va générer X_1 et X_2 selon cette densité g_j, et finalement on va calculer $X_{\text{is}} = (1 - F_3(b - X_2 - X_1))\, L$ comme estimateur, où $L = (\pi_1(X_1)/g_1(X_1))\,(\pi_2(X_2)/g_2(X_2))$ et $F_3(x) = \mathbb{P}(X_3 \le x)$. Cela combine IS avec CMC.

a) Montrez que $\mathbb{E}_g[X_{\text{is}}] = p$.

$$
\begin{aligned}
\mathbb{E}_g[X_{\text{is}}] &= \mathbb{E}_g[(1 - F_3(b - X_2 - X_1))\, L] \\
&= \mathbb{E}_g[(1 - \mathbb{P}[X_3 \le b - X_2 - X_1])\, L] \\
&= \mathbb{E}_g[(1 - \mathbb{P}[X_1 + X_2 + X_3 \le b])\, L] \\
&= \mathbb{E}_g[(1 - \mathbb{P}[Y \le b])\, L] \\
&= \mathbb{E}_g[\mathbb{P}[Y > b] L] \\
&= \mathbb{E}_g[pL] \\
&= p \iint \underline{\pi_1(X_1)\pi_2(X_2)} \\
&= g_1(X_1) g_2(X_2) g(X_1, X_2)\; dX_1\, dX_2 \\
&= p \iint \pi_1(X_1)\pi_2(X_2)\; dX_1\, dX_2 \\
&= p \int \pi_1(X_1)\; dX_1 \cdot \int \pi_2(X_2)\; dX_2 \\
&= p
\end{aligned}
$$

b) Supposons maintenant que chaque X_j suit une loi exponentielle de moyenne $1/\lambda = 2$ et que $b = 15$. Comme on a $\mathbb{E}[Y] = 6$ et $(Var[Y])^{1/2} = \sqrt{12}$, on sait que $Y > 15$ ne se produira pas souvent. Pour approximer la loi de (X_1, X_2, X_3) conditionnelle à $Y > 15$, on pourrait prendre g_j comme une exponentielle de moyenne 5 au lieu 2. L'idée (heuristique) est que sous cette loi conditionnelle, on s'attend à ce que les X_j soient autour de 5 ou un peu plus pour que leur somme dépasse 15. On peut alors les générer avec une moyenne de 5 au lieu de 2. Cette valeur n'est certainement pas optimale, mais c'est un choix simple et raisonnable. Trouvez la formule pour l'estimateur X_{is} correspondant, implantez-le, et effectuez une expérience qui génére n répétitions indépendantes de la simulation pour $n = 1000$, d'abord avec MC sans utiliser IS, puis en utilisant IS. Dans les deux cas, calculez l'estimateur p avec un intervalle de confiance à 95%, calculez aussi les variances, et comparez.

Nous avons les densités π_j et g_j suivantes :

$$\pi_j = \lambda \exp\left[-\lambda x_j\right]$$
$$g_j = \lambda_b \exp\left[-\lambda_b x_j\right]$$

où $\lambda = \frac{1}{2}$ et $\lambda_b = \frac{1}{5}$. Ainsi, la fonction de répartition pour X_3 est :

$$F_3\left(b - X_2 - X_1\right) = 1 - \exp\left[-\lambda\left(b - x_2 - x_1\right)\right]$$

À partir de l'équation définit dans l'énoncé, on retrouve la formule de notre estimateur X_{is} :

$$\begin{aligned} X_{\text{is}} &= \begin{cases} \dfrac{\lambda^2}{\lambda_b^2}\dfrac{\exp\left[\lambda_b(x_1+x_2)\right]}{\exp[b\lambda]} \\ \dfrac{\lambda^2}{\lambda_b^2}\dfrac{\exp\left[\lambda_b(x_1+x_2)\right]}{\exp\left[\lambda(x_1+x_2)\right]} \end{cases} \\ &= \begin{cases} \dfrac{25}{4}\dfrac{\exp\left[\frac{1}{5}(x_1+x_2)\right]}{\exp\left[\frac{15}{2}\right]} \\ \dfrac{25}{4}\dfrac{\exp\left[\frac{1}{5}(x_1+x_2)\right]}{\exp\left[\frac{1}{2}(x_1+x_2)\right]} \end{cases} \\ &= \begin{cases} \dfrac{25\exp\left[\frac{1}{5}(x_1+x_2)-\frac{15}{2}\right]}{4}, & \text{si } x_1+x_2 \le b \\ \dfrac{25}{4\exp\left[-\frac{3}{10}(x_1+x_2)\right]}, & \text{autrement} \end{cases} \end{aligned}$$

Le code se retrouve dans le fichier `NumeroUn`. Avec MC sans IS, j'obtiens un estimateur p de 0.0170, une variance correspondante de 0.0167 et un intervalle de confiance de [0.0090, 0.0250]. Cet intervalle de confiance relativement large met en évidence une incertitude importante dans les résultats. Bien que cette méthode soit facile à implémenter, elle produit des estimations trop

dispersées pour être fiable dans un contexte où l'objectif est de réaliser un échantillonnage précis. Quant à l'implémentation avec IS, j'obtiens un estimateur p de 0.0204, une variance correspondante de 0.0003 et un intervalle de confiance de $[0.0193, 0.0214]$. Par rapport à la méthode précédente, cet intervalle de confiance est bien plus étroit ce qui démontre qu'il y a une augmentation de la précision des valeurs échantillonnées. Ce qui est tout à fait normale puique le importance sampling se concentre sur les évènements les plus susceptibles de se produirent.

c) Question bonus (+5): Au lieu de prendre g_j exponentielle de moyenne $1/\lambda_0 = 5$, on pourrait essayer d'optimiser $1/\lambda_0$ en cherchant la valeur qui minimise la variance, par exemple dans l'intervelle $[4, 10]$, un peu comme dans l'exemple de la diapo 62. Pouvez-vous approximer le $1/\lambda_0$ optimal?

Question 2

On considère le modèle de déflexion d'une poutre en porte-à-faux vu à la page 75 des diapos. On a

$$X = X(\sigma_1) = h(Y_1, Y_2, Y_3) = \frac{\kappa}{Y_1}\sqrt{\frac{Y_2^2}{w^4} + \frac{Y_3^2}{t^4}}$$

où Y_1, Y_2, Y_3 sont normales indépendantes, $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Prenons $w = 4, t = 2$, et $\kappa = 5\times10^5$ pour les constantes et $\mu_1 = 2.9\times10^7, \sigma_1 = 1.45\times10^6, \mu_2 = 500, \sigma_2 = 100$, $\mu_3 = 1000, \sigma_3 = 100$, pour les paramètres des lois normales.

a) Supposons que l'on veut estimer $\mu = \mathbb{P}[X \leq x]$ pour $x = 5$. Vous allez faire cela avec MC ordinaire, puis avec CMC en conditionnant sur $\mathcal{G} = \{Y_2, Y_3\}$ comme sur les diapos. Cela donne les estimateurs I et J sur les diapos. Dans les deux cas, faites $n = 1000$ simulations, estimez μ, la variance de votre estimateur, et calculez un intervalle de confiance à 95% sur μ. Comparez MC vs CMC.

Avec MC, j'obtiens un estimateur $\mu = 0.898$, une variance $\mathbb{V}[\mu] = 0.092$ et un intervalle de confiance $[0.879,\ 0.917]$. Pour CMC, j'obtiens un estimateur $\mu = 0.904$, une variance $\mathbb{V}[\mu] = 0.039$ et un intervalle de confiance $[0.892,\ 0.917]$. Je constate qu'avec CMC, il y a une diminution de la variance et donc un resserement de l'intervalle de confiance. Il n'est pas anormal d'observer une telle chose car cette méthode fixe une variable afin de réduire la variance qu'elle pourrait générée. C'est exactement ce qui se passe dans notre cas, où la variable X_3 est fixée. Cet exemple démontre très bien pourquoi CMC est une méthode d'échantillonage qui est préférée lorsque l'objectif est d'améliorer la précision. D'ailleurs, les estimateurs sont proches ce qui annoncent que les deux approches convergent vers une estimation correcte de la moyenne.

b) Supposons maintenant que l'on fait varier le paramètre σ_3 et que l'on veut estimer la dérivée de μ par rapport à σ_3. On peut noter $\mu = \mu(\sigma_3), X = X(\sigma_3), I = I(\sigma_3)$ et $J = J(\sigma_3)$ pour indiquer la dépendance en σ_3. On veut donc estimer $\mu'(\sigma_3) = d\mu(\sigma_3)/d\sigma_3$ au point $\sigma_3 = 100$. Essayez cela avec $n = 1000$ répétitions avec chacune des méthodes suivantes:

1. Différences finies avec MC et des variables aléatoires indépendantes (IRN), avec $\delta = 1$;
2. Différences finies avec CMC et des variables aléatoires indépendantes (IRN), avec $\delta = 1$;
3. Différences finies avec CMC et des variables aléatoires communes (CRN), avec $\delta = 1$;
4. Dérivée stochastique avec CMC;

Dans chaque cas, expliquez comment vous faites, calculez un intervalle de confiance à 95% pour la dérivée et donnez une estimation de la variance de votre estimateur de dérivée. Comparez les variances. Pour le cas de la dérivée stochastique, il faut montrer comment on trouve l'estimateur. On ne demande pas de "prouver" qu'il est sans biais, mais vous aurez des points supplémentaires si vous le faites (correctement), et vous pouvez aussi le comparer avec celui utilisé en (3).

Le code pour l'ensemble de ces questions se retrouvent dans le fichier `NumeroDeux` et les méthodes respectives sont `estimerDeriveeIRNMC`, `estimerDeriveeIRNCMC`, `estimerDeriveeCRNCMC` et `estimerDeriveeStochastique`.

1. Pour implémenter cette méthode, je vais créer deux distributions gaussiennes ; l'une avec $\sigma_3 = 100$ et l'autre avec $\sigma_3 = 101$. Je génére par inversion une valeur de ces distributions puis je calcule la dérivée. J'effectue ce processus 1000 fois et je calcule finalement la moyenne, la variance et l'intervalle de confiance de ces dérivées. À noter que les valeurs utilisées lors de l'inversion proviennent du même stream, c'est-à-dire que les valeurs générées par `RandomStream` proviennent de la même instance. De cette façon, je respecte bien la condition que les variables aléatoires Y_1, Y_2 et Y_3 sont indépendantes. Voici les valeurs obtenues : moyenne de -0.0070, variance de 0.1570 et intervalle de confiance de $[-0.0316,\ 0.0176]$.

2. L'implémentation de cette méthode ne diffère pas de beaucoup que la précédente. Ici, la seule chose qui change est que je fais appel à une méthode auxiliaire différente, soit celle qui effectue CMC. Voici les valeurs obtenues : moyenne de 0.0072, variance de 0.0800 et un intervalle de confiance de $[-0.0104,\ 0.0247]$.

3. Pour implémenter CMC mais désormais avec des variables communes (CRN), il faut que les valeurs utilisées durant l'inversion (générées par `MRG32k3a`) proviennt d'un stream différent. Si on crée deux `RandomStream` qui part du même germe alors à chaque itération, la méthode `nextDouble()` générera les mêmes valeurs contrairement à IRN où on utilisait un même stream, en d'autres mots, les valeurs passées par inversion étaient différentes pour Y_1, Y_2 et Y_3. Voici les valeurs obtenues après 1000 itérations : moyenne de -0.0017, variance de 0.000010 et un intervalle de confiance de $[-0.0019, -0.0015]$.

Pour la dérivée stochastique, le processus pour estimer la valeur de σ_3 débute en dérivant la probabilité conditionnelle $J = \mathbb{P}[X \leq x \mid Y_2, Y_3]$ selon σ_3. Voici à quoi cela ressemble :

$$\begin{aligned}\frac{\mathrm{d}J(\sigma_3)}{\mathrm{d}\sigma_3} &= \frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[1 - \Phi\left(\frac{W_1(x) - \mu_1}{\sigma_1}\right)\right] \\ &= \Phi\left(\frac{W_1(x) - \mu_1}{\sigma_1}\right) \cdot \frac{1}{\sigma_1}\frac{\mathrm{d}}{\mathrm{d}\sigma_3}[W_1(x)]\end{aligned}$$

Pour continuer, nous allons calculer la dérivée de $W_1(x)$ selon σ_3 :

$$\begin{aligned}
\frac{\mathrm{d}W_1(x)}{\mathrm{d}\sigma_3} &= \frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\frac{\kappa}{x}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\,\right] \qquad (1)\\
&= \frac{\kappa}{2x}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\frac{Y_2^2}{w^4}+\frac{(\mu_3+\sigma_3\Phi^{-1}(U_3))^2}{t^4}\right]\\
&= \frac{\kappa}{2x}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\frac{(\mu_3+\sigma_3\Phi^{-1}(U_3))^2}{t^4}\right]\\
&= \frac{\kappa}{2x}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\frac{\left(\mu_3+\sigma_3\left(\mu_3+\sigma_3\sqrt{2}\mathrm{erf}^{-1}(2U_3-1)\right)\right)^2}{t^4}\right]\\
&= \frac{\kappa}{2xt^4}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\left(\sigma_3\mu_3+\sigma_3^2\underbrace{\sqrt{2}\mathrm{erf}^{-1}(2U_3-1)}_{c}\right)^2\right]\\
&= \frac{\kappa}{2xt^4}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\frac{\mathrm{d}}{\mathrm{d}\sigma_3}\left[\sigma_3^2\mu_3^2+2\sigma_3^3\mu_3 c+\sigma_3^4c^2\right]\\
&= \frac{\kappa}{2xt^4}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\cdot\left(2\sigma_3\mu_3^2+6\sigma_3^2\mu_3 c+4\sigma_3^3c^2\right)
\end{aligned}$$

Dans (1), $Y_3 = \mu_3 + \sigma_3\Phi^{-1}(U_3)$ à cause de $\Phi\left(\frac{Y_3-\mu_3}{\sigma_3}\right) = U_3$. Donc, pour estimer $\mu'(\sigma_3)$, on utiliserait l'équation suivante :

$$\mu'(\sigma_3) = \frac{\Phi\left(\frac{W_1(x)-\mu_1}{\sigma_1}\right)\left(\frac{\kappa}{2xt^4}\sqrt{\frac{Y_2^2}{w^4}+\frac{Y_3^2}{t^4}}\left(2\sigma_3\mu_3^2+6\sigma_3^2\mu_3 c+4\sigma_3^3c^2\right)\right)}{\sigma_1} \qquad (2)$$

où $c = \sqrt{2}\mathrm{erf}^{-1}(2U_3-1)$ avec erf^{-1} étant l'inverse de la fonction d'erreur. Voici les valeurs obtenues lors de cette estimation : moyenne de -0.0025, variance de 0.000051 et un intervalle de confiance de $[-0.0029, -0.0021]$. Très proche des valeurs obtenues avec CMC CRN.

Au final, on constate qu'à chaque nouvelle méthode, la précision de l'échantillonnage continue à augmenter. Cela est particulièrement visible à travers la réduction progressive des variances et le rétrécissement des intervalles de confiance. En effet, en passant des méthodes classiques d'inversion à celles intégrant des techniques telles que le contrôle des variables communes (CRN) ou la pondération explicite via la fonction X, les estimations deviennent de plus en plus fiables.