

Quelques résultats en Importance Sampling pour des sommes de variables i.i.d.

Virgile Caron

5 février 2013

Résumé

Dans cet article, nous présentons certaines méthodes d'Importance Sampling dans le cadre de l'estimation d'événements rares qui s'écrivent sous la forme de sommes de variables aléatoires réelles i.i.d. dépassant un niveau. Nous rappelons les méthodes classiques de l'Importance Sampling à états indépendants basés sur des résultats de grandes déviations, puis nous présentons une méthode d'approximation de la densité conditionnelle optimale. Nous traitons en détails un exemple permettant de comprendre les avantages et inconvénients de ces méthodes. Les notions clés sont présentées de telle sorte qu'elles restent accessibles aux lecteurs débutants. Pour approfondir les notions, le lecteur intéressé pourra se reporter aux nombreuses références.

1 Introduction

1.1 Contexte

L'Importance Sampling (ou échantillonnage préférentiel, noté IS dans la suite) est une méthode de Monte-Carlo qui permet de réduire la variance dans le calcul d'une intégrale, ou d'une probabilité d'un événement rare. Plus précisément, l'IS permet de réduire la variance (ou l'erreur relative associée) par rapport à la méthode de Monte-Carlo pour un nombre d'échantillons fixés. Ou inversement, de réduire le nombre d'échantillons pour une variance fixée. Les applications des méthodes d'IS sont nombreuses : dans les modèles de files d'attentes, il est important d'estimer les possibles pertes de données causés par un overflow (voir [19], [44] ou [55]); pour une compagnie d'assurance, il est important d'estimer la probabilité de ruine (voir [1], [2] ou [3]); en théorie du risque, il faut estimer la probabilité que le système s'effondre (voir [42]).

Cet article ne se veut pas un état de l'art exhaustif de toutes les méthodes d'IS. Le but est de donner quelques idées importantes sur l'estimation des événements rares définis comme le dépassement d'un niveau pour des sommes de variables aléatoires indépendantes et identiquement distribuées. Nous proposons deux grandes types de méthodes, chacune relié à des résultats fondamentaux des statistiques. De nombreuses références seront données dans chaque section, permettant d'approfondir toutes les notions développées dans cet article.

1.2 Méthode de Monte-Carlo naïve

Pour démontrer la difficulté qu'implique l'estimation d'une probabilité d'un événement rare, étudions un exemple simple. Soit \mathbf{X} une variable aléatoire réelle de densité p par rapport à la mesure de Lebesgue et notons

$$P_1 := P[\mathbf{X} \in A] = E_p[\mathbb{1}_A(\mathbf{X})]$$

où A est un borélien de \mathbb{R} . E_p indique que l'espérance est prise sous p et $\mathbb{1}_A(\cdot)$ est la fonction indicatrice

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon} \end{cases}$$

On veut estimer P_1 par simulation, en utilisant la méthode la plus simple à notre disposition, la méthode de Monte-Carlo standard (ou naïve). On simule un échantillon $(X(1), \dots, X(L))$ de taille L , de densité p et on fabrique l'estimateur suivant :

$$\widehat{P}_1 = \frac{1}{L} \sum_{l=1}^L \mathbb{1}_A(X(l))$$

L'espérance de \widehat{P}_1 est P_1 tandis que sa variance est $P_1(1 - P_1)/L$. Par le théorème central limite, on peut fabriquer un intervalle de confiance à $100(1 - \alpha)\%$ pour P_1 de la forme

$$[\widehat{P}_1 - z_{\alpha/2} \sqrt{\frac{P_1(1 - P_1)}{L}}, \widehat{P}_1 + z_{\alpha/2} \sqrt{\frac{P_1(1 - P_1)}{L}}]$$

où $z_{\alpha/2}$ est le quantile défini par $P[Z \geq z_{\alpha/2}] = \alpha/2$ avec Z qui suit la loi normale centrée réduite.

On veut estimer P_1 à 10% près (deux chiffres significatifs), c'est à dire que l'on veut que la demi-largeur relative de l'intervalle de confiance pour P_1 soit plus petite que 0.1. Ainsi, pour un $\alpha = 0.01$, on a

$$2.576 \sqrt{\frac{P_1(1 - P_1)}{L}} \frac{1}{\widehat{P}_1} \leq 0.1$$

Nous voulons déterminer pour quelle valeur de L , cette dernière égalité est valide. Par la LGN, $\widehat{P}_1 \rightarrow P_1$ presque sûrement quand n tends vers l'infini, donc

$$L = \frac{100 * 2.576^2 (1 - P_1)}{P_1}$$

et la taille de l'échantillon est proportionnel à $1/P_1$.

Plus P_1 est petit, plus L doit être grand. Ainsi si $P_1 = 10^{-6}$, alors $L = 6.64 \times 10^8$; si $P_1 = 10^{-9}$, alors $L = 6.64 \times 10^{11}$. Clairement, si P_1 est très petit, la taille de l'échantillon requis pour obtenir une variance fixée, est trop grand pour envisager d'appliquer cette méthode. Il faut envisager une méthode alternative qui permet de réduire ce nombre d'échantillons à un nombre raisonnable sans sacrifier la variance. C'est le cas de l'IS.

1.3 Importance Sampling

L'IS a été introduite par [43] et consiste à changer de densité d'échantillonnage pour réduire le nombre de réplifications dans le calcul d'une probabilité. Considérons un cadre assez général. Soit $\mathbf{X}_1^n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ n variables aléatoires réelles de densité $f_n(x_1, \dots, x_n)$. On veut estimer

$$P_n := P[\mathbf{X}_1^n \in \mathcal{E}_n] \quad (1)$$

où \mathcal{E}_n est un ensemble mesurable de \mathbb{R}^n . L'IS consiste à modifier la loi d'échantillonnage. Au lieu de simuler l'échantillon sous la vraie loi f_n , on fabrique l'échantillon sous une loi alternative $f_n^*(x_1, \dots, x_n)$ (sous la seule condition que la restriction de f_n à l'ensemble \mathcal{E}_n soit absolument continue par rapport à f_n^*) en pondérant, par le rapport de vraisemblance entre la densité de départ et la nouvelle densité, calculé sur les chemins simulés. On fabrique alors l'estimateur suivant

$$\widehat{P}_n = \frac{1}{L} \sum_{l=1}^L \mathbf{1}_{\mathcal{E}_n}(X_1(l), \dots, X_n(l)) w_n(X_1(l), \dots, X_n(l)) \quad (2)$$

où $X_1(l), \dots, X_n(l)$ est un échantillon simulé sous $f_n^*(x_1, \dots, x_n)$ et w_n est appelé le rapport de vraisemblance (ou facteur d'importance)

$$w_n(x_1, \dots, x_n) = \frac{f_n(x_1, \dots, x_n)}{f_n^*(x_1, \dots, x_n)}. \quad (3)$$

L'objectif de l'IS est de choisir la *meilleure* densité f_n^* telle que la variance soit la plus faible possible. Etudions les propriétés de cet estimateur, en commençant par calculer l'espérance.

$$\begin{aligned} E_{f_n^*}[\widehat{P}_n] &= E_{f_n^*}[\mathbf{1}_{\mathcal{E}_n}(\mathbf{X}_1, \dots, \mathbf{X}_n) w_n(\mathbf{X}_1, \dots, \mathbf{X}_n)] \\ &= \int_{\mathcal{E}_n} \frac{f_n(x_1, \dots, x_n)}{f_n^*(x_1, \dots, x_n)} f_n^*(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= E_{f_n}[\mathbf{1}_{\mathcal{E}_n}(\mathbf{X}_1, \dots, \mathbf{X}_n)] \\ &= P_n \end{aligned}$$

Par définition, tous les estimateurs d'IS sont sans biais quelque soit la densité d'échantillonnage f_n^* choisie. Ainsi, l'espérance ne peut pas être un critère de choix. Etudions à présent la variance.

$$\text{Var}_{f_n^*}(\widehat{P}_n) = \frac{1}{L} (\eta_n(f_n^*) - P_n^2)$$

où

$$\eta_n(f_n^*) = E_{f_n^*}[(\widehat{P}_n)^2] = E_{f_n^*}[(\mathbb{1}_{\mathcal{E}_n}(\mathbf{X}_1, \dots, \mathbf{X}_n) w_n(\mathbf{X}_1, \dots, \mathbf{X}_n))^2] \quad (4)$$

Pour obtenir la variance (en fait, le moment d'ordre 2, puisque les estimateurs sont sans biais) la plus petite possible, le terme $\eta_n(f_n^*)$ doit être minimiser. La densité *optimale* est présentée dans le lemme suivant.

Lemme 1 ([43], p.57-58, [56], p.122-124) *La densité d'échantillonnage optimale est*

$$f_n^*(x_1, \dots, x_n) = \frac{\mathbb{1}_{\mathcal{E}_n}(x_1, \dots, x_n)}{P_n} f_n(x_1, \dots, x_n) \quad (5)$$

Dans ce cas,

1. $\eta_n(f_n^*) = P_n^2$
2. $\text{Var}_{f_n^*}(\widehat{P}_n) = 0$, pour tout $L \geq 1$.
3. Les facteurs d'importance sont tous égaux à P_n avec proba 1.

Démonstration. Pour obtenir ce résultat, il suffit d'appliquer l'inégalité de Jensen à (4) et d'étudier le cas d'égalité.

■

Evidemment (5) dépend de P_n , quantité inconnue. Ainsi la solution optimale n'est pas utilisable.

Au regard de (5), une bonne densité d'échantillonnage f_n^* doit minimiser les valeurs de $w_n(x_1, \dots, x_n)$. Ce sera vrai si f_n^* met un grand pourcentage de sa masse sur \mathcal{E}_n , en concentrant sa masse à l'endroit où f_n est plutôt grand. Pour réduire la variance, on veut que le ratio $w_n(x_1, \dots, x_n)$ soit petit sur l'ensemble \mathcal{E}_n . \mathcal{E}_n étant un ensemble rare sous f_n , $f_n(x_1, \dots, x_n)$ est petit quand $(x_1, \dots, x_n) \in \mathcal{E}_n$, donc il faut que $f_n^*(x_1, \dots, x_n)$ soit grand sur \mathcal{E}_n . Le changement de mesure doit permettre à l'événement $(x_1, \dots, x_n) \in \mathcal{E}_n$ d'être plus souvent réalisé.

Pour résoudre ce problème, deux choix bien différents s'offrent à nous. Premièrement, nous pouvons réduire le champ des recherches en limitant les densités possibles à une classe de densité plus petite, typiquement, en imposant à $f_n^*(x_1, \dots, x_n)$ d'être le produit de n densités identiques (même si (5) ne préserve pas l'indépendance) ou en imposant à $f_n^*(x_1, \dots, x_n)$ d'appartenir à certaine une classe de densité paramétrique. La seconde possibilité est d'essayer d'approximer le mieux possible la densité d'échantillonnage optimale. Dans cet article, nous étudierons plusieurs de ces méthodes.

1.4 Contre-Exemple

Il faut cependant faire attention à l'application de ces méthodes qui n'améliorent pas toujours la méthode de Monte-Carlo naïve. Considérons un exemple simple. Soit \mathbf{X} de loi exponentielle de paramètre λ strictement positif. Notons p sa densité. On veut estimer

$$P_1 := P[\mathbf{X} > \gamma]$$

avec $\gamma > 1/\lambda$. On se propose de choisir une nouvelle densité d'échantillonnage appartenant toujours à la famille exponentielle, de paramètre λ_1 . Notons q sa densité. L'estimateur \widehat{P}_1 de P_1 est défini par (2). Nous pouvons déterminer sa variance.

$$\begin{aligned} E_q[\widehat{P}_1^2] &= E_q\left[\left(\frac{p(\mathbf{X})}{q(\mathbf{X})} \mathbb{1}_{(t, +\infty)}(\mathbf{X})\right)^2\right] = \int_t^{+\infty} \left(\frac{p(x)}{q(x)}\right)^2 q(x) dx \\ &= \frac{\lambda^2}{\lambda_1} \int_{x-t}^{+\infty} \exp((\lambda_1 - 2\lambda)x) dx \end{aligned}$$

Ainsi, si $\lambda_1 \geq 2\lambda$, alors la variance est infinie. Quand $\lambda_1 \rightarrow 0$, alors la variance tends vers $+\infty$.

Ce petit exemple permet de voir que l'IS n'est pas toujours une meilleure solution que la méthode naïve, et même parfois pire. Il faut donc choisir soigneusement la nouvelle loi d'échantillonnage pour obtenir une réduction drastique de la variance.

1.5 Notions d'efficacité

Dans les premières sections, nous avons parlé du contrôle de la variance, ou du moment d'ordre 2 puisque les estimateurs d'IS sont sans biais. Nous étudierons aussi le coefficient de variation défini par

Définition 2 *Le coefficient de variation d'un estimateur R_n est défini par*

$$CV(R_n) = \frac{\sqrt{\text{Var}(R_n)}}{E[R_n]}$$

Nous définissons maintenant les notions d'efficacité utilisées dans la suite.

Définition 3 *Un estimateur R_n est dit fortement efficace si*

$$E[R_n^2] = O(P_n^2) \quad (6)$$

quand $n \rightarrow \infty$. En d'autres termes, un estimateur est dit fortement efficace si la taille de l'échantillon pour obtenir α_n à une erreur fixée est bornée comme fonction de n .

Définition 4 *Un estimateur R_n est dit faiblement efficace (ou asymptotiquement efficace) si, pour tout $\epsilon > 0$,*

$$E[R_n^2] = O(P_n^{2-\epsilon}) \quad (7)$$

quand $n \rightarrow \infty$.

Définition 5 *Un estimateur R_n a une complexité polynomiale d'ordre $l \geq 0$ si*

$$E[R_n^2] = O\left(P_n^2 (\log(1/P_n))^{2l}\right) \quad (8)$$

quand $n \rightarrow \infty$. En d'autres termes, si le coefficient de variation augmente au plus à la vitesse polynomiale au degré l en $\log(1/P_n)$. L'estimateur est fortement efficace si $l = 0$ dans (8).

Dans la plupart des cas (Section 3 ou Section 4), l'analyse des estimateurs d'IS permet de conclure uniquement à l'efficacité asymptotique des estimateurs. Dans certains cas très particulier (voir [29]), on peut définir une notion plus forte que celle définie ci-dessous, la notion d'erreur relative asymptotiquement négligeable.

Définition 6 *Un estimateur R_n est dit d'erreur relative asymptotiquement négligeable si, pour tout $\epsilon > 0$,*

$$\limsup_{n \rightarrow \infty} \frac{E[R_n^2]}{P_n^2} \leq 1 + \epsilon \quad (9)$$

En pratique, l'effort de simulation requis pour générer un échantillon sous une loi alternative est très souvent plus grand, comparé à la simulation naïve. Ainsi le ratio des variances ne peut pas être la seule quantité d'intérêt. La comparaison entre deux estimateurs doit prendre en compte l'effort de simulations requis pour obtenir cet estimateur. Dans la plupart des cas, la réduction de la variance est tellement grand que même si l'augmentation du nombre de simulations est non négligeable, le gain est toujours important. Voir [41] pour un indice qui est le produit de la variance et l'espérance de l'effort computationnel requis. Une série d'indices de performance sur la stabilité des moments d'ordre supérieur est considérée dans [50].

1.6 Cas considéré

Dans tout cet article, nous allons nous intéresser à un cas très spécifique d'événements rares permettant d'aborder un certains nombres de notions qui nécessitent peu de bagages théoriques. Le vecteur aléatoire $\mathbf{X}_1^n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ est composé de n variables aléatoires réelles i.i.d. chacune de densité p . Dans ce cas,

$$f_n(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i).$$

On veut estimer la probabilité suivante

$$P_n := P\left[\frac{\mathbf{U}_{1,n}}{n} \geq \gamma\right] \quad (10)$$

pour γ fixé tel que $\gamma > E[u(\mathbf{X})]$ et

$$\mathbf{U}_{1,n} = \sum_{i=1}^n u(\mathbf{X}_i).$$

On peut réécrire (10) sous la forme (1) en notant

$$\mathcal{E}_n := \{(x_1, \dots, x_n) : \frac{1}{n} \sum_{i=1}^n u(x_i) > \gamma\}.$$

On suppose que $\mathbf{U} = u(\mathbf{X})$ a une densité $p_{\mathbf{U}}$ par rapport à la mesure de Lebesgue sur \mathbb{R} et que la fonction caractéristique de \mathbf{U} appartient à L^r pour $r \geq 1$.

Nous considérons uniquement des variables aléatoires à queues légères (les queues de distribution décroissent à une vitesse exponentielle au plus vite), c'est à dire que la fonction génératrice des moments de \mathbf{U} satisfait

$$\phi_{\mathbf{U}}(t) := E[\exp(t\mathbf{U})] < \infty$$

pour t dans un voisinage non vide de 0. On définit alors les fonctions $m(t)$, $s^2(t)$ et $\mu_3(t)$ comme la première, deuxième et troisième dérivée de $\log \phi_{\mathbf{U}}(t)$ et m^{-1} la fonction réciproque de m .

La plupart des événements rares considérés dans la littérature, ne s'écrivent pas sous cette forme. En général, le processus considéré sera plutôt modélisé sous la forme d'une chaîne de Markov ou, plus généralement, d'un processus stochastique, et non pas sous la forme d'une suite de variables aléatoires i.i.d.. Plusieurs états de l'art consacrés à l'IS existent déjà dans la littérature dans ce cadre ([6], [44], [46], ou [15]). Cependant, l'estimation de la probabilité de grandes déviations dans le cas des sommes i.i.d. est une étape indispensable pour des problèmes plus complexes (en file d'attente ou en modélisation du risque). Cela nous permet d'introduire les notions principales les plus importantes, de donner des méthodes effectives simples et des pistes pour des modèles plus complexes. Des sujets variés concernant l'IS sont présentés dans [48]; pour les premiers temps de passage, voir [52] ou [34]; pour l'étude de P_n dans le cas des variables aléatoires à queue lourde, voir [33].

On introduit aussi la famille de densité *tiltée*

$$\pi_u^\alpha(x) := \frac{\exp(tu(x))}{\phi_{\mathbf{U}}(t)} p_{\mathbf{X}}(x). \quad (11)$$

avec Π_u^α la fonction de distribution associée. Dans (11), $m(t) = \alpha$ pour tout α appartenant au support de $P_{\mathbf{U}}$, la fonction de répartition de \mathbf{U} . On suppose que la fonction $\phi_{\mathbf{U}}(t)$ est *steep*, c'est à dire que t défini à travers la fonction m existe pour tout α dans le support de $P_{\mathbf{U}}$ (voir [9], p153).

2 Méthodes générales

2.1 De l'IS à états indépendants à l'IS à états dépendants.

Nous passerons en revue les deux principales méthodes existantes : l'Importance Sampling à état indépendant et l'Importance Sampling à état dépendant. La première consiste à fabriquer une densité d'échantillonnage i.i.d. tandis que la seconde consiste à adapter le changement de mesure aux réalisations précédentes.

Dans le contexte des grandes déviations, les premiers résultats obtenus permettant à \widehat{P}_n d'être asymptotiquement efficace sont donnés par [64]. Il propose que le changement de mesure (appelée *densité tiltée optimale*) utilisé pour prouver la borne inférieure du principe de grande déviation associé doit être une bonne loi d'échantillonnage à utiliser dans le cadre de l'IS. Cette idée a été reprise et étudiée, en particulier par [52], [61], [62] ou [16] dans différents contextes. Nous étudierons en détails cette méthode dans le cadre défini par la section 1.6. Des conditions nécessaires et suffisantes pour lesquelles la densité tiltée optimale permet l'efficacité asymptotique sont exposées dans [60] et dans [63]. Ces conditions ont été améliorées par [31] utilisant le lemme de Varadhan. Il est démontré que la densité tiltée optimale est le seul schéma i.i.d. qui permet l'efficacité asymptotique. Ce résultat n'est pas étonnant au sens où le tilting optimal coïncide avec la distribution de \mathbf{X}_1 conditionnellement à $(\mathbf{X}_1^n \in \mathcal{E}_n)$, voir [22] ou [37]; cette loi est la loi d'échantillonnage idéale permettant de fabriquer un estimateur de variance nulle. Ces résultats sont généralisés à des ensembles A convexes de \mathbb{R}^d . Dans ces articles, l'identification d'un point appelé *point dominant* ou de plusieurs *points à taux minimaux* (voir [54] et section 3.3) permet de fabriquer soit la densité tiltée optimale associée à ce point dominant, soit une combinaison convexe de lois tiltées associées chacune à un point à taux minimal. Cette méthode est appelée Importance Sampling à états indépendants (ISEI dans la suite) car, à chaque itération, X_i est simulé indépendamment de (X_1, \dots, X_{i-1}) et de la cible $(\mathbf{X}_1^n \in \mathcal{E}_n)$.

Cette méthode a cependant quelques imperfections. En effet, dans [39], des ensembles A sont exhibés pour lesquels l'ISED donne de très mauvais résultats, parfois pire que la méthode Monte-Carlo naïve. Nous étudierons un de ces exemples dans la section 5 pour comparer cette méthode avec la méthode proposée dans la section 4.

Ces remarques ont permis de remettre en cause l'idée selon laquelle le changement de mesure permettant de prouver la borne supérieure est un bon changement de mesure. La question posée est la suivante : existe-t-il d'autres changements de mesure permettant d'améliorer la performance de \widehat{P}_n ? Nous avons présenté, ci-dessus, un changement de mesure qui considère toutes les itérations de manière identique, mais nous pourrions proposer à la place un changement de mesure adaptatif, voir [21] ou [32]. Ainsi, la simulation de X_i dépend à la fois du passé (X_1, \dots, X_{i-1}) et de la cible $(\mathbf{X}_1^n \in \mathcal{E}_n)$. Cette méthode est appelée IS à états dépendants (ISED dans la suite).

Une autre approche a été proposée dans [35] qui fait le lien entre IS et théorie des jeux. Une approche alternative aux principes de grandes déviations est basée sur la méthodologie du contrôle stochastique dans lequel on peut prouver des bornes supérieures et inférieures simultanément. Dans ce cadre, la quantité d'intérêt $-(1/n)\log(p_n)$ est représenté comme le coût minimal d'un problème de contrôle stochastique. Les auteurs proposent d'appliquer l'IS en accord avec la solution de l'équation d'Isaac correspondante (section 3.4 de [35]). Ils prouvent que si une solution est continument différentiable, alors l'estimateur correspondant est faiblement efficace. Dans [4], la solution à l'équation d'Isaac correspondante est deux fois continument différentiable. En appliquant le tilting optimal à état dépendant dans une région où le principe de grande déviation est vérifié, la forte efficacité de leur estimateur est prouvée. En d'autres termes, ils prouvent que le nombre de simulations nécessaires pour obtenir à une erreur fixé près reste borné comme fonction de n . Fabriquer des estimateurs fortement efficaces n'est pas chose aisée. Ils sont pour l'instant limités aux systèmes à queue lourde (voir [8]) ou aux incréments gaussiens (voir [7]). D'autres méthodes adaptatives ont été développées dans le cadre de l'IS, voir [17] ou [20].

Depuis le début, les méthodes développés ici permettaient d'estimer P_n pour $A = (\gamma, +\infty)$ avec γ fixé. Dans les applications, il est souvent utile d'estimer plusieurs probabilités en même temps, ou même une loi entière. Le tilting exponentiel produit des estimateurs asymptotiquement efficaces pour un seul point dans les queues de distributions. Si on veut obtenir des estimateurs asymptotiquement efficaces pour plusieurs points simultanément, les méthodes standard ne fonctionnent pas. Dans [40], les auteurs fabriquent des estimateurs fortement efficaces de $P[\mathbf{U}_{1,n}/n \geq \gamma]$ pour n grand et pour tout $\gamma > E[u(\mathbf{X})]$ où γ est un point d'un intervalle A .

2.2 Approximation de la densité optimale

Un autre type de méthodes existe et consiste à approximer la densité optimale définie dans (5) et qui peut être réécrite sous la forme suivante

$$p(x_1, \dots, x_n | \mathcal{E}_n).$$

Nous nous intéressons à l'approximation de cette densité non pas sur tout \mathbb{R}^n mais sur \mathbb{R}^k pour k pouvant dépendre de n mais plus petit que n .

Dans le cas où $k = 1$, c'est une version du *Principe conditionnel de Gibbs* qui a été étudiée de manière intensive dans le cas où γ est fixé et différent de $E[u]$. Dans [30], les auteurs ont généralisé ce résultat au cas où $k/n \rightarrow \theta$ pour $0 \leq \theta < 1$, en connections avec le théorème de Finetti pour des suites échangeables. Ils ont démontré que cette densité conditionnelle peut-être approximée par le produit des densités des \mathbf{X}_i , démontrant alors l'indépendance asymptotique des \mathbf{X}_i sous le conditionnement. Dans [65], les auteurs ont aussi étudié le même type de problème. Ce résultat est à mettre en parallèle avec les résultats obtenues par [22], c'est à dire l'indépendance asymptotique dans le cas où l'événement conditionnant s'écrit $(\mathbf{S}_{1,n} > na)$. On peut aussi noter les travaux de [25] qui ont également considérés des problèmes similaires. En physique statistique, l'étude des polymères soulève des questions proches, et, dans ce contexte, des résultats ont été obtenues dans [27] et [28]. Dans le cas des déviations modérées, c'est à dire quand $\gamma = \gamma_n$ converge lentement vers $E[u(\mathbf{X})]$, [37] a aussi considéré le problème pour $k = 1$.

L'approximation de la densité optimale doit avoir plusieurs propriétés importantes. Elle doit elle-même être une densité sur \mathbb{R}^k , elle doit être relativement facile à simuler et être une approximation très fine. Le résultat présenté ici a plusieurs différences majeures avec les résultats obtenus dans la littérature. D'une part, le résultat est obtenu non pas sur tout \mathbb{R}^k , mais uniquement sur les chemins simulés ou *chemins typiques*. Pour développer des méthodes d'IS, l'approximation doit être fine pour les chemins simulés sous cette approximation. Cependant, des résultats obtenues sur tout \mathbb{R}^k sont obtenus. En effet, la distance en variation totale entre la densité conditionnelle et son approximation tends vers 0 sur tout \mathbb{R}^k .

2.3 Autres méthodes

Nous proposons dans cette section quelques références sur deux méthodes alternatives permettant d'étudier les événements rares.

L'importance splitting (ou méthode multi-niveaux) est une méthode alternative à l'IS permettant d'augmenter l'occurrence de l'événement rare. Elle a été introduite par [47]. Dans ce cas, la loi de probabilité qui régit le modèle n'est pas modifiée, à la place, nous modifions l'événement cible. L'idée principale est de décomposer l'événement rare en une suites d'événements pour lesquels la probabilité d'atteinte est plus grande. Les chemins qui vont dans la mauvaise direction sont éliminés tandis que les chemins qui atteignent le plus petit niveau sont répliqués ; pour un tutoriel sur ces méthodes, voir [23] ou [48].

La méthode de l'entropie croisée, proposée par [57] et [58], peut-être aussi considérée comme une méthode d'IS, dans le sens où elle permet de déterminer la densité d'échantillonnage de manière adaptative en minimisant l'entropie croisée entre la densité paramétrique proposée et la densité optimale. Nous obtenons alors comme densité d'échantillonnage la densité la plus proche de la densité optimale dans la classe des densités paramétriques choisie au sens de la divergence de Kullback-Leibler ; voir [10] ou [24] pour un tutoriel.

3 Importance Sampling et grande déviation

Cette section est consacrée aux connexions entre l'IS et la théorie des grandes déviations. Les hypothèses sont les plus élémentaires possibles, laissant la généralisation de ces théorèmes en référence. Le résultat principal de cette section est le suivant, parmi toutes les lois d'échantillonnage i.i.d., la seule permettant d'obtenir l'efficacité asymptotique est le tilting exponentiel optimal. Le résultat est basé sur les principes de grandes déviations et, notamment, le théorème de Cramer.

3.1 Principes de grandes déviations

Les résultats présentés dans cette section est une version du théorème de Cramer portant sur la vitesses de convergence de P_n . De nombreuses versions de ce théorème existent et des généralisations peuvent être trouvées dans [26] ou [36]. L'idée consiste à déterminer une borne supérieure et une borne inférieure pour obtenir la vitesse de convergence de P_n vers 0 quand n tends vers l'infini. L'étude de la démonstration de la borne supérieure permet d'obtenir des éléments de preuve pour démontrer l'efficacité asymptotique de \widehat{P}_n . On a,

$$P_n = E_p[\mathbb{1}_{(\gamma, +\infty)}\left(\frac{\mathbf{U}_{1,n}}{n}\right)] \leq E_p[\exp\left(n\alpha\left(\frac{\mathbf{U}_{1,n}}{n} - \gamma\right)\right)]$$

pour tout $\alpha \geq 0$, en utilisant l'inégalité suivante $\mathbb{1}_{(\gamma, +\infty)}(x) \leq \exp(\alpha(x - \gamma))$. Ainsi,

$$\begin{aligned} \mathbb{1}_{\mathcal{E}_n}(x_1, \dots, x_n) &\leq \exp\left(\alpha \sum_{i=1}^n (u(x_i) - \gamma)\right) \\ &= \prod_{i=1}^n \exp(\alpha(u(x_i) - \gamma)). \end{aligned}$$

On obtient ainsi la borne supérieure suivante pour P_n

$$\begin{aligned} P_n &\leq E_p\left[\prod_{i=1}^n \exp(\alpha(u(\mathbf{X}_i) - \gamma))\right] \\ &\leq (E_p[\exp(\alpha(u(\mathbf{X}) - \gamma))])^n \\ &\leq \exp(-n(\alpha\gamma - \log(\phi(\alpha)))) \end{aligned} \tag{12}$$

avec $\phi(\alpha)$ la fonction génératrice des moments (ou $\Lambda(\alpha) := \log(\phi(\alpha))$ la fonction génératrice des cumulants). Pour obtenir la plus petite borne dans (12), nous devons minimiser l'exposant par rapport au paramètre α . Ainsi

$$P_n \leq \exp(-nI(\gamma)) \tag{13}$$

avec

$$I(\gamma) = \sup_{\alpha \in \mathbb{R}} \{\alpha\gamma - \log(\phi(\alpha))\}. \tag{14}$$

La maximisation dans (14) devrait être seulement sur $\alpha \geq 0$. Nous avons dans la section 1.6 réduit les hypothèses au cas où $\gamma \geq E[u(\mathbf{X})]$. Ainsi, la maximisation dans (13) a lieu soit quand $\alpha \geq 0$, soit à la limite $\alpha \rightarrow +\infty$. La

maximisation peut être étendue à tout l'ensemble \mathbb{R} . En remplaçant l'hypothèse précédente par $\gamma \leq E[u(\mathbf{X})]$, la maximisation aura lieu soit quand $\alpha \leq 0$, soit à la limite $\alpha \rightarrow -\infty$. On peut aussi dans ce cas étendre la maximisation à tout l'ensemble \mathbb{R} .

La fonction $I(\gamma)$, appelée fonction de taux de grande déviation, est la transformée de Legendre-Fenchel de $\log(\phi(\alpha))$. La borne supérieure (13) décroît exponentiellement vite comme fonction de n . Nous devons aussi déterminer la borne inférieure. Nous renvoyons à [26] pour cette démonstration. La vitesse exponentielle de décroissance de P_n est exactement $I(\gamma)$.

Théorème 7 Soit \mathbf{X}_1^n n variables aléatoires i.i.d.. On suppose satisfaites les conditions mentionnées dans la section 1.6. Alors

1. $I(\gamma)$ est une fonction convexe.
2. $I(\gamma) \geq 0$
3. $I(\gamma) = 0$ si et seulement si $\gamma = m(0) = E[\frac{\mathbf{U}_{1,n}}{n}]$

Soit $I = \{\gamma : I(\gamma) < \infty\}$ un intervalle, alors

$$\forall \gamma \in \overset{\circ}{I}, \quad \lim_{n \rightarrow \infty} \frac{1}{n} \log(P_n) = -I(\gamma) \quad (15)$$

et il existe un unique point α_γ tel que

$$m(\alpha_\gamma) = \gamma$$

et à ce point $I(\gamma) = \alpha_\gamma \gamma - \log(\phi(\alpha_\gamma))$.

Remarque 8 En général, on trouvera une version de ce théorème faisant référence aux limites supérieure et inférieure. Les hypothèses, sous lesquelles nous travaillons, nous permettent de nous affranchir de ce double énoncé.

Remarque 9 Dans le cas plus général où $P_n := P[\mathbf{Y}_n \in A]$ avec Y_n une chaîne de Markov, des résultats sont obtenues dans [26] ou [36].

3.2 Application à l'IS

Nous considérons toutes les densités d'échantillonnage de la forme

$$f_n^*(x_1, \dots, x_n) = \prod_{i=1}^n q(x_i).$$

Nous rappelons dans ce cas particulier l'écriture de la variance de l'estimateur d'IS

$$\text{Var}_q(\widehat{P}_n) = \frac{1}{L} (\eta_n(q) - P_n^2)$$

avec

$$\eta_n(q) = E_q \left[\mathbb{1}_{\mathcal{E}_n}(\mathbf{X}_1^n) \left(\prod_{i=1}^n \frac{p(\mathbf{X}_i)}{q(\mathbf{X}_i)} \right)^2 \right].$$

En utilisant les mêmes arguments utilisés pour démontrer la borne supérieure pour P_n , on obtient,

$$\eta_n(q) \leq \exp(-n(\alpha_q - \Lambda_q(\alpha))) \quad (16)$$

avec

$$\Lambda_q(\alpha) = \log \left(E_q \left[\left(e^{\alpha g(\mathbf{X})} \frac{p(\mathbf{X})}{q(\mathbf{X})} \right)^2 \right] \right). \quad (17)$$

Une nouvelle fois, nous minimisons le membre de droite de (16) pour obtenir

$$\eta_n(q) \leq e^{-nI_q(\gamma)} \quad (18)$$

avec

$$I_q(\gamma) = \sup_{\alpha \in \mathbb{R}} \{\alpha \gamma - \Lambda_q(\alpha)\}. \quad (19)$$

Théorème 10 *Sous les mêmes hypothèses que le théorème 7, $I_q(\gamma)$ est convexe et pour tout $\gamma \in \dot{I}_q$ telle que $\Lambda'_q(0) \leq \gamma$, alors*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log(\eta_n(q)) = -I_q(\gamma).$$

En utilisant ce résultat et le fait que $\eta_n(q) \geq P_n^2$ pour tout n (car la variance est toujours positive), on obtient alors l'inégalité suivante

$$\forall t \in \mathbb{R}, \quad I_q(t) \leq 2I(t). \quad (20)$$

Nous pouvons maintenant reformuler la définition 4 concernant l'efficacité asymptotique sous une forme faisant apparaître les fonctions de taux précédentes.

Définition 11 *L'estimateur d'IS basé sur la densité q est asymptotiquement efficace pour P_n si et seulement si $I_q(\gamma) = 2I(\gamma)$.*

L'objectif est de caractériser q dans le cas où $I_q(t) = 2I(t)$. Pour ce faire, nous utilisons des résultats de dualité convexe entre les relations (14) et (19) (voir [36]), on obtient l'équivalence suivante

$$I_q(\gamma) = 2I(\gamma) \text{ si et seulement si } \Lambda_q(2\alpha_\gamma) = 2\Lambda(\alpha_\gamma). \quad (21)$$

En utilisant la définition de Λ_q et l'inégalité de Jensen, on obtient une inégalité similaire à celle déjà obtenue pour la fonction I_q .

$$\begin{aligned} \Lambda_q(2\alpha_\gamma) &\geq \log \left(E_q \left[\left(e^{\alpha_\gamma g(\mathbf{X})} \frac{p(\mathbf{X})}{q(\mathbf{X})} \right)^2 \right] \right) \\ &= 2 \log \left(\int \left(e^{\alpha_\gamma g(x)} \frac{p(x)}{q(x)} \right) q(x) dx \right) \\ &= 2 \log \left(\int \left(e^{\alpha_\gamma g(x)} p(x) \right) dx \right) \\ &= 2\Lambda(\alpha_\gamma). \end{aligned}$$

Grâce à l'inégalité de Jensen, nous obtenons le cas d'égalité. $\Lambda_q(2\alpha_\gamma) = 2\Lambda(\alpha_\gamma)$ si et seulement si

$$\exp \left(e^{\alpha_\gamma g(x)} \right) \frac{p(x)}{q(x)} = \text{cst}$$

avec q -probabilité 1. Et on obtient l'efficacité asymptotique pour une seule et unique densité d'échantillonnage.

Théorème 12 *Sous les hypothèses du Théorème 7 et pour tout γ dans l'intérieur de I , alors la densité d'échantillonnage tiltée est l'unique densité permettant d'atteindre l'efficacité asymptotique pour P_n .*

Remarque 13 *Dans le cas plus général où $P_n := P[\mathbf{Y}_n \in A]$ avec Y_n une chaîne de Markov satisfaisant une certaine relation de récurrence, des résultats sont obtenues dans [16].*

Pour fabriquer l'estimateur d'IS défini par 2, nous devons fabriquer L échantillons de taille n . Nous proposons une procédure en plusieurs étapes :

- Résoudre en α_γ l'équation $m(\alpha_\gamma) = \gamma$.
- Simuler Y_1, \dots, Y_n de densité π^{α_γ} définie par (11).
- Itérer l'opération L fois.

3.3 Points dominants

Quand l'événement étudié ne s'écrit plus sous la forme P_n mais sous une forme plus général

$$P_n = P[\mathbf{U}_{1,n}/n \in A]$$

pour un ensemble A mesurable de \mathbb{R} , la méthode précédente doit être légèrement modifiée. En effet, nous devons choisir un point particulier de l'ensemble A pour lequel nous pouvons définir le tilting optimal. C'est pourquoi la notion de point dominant est centrale en théories des grandes déviations, voir [54] pour une formalisation théorique. Notons $I(A)$ la quantité suivante

$$I(A) := \inf_{t \in A} \{I(t)\}$$

Définition 14 On suppose $0 < I(A) < \infty$.

1. v est un point à taux minimal de A si $v \in \delta A$ et $I(v) = I(A)$.
2. v est un point dominant de A si v est l'unique point tel que
 - (a) $v \in \delta A$
 - (b) $\exists v \in \mathbb{R}^d$ tel que $\nabla \mu(s_v) = v$
 - (c) $A \subset \mathcal{H}(v) := \{y \in \mathbb{R}^d : \langle s_v, y - v \rangle \geq 0\}$

Remarque 15 1. Si $X \sim \mathcal{N}(0, 1)$ et $A = (v, +\infty)$, alors v est un point dominant de A . Si $A = (-\infty, -v) \cup (v, +\infty)$, alors v et $-v$ sont des points à taux minimaux.

2. Une condition suffisante (mais pas nécessaire) pour l'existence d'un point dominant est

- (a) A convexe
- (b) $A \cap \mathring{I} \neq \emptyset$.

Le cas où A a un point dominant est évidemment un cas très particulier. Plus généralement, dans le cas où il n'existe pas de point dominant mais plusieurs points à taux minimal, une loi d'échantillonnage doit prendre en compte tous ces points mais la loi définie par (11) ne prends en compte qu'un seul point. On élimine ce problème en fabriquant pour la loi d'échantillonnage des combinaisons de lois tiltées définies par (11), chacune associée à un point à taux minimal.

3.4 Limitation de l'ISEI

Un des problèmes majeurs de cette méthode est de déterminer les points dominants. Dans [39], les auteurs montrent que, dans des cas simples, il faut recourir à un mélange de lois tiltées pour fabriquer un estimateur faiblement asymptotique. Donnons un exemple simple qui sera traité dans la Section 5. Soit $A = (-\infty, a) \cup (b, +\infty)$, $E[\mathbf{X}_1] \in (a, b)$ et $u(x) = x$. Dans ce cas, l'ensemble A est non convexe. Pour des valeurs bien choisies de a et b telles que $0 < I(b) < I(a)$, la nouvelle densité d'échantillonnage est centrée en b , la partie gauche de l'ensemble A est ignoré. Evidemment, deux solutions s'offrent à nous, soit faire un mélange de deux lois en appliquant des poids correspondant à chacun de deux ensembles, soit réduire l'ensemble A en deux sous-ensembles et les traiter en suivant le schéma introduit ci-dessus. Mais on pourrait toujours réduire l'ensemble A d'intérêt à des sous-ensembles élémentaires, même si A est un ensemble de \mathbb{R}^d . On voudrait éviter de fabriquer une série de sous-ensembles, ce qui peut être compliqué. Une des nombreux avantages de la méthode présentée dans la section suivante est l'absence de la notion de point dominant.

4 Approximation de la densité d'échantillonnage optimal

4.1 Contexte

Cette section est consacrée à l'étude de la densité de longue suite d'une marche aléatoire conditionnée à sa valeur final quand le nombre de termes augmente.

Comme explicité précédemment, le choix optimal pour la densité d'Importance Sampling est la densité conditionnelle définie dans (5). Dans cette section, nous allons présenter comment obtenir une approximation fine de cette densité conditionnelle et comment l'utiliser pour obtenir un estimateur faiblement efficace. Dans des contextes différents, des approximations de densité conditionnelle ont aussi été obtenues, voir [51] dans le contexte de la fiabilité.

Considérons une nouvelle fois $\mathbf{X}_1, \dots, \mathbf{X}_n$ n variables aléatoires réelles i.i.d. de densité $p_{\mathbf{X}}$ et u une fonction réelle mesurable à valeurs dans \mathbb{R} . On définit à présent $\mathbf{U} := u(\mathbf{X})$ de densité $p_{\mathbf{U}}$ et

$$\mathbf{U}_{1,n} := \sum_{i=1}^n \mathbf{U}_i.$$

Nous voulons estimer

$$P_n := P(\mathbf{U}_{1,n} \in nA_n)$$

pour n grand et fixe où

$$A := (\gamma, \infty) \tag{22}$$

Les résultats présentés ici restent vrai dans le cas où $A = A_n := (\gamma_n, +\infty)$. La limite de la suite γ_n peut-être

1. soit $E[\mathbf{U}]$. Dans ce cas, on supposera que γ_n converge lentement vers $E[\mathbf{U}]$ et P_n ne peut pas être obtenu par le théorème de la limite central. Ce cas est souvent appelé *moyenne déviation*.
2. soit être plus grand que $E[\mathbf{U}]$. Ce cas sera appelé *grande déviation*.

L'intérêt d'obtenir un résultat valide dans le cas moyenne déviation est double. D'une part, des méthodes d'IS ont été proposées par [37] permettant l'évaluation de p-values pour des M-estimateurs. En effet, les M-estimateurs peuvent être approximés par linéarisation de leur fonction d'influence (voir [38] pour les détails) permettant d'utiliser les méthodes décrites ici. D'autre part, obtenir des évaluations des probabilités dans le cas de déviations modérées est aussi important en statistiques, dans le cas des fonctions puissances des tests. En effet, [13] ont utilisé cette méthode utilisée dans le cadre des tests de Monte-Carlo.

Les résultats présentés dans cette section sont grandement tirés de plusieurs articles (voir [12] où les démonstrations des théorèmes peuvent être trouvées, ainsi que [14] pour une étude exhaustive). Le cas où les variables aléatoires sont d -dimensionnelles est traité dans [18]. Le cas où A est une union dénombrable d'intervalles peut être facilement déduit du cas présenté ici (voir Section 5).

4.2 Notations et Hypothèses

Pour les raisons explicitées ci-dessus, la valeur de la densité $p_{\mathbf{Z}}$ d'un vecteur aléatoire \mathbf{Z} au point z sera noté $p_{\mathbf{Z}}(z)$ or $p(\mathbf{Z} = z)$. La densité normale réelle de moyenne μ et de variance τ sera noté $\mathbf{n}(\mu, \tau, x)$.

La densité de $\mathbf{X}_1^{k_n}$ sous le conditionnement local ($\mathbf{U}_{1,n} = nv$) sera noté

$$p_{nv}(\mathbf{X}_1^{k_n} = Y_1^{k_n}) := p(\mathbf{X}_1^k = Y_1^{k_n} \mid \mathbf{U}_{1,n} = nv) \quad (23)$$

où $Y_1^{k_n}$ appartient à \mathbb{R}^{k_n} .

Nous allons aussi considérer la densité p_{nA} de $\mathbf{X}_1^{k_n}$ sous le conditionnement global ($\mathbf{U}_{1,n} \in nA$)

$$p_{nA}(\mathbf{X}_1^{k_n} = Y_1^{k_n}) := p(\mathbf{X}_1^{k_n} = Y_1^{k_n} \mid \mathbf{U}_{1,n} \in nA). \quad (24)$$

La densité approximante de p_{nv} est noté g_{nv} tandis que la densité approximante correspondante à p_{nA} est noté g_{nA} . Les fonctions de répartition sont notés, respectivement, P_{na} , G_{na} , P_{nA} and G_{nA} .

4.3 Principes de la méthode.

Nous voulons approximer la densité p_{nA} sur \mathbb{R}^{k_n} afin d'obtenir des estimateurs d'IS les plus efficaces possibles. Pour cela, nous réécrivons p_{nA} sous la forme suivante

$$p_{nA}(x_1^{k_n}) = \int_A p_{nv}(\mathbf{X}_1^{k_n} = x_1^{k_n}) p(\mathbf{U}_{1,n}/n = v \mid \mathbf{U}_{1,n} \in nA) ds. \quad (25)$$

Cette écriture permet de nous affranchir de la notion de point dominant introduite dans la section précédente. Nous réalisons une approximation de la densité sous le conditionnement ponctuel pour tous les points de A , et nous considérons la contribution dominante de toutes ces distributions dans l'évaluation de la densité conditionnelle p_{nA} .

Afin d'expliquer clairement le fonctionnement de cette méthode, nous l'illustrons à travers un simple exemple de simulation d'un échantillon sous l'approximation de la densité conditionnelle p_{nA} . Soit \mathbf{X}_1^n un vecteur aléatoire i.i.d. où chaque X_i est une variable aléatoire réelle de loi normale standard. On s'intéresse à l'estimation de $P_n := P(\mathbf{S}_{1,n} > n\gamma)$ où $\mathbf{S}_{1,n} := \mathbf{X}_1 + \dots + \mathbf{X}_n$ et $\gamma > 0$ fixé.

1. Pour tout $v > a$, la densité jointe p_{nv} de $\mathbf{X}_1, \dots, \mathbf{X}_{n-1}$ conditionnellement à ($\mathbf{S}_{1,n} = nv$) est connue analytiquement (un produit de loi normale) et, donc, simuler Y_1^n sous p_{nv} devient simple pour tout v .
2. La densité optimale p_{nA} est similaire à p_{nv} sous l'événement conditionnant ($\mathbf{S}_{1,n} > n\gamma$). La densité g_{nA} est obtenue par intégration par rapport à la loi de $\mathbf{S}_{1,n}/n$ sous le conditionnement ($\mathbf{S}_{1,n} > n\gamma$), qui est bien approximée par une loi exponentielle sur (γ, ∞) de paramètre $\gamma + (1/n\gamma)$ (voir Proposition 3 de [4] ou Corollaire 6.4.1 de [45]). Ainsi, un échantillon de loi g_{nA} est obtenu par une méthode Monte-Carlo de la manière suivante :
 - (a) Simuler \mathbf{V} selon la loi exponentielle précédente.
 - (b) Simuler Y_1^n de densité $p_{n\mathbf{V}}$.

En répétant cette opération L fois, nous obtenons \widehat{P}_n en insérant ces termes dans (2).

L'exemple présenté est idéal, au sens où, dans le cas gaussien, nous connaissons p_{nv} analytiquement et nous pouvons approximer facilement $p(\mathbf{U}_{1,n}/n = v | \mathbf{U}_{1,n} \in nA)$. La densité p_{nv} n'est évidemment pas analytiquement connu dans la plupart des cas et doit donc être approximé. Par construction, l'approximation proposée ici est valide uniquement sur les k_n premières composantes avec k_n très grand, typiquement $k_n/n \rightarrow 1$, mais doit cependant satisfaire la condition $(n - k_n) \rightarrow \infty$. Le choix de ce paramètre k_n est crucial pour assurer la qualité de l'estimateur d'IS. La qualité de l'approximation et la valeur de k_n sont inversement proportionnelles. Ce choix est discuté dans la section 3.2 de [12]. La densité optimale de $\mathbf{X}_1, \dots, \mathbf{X}_n$ est obtenu en multipliant cette approximation par le produit de $n - k_n$ densités π^{a_n} . Ce choix est en accord avec les résultats de grande déviation présenté dans la section 3. Les résultats d'efficacité sont présentés dans la section 4.7 et montre que l'erreur relative de l'estimateur IS est proportionnel à $\sqrt{n - k_n}$ permettant une diminution d'un facteur $\sqrt{n - k_n}/\sqrt{n}$ par rapport à la méthode ISEI. Une amélioration est de compléter par $n - k_n$ variables aléatoires simulées de manière adaptative en accord avec [4] et permettant la fabrication d'un estimateur fortement efficace.

La section 4.4 présente l'approximation dans le cas d'un conditionnement ponctuel ainsi que des éléments de preuve et des remarques sur le choix de k_n . La section 4.5 présente l'approximation de la densité optimale et la section 4.7 la mise en place pratique de l'estimateur IS associé.

Le résultat présenté peut aisément être généralisé (voir [18] pour détails) au cas où la fonction u est une fonction mesurable de \mathbb{R}^d dans \mathbb{R}^s (d et s deux entiers plus grand que 1) et où les \mathbf{X}_i sont des vecteurs aléatoires d -dimensionnels.

Pour simplifier les notations, la suite positive d'entiers k_n sera noter k .

4.4 Echantillons conditionnels

Le point de départ de l'obtention d'une approximation de la densité optimale est l'approximation de p_{nv} définie par (23) sur \mathbb{R}^k pour de grandes valeurs de k sous le conditionnement ponctuel ($\mathbf{U}_{1,n} = nv$) quand v appartient à A .

Soit ϵ_n une suite positive qui satisfait

$$\lim_{n \rightarrow \infty} \epsilon_n \sqrt{n - k} = \infty \quad (26)$$

$$\lim_{n \rightarrow \infty} \epsilon_n (\log n)^2 = 0. \quad (27)$$

Définissons à présent la densité approximante $g_{nv}(y_1^k)$ de la façon suivante. Soit

$$g_0(y_1 | y_0) := \pi_u^v(y_1) \quad (28)$$

où y_0 est arbitraire et, pour $1 \leq i \leq k - 1$, nous définissons $g(y_{i+1} | y_1^i)$ de manière récursive de la façon suivante.

Soit t_i l'unique solution de l'équation

$$m_i := m(t_i) = \frac{n}{n - i} \left(v - \frac{u_{1,i}}{n} \right) \quad (29)$$

avec $u_{1,i} := u(y_1) + \dots + u(y_i)$.

Définissons

$$g(y_{i+1} | y_1^i) = C_i p_{\mathbf{X}}(y_{i+1} | \mathbf{n}(\alpha\beta + v, \beta, u(y_{i+1}))) \quad (30)$$

où C_i est une constante de normalisation et α et β sont définis par

$$\alpha = t_i + \frac{\mu_3(t_i)}{2s^4(t_i)(n - i - 1)}. \quad (31)$$

$$\beta = s^2(t_i)(n - i - 1) \quad (32)$$

Finalement,

$$g_{nv}(y_1^k) := g_0(y_1 | y_0) \prod_{i=1}^{k-1} g(y_{i+1} | y_1^i). \quad (33)$$

Nous pouvons maintenant énoncer le théorème suivant.

Théorème 16 *Supposons (26) et (27). Alors (i)*

$$p_{nv}(\mathbf{X}_1^k = Y_1^k) = g_{nv}(Y_1^k)(1 + o_{P_{nv}}(\epsilon_n(\log n)^2)) \quad (34)$$

et (ii)

$$p_{nv}(\mathbf{X}_1^k = Y_1^k) = g_{nv}(Y_1^k)(1 + o_{G_{nv}}(\epsilon_n(\log n)^2)). \quad (35)$$

Remarque 17 *Dans le cas particulier où les \mathbf{X}_i sont i.i.d. de loi normale standard, les résultats d'approximation du théorème précédent sont vrais pour $k = n - 1$ avec $p_{nv}(\mathbf{X}_1^{n-1} = x_1^{n-1}) = g_{nv}(x_1^{n-1})$ pour tout x_1^{n-1} dans \mathbb{R}^{n-1} sans terme d'erreur.*

Remarque 18 *Sous les hypothèses du théorème précédent, la distance en variation totale entre P_{nv} and G_{nv} tend vers 0 quand n tend vers l'infini.*

4.5 Approximation de la densité optimale

Comme explicité dans l'introduction de cette section, le choix optimal pour la densité d'IS est p_{nA} pour laquelle nous énonçons un théorème d'approximation. En observant (25), nous devons obtenir une approximation pour le terme

$$p(\mathbf{U}_{1,n}/n = v | \mathbf{U}_{1,n} \in nA) = \frac{p(\mathbf{U}_{1,n}/n = v) \mathbb{1}_A(v)}{P(\mathbf{U}_{1,n} \in nA)} \quad (36)$$

avec uniformité pour v dans A .

Pour obtenir l'approximation de p_{nA} , il est nécessaire d'introduire un intervalle $(\gamma, \gamma + c_n)$, c'est à dire sur la partie principale de l'intégrale (25). On définit alors la densité g_{nA} sur \mathbb{R}^k de la façon suivante

$$\begin{aligned} g_{nA}(y_1^k) \\ := \frac{nm^{-1}(\gamma) \int_{\gamma}^{\gamma+c} g_{nv}(y_1^k) (\exp -nm^{-1}(\gamma)(v - \gamma)) dv}{1 - \exp -nm^{-1}(\gamma)c}. \end{aligned} \quad (37)$$

La densité

$$\frac{nm^{-1}(\gamma) (\exp -nm^{-1}(\gamma)(v - \gamma)) \mathbb{1}_{(\gamma, \gamma+c)}(v)}{1 - \exp -nm^{-1}(\gamma)c} \quad (38)$$

qui apparait dans (37) approxime $p(\mathbf{U}_{1,n}/n = v | \gamma < \mathbf{U}_{1,n}/n < \gamma + c)$.

La fonction variance V de \mathbf{U} est définie sur le support de la loi de \mathbf{U} par

$$v \rightarrow V(v) := s^2(m^{-1}(v))$$

Théorème 19 *Soit c_n une suite positive satisfaisant les conditions suivantes*

$$\begin{aligned} \lim_{n \rightarrow \infty} nc_n &= \infty \\ \sup_{n \geq 1} \frac{nc_n}{(n-k)} &< \infty. \end{aligned}$$

De plus, la condition suivante est supposée satisfaite

$$\sup_{n \geq 1} \sqrt{n} \int_{\gamma}^{\infty} V'(v) (\exp -nm^{-1}(\gamma)(v - \gamma)) dv < \infty. \quad (V)$$

Alors, pour tout $0 < \delta < 1$,

$$p_{nA}(\mathbf{X}_1^k = Y_1^k) = g_{nA}(Y_1^k)(1 + o_{P_{nA}}(\delta_n)) \quad (39)$$

$$p_{nA}(\mathbf{X}_1^k = Y_1^k) = g_{nA}(Y_1^k)(1 + o_{G_{nA}}(\delta_n)) \quad (40)$$

où

$$\delta_n := \max \left(\epsilon_n(\log n)^2, (\exp -nc)^\delta \right). \quad (41)$$

Remarque 20 La plupart des lois utilisées en statistiques utilisent (V). Les propriétés de la fonction variance ont été l'objet de nombreux articles, voir [53] et les références.

Dans le cas plus général où A est une union dénombrable d'intervalles, aucune approximation ne peut être obtenue. Pour simuler des échantillons de la loi de $\mathbf{U}_{1,n}/n$ sous le conditionnement $(\mathbf{U}_{1,n} \in nA)$, nous proposons d'utiliser l'algorithme de Metropolis-Hastings car le ratio suivant

$$r(v, v') := \frac{p(\mathbf{U}_{1,n}/n = v | \mathbf{U}_{1,n} \in nA)}{p(\mathbf{U}_{1,n}/n = v' | \mathbf{U}_{1,n} \in nA)}$$

est indépendant de la quantité inconnue $P(\mathbf{U}_{1,n} \in nA)$. La densité de proposition doit évidemment avoir comme support l'ensemble A .

4.6 Remarque importante

Contrairement à tous les résultats mentionnés précédemment, le résultat présenté dans cette section permet d'étudier des cas plus généraux. En effet, il arrive souvent que l'événement rare s'écrive sous la forme d'une intersection de contraintes, par exemple,

$$u(\mathbf{X}_1) + \dots + u(\mathbf{X}_n) \in nA \quad (42)$$

et

$$h(\mathbf{X}_1^n) \in B_n \quad (43)$$

pour une certaine fonction h et un ensemble mesurable B_n . Clairement, dans la plupart des cas l'approximation de la densité de \mathbf{X}_1^k sous les deux contraintes est impossible à obtenir. Dans les cas précédents, les méthodes proposées permettent de simuler des vecteurs $\mathbf{U}_1^k := (u(\mathbf{X}_1), \dots, u(\mathbf{X}_k))$ sous $(\mathbf{U}_{1,n} \in nA)$. Dans cette section, la méthode permet de simuler un vecteur \mathbf{X}_1^k conditionnée à $(\mathbf{X}_1^n \in \mathcal{E}_n)$ et étendu à \mathbb{R}^n comme indiqué ci-dessus. Ainsi l'estimation de

$$P((u(\mathbf{X}_1) + \dots + u(\mathbf{X}_n) > na_n) \cap (h(\mathbf{X}_1^n) \in B_n)). \quad (44)$$

en est grandement facilité puisqu'elle évite l'inversion de la fonction u .

4.7 Estimateur d'Importance Sampling adaptatif et ses propriétés.

La densité d'IS g_{nA} , qui est un mélange continu de densité g_{nv} comme dans (33) avec une loi de mélange exponentielle de paramètre $nm^{-1}(a)$ sur (γ, ∞)

$$p_{\mathbf{V}}(v) \mathbb{1}_{(\gamma, \infty)}(\gamma) nm^{-1}(a) \exp(-nm^{-1}(\gamma)(v - \gamma)) \quad (45)$$

permet la simulation de $Y_1^k := (Y_1, \dots, Y_k)$.

Nous considérons des versions non tronquées de g_{nA} définie dans (37), en intégrant sur (γ, ∞) à la place de $(\gamma, \gamma + c_n)$. Ce choix évite d'ajouter un nouveau paramètre à notre étude et ne change pas les résultats numériques. Simuler v sous (45) à la place de (38) est légèrement sous optimal mais bien plus simple. Nous gardons la notation g_{nA} pour la version non-tronquée.

La densité g_{nA} est prolongée de \mathbb{R}^k à \mathbb{R}^n en complétant les $n - k$ coordonnées restantes Y_{k+1}, \dots, Y_n de manière i.i.d. sous la densité tiltée suivante

$$g_{nA}(y_{k+1}^n | y_1^k) := \prod_{i=k+1}^n \pi_u^{m_k}(y_i) \quad (46)$$

avec $m_k := m(t^k) = \frac{n}{n-k} \left(v - \frac{u_{1,k}}{n} \right)$ définie de la même manière que dans (29) et

$$u_{1,k} = \sum_{i=1}^k u(y_i).$$

Ainsi, Y_{k+1}, \dots, Y_n sont simulés en accord avec l'ISEI introduit dans la section 3. Nous pouvons à présent définir notre estimateur de P_n . Soit $Y_1^n(l) := Y_1(l), \dots, Y_n(l)$ simulés sous g_{nA} . Nous définissons

$$\widehat{P}_n(l) := \frac{\prod_{i=0}^n p_{\mathbf{X}}(Y_i(l))}{g_{nA}(Y_1^n(l))} \mathbb{1}_{\mathcal{E}_n}(Y_1^n(l)) \quad (47)$$

ainsi,

$$\widehat{P}_n := \frac{1}{L} \sum_{l=1}^L \widehat{P}_n(l). \quad (48)$$

Pour obtenir un échantillon de densité g_{nA} , nous proposons une procédure en plusieurs étapes :

- Simuler v de densité $p_{\mathbf{V}}$ définie dans (45).
- Simuler Y_1, \dots, Y_k de densité g_{nv} définie par (33).
- Simuler Y_{k+1}, \dots, Y_n de manière i.i.d. de densité (46).
- Itérer l'opération L fois.

Il reste maintenant à étudier les propriétés d'efficacité de notre estimateur. Nous ne pouvons étudier directement l'ordre de la variance puisque notre approximation a été obtenu uniquement sur les chemins typiques et pas sur tout l'espace entier \mathbb{R}^n . Nous obtenons l'ordre du coefficient de variation sur une classe de sous ensembles de \mathbb{R}^n dont la probabilité tends vers 1 sous la densité g_{nA} . Ainsi, la différence de performance entre la densité d'échantillonnage optimal et la densité d'IS proposée est du au $n - k$ simulations i.i.d. restantes, mène à un coefficient de variation de l'ordre de $\sqrt{n - k}$ à la place de \sqrt{n} dans le cas de la méthode d'ISEI. Ces résultats sont démontrés dans [14] sous des conditions de régularité de la densité de \mathbf{X} .

5 Comparaison sur un exemple

Nous allons comparer les deux méthodes développées dans cet article sur un exemple simple tiré de [39].

On considère un échantillon aléatoire X_1, \dots, X_{100} i.i.d. où X_1 suit une loi normale de moyenne 0.05 et de variance

1. Définissons

$$\mathcal{E}_{100} := \left\{ x_1^{100} : \frac{|x_1 + \dots + x_{100}|}{100} > 0.28 \right\}$$

pour lequel

$$P_{100} = P((X_1, \dots, X_{100}) \in \mathcal{E}_{100}) = 0.01120.$$

Dans cet exemple dissymétrique, l'ISEI ignore une partie de l'ensemble \mathcal{E}_{100} tandis que la méthode développée dans la section 4 permet l'obtention d'un bon estimateur. En utilisant l'ISEI, le point dominant de l'ensemble considéré est $\gamma = 0.28$ et la densité d'échantillonnage est $\mathcal{N}(\gamma, 1)$. La valeur de l'estimateur P_{100} résultant est 0,01074 (avec $L = 1000$). Cette valeur nous indique que l'événement $S_{1,100}/100 < -0.28$ est ignoré dans l'évaluation de P_{100} . En effet, cet événement n'est jamais visité durant la procédure. La raison pour laquelle cette méthode fonctionne mal est la présence de deux points à taux minimaux. Evidemment, nous pourrions proposer de diviser l'événement en deux sous événements, puis traiter chacun de façon indépendante, ou de proposer un mélange de lois (voir [11]). Dans le cas traitée ici, cette méthode permettrait d'obtenir de bon résultats. Cependant, si l'événement est plus compliqué, trouver des points dominants n'est pas toujours évident. C'est la raison pour laquelle la méthode d'approximation obtenue précédemment est la plus à même de fonctionner.

En effet, explorer l'ensemble d'intérêt sous la loi de $(x_1 + \dots + x_{100})/100$ conditionnée à \mathcal{E}_{100} permet d'éviter la recherche de points dominants. Nous simulons L points i.i.d. v_1, \dots, v_L selon la loi de $S_{1,100}/100$ conditionnellement $|S_{1,100}|/100 > 0.28$. Nous évaluons alors P_{100} avec $k = 99$ car, dans le cas gaussien, le résultat du Théorème 16 est exact pour le conditionnement ponctuel, et ainsi l'approximation fourni par le Théorème 19 est très fine. La valeur obtenue de l'estimateur est 0.01125 pour $L = 1000$, ce qui est très proche de P_{100} .

Nous proposons à présent d'étudier le gain en erreur relative quand la dimension de l'ensemble d'intérêt augmente. Notons $B := (\mathcal{E}_{100})^d$. Les variables aléatoires X_1, \dots, X_n sont des vecteurs aléatoires de dimension d de loi $\mathcal{N}(0.05, 1)$. Le point dominant de B est le point avec toutes ces coordonnées égale à 0.28. Nous observons sur la figure suivante que le phénomène décrit plus haut est amplifié par l'augmentation de la dimension. La figure suivante montre le gain en erreur relative de la méthode présentée dans la section 4 par rapport à l'ISEI en fonction de d . La valeur de P_{100} associée est 10^{-2d} .

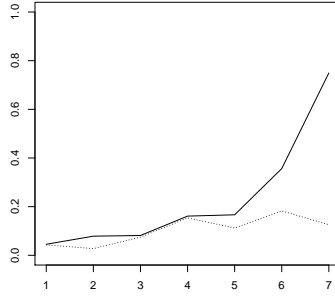


FIGURE 1 – Erreur relative de l’estimateur adaptatif (ligne pointillée) par rapport à l’ISEI (ligne continue) comme fonction de la dimension d pour $L = 1000$.

Références

- [1] Asmussen, S. (1989). Risk theory in a Markovian environment. *Scand. Actuarial J.*, 69-100.
- [2] Asmussen, S. (2000). Ruin Probabilities, *World Scientific Publishing Co. Ltd. London*.
- [3] Asmussen, S. (2003). Applied Probability and Queues, *Springer-Verlag, New York*.
- [4] Blanchet, J.H., Leder K. and Glynn, P. W. (2009). Efficient simulation of light-tailed sums : an old-folk song sung to a faster new tune. . . . *Monte Carlo and quasi-Monte Carlo methods 2008*, P. L’Ecuyer and A.B. Owen Eds, Springer-Verlag, 227-248.
- [5] Blanchet, J. and Glynn, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *The Annals of Applied Probability*, 18, 4, 1351-1378.
- [6] Blanchet, J., and H. Lam (2011). State-dependent importance sampling for rare-event simulation : An overview and recent advances. *Surveys in Operations Research and Management Sciences*, 17, 38-59.
- [7] Blanchet, J. and Glynn, P. (2006). Strongly efficient estimators for light-tailed sums, in *Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*.
- [8] Blanchet, J. and Liu, J.C. (2008). State-dependent importance sampling for regularly varying random walks. *Adv. in Appl. Probab.*, 40(4) :1104-1128.
- [9] Barndorff-Nielsen, O. (1978). Information and Exponential Families in Statistical Theory. *Wiley, New-York*.
- [10] Botev, Z.I. and Kroese, D.P. (2010). Efficient Monte Carlo simulation via the Generalized Splitting Method. *Statistics and Computing*, 22(1), 1-16.
- [11] Brereton, T.J., Chan, J.C.C. and Kroese, D.P., (2011). Fitting mixture importance sampling distributions via improved cross-entropy. *Proceedings of the 2011 Winter Simulation Conference, Phoenix, AZ*.
- [12] Broniatowski M. and Caron V., (2011). Long runs under a conditional limit distribution. arxiv :1010.3616.
- [13] Broniatowski M. and Caron V., (2012). Conditional inference in parametric models. arXiv :1202.0944.
- [14] Broniatowski M. and Caron. V., (2013). Small variance estimators for rare event probabilities. *ACM Transactions on Modeling and Computer Simulation*, 23(1) :Article 6, 2013.
- [15] Bucklew, J.A. (2004). Introduction to rare event simulation. Springer Series in Statistics, Springer-Verlag, New York.
- [16] Bucklew, J.A., Ney, P., and Sadowsky, J.S., 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Ann. Probab.*, 27 :44-59.
- [17] Cappé, O., Guillin, A., Marin, J.M. and Robert, C.P. (2004). Population Monte Carlo, *Journ. of Comp. and Graph. Stat.*, 13(4), 907-929
- [18] Caron, V. (2012). A conditional limit theorem. Applications to conditional inference and Importance Sampling methods. *tel-00763369*.

- [19] Chang, C. S., P. Heidelberger, S. Juneja and P. Shahabuddin (1994). Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation* 20, 45-65.
- [20] Cornuet, J.M., Marin, J.M., Mira A. and Robert, C.P. (2012). Adaptive Multiple Importance Sampling, *Scandinavian Journal of Statistics*, 39(4), 798-812
- [21] Cottrel, M., Fort, J.C., and Malgouyres, G. (1983). Large deviations and rare event in the study of stochastic algorithms. *IEEE Trans. Automat. Control*, AC-28 :907-920.
- [22] Csizsár, I. (1984). Sanov property, generalized I -projection and a conditional limit theorem. *Ann. Probab.*, 12, 768–793.
- [23] Dean, T. and Dupuis, P. (2009). Splitting for rare event simulation : A large deviation approach to design and analysis. *Stoch. Proc. App.* 119 :562–587
- [24] de Boer, P. T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1) :19-67.
- [25] Dembo, A. and Zeitouni, O. (1996). Refinements of the Gibbs conditioning principle. *Probab. Theory Related Fields* **104** 1-14.
- [26] Dembo, A., and Zeitouni O. (1998). Large Deviations Techniques and Applications. *Springer, New York, NY*.
- [27] den Hollander, W.Th.F. and Weiss, G.H. (1988). On the range of a constrained random walk. *J. Appl. Probab.* **25** 451–463.
- [28] den Hollander, W.Th.F. and Weiss, G.H. (1988). A note on configurational properties of constrained random walks. *J. Phys. A* **21** 2405–2415.
- [29] Dey, S., Juneja, S. and Agarwal, A. (2012). Efficient simulation of density and probability of large deviations of sum of random vectors using saddle point representations. arXiv :1203.0817.
- [30] Diaconis, P. and Freedman, D.A. (1988). Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. *J Theoret Probab* 1(4), 381-410.
- [31] Dieker, A.B. and Mandjes M. (2005). On asymptotically efficient simulation of large deviation probabilities. *Advances in Applied Probability*, Vol. 37, No 2, 539-552.
- [32] Dupuis, P. and Ellis, R.S. (1997). A weak convergence approach to the theory of large deviations. John Wiley & Sons, New-York.
- [33] Dupuis, P., Leder, K. and Wang, H. (2007). Importance Sampling for Sums of Random Variables with Regularly Varying Tails. *ACM Trans. Model. Comput. Simul.* 17(3) :Article 14.
- [34] Dupuis, P., Sezer, A.D. and Wang, H. (2007). Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.* 17(4) :1306-1346.
- [35] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. Stoch. Rep.*, 76 :481-508.
- [36] Ellis, R.S. (1985). Entropy, Large Deviations and Statistical Mechanics. *Springer-Verlag, New-York*
- [37] Ermakov, M. (2007). Importance sampling for simulations of moderate deviation probabilities of statistics. *Statist Decisions* 25(4), 265-284.
- [38] Fernholz, L. T. (1983). Von Mises Calculus for Statistical Functionals. *Lecture Notes in Statistics* 19, New York : Springer-Verlag.
- [39] Glasserman, P. and Wang, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.* 7(3), 731-746.
- [40] Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. (1999). Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47, 4, 585-600.
- [41] Glynn P.W. and Whitt W. (1992). The asymptotic efficiency of simulation estimators, *Oper. Res.*, 40, 505–520.
- [42] Goyal, A., P. Shahabuddin, P. Heidelberger, V.F. Nicola and P.W. Glynn. 1992. A unified framework for simulating Markovian models of highly reliable systems. *IEEE Transactions on Computers*, C-41, 36-51.
- [43] Hammersley, J.M. and Handscomb, D.C. (1964). Monte Carlo Methods. *Chapman and Hall, New-York*.
- [44] Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5, 1, 43-85.
- [45] Jensen, J.L. (1995). Saddlepoint Approximations. *Oxford Statistical Science Series, vol 16*. The Clarendon Press Oxford University Press, New York, Oxford Science Publications.

- [46] Juneja, S. and Shahabuddin, P. (2006). Rare event simulation techniques : An introduction and recent advances. In *Handbook on Simulation* (S. Henderson and B. Nelson, eds) 291–350. North-Holland, Amsterdam.
- [47] Kahn H. and Harris T. E. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards Appl. Math. Series*, 12 :27-30.
- [48] L'Ecuyer, P., F. L. Gland, P. Lezard, and B. Tuffin (2009). Splitting techniques. In *Rare Event Simulation Using Monte Carlo Methods*, edited by G. Rubino and B. Tuffin, 39-62. West Sussex, United Kingdom : Wiley. Chapter 3.
- [49] L'Ecuyer P., Mandjes M. and Tuffin B. (2009). Importance sampling in rare event simulation, in *Rare Event Simulation Using Monte-Carlo Methods*, Editors Rubino G. and Tuffin B., Wiley, Chichester.
- [50] L'Ecuyer P., Blanchet J. , Tuffin B. and Glynn P. (2010). Asymptotic Robustness of Estimators in Rare-Event Simulation. *ACM Transactions on Modeling and Computer Simulation* (TOMACS), 20(1), 6 :1-6 :41.
- [51] L'Ecuyer P., Rubino G., Saggadi S. and Tuffin B. (2011). Approximate zero-variance importance sampling for static network reliability estimation. *IEEE Transactions on Reliability* 8(4), 590-604.
- [52] Lehtonen, T., and Nyrhinen H. (1992). Simulating level-crossing probabilities by importance sampling. *Advances in Applied Probability* 24, 858-874.
- [53] Letac, G. et Mora, M. (1990). Natural real exponential families with cubic variance functions. *Ann. Statist.* 18, no. 1, 1-37.
- [54] Ney, P. (1983). Dominating points and the asymptotics of large deviations for random walks on \mathbb{R}^d . *Annals of Probability* 11, 158-167.
- [55] Parekh, S., and J. Walrand (1989). A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34, 1, 54-66.
- [56] Rubinstein, R.Y. (1981). Simulation and the Monte-Carlo method. *John Wiley & Sons*, New-York, United States.
- [57] Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99 :89-112.
- [58] Rubinstein, R. Y. (1999). Rare-event simulation via cross-entropy and importance sampling. *Second Workshop on Rare Event Simulation, RESIM 99*, 1–17.
- [59] Rihter, V. (1957). Local limit theorems for large deviations. *Dokl Akad Nauk SSSR (NS)* : 115 :53-56.
- [60] Sadowsky, J.S. and Bucklew, J.A. (1990). On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans Inform Theory* 36(3), 579-588.
- [61] Sadowsky, J.S. (1991). Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Transactions on Automatic Control* 36, 12, 1383–1394.
- [62] Sadowsky J.S. (1993). On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Trans. Inform. Theory*, 39, 119–128.
- [63] Sadowsky, J. S. (1996). On Monte Carlo estimation of large deviations probabilities. *The Annals of Applied Probability* 6, 2, 399-422.
- [64] Siegmund, D. (1976). Importance sampling in the Monte-Carlo study of sequential tests. *Annals of Statistics*, 4, 673-684.
- [65] Van Campenhout, J.M. and Cover, T.M. (1981) Maximum entropy and conditional probability. *IEEE Trans. Inform. Theory* **27** 483–489.