Arnaud Guyader
Rennes 2, INRIA et IRMAR

GdT Théorie des Valeurs Extrêmes
Décembre 2013

Simulation d'événements rares

Bibliographie

1 Introduction

Les deux principales méthodes de réduction de variance pour la simulation d'événements rares sont l'échantillonnage préférentiel (*Importance Sampling*) et les méthodes multi-niveaux (*Multilevel Splitting*, ou *Importance Splitting*, ou encore parfois *Subset Simulation*). Tout porte à croire qu'elles ont été proposées à peu près en même temps (fin des années 40) en deux endroits différents et pour répondre à différents problèmes en physique des particules : à Los Alamos par Ulam et von Neumann dans la foulée du projet Manhattan, et à Santa Monica par Kahn et Harris dans le cadre du projet RAND.

- S. Ulam and J. von Neumann (+ R.D. Richtmyer). Statistical Methods in Neutron Diffusion, *Report LAMS-551*, Los Alamos Scientific Laboratory, New Mexico, 1947.
- H. Kahn and T.E. Harris. Estimation of Particle Transmission by Random Sampling, *National Bureau of Standards, Appl. Math. Series*, 12: 27-30, 1951.

La suite de cette bibliographie propose surtout quelques références théoriques sur ces deux familles de méthodes, lesquelles ont par ailleurs donné lieu à des applications dans beaucoup de domaines : télécommunications, dynamique moléculaire, mécanique probabiliste, contamination alimentaire, dénombrement, etc. Notons aussi que l'accent est mis sur les méthodes multi-niveaux, ce qui ne reflète pas du tout l'état de l'art, les contributions méthodologiques étant pour l'heure bien plus nombreuses en échantillonnage préférentiel.

2 Importance Sampling

L'échantillonnage préférentiel n'est pas propre à l'estimation d'événements rares. C'est une méthode générale de réduction de variance. De fait, on la trouve exposée dans tout ouvrage traitant de la simulation Monte-Carlo, par exemple en Section 3.3 de :

- C. Robert and G. Casella. *Monte Carlo Statistical Methods*, Springer, 2004.

En estimation d'événements rares proprement dit, elle a donné lieu à un certain nombre de développements théoriques en raison du lien avec la théorie des grandes déviations. Une référence classique sur le sujet :

- J. Bucklew. *Introduction to Rare Event Simulation*, Springer, 2004.

On pourra aussi se reporter au Chapitre 6 de :

- S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, Springer, 2007, ainsi qu'au Chapitre 2 de :

– G. Rubino and B. Tuffin (Editors). *Rare Event Simulation using Monte Carlo Methods*, Wiley, 2009.

3 Multilevel Splitting

Si la méthode multi-niveaux pour événements rares a été appliquée dans différents domaines depuis l'article de Kahn et Harris en 1951, on peine à trouver le moindre travail théorique durant près de 50 ans. En particulier, le lien avec les processus de branchement était déjà mentionné dans leur article, mais il a fallu attendre la fin des années 90 pour qu'il soit dûment explicité :

– P. Glasserman, P. Heidelberger, P. Shahabuddin and T. Zajic. Multilevel splitting for estimating rare event probabilities, *Operations Research*, 47(4), 585-600, 1999.

Cette approche est aussi détaillée dans la thèse suivante :

– A. Lagnoux Renaudie. *Analyse des modèles de branchement avec duplication des trajectoires pour l'étude des événements rares*, Université Toulouse III, 2006.

Pour garder un nombre de trajectoires constant au cours du temps et par là un contrôle de la complexité algorithmique, une variante est ensuite apparue, faisant le lien avec les méthodes de Monte-Carlo en interaction et le cadre théorique des formules de Feynman-Kac :

– F. Cérou, P. Del Moral, F. Le Gland and P. Lezaud. Genetic Genealogical Models in Rare Event Analysis, *ALEA Latin American Journal of Probability and Mathematical Statistics*, 181-203, 2006.

Une méthode adaptative pour fixer les niveaux au fil de l'eau est alors proposée et analysée :

– F. Cérou and A. Guyader. Adaptive Multilevel Splitting for Rare Event Analysis, *Stochastic Analysis and its Applications*, Vol. 25, 417-443, 2007.

Les travaux précédents ont pour point commun de simuler et d'analyser des événements rares pour des processus stochastiques, autrement dit dans un cadre dynamique. Dans un cadre statique, on cherche plutôt à évaluer une probabilité de la forme $\mathbb{P}(S(X) > q)$ où, typiquement, X est un vecteur aléatoire simulable à valeurs dans $\mathbb{R}^d$, $S : \mathbb{R}^d \to \mathbb{R}$ une fonction boîte noire et q un réel se situant dans la queue de la distribution de la variable aléatoire $S(X)$. L'introduction de méthodes multi-niveaux dans ce contexte est plus récente. L'une des premières références viendrait de la mécanique probabiliste :

– S.K. Au and J.L. Beck. Estimation of Small Failure Probabilities in High Dimensions by Subset Simulation, *Journal of Probabilistic Engineering Mechanics*, Vol. 16, 263-277, 2001.

Noter que, d'une part, les auteurs dénomment *Subset Simulation* ce qu'on appelle usuellement *Multilevel Splitting* et, d'autre part, qu'aucune référence n'est faite à l'article de Kahn et Harris. Ceci explique que les papiers de Au et Beck ne furent pas toujours cités par la suite.

Toujours dans le cadre statique, l'algorithme de la dernière particule a été introduit et analysé plus récemment. Tous les résultats présentés en groupe de travail sont détaillés dans l'article suivant :

– A. Guyader, N. Hengartner and E. Matzner-Løber. Simulation and Estimation of Extreme Quantiles and Extreme Probabilities, *Applied Mathematics & Optimization*, Vol. 64 (2), 171-196, 2011.

Comme l'a fait remarquer l'un des rapporteurs de cet article, il existe en fait un algorithme voisin pour la comparaison de modèles en statistique bayésienne, connu sous le nom de *Nested Sampling* et qui fut introduit en 2006 :

– J. Skilling. Nested Sampling for General Bayesian Computation, *Bayesian Analysis*, 1 (4), 833-859, 2006.

Une analyse asymptotique de ses propriétés a par ailleurs été proposée peu après :

– N. Chopin and C. Robert. Properties of Nested Sampling, *Biometrika*, Vol. 97 (3), 741-755, 2010.

4 Conclusion

A l'heure actuelle, le seul ouvrage sur la simulation d'événements rares présentant à la fois l'échantillonnage préférentiel et les méthodes multi-niveaux, que ce soit d'un point de vue méthodologique ou applicatif, est le suivant :

– G. Rubino and B. Tuffin (Editors). *Rare Event Simulation using Monte Carlo Methods*, Wiley, 2009.

Pour un exemple d'algorithme combinant les deux méthodes, on pourra se reporter à :

– P. Del Moral and J. Garnier. Genealogical Particle Analysis of Rare Events, *Annals of Applied Probability*, Vol. 15, 2496-2534, 2005,

ou encore au second jeu de transparents suivants :

– J. Garnier. *Rare Events: Models and Simulations* [1], *Interacting Particle Systems for the Analysis of Rare Events* [2], CEMRACS 2013.

Au-delà de l'échantillonnage préférentiel et des méthodes multi-niveaux, cet exposé offre une vue d'ensemble sur la simulation d'événements rares. Ainsi, il évoque également les techniques à base de métamodèles, de stratification, etc.

Pour finir, il reste à mentionner la méthode de l'entropie croisée (ou *Cross-entropy method*) qui, par une approche paramétrique, permet d'améliorer de façon itérative la loi d'échantillonnage en Importance Sampling :

– R.Y. Rubinstein and D.P. Kroese. *The Cross-Entropy Method*, Springer, 2004.

Initialement proposée pour l'estimation d'événements rares, elle a ensuite été appliquée à des problèmes d'optimisation et de dénombrement. Cette généralisation n'est pas propre à la méthode d'entropie croisée, mais consiste simplement à réécrire un problème d'optimisation ou de dénombrement en terme d'événement rare. Supposons par exemple qu'on veuille maximiser une fonction f définie sur $[0,1]$. De façon approchée, ceci revient à trouver M tel que la probabilité $\mathbb{P}(f(X) \geq M)$ soit très faible, avec X suivant par exemple une loi uniforme sur $[0,1]$.