

REVUE DE STATISTIQUE APPLIQUÉE

N. F. STEWART

Sur une méthode de Monte-Carlo adaptative

Revue de statistique appliquée, tome 21, n° 4 (1973), p. 53-58

http://www.numdam.org/item?id=RSA_1973__21_4_53_0

© Société française de statistique, 1973, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

SUR UNE MÉTHODE DE MONTE-CARLO ADAPTATIVE (1)

N.F. STEWART

Professeur adjoint, Université de Montréal

This paper is concerned with a gradient technique of adaptive Monte Carlo. It is shown how to increase efficiency by rearranging the order of the steps in the algorithm and modifying the estimator of the direction of steepest descent.

Cet article traite d'une méthode de Monte Carlo adaptative qui emploie un estimateur du gradient de la variance. Nous montrons comment améliorer l'efficacité de la méthode en modifiant l'ordre des étapes de l'algorithme et en modifiant l'estimateur du gradient.

1. – INTRODUCTION

Une méthode bien connue qui réduit l'erreur d'une estimation de Monte Carlo est la méthode d'échantillonnage suivant l'importance [Réf. 2, p. 74]. A part la réduction de l'erreur, cette méthode permet parfois d'utiliser plusieurs fois le même échantillon [Réf. 1], mais c'est la réduction de l'erreur qui nous intéresse ici.

La méthode d'échantillonnage suivant l'importance (et d'ailleurs, plusieurs autres méthodes qui ont pour but la réduction de la variance), peut être incorporée dans un algorithme dit adaptatif, c'est-à-dire, un algorithme qui utilise l'information contenue dans l'échantillon afin d'améliorer la méthode d'échantillonnage elle-même. Une façon de le faire [Réf. 3] est d'utiliser un estimateur du gradient de la variance. Le but du présent article est de décrire une modification de cet algorithme qui est plus efficace [Réf. 2, p. 65].

2. – L'ALGORITHME ADAPTATIF

On cherche un estimateur pour

$$\xi = \int \Phi \, d\mu, \quad (1)$$

(1) Article remis le 5/9/72.

où le domaine de définition de la fonction Φ est un sous-ensemble d'un espace Euclidien, et μ est une mesure sur le domaine de Φ . Soit ν une mesure de probabilité telle que μ est absolument continue par rapport à ν . Alors

$$\xi = \int \Phi \frac{d\mu}{d\nu} d\nu,$$

et si x_1, x_2, \dots, x_n sont des observations indépendantes d'une variable x suivant la loi ν , alors

$$\hat{\xi} = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \frac{d\mu}{d\nu}(x_i)$$

est un estimateur non-biaisé de ξ . Nous voudrions choisir ν de façon telle que $\text{Var}(\hat{\xi})$ soit aussi petite que possible, ou ce qui l'équivaut, telle que

$$\Gamma(\nu) = \int \Phi^2 \left\{ \frac{d\mu}{d\nu} \right\}^2 d\nu \quad (2)$$

soit aussi petit que possible. C'est la méthode appelée "échantillonnage suivant l'importance".

Pour utiliser le gradient [Réf. 3], on considère une suite de distributions $\nu_1, \nu_2, \dots, \nu_{k_0}$, où ν_k est déterminée par $\alpha^{(k)} = (\alpha_1^{(k)}, \dots, \alpha_m^{(k)})$, $k = 1, \dots, k_0$. La distribution initiale ν_1 est choisie de façon arbitraire. Pour $1 \leq k \leq k_0 - 1$, on prend n observations indépendantes x_{k1}, \dots, x_{kn_k} issues de ν_k , et on les emploie pour calculer, d'une part un estimateur non-biaisé de ξ , à savoir

$$\hat{\xi}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \Phi(x_{ki}) \frac{d\mu}{d\nu_k}(x_{ki}),$$

et d'autre part un estimateur non-biaisé de

$$V_j = - \int \Phi^2 \left[\frac{\partial}{\partial \alpha_j^{(k+1)}} \frac{d\mu}{d\nu_{k+1}} \right]_{\alpha^{(k)}} \cdot d\mu,$$

$$j = 1, \dots, m$$

à savoir

$$\hat{V}_j = - \frac{1}{n_k} \sum_{i=1}^{n_k} \Phi^2(x_{ki}) \left[\frac{\partial}{\partial \alpha_j^{(k+1)}} \cdot \frac{d\mu}{d\nu_{k+1}} \right]_{\alpha^{(k)}}(x_{ki}) \frac{d\mu}{d\nu_k}(x_{ki})$$

Remarquons que V_j est la j^{me} composante du gradient de la fonction Γ . On obtient alors la distribution ν_{k+1} à partir de l'équation

$$\alpha^{(k+1)} = \alpha^{(k)} + \frac{\delta}{\|\hat{V}\|} (\hat{V}_1, \dots, \hat{V}_m)$$

où δ est une petite constante.

On continue de la même façon jusqu'au moment où on a trouvé $\hat{\xi}_{k_0}$; finalement, on prend comme estimateur non-biaisé de ξ la somme pondérée

$$\sum_{k=1}^{k_0} a_k \hat{\xi}_k ; \quad (3)$$

où $\sum_{k=1}^{k_0} a_k = 1$, $a_k > 0$, $k = 1, \dots, k_0$. Notons G l'algorithme que nous venons de décrire.

3. – MODIFICATION DE L'ALGORITHME

Nous proposons que la recherche du minimum relatif de Γ soit effectuée en employant uniquement les n_1 premières observations x_{11}, \dots, x_{1n_1} , et en employant

$$-\frac{1}{n_1} \sum_{i=1}^{n_1} \Phi^2(x_{1i}) \left[\frac{\partial}{\partial \alpha_j^{(k+1)}} \frac{d\mu}{d\nu_{k+1}} \right]_{\alpha^{(k)}}(x_{1i}) \frac{d\mu}{d\nu_1}(x_{1i}), \quad (4)$$

comme estimateur de V_j , $j = 1, \dots, m$. Donc, les n_1 premières observations seront employées pour trouver et $\hat{\xi}_1$, et la suite de distributions

$$\nu'_1 = \nu_1, \nu'_2 = \nu_2, \nu'_3, \nu'_4, \dots, \nu'_{k'_0},$$

où on obtient ν'_{k+1} de ν'_k en utilisant (4), $k = 1, \dots, k'_0 - 1$. La distribution $\nu'_{k'_0}$ pourra être utilisée pour le reste de l'échantillonnage. Bien entendu, on peut terminer l'échantillonnage quand on veut – par exemple, il n'est pas nécessaire d'utiliser $\sum_{k=1}^{k'_0} n_k$ observations. Notons N le nombre d'observations tirées de $\nu'_{k'_0}$, et notons GM l'algorithme modifié.

Afin de montrer nettement la différence entre les deux algorithmes, décrivons chaque algorithme pour des valeurs fixes des paramètres. Supposons que l'on ait choisi $k_0 = 20$, et $n_k = 50$ pour $k = 1, \dots, 20$ (comme dans l'exemple de la Réf. [3] ; voir le § 4, ci-dessous). Avec l'algorithme G, pour $k = 1, \dots, 20$, on tire 50 observations de ν_k , et on les utilise pour trouver $\hat{\xi}_k$ et \hat{V} . Alors, on trouve ν_{k+1} , et on répète la même chose pour $k + 1$. Au total, 1 000 observations sont utilisées pour l'estimateur de ξ .

Par contre, pour l'algorithme GM, on tire 50 observations de ν_1 . Elles sont utilisées pour trouver $\hat{\xi}_1$, et les mêmes 50 observations sont utilisées pour trouver $\nu'_2, \nu'_3, \dots, \nu'_{k'_0}$. Alors, on tire N observations de $\nu'_{k'_0}$, et on emploie $50 + N$ observations pour l'estimateur de ξ .

Si $k'_0 = 20$ et $N = 950$, les deux algorithmes exigent le même travail ; c'est uniquement l'ordre des étapes qui est différent. Cependant l'estimateur indiqué dans (4) est un estimateur biaisé de V_j , parce que $\alpha^{(k)}$ dépend de x_{11}, \dots, x_{1n_1} ;

donc, la recherche de l'optimum de distribution sera moins sûre. Mais notre méthode a l'avantage que l'on peut utiliser ν'_{k_0} pour toutes les observations, à part les n_1 premières, tandis que la méthode G ne peut utiliser qu'une distribution ν_k qui s'améliore au fur et à mesure que k s'accroît. Soulignons que ce n'est que l'estimateur de V_j qui est biaisé :

$$a_1 \xi_1 + \sum_{k=2}^{k_0} \frac{a_k}{n_k} \sum_{i=1}^{n_k} \Phi(x_{k_0 q}) \frac{d\mu}{d\nu'_{k_0}}(x_{k_0 q}), \quad (5)$$

où $q = \sum_{j=2}^{k-1} n_j + i$ et où $x_{k_0 s}$ est une observation issue de ν'_{k_0} , $s = 1, \dots, N$, reste un estimateur non-biaisé de ξ .

4. — COMPARAISON POUR k_0 , δ ET n_k FIXES

Supposons que k_0 , δ et n_k , $k = 1, \dots, k_0$, soient fixés au début de l'algorithme, et supposons que $N = \sum_{k=2}^{k_0} n_k$. Alors, le travail exigé par notre modification GM n'est pas plus grand que celui exigé par G si $k_0 = k'_0$. (En fait, si évaluer Φ implique beaucoup de travail, le travail exigé par GM n'est que négligeablement plus grand que celui exigé par G même si $k'_0 > k_0$). Donc, si l'estimateur indiqué dans (5) est finalement utilisé comme l'estimateur non-biaisé de ξ , alors chacune des suivantes est une condition suffisante pour que GM soit plus efficace que G :

$$\text{i) } \Gamma(\nu'_{k_0}) \leq \Gamma(\nu_k), \text{ pour } k = 2, \dots, k_0, \text{ et} \quad (6)$$

$$\Gamma(\nu'_{k_0}) < \Gamma(\nu_k) \text{ pour quelque } k.$$

$$\text{ii) } a_k \text{ et } n_k = r \text{ ne dépendent pas de } k, k = 2, \dots, k_0, \text{ et}$$

$$\frac{1}{r} \text{Var} \left[\Phi(X_{k_0}) \frac{d\mu}{d\nu'_{k_0}}(X_{k_0}) \right] < \frac{1}{(k_0 - 1)} \sum_{k=2}^{k_0} \text{Var}(\hat{\xi}_k), \quad (7)$$

où X_{k_0} est une variable qui suit la loi ν'_{k_0} .

On peut s'attendre à la validité de (7).

Dans le § 4 de la Réf. [3] on trouvera les résultats d'une expérience qui implique la durée T de la vie d'une certaine machine, pour laquelle les distributions ν dépendent de trois paramètres. Les valeurs des constantes sont

$$k_0 = 20, n_k = 50, \delta = .01,$$

et donc, 1 000 observations sont employées. Comme il est dit dans la réf. [3], l'écart-type de l'estimateur ordinaire de T est .0122 quand on emploie 1 000 observations. Cependant, la comparaison dans la Réf. [3], entre l'estimateur ordinaire et celui de G nous égare : il faut [Réf. 2] une comparaison entre

.0122 et une approximation à l'écart-type de la méthode G, et non pas une comparaison entre .0122 et l'erreur entraînée par la méthode G pour un choix particulier d'observations.

Nous avons fait l'expérience décrite dans la Réf. [3], et nous avons trouvé $s(G) = .0094$, où s^2 est l'estimateur ⁽¹⁾ non-biaisé de la variance de l'estimateur de G. Bien sûr, la suite de distributions ν_1, \dots, ν_{20} était légèrement différente de la suite donnée dans la Réf. [3], parce que nous avons certainement utilisé une génération différente de nombres pseudo-aléatoires. Cependant, nous avons tiré 50 observations de chacune des distributions données dans la Réf. [3], et nous avons calculé la somme pondérée qui y est utilisée. Encore une fois nous avons trouvé $s(G) = .0094$. En utilisant cette dernière suite de distributions avec $a_k = .05$, $k = 1, \dots, 20$, au lieu de la somme pondérée, nous avons trouvé $s(G) = .0093$.

En utilisant la méthode GM avec $k' = 20$, $n_k = 50$, $N = 950$ et

$$a_k = .05, k = 1, \dots, 20,$$

nous avons trouvé $s(GM) = .0085$. Donc l'efficacité de GM par rapport à G est approximativement 1.20 dans le cas du problème étudié ici.

La méthode GM exige plus de mots-mémoire que G. Par exemple, pour le problème que nous venons de décrire, il fallait 200 mots-mémoire pour emmagasiner $x_{11}, \dots, x_{1n_1}, \Phi(x_{11}), \dots, \Phi(x_{1n_1})$, afin de les utiliser plus tard pour trouver les v'_k .

Les expériences décrites ci-dessus étaient réalisées sur un ordinateur type IBM 360, avec la génération de nombres pseudo-aléatoires

$$\begin{aligned} x_0 &= 314\,159 \\ x_{i+1} &= 65\,539 \cdot x_i \pmod{2^{32}}. \end{aligned}$$

5. — REMERCIEMENTS

L'auteur voudrait remercier K. Hastings pour ses commentaires très utiles. Bien entendu, toute responsabilité pour des erreurs dans l'article incombe à l'auteur.

Cette recherche était subventionnée par le Conseil National de Recherche du Canada. Une partie de la recherche a été faite pendant que l'auteur faisait un séjour au L.A.A.S. (C.N.R.S.) à l'Université Paul Sabatier à Toulouse, séjour subventionné par la Sous-commission Franco-Québécoise à la Recherche Scientifique et Technologique.

(1) Il s'agit bien entendu de l'estimateur au sens de l'erreur quadratique moyenne minimale.

BIBLIOGRAPHIE

- [1] EVANS D.H. — Applied Multiplex Sampling (*Technometrics* Vol. 5, 1963, pp. 341-359).
- [2] HAMMERSLEY J.M. and HANDSCOMB D.C. — *Les Méthodes de Monte-Carlo* (Dunod, 1967).
- [3] PUGH E.L. — A gradient technique of adaptive Monte Carlo (*SIAM Review* Vol. 3, 1966, pp. 346-355).