

Récents développements et problèmes dans le domaine de la génération aléatoire

Herbert. S. Wilf*

University of Pennsylvania, Philadelphia, Pa 19104-6395, USA

Je suis très heureux d'être parmi vous aujourd'hui, et d'observer les progrès en cours dans le domaine de la sélection aléatoire d'objets combinatoires. J'essayerai de vous lire cette présentation en français. J'espère sincèrement éviter de multier votre belle langue. Cela fait presque vingt ans que j'ai parlé ici à Bordeaux. J'étais venu ici avec le Professeur Schützenberger, et j'ai parlé de la génération des partages d'un entier donné, uniformément au hasard. La présentation d'aujourd'hui est dédiée à Marcel Shützenberger, avec respect et affection.

J'aimerais tout d'abord vous donner certains renseignements autobiographiques afin de vous expliquer comment j'en suis arrivé à m'intéresser à ce sujet. Avant de devenir combinatorialiste, dans les années soixante, je travaillais pour un cabinet de conseils dont les réacteurs possèdent la configuration géométrique suivante (Fig. 1).

Au centre, les neutrons sont créés à la suite d'une fission. Ils sont créés sous haute énergie. Maintenant nous désirons suivre les neutrons, par rapport à leur énergie, leur emplacement, et le temps, à travers leur migration dans le réacteur. Suivons un neutron depuis sa naissance. S'il est né au point A du centre, il a une certaine énergie E , et il voyage dans la direction ω . Comment décidons-nous de la valeur énergétique de ce neutron? Il y a une distribution probabiliste déterminée de façon expérimentale qui est la probabilité qu'un neutron qui vient de naître possède l'énergie E (voir Fig. 2).

Ainsi, afin d'équiper notre nouveau neutron avec une énergie, nous devons choisir une valeur E de la distribution $f(E)$. Comment procédons-nous? Eh bien si $f(E)$ a une forme analytique simple, on peut y arriver par diverses méthodes rusées. Toutefois, dans ce cas-ci, f est déterminé de façon expérimentale, donc nous choisissons E selon une recherche binaire dans la table de la fonction de distribution cumulative. OK, maintenant nous avons l'énergie E et la position de notre neutron. Comment choisissons nous sa direction de vol? Les lois de la physique nous disent que la direction est

* E-mail: wilf@central.cis.upenn.edu.

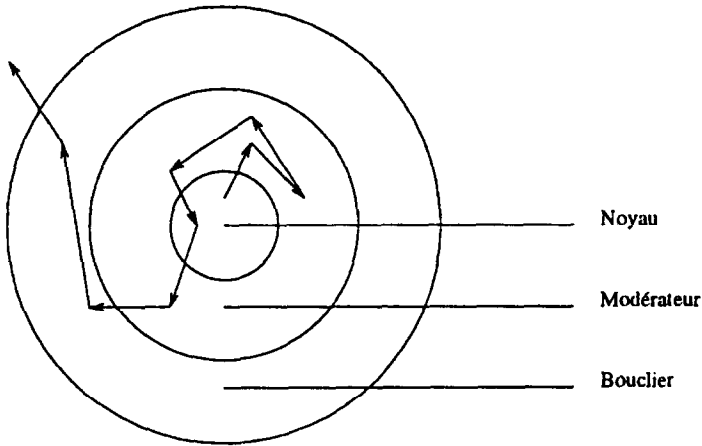


Fig 1.

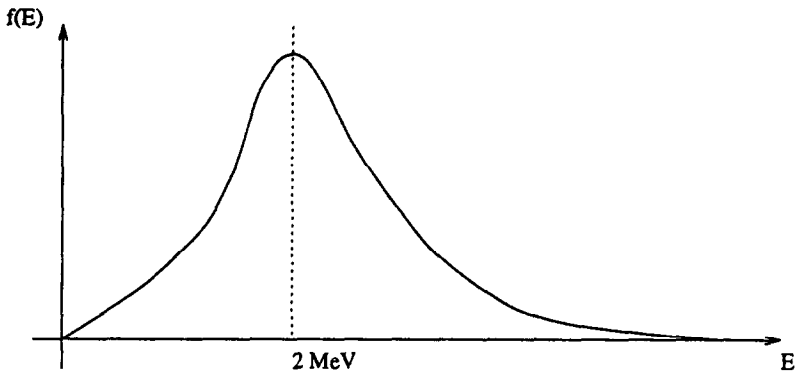


Fig 2.

isotropique. C'est-à-dire que toutes les directions sur la sphère ont les mêmes probabilités. Alors, nous choisissons une direction de vol au hasard uniformément sur la sphère. Maintenant notre particule est complètement déterminée avec ses coordonnées (x, ω, E) . Elle voyagera sur une certaine distance, d , qui, lorsqu'elle est mesurée en *distance moyenne sans obstacle* ("mean free paths"), a une distribution exponentielle. C'est-à-dire que la probabilité pour qu'elle parcoure d distances moyennes est e^{-d} . Donc, la première chose que nous devons faire est de choisir une valeur X de la distribution exponentielle. Dans ce but nous devrions utiliser le principe général de tirage aléatoire: Afin de choisir une valeur X d'une fonction de distribution cumulative $F(X)$,

1. Choisissez un nombre t , au hasard dans $(0, 1)$.
2. Retournez $X = F^{-1}(t)$.

Dans ce cas, nous pouvons prendre $X = -\log t$, où t est un nombre aléatoire, et X aura alors la distribution exponentielle exacte.

A peu près au même moment, nous apprenons la technique ingénieuse de rejet de Von Neumann, qui permet de choisir suivant la distribution exponentielle. On ne sait pas vraiment si cette technique est plus efficace que le calcul du logarithme d'un nombre aléatoire, mais c'est certainement plus élégant. Alors nous l'utilisons. Maintenant nous avons une seule particule, complètement déterminée avec ses coordonnées (x, E, ω) , et nous savons qu'elle parcourt X distances moyennes jusqu'à ce que quelque chose lui arrive. Nous devons maintenant suivre la particule jusqu'à ce que la portion X soit épuisée, à travers plusieurs media différents, chacun ayant une relation différente entre la distance moyenne sans collision et la distance parcourue. Supposons que la particule parcourt une distance linéaire de d centimètres avant d'arriver sur la frontière du prochain medium. En faisant cela, le nombre de distances moyennes qu'elle aura accomplies sera $d\mu$, où μ est une certaine propriété du medium. Alors nous réduisons le nombre X de distances moyennes qu'il reste à parcourir de $d\mu$, nous transportons la particule sur la frontière, et nous continuons la trajectoire dans la région suivante.

Cette recherche, suivre une trajectoire linéaire à travers un espace à trois dimensions dans lequel habitent plusieurs objets différents, de formes différentes, contient des problèmes géométriques très beaux. Comment peut-on décrire les objets d'une manière compacte? Etant données les équations de la surface des objets, quels sont les bons algorithmes pour décider quel est l'objet que la trajectoire croisera ensuite?

Quand on arrive au moment du prochain événement, nous devons nous demander ce qui va arriver à la particule. Il y a plusieurs possibilités:

1. Elle a subi une collision avec un noyau et elle a ricoché dans une autre direction. Dans ce cas nous devons changer deux choses. D'abord nous devons choisir une nouvelle direction pour la particule, de façon isotropique. Ensuite nous devons choisir une nouvelle énergie, plus basse, pour la particule, au moyen d'une formule explicite qui lie la nouvelle énergie au changement de direction. L'ensemble du processus continue ensuite dans la nouvelle direction.

2. Elle est absorbée, et elle a disparu du calcul. Dans ce cas nous inscrivons une petite augmentation de la chaleur qui est générée au point où l'absorption est arrivée, et ensuite nous nous penchons sur la prochaine particule. Il faut se rappeler que la seule raison d'être du réacteur est la production de cette chaleur.

3. Elle est complètement sortie de la configuration (hélas!), et ayons pitié pour les êtres humains dans le voisinage.

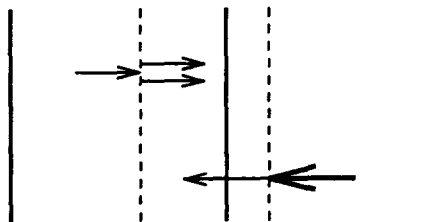
4. Elle est absorbée et elle a créé une fission et donc plusieurs particules nouvelles. Dans ce cas nous devons choisir la quantité de particules nouvelles produites, en utilisant la distribution observée de façon expérimentale (qui était, à cette époque, un renseignement très secret!)

Généralement on doit étudier des millions de particules afin d'obtenir des statistiques modestement justes sur la position des particules, sur la quantité qui s'échappe, et la quantité de chaleur produite.

Beaucoup d'artifices très brillants ont été inventés pour augmenter la sûreté des statistiques observées. Des ruses telles que le doublage et le pesage des particules. Ce sont des artifices qui permettent de réduire la variance des estimations. Supposons que nous essayions d'estimer le flux des particules qui s'échappent. Il s'agit d'une petite fraction de toutes les particules, donc nous devons générer beaucoup de particules afin de trouver le peu qui s'échappe. Afin d'obtenir une estimation plus sûre des particules qui s'échappent, nous créons plusieurs "périmètres de doublage" artificiels (voir Fig. 3).

Lorsqu'une particule traverse un de ces périmètres vers l'extérieur, nous l'encourageons en la doublant, chacune possédant la moitié du poids de la particule originale. Lorsqu'une particule traverse un de ces périmètres vers l'intérieur, on la tue, avec une probabilité d'un demi (c'est la "roulette russe"). Beaucoup d'autres artifices ont été développés pour réduire les variances, et beaucoup d'autres méthodes de sélection ont été trouvées à partir de distributions multivariées données. Ceci est, bien sûr, la fameuse méthode de Monte Carlo, qui a eu beaucoup de succès. Elle permettait d'analyser des réacteurs de géométrie très complexe. Elle réussissait mieux que les méthodes conventionnelles, fondées sur l'intégration des systèmes d'équations intégral-différentielles.

En parlant autobiographiquement à nouveau, j'ai fait toutes mes recherches sur la méthode de Monte Carlo au cours d'un travail à mi-temps pendant mes études supérieures. Je dois vous confesser que mon travail était si intéressant, que j'estimais mes études comme un obstacle à ma vraie formation mathématique. Etant donné que mes premières expériences mathématiques étaient dans le domaine de la méthode de Monte Carlo, il n'est peut-être pas suprenant qu'en 1975, quand la combinatoire était mon principal intérêt, Albert Nijenhuis et moi, nous ayons écrit "Combinatorial Algorithms" [4], un livre qui contient de nombreuses méthodes de génération aléatoire d'objets combinatoires. A l'époque, la seule méthode connue était la suivante: lorsqu'on avait un procédé récursif pour construire une famille d'objets combinatoires et un procédé qui nous permettait de compter ces objets, alors nous pouvions aussi les générer au hasard uniformément. Un cadre très général, basé sur la théorie des marches dans un graphe dirigé, a été formulé en 1978. Ce cadre décrivait comment



Doublage.

Fig 3.

nous pouvons mener à bien quatre opérations de base sur les familles combinatoires graphiques (voir Fig. 4).

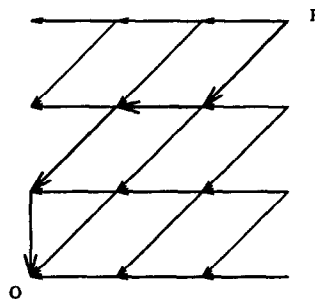
1. “Listing”: faire une liste de tous les objets d’une taille donnée.
2. “Un-ranking”: Dans la liste lexicographique de tous les objets d’une taille donnée, trouver le $r^{\text{ème}}$ membre de la liste, où r est donné.
3. “Rank”: Etant donné un objet de la famille, trouver son rang dans la liste de tous les objets de sa taille ordonnée de façon lexicographique.
4. “Random”: Tirer uniformément et au hasard un objet de l’ensemble de tous les objets d’une taille donnée.

Je voudrais mentionner deux des développements dans ce domaine depuis cette époque; deux que je pense être particulièrement beaux et importants.

D’abord, je voudrais parler de l’algorithme de David Aldous [1] et de celui d’Andrei Broder [2], qui ont travaillé indépendamment, pour la génération uniforme aléatoire des arbres maximaux d’un graphe donné. Il y a eu plusieurs tentatives précédemment pour produire un bon algorithme pour ce problème, mais la méthode A–B est certainement la meilleure.

L’algorithme: Etant donné un graphe G , commencez à un certain sommet X_0 . En général, après être arrivé à un certain sommet X , choisissez un voisin X' de X uniformément parmi les voisins de X , et ensuite allez à X' . Arrêtez-vous dès que vous aurez visité tous les sommets au moins une fois. Lorsque la marche est terminée prenez alors l’arbre maximal désiré de manière suivante: l’ensemble de toutes les arêtes qui arrivent au sommet pour la première fois.

Il est clair que cet arbre T est un arbre maximal de G . Mais il n’est pas aussi clair que l’arbre T est uniformément distribué parmi les arbres maximaux de G . Dans le cas où G est le graphe complet, alors l’algorithme produit, bien sûr, un arbre de n sommets, uniformément distribué parmi tous les arbres étiquetés de n sommets. Dans ce



“Objet” = chemin de P à O.

Fig 4.

cas l'algorithme peut être reformulé et condensé comme suit:

1. Pour chaque $i = 2, 3, \dots, n$, joignez le sommet i avec le sommet

$$V_i = \text{Min}[U_i, i - 1],$$

où les U_i sont choisis indépendamment et uniformément au hasard dans l'ensemble $1, 2, \dots, n$.

2. Choisissez une permutation aléatoire de $1, 2, \dots, n$ et réétiquetez les sommets $1, \dots, n$ avec cette permutation.

Des marches de ce type nous sont familières depuis de nombreuses années parce qu'elles sont utilisées dans la démonstration du théorème "matrix-tree", qui énumère les marches eulériennes sur un graphe dirigé en termes de nombre d'arborescences maximales avec une racine donnée. On n'avait pas soupçonné, toutefois, que de telles marches et de tels arbres étaient uniformes dans l'ensemble des arbres maximaux d'un graphe non-dirigé.

Voici un développement relativement récent dont j'aimerais parler. Il s'agit de la question suivante: étant donné un graphe G biparti, tirez uniformément et au hasard un couplage parfait pour G . Ce problème a été traité par Mark Jerrum et Alistair Sinclair [3]. La méthode diffère des précédentes de plusieurs manières significatives. D'abord, nous ne savons pas comment compter les couplages parfaits d'un graphe biparti. Précisément, la question de leur énumération est équivalente à l'évaluation de la fonction *permanent*, et ce problème est bien connu comme étant $\#$ P-complet. Donc, nous ne pouvons pas formuler une méthode pour la sélection uniforme basée sur un procédé énumératif et récursif. Malgré tout nous pouvons créer un procédé de sélection qui est approximativement uniforme. Cela veut dire que nous pouvons tirer un couplage en commençant avec un couplage quelconque, et mener à bien des perturbations aléatoires. Pendant que nous le perturbons, le couplage que nous avons devient de mieux en mieux uniformément distribué, et après peu de perturbations on pourra démontrer qu'il est presque uniforme. De telles méthodes dépendent de la propriété qu'ont certaines chaînes Markoviennes de *se mélanger* rapidement. Ce qui

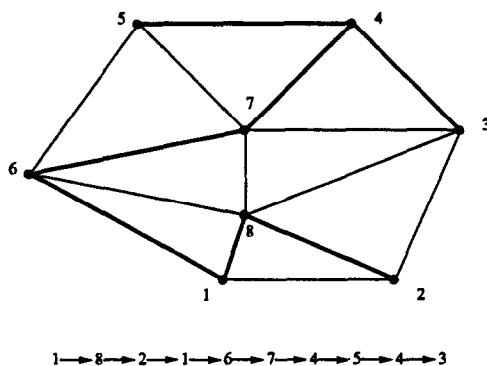


Fig 5.

est difficile est de toujours prouver qu'une chaîne donnée se mélange en fait rapidement. Dans ce problème, de tirage d'un couplage parfait, il y a eu de faux départs. C'est à dire que certains algorithmes basés sur de prétendues propriétés de mélanges rapides de certaines chaînes Markoviennes ont été proposés, mais les démonstrations sur leur nature de mélange rapide étaient fausses. Je dis cela seulement pour souligner le fait qu'il peut être très difficile de prouver qu'une chaîne se mélange rapidement. Nous postulons que notre graphe biparti donné, G , a un couplage parfait. Nous relevons, en passant, que cette propriété de G peut être testée en temps polynomial. Alors, supposez que $G = (U, V, E)$, où $U = V = \{0, 1, \dots, n - 1\}$. Nous produirons un couplage parfait presque uniformément au hasard, pour avoir mené à bien suffisamment de fois la marche aléatoire suivante: Supposez que l'on nous donne un couplage qui est soit parfait soit parfait à une arête près. C'est-à-dire que l'on nous donne un couplage M qui consiste en soit n soit $n - 1$ arêtes. Alors nous faisons ceci:

Commençons avec un couplage parfait quelconque M de G (trouvé en temps polynomial) Ensuite, en général, après être arrivé à un couplage quelconque M , nous faisons ce qui suit (voir Fig. 6).

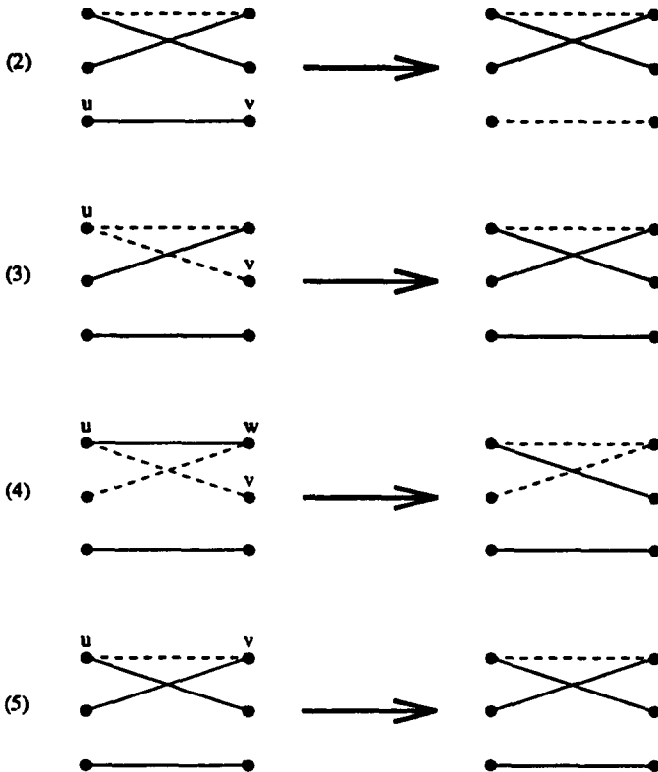


Fig 6.

1. Choisissons une arête $e = (u, v)$ parmi l'ensemble des toutes les arêtes de G , uniformément au hasard.
2. Si notre couplage est parfait et si e est une des arêtes du couplage, alors nous éliminons e de M .
3. Si notre couplage M contient seulement $n - 1$ arêtes, et si aucune des extrémités de e n'appartient à une arête du couplage, alors nous ajoutons e à M .
4. Si notre couplage M contient seulement $n - 1$ arêtes, et si, dans M , u est accouplé à un sommet w quelconque et v n'est pas accouplé, alors vous allez au couplage $M' = M + e - (u, w)$. De même, si v est accouplé à w et u n'est pas accouplé.
5. Dans tous les autres cas, ne faites rien.

Théorème (Jerrum et Sinclair [3]). *Soit G un graphe biparti à $2n$ sommets tel que que le degré de chaque sommet est au moins $n/2$. Si $\epsilon > 0$ est donné alors il y a un polynôme P tel que si nous menons a bien les P transitions, la chaîne Markovienne ci-dessus décrite produira un couplage parfait de G tel que la probabilité d'obtenir un couplage parfait particulier diffère de la distribution uniforme de moins de ϵ .*

Application: Les permutations avec des positions restreintes. Nous allons parler des conséquences des permutations avec des positions restreintes. Nous imaginons qu'on a donné un tableau $n \times n$ de cellules C , dont certaines sont rouges, et d'autres sont noires. Un tel damier définit un ensemble de permutations de n lettres, d'une façon bien connue: une permutation f appartient à l'ensemble si et seulement si pour chaque $i = 1, \dots, n$ il est vrai que la cellule $(i, f(i))$ sur le damier C est rouge (voir Fig. 7).

Nous associons à un tel damier C un graphe biparti G comme suit: G a $2n$ sommets: $1, \dots, n; 1', \dots, n'$. La paire (i, j') est une arête de G si la cellule (i, j') de C est rouge. Clairement, les couplages parfaits de $G(C)$ sont en bijection avec les permutations dont le damier des positions restreintes est C . Le théorème de Jerrum et Sinclair nous montre comment tirer presque uniformément une permutation avec le damier C , dans le cas où chaque rangée de C a au moins $n/2$ cellules rouges.

D'une part, les problèmes suivants sont tous résolus par l'algorithme de Jerrum-Sinclair: Etant donné n , choisissez presque uniformément au hasard une permutation

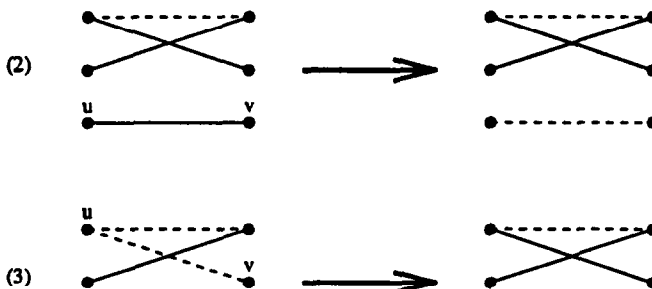


Fig 7.

f qui n'a pas de point fixes, ou telle que $|f(i) - i|$ est au moins 2 pour tous les i (cycliquement ou non), ou bien au moins 3, etc.

D'autre part, les problèmes suivants ne sont pas résolus par l'algorithme, puisque les graphes ne sont pas assez denses: étant donné n , choisissez presque uniformément au hasard une permutation telle que $|f(i) - i| < 5$ pour chaque i , etc.

Bibliographie

- [1] D.J. Aldous, The random walk construction of uniform spanning trees and uniform labelled trees, *SIAM J. Discrete Math.* **3** (1990) 450–465.
- [2] A Broder, Generating random spanning trees, in: *Proc. Symp. Foundations of Computer Sci.* (Institute for Electrical and Electronics Engineers, New York, 1989 442–447).
- [3] M. Jerrum and A. Sinclair, Approximating the permanent, *SIAM J. Comput.* **18** (1989) 1149–1178.
- [4] A. Nijenhuis and H.S. Wilf, *Combinatorial Algorithms* (Academic Press, New York, 1978).