Chemins du plan contraints.

6 septembre 2019

On présente, à l'intention de lecteurs déjà exposés à la théorie des probabilités, quelques thèmes choisis à l'intersection des probabilités et de la combinatoire en rapport avec des marches du plan parmi les plus simples : les marches vers le haut et vers la droite, encore appelées marches Nord Est (NE). Dans le cas de marche finies avec autant de pas N que de pas E (c'est-à-dire dont le dernier sommet est sur la diagonale), en essaie (entre autre) de comprendre la probabilité de rester au dessus de la diagonale ou encore la probabilité de partager le plan en deux parties d'aire égale. La première question renvoie aux chemins de Dyck, à la combinatoire de Catalan et à la théorie des arbres aléatoires, la seconde question renvoie aux partitions d'entiers.

1 Deux amuse-bouches et le LCLT

1.1 Deux problèmes semblables mais distincts

1. (RMS158) Une urne équilibrée : Une urne qui comprend initialement N boules blanches et N boules noires. On tire successivement et sans remise des boules jusqu'à ce qu'il ne reste que des boules d'une seule couleur dans l'urne. Combien de boules reste-t-il alors : $O(1), O(\log(N)), O(\sqrt{N})$ [1] ?
2. Le problème des boites d'allumettes de Banach [2] : un professeur a deux boîtes d'allumettes, une dans sa poche droite et une dans sa poche gauche. Chacune contient initialement N allumettes. Chaque matin, le professeur choisit une poche au hasard, puis retire une allumette de la boîte ; un certain matin, il trouvera une boîte vide ; quelle sera alors la loi du nombre d'allumettes dans l'autre poche ?

Les deux problèmes sont distincts : les modèles de chemins du plan sont différents : dans le premier cas,, c'est simplement la mesure uniforme sur les chemins NE de (0,0) à (N,N) (qui comptent donc autant de pas N que de pas E) : il y a $\binom{2N}{N}$ tels chemins.

Dans le second cas, c'est plus compliqué ; on considère les chemins NE jusqu'au premier temps d'atteinte de $\{N+1\}\times\mathbb{N}\cup\mathbb{N}\times\{N+1\}$, et on munit ces chemins de la probabilité 2^{-L} où L est la longueur du chemin, cette fois variable (=nombre d'arêtes). Noter qu'il n'est pas complètement évident que ceci définisse une mesure de probabilité.

Dans le cas de l'urne équilibrée, notons $K = K_N$ la variable aléatoire égale au nombre de boules dans la boîte la première fois que celle-ci est "monochromatique" : pour $k \geq 1$ (notons que la loi ne charge pas 0), notons que l'événement $\{K = k\}$ détermine les $k+1$ derniers pas : ou bien il s'agit de NE...E, où E apparaît k fois, ou bien il s'agit de EN... N, où N apparaît k fois. Les 2 situations étant symétriques, elles ont même probabilité et on considère la seule première situation. Alors parmi les $2N-k-1$ premiers pas, on doit choisir $N-1$ pas N et partant :

$$P(K=k) = 2\frac{\binom{2N-k-1}{N-1}}{\binom{2N}{N}} = \frac{(2N-k-1)!}{(N-1)!(N-k)!}\frac{(N!)^2}{(2N)!} = 2\frac{N\cdot(N)\ldots(N-k+1)}{(2N)\cdot(2N-k)} \to 2\cdot(\frac{1}{2})^{k+1} = (\frac{1}{2})^k$$

où l'on notera que le premier facteur 2 provient de la symétrie de la situation. La loi limite est donc une loi $\text{Géom}_{\mathbb{N}^*}(1/2)$,

1. en vérité, il faudrait écrire $O_P(1), O_P(\log(N)), O_P(\sqrt{N})$ puisqu'il s'agit de variables aléatoires
2. Banach matchbox problem, voir `https://en.wikipedia.org/wiki/Banach%27s_matchbox_problem`

Exercice 1. *Vérifier que la relation de domination stochastique*[3] *suivante est vérifiée :*

$$\forall k \in \mathbb{N}^\star, P(K \geq k) \leq (\frac{1}{2})^k$$

En particulier, il y a donc $O_P(1)$ boules de même couleur restant dans l'urne.

Pour le nombre d'allumettes restantes, ègalement noté K, dans le problème des boîtes d'allumettes de Banach, la première fois qu'une poche vide est découverte : cette variable aléatoire vérifie $P(0 \leq K \leq N) = 1$. Ensuite, pour $k = O(\sqrt{N})$: $P(K = k) = 2 \cdot 2^{-(2N+1-k)} \cdot \binom{2N-k}{N} \sim 2^{-(2N-k)} \cdot \frac{1}{\sqrt{\pi N}} 2^{2N-k} e^{-\frac{k^2}{4N}} = \frac{1}{\sqrt{\pi N}} e^{-\frac{k^2}{4N}}$ où on a utilisé le LCLT pour un équivalent précis.

On peut aussi en déduire, mais c'est moins fort, que

$$K_N / \sqrt{4N} \Rightarrow |G|$$

où $G \sim \mathcal{N}_1(0,1)$ suit une loi normale centrée réduite. En particulier, il y a $O_P(\sqrt{N})$ allumettes restantes. On note donc que la loi de trajectoire aléatoire interrompue lorsqu'elle *quitte* le carré $[0,N] \times [0,N]$ est "similaire" à celle de la loi de la trajectoire lorsqu'elle coupe l'axe $x+y = 2N$, au sens où on obtient encore une loi normale à la limite.

Enfin on peut montrer que $E[K] \sim \frac{2}{\sqrt{\pi}} \sqrt{N}$ (NB : ceci ne découle pas directement de la convergence en loi).

1.2 Limite locale 1

La limite locale de la loi uniforme des chemins NE de $(0,0)$ à (N,N) est la loi produit $Ber(1/2) \otimes Ber(1/2) \otimes \dots$ sur les symboles NE. Pour le voir on se fixe un chemin NE de taille finie issu de $(0,0)$. Supposons qu'il compte i_1 pas E et i_2 pas N (il est donc de longueur $i_1 + i_2$) ; on observe que la probabilité qu'un chemin uniforme NE $(0,0)$ à (N,N) commence par ce chemin vaut :

$$\frac{\binom{2N-i_1-i_2}{N-i_1}}{\binom{2N}{n}} \sim \frac{\frac{2^{2N-i_1-i_2}}{\sqrt{\pi N}} e^{-\frac{(i_1-i_2)^2}{2N-(i_1+i_2)}}}{\frac{2^{2N}}{\sqrt{\pi N}}}$$
$$= 2^{-i_1-i_2} e^{-\frac{(i_1-i_2)^2}{2N-(i_1+i_2)}}$$

En particulier, pour i_1, i_2 fixés, on a donc bien que la limite $N \to \infty$ vaut $2^{-i_1-i_2}$, c'est-à-dire la probabilité du chemin constitué d'une suite de Ber(1/2) indépendantes à valeurs dans les deux symboles NE. La plus-value de la formule précédente est qu'en fait cette approximation vaut dès lors que $(i_1 - i_2)^2 = o(N)$, c'est-à-dire tant que la distance à la diagonale n'excède pas $\sqrt{N}$.

Notons la bijection naturelle entre les chemins NE $(0,0)$ à (N,N) et l'ensemble $\{(i_0, \dots, i_N) \in \mathbb{N}^{N+1} : \sum i_j = N\}$ où i_j compte le nombre de pas N d'abscisse j. Muni de la loi uniforme sur les chemins, la variable aléatoire I_j associée converge vers la loi $\mathrm{Geom}_{\mathbb{N}}(1/2)$ (on pourrait aussi prendre des veceturs composée d'un nombre fini de coordonnées, ou même étudier sous quelles conditions le résultat reste vrai si l'on fait croître le nombre de coordonnées avec N).

Exercice 2. *On peut aussi considérer l'ensemble*

$$\{k \in \mathbb{N}, (i_1, \dots, i_k) \in (\mathbb{N}^\star)^k : \sum_{j=1}^{k} i_j = N\}$$

où cette fois, la longueur k de la suite d'entiers (désormais tous non nuls) est aussi laissée libre. Quelle est la loi limite de $K = K(N)$ la variable aléatoire associée lorsqu'on met la loi uniforme sur cet ensemble ? Quelle sont les lois limites des I_1, I_2 (conditionnellement à ce que I_2 soit défini) ?

3. les relations de domination stochastique s'expriment forcément en fonction des queues des variables aléatoires ; on ne peut avoir des dominations sur la fonction de masse, puisque les fonctions de masse somment à 1.

Réponse : On note que $i_j \mapsto i_j - 1$ met cet ensemble en bijection cet ensemble avec

$$\{k \in \mathbb{N}, (i_1, \dots, i_k) \in (\mathbb{N}^\star)^k : \sum_{j=1}^{k} i_j = N - k\}$$

Et ce dernier ensemble est de cardinal $\binom{N}{N-k} = \binom{N}{k}$, maximisé pour $k = N/2$ (adapter dans le cas où ce coefficient est impair).

Il existe d'autres façons naturelles d'induire une loi sur les chemins NE de $(0,0)$ à (N,N).

Exercice 3. *Considérons par exemple la mesure probabilité uniforme sur $\{f : [N] \to [N]\}$, puis posons $i = (i_1, \dots, i_N)$ avec $i_j = \sum \mathbf{1}_{f(k)=j}$. Notons que $i_1 + \dots + i_N = N$ par définition. On note $(I_1, \dots, I_N)$ les variables aléatoires associées. Quelle est la loi de $(I_1, \dots, I_N)$? Pour $1 \leq j \leq N$ fixé, vers quelle loi I_j converge ?*

Réponse : $(I_1, \dots, I_N)$ suit une loi multinomiale, et la loi marginale $I_j \Rightarrow \mathrm{Poi}(1)$ quand $N \to \infty$ (et si on prend fixe de coordonnées, la convergence est vers la loi du vecteur de ces variables indépendantes).

En dehors du problème de la limite locale, on peut aussi s'intéresser à la limite d'échelle : cette dernière consiste à trouver une limite pour l'objet global remis à l'échelle : c'est le sujet du mouvement Brownien, abordé dans un cours de probabilité de niveau M2.

1.3 LCLT

Soit $X_1, X_2, \dots$ des variables aléatoires iid de carré intégrable. On pose $S_N = \sum_{i=1}^{N} X_i$. Notons $\mu = E[X_1]$ et $\sigma^2 = \mathrm{Var}(X_1)$. On sait que

$$\frac{S_N - E[S_N]}{\sqrt{\mathrm{Var}(S_N)}} = \frac{S_N - \mu N}{\sigma\sqrt{N}} \Rightarrow \mathcal{N}_1(0,1)$$

Et ceci se traduit concrètement de la façon suivante :

$$P(\mu N + a\sigma\sqrt{N} \leq S_N < \mu N + b\sigma\sqrt{N}) = P(a \leq \frac{S_N - N\mu}{\sigma\sqrt{N}} < b) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

On a donc un contrôle sur la probabilité des intervalles de taille $\sqrt{N}$ proches de $N\mu$. Une bien meilleure approximation serait la suivante :

$$S_N \sim \mathcal{N}_1(E[S_n], \mathrm{Var}(S_n)) = \mathcal{N}_1(n\mu, \sigma^2 n))$$

Quel sens peut on donner à un tel énoncé ?

$$P(a < S_N < b) \sim \int_a^b \frac{1}{\sqrt{2\pi\sigma^2 n}} e^{-\frac{(x-n\mu)^2}{2\sigma^2 n}} dx$$

pour a et b à distance $O(\sqrt{N})$ de $n\mu$, de sorte que le membre de droite soit de l'ordre $N^{-1/2}$. C'est-à-dire qu'on est passé d'intervalles de taille $O(\sqrt{N})$ à des intervalles de taille $O(1)$. Il y a bien sûr des obstructions à un tel énoncé. Si X_1 est une variable aléatoire discrète, alors S_n est encore une variable aléatoire discrète, tandis que la variable aléatoire gaussienne est à densité : aussi suprenant que cela puisse paraître, c'est essentiellement la seule obstruction ; donnons maintenant des énoncés précis.

Si les variables entières à valeurs dans $\mathbb{Z}$, et on suppose on note

$$h := \max\{k | \exists a \in \mathbb{Z}, X \subset a + k\mathbb{Z}\} < \infty$$

(h définit la période, et la marche est dite apériodique si $h > 1$). Alors

$$P(S_N = k) = \frac{h}{\sqrt{2\pi\sigma^2 N}} e^{-\frac{(k-N\mu)^2}{2N\sigma^2}} + o(N^{-1/2})$$

pour $k \in \mathbb{Z}$, et le terme d'erreur $o(N^{-1/2})$ est *uniforme* en k. En d'autres termes,

$$\sup_{k\in\mathbb{Z}} N^{1/2}|P(S_N = k) - \frac{h}{\sqrt{2\pi\sigma^2 N}} e^{-\frac{(k-N\mu)^2}{2N\sigma^2}}| \to 0$$

quand $N \to \infty$.

Comme exemple élémentaire de marche aléatoire périodique, notons que la marche aléatoire simple sur $\mathbb{Z}$ est périodique de période 2, avec $X_1 \in 1 + 2\mathbb{Z}$ donc $S_N \in N + 2\mathbb{Z}$., c'està-dire que $S_N - N$ est pair ou encore que S_N est de la parité de N : les marches périodiques ne sont donc pas atypiques.

Exercice 4. *Supposons le théorème établi pour $h = 1$. Montrer que le cas de h quelconque s'en déduit.*

Il existe encore des versions pour des variables qui ne sont pas identiquement distribuées, ou même qui ne sont pas indépendantes mais qui possèdent une autre structure de type martingale.

1.4 Applications : les coefficients binomiaux

Dans le cas de variables binomiales cela redonne l'équivalent bien connu de Stirling :

$S_N \sim \mathrm{Po}(N)$ peut être réalisé comme la somme de variables de Poisson indépendantes de paramètre 1, et alors

$$P(S_N = N) = e^{-\lambda}\frac{\lambda^N}{N!}_{|\lambda=N} = e^{-N}\frac{N^N}{N!} \sim \frac{1}{\sqrt{2\pi N}}, \quad \text{soit } N! \sim \sqrt{2\pi N}(\frac{N}{e})^N.$$

car espérance et variance de la loi de $\mathrm{Po}(\lambda)$ sont toutes deux égales à λ (penser à la limite de variables aléatoires binomiales peut aider, car espérance et variance, Np et $Np(1-p)$, convergent vers λ quand $Np \to \lambda$). L'équivalent LCLT redonne bien Stirling.

On peut aussi choisir k fonction de N, avec $k = O(N^{1/2})$

$$P(S_N = N + k) = e^{-\lambda}\frac{\lambda^{N+k}}{(N+k)!}_{|\lambda=N} = e^{-N}\frac{N^{N+k}}{(N+k)!} \sim \frac{1}{\sqrt{2\pi N}} e^{-\frac{k^2}{2N}}$$

Ainsi $(N+k)! \sim (\frac{N}{e})^N\sqrt{2\pi N}e^{\frac{k^2}{2N}} \sim N!N^k e^{\frac{k^2}{N}}$ i.e.

$$\lim_{N\to\infty} \frac{1}{N^k}\frac{(N+k)!}{N!} \sim e^{\frac{k^2}{2N}}$$

$S_N \sim \mathrm{Bin}(N, 1/2)$ permet d'estimer le coefficient binomial médian (ou milieu) :

$$P(S_{2N} = N) = 2^{-2N}\binom{2N}{N} \sim \frac{1}{\sqrt{2\pi(2N)/4}} \text{ soit } \binom{2N}{N} \sim \frac{2^{2N}}{\sqrt{\pi N}}$$

et aussi :

$$P(S_{2N} = N + k) = 2^{-2N}\binom{2N}{N+k} \sim \frac{1}{\sqrt{2\pi(2N)/4}} e^{-\frac{k^2}{2(2N)/4}} \text{ soit } \binom{2N}{N+k} \sim \frac{2^{2N}}{\sqrt{\pi N}} e^{-\frac{k^2}{N}}$$

et plus généralement, pour $\alpha \in (0, 1)$ et $S_N \sim \mathrm{Bin}(N, \alpha)$:

$$P(S_N = \alpha N) = \alpha^{\alpha N}(1-\alpha)^{(1-\alpha)N}\binom{N}{\alpha N} \sim \frac{1}{\sqrt{2\pi N\alpha(1-\alpha)}}$$

c'est-à-dire [4]

$$\binom{N}{\alpha N} \sim (\alpha^\alpha(1-\alpha)^{1-\alpha})^{-N}\frac{1}{\sqrt{2\pi N\alpha(1-\alpha)}}$$

4. ce calcul est lié à l'entropie de la loi de Bernoulli de paramètre α

et même, pour $k = O(N^{1/2})$,

$$P(S_N = \alpha N + k) = \alpha^{\alpha N+k}(1-\alpha)^{(1-\alpha)N-k}\binom{N}{\alpha N+k} \sim \frac{1}{\sqrt{2\pi N\alpha(1-\alpha)}} e^{-\frac{k^2}{2N\alpha(1-\alpha)}}$$

c'est-à-dire :

$$\binom{N}{\alpha N+k} \sim (\alpha^\alpha(1-\alpha)^{(1-\alpha)})^{-N} \frac{1}{\sqrt{2\pi N\alpha(1-\alpha)}} (\frac{1-\alpha}{\alpha})^k e^{-\frac{k^2}{2N\alpha(1-\alpha)}}$$

Pour imaginer de nouveaux exemples il suffit essentiellement de trouver de nouveaux exemples de distributions de probabilité dont les convolées (lois des sommes) sont explicites ; dès lors que les variances sont finies, on sera en position d'appliquer le LCLT.

1.5 Applications

Il existe enfin une version vectorielle ; on donne celle qui concerne les marches à valeurs dans $\mathbb{Z}^d$. Soit $X_1, X_2, \dots$ des vecteurs aléatoires iid à valeurs dans $\mathbb{Z}^d$ de carré intégrable. On pose $S_N = \sum_{i=1}^N X_i$. On note $\mu = E[X_1]$ et $\Sigma = \mathrm{Cov}(X)$. On suppose la marche apériodique, c'est-à-dire que $\{x \in \mathbb{Z}^d : P(S_n = 0) > 0\}$ contient tous les enteirs n'est pas contenu dans un sous-réseau strict de $\mathbb{Z}^d$. Alors, quelque soit $k \in \mathbb{Z}^d$

$$P(S_N = k) = \frac{1}{(2\pi)^{d/2} N^{d/2} \sqrt{\mathrm{Det}(\Sigma)}} e^{-\frac{1}{2}(k-N\mu)^t \Sigma^{-1}(k-N\mu)} + o(N^{-1/2})$$

pour $k \in \mathbb{Z}^d$, et le terme d'erreur $o(N^{-1/2})$ est uniforme en k.

Dans un graphe, on appelle chemin une suite de sommets adjacents (c'est-à-dire reliés par une arête), Dans le cas d'un chemin fini $\omega = (\omega_0, \omega_1, \dots, \omega_N)$, N est la longueur du chemin, et le chemin est dit fermé si $\omega_0 = \omega_N$.

On s'intéresse au nombre chemins fermés sur le réseau carré, c'est-à-dire $\mathbb{Z}^2$. Nécessairement, la longueur d'un tel chemin est paire. Il existe une fornule fermée pour compter les chemins de longueur $2N$:

$$\sum_{k=0}^{N} \binom{2k}{k}\binom{2(N-k)}{N-k}\binom{2N}{2k} \tag{5}$$

Et il est d'ailleurs intéressant d'essayer de comprendre sans outils l'ordre de grandeur de cette somme. Nous procédons ici de procéder à l'aide du théorème central limite local, pour illustrer la puissance de cet outil. On se donne donc une suite $(B_i)_{i=1}^{2N}$ une suite de vecteurs aléatoires indépendants de loi uniforme sur l'ensemble à 4 éléments

$$\{(0,1),(1,0),(-1,0),(0,-1)\}.$$

On note que $\Sigma := \mathrm{Cov}(B) = 1/2 \cdot I_2$. Donc, du théorème central limite local (avec ici $d = 2$),

$$4^{-2N}\mathrm{Card}\{\omega : \omega_0 = \omega_{2N} = (0,0)\} = P\left(\sum_{i=1}^{2N} B_i = (0,0)\right) \sim 2 \cdot \frac{1}{(2\pi)^{d/2}(2N)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}} = \frac{2}{\pi} \cdot \frac{1}{2N}$$

Donc le nombre de chemins fermés de longueur N est équivalent à $\frac{2}{\pi} \cdot \frac{4^N}{N}$ si N est pair, et nul sinon.

Exercice 6. *Montrer que le nombre de chemins fermés de longueur $2N$ est équivalent à*
- $\frac{\sqrt{3}}{2\pi} \cdot \frac{6^N}{N}$ *sur le réseau triangulaire.*
- $\frac{3}{\pi} \cdot \frac{3^N}{N}$ *sur le réseau hexagonal.*

2 Positivité

On étudiera ici deux attaques pour traiter la contrainte de positivité : le lemme du scrutin [5] (originellement formulé pour pour des marches NE adaptés dans le cas d'un problème de vote), mais aussi le lemme cyclique (qu'on peut voir comme une généralisation du lemme du scrutin aux marches continues à gauche).

5. le "ballot lemma" fut publié par W. A. Whitworth en 1878, est redécouvert par Louis François Bertrand en 1887

2.1 Principe de réflexion 1 : le lemme du scrutin (ballot lemma)

On considère la mesure de probabilité uniforme sur les chemins NE issus de $(0,0)$ et qui terminent en (a,b) avec $a<b$ deux entiers. Alors la probabilité que le chemin ne croise la diagonale qu'en son point de départ $(0,0)$ vaut :

$$\frac{b-a}{b+a}.$$

Interprétation en terme de dépouillement de vote : étant donné que le candidat B a battu le candidat A par b voix contre a, la probabilité qu'il ait toujours strictement mené lors du dépouillement (supposé aléatoire et uniforme sur les dépouillements possibles, c'est-à-dire des chemins NE qui aboutissent en (a,b)) vaut $\frac{b-a}{b+a}$.

D'abord, les chemins NE de $(0,0)$ à (a,b) qui ne coupent pas la diagonale commencent nécessairement par un pas N : de ceux-ci il faut enlever ceux qui recroisent la diagonale : or pour ces derniers, la réflexion de la partie après croisement (on remplace N par E et E par N) les met en bijection aves les chemins NE de $(0,1)$ a (b,a). Ainsi la probabilité cherchée peut sécrire :

$$\left(\binom{b+a-1}{a}-\binom{b+a-1}{b}\right)/\binom{b+a}{a}$$

Or

$$\binom{b+a-1}{a}-\binom{b+a-1}{b}=\frac{(b+a-1)!}{(b-1)!a!}-\frac{(b+a-1)!}{(b)!(a-1)!}=(b+a-1)!\frac{b-a}{(b)!a!}=\frac{b-a}{b+a}\frac{(b+a)!}{b!a!}=\frac{b-a}{b+a}\binom{b+a}{a}$$

L'intérêt du lemme est qu'il est valable pour tout couple (a,b), donc dans tout régime ; si N votes pour le candidat a, et $N+k$ pour le candidat b, on obtient par exemple, et si $k=o(N)$

$$\frac{k}{2N+k}\sim\frac{k}{2N}$$

2.2 Principe de réflexion 2 : temps d'atteinte de la marche aléatoire simple

Soit $0=S_0,S_1,S_2,\dots$ une marche aléatoire simple, c'est-à-dire une marche dont les incréments sont indépendants, égaux à ±1 avec probabilité $1/2$. On note $T_0(S)$ le temps d'atteinte de 0 par S, c'est-à-dire $T_0(S)=\min\{n,S_n=0\}\in\mathbb{N}\cup\{\infty\}$. On s'intéresse alors à la quantité $P_1(T_0(S)=N)$.

$$\begin{aligned}
P_1(T_0(S)=N)&=P_1(T_0(S)=N,S_{N-1}=1,S_N=0)\\
&=\frac{1}{2}\cdot P_1(T_0(S)>N-1,S_{N-1}=1)\\
&=\frac{1}{2}\cdot(P_1(S_{N-1}=1)-P_1(T_0(S)<N-1,S_{N-1}=1))\text{ du principe de réflexion}\\
&=\frac{1}{2}\cdot(P_1(S_{N-1}=1)-P_1(S_{N-1}=-1))
\end{aligned}$$

NB : on ne peut utiliser le LCLT pour estimer cette différence : en effet, la différence se trouve majorée par le terme d'erreur ; il faut donc passer par l'expression exacte avec les coefficients binomiaux.

On s'intéresse ensuite à la quantité $P_k(T_0(S)>N)$ pour $N\geq k$.

$$\begin{aligned}
P_k(T_0(S)>N)&=P_k(S_N>0,T_0(S)>N)=P_k(S_N>0)-P_k(S_N>0,T_0(S)\leq N)\\
&=P_k(S_N>0)-P_k(S_N>0,T_0(S)\leq N)\\
&=P_k(S_N>0)-P_k(S_N<0)\text{ du principe de réflexion}\\
&=P_0(S_N>-k)-P_0(S_N<-k)\text{ par translation}\\
&=P_0(-k<S_N\leq k)+P_0(S_N>k)-P_0(S_N<-k)\\
&=P_0(-k<S_N\leq k)
\end{aligned}$$

Et il est facile d'estimer précisément cette dernière quantité à l'aide du théorème central limite, ou de sa version locale (suivant que l'on veut faire dépendre k de N).

Le principe de réflexion admet une généralisation aux marches dont les incréments sont symétriques.

2.3 Lemme cyclique 1

Soit $x_1, \dots, x_N$ des entiers relatifs ≥ -1. On pose $x_i^{(k)} = x_{i+k[N]}$ le shift cyclique du vecteur x, puis on définit les sommes partielles $s_0^{(k)} = 0$ et pour $j \in \{1, \dots, N\}$, $s_j^{(k)} = \sum_{i=1}^{j} x_i^{(k)}$. Enfin pour $k \in \mathbb{Z}$, on pose $t_\ell(s) = \min\{j : s_j = k\} \in \mathbb{N} \cup \{\infty\}$ le temps d'atteinte de k par la marche s. On suppose $s_N = -\ell$ avec $\ell \in \mathbb{N}^\star$. Alors

$$\mathrm{Card}\{k \in \{0, \dots, N-1\} : t_{-\ell}(s^{(k)}) = N\} = \ell \tag{7}$$

On peut supposer sans perte de généralité que $t_{-\ell}(s) = N$, quitte à travailler avec $s^{(k)}$ pour k le plus petit entier k tel que s_k soit minimal. Alors les entiers comptés dans le membre de gauche de (7) sont exactement les entiers $0, t_{-1}(s), t_{-2}(s), \dots, t_{-(\ell-1)}(s)$.

Le lemme cyclique est dû à Feller.

L'hypothèse que les accroissements sont minorés par -1 est ici cruciale : on peut facilement mettre le lemme en défaut dans le cas contraire. En revanche, il n'y a pas de restriction sur la taille des accroissements positifs, ce qui représente une généralisation importante du lemme du scrutin.

Exercice 8. *Montrer que le lemme du ballot découle du lemme cyclique.*

On peut encore écrire (7) sous la forme équivalente suivante :

$$\sum_{k=0}^{N-1} \mathbf{1}_{\{t_{-\ell}(s^{(k)})=N\}} = \ell \mathbf{1}_{\{s_N=-\ell\}}$$

Si on considère alors une marche aléatoire $0 = S_0, S_1, S_2, \dots$ dont la loi des incréments vérifie $P(X \in \{-1, 0, 1, \dots\}) = 1$, alors on obtient la verison probabiliste suivante du lemme cyclique, très utile en pratique :

$$\begin{aligned}
P(T_{-\ell}(S) = N) &= \sum_{(x_1,\dots,x_N)\in(\{-1,0,1,\dots\})^N} P(X_1 = x_1, \dots, X_n = x_n)\mathbf{1}_{\{t_{-\ell}(s)=N\}} \\
&= \sum_{(x_1,\dots,x_N)} \frac{1}{N} \sum_{k=0}^{N-1} P(X_1 = x_1^{(k)}, \dots, X_n = x_n^{(k)})\mathbf{1}_{\{t_{-\ell}(s^{(k)})=N\}} \\
&= \sum_{(x_1,\dots,x_N)} P(X_1 = x_1, \dots, X_n = x_n)\frac{1}{N} \sum_{k=0}^{N-1} \mathbf{1}_{\{t_{-\ell}(s^{(k)})=N\}} \\
&= \sum_{(x_1,\dots,x_N)} P(X_1 = x_1, \dots, X_n = x_n)\frac{\ell}{N}\mathbf{1}_{\{s_N=-\ell\}} \\
&= \frac{\ell}{N} P(S_N = -\ell)
\end{aligned}$$

L'idée du calcul est de regrouper par paquets les $(x_1, \dots, x_N)$ qui se correspondent via une permutation cyclique et d'analyser chacun de ces paquets à l'aide du lemme cyclique. Le lemme cyclique admet une généralisation aux marches dont les incréments sont cycliquement échangeables (c'est-à-dire que les vecteurs $(X_1, \dots, X_N)$ et $(X_1^{(k)}, \dots, X_N^{(k)})$ ont même loi quelque soit k).

2.4 Lemme cyclique 2 : Fonction de parking

On appelle fonction de parking sur $[n]$ une fonction $f : [M] \to [N]$ [6] telle que

$$\text{Card}\{i \in [M] : f(i) \leq j\} \leq j \text{ pour tout } j \in [N].$$

On peut modéliser la situation comme suit : il y a M voitures, N places de parking et $f(i)$ est la place de parking préférée de la voiture i ; le procédé de parking est alors récursif : la voiture 1 choisit la position $f(1)$, puis la voiture 2 la position $f(2)$, sauf si celle ci est prise, auquel cas elle tente sa chance dans l'ordre décroissant des places, c'est-à-dire en $f(2) - 1$ puis en $f(2) - 2$, ... jusque 1. Si toutes ces places sont occupées, et que la voiture ne peut se garer, la fonction n'est pas une fonction de parking.

Puisque chaque voiture nécessite une place, il n'y a donc de fonctions de parking que pour $M \leq N$. Aussi, puisque la restriction à $[M-1]$ d'une fonction de parking : $[M] \to [N]$ est encore une fonction de parking, le nombre de fonctions de parking est une fonction croissante de M à N fixé.

On note $F_{N,M}$ l'ensemble des fonctions de parking. On construit une bijection entre les ensembles :

$$[N+1]^M \times [N+1-M] \leftrightarrow [N+1] \times F_{N,M}$$

L'idée est d'arranger N+1 places de parking dans l'ordre cyclique si bien que la place $N+1$ est suivie par la place 1 ; le processus de parking est alors le même que précédemment, dans le sens où une voiture qui n'obtient pas une place tente sa chance à la place précédente (et donc $N+1$ si 1 est occupée). Une fonction de $[M]$ dans $[N+1]$ définit alors une place de parking mais aussi $N+1-M$ places libres : renseigner le numéro d'une place libre donne alors une place de parking shiftée, et la valeur du shift est lélément de $[N+1]$ qui apparaît dans la bijection. De cette bijection on tire :

$$|F_{N,M}| = (N+1-M)(N+1)^{M-1}$$

Ainsi la probabilité qu'une fonction soit de parking est :

$$\frac{|F_{N,M}|}{N^M} = \frac{(N+1-M)}{N+1}(1+\frac{1}{N})^M \sim e \cdot \frac{k}{N} \text{ si } k = N+1-M \ll N$$

Pour comprendre cette bijection on étudiera d'abord le cas où $M = N$: il y a alors une seule place libre. On note qu'il existe un parallèle net avec le lemme cyclique : on pourrait peut-être aussi bien se ramener à ce lemme.

Exercice 9. *Réfléchir à un lien direct avec le lemme cyclique.*

2.5 Principe de réflexion 3 : une identité combinatoire et la loi de l'arcsinus

On a l'identité

$$\sum_{k=0}^{N} \binom{2k}{k}\binom{2(N-k)}{N-k} = 2^{2N} \tag{10}$$

Cette identité ressemble un peu à l'identité (5) mais il lui manque le troisième terme [7] Observons d'abord que :

$$\sum_{k\geq 0} 2^{2k} z^k = \frac{1}{1-4z} = \left(\frac{1}{\sqrt{1-4z}}\right)^2 \tag{11}$$

6. la notion générique de fonction de parking est plutôt associé au cas $M = N$

7. aussi, elle ne doit pas être confondue avec l'identité

$$\binom{2N}{N} = \sum_{k=0}^{N} \binom{N}{k}^2 = \sum_{k=0}^{N} \binom{N}{k}\binom{N}{N-k}$$

dont la preuve et l'interprétation combinatoire sont bien plus aisées (exercice !) ; cette dernière identité est un cas particulier de l'identité de Chu-Vandermonde

Rappelons la définition suivante de la double factorielle $(2k-1)!! = 1 \cdot 3 \cdot \ldots \cdot 2k-1$. Notons l'identité suivante sur le coefficient binomial *généralisé* :

$$\binom{-1/2}{k} = \frac{(-1/2)\ldots(-1/2-k+1)}{k!} = (-1)^k \frac{(2k-1)!!}{2^k k!} = (-1)^k \frac{2^k k!(2k-1)!!}{(2^k k!)^2} = (-1)^k 4^{-k} \frac{2k!}{k!k!} = (-1)^k 4^{-k} \binom{2k}{k}$$

À l'aide du théorème binomial généralisé (revenir à la formule de Taylor sinon), on déduit donc que

$$(1-4z)^{-1/2} = \sum_{k\geq 0} \binom{-1/2}{k} (-4z)^k = \sum_{k\geq 0} \binom{2k}{k} z^k$$

comme attendu : on déduit l'identité annoncée par identification des coefficients.

Nous savons aussi que $\binom{2k}{k}$ compte les marches NE de $(0,0)$ à (k,k) avec autant de pas N que de pas E, et aussi par rotation, les ponts de longueur $2k$ de marches de $(0,0)$ à $(2k,0)$ de pas ± 1 (c'est ce dernier point de vue qui sera ici choisi).

On recherche désormais une preuve bijective de cette identité. À gauche on a les paires de ponts de longueur $2k$ et $2(N-k)$ et à droite les marches de pas ± 1 de longueur $2N$.

Il existe une décomposition naturelle d'une telle marche en un pont et une marche non nulle, en considérant le dernier temps d'atteinte de 0 par la marche. Nous allons montrer que l'ensemble C_k des marches "non nulles" de longueur $2k$ et celui B_k des ponts de longueur $2k$, respectivement définis par :

$$C_k = \{(n_1,\ldots,n_{2k}) \in \{-1,1\}^{2k} : \ell \leq 2k \Rightarrow \sum_{j\leq \ell} n_j \neq 0\} \text{ et } B_k = \{(b_1,\ldots,b_{2k}) \in \{-1,1\}^{2k} : \sum_{j=1}^{2k} b_j = 0\}$$

sont en bijection, ce qui conclura la preuve.

2.5.1 Trois ensembles équivalents

La bijection, qui semble remonter à E. Nelson [8] est la suivante :

On partitionne B_k en deux sous-ensembles, B_k^+ et B_k^- en fonction du signe du premier pas, et de même on partitionne C_k en deux sous-ensembles C_k^+ et C_k^- , selon le signe des sommes partielles.

On appelle partie initiale d'un pont de B_k^- la partie du pont jusqu'au premier temps d'atteinte du minimum. On appelle partie initiale d'un chemin de C_k^+ la partie du chemin jusqu'au dernier temps d'atteinte de la valeur moitié *par un pas de type "+1"*. La bijection entre B_k^- et C_k^+renverse l'ordre et le signe des accroissements des parties initiales.

La bijection entre B_k^+ et C_k^- est similaire et laissée au lecteur (passer à l'opposé).

Exercice 12. *1. Proposer une preuve bijective (similaire à celle qu'on vient d'esquisser) au fait que l'ensemble D_k suivant est équipotent à B_k.*

$$D_k = \{(\ell_1,\ldots,\ell_{2k}) \in \{-1,1\}^{2k} : j \leq 2k \Rightarrow \sum_{i\leq j} \ell_i \geq 0\}$$

2. On définit ainsi une application $\phi : \{-1,1\}^k \to \{-1,1\}^k$ comme suit :

$$(i_1,\ldots,i_k) \mapsto (\ell_1,\ldots,\ell_k) = \phi(i_1,\ldots,i_k)$$

on pose $m = \max\{\sum_{\ell=1}^j i_\ell, j = 1\ldots k\}$ puis $t_m = \min\{j, \sum_{\ell=1}^j i_\ell = m\}$ et enfin $\ell_{t_m-1} = -i_{t_m-1}$, tandis que les autres coordonnés sont inchangées : $\ell_j = i_j$ si $j \neq t_m - 1$. L'application réciproque est alors ainsi définie : on pose $\theta(i_1,\ldots,i_k) = (i_k,\ldots,i_1)$ puis on note que

$$\theta(\phi(\theta(\ell_1,\ldots,\ell_k))) = (i_1,\ldots,i_k), \textit{i.e. } \theta\circ\phi\circ\theta = \phi^{-1}$$

Interpréter géométriquement ces opérations et en déduire une nouvelle preuve que l'ensemble D_k est équipotent à B_k

8. E. Nelson, 1932-2014, Professeur a Princeton, specialiste de physique mathématique et de logique

De l'identité (10), on déduit une propriété de symétrie intéressante de la marche aléatoire simple issue de 0 : Si $S = (0 = S_0, S_1, \dots, S_{2n})$ est une marche aléatoire de longueur $2N$, c'est à dire possède la loi uniforme sur les chemins de longueur 2N dont les incréments sont dans ± 1 , alors le dernier temps d'atteinte de 0 avant l'instant $2N$, $L_{2N}(S) = \max\{j \in 0, \dots, 2N : S_j = 0\}$ vérifie

$$P(L_{2N}(S) = 2k) = P(L_{2N}(S) = 2(N-k)) = \frac{\binom{2k}{k}\binom{2(N-k)}{N-k}}{\binom{2N}{N}}$$

Si l'on choisit $k = \alpha n$ entier, qu'obtient-t-on ? En s'aidant de l'équivalent sur le coefficient binomial milieu, on montre que, si $2\alpha N$ est un entier pair :

$$P(L_0(S) = 2\alpha N) = \frac{\binom{2\alpha N}{\alpha N}\binom{2(1-\alpha)N}{(1-\alpha)N}}{2^{2N}} \sim \frac{\frac{2^{2\alpha N}}{\sqrt{\pi \alpha N}} \frac{2^{2(1-\alpha)N)}}{\sqrt{\pi(1-\alpha)N}}}{2^{2N}} = \frac{1}{N} \frac{1}{\pi\sqrt{\alpha(1-\alpha)}}$$

ce qui fait apparaître la mesure de probabilité de densité $\frac{1}{\pi\sqrt{x(1-x)}}\mathbf{1}_{(0,1)}(x)$, encore appelée loi de l'Arcsinus. Menons en effet le calcul de la fonction de répartition de cette loi :

$$\int_0^y \frac{1}{\sqrt{x(1-x)}}dx = \int_{-1}^{2y-1} \frac{1}{\sqrt{\frac{1+z}{2}\frac{1-z}{2}}}dz/2 = \int_{-1}^{2y-1} \frac{1}{\sqrt{1-z^2}}dz = \text{Arcsin}(2y-1) - \text{Arcsin}(-1) = \text{Arcsin}(2y-1) + \frac{\pi}{2}$$

2.6 Application 1 : arbres plans enracinés, version 1

Un arbre au sens de la théorie des graphe est un graphe connexe et sans cycle (on considère des graphes non-dirigés sans boucles). Un arbre enraciné est alors un tel graphe avec un sommet dit marqué ou distingué. Enfin un arbre enraciné est dit plan lorsqu'un ordre est défini sur les arêtes adjacentes à tout sommet : ceci équivaut en fait à la donnée d'un plongement de l'arbre dans le plan en plus de la donnée d'un coin adjacent à la racine dans le cas.

Un arbre plan enraciné est alors défini de façon unique par son chemin de Dyck : celui-ci associe à chaque arête un signe $+1$ ou -1 selon que l'arête parcourue dans l'ordre dit de contour éloigne ou rapproche de la racine.

Ainsi les arbres plans enracinés à N arêtes, et donc $N-1$ sommets, sont en bijection avec l'ensemble des chemins de Dyck ainsi définis ; pour alléger les notations, on pose $s_k = \sum_{j=1}^{k} i_j$:

$$\{(i_1, i_2, \dots, i_{2N}) \in \{-1, 1\}^{2N} : s_k \geq 0, k = 1 \dots, 2N-1; s_{2N} = 0\}$$

Ajoutant un dernier -1 à la fin de chaque tel $2N$-uplet, on voit qu'on peut appliquer le lemme cyclique pour déduire que la taille de cet ensemble est égale au N-ième nombre de Catalan C_N, puisque :

$$\frac{1}{2N+1}\binom{2N+1}{N} = \frac{1}{2N+1}\frac{(2N+1)!}{N!(N+1)!} = \frac{1}{N+1}\frac{(2N)!}{N!N!} = \frac{1}{N+1}\binom{2N}{N} = C_N.$$

Exercice 13. *Observer que* $C_N = \binom{2N}{N} - \binom{2N}{N-1}$ *et proposer une interprétation combinatoire de cette différence à l'aide du principe de réflexion.*

On remarque aussi que la fonction génératrice a une forme close particulièrement explicite : ce n'est pas un miracle, nous y reviendrons quand nous détaillerons la méthode symbolique qui permet d'expliquer de manière clinique de nombreux résultats de dénombrement :

$$\sum_{k \geq 1} \frac{1}{k+1}\binom{2k}{k} z^k = \frac{1}{2z}(1 - \sqrt{1-4z}) \tag{14}$$

Exercice 15. *Retrouver cette identité par théorème binomial généralisé ou intégration de* (11).

2.7 Application 2 : Formule de Lagrange et arbres plans enracinés, version 2

On s'intéresse à une équation posée sur l'ensemble des séries formelles : soit ϕ analytique au voisinage de 0 avec $\phi(0) \neq 0$ donné. Alors l'équation

$$T(z) = z\phi(T(z))$$

admet une unique solution analytique dans un voisinage de 0, dont le terme général en n est donné par :

$$[z^N](T(z)) = \frac{1}{N}[z^{N-1}](\phi^N(z))$$

ou même, de façon un peu plus générale :

$$[z^N](T(z)^k) = \frac{k}{N}[z^{N-k}](\phi^N(z))$$

En général, on peut résoudre ces équations par identification des termes des deux membres du DSE : ceci donne un système déquations qui se résout de proche en proche ; il est bon de faire un essai avec une fonction ϕ simple pour s'assurer que la méthode fonctionne. Le théorème de Lagrange admet aussi une généralisation aux séries entières formelles. Il en existe des preuves reposant uniquement sur l'analyse complexe, cependant nous allons en donner une qui repose sur et montre le parallèle avec le lemme cyclique.

On commence par remarquer que la réponse est déjà connue dans un cas particulier : celui où ϕ est la fonction génératrice d'une loi de probabilité sur $\mathbb{N}$. Soit en effet μ une loi de probabilitié sur $\mathbb{N}$ de fonction génefatrice $\phi(z) = \sum \mu\{k\}z^k$. Soit T un μ-arbre de Bienaymé-Galton-Watson (BGW), alors le nombre total de sommets $|T|$ de T vérifie l'équation de récurrence suivante pour X variable aléatoire de loi μ indépendante de la suite i.i.d $T_1, T_2, \ldots$ d'arbres de μ-BGW :

$$T(z) = E[z^{|T|}] = E[z^{1+|T_1|+\ldots+|T_X|}] = z\phi(T(z))$$

et on sait aussi par la bijection dite de Lukaciewicz et du lemme cyclique que

$$[z^N](T(z)) = P(|T| = N) = \frac{1}{N}P(S_N = N-1) = \frac{1}{N}[z^{N-1}](\phi^N(z))$$

Si $T(z) = \sum_k p_k z^k$, alors les deux membres de l'égalité ci-dessus sont des fonctions polynomiales de $p_0, \ldots, p_N$, qui coïncident pour des réels $p_0, \ldots, p_N \geq 0$ tels que $\sum_{k=0}^N p_k \leq 1$. Le principe de prolongement polynomial garantit alors que les deux membres coincident pour tout choix de $p_0, \ldots, p_N$, ce qui achève la preuve.

2.7.1 Dénombrement des arbres : la méthode symbolique

On peut à nouveau considérer le problème du dénombrement des arbres plans enracinés. La somme suivante porte sur les arbres plans enracinés t, qu'on décompose dans la seconde égalité en la suite finie (éventuellement vide) de ses sous-arbres issus de la racine $t_1, t_2, \ldots$:

$$\begin{aligned} T(z) &= \sum_t z^{|t|} = \sum_t z^{1+|t_1|+|t_2|+\ldots} \\ &= z\sum_{k\geq 0}\sum_{(t_1,\ldots,t_k)}\prod_{i=1}^k z^{|t_i|} \\ &= z(1 + T(z) + T(z)^2 + T(z)^3 + \ldots) \\ &= \frac{z}{1-T(z)} \end{aligned}$$

On peut bien sûr résoudre cette équation du second degré et développer - ce faisant on obtient la solution (14). Cependant si on suppose connu le théorème de Lagrange, il est bien plus commode de l'utiliser :

$$[z^N]T(z) = \frac{1}{N}[z^{N-1}]\phi^n(z) = \frac{1}{N}[z^{N-1}]\frac{1}{(1-z)^N} = \frac{1}{N}\binom{2(N-1)}{N-1} = C_{N-1}$$

ayant noté pour le dernier calcul que

$$(1-z)^{-N} = (\sum_{k\geq 0} z^k)^N = \sum_{k_1,\dots,k_N\geq 0} z^{k_1+\dots+k_N} = \sum_{k\geq 0} z^k \Big(\sum_{k_1,\dots,k_N\geq 0} \mathbf{1}_{k_1+\dots+k_N=k} \Big) = \sum_{k\geq 0} \binom{N-1+k}{k} z^k$$

et le dernier calcul peut être obtenu à l'aide du théorème binomial généralisé puisque :

$$\binom{-N}{k} = \frac{-N\dots-N-k+1}{k!} = (-1)^k \frac{(N+k-1)\cdots N}{k!} = (-1)^k \binom{N+k-1}{k}$$

On retombe donc bien sur les nombres de Catalan. Bien entendu de notre point de vue, le raisonnement n'est pas du tout simplifié : on a utilisé le théorème de Lagrange qui repose sur le lemme cyclique et de nombreux calculs algébriques supplémentaires au lieu d'une simple application de ce lemme ! Dans la pratique cependant, le théorème de Lagrange est un instrument souple et puissant.

2.8 Retour sur les asymptotiques vues jusqu'à présent.

A ce moment du cours, il est bon de faire un point sur les asymptotiques rencontrées : d'abord le facteur le plus intéressant n'est pas le facteur exponentiel, qui est modèle dépendant, mais le facteur polynômial, qui révèle ce que "coûte" chacune des contraintes.

Les ponts, marches positives (on dirait non-négatives en anglais) et marches non nulles de longueur $2N$ sont tous trous comptés par $\binom{2N}{N}$, d'où un facteur polynomial en $N^{-1/2}$: c'est celui qui apparaît dans le TCL (qui peut lui-même être vu comme une généralisation de l'asymptotique du coefficient binomial médian via Stirling).

En revanche, les chemins de Dyck de longueur $2N$, c'est-à-dire les excursions positives sont comptées par les nombres de Catalan $\frac{1}{N+1}\binom{2N}{N}$, et ont donc une asymptotique en $N^{-3/2}$: le facteur N supplémentaire peut être vue comme l'application du lemme cyclique (une seule des N rotations convient). On peut enfin donner une dernière interprétation de ce facteur $N^{-3/2}$: les deux contraintes de positivité donnent un facteur $N^{-1/2}$ chacun, puis le fait que les deux excursions positives se rejoignent au même point implique un dernier facteur $N^{-1/2}$ du TCL.

2.9 Application 3 : limite locale 2

On peut s'intéresser à la limite locale de l'ensemble des chemins de Dyck :

$$\left\{ (i_1, i_2, \dots, i_{2N}) \in \{-1,1\}^{2N} : \sum_{j=1}^{2N} i_j = 0, \Big(k \leq 2N \Rightarrow \sum_{j=1}^{k} i_j \geq 0 \Big) \right\}$$

On cherche à caractériser la limite locale de la loi uniforme sur les chemins de Dyck. Pour le voir on se fixe un chemin de taille finie $(i_1^0, i_2^0, \dots, i_k^0)$ et on pose comme à l'accoutumée $s_k^0 = \sum_{j=1}^k i_j^0$ sa somme partielle. On rappelle la notation $t_{-\ell}(s)$ pour le premier temps d'atteinte de $-\ell$ par s. La probabilité qu'un chemin de Dyck uniforme de longueur $2N$ commence comme par ce chemin est alors :

$$\begin{aligned}
\mathbf{1}_{\cap_{j=1\dots k}\{s_j^0\geq 0\}} \cdot \frac{\mathrm{Card}\{s : t_{-s_k^0-1}(s) = 2N+1-k\}}{\mathrm{Card}\{s : t_{-1}(s) = 2N+1\}} &= \mathbf{1}_{\cap_{j=1\dots k}\{s_j^0\geq 0\}} \cdot \frac{\frac{s_k^0+1}{2N-k}\mathrm{Card}\{s_{2N+1-k} = -s_k^0 - 1\}}{\frac{1}{2N+1}\mathrm{Card}\{s_{2N+1} = -1\}} \\
&= \mathbf{1}_{\cap_{j=1\dots k}\{s_j^0\geq 0\}} \cdot (s_k^0+1) \cdot \frac{2N+1}{2N-k} \cdot \frac{\binom{2N+1-k}{N+1-\frac{k-s_k^0}{2}}}{\binom{2N+1}{N+1}} \\
&\sim \mathbf{1}_{\cap_{j=1\dots k}\{s_j^0\geq 0\}} \cdot (s_k^0+1) \cdot 2^{-k} \cdot e^{-\frac{(k-s_k^0)^2}{2.2N}}
\end{aligned}$$

pour k et s_k $O(\sqrt{N})$. A k fixé, ceci définit la loi d'une marche aléatoire biaisée ; on peut calculer les taux de transition de cette marche et constater que les probas de transition de k vers $k+1$ et de k vers $k-1$, pour $k \geq 1$ entier, valent respectivement $\frac{1}{2}\frac{k+2}{k+1} = \frac{1}{2}(1+\frac{1}{k+1})$ et $\frac{1}{2}\frac{k}{k+1} = \frac{1}{2}(1-\frac{1}{k+1})$. Cette chaîne de Markov est une version discrète du processus de Bessel 3.

NB : on aurait pu considérer aussi les marches qui reviennent en 0 pour la première fois à l'instant $2N$ ($t_0(s) = 2N$) plutôt que les chemins de Dyck, pour un résultat légèrement plus agréable : le facteur s_k^0+1 ci-dessus serait alors remplacé par s_k^0, ce qui est plus agréable.

3 Aire fixe

3.1 Partitions équitables

Tout d'abord un problème (issu de la RMS, numéroté RMS 772) : de combien de façons peut-on partionner l'ensemble $\{1, \ldots, 2n\}$ en deux sous-ensembles de taille égale et de même somme ? Comme d'habitude, on ne s'intéresse pas tant à une fomule fermée qu'à une estimée de bonne qualité valable pour N grand.

On commence par noter que la somme sur tous les éléments de l'ensemble $\sum_{i=1}^{2N} i = 2N(2N+1)/2 = N(2N+1)/2$ ne peut être un entier que si N est pair : on le supposera désormais.

Quelques notations : il s'agit d'évaluer le cardinal $q(N)$ de l'ensemble

$$\{1 \leq i_1 < \ldots < i_N \leq 2N : \sum_{j=1}^{N} i_j = \frac{2N(2N+1)}{4}\}$$

Cet ensemble est l'ensemble des partitions de $\frac{N(2N+1)}{2}$ en N parts distinctes (les parts sont les nombres $i_1, i_2, \ldots$ et l'inégalité stricte dans la définition implique bien qu'elles sont distinctes). La version probabiliste du problème est la suivante : si $(B_i, 1 \leq i \leq 2N)$ est une suite de variables de Ber(1/2) indépendantes, on cherche à calculer la probabilité :

$$P(\sum_{i=1}^{2N} B_i = N, \sum_{i=1}^{2N} iB_i = \frac{2N(2N+1)}{2}) = 2^{-2N} q(N)$$

Par l'application $i = (i_j)_{j=1}^N \mapsto (i_j - j)_{j=1}^N$, on a une *bijection* avec l'ensemble

$$\{0 \leq i_1 \leq \ldots \leq i_N \leq N : \sum_{j=1}^{N} i_j = \frac{N^2}{2}\}$$

c'est-à-dire l'ensemble des partitions de $N^2/2$ en au plus N parts de taille au plus N (noter que cette fois-ci les parts *peuvent être* égales) : ces partitions sont en bijection avec les chemins NE de $(0,0)$ à (N,N) qui découpent le carré en deux parties d'aire d'égale : ceci nous ramène à notre sujet des chemins NE et introduit une nouvelle contrainte intéressante.

La solution dérive d'une application du LCLT ; en effet on note que'on cherche à évaluer la probabilité qu'un vecteur aléatoire à valeurs prenne précisément la valeur égale à l'espérance de chacune de ses composantes :

$$E[\sum_{i=1}^{2N} B_i] = \frac{1}{2} \cdot 2N, \quad E[\sum_{i=1}^{2N} iB_i] = \frac{1}{2} \cdot \frac{2N(2N+1)}{2}$$

Néanmoins, il est délicat d'appliquer directement le LCLT : les variables aléatoires que l'on somme ne sont pas identiquement distribuées ; un soin particulier est requis pour vérifier la condition d'apériodicité dans ce cas. Calculons les éléments de la matrice de covariance Σ du vecteur aléatoire $\sum_i (B_i, iB_i)$; il s'agit de

$$\mathrm{Var}(\sum_{i=1}^{2N} B_i) = \frac{2N}{4} = \frac{N}{2}, \mathrm{Var}(\sum_{i=1}^{2N} iB_i) = \sum_{i=1}^{2N} \frac{i^2}{4} = \frac{1}{4}\frac{(2N)(2N+1)(4N+1)}{6}, \mathrm{Cov}(\sum_{i=1}^{2N} B_i, \sum_{i=1}^{2N} iB_i) = \frac{1}{4}\frac{2N(2N+1)}{2}$$

Donc :

$$\mathrm{Det}(\Sigma) \sim \Big(\frac{1}{3} - (\frac{1}{2})^2\Big)N^4 = \frac{1}{12}N^4$$

Une version non ici détaillée du LCLT (il y a de multiples hypothèses à vérifier) donne :

$$P(\sum_{i=1}^{2N} B_i = N, \sum_{i=1}^{2N} iB_i = \frac{N(2N+1)}{4}) \sim \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}} = \frac{\sqrt{3}}{\pi} \cdot \frac{1}{N^2} \Rightarrow q(N) \sim \frac{\sqrt{3}}{\pi} \cdot \frac{4^N}{N^2}$$

3.2 Les partitions d'entiers, et la formule de Hardy-Ramanajuan

Le problème plus classique consiste à dénombrer l'ensemble des partitions de N [9] :

$$p(N) = |\{i_1 \geq i_2 \geq \ldots \geq 1 : \sum_{j=1}^{N} i_j = N\}| = |\{(\ell_1, \ell_2, \ldots) \in \mathbb{N}^{\mathbb{N}} : \sum_{j=1}^{N} j\ell_j = N\}|$$

L'équivalence entre ces deux façons de représenter des partitions de N peut être vue comme suit : $\ell_k = \sum_j \mathbf{1}_{i_j=k}$ compte le nombre de parts égales à k dans la partition de N.

On détaille maintenant une méthode probabiliste à nouveau issue d'un exercice paru dans la RMS, RSM188, et dont la solution ici rapportée est dûe à Julien Bureaux. On commence par introduire les variables aléatoires suivantes : $X_j \sim \mathrm{Geom}(1 - e^{-\beta j})$, $j \in \mathbb{N}$, indépendantes. On pose alors $S = \sum_{j=1}^{\infty} jX_j$. On calcule alors les deux premiers moments de cette variable aléatoire, en calculant les valeurs de $\sum jz^j$ et $\sum j^2 z^j$ par dérivation de la série entière $\sum z^j = 1/(1-z)$:

$$E[S] = \sum_{j\geq 1} jE[X_j] = \sum_{j\geq 1} \frac{je^{-\beta j}}{1 - e^{-\beta j}} = \sum_{j\geq 1}\sum_{k\geq 1} je^{-\beta jk} = \sum_{k\geq 1} \frac{e^{-\beta k}}{(1 - e^{-\beta k})^2}$$

Ainsi, du théorème de convergence dominée,

$$\beta^2 E[S] \to \zeta(2) = \sum_{k\geq 1} \frac{1}{k^2} = \frac{\pi^2}{6}$$

$$\mathrm{Var}(S) = \sum_{j\geq 1} j^2 \mathrm{Var}[X_j] = \sum_{j\geq 1} \frac{j^2 e^{-\beta j}}{(1 - e^{-\beta j})^2} = \sum_{j\geq 1} j^2 \sum_{k\geq 1} ke^{-\beta jk} = \sum_{k\geq 1} k \sum_{j\geq 1} j^2 e^{-\beta jk} = \sum_{k\geq 1} ke^{-\beta k} \frac{1 + e^{-\beta k}}{(1 - e^{-\beta k})^3}$$

Ainsi, du théorème de convergence dominée,

$$\beta^3 \mathrm{Var}[S] \to 2\zeta(2) = \sum_{k\geq 1} \frac{2}{k^2} = \frac{\pi^2}{3}$$

9. On notera que le résultat est très différent selon qu'on ordonne ou non les entiers i_j - le cas non ordonné traité dans la première partie du cours est bien plus simple

et on note :

$$\begin{aligned}
P(S=N) &= \sum_{\ell_1,\ell_2,\ldots} \mathbf{1}_{\sum_j j\ell_j=N} P(\bigcap_{j=1}^{\infty} X_{\ell_j}=\ell_j) \\
&= \sum_{\ell_1,\ell_2,\ldots} \mathbf{1}_{\sum_j j\ell_j=N} \prod_{j=1}^{\infty} e^{-\beta j\ell_j}(1-e^{-\beta j}) \\
&= e^{-\beta N}\prod_{j=1}^{\infty}(1-e^{-\beta j}) \sum_{\ell_1,\ell_2,\ldots} \mathbf{1}_{\sum_j j\ell_j=N} \\
&= \mathrm{e}^{-\beta N}\prod_{j=1}^{\infty}(1-e^{-\beta j})p(N)
\end{aligned}$$

id est :

$$p(N) = P(S=N)\mathrm{e}^{\beta N}(\prod_{j=1}^{\infty}(1-e^{-\beta j}))^{-1}$$

Ce calcul contient l'idée clef qui justifie l'introduction de ces variables aléatoires : à savoir que la loi induite sur les (parts des) partitions de N est simplement la loi uniforme. En termes plus précis, la loi conditionnelle de $(X_1, X_2, \ldots)$ sachant $\sum kX_k = N$ est la loi uniforme sur l'ensemble des parts des partitions de N. Posons $f(\beta) = \prod_{j=1}^{\infty}(1-e^{-\beta j})$. Cette fonction vérifie :

$$-\log(f(\beta)) = -\sum_{j\geq 1}\log(1-e^{-\beta j}) = \sum_{j\geq 1}\sum_{k\geq 1}\frac{e^{-\beta jk}}{k} = \sum_{k\geq 1}\frac{e^{-\beta k}}{k(1-e^{-\beta k})}$$

Donc

$$-\beta\log(f(\beta)) \to \zeta(2)$$

Si l'on choisit $\beta = \sqrt{\frac{\zeta(2)}{N}}$ alors

$$p(N) = \log(P(S=N)) + \beta N - \log(f(\beta)) \leq \beta N - \log(f(\beta)),$$

(une probabilité étant plus petite que 1), et on a par ailleurs l'équivalent :

$$\beta N - \log(f(\beta)) \sim \sqrt{\frac{\zeta(2)}{N}}N + \frac{\zeta(2)}{\sqrt{\frac{\zeta(2)}{N}}} = 2\sqrt{\zeta(2)N} = \pi\sqrt{\frac{2}{3}N}$$

Pour la minoration, on fixe ϵ indépendant de N et on pose $\beta_\epsilon = \sqrt{\frac{\zeta(2)}{(1-\epsilon)N}}$, et également

$$I = [N_-, N_+] = [N(1-\epsilon) - kN^{3/4}, N(1-\epsilon) + kN^{3/4}]$$

avec N suffisamment grand pour que N majore le dernier terme. Par Bienaymé-Chebychev et nos estimées sur espérance et variance, on peut choisir k de sorte que la première inégalité ci dessous soit réalisée

$$1/2 \leq \sum_{\ell\in I} P(S=\ell) \leq \sum_{\ell\in I} e^{-\beta\ell}f(\beta)p(\ell) \leq 2kN^{3/4}e^{-\beta N_-}f(\beta)p(N)$$

où l'on a aussi utilisé la croissance de p. On en déduit

$$\log(p(N)) \geq \beta N_- - \log(f(\beta)) - \log(2kN^{3/4}) - \log(2)$$

Ceci implique :

$$\liminf_N \frac{\log(p(N))}{\sqrt{N}} \geq \liminf_N \frac{\beta_\epsilon N_- - \log(f(\beta_\epsilon))}{\sqrt{N}}$$

et cette dernière quantité est arbitrairement proche de $\sqrt{\frac{2}{3}}\pi$ lorsque $\epsilon \downarrow 0$.

La formule de Hardy-Ramanajuan implique l'équivalent plus fort suivant :

$$p(N) \sim \frac{1}{4\sqrt{3}N} e^{\pi\sqrt{2/3N}}$$

Il est possible de l'obtenir en suivant cette méthode avec un petit effort supplémentaire : précisément il faut en plus un équivalent de la fonction $f(\beta)$ (et non de son seul logarithme) et un TCL local pour la quantité $P(S = N)$.

3.2.1 Approche par les fonctions génératrices

La fonction génératrice de $p(N)$ est la suivante :

$$\phi(z) = \sum_{n\geq 1} p(n) z^n = \sum_{\ell_1,\ell_2,\ldots} z^{\sum_{k\geq 1} k\ell_k} = \sum_{\ell_1,\ell_2,\ldots} \prod_{k\geq 1} z^{k\ell_k} = \prod_{k\geq 1} \sum_{\ell_k \geq 0} z^{k\ell_k} = \prod_{k\geq 1} \frac{1}{1-z^k}$$

(on notera que $p(0) = 1$). Il n'est pas difficile de voir que le rayon de convergence de cette fonction génératrice vaut 1, il est en revanche très délicat d'obtenir à l'aide de cette fonction le développement asymptotique de $p(N)$: c'est un résultat célèbre de Hardy-Ramanajuan, qui passe par une analyse très technique des singularités des fonctions complexes.

Une manière de revisiter la méthode que nous venons de voir consiste à choisir un réel $t < 1$, puis à définir une variable aléatoire S_t qui vérifie :

$$P(S_t = n) = \frac{p(n)t^n}{\phi(t)} = p(n)t^n \prod_{k\geq 1}(1-t^k)$$

Si maintenant $t = e^{-\beta}$, alors on reconnaît bien la loi précédente :

$$P(S_t = n) = p(n)e^{-\beta n} \prod_{k\geq 1}(1-e^{-\beta k})$$

que l'on peut alors décomposer comme suit : $S_t = \sum kX_k$ avec $X_k \sim \mathrm{Geom}_{\mathbb{N}}(1-e^{-\beta k})$ des variables indépendantes. C'est en partant de cette considération que Baez-Duarte a dérivé la formule de Hardy-Ramanajuan. Le fait de se placer à l'intérieur du disque de convergence permet de travailler avec des suites sommables positives, et donc des probabilités.

3.2.2 Représentation des parts

Si l'on met la mesure uniforme P_N sur les partitions de N, alors les nombres de parts $(j_1, j_2, \ldots)$ deviennent des variables aléatoires (dépendant de N), que nous noterons $(J_1, J_2, \ldots)$.

Un corollaire immédiat de la méthode que nous venons de voir est que

$$(J_1, J_2, J_3) \text{ sous } P_N = (X_1, X_2, \ldots) \text{ sachant } \sum jX_j = N$$

où les $X_k \sim \mathrm{Geom}_{\mathbb{N}}(1-e^{-\beta k})$ sont des variables indépendantes. Ceci vaut pour tout paramètre β, mais le choix le plus intéressant est de poser $\beta = \sqrt{\frac{\zeta(2)}{N}}$, car alors $N = E[\sum jX_j]$, car alors on peut montrer que le conditionnement est de peu d'effet, au sens où la loi de $(X_1, X_2, \ldots, X_k)$ sachant $\sum jX_j = N$ est proche de la loi (non conditionnelle) $(X_1, X_2, \ldots)$