Constant time estimation of ranking statistics by analytic combinatorics

Cyril Banderier, Pierre Nicodème

Constant time estimation of ranking statistics by analytic combinatorics

C. Banderier*and P. Nicodème†

December 14, 2010

Abstract

We consider [1] i.i.d. increments (or jumps) X_i that are integers in $J \subseteq [-c, \dots, +d]$ for $c, d \in \mathbb{N}$, the partial sums $S_j = \sum_{1 \le i \le j} X_i$, and the discrete walks $((j, S_j))_{1 \le j \le n}$. Late conditionning by a return of the walk to zero at time n provides discrete bridges that we note $(B_j)_{1 \le j \le n}$. We give in this extended abstract the asymptotic law in the central domain of the height $(\max_{1 \le j \le n} B_j)$ of the bridges as n tends to infinity. As expected, this law converges to the Rayleigh law which is the law of the maximum of a standard Brownian bridge. In the case where $c = 1$ (only one negative jump), we provide a full expansion of the asymptotic limit which improves upon the rate of convergence $\mathcal{O}(\log(n)/\sqrt{n})$ given by Borisov [4] for lattice jumps; this applies in particular for the case where $X_i \in \{-1, +d\}$, in which case the expansion is expressible as a function of n, d and of the height of the bridge. Applying this expansion for $X_i \in \{-1, d/c\}$ gives an excellent approximation of the case $X_i \in \{-d, +c\}$ and provides in constant time an indicator used in ranking statistics; this indicator can be used for medical diagnosis and bioinformatics analysis (see Keller *et al.* [8] who compute it in time $\mathcal{O}(n \times \min(c, d))$ by use of dynamical programming).

1 Generating function of upper bounded bridges

We use in this article generating functions as main tool for the precise analysis of the behaviour of the walks. Asymptotics methods provide then excellent results for bioinformatics applications (see Section 3 and Figure 2).

We consider first the characteristic (Laurent) polynomial of the jumps

$$P(u) = p_d u^d + \cdots + p_{-c} \frac{1}{u^c},$$

where the p_i are weights. Then we define the generating function of the altitude of the walk at time k as

$$f_k(u) = \sum_{-kc \le j \le kd} f_{k,j} u^j, \tag{1}$$

where $f_{k,j}$ is the number of "walks at altitude j at time k" if $p_i = 1$ for all i, or the probability of this last event if $P(1) = 1$.

We consider walks that are forbidden to go upon a barrier h (the level h is permitted). We can write a recurrence for the Laurent polynomials $f_k(u)$, by removing the cases that make the walks go upon the barrier,

$$f_{k+1}(u) = f_k(u)P(u) - \sum_{i=1}^{d} u^{h+d}[u^{h+d}]f_k(u)P(u); \tag{2}$$

*LIPN, CNRS-UMR 7030, Université Paris-Nord, 93430 Villetaneuse, France. http://lipn.fr/~banderier

†LIX, CNRS-UMR 7161, École polytechnique, 91128 Palaiseau and AMIB Team, INRIA-Saclay, France. http://www.lix.polytechnique.fr/Labo/Pierre.Nicodeme

[1] This extended abstract summarizes a recently published article of Banderier and Nicodème [3]. See [5] for similar techniques.

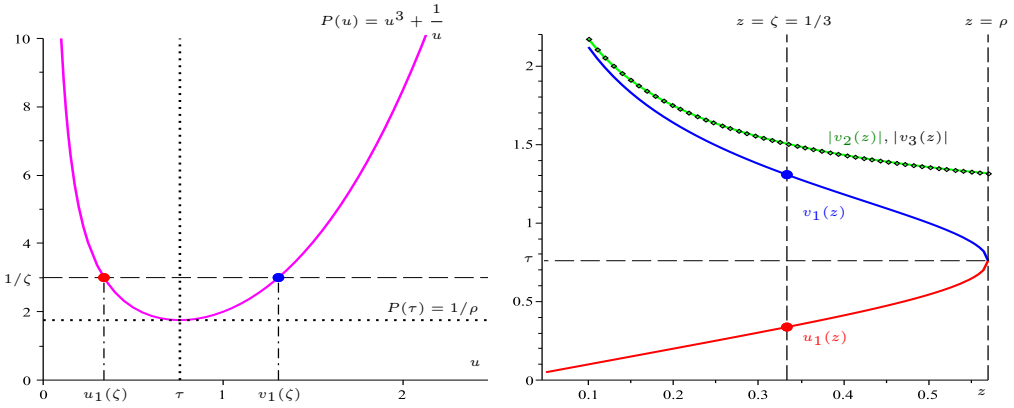

Figure 1: See [1, 2] for the relevant proofs (in particular, we do not comment upon the fact that some walks are *periodic* while others are not; see also [3]). If the equation $1 - zP(u) = 0$ has c small roots $u_i(z)$ (that tend to zero at the origin) and d large roots $v_j(z)$ (that tend to infinity at the origin), there are two real dominant roots of $1 - zP(u) = 0$ for $z \in]0, \rho]$, a small root (that we call $u_1(z)$) and a large root (that we call $v_1(z)$), such that $\max_{2\le i\le c} |u_i(z)| < u_1(|z|) < v_1(|z|) < \min_{2\le j\le d} |v_j(z)|$ in the disk $|z| < \rho$. (Figure 1 Left): behaviour of the particular characteristic polynomial $P(u) = u^3 + \frac{1}{u}$. (Figure 1 Right): a visual rendering of the domination property of the roots of $1 - zP(u) = 1 - z\left(u^3 + \frac{1}{u}\right)$ in the real interval $]0, \rho]$, where the number τ is the unique positive solution of $P'(z) = 0$ and $\rho = 1/P(\tau)$. For $\frac{1}{z} > \frac{1}{\rho}$ or $z < \rho$ the two dominant *real* solutions $u_1(z)$ and $v_1(z)$ are such that (i) $\lim_{z\to 0^+} u_1(z) = 0$ (small root) and $\lim_{z\to 0^+} v_1(z) = +\infty$ (dominant large root) and (ii) $u_1(z) < v_1(z)$. As a consequence of the identity $P'(\tau) = 0$, we have $u_1(\rho) = v_1(\rho)$. The non-dominant large roots $v_2(z)$ and $v_3(z)$ are algebraically conjugate and we have $u_1(z) < v_1(z) < |v_2(z)| = |v_3(z)|$ for $z \in]0, \rho[$.

We can multiply both sides of Equation 2 by z^{k+1} and sum up from $k = 0$ to $k = \infty$; this gives, (assuming that the walk starts at zero at time zero, or $f_0(u) = u^0 = 1$),

$$F^{[\le h]}(z, u)(1 - zP(u)) = 1 - zu^{h+1}F_{h+1}(z) - \cdots - zu^{h+d}F_{h+d}(z), \tag{3}$$

where $F^{[\le h]}(z, u) = \sum_{k\ge 0} z^k f_k(u)$ and the functions $F_{h+1}(z), \ldots, F_{h+d}(z)$ are unknown functions.

The popular *kernel* method [2] uses the fact that the roots $v(z)$ of the equation $1 - zP(u) = 0$ cancels the left member of Equation (3); there are d such solutions or *large roots* $v_j(z)$ $(1 \le j \le d)$ that tend to infinity as z tends to zero. These solutions then provide a linear system of d equations of the type

$$v_j(z)^{h+1}F_{h+1}(z) + \cdots + v_j(z)^{h+d}F_{h+d}(z) = 1/z.$$

By solving the system we obtain explicit expressions $F_{h+j}(z) = D_j(z)/V(z)$ where $V(z)$ is a Vandermonde determinant upon the roots $v_j(z)$ and the $D_j(z)$ are variants of it. We obtain

$$F^{[>h]}(z, u) = \frac{1}{1 - zP(u)} \sum_{j=1}^{d} \frac{u^{h+1}}{v_j(z)^{h+1}} \frac{Q_j(u)}{Q_j(v_j)}, \qquad \text{where} \qquad Q_j(t) = \prod_{\substack{1\le i\le d\\ i\ne j}} (t - v_i(z)). \tag{4}$$

Considering the *small* roots $u_j(z)$ of $1 - zP(u) = 0$ that tend to zero at the origin, we have from [2]

$$(-k < -c) \qquad [u^{-k}]\frac{1}{1 - zP(u)} = z\sum_{j=1}^{c} \frac{u_j'(z)}{u_j(z)^{-k+1}} = [u^0]\frac{u^k}{1 - zP(u)},$$

which gives for the generating function of bridges

$$[u^0]F^{[>h]}(z, u) = z\sum_{j=1}^{d} \frac{1}{v_j(z)} \sum_{i=1}^{c} \left(\frac{u_i(z)}{v_j(z)}\right)^h u_i'(z) \frac{Q_j(u_i(z))}{Q_j(v_j(z))}. \tag{5}$$

[2] For more informations about this method, see [1, 2, 6].

2 Asymptotics

We consider now that (i) $P(1) = 1$, (ii) $\mathbf{E}(X_i) = 0$ (or $P'(1) = 0$); these two conditions imply that $\tau = \rho = 1$ (see the legend of Figure 1 for relevants definitions). We let now n and h tend to infinity with the condition that $h = x\sigma\sqrt{n}$, where $x \in \mathbb{R}$ and $\sigma^2 = P''(1)$ is the variance of the jumps.

Using the properties of domination of the roots (see Figure 1), we obtain asymptotically

$$[u^0]F^{[>h]}(z,u) = z\left(\frac{u_1(z)}{v_1(z)}\right)^h \times \frac{u_1'(z)Q_1(u_1(z))}{v_1(z)Q_1(v_1(z))} \times \left(1+\mathcal{O}(C^h)\right) \qquad \text{for } |z|<1, \tag{6}$$

with

$$C = \max\left(\max_{j\geq 2}\sup_{|z|<\rho-\epsilon}\frac{|v_1(z)|}{|v_j(z)|},\ \max_{j\geq 2}\sup_{|z|<\rho-\epsilon}\frac{|u_1(z)|}{|u_j(z)|}\right).$$

Considering the expansion of $1/(P(u(z)) = z$ at $z = \rho = 1$, we get

$$z\sim 1^-\begin{cases} u_1(z) = 1-\sqrt{\frac{2}{\sigma^2}(1-z)}+\mathcal{O}(1-z), \qquad v_1(z) = 1+\sqrt{\frac{2}{\sigma^2}(1-z)}+\mathcal{O}(1-z)\\ \dfrac{Q_1(u_1(z))}{Q_1(v_1(z))} = \dfrac{Q_1(1)+\mathcal{O}(\sqrt{1-z})}{Q_1(1)+\mathcal{O}(\sqrt{1-z})} = 1+\mathcal{O}(\sqrt{1-z})\end{cases}$$

Equation (6) is valid in a Delta-domain (see [7]), and we have in such a domain

$$[u^0]F^{[>x\sigma\sqrt{n}]}(z,u) = \frac{z}{\sigma\sqrt{2}}\frac{\left(1-2\sqrt{\frac{2}{\sigma^2}(1-z)}\right)^{x\sigma\sqrt{n}}}{\sqrt{1-z}} \times \left(1+\mathcal{O}(\sqrt{1-z})\right)\times\left(1+\mathcal{O}(C^n)\right).$$

Using now the Semi-Large powers theorem (see [7] again), we obtain $[z^n][u^0]F^{[>x\sigma\sqrt{n}]} = \frac{\sqrt{n}}{\sigma\sqrt{2}}\times e^{-2x^2}\times\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$. Since [2] obtained previously $[z^n][u^0]F^{]-\infty,+\infty[} = \frac{\sqrt{n}}{\sigma\sqrt{2}}\times\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$, we get to

$$\Pr\left(\max_{0\leq i\leq n} B_i > x\sigma\sqrt{n}\right) = \text{Rayleigh}(x)\times\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \qquad \text{where} \qquad \text{Rayleigh}(x) = e^{-2x^2}.$$

Łukasiewicz bridges: $J = \{-1,\dots,c\}$.
We have in this case

$$Q_1(u_1(z)) = \frac{1}{p_d z}\frac{\partial}{\partial u}\frac{u(1-zP(u))}{u-v_1(z)}\bigg|_{u=u_1(z)} = \frac{1}{p_d z^2}\frac{u_1(z)}{u_1'(z)(u_1(z)-v_1(z))}. \tag{7}$$

The value of $Q_1(v_1(z))$ follows by interchanging the rôles of u_1 and v_1. This leads to

$$\begin{aligned}\frac{\Pr\left(\max_{1\leq j\leq n} B_j > x\sigma\sqrt{n}\right)}{\exp(-2x^2)} = 1 &+ \frac{(-(2/3)x\xi/\zeta^{3/2}-6x/\sqrt{\zeta})}{\sqrt{n}} + \frac{1}{n}\left((-2-\frac{10}{9}\frac{\xi^2}{\zeta^3}+\frac{2}{3}\frac{\theta}{\zeta^2}-\frac{16}{3\zeta}-\frac{8}{3}\frac{\xi}{\zeta^2})x^4\right.\\ &\left.+(\frac{24}{\zeta}+\frac{5}{3}\frac{\xi^2}{\zeta^3}+3-\frac{\theta}{\zeta^2}+\frac{20}{3}\frac{\xi}{\zeta^2})x^2-\frac{5}{\zeta}-\frac{3}{8}-\frac{7}{6}\frac{\xi}{\zeta^2}-\frac{5}{24}\frac{\xi^2}{\zeta^3}+\frac{1}{8}\frac{\theta}{\zeta^2}+\frac{5}{24}\frac{\xi^3}{\zeta^3}-\frac{1}{8}\frac{\theta^2-3\zeta^2}{\zeta^2}\right)\\ &+\mathcal{O}\left(\frac{1}{n^{3/2}}\right),\end{aligned} \tag{8}$$

where $\zeta = \sigma^2 = P''(1)$, $\xi = P'''(1)$ and $\theta = P''''(1)$. The *algolib* Maple package (more precisely, the *gdev* and *equivalent* functions developed by Bruno Salvy, see algo.inria.fr/librairies) can naturally push the expansion to higher orders.

The particular case: $J = \{-1,+d\}$.
In this case, it is possible to compute the values of the derivatives of $P(u)$ evaluated at 1 as functions of d. For $P(u) = \frac{u^d}{d+1}+\frac{d}{d+1}\frac{1}{u}$, we have in particular $P''(1) = d, P'''(1) = d(d-4),\dots$; substituting these values in Equation (8) leads to an efficient formula (which has a constant time complexity in n, h, and d).

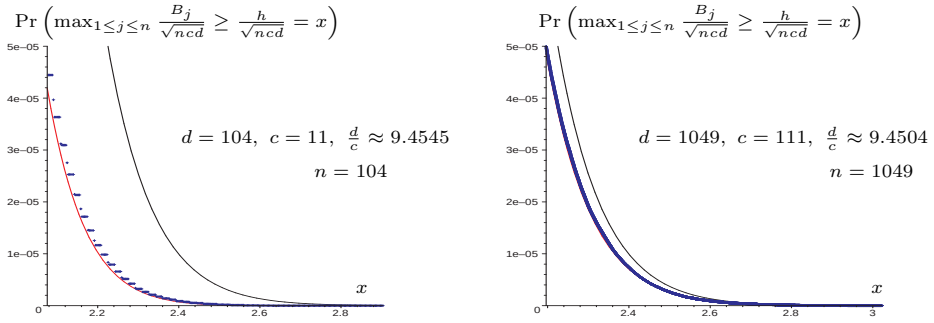

Figure 2: Convergence of the height of discrete bridges to the Rayleigh distribution (top continuous curve); the dotted curves correspond to simulations; the lower continuous curve follows from a refinement of Equation (8), computed up to $\mathcal{O}(1/n^2)$, with $P(u) = \frac{1}{1+d/c}u^{d/c} + \frac{d/c}{1+d/c}\frac{1}{u}$ (heuristics). We emphasize the excellent precision of our asymptotics for small n.

3 Bioinformatics application

A set of G genes is expressed in a given tissue; this provides a ranking of level of expression of these genes. Considering now the same ranking and a subset of specific interest of g genes, if these g genes have a high level of expression, they will mostly appear at the top of the ranking. The aim is to provide a statistical estimator for exceptional behaviours. Keller *et al.* [8] proposed the following approach: while scanning from left to right the ranking of the G genes, build a random walk $(B_i)_{0\le i\le G}$ starting at zero and such that its jump at time i is

$$\begin{cases} G-g & \text{if the gene at rank } i \text{ belongs to } g,\\ -g & \text{if the gene at rank } i \text{ belongs to } G-g. \end{cases}$$

By construction, we have $B_0 = B_G = 0$ and these walks are therefore bridges. The tail probability (referred to as p-value) that Keller *et al.* choose as statistical indicator is p-value $= \Pr(\max_{1\le i\le G}\left|B_i\right| > h)$, for any chosen h. They provide a dynamic programming algorithm computing this indicator in complexity $\mathcal{O}(G\times g)$. We compute heuristically the indicator in constant time by setting $n = G$ and $P(u) = \frac{1}{1+d/c}u^{d/c} + \frac{d/c}{1+d/c}\frac{1}{u}$ with $d = G-g$ and $c = g$, and applying Equation (8); see Figure 2 for an example.

References

[1] Banderier, C. *Combinatoire analytique des chemins et des cartes.* Ph.D. thesis, Université Paris VI, 2001.

[2] Banderier, C., and Flajolet, P. Basic analytic combinatorics of directed lattice paths. *Theoretical Computer Science 281, Issue 1-2* (2002), 37–80. (Special volume dedicated to M. Nivat).

[3] Banderier, C., and Nicodème, P. Bounded discrete walks. In *Proc. AofA 2010* (2010), DMTCS, pp. 35–48. Vienna, June 2010.

[4] Borisov, I. S. On the rate of convergence in the "conditionnal" invariance principle. *Theory of Probability and its applications 23*, 1 (1978), 63–76.

[5] Bousquet-Mélou, M. Discrete Excursions. *Sémin. Lothar. Comb. 57* (2008).

[6] Bousquet-Mélou, M., and Petkovšek, M. Linear recurrences with constant coefficients: The multivariate case. *Discrete Math. 225*, 1-3 (2000), 51–75.

[7] Flajolet, P., and Sedgewick, R. *Analytic combinatorics.* Cambridge University Press, 2009. (810 pages). See also `http://algo.inria.fr/flajolet/Publications/books.html`.

[8] Keller, A., Backes, C., and Lenhof, H. P. Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics 290*, 8 (2007). Also available as `http://www.biomedcentral.com/1471-2105/8/290`.