Wiener-Hopf factorization and maximal scores in biological sequences

Pierre Nicodème
INRIA-Rocquencourt

February 10, 1997

[summary by Philippe Robert]

1. Introduction

In this talk we study a matching problem for two sequences $S = (s_1, \ldots, s_n)$, $T = (t_1, \ldots, t_p)$ where $s_1, \ldots, s_n, t_1, \ldots, t_p$ are elements of some alphabet A. This mathematical model is used in the analysis of some biological sequences. To any couple $(x, y) \in A \times A$, one associates a score $H(x, y) \in \mathbb{R}$ which is negative if the two letters do not agree or positive if their affinity is significant. A local matching of length l of these sequences is given by a sequence $((s_{i_1}, t_{j_1}), (s_{i_1}, t_{j_1}), \ldots, (s_{i_1}, t_{j_1}))$ with $1 \leq i_1 < i_2 < \cdots < i_l \leq n$ and $1 \leq j_1 < \cdots < j_l \leq p$. The score of this matching is then defined as

$$\sum_{k=1}^{l} H(s_{i_k}, t_{j_k}).$$

The main problem considered in this talk is to estimate the maximal score among all the possible matchings of these sequences.

A probabilistic setting is used to give estimates of this optimal score. The letters are assumed to be drawn independently from the alphabet. This hypothesis leads to a formulation of the problem in terms of random walk. The optimal score $M(n)$ for two sequences of size n can be represented as

$$M(n) = \sup_{0 \leq j \leq k \leq n} (S_k - S_j), \tag{1}$$

where $S_n = \sum_{i=1}^{n} X_i$ ($\sum_1^0 = 0$); The variables (X_i) are assumed to be independent and identically distributed. The sequence (S_n) is the random walk starting from 0 associated to the distribution of X_1. Clearly the sequence $(M(n))$ is non decreasing with n, and as we will see, it converges to infinity as $n \to +\infty$. Our goal is to find an asymptotic estimate of $M(n)$ for n large. We prove that if $E(X_1) < 0$, and some other technical conditions, there exists some constant α such that the renormalized sequence $M(n) - \alpha \log(n)$ converges in distribution.

2. The relation with a reflected random walk

For $n \geq 1$ we denote by

$$W_n = \sup_{0 \leq k \leq n} (S_n - S_k),$$

then $M(n)$ can be expressed as

$$M(n) = \sup_{0 \leq k \leq n} W_k. \tag{2}$$

It is easy to see that the sequence (W_n) satisfies the following relation

$$W_{n+1} = (W_n + X_{n+1})^+, \qquad n \geq 0 \tag{3}$$

where $x^+ = \max(x, 0)$. Now define

$$\nu_- = \inf\{n > 0 / S_n \leq 0\},$$

which is the first time the random walk visits the negative axis. The law of large numbers gives that almost surely $\lim_{n\to+\infty} \frac{1}{n}\sum_1^n X_i = E(X_1)$, and because $E(X_1) < 0$, we have $\lim_{n\to+\infty} \sum_1^n X_i = -\infty$, thus the quantity ν_- is always finite.

By induction, using (3), one can check that $W(n) = S_n$ for $n < \nu_-$. Furthermore, we have $W_{\nu_-} = (S_{\nu_-})^+ = 0$, by definition of ν_-. It is easy to prove that starting from $t = \nu_-$, the sequence W_n performs another similar excursion above 0 independently of the previous excursion, and so on. The sequence (W_n) is the reflected random walk at 0.

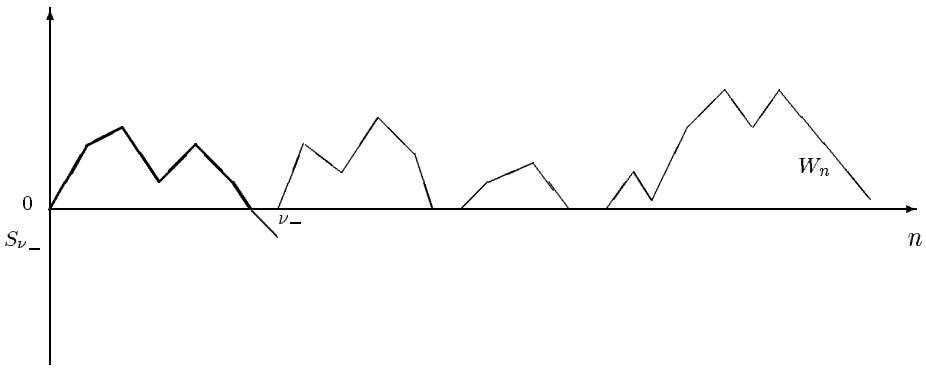

FIGURE 1. The path of $n \to W_n$

The method of resolution. To estimate $M(n)$, the maximum of $k \to W_k$ on the interval $\{0, \ldots, n\}$, we proceed as follows

1. Estimate the distribution of M_{exc}, the maximum of the random walk during an excursion above 0.
2. Count the number of excursions of $k \to W_k$ above 0 up to time n.
3. $M(n) = \max\{M_{exc_i}\}$ where the maximum is taken on all excursions exc_i before time n. The excursions being independent, the same is true for the (M_{exc_i}). Estimating the maximum of independent random variables is easy.

Technically, the main tool used to prove convergences is the renewal theorem. The explicit calculation of constants involved in the limiting distribution requires the Wiener-Hopf factorization associated to this random walk. We give a formulation of these results in the next section.

3. The probabilistic tools

3.1. **The renewal theorem.** We consider a sequence of non negative i.i.d. integrable random variables (Y_i) and denote by $T_n = \sum_1^n Y_i$, the non decreasing random walk associated to the distribution of Y_1. For $t \in \mathbb{R}_+$, let N_t be the number of T_k between 0 and t. Thus, $T_{N_t+1} - t$ is the length of the interval between t and the first T_k after t.

Proposition 1. *Almost surely*

$$\lim_{t\to+\infty} \frac{N_t}{t} = \frac{1}{E(Y_1)},$$

where $E(Y_1)$ denotes the expected value of Y_1.

As we can remark at that point the above proposition will solve the second point of our program above. In this case the beginnings of the excursions will define the renewal process.

The main result in renewal theory concerns the solution of the so-called renewal equation. If f is some function, the function Z is the solution of the renewal equation associated to f if, for all $x \geq 0$,

$$Z(x) = f(x) + \int_0^x Z(x-y)P(X_1 \in dy). \tag{4}$$

The main theorem is the following

Theorem 1. *If f is Riemann integrable, there is a unique solution Z_f of (4) and Z_f satisfies*

$$\lim_{x\to+\infty} Z_f(x) = 1/E(X_1) \int_0^{+\infty} f(u)du.$$

This analytical formulation of the renewal theorem can be seen as a consequence of a probabilistic result: the variable $T_{N_t+1} - t$ converges in distribution as $t \to +\infty$.

3.2. **The Wiener-Hopf factorization.** This technique concerns the calculation of the distribution of the hitting times of the positive, negative axis by a random walk and the position of the random walk at these times. We have already seen ν_-, we define its positive counterpart ν_+,

$$\nu_+ = \inf\{n/S_n > 0\}.$$

Theorem 2. *For $u \in \mathbb{C}$, such that $|u| < 1$, there exist $\phi_+(u,.)$, $\phi_-(u,.)$ such that*

1.

$$\frac{1}{1 - uE(e^{-\xi X})} = \phi_+(u,\xi)\phi_-(u,\xi), \qquad \Re(\xi) = 0. \tag{5}$$

2. *The function $\phi_+(u,.)$ [resp. $\phi_-(u,.)$] is analytic on $\{\Re(\xi) > 0\}$ [resp. $\{\Re(\xi) < 0\}$], continuous, bounded away from 0 and ∞ on $\{\Re(\xi) \geq 0\}$ [resp. $\{\Re(\xi) \leq 0\}$]. Moreover*

$$\lim_{\Re(\xi)\to+\infty} \phi_+(u,\xi) = 1.$$

Such a decomposition is unique.

The following corollary is the probabilistic interpretation of the above theorem.

Corollary 1. *The functions of the Wiener-Hopf factorization can be expressed as*

$$\phi_+(u,\xi) = \frac{1}{1 - E(u^{\nu_+}e^{-\xi S_{\nu_+}})}, \qquad |u| < 1, \quad \Re(\xi) \geq 0,$$

$$\phi_-(u,\xi) = \frac{1}{1 - E(u^{\nu_-}e^{-\xi S_{\nu_-}})} \qquad |u| < 1, \quad \Re(\xi) \leq 0.$$

Thus, if we are able to decompose the function $1/(1 - uE(e^{-\xi X}))$, the joint distributions of (ν_+, S_{ν_+}) and (ν_-, S_{ν_-}) are known through their Fourier-Laplace transforms, $E(u^{\nu_+}e^{-\xi S_{\nu_+}})$, $E(u^{\nu_-}e^{-\xi S_{\nu_-}})$.

4. The main results

Using theorem 1, one can prove the proposition about the tail distribution of the maximum of the random walk during an excursion.

Proposition 2. *If the following conditions are satisfied,*

- $E(X_1) < 0$, $P(X_1 > 0) > 0$ *and* X_1 *is non arithmetic;*
- *There exists* $\theta > 0$ *such that* $E(e^{\theta X_1}) < +\infty$*;*
- $E(|X_1| e^{\gamma X_1}) < +\infty$*, where* γ *is the positive solution of* $E(e^{\gamma X_1}) = 1$*;*

then

$$\lim_{x \to +\infty} e^{\gamma x} P\left(M_{exc} \geq x\right) = C_{exc} = \frac{P(\nu_+ = +\infty)(1 - E(e^{\gamma S_{\nu_-}}))}{\gamma E(X_1 e^{\gamma X_1}) E(\nu_+ e^{\gamma S_{\nu_+}} 1_{\{\nu_+ < +\infty\}})}.$$

Notice that to make the constant C_{exc} explicit, one has to know some functionals of ν_+, ν_-. This is the place where the Wiener-Hopf decomposition is useful.

At that point cases 1 and 2 of our program are solved. For point 3 it remains to integrate these results. This gives our final theorem.

Theorem 3. *Under the hypotheses of the proposition 2, there exist two constant* K, λ *such that*

$$\lim_{n \to +\infty} P(M(n) - \frac{\log(n)}{\lambda} \leq x) = e^{-Ke^{-\lambda x}}.$$

References

[1] Asmussen (Søren). – *Applied Probability and Queues.* – John Wiley & Sons, Chichester, 1987, *Wiley Series in Probability and Mathematical Statistics.*

[2] Feller (William). – *An introduction to probability theory and its applications.* – John Wiley & Sons, New York, 1971, 2nd edition, vol. II.

[3] Iglehart (Donald L.). – Extreme values in the $GI/G/1$ queue. *Annals of Mathematical Statistics*, vol. 43, n° 2, 1972, pp. 627–635.

[4] Karlin (Samuel) and Altschul (Stephen F.). – Methods for assesing the statistical significance of molecular sequences features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA*, vol. 87, 1990, pp. 2264–2268.

[5] Karlin (Samuel) and Dembo (Amir). – Strong limit theorems of empirical functionals for large exedances of partial sums of i.i.d. variables. *Annals of Probability*, vol. 19, n° 4, 1991, pp. 1737–1755.