Local Maximal Stack Scores with General Loop Penalty Function

EVA 2005, Gothenburg

Niels Richard Hansen

Local Maximal Stack Scores with General Loop Penalty Function

EVA 2005, Gothenburg

Niels Richard Hansen

This talk is based on two papers:

- Asymptotics for Local Maximal Stack Scores with General Loop Penalty Function. *To be submitted shortly*.
- The Maximum of a Random Walk Reflected at a General Barrier. *To appear in Ann. Appl. Probab.*

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.
- Its (secondary) structure is a graph with vertex set $\{1, \ldots, n\}$.

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.
- Its (secondary) structure is a graph with vertex set $\{1, \ldots, n\}$.
- The graph is a partial matching: A vertex can enter in at most one edge and no loops.

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.
- Its (secondary) structure is a graph with vertex set $\{1, \ldots, n\}$.
- The graph is a partial matching: A vertex can enter in at most one edge and no loops.
- Typically edges between near neighbours (sharp turns) are not allowed.

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

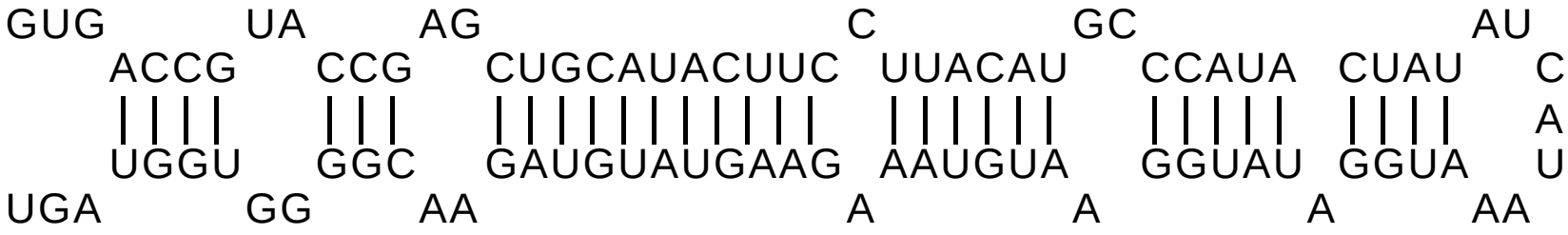
- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.
- Its (secondary) structure is a graph with vertex set $\{1, \ldots, n\}$.
- The graph is a partial matching: A vertex can enter in at most one edge and no loops.
- Typically edges between near neighbours (sharp turns) are not allowed.
- Typically pseudo-knots are not allowed: Pairs of edges of the form $\{i_1, j_1\}$ and $\{i_2, j_2\}$ with $i_1 < i_2 < j_1 < j_2$ are not allowed.

RNA-structures

RNA molecules are sequences of nucleotides – some forming functionally important structures.

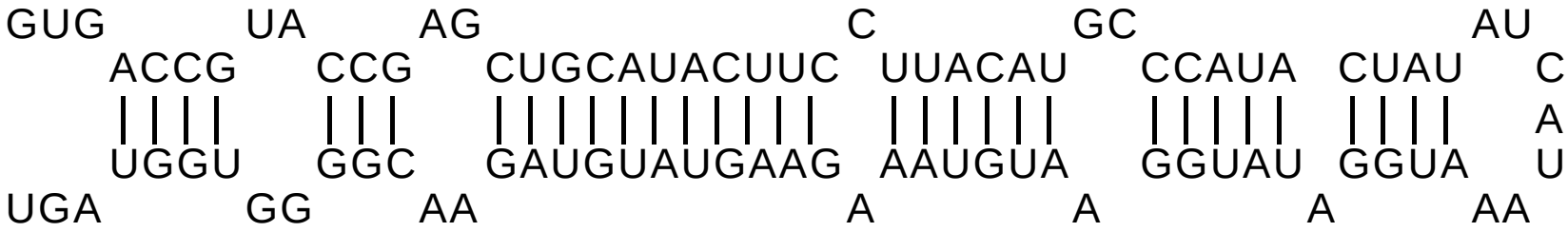
- An RNA-molecule is represented as a sequence, $X_1 \ldots X_n$, of letters from the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$.
- Its (secondary) structure is a graph with vertex set $\{1, \ldots, n\}$.
- The graph is a partial matching: A vertex can enter in at most one edge and no loops.
- Typically edges between near neighbours (sharp turns) are not allowed.
- Typically pseudo-knots are not allowed: Pairs of edges of the form $\{i_1, j_1\}$ and $\{i_2, j_2\}$ with $i_1 < i_2 < j_1 < j_2$ are not allowed.
- An edge represents a hydrogen bond between nucleotides.

RNA-structures



An example RNA-molecule from the nematode *C. elegans*.

RNA-structures



An example RNA-molecule from the nematode *C. elegans*.

Xiong and Waterman (1997) show strong limit results for the maximum of (minus) the free energy score of RNA-structures. The free energy score being

- an additive score of the hydrogen bonded nucleotides (edges) plus
- linear penalties on the length of the loops (unpaired vertices).

The score depends on a parameter vector α.

Strong Limits

Let $X_1, \ldots, X_n$ be an iid RNA-sequence. Let $T_{i,j}$ denote the maximal structure score for $X_i, \ldots, X_j$ for $i < j$ and

$$M_n = \max\{\max_{1 \leq i < j \leq n} T_{i,j}, 0\}.$$

Strong Limits

Let $X_1, \ldots, X_n$ be an iid RNA-sequence. Let $T_{i,j}$ denote the maximal structure score for $X_i, \ldots, X_j$ for $i < j$ and

$$M_n = \max\{\max_{1 \le i < j \le n} T_{i,j}, 0\}.$$

Relying on subadditive techniques Xiong and Waterman show that

$$\lim_{n \to \infty} \frac{1}{n} T_{1,n} = a(\alpha) \quad a.s.$$

Strong Limits

Let $X_1, \ldots, X_n$ be an iid RNA-sequence. Let $T_{i,j}$ denote the maximal structure score for $X_i, \ldots, X_j$ for $i < j$ and

$$M_n = \max\{\max_{1 \le i < j \le n} T_{i,j}, 0\}.$$

Relying on subadditive techniques Xiong and Waterman show that

$$\lim_{n \to \infty} \frac{1}{n} T_{1,n} = a(\alpha) \quad a.s.$$

If $a(\alpha) > 0$,

$$\lim_{n \to \infty} \frac{1}{n} M_n = a(\alpha) \quad a.s.$$

and if $a(\alpha) < 0$

$$\lim_{n \to \infty} \frac{1}{\log n} M_n = b(\alpha) \quad a.s.$$

A Conjecture

In the logarithmic phase, $a(\alpha) < 0$, Xiong and Waterman conjecture that

$$\mathbb{P}(M_n > t) \simeq 1 - \exp(-K(\alpha) n \exp(-t/b(\alpha))) \tag{1}$$

for suitable large n and t.

A Conjecture

In the logarithmic phase, $a(\alpha) < 0$, Xiong and Waterman conjecture that

$$\mathbb{P}(M_n > t) \simeq 1 - \exp(-K(\alpha) n \exp(-t/b(\alpha))) \tag{1}$$

for suitable large n and t.

For a (quite restrictive) class of stack/hairpin-loop structures we show such a result. Our result contains situations corresponding to $a(\alpha) = 0$ but where (1) holds.

Local scores

We proceed as follows:

- Choose functions $f : \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^2 \to \mathbb{R}$ (non-lattice) and $g : \mathbb{N}_0 \to (-\infty, 0]$.

Local scores

We proceed as follows:

- Choose functions $f : \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^2 \to \mathbb{R}$ (non-lattice) and $g : \mathbb{N}_0 \to (-\infty, 0]$.
- For $1 \leq i < j \leq n$ define

$$T_{i,j} = \max_{-2 \leq 2\delta < j-i} \left\{ \sum_{k=0}^{\delta} f(X_{i+k}, X_{j-k}) + g(j - i - 2\delta - 1) \right\}.$$

$$X_1 \dots X_{i-1} \quad \underbrace{X_i \dots X_{i+\delta}}_{\delta+1} \quad \underbrace{X_{i+\delta+1} \dots X_{j-\delta-1}}_{j-i-2\delta-1} \quad \underbrace{X_{j-\delta} \dots X_j}_{\delta+1} \quad X_{j+1} \dots X_n.$$

stack hairpin-loop stack

Local scores

We proceed as follows:

- Choose functions $f : \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^2 \to \mathbb{R}$ (non-lattice) and $g : \mathbb{N}_0 \to (-\infty, 0]$.
- For $1 \leq i < j \leq n$ define

$$T_{i,j} = \max_{-2 \leq 2\delta < j-i} \left\{ \sum_{k=0}^{\delta} f(X_{i+k}, X_{j-k}) + g(j - i - 2\delta - 1) \right\}.$$

stack hairpin-loop stack

$$X_1 \dots X_{i-1} \quad \underbrace{X_i \dots X_{i+\delta}}_{\delta+1} \quad \underbrace{X_{i+\delta+1} \dots X_{j-\delta-1}}_{j-i-2\delta-1} \quad \underbrace{X_{j-\delta} \dots X_j}_{\delta+1} \quad X_{j+1} \dots X_n.$$

- Let $M_n = \max_{1 \leq i < j \leq n} T_{i,j}$.

The Recursion

The scores $T_{i,j}$ fulfill the recursion

$$T_{i,j} = \max\{T_{i+1,j-1} + f(X_i, X_j), g(j-i+1)\}.$$

	X_1	X_2	X_3	X_4	X_5
X_1	$g(1)$				
X_2	0	$g(1)$			
X_3		0	$g(1)$		
X_4			0	$g(1)$	
X_5				0	$g(1)$

The Recursion

The scores $T_{i,j}$ fulfill the recursion

$$T_{i,j} = \max\{T_{i+1,j-1} + f(X_i, X_j), g(j-i+1)\}.$$

	X_1	X_2	X_3	X_4	X_5
X_1	$g(1)$	$T_{1,2}$			
		↗			
X_2	0	$g(1)$			
X_3		0	$g(1)$		
X_4			0	$g(1)$	
X_5				0	$g(1)$

The Recursion

The scores $T_{i,j}$ fulfill the recursion

$$T_{i,j} = \max\{T_{i+1,j-1} + f(X_i, X_j), g(j-i+1)\}.$$

	X_1	X_2	X_3	X_4	X_5
X_1	$g(1)$	$T_{1,2}$	$T_{1,3}$		
		↗	↗		
X_2	0	$g(1)$			
X_3		0	$g(1)$		
X_4			0	$g(1)$	
X_5				0	$g(1)$

The Recursion

The scores $T_{i,j}$ fulfill the recursion

$$T_{i,j} = \max\{T_{i+1,j-1} + f(X_i, X_j), g(j-i+1)\}.$$

	X_1	X_2	X_3	X_4	X_5
X_1	$g(1)$	$T_{1,2}$	$T_{1,3}$	$T_{1,4}$	
		↗	↗	↗	
X_2	0	$g(1)$	$T_{2,3}$		
			↗		
X_3		0	$g(1)$		
X_4			0	$g(1)$	
X_5				0	$g(1)$

The Recursion

The scores $T_{i,j}$ fulfill the recursion

$$T_{i,j} = \max\{T_{i+1,j-1} + f(X_i, X_j), g(j-i+1)\}.$$

	X_1	X_2	X_3	X_4	X_5
X_1	$g(1)$	$T_{1,2}$	$T_{1,3}$	$T_{1,4}$	$T_{1,5}$
X_2	0	$g(1)$	$T_{2,3}$	$T_{2,4}$	$T_{2,5}$
X_3		0	$g(1)$	$T_{3,4}$	$T_{3,5}$
X_4			0	$g(1)$	$T_{4,5}$
X_5				0	$g(1)$

The Diagonals

Suppose $(X_k)_{k\in\mathbb{Z}}$ is a doubly infinite sequence of iid variables. Define recursively

$$T_k^0 = \max\{T_{k-1}^1 + f(X_{-k}, X_k), g(2k)\}, \quad T_0^0 = 0$$

and

$$T_k^1 = \max\{T_{k-1}^2 + f(X_{-k}, X_k), g(2k+1)\}, \quad T_0^1 = g(1).$$

The Diagonals

Suppose $(X_k)_{k\in\mathbb{Z}}$ is a doubly infinite sequence of iid variables. Define recursively

$$T_k^0 = \max\{T_{k-1}^1 + f(X_{-k}, X_k), g(2k)\}, \quad T_0^0 = 0$$

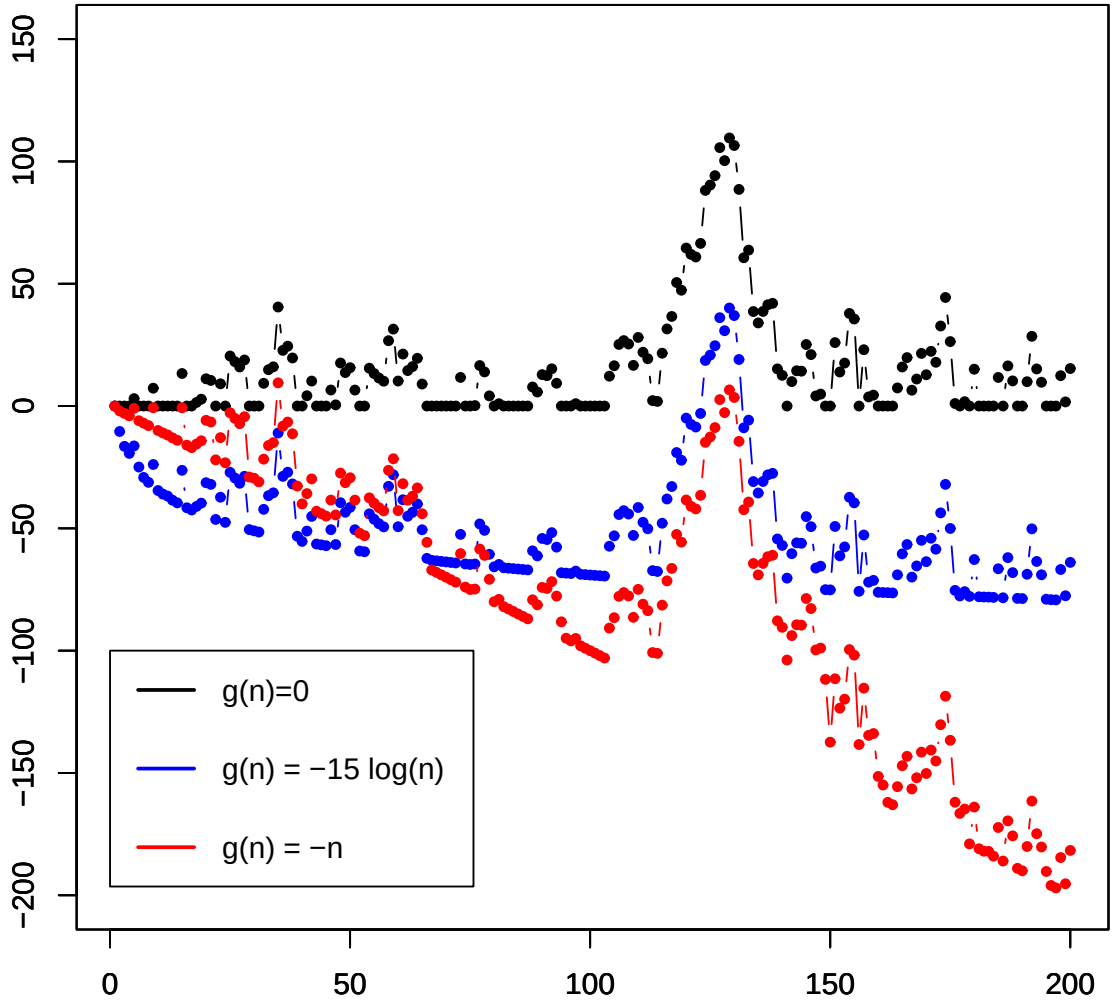
and

$$T_k^1 = \max\{T_{k-1}^2 + f(X_{-k}, X_k), g(2k+1)\}, \quad T_0^1 = g(1).$$

$$T_{i,j} \stackrel{\mathcal{D}}{=} \begin{cases} T^0_{(j-i+1)/2} & \text{if } j-i \text{ is odd} \\ T^1_{(j-i)/2} & \text{if } j-i \text{ is even} \end{cases}$$

Reflected Random Walks

The processes $(T_k^i)_{k\geq 0}$, $i = 0, 1$ are random walks reflected at g.



Reflected Random Walks

If $M^i := \sup_{k \geq 0} T_k^i < \infty$ a.s. and $\theta^* > 0$ solves

$$\mathbb{E} \exp(\theta f(X_{-1}, X_1)) = 1.$$

then

$$\mathbb{P}(M^i > x) \sim K_i^* \exp(-\theta^* x)$$

for $x \to \infty$.

Reflected Random Walks

If $M^i := \sup_{k \geq 0} T_k^i < \infty$ a.s. and $\theta^* > 0$ solves

$$\mathbb{E} \exp(\theta f(X_{-1}, X_1)) = 1.$$

then

$$\mathbb{P}(M^i > x) \sim K_i^* \exp(-\theta^* x)$$

for $x \to \infty$.

Its necessary that

$$\mu := \mathbb{E} f(X_{-1}, X_1) < 0$$

in which case

$$\sum_{k=1}^{\infty} \exp(\theta^* g(k)) < \infty$$

is sufficient for $M^i < \infty$ a.s.

The Main Result

Define

$$C(t) = \sum_{i=1}^{n} 1(\exists \delta : T_{i-\delta,i+\delta} > t) + 1(\exists \delta : T_{i-\delta,i+1+\delta} > t).$$

Theorem: *With*

$$t_n = \frac{\log(K_0^* + K_1^*) + \log n + x}{\theta^*}, \tag{1}$$

for $x \in \mathbb{R}$ *then*

$$||\mathcal{D}(C(t_n)) - \mathrm{Poi}(\exp(-x))||_{tv} \to 0 \tag{1}$$

for $n \to \infty$. *In particular*

$$\mathbb{P}(M_n \leq t_n) \to \exp(-\exp(-x)) \tag{1}$$

for $n \to \infty$.

A consequence of the theorem is that

$$\frac{1}{\log n} M_n \xrightarrow{\mathbb{P}} \frac{1}{\theta^*}.$$

The "parameters" involved are the functions f and g and

$$b(f, g) = \frac{1}{\theta^*}$$

where $\theta^* > 0$, solving $\mathbb{E} \exp(\theta f(X_{-1}, X_1)) = 1$, does not depend upon g.

A consequence of the theorem is that

$$\frac{1}{\log n} M_n \xrightarrow{\mathbb{P}} \frac{1}{\theta^*}.$$

The "parameters" involved are the functions f and g and

$$b(f, g) = \frac{1}{\theta^*}$$

where $\theta^* > 0$, solving $\mathbb{E} \exp(\theta f(X_{-1}, X_1)) = 1$, does not depend upon g.

Moreover, for suitable n and t

$$\mathbb{P}(M_n > t) \simeq 1 - \exp(-(K_0^* + K_1^*) n \exp(-\theta^* t))$$

The Proof

Apply Arratia et al. (1989) “Two moments suffice for Poisson approximations: the Chen-Stein method”. It involves:

- Localisation of dependencies by band-limitation: Consider only $T_{i,j}$ with $j - i \leq h(n)$ where

$$\lim_{n\to\infty} h(n)^{-1} \log n = \lim_{n\to\infty} n^{-\epsilon} h(n) = 0.$$

The Proof

Apply Arratia et al. (1989) “Two moments suffice for Poisson approximations: the Chen-Stein method”. It involves:

- Localisation of dependencies by band-limitation: Consider only $T_{i,j}$ with $j - i \leq h(n)$ where

$$\lim_{n\to\infty} h(n)^{-1} \log n = \lim_{n\to\infty} n^{-\epsilon} h(n) = 0.$$

- Handling of the tail-behavior of partial maxima of reflected random walks due to band-limitation.

The Proof

Apply Arratia et al. (1989) “Two moments suffice for Poisson approximations: the Chen-Stein method”. It involves:

- Localisation of dependencies by band-limitation: Consider only $T_{i,j}$ with $j - i \leq h(n)$ where

$$\lim_{n\to\infty} h(n)^{-1} \log n = \lim_{n\to\infty} n^{-\epsilon} h(n) = 0.$$

- Handling of the tail-behavior of partial maxima of reflected random walks due to band-limitation.
- Bounding probabilities of the form

$$\mathbb{P}(T_{i,j} > t, T_{i',j'} > t)$$

by the Azuma-Hoeffding inequality and exponential change of measure.

Back to Xiong and Waterman

The variables $T_{1,n}$ do not form a subadditive sequence.

By other means one can sometimes establish that

$$\lim_{n\to\infty} \frac{1}{n} T_{1,n} = a(f,g).$$

Back to Xiong and Waterman

The variables $T_{1,n}$ do not form a subadditive sequence.

By other means one can sometimes establish that

$$\lim_{n\to\infty} \frac{1}{n} T_{1,n} = a(f,g).$$

Using that $g \leq 0$ and $\mu < 0$

$$\limsup_{n\to\infty} \frac{1}{n} T_{1,n} \leq 0.$$

If g is sublinear, $g(n)/n \to 0$,

$$\frac{1}{n} T_{1,n} \geq \frac{g(n)}{n} \to 0,$$

hence $a(f,g) = 0$.

Example I

Let $g(n) = \rho n$ for $\rho < 0$. Then

$$\sum_{k=1}^{\infty} \exp(\theta^* \rho k) < \infty.$$

and $M^i < \infty$ a.s.

Example I

Let $g(n) = \rho n$ for $\rho < 0$. Then

$$\sum_{k=1}^{\infty} \exp(\theta^* \rho k) < \infty.$$

and $M^i < \infty$ a.s. If $\rho < \mu$

$$a(f, g) = \mu$$

and if $\rho > \mu$

$$a(f, g) = \rho.$$

Example II

Let $g(n) = \rho \log n$ for $\rho < 0$. Then

$$\sum_{k=1}^{\infty} \exp(\theta^* \rho \log k) = \sum_{k=1}^{\infty} k^{\theta^* \rho} < \infty$$

iff $\rho < -1/\theta^*$ and $a(f, g) = 0$.

Example II

Let $g(n) = \rho \log n$ for $\rho < 0$. Then

$$\sum_{k=1}^{\infty} \exp(\theta^* \rho \log k) = \sum_{k=1}^{\infty} k^{\theta^* \rho} < \infty$$

iff $\rho < -1/\theta^*$ and $a(f, g) = 0$.

It is possible to show that for $\rho > -1/\theta^*$ then $M^i = \infty$ a.s. for $i = 0, 1$. What happens here is an open question.

Example II

Let $g(n) = \rho \log n$ for $\rho < 0$. Then

$$\sum_{k=1}^{\infty} \exp(\theta^* \rho \log k) = \sum_{k=1}^{\infty} k^{\theta^* \rho} < \infty$$

iff $\rho < -1/\theta^*$ and $a(f, g) = 0$.

It is possible to show that for $\rho > -1/\theta^*$ then $M^i = \infty$ a.s. for $i = 0, 1$. What happens here is an open question.

When $g \equiv 0$ (the limiting case $\rho \to 0$) is understood and

$$\frac{1}{\log n} M_n = \frac{2}{\theta^*} \ a.s.$$

with a corresponding asymptotic extreme value distribution of M_n.

Concluding Remarks

- The use of extreme value distributions in local sequence alignment for significance evaluation of the alignment score is much used (BLAST) with a theoretical justification for special cases.

Concluding Remarks

- The use of extreme value distributions in local sequence alignment for significance evaluation of the alignment score is much used (BLAST) with a theoretical justification for special cases.
- We have provided a result for sequence structure where one finds that the structure score follows asymptotically an extreme value distribution.

Concluding Remarks

- The use of extreme value distributions in local sequence alignment for significance evaluation of the alignment score is much used (BLAST) with a theoretical justification for special cases.
- We have provided a result for sequence structure where one finds that the structure score follows asymptotically an extreme value distribution.
- Our result is particular useful when searching large sequences for local parts containing “a lot of structure”.

Concluding Remarks

- The use of extreme value distributions in local sequence alignment for significance evaluation of the alignment score is much used (BLAST) with a theoretical justification for special cases.
- We have provided a result for sequence structure where one finds that the structure score follows asymptotically an extreme value distribution.
- Our result is particular useful when searching large sequences for local parts containing “a lot of structure”.
- The result confirms to some extend the conjecture by Xiong and Waterman – and extends the conjecture in one direction.