A SURVEY ON STRONG APPROXIMATION TECHNIQUES IN CONNECTION WITH RECORDS

D. Pfeifer and Y.-S. Zhang

University of Oldenburg and Beijing University of Iron and Steel

Abstract. The intention of the paper is to review various techniques for the strong approximation of record times, inter-record times and record values (essentially by Poisson and Wiener processes) which have been considered in the recent years by different authors. Besides the iid case, we also discuss which of the methods described are suited to treat corresponding problems in more general (non-iid) settings.

1. Introduction.

Let $\{X_n; n \geq 1\}$ be an iid sequence of random variables (r.v.'s) on some probability space $(\Omega, \mathcal{A}, P)$, with a joint continuous cumulative distribution function F. *Record times* $\{U_n; n \geq 0\}$, *inter-record times* $\{\Delta_n; n \geq 0\}$ and *record values* $\{X_{U_n}; n \geq 0\}$ for this sequence are recursively (and by our assumptions, a.s. well-)defined by

$$(1) \qquad U_0 = \Delta_0 = 1,\ U_{n+1} = \inf\{k; X_k > X_{U_n}\},\ \Delta_{n+1} = U_{n+1} - U_n \quad (n \geq 0).$$

These sequences, related to the partial extremes of the original sequence, have been of increasing interest since their first exploration by Chandler (1952). Several survey articles have been published since then, e.g. by Glick (1978) or Nevzorov (1988), to mention some. Besides structural properties of these sequences, asymptotic features (in various meanings) have been the subject of research very early, pointing out relationships with the law of large numbers, the central limit theorem and the law of the iterated logarithm for iid sequences of r.v.'s. In particular, it was shown by several authors that

$$(2) \qquad \frac{1}{n} Z_n \to 1 \text{ a.s. } (n \to \infty)\ ,\quad n^{-1/2}(Z_n - n) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ and } \lim \begin{matrix}\sup\\ \inf\end{matrix} \frac{Z_n - n}{\sqrt{2n \log\log n}} = \pm 1 \text{ a.s.}$$

where $\xrightarrow{\mathcal{D}}$ means convergence in distribution, and Z_n may be replaced by any of the r.v.'s $\log U_n$, $\log \Delta_n$ or $-\log(1-F(X_{U_n}))$, $n \geq 1$. (For further details, c.f. e.g. Rényi (1962), Neuts (1967), Strawderman and Holmes (1970), and Shorrock (1972a, 1972b).) Due to the fact that for the proofs of the forementioned results, mostly different techniques were used, tailored to the specific situation, there was a need for a unifying approach explaining the similarity of the asymptotic behaviour of the three different random sequences in (2). It turned out that certain strong approximation techniques in the spirit of Komlós, Major and Tusnády (1976) were suitable tools for such an explanation. Their main result states that under certain regularity conditions (existence of exponential moments) it is possible to approximate a partial sum process $S_n = \sum_{k=1}^{n} \xi_k$ (with an iid sequence $\{\xi_k\}$) by a partial sum process $T_n = \sum_{k=1}^{n} \eta_k$ where $\{\eta_k\}$ is a sequence of iid normally distributed r.v.'s with the same mean and variance as the $\{\xi_k\}$ sequence, on the same probability space (if rich enough to carry sufficiently many iid sequences) such that

$$(3) \qquad |S_n - T_n| = O(\log n) \text{ a.s.} \qquad (n \to \infty).$$

In this paper, such techniques for the strong approximation of record times, inter-record times and record values are reviewed. The possibility of extending some of the results to more general than iid cases is also discussed.

However, before doing so, it will be necessary to expose some well-known structural properties of the sequences involved.

<u>**Theorem 1.**</u> (Rényi (1962)) Under the conditions specified above, the record time sequence $\{U_n\}$ forms a homogeneous Markov chain (MC) with transition probabilities given by

$$(4) \qquad P(U_{n+1} > k \mid U_n = j) = \frac{j}{k}, \qquad 1 \leq j \leq k,\ n \geq 0.$$

<u>**Theorem 2.**</u> (Shorrock (1972a, 1972b)) The inter-record time sequence $\{\Delta_n\}$ is conditionally independent given the sequence of record values, with a (conditional) geometric distribution of the form

$$(5) \qquad P(\Delta_n = k \mid X_{U_{n-1}}) = \{1-F(X_{U_{n-1}})\}F^{k-1}(X_{U_{n-1}}) \qquad \text{a.s.} \qquad (n,k \geq 1).$$

Furthermore, the sequence $\{-\log(1-F(X_{U_n}))\}$ forms the arrival time sequence of a unit rate Poisson process (i.e., has independent exponentially distributed increments with unit mean).

From Theorem 2 it follows immediately that the record value sequence itself forms the arrival time sequence of such a Poisson process if the underlying distribution is exponential with unit mean.

In what follows we shall discuss in more detail the different strong approximation approaches for record times, inter-record times and record values, both individually and also jointly.

2. The conditional independence approach.

Deheuvels (1982, 1983) proved that it is possible to define an iid sequence $\{Y_n\}$ on the same probability space $(\Omega, \mathcal{A}, P)$ (if rich enough) such that

$$(6) \quad \Delta_n = \mathrm{int}\{Y_n/-\log(1-\exp(-S_n))\}+1 \quad \text{with } S_n = -\log(1-F(X_{U_{n-1}})),\ n \geq 1,$$

where $\{Y_n\}$ is exponentially distributed with unit mean, and independent of the record value sequence. This reflects precisely the conditional independence property of Theorem 2 (note that a geometrically distributed r.v. can be generated from an exponentially distributed one by appropriate rounding off). By a suitable Taylor expansion, relation (6) leads to relation

$$(7) \quad \log \Delta_n = \log Y_n + S_n + o(1) \quad \text{a.s.} \quad (n \to \infty)$$

which in turn proves relation (2) for inter-record times via (3), jointly with the transformed record value sequence (i.e. with the same strong approximand $\{T_n\}$ from the Komlós-Major-Tusnády construction). This is due to the fact that by a simple Borel-Cantelli argument,

$$(8) \quad \log Y_n = O(\log n) \quad \text{a.s.} \quad (n \to \infty).$$

(A slightly refined expansion of (7) is given in Pfeifer (1985).)

Unfortunately, this approach does not immediately lead to a nice strong approximation of record times from where a direct proof of (2) could be read off easily (cf. Deheuvels (1982)). The following approach provides such an approximation.

3. The Markov chain approach.

This approach is based on Theorem 1, expanding ideas of Williams (1973) and Westcott (1977). The following result (Pfeifer (1987)) is a key to the procedure.

Theorem 3. Let $\{M_n\}$ be a homogeneous MC with conditional cumulative distribution function $F(.\mid.)$. Let $F_-(.\mid.)$ denote the corresponding left-continuous version. Then there exists an iid sequence of uniformly distributed r.v.'s (over $(0,1)$) on the same probability space (if rich enough), $\{V_n\}$, say, such that

$$(9) \quad V_{n+1} = (1-W_{n+1})F(M_{n+1} \mid M_n) + W_{n+1}F_-(M_{n+1} \mid M_n) \quad (n \geq 0)$$

where $\{W_n\}$ is also iid uniformly distributed, and independent of the MC $\{M_n\}$.

Using relation (4), this translates in our case into

$$(10) \qquad V_{n+1} = (1-W_{n+1})\frac{U_n}{U_{n+1}} + W_{n+1}\frac{U_n}{U_{n+1}-1} \qquad (n \geq 0)$$

using the fact that with V_n also $1-V_n$ is uniformly distributed. Taking logarithms in (10) one obtains the following strong approximation result (Pfeifer (1987)).

<u>**Theorem 4.**</u> There exists on the same probability space $(\Omega, \mathcal{A}, P)$ (if rich enough) an arrival time sequence $\{S_n^*\}$ of a unit-rate Poisson process and a non-negative r.v. Z possessing all positive moments, with mean $E(Z) = 1-C$ ($C = .577216$ denoting Euler's constant), such that

$$(11) \qquad Z \text{ and } \{(S_n^*-n)/\sqrt{n}\} \text{ are asymptotically independent}$$

$$(12) \qquad \log U_n = Z+S_n^*+o(1) \text{ a.s.}, \ \log \Delta_n = Z+S_n^*+\log(1-\exp(S_{n-1}^*-S_n^*))+o(1) \text{ a.s.} (n \to \infty)$$

where $\{-\log(1-\exp(S_{n-1}^*-S_n^*))\}$ again is an iid sequence of exponentially distributed r.v.'s with unit mean.

This again proves relation (2) via (3), this time jointly for record times and inter-record times.

It should be pointed out that the Poisson process in Theorem 4 does not coincide with the one in Deheuvels' approach. Actually, we have

$$(13) \qquad S_n^* = \sum_{k=1}^{n} -\log V_k$$

with $\{V_k\}$ as in (10).

We should like to mention that with the approaches 2. and 3., also moment estimations of the logarithms of record times and inter-record times are readily obtained (cf. Pfeifer (1984a), and Nevzorov (1988)).

Recently, Deheuvels (1988) has extended the one-dimensional MC approach to the two-dimensional case (record times and record values jointly form a MC, too). With this approach, he was able to extend the strong approximation by a Poisson process to all three sequences in (2) simultaneously.

Instead of expressing everything in terms of Poisson processes, the Komlós-Major-Tusnády construction also allows for a formulation in terms of (standard) Wiener processes. Namely, if $\{W(t); t \geq 0\}$ stands for such a process, the forementioned results show that it is possible to establish on the same probability space (if rich enough) the following strong relationship, which is even closer to (2):

$$(14)\quad \log U_n = n + W(n) + O(\log n) \quad a.s.\ (n \to \infty)$$
$$\log \Delta_n = n + W(n) + O(\log n) \quad a.s.\ (n \to \infty)$$
$$-\log(1-F(X_{U_n})) = n + W(n) + O(\log n) \quad a.s.\ (n \to \infty)$$

(see Deheuvels (1988), and Pfeifer (1986).)

4. The embedding approach.

Going back to Resnick (1973, 1974, 1975) and Resnick and Rubinovitch (1973), the basic idea here is an appropriate embedding of the partial maxima sequence $\{\max(X_1,\ldots,X_n);\ n \geqq 1\}$ derived from $\{X_n\}$ into so-called *extremal processes*. Any such process $\{E(t);t>0\}$ is a pure jump Markov process with right continuous paths and finite-dimensional marginal distributions which in our case are given by

$$(15)\quad P\Big(\bigcap_{i=1}^{k}\{E(t_i)\leq x_i\}\Big) = F^{t_1}(\min\{x_1,\ldots,x_k\})\prod_{i=2}^{k} F^{t_i-t_{i-1}}(\min\{x_i,\ldots,x_k\})$$

for all selections $0<t_1<t_2<\ldots<t_k$ of time points, and values $x_1,\ldots,x_k \in \mathbb{R}$. Such an extremal process 'interpolates' the partial maxima process in that we have

$$(16)\quad \{\max(X_1,\ldots,X_n)\} \overset{\mathcal{D}}{=} \{E(n)\},$$

where $\overset{\mathcal{D}}{=}$ means equality in distribution. The interesting point is here that the jump time sequence $\{\tau_n\}$ of the extremal process in the interval $(1,\infty)$ forms a non-homogeneous Poisson point process with intensity $1/t$, $t \geqslant 1$, such that the points are a.s. clustering in the intervals (U_n-1,U_n), $n \geqq 1$. It follows that the surplus number S of extremal jumps over the record times is a.s. finite, with $E(S|\Sigma) = E(Z|\Sigma)$, where Σ denotes the σ-field generated by the record values (Pfeifer (1986)). Here Z is the r.v. from Theorem 4, such that we also have

$$(17)\quad E(S) = 1 - C.$$

An application of the log function now shows that

$$(18)\quad \log U_n = \log \tau_{n+S} + o(1) = \log \tau_n + O(\log n) \quad a.s.\ (n \to \infty),$$

where now $\{\log \tau_n\}$ forms a homogeneous Poisson point process on $(0,\infty)$ with unit intensity. This again proves (2) via (3), for $Z_n = \log U_n$.

5. The time change approach.

It is well-known that if the underlying distribution is doubly-exponential, then $\max(X_1,\ldots,X_n) - \log n$, $n \geq 1$, is also doubly-exponentially distributed (see e.g. Leadbetter et al. (1983)). The same holds true if the indices n are replaced by the random times U_n, i.e. $X_{U_n} - \log U_n$, $n \geq 1$, is also doubly-exponentially distributed (Pfeifer (1986)). Since doubly-exponential and exponential distributions are tail-equivalent, and any doubly-exponentially distributed sequence is $O(\log n)$ a.s., it follows that, in general,

(19) $-\log(1-F(X_{U_n})) - \log U_n$ is asymptotically doubly exponentially distributed

(20) $$-\log(1-F(X_{U_n})) - \log U_n = O(\log n) \quad \text{a.s.} \quad (n \to \infty).$$

Since by (12), we always have $\log U_n - \log \Delta_n = O(\log n)$ a.s. $(n \to \infty)$, relation (20) is an elegant tool for again proving (2) (or (14), resp.), for all three sequences simultaneously.

6. Generalizations and open problems.

Here we shall shortly discuss which of the above approaches are suitable for treating more general situations than the iid case.

6.1. The Markov case. Not much is known in the case where the underlying r.v.'s form a (homogeneous, say) MC; cf. Biondini and Siddiqui (1975). However, it was shown in Pfeifer (1984b) that in this case, the corresponding inter-record times are still conditionally independent given the σ-field of record values, such that the conditional independence approach 2. is potentially applicable. The main problem here lies in the fact that neither the general structure of the record value process is completely clear, nor is the (conditional) waiting time distribution between successive record values geometric in general. Research in this direction is in progress.

6.2. The non-homogeneous record model. Here one considers the case that after the occurence of a new record value, the underlying distribution is allowed to change, keeping however the independence assumption (Pfeifer (1982a, 1982b)). Here, too, the conditional independence of inter-record times given the record values is preserved, the (conditional) waiting time distributions between records being still (but possibly different) geometric distributions. Here the record value process is connected with general pure birth processes, such that strong approximations in the spirit of approach 2. become available. Some possibilities for this procedure are outlined in Pfeifer (1984c).

Unfortunately, the Markov property for record times goes lost in general for this model, such that approach 3. is not applicable.

6.3. Nevzorov's record model. Here one considers the case that the underlying distributions are of the form $F_n = F^{\alpha_n}$, where $\{\alpha_n\}$ is a sequence of positive real numbers with $\sum_{n=1}^{\infty} \alpha_n = \infty$ (see Nevzorov (1988) and further references therein). Letting $A(n) = \sum_{k=1}^{n} \alpha_k$, $n \geqq 1$ it can be proved that the corresponding record times again form a homogeneous MC with transition probabilities given by

$$(21) \qquad P(U_{n+1} > k \mid U_n = j) = \frac{A(j)}{A(k)}, \qquad 1 \leq j \leq k, \; n \geq 0.$$

Hence the Markov chain approach 3. is applicable, and shows that an analogue of Theorem 4 is valid for $\log A(U_n)$ if the conditions $\sum_{k=1}^{\infty} (\alpha_k/A(k))^2 < \infty$, $\sum_{k=1}^{\infty} (\alpha_k/A(k)) = \infty$ are met (see Pfeifer (1988) and Zhang (1988)). Zhang also proved that an embedding approach with non-homogeneous extremal processes works under these conditions, and that similarly the surplus number S of extremal jumps over the record times is a.s. finite with $E(S) \leq \sum_{k=1}^{\infty} (\alpha_k/A(k))^2$. (Similar embeddings have been considered earlier by Ballerini and Resnick (1985), however with a different emphasis.)
Hence for Nevzorov's model, relation (2) is valid for the sequence $Z_n = \log A(U_n)$.

Finally, we should like to mention that a time change approach similar to 5. also applies here, however for the specific situation that F is doubly-exponential. By imitation of the proof of Theorem 3 in Pfeifer (1986), it can be seen that here still $X_{U_n} - \log A(U_n)$ is doubly-exponentially distributed, hence again

$$(22) \qquad X_{U_n} = \log A(U_n) + O(\log n) \quad \text{a.s.} \quad (n \to \infty).$$

Note that the latter relation also holds without the above-mentioned regularity conditions. For instance, if $\alpha_n = e^{cn}$, $n \geqq 1$ for some $c > 0$, then $\log A(n) = nc - \log(1-e^{-c}) + O(e^{-nc})$ $(n \to \infty)$, hence we have here

$$(23) \qquad X_{U_n} = cU_n + O(\log n) \quad \text{a.s.} \quad (n \to \infty)$$

with

(24) $$U_n = \sum_{k=1}^{n} I_k + O(1) \quad \text{a.s.} \quad (n \to \infty)$$

where the I_k are iid with $P(I_k=0) = 1-P(I_k=1) = e^{-c}$.

This gives immediately rise to results like (2), with the proper normalizations, via (3), for U_n and X_{U_n}.

References

R. Ballerini and S.I. Resnick (1985): Records from improving populations. J. Appl. Prob. 22, 487 - 502.

R.M. Biondini and M.M. Siddiqui (1975): Record values in Markov sequences. In: Statistical Inference and Related Topics, Vol. 2, 291 - 352. Ac.Press.

K.N. Chandler (1952): The distribution and frequency of record values. J. Roy. Statist. Soc. Ser. B, 14, 220 - 228.

P. Deheuvels (1982): Strong approximation in extreme value theory and applications. Coll. Math. Soc. János Bolyai 36. In: Limit Theorems in Probability and Statistics, 369 - 404. North Holland.

________ (1983): The complete characterization of the upper and lower class of the record and inter-record times of an i.i.d. sequence. Z. Wahrscheinlichkeitsth. verw. Geb. 62, 1 - 6.

________ (1988): Strong approximations of k-th records and k-th record times by Wiener processes. Prob. Th. Rel. Fields 77, 195 - 209.

N. Glick (1978): Breaking records and breaking boards. Amer. Math. Monthly 85, 2 - 26.

J. Komlós, P. Major, and G. Tusnády (1976): An approximation of partial sums of independent rv's, and the sample df. II. Z. Wahrscheinlichkeitsth. verw. Geb. 34, 33 - 58.

M.R. Leadbetter, G. Lindgren, and H. Rootzén (1983): Extremes and Related Properties of Random Sequences and Processes. Springer.

M.F. Neuts (1967): Waiting times between record observations. J. Appl. Prob. 4, 206 - 208.

V.B. Nevzorov (1988): Records. Th. Prob. Appl. 32, 201 - 228.

D. Pfeifer (1982a): Characterizations of exponential distributions by independent non-stationary record increments. J. Appl. Prob. 19, 127 - 135. Corr., 906.

________ (1982b): The structure of elementary pure birth processes. J. Appl. Prob. 19, 664 - 667.

________ (1984a): A note on moments of certain record statistics. Z. Wahrscheinlich-keitsth. verw. Geb. 66, 293 - 296.

________ (1984b): A note on random time changes of Markov chains.

Scand. Act. J. , 127 - 129.

D. Pfeifer (1984c): Limit laws for inter-record times from non-homogeneous record values. J. Org. Behavior Stat. 1, 69 - 74.

_______ (1985): On the rate of convergence for some strong approximation theorems in extremal statistics. European Meeting of Statisticians, Marburg 1984. In: Stat. & Decisions Supp. Iss. No. 2, 99 - 103.

_______ (1986): Extremal processes, record times and strong approximation. Pub. Inst. Stat. Univ. Paris XXXI, 47 - 65.

_______ (1987): On a joint strong approximation theorem for record and inter-record times. Th. Prob. Rel. Fields 75, 212 - 221.

_______ (1988): Extremal processes, secretary problems and the 1/e-law. To appear in: J. Appl. Prob.

A. Rényi (1962): Théorie des éléments saillants d'une suite d'observations. Coll. Comb. Meth. in Prob. Th., August 1 - 10, 1962, Math. Inst., Aarhus Univ., Denmark, 104 - 117.

S.I. Resnick (1973): Extremal processes and record value times. J. Appl. Prob. 10, 863 - 868.

__________ (1974): Inverses of extremal processes. Adv. Appl. Prob. 6, 392 - 406.

__________ (1975): Weak convergence to extremal processes. Ann. Prob. 3, 951 - 960.

S.I. Resnick and M. Rubinovitch (1973): The structure of extremal processes. Adv. Appl. Prob. 5, 287 - 307.

R.W. Shorrock (1972a): On record values and record times. J. Appl. Prob. 9, 316 - 326.

_______ (1972b): A limit theorem for inter-record times. J. Appl. Prob. 9, 219 - 223.

W.E. Strawderman and P.T. Holmes (1970): On the law of the iterated logarithm for inter-record times. J. Appl. Prob. 7, 432 - 439.

M. Westcott (1977): A note on record times. J. Appl. Prob. 14, 637 - 639.

D. Williams (1973): On Rényi's record problem and Engel's series. Bull. London Math. Soc. 5, 235 - 237.

Y.S. Zhang (1988): Strong approximations in extremal statistics (in German). Ph.D. Thesis, Technical University Aachen.