

ON INDEPENDENCE OF k -RECORD PROCESSES: IGNATOV'S THEOREM REVISITED

BY YI-CHING YAO

Academia Sinica and Colorado State University

For an infinite sequence of independent and identically distributed (i.i.d.) random variables, the k -record process consists of those terms that are the k th largest at their appearance. Ignatov's theorem states that the k -record processes, $k = 1, 2, \dots$, are i.i.d. A new proof is given which is based on a "continualization" argument. An advantage of this fairly simple approach is that Ignatov's theorem can be stated in a more general form by allowing for different tiebreaking rules. In particular, three tiebreakers are considered and shown to be related to Bernoulli, geometric and Poisson distributions.

1. Introduction. For an infinite sequence (X_1, X_2, \dots) of independent and identically distributed (i.i.d.) random variables (with common distribution F), define the *initial rank* of the n th term X_n as

$$(1) \quad r'_n = \sum_{i=1}^n I\{X_i \geq X_n\},$$

where $I(A)$ denotes the indicator of event A . For each $k = 1, 2, \dots$, denote by $\mathbf{Y}_k = (Y_{k,1}, Y_{k,2}, \dots)$ the collection of those X_n with $r'_n = k$, which is called the k -record process for the sequence (X_1, X_2, \dots) . Thus, $Y_{k,j}$ is the j th X_n with initial rank equaling k , and $Y_{k,1} < Y_{k,2} < \dots$. Collectively, these processes are referred to as partial record processes. Ignatov's theorem states that $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are i.i.d.

In the literature, several authors have taken different approaches to proving this surprising result. Ignatov (1977), Deheuvels (1983) and Stam (1985) considered only the case of continuous F and showed that $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are i.i.d., each being a Poisson process. Goldie and Rogers (1984) used an interesting martingale argument to establish the general case. Engelen, Tommassen and Vervaat (1988) first proved the discrete case directly, and then passed to the general case by a discretization device. [See also Resnick (1987), Section 4.6, and Rogers (1989) for related discussions.] Samuels' (1992) proof goes "backwards" by constructing the X_n recursively from the k -record processes. While none of the proofs is very long, the "discretization" and "backward" approaches provide good insight into why the k -record processes are i.i.d. Furthermore, although most of the proofs assume the

Received October 1996.

AMS 1991 subject classification. Primary 60G55.

Key words and phrases. Record values, Poisson process.

condition $P(F(X_1) = 1) = 0$ to avoid the possibility that each k -record process has only finitely many terms, Samuels (1992) observed that Ignatov's theorem still holds with $P(F(X_1) = 1) > 0$.

In this article, we present another simple proof based on a “continualization” argument, which is in direct contrast to that of Engelen, Tommassen and Vervaat (1988). More precisely, by expanding each discrete point (discontinuity of F) into an interval, the k -record processes are shown to form a Poisson random measure on the expanded space, from which Ignatov's theorem follows easily. [The “continualization” idea was also mentioned in the final remark of Samuels (1992) concerning the case $P(F(X_1) = 1) > 0$.] An advantage of this fairly simple approach is that Ignatov's theorem can be stated in a more general form by allowing for different tiebreaking rules. In particular, we consider three tiebreakers resulting in Bernoulli, geometric and Poisson distributions at each discrete point.

2. Ignatov's theorem with general tiebreaking rules. For a pair of ties, the tiebreaking rule associated with (1) treats the “later” observation as “smaller”. We will consider a general definition of initial rank r_n of X_n which satisfies the following two requirements:

$$(2) \quad r_n \text{ depends on } X_1, \dots, X_n \text{ only through } \alpha_n \text{ and } \beta_n; \text{ and}$$

$$(3) \quad \alpha_n + 1 \leq r_n \leq \alpha_n + \beta_n;$$

where

$$\alpha_n := \#\{i: X_i > X_n, 1 \leq i \leq n\} \quad \text{and} \quad \beta_n := \#\{i: X_i = X_n, 1 \leq i \leq n\}.$$

Indeed, r_n is allowed to be a random variable whose (conditional) distribution, given X_1, \dots, X_n , depends only on α_n and β_n . In addition to (1), which is equivalent to $r'_n = \alpha_n + \beta_n$, we will also discuss two other simple definitions r''_n and r'''_n of initial rank, given by

$$(4) \quad r''_n = \alpha_n + 1,$$

and

$$(5) \quad P(r'''_n = \alpha_n + i \mid X_1, \dots, X_n) = 1/\beta_n \quad \text{for } 1 \leq i \leq \beta_n.$$

For a pair of ties, the tiebreaking rule associated with r''_n treats the “later” observation as “larger” while that associated with r'''_n is a randomized rule.

It is instructive to think of the processes $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ collectively as a random (point) measure Π on the space $R \times N$; that is, $\mathbf{Y}_k = (Y_{k,1}, Y_{k,2}, \dots)$ is identified with $\tilde{\mathbf{Y}}_k = ((Y_{k,1}, k), (Y_{k,2}, k), \dots)$ (a point process on $R \times \{k\}$), and Π is induced by $\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2, \dots$. We will also identify the restricted random measure $\Pi|_{R \times \{k\}}$ with \mathbf{Y}_k and $\tilde{\mathbf{Y}}_k$. Denote by $D = \{\delta_1, \delta_2, \dots\}$ the set of all discontinuities of F . Let $D^c := R \setminus D$, the complement of D .

THEOREM 1. *For a general definition of initial rank r_n satisfying (2) and (3), the partial record processes $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ (for an i.i.d. sequence (X_1, X_2, \dots))*

with common distribution F) induce a random measure Π on $R \times N$ with the following properties:

- (i) $\Pi|_{D^c \times N}, \Pi|_{\{\delta_i\} \times N}, i = 1, 2, \dots$ are independent;
- (ii) $\Pi|_{D^c \times N}$ is a Poisson process with mean measure given by $m(dx \times \{k\}) = F(dx)/(1 - F(x))$ for $x \in D^c$ and $k \in N$.

We first prove two lemmas.

LEMMA 1. For continuous F , Π is a Poisson process on $R \times N$ with mean measure given by $m(dx \times \{k\}) = F(dx)/(1 - F(x))$, for $x \in R$ and $k \in N$.

PROOF. Using the memoryless property of the exponential distribution, Deheuvels (1983) gave a very short proof that if F is exponential with mean 1, then $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are i.i.d. Poisson processes with rate 1; that is, Π is a Poisson process on $R \times N$ with mean measure given by $m(dx \times \{k\}) = I\{x > 0\} dx$. For a general continuous distribution F , using the monotone transformation $g(x) = -\log(1 - F(x))$, we have $g(X_1), g(X_2), \dots$ i.i.d. exponential with common mean 1. The lemma follows easily. \square

LEMMA 2. For an arbitrary sequence of real numbers (x_1, x_2, \dots) , let $\mathbf{y}_k = (y_{k,1}, y_{k,2}, \dots)$ be the k -record sequence (according to r'_n). For given $-\infty \leq a < b \leq \infty$, let $\mathbf{y}_k|_{[a,b]}$ be a subsequence of \mathbf{y}_k consisting of those $y_{k,j}$ with $a \leq y_{k,j} < b$. Then from $\mathbf{y}_k|_{[a,b]}, k = 1, 2, \dots$, one can construct the sequence (x'_1, x'_2, \dots) which is derived from (x_1, x_2, \dots) with terms less than a removed and terms greater than or equal to b replaced by b ; that is, $x'_i = \min\{x_i^*, b\}$ where x_i^* is the i th $x_n \geq a$.

PROOF. We first consider the case $a = -\infty$ and $b = \infty$, so that $\mathbf{y}_k = \mathbf{y}_k|_{[a,b]}$ and $(x_1, x_2, \dots) = (x'_1, x'_2, \dots)$. To construct the x_n recursively from the \mathbf{y}_k , note that $x_1 = y_{1,1}$. The next term x_2 must be one of the first available terms in $\mathbf{y}_k, k = 1, 2$; that is, x_2 must be either $y_{2,1}$, or $y_{1,2}$. If $y_{2,1} \leq x_1$, then $x_2 = y_{2,1}$; otherwise $x_2 = y_{1,2}$. (If $y_{2,1} \leq x_1$ and if $x_2 = y_{1,2}$, then $y_{2,1}$ must be some x_m with $m > 2$ whose initial rank would be at least 3, a contradiction.) In general, suppose x_1, \dots, x_n have been identified (constructed). Then the next term x_{n+1} must be one of the first available terms in $\mathbf{y}_k, k = 1, \dots, n + 1$. Denote these first available terms by $z_k, k = 1, \dots, n + 1$. We say that z_k is a valid candidate for x_{n+1} if $\sum_{i=1}^n I\{x_i \geq z_k\} + 1 = k$. Then x_{n+1} must be that valid candidate with the largest index. (If both z_k and $z_{k'}$ are valid candidates with $k < k'$ and if $x_{n+1} = z_k$, then $z_{k'}$ must be some x_m with $m > n + 1$ whose initial rank would be at least $k' + 1$ since $z_k > z_{k'}$.) This proves the lemma for the case $a = -\infty$ and $b = \infty$.

To deal with the general case, attach a "b" to $\mathbf{y}_k|_{[a,b]}$ (if it is a finite sequence), resulting in a new sequence (denoted \mathbf{y}'_k). Set $\mathbf{y}'_k := \mathbf{y}_k|_{[a,b]}$ if it is an infinite sequence. Clearly, the k -record sequence for (x'_1, x'_2, \dots) is either \mathbf{y}'_k or $\mathbf{y}_k|_{[a,b]}$. It is not difficult to show that when applied to $\mathbf{y}'_k, k = 1, 2, \dots$, the above algorithm yields the sequence (x'_1, x'_2, \dots) , since (i) at the time

when the “ b ” in \mathbf{y}'_1 (if \mathbf{y}'_1 is a finite sequence) becomes a valid candidate for x'_n , say, we have $x'_n = b$ only if no other valid candidate is available, and (ii) the “ b ” in \mathbf{y}'_k (if \mathbf{y}'_k is a finite sequence) cannot be selected unless the “ b ” in each \mathbf{y}'_l , $l = 1, \dots, k - 1$ has already been selected. The proof is complete. \square

Note. In the above proof, the algorithm for constructing a sequence (x_1, x_2, \dots) from its k -record sequences \mathbf{y}_k is due to Samuels (1992), who used this algorithm to show that an i.i.d. sequence can be constructed from its partial record processes. His description of the algorithm is slightly different (but equivalent): among all the first available terms z_k in \mathbf{y}_k , $k = 1, \dots, n + 1$ when x_{n+1} is to be determined, choose z_k for x_{n+1} with

$$k = \max \left\{ k' : \sum_{i=1}^n I\{x_i \geq z_{k'}\} + 1 \geq k' \right\}.$$

Our description used the fact that the \mathbf{y}_k are assumed to be the k -record sequences for some unknown (unobserved) sequence (x_1, x_2, \dots) .

PROOF OF THEOREM 1. Consider an i.i.d. sequence (U_1, U_2, \dots) of uniform(0, 1) random variables, which is independent of (X_1, X_2, \dots) . Define

$$\begin{aligned} X_n^* &:= (X_n, 0) && \text{if } X_n \in D^c \\ &:= (X_n, U_n) && \text{if } X_n \in D. \end{aligned}$$

Thus, the X_n^* take values in the space $S := D^c \times \{0\} \cup D \times [0, 1]$ with common continuous distribution G given by

$$\begin{aligned} G(dx \times \{0\}) &= F(dx) && \text{for } x \in D^c, \\ G(\{\delta_i\} \times du) &= F\{\delta_i\} du && \text{for } 0 < u < 1. \end{aligned}$$

The ordering $<$ on R induces an ordering $<^*$ on S : for $a = (a_1, a_2) \in S$ and $b = (b_1, b_2) \in S$, we write $a <^* b$ if either $a_1 < b_1$ or $a_1 = b_1$ and $a_2 < b_2$. Define $r_n^* := \sum_{i=1}^n I\{X_n^* \leq^* X_i^*\}$, and let \mathbf{Y}_k^* be the k -record process for (X_1^*, X_2^*, \dots) (according to r_n^*). Let Π^* be the random measure on $S \times N$ induced by $\mathbf{Y}_1^*, \mathbf{Y}_2^*, \dots$. Since G is continuous (and S can be suitably embedded into R with the orderings $<^*$ and $<$ preserved), it follows from Lemma 1 that Π^* is a Poisson process with mean measure given by

$$\begin{aligned} (6) \quad m^*((dx \times \{0\}) \times \{k\}) &= \frac{F(dx)}{1 - F(x)}, && x \in D^c; \\ (7) \quad m^*((\{\delta_i\} \times du) \times \{k\}) &= \frac{G(\{\delta_i\} \times du)}{P((\delta_i, u) <^* X_1^*)}, && 0 < u < 1 \\ &= \frac{F\{\delta_i\} du}{1 - F(\delta_i) + (1 - u)F\{\delta_i\}}. \end{aligned}$$

Since $(D^c \times \{0\}) \times N$, $(\{\delta_i\} \times [0, 1]) \times N$, $i = 1, 2, \dots$ are disjoint, it follows that $\Pi^*|_{(D^c \times \{0\}) \times N}$, $\Pi^*|_{(\{\delta_i\} \times [0, 1]) \times N}$, $i = 1, 2, \dots$ are independent. Note that

with probability 1, no two X_n can take the same value in D^c , implying that $\beta_n = 1$ if $X_n \in D^c$. So, if $X_n \in D^c$, then its initial rank is not affected by the continualization device; that is,

$$r_n = r_n^* = \sum_{i=1}^n I\{X_i > X_n\} + 1 = \alpha_n + 1 \quad \text{if } X_n \in D^c.$$

Hence the two random measures $\Pi|_{D^c \times N}$ and $\Pi^*|_{(D^c \times \{0\}) \times N}$ are the same if we identify the space $D^c \times N$ with $(D^c \times \{0\}) \times N$. It follows that $\Pi|_{D^c \times N}$ is a Poisson process with mean measure given by $m(dx \times \{k\}) = F(dx)/(1 - F(x))$ for $x \in D^c$ and $k \in N$.

The proof will be complete if we can show that for fixed i , and for each $X_n = \delta_i$, one can determine α_n and β_n from $\Pi^*|_{((\delta_i) \times [0, 1]) \times N}$. Note that $\Pi^*|_{((\delta_i) \times [0, 1]) \times N}$ is identified with the collection of \mathbf{Y}'_k , $k = 1, 2, \dots$ where $\mathbf{Y}'_k := \mathbf{Y}_k^*|_{((\delta_i, 0), (\delta_i, 1))}$ is the subsequence of \mathbf{Y}_k^* with terms $<^*(\delta_i, 0)$ and terms $\geq^*(\delta_i, 1)$ all removed. [Here we have assumed that all U_n are strictly less than 1 so that no X_n^* equals $(\delta_i, 1)$.] While it considers only the definition r'_n , Lemma 2 applies here in the absence of ties (since G is continuous). Thus, from \mathbf{Y}'_k , $k = 1, 2, \dots$, one can construct the sequence (X'_1, X'_2, \dots) , which is derived from (X_1^*, X_2^*, \dots) with terms $<^*(\delta_i, 0)$ removed and terms $\geq^*(\delta_i, 1)$ replaced by $(\delta_i, 1)$. In this sequence, if the m th term is (δ_i, u) with $0 \leq u < 1$, then it corresponds to an X_n equaling δ_i for which α_n is the number of times $(\delta_i, 1)$ appears before the m th term and $\beta_n = m - \alpha_n$. The proof is complete. \square

For any initial rank r_n satisfying (2) and (3), Theorem 1 enables one to separate the continuous part and individual discrete points from one another. The remaining task is to find the distribution of $\Pi|_{\{\delta_i\} \times N}$ for each $\delta_i \in D$. Let $\nu_k(\delta_i) := \Pi\{(\delta_i, k)\} (= \#\{j: Y_{k,j} = \delta_i\})$. Then $\Pi|_{\{\delta_i\} \times N}$ can be identified with $(\nu_1(\delta_i), \nu_2(\delta_i), \dots)$. To determine the joint distribution of $\nu_1(\delta_i), \nu_2(\delta_i), \dots$, we need to pay attention only to the initial ranks of those X_n equal to δ_i . So it suffices to consider a 0–1 sequence (B_1, B_2, \dots) which is derived from (X_1, X_2, \dots) with terms $< \delta_i$ removed, terms $= \delta_i$ replaced by 0, and terms greater than δ_i replaced by 1. Clearly, the B_k are i.i.d. with $P(B_k = 0) = 1 - P(B_k = 1) = p_i$ where $p_i := F\{\delta_i\}/(1 - F(\delta_i -))$.

THEOREM 2. (i) *If $r_n = r'_n$ for all n , then $\nu_k(\delta_i)$, $k = 1, 2, \dots$ are i.i.d. (Bernoulli) with*

$$P(\nu_k(\delta_i) = 1) = 1 - P(\nu_k(\delta_i) = 0) = p_i.$$

(ii) *If $r_n = r''_n$ for all n and $F(\delta_i) < 1$, then $\nu_k(\delta_i)$, $k = 1, 2, \dots$ are i.i.d. (geometric) with*

$$P(\nu_k(\delta_i) = m) = p_i^m(1 - p_i), m = 0, 1, \dots$$

(iii) *If $r_n = r''_n$ for all n and $F(\delta_i) = 1$, then with probability 1, $\nu_1(\delta_i) = \infty$, and $\nu_k(\delta_i) = 0$, $k = 2, 3, \dots$*

(iv) If $r_n = r_n'''$ for all n , then $\nu_k(\delta_i)$, $k = 1, 2, \dots$ are i.i.d. (Poisson) with common mean equal to $-\log(1 - p_i)$ where $-\log 0 := \infty$ and a Poisson random variable with infinite mean is defined to be infinite with probability 1.

PROOF. If $r_n = r_n'$ for all n , then $\nu_k(\delta_i)$ equals 0 or 1 according as B_k equals 1 or 0. This proves (i). If $r_n = r_n''$ for all n , then $\nu_k(\delta_i)$ equals the number of 0's between the $(k - 1)$ th and k th 1 in the sequence (B_1, B_2, \dots) . If $F(\delta_i) < 1$, then $p_i < 1$ and clearly $\nu_k(\delta_i)$, $k = 1, 2, \dots$ are i.i.d. geometric, proving (ii). If $F(\delta_i) = 1$, then $p_i = 1$ so that the B_k are all 0 (with probability 1). Hence $\nu_1(\delta_i) = \infty$ and $\nu_k(\delta_i) = 0$, $k = 2, 3, \dots$, proving (iii). The case of $r_n = r_n'''$ for all n is equivalent to assigning an independent uniform(0, 1) random variable to each $X_n = \delta_i$ and breaking the ties by comparing the values of the assigned variables. But this is exactly what the continualization device did in the proof of Theorem 1. Therefore, $\nu_k(\delta_i)$ may be identified with $\Pi^*({\{\delta_i\}} \times [0, 1] \times \{k\})$, $k = 1, 2, \dots$, which are i.i.d. Poisson with common mean equal to [see (7)]

$$\int_0^1 m^*({\{\delta_i\}} \times du) \times \{k\} = \int_0^1 \frac{F\{\delta_i\} du}{1 - F(\delta_i) + (1 - u)F\{\delta_i\}} = -\log(1 - p_i).$$

This proves (iv). \square

By Theorems 1 and 2, we have the following corollary.

COROLLARY 1. Let \mathbf{Y}_k be the k -record process for the sequence (X_1, X_2, \dots) according to some definition of initial rank r_n .

(i) If either $r_n = r_n'$ for all n or $r_n = r_n'''$ for all n , then $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are i.i.d.
(ii) Assume that $r_n = r_n''$ for all n . If $P(F(X_1) = 1) = 0$, then $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are i.i.d. If $P(F(X_1) = 1) > 0$, then $\mathbf{Y}_1|_{(-\infty, c)}$, $\mathbf{Y}_2, \mathbf{Y}_3, \dots$ are i.i.d. where c is determined by $F(c -) < 1 = F(c)$ and $\mathbf{Y}_1|_{(-\infty, c)}$ consists of those terms in \mathbf{Y}_1 that are less than c .

Part (ii) of Corollary 1 can be restated as follows:

If $r_n = r_n''$ for all n , then $\mathbf{Y}_1|_{(-\infty, \sup \mathbf{Y}_1)}$, $\mathbf{Y}_2, \mathbf{Y}_3, \dots$ are i.i.d. where $\sup \mathbf{Y}_1 := \sup_j Y_{1,j}$.

3. Some remarks.

REMARK 1. Let (X_1, X_2, \dots) be an exchangeable sequence and let \mathbf{Y}_k be the k -record process according to some definition of initial rank r_n . By de Finetti's theorem and Corollary 1, if either $r_n = r_n'$ for all n or $r_n = r_n'''$ for all n , then $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ are exchangeable. If $r_n = r_n''$ for all n , then $\mathbf{Y}_1|_{(-\infty, \sup \mathbf{Y}_1)}$, $\mathbf{Y}_2, \mathbf{Y}_3, \dots$ are exchangeable.

REMARK 2. For an i.i.d. sequence (X_1, X_2, \dots) with common distribution F , the 1-record process \mathbf{Y}_1 (according to initial rank r_n') is said to have the

record distribution associated with F . Samuels' (1992) proof of Ignatov's theorem implies that if \mathbf{Y}_k , $k = 1, 2, \dots$ are i.i.d. processes each having the record distribution associated with some F , then there exists an i.i.d. sequence (X_1, X_2, \dots) with common distribution F such that \mathbf{Y}_k is the k -record process for (X_1, X_2, \dots) , $k = 1, 2, \dots$. It would be of interest to characterize the class of sequences of random variables for which the k -record processes are i.i.d. The following simple example shows that there exist non-i.i.d. sequences for which the k -record processes are i.i.d.

Consider an i.i.d. sequence (Z_1, Z_2, \dots) with common continuous distribution G satisfying $G(0) = 0$ and $G(1) = 1$. Let $Y_{k,j} := Z_k + j - 1$, for $k, j = 1, 2, \dots$ and let $\mathbf{Y}_k := (Y_{k,1}, Y_{k,2}, \dots)$, $k = 1, 2, \dots$. Clearly, the \mathbf{Y}_k are i.i.d. We claim that there exists a sequence (X_1, X_2, \dots) for which \mathbf{Y}_k is the k -record process, $k = 1, 2, \dots$. [The sequence (X_1, X_2, \dots) is necessarily non-i.i.d.] By Samuels' algorithm applied to the \mathbf{Y}_k , all we have to show is that all the $Y_{k,j}$ get used (with probability 1). The details are omitted.

REFERENCES

- DEHEUVELS, P. (1983). The strong approximation of extremal processes II. *Z. Wahrsch. Verw. Gebiete* **62** 7–15.
- ENGELN, R., TOMMASSEN, P. and VERVAAT, W. (1988). Ignatov's theorem: a new and short proof. *J. Appl. Probab.* **25A** 229–236.
- GOLDIE, C. M. and ROGERS, L. C. G. (1984). The k -record processes are i.i.d. *Z. Wahrsch. Verw. Gebiete* **67** 197–211.
- IGNATOV, Z. (1977). Ein von der Variationsreihe erzeugter Poissonscher Punktprozess. *Annuaire Univ. Sofia Fac. Math. Méch.* **71** 79–94 (published 1986).
- RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.
- ROGERS, L. C. G. (1989). Ignatov's theorem: an abbreviation of the proof of Engelen, Tommassen and Vervaat. *Adv. in Appl. Prob.* **21** 933–934.
- SAMUELS, S. M. (1992). An all-at-once proof of Ignatov's theorem. *Contemp. Math.* **125** 231–237.
- STAM, A. J. (1985). Independent Poisson processes generated by record values and interrecord times. *Stochastic Process Appl.* **19** 315–325.

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI
TAIWAN