

## THE NOISY SECRETARY PROBLEM AND SOME RESULTS ON EXTREME CONCOMITANT VARIABLES

ABBA M. KRIEGER,\* *University of Pennsylvania*

ESTER SAMUEL-CAHN,\*\* *The Hebrew University of Jerusalem*

### Abstract

The classical secretary problem for selecting the best item is studied when the actual values of the items are observed with noise. One of the main appeals of the secretary problem is that the optimal strategy is able to find the best observation with a nontrivial probability of about 0.37, even when the number of observations is arbitrarily large. The results are strikingly different when the qualities of the secretaries are observed with noise. If there is no noise then the only information that is needed is whether an observation is the best among those already observed. Since the observations are assumed to be independent and identically distributed, the solution to this problem is distribution free. In the case of noisy data, the results are no longer distribution free. Furthermore, we need to know the rank of the noisy observation among those already observed. Finally, the probability of finding the best secretary often goes to 0 as the number of observations,  $n$ , goes to  $\infty$ . The results heavily depend on the behavior of  $p_n$ , the probability that the observation that is best among the noisy observations is also best among the noiseless observations. Results involving optimal strategies if all that is available is noisy data are described and examples are given to elucidate the results.

*Keywords:* Optimal stopping rule; best-choice secretary problem; noisy data

2010 Mathematics Subject Classification: Primary 62L15

Secondary 60F99

### 1. Introduction

The classical ‘best-choice secretary problem’ is surprising in that there is a rule which enables finding the best secretary with nonzero probability even if the number of secretaries that are considered is arbitrarily large. But, what happens if the qualities of the secretaries at the time of decision are known only subject to noise? In this paper we prove various aspects of this problem. First, the optimal rule when there is no noise is no longer optimal when measurements are made with noise. Second, in many cases the probability of finding the best secretary goes to 0, albeit slowly, as the number of secretaries,  $n$ , increases. Third, the results are sensitive to distributional assumptions, unlike the classical secretary problem, and there are distributions for which the probability of finding the best goes to the same limit as in the noiseless case, and other distributions where the probability goes to 0 as  $n$  goes to  $\infty$ .

---

Received 6 October 2011; revision received 2 February 2012.

\* Postal address: Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.  
Email address: krieger@wharton.upenn.edu

Supported by funds from the Marcy Bogen Chair of Statistics at the Hebrew University of Jerusalem.

\*\* Postal address: Department of Statistics and Center for Rationality, The Hebrew University of Jerusalem, Jerusalem, 91905, Israel. Email address: schan@mssc.huji.ac.il

Supported by the Israel Science Foundation, grant no. 467/04.

In the classical best-choice secretary problem the underlying assumption is that ranks are sequentially obtained from  $n$  independent and identically distributed (i.i.d.) continuous random variables,  $X_i, i = 1, \dots, n$ . This total number,  $n$ , is called the horizon and is assumed known. Only the relative ranks,

$$RR(X_i) := \sum_{j=1}^i \mathbf{1}_{\{X_i \leq X_j\}},$$

are observed. The goal is to maximize the probability of picking the  $X_i$  that are maximal, i.e. the  $i$  for which the absolute rank

$$AR(X_i) := \sum_{j=1}^n \mathbf{1}_{\{X_i \leq X_j\}}$$

equals 1. The well-known optimal solution is to let a certain number,  $N(n)$ , go by, and pick the first item thereafter,  $i > N(n)$ , for which  $RR(X_i) = 1$ . If no such  $i$  exists, stop at  $n$  anyway. It is well known that  $N(n)/n$  tends to  $e^{-1}$  as  $n \rightarrow \infty$ , and that the optimal probability,  $W_n$ , of picking the best tends to  $e^{-1}$  as  $n \rightarrow \infty$ . See, e.g. Gilbert and Mosteller (1966) or Samuels (1991) with generalizations in Bruss (2000) and Gneden (2007). When applying this rule, it suffices to know whether the present item is relatively best, i.e. if  $RB(X_i) = 1$ , where  $RB(X_i) = 1$  if  $RR(X_i) = 1$  and  $RB(X_i) = 0$  otherwise. Clearly, the solution to this classical problem is distribution free.

In the present paper we consider the case where the relative ranks are not those of the  $X_i$ s themselves, but of i.i.d.  $Y_i$ s, where  $Y_i = X_i + \varepsilon_i$ , and the  $\varepsilon_i$  are i.i.d. noise (or error) variables, independent of the  $X_i$ s. The goal is the same as before, that is, to maximize the probability of selecting the  $i$  for which the  $X$  value is maximal. The optimal rule, and the optimal probability of picking the best  $X_i$  are no longer distribution free. Denote the optimal probability when the  $RR(Y_i)$  are known by  $W_n(X, \varepsilon)$ , where  $n$  is the known horizon, and the optimal probability of selecting the best when only the  $RB(Y_i)$  are known by  $W_n^*(X, \varepsilon)$ .

If we use the classical rule on the noisy data then, clearly, there is a probability that goes to  $e^{-1}$  of finding the best  $Y_i$  as  $n \rightarrow \infty$ . However, if we find the best  $Y$ , have we found the best  $X$ , which is what is desired? The difference between the classical secretary problem (i.e. without noise) and the noisy secretary problem heavily depends on the value of  $p_n$ , where

$$p_n = p_n(X, \varepsilon) = P\left(\arg \max_{i \leq n} X_i = \arg \max_{i \leq n} Y_i\right), \tag{1.1}$$

i.e.  $p_n$  is the probability that the location of the maximal  $X$  is the same as the location of the maximal  $Y$ . The behavior of  $p_n$ , as mentioned above, is crucial to the values of  $W_n(X, \varepsilon)$  and  $W_n^*(X, \varepsilon)$ .

Our main results concerning the identification of the best  $X$  from noisy data are the following.

- (S1) For any  $X$  and  $\varepsilon$ , if the observed values are the  $RB(Y_i)$  only,  $W_n^*(X, \varepsilon) \geq W_n p_n$ .
- (S2) If only the RBs are observed, the optimal value  $N(n, X, \varepsilon)$  after which one should pick the first item for which  $RB(Y_i) = 1$  satisfies  $N(n, X, \varepsilon) \leq N(n)$ , i.e. one should stop earlier than in the classical case.
- (S3) If the  $RR(Y_i)$  are sequentially available, it is no longer optimal to base the stopping rule on the  $RB(Y_i)$ s only.

Some of the results depend on the probability that the best  $X$  in  $n$  items is the  $m$ th best  $Y$ . To this end, for any  $i$ , let

$$\begin{aligned}
 p_{nm} &= P(\text{AR}(Y_i) = m \mid \text{AR}(X_i) = 1) \\
 &= P(\text{AR}(X_i) = 1 \mid \text{AR}(Y_i) = m) \\
 &= \sum_{i=1}^n P([\text{AR}(X_i) = 1] \cap [\text{AR}(Y_i) = m]).
 \end{aligned}
 \tag{1.2}$$

- (S4) The optimal rule, which can in principle be found by backward induction once  $p_{nm}$  is determined, is of the following form. There exist integer values  $1 \leq k_1 \leq \dots \leq k_n = n$ , not necessarily distinct, such that one should stop with the smallest  $i$  such that  $i < k_j$  and  $\text{RR}(Y_i) < j$ .
- (S5)  $\lim_{n \rightarrow \infty} W_n(X, \varepsilon) = 0$  if and only if  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ .

Results (S1)–(S5) are proved in Section 4.

In the next section we consider the important example where  $X$  and  $\varepsilon$  are normally distributed. This example elucidates (S1)–(S4) above and (R1) below. As is apparent from the above list of results and will be even more apparent from the example, the  $p_n(X, \varepsilon)$  values play an important role. These values are also of intrinsic interest. Hence, we discuss  $p_n(X, \varepsilon)$  in Section 3. Denote the distribution of  $X$  by  $F$  with density  $f$  and the distribution of  $\varepsilon$  by  $G$  with density  $g$ . We show that the following statements hold.

- (R1) Suppose that  $\sup\{x : F(x) < 1\} = \infty$ , and that the  $\lim_{x \rightarrow \infty} f(x + d)/f(x)$  exists for all  $d > 0$ . Then a necessary and sufficient condition for  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$  for all distributions  $G$  is that  $\lim_{x \rightarrow \infty} f(x + d)/f(x) = 0$  for every fixed  $d$ .
- (R2) If  $\sup\{x : F(x) < 1\} = c < \infty$  then  $\lim_{n \rightarrow \infty} p_n = 0$  for all  $G$ .
- (R3) For any given  $F$ , there exists a distribution  $G$  such that  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ .
- (R4) There exist distributions  $F$  and  $G$  such that  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 1$ .
- (R5) The  $p_n(X, \varepsilon)$  values are *not* necessarily monotone in  $n$ .

Additional examples are given in Section 5. Because noise (or error) is often assumed to be normal, special attention is given to the case where  $G$  is normal. The examples include cases with exponential and Pareto with parameter 1 distributions, and the case where both  $F$  and  $G$  are normal.

### 2. Normal–normal example

In order to illustrate the results, we consider the case where  $X_i \sim \mathcal{N}(0, \rho^2)$  and  $Y_i = X_i + \varepsilon_i$ , where  $\varepsilon_i \sim \mathcal{N}(0, 1 - \rho^2)$  and all the  $X_i$  and  $\varepsilon_i$  are independent. Hence, the concomitant variable  $Y_i$  (cf. David and Nagaraja (2003)) is  $\mathcal{N}(0, 1)$  and the correlation between  $X$  and  $Y$  is  $\rho$ .

There are two kinds of result mentioned in the introduction that we need to consider. First we consider the behavior of  $p_n$ , the probability that the index for which  $Y_i$  is maximum agrees with the index for which  $X_i$  is maximum. Since a normal distribution satisfies the condition that  $\lim_{x \rightarrow \infty} f(x + d)/f(x) = 0$  for all  $d$ ,  $p_n$  goes to 0 (see (R1)). The results in Table 1, calculated by simulation with 10 000 replications, show how  $p_n$  varies as a function of  $n$  and  $\rho$ .

TABLE 1: The probability that the observation with the largest  $X$  value also yields the largest  $Y$  value.

$\varphi$	$n$				
	10	50	100	1000	10000
0.5	0.291	0.138	0.107	0.044	0.018
0.6	0.353	0.191	0.155	0.076	0.039
0.7	0.423	0.261	0.222	0.125	0.073
0.8	0.519	0.360	0.324	0.210	0.151
0.9	0.647	0.512	0.481	0.370	0.303

TABLE 2: The probability that the index for which the best  $X$  value is attained yields a  $Y$  value that is  $m$ th absolute best. Here  $n = 10000$  and  $\rho = 0.9$ .

$m$	1	2	3	4	5	6	7	8	9	10
$p_{nm}$	0.303	0.137	0.083	0.065	0.050	0.036	0.032	0.028	0.022	0.018

It is interesting to note how sensitive  $p_n$  is to  $\rho$ . In fact, in Ledford and Tawn (1998) it was shown that

$$\lim_{n \rightarrow \infty} D_n p_n = \left( \Gamma \left( \frac{1}{1 + \rho} \right) \right)^2, \tag{2.1}$$

where  $D_n = (1 - \rho^2)^{1/2} \{ (4\pi \log n)^\rho n^{1-\rho} \}^{1/(1+\rho)}$ .

Second we illustrate by simulation what occurs with a secretary-like decision rule, which depends on  $p_{nm}$  (see (1.2)), the probability that the  $m$ th largest  $Y$  corresponds to the observation with the largest  $X$  value. In Table 2 we present simulation results for  $p_{nm}$  using 10000 replications. If we employed the secretary rule, with  $N(n) = n/e$ ,  $n = 10000$ , and  $\rho = 0.9$ , we would find the largest  $X$ , observing only  $RB(Y_i)$ , with a probability of 0.138, as compared to  $W_n$  of approximately  $e^{-1} = 0.368$ , the value if there was no noise (or, equivalently,  $\rho = 1$ ). The probability of 0.138 is greater than  $W_n p_n = 0.303e^{-1} = 0.112$ . The reason is that the secretary rule, when it stops at the relative best  $Y$ , might be stopping at an observation that is, say, the second best  $Y$  in absolute rank. The second best absolute rank of  $Y$  has a probability of 0.137 of being the observation with best  $X$ . This would add to the probability that the rule chooses a  $Y$  with the best  $X$ ; in fact, the probability of  $0.138 - 0.112 = 0.026$  is attributable to stopping at a  $Y$  which ultimately is not the best  $Y$ , but corresponds to the  $X$  which is the best.

As mentioned in (S2), choosing  $N(n) = n/e$  in the secretary-like rule might not be the best choice. In fact, it is stated that the value of  $N(n, X, \varepsilon) \leq N(n)$ . In the present example,  $N(10000, X, \varepsilon) = 2740 < 3678 = 10000/e$ . When  $n = 10000$  and  $\rho = 0.9$ , the optimal secretary-like rule has probability 0.141 of finding the best  $X$ .

In order to show that the secretary rule is not necessarily optimal when there is noise, as mentioned in (S4), we consider  $n = 5$  items. The optimal classical secretary rule on the  $Y$  values has  $N(5) = 2$  (i.e. two items are allowed to pass before selection). This results in stopping at the  $i$ th best  $Y$  with respective probabilities  $\frac{13}{30}, \frac{7}{30}, \frac{4}{30}, \frac{3}{30}$ , and  $\frac{3}{30}$ . On the other hand, if we apply the same rule, but also stop at the next to last item if it is either the relative best or second best, then this results in stopping at the  $i$ th best  $Y$  with respective probabilities  $\frac{12}{30}, \frac{9}{30}, \frac{5}{30}, \frac{2}{30}$ , and  $\frac{2}{30}$ . These probabilities are derived by simple calculation.

The five probabilities for each of the cases need to be weighted by the probability that the  $i$ th best  $Y$  corresponds to the best  $X$ . We find the respective probabilities when  $\rho = 0.5$  to be

0.4110, 0.2490, 0.1675, 0.1104, and 0.0621. These probabilities are found by simulation with ten million replications to ensure accuracy. Finally, the optimal secretary rule finds the best  $X$  with probability 0.2758 as compared to 0.2785 if we stop with relative rank of two on the next to last observation, conditional on not having stopped earlier.

**3. Probability that the best concomitant observation is the best observation**

In this section we prove (R1)–(R5).

Let  $X_1, \dots, X_n$  and  $\varepsilon_1, \dots, \varepsilon_n$  be independent random variables, where the  $X_i$ s are i.i.d. with distribution  $F$ , and the  $\varepsilon_i$  are i.i.d. with distribution  $G$ . Let  $Y_i = X_i + \varepsilon_i$ . Our interest is in the behavior of  $p_n$  given in (1.1).

Let  $X_{(1)}^n \geq X_{(2)}^n \geq \dots \geq X_{(n)}^n$  be the order statistics of  $X_1, \dots, X_n$ . Let  $Y_{[j]}^n = X_{(j)}^n + \varepsilon_j^*$ , where the  $\varepsilon_j^*$  are i.i.d. and a random permutation of  $\varepsilon_1, \dots, \varepsilon_n$ . The variables  $Y_{[j]}^n$  are called the concomitant random variables, i.e. the random variable  $Y_i$  that ‘belongs to’  $X_{(j)}^n$ .

For any (cumulative) distribution  $H$ , let  $x_H = \sup\{x : H(x) < 1\}$ .

**Theorem 3.1.** *If  $F$  is such that, for every fixed  $c > 0$  and fixed integer  $k$ ,*

$$\lim_{n \rightarrow \infty} P(X_{(1)}^n - X_{(k)}^n < c) = 1 \tag{3.1}$$

then  $\lim_{n \rightarrow \infty} p_n = 0$  for all  $G$ .

*Proof.* Let  $n > k$ , and fix  $c$  of (3.1). Then

$$\begin{aligned} p_n &= P\left(Y_{[1]}^n = \max_{i=1, \dots, n} Y_i\right) \\ &< P\left(Y_{[1]}^n > \max_{j=2, \dots, k} Y_{[j]}^n\right) \\ &= P\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2, \dots, k} (X_{(j)}^n + \varepsilon_j^*)\right). \end{aligned} \tag{3.2}$$

We will show that, for any  $\delta > 0$  and sufficiently large  $n$ ,  $p_n < \delta$ . Let  $x_0$  be such that  $G(x_0) = 1 - \delta/4$ . Then

$$\begin{aligned} P\left(\max_{j=2, \dots, k} \varepsilon_j^* < \varepsilon_1^* + c\right) &\leq \int_{-\infty}^{x_0} [G(x+c)]^{k-1} g(x) dx + \frac{\delta}{4} \\ &< [G(x_0+c)]^{k-1} + \frac{1}{4}\delta \\ &< \frac{1}{2}\delta, \end{aligned} \tag{3.3}$$

provided we choose  $k$  large enough that  $[G(x_0+c)]^{k-1} < \delta/4$ . If  $x_G = \infty$ , this is always possible, but if  $x_G < \infty$ , we may have to replace the original  $c$  by a smaller value,  $c_0$ , such that  $G(x_0+c_0) < 1$ .

Now, using (3.1) with  $c$  and  $k$  satisfying (3.3), choose  $n$  sufficiently large that

$$P(X_{(1)}^n - X_{(k)}^n < c) > 1 - \frac{1}{2}\delta \quad \text{for all } n > N(\delta).$$

Let  $A_n$  denote the event  $\{X_{(1)}^n - X_{(k)}^n < c\}$ , and let  $\bar{A}_n$  be its complement. Then, for that  $c$

and  $k$ , (3.2) becomes

$$\begin{aligned} & \mathbb{P}\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2,\dots,k} (X_{(j)}^n + \varepsilon_j^*)\right) \\ & \leq \mathbb{P}\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2,\dots,k} (X_{(j)}^n + \varepsilon_j^* \mid A_n)\right) \mathbb{P}(A_n) + \mathbb{P}(\bar{A}_n) \\ & < \mathbb{P}\left(X_{(1)}^n + \varepsilon_1^* > X_{(1)}^n - c + \max_{j=2,\dots,k} \varepsilon_j^*\right) + \frac{1}{2}\delta \\ & = \mathbb{P}\left(\varepsilon_1^* + c > \max_{j=2,\dots,k} \varepsilon_j^*\right) + \frac{1}{2}\delta \\ & < \frac{1}{2}\delta + \frac{1}{2}\delta \\ & = \delta, \end{aligned}$$

where we have used (3.3) in the last inequality.

**Corollary 3.1.** *Statement (R2) holds.*

*Proof.* If  $x_F < \infty$  then, clearly, (3.1) holds.

**Theorem 3.2.** *A sufficient condition for  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$  for all  $G$  is that  $x_F = \infty$  and that, for any fixed  $d > 0$ ,*

$$\lim_{x \rightarrow \infty} \mathbb{P}(X \geq x + d \mid X > x) = 0, \tag{3.4}$$

or, equivalently,

$$\lim_{x \rightarrow \infty} \frac{f(x + d)}{f(x)} = 0. \tag{3.5}$$

*Proof.* We will show that (3.4) implies (3.1). For a given  $c$  and  $k$ , let  $d = c/(k - 1)$ . Suppose that (3.4) holds, and let  $x_0$  be such that

$$\mathbb{P}(X \geq x + d \mid X > x) < \frac{\delta}{2(k - 1)} \quad \text{for all } x > x_0. \tag{3.6}$$

Let  $N$  be sufficiently large such that, for all  $n \geq N$ ,

$$\mathbb{P}(X_{(k)}^n > x_0) > 1 - \frac{1}{2}\delta. \tag{3.7}$$

Let  $B_n$  be the event  $\{X_{(k)}^n > x_0\}$ . Then

$$\begin{aligned} \mathbb{P}(X_{(1)}^n - X_{(k)}^n < c) &= \mathbb{P}\left(\sum_{i=1}^{k-1} (X_{(i)}^n - X_{(i+1)}^n) < c\right) \\ &> \mathbb{P}\left(\bigcap_{i=1}^{k-1} \{X_{(i)}^n - X_{(i+1)}^n < d\}\right) \\ &> \mathbb{P}\left(\bigcap_{i=1}^{k-1} \{X_{(i)}^n - X_{(i+1)}^n < d\} \mid B_n\right) \mathbb{P}(B_n) \\ &> \left[1 - \sum_{i=1}^{k-1} \mathbb{P}(\{X_{(i)}^n - X_{(i+1)}^n \geq d\} \mid B_n)\right] \left(1 - \frac{\delta}{2}\right) \quad (\text{by (3.7)}) \\ &> \left(1 - (k - 1) \frac{\delta}{2(k - 1)}\right) \left(1 - \frac{\delta}{2}\right) \\ &> 1 - \delta \quad (\text{by (3.6)}). \end{aligned}$$

Since  $\delta > 0$  was arbitrarily small, (3.1) holds, and the result follows.

Since (3.4) can be written as

$$\lim_{x \rightarrow \infty} \frac{1 - F(x + d)}{1 - F(x)},$$

we can, by l'Hôpital's rule, take the limit of derivatives, which yields (3.5).

**Theorem 3.3.** *Let  $x_F = \infty$  and  $d > 0$ , and assume that  $\lim_{x \rightarrow \infty} P(X \geq x + d \mid X > x)$  exists and is equal to  $a$ , where  $a > 0$ . Then there exists a  $G$ , and  $\varepsilon \sim G$ , such that  $\liminf_{n \rightarrow \infty} p_n(X, \varepsilon) > a - \gamma$  for any  $\gamma > 0$ .*

Theorem 3.3 establishes the necessary statement (R1).

*Proof of Theorem 3.3.* Fix  $\gamma > 0$ , and let  $\delta = \gamma/(1 + a)$ . There exists an  $x_0$  such that  $P(X \geq x + d \mid X > x) \geq a - \delta$  for all  $x > x_0$ . Since  $X_{(2)}^n$  goes to  $\infty$  with probability 1 as  $n \rightarrow \infty$  (and it is stochastically increasing in  $n$ ), there exists an  $N$  and an  $x_1 > x_0$  such that  $P(X_{(2)}^n > x_1) \geq 1 - \delta$  for all  $n > N$ . Let  $\varepsilon$  have a uniform distribution on  $[0, d]$ . We make use of the following result which is straightforward to verify. Let  $X_1, \dots, X_n$  be i.i.d. continuous random variables with distribution  $F$ . Then

$$P(X_{(1)}^n \geq x + d \mid X_{(2)}^n = x) = P(X \geq x + d \mid X > x),$$

where  $X \sim F$ . Hence,

$$\begin{aligned} p_n(X, \varepsilon) &= P\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2, \dots, n} (X_{(j)}^n + \varepsilon_j^*)\right) \\ &\geq P(X_{(1)}^n \geq X_{(2)}^n + d) \\ &= \int_{w=-\infty}^{\infty} P(X_{(1)}^n \geq w + d \mid X_{(2)}^n = w) f_{X_{(2)}^n}(w) dw \\ &\geq \int_{w=x_1}^{\infty} P(X_{(1)}^n \geq w + d \mid X_{(2)}^n = w) f_{X_{(2)}^n}(w) dw \\ &\geq (a - \delta) P(X_{(2)}^n > x_1) \\ &\geq (a - \delta)(1 - \delta) \\ &> a - \gamma. \end{aligned}$$

In Section 5 we consider two examples, Examples 5.1 and 5.2, where  $F$  has an exponential distribution. Depending on  $G$ ,  $\lim_{n \rightarrow \infty} p_n = 0$  or  $\liminf_{n \rightarrow \infty} p_n > 0$ . The exponential distribution is of special interest because

$$P(X > x + d \mid X > x) = e^{-d},$$

i.e. it is independent of  $x$ , and, thus, can be considered as a borderline case. For many well-known distributions, if (3.4) fails, the limit on the left-hand side of (3.4) will be 1.

To show (R3), we need the following lemma.

**Lemma 3.1.** *For any continuous distribution  $F$  and  $X_i$  i.i.d. distributed  $F$ , there exists a distribution  $H$  and i.i.d.  $Z_i$ , independent of the  $X_i$ s, distributed as  $H$ , such that*

$$\lim_{n \rightarrow \infty} P(X_{(1)}^n > Z_{(1)}^n) = 0. \tag{3.8}$$

*Proof.* We will first prove the statement when  $F$  is the uniform distribution on  $[0, 1]$ . Then, for any continuous  $H$ ,

$$P(X_{(1)}^n > Z_{(1)}^n) = n \int_0^1 [H(x)]^n x^{n-1} dx.$$

Now let  $H(x) = 1 - (1 - x)^{1/2}$  for  $0 \leq x \leq 1$ . We will show that (3.8) holds. We have

$$P(X_{(1)}^n > Z_{(1)}^n) = n \int_0^1 [1 - (1 - x)^{1/2}]^n x^{n-1} dx = 2n \int_0^1 (1 - y)^n (1 - y^2)^{n-1} y dy, \tag{3.9}$$

where we have made the change of variable  $y = (1 - x)^{1/2}$ . The right-hand side of (3.9) equals

$$\begin{aligned} & 2n \int_0^{n^{-2/3}} (1 - y^n)(1 - y^2)^{n-1} y dy + 2n \int_{n^{-2/3}}^1 (1 - y)^n (1 - y^2)^{n-1} y dy \\ & < 2n \int_0^{n^{-2/3}} y dy + 2n \int_{n^{-2/3}}^1 (1 - y)^n dy \\ & = n^{-1/3} + \frac{2n}{n + 1} (1 - n^{-2/3})^{n+1}. \end{aligned} \tag{3.10}$$

The first term on the right-hand side of (3.10) clearly tends to 0. The last term is less than  $2(1 - n^{-2/3})^n = 2[(1 - 1/m)^m]^{\sqrt{m}}$ , where  $m = n^{2/3}$ , which also goes to 0, since the value in the bracket is arbitrarily close to  $e^{-1}$ . This proves (3.8) for  $F$  uniformly distributed on  $[0, 1]$ .

To generalize, add asterisks to all of the previous variables, i.e.  $X_1^*, X_2^*, \dots, Z_1^*, Z_2^*, \dots$ , and  $H^*$ , and note that  $H^*$  has all of its mass on  $[0, 1]$ . Now consider any continuous  $F$ , and its inverse  $F^{-1}$ . Let  $X_i = F^{-1}(X_i^*)$  and  $Z_i = F^{-1}(Z_i^*)$ . Then the  $X_i$ s are i.i.d. with distribution  $F$  and the  $Z_i$  are i.i.d. with distribution  $H(x) = H^*(F(x))$ . But, since  $F^{-1}(x)$  is monotone increasing, clearly,

$$P(X_{(1)}^n > Z_{(1)}^n) = P(X_{(1)}^{*n} > Z_{(1)}^{*n})$$

and (3.8) follows.

**Theorem 3.4.** *For any  $F$ , there exists  $G$  such that  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ .*

*Proof.* If  $x_F < \infty$ , the result follows from Corollary 3.1. Thus, assume that  $x_F = \infty$ .

We first consider  $P(X_i \geq 0) = 1$ . Note that  $\varepsilon_i^*$  is the  $\varepsilon$  that is associated with the  $i$ th largest  $X_i$ . Then  $Y_{[1]}^n = X_{(1)}^n + \varepsilon_1^*$ . We choose  $G$  so that  $\varepsilon^* > 0$ . We want to show that  $p_n := P(Y_{[1]}^n = \max_{j=1, \dots, n} Y_j) \rightarrow 0$ . Now

$$\begin{aligned} P\left(Y_{[1]}^n > \max_{j=2, \dots, n} Y_{[j]}^n\right) &= P\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2, \dots, n} (X_{(j)}^n + \varepsilon_j^*)\right) \\ &\leq P\left(X_{(1)}^n + \varepsilon_1^* > \max_{j=2, \dots, n} \varepsilon_j^*\right) \\ &= P(X_{(1)}^n > \varepsilon_{(1)}^{n-1} - \varepsilon_1^*) \\ &\leq P(X_{(1)}^n > \varepsilon_{(1)}^n - 2\varepsilon_1^*), \end{aligned} \tag{3.11}$$

where we have used  $X_j \geq 0$  in the first inequality, and  $\varepsilon_1^* \geq 0$  and  $\varepsilon_{(1)}^n \leq \varepsilon_{(1)}^{n-1} + \varepsilon_1^*$  in the second inequality. Consider  $Z_i$  of Lemma 3.1. Since  $X_{(1)}^n \rightarrow \infty$  almost surely as  $n \rightarrow \infty$ , it follows that  $Z_{(1)}^n \rightarrow \infty$ , and we may take  $Z_i \geq 0$ . Let  $\delta > 0$ . For a proper choice of  $G$ , we will show that  $p_n < \delta$  for all sufficiently large  $n$ . Let  $\varepsilon_i = 2Z_i$ , where  $Z_i$  satisfies (3.8), let



$z_0$  be a constant such that  $P(4Z_i > z_0) < \delta/4$ , and let  $n$  be so large that  $P(Z_{(1)}^n < z_0) < \delta/4$ . With these choices, the inequality in (3.11) can be continued:

$$\begin{aligned} P(X_{(1)}^n > 2Z_{(1)}^n - 4Z_{i_n}) &< \frac{\delta}{4} + P(X_{(1)}^n > 2Z_{(1)}^n - z_0) \\ &< \frac{\delta}{4} + P(Z_{(1)}^n \leq z_0) + P(X_{(1)}^n > Z_{(1)}^n) \\ &< \frac{2\delta}{4} + P(X_{(1)}^n > Z_{(1)}^n). \end{aligned}$$

Now, by (3.8) we can choose  $n$  sufficiently large that the last term is less than  $\delta/4$ ; thus,  $p_n < \frac{3}{4}\delta$  for all sufficiently large  $n$ .

If  $X_i$  can take on negative values, but is bounded from below by some  $c < 0$ , shift  $X_i$  by  $c$  to obtain  $\hat{X}_i = X_i - c \geq 0$ . Let  $\hat{Y}_j = \hat{X}_j + \varepsilon_j$ . Then

$$\begin{aligned} \hat{p}_n &:= P(\hat{Y}_{[1]}^n > \max_{j=2,\dots,n} \hat{Y}_{[j]}^n) \\ &= P(\hat{X}_{(1)}^n + \varepsilon_1^* > \max_{j=2,\dots,n} (\hat{X}_{(j)} + \varepsilon_j^*)) \\ &= P(X_{(1)}^n - c + \varepsilon_1^* > \max_{j=2,\dots,n} (X_{(j)} + \varepsilon_j^*) - c) \\ &= P(X_{(1)}^n + \varepsilon_1^* > \max_{j=2,\dots,n} (X_{(j)} + \varepsilon_j^*)) \\ &= p_n. \end{aligned}$$

But,  $\hat{X}_i \geq 0$ , so if the  $\varepsilon_i$  are chosen so that  $\hat{p}_n \rightarrow 0$ , the same  $\varepsilon_i$  will do for the original  $X_i$ , and  $p_n < \frac{3}{4}\delta$  for all sufficiently large  $n$ .

Now consider the  $X_i$  which are not bounded below. Choose  $n$  sufficiently large so that  $P(X_{(1)}^n < c) < \delta/4$ . On  $\{X_{(1)}^n \geq c\}$  we may replace  $X_i$  by  $\check{X}_i = \max(X_i, c)$ , which are bounded, and obtain  $p_n < \delta/4 + 3\delta/4 = \delta$ .

Note that Theorem 3.4 establishes (R3). Statement (R4) is established through Example 5.4 below where  $F(x) = (1 - x^{-1}) \mathbf{1}_{\{x>1\}}$  and  $G$  is  $N(0, 1)$ . Statement (R5) also follows from Example 5.4.

### 4. Results for the noisy secretary problem

In this section we prove or illustrate the five results, (S1)–(S5), stated in the introduction. There are two versions of the secretary problem in the presence of noise. In one problem we only observe whether a noisy observation is the relative best  $Y$ . In the other problem we observe  $RR(Y_i)$ , that is, the relative rank of the noisy observation amongst those observed so far. It is important to note that if we observed  $X_i$ , or, equivalently, there was no noise, then there would not be a distinction between these two problems. If there is no noise then it is obvious that we should not stop if an observation is not the relative best.

Result (S1) relates the probability of finding the best  $X$  in the two versions of the problem, that is, knowing  $RB(X_i)$  as compared to knowing  $RB(Y_i)$ . Result (S2) considers the rule that only uses  $RB(Y_i)$ . Specifically, let a certain number of observations go by and then stop at the first  $i$  such that  $RB(Y_i) = 1$ . The main finding is that it is optimal to let fewer observations go by when the data are noisy than in the classical secretary problem (where it is optimal to let approximately  $n/e$  observations go by). Result (S3) indicates that the two versions of the problem do not necessarily have the same solution. This is shown by an example.

In the last two results we consider the problem where the relative ranks of the noisy data are available. An algorithm that produces the optimal solution for this problem, based on dynamic programming, is described in the discussion of (S4), which largely parallels the treatment in Ferguson (2008). Finally, it is shown that the probability of finding the best  $X$  goes to 0 when and only when  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ . This is in contrast to the classical secretary problem where the probability of choosing the best  $X$  goes to  $e^{-1}$ . The results that follow depend on

$$q_{nijk} = P(\text{AR}(Y_i) = k \mid \text{RR}(Y_i) = j) \tag{4.1}$$

for  $1 \leq j \leq i$  and  $j \leq k \leq n + j - i$ . This probability, which is negative hypergeometric, requires that the first  $i$  items include  $j - 1$  of the observations that have absolute rank of at most  $k - 1$  and the remaining  $i - j$  items from among those with absolute rank exceeding  $k$ . Hence,

$$q_{nijk} = \frac{\binom{k-1}{j-1} \binom{n-k}{i-j}}{\binom{n}{i}}.$$

This is given in Ferguson (2008, Chapter 2, p. 2.4).

In order to show that (S1) holds, suppose that we use the simple classical rule which maximizes the probability of finding the maximal  $Y$ . The probability of finding the maximal  $Y$  is  $W_n$ . The probability that this is also the maximal  $X$  is  $p_n$ . In addition, this rule may pick a  $Y$  value which turns out not to be the maximal  $Y$ , but could still be the  $Y$  value that corresponds to the maximal  $X$ , i.e.  $Y_{[1]}^n$ . Thus, with this rule, we achieve a value which is at least  $W_n p_n$ . As this rule may not be optimal (see (S2)), an optimal rule may achieve an even higher value.

To show that (S2) holds, we consider optimal rules subject to the assumption that we know only whether an observation is the relative best. These rules can be characterized by an integer  $S(n)$  which allows us to state that the stopping time is the first time that  $\text{RB}(Y_i) = 1$  for  $i > S(n)$ . This is akin to the elegant result in Bruss (2000), who showed how to obtain the secretary rule by summing odds. The difference here is that we do not observe the variables  $I_i$  which indicate whether we have a relative record among the  $X$  values at the  $i$ th observation. Rather we observe the noisy data, which indicates whether we have a relative record among the  $Y$  values. We prove the following theorem that relates the best secretary rule in the classical problem to that in the noisy problem.

**Theorem 4.1.** *Let  $N(n)$  be the number of observations in the classical secretary rule such that we stop the first time,  $i$ , that  $i > N(n)$  and  $\text{RB}(X_i) = 1$ . The optimal value  $N(n, X, \varepsilon)$  after which we should pick the first item for which  $\text{RB}(Y_i) = 1$  in the noisy case satisfies  $N(n, X, \varepsilon) \leq N(n)$ .*

Let  $S^*(n) = N(n, X, \varepsilon)$  denote the optimal stopping rule in the noisy case, which depends on the horizon,  $n$ , and the distributions of  $X$  and  $\varepsilon$ . Let  $N(n) \approx n/e$  be the analog to  $S^*(n)$  in the classical secretary rule, which is based on  $\text{RB}(X_i)$ . Then Theorem 4.1 shows that  $S^*(n) \leq N(n)$  and, hence, when there is noise, we should stop no later than when there is no noise.

*Proof of Theorem 4.1.* Consider the rule in the noisy case where  $S$  (for  $1 \leq S \leq n - 1$ ) items are allowed to go by and we stop at the first  $i > S$  for which  $\text{RB}(Y_i) = 1$  (otherwise, stop at  $n$  anyway). Let

$$r_{in} = P(\text{AR}(X_i) = 1 \mid \text{RB}(Y_i) = 1) = \sum_{k=1}^{n-i+1} q_{ni1k} p_{nk}, \tag{4.2}$$

where  $q_{ni1k}$  is given in (4.1) and  $p_{nk}$  is given in (1.2). Hence, the probability that this rule chooses the best  $X$  is

$$P(S) := \sum_{i=S+1}^n \frac{1}{i} \frac{S}{i-1} r_{in},$$

where the  $1/i$  term is the probability that  $Y_i$  has a relative rank of 1 and  $S/(i-1)$  is the probability that the best in the first  $i-1$  observations is among the first  $S$  items (so that one does not stop before the  $i$ th observation).

Let  $r_{in}^* = nr_{in}/i$ . This implies that

$$P(S) = \frac{S}{n} \sum_{i=S+1}^n \frac{r_{in}^*}{i-1}.$$

Consider

$$\begin{aligned} P(S+1) - P(S) &= \frac{S+1}{n} \sum_{i=S+2}^n \frac{r_{in}^*}{i-1} - \frac{S}{n} \sum_{i=S+1}^n \frac{r_{in}^*}{i-1} \\ &= \frac{S+1}{n} \sum_{i=S+2}^n \frac{r_{in}^*}{i-1} - \frac{S}{n} \left( \sum_{i=S+2}^n \frac{r_{in}^*}{i-1} + \frac{r_{S+1,n}^*}{S} \right) \\ &= \frac{1}{n} \left( \sum_{i=S+2}^n \frac{r_{in}^*}{i-1} - r_{S+1,n}^* \right). \end{aligned}$$

The above expression is nonnegative if and only if

$$\sum_{i=S+2}^n \frac{1}{i-1} \frac{r_{in}^*}{r_{S+1,n}^*} \geq 1.$$

Note that, for the classical secretary problem,  $r_{in}^* = 1$ .

If we can show that  $r_{in}^*$  decreases as  $i$  increases then the optimal  $S$  must necessarily be smaller than the corresponding value for the classical secretary problem, as desired. Note that  $p_{nk}$  does not depend on  $i$ , and, as  $i$  increases, the number of terms in (4.2) decreases. Hence, it is sufficient to show that  $nq_{ni1k}/i$  decreases in  $i$  for any  $n$  and  $k$ . But,  $q_{ni11} = i/n$  since we need the best  $Y$  to be among the first  $i$  items. In general, for  $k > 1$ ,

$$q_{ni1k} = \frac{\binom{n-k}{i-1}}{\binom{n}{i}}.$$

Hence,

$$\frac{q_{ni1k}}{i/n} = \frac{(n-k)!}{(n-1)!} \frac{(n-i)!}{(n+1-i-k)!}.$$

The above expression clearly decreases in  $i$  for any  $k > 1$  and  $n$ .

We now provide an example that shows that (S3) holds, that is, it is better to stop in some cases with  $RR(Y_i) > 1$ . The smallest number of observations for which this can occur is  $n = 4$ , where it might be better to take the third observation if it is second best among the first three  $Y$  values. We present such an example below, involving exponential random variables.

**Example 4.1.** (*The solution based on RR is better than the solution based on RB.*) Let  $X_i$  be i.i.d. exponential with  $\mu = 1$ , and let  $\varepsilon_i$  be i.i.d. exponential with  $\mu = c$ , where  $\mu$  is the mean. Let  $Y_{[i]}^3 = X_{(i)}^3 + \varepsilon_i^*$ . We want to show that it is better to stop at  $n = 3$  if  $RR(Y_3) = 2$ . To this end, we consider

$$\gamma_{ci} = P(Y_{[1]}^3 = Y_{(i)}^3) \quad \text{for } i = 1, 2, 3.$$

Specifically, since there is a  $\frac{3}{4}$  chance that the best  $X$  in four observations is among the first three observations, we need to show that  $3\gamma_{c2}/4 > \frac{1}{4}$  or  $\gamma_{c2} > \frac{1}{3}$ .

We need the following lemma.

**Lemma 4.1.** *Let  $c$  be the mean in the exponential distribution of the  $\varepsilon$ s. Then*

$$\begin{aligned} \gamma_{c1} &= 1 + \frac{2c^2}{3(2c + 1)(c + 2)} - \frac{c(4c + 1)}{2(2c + 1)(c + 1)}, \\ \gamma_{c2} &= \frac{c(4c + 1)}{2(2c + 1)(c + 1)} - \frac{4c^2}{3(2c + 1)(c + 2)}, \\ \gamma_{c3} &= \frac{2c^2}{3(2c + 1)(c + 2)}. \end{aligned}$$

Before we prove the lemma we observe that if  $c = 3$  then  $\gamma_{c1} = \frac{133}{280}$ ,  $\gamma_{c2} = \frac{99}{280}$ , and  $\gamma_{c3} = \frac{48}{280}$ . Since  $\gamma_{c2}$  is approximately  $0.35357 > \frac{1}{3}$ , this is an example where it is better to stop at  $n = 3$  if we observe the second largest  $Y$  value from among the first three  $Y$  values. The largest value that  $\gamma_{c2}$  can achieve is  $0.36275$ . This occurs when  $c = 5.535$ .

*Proof of Lemma 4.1.* We begin with  $\gamma_{c3}$  as it is the easiest and highlights the argument. The probability of interest is

$$P([X_{(1)}^3 + \varepsilon_1^* < X_{(2)}^3 + \varepsilon_2^*] \cap [X_{(1)}^3 + \varepsilon_1^* < X_{(3)}^3 + \varepsilon_3^*]).$$

First note that since the  $X_i$  are i.i.d. standard exponential random variables, we can express the resulting order statistics as  $X_{(3)}^3 = E_3$ ,  $X_{(2)}^3 = E_3 + E_2$ , and  $X_{(1)}^3 = E_3 + E_2 + E_1$ , where the  $E_i$  are independent exponential random variables with mean  $1/i$ . In order to evaluate the above probability, consider the two events  $A(x, y) = \{X_{(2)}^3 + \varepsilon_2^* > X_{(1)}^3 + \varepsilon_1^* \mid ([\varepsilon_1^* = x] \cap [E_1 = y])\}$  and, similarly,  $B(x, y) = \{X_{(3)}^3 + \varepsilon_3^* > X_{(1)}^3 + \varepsilon_1^* \mid ([\varepsilon_1^* = x] \cap [E_1 = y])\}$ . First,  $A(x, y)$  and  $B(x, y)$  are independent conditional on  $E_1$ . This follows since  $X_{(1)}^3 - X_{(2)}^3 = E_1$  and  $X_{(1)}^3 - X_{(3)}^3 = E_1 + E_2$ ; hence,  $A(x, y)$  depends only on  $\varepsilon_2^*$ ,  $x$ , and  $y$ , and  $B(x, y)$  depends only on  $E_2$ ,  $\varepsilon_3^*$ ,  $x$ , and  $y$ . Second,  $P(A(x, y)) = e^{-(y+x)/c}$  and

$$P(B(x, y)) = \int_{v=0}^{\infty} e^{-(y+x+v)/c} 2e^{-2v} dv = \frac{2ce^{-(y+x)/c}}{2c + 1}.$$

Therefore,

$$\gamma_{c3} = \frac{2c}{2c + 1} \int_{x=0}^{\infty} \int_{y=0}^{\infty} e^{-2(y+x)/c} e^{-x} \frac{1}{c} e^{-y/c} dy dx = \frac{2c^2}{3(2c + 1)(c + 2)}.$$

To obtain  $\gamma_{c1}$ , we use a similar argument. Hence,

$$\gamma_{c1} = \int_{x=0}^{\infty} \int_{y=0}^{\infty} P(\overline{A(x, y)}) P(\overline{B(x, y)}) e^{-x} \frac{1}{c} e^{-y/c} dy dx.$$

Finally,  $\gamma_{c2} = 1 - (\gamma_{c1} + \gamma_{c3})$ .

In (S4), the method for finding the optimal rule when relative ranks are observed is described. The optimal rule, which can in principle be found by backward induction, is of the following form. There exist integer values  $0 \leq k_1 \leq \dots \leq k_n = n$  not necessarily distinct such that we should stop on the smallest  $i$  satisfying  $RR(Y_i) \leq k_i$ . The obvious way to proceed, which we programmed, is by backward induction. Once we determine  $p_{nm}$  by simulation, the backward induction, which we outline below, is distribution free.

At observation  $i$ , we need to decide whether we should stop or not if  $RR(Y_i) = j$  for  $1 \leq j \leq i$ . If we were to stop when  $RR(Y_i) = j$  then the probability that  $AR(X_i) = 1$  is

$$f_{nij} = \sum_{v=j}^{n-i+j} q_{nijv} p_{nv},$$

i.e.  $f_{nij}$  is the probability that  $X_i$  is the best among all of the  $n$   $X$  observations, conditional on  $Y_i$  being the  $j$ th best from among the first  $i$   $Y$  observations.

We need to keep track of  $R_{ni}$ , which is the probability of getting the best  $X$  if the optimal rule is followed from observation  $i$  and thereafter. To complete the discussion, we need to show how  $R_{ni}$  is determined recursively, beginning with the last observation,  $n$ , and going backwards.

Note that  $R_{nn} = 1/n$ . For any  $i$ , let  $k_i$  be the largest  $j$  such that  $f_{nij} > R_{n,i+1}$ . Then

$$R_{ni} = \frac{1}{i} \sum_{j=1}^{k_i} f_{nij} + R_{n,i+1} \frac{i - k_i}{i}.$$

The form of the above solution, which states that the maximum  $RR(Y_i)$  for which we stop at observation  $i$  is nondecreasing in  $i$ , is intuitive. It also follows from the solution described above since  $R_{ni}$  is nonincreasing in  $i$ ,  $p_{nj}$  is clearly nonincreasing in  $j$ , and there exists a  $k_0$  which depends on  $i$  such that  $q_{nijk} \leq q_{n,i+1,j,k}$  only when  $k \leq k_0$ . The last two statements imply that  $f_{nij}$  increases as  $i$  increases.

Result (S5) is straightforward. No rule can be better than the rule that finds the observation that is best amongst the  $Y$  values with certainty. But, if  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ , even this rule satisfies  $\lim_{n \rightarrow \infty} W_n(X, \varepsilon) = 0$ . The converse follows from (S1).

### 5. Examples

In this section we present four examples to illustrate interesting findings concerning the behavior of  $p_n$ . It follows from (2.1) that when  $X$  is normal and  $\varepsilon$  is normal, the probability that the best concomitant observation is the best observation goes to 0 if  $\rho < 1$ . In Example 5.1 below we show that the probability does not go to 0 if the  $X$  distribution is exponential.

**Example 5.1.** (*Exponential  $F$ , normal  $G$ , and  $\liminf_{n \rightarrow \infty} p_n > 0$ .)* We want to know how likely it is that the largest  $Y_i$  has the same index as the largest  $X_j$ . Hence, we need to consider

$$P(X_{(1)}^n + \varepsilon_1^* > X_{(j)}^n + \varepsilon_j^*, j = 2, \dots, n),$$

where the  $\varepsilon_j^*$  are i.i.d. standard normal. Let  $A_j = \{X_{(1)}^n - X_{(j)}^n > \varepsilon_j^* - \varepsilon_1^*\}$ . Note that  $A_j$  depends on  $n$ , but, for ease of notation and because the value of  $n$  remains fixed in this argument, we do not include  $n$  as a superscript. We want to show that  $P(A_2 \cap A_3 \cap \dots \cap A_n)$  goes to a constant greater than 0.

It suffices to show that  $P(A_2 \cap A_3 \cap \dots \cap A_n \mid \varepsilon_1^* = z)$  goes to a constant greater than 0 as  $n \rightarrow \infty$  for any  $z$ . This is the case since  $P(A_2 \cap A_3 \cap \dots \cap A_n) = \int_z P(A_2 \cap A_3 \cap \dots \cap A_n \mid \varepsilon_1^* = z) \times \phi(z) dz$ , where  $\phi$  is the density of the standard normal. Since the conditional probability in the

integral increases in  $z$ ,  $P(A_2 \cap A_3 \cap \dots \cap A_n) \geq P(A_2 \cap A_3 \cap \dots \cap A_n \mid \varepsilon_1^* = z)P(Z > z)$ , where  $Z$  is a standard normal random variable independent of the  $E_i$ s. But,

$$P(A_2 \cap A_3 \cap \dots \cap A_n \mid \varepsilon_1^* = z) = 1 - P(B_1 \cup \dots \cup B_{n-1}) \geq 1 - \sum_{j=1}^{n-1} P(B_j),$$

where  $B_j = \{X_{(1)} - X_{(j+1)} \leq \varepsilon_{j+1}^* - z\}$ .

To evaluate  $P(B_j)$ , note that, for i.i.d. exponential random variables,  $E_i = X_{(i)}^n - X_{(i+1)}^n$  are independent exponential with mean of  $1/i$ . Hence,  $P(B_j) = P(E_1 + \dots + E_j + Z \leq -z)$ .

It suffices to show that  $\sum_{i=1}^{n-1} P(B_i) < 1$  as  $n \rightarrow \infty$ . We use the following Chernoff bound:  $P(H \leq a) \leq e^{-ta}M(t)$  for all  $t < 0$ , where  $M$  is the moment generating function of  $H$ . The moment generating function of the random variable  $E_1 + \dots + E_j + Z$  is

$$M_j(t) = e^{t^2/2} \prod_{i=1}^j \frac{i}{i-t}.$$

We now choose  $z = -a$  to be 2 and  $t$  to be  $-2$ . This implies that

$$P(B_j) \leq e^{-4}e^2 \prod_{i=1}^j \frac{i}{i+2} = \frac{2e^{-2}}{(j+1)(j+2)}.$$

Hence,

$$\sum_{j=1}^{n-1} P(B_j) \leq 2e^{-2} \sum_{j=1}^{n-1} \frac{1}{(j+1)(j+2)} = e^{-2} \left(1 - \frac{2}{n+1}\right) < 1.$$

We have shown that, when  $X$  and  $\varepsilon$  have normal distributions,  $p_n$  goes to 0. This is intuitive because the differences between the largest  $X_i$  values becomes arbitrarily small in probability as  $n$  gets large. In the exponential case, however, the expected difference between the largest and the second largest observations is 1 and, hence, the largest  $X$  values do not become indistinguishable as in the normal case. Nevertheless, Example 5.2 below shows that, when the distributions of  $X$  and  $\varepsilon$  are both exponential, and, hence, the error term is sufficiently large,  $\lim_{n \rightarrow \infty} p_n(X, \varepsilon) = 0$ .

**Example 5.2.** (*Exponential  $F$ , exponential  $G$ , and  $\lim_{n \rightarrow \infty} p_n = 0$ .) Assume that we observe  $X_1, \dots$  and  $\varepsilon_1, \dots$  as i.i.d. exponential with equal means taken, without loss of generality, to be 1. Let  $Y_i = X_i + \varepsilon_i$ . Let  $x_N = X_{(1)}^N$  and  $y_N = Y_{(1)}^N$ . Let  $n > N$  be an observation after  $N$ . The probability that we have a record at  $n$  in at least one of the sequences  $X$  and  $Y$  is*

$$P([X_n > x_N] \cup [Y_n > y_N]).$$

We want to determine

$$\begin{aligned} &P([X_n > x_N] \cap [Y_n > y_N] \mid [X_n > x_N] \cup [Y_n > y_N]) \\ &= \frac{P([X_n > x_N] \cap [Y_n > y_N])}{P([X_n > x_N] \cup [Y_n > y_N])} \\ &\leq \frac{P([X_n > x_N] \cap [Y_n > y_N])}{\max[P(X_n > x_N), P(Y_n > y_N)]}. \end{aligned} \tag{5.1}$$

Since  $X_n$  is exponential with mean 1 and  $Y_n$  is Gamma (2,1), it follows that

$$P(X_n > x_N) = e^{-x_N}$$

and

$$P(Y_n > y_N) = (1 + y_N)e^{-y_N}.$$

Furthermore,

$$\begin{aligned} P([X_n > x_N] \cap [Y_n > y_N]) &= \int_{u=x_N}^{u=y_N} (e^{-u})(e^{u-y_N}) du + \int_{u=y_N}^{\infty} e^{-u} du \\ &= e^{-y_N}(y_N - x_N) + e^{-y_N} \\ &= e^{-y_N}(y_N - x_N + 1). \end{aligned}$$

For any  $\delta > 0$ , let

$$y(\delta) = \min \left\{ y \mid \frac{\log(1 + y) + 1}{y + 1} < \frac{\delta}{2} \right\}.$$

Let  $N$  be sufficiently large so that  $P(Y_{(1)}^m > y(\delta)) > 1 - \delta/2$  for all  $m > N$ . We need to consider two cases.

1. If  $e^{-x_N} > (1 + y_N)e^{-y_N}$ , or, equivalently,  $e^{y_N - x_N} > (1 + y_N)$ , then the right-hand side of (5.1) is  $e^{-y_N + x_N}(y_N - x_N + 1) < (\log(1 + y_N) + 1)/(1 + y_N) < \delta/2$ . The first inequality in the line above follows because  $e^{-u}(u + 1)$  is a decreasing function.
2. If  $e^{-x_N} \leq (1 + y_N)e^{-y_N}$ , or, equivalently,  $e^{y_N - x_N} \leq (1 + y_N)$ , then the right-hand side of (5.1) is  $(y_N - x_N + 1)/(y_N + 1) \leq (\log(1 + y_N) + 1)/(1 + y_N) < \delta/2$ .

Finally, if  $N$  is sufficiently large then

$$\begin{aligned} &P([X_n > x_N] \cap [Y_n > y_N] \mid ([X_n > x_N] \cup [Y_n > y_N])) \\ &\leq P([X_n > x_N] \cap [Y_n > y_N] \mid \{[X_n > x_N] \cup [Y_n > y_N]\} \cap [Y_{(1)}^n > y(\delta)]) \left(1 - \frac{1}{2}\delta\right) \\ &\quad + \frac{1}{2}\delta \\ &< \left(\frac{1}{2}\delta\right)\left(1 - \frac{1}{2}\delta\right) + \frac{1}{2}\delta \\ &< \delta. \end{aligned}$$

One might conjecture that if  $X$  and  $\varepsilon$  have the same distributions then  $p_n$  goes to 0 as in the normal and exponential cases. But, if the tail of  $X$  is sufficiently fat then  $p_n$  need not go to 0. The intuition is that the largest  $X$  is likely to be a lot larger than the second largest  $X$ . This is in essence what is shown in the following example.

**Example 5.3.** (*Pareto F, Pareto G, and  $\liminf_{n \rightarrow \infty} p_n > 0$ .*) Let  $F(x) = G(x) = 1 - 1/x^\alpha$  for  $x \geq 1$ . (Note that, for  $\alpha = 1$ , the tail behavior of this distribution is the same as that of a Cauchy distribution.) We want to show that  $\liminf_{n \rightarrow \infty} p_n > 0$ . If we can show that

$$P(X_{(1)}^n > X_{(2)}^n + \varepsilon_{(2)}^n) > \delta \tag{5.2}$$

for some  $\delta > 0$  as  $n \rightarrow \infty$ , we are done. The reason for this is that if the second largest  $\varepsilon$  coincides with the second largest  $X$  then the only other possible observation that has a higher  $Y$  than the  $Y$  with index corresponding to the largest  $X$  is the one with the largest  $\varepsilon$ . But it is just as likely that the index with the largest  $X$  and the index with the largest  $\varepsilon$  has the largest  $Y$  value.

Let  $X_i^\alpha = 1/U_i$ , where the  $U_i$  are i.i.d. uniform (0,1), and, similarly, let  $\varepsilon_i^\alpha = 1/V_i$ , where  $V_i$  are i.i.d. uniform (0,1).

We will show that  $\lim P(X_{(1)}^n - X_{(2)}^n > (n + 1)/2) > 0$  and that  $\lim P(\varepsilon_{(2)}^n < (n + 1)/2) > 0$  from which (5.2) follows. However,

$$\begin{aligned} &P\left([X_{(1)}^n > (n + 1)^{1/\alpha}] \cap \left[X_{(2)}^n < \left(\frac{n + 1}{2}\right)^{1/\alpha}\right]\right) \\ &= P\left(\left[U_{(n)}^n < \frac{1}{n + 1}\right] \cap \left[U_{(n-1)}^n > \frac{2}{n + 1}\right]\right) \\ &= n \left[\frac{1}{n + 1} \left(1 - \frac{2}{n + 1}\right)^{n-1}\right] \\ &\rightarrow e^{-2}, \end{aligned}$$

where  $U_{(i)}^n$  is the  $i$ th largest uniform that generates  $X_{(n+1-i)}$ . Note that

$$(n + 1)^{1/\alpha} - \left(\frac{n + 1}{2}\right)^{1/\alpha} = \frac{(n + 1)^{1/\alpha}}{c},$$

where  $c = 2^{1/\alpha}/(2^{1/\alpha} - 1)$ . Finally, since  $P(\varepsilon_{(2)}^n < ((n + 1)/c)^{1/\alpha})$  is  $P(V_{(n-1)}^n > c/(n + 1))$ , which goes to  $(c + 1)e^{-c}$ , where  $V_{(i)}^n$  is the  $i$ th largest uniform that generates  $\varepsilon_{(n+1-i)}^n$ , we are done.

It is somewhat intuitive that, as  $n$  increases, the probability that the observation with the largest  $X$  value also has the largest  $Y$  value decreases. However, this is not necessarily the case. In fact, Example 5.4 below says more. The above probability in Example 5.4 goes to 1 as  $n$  goes to  $\infty$ .

**Example 5.4.** (Pareto  $F$ , normal  $G$ , and  $\lim_{n \rightarrow \infty} p_n = 1$ .) Let  $X_1, \dots, X_n$  be i.i.d. with  $f(x)$  as in Example 5.3 and  $\alpha = 1$ , that is,  $F(x) = 1 - 1/x$  for  $x \geq 1$ . Assume that we observe  $Y_i = X_i + \varepsilon_i$ , where  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d. and normally distributed. We claim that  $P(X_{(1)}^n - X_{(2)}^n \geq z_n) \rightarrow 1$  as  $n \rightarrow \infty$ , where  $z_n = n^{1-\delta}$  for any  $\delta > 0$ .

Since the maximum of  $n$  normally distributed random variables is of the order  $(\log n)^{1/2}$  for large  $n$ , the above claim states that the observation with the largest  $X$  value must also have the largest  $Y$  value with a probability tending to 1.

*Proof of the claim.* Let  $X_{(i)}^n = 1/U_{(n+1-i)}^n$ . Then

$$\begin{aligned} P(X_{(1)}^n - X_{(2)}^n \geq z_n) &= P\left(\frac{1}{U_{(n)}^n} - \frac{1}{U_{(n-1)}^n} \geq z_n\right) \\ &= P\left(\frac{1}{U_{(n)}^n} \geq z_n + \frac{1}{U_{(n-1)}^n}\right) \\ &= P\left(U_{(n)}^n \leq \frac{U_{(n-1)}^n}{1 + z_n U_{(n-1)}^n}\right) \\ &= \int_{t=0}^1 P\left(U_{(n)}^n \leq \frac{U_{(n-1)}^n}{1 + z_n U_{(n-1)}^n} \mid U_{(n-1)}^n = t\right) f_{n-1}(t) dt \\ &= \int_{t=0}^1 \frac{1}{1 + z_n t} f_{n-1}(t) dt, \end{aligned}$$



where  $f_i$  is the density of the  $i$ th largest order statistic from a uniform (0,1) distribution which is Beta( $\alpha = n + 1 - i, \beta = i$ ). But,  $U_{(n-1)} \sim \text{Beta}(\alpha = 2, \beta = n - 1)$ . Hence,  $E(U_{(n-1)}^n) = 2/(n + 1)$  and  $\text{var}(U_{(n-1)}^n) = 2(n - 1)/(n + 1)^2(n + 2) = O(1/n^2)$ . So  $P(U_{(n-1)}^n < 1/n^{1-\delta/2}) \rightarrow 1$  as  $n \rightarrow \infty$ . This implies that

$$\begin{aligned} P(X_{(1)}^n - X_{(2)}^n \geq n^{1-\delta}) &\geq \frac{1}{1 + n^{1-\delta}n^{-(1-\delta/2)}} P\left(U_{(n-1)}^n < \frac{1}{n^{1-\delta/2}}\right) \\ &= \frac{1}{1 + n^{-\delta/2}} P\left(U_{(n-1)}^n < \frac{1}{n^{1-\delta/2}}\right) \\ &\rightarrow 1 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

**Remark 5.1.** The assumption that  $G$  has a normal distribution in Example 5.4 is easily relaxed. All that is needed in the proof is that the maximum of  $n$  random variables be  $o(n^{1-\delta})$  for any  $\delta > 0$ .

**Remark 5.2.** Note that Example 5.4 establishes (R5). Clearly, for small  $n$ , the value of  $p_n$  here is not equal to 1. But, if  $\lim_{n \rightarrow \infty} p_n = 1$ , the  $p_n$  sequence cannot be monotone.

### Acknowledgements

The authors would like to thank Moshe Pollak for the invaluable insights he provided during the early stages of this research. The authors would also like to thank an anonymous referee for extremely helpful comments.

### References

BRUSS, F. T. (2000). Sum the odds to one and stop. *Ann. Prob.* **28**, 1384–1391.  
 DAVID, H. A. AND NAGARAJA, H. N. (2003). *Order Statistics*, 3rd edn. John Wiley, Hoboken, NJ.  
 FERGUSON, T. S. (2008). *Optimal Stopping and Applications*. Available at <http://www.math.ucla.edu/~tom/Stopping?contents.html>.  
 GILBERT, J. P. AND MOSTELLER, F. (1966). Recognizing the maximum of a sequence. *J. Amer. Statist. Assoc.* **61**, 35–73.  
 GNEDIN, A. V. (2007). Optimal stopping with rank-dependent loss. *J. Appl. Prob.* **44**, 996–1011.  
 LEDFORD, A. W. AND TAWN, J. A. (1998). Concomitant tail behaviour for extremes. *Adv. Appl. Prob.* **30**, 197–215.  
 SAMUELS, S. M. (1991). Secretary problems. In *Handbook of Sequential Analysis* (Textbooks Monogr. **118**), eds B. K. Ghosh and P. K. Sen, Marcel Dekker, New York, pp. 381–405.