Combinatorics, Probability and Computing (2000) 9, 13–17. Printed in the United Kingdom
© 2000 Cambridge University Press

A Note on Sequential Selection from Permutations

ALEXANDER V. GNEDIN

Institute of Mathematical Stochastics, University of Göttingen, Lotzestrasse 13, 37083 Göttingen, Germany
(e-mail: gnedin@math.uni-goettingen.de)

Received 28 April 1998; revised 22 October 1998

The maximum expected length of an increasing subsequence which can be selected by a non-anticipating policy from a random permutation of $1,\dots,n$ is known to be asymptotic to $\sqrt{2n}$. We give a new proof of this fact and demonstrate a policy which achieves this value.

1. Let $x_1,\dots,x_n$ be an unknown random permutation of n integers. The x_js are revealed in strict succession. As soon as x_j is inspected it can be either selected or rejected. If x_j is selected it cannot be discarded later and if rejected cannot be recalled. The objective is to choose a long increasing subsequence of the permutation.

Because any decision is related to some kind of risk, we cannot always select the longest increasing subsequence of $x_1,\dots,x_n$, which has expected length $\sim 2\sqrt{n}$ (see [3] for the history of this result). A question of interest is about the maximum expected length which can be achieved by a selection policy without prior knowledge of the permutation.

This selection problem appeared in Section 3 of [1]. Baer and Brock derived a recursion for the maximum expected length and on the basis of substantial computation conjectured that it is asymptotic to $\sqrt{2n}$. The conjecture was settled by Samuels and Steele [2], who showed first that the $\sqrt{2n}$ asymptotics holds in a more complex problem of selection from a sequence of n independent random points and then used a coupling argument (attributed to Burgess Davis) to establish the asymptotic similarity of the two problems. This $\sqrt{2n}$ result is exciting and worth generalization, so it is highly desirable to have simple proofs. In what follows we give an elementary direct proof based on a dynamic programming equation.

The problem remains essentially the same if we assume that $x_1,\dots,x_n$ is a random permutation of known different numbers, say $y_1 < \cdots < y_n$. Another type of problem appears if the y_js are unknown: because of uncertainty about the numbers we cannot perform as well as in the first case. We shall show, however, that the lack of information is asymptotically negligible in the sense that there is a selection policy which does not use

https://doi.org/10.1017/S0963548399004149 Published online by Cambridge University Press

any prior knowledge of the numbers but still yields an increasing subsequence of expected length asymptotic to $\sqrt{2n}$.

The central subject of the Samuels and Steele paper is an instance of the selection problem with known probability law for $y_1,\dots,y_n$. This situation is intermediate between (a) the setting where no prior information about the numbers is available, and (b) the setting where $y_1,\dots,y_n$ are completely known. Since (a) and (b) are asymptotically equivalent, the $\sqrt{2n}$ result of [2] becomes a consequence.

We will not survey here a recent cascade of papers on the longest increasing subsequence. Many earlier references may be found in [3].

2. We concentrate first on selection from a random permutation of $1,\dots,n$. Denote by v_n the maximum expected length which could be achieved by a selection policy.

Let us pretend that we have just inspected x_1, which turned out to be k. If we reject it and proceed optimally, we select a subsequence of expected length v_{n-1}, because the choice from $x_2,\dots,x_n$ becomes equivalent, after obvious identifications, to the selection from a random permutation of $n-1$ integers. Alternatively, accepting k we get one element for the to-be-constructed subsequence but lose the possibility of selecting $1,\dots,k-1$; thus the expected length by an optimal continuation will be $1+v_{n-k}$. Comparing the outcomes and averaging over k yield an equation

$$v_n = n^{-1}\sum_{k=1}^{n} \max\,(v_{n-1},\, v_{n-k}+1), \tag{1}$$

where $v_0 = 0$. (Baer and Brock used a more complicated recursion for computation.)

For ease of notation substitute n for $n+1$ and write (1) as

$$\Delta v_n = \Sigma v_n, \tag{2}$$

where $\Delta u_n = u_{n+1} - u_n$ is the forward difference operator and

$$\Sigma u_n = (n+1)^{-1}\sum_{k=1}^{n+1}(u_{n-k+1}+1-u_n)^+.$$

Observe the monotonicity: $u_j \leqslant w_j$ for $j \leqslant n$ implies $u_n + \Sigma u_n \leqslant w_n + \Sigma w_n$. Another important property to keep in mind is that adding a constant sequence to $\{u_n\}$ affects neither Δu_n nor Σu_n.

Now test the functions $h_n = \alpha\sqrt{n} + c$ with $\alpha > 0$. Elementary calculus shows that $\Delta h_n \sim \alpha/(2\sqrt{n})$ while $\Sigma h_n \sim 1/(\alpha\sqrt{n})$. Because $\alpha/2$ is increasing and $1/\alpha$ decreasing we have

$$\alpha \bowtie \sqrt{2} \iff \lim_{n\to\infty} \frac{\Delta h_n}{\Sigma h_n} \bowtie 1, \tag{3}$$

where $\bowtie$ is any of the relations $=, <$ or $>$. The match occurs at $\alpha = \sqrt{2}$; thus we are led to believe that this is the right factor. The idea is to compare v_n and h_n with various α and c.

Fix $\alpha > \sqrt{2}$. We claim that $v_n - \alpha\sqrt{n}$ is bounded from above. Assume not: then, for any c, there is an n with $v_{n+1} > \alpha\sqrt{n+1} + c$. Let $n = n(c)$ be the minimum n with this property. Since $n(c)$ goes to infinity with c, by (3) we see that for c sufficiently large

https://doi.org/10.1017/S0963548399004149 Published online by Cambridge University Press

$\Delta h_{n(c)} > \Sigma h_{n(c)}$, which is equivalent to $h_{n(c)+1} > h_{n(c)} + \Sigma h_{n(c)}$. In view of $v_j \leqslant h_j$ for $j \leqslant n(c)$, the monotonicity implies that

$$v_{n(c)} + \Sigma v_{n(c)} \leqslant h_{n(c)} + \Sigma h_{n(c)}.$$

By (2) the left-hand side is equal to $v_{n(c)+1}$, while the right-hand side is less than $h_{n(c)+1}$, which contradicts the choice of $n(c)$.

Setting $\alpha \downarrow \sqrt{2}$ we obtain $\limsup_n v_n/\sqrt{n} \leqslant \sqrt{2}$. The lower bound is derived in the same manner, by taking $\alpha < \sqrt{2}$. Putting the two sides together yields

$$v_n \sim \sqrt{2n}. \tag{4}$$

The numerical results of Baer and Brock demonstrate that the difference of v_n and $\sqrt{2n}$ remains surprisingly small even for moderate n; the last computed value in their Table 2 is $v_{10000} = 142.04\cdots$, which should be compared with $141.42\cdots$.

Noting that v_n is increasing we see that the inequality $v_{n-k} + 1 \geqslant v_{n-1}$ holds for k not exceeding some threshold $k^*(n)$. Thus the optimal policy prescribes selecting the first inspected number if and only if $x_1 \leqslant k^*(n)$. In view of the recursive character of the selection problem, the optimal policy for n numbers is completely determined by thresholds $k^*(j), j \leqslant n$. The asymptotics (4) and the computations of Baer and Brock strongly suggest that $k^*(n)$ is very close to $\sqrt{2n}$ (e.g., $k^*(10000) = 141$). It seems hard, however, to prove this along the lines of the above argument.

In the next section we construct a policy which is only suboptimal, but approaches optimality as n becomes large.

3. Now suppose $x_1, \ldots, x_n$ to be a random permutation of the numbers $y_1 < \cdots < y_n$. The key property of $x_1, \ldots, x_n$ is the exchangeability which says that the probability law of $x_{\pi_1}, \ldots, x_{\pi_n}$ is the same for any fixed permutation of n integers. Even if we have absolutely no idea about the actual values of the numbers, we can still effectively exploit the property that y_j is equally likely to appear in any position of the observed sequence $x_1, \ldots, x_n$.

Set $\delta = \sqrt{2/n}$, $z_0 = 0$. The policy we consider amounts to the following rule. *For $j \geqslant 1$, select x_j if and only if* (i) $x_j > z_{j-1}$ *and* (ii) $\#\{k \leqslant j : z_{j-1} < x_k \leqslant x_j\} < \delta j$. *If x_j is selected, set $z_j = x_j$; otherwise $z_j = z_{j-1}$.*

The variable z_j is the value of the last choice from the first j numbers to appear. While (i) guarantees that the selected sequence is ascending, condition (ii) is imposed to regulate its growth. Note that no selection is made from the first $\lfloor \sqrt{n/2} \rfloor$ observations.

To evaluate the policy we introduce two types of rank. Define the initial rank of x_j by setting $r_j = k$ if x_j is the kth smallest among $x_1, \ldots, x_j$. Similarly, set $a_j = k$ if z_j is the kth smallest among $x_1, \ldots, x_j$, with the convenience $a_0 = 0$.

Alternatively, a_j can be introduced recursively via initial ranks as

$$a_j = a_{j-1} + 1_{G_j} + (r_j - a_{j-1}) 1_{H_j},$$

with events $G_j = \{r_j \leqslant a_{j-1}\}$ and $H_j = \{a_{j-1} < r_j < a_{j-1} + \delta j\}$. This follows from the definition of our policy which, in the language of ranks, says that x_j should be selected if and only if H_j occurs. The total length of the selected sequence is equal to the number of occurrences in the series $H_1, \ldots, H_n$, whence the average length is $\sum_j P(H_j)$.

https://doi.org/10.1017/S0963548399004149 Published online by Cambridge University Press

A well-known consequence of the exchangeability is that the initial ranks are independent, with r_j uniformly distributed on $1,\dots,j$ (this is just another way of saying that $y_1,\dots,y_n$ appear in random order). It follows easily that, for $\delta j \to \infty$,

$$P(H_j|A_j) \to \delta, \tag{5}$$

where $A_j = \{a_{j-1} + \delta j < j\}$. Our plan is to show that the events A_j have probability close to 1.

By the independence, the sequence of random variables $a_j/(j+1), j \leqslant n$, is Markovian and an easy computation shows that it is also a submartingale. Using a standard method write $a_j/(j+1) = m_j + q_j$, where

$$m_j = \sum_{i=1}^{j} \left[\frac{a_i}{i+1} - E\left(\frac{a_i}{i+1} \middle| \frac{a_{i-1}}{i} \right) \right]$$

is a martingale and

$$q_j = \sum_{i=1}^{j} \left[E\left(\frac{a_i}{i+1} \middle| \frac{a_{i-1}}{i} \right) - \frac{a_{i-1}}{i} \right]$$

is increasing in j. Keep in mind that all terms with $i < \sqrt{n/2}$ are trivial.

Conditionally on a_{j-1} and H_j, $r_j - a_{j-1}$ is uniformly distributed on an initial segment of $1,2,\dots,\lfloor \delta j \rfloor$. Using $P(H_j) < \delta$ one can bound from above the variance of each nonzero term in m_j by $9i^{-2} + \delta^3$. Since the terms are uncorrelated, we have $\operatorname{Var} m_j = O(1/\sqrt{n})$. Applying Kolmogorov's inequality we see that

$$P\left(\max_{j \leqslant n} |m_j| > \epsilon/2 \right) \to 0 \text{ as } n \to \infty$$

for any $\epsilon > 0$. Similarly, using independence of ranks we compute

$$E\left(\frac{a_i}{i+1} \middle| \frac{a_{i-1}}{i} \right) - \frac{a_{i-1}}{i} = E\left(\frac{r_i - a_{i-1}}{i+1} 1_{H_j} \middle| \frac{a_{i-1}}{i} \right) < \frac{\delta^2}{2} = \frac{1}{n},$$

whence $q_j < j/n$. Thus, for $n \to \infty$,

$$P\left(a_j/(j+1) < 1 - \epsilon/2, \ \forall\ j < n(1-\epsilon) \right) \to 1,$$

which implies that the probability of A_j goes to 1 uniformly in $j < n(1-\epsilon)$. Using (5) we have, for sufficiently large n,

$$\sum_j P(H_j) = \sum_j P(H_j|A_j)P(A_j) > (1-2\epsilon)\sqrt{2n}.$$

Letting $\epsilon \to 0$ we see that, asymptotically, the expected length of selected subsequences is at least $\sqrt{2n}$. But the expected length cannot exceed v_n, which could be achieved only if $y_1,\dots,y_n$ were known before the observation. It follows that our policy selects on average $\sim \sqrt{2n}$ elements and is therefore asymptotically optimal.

https://doi.org/10.1017/S0963548399004149 Published online by Cambridge University Press

References

[1] Baer, R. M. and Brock, P. (1968) Natural sorting over permutation spaces. *Math. Comput.* **22** 385–410.
[2] Samuels, S. M. and Steele, J. M. (1981). Optimal sequential selection of a monotone sequence from a random sample. *Ann. Prob.* **9** 937–947.
[3] Steele, J. M. (1995). Variations on the monotone subsequence theme of Erdős and Szekeres. In *Discrete Probability and Algorithms* (D. Aldous, *et al.*, eds), Springer, New York.

https://doi.org/10.1017/S0963548399004149 Published online by Cambridge University Press