Analysis of “hiring above the median”: a “Lake Wobegon” strategy for the hiring problem *

Ahmed Helmi † Alois Panholzer ‡

October 7, 2011

Abstract

The hiring problem is a recent research problem, which has been introduced and studied first by Broder et al. [2] in 2008. It belongs to the category of on-line decision making under uncertainty. In such kind of research, the input is a sequence of instances and a decision must be taken for each instance depending on the instances seen so far while no information on the future is available. The hiring problem can be considered as a natural extension of the well-known secretary problem [3], where the employer is now looking for many candidates rather than only one (as it is the case for the secretary problem). Here the goal is to design some hiring strategy to meet the demands of the employer, which essentially are to obtain a good quality staff at a reasonable hiring rate. Broder et al. introduced two so-called “Lake Wobegon strategies”, namely “hiring above the current mean” and “hiring above the current median”, applied for a continuous probabilistic model for the sequence of scores of the candidates.

Archibald and Martínez [1] have reformulated the problem for a discrete model that considers the relative ranks amongst candidates as it is the case in the secretary problem. For this model in [1] the authors studied two general strategies, namely “hiring above the m-th best candidate”, and “hiring above the median” (and other quantiles). In this work we give a detailed study of the “hiring above the median” strategy under this discrete model for the input sequence of scores of the candidates. This strategy processes the sequence of candidates as follows: hire the first interviewed candidate, and then any coming candidate is hired if and only if his rank is better than the rank of the median of the already hired staff, and discarded otherwise.

Compared to the previous work of [1] we use a somewhat different recursive approach for a study of the “hiring above the median” strategy leading to rather explicit results. The key ingredients are to take into account the score of the median (the so-called decision maker) of the hired staff during the hiring process as well as to distinguish between two cases according to the parity of the size of the hiring set. Considering the transition probabilities during the hiring process yields, for fundamental hiring quantities, a system of linear recurrences that can be translated into a system of partial differential equations for the corresponding generating functions. In order to solve the PDEs appearing it turned out to be crucial to find suitable normalization factors of the studied recursive sequences, such that one of the corresponding generating functions itself reduces to a first order linear PDE.

The exact solutions obtained for the differential equations yield to a rather detailed description of the exact probability distributions together with limiting distribution results for various hiring quantities, which might lead to a fairly good understanding of the “hiring above the median” hiring process. In particular we obtained results for the number of hired candidates, the score of the last hired candidate, the index of the last hired candidate, the distance between the last two hirings, the score of the best discarded candidate, and the number of hired candidates conditioned on the score of the first candidate. We also give the probability that a given score is getting hired in a sequence of n candidates.

1 Introduction

The hiring problem Problems related to on-line decision making under uncertainty arise in various contexts, e.g., in computer science and economics, and thus mak-

*This work started when the first author was visiting the second author in a short stay supported by an FPI grant from the Spanish Ministry of Science and Innovation. The research of the first author was supported by a project TIN2010-17254 (FRADA) from the Spanish Ministry of Science and Innovation. The second author was supported by the Austrian Science Foundation FWF, grant S9608-N23.

†Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, Jordi Girona, 1-3, E-08034 Barcelona, Spain. ahelmi-at-lsi.upc.edu.

‡Institut für Diskrete Mathematik und Geometrie. Technische Universität Wien, Wiedner Hauptstr, 8-10/104, A-1040 Wien, Austria. Alois.Panholzer-at-tuwien.ac.at.

ing this subject an active research discipline. A specific simple model for on-line decision making under uncertainty has been introduced by Broder et al. [2] called the "hiring problem". Here, in analogy to the famous "secretary problem", see, e.g., [3], a sequence of candidates is interviewed sequentially and for each candidate one has to make an immediate decision whether to hire him or not based on the subsequence examined so far, while nothing is known about the future. However, unlike in the secretary problem, the number of candidates is not known in advance (and might be infinite) and the goal for the employer is different: the aim is not to hire one candidate, namely the best of the whole sequence, but, informally speaking, rather to hire many candidates of a reasonable good quality at a reasonable hiring rate (i.e., the ratio between the number of hired candidates and the number of interviewed candidates). Any hiring strategy has to meet the following common rules: a decision must be taken whether to hire the i-th candidate or not at step i, decisions are irrevocable, and there is no information on the future.

In the probabilistic model introduced and studied in [2] it is assumed that the absolute quality scores of the candidates are given by independent identically distributed random variables (r.v.), which are uniformly distribution on $[0, 1]$. Two natural strategies for recruiting candidates which are called "Lake Wobegon strategies"[1] have been considered; they have the common feature that a new candidate is recruited only if he is better (i.e., if he has a larger score) than the "average employee", where for the score of the average employee they either refer to the mean or to the median of the scores of the already hired staff. Using this continuous model the authors studied, for both of these hiring strategies called "hiring above the mean" and "hiring above the median", as main hiring parameters the number of interviews required to hire n candidates (yielding thus the hiring rate) as well as a quality measure for the hired staff called the gap of the n-th employee, which is defined by the difference between the maximum possible score (i.e., 1) and the score of the n-th hired candidate.

Whereas decision making for the strategy "hiring above the mean" crucially depends on the absolute values of the scores this is not the case for the strategy "hiring above the median", where only the relative ranks of the candidates are important (as it is also the setting in the secretary problem). For such "rank-based hiring strategies" Archibald and Martínez [1] introduced a natural combinatorial (discrete) model for the distribution of the ranks of the sequence of candidates, which will be also the underlying probability model for all of the studies in this work. Basically, in this model each permutation σ_n of $\{1, 2, \dots, n\}$ appears with equal probability for the ranks of a sequence of n candidates, where we assume that the worst-ranked (i.e., rank 1) candidate gets value 1 and the best-ranked (i.e., rank n) candidate value n; thus we use the term "random permutation model". However, since we might deal also with infinite sequences of candidates, the following definition, which matches with the preceding one for finite sequences, is advantageous.

As mentioned we assume that the sequence of candidates may be infinite and that we can rank candidates from best to worst without ties. So we start giving the first interviewed candidate a rank 1 while we assume that at step i all values from 1 (worst) to i (best) appear equally likely (and are independent of the choices of the other candidates) as relative rank of the i-th candidate. More formally, the input is a sequence of relative scores $S = s_1, s_2, \dots, s_i, \dots$, with $1 \le s_i \le i$, of the candidates. For a candidate with score s_i, exactly $s_i - 1$ previous candidates rank worse than that candidate. The relative score s_i of the i-th candidate is uniformly distributed on $\{1, 2, \dots, i\}$ and independent of $s_j, j \neq i$. Then each finite subsequence of candidates represents a random permutation σ_n of $\{1, 2, \dots, n\}$. More precisely, σ_n can be obtained recursively as follows: given a permutation σ_{n-1} (of size $n-1$) and a value (relative rank) j, $1 \le j \le n$, $\sigma_n = \sigma_{n-1} \circ j$ denotes the resulting permutation after relabelling $j, j+1, \dots, n-1$ in σ_{n-1} as $j+1, j+2, \dots, n$, and appending j to the end. For example, let $S_8 = 1, 2, 1, 4, 1, 5, 4, 6$ represent the input sequence of relative ranks of the candidates. Then $\sigma_1 = 1$, $\sigma_2 = \sigma_1 \circ 2 = 12$, $\sigma_3 = \sigma_2 \circ 1 = 231$ and so on until $\sigma_8 = 35281746$.

The strategy "hiring above the median" In the "Lake Wobegon" strategy "hiring above the median" the first candidate in the sequence is always recruited and, after that, a new candidate is recruited exactly if his quality score is better than the median score of the staff hired so far. We use here the convention (which is then in accordance with [2]) that the median of a set of k (distinct) elements $r_1 < r_2 < \dots < r_k$ is the ℓ-th largest element, i.e., $r_{k+1-\ell}$, with $\ell = \lceil \frac{k+1}{2} \rceil$. As an example, if we apply the "hiring above the median" strategy to the sequence of (absolute) scores $\sigma = \underline{3}\,\underline{5}\,2\,\underline{8}\,1\,\underline{7}\,4\,\underline{6}$ the hiring set (i.e., the set of hired candidates) consists of $\{3, 5, 6, 7, 8\}$ (which are underlined in σ), since each of these candidates has, at the time of hiring, a score better than the median of the previously hired candidates.

[1] As pointed out in [2], Lake Wobegon is a fictional town, where "the women are strong, the men are good looking, and all the children are above average". The considered strategies match this term in the sense that each recruited candidate, at least at the time when he is hired, is above "average".

The number h_n of hired candidates, i.e., the size of the hiring set, of a sequence of n candidates for the random permutation model using the strategy "hiring above the median" has been studied in [1]; in fact, they even consider more general strategies called "hiring above the α-quantile", where a new candidate is only recruited if he has a score larger than the α-quantile, $0 < \alpha < 1$, of the scores of the already hired staff. For all of these strategies they were able to determine the exact growth order of the r-th integer moments of h_n:

$$\mathbb{E}(h_n^r) = \Theta(n^{\alpha r}), \quad r \in \mathbb{N}.$$

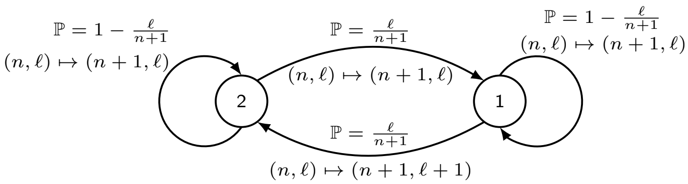
In particular, when specializing to the "hiring above the median" strategy, i.e., $\alpha = \frac{1}{2}$, they get $\mathbb{E}(h_n^r) = \Theta(n^{\frac{r}{2}})$.

As is somehow inherent in the problem and already observed in [1, 2] there is always a trade-off between the hiring rate and the quality of the recruited staff (i.e., the more candidates are hired, the worse is the staff quality and vice versa). Thus, besides the fundamental quantity h_n describing the size of the hiring set, it is also of interest to study quantities measuring the quality of the hired staff as well as quantities which help to understand the dynamics of the hiring process. One parameter yielding such a quality measure is the gap g_n, which has been defined in [1] for the discrete model as $1 - \frac{R_n}{n}$, where R_n denotes the rank of the last hired candidate in a sequence of n candidates. Again, for "hiring above the α-quantile", Archibald and Martínez obtained in [1] the exact growth order for the expectation of g_n: $\mathbb{E}(g_n) = \Theta(n^{\alpha-1})$.

It is the aim of this paper to give a detailed study of the "hiring above the median" strategy for a sequence of candidates following the random permutation model. We do this by first giving a distributional analysis of the number h_n of hired candidates, characterizing the exact and limiting distribution, which, as a consequence, also yields exact and asymptotic results for a couple of further quantities of interest when describing the dynamics of the hiring process as the "index of the last hired candidate", the "distance between the last two hirings", and the "probability of getting hired for the n-th candidate". Furthermore, we introduce and study quality measures as the "score of the last hired candidate", the "score of the best discarded candidate", and the "probability of getting hired depending on the score of the candidate", for which we present exact and (in most cases) asymptotic results. A detailed description of the quantities studied and a presentation of the results is given in the following section.

A recursive approach to treat "hiring above the median" Archibald and Martínez introduced a systematic approach using analytic combinatorics techniques to analyze quantitative properties of hiring strategies, which they applied to the strategies "hiring above the α-quantile" as well as to another strategy called "hiring above the m-th best". Using this approach, a direct treatment of quantities for the strategy "hiring above the median" turns out to be involved (mainly due to the appearance of the ceiling-function in the definition of the median), and the authors sticked to compute upper and lower bounds leading to the results mentioned before. To resolve this problem and to carry out a distributional analysis of the quantities of interest we use a slightly different approach, where we take care of the rank of the median of the hiring set during the hiring process. During the process, the median of the current hiring set could be considered as the "decision maker", who actually makes the decision, whether a new candidate is recruited (if he has a score larger than the decision maker) or not (otherwise). It is a simple but quite useful observation that, when applying the strategy "hiring above the median" (note that this property even holds for the "hiring above the α-quantile" and also for the "hiring above the m-th best" strategy), at each time of the hiring process all candidates seen so far with a score larger than the current decision maker are contained in the hiring set (i.e., have been recruited). Thus there is a simple relation between the score of the decision maker and the size of the hiring set. Let us assume that in a sequence of n candidates k candidates are eventually recruited and let us further assume that the decision maker has the ℓ-th largest score amongst all candidates in this sequence. It follows then that $\ell = \frac{k+1}{2}$ if k is odd and $\ell = \frac{k}{2}+1$ if ℓ is even, i.e., $\ell = \lfloor \frac{k+2}{2} \rfloor$. And this yields the basis of the recursive approach, where we thus have to distinguish according to the parity of the size of the hiring set and to take into account the score of the decision maker.

Many of the quantities considered here can be expressed by the sequences of numbers $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$, which give the probabilities that, after interviewing n candidates, the decision maker has the ℓ-th largest score amongst all candidates and an odd or even number of candidates has been hired, respectively. Moreover, the remaining quantities are obtained by studying extensions of them to take into account further parameters. The following "automaton" describes then the "transition probabilities", when at time $n+1$ a new candidate appears, whose score will be compared with the score of the decision maker, which is the ℓ-th largest amongst the first n candidates (please note that, of course, the hiring process is not described by a finite two-state automaton, but it shall give a simplified picture of the underlying two-type Markov chain); here states 1 and 2 correspond to an odd and even number of hired candidates, respectively.



For the fundamental quantities studied in this work this naturally leads to systems of two double- or triple-indexed linear recurrences, which can be "translated" into systems of linear partial differential equations (PDEs) for the corresponding generating functions. However, here a second aspect of the present approach comes into play: a direct treatment of the recurrences obtained via generating functions always leads to pairs of first order PDEs, which then yield second order linear PDEs for each of the generating functions corresponding to an odd or even number of hired candidates, respectively. Since it seems quite involved to get the desired solutions of these second order PDEs in a systematic way we used a "trick" similar to one applied (in a slightly different context) in a PDE approach for the study of diminishing urn models in [5]. Namely, we were successful in finding suitable normalization factors of the studied recursive sequences, such that one of the corresponding generating functions itself reduces to a first order linear PDE (or even to an ordinary differential equation). The explicit solutions of these differential equations also lead to explicit results for the exact distribution of the fundamental quantities considered, from which the limiting distribution results can be obtained in a rather straightforward way.

2 Hiring quantities and results

2.1 Definition of hiring quantities We give here a precise description of the parameters studied for the strategy "hiring above the median" using the random permutation model for the input sequence of candidates. As already mentioned the basic quantity is the r.v. h_n, which gives the number of hired candidates (i.e., the size of the hiring set) for an input sequence of n candidates thus yielding the hiring rate. Obviously closely related to this quantity is the probability p_n that the n-th candidate in a sequence of candidates is getting hired.

Two quantities, which shall give some insight into the dynamics of the hiring process, are the r.v. L_n giving the index of the last hired candidate and the r.v. Δ_n giving the distance (i.e., difference) between the indices of the last two recruited candidates in a sequence $x_1, x_2, \dots, x_n$ of absolute scores (ranks) of n candidates. Thus, $L_n = i$ if x_i is recruited, but x_j, for all $j > i$ aren't. Also, if there are at least two recruited candidates, then $\Delta_n = d$ if x_{i-d} and x_i are recruited, but x_j, $i-d < j < i$ or $j > i$ aren't; moreover, if there is only one recruited candidate (which happens if the one with largest score appears at first position), then we define the distance Δ_n as 0. It has been noticed in [2] that the hiring process is quite sensitive to the score of the first candidate (and thus the first hired candidate) in the sequence. To get a quantitative result in this direction we also study the r.v. $h_{n,q}$, which gives the number of hired candidates h_n conditioned on the event that the candidate with score q (i.e., the q-th smallest candidate) in a sequence of n candidates appears at first position. Of course, if we denote by U_n the absolute score of the first candidate, it holds $h_{n,q} = h_n | U_n = q$, where, for the considered random permutation model, U_n is uniformly distributed on $\{1, 2, \dots, n\}$.

Let us now define the quantities concerning quality measures of the hired staff in a sequence $x_1, x_2, \dots, x_n$ of scores of n candidates. The r.v. R_n gives, for a sequence of n candidates, the score of the last hired candidate, i.e., $R_n = x_i$ if x_i is the last candidate which is recruited. Directly related is the gap $g_n = 1 - \frac{R_n}{n}$, which has been introduced in [1] in order to measure how close the quality of the last hired candidate is compared to the topmost one; thus when presenting our results we entirely stick to R_n. The r.v. M_n gives, for a sequence of n candidates, the score of the best discarded candidate, i.e., the maximum score amongst all candidates which are not contained in the hiring set. In a sense, this quantity describes how selective the hiring process is and thus yielding also a quality measure for the hired staff. Furthermore, we introduce the probabilities $p_{n,q}$, which give the probability that the candidate with score q (i.e., the one with rank q and thus the q-th smallest) in a sequence of n candidates is getting hired. Note that trivially $\frac{1}{n} = p_{n,1} \le p_{n,2} \le \cdots \le p_{n,n} = 1$ by considering the probabilities conditioned on the event that the candidate with rank q appears at first position, second position, etc. in the sequence. Note that these probabilities can also be used to give a first result for the r.v. S_n measuring the total score of the hiring set, i.e., the sum of the scores of the elements in the hiring set (a quantity, which seems difficult to treat directly via the proposed recursive approach), since $\mathbb{E}(S_n) = \sum_{q=1}^{n} q \cdot p_{n,q}$; however, since so far we didn't evaluate the triple sum appearing for this expression asymptotically, here we do not state results for this quantity apart from the trivial bound $\mathbb{E}(S_n) = \Theta(n^{\frac{3}{2}})$.

We illustrate these quantities on the example sequence $\sigma = \underline{3}\,\underline{5}\,2\,\underline{8}\,1\,\underline{7}\,4\,\underline{6}$ of scores of the candidates, for which we get $h = 5$, $L = 8$, $\Delta = 2$, $R = 6$, $M = 4$ and

$S = 29$.

2.2 Results for “hiring above the median”

Theorem 1. *Let h_n denote the number of hired candidates for n candidates and using the strategy “hiring above the median”. Then the exact distribution of h_n is given as follows, with $1 \le k \le n$:*

$$\mathbb{P}\{h_n = k\} = \frac{\binom{n-1-\lfloor \frac{k}{2} \rfloor}{\lceil \frac{k}{2} \rceil - 1}}{\binom{n}{\lceil \frac{k}{2} \rceil}} = \begin{cases} \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}}, & \text{for } k = 2\ell - 1 \text{ odd}, \\ \frac{\binom{n-\ell}{\ell-2}}{\binom{n}{\ell-1}}, & \text{for } k = 2\ell - 2 \text{ even}. \end{cases}$$

After normalization, the r.v. h_n is asymptotically, for $n \to \infty$, Rayleigh distributed with parameter $\sigma = \sqrt{2}$, i.e., $\frac{h_n}{\sqrt{n}} \xrightarrow{(d)} \hat{R}$, where $\hat{R}$ has the density function

$$\hat{f}(x) = \frac{x}{2} e^{-\frac{x^2}{4}}, \quad \text{for } x > 0.$$

Furthermore, the expectation of h_n satisfies: $\mathbb{E}(h_n) = \sqrt{\pi}\sqrt{n} + O(1)$.

Corollary 1. *Let p_n denote the probability that the n-th interviewed candidate is getting hired by using the strategy “hiring above the median”. Then the probability p_n is given by the following exact and asymptotic formula:*

$$p_n = \sum_{\ell=1}^{n-1} \frac{(2\ell-1)n + \ell(2-3\ell)}{(n-\ell)^2} \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}} = \frac{\sqrt{\pi}}{2\sqrt{n}} \cdot \left(1 + O(n^{-\frac{1}{2}})\right).$$

Theorem 2. *Let L_n denote the index of the last hired candidate for n candidates and using the strategy “hiring above the median”. Then the exact distribution of L_n is given as follows: for $1 < m \le n$,*

$$\mathbb{P}\{L_n = m\} = \sum_{\ell=1}^{m-1} \frac{\binom{m-1-\ell}{\ell-2}}{\binom{n}{\ell}} + \sum_{\ell=1}^{m} \frac{\ell-1}{\ell} \frac{\binom{m-\ell}{\ell-2}}{\binom{n}{\ell}},$$

and $\mathbb{P}\{L_n = 1\} = \frac{1}{n}$.

Asymptotically, for $n \to \infty$, the r.v. L_n is distributed as follows: $\frac{n-L_n}{\sqrt{n}} \xrightarrow{(d)} X$, where the density function $f(x)$ of X is given by

$$f(x) = 2\int_0^\infty t^2 e^{-t(x+t)} dt, \quad x > 0.$$

(Note that the moments $\mathbb{E}(X^r)$ of X do not exist for $r \ge 2$.) Furthermore, the expectation of L_n satisfies: $\mathbb{E}(L_n) = n - \sqrt{\pi}\sqrt{n} + O(\log n)$.

Theorem 3. *Let Δ_n denote the distance between the last two hirings for n candidates and using the strategy “hiring above the median”. Then the exact distribution of Δ_n is given as follows:*

$$\mathbb{P}\{\Delta_n = d\} = \begin{cases} \frac{1}{n}, & d = 0, \\ \sum_{m=1}^{n-d-1} \sum_{\ell=1}^{m} \frac{\binom{m-\ell}{\ell-1}}{\binom{n}{\ell+1}} \cdot \frac{\ell}{m+d-\ell} \\ + \sum_{m=1}^{n-d-1} \sum_{\ell=1}^{m} \frac{\binom{m-\ell}{\ell-2}}{\binom{n}{\ell+1}} \cdot \frac{\ell}{\ell+1} \\ + \frac{1}{n(n-1)}, & 1 \le d \le n-1. \end{cases}$$

Asymptotically, for $n \to \infty$, the r.v. Δ_n is distributed as follows: $\frac{\Delta_n}{\sqrt{n}} \xrightarrow{(d)} X$, where the density function $f(x)$ of X is given by

$$f(x) = 2\int_0^\infty t^2 e^{-t(x+t)} dt, \quad x > 0.$$

Theorem 4. *Let $h_{n,q} := h_n | U_n = q$ denote the number of hired candidates for n candidates and using the strategy “hiring above the median” conditioned on the event that the score U_n of the first candidate is q. Then the exact distribution of $h_{n,q}$ is given as follows, with $1 \le k, q \le n$:*

$$\mathbb{P}\{h_{n,q} = k\} = \begin{cases} (\ell-1) \frac{\binom{n-q-\ell}{\ell-2}}{\binom{n-q}{\ell}}, & \text{for } k = 2\ell - 1 \text{ odd}, \\ \frac{(\ell-1)(\ell-2)}{\ell} \frac{\binom{n-q+1-\ell}{\ell-2}}{\binom{n-q}{\ell}}, & \text{for } k = 2\ell - 2 \text{ even}. \end{cases}$$

The r.v. $h_{n,q}$ converges, for $n \to \infty$ and provided that $n - q \to \infty$, to a limiting distribution Q, i.e., $\frac{h_{n,q}}{\sqrt{n-q}} \xrightarrow{(d)} Q$, where Q has the density function

$$h(x) = \frac{x^3}{8} e^{-\frac{x^2}{4}}, \quad \text{for } x > 0.$$

Theorem 5. *Let R_n denote the score of the last hired candidate for n candidates and using the strategy “hiring above the median”. Then the exact distribution of R_n is given as follows:*

$$\mathbb{P}\{R_n = r\} = \sum_{\ell=n+1-r}^{n-1} \frac{\binom{n-\ell}{\ell}}{\binom{n}{\ell+1}} \frac{1}{\ell+1} + \sum_{\ell=n+1-r}^{n-1} \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}} \frac{1}{\ell}, \quad \text{for } 1 \le r \le n.$$

Asymptotically, for $n \to \infty$, the r.v. R_n is distributed as follows: $\frac{n-R_n}{\sqrt{n}} \xrightarrow{(d)} Y$, where the density function $g(x)$ of Y is given by

$$g(x) = 2\int_0^\infty e^{-(x+t)^2} dt, \quad x > 0.$$

Theorem 6. *Let M_n denote the score of the best discarded candidate for n candidates and using the strategy "hiring above the median". Then the exact distribution of M_n is given as follows, for $1 \le r \le n-1$:*

$$\mathbb{P}\{M_n = r\} = \sum_{\ell=1}^{n-r} \frac{\binom{\ell-1}{n-\ell-r}\binom{2n-2\ell-r}{n-r-1}}{\binom{n-r}{\ell}\binom{n}{r}} \cdot \left(1 + \frac{(n-2\ell-r+1)(2n-2\ell-r+1)}{r(n-r)}\right) + \sum_{\ell=1}^{n-r} \frac{\binom{\ell-1}{n-\ell-r+1}\binom{2n-2\ell-r+1}{n-r-2}}{\binom{n-r}{\ell-1}\binom{n}{r}} \cdot \left(1 + \frac{(n-2\ell-r+2)(2n-2\ell-r+2)}{r(n-r)}\right),$$

and further $\mathbb{P}\{M_n = 0\} = \frac{1}{\binom{n}{\lceil \frac{n}{2} \rceil}}$. After normalization, the r.v. M_n is asymptotically, for $n \to \infty$, Rayleigh distributed with parameter $\sigma = 1/\sqrt{2}$, i.e., $\frac{n-M_n}{\sqrt{n}} \xrightarrow{(d)} \tilde{R}$, where $\tilde{R}$ has the density function

$$\tilde{f}(x) = 2xe^{-x^2}, \quad \text{for } x > 0.$$

Theorem 7. *Let $p_{n,q}$ denote the probability that the candidate with score q of n candidates using the strategy "hiring above the median" is getting hired. Then the probabilities $p_{n,q}$ are given as follows, with $1 \le q \le n$:*

$$p_{n,q} = \sum_{\ell=1}^{n-q} \left[\frac{(\ell-1)}{n\binom{n-1}{\ell}\binom{n-\ell-1}{n-\ell-q}} \cdot \sum_{k=0}^{\ell-2} \binom{\ell-2}{k}\binom{n-\ell-q+k}{\ell-2}\binom{n-\ell+1}{n-\ell-q+k+2} + \frac{(\ell-1)}{n\binom{n-1}{\ell-1}\binom{n-\ell}{n-\ell-q+1}} \cdot \sum_{k=0}^{\ell-2} \binom{\ell-2}{k}\binom{n-\ell-q+k}{\ell-3}\binom{n-\ell+1}{n-\ell-q+k+2} \right] + \sum_{\ell=n-q+1}^{n} \left[\frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}} + \frac{\binom{n-\ell}{\ell-2}}{\binom{n}{\ell-1}} \right].$$

3 Comments on the proof of the results

It is here only possible to very briefly sketch the proof of the results stated in Subsection 2.2. We will focus there on showing how to derive the explicit results characterizing the exact probability distributions of the quantities considered, since, due to the explicit nature of these exact formulas, the stated asymptotic results follow from them essentially by applying Stirling's formula for the factorials in connection with standard techniques as Euler's summation formula and the Mellin-transform.

3.1 Proof of Theorem 1

As follows from the remarks given in Section 1, the there defined sequences $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$ fully determine the probability distribution of h_n according to

$$\mathbb{P}\{h_n = k\} = \begin{cases} a_{n,\frac{k+1}{2}}^{[1]}, & k \text{ odd}, \\ a_{n,\frac{k}{2}+1}^{[2]}, & k \text{ even}. \end{cases}$$

From the description of the hiring process via the transition probabilities of the automaton given in Section 1 the following recurrences for the probabilities $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$ are deduced immediately (with initial values $a_{1,1}^{[1]} = 1$ and $a_{1,1}^{[2]} = 0$ and where we define $a_{n,\ell}^{[1]} = a_{n,\ell}^{[2]} = 0$ outside the range $1 \le \ell \le n$):

$$a_{n,\ell}^{[1]} = \frac{\ell}{n} \cdot a_{n-1,\ell}^{[2]} + \left(1 - \frac{\ell}{n}\right) \cdot a_{n-1,\ell}^{[1]}, \quad n \ge 2,\ 1 \le \ell \le n,$$

$$a_{n,\ell}^{[2]} = \frac{\ell-1}{n} \cdot a_{n-1,\ell-1}^{[1]} + \left(1 - \frac{\ell}{n}\right) \cdot a_{n-1,\ell}^{[2]}, \quad n \ge 2,\ 1 \le \ell \le n.$$

In order to treat this system of recurrences we introduce suitable generating functions using the normalization factor $\binom{n}{\ell}$, which reduces the occurring system of linear PDEs to a single first order linear PDE (instead of second order without normalization); i.e., we consider $b_{n,\ell}^{[\cdot]} := \binom{n}{\ell} a_{n,\ell}^{[\cdot]}$ and the corresponding generating functions $B^{[\cdot]}(z,u) := \sum_{n \ge 1} \sum_{1 \le \ell \le n} b_{n,\ell}^{[\cdot]} z^n u^\ell$. This yields the following linear first order PDE for $B^{[1]}(z,u)$:

$$z(1-z)B_z^{[1]}(z,u) + \left(zu - u - \frac{u^2z^2}{1-z}\right) B_u^{[1]}(z,u) - zB^{[1]}(z,u) = 0,$$

which, after adapting the general solution to the initial conditions (which for PDEs is not always trivial) leads to the following astonishing simple solution of the problem:

$$B^{[1]}(z,u) = \frac{zu}{1-z-z^2u}.$$

Extracting coefficients via $a_{n,\ell}^{[1]} = \frac{1}{\binom{n}{\ell}}[z^n u^\ell]B^{[1]}(z,u)$ and then expressing $a_{n,\ell}^{[2]}$ via the above recursive description yields to explicit results for the required probabilities:

$$a_{n,\ell}^{[1]} = \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}}, \quad \text{and}$$

$$a_{n,\ell}^{[2]} = \frac{n+1}{\ell}\left(a_{n+1,\ell}^{[1]} - \left(1 - \frac{\ell}{n+1}\right) a_{n,\ell}^{[1]}\right) = \frac{\binom{n-\ell}{\ell-2}}{\binom{n}{\ell-1}}.$$

This leads to the exact probability distribution of h_n as stated in the theorem.

The corresponding limiting distribution result follows immediately from the local expansions (obtained by applying Stirling's formula)

$$a_{n,\ell}^{[1]} \sim a_{n,\ell}^{[2]} = \frac{\ell}{n} e^{-\frac{\ell^2}{n}} \cdot \left(1 + O\big(\frac{1}{\ell}\big) + O\big(\frac{\ell}{n}\big) + O\big(\frac{\ell^3}{n^2}\big)\right),$$

uniformly for $1 \le \ell \le n^{\frac{1}{2}+\epsilon}$,

whereas these numbers are exponentially small for $\ell \ge n^{\frac{1}{2}+\epsilon}$, yielding thus $\mathbb{P}\{h_n = k\} \sim \frac{k}{2n} e^{-\frac{k^2}{4n}}$ (valid, in particular, for $k \in [n^{\frac{1}{2}-\epsilon}, n^{\frac{1}{2}+\epsilon}]$).

3.2 Consequences of Theorem 1 The computations and results for the sequences $a_{n,\ell}^{[\cdot]}$ given in Subsection 3.1 lead also to results for the quantities p_n, L_n, Δ_n, R_n, and $h_{n,q}$ (although some of these derivations are a bit less immediate), thus showing the corresponding theorems stated in Section 2.2; we just comment very briefly on two of these parameters.

The probability that the m-th interviewed candidate is the last one hired satisfies:

$$\begin{aligned}\mathbb{P}\{L_n = m\} &= \mathbb{P}\{\text{We hire at position } m\} \\ &\quad \cdot \mathbb{P}\{\text{No hirings from position } (m+1) \text{ till } n\} \\ &= \sum_{\ell=1}^{m} a_{m-1,\ell-1}^{[1]} \frac{\ell-1}{m} \prod_{j=m}^{n-1} \left(1 - \frac{\ell}{j+1}\right) \\ &\quad + \sum_{\ell=1}^{m-1} a_{m-1,\ell}^{[2]} \frac{\ell}{m} \prod_{j=m}^{n-1} \left(1 - \frac{\ell}{j+1}\right),\end{aligned}$$

for $1 < m \le n$, whereas $\mathbb{P}\{L_n = 1\} = \frac{1}{n}$, yielding the exact result stated in the theorem. The limiting distribution result follows then by setting $m := n-k$ and the local approximation (obtained by Stirling's formula) valid for $k = O(n^{\frac{1}{2}+\epsilon})$:

$$\mathbb{P}\{L_n = n-k\} \sim 2 \sum_{\ell \ge 1} \frac{\ell^2}{n^2} e^{-\frac{\ell(k+\ell)}{n}}.$$

The stated asymptotic result for the expectation $\mathbb{E}(L_n) = \sum_m m\,\mathbb{P}\{L_n = m\}$ follows from a more detailed study of the occurring sums applying basic Mellin-transform techniques, i.e., using the representation $\sum_{\ell \ge 1} \ell^j e^{-\frac{\ell^2}{n}} = \frac{1}{2\pi i} \int_{C-i\infty}^{C+i\infty} \Gamma(s) n^s \zeta(2s-j) ds$, with $C > \frac{j+1}{2}$, for $j \ge -1$.

Furthermore, the probabilities p_n can be obtained easily due to the relation (there are ℓ possibilities (out of n) for the n-th candidate being hired, if the decision maker has the ℓ-th largest score after $n-1$ interviews):

$$p_n = \sum_{\ell=1}^{n-1} \left(\frac{\ell}{n} a_{n-1,\ell}^{[1]} + \frac{\ell}{n} a_{n-1,\ell}^{[2]}\right), \quad n \ge 2, \quad p_1 = 1.$$

3.3 Proof of Theorem 6 A direct recursive study of M_n seems to be involved (e.g., a PDE approach leads to equations, where the unknown boundary values explicitly appear). We resolve the problem by considering auxiliary quantities, namely $\hat{a}_{n,\ell,q}^{[1]}$ and $\hat{a}_{n,\ell,q}^{[2]}$, $0 \le q \le n-\ell$, which give the probabilities that, for n interviewed candidates, the decision maker has the ℓ-th largest score in the sequence, that an odd or even number of candidates, respectively, has been recruited and that all of the $\ell+q$ highest ranked candidates are hired (and maybe others). But the probability that the best discarded candidate has rank r is simply given by the difference between the probability that all candidates with a rank higher than r are recruited and the probability that all candidates with a rank higher than $r-1$ are recruited, which yields the following relation eventually characterizing the distribution of M_n:

$$\begin{aligned}\mathbb{P}\{M_n = r\} &= \sum_{\ell=1}^{n-r} \left(\hat{a}_{n,\ell,n-\ell-r}^{[1]} - \hat{a}_{n,\ell,n-\ell-r+1}^{[1]}\right) \\ &\quad + \sum_{\ell=1}^{n-r} \left(\hat{a}_{n,\ell,n-\ell-r}^{[2]} - \hat{a}_{n,\ell,n-\ell-r+1}^{[2]}\right).\end{aligned}$$

Of course, $\hat{a}_{n,\ell,0}^{[\cdot]} = a_{n,\ell}^{[\cdot]}$, where the latter numbers are defined in Section 1, since the ℓ highest ranked candidates are always hired if the decision maker has the ℓ-th largest score. By an extension of the automaton for the transition probabilities as given in Section 1 one gets that the quantities $\hat{a}_{n,\ell,q}^{[\cdot]}$ satisfy (for $n \ge 2$, $1 \le \ell \le n$ and $1 \le q \le n-\ell$) the following system of recurrences:

$$\begin{aligned}\hat{a}_{n,\ell,q}^{[1]} &= \frac{\ell}{n} \cdot \hat{a}_{n-1,\ell,q-1}^{[2]} + \left(1 - \frac{\ell+q}{n}\right) \cdot \hat{a}_{n-1,\ell,q}^{[1]}, \\ \hat{a}_{n,\ell,q}^{[2]} &= \frac{\ell-1}{n} \cdot \hat{a}_{n-1,\ell-1,q}^{[1]} + \left(1 - \frac{\ell+q}{n}\right) \cdot \hat{a}_{n-1,\ell,q}^{[2]}.\end{aligned}$$

To treat these recurrences it turns out that the normalization factors $\frac{n!}{\ell!(n-q-\ell)!}$ are appropriate and lead to a reduction of the system of PDEs for the corresponding generating functions; one even gets the following ordinary differential equation:

$$\hat{B}_u^{[1]}(z,u,v) - \frac{(1-z)^2}{z^2u^2v} \hat{B}^{[1]}(z,u,v) + \frac{z(1-z)}{(1-z-z^2u)^2} = 0$$

for $\hat{B}^{[\cdot]}(z,u,v) := \sum_{n \ge 1} \sum_{1 \le \ell \le n} \sum_{1 \le q \le n-\ell} \hat{b}_{n,\ell,q}^{[\cdot]} z^n u^\ell v^q$ with $\hat{b}_{n,\ell,q}^{[\cdot]} := \frac{n!}{\ell!(n-\ell-q)!} \hat{a}_{n,\ell,q}^{[\cdot]}$.

Extracting coefficients from the proper solution of this differential equation eventually leads to the following explicit results for the required probabilities:

$$\hat{a}_{n,\ell,q}^{[1]} = \frac{\binom{\ell-1}{q}\binom{n-\ell+q}{\ell+q-1}}{\binom{n}{q}\binom{n-q}{\ell}}, \quad \text{and}$$

$$\hat{a}^{[2]}_{n,\ell,q} = \frac{\binom{\ell-1}{q+1}\binom{n-\ell+q+1}{\ell+q-1}}{\binom{n}{q+1}\binom{n-q-1}{\ell-1}}, \quad 0 \le q \le n-\ell,$$

which show the result for the distribution of M_n stated in the theorem.

3.4 Proof of Theorem 7 It suffices to study the numbers $a^{[1]}_{n,\ell,q}$ and $a^{[2]}_{n,\ell,q}$, $0 \le q \le n-\ell$, which give the probabilities that, for n interviewed candidates, the decision maker has the ℓ-th largest score in the sequence, that an odd or even number of candidates, respectively, has been recruited and that the candidate with the $(\ell+q)$-th largest score amongst all candidates is hired. The probabilities $p_{n,q}$ can be obtained from these numbers via the following relation (with $a^{[\cdot]}_{n,\ell}$ the sequences defined in Section 1):

$$p_{n,q} = \sum_{\ell=1}^{n-q}(a^{[1]}_{n,\ell,n-\ell-q+1} + a^{[2]}_{n,\ell,n-\ell-q+1}) + \sum_{\ell=n-q+1}^{n}(a^{[1]}_{n,\ell} + a^{[2]}_{n,\ell}).$$

But the probabilities $a^{[1]}_{n,\ell,q}$ and $a^{[2]}_{n,\ell,q}$ satisfy (for $n \ge 2$, $1 \le \ell \le n$ and $1 \le q \le n-\ell$) the following system of recurrences, which can be obtained by an extension of the automaton for the transition probabilities presented in Section 1:

$$a^{[1]}_{n,\ell,q} = \frac{\ell}{n}\cdot a^{[2]}_{n-1,\ell,q-1} + \frac{q-1}{n}\cdot a^{[2]}_{n-1,\ell,q-1} + \left(1-\frac{\ell+q}{n}\right)\cdot a^{[1]}_{n-1,\ell,q},$$

$$a^{[2]}_{n,\ell,q} = \frac{\ell-1}{n}\cdot a^{[1]}_{n-1,\ell-1,q} + \frac{q-1}{n}\cdot a^{[1]}_{n-1,\ell,q-1} + \left(1-\frac{\ell+q}{n}\right)\cdot a^{[2]}_{n-1,\ell,q}.$$

It turns out that the normalization factor $\frac{n!}{(\ell-1)!(q-1)!(n-\ell-q)!}$ yields a reduction of the system of PDEs for the corresponding generating functions and one gets

$$v(1-z-zv)B^{[1]}_v(z,u,v) - \frac{z^2u^2v}{1-z-zv}B^{[1]}_u(z,u,v) - \big(1-z+\frac{z^2uv}{1-z-zv}\big)B^{[1]}(z,u,v) = 0,$$

for $B^{[\cdot]}(z,u,v) := \sum_{n\ge1}\sum_{1\le\ell\le n}\sum_{1\le q\le n-\ell} b^{[\cdot]}_{n,\ell,q} z^n u^\ell v^q$ with $b^{[\cdot]}_{n,\ell,q} := \frac{n!}{(\ell-1)!(q-1)!(n-\ell-q)!}a^{[\cdot]}_{n,\ell,q}$. However, we note that it turned out to be rather involved to adapt the general solution of this PDE to the boundary values, i.e., to find the proper solution, which eventually yields the result stated in the theorem.

4 Conclusion and outlook

We provided a rather detailed study of various hiring quantities extending the analysis started in [1] yielding insight into the hiring process when applying the "hiring above the median" strategy. One further quantity, which might be called the number of replacements, seems to be of interest in a variation of the problem, where it is possible not only to hire candidates, but also to fire already recruited candidates (of a low score) as discussed in [2, 4]. A technique for treating this variant and to achieve the best quality of the hired staff is to allow replacements of "low performers" in basic hiring strategies such as "hiring above the median". Briefly, it might happen that at some step a good candidate is discarded (because his score is less than the decision maker), but he is better than the worst hired candidate. Then the hiring set is missing such a good candidate. To resolve this situation we extend the hiring strategy as follows: each candidate has two possibilities to get hired, namely either the basis strategy will hire him, or he replaces the worst candidate amongst all hired ones. We notice that the size of the hiring set does not change by using replacements, since we hire one candidate but we fire another one instantly. Our recursive approach seems promising to also obtain results for this quantity.

The analysis of the hiring quantities given in this work is based on a recursive approach, where we always took into account the rank of the decision maker. It is a natural question to ask whether the present approach could be applied to treat other interesting hiring strategies also. First we note that for the strategy "hiring above the m-th best candidate" this point of view indeed turned out to be fruitful [4]. Furthermore, at present we work on treating instances (with $0 < \alpha < 1$ a rational number) of the class of "hiring above the α-quantile" strategies other than the present $\alpha = \frac{1}{2}$. E.g., by distinguishing whether the number of hired candidates is congruent 0, 1, or 2 modulo 3, the instances $\alpha = \frac{1}{3}$ and $\alpha = \frac{2}{3}$ naturally lead to systems of equations for three recursive sequences; at least for these cases we also obtained suitable normalization factors to reduce the corresponding system of generating functions and to get explicit solutions.

Another approach in getting further insight into the "hiring above the α-quantile" strategies, but which might be interesting in its own, is to consider a "probabilistic relaxation" of the hiring process in the following sense. Let us consider a hiring strategy of the following type.

- The first M candidates are recruited.
- Then one of these candidates is selected (by a certain rule) as the first decision maker.

- Each time a new candidate is "examined" his score will be compared with the score of the decision maker.
  - If the new candidate does not have a larger score he will not be hired and the decision maker remains the same.
  - If the new candidate has a score larger than the decision maker he will be hired and furthermore with a certain probability $1 - p$ (which might depend on certain quantities) the decision maker remains the same, but with probability p the decision maker changes to the recruited candidate with the lowest score larger than the actual decision maker, i.e., the "next better candidate" will be the new decision maker.

"Hiring above the median" (and also "hiring above the α-quantile") is falling into this class of strategies, where, of course, the probabilities p are given deterministically as 1 or 0 depending on the parity of the size of the hiring set. However, e.g., one could consider this probabilistic strategy for a fixed probability $0 < p < 1$ (thus yielding a "relaxed hiring above the α-quantile" strategy), for which a PDE approach seems to be feasible.

From a combinatorial point of view our recursive approach does not give a satisfying answer to the question, why there are occurring so simple formulas, in particular for the size of the hiring set. Although meanwhile we also obtained a non-recursive derivation of the results for the size of the hiring set, we still cannot provide a simple combinatorial explanation for this fact. Of course, it would be interesting to find one, which then either could turn out to be useful in treating other instances of "hiring above the α-quantile", or alternatively could give an "explanation", why the case $\alpha = \frac{1}{2}$ is so special leading to nice formulas.

References

[1] M. Archibald and C. Martinez, The Hiring Problem and Permutations, in: Proceedings of the 21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2009), *Discrete Mathematics and Theoretical Computer Science*, proceedings AK, 63–76, 2009.

[2] A. Broder, A. Kirsch, R. Kumar, M. Mitzenmacher, E. Upfal and S. Vassilvitskii, The Hiring Problem and Lake Wobegon Strategies, in: Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 1184–1193, ACM-SIAM, 2008.

[3] P. R. Freeman, The Secretary Problem and its Extensions: A Review, *International Statistical*, 51(2), 189–206, 1983.

[4] A. Helmi, C. Martínez and A. Panholzer, Hiring above the m-th best candidate: a generalization of records in permutations, submitted.

[5] H.-K. Hwang, M. Kuba and A. Panholzer, Analysis of some exactly solvable diminishing urn models, in: *Proceedings of the "19th International Conference on Formal Power Series and Algebraic Combinatorics"*, Nankai University, Tianjin, 2007. available at `http://dmg.tuwien.ac.at/kuba/urnsub.pdf`