Hiring above the m-th best candidate: a generalization of records in permutations*

Ahmed Helmi[1], Conrado Martínez[1], and Alois Panholzer[2]

[1] Dept. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain. {ahelmi,conrado}-at-lsi.upc.edu
[2] Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien, 1040 Wien, Austria. Alois.Panholzer-at-tuwien.ac.at

Abstract. The *hiring problem* is a simple model of on-line decision-making under uncertainty. As in many other such models, the input is a sequence of instances and a decision must be taken for each instance depending on the subsequence examined so far, while nothing is known about the future. One famous example of on-line decision-making is the selection of the maximum of a sequence, when the instances of this sequence are serviced sequentially and a decision must be taken to select or discard the current instance. Such model was first introduced in the early sixties as the *secretary problem* (Freeman, 1983). Broder et al. (2008) introduced the *hiring problem* as an extension of the secretary problem. Instead of selecting only one candidate, we are looking for selecting (hiring) many candidates to grow up a small company. In this context, a hiring strategy should meet two demands: to hire candidates at some reasonable rate and to improve the average quality of the hired staff. Soon afterwards, Archibald and Martínez (2009) introduced a discrete model of the hiring problem where candidates seen so far could be ranked from best to worst without the need to know their absolute quality scores. Hence the sequence of candidates could be modeled as a random permutation. Two general families of hiring strategies were introduced: *hiring above the m-th best candidate* and *hiring in the top P% quantile* (for instance, $P = 50$ is hiring above the median). In this paper we consider only hiring above the m-th best candidate. The hiring process under this strategy goes as follows: hire the first interviewed m candidates whatever their relative ranks, then hire the current candidate if and only if his rank is better than the current m-th best one (i.e., better than the current m-record) of the already hired staff, and discard him otherwise. We introduce new hiring parameters that describe the dynamics of the hiring process, like the *distance between the last two hirings*, and the quality of the hired staff, like the *score of the best discarded candidate*. While Archibald and Martínez made systematic use of analytic combinatorics techniques (Flajolet, Sedgewick, 2008) in their analysis, we use here a different approach to study the various hiring parameters related associated to the hiring process. We are able to obtain explicit formulas for the probability distribution or the probability generating function of the random variables of interest in a rather direct way. The explicit nature of the results also allows a very detailed study of their asymptotic behaviour. Adding our new results to those of Archibald and Martínez leads to a very precise quantitative characterization of the hiring above the m-th best candidate strategy. This might prove very useful in applications of the hiring process, e.g. in data stream algorithms.

1 Introduction

On-line decision making under uncertainty is a rich discipline of research. It arises in diverse fields such as Computer Science and Economics, where the input is a sequence of instances and a decision must be taken for each instance depending on the subsequence examined so far, while nothing is known about the future. The goal is often to design an algorithm or a strategy that meets the desired requirements. There are many real world and theoretical situations that share the aspects of decision making under uncertainty.

The famous secretary problem [5] involves many of the main features of decision making under uncertainty. In the standard secretary problem, the employer is looking for only one candidate to fill one secretarial

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

position under the following conditions: the number n of applicants is known, the applicants are interviewed sequentially in random order, each order being equally likely, it is assumed that one can rank all the applicants from best to worst without ties, the decision to accept or reject an applicant must be based only on the relative ranks of those applicants interviewed so far, decisions are taken on-line and are irrevocable, an applicant once rejected cannot be recalled later and the employer will be satisfied with nothing but the very best. Thus the goal is to maximize the probability of choosing the best candidate in the sequence.

The secretary problem has many extensions and generalizations (see [5]), including the relaxation of some of the conditions described above. One important extension is to consider the case when the employer is looking for many employees to grow her company. Broder et al. [2] introduced this extension as the hiring problem. The hiring problem has the same spirit as the secretary problem but with some changes. One difference is the number of candidates, which is unknown (potentially infinite) in the hiring problem, whereas this number is known in advance in the secretary problem. Another one is the measure of quality: this measure is clear for the secretary problem where the optimal strategy is the one maximizing the probability of choosing the best candidate. On the contrary, the number of applicants to hire in the hiring problem is not fixed in advance, and there are two —conflicting— goals in the hiring problem: to hire candidates at some reasonable rate and to improve the "average quality" of the hired staff.

Broder et al. presented their continuous probabilistic model of the hiring problem in [2]. They considered the quality scores of the candidates as i. i. d. random variables with common distribution $\mathrm{Unif}(0,1)$ rather than their relative ranks as in secretary problem. They presented some natural hiring strategies which they called *Lake Wobegon* strategies: *hiring above the current mean* and *hiring above the current median*. For instance, in hiring above the current mean, the next candidate is hired if and only if his quality score is better than the mean score of all previous hired candidates, and discarded otherwise. Broder et al. use the *number of interviews required to hire n candidates* and the *gap between the score of the last hired candidate and the maximum score* (which is 1) as the hiring parameters of interest.

Archibald and Martínez [1] handled the hiring problem from another point of view. They introduced a combinatorial (discrete) model of the problem. They 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 at step j all ranks from 1 (worst) to j (best) are equally likely. Then each finite subsequence of candidates represents a random permutation. More precisely, given a permutation σ_{n-1} (of size $n-1$) and a value (relative rank) j, $1 \leq j \leq n$, $\sigma_n = \sigma_{n-1} \circ j$ denotes the resulting permutation after relabelling $j, j+1, \ldots, n-1$ in σ_{n-1} as $j+1, \ldots, n$, and appending j to the end. For example, let $S_7 = 1,2,1,4,2,4,2$ represent the input sequence of candidates. Then $\sigma_1 = 1$, $\sigma_2 = \sigma_1 \circ 2 = 12$, $\sigma_3 = \sigma_2 \circ 1 = 231$ and so on until $\sigma_7 = 4617352$.

More formally, the input is a sequence of relative scores $S = s_1, s_2, \ldots, s_i, \ldots$, with $1 \leq s_i \leq i$, 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 $\{1, 2, \ldots, i\}$. Furthermore, we have the other common rules: a decision must be taken whether to hire the i-th candidate or not at step i; decisions are irrevocable; there is no information about the future candidates.

Hiring above the m-th best candidate strategy processes the sequence of candidates in two phases. In the initial phase, the first m interviewed candidates are hired regardless of their ranks. After that, there comes a selection phase, in which any coming candidate will be hired if and only if he ranks better than the m-th best already hired candidate. So the m-th best hired candidate (i.e., the current m-record) is the decision maker for this strategy and at any time step n there are m choices for hiring a new candidate which must have one of the relative ranks $n, n-1, \ldots, n-m+1$. For example, let $m = 3$ and we have already seen seven candidates represented by the permutation $\sigma_7 = 4617352$. Then candidates with scores $\{4, 6, 1, 7, 5\}$ are hired, whereas the ones with scores $\{3, 2\}$ are discarded. A candidate coming after σ_7 gets hired if he has a rank in the set $\{8, 7, 6\}$, whereas he gets discarded otherwise. For this hiring strategy it holds that, for any $n \geq m$, the hiring set always contains the m best candidates seen so far (and maybe others). To be more precise, the set of hired candidates $R_{\leq m}$ can be described as the set of left-to-right ($\leq m$)-maxima (or ($\leq m$)-records); of course, the particular case $m = 1$ (*hiring above the best strategy*)

coincides with the usual notion of records in a sequence. Let us explore the close connections between this hiring strategy and records in more detail. Consider the sequence $x_1, x_2, \ldots, x_n$ of n different scores, which are ranked $x_{i_1} < x_{i_2} < \cdots < x_{i_n}$. In the usual definition of m-records (see [7]) and references therein), an element x_i is contained in the set R_m of m-records if there exists an index $j \geq i$, such that x_i is the m-th largest element in the set $\{x_1, \ldots, x_j\}$ (i.e., if x_i is the m-th largest element seen so far at time j). It holds now that the set $R_{\leq m}$ of hired candidates of this sequence using the "hiring above the m-th best strategy" exactly consists of the $m-1$ candidates with largest score together with the set R_m of m-records, i.e.,

$$R_{\leq m} = R_m \ \dot\cup \ \{x_{i_n}, x_{i_{n-1}}, \ldots, x_{i_{n-m+2}}\}.$$

In particular, it easily follows that, for distinct scores of the candidates, the size of the hiring set is always $m-1$ plus the number of m-records in this sequence. Therefore, results for m-records in permutations as obtained, e.g., by Prodinger [7] are of interest here also, and vice versa, our detailed studies of this hiring strategy might lead to new insights in connection with record statistics.

Archibald and Martínez used analytic combinatorics techniques [4] to analyze the quantitative properties of hiring strategies. We review some of their results in Sect. 2. While still combinatorial, our approach in this work is significantly different. Since the behaviour of "hiring above the m-th best" is quite simple, the definition of each parameter can be used to directly obtain explicit formulas for the probability distribution or the probability generating function of the quantity of interest. The explicit nature of the results allows a very detailed study of their asymptotic behaviour. In particular we are able to characterize the limiting behaviour of the quantities depending on the size relation between m (the parameter of "rigidity" for hiring) and the number n of candidates, and thus get results not only for m fixed and $n \to \infty$. To clarify this point: the value m is always fixed during the application of this hiring strategy to a given sequence of candidates, but we can stop the hiring process after n candidates, where n might depend on m; e.g., it might be natural to ask for the asymptotic behaviour of the number of hired candidates if $n = 2m$, $n = m^2$, or $n = \exp(m)$, where $m \to \infty$ (and thus also $n \to \infty$). The results given in Sect. 3 will answer such questions; to cover the whole range $1 \leq m \leq n$ we state our asymptotic results in an equivalent way by expressing $m = m(n)$.

For the readers' convenience we collect here some notation used throughout this paper. We denote by $H_n := \sum_{k=1}^{n} \frac{1}{k}$ the harmonic numbers and by $H_n^{(r)} := \sum_{k=1}^{n} \frac{1}{k^r}$ the r-th order harmonic numbers. The signless Stirling numbers of first kind, which enumerate, e.g., the number of permutations of size n with exactly k cycles, are denoted by $\left[{n \atop k}\right]$. Furthermore, we use the Iverson's bracket notation $[\![P]\!]$: $[\![P]\!]$ evaluates to 1 if P is true and to 0 otherwise. Moreover, we write $X_n \xrightarrow{(d)} X$ for the weak convergence (i.e., convergence in distribution) of a sequence of random variables (r.v.) X_n to a r.v. X. The normal distribution with expectation μ and standard deviation σ is denoted by $\mathcal{N}(\mu, \sigma^2)$.

The sequel of this paper is structured as follows: Sect. 2 contains some of previous results of Archibald and Martínez for the hiring above the m-th best candidate strategy. Sect. 3 collects our new results about this strategy and constitute the main contribution of the paper. Sect. 4 is devoted to the proofs of the theorems given in Sect. 3. Finally, Sect. 5 ends with conclusions and a discussion about future work.

2 Previous work

The framework introduced by Archibald and Martínez in [1] considers random permutations to model the sequence of candidates as explained in the introduction. For each hiring parameter, they define a bivariate exponential generating function (BEGF) of the form $B(z,u) = \sum_{p \in \mathcal{P}} u^{\mathrm{cost}(p)} z^{|p|}/|p|!$, with $\mathcal{P}$ the family of permutations and $\mathrm{cost}(\cdot)$ a certain cost function. Then using the symbolic method, they derive a PDE for $B(z,u)$ by combining the corresponding recurrence of that parameter with the BEGF. Solving the PDE and using some analytic techniques often leads to a closed form for $B(z,u)$, from which one gets the probability distribution and (factorial) moments of the studied parameter by extracting the coefficients $[z^n u^k]B(z,u)$ or $[z^n] \frac{\partial^r}{\partial u^r} B(z,u)\big|_{u=1}$, respectively.

2.1 Size of the hiring set

This fundamental parameter counts the number of hired candidates in the sequence after n interviews, applying some given strategy. Let $h_{n,m}$ be the random variable that denotes the size of the hiring set, i.e., the number of hired candidates, after n interviews when the strategy "hiring above the m-th best candidate" is applied. Recall the example mentioned in the introduction: for $m = 3$ and $\sigma_7 = 4617352$, the hiring set contains $\{4, 6, 1, 7, 5\}$ hence $h_{7,3} = 5$. Then, for $1 \leq m \leq n$, Archibald and Martínez [1] showed (see also [7] for corresponding results on m-records) the following exact result for the expectation of $h_{n,m}$, where the given asymptotic expansion holds uniformly for $1 \leq m \leq n$:

$$\mathbb{E}\{h_{n,m}\} = m(H_n - H_m + 1) = m \ln\left(\frac{n}{m}\right) + m + O(1).$$

For hiring above the best, $m = 1$, it holds $\mathbb{E}\{h_{n,1}\} = \ln n + O(1)$ and $n!\,\mathbb{P}\{h_{n,1} = k\}$ is given by the signless Stirling number of the first kind $\left[{n \atop k}\right]$, which coincides with the number of permutations of size n that have exactly k left-to-right maxima [6]. For $m = \Theta(1)$ (fixed m), Archibald and Martínez have shown that the asymptotic behaviour of the variance is also $\mathbb{V}\{h_{n,m}\} = m \ln n + O(1)$. Furthermore, by applying Hwang's quasi-power theorem [4], they have proved a central limit theorem for $h_{n,m}$, which we restate here.

Theorem 1. *Let $h_{n,m}$ denote the size of hiring set for n candidates and the strategy "hiring above the m-th best candidate". Then, for m fixed and $n \to \infty$, it holds:*

$$\frac{h_{n,m} - m \ln n}{\sqrt{m \ln n}} \xrightarrow{(d)} \mathcal{N}(0, 1).$$

2.2 Gap of the last hired candidate

Let $r_{n,m}$ denote the rank of the last hired candidate for a permutation of size n when the strategy "hiring above the m-th best candidate" is applied. We consider $g_{n,m} := 1 - r_{n,m}/n$, the *gap* of the last hired candidate. The random variable $g_{n,m}$ hints at the quality of the hired staff, and a good hiring strategy should have $\mathbb{E}\{g_{n,m}\} \to 0$ as $n \to \infty$. For example, let $m = 3$ and $\sigma_7 = 4617352$, then the candidate with score 5 is the last to be hired and $g_{7,3} = 1 - 5/7 = 2/7$. By definition, $r_{n,m}$ is uniformly distributed over the best m ranks seen so far: $n, n-1, n-2, \ldots, n-m+1$. Thus, $g_{n,m}$ is uniformly distributed over the values $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{m-1}{n}$, as stated in the following theorem.

Theorem 2. *Let $g_{n,m}$ denote the gap of the last hired candidate for n candidates under the strategy "hiring above the m-th best candidate". Then, for $1 \leq m \leq n$,*

$$\mathbb{P}\left\{g_{n,m} = \frac{k}{n}\right\} = \frac{1}{m}, \quad \text{for } k \in \{0, 1, \ldots, m-1\}.$$

As an immediate consequence, we have that $\mathbb{E}\{g_{n,m}\} = \frac{m-1}{2n}$.

In particular, it follows that $g_{n,m} = 0$ for hiring above the best ($m = 1$) since this strategy stops recruiting candidates once the best candidate has been hired; for $m > 1$ fixed, $g_{n,m}$ quickly goes to 0 as n grows.

3 Results

3.1 Size of the hiring set, $h_{n,m}$

In addition to the already mentioned results for the size $h_{n,m}$ of the hiring set obtained in [1], where the authors mainly focused on a study of $h_{n,m}$ for m fixed, we give here a characterization of the exact and limiting behaviour of this fundamental quantity, which is valid for any size relation between m and n.

Theorem 3. *Let $h_{n,m}$ denote the number of hired candidates for n candidates and the strategy "hiring above the m-th best candidate". Then the exact distribution of $h_{n,m}$ is given as follows:*

$$\mathbb{P}\{h_{n,m}=j\}=\begin{cases}[\![n=j]\!], & \text{if } m>n,\\ \frac{m^{j-m}}{\binom{n}{m}}\sum_{\ell=0}^{n-j}\frac{\left[{\ell+j-m \atop j-m}\right]}{(\ell+j-m)!}, & \text{if } m\le j\le n.\end{cases}$$

For $1\le m\le n$ the expectation and the variance of $h_{n,m}$ are given as follows, where the asymptotic expansions hold uniformly for $1\le m\le n$ and $n\to\infty$:

$$\mathbb{E}\{h_{n,m}\}=m(H_n-H_m+1)=m\big(\log n-\log m+1\big)+O(1),$$
$$\mathbb{V}\{h_{n,m}\}=m(H_n-H_m)-m^2(H_n^{(2)}-H_m^{(2)})=m\big(\log n-\log m-1+\frac{m}{n}\big)+O(1).$$

The limiting distribution of $h_{n,m}$ is, for $n\to\infty$ and depending on the size relation between m and n, characterized as follows:

- *$n-m\gg\sqrt{n}$: Suitably normalized, $h_{n,m}$ is asymptotically standard normal distributed, i.e.,*

$$\frac{h_{n,m}-m\big(\log n-\log m+1\big)}{\sqrt{m\big(\log n-\log m-1+\frac{m}{n}\big)}}\xrightarrow{(d)}\mathcal{N}(0,1).$$

- *$n-m\sim\alpha\sqrt{n}$, with $\alpha>0$: $n-h_{n,m}$ is asymptotically Poisson distributed with parameter $\frac{\alpha^2}{2}$, i.e.,*

$$n-h_{n,m}\xrightarrow{(d)}\text{Poisson}\,\big(\frac{\alpha^2}{2}\big).$$

- *$n-m=o(\sqrt{n})$: $n-h_{n,m}$ converges in distribution to 0, i.e., $n-h_{n,m}\xrightarrow{(d)}0$.*

3.2 Index of the last hired candidate

The index $L_{n,m}$ of the last hired candidate can be seen as the *time* of the last hiring in a permutation of size n. Its behavior helps us to better understand the dynamics of the hiring process and it gives a measure of the hiring rate. Archibald and Martínez already introduced $L_{n,m}$ in [1] and gave a general PDE that applies to "hiring above the m-th best candidate" and many other (in particular, to all those strategies were decisions depend exclusively in the relative ranks of the candidates, not on their absolute scores). For example, for $m=3$ and $\sigma_7=4617352$, we have $L_{7,3}=6$ since the candidate with score 5 is the last one to be hired. The following theorem contains our results for $L_{n,m}$, which characterize its probability distribution and the corresponding limiting distribution.

Theorem 4. *Let $L_{n,m}$ denote the index of the last hired candidate for n candidates under the strategy "hiring above the m-th best candidate". Then the exact distribution of $L_{n,m}$ is given as follows:*

$$\mathbb{P}\{L_{n,m}=j\}=\begin{cases}[\![j=n]\!], & \text{if } m>n,\\ \frac{\binom{j-1}{m-1}}{\binom{n}{m}}, & \text{if } m\le n \text{ and } 1\le j\le n.\end{cases}$$

For $m\le n$ the expectation of $L_{n,m}$ is $\mathbb{E}\{L_{n,m}\}=\frac{m(n+1)}{m+1}$.

The limiting distribution of $L_{n,m}$ is, for $n\to\infty$ and depending on the size relation between m and n, characterized as follows:

- *m fixed: Suitably normalized, $L_{n,m}$ is asymptotically beta distributed with parameters m and 1, i.e.,*

$$\frac{L_{n,m}}{n}\xrightarrow{(d)}\text{Beta}(m,1).$$

– *$m \to \infty$, but $m = o(n)$: Suitably normalized, $n - L_{n,m}$ is asymptotically exponential distributed with parameter 1, i.e.,*

$$\frac{m}{n}(n - L_{n,m}) \xrightarrow{(d)} \mathrm{Exp}(1).$$

– *$m \sim \alpha n$, with $0 < \alpha < 1$: $n - L_{n,m}$ is asymptotically geometrically distributed with success probability α, i.e.,*

$$n - L_{n,m} \xrightarrow{(d)} \mathrm{Geom}(\alpha).$$

– *$n - m = o(n)$: $n - L_{n,m}$ converges in distribution to 0, i.e., $n - L_{n,m} \xrightarrow{(d)} 0$.*

3.3 Distance between the last two hirings

We define the distance $\Delta_{n,m}$ between the last two hirings as the number of interviews between the last two hired candidates plus one. By convention we take $\Delta_{n,m} = 0$ if $h_{n,m} < 2$. A reasonable hiring rate requires $L_{n,m}$ to be close to n and $\Delta_{n,m}$ to be relatively small compared to n. In the initial phase ($2 \le n \le m$) of hiring above the m-th best candidate, $\Delta_{n,m}$ takes the value 1 because every candidate is hired. For $n > \max(m, 2)$, $\Delta_{n,m}$ can take any value between 1 and $n - m$; if only one candidate is hired, which holds for $n = 1$ and can occur also for the particular instance $m = 1$, we set $\Delta_{n,m} = 0$. For example, let $m = 3$ and $\sigma_7 = 4617352$, then $\Delta_{7,3} = 2$ because the last two hired candidates are those with scores 7 and 5. The following theorem gives a characterization of the exact and limiting probability distribution of $\Delta_{n,m}$.

Theorem 5. *Let $\Delta_{n,m}$ denote the distance between the last two hirings for n candidates for the strategy "hiring above the m-th best candidate". Then the exact distribution of $\Delta_{n,m}$ is given as follows (for all other values of the parameters the probabilities are zero):*

– *$m > n$: $\mathbb{P}\{\Delta_{n,m} = 1\} = 1$ if ($d = 1$ and $n > 1$) or ($d = 0$ and $n = 0$).*
– *$m = 1 \le n$:*

$$\mathbb{P}\{\Delta_{n,1} = d\} = \begin{cases} \frac{1}{n}, & \text{if } d = 0, \\ \frac{1}{n}\big(H_{n-1} - H_{d-1}\big), & \text{if } 1 \le d \le n-1. \end{cases}$$

– *$2 \le m \le n$:*

$$\mathbb{P}\{\Delta_{n,m} = d\} = \begin{cases} \frac{1}{m-1}\big(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\big), & \text{if } d = 1, \\ \frac{m}{\binom{n}{m}} \displaystyle\sum_{j=m+d}^{n} \frac{1}{j-m}\binom{j-d-1}{m-1}, & \text{if } 2 \le d \le n-m. \end{cases}$$

For $2 \le m \le n$ the expectation of $\Delta_{n,m}$ is given as follows, where the asymptotic equivalent holds for $m = o(n)$ and $n \to \infty$:

$$\mathbb{E}\{\Delta_{n,m}\} = \frac{m(n+1)}{(m+1)^2} - \frac{m^2}{n(m-1)} + \frac{2m}{(m^2-1)\binom{n}{m}} \sim \frac{m(n+1)}{(m+1)^2}.$$

The limiting distribution of $\Delta_{n,m}$ is, for $n \to \infty$ and depending on the size relation between m and n, characterized as follows:

– *m fixed: Suitably normalized, $\Delta_{n,m}$ converges in distribution to a continuous r.v., which is characterized by its density function: $\frac{\Delta_{n,m}}{n} \xrightarrow{(d)} X_m$, where X_m has the density function*

$$f_m(x) = m^2\Big((-1)^m x^{m-1}\log x + (-1)^{m-1}H_{m-1}x^{m-1} + \sum_{\ell=0}^{m-2}\frac{(-1)^\ell}{m-1-\ell}\binom{m-1}{\ell}x^\ell\Big), \quad 0 < x < 1.$$

– *$m \to \infty$, but $m = o(n)$: Suitably normalized, $\Delta_{n,m}$ is asymptotically exponential distributed with parameter 1, i.e.,*

$$\frac{m}{n}\Delta_{n,m} \xrightarrow{(d)} \mathrm{Exp}(1).$$

– $m \sim \alpha n$, *with* $0 < \alpha < 1$*:* $\Delta_{n,m} - 1$ *is asymptotically geometrically distributed with success probability* α*, i.e.,*

$$\Delta_{n,m} - 1 \xrightarrow{(d)} \mathrm{Geom}(\alpha).$$

– $n - m = o(n)$*:* $\Delta_{n,m} - 1$ *converges in distribution to* 0*, i.e.,* $\Delta_{n,m} - 1 \xrightarrow{(d)} 0$*.*

3.4 Score of the best discarded candidate,

As with the gap $g_{n,m}$, this random variable $M_{n,m}$ provides a measure of the quality of the hired staff. For example, let $m = 3$ and $\sigma_7 = 4617352$ then $M_{7,3} = 3$, since all larger ranks are hired in this instance. A high value (close to n) of $M_{n,m}$ means that the hiring strategy is very selective, whereas a low value of $M_{n,m}$ means that the strategy is hiring too many candidates. For $m \le n$, $M_{n,m}$ can take values between 0 (all candidates have been hired) and $n - m$ because as mentioned before the best m candidates in the sequence are always hired; if $n < m$ then all candidates are hired and $M_{n,m} = 0$ holds. Explicit formulæ for the probability distribution and the limiting distribution of $M_{n,m}$ are stated in the following theorem.

Theorem 6. *Let* $M_{n,m}$ *denote the score of the best discarded candidate for* n *candidates under the strategy "hiring above the* m*-th best candidate". Then the exact distribution of* $M_{n,m}$ *is given as follows:*

$$\mathbb{P}\{M_{n,m} = b\} = \begin{cases} [\![b = 0]\!], & \text{if } n > m, \\ \frac{m!}{n!} m^{n-m}, & \text{if } b = 0 \text{ and } 1 \le m \le n, \\ \frac{m!}{(n-b+1)!} \cdot (n - m - b + 1) \cdot m^{n-m-b}, & \text{if } 1 \le b \le n - m \text{ and } 1 \le m \le n. \end{cases}$$

For $1 \le m \le n$*, the expectation of* $M_{n,m}$ *is*

$$\mathbb{E}\{M_{n,m}\} = n - m - \frac{(n-m)\, m!\, m^{n-m+1}}{(n+1)!} - \sum_{j=0}^{n-m} \frac{j\,(j+1)\, m^j\, m!}{(m+j+1)!} = n - m + O(\sqrt{m}),$$

where the asymptotic expansion holds uniformly for $1 \le m \le n$ *and* $n \to \infty$*.*

The limiting distribution of $M_{n,m}$ *is, for* $n \to \infty$ *and depending on the size relation between* m *and* n*, characterized as follows:*

– m *fixed:* $n - m - M_{n,m}$ *converges in distribution to a discrete r.v., which is characterized by its probability function:* $n - m - M_{n,m} \xrightarrow{(d)} Y_m$*, where* Y_m *has the probability function*

$$\mathbb{P}\{Y_m = j\} = \frac{(j+1)\, m^j\, m!}{(m+j+1)!}, \quad j \in \mathbb{N}.$$

– $m \to \infty$*, but* $n - m \gg \sqrt{m}$*: Suitably normalized,* $n - m - M_{n,m}$ *is asymptotically Rayleigh distributed with parameter* 1*, i.e.,*

$$\frac{n - m - M_{n,m}}{\sqrt{m}} \xrightarrow{(d)} \mathrm{Rayleigh}(1).$$

– $n - m \sim \alpha\sqrt{m}$*, with* $\alpha > 0$*: Suitably normalized,* $n - m - M_{n,m}$ *converges in distribution to the minimum between* α *and a Rayleigh distributed r.v., i.e.,*

$$\frac{n - m - M_{n,m}}{\sqrt{m}} \xrightarrow{(d)} \min\big(\alpha, \mathrm{Rayleigh}(1)\big).$$

– $n - m = o(\sqrt{m})$*:* $M_{n,m}$ *converges in distribution to* 0*, i.e.,* $M_{n,m} \xrightarrow{(d)} 0$*.*

4 Proofs

We give here the analytical proofs of the theorems in Sect. 3. We focus here on deriving the explicit results characterizing the exact probability distributions of the quantities considered, since, due to the explicit nature of these exact formulas, the asymptotic results follow from them essentially by applying Stirling's formula for the factorials (or the Gamma function)

$$\log x! = \left(x + \frac{1}{2}\right) \log x - x + \frac{1}{2} \log(2\pi) + O(x^{-1}) \tag{1}$$

in connection with standard techniques, which allow us to be more brief here.

4.1 Proof of Theorem 3

Since the instance $m > n$ is trivial (all candidates are hired), we can focus on the case $1 \leq m \leq n$. From the definition of this hiring strategy it follows immediately that

$$h_{n,m} = \chi_1 + \chi_2 + \cdots + \chi_n,$$

where the indicator variables χ_j, which are 1 if the j-th candidate of the sequence is hired, and 0 otherwise, are mutually independent with distribution

$$\mathbb{P}\{\chi_j = 1\} = \begin{cases} 1, & \text{for } 1 \leq j \leq m, \\ \frac{m}{j}, & \text{for } m < j \leq n. \end{cases}$$

Thus, the probability generating function $h_{n,m}(v) := \sum_{\ell \geq 0} \mathbb{P}\{h_{n,m} = \ell\} v^\ell$ is given by the following explicit formula (note that the corresponding probability generating function for m-records in permutations already appears in [7]), which will be the starting point to derive the exact and asymptotic results:

$$h_{n,m}(v) = v^m \prod_{j=m+1}^{n} \frac{mv + (j-m)}{j} = v^m \frac{(mv+n-m)! \cdot m!}{(mv)! \cdot n!} = v^m \frac{\binom{n+m(v-1)}{n}}{\binom{mv}{m}}. \tag{2}$$

To get an explicit result for the probabilities and thus the connection to signless Stirling numbers of first kind we introduce the generating function $h_m(z,v) := \sum_{n \geq m} \binom{n}{m} h_{n,m}(v) z^n$. A simple computation shows then

$$h_m(z,v) = \frac{(zv)^m}{(1-z)^{mv+1}}.$$

Using the well-known generating function [6] of the Stirling numbers $\sum_{n,k} \left[{n \atop k}\right] \frac{z^n}{n!} v^k = \frac{1}{(1-z)^v}$ the explicit result for the distribution of $h_{n,m}$ easily follows. Furthermore, the result for $h_m(z,v)$ easily gives, via differentiating r times with respect to v, evaluating at $v = 1$ and extracting coefficients $[z^n]$, explicit results for the r-th factorial moments of $h_{n,m}$ and, as a consequence, the formulas for the expectation and the variance stated in the theorem. The corresponding asymptotic results follow from the asymptotic expansion of the first and second order harmonic numbers, $H_n = \log n + \gamma + O(n^{-1})$ and $H_n^{(2)} = \frac{\pi^2}{6} - n^{-1} + O(n^{-2})$.

It remains to show the limiting distribution results, which we will only sketch here very briefly. Basically we will show that the moment generating function $\mathbb{E}\left\{e^{h^*_{n,m} s}\right\}$ of a suitably normalized version $h^*_{n,m}$ of $h_{n,m}$ converges pointwise for each real s to the moment generating function $\mathbb{E}\left\{e^{Xs}\right\}$ of a certain r.v. X. An application of the theorem of Curtiss [3] shows then the weak convergence of $h^*_{n,m}$ to X.

For the main region $n - m \gg \sqrt{n}$ we consider the normalized r.v. $h^*_{n,m} := \frac{h_{n,m} - \mu}{\sigma}$, with $\mu := \mu_{n,m} = m(\log n - \log m + 1)$ and $\sigma := \sigma_{n,m} = m(\log n - \log m - 1 + \frac{m}{n})$, yielding thus the moment generating function $\mathbb{E}\left\{e^{h^*_{n,m} s}\right\} = e^{-\frac{\mu}{\sigma} s} \cdot h_{n,m}(e^{\frac{s}{\sigma}})$, with $h_{n,m}(v)$ the probability generating function (2) given above. For simplicity we consider here only $m \to \infty$, since for the region m fixed the central limit theorem has

been shown already in [1]. An application of Stirling's formula (1) shows then, after some computations, the following expansion (which holds for any fixed real s):

$$\log\left(\mathbb{E}\left\{e^{h^*_{n,m}s}\right\}\right) = \frac{s^2}{2} + O\Big(\frac{m(1-\frac{m}{n})^2}{\sigma^3}\Big) + O(\sigma^{-1}) + O(m^{-1}),$$

which implies that $\mathbb{E}\left\{e^{h^*_{n,m}s}\right\} \to e^{\frac{s^2}{2}}$, pointwise for each real s, provided that $n - m \gg \sqrt{n}$. Since $e^{\frac{s^2}{2}}$ is the moment generating function of a standard normal distribution the theorem of Curtiss shows the stated central limit theorem.

For the region $n - m = O(\sqrt{n})$ we consider the r.v. $h^*_{n,m} := n - h_{n,m}$, yielding the moment generating function $\mathbb{E}\left\{e^{h^*_{n,m}s}\right\} = e^{ns} \cdot h_{n,m}(e^{-s})$. Again, an application of Stirling's formula shows the expansion

$$\mathbb{E}\left\{e^{h^*_{n,m}s}\right\} = e^{\frac{(n-m)^2}{2n}(e^s-1)} \cdot \Big(1 + O\big(\frac{n-m}{n}\big) + O\big(\frac{(n-m)^3}{n^2}\big)\Big).$$

Since $e^{\lambda(e^s-1)}$ is the moment generating function of a Poisson distributed r.v. with parameter λ the limiting distribution result for $n - m \sim \alpha\sqrt{n}$ follows. For $n - m = o(\sqrt{n})$ the moment generating function of $h^*_{n,m}$ converges to 1, which shows the stated theorem for this region also.

4.2 Proof of Theorem 4

Trivially, for $n > m$ one gets $\mathbb{P}\{L_{n,m} = n\} = 1$, thus we only have to consider the range $1 \le m \le n$. It is immediate from the definition of "hiring above the m-th best candidate" (see also Subsect. 4.1) that the probability of hiring at any position $j > m$ equals $\frac{m}{j}$. Thus we get the stated exact result for the probability distribution of $L_{n,m}$:

$$\begin{aligned}\mathbb{P}\{L_{n,m} = j\} &= \mathbb{P}\{\text{We hire at position } j\} \cdot \mathbb{P}\{\text{No hirings from position } (j+1) \text{ till } n\} \\ &= \frac{m}{j} \cdot \prod_{\ell=j+1}^{n} \big(1 - \frac{m}{\ell}\big) = \frac{\binom{j-1}{m-1}}{\binom{n}{m}}.\end{aligned}$$

The result for the expectation can be obtained by applying a simple summation formula.

The limiting distribution results can be obtained by applying Stirling's formula to the exact formula for the probabilities. We just give here the following expansion valid for the main region $m \to \infty$ and $m, k = o(n)$:

$$\mathbb{P}\{L_{n,m} = n - k\} = \mathbb{P}\{n - L_{n,m} = k\} = \frac{m}{n} e^{-\frac{km}{n}} \cdot \Big(1 + O\big(\frac{k^2 m}{n^2}\big) + O\big(\frac{km^2}{n^2}\big)\Big),$$

from which one immediately gets that $\frac{m}{n}(n - L_{n,m}) \xrightarrow{(d)} \mathrm{Exp}(1)$, provided that $m \to \infty$, but $m = o(n)$. The other regions are completely straightforward and thus we will not discuss them here.

4.3 Proof of Theorem 5

We only comment on the non-trivial case $1 \le m \le n$. Let us first consider the generic instance $2 \le m \le n$ and $d \ge 2$. By considering the position $j \ge m + d$ of the last hiring we immediately get the following formula:

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = d\} = \sum_{j=m+d}^{n} \Big[&\mathbb{P}\{\text{We hire at position } (j-d)\} \cdot \mathbb{P}\{\text{No hirings from position } (j-d+1) \text{ till } (j-1)\} \\ &\cdot \mathbb{P}\{\text{We hire at position } j\} \cdot \mathbb{P}\{\text{No hirings from position } (j+1) \text{ till } n\}\Big] \\ = \sum_{j=m+d}^{n} \frac{m}{j-d} \cdot \prod_{\ell=j-d}^{j-1} \big(1 - \frac{m}{\ell}\big) \cdot \frac{m}{j} \cdot \prod_{\ell=j+1}^{n} \big(1 - \frac{m}{\ell}\big) &= \frac{m}{\binom{n}{m}} \sum_{j=m+d}^{n} \frac{1}{j-m} \binom{j-d-1}{m-1}. \qquad (3)\end{aligned}$$

The other cases can be obtained from this generic instance by simple modifications. For $2 \le m \le n$ and $d = 1$ one has to add the contribution of the event that the last hiring occurs at position $j = m$, thus

$$\mathbb{P}\{\Delta_{n,m} = 1\} = \frac{m}{\binom{n}{m}} \sum_{j=m+1}^{n} \frac{1}{j-m}\binom{j-2}{m-1} + \mathbb{P}\{L_{n,m} = m\} = \frac{m}{\binom{n}{m}} \sum_{j=m+1}^{n} \frac{1}{j-m}\binom{j-2}{m-1} + \frac{1}{\binom{n}{m}}$$
$$= \frac{1}{m-1}\Big(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\Big),$$

where the last simplification follows from a summation formula. Finally, for the instance $m = 1$ the formula (3) holds for $d \ge 1$, but simplifies to the result stated in the theorem; additionally one has to consider here the case $d = 0$, i.e., there is only one hired candidate, namely the one with highest rank, which thus has to appear at the first position, yielding $\mathbb{P}\{\Delta_{n,1} = 0\} = \frac{1}{n}$.

As in previous cases, the exact result for the expectation $\mathbb{E}\{\Delta_{n,m}\}$ as stated in the theorem follows by applying a simple summation formula, and the asymptotic result immediately follows.

The asymptotic results for $\Delta_{n,m}$ are also a direct consequence of Stirling's formula applied to the exact probabilities, but, due to the summation occurring in the formula, they require slightly more care. For the most interesting region $m \to \infty$, but $m = o(n)$, we get for $d = O(\frac{n}{m})$ the local approximation $\mathbb{P}\{\Delta_{n,m} = d\} \sim \frac{m}{n} e^{-\frac{md}{n}}$, thus yielding the stated limiting distribution result for this region. For m fixed we get the local approximation $\mathbb{P}\{\Delta_{n,m} = d\} \sim \frac{m^2}{n} \int_{\frac{d}{n}}^{1} \frac{1}{t}\big(t - \frac{d}{n}\big)^{m-1} dt$, thus showing that $\frac{\Delta_{n,m}}{n} \xrightarrow{(d)} X_m$, where X_m has density function $f_m(x) = m^2 \int_x^1 \frac{1}{t}(t-x)^{m-1} dt$, $0 < x < 1$. The expression for $f_m(x)$ can be expressed also in the more explicit form stated in the theorem. The remaining regions are straightforward and we do not comment on them here.

4.4 Proof of Theorem 6

Again we only comment on the non-trivial case $1 \le m \le n$. To show the explicit result for the exact distribution of $M_{n,m}$ we will consider an auxiliary quantity, namely the probability $a_{n,m,j}$, with $0 \le j \le n-m$, that all of the $m + j$ highest ranked candidates are hired (and maybe others). Of course, $a_{n,m,0} = 1$, since the m highest ranked candidates are always hired. Since the candidate with the $(m + \ell)$-th highest rank, $1 \le \ell \le j$, is hired exactly when at most $m-1$ (i.e., $0, 1, \ldots, m-1$) of the (in total $m+\ell-1$) higher ranked candidates occur earlier in the sequence, the probability that this happens is thus given by $\frac{m}{m+\ell}$, and these events are independent, we get

$$a_{n,m,j} = \prod_{\ell=1}^{j} \frac{m}{m+\ell} = \frac{m!\, m^j}{(m+j)!}, \quad 0 \le j \le n-m.$$

But the probability that the best discarded candidate has rank $1 \le b \le n-m$ is thus simply given by the difference between the probability that all candidates with a rank higher than b are recruted and the probability that all candidates with a rank higher than $b-1$ are recruted, i.e.,

$$\mathbb{P}\{M_{n,m} = b\} = a_{n,m,n-m-b} - a_{n,m,n+1-m-b} = \frac{m!\, m^{n-m-b}}{(n-b)!} - \frac{m!\, m^{n+1-m-b}}{(n+1-b)!} = \frac{(n-m-b+1)\, m!\, m^{n-m-b}}{(n-b+1)!}.$$

Additionally, we have $\mathbb{P}\{M_{n,m} = 0\} = \mathbb{P}\{h_{n,m} = n\} = \frac{m!\, m^{n-m}}{n!}$, thus completing the results for the exact probability distribution of $M_{n,m}$ as stated in the theorem. Moreover, the exact result for the expectation $\mathbb{E}\{M_{n,m}\}$ follows by summation.

From this explicit formulas for the exact probabilities the limiting behaviour can readily be obtained by applying Stirling's formula. E.g., when considering the main region $m \to \infty$, but $n - m \gg \sqrt{m}$, the local expansion

$$\mathbb{P}\{M_{n,m} = n-m-j\} = \mathbb{P}\{n-m-M_{n,m} = j\} = \frac{j}{m} e^{-\frac{j^2}{2m}} \cdot \Big(1 + O\big(\frac{j}{m}\big) + O\big(\frac{j^3}{m^2}\big)\Big)$$

immediately entails that $\frac{n-m-M_{n,m}}{\sqrt{m}} \xrightarrow{(d)} Y$, where Y has the density function $f(x) = xe^{-\frac{x^2}{2}}$, $x > 0$, thus Y is Rayleigh distributed with parameter 1. We omit here the details for the remaining regions.

The asymptotic result for the expectation stated in the theorem can be obtained from this local expansion of the probabilities $\mathbb{P}\{M_{n,m} = n - m - j\}$. However, there is also an alternative approach expressing the formula in terms of hypergeometric functions leading to the following expansion, which holds uniformly for $1 \leq m \leq n$:

$$\mathbb{E}\{M_n\} = n - m - \sqrt{2\pi m}\left(1 - \frac{\Gamma(m+1,m)}{\Gamma(m+1)}\right) + O(1),$$

where $\Gamma(s,x)$ is the incomplete Gamma function, which is defined as $\Gamma(s,x) = \int_x^\infty t^{s-1}e^{-t}dt$, while $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$ is the ordinary Gamma function.

Since $\frac{\Gamma(m+1,m)}{\Gamma(m+1)} \leq 1$, for $m \geq 1$, the asymptotic expansion given in the theorem immediately follows.

5 Conclusions and Future Work

We have presented various theorems that describe the properties of the hiring process when applying the "hiring above the m-th best candidate" strategy. These results provide a very detailed picture of this natural hiring strategy. It is obvious from Theorem 2 that the quality of the hiring set improves along time, as the gap of the last hired candidate goes to zero as n becomes large. The hiring rate is relatively slow, with the index of the last hiring satisfying $L_{n,m}/n < 1$ (Theorem 4). In particular, for m fixed, this entails an exponential number of interviews to hire n candidates, although the base of the exponential growth approaches 1 as m is larger.

As already pointed out by Broder et al. [2] and Archibald et al. [1], non-degenerate hiring strategies[3] always exhibit trade-offs between the quality of the hired staff and the rate at which they hire. "Hiring above the m-th best candidate" provides an excellent example. By playing around with the value of m (the "rigidity"), we can give priority to a faster hiring rate or to a more selective process. If we make m bigger, then the distance between consecutive hirings decreases (better hiring rate), but the gap of the last hired candidate gets bigger too (worse staff quality). Similar trade-offs show up if we consider other combinations of the parameters that we have studied, like the size of the hiring set $h_{n,m}$ and the score $M_{n,m}$ of the best discarded candidate.

Despite these trade-offs arise very naturally, it seems very difficult to define a natural yardstick with which to compare different hiring strategies, and thus to come up with a clear notion of optimality. Intuitively, an "optimal" hiring strategy should achieve the perfect balance between the quality of the hired staff and the rate of hiring, but quantifying this balance remains as an elusive open problem.

A strong candidate to form part of the definition of optimality among hiring strategies is a new parameter which we have not considered here, the *number of replacements*. Take some ordinary hiring strategy and process the sequence of candidates as usual. For a candidate with rank j apply the hiring strategy and decide whether to hire or discard her. But here comes the difference: if the candidate was to be discarded **and** there is some candidate in the hired staff with lower rank, replace the worst hired candidate by the new candidate. The size h of the hiring set will remain the same, but it will contain the best h candidates in the sequence. The number of replacement F_n gives thus a measure of the effort that a given hiring strategy needs to "build" the perfect staff, and combines both quality and quantity aspects. We have preliminary results on the expectation of $F_{n,m}$, namely,

$$\mathbb{E}\{F_{n,m}\} = \frac{m}{2}\left(H_n^2 + H_{m-1}^2 + H_{m-1}^{(2)} - H_n^{(2)}\right) - mH_nH_m,$$

for "hiring above the m-th best candidate", but we are still working to obtain the probability distribution of $F_{n,m}$, as we have done for the other parameters studied in this work.

[3] here, by a non-degenerate hiring strategy, we mean a hiring strategy that is not hiring everybody nor discarding everybody.

Last but not least, we are currently studying the application of the results in this paper and in [1] to the analysis of data stream algorithms. The nature of data stream algorithms requires very little memory, simple computations and reasonably accurate results. Processing a sequence with "hiring above the m-th best candidate" is simple and can be efficiently done with little memory: we need only the m best values seen so far. Tracking the corresponding hiring parameters is very easy too. The usefulness of our results stems from the fact that the observed realization of the hiring parameters can be used to infer global properties of the underlying sequence. In particular, our exact and asymptotic formulas for the probability distribution of several hiring parameters is very useful to define estimators of global quantities of interest (e.g., the number of distinct elements in the sequence) and to show that these estimators are unbiased and have low variance.

References

[1] M. Archibald and C. Martínez. The hiring problem and permutations. In *Proc. of the 21st Int. Col. on Formal Power Series and Algebraic Combinatorics (FPSAC)*, volume AK of *Discrete Mathematics & Theoretical Computer Science Proceedings*, pages 63–76, 2009.

[2] A. Z. Broder, A. Kirsch, R. Kumar, M. Mitzenmacher, E. Upfal, and S. Vassilvitskii. The hiring problem and Lake Wobegon strategies. In *Proceedings of the 19textth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'08)*, pages 1184–1193, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[3] J. H. Curtiss. A note on the theory of moment generating functions. *Annals of Mathematical Statistics*, 13(4):430–433, 1942.

[4] Ph. Flajolet and R. Sedgewick. *Analytic Combinatorics.* Cambridge Univ. Press, 2008.

[5] P. R. Freeman. The secretary problem and its extensions: A review. *International Statistical*, 51(2):189–206, 1983.

[6] D. E. Knuth, R. L. Graham, and O. Patashnik. *Concrete Mathematics.* Addision Wesley, Reading, Mass., 2nd edition, 1994.

[7] H. Prodinger. d-records in geometrically distributed random variables. *Discrete Mathematics & Theoretical Computer Science*, 8(1):273–284, 2006.