The Hiring Problem and its Algorithmic Applications

Ahmed Mohamed Helmi Mohamed Elsadek

Barcelona, April, 2013

The Hiring Problem and its Algorithmic Applications

PhD Dissertation

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy to

Universitat Politècnica de Catalunya
Departament de Llenguatges i Sistemes Informàtics

By
Ahmed Mohamed Helmi Mohamed Elsadek

Under the supervision of
Prof. Dr. Conrado Martínez Parra

Barcelona, April, 2013

Resumen

El problema de la contratación es un modelo simple para la toma de decisiones secuencial en condiciones de incertidumbre, recientemente introducido en la literatura. Este tipo de problemas se presenta en diversas campos, como las Ciencias de la Computación y la Economía. El problema fue introducido explícitamente por primera vez por Broder et al. [15] en 2008 como una extensión natural del bien conocido problema de la secretaria (véase [38] y las referencias citadas por éste). Poco después, Archibald y Martínez [5] en 2009 introdujeron un modelo discreto combinatorio del problema de la contratación, donde los candidatos vistos hasta un momento dado podrían ser clasificados de mejor a peor sin la necesidad de conocer sus puntuacíones de calidad en términos absolutos. En esta tesis se presenta un extenso estudio para el problema de la contratación bajo la formulación propuesta por Archibald y Martínez, se exploran las conexiones con otros procesos de selección secuenciales, y se desarrolla un aplicación interesante de nuestros resultados en el campo de los algoritmos sobre flujos de datos.

En el modelo combinatorio del problema de la contratación [5], la secuencia de candidatos se puede modelar como una permutación aleatoria. Más precisamente, los candidatos están representados por rangos relativos de acuerdo con la siguiente clasificación o esquema: el mejor tiene rango n, mientras que el peor es de rango 1, entre los n candidatos. Una decisión debe ser tomada inmediatamente ya sea para contratar o descartar el actual candidato sobre la base de su rango relativo a todos los candidatos vistos hasta el momento.

En el problema de la contratación, estamos interesados en el diseño y análisis de las estrategias de la contratación. Estudiamos en detalle dos estrategias, a saber, "la contratación por encima de la mediana" y "contratar por encima del m-ésimo mejor". En "Contratar por encima de la mediana": se contrata al primer candidato entrevistado y a partir de entonces cualquier candidato que viene es contratado si su rango relativo es mayor que la mediana de los rangos de los candidatos previamente contratados, en caso contrario se descarta a dicho candidato. "Contratar por encima del m-ésimo mejor" contrata a los primeros m candidatos en la secuencia, y acontinuación cualquier candidato que viene es contratado si su rango relativo es mayor que el m-ésima mejor entre todos candidatos contratado, en caso contrario se descarta al candidato.

Para ambas estrategias, hemos sido capaces de obtener resultados exactos y la distribución de probabilidad asintótica para varios cantidades de interés (lo que llamamos los parámetros de la contratación). Nuestra parámetro fundamental es el número de candidatos contratados. Otros parámetros incluyen el tiempo de espera, el índice de último candidato contratado y la distancia entre las dos últimas contrataciones. Estos cuatro parámetros nos dan una idea clara del ritmo de la contratación o la dinámica de el proceso de la contratación para la estrategia particular que se estudia. Hay otro grupo de parámetros como la puntuación del último candidato contratado, la puntuación del mejor candidato descartado y el número de sustituciones (al acoplar un mecanismo de reemplazo a la estrategia estudiada) nos dan una indicador de la calidad del grupo contratado. Para la estrategia de "contratar por encima de la mediana", se estudian más cantidades como el número de candidatos contratados condicionado al rango del primer candidato y la probabilidad de que el candidato con puntuación q sea contratado.

También estudiamos la regla de selección del "$\frac{1}{2}$-percentil" introducida por Krieger et al. [59] en 2007, y la distribucin de comensales en el proceso del restaurante chino (CRP) con el plan $(\frac{1}{2}, 0)$

introducido por Pitman [77]. Ambos procesos estocásticos son muy similares a "contratar por encima de la mediana". Las conexiones entre "la contratación por encima del m-ésimo mejor" y la noción de m-records (máximos de izquierda a derecha.) [6], y el plan $(0, m)$ de CRP se investiguan también.

También presentamos los resultados preliminares para el número de candidatos contratados por la generalización de "contratar por encima la mediana" llamada "contratar por encima del α-cuantil (del los candidatos contratados)", que se introduce en [5]. Nuestros resultados sobre la distributicin de probabilidad se aplican al caso $\alpha = \frac{1}{d}$, donde $d \in \mathbb{N}$. Para el caso general, $0 < \alpha < 1$, hemos sido capaces de dar el orden de crecimiento de la nmero medio de candidatos contratados, el rango medio del último candidato contratado, y el número medio de sustituciones.

Los resultados explícitos para el número de candidatos contratados nos han permitido diseñar un estimador, llamado RECORDINALITY, para el número de elementos distintos que hay en una gran secuencia de datos que pueden contener repeticiones; este problema se conoce en la literatura como "el problema de estimación de la cardinalidad" (ver [33]). RECORDINALITY tiene varias propiedades interesantes, por ejemplo es el primer algoritmo de estimación de la cardinalidad —por lo que sabemos—que, en el modelo de orden aleatorio, no necesita ni muestreo (sampling) ni usar funciones de *hashing*. El algoritmo propuesto también proporciona una muestra aleatoria de elementos distintos de la secuencia. Se demuestra que otro parámetro de contratación, la puntuación del mejor candidato descartado, también se puede utilizar para diseñar un estimador de cardinalidad, al que llamamos DISCARDINALITY. En la práctica, DISCARDINALITY no es tan interesante como RECORDINALITY, pero este nuevo parámetro puede resultar útil para abordar otros problemas tales como la "estimación del índice de la similitud" [14] entre dos documentos o conjuntos de datos.

La mayoría de los resultados presentados aquí han sido publicados o presentados para su publicación. Nuestros resultados sobre la estrategia de "contratar por encima de la mediana" en el capítulo 4 aparecen en [51, 52]. El capítulo 6 se refiere a los resultados de [48, 50] para la estrategia de "contratar por encima del m-ésimo mejor". El capítulo 7 contiene nuestras aplicaciones a los algoritmos sobre flujos de datos, publicados en [47]. Nuestros resultados en "contratar por encima de la α-cuantil" en el capítulo 5 son aún un trabajo en curso: el informe técnico [49] contiene nuestros resultados hasta el momento.

La tesis deja algunas preguntas abiertas, así como muchas ideas prometedoras para el trabajo futuro. Por ejemplo, una pregunta interesante es cómo comparar dos estrategias diferentes, que requiere de una definición adecuada de la noción de "optimalidad"; tal definición parece muy compleja en el contexto de la problema de la contratación. Además de los resultados actuales de "contratar por encima de la α-cuantil", estamos tratando de ampliar nuestro resultados al caso de cualquier valor de α racional. Esta clase de estrategias junto con "la contratación por encima del m-ésimo mejor" puede ser útil para desarrollar algoritmos de muestreo que generan muestras aleatorias de elementos distintos, muestras cuyo tamaño depende del número (desconocido) de elementos distintos en el flujo de datos. También queremos completar el análisis del parámetro número de sustituciones, del que hasta ahora sólo hemos obtenido su valor esperado para varias estrategias de la contratación. Estamos también interesados en la investigación de otras variantes del problema como podrían ser las "estrategias probabilistas de la contratación", es decir, cuando el criterio de la contratación no es determinista.

Resum

El *problema de la contractació* és un model simple per a la presa de decisions seqüencial en condicions d'incertesa, recentment introduït a la literatura. Aquest tipus de problemes es presenta en diverses camps, com les Ciències de la Computació i l'Economia. El problema va ser introduït explícitament per primera vegada per Broder et al. [15] al 2008, com una extensió natural del bén conegut problema de la secretària (vegeu [38] i les referències citades per aquest). Poc després, Archibald i Martínez [5] el 2009 van introduir un model discret combinatori del problema de la contractació, on els candidats vistos fins un moment donat podrien ser classificats de millor a pitjor sense la necessitat de conèixer les seves puntuacions de qualitat en termes absoluts. En aquesta tesi es presenta un extens estudi per al problema de la contractació sota la formulació proposada per Archibald i Martínez, s'exploren les connexions amb altres processos de selecci on seqüencials, i es desenvolupa una aplicació interessant dels nostres resultats en el camp dels algorismes sobre fluxos de dades.

En el model combinatori del problema de la contractació [5], la seqüència de candidats es pot modelar com una permutació aleatòria. Més precisament, els candidats estan representats per rangs relatius d'acord amb la següent classificació o esquema: el millor té rang n, mentre que el pitjor és de rang 1 entre els primers n candidats. Una decisió ha de ser presa immediatament ja sigui per contractar o descartar l'actual candidat sobre la base del seu rang relatiu a tots els candidats vistos fins ara.

Al problema de la contractació, estem interessats en el disseny i l'anàlisi de les estratègies de contractació. Estudiem en detall dues estratègies, a saber, la "contractar per sobre de la mitjana" i "contractar per sobre del m-èsim millor". En la estratègia"contractar per sobre de la mitjana" es contracta el primer candidat entrevistat i a partir de llavors qualsevol candidat que ve és contractat si el seu rang relatiu és major que la mitjana dels rangs dels candidats prèviament contractats, en cas contrari es descarta a aquest candidat. "Contractar per sobre del m-èsim millor" contracta els primers m candidats de la seqüència, i a continuació qualsevol candidat que ve és contractat si el seu rang relatiu és més gran que el m-èsim millor entre tots els candidats contractats, en cas contrari es descarta el candidat.

Per ambdues estratègies, hem estat capaços d'obtenir resultats exactes i la distribució de probabilitat asimptòtica per diverses quantitats d'interès (el que anomenem els paràmetres de la contractació). El nostre paràmetre fonamental és el nombre de candidats contractats. Altres paràmetres inclouen el temps d'espera, l'índex de lúltim candidat contractat i la distància entre les dues últimes contractacions. Aquests quatre paràmetres ens donen una idea clara del ritme de la contractació o dinàmica del procés de la contractació per l'estratégia particular que s'estudia. Hi ha un altre grup de paràmetres com ara el rang de l'últim candidat contractat, el rang del millor candidat descartat i el nombre de substitucions (aquest paràmetre s'estudia al acoblar un mecanisme de reemplaçament amb l'estratègia del nostre interès) ens donen indicadors de la qualitat del grup contractat. Per l'estratègia "contractar per sobre de la mitjana", estudiem altres quantitats addicionals: el nombre de candidats contractats condicionat al rang del primer candidat, i la probabilitat que el candidat amb rang q sigui contractat.

També estudiem la regla de selecció del "$\frac{1}{2}$-percentil" introduïda per Krieger et al. [59] al 2007, i la distribuci de comensals en el procés del restaurant xinès (CRP), introduït per Pitman [77], amb el pla $(\frac{1}{2}, 0)$. Tots dos processos estocàstics són molt similars a"contractar per sobre de la mitjana". Les connexions entre "contractar per sobre del m-èsim millor" i la noció de m-records (màxims d'esquerra a dreta.) [6], i el pla $(0, m)$ de CRP s'investiguen també.

També presentem els resultats preliminars per al nombre de candidats contractats per la gen-

eralització de "contractar per sobre la mitjana" anomenada "contractar per sobre de l'α-quantil (dels candidats contractats)", que s'introdueix en [5]. Els nostres resultats sobre la distributicin de probabilitat s'apliquen al cas $\alpha = \frac{1}{d}$, on $d \in \mathbb{N}$. Per al cas general, $0 < \alpha < 1$, hem estat capaços de donar l'ordre de creixement del nombre mitjà de candidats contractats, del rang mitjà de l'últim candidat contractat, i del nombre mitjà de substitucions.

Els resultats explícits per al nombre de candidats contractats ens han permés dissenyar un estimador, anomenat RECORDINALITY, per al nombre d'elements diferents que hi ha en una gran seqüència de dades que pot contenir repeticions; aquest problema es coneix a la literatura com "el problema de l'estimació de la cardinalitat" (veure [33]). RECORDINALITY té diverses propietats interessants, per exemple és el primer algorisme d'estimació de la cardinalitat—pel que sabem—que, en el model d'ordre aleatori, no necessita ni mostreig (*sampling*) ni utilitzar funcions de *hashing*. L'algorisme proposat també proporciona una mostra aleatòria d'm elements diferents de la seqüència. Es demostra que un altre paràmetre de contractació, el rang del millor candidat descartat, també es pot utilitzar per dissenyar un estimador de cardinalitat, que anomenem DISCARDINALITY. A la pràctica, DISCARDINALITY no és tan interessant com RECORDINALITY, però aquest nou parémetre pot ser útil per abordar altres problemes com ara l'estimació de l'índex de similitud [14] entre dos documents o conjunts de dades.

La majoria dels resultats presentats aquí han estat publicats o presentats per a la seva publicació. Els nostres resultats sobre l'estratègia de "contractar per sobre de la mitjana" del capítol 4 apareixen a [51, 52]. El capítol 6 conté els resultats publicats a [48, 50] per a l'estratègia de "contractar per sobre del m-èsim millor". El capítol 7 està dedicat a les nostres aplicacions als algorismes sobre fluxos de dades; bona part dels resultats van ser publicats en [47]. Els nostres resultats per a l'estratègia "contractar per sobre de l'α-quantil" del capítol 5 són encara un treball en curs presentat en l'informe tècnic [49].

La tesi deixa algunes preguntes obertes, així com moltes idees prometedores per al treball futur. Per exemple, una pregunta interessant és com comparar dues estratègies diferents, la qual cosa portaria a una noció adequada d'"optimalitat"; tal definició sembla molt complexa en el context del problema de la contractació. A més dels resultats actuals de "contractar per sobre de l'α-quantil", estem tractant d'ampliar els nostres resultats al cas de qualsevol valor d'α racional. Aquesta classe d'estratègies juntament amb "contractar per sobre del m-èsim millor" pot ser útil per desenvolupar algorismes de mostreig que generin mostres aleatòries de talla variable d'elements diferents, és a dir, mostres la talla de les quals depén del nombre (desconegut) d'elements diferents en el flux de dades. També volem completar l'anàlisi del paràmetre nombre de substitucions, del qual fins ara només hem obtingut el valor esperat per diverses estratègies de contractació. Estem també interessats en la investigació d'altres variants del problema, com podrien ser les estratègies probabilistes de contractació, és a dir, quan el criteri de contractació no és determinista.

Abstract

The hiring problem is a simple model for on-line decision-making under uncertainty, recently introduced in the literature. Despite some related work dates back to 2000, the name and the first extensive studies were written in 2007 and 2008. This kind of problems arises in various fields, like Computer Science and Economics. The problem has been introduced explicitly first by Broder et al. [15] in 2008 as a natural extension to the well-known secretary problem (see [38] and references therein). Soon afterwards, Archibald and Martínez [5] in 2009 introduced a discrete (combinatorial) model of the hiring problem, where the candidates seen so far could be ranked from best to worst without the need to know their absolute quality scores. This thesis introduces an extensive study for the hiring problem under the formulation given by Archibald and Martínez, explores the connections with other on-line selection processes in the literature, and develops one interesting application of our results to the field of data streaming algorithms.

In the combinatorial model of the hiring problem [5], there is a potentially infinite sequence of candidates that arrive sequentially. It is assumed that we can rank all candidates from best to worst without ties and all orders are equally likely. Then the sequence of candidates is modeled as a random permutation. More precisely, candidates are represented by *relative ranks* according to the following ranking scheme: the best has rank n while the worst has rank 1, among n candidates. A decision must be taken immediately either to hire or discard the current candidate based on his relative rank among all candidates previously seen. In this context, the goals for a reasonable hiring strategy are to hire candidates at some reasonable rate and to improve the average quality of the hired staff.

In the hiring problem we are interested in the design and analysis of hiring strategies. We study in detail two hiring strategies, namely "hiring above the median" and "hiring above the m-th best". Hiring above the median was introduced originally by Broder et al. [15] and processes the sequence of candidates as follows: hire the first interviewed candidate then any coming candidate is hired if and only if his relative rank is better than the median rank of the already hired staff, and others are discarded. Hiring above the m-th best was introduced by Archibald and Martínez [5] and hires the first m candidates in the sequence whatever their relative ranks, then any coming candidate is hired if and only if his relative rank is larger than the m-th best among all hired candidates, and others are discarded.

For both strategies, we were able to obtain exact and asymptotic distributional results for various quantities of interest (which we call *hiring parameters*). Our fundamental parameter is the *number of hired candidates*, together with other parameters like *waiting time*, *index of last hired candidate* and *distance between the last two hirings* give us a clear picture of the hiring rate or the dynamics of the hiring process for the particular strategy under study. There is another group of parameters like *score of last hired candidate*, *score of best discarded candidate* and *number of replacements* that give us an indicator of the quality of the hired staff. For the strategy "hiring above the median", we study more quantities like *number of hired candidates conditioned on the first one* and *probability that the candidate with score* q *is getting hired*.

We study the selection rule "$\frac{1}{2}$-percentile rule" introduced by Krieger et al. [59], in 2007, and the seating plan $(\frac{1}{2}, 1)$ of the *Chinese restaurant process* (CRP) introduced by Pitman [77], which are very similar to "hiring above the median". The connections between "hiring above the m-th best"

and the notion of m-records [6], and the seating plan $(0, m)$ of the CRP are also investigated here. Moreover, we obtain the explicit and asymptotic distributions of the parameter *number of retained items* of the "$\frac{1}{2}$-percentile rule", that completes some results already given in [59]. For both mentioned seating plans, as well as the $\frac{1}{2}$-percentile rule, we analyze a new parameter which is the *waiting time* where we characterize its probability distribution and expectation.

We report preliminary results for the *number of hired candidates* for a generalization of "hiring above the median"; called "hiring above the α-quantile (of the hired staff)", which is introduced in [5]. Our distributional and asymptotic results apply for $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$. For the general case, $0 < \alpha < 1$, we were able to give the order of growth of the expectation for the *number of hired candidates*, the *gap of last hired candidate*, and the *number of replacements*.

We also introduce one application of the results obtained for the strategy "hiring above the m-th best" to the field of data streaming. The explicit results for the *number of hired candidates* enable us to design an estimator, called RECORDINALITY, for the number of distinct elements in a large sequence of data which may contain repetitions; this problem is known in the literature as "cardinality estimation problem" (see [33]). RECORDINALITY has several interesting properties, namely, it is the first cardinality estimation algorithm—as far as we know— which, in the random-order-model, would not need neither sampling nor hashing. It also provides a random sample of distinct elements in the stream. We show that another hiring parameter, the *score of best discarded candidate*, can also be used to design a new cardinality estimator, which we call DISCARDINALITY. DISCARDINALITY is not as interesting as RECORDINALITY from a practical point of view, but the idea may help to investigate other problems such as the "similarity index estimation" [14] between two documents or data sets.

Most of the results presented here have been published or submitted for publication. Our results on the strategy "hiring above the median" in Chapter 4 appear in [51, 52]. Chapter 6 covers the results of [48, 50] for the strategy "hiring above the m-th best". Chapter 7 contains the results on applications to data streaming algorithms published in [47]. Our results on "hiring above the α-quantile" in Chapter 5 are still on-going work; the technical report [49] contains our findings so far.

The thesis leaves some open questions, as well as many promising ideas for future work. For instance, one interesting question is how to compare two different strategies; that requires a suitable definition of the notion of "optimality", which is still missing in the context of the hiring problem. Besides the current results on "hiring above the α-quantile", we are trying to extend our results to the case of any rational α. This class of strategies together with "hiring above the m-th best" may be helpful to develop sampling algorithms that generate random samples of distinct elements, whose size (of the sample) depends on the actual, but unknown, number of distinct elements in the data stream. We also wish to complete the analysis of the novel hiring parameter *number of replacements*; so far we have only obtained its expectation for several hiring strategies. Least but not last, we are interested in investigating other variants of the problem like "probabilistic hiring strategies", that is when the hiring criteria is not deterministic, unlike all the studied strategies here.

Acknowledgements

I am very grateful to the many people who have supported me and encouraged me until this dissertation becomes ready to be defended.

First of all, Conrado Martínez who has always been a gentleman, and who has left a strong finger-print not only at the scientific level but also on moral and cultural aspects. I feel very lucky to have such a friend. One who has respected my beliefs, helped me countless times without weariness, and has made me concentrate only on my research.

I would like to express my deep thanks to Alois Panholzer for hosting me in the university of Vienna in 2011 and 2012. Working with him has improved my mathematical skills a lot, and together with the bright algorithmic ideas of Conrado has helped to achieve a high-quality joint work. I have enjoyed the blackboard-meetings with Alfredo Viola. I have learned not only many technical tricks from him, but have also received good advice for my future career. I will never forget the interesting cultural and philosophical discussions with him.

A special acknowledgment is due to the friends Rosa Jimenez and Amalia Duch. They have supported and motivated me during my long stay in Barcelona. We spent very nice times during conferences and in coffee breaks in the university.

I am also thankful to all members of the Departament de Llenguatges i Sistemes Informàtics. In particular, María J. Serna whose office was always open for me, answering many scientific questions and helping to solve bureaucratic issues. Many thanks to Josep diaz, Joaquim Gabarró, Jordi Petit and Christian Blum for their efforts during teaching the courses of the Masters. Also thanks to Mercé Juan who has simplified many things and who makes bureaucracy easier.

It is worth thanking one anonymous referee for his valuable corrections and suggestions, especially his nice contribution in Theorem 6.1.

Finally, a great recognition for my dear father Helmi and my lovely mother Wedad. They have prayed for me all the time and there are no words that can reward their graces upon me. I would like also to thank my wife Amani for her great patience and non-stop assistance, and my beautiful daughter Salma who gives me the hope for tomorrow. Thanks a lot to my brothers Hassan, Hossam, Hani, and Samir, and my sisters Nelli and Heba for your support and encouragement.
I wish to thank many friends in Egypt, Spain and Vienna for their prayers and their best wishes. However, there is no enough space to mention all of you, but for sure you are all in the memory.

Funding. I would like to thank the Spanish ministry of Science and Technology to offer me an FPI grant (projects references: TIN2006-11345 and TIN2010-17254 (FRADA)) to obtain the MSc and PhD degrees. I was also supported through this scholarship to visit the Institute of Discrete Mathematics and Geometry, University of Vienna, and work with Alois Panholzer for three months in each 2011 and 2012 (six months in total).

سبحانك اللهم خير معلم
علمت بالقلم القرون الأولى
أرسلت بالتوارة موسى مرشدا
وابن البتول فعلم الانجيلا
وفجرت ينبوع البيان محمدا
فسقى الحديث وناول التنزيلا

أحمد شوقي، شاعر مصري (1868–1932)

To my father and my mother.
To my wife and my beautiful Salma.

Contents

Part I

Introduction

On-line decision-making under uncertainty is a rich area of research. It arises in diverse fields such as Computer Science and Economics. In this area, we consider processes 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 requirements of the decision maker. There are many real world and theoretical situations where decision-making under uncertainty arises. One simple such situation is selecting the maximum of a sequence where the instances of this sequence are serviced sequentially and a decision must be taken to select or discard the current instance. This model was first introduced in the early sixties as *the secretary problem* [38].
The secretary problem involves many of the main features of decision-making under uncertainty. In the secretary problem, the employer is looking for only one candidate to fill one secretarial 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 to 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 later be recalled, 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.

In the problem addressed here, *the hiring problem*, we are looking for selecting many good candidates from the input sequence instead of only one. The hiring problem has the same spirit as the secretary problem but with some major differences. One difference is that the number of candidates is unknown in the hiring problem, whereas this number is known in advance in the secretary problem. Another important difference is the measure of quality of the selection rule or strategy; this measure is clear in the secretary problem where the optimal strategy should maximize the probability of hiring the best applicant, as mentioned before. On the other hand, there are two main goals in the hiring problem: to hire candidates at some reasonable rate and to improve the mean quality of the hired staff. Due to the trade-off between these two goals (i.e., the more candidates are hired, the worse is the staff's average quality and vice-versa), the notion of optimality is not clear. But various quantities of interest (*hiring parameters*) can help to characterize the behaviour of hiring strategies. These quantities measure the hiring rate and the average quality of the hired staff. It is important to emphasize that the hiring problem cannot be regarded as an extension of the secretary problem; but rather it represents an independent and different class of sequential multiple selection, despite that it is inspired by the secretary problem and shares some common features.

History of the hiring problem. To the best of our knowledge, Preater [82] introduced in 2000, for the first time, a selection rule in the context of sequential multiple selection, namely "better-than-average rule" and considered the setup of the hiring problem, despite he did not formalize it nor give it a name. Seven years later, Krieger, Pollak and Samuel-Cahn [59] introduced a general class of selection rules called "p-percentile rules", $0 < p \leq 1$, that consider only relative ranks between candidates, thus those rules work in the *random permutation model* of the sequential multiple selection problem. Krieger et al. studied also other incarnations of the problem (see [60, 61]) where they considered different distributions of the absolute scores of candidates, introducing the "β-better-than-average rule", $\beta > 0$, a generalization of the rule given by Preater.

Soon after that, in 2008, Broder, Kirsch, Kumar, Mitzenmacher, Upfal and Vassilvitskii [15]

introduced explicitly the notion of the hiring problem, motivated by the secretary problem, and independently of the work of Krieger et al. and Preater. Broder et al. considered the absolute scores of candidates as uniformly distributed independent random variables (r.v.'s) in $(0,1)$. They introduced a reasonable class of hiring strategies; namely "Lake Wobegon strategies" which include "hiring above the mean" and "hiring above the median" strategies. Hiring above the mean behaves exactly like the better-than-average rule given by Preater, where it hires the first candidate in the sequence, then any next candidate is hired if his quality measure (absolute score) is better than the average score of all hired candidates so far. Hiring above the median cares only about the *rank* of the current candidate among all those seen so far, regardless of his absolute score; it hires candidates who are better than the current median of the hired staff.

Archibald and Martínez, in 2009, introduced in [5] the *random permutation model of the hiring problem* motivated by the work of Broder et al. [15], while the former work of Krieger et al. [59] went also unnoticed in [5]. They introduced a framework to analyze *rank-based* hiring strategies which are working in the random permutation model. They studied two hiring strategies: "hiring above the m-th best candidate" and "hiring above the α-quantile of the hired staff", $0 < \alpha < 1$. The strategy hiring above the m-th best is a selection rule closely related to records in permutations, where it hires the best $m-1$ candidates together with the m-records (see [6]) from the input sequence. The strategy hiring above the α-quantile is a generalization of "hiring above the median" (when $\alpha = \frac{1}{2}$) introduced by Broder et al.

Goals of the thesis. This thesis builds upon the combinatorial formulation of the hiring problem given by Archibald and Martínez [5], in order to fully analyze rank-based hiring strategies. A *hiring strategy* is simply an algorithm that: i) receives as input a sequence of values which represent the quality measures of candidates, ii) defines a selection criterion that determines whether an incoming candidate gets hired or discarded; the decisions can only take into account the candidates seen so far. A special class of such strategies is the so-called "pragmatic hiring strategies," in which the selection criteria is defined by what we call a *threshold candidate*: candidates who are above this threshold get hired, and others are discarded. The threshold candidate may change along the hiring process, always to increasingly better candidates.

Concerning the modeling of the sequence of candidates, then we find that there are two general models. In the first model, we have the sequence of the *absolute quality measures* or simply *absolute scores* of candidates; that requires knowing the distribution of those scores, e.g., Uniform, Normal, Exponential, etc. The second model considers the *relative ranks* of candidates without the need to have their actual absolute scores. In the later model, assuming that we can rank candidates from best to worst without ties leads to the random permutation model. Then, among n candidates, the best candidate is given rank 1 while the worst is given rank n, as in the secretary problems and the model of Krieger et al. [59].

Archibald and Martínez considered the dual (and equivalent) ranking scheme where the best candidate has rank n and the worst has rank 1, after n interviews. More formally, the input of candidates is represented by a sequence S of their initial ranks, $S = s_1, s_2, \ldots, s_i, \ldots$, with $1 \leq s_i \leq i$. The rank s_i of the i-th coming candidate is uniformly distributed on $\{1, 2, \ldots, i\}$ and independent of $s_j, j \neq i$. Then the initial prefix of length n of S represents a random permutation $\sigma^{(n)} = \left(\sigma_1^{(n)}, \sigma_2^{(n)}, \ldots, \sigma_n^{(n)}\right)$ of $\{1, 2, \ldots, n\}$. Notice that the initial rank of any candidate may remain the same or increase later depending on the ranks of the subsequent candidates. So we

say that, after processing n candidates, $\sigma^{(n)}$ represents the "final scores" (or just "scores") of candidates. More precisely, $\sigma^{(n)}$ can be obtained recursively as follows: given a permutation $\sigma^{(n-1)}$ (of size $n-1$) and a 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 $\sigma^{(n-1)}$ as $j+1, j+2, \ldots, n$, and appending j to the end. For example, let $S = 1,2,1,4,1,5,4,6$ be the input sequence of 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 $\sigma^{(8)} = 35281746$.

We study here in detail two rank-based hiring strategies: "hiring above the median" and "hiring above the m-th best". Hiring above the median processes the sequence of candidates as follows: i) hire the first coming candidate, ii) hire any candidate after that if and only if his rank is better than the current median in the set of scores hired candidates, and discard otherwise. Hired candidates are represented by the *hiring set*, $\mathcal{H}(\sigma)$ which is the set of their indices (arrival times), and $\mathcal{Q}(\sigma)$ which is the associated set of their scores. Since we are talking about the median of a set, then we have to be precise about how we define the median. In case of odd size of the hiring set, it is clear that there is one median (that is the median score in $\mathcal{Q}(\sigma)$) and it is the threshold candidate for this strategy. But if the hiring set has an even size, we can say that there are two medians and "hiring above the median" takes the lower one as its threshold candidate. Formally speaking, the median of a set of k (distinct) elements $x_1 < x_2 < \cdots < x_k$ is the ℓ-th largest element, i.e., $x_{k+1-\ell}$, with $\ell = \lceil \frac{k+1}{2} \rceil$, where $\lceil x \rceil = \min\{j \in \mathbb{Z} : j \geq x\}$ denotes the ceiling function. As an example, if we apply this strategy to the sequence $\sigma^{(8)} = \underline{3}\,\underline{5}\,2\,\underline{8}\,1\,\underline{7}\,4\,\underline{6}$, then $\mathcal{H}(\sigma^{(8)}) = \{1,2,4,6,8\}$ and $\mathcal{Q}(\sigma^{(8)})$ contains the underlined scores in $\sigma^{(8)}$. "Hiring above the α-quantile", which is a generalization of "hiring above the median" (corresponds to $\alpha = \frac{1}{2}$), is also considered in the thesis.

Hiring above the m-th best processes the sequence of candidates as follows: i) hire the first coming m candidates, ii) hire any next candidate if and only if his rank is larger than the m-th largest hired one, and discard otherwise. For example, let $m = 3$, then, processing the sequence $\sigma^{(8)} = \underline{4}\,\underline{6}\,\underline{1}\,\underline{7}\,3\,\underline{5}\,2\,\underline{8}$ using this strategy results in $\mathcal{H}(\sigma^{(8)}) = \{1,2,3,4,6,8\}$ and $\mathcal{Q}(\sigma^{(8)})$ contains the underlined scores in $\sigma^{(8)}$. Thus the threshold candidate for this strategy is what is known in the literature as a `Type2` m-record [6] (which has been studied several times under the name m-record, see for example [13, 83]), and $\mathcal{Q}(\sigma)$ consists of the $m-1$ largest scores together with the set of m-records in the input sequence.

Each hiring strategy has its unique hiring criteria which determines its potential of hiring candidates. For hiring above the median, as the size of the hiring set grows, the choices of hiring the next candidate increase. This is the case also for the class of hiring above the α-quantile strategies and the p-percentile rules in [59]. But for hiring above the m-th best, there are always m choices for hiring a new candidate at any step, after the first m interviews, regardless of the number of hired candidates so far. However, in all these strategies, the hiring threshold rises all the time and never goes down, that is, the score of the threshold candidate always increases during the hiring process. In fact, this property holds also for other hiring strategies like hiring above the mean and the β-better-than-average rules in general (which are non rank-based strategies). The strategies that have such property were called "locally subdiagonal" (LsD) by Krieger et al. [59] and, later, "pragmatic" by Archibald and Martínez [5].

When analyzing some hiring strategy we care about the behaviour of the strategy from the point of view of the hiring rate and the quality of the hired staff. So that we introduce several *hiring parameters* that describe the hiring process. The most important and fundamental parameter is the *number of hired candidates* or *size of the hiring set*, denoted by the r.v. h_n, which characterizes

the hiring rate of the applied strategy. The hiring rate can be studied also from the dual point of view; that is, the number of interviewed candidates in order to hire exactly N candidates, which we call the *waiting time*, W_N. This group of *dynamics indicators* contains also the *index of last hired candidate* or time of last hiring, L_n, and the *distance between last the two hirings*, Δ_n, which denotes the number of interviews between the last two hirings.

Another group of hiring parameters relates to the quality of the hired staff, thus we call them *quality indicators*. The *score of last hired candidate*, R_n, is the score of the last hired candidate after processing n candidates. This parameter is directly related to the *gap of last hired candidate* parameter, $g_n = 1 - \frac{R_n}{n}$, a normalization of R_n which is convenient for the case when we assume the random permutation model. The *score of best discarded candidate*, M_n, denotes the maximum score that is not contained in $\mathcal{Q}(\sigma^{(n)})$ after processing the whole sequence of candidates. This quantity describes how selective the hiring strategy is (thus yielding a quality measure for the hired staff). Another parameter is the *number of replacements*, f_n, a quantity naturally arising when we consider one interesting variant of the hiring problem, namely *hiring with replacements*, that is, when candidates can be hired directly using the applied strategy, hired to replace some previously hired candidate, or discarded. f_n combines the dynamical and quality aspects of the hiring process, because a good hiring strategy should use fewer replacements to gather exactly the h_n best candidates.

As mentioned in our short review of the literature related to the hiring problem, Krieger et al. introduced a similar work to ours here. Moreover, there is an another process which has a similar setup as the hiring problem: the Chinese restaurant process (CRP) introduced by Pitman [77]. In the CRP, a class of probabilistic rules that work in the so-called two-parameter model, called "seating plans", are analyzed. We find that the seating plan $(0, m)$ is equivalent to the strategy "hiring above the m-th best", while the seating plan $(\frac{1}{2}, 1)$ is very close to the strategy "hiring above the median" although not equivalent. We have used our methods to obtain new results for both seating plans $(0, m)$ and $(\frac{1}{2}, 1)$.

We have also been interested in applications of the hiring problem. As a first result, the algorithmic ideas and the results obtained for "hiring above the m-th best" have been very useful in the design and analysis of algorithms to estimate the cardinality of a stream and other common tasks in data streaming analysis.

Overview of our approach. As mentioned before, Archibald and Martínez [5] introduced a combinatorial framework to analyze rank-based hiring strategies. Their choice was reasonable because sequences are the fundamental structures in on-line decision-making. Then many techniques from Analytic Combinatorics [37] are useful here. It turned out that treating the quantities of interest directly using the framework in [5] is a bit complicated for either hiring above the m-th best or hiring above the median strategies. For hiring above the m-th best, simple reasonings from the definition of the studied parameters are enough in some cases to carry out the distributional analysis, while for other parameters like score of best discarded candidate we need to define auxiliary quantities to obtain the results.

For hiring above the median, the key solution is to keep track of the *median* of $\mathcal{Q}(\sigma)$ during the hiring process. Then we make use of a simple but quite useful observation in [5], that, when applying hiring above the median (and any pragmatic hiring strategy), at each time of the hiring process all candidates seen so far with a score larger than the current threshold candidate must be

part of the hiring set (this has also been proven by Krieger et al. in [59]). Thus there is a simple relation between the score of the threshold candidate and the size of the hiring set. And this yields the basis of the recursive approach used, where we thus have to distinguish cases according to the parity of the size of hiring set and to take into account the score of the threshold candidate.

We setup an automaton that describes the underlying Markov chain of the transition probabilities during the hiring process and switching from odd to even number of hired candidates and vice-versa. Then we can write down easily the recurrences for two fundamental quantities: $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$, which give the probabilities that, after interviewing n candidates, the threshold candidate has the ℓ-th largest score amongst all candidates seen so far and an odd or even number of candidates has been hired, respectively. Then the results for the *number of hired candidates* follow directly, and the remaining parameters are obtained by studying extensions of this approach. It is natural that those recurrences can be translated into systems of linear partial differential equations (PDEs) for the corresponding generating functions, but it seems very complicated to get a closed form solution. To avoid this, we use a trick (introduced originally in [54]) which is finding suitable normalization factors of the studied recursive sequences, such that the system of differential equations reduces to a first order linear PDE. Finally adapting the initial conditions carefully gives us the desired results. We still make use of this approach in studying the general case, hiring above the α-quantile. In case of $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$, the results follow easily, but for $\alpha = \frac{p}{q}$ with $\gcd(p, q) = 1$, it becomes more complicated. We have used the systematic approach given in [5] to obtain some useful information for general α, $0 < \alpha < 1$, in particular the order of growth of the expectation of many hiring parameters.

We have explicit results for the probability distributions of many parameters for the studied hiring strategies. Basic techniques like Stirling's formula, Euler-Maclaurin formula and Curtiss' theorem [22] for the weak convergence of random variables, are enough to study the asymptotic regime for many parameters when $n \to \infty$. In the case of hiring above the m-th best, we have the additional parameter, m which we call "rigidity" of hiring. Thus, the asymptotic behaviour of the studied parameters depends on the relation between n and m. We study different regimes: m is fixed (i.e., $m = \Theta(1)$) and $n \to \infty$, and other cases in which we might stop the hiring process after some number of interviews n, where n depends on m. For example, $m = \lceil \log n \rceil$, $m = \lceil \sqrt{n} \rceil$, or $m = \lceil \alpha n \rceil$ with fixed $0 < \alpha < 1$. Here $m \to \infty$ (and thus also $n \to \infty$).

Contributions of the thesis

This dissertation is devoted to the analysis and applications of the hiring problem. We give a detailed study for "hiring above the median strategy" introduced originally in [15] and "hiring above the m-th best candidate strategy" introduced in [5]. We give also interesting applications of some results obtained for hiring above the m-th best in the field of data streaming algorithms.

We give explicit distributional results for many hiring parameters like number of hired candidates, waiting time to hire N candidates, index of last hired candidate, distance between the last two hirings, score of best discarded candidate and number of replacements, for both strategies. For example, we show that the number of hired candidates under hiring above the median, has the expectation: $\mathbb{E}\{h_n\} = \sqrt{\pi n} + O(1)$, and a suitable normalization of h_n has a limit distribution which is a *Rayleigh* distribution with parameter $\sqrt{2}$, i.e.,

$$\frac{h_n}{\sqrt{n}} \xrightarrow{(d)} \hat{R} \sim \text{Rayleigh}\left(\sqrt{2}\right),$$

where $\hat{R}$ has density:

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

For hiring above the m-th best, the expectation of the number of hired candidates is

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

where H_n denotes the n-th harmonic number of first order and the asymptotic estimate holds uniformly for $1 \le m \le n$ and $n \to \infty$. In the main region $n - m \gg \sqrt{n}$; a suitably normalization of $h_{n,m}$ is asymptotically standard *Normal* (d refers to weak convergence):

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

More results are obtained for hiring above the median like the *number of hired candidates conditioned on the first one*, this r.v. is interesting since this strategy is sensitive to the first candidate in the sequence, and the *probability that the candidate with score* q *is getting hired* which gives some indication of the quality of the hired staff. In most cases we were able to give also the corresponding limiting distributions of those parameters.

We have been successful to obtain the distributional and the asymptotic results for h_n for the strategy "hiring above the α-quantile" with rational $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$, together with the order of growth of the expectation of h_n, g_n and f_n for the general case, $0 < \alpha < 1$.

As a fruit of our recursive approach for the study of the strategy "hiring above the median", we were able to complete a previous work in [59] where we characterize explicitly the distribution of the main quantity there, which is the *number of selected items* for the "$\frac{1}{2}$-percentile rule". Results for other quantities, like the *waiting time*, also become in hand.

Moreover, we explore the relationship between hiring above the median and the seating plan $(\frac{1}{2}, 1)$. We add some results to those in [77] related to the main parameter studied there, which is the *number of occupied tables after receiving* n *customers,* K_n. A normalization of K_n converges, as $n \to \infty$, to a *Maxwell-Boltzmann* distribution with parameter $\sqrt{2}$. We give also novel results for the *waiting time* parameter for the seating plan $(\frac{1}{2}, 1)$: its explicit distribution, expectation and limiting distribution.

The hiring set (and thus the set of scores of hired candidates) under the strategy "hiring above the m-th best" is closely related to m-records. The results obtained for this strategy are of interest in the context of statistics of m-records and vice-versa. The connection between this strategy and the seating plan $(0, m)$ of the CRP is also presented, as well as some novel results for the seating plan $(0, m)$, namely, the explicit distribution and the expectation for the *waiting time*.

Another set of contributions are the applications of some of our results to data stream algorithms. We were able to make use of the explicit probability distribution of the *number of hired candidates* for "hiring above the m-th best" to derive a new cardinality estimator of the number of distinct elements in a large data sequence that may contain repetitions. This is known as "cardinality estimation problem" (see [33]). Our approach to study this problem is novel, as our estimator is the first one that exploits the random-order model. The new cardinality estimator, called RECORDINALITY

does not need neither the use of hash functions nor sampling. [1] We show also that our results for other hiring parameters might be useful in this context. We introduce another cardinality estimator, called DISCARDINALITY, that is built upon the parameter *score of best discarded candidate*, again using hiring above the m-th best. In practice, DISCARDINALITY is less interesting than RECORDINALITY, but the ideas behind its design might be useful for the *similarity index estimation* (or "Jaccard similarity") [14] of two data sets, which is another interesting problem in the data streaming field.

In the conclusions we discuss some promising lines of research that we have left open, and others which are still on-going work. One interesting question is how to compare two different hiring strategies; that requires a suitable definition of the notion of optimality, which is still missing in the context of the hiring problem. Besides the preliminary results obtained for the strategy "hiring above the α-quantile" for rational $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$, we wish to continue studying more general cases of this strategy, e.g., rational $\alpha = \frac{p}{q}$ where gcd(p,q)=1. The general most case, with irrational α is a more challenging problem, and our combinatorial approach breaks down. We aim also to complete the analysis of the important parameter *number of replacements*; we could only derive its expectation for the studied strategies. In the context of applications, we are investigating sampling algorithms that generate random samples of distinct elements, whose size (of the sample) depends on the actual, but unknown, number of distinct elements in the data stream.

Other natural variants or extensions of the hiring problem might be worth studying, like "probabilistic hiring", in which the hiring strategy uses randomness to make decisions, i.e., determining the threshold candidate probabilistically (in contrast to all previously studied strategies here). One important challenge is that most probabilistic strategies are not pragmatic. Other extensions that might be worth being analyzed include "hiring with sliding-window", in which the final decision to hire or discard some candidate is delayed until the next $w-1$ candidates are interviewed. Also "multicriteria hiring", in which each candidate has more than one quality measure.

Organization of this document

This dissertation is structured into three main parts: Part II reviews all necessary mathematical techniques and the previous work on the hiring problem and related problems. Part III contains the main contributions of this thesis. Part IV presents the conclusions of the work done and discusses the open problems and future work.

Part II includes two chapters: Chapter 1 introduces some mathematical background covering the main ingredients of the analysis of combinatorial structures, i.e., the symbolic method, generating functions and singularity analysis. Chapter 2 reviews in some depth the history of the hiring problem in the literature. We summarize there the work of Krieger et al., Broder et al., and Archibald and Martínez, where we highlight the formulation of the problem in each work, the approach used in the analysis and their main results. We devote a section also to define the Chinese restaurant process, explain the similarities with the hiring problem, and recall the important results for this problem.

Part III starts with Chapter 3, where we formally define the model of the hiring problem that we will focus on, and the various hiring parameters analyzed in this work. Chapter 4 is devoted

[1] It can benefit from the use of hash functions. In particular, by using hash functions, we need not assume the random-order model.

to the analysis and results of the strategy hiring above the median. It presents the distributional results for the studied parameters, a detailed explanation of the recursive approach used in the analysis, and the proofs of all results, ending with a thoroughly discussion of the relationship of hiring above the median with the $\frac{1}{2}$-percentile rule of Krieger et al. and with the seating plan $(\frac{1}{2}, 1)$ of the CRP. It presents some new results for both the $\frac{1}{2}$-percentile rule and the seating plan $(\frac{1}{2}, 1)$. Most of the results in Chapter 4 appear in the following publications,

- [52] A. Helmi and A. Panholzer. Analysis of "hiring above median" selection strategy for the hiring problem. *Algorithmica*, pages 1-42, 2012.
- [51] A. Helmi and A. Panholzer. Analysis of "hiring above the median": a "Lake Wobegon" strategy for the hiring problem. In *Proceedings of the ACM-SIAM Meeting on Analytic Algorithmics and Combinatorics (ANALCO'12)*, 75-83, 2012.

Chapter 5 is devoted to our study of the strategy hiring above the α-quantile. The following technical report contains the current results and the analysis of this strategy:

- [49] A. Helmi, C. Martínez, and A. Panholzer. Analysis of the "hiring above the α-quantile" strategy. *Technical report*, LSI-12-15-R, 2012.

Chapter 6 is dedicated to the analysis of the strategy hiring above the m-th best. It shows the relationship between this strategy and m-records, contains our results for the studied hiring parameters, and discusses the equivalence of this strategy and the seating plan $(0, m)$ of the CRP. It also gives some novel results for the seating plan $(0, m)$. The following two publications contain the main results that appear in Chapter 6:

- [50] A. Helmi, C. Martínez, and A. Panholzer. Hiring above the m-th best candidate: a generalization of records in permutations. In D. Fernndez-Baca, editor, *Proceedings of the 10th Latin American Symposium on Theoretical Informatics (LATIN'12)*, volume 7256 of LNCS, pages 470-481. Springer, Berlin, Heidelberg, 2012.
- [48] A. Helmi, C. Martínez, and A. Panholzer. Analysis of "hiring above the m-th best candidate strategy", 2012. Submitted to *Algorithmica*.

Finally, Chapter 7 is devoted to the applications of the results obtained in Chapter 6 to the field of data streaming algorithms. In particular we concentrate in the cardinality estimation problem. We describe the estimator, RECORDINALITY, and give its precise analysis, showing that it is unbiased and quantify its accuracy. We also report experimental results for this estimator. This chapter also presents the other estimator, DISCARDINALITY, and its full analysis. We also give preliminary results on how the same underlying hiring parameter (*score of best discarded candidate*) can be used for the estimation of the similarity index of two data sets. The following publication contains the main results about RECORDINALITY:

- [47] A. Helmi, J. Lumbroso, C. Martínez, and A. Viola. Data Streams as Random Permutations: the Distinct Element Problem. In *DMTCS Proceedings, the 23rd International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, number 1, 2012.

Part IV ends the thesis with the conclusions of the work done, as well as our preliminary results and on-going work for some extensions and generalizations related to the hiring problem; we would like to investigate these problems in the future. Part IV also discusses some interesting open problems in connection with the hiring problem.

Part II

Preliminaries and Previous Work

Chapter 1

Mathematical preliminaries

In this chapter we review briefly the main ingredients of Analytic Combinatorics, as our study of the hiring problem is essentially combinatorial. We discuss the symbolic method and the principles of generating functions in Section 1.3. Generating functions are the main tool of the analysis in the framework given later in Section 2.4. We will use them also in the context of the recursive approach used in Chapter 4 to translate the obtained recurrences for the quantities of interest into differential equations. We rely on the generating functions again in Chapter 5 when we discuss the strategy "hiring above the α-quantile". We will use also singularity analysis in Chapter 5 to extract the desired information like the expectation of the parameters studied, so the main theorems of this technique are discussed in Section 1.4. We start with the following section that reviews some mathematical background necessary for our analysis.

1.1 Background and notation

The material of this section can be found in [28, 29, 57] and others.

1.1.1 Probability distributions

Table 1.1 shows the notation used for many distributions and their density functions.

Distribution	Notation	pdf
Beta	Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0 \le x \le 1$
Exponential	Exp(λ)	$\lambda e^{-\lambda x}$, $x \ge 0$
Geometric	Geom(p)	$(1-p)^{x-1}p$, $x = 1, 2, \ldots$
Normal	$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$, $x \in \mathbb{R}$
Poisson	Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, \ldots$
Rayleigh	Rayleigh(α)	$\frac{x}{\alpha^2} e^{-\frac{-x^2}{2\alpha^2}}$, $x > 0$
Uniform	Unif(a,b)	$\frac{1}{b-a}$, $x \in (a, b)$

Table 1.1: Notation used for different distributions. "pdf" refers to probability density function.

1.1.2 Curtiss' theorem

We use the following theorem [22] in the proof of the convergence of random variables (r.v.'s) using the moment generating function approach,

Theorem 1.1 (Curtiss, 1942) *Let $F_n(x)$ and $G_n(\alpha)$ be respectively the pdf and the MGF of a variate X_n. If $G_n(\alpha)$ exists for $|\alpha| < \alpha_1$ and for all $n \geqq n_o$, and if there exists a finite-valued function $G(\alpha)$ defined for $|\alpha| \leqq \alpha_2 < \alpha_1$, $\alpha_2 > 0$, such that $\lim_{n\to\infty} G_n(\alpha) = G(\alpha)$, $|\alpha| \leqq \alpha_2$, then there exists a variate X with pdf $F(x)$ such that $\lim_{n\to\infty} F_n(x) = F(x)$ at each continuity point and uniformly in each finite or infinite interval of continuity of $F(x)$. The MGF of X exists for $|\alpha| \leqq \alpha_2$ and is equal to $G(\alpha)$ in that interval.*

1.1.3 Convergence of random variables

There are three cases of convergence of r.v.'s (see [12]). For a sequence of r.v.'s X_n and a r.v. X:

- If $\lim_{n\to\infty} \mathbb{P}\{X_n \leq t\} = \mathbb{P}\{X \leq t\}$, for all points of continuity $t \in \mathbb{R}$, then it is said that X_n converges *in distribution* (or converges in distribution, or converges in law) to X, and we write $X_n \xrightarrow{(d)} X$ to denote it.
- If $\lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \epsilon\} = 0$, for all $\epsilon > 0$, then it is said that X_n converges *in probability* to X and we write $X_n \xrightarrow{(P)} X$ to denote it.
- If $\mathbb{P}\{\lim_{n\to\infty} X_n = X\} = 1$, then it is said that X_n converges *almost surely* (or almost everywhere) to X and we write $X_n \xrightarrow{(a.s.)} X$ to denote it.

Moreover, suppose that $g : \mathbb{R} \to \mathbb{R}$ is a continuous function, then it is true that

$$X_n \xrightarrow{(d)} X \quad \Rightarrow \quad g(X_n) \xrightarrow{(d)} g(X), \tag{1.1}$$

and for other cases above also.

1.1.4 Stirling's formula for the factorials

This formula is our main tool in the asymptotic analysis:

$$\log x! = \left(x + \frac{1}{2}\right)\log x - x + \frac{1}{2}\log(2\pi) + \mathcal{O}\left(\frac{1}{x}\right), \tag{1.2}$$

used together with the asymptotic expansion of the logarithmic function for small values of x,

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \mathcal{O}(x^4), \qquad \text{as } x \to 0. \tag{1.3}$$

1.1.5 Euler-Maclaurin formula

The most useful form of this formula, since we use it to obtain asymptotic expansions of sums, is the following:

$$\sum_{k=a}^{b} f(k) \sim \int_a^b f(x)dx + \frac{f(a)+f(b)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{2k!}\left(f^{(2k-1)'}(b) - f^{(2k-1)'}(a)\right), \tag{1.4}$$

where $f(k)$ is a smooth function (i.e., continuously differentiable), a and b are integers, and B_k are Bernoulli numbers.

1.1.6 Unsigned Stirling numbers of the first kind

These special numbers are denoted by $\begin{bmatrix} n \\ k \end{bmatrix}$ and count the number of permutations of size n with exactly k cycles. The following recurrence holds for $n > 0$:

$$\begin{bmatrix} n+1 \\ k \end{bmatrix} = n \begin{bmatrix} n \\ k \end{bmatrix} + \begin{bmatrix} n \\ k-1 \end{bmatrix}, \tag{1.5}$$

with the initial conditions: $\begin{bmatrix} n \\ n \end{bmatrix} = \begin{bmatrix} n \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ n \end{bmatrix} = 1$ and $\begin{bmatrix} n \\ k \end{bmatrix} = 0$ for $k > n$. It turned out that $\begin{bmatrix} n \\ k \end{bmatrix}$ also count the number of permutations of size n with exactly k left-to-right maxima. Those numbers have the following useful identity:

$$\sum_{j=0}^{N} \begin{bmatrix} N \\ j \end{bmatrix} z^j = z(z+1)\dots(z+N-1) = z^{\overline{N}}, \tag{1.6}$$

where $z^{\overline{N}}$ denotes the *rising factorial* (defined later in Subsection 1.1.9).

1.1.7 Stirling numbers of the second kind

These numbers are denoted by $\begin{Bmatrix} n \\ k \end{Bmatrix}$ and count the number of ways to partition a set of n objects into k non-empty subsets. They obey the following recurrence for $n > 0$:

$$\begin{Bmatrix} n+1 \\ k \end{Bmatrix} = k \begin{Bmatrix} n \\ k \end{Bmatrix} + \begin{Bmatrix} n \\ k-1 \end{Bmatrix},$$

with the initial conditions: $\begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = 1$ and $\begin{Bmatrix} n \\ 0 \end{Bmatrix} = \begin{Bmatrix} 0 \\ n \end{Bmatrix} = 0$ for $n > 0$. For $k > n$, $\begin{Bmatrix} n \\ k \end{Bmatrix} = 0$. They can also be computed using this explicit formula:

$$\begin{Bmatrix} n \\ k \end{Bmatrix} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n. \tag{1.7}$$

1.1.8 Special functions

Complete and *Incomplete Gamma functions* are defined as

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt \quad \text{and} \quad \Gamma(s,x) = \int_x^\infty t^{s-1} e^{-t} dt, \quad \text{respectively.}$$

The *ceiling function* is defined as

$$\lceil x \rceil = \min \{ j \in \mathbb{Z} : j \geq x \},$$

that is the smallest integer greater than or equal to x. The *floor function* is defined as

$$\lfloor x \rfloor = \max \{ j \in \mathbb{Z} : j \leq x \},$$

that is the greatest integer smaller than or equal to x.

1.1.9 Other notation

Harmonic numbers are denoted by $H_n = \sum_{k=1}^{n} \frac{1}{k}$ and $H_n^{(r)} = \sum_{k=1}^{n} \frac{1}{k^r}$ denotes the r-*th order harmonic numbers*. *Natural logarithm* (with base e) is always denoted by $\log n$. The r-*th falling factorial*, for $r \geq 0$, is denoted by $x^{\underline{r}} = x(x-1)\dots(x-r+1)$. The *rising factorial*, for $r \geq 0$, is denoted by $x^{\overline{r}} = x(x+1)\dots(x+r-1)$. The multifactorial of x is denoted by $x!^{(r)}$, defined as

$$x!^{(r)} = \begin{cases} 1, & \text{if } 0 \leq x < r, \\ x \cdot \left((x-r)!^{(r)}\right), & \text{if } x \geq r. \end{cases} \tag{1.8}$$

The *greatest common divisor* of two integers p and q, is denoted by $\gcd(p, q)$. *Iverson's bracket notation* is defined as $[\![P]\!]$, and $[\![P]\!]$ evaluates to 1 if the predicate P is true and to 0 otherwise.

1.2 Analytic Combinatorics

Analytic Combinatorics [37] studies combinatorial (finite) structures using methods from complex and asymptotic analysis. It involves two main fields. The first is *combinatorial enumeration* which is concerned with the enumeration of combinatorial structures (the number of structures of some given size n). The corner stone of this field is the symbolic method and generating functions. The second field is *complex analysis* which presents the mathematical tools, like singularity analysis, required to extract the asymptotic behaviour of the coefficients of generating functions. Complex analysis helps also to obtain asymptotic estimates of complex quantities in terms of elementary functions, the most famous example is Stirling's formula

$$n! \sim n^n e^{-n} \sqrt{2\pi n}.$$

We introduce the following example to show the interplay between combinatorics and analysis. Suppose we are interested in C_n, the number of binary trees that have n internal nodes, hence $n+1$ external nodes. First, we write the combinatorial equation for the class of binary trees as follows

$$\mathcal{C} \quad = \quad \{\square\} \quad \uplus \quad \mathcal{C} \times \{\bullet\} \times \mathcal{C}, \quad \textit{where } \uplus \textit{ is disjoint union.} \tag{1.9}$$

which reflects the definition of a binary tree: any binary tree is the empty tree or two binary trees attached to one root.
The next step is to define the following generating function,

$$C(z) = \sum_{n \geq 0} C_n z^n \quad = \quad \sum_{t \in \mathcal{C}} z^{|t|}.$$

The symbolic method allows us to translate the combinatorial equation (1.9) into a functional equation

$$C(z) \quad = \quad 1 \quad + \quad zC(z)^2,$$

whose solution is

$$C(z) = \frac{1 - \sqrt{1-4z}}{2z}.$$

Then, by means of Newton's theorem, one finds easily the closed form expression

$$C_n = \frac{1}{n+1} \binom{2n}{n}.$$

Now, Stirling's asymptotic formula gives us the following approximation

$$C_n \sim C_n^\star \quad \text{where} \quad C_n^\star = \frac{4^n}{\sqrt{\pi n^3}}.$$

It is worth mentioning that C_n are the most famous numbers in Combinatorics. They are known as Catalan numbers and there are about 66 different types of combinatorial structures that are enumerated by the Catalan numbers [87].

1.3 Symbolic method

We can think of the symbolic method [37] as a general approach to translate the set-theoretic and algorithmic operations on the combinatorial structures into functional equations over generating functions. The derivation of such equations is made by applications of translation rules that establish a "one-to-one" correspondence between algorithmic and set-theoretic constructions and operators over generating functions. The coefficients of the generating functions represent the quantities that we want to analyze.

Definition 1.1 *The counting sequence of a combinatorial class* $(\mathcal{A}_n)$ *is the sequence of integers* $(A_n)_{n\geq 0}$ *where* $A_n = card\ (\mathcal{A}_n)$ *is the number of objects in the class* $\mathcal{A}$ *that have size* n.

Definition 1.2 *The ordinary generating function* (OGF) *of a sequence* (A_n) *is the formal power series*

$$A(z) = \sum_{n\geq 0} A_n z^n.$$

The variable z is purely formal; $A(z)$ is also a formal object which can be manipulated algebraically; we will not be concerned about convergence at this step.
This generating function is called ordinary generating function, to distinguish it from the exponential generating function, the Dirichlet generating function, etc. Ordinary generating functions usually enumerate unlabelled structures, while exponential generating functions enumerate labelled structures. Also, the OGF of class $\mathcal{A}$ admits the following combinatorial form

$$A(z) = \sum_{\alpha\in\mathcal{A}} z^{|\alpha|}.$$

Definition 1.3 *The exponential generating function* (EGF) *of a sequence* (A_n) *is the formal power series*

$$A(z) = \sum_{n\geq 0} A_n \frac{z^n}{n!}.$$

But the full power of generating functions comes when we consider their dual nature. First they are considered as a formal power series. Second, when we see them as functions of complex variable in the complex plane, analytic in a disk around the origin.

The n-th coefficient of a generating function $F(z)$ will be denoted by $[z^n]F(z)$. Thus, $F(z) = \sum_{n\geq 0} f_n z^n$ implies $[z^n]F(z) = f_n$. We can expand a generating function using Taylor's expansion theorem. Expanding here means finding the associated sequence.

We show also that elementary operations over sequences translate to operations over the corresponding generating functions. For example, let $A(z)$, $B(z)$ and $C(z)$ be the OGFs for the sequences a_n, b_n and c_n respectively, then

i) $c_n = a_n \pm b_n \quad \Longrightarrow \quad C(z) = A(z) \pm B(z)$.

ii) $c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad \Longrightarrow \quad C(z) = A(z) \cdot B(z)$.

iii) $c_n = a_{n-1} \quad \Longrightarrow \quad C(z) = zA(z)$.

iv) $c_n = a_{n+1} \quad \Longrightarrow \quad C(z) = (A(z) - A(0))/z$.

v) $c_n = na_n \quad \Longrightarrow \quad C(z) = z\frac{d}{dz}A(z)$.

vi) $c_n = a_{n-1}/n \quad \Longrightarrow \quad C(z) = \int_0^z A(t)dt$.

Operations i), iii) and iv) are known as sum, backward shift and forward shift; (ii) is called the convolution of sequences. Differentiation and integration are specified by (v) and (vi).

Generating functions are also useful to study many interesting constructions of combinatorial classes, such as Cartesian product, disjoint union, sequences, multisets and powersets. These different constructions translate to operators over generating functions. Now, we need to define what is a class of combinatorial structures and an admissible combinatorial construction.

Definition 1.4 *A combinatorial class is a finite or denumerable set on which a size function is defined, satisfying the following conditions:*

i) the size of an element is a non-negative integer;

ii) the number of elements of any given size is finite.

Definition 1.5 *A combinatorial construction Φ is admissible if there exists an operator Ψ over generating functions such that:*

$$\mathcal{A} = \Phi(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k) \implies A(z) = \Psi(C_1(z), C_2(z), \ldots, C_k(z)),$$

where $A(z), C_1(z), \ldots, C_k(z)$ are the counting generating functions corresponding to the classes $\mathcal{A}, \mathcal{C}_1, \ldots, \mathcal{C}_k$.

We are not going to define all of the admissible constructions, but we give only two of them in the following definitions, by a way of example.

Definition 1.6 *A class $\mathcal{C}$ is the Cartesian product of $\mathcal{A}$ and $\mathcal{B}$, denoted $\mathcal{C} = \mathcal{A} \times \mathcal{B}$ if*

1. *$\mathcal{C} = \mathcal{A} \times \mathcal{B}$ (in the set-theoretic sense)*
2. *$|(a, b)|_{\mathcal{C}} = |a|_{\mathcal{A}} + |b|_{\mathcal{B}}$.*

Definition 1.7 *A class $\mathcal{C}$ is the sequence class of $\mathcal{A}$, denoted $\mathcal{C} = \mathcal{A}^\star$ if*

$$\mathcal{C} = \{\epsilon\} + \mathcal{A} + (\mathcal{A} \times \mathcal{A}) + (\mathcal{A} \times \mathcal{A} \times \mathcal{A}) + \cdots$$

where ϵ is the empty structure (of size 0), in other words

$$\mathcal{C} = \left\{(\beta_1, \ldots, \beta_\ell) \quad | \quad \ell \geq 0, \quad \beta_j \in \mathcal{A}\right\}.$$

Then, the following theorem introduces the basic admissible combinatorial constructions and the corresponding operators on OGFs.

Theorem 1.2 *For unlabelled structures, these are the basic admissible constructions and the associated operators over generating functions:*

Disjoint union: $\mathcal{A} = \mathcal{B} \uplus \mathcal{C} \quad \Longrightarrow \quad A(z) = B(z) + C(z)$,

Cartesian product: $\mathcal{A} = \mathcal{B} \times \mathcal{C} \quad \Longrightarrow \quad A(z) = B(z) \cdot C(z)$,

Sequence: $\mathcal{A} = \mathrm{SEQ}(\mathcal{B}) \quad \Longrightarrow \quad A(z) = \frac{1}{1-B(z)}$,

Powerset: $\mathcal{A} = \mathrm{PSET}(\mathcal{B}) \quad \Longrightarrow A(z) = \exp\Big(\sum_{k\geq 1} \frac{(-1)^{k-1}}{k} B(z^k)\Big)$,

Multiset: $\mathcal{A} = \mathrm{MSET}(\mathcal{B}) \quad \Longrightarrow A(z) = \exp\Big(\sum_{k\geq 1} \frac{1}{k} B(z^k)\Big)$.

A simple application to the SEQ construction is binary words. Let $\mathcal{B} = \{0, 1\}$, then its OGF is $B(z) = 2z$, since it has two objects each of size 1. Now, if we define a class $\mathcal{A}$ as the class of binary words, then $A(z) = 1/(1-2z)$ which gives us the number of binary words of length n, 2^n.

The symbolic method can also be applied to study random structures. We are interested in the probability that a certain measure X has a value k for a random structure of size n. A new type of generating function, the multivariate generating functions (MGFs) is useful here. MGFs keep track of a collection of parameters defined over combinatorial structures. MGFs allows us to derive results about probability distribution or, at least, mean and variance.

BGF (bivariate generating function) are a particularly important instance of MGFs, when we have two formal variables. They are defined as follows

Definition 1.8 *Given a doubly indexed sequence* $\{a_{nk}\}$*, the function*

$$A(z, u) = \sum_{n\geq 0} \sum_{k\geq 0} a_{nk} u^k z^n$$

is called the bivariate generating function (BGF) *of the sequence. We use the notation* $[u^k z^n]A(z, u)$ *to refer to* a_{nk}*;* $[z^n]A(z, u)$ *to refer to* $\sum_{k\geq 0} a_{nk} u^k$*; and* $[u^k]A(z, u)$ *to refer to* $\sum_{n\geq 0} a_{nk} z^n$.

If a_{nk} denotes the number of combinatorial objects of $\mathcal{A}_n$ such that some parameter or cost measure is k, then we can write

$$A(z, u) = \sum_{\alpha\in\mathcal{A}} u^{\mathrm{cost}(\alpha)} z^{|\alpha|}.$$

Thus we say that variable z marks the problem size, while variable u marks the value of the parameter being analyzed.
Now we can use BGF to compute moments and the average cost.

Definition 1.9 *Let* $\mathcal{P}$ *be a class of combinatorial structures with* BGF $P(z, u)$*. Then the function*

$$p(z) = \frac{\partial P(z, u)}{\partial u}\Big|_{u=1} = \sum_{p\in\mathcal{P}} \mathit{cost}(p) z^{|p|}$$

is defined as the cumulative generating function (CGF) *for the cost measure. Also, let* $\mathcal{P}_n$ *denote the class of all the structures of size* n *in* $\mathcal{P}$*. Then the sum*

$$\sum_{p\in\mathcal{P}_n} cost(p) = [z^n]p(z)$$

is defined to be the cumulated cost for the structures of size n*.*

Theorem 1.3 *(BGFs and average costs) Given a* BGF $P(z,u)$ *for a class of combinatorial structures, the average cost of all the structures of a given size is given by the cumulated cost divided by the number of structures, or*

$$\frac{[z^n]\frac{\partial P(z,u)}{\partial u}\Big|_{u=1}}{[z^n]P(1,z)} = \frac{p_n'(1)}{p_n(1)} = \frac{[z^n]p(z)}{p_n(1)},$$

where $p_n(u) = [z^n]P(z,u)$*. Also, the variance is*

$$\frac{p_n''(1)}{p_n(1)} + \frac{p_n'(1)}{p_n(1)} - \left(\frac{p_n'(1)}{p_n(1)}\right)^2.$$

Analogous definitions and results hold for EGFs, which we shall use when dealing with labelled structures.
For permutations (which are labelled objects), we can analyze their properties using the following systematic method:

- Define an exponential CGF of the form $B(z) = \sum_{p\in\mathcal{P}} cost(p)z^{|p|}/|p|!$ or the bivariate EGF $B(z,u) = \sum_{p\in\mathcal{P}} u^{cost(p)}z^{|p|}/|p|!$.
- Derive a functional equation for $B(z)$ (or $B(z,u)$) using the symbolic method.
- Solve the equation and use analytic techniques to find $[z^n]B(z)$, $[z^nu^k]B(z,u)$, etc.

Now we are giving an example from the hiring problem (following [5]), that follows the mentioned method.

Size of hiring set. Let σ denote the permutation of scores of the incoming candidates. Then, $h(\sigma)$ is the size of the hiring set $\mathcal{H}(\sigma)$, or the number of hired candidates in σ. Then we have $h(\sigma) = 0$ if σ is the empty permutation and $h(\sigma\circ j) = h(\sigma) + X_j(\sigma)$, where

$$X_j(\sigma) = \begin{cases} 1, & \text{if the last candidate of } \sigma\circ j \text{ is hired,} \\ 0, & \text{otherwise.} \end{cases}$$

and we can obtain the following result, which applies to any rank-based strategy.

Theorem 1.4 (Archibald and Martínez, 2009) *Let* $H(z,u)$ *be the generating function*

$$H(z,u) = \sum_{\sigma\in\mathcal{P}} \frac{z^{|\sigma|}}{|\sigma|!}u^{h(\sigma)},$$

where $h(\sigma)$ *is the size of the hiring set in* σ *and* $\mathcal{P}$ *is the class of all permutations.*
Then

$$(1-z)\frac{\partial}{\partial z}H(z,u) - H(z,u) = (u-1)\sum_{\sigma\in\mathcal{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}u^{h(\sigma)}.$$

Proof: Let $\mathcal{P}_n$ denote the set of permutations of size n. We can write thus

$$\begin{aligned} H(z,u) &= \sum_{\sigma\in\mathcal{P}} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} \\ &= 1 + \sum_{n>0} \sum_{\sigma\in\mathcal{P}_n} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} \\ &= 1 + \sum_{n>0} \sum_{1\le j\le n} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma\circ j|}}{|\sigma\circ j|!} u^{h(\sigma\circ j)}, \end{aligned}$$

where we have used the decomposition of any permutation σ' of size $n > 0$ as the product of a permutation σ of size $n-1$ times a value j between 1 and n (see Section 2.4). Hence

$$\begin{aligned} H(z,u) &= 1 + \sum_{n>0} \sum_{1\le j\le n} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|+1}}{(|\sigma|+1)!} u^{h(\sigma)+X_j(\sigma)} \\ &= 1 + \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|+1}}{(|\sigma|+1)!} u^{h(\sigma)} \sum_{1\le j\le n} u^{X_j(\sigma)}. \end{aligned}$$

Since $X_j(\sigma)$ is either 0 or 1 for all j and all σ, we have

$$\sum_{1\le j\le n} u^{X_j(\sigma)} = \big(|\sigma| + 1 - X(\sigma)\big) + uX(\sigma),$$

where $X(\sigma) = \sum_{1\le j\le|\sigma|+1} X_j(\sigma)$. Note that $X(\sigma)$ is the number of relative ranks such that a candidate with such a rank would be hired right after processing σ.

Hence,

$$H(z,u) = 1 + \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|+1}}{(|\sigma|+1)!} u^{h(\sigma)} \Big(\big(|\sigma| + 1 - X(\sigma)\big) + uX(\sigma) \Big).$$

Taking derivatives w.r.t. z, we obtain

$$\begin{aligned} \frac{\partial}{\partial z} H(z,u) &= \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} \Big(\big(|\sigma| + 1 - X(\sigma)\big) + uX(\sigma) \Big) \\ &= \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z\frac{d}{dz} z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} + \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} \\ &\quad + (u-1) \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} X(\sigma) \\ &= z\frac{\partial}{\partial z} H(z,u) + H(z,u) + (u-1) \sum_{n>0} \sum_{\sigma\in\mathcal{P}_{n-1}} \frac{z^{|\sigma|}}{|\sigma|!} u^{h(\sigma)} X(\sigma). \end{aligned}$$

After reorganizing the terms in the equation above and simplifying, we obtain the statement of the theorem. ■

Once we have the solution (i.e., a closed form) for $H(z,u)$, then $\mathbb{P}\{h_n = k\} = [u^k z^n] H(z,u)$. We can

also obtain the generating functions of the moments of h_n; we have to take successive derivatives of $H(z,u)$ w.r.t. u and set $u=1$.

$$h_r(z) = \frac{\partial^r}{\partial u^r} H(z,u)\Big|_{u=1} = \sum_{\sigma\in\mathcal{P}} \mathbb{E}\{h_n^r\} z^n \,,$$

The first moment gives us the expected value of the number of hired candidates as follows

$$h(z) = \frac{\partial}{\partial u} H(z,u)\Big|_{u=1} = \sum_{\sigma\in\mathcal{P}} h(\sigma)\frac{z^{|\sigma|}}{|\sigma|!} \,,$$

that is, $\mathbb{E}\{h_n\} = [z^n]h(z)$.

1.4 Singularity analysis

The other fundamental component in Analytic Combinatorics is the use of complex analysis techniques to extract information about the coefficients of generating functions. One of the most important techniques is *singularity analysis* [37].

Definition 1.10 *Given a function $f : \mathbb{C} \to \mathbb{C}$ defined in the region interior to a simple closed curve γ, a point z_0 on the boundary (γ) of the region is a singular point (a singularity) if f is not analytically continuable at z_0.*

We can summarize the singularity analysis process for the simple case where there is only a single dominant singularity: Let $f(z)$ be a function analytic at 0 whose coefficients are to be asymptotically analyzed.

i) *Locate singularities*. Determine the dominant singularities of $f(z)$. Check that $f(z)$ has a single singularity ζ on its circle of radius of convergence.

ii) *Check continuation*. Establish that $f(z)$ is analytic in some domain, larger than the disk of convergence.

iii) *Singular expansion*. Analyze the function $f(z)$ as $z \to \zeta$ in its domain of analyticity and determine in that domain an expansion of the form

$$f(z) \underset{z\to\zeta}{=} w(z/\zeta) + o\big(w(z/\zeta)\big).$$

For the method to succeed, the functions w and τ should belong to the standard scale of functions $\mathcal{S} = \{(1-z)^{-\alpha}\lambda^{\beta}\}$, with $\lambda = z^{-1}\log(1-z)^{-1}$, $\alpha \notin \mathbb{Z}_{\leq 0}$.

iv) *Transfer*. Translate the main term $w(z)$ and the error term using Theorems 1.8 and 1.7 below. Conclude that

$$[z^n]f(z) \underset{n\to+\infty}{=} \zeta^{-n} w_n + \mathcal{O}\big(\zeta^{-n}\tau_n^\star\big),$$

where $w_n = [z^n]w(z)$ and $\tau^\star = [z^n]\tau(z)$.

In the case of multiple singularities, the separate contribution from each of the singularities, as given by the basic singularity analysis process, are to be added up. For this case, there are some theorems like those that we have illustrated here for the case where there is only a single dominant singularity.

The method of singularity analysis applies to functions the behaviour of which around its singularities involves fractional powers and logarithms—one times refers to such singularities as "algebraic-logarithmic". The technology of singularity analysis is based on Cauchy's Coefficients formula, used in conjunction with special contours of integration known as Hankel contours.

Now, we are looking for the coefficients asymptotics. We can express the n-th coefficients of $f(z)$ as $A^n\theta(n)$, where A^n is called the exponential growth rate and $\theta(n)$ is called the subexponential one. It follows two principles of coefficient asymptotics.

First Principle of Coefficient Asymptotic. The location of a function's singularities dictates the exponential growth (A^n) of its coefficients.
Second Principle of Coefficient Asymptotic. The nature of a function's singularities determines the associate exponential factor (θ^n).

Hence, the most appropriate tool to investigate the asymptotic behaviour of the function near its dominant singularity is Cauchy's Integral Formula or Cauchy's Coefficient Formula as stated in the following theorem,

Theorem 1.5 *(Cauchy's Coefficient Formula). Let $f(z)$ be analytic in a region Ω containing 0 and let λ be a simple loop around 0 in Ω that is positively oriented. Then, the coefficients $[z^n]f(z)$ admits the integral representation*

$$f_n \equiv [z^n]f(z) = \frac{1}{2i\pi}\int_\lambda f(z)\frac{dz}{z^{n+1}}.$$

It follows also Cauchy's Residue Theorem.

Theorem 1.6 *(Cauchy's Residue Theorem). Let z_0 be an isolated singularity of $f(z)$ and let C be a circle centered at z_0 such that $f(z)$ is analytic in C and its interior, except possibly at z_0. Then,*

$$\int_C f(z)dz = 2i\pi \mathrm{Res}[f(z); z = z_0],$$

where $\mathrm{Res}[f(z); z = z_0]$ *denotes the residue of* $f(z)$ *at* z_0.

The coefficient formula allows us to deduce information about the coefficients from the function itself, using suitable chosen contours of integration. It becomes possible to estimate the coefficients $[z^n]f(z)$ in the expansion of $f(z)$ near 0 by using information on $f(z)$ away from 0.

Now, suppose we have the generating function $f(z)$ of a certain sequence. Since the power series of the function is analytic in the largest disk centered at the origin containing no singularities, one can look for singularities of the function that are nearest to the origin. The distance from the origin to the the nearest singularity is called *radius of convergence* of the power series and such singularity is called dominant.

Singularity analysis theory considers functions whose expansion at a singularity ζ behaves like

$$\left(1-\frac{z}{\zeta}\right)^{-\alpha}\left(\log\frac{1}{1-\frac{z}{\zeta}}\right)^{\beta}.$$

Under suitable conditions, we can show that such a singularity contributes a term of the form

$$\zeta^{-n}n^{\alpha-1}(\log n)^{\beta}, \quad \text{where } \alpha \text{ and } \beta \text{ can be arbitrary complex numbers.}$$

Singularity analysis theory considers also some *Transfer Theorems* which help to translate the asymptotic behaviour of a function near a singularity into an asymptotic approximation of its coefficients. The notation of such "Transfer" process is given in the following definition and theorem.

Definition 1.11 *Given two numbers ϕ, R with $R > 1$ and $0 < \phi < \frac{\pi}{2}$, then the open domain $\Delta(\phi, R)$ is defined as*

$$\Delta(\phi, R) = \{z \quad | \quad |z| < R,\, z \neq 1,\, |arg(z-1)| > \phi\}.$$

A domain is a Δ–domain at 1 if it is a $\Delta(\phi, R)$ for some R and ϕ. For a complex number $\zeta \neq 0$, a Δ–domain at ζ is the image by the mapping $z \mapsto \zeta z$ of a Δ–domain at 1. A function is Δ–analytic if it is analytic in some Δ–domain.

Theorem 1.7 *(Transfer, Big-Oh and little-oh). Let α, β be arbitrary real numbers, $\alpha, \beta \in \mathbb{R}$ and let $f(z)$ be a function that is analytic in the disk $\{z : |z| < 1\}$,*

i) Assume that $f(z)$ satisfies in the intersection of a neighborhood of $z = 1$ with Δ–domain the condition

$$f(z) = \mathcal{O}\left((1-z)^{-\alpha}\Big(\log\frac{1}{1-z}\Big)^{\beta}\right).$$

Then one has $[z^n]f(z) = \mathcal{O}\big(n^{\alpha-1}(\log n)^{\beta}\big)$.

ii) Assume that $f(z)$ satisfies in the intersection of a neighborhood of $z = 1$ with Δ–domain the condition

$$f(z) = o\left((1-z)^{-\alpha}\Big(\log\frac{1}{1-z}\Big)^{\beta}\right).$$

Then one has $[z^n]f(z) = o\big(n^{\alpha-1}(\log n)^{\beta}\big)$.

iii) Assume that $f(z)$ satisfies in the intersection of a neighborhood of $z = 1$ with Δ–domain the condition

$$f(z) \sim (1-z)^{-\alpha}\Big(\log\frac{1}{1-z}\Big)^{\beta}.$$

Then one has $[z^n]f(z) \sim n^{\alpha-1}(\log n)^{\beta}$.

Together with Theorem 1.7, the following theorem gives us a very powerful tool to obtain very precise asymptotic estimates of the coefficients of the generating functions we are interested in.

Theorem 1.8 *(Standard function scale, logarithms). Let α be an arbitrary complex number in $\mathbb{C} \setminus \mathbb{Z}_{\leq 0}$. The coefficient of z^n in the function*

$$f(z) = (1-z)^{-\alpha}\left(\frac{1}{z}\log\frac{1}{1-z}\right)^{\beta}$$

admits for large n a full asymptotic expansion in descending power of $\log n$,

$$f_n = [z^n]f(z) \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}(\log n)^{\beta}\left[1 + \frac{C_1}{\log n} + \frac{C_2}{\log^2 n} + \dots\right],$$

where $C_k = \binom{\beta}{k}\frac{d^k}{ds^k}\frac{1}{\Gamma(s)}\Big|_{s=\alpha}$.

As an example, if we go back to the binary trees mentioned in Section 1.2, then the OGF for the binary trees is

$$C(z) = \frac{1-\sqrt{1-4z}}{2z}.$$

Hence, $C(z)$ has a singularity at $z = \frac{1}{4}$, then the exponential growth for C_n is $A^n = 4^n$. That means that, for some $n_0 \in \mathbb{N}$, for all $n \geq n_0$ and for all $\epsilon > 0$,

$$(4-\epsilon)^n \leq C_n \leq (4+\epsilon)^n$$

After that we can see that

$$C(z) \sim -2\sqrt{1-4z} \quad \text{as } z \to \frac{1}{4}.$$

Thus the nature of this singularity is square-root type which gives us the subexponential or the polynomial factor, $\theta(n) = n^{-3/2}$. In particular

$$\begin{aligned} C_n = [z^n]C(z) &\sim 4^n \cdot (-2)[z^n]\sqrt{1-z} \\ &\sim 4^n \cdot (-2)\frac{n^{-\frac{1}{2}-1}}{\Gamma\left(-\frac{1}{2}\right)} \\ &\sim \frac{4^n}{\sqrt{\pi n^3}}. \end{aligned}$$

Chapter 2

A review of the hiring problem and related problems

In this chapter we review the work done in the hiring problem in the past twelve years. We briefly review the antecedents of the hiring problem, then we move on to discussion of the few papers on the hiring problem. In particular, we devote three sections, to discuss the articles of Preater [82], Krieger et al. [59, 60, 61], Broder at al. [15], and of Archibald and Martínez [5], respectively. We show each point of view and review their results on the analysis of various hiring strategies. We give also a brief discussion of the Chinese restaurant process [77] pinpointing its relation to the hiring problem. Along this chapter, we try to use the same notation and names as they appear in the original papers.

2.1 History of the hiring problem

The sequential multiple selection problem has been introduced several times, in particular in the probabilistic and Computer Science related literature. This section gives a short overview about this class of problems and the former studies of the hiring problem.

Literature related to *on-line decision-making under uncertainty* starts in the early sixties when Gardner [41] introduced the secretary problem, which has been solved first by Gilbert and Mosteller [43]. In the sequel a lot of papers addressed and studied extensions and variations of the secretary problem. One such natural and important extension is to choose many candidates and not only one. Such extensions have been studied extensively under various formulations with different goals and thus bear different titles in the literature, e.g., multiple-choice secretary problems [55, 72, 81], multiple optimal stopping rules [58, 75, 80], d-choice secretary problem [45], knapsack secretary problem [7], a generalization of the secretary problem [2, 10, 78] and others. Beside these extensions of the secretary problem, the *hiring problem* has received recently special interest as a close relative of the secretary problem but with major changes and different goals.

To the best of our knowledge, the name of the "hiring problem" has appeared for the first time in Chen et al. [19] in 1984. In fact, the authors introduced the following novel extension of the secretary problem, namely, a sequential multiple selection problem with constraints: suppose that one wants to hire N secretaries, where the secretary's salary demands are independent and identically distributed (i.i.d.) random variables (r.v.'s) with a known distribution, under the condition that their total salary must not exceed some value C. The question addressed and solved in that work

is how to set the thresholds (which "decide", whether a new secretary with salary demand X will be hired, if already $N - m$ secretaries are hired requiring together a salary S_{N-m}) in the sequential selection procedure to minimize the expected number of secretaries to be interviewed. However, this problem bears little resemblance with the "hiring problem" we study here.

In 2000, Preater [82] addressed the sequential selection problem and introduced for the first time a "hiring strategy" which is "better-than-average rule". In fact, the problem studied by Preater is exactly the hiring problem, although, he did not use the name. He assumed that the absolute scores of the candidates are forming a sequence of exponentially i.i.d. r.v.'s less than 1. The proposed rule works as follows: $i)$ the first candidate will be hired anyway, $ii)$ then a new candidate will be hired only if the average score of hired candidates will be increased (i.e., if the score of the new candidate is higher than the average score of the already hired staff).

Seven years later, Krieger, Pollak and Samuel-Cahn [59] introduced the *random permutation model* of the sequential multiple selection problem, under the name *select sets*. They gave a general class of selection rules called "p-percentile rules" that consider only relative ranks for the input sequence of items in analogy to the secretary problem. In this general strategy, fix p, $0 < p \leq 1$; then a new item will be selected, if it is amongst the best $100 \cdot p$ percent of those items that have been already retained.

After that, in a couple of publications [60, 61] Krieger et al. studied other configurations of the problem where they considered the absolute scores model under different distributions (i.e., Exponential, Pareto, Beta, Normal, Lognormal and Gamma) and a generalization of the selection rule given in [82]; namely "β-better-than-average rule" with $\beta > 0$.

Broder, Kirsch, Kumar, Mitzenmacher, Upfal and Vassilvitskii [15] introduced explicitly the name "hiring problem" in 2008; independently and unaware of the previous work of Preater [82] and Krieger et al. [59]. Broder et al. dealt with the absolute scores of candidates as uniform i.i.d. r.v.'s in $(0, 1)$. They introduced a natural and reasonable class of hiring strategies; namely "Lake Wobegon strategies" which includes "hiring above the mean" and "hiring above the median" strategies. Hiring above the mean processes the sequence of candidates exactly like the better-than-average rule in [82]. For hiring above the median, after hiring the first candidate, if the next candidate ranks better than the *median* score of all those hired before, then gets hired, and others are discarded. They discussed also the strategy called "max strategy" which hires only the records (i.e., left-to-right maxima) of the sequence of candidates.

Archibald and Martínez [5] introduced, in 2009, the discrete (combinatorial) model of the hiring problem, inspired by the work of Broder et al. [15]. Again they did so independently of the work of Preater and Krieger et al. They introduced in [5] two general hiring strategies, "hiring above the m-th best candidate" and "hiring above the α-quantile of hired candidates". Notice that the max strategy given in [15] is a special case of hiring above the m-th best when $m = 1$; that is, "hiring above the best". The strategy hiring above the α-quantile, with $0 < \alpha < 1$, is also a generalization for hiring above the median ($\alpha = \frac{1}{2}$) introduced in [15].

The various studies of the hiring problem might be classified according to two main features. First, *the model of the sequence of candidates*:

(i) **Random permutation model**: Krieger et al. [59] considered the relative ranks of candidates where the best one is given rank 1 while the worst is given rank n. Archibald and Martínez

[5] model the sequence of candidates as a random permutation also, but they considered the opposite ranking scheme where the best candidate has rank n while the worst one has rank 1. At any given moment, after i interviews, we can rank the i candidates seen so far from worst to best without ties.

(ii) **Absolute quality scores model**: Here, the input is a sequence of real numbers, say in $(0, 1)$ representing the actual scores of candidates. Preater [82] studied the case where the scores are given by an Exponential distribution. Broder et al. [15] considered a Uniform distribution of the scores in $(0, 1)$. Krieger et al. [60, 61] studied different distributions of the scores of candidates, namely, Exponential, Pareto, Beta, Normal, Lognormal and Gamma distributions.

The second feature is *the type of hiring strategies*: we have two main types of strategies.

(i) **Rank-based strategies**: these strategies take decisions based only on the ranks of candidates whether the actual quality scores are available or not. This category includes the p-percentile rules introduced by Krieger et al. [59], hiring above the median introduced by Broder et al. [15], and hiring above the α-quantile and hiring above the m-th best studied by Archibald and Martínez [5].

(ii) **Score-based strategies**: those take into account the absolute scores of candidates. They include the better-than-average rule introduced by Preater [82], the β-better-than-average rule (under different models of the sequence of candidates) studied by Krieger et al. [60, 61]. Also hiring above the mean introduced by Broder et al. in [15].

In the different studies of the problem, the main quantity of interest is the *number of selected/hired items* for a sequence of size n, or a closely related quantity, the *number of observations to select* k *items*. We call the latter one the *waiting time*. Such quantities help to study the *hiring rate* of the applied strategy, while other quantities are used to indicate the *quality of the selected/hired items* like the *average rank (score) of selected items* and others. In the following sections we give precise explanations for all the hiring strategies mentioned above and show the results obtained for various quantities of interest in each model of the problem.

2.2 Select sets

Krieger et al. have introduced a pioneering work in the sequential multiple selection problem, under the name *select sets*. That was an important departure from all previous attempts inherited from the secretary problem. No doubt that they got inspired by the secretary problem but they introduced a different setup for one variant of the secretary problem; that is, selecting many candidates instead of only one. As we have seen in the previous historical review, the main goal of secretary problems is to maximize the probability of selecting the best candidate(s) from the input sequence. But Krieger et al. care about different issues like designing "reasonable" selection rules and characterizing their behaviour according to the speed of selection and the quality of the retained group.

Krieger et al. introduced an equivalent formulation of the hiring problem covering more aspects of the problem; the most important is the formal definition of "reasonable" selection rules.

They have studied both models of the problem; the random permutation model that is investigated using the "p-percentile rules" [59], and the absolute quality scores model under different distributions of the scores, and using the class of "β-better-than-average rules" [60, 61].

We discuss here the main aspects of the framework of Krieger et al. in [59] to design and analyze selection rules based on ranks, and review the main results obtained. We also highlight their rich study of the "β-better-than-average rule" under different distributions of the quality scores of items. Their results characterize the asymptotic behaviour of various quantities of interest.

2.2.1 Percentile rules

In this class of selection rules, the decision of selecting or discarding an item depends only on the number of retained items so far and the rank of the current item among all interviewed items. Krieger et al. considered the following ranking scheme for the input sequence: the better is equivalent to smaller rank so the best item is given rank 1 while the worst one is given n among n items, in analogy to the secretary problem.
Let us define some notations used here: R_i^n denotes the rank of the i-th item in a sequence of size n and L_n is a r.v. that denotes the *number of retained items* or the *size of retained group* after observing n items.
Then they discuss that reasonable selection rules should fulfill the conditions of being *locally subdiagonal* according to the following definition:

Definition 2.1 *A locally subdiagonal rank selection scheme (LsD) is a rule determined by an integer-valued function $r(\cdot)$ with the following properties:*

i) r *is nondecreasing.*

ii) $r(0) = 1$ *and* $L_0 = 0$.

iii) $r(k+1) \leq r(k) + 1$, *where* k *is the number of retained items so far.*

iv) The first item is retained, then the n*-th item is retained if and only if* $R_n^n \leq r(L_{n-1})$.

This class contains many selection rules that make sense; two main families of them are: "p-percentile rules" and "m-record rules". The latter one, with $r(k) = \min(k+1, m)$, has been studied extensively—as Krieger et al. put it—. In fact, there are numerous publications related to m-records (i.e., [3, 13, 26, 63, 74, 85]) concerning some statistics like their *values* and *times* (*positions*) under different distributions. But the first time that m-records were considered in the context of the hiring problem was by Archibald and Martínez when they introduced the strategy "hiring above the m-th best" (see Section 2.4). The p-percentile rules are defined as follows:

Definition 2.2 *A "*p*-percentile rule",* $0 < p \leq 1$, *is a LsD rule with* $r(k) = \lceil pk \rceil$ *for* $k \geq 1$, *then the* n*-th item is retained if* $R_n^n \leq \lceil pL_{n-1} \rceil$.

The most important instance of this family is the "$\frac{1}{2}$-percentile rule" or the median rule, where $r(1) = r(2) = 1$ and generally $r(2j-1) = r(2j) = j$ for odd and even sizes of retained group. In this class of rules, there is a *cutoff (or threshold) rank*, that is $j_n \equiv \lceil pL_n \rceil$, where the $(n+1)$-th item is retained if its rank is less than or equal to j_n, and others are discarded.

Krieger at al. went through a purely probabilistic approach, using the theory of martingales, and were able to study two main quantities: the *number of retained items* L_n and the *average rank of retained group* A_n. The results of the order of L_n, as well as the limiting behaviour is given in the following theorem:

Theorem 2.1 (Krieger et al., 2007) *For the "p-percentile rules", $0 < p \leq 1$, let L_n denote the number of retained items after n observations, then as $n \to \infty$*

$$\frac{\mathbb{E}\{L_n\}}{n^p} \to c_p,$$
$$\lim_{n\to\infty} \mathbb{E}\left\{\frac{L_n^2}{n^{2p}}\right\} \text{ and } \lim_{n\to\infty} \mathbb{V}\left\{\frac{L_n}{n^p}\right\} \text{ exist and are finite},$$
$$\frac{L_n}{(n+1)^p} \xrightarrow{\text{(a.s.)}} \text{nondegenrate finite r.v. } \Lambda \text{ such that:}$$
$$\lim_{n\to\infty} \mathbb{E}\left\{\frac{L_n}{(n+1)^p}\right\} = \mathbb{E}\{\Lambda\} = c_p, \text{ with } 0 < c_p < \infty.$$

That means that L_n for p-percentile rules is of order n^p which is a very important result. However the leading factor and the explicit limiting distribution are missing for even some particular rules like $\frac{1}{2}$-percentile rule. They have said *"It seems impossible to determine c_p analytically, except for $p = 1$"*, which is debatable as we will see in Chapter 4 when we discuss the strategy "hiring above the median" showing the results for $c_{1/2}$. There is also one recent result by Gaither and Ward [40] for general p:

Theorem 2.2 (Gaither and Ward, 2012) *For the "p-percentile rules", $0 < p \leq 1$, let L_n denote the number of retained items after n observations. Then, as $n \to \infty$, we have $\frac{\mathbb{E}\{L_n\}}{n^p} \to c_p$, where*

$$c_p = \frac{1 + \sum_{k\geq 1} \frac{\lceil pk\rceil - pk}{\lceil pk\rceil} \prod_{j=1}^{k} \frac{1}{1+\frac{p}{\lceil pj\rceil}}}{(p+1)\Gamma(p+1)}.$$

The first result regarding the quality of the set of selections follows for LsD rules by definition (but also proved in [59]):

Theorem 2.3 (Krieger et al., 2007) *Consider an LsD rule defined by $r(\cdot)$. Let L_n denote the number of selected items after n observations, then the set of selections contains the best $r(L_n)$ items.*

For example, almost half of the set of selected items are the very best seen using the $\frac{1}{2}$-percentile rule. Krieger et al. introduced another quantitative measure of the quality of retained items, that is $A_n = \frac{Q_n}{L_n}$, where Q_n is the sum of the ranks of retained items. The following theorem introduces the order of growth of A_n for different families of rules:

Theorem 2.4 (Krieger et al., 2007) *For the "p-percentile rules", $0 < p \leq 1$, let A_n denote the average rank of retained group after n observations, then*

$$\frac{\mathbb{E}\{A_n\}}{a_n(p)} \to b_p, \text{ as } n \to \infty,$$

where

$$a_n(p) = \begin{cases} n^{1-p}, & \text{if } p < \frac{1}{2}, \\ \sqrt{n}\cdot \log n, & \text{if } p = \frac{1}{2}, \\ n^p, & \text{if } p > \frac{1}{2}. \end{cases}$$

The limiting behaviour of A_n suitably normalized, depends on p and is characterized in the following theorem:

Theorem 2.5 (Krieger et al., 2007) *For the "p-percentile rules", $0 < p \leq 1$, let A_n denote the average rank of retained group after n observations, then as $n \to \infty$*

- *If $0 < p < \frac{1}{2}$, then*
$$\frac{A_n}{n^{1-p}} \xrightarrow{\text{(a.s.)}} \text{nondegenerate r.v.},$$
- *If $\frac{1}{2} < p \leq 1$, then*
$$\frac{A_n}{L_n} \xrightarrow{\text{(a.s.)}} q_p, \text{ and } \mathbb{E}\left\{\frac{A_n}{L_n}\right\} \to q_p, \text{ with } q_p = \frac{p^2}{2(2p-1)},$$
- *If $p = \frac{1}{2}$, then*
$$\frac{A_n}{L_n \cdot \log n} \xrightarrow{\text{(a.s.)}} \frac{1}{8}, \text{ and } \mathbb{E}\left\{\frac{A_n}{L_n \cdot \log n}\right\} \to \frac{1}{8}.$$

Moreover, Krieger at al. have pointed out to other useful quantities like the *number of observations made from the instance that the size of retained group became $i-1$ until its size became i, called Z_n*, and the *waiting time, called N_n* that is the *number of observations made until n items have been retained*. Z_n is simply the distance between the last two retained items after n observations. For N_n, they have mentioned that for $0 < p \leq 1$, and $n \geq 1$, then $\mathbb{E}\{N_n\} = \infty$ since $\mathbb{E}\{Z_2\} = \infty$, which is true for well understood reasons.

2.2.2 Better-than-average rules

The most important instance in this class of rules is the "better-than-average rule", which was introduced first by Preater [82] but generalized and studied widely by Krieger et al. [60, 61]. We start defining the general case.

Definition 2.3 *For $\beta > 0$, "β-better-than average rule" selects the first item in the sequence, then any further item is selected if and only if its score is better than β times the present average of the retained group.*

The selection criteria are based on the *absolute quality scores* (*scores*, for short), of the present item and the corresponding *average* score of the retained group.

In their paper [60], Krieger at al. have studied three distributions of the scores: Exponential, Pareto and Beta. Four quantities were considered,

- T_k: the *waiting time*, that is the *number of observations until the size of retained group is* k,
- M_n: the *number of selected items after* n *observations*,
- A_n: the *average score of selected items after* n *observations*, and
- Y_k: the *average score of the first* k *selected items*.

They made use of the Martingale Convergence Theorem and show many interesting results for different distributions of the scores. We review the main results related to the limiting behaviour of the studied quantities in the following theorems:

Theorem 2.6 (Krieger et al., 2008) *For the "better-than-average rule", assume that the observations are i.i.d. r.v.'s from an* Exponential *distribution with mean 1. Let* T_k, M_n, A_n *and* Y_k *be defined as above, and* G *denote a r.v. that has the* Gumbel *distribution, with cumulative function (c.f.)* $\exp(-e^{-x})$*, then as* $n, k \to \infty$

i) $\frac{T_k}{k^2} \xrightarrow{(a.s.)} e^G/2$,

ii) $\frac{M_n}{\sqrt{n}} \xrightarrow{(a.s.)} \sqrt{2}e^{-G/2}$,

iii) $A_n - (\log n)/2 \xrightarrow{(a.s.)} (G + \log 2)/2$,

iv) $Y_k - \log k \xrightarrow{(a.s.)} G$.

Preater was the first to give the result for Y_k in [82]. Theorem 2.6 extends Preater's results for the Exponential distribution of scores.

Theorem 2.7 (Krieger et al., 2008) *For the "β-better-than-average rule", assume that the observations are i.i.d. r.v.'s from* Pareto(α) *distribution with* $\alpha > 1$*. Let* T_k, M_n, A_n *and* Y_k *be defined as above, then for* $\beta > \frac{\alpha-1}{\alpha}$ *and as* $n, k \to \infty$

$$\frac{T_k}{k^{\frac{(\beta-1)\alpha^2+2\alpha-1}{\alpha-1}}}, \quad \frac{M_n}{n^{\frac{\alpha-1}{(\beta-1)\alpha^2+2\alpha-1}}}, \quad \frac{A_n}{n^{\frac{(\beta-1)\alpha+1}{(\beta-1)\alpha^2+2\alpha-1}}} \quad \text{and} \quad \frac{Y_k}{k^{\frac{(\beta-1)\alpha+1}{\alpha-1}}}$$

converge a.s. to a positive finite r.v.

Pareto distribution with parameter α has a c.f. $F_\alpha(x) = 1 - x^{-\alpha}$, $x \geq 1$.

Theorem 2.8 (Krieger et al., 2008) *For the "β-better-than-average rule", $\beta > 0$, assume that the observations are i.i.d. r.v.'s from* Pareto(α) *distribution. Let* Y_k *denote the average score of the first* k *selected items, then as* $k \to \infty$

i) If $0 < \alpha < 1$ *then*

$$\frac{\log Y_k}{k^{1-\alpha}} \quad \textit{converges a.s.}$$

ii) If $\alpha = 1$ *then*

$$\frac{\log Y_k}{(\log k)^2} \quad \textit{converges a.s.}$$

The corresponding results for the other quantities T_k, M_n and A_n were complicated.

Theorem 2.9 (Krieger et al., 2008) *For the "β-better-than-average rule", assume that the observations are i.i.d. r.v.'s from* Beta(α, 1) *distribution. Let* T_k, M_n, A_n *and* Y_k *be defined as above, then for* $\beta < \frac{\alpha+1}{\alpha}$ *and as* $n, k \to \infty$

$$\frac{T_k}{k^{\frac{(\alpha+1)^2-\alpha^2\beta}{\alpha+1}}}, \quad \frac{M_n}{n^{\frac{\alpha+1}{(\alpha+1)^2-\alpha^2\beta}}}, \quad A_n \cdot n^{\frac{\alpha+1-\alpha\beta}{(\alpha+1)^2-\alpha^2\beta}} \quad \text{and} \quad Y_k \cdot k^{1-\frac{\alpha\beta}{\alpha+1}}$$

converge a.s.

Beta distribution with parameters $(\alpha, 1)$ has a c.f. $F_\alpha(x) = x^\alpha$ for $0 \leq x \leq 1$ and $F_\alpha(x) = 1$ for $1 \leq x$, with $\alpha > 0$.

Krieger et al. continued studying the β-better-than-average rule in [61] but considering different class of distributions for the scores, namely the Gumbel domain of attraction of extreme value distribution with c.f. $\exp(-e^{-x})$. In particular, they presented many results for a specific subset of the Gumbel family, which is called "stretch exponential" distributions. The later class contains many interesting distributions, i.e. Normal, Lognormal, Gamma and Weibull.

Definition 2.4 *Consider the class of distributions $\mathcal{G}_\alpha$ defined by:*

$$\mathcal{G}_\alpha(x) = 1 - \exp\big(-H(x)\big), \quad H(x) = c \cdot x^\alpha + h(x),$$

where

- $h''(x)$ *exists, and* $c, \alpha > 0$ *are constants.*
- $\lim_{x\to\infty} \dfrac{h(x)}{x^\alpha} = 0.$
- $\lim_{x\to\infty} \dfrac{h'(x)}{x^{\alpha-1}} = 0.$

When $c = 1$, we have the "stretch exponential" distributions, that contains the Normal (that is $\mathcal{G}_2$ with $h(x) = \log x$) and Gamma distributions as particular cases.

They considered here two quantities: T_k that is the *waiting time for selecting* k *items* and Y_k is the *average score of the first* k *selected items*. We review first the main results for Y_k.

Theorem 2.10 (Krieger et al., 2010) *For the "better-than-average rule", assume that the underlying distribution of the scores of items belongs to the* stretch exponential *family. Let Y_k denote the average score of the first k selected items and G denote a r.v. that has a Gumbel distribution, then as $k \to \infty$*

$$Y_k - G^{-1}(\log k) \xrightarrow{(a.s.)} \textit{a finite r.v.},$$

under the conditions

- $\mathbb{E}\{Z^2(a)\} < a^\gamma$ *for some* $0 < \gamma < \infty$ *and all* $a > a_0$, *and*
- $f'(a) \leq 0$ *for all* $a \geq a_0$, *for some* $a_0 < \infty$.

Here $Z_k = Y_k - Y_{k-1}$, is the "overshoot" over Y_{k-1} after having the retained group size is equal to k. Then $Z(a)$ is distributed like $X - a | X > a$, and $f(a) = \mathbb{E}\{Z(a)\}$.

Theorem 2.11 (Krieger et al., 2010) *For the "β-better-than-average rule", $\beta > 1$, assume that the underlying distribution of the scores of items belongs to the* stretch exponential *family. Consider the normalized r.v. $\frac{Y_k}{k^{\beta-1}}$, where Y_k denotes the average score of the first k selected items, and $f(x) = \mathbb{E}\{X - x | X > x\}$, then*

i) If $f(x) < \frac{cx}{(\log x)^{1+\epsilon}}$, where $c, \epsilon > 0$, then

$$\frac{Y_k}{k^{\beta-1}} \xrightarrow{(a.s.)} \textit{a nondegenerate positive r.v.},$$

ii) If $\frac{Y_k}{k^{\beta-1}}$ *converges a.s.,* f *is monotone and* $\lim_{k\to\infty} \mathbb{E}\left\{\frac{Y_k}{k^{\beta-1}}\right\} < \infty$ *then for some constant* $x_0 > 0$

$$\int_{x_0}^{\infty} \frac{f(x)}{x^2} dx < \infty.$$

Moreover, in general for $\mathcal{G}_\alpha$ *distribution of scores (Definition 2.4),* $\alpha > 0$*, then*

$$\mathbb{E}\left\{\frac{Y_k}{k^{\beta-1}}\right\} \text{ and } \mathbb{V}\left\{\frac{Y_k}{k^{\beta-1}}\right\} \xrightarrow{(a.s.)} \text{nondegenerate positive r.v.'s.}$$

Now we move to the other quantity, T_k, and show the main results for its asymptotic behaviour. For the general rule, $\beta > 1$, no general results are available for T_k, but for a standardized version which is

$$T_k^* = \frac{T_k}{\sum_{j=1}^{k-1}(1 - F(\beta Y_j))^{-1}}. \tag{2.1}$$

Here they considered only $F \in \mathcal{G}_\alpha$, $\alpha > 0$ with $h(x) \equiv 0$ and $H(x) = x^\alpha$. Then it is nice to show that the asymptotic behaviour of T_k^* depends on β as shown in the following theorems:

Theorem 2.12 (Krieger et al., 2010) *For the "*β*-better-than-average rule", consider* F *as the distribution of scores where* $F(x) = 1 - e^{-x^\alpha}$ *with* $\alpha > 0$*. Let* T_k^* *be defined as in* (2.1) *and* Y_k *denote the average score of the first* k *selected items , then as* $k \to \infty$

i) If $1 < \beta < 1 + \frac{1}{2\alpha}$ *then* $T_k^* \xrightarrow{(a.s.)} 1$*,*

ii) If $1 + \frac{1}{2\alpha} < \beta < 1 + \frac{1}{\alpha}$ *then* $T_k^* \xrightarrow{(P)} 1$*,*

iii) If $1 + \frac{1}{\alpha} < \beta$ *then* $T_k^* \xrightarrow{(d)} \text{Exp}(1)$*, and* $\frac{T_k}{W_k} \xrightarrow{(d)} \text{Exp}(1)$ *where* $W_k = e^{\beta\alpha Y_{k-1}^\alpha}$*.*

Theorem 2.13 (Krieger et al., 2010) *For the "*β*-better-than-average rule", consider* F *as the distribution of scores where* $F(x) = 1 - e^{-x^\alpha}$ *with* $\alpha > 0$ *and* $\beta = 1 + \frac{1}{\alpha}$*. Let* $W = \lim_{k\to\infty} \frac{Y_k}{k^{1/\alpha}}$*, where* Y_k *denotes the average score of the first* k *selected items and* T_k^* *is defined as in* (2.1)*, then as* $k \to \infty$

$$T_k^* \xrightarrow{(d)} \sum_{j=1}^{\infty} R_j,$$

where, conditionally on $W = w$*, the* R_j *are independent, exponentially distributed with mean* μ_j*, where*

$$\mu_j = \frac{e^{(\beta w)^\alpha} - 1}{e^{j(\beta w)^\alpha}}.$$

2.3 Lake Wobegon strategies

Broder et al. [15] introduced explicitly for the first time the terminology of the "hiring problem" in 2008. They got motivated by the secretary problem considering the extension of hiring many employees instead of hiring only one. They considered the absolute scores of candidates as uniformly distributed on the interval $(0, 1)$; each candidate has a quality score Q_i. Thus these Q_i's are i.i.d. r.v.'s with common distribution $\text{Unif}(0, 1)$.

In this model they introduced some hiring strategies, the most interesting are what they called "Lake Wobegon strategies"[1]. Broder et al. have borrowed this colorful name from Peter Norvig [76] who claimed that Google actually uses such kind of strategies to hire new employees. Lake Wobegon strategies hire candidates that are *better than the "average"* candidate already hired, where the average may refer to either *mean* or *median*.

Thus, this class of strategies includes "hiring above the mean" and "hiring above the median". They also considered briefly the max strategy or hiring above the best and hiring above a fixed threshold. We review in this section the results obtained in [15] for two quantities: the *waiting time* in terms of the number of interviews until n candidates are getting hired, and the *gap between the average score of hired candidates and the maximum score (i.e.,* 1*)*.

Hiring above the mean

This strategy behaves exactly like the "better-than-average rule" introduced first by Preater [82]. It is also one member in the class of "β-better-than average rule" when $\beta = 1$ (Subsection 2.2.2). We define it anew before showing the results,

Definition 2.5 *The strategy "hiring above the mean" hires the first candidate in the sequence, then any further candidate is hired if and only if his score is larger than the current average score of all hired candidates so far.*

In this strategy, let A_i denote the average quality after i hirings, with $A_0 = q$ being the quality of the initial candidate, so that A_i refers to the *average quality of* $i + 1$ *hired candidates*. Then, at any step, this strategy will hire only scores that are above the *mean* quality of the hired staff. We summarize the results obtained for this strategy in the following theorem:

Theorem 2.14 (Broder et al., 2008) *For the strategy "hiring above the mean", assume that the scores of candidates are i.i.d. r.v.'s from* $\text{Unif}(0, 1)$ *distribution.*

- *Let* A_i *denote the average quality of* $i + 1$ *hired candidates, then with probability 1, infinitely many candidates will be hired and* $\lim_{i\to\infty} A_i = 1$.
- *The expected gap after* n *hirings, where gap is defined as* $G_n = 1 - A_n$*, is*

$$\mathbb{E}\{G_n\} = \Theta\Big(\frac{1}{\sqrt{n}}\Big).$$

- *Let* T_n *be the number of candidates that have been interviewed when the number of hired candidates reaches* n*, then*

$$\mathbb{E}\{T_n\} = \Theta\Big(n^{3/2}\Big).$$

[1] As pointed out in [15], 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 distribution of the gap (under suitable initial conditions) weakly converges to a* Lognormal *distribution, that means that the body of the distribution converges to a* lognormal *distribution, but there may be larger error at the tails.*

Thus Broder et al. have shown that the quality of the hired staff is improved all the time and also the hiring rate will be reasonable. The quantity G_i gives us an indicator of the quality of hired staff. So, for large values of n, the quality will be close to 1 (the maximum quality). Following the derivation in their paper, we find that the initial starting gap has a multiplicative effect on the expected gap and the expected number of hired candidates. And this is true also when starting with more than one employee.

Hiring above the median

This strategy has the same flavor as the "$\frac{1}{2}$-percentile rule" (Subsection 2.2.1) , but with a small change in the hiring criteria as we shall explain soon. The following definition clarifies well the behaviour of this strategy

Definition 2.6 *The strategy "hiring above the median" hires the first candidate in the sequence, then any further candidate is hired if and only if his score is larger than the current median score of the hired staff. The median of a set of* k *(distinct) elements* $x_1 < x_2 < \cdots < x_k$ *is the* ℓ*-th largest element, i.e.,* $x_{k+1-\ell}$*, with* $\ell = \lceil \frac{k+1}{2} \rceil$.

Thus when the number of hired candidates is *odd*, there is only one *unique median* and this is the threshold candidate for both hiring above the median and the $\frac{1}{2}$-percentile rule. But when the number of hired candidates is *even*, then we say that there are *two medians*, then hiring above the median considers the *lowest median* as the threshold, while the $\frac{1}{2}$-percentile rule takes the *highest median* to be its threshold.

The set of hired candidates here starts with one candidate with quality $q \in (0, 1)$ and whenever we have $2k + 1$ hired candidates, the next two hired candidates must have at least the median score M_k of the $2k + 1$ candidates. The results of this strategy are summarized in the following theorem:

Theorem 2.15 (Broder et al., 2008) *For the strategy "hiring above the median", assume that the scores of candidates are i.i.d. r.v.'s from* Unif$(0, 1)$ *distribution.*

- *Let* M_k *denote the median score of* $2k + 1$ *hired candidates, then with probability 1, infinitely many candidates will be hired and* $\lim_{k\to\infty} M_k = 1$.
- *The gap is defined as* $G'_k = 1 - M_k$*, hence it converges to 0 as* $k \to \infty$*. The gap expectation is*

$$\mathbb{E}\{G'_k\} = \Theta\Big(\frac{1}{k}\Big).$$

- *Let* T'_k *be the number of interviews until there are* $2k + 1$ *hired candidates. Then*

$$\mathbb{E}\{T'_k\} = \frac{k(k+1)}{g},$$

where $g = 1 - q$*, the initial starting gap.*

- *Let A'_n denote the mean quality score of the first n hired candidates. Then*

$$\mathbb{E}\{A'_n\} = 1 - \Theta\Big(\frac{\log n}{n}\Big).$$

- *The distribution of the gap also converges weakly to a* Lognormal *distribution as in hiring above the mean.*

Thus it is also true for this strategy that the quality of the hired staff is improved all the time —as G'_k and A'_n say—and also the hiring rate will be reasonable. Moreover, hiring above the median leads to *smaller* gap (*higher* quality) than hiring above the mean, but with *fewer* hirings (*slower* rate of growth) because the number of interviews between hirings is much larger.
Finally, Broder et al. claim that both strategies (hiring above the mean and hiring above the median) are within a constant factor of optimal. However, they do not explicitly state what do they mean by "optimal" in this context, and there seems not to be an obvious notion of optimality.

Broder et al. proposed some generalization of the strategy hiring above the median that is, for two integers a and b, "hire a candidates, then move up the threshold b candidates in the rank order". So hiring above the median strategy is "hire 2, move up 1". They have shown in the appendix of [15] the following results for two quantities: G^*_k which denotes the *quality at the threshold candidate after* ka *hirings*. And T^*_k which denotes the *total number of interviews before there are* $ak + 1$ *hirings*. They show that:

$$\mathbb{E}\{G^*_k\} = \Theta\big(k^{-\frac{b}{a-b}}\big), \quad \text{and} \quad \mathbb{E}\{T^*_k\} = \Theta\big(k^{\frac{a}{a-b}}\big).$$

2.4 The hiring problem and permutations

Archibald and Martínez [5] proposed the *random permutation model* for the hiring problem as an alternative to the model used by Broder et al. in [15]. Krieger et al. had worked in the random permutation model of the problem (review Subsection 2.2.1), but the independent study by Archibald and Martínez is purely discrete and opens the door to obtain various interesting results for the hiring problem as we will see in the rest of this section.

Archibald and Martínez considered a class of hiring strategies that work in the random permutation model, so such strategies are *rank-based*. Considering relative ranks is one similarity between this model and the secretary problem, which does not hold for the absolute scores model. "Hiring above the median" strategy (introduced originally by Broder et al.) is one member in this class, as well as its generalization, introduced here, "hiring above the α-quantile of the hired staff", with $0\alpha < 1$. They proposed also another rank-based strategy, namely "hiring above the m-th best candidate". As we shall see, the later strategy is closely linked to m-records [6] in permutations. Two quantities of interest were studied in [5]: h_n, that denotes the *number of hired candidates after receiving* n candidates or the *size of the hiring set*, and g_n which represents the *gap of last hired candidate after receiving* n *candidates*.

We review in this section the framework given by Archibald and Martínez showing how they were able to use basic techniques of Analytic Combinatorics [37] in order to analyze rank-based hiring strategies. We discuss their important definition of the notion of "pragmatic" hiring strategies, together with the results for h_n and g_n under the studied hiring strategies.

Random permutation model

Here, the interviewed candidate is given a *rank* which is relative to the ranks of the previous candidates. The ranking scheme considers the *larger* rank as *better* than *lower* one, that is the contrary to ranking scheme in secretary problems and the p-percentile rules. But of course both ways are totally equivalent and lead to the random permutation model for the sequence of candidates. Thus, for a sequence $S = s_1, \ldots, s_i, \ldots$ of candidates, s_i denotes the relative rank of the i-th candidate among all interviewed ones, where the best candidate seen so far among the n gets a rank n, while the worst one gets rank 1. Then the n ranks (or scores) of the n candidates form finally a permutation $\sigma^{(n)} = \{1, \ldots, n\}$. Then any permutation σ representing the scores is equally likely. In this context, the *hiring set* of a permutation σ is the set of indices (arrival times) of candidates that would be hired by applying a specific strategy to σ.

General framework

Before introducing the main tool used in this model, it is better to be familiar with the meaning of some terms. Given a permutation $\sigma^{(n-1)}$ of length $n-1$ and a value (relative rank) j, $1 \leq j \leq n$, then $\sigma^{(n-1)} \circ j$ denotes the resulting permutation of size n after relabelling $j, j+1, \ldots, n$ and appending j to the end. For example, if we have this sequence of relative ranks: $S = 1, 1, 3, 2, 2$ then the corresponding permutations are $\sigma^{(1)} = 1$, $\sigma^{(2)} = \sigma^{(1)} \circ 1 = 21$, $\sigma^{(3)} = \sigma^{(2)} \circ 3 = 213$, $\sigma^{(4)} = \sigma^{(3)} \circ 2 = 3142$ and $\sigma^{(5)} = \sigma^{(4)} \circ 2 = 41532$. The notation $\mathcal{H}(\sigma)$ will denote the set of indices of the hired candidates or the *hiring set* of the permutation σ. This hiring set has some parameters to be studied w.r.t. a given hiring strategy such as its size $h(\sigma)$, the *gap of last hired candidate* $g(\sigma)$, the *index of last hired candidate* $L(\sigma)$ and other useful parameters as we will see later. The letters $h_n, g_n, L_n, \ldots$ will denote the corresponding r.v.'s of these parameters. The gap of last hired candidate is defined as $g(\sigma) = 1 - \frac{R(\sigma)}{|\sigma|}$ where $R(\sigma)$ is the *score of last hired candidate*.

The main tool in this framework is the generating functions which are essential for analyzing combinatorial structures as permutations. For each hiring parameter, they define a bivariate exponential generating function (BEGF) of the form $B(z,u) = \sum_{p \in \mathcal{P}} u^{\text{cost}(p)} z^{|p|}/|p|!$, with $\mathcal{P}$ the family of permutations and $\text{cost}(\cdot)$ 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.

There is an important r.v. indicator called $X_j(\sigma)$ which is defined as

$$X_j(\sigma) = \begin{cases} 1, & \text{if a candidate with score } j \text{ is hired after } \sigma, \\ 0, & \text{otherwise.} \end{cases}$$

and the quantity

$$X(\sigma) = \sum_{1 \leq j \leq |\sigma|+1} X_j(\sigma),$$

tells us how many candidates from the $|\sigma|+1$ possible ones, would be hired after processing the permutation σ under the applied strategy. Each hiring strategy is characterized by its corresponding $X(\sigma)$. So, for each parameter, we will obtain a differential equation in the function representing

this parameter involving $X(\sigma)$. Then, for a particular hiring strategy, one has to use its characteristic $X(\sigma)$ and solve.
Archibald and Martínez introduced the concept of "pragmatic" strategies which is equivalent to the LsD rules by Krieger et al. (Definition 2.1).

Definition 2.7 *A rank-based strategy is* pragmatic *if the following two conditions are met:*

1. *For all* σ *and all* j, $X_j(\sigma) = 1$ *implies* $X_{j'}(\sigma) = 1$ *for all* $j' \geq j$.
2. *For all* σ *and all* j, $X(\sigma \circ j) \leq X(\sigma) + X_j(\sigma)$

The first condition states that whenever a pragmatic strategy hires a candidate with score j, it would hire a candidate with a higher score. The second condition bounds the hiring rate and guarantees that the potential of hiring $X(\cdot)$ does not change if no new candidate is hired. Archibald and Martínez proved the following result, which applies for all pragmatic strategies, and is completely equivalent to Theorem 2.3:

Theorem 2.16 (Archibald and Martínez, 2009) *For any pragmatic hiring strategy and any permutation* σ, $\mathcal{H}(\sigma)$ *contains at least the* $X(\sigma)$ *best candidates of* σ, *that is, the candidates with scores* $|\sigma|, |\sigma|-1, \ldots, |\sigma| + 1 - X(\sigma)$.

We have also the following general result:

Theorem 2.17 (Archibald and Martínez, 2009) *For any pragmatic hiring strategy, let* g_n *denote the gap of last hired candidate, then*

$$\mathbb{E}\{g_n\} = \frac{1}{2n}(\mathbb{E}\{X_n\} - 1),$$

where $\mathbb{E}\{X_n\} = [z^n] \sum_{\sigma \in \mathcal{P}} X(\sigma) \frac{z^{|\sigma|}}{|\sigma|!}$.

Now we can show the results of Archibald and Martínez for hiring above the m-th best, and hiring above the α-quantile strategies.

Hiring above the m-th best candidate

The following definition explains the behaviour of this strategy,

Definition 2.8 *The strategy "hiring above the* m*-th best" hires the first* m *candidates in the sequence regardless of their relative ranks, then any further candidate is hired if and only if his relative rank is larger than the current* m*-th largest one between all previously hired candidates.*

Thus $X(\sigma)$ for this strategy is defined as follows

$$X(\sigma) = \begin{cases} |\sigma| + 1, & \text{if } |\sigma| < m, \\ m, & \text{if } |\sigma| \geq m. \end{cases}$$

Notice that the value m is fixed along the hiring process and it might be some fixed integer or being determined according to some function of the number of candidates n, i.e. $m = \lceil \sqrt{n} \rceil$, $m = \lceil \log n \rceil$, ... The subscript notation $\{n, m\}$ is used with the studied parameters here to refer to the strategy.

Theorem 2.18 (Archibald and Martínez, 2009) *For the strategy "hiring above the m-th best", let $h_{n,m}$ denote the size of the hiring set after n interviews. Then for $1 \leq m \leq n$, the expectation is*

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

where the asymptotic expansion holds uniformly for $1 \leq m \leq n$. The variance for fixed m ($m = \Theta(1)$) is asymptotically $\mathbb{V}\{h_{n,m}\} = m \ln n + \mathcal{O}(1)$. Moreover, the following central limit holds for fixed m and $n \to \infty$,

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

With $m = 1$, hiring above the best, $\mathbb{E}\{h_n\} = \ln n + \mathcal{O}(1)$ and $\mathbb{P}\{h_n = k\}$ is given by the unsigned Stirling numbers of the first kind $\begin{bmatrix} n \\ k \end{bmatrix}$ which coincides with the number of permutations of size n that have exactly k left-to-right maxima [57]. That is because hiring above the best hires only the left-to-right maxima ranks. It is worth to mention here that they were able to proof the central limit theorem for $h_{n,m}$ directly from the closed form of the generating function using Hwang's quasi-powers theorem [37]. Prodinger [83] studied the number of m-records for n independent r.v.'s drawn from a Geometric distributed with $\mathbb{P}\{X = x\} = pq^{x-1}$ and $p + q = 1$. He obtained same results for $\mathbb{E}\{h_{n,m}\}$ and $\mathbb{V}\{h_{n,m}\}$, as given in Theorem 2.18, where the random permutation model results by considering the limit $q \to 1$.

Theorem 2.19 (Archibald and Martínez, 2009) *For the strategy "hiring above the m-th best", let $g_{n,m}$ denote the gap of last hired candidate after n interviews. Then, for $1 \leq m \leq n$,*

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

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

It follows that $g(\sigma) = 0$ for hiring above the best as the hiring set contains the best candidate seen ever, and goes quickly to 0 as n grows for $m \geq 1$.

Hiring above the α-quantile of the hired staff

Archibald and Martínez introduced a generalization of "hiring above the median" as defined here

Definition 2.9 *The strategy "hiring above the α-quantile" hires the first candidate in the sequence, then any further candidate is hired if and only if his relative rank is larger than the α-quantile, $0 < \alpha < 1$, of the already hired staff. The α-quantile of a sequence $x_1 < x_2 < \cdots < x_k$ of k elements is the element x_j with $j = \lceil \alpha k \rceil$.*

Thus $X(\sigma) = k - \lceil \alpha h(\sigma) \rceil + 1$, $|\sigma| \geq 1$. The results given in [5] for this general strategy are mentioned as follows:

Theorem 2.20 (Archibald and Martínez, 2009) *For the strategy "hiring above the α-quantile", $0 < \alpha < 1$, let h_n and g_n denote the size of the hiring set and the gap of last hired candidate after n interviews, respectively. Then the exact growth order of the r-th integer moments of h_n:*

$$\mathbb{E}\{h_n^r\} = \Theta\left(n^{(1-\alpha)r}\right).$$

The exact growth order for the expectation of g_n *is*

$$\mathbb{E}\{g_n\} = \Theta(n^{-\alpha}).$$

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

Archibald and Martínez also mentioned another parameter, the *index of last hired candidate*. This parameter helps us to study the dynamics of the hiring problem. They gave only the differential equation that describes the behaviour of this parameter.

Theorem 2.21 (Archibald and Martínez, 2009) *Let* $L(z, u)$ *be the generating function*

$$L(z, u) = \sum_{\sigma \in \mathcal{P}} \frac{z^{|\sigma|}}{|\sigma|!} u^{L(\sigma)},$$

where $L(\sigma)$ *is the index of last hired candidate in* σ.
Then

$$(1 - z)\frac{\partial L(z, u)}{\partial z} - L(z, u) = u \sum_{\sigma \in \mathcal{P}} X(\sigma) \frac{(zu)^{|\sigma|}}{|\sigma|!} - \sum_{\sigma \in \mathcal{P}} X(\sigma) \frac{z^{|\sigma|}}{|\sigma|!} u^{L(\sigma)}.$$

The proof of this theorem is similar to that one of Theorem 1.4.

2.5 The Chinese restaurant process

Here we review a similar process to those having on-line decision-making fashion, so it is of interest when discussing the hiring problem. Pitman introduced, in his notes: *Combinatorial Stochastic Processes* [77], the *Chinese restaurant process* (CRP) under the so-called two-parameter model (α, θ). In fact, these notes [77] give a rich study of the properties of some combinatorial models of random partitions and random trees and their relations with stochastic processes with independent increments. Let us focus on the formulation of the CRP and its distributional and asymptotic results related to the main studied quantity: the *number of blocks*, or this consistent name: the *number of occupied tables after receiving* n *customers*, K_n.

Definition 2.10 *Suppose that initially (time* 0*) there is an empty restaurant with an infinite number of unlimited circular tables. Customers arrive at discrete time events. Let the first customer (arriving at time* 1*) be seated at table* $\#1$. *Assume that after time* n*, we have* n *customers seated at* k *tables. Then the* $(n + 1)$*-th customer is seated according to the following probabilistic rule, that is called "seating plan* (α, θ)*":*

- *he is placed at an unoccupied table with probability* $\frac{k\alpha+\theta}{n+\theta}$,
- *he is placed at the occupied table* $\#i$ *with probability* $\frac{n_i-\alpha}{n+\theta}$*, if table* $\#i$ *has* n_i *customers (note that* $\sum_{i=1}^{k} n_i = n$*).*

According to this rule, there are three possible classes (that satisfy the conditions of probability) of seating plans can be induced:

- `Case` #1: $\alpha = -\kappa < 0$ and $\theta = m\kappa$ for $m = 1, 2, \ldots$

- `Case` #2: $\alpha = 0$ and $\theta > 0$.
- `Case` #3: $0 < \alpha < 1$ and $\theta > -\alpha$.

Thus under any particular seating plan, the sequence $(K_n)_{n\geq 1}$ is a Markov chain, with initial value $K_1 = 1$ and increments in $\{0, 1\}$, and inhomogeneous transition probabilities

$$\mathbb{P}\{K_{n+1} = k+1 | K_1, \ldots, K_n = k\} = \frac{k\alpha + \theta}{n+\theta},$$
$$\mathbb{P}\{K_{n+1} = k | K_1, \ldots, K_n = k\} = \frac{n - k\alpha}{n+\theta}.$$

The distribution of the r.v. $K_n^{(\alpha,\theta)}$, under a seating plan (α, θ), is given by:

$$\mathbb{P}\left\{K_n^{(\alpha,\theta)} = k\right\} = \alpha^{k-1}\frac{\Gamma(k + \frac{\theta}{\alpha})\Gamma(\theta+1)}{\Gamma(n+\theta)\Gamma(\frac{\theta}{\alpha}+1)} S_{n,k}^{-1,-\alpha}, \tag{2.2}$$

where $S_{n,k}^{-1,-\alpha}$ represents a generalization of Stirling numbers of the first kind (i.e., $S_{n,k}^{-1,0}$ is the unsigned Stirling numbers of the first kind, refer to [77] for more details), and it can be computed after extracting the coefficients of the generating function below:

$$S_{n,k}^{-1,-\alpha} = \frac{n!}{k!}[\zeta^n]\big(w^{-1,-\alpha}(\zeta)\big)^k,$$

where

$$w^{-1,-\alpha}(\zeta) = \begin{cases} \frac{1}{\alpha}\big(1 - (1-\zeta)^\alpha\big), & \text{if } \alpha \neq 0, \\ \log\frac{1}{1-\zeta}, & \text{if } \alpha = 0. \end{cases}$$

The expected value of $K_n^{(\alpha,\theta)}$ is given as follows:

$$\mathbb{E}\left\{K_n^{(\alpha,\theta)}\right\} = \begin{cases} \displaystyle\sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}, & \text{if } \alpha = 0, \\ \frac{\Gamma(n+\theta+\alpha)\Gamma(\theta+1)}{\alpha\Gamma(n+\theta)\Gamma(\theta+\alpha)} - \frac{\theta}{\alpha}, & \text{if } \alpha \neq 0. \end{cases} \tag{2.3}$$

Pitman introduced also the asymptotic properties of K_n for different cases as follows:

- `Case` #1: $\alpha < 0$. Then $\theta = -m\alpha$, and for large n, $K_n \xrightarrow{(a.s.)} m$.
- `Case` #2: $\alpha = 0$. If we consider r.v. indicators X_i at the i-th arriving customer, then X_i are Bernoulli$\big(\frac{\theta}{\theta+i-1}\big)$ variables, hence the following theorem characterizes the limit distribution of $K_n^{(0,\theta)}$ in this case:

  Theorem 2.22 (Pitman, 2006) *For the seating plan $(0, \theta)$, let $K_n^{(0,\theta)}$ denote the number of occupied tables after receiving n customers, then as $n \to \infty$:*

$$\frac{K_n^{(0,\theta)} - \theta \log n}{\sqrt{\theta \log n}} \xrightarrow{(a.s.)} \mathcal{N}(0, 1).$$

- `Case` #3: $0 < \alpha < 1$. The expectation of K_n follows from (2.3):

$$\mathbb{E}\{K_n\} \sim \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)} n^{\alpha},$$

which indicates the suitable normalization for a limit law as stated in the following theorem:

Theorem 2.23 (Pitman, 2006) *For the seating plan (α, θ), with $0 < \alpha < 1$ and $\theta > -\alpha$, let K_n denote number of occupied tables after receiving n customers, then as $n \to \infty$:*

$$\frac{K_n}{n^{\alpha}} \xrightarrow{(a.s.)} S_{\alpha},$$

with continuous distribution, for $s > 0$:

$$\frac{d}{ds}\mathbb{P}\{S_{\alpha} \in ds\} = g_{\alpha,\theta}(s) = \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} s^{\theta/\alpha} g_{\alpha}(s),$$

where $g_{\alpha} = g_{\alpha,0}$ is the Mittage-Leffler *density defined as follows:*

$$g_{\alpha}(s) = \frac{1}{\pi\alpha} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \Gamma(\alpha k + 1) s^{k-1} \sin(\pi\alpha k).$$

In this sense, it is said that the normalized r.v. $\frac{K_n}{n^{\alpha}}$ is a "variant" of the Mittage-Leffler distribution.

We emphasize that a seating plan (α, θ) is not an ordinary selection rule or hiring strategy but rather an on-line decision-making procedure which is defined by the transition probabilities of increment of the parameter K_n. Thus K_n and the *number of hired candidates* in the hiring problem, h_n, are two Markovian r.v.'s with increments in $\{0, 1\}$. According to Definition 2.10, since the probability of opening a new table depends directly on the number of open tables so far, then it is obvious to check similarity between some seating plans and the corresponding "rank-based" hiring strategies. In some cases like the seating plan $(0, m)$ and the strategy "hiring above the m-th best", both $K_n^{(0,m)}$ and $h_{n,m}$ are equivalent. In other cases like the seating plan $(\frac{1}{2}, 1)$ and the strategy "hiring above the median" (also the "$\frac{1}{2}$-percentile rule"), $K_n^{(\frac{1}{2},1)}$ and h_n are very similar but not equivalent in neither the distribution nor their asymptotic behaviour. We will investigate such relationships in detail in the next chapters.

Some simple **Pólya's urn schemes** [67] share similar aspects with the hiring problem. For example (see [77]), let U denote an urn and we have the following scheme:

1. Initially U contains two balls with two distinct colors c_1 and c_2.
2. Draw one ball from U, call it b then:
   - The color of b is noted then it is placed back in U.
   - If the color of b was never drawn before, then place two balls with new distinct colors c_+, c_{++} in U.

- Otherwise, place two balls of the same color of b in U.

3. Repeat step # 2.

Thus under this probabilistic model, one quantity of interest is the *number of distinct colors in the urn*, C_n. If we stop this process after drawing n balls from the urn, then C_n behaves exactly like $K_n^{(\frac{1}{2},0)}$ in the CRP. It is, in turn, close to h_n for hiring above the median.

In the same context, some schemes in the **balls and bins** problem are closely related to both CRP and the hiring problem. Consider the following process involving balls and bins:

- Initially there are m bins, for a positive integer $m > 1$, each contains one ball.
- Balls arrive one at a time. For each new ball:
  - With probability p, create a new bin and place the ball in it.
  - Place the ball in the i-th bin with probability $1 - p$, which is proportional to n_i^γ, $\gamma \in \mathbb{R}$ where n_i is the number of balls in the i-th bin.

This process involves a generalization for Pólya's urn problem (see [20] for more details). For the particular setting $\gamma = 1$ and $p = \frac{m}{n+m}$, the n-th ball has probability $\frac{n_i}{n+m}$ to be placed in the i-th bin. Now if we consider the parameter: *number of bins after placing n balls*, B_n under this special process, then B_n is equivalent to $K_n^{(0,m)}$ the number of tables of the CRP with seating plan $(0, m)$ and $h_{n,m}$ the number of hired candidates for hiring above the m-th best.

2.6 General discussion

In this section we give some important conclusions and remarks concerning all the studies of the hiring problem so far. All analyzed hiring strategies in the context of the hiring problem are "pragmatic". The notion of pragmaticity (not under that name) was introduced by Krieger et al. [59], as stated in Definition 2.1, where a rank-based strategy is pragmatic if it satisfies the LsD scheme. Archibald and Martínez introduced explicitly an equivalent definition for pragmaticity (Definition 2.7). For pragmatic rank-based strategies, the *threshold candidate* at any time of the hiring process should be one candidate in the hiring set, the score of this threshold always goes up (its quality is improving all the time), and the number of choices to hire the next candidate is an integer for any finite sequence of candidates. Score-based strategies like the "β-better-than-average" family studied in [15, 60, 61] are also pragmatic, although a slightly different technical definition of pragmaticity is needed. The *hiring threshold* for such strategies cannot decrease during the hiring process, but this threshold (which is essentially a real number in $(0, 1)$) may be a score of one hired candidate before or not. The "potential of hiring" for score-based strategies is the length of the gap between the threshold score and the maximum score (say, 1).

On the other hand, similar on-line decision-making procedures like the seating plans (α, θ) of the CRP are defined by the *transition probabilities of increment* for the quantities of interest. Thus one feasible or valid seating plan should obey the general conditions of probability, i.e., the probabilities of all taken decisions at some moment sum up to 1.

We point out that, for rank-based strategies, the absolute scores model when the scores are i.i.d. r.v.'s from a continuous distribution, and the random permutation model are equivalent. The behaviour of any rank-based hiring strategy, in terms of the number of hired candidates, waiting time, etc., is "distribution-free" when only ranks of candidates are considered. This remark was already given by Krieger et al. [60] and also has been discussed by others, e.g. [5]. As we have seen, the number of hired candidates under the "$\frac{1}{2}$-percentile rule" (also for "hiring above the median") is $\Theta(\sqrt{n})$ regardless of the distribution. For "better-than-average rule ", which is not a rank-based strategy, the number of hired candidates grows with order $\Theta(\sqrt{n})$ for the Exponential distribution, but with order $\Theta(n^{2/3})$ if the scores follow a Uniform distribution (Theorem 2.9, where the Uniform distribution is exactly Beta$(1, 1)$ distribution).

We gave some examples showing the connections between related sequential selection processes like the hiring problem, the CRP, Polya's urn schemes, and balls and bins models. Establishing such connections is useful where the results obtained for some problem may give new insights for another related one and vice-versa, as we will see in next chapters. We have seen various approaches to analyze those processes, but since there are many instances of isomorphism between particular cases of them regarding the main parameter the *number of hired/selected/distinct/... items*, then it is worth to think of a "unified" framework which can analyze similar selection rules. This point will be discussed in Chapter 5.

Part III

Results and Applications of the Hiring Problem

This part contains the main contributions of the thesis. Chapter 4 introduces a detailed study of the strategy "hiring above the median". Definition 2.9 explains the behaviour of this strategy which is a special case of "hiring above the α-quantile" when $\alpha = \frac{1}{2}$. We investigate this hiring strategy under the setup of Archibald and Martínez, but we use a direct approach to obtain our results, rather than relying on more general techniques of Analytic Combinatorics. We consider the relation between the score of the threshold candidate (current median) and the *number of hired candidates* (the size of the hiring set), then our recursive approach depends on distinguishing between two cases according to the parity of the size of the hiring set, whether it is odd or even. Thus we obtain two fundamental quantities, namely $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$, the probabilities that, after interviewing n candidates, the threshold candidate has the ℓ-th largest score amongst all candidates seen so far and an odd or an even number of candidates has been hired, respectively. Many quantities studied for this strategy can be expressed easily in terms of $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$, while other quantities are obtained by appropriate generalization of them.

We obtain many results for this strategy involving the exact and limiting probability distributions of many hiring parameters that give us a precise characterization of the strategy. The connections between hiring above the median, and similar on-line selection processes like the "$\frac{1}{2}$-percentile rule" (see Subsection 2.2.1) and the seating plan $(\frac{1}{2}, 1)$ of the Chinese restaurant process (CRP) (see Section 2.5) are also investigated.

Moreover, our approach proves useful to obtain explicitly the exact and limiting probability distributions of the *number of retained items* for the $\frac{1}{2}$-percentile rule, which are slightly different from those of hiring above the median. We also obtain the results of an interesting quantity that is the *waiting time* for both the $\frac{1}{2}$-percentile rule and the seating plan $(\frac{1}{2}, 1)$. The main contributions and results of hiring above the median have been published in [51], and a journal version of that paper in [52].

Chapter 5 contains our study of the strategy "hiring above the α-quantile", see Definition 2.9. For the general case, $0 < \alpha < 1$, we show that the framework given in Section 2.4 can give us at least the order of growth of several basic quantities, namely the expectation of many parameters like the *number of hired candidates*, the *gap of last hired candidate*, and the *number of replacements*. Using the recursive approach, used in Chapter 4, leads to explicit results for the *number of hired candidates* for rational α, where $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$. The work on this strategy is still on-going, where we try to obtain similar results for other particular cases, i.e., $\alpha = \frac{2}{3}, \frac{3}{4}, \ldots$, that might give us some intuition of the case of rational α, where $\alpha = \frac{p}{q}$, with $\gcd(p, q) = 1$. The current results and contributions of hiring above the α-quantile are available in the technical report [49].

Chapter 6 is dedicated to the analysis of "hiring above the m-th best" strategy, which is defined in Definition 2.8. The behaviour of this strategy is quite simple, so that in most cases starting from the definition of each studied parameter leads in a straightforward manner to the desired analysis. The contributions of this chapter include the distributional results for many parameters, clarifying the connections with m-records, the relationship between this strategy and the seating plan $(0, m)$ of the CRP, and the results of the *waiting time* parameter for the seating plan $(0, m)$. We have published our preliminary results of hiring above the m-th best in [50], and submitted a journal version, [48] containing all the details and additional results.

Chapter 7 introduces one application of the results obtained for "hiring above the m-th best" to

the design and analysis of data streaming algorithms. More specifically, the explicit distributional results for the *number of hired candidates* when the strategy hiring above the m-th best is used to process the sequence of candidates lead us to the design of a novel estimator, called RECORDINALITY, of the number of distinct elements in a very large sequence which may contain repetitions; this problem is known in the literature as "cardinality estimation". We prove that RECORDINALITY is an unbiased cardinality estimator and quantify precisely its accuracy, in terms of the standard error. The estimator needs $\Theta(m \log n)$ bits of memory and we show how the accuracy improves as we make the parameter m larger. We provide also some experimental results that support our theoretical findings, compare the estimator with other existing ones and show its reliability. Moreover, we discuss some promising ideas around RECORDINALITY and our new approach to data streaming analysis. The contributions related to RECORDINALITY have been published in [47].

Besides RECORDINALITY, we show that other hiring parameters can be useful also in such applications; we introduce another cardinality estimator called DISCARDINALITY which is based upon the *best discarded candidate*, again using hiring above the m-th best. We give the analysis of DISCARDINALITY and compute its standard error. DISCARDINALITY is not as interesting as RECORDINALITY from a practical point of view, but we think that it might be useful as a basis for the estimation of the *similarity index* of two data sets.

Chapter 3 contains the necessary introductory material before discussing the analysis of the studied hiring strategies. All strategies studied in this thesis are rank-based, and we assume the random permutation model. We review the combinatorial model of the hiring problem in Section 3.1. We give the formal definition of all hiring parameters (the random variables of interest) in Section 3.2. These parameters can be defined for any rank-based hiring strategy not only the studied strategies here. For the strategy hiring above the median we study two more quantities: the *number of hired candidates conditioned on the first one*, as hiring above the α-quantile strategy is sensitive to the first hired candidate; and the *probability that the candidate with score* q *is getting hired* with $1 \leq q \leq n$. Those two quantities are explained in Chapter 4.

Since improving the average quality of the hired staff is one essential goal for any hiring strategy then we introduce in Section 3.3, "hiring with replacements" which combines a basis hiring strategy with a replacement mechanism. Hiring with replacements yields the ultimate best quality of the hired staff. For each new candidate, we use the basis strategy to decide if he is to be hired or not. However, if the candidate does not rank well enough to be hired but he is better than the worst hired one then this new candidate *replaces* the worst hired one. The quantity of interest in hiring with replacements is obviously the *number of replacements* done, so we analyze this parameter also for the studied strategies. The number of replacements depends on the number of hired candidates, thus it also gives implicitly some indication of the dynamics of the hiring process.
Hiring with replacements opens the door for simple and efficient distinct sampling algorithms in data streaming analysis, with added benefit that the size of the generated sample smoothly adapts to the (unknown) number of distinct elements.

Chapter 3

Preliminaries

3.1 Formal statement of the problem

As mentioned in Section 2.4, Archibald and Martínez introduced the combinatorial or discrete model of the hiring problem. This formulation of the problem opens the door to study many useful hiring parameters that characterize the hiring process when applying a certain hiring strategy. We specify here the mathematical formulation of the hiring problem.

i) Input: a sequence of relative ranks $S = s_1, s_2, \ldots, s_i, \ldots$ of the candidates. For a candidate with rank s_i, exactly $s_i - 1$ previous candidates rank worse than that candidate.

ii) The rank of the i-th candidate, $1 \leq s_i \leq i$, is uniformly distributed.

iii) Each finite sequence $s_1, \ldots, s_n$ represents a random permutation $\sigma^{(n)}$ of length n.

iv) A decision must be taken whether to hire the i-th candidate or not at step i.

v) Decisions are irrevocable.

vi) We have no information about the future.

Recall from Section 2.4 that the *indices* (arrival times) of the hired candidates are forming the *hiring set* denoted by $\mathcal{H}(\sigma)$. We use $\mathcal{Q}(\sigma)$ to denote the set of scores of hired candidates after processing the permutation σ of candidates using a specific hiring strategy. There is always a trade-off between two demands: hiring candidates at some reasonable rate and improving the "average" quality of the hired staff (i.e., the more candidates are hired, the worse could be the quality of the hired staff and vice-versa). We focus here on studying the behaviour of the hiring strategies via several hiring parameters.

3.2 Hiring parameters

There are two groups of parameters to investigate the two corresponding general aspects of the hiring process, namely, the hiring rate (the dynamics of the hiring process), and the quality of the hired staff. We have already seen some parameters like the *waiting time*, the *number of hired candidates* and the *gap of last hired candidate* in Chapter 2. Here we introduce more parameters, giving formal definitions of all of them. For each introduced parameter, we review related quantities

that were studied previously. We begin with the parameters related to the dynamics of the hiring process, we call them *dynamics indicators*.

Number of hired candidates. Our basic quantity is the random variable (r.v.) h_n, which gives the number of hired candidates (i.e., the size of the hiring set) for an input sequence of length n. We use the same notation, h_n, as Archibald and Martínez (see Section 2.4). This quantity was studied by Krieger et al. as L_n in the analysis of the p-percentile rules, (Subsection 2.2.1). Also, they used the r.v. M_n to denote the number of retained items for the β-better-than average rule, (Subsection 2.2.2).

Waiting time. Here we use the r.v. W_N to give the waiting time in terms of the number of interviews required to hire exactly N candidates. One might say that h_n and W_n are two faces of the same coin. Krieger et al. used the r.v. T_k to denote the waiting time for the β-better-than average rule. Broder et al. used also T_k to represent the number of interviews to hire k candidates for Lake Wobegon strategies (Section 2.3).

Index of last hired candidate. The r.v. L_n denotes the index of last hired candidate of a sequence $s_1, \ldots, s_n$ of length n, where $L_n = i$ if the i-th candidate is recruited, and no subsequent candidate j, $j > i$, is recruited. In other words, L_n represents the maximum index in $\mathcal{H}(\sigma^{(n)})$. If $\mathcal{H} = \emptyset$ then $L_n = 0$ by convention. Again we use same notation, L_n, as Archibald and Martínez.

Distance between the last two hirings. The r.v. Δ_n gives the distance (i.e., difference) between the indices of the last two recruited candidates in the input sequence. If we look at the hiring set after scanning the input sequence then Δ_n is the difference between the two maximum indices in $\mathcal{H}(\sigma^{(n)})$. By convention, if $h_n \leq 1$ then $\Delta_n = 0$.

Next we give a description of the quantities that measure the quality of the hired staff, we call them *quality indicators*.

Score of last hired candidate. The r.v. R_n gives the score of last hired candidate for a sequence of n candidates (i.e., the last "score" in $\mathcal{Q}(\sigma^{(n)})$ which corresponds to the maximum index in $\mathcal{H}(\sigma^{(n)})$). Directly related is the gap $g_n = 1 - \frac{R_n}{n}$ (introduced by Archibald and Martínez), which helps to measure how close the quality of the last hired candidate is compared to the topmost one. Krieger et al. have considered a related quantity A_n; the average rank/score of retained group, after n observations, for both the p-percentile rules and β-better-than average rule. Also the r.v. Y_k the average score of the first k selected items, was analyzed for the β-better-than average rule. Broder et al. analyzed the gap $G_i = 1 - A_i$, where A_i is the average quality of i hirings, for hiring above the mean. For hiring above the median, they defined the gap as $G_k = 1 - M_k$, where M_k is the median score of $2k + 1$ selected candidates.

Score of best discarded candidate. The r.v. M_n gives, for a sequence of n candidates, the score of best discarded candidate. After scanning the sequence of candidates, the strategy generates two sets of candidates: the set of selections (defined by $\mathcal{H}(\sigma^{(n)})$ and its related set of scores $\mathcal{Q}(\sigma^{(n)})$) and the rest of the sequence which forms the discarded set; thus M_n represents the maximum score of the discarded set. This parameter describes also how selective the hiring process is: a high value (close to n) of M_n means that the hiring strategy is very selective, whereas a low value

of M_n means that the strategy is hiring too many candidates.

3.3 Hiring with replacements

One interesting extension of the hiring problem is "hiring with replacements". This extension violates one condition of the problem statement, namely the restriction that "decisions are irrevocable". Briefly, it might happen that, at some step, a good candidate is discarded (because his rank is not good enough to become selected by the standard strategy), but he is better than the worst already 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 standard strategy will hire him, or he replaces the worst candidate amongst all hired ones. Thus, after interviewing a new candidate, the following three cases may appear:

i) we hire the candidate by a direct application of the underlying standard strategy;

ii) we hire the candidate despite the standard strategy would not, because he is better than the worst already hired candidate, but in this case the new candidate *replaces* the worst candidate of the hiring set (thus the number of hired candidates remains the same);

iii) we discard the new candidate, because his rank is worse than the score of the worst hired one.

We have the notation $\mathcal{H}(\sigma)$ and $\mathcal{Q}(\sigma)$ that denote the hiring set and the set of scores of hired candidates, respectively, for the standard strategy. Let $\mathcal{H}_R(\sigma)$, $\mathcal{Q}_R(\sigma)$ and $h_n^{(R)}$ denote the hiring set, the set of scores of hired candidates and its size when we combine the standard strategy with the replacement mechanism. As noticed above, for case $ii)$ above, the size of the hiring set does not change (we hire a candidate, but we fire another one), which implies $h_n^{(R)} = h_n$ (although, of course, in general $\mathcal{H}_R(\sigma) \neq \mathcal{H}(\sigma)$ and $\mathcal{Q}_R(\sigma) \neq \mathcal{Q}(\sigma)$). A direct result of hiring with replacements is stated as follows:

Theorem 3.1 *For any pragmatic strategy hiring h_n candidates, its combinations with the proposed replacement mechanism will hire exactly the best h_n candidates in the sequence. If $\sigma^{(n)}$ represents the scores of n candidates and $h_n = k$, then*

$$\mathcal{Q}_R\big(\sigma^{(n)}\big) = \{n-k+1, n-k+2, \ldots, n-1, n\}.$$

We introduce now the r.v. f_n, which measures the *number of replacements* done (i.e., the number of applications of case $ii)$ above) using the replacement mechanism together with some hiring strategy. f_n gives a measure about the "quality" of the standard strategy, since, if the hiring set obtained contains good candidates then we do not need many replacements to obtain the set of the best candidates and vice-versa. f_n also gives an implicit indication of the dynamics of the hiring process, where it depends directly on h_n. So that this quantity combines the dynamical and quality aspects of the hiring process.

Comment

As already hinted out in Section 2.6, the random permutation model and the absolute quality scores model of the input sequence of candidates, are *equivalent* if we take only into consideration

the relative ranks between candidates. Thus the *dynamics indicators* and the *number of replacements* associated to any applied rank-based hiring strategy are "distribution-free", regardless of the underlying distribution of scores. On the other hand, it is clear that the values of the *quality indicators* depend directly on the probability distribution from which the scores are drawn.

Chapter 4

Hiring above the median

4.1 Introduction

This chapter is devoted to the detailed study of the rank-based strategy "hiring above the median". The strategy was introduced originally by Broder et al. [15], but we study it under the combinatorial model of the hiring problem introduced by Archibald and Martínez in [5] (explained in Section 2.4). Given a sequence of candidates of length n that is modeled as a random permutation (best candidate has a rank n whereas the worst one has 1), then, according to Definition 2.9, the strategy hires the first candidate in the sequence and thereafter any coming candidate is hired if and only if his rank is better than the median score of all previously hired candidates. We use here the convention (refer to Section 2.3) that the median of a set of k (distinct) elements $x_1 < x_2 < \cdots < x_k$ is $x_{\lceil \frac{k}{2} \rceil}$.

As an example, if we process the sequence of scores $\sigma^{(8)} = \underline{3}\,\underline{5}\,2\,\underline{8}\,1\,\underline{7}\,4\,\underline{6}$ using "hiring above the median" then $\mathcal{H}(\sigma^{(8)}) = \{1, 2, 4, 6, 8\}$ and $\mathcal{Q}(\sigma^{(8)}) = (\text{the underlined scores in } \sigma^{(8)})$, since each of these candidates has, at the time of hiring, a rank better than the median of the previously hired candidates. The number of hired candidates $h_8 = 5$, the index of last hired candidate $L_8 = 8$, the distance between the last two hirings $\Delta_8 = 2$, the score of last hired candidate $R_8 = 6$, and the score of best discarded candidates $M_8 = 4$. If we apply the proposed replacement mechanism in Section 3.3, then we have $\sigma^{(8)} = 3\,\underline{5}\,2\,\underline{8}\,1\,\underline{7}\,\underline{4}\,\underline{6}$ and the underlined scores represent $\mathcal{Q}_R(\sigma^{(8)})$ with hiring set $\mathcal{H}_R(\sigma^{(8)}) = \{2, 4, 6, 7, 8\}$. In this example we have the number of replacements $f_8 = 1$, since the candidate with score 4 *replaces* the one with score 3 during the hiring process.

Besides those hiring parameters explained in Sections 3.2 and 3.3, we study the following quantities for hiring above the median:

1) *Probability p_n that the n-th candidate in a sequence of candidates is getting hired.* This quantity is closely related to the number of hired candidates, h_n, and gives also some insight about the hiring rate of this strategy.

2) *Size of the hiring set conditioned on the first candidate.* It has been noticed in [15] 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 the first position.

Of course, if we denote by U_n the score of the first candidate, it holds $h_{n,U_n} = h_n$, where, for the considered random permutation model, U_n is uniformly distributed on $\{1, 2, \ldots, n\}$.

3) *Probability that the candidate with score* q *is getting hired.* We introduce the probabilities $p_{n,q}$, which give the probability that the candidate with score q in a sequence of n candidates is getting hired. Notice that trivially $\frac{1}{n} = p_{n,1} \leq p_{n,2} \leq \cdots \leq p_{n,n} = 1$ by considering the probabilities conditioned on the event that the candidate with score q appears at first position, second position, etc., in the sequence. These probabilities can also be used to give a first result for the r.v. S_n measuring the *total score of the set of hired candidates*, i.e., the sum of the scores of $\mathcal{Q}(\sigma)$ (a quantity, which seems difficult to treat directly via the proposed recursive approach), since $\mathbb{E}\{S_n\} = \sum_{q=1}^{n} q \cdot p_{n,q}$.

The sequel of this chapter is organized as follows: Section 4.2 introduces the results of many studied parameters under this strategy. Section 4.3 gives the analysis and proofs of all theorems, starting with the explanation of the used recursive approach. Section 4.4 explains in detail the relationship between hiring above the median strategy and two similar selection processes: the $\frac{1}{2}$-percentile rule and the seating plan $(\frac{1}{2}, 1)$ of the CRP; it also contains new results for both mentioned processes. The chapter ends with some conclusions and future work in Section 4.5. The results of this chapter have appeared in [51, 52].

4.2 Results

Theorem 4.1 *Let* h_n *denote the size of the hiring set after* n *interviews. Then the exact distribution of* h_n *is given as follows, with* $1 \leq k \leq n$*:*

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

Asymptotically, as $n \to \infty$, $\frac{h_n}{\sqrt{n}} \xrightarrow{(d)} \hat{R}$, *where* $\hat{R}$ *is* Rayleigh *distributed with parameter* $\sigma = \sqrt{2}$*. Furthermore, the expectation of* h_n *satisfies:* $\mathbb{E}\{h_n\} = \sqrt{\pi n} + \mathcal{O}(1)$.

Corollary 4.1 *Let* p_n *denote the probability that the* n*-th interviewed candidate is getting hired. Then the probability* p_n *is given by the following exact formula (valid for* $n \geq 2$*, whereas* $p_1 = 1$*), for which the stated asymptotic expansion holds (for* $n \to \infty$*):*

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

Theorem 4.2 *Let* W_N *denote the "time" (i.e., the number of candidates that have to be interviewed) until exactly* N *candidates are hired, then* W_N *is distributed as follows:*

$$\mathbb{P}\{W_N = t\} = \begin{cases} \frac{\ell}{t} \cdot \frac{\binom{t-1-\ell}{\ell-2}}{\binom{t-1}{\ell-1}}, & \text{for } N = 2\ell - 1 \text{ odd and } N \geq 3, \\ \frac{\ell}{t} \cdot \frac{\binom{t-1-\ell}{\ell-1}}{\binom{t-1}{\ell}}, & \text{for } N = 2\ell \text{ even and } N \geq 2, \end{cases}$$

with $\mathbb{P}\{W_1 = 1\} = 1$.
Asymptotically, as $N \to \infty$, $\frac{W_N}{N^2} \xrightarrow{(d)} \hat{W}$, *where* $\hat{W}$ *has the following density function:*

$$f_W(x) = \frac{1}{4x^2} \cdot e^{-\frac{1}{4x}}, \quad \textit{for } x > 0.$$

(Note that the moments of $\hat{W}$ *do not exist.)*

Theorem 4.3 *Let* L_n *denote the index of last hired candidate after* n *interviews. Then the exact distribution of* L_n *is given as follows:*

$$\mathbb{P}\{L_n = m\} = \sum_{\ell=1}^{m-1} \frac{\binom{m-1-\ell}{\ell-2}}{\binom{n}{\ell}} + \sum_{\ell=1}^{m} \frac{\ell-1}{\ell} \cdot \frac{\binom{m-\ell}{\ell-2}}{\binom{n}{\ell}}, \quad \textit{for } 1 < m \leq n,$$

and $\mathbb{P}\{L_n = 1\} = \frac{1}{n}$.
Asymptotically, as $n \to \infty$, $\frac{n-L_n}{\sqrt{n}} \xrightarrow{(d)} \hat{L}$, *where* $\hat{L}$ *has the following density function*

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

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

Theorem 4.4 *Let* Δ_n *denote the distance between the last two hirings after* n *interviews. Then the exact distribution of* Δ_n *is given as follows:*

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

Asymptotically, as $n \to \infty$, $\frac{\Delta_n}{\sqrt{n}} \xrightarrow{(d)} \hat{\Delta}$, *where* $\hat{\Delta}$ *has the following density function*

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

Theorem 4.5 *Let* $h_{n,q}$ *denote the size of the hiring set after* n *interviews, conditioned on the event that the score of the first candidate is* q*. Then the exact distribution of* $h_{n,q}$ *is given as follows, with* $1 \leq k, q \leq n$*:*

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

Asymptotically, as $n \to \infty$, *and provided that* $n - q \to \infty$, $\frac{h_{n,q}}{\sqrt{n-q}} \xrightarrow{(d)} \hat{h}$, *where* $\hat{h}$ *has the following density function*

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

Theorem 4.6 *Let R_n denote the score of last hired candidate after n interviews. Then the exact distribution of R_n is given as follows:*

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

Asymptotically, as $n \to \infty$, $\frac{n-R_n}{\sqrt{n}} \xrightarrow{(d)} \hat{R}$, where $\hat{R}$ has the following density function

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

Theorem 4.7 *Let M_n denote the score of best discarded candidate after n interviews. Then the exact distribution of M_n is given as follows, for $1 \le r \le n-1$:*

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

and further $\mathbb{P}\{M_n = 0\} = \frac{1}{\binom{n}{\lceil \frac{n}{2} \rceil}}$.

Asymptotically, for $n \to \infty$, $\frac{n-M_n}{\sqrt{n}} \xrightarrow{(d)} \hat{M}$, where $\hat{M}$ is Rayleigh *distributed with parameter $\sigma = \frac{1}{\sqrt{2}}$.*

Theorem 4.8 *Let $p_{n,q}$ denote the probability that the candidate with score q of n interviewed candidates is getting hired. Then the probabilities $p_{n,q}$ are, for $1 \le q \le n$, given as follows:*

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

Theorem 4.9 *Let f_n denote the number of replacements done after processing n candidates using the mechanism "hiring with replacements". Then asymptotically, as $n \to \infty$, the expectation of f_n is given as follows:*

$$\mathbb{E}\{f_n\} = \sqrt{\pi n} + \mathcal{O}(\log n).$$

4.3 Analysis

Here we give the detailed calculations leading to the results stated in Section 4.2. Due to the explicit nature of the obtained exact formulas, the stated asymptotic results follow by applying Stirling's formula for the factorials (1.2) in connection with standard techniques as Euler-Maclaurin formula (1.4).

It is also worth to mention that we have made a few "sanity checks" of our theoretical findings on the one hand with the exact probabilities if the number n of candidates is small and in the other hand by carrying out experimental studies for n large and they match well [46].

4.3.1 Outline of the analytical approach

As explained previously in Section 4.1, the median (i.e., the "lower" median) of a set of k (distinct) elements is the ℓ-th largest element with $\ell = \lceil \frac{k+1}{2} \rceil = \lfloor \frac{k+2}{2} \rfloor$. In this strategy, each of the selected candidates has, at the time of hiring, a rank better than the *median score* of the previously hired candidates. Thus, during the hiring process, the median of the current set of scores of hired candidates could be considered as the *threshold candidate*, who is actually used to make the decision, whether a new candidate is recruited (if he has a rank larger than the score of the threshold candidate) or not (otherwise).

It is a simple but quite useful observation that, when applying "hiring above the median", at each time of the hiring process all candidates seen so far with a score larger than the current threshold candidate are contained in the hiring set (Theorem 2.16). Thus there is a simple relation between the score of the threshold candidate 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 threshold candidate 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 k is even, i.e., $\ell = \lceil \frac{k+1}{2} \rceil$. And this yields the basis of the recursive approach used here, 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 threshold candidate.

Many of the parameters considered here can be expressed using the following two sequences of numbers:

- $a^{[1]}_{n,\ell}$: the probability that, after n interviews, the threshold candidate has the ℓ-th largest score amongst all candidates seen so far and an *odd* number of candidates has been hired.
- $a^{[2]}_{n,\ell}$: the probability that, after n interviews, the threshold candidate has the ℓ-th largest score amongst all candidates seen so far and an *even* number of candidates has been hired.

Moreover, the remaining parameters are obtained by studying extensions of this approach (as we will see later, e.g., M_n in Subsection 4.3.8).

Evolution of the median. During the hiring process, after the first hiring, every candidate becomes the median of all hired candidates remains as the threshold of this strategy for two consecutive hirings, then the next best one in the rank order becomes the threshold. The following table shows the evolution of the median (the ℓ-th largest) against the first few values of the size of the hiring set k.

k	1	2	3	4	5	6	7	$\cdots$
ℓ	1	2	2	3	3	4	4	$\cdots$

Thus, if a new candidate (better than the threshold) is hired and the size of the hiring set is odd, then the threshold candidate remains the same, and his rank has to be increased (because there

is one more hired candidate better than him). While for even size of the hiring set, hiring a new candidate leads to change the current threshold candidate to be the next best one in the rank order, that means the rank of the threshold candidate is the same but the candidate himself has been changed. This is simply explained in Figure 4.1. The "automaton" stated in Figure 4.2 de-

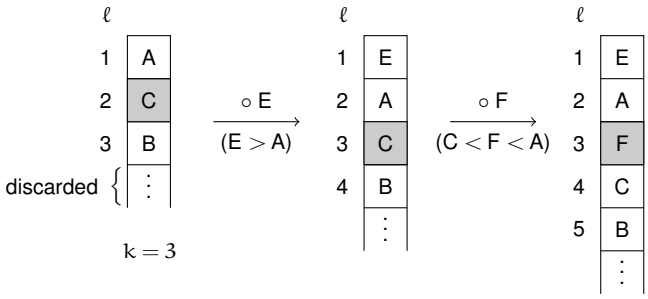


Figure 4.1: A "snapshot" of the hiring process under hiring above the median. We start after hiring three candidates ($k = 3$); then two consecutive hirings occurred. The threshold candidate (i.e., the median) is marked in gray. "ℓ" represents the rank of each candidate.

scribes then the "transition probabilities", when at time $n + 1$ a new candidate appears, whose rank will be compared with the score of the threshold candidate, which is the ℓ-th largest amongst the first n candidates. 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 Markov chain with two states; 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 [54].

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.

4.3.2 Size of the hiring set

As follows from the remarks given in Subsection 4.3.1, the defined sequences $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$, fully determine the probability distribution of the r.v. h_n measuring the number of hired candidates

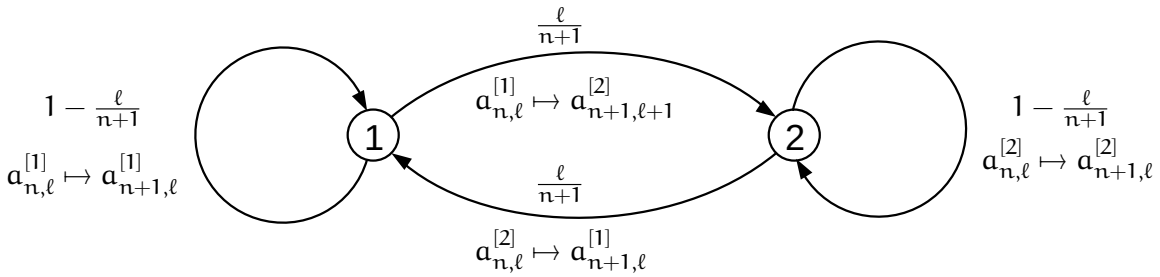


Figure 4.2: The "automaton" describing the transition probabilities of the underlying Markov chain for the numbers $a^{[1]}_{n,\ell}$ (state 1: *odd* size of the hiring set) and $a^{[2]}_{n,\ell}$ (state 2: *even* size of the hiring set). For example, moving from state 1 to state 2: at time n, the size of the hiring set is odd and the threshold candidate is the ℓ-th best hired, then the probability of increment (new hiring) at time $n+1$ (thus the size of the hiring set becomes even) is $\frac{\ell}{n+1}$ and the *rank* of the threshold candidate will increase ($\ell \to \ell+1$).

according to

$$\mathbb{P}\{h_n = k\} = \begin{cases} a^{[1]}_{n,\frac{k+1}{2}}, & \text{for } k \text{ odd}, \\ a^{[2]}_{n,\frac{k}{2}+1}, & \text{for } k \text{ even}. \end{cases} \tag{4.1}$$

From the description of the hiring process via the transition probabilities of the automaton given in Figure 4.2, the following recurrences for the probabilities $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ are deduced immediately (with initial values $a^{[1]}_{1,1} = 1$ and $a^{[2]}_{1,1} = 0$, and where we define $a^{[1]}_{n,\ell} = a^{[2]}_{n,\ell} = 0$ outside the range $1 \le \ell \le n$):

$$a^{[1]}_{n,\ell} = \frac{\ell}{n} \cdot a^{[2]}_{n-1,\ell} + \left(1 - \frac{\ell}{n}\right) \cdot a^{[1]}_{n-1,\ell}, \quad n \ge 2,\ 1 \le \ell \le n, \tag{4.2a}$$

$$a^{[2]}_{n,\ell} = \frac{\ell-1}{n} \cdot a^{[1]}_{n-1,\ell-1} + \left(1 - \frac{\ell}{n}\right) \cdot a^{[2]}_{n-1,\ell}, \quad n \ge 2,\ 1 \le \ell \le n. \tag{4.2b}$$

Since introducing (ordinary) generating functions for the numbers $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ would lead to systems of PDEs, for which a treatment seems to be rather involved, we proceed as follows. Either via algebraic manipulations or, alternatively, by considering the automaton given in Figure 4.2 (i.e., by considering the "time" when we changed from state 1 to state 2 until we are back again in state 1), we find that $a^{[2]}_{n,\ell}$ can be eliminated from the system of recurrences (4.2a)-(4.2b). This yields

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

and after simple manipulations we obtain the following recurrence for $a^{[1]}_{n,\ell}$:

$$(n-\ell) \cdot \binom{n}{\ell} a^{[1]}_{n,\ell} = (n-\ell) \cdot \binom{n-1}{\ell} a^{[1]}_{n-1,\ell} + \sum_{m=1}^{n-2} (\ell-1) \cdot \binom{m}{\ell-1} a^{[1]}_{m,\ell-1}. \tag{4.3}$$

The form of this recurrence for $a^{[1]}_{n,\ell}$ suggests to introduce suitable "normalizations" of these numbers via

$$b^{[1]}_{n,\ell} = \binom{n}{\ell} a^{[1]}_{n,\ell} \tag{4.4}$$

and to consider the corresponding generating function

$$B^{[1]}(z,u) = \sum_{n\geq 1} \sum_{1\leq \ell\leq n} b^{[1]}_{n,\ell} \cdot z^n u^\ell. \tag{4.5}$$

It is a routine task to show that the recurrence (4.3) yields the following linear first order PDE for $B^{[1]}(z,u)$:

$$z(1-z)\frac{\partial}{\partial z}B^{[1]}(z,u) + \left(zu - u - \frac{u^2z^2}{1-z}\right)\frac{\partial}{\partial u}B^{[1]}(z,u) - zB^{[1]}(z,u) = 0. \tag{4.6}$$

To get the general solution of a linear first order PDE one can apply the so-called "method of characteristics" (see, e.g., [90] for a description of this method), which also is a standard tool in computer algebra systems. One obtains that the general solution of (4.6) is given by

$$B^{[1]}(z,u) = \frac{1}{1-z} \cdot F\left(\frac{1-z-zu}{zu(1-z)}\right), \tag{4.7}$$

with an arbitrary differentiable function $F(x)$. To characterize the unknown function $F(x)$ in (4.7) we have to adapt the general solution to the initial conditions, which is not always trivial for PDEs. We are successful by introducing

$$\tilde{B}(z,u) = B^{[1]}\left(zu, \frac{1}{u}\right) = \sum_{n\geq 1} \sum_{0\leq \ell\leq n-1} b^{[1]}_{n,n-\ell} \cdot z^n u^\ell, \tag{4.8}$$

and considering the "diagonal" of the sequence $b^{[1]}_{n,\ell}$. Due to the combinatorial description of $a^{[1]}_{n,\ell}$ it follows that $a^{[1]}_{n,n} = 0$, for $n \geq 2$, which, together with the initial value $a^{[1]}_{1,1} = 1$, yields

$$b^{[1]}_{n,n} = \begin{cases} 1, & n = 1, \\ 0, & n \geq 2. \end{cases}$$

The latter values yield thus the useful initial condition $\tilde{B}(z,0) = z$ for $\tilde{B}(z,u)$, which, due to (4.7) and (4.8), can be written as follows:

$$\tilde{B}(z,u) = \frac{1}{1-zu} \cdot F\left(\frac{1-zu-z}{z(1-zu)}\right). \tag{4.9}$$

Plugging $u = 0$ into (4.9) gives

$$z = \tilde{B}(z,0) = F\left(\frac{1}{z} - 1\right),$$

and thus characterizes the function $F(x) = \frac{1}{x+1}$ occurring in (4.7). Therefore, we obtain from (4.7) the following astonishing simple solution for the generating function of the sequence $b^{[1]}_{n,\ell}$:

$$B^{[1]}(z,u) = \frac{1}{1-z} \cdot \frac{1}{1+\frac{1-z-zu}{zu(1-z)}} = \frac{zu}{1-z-z^2u}. \tag{4.10}$$

Extracting coefficients from the solution given in (4.10) is an easy task:

$$\begin{aligned} b^{[1]}_{n,\ell} &= [z^n u^\ell] B^{[1]}(z,u) \\ &= [z^n u^\ell] \frac{zu}{(1-z)\left(1-\frac{z^2}{1-z}u\right)} \\ &= [z^n] \frac{z}{1-z} [u^{\ell-1}] \frac{1}{1-\frac{z^2}{1-z}u} \\ &= [z^n] \frac{z}{1-z} \left(\frac{z^2}{1-z}\right)^{\ell-1} \\ &= [z^{n-2\ell+1}] \frac{1}{(1-z)^\ell} = \binom{n-\ell}{\ell-1}. \end{aligned}$$

Using (4.4), this immediately yields the following explicit formula for the numbers $a^{[1]}_{n,\ell}$:

$$a^{[1]}_{n,\ell} = \frac{b^{[1]}_{n,\ell}}{\binom{n}{\ell}} = \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}}. \tag{4.11}$$

Furthermore, by using the recurrence (4.2a) and plugging (4.11) into it, we also get an explicit expression for the numbers $a^{[2]}_{n,\ell}$:

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

which, after simple manipulations that are omitted here, yields

$$a^{[2]}_{n,\ell} = \frac{\binom{n-\ell}{\ell-2}}{\binom{n}{\ell-1}}. \tag{4.12}$$

Combining the results of (4.11) and (4.12) leads to the exact probability distribution of h_n as stated in Theorem 4.1.

Limit distribution. To characterize the limiting distribution of h_n we use the exact formulas for $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ given in (4.11) and (4.12) and show suitable asymptotic expansions. To get them we require the following expansion, which can be obtained easily by applying Stirling's formula:

$$c(n,\ell) = \frac{\binom{n-\ell}{\ell}}{\binom{n}{\ell}} = e^{-\frac{\ell^2}{n}}\cdot\left(1+\mathcal{O}\Big(\frac{\ell}{n}\Big)+\mathcal{O}\Big(\frac{\ell^3}{n^2}\Big)\right), \tag{4.13}$$

uniformly for $1 \le \ell \le n^{\frac{1}{2}+\epsilon}$, whereas these numbers are exponentially small for $\ell \ge n^{\frac{1}{2}+\epsilon}$. Since

$$a^{[1]}_{n,\ell} = \frac{\ell}{n+1-2\ell}\cdot c(n,\ell) \quad \text{and} \quad a^{[2]}_{n,\ell} = \frac{(\ell-1)(n-\ell+1)}{(n+1-2\ell)(n+2-2\ell)}\cdot c(n,\ell),$$

we furthermore get from (4.13) the following asymptotic expansions:

$$a_{n,\ell}^{[1]} \sim a_{n,\ell}^{[2]} = \frac{\ell}{n} e^{-\frac{\ell^2}{n}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell}\Big) + \mathcal{O}\Big(\frac{\ell}{n}\Big) + \mathcal{O}\Big(\frac{\ell^3}{n^2}\Big)\right), \tag{4.14}$$

uniformly for $1 \le \ell \le n^{\frac{1}{2}+\epsilon}$, whereas these numbers are exponentially small for $\ell \ge n^{\frac{1}{2}+\epsilon}$. Therefore, using (4.1), one obtains the following asymptotic equivalent of the probabilities $\mathbb{P}\{h_n = k\}$ valid, in particular, for $k \in [n^{\frac{1}{2}-\epsilon}, n^{\frac{1}{2}+\epsilon}]$:

$$\mathbb{P}\{h_n = k\} \sim \frac{k}{2n} e^{-\frac{k^2}{4n}}. \tag{4.15}$$

Setting $k = x\sqrt{n}$ implies that $\frac{h_n}{\sqrt{n}}$ converges, as $n \to \infty$, in distribution to a limiting r.v. $\hat{R}$ with density

$$\hat{f}(x) = \frac{x}{2} e^{-\frac{x^2}{4}}, \quad \text{for } x > 0, \tag{4.16}$$

i.e., to a Rayleigh distributed r.v. with parameter $\sigma = \sqrt{2}$.

The asymptotic result for the expectation

$$\begin{aligned} \mathbb{E}\{h_n\} &= \sum_{k=1}^{n} k\, \mathbb{P}\{h_n = k\} \\ &= \sum_{\ell=1}^{n} \left((2\ell-1) a_{n,\ell}^{[1]} + (2\ell-2) a_{n,\ell}^{[2]}\right), \end{aligned} \tag{4.17}$$

given in Theorem 4.1 follows from the uniform asymptotic expansion (4.14) for $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$ via

$$\begin{aligned} \mathbb{E}\{h_n\} &= \sum_{\ell=1}^{n^{\frac{1}{2}+\epsilon}} \frac{4\ell^2}{n} e^{-\frac{\ell^2}{n}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell}\Big) + \mathcal{O}\Big(\frac{\ell}{n}\Big) + \mathcal{O}\Big(\frac{\ell^3}{n^2}\Big)\right) \\ &= 4\sqrt{n} \int_0^\infty x^2 e^{-x^2}\, dx \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)\right) \\ &= \sqrt{\pi n} + \mathcal{O}(1). \end{aligned}$$

We remark that plugging (4.11) and (4.12) into (4.17) leads, after simple manipulations, to the following exact formula for $\mathbb{E}\{h_n\}$:

$$\mathbb{E}\{h_n\} = \sum_{\ell=1}^{n} \frac{\big(\ell(2\ell-1)(n+2-2\ell) + 2(\ell-1)^2(n+1-\ell)\big)}{\ell(n+1-\ell)} \frac{\binom{n-\ell+1}{\ell-1}}{\binom{n}{\ell}}. \tag{4.18}$$

We state as an immediate consequence of the preceding studies exact and asymptotic results for the quantity p_n, which gives the *probability that the n-th coming candidate in the sequence is getting hired*. Namely, since there are ℓ possibilities (out of n) for the n-th candidate being hired, if the threshold candidate has the ℓ-th largest score after $n-1$ interviews, we obtain

$$p_n = \sum_{\ell=1}^{n-1} \left(\frac{\ell}{n} \cdot a_{n-1,\ell}^{[1]} + \frac{\ell}{n} \cdot a_{n-1,\ell}^{[2]}\right), \quad \text{for } n \ge 2 \quad \text{and} \quad p_1 = 1. \tag{4.19}$$

Thus, using (4.11) and (4.12), we get from (4.19) after simple manipulations the exact formula for p_n stated in Corollary 4.1. Furthermore, by using the asymptotic expansion (4.14), one can easily evaluate p_n asymptotically, for $n \to \infty$, and also obtains the corresponding asymptotic result of Corollary 4.1.

4.3.3 Waiting time

Of course, the distribution of the quantity W_N, which measures the number of candidates that have to be interviewed until exactly N candidates are hired, is closely related to the distribution of h_n studied in Subsection 4.3.2. By considering the probability that exactly $N-1$ candidates amongst the first $t-1$ interviewed candidates are recruited and that the t-th candidate is also recruited (if he ranks better than the threshold candidate with the ℓ-th largest score; this happens with probability $\frac{\ell}{t}$), one immediately gets:

$$\mathbb{P}\{W_N = t\} = \mathbb{P}\{h_{t-1} = N-1\} \cdot \frac{\ell}{t} = \begin{cases} a^{[2]}_{t-1,\ell} \cdot \frac{\ell}{t}, & \text{for } N = 2\ell - 1 \text{ and } N \geq 3, \\ a^{[1]}_{t-1,\ell} \cdot \frac{\ell}{t}, & \text{for } N = 2\ell \text{ and } N \geq 2, \end{cases} \tag{4.20}$$

with $\mathbb{P}\{W_1 = 1\} = 1$, and where the quantities $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ are given in Subsection 4.3.2. Plugging (4.11) and (4.12) into (4.20) yields the explicit formula for the probability distribution stated in Theorem 4.2. Furthermore, by using the asymptotic expansion (4.13) we obtain the following expansion:

$$\mathbb{P}\{W_N = t\} = \frac{N^2}{4t^2}\, e^{-\frac{N^2}{4t}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{N}\Big) + \mathcal{O}\Big(\frac{N}{t}\Big) + \mathcal{O}\Big(\frac{N^2}{t^3}\Big)\right), \tag{4.21}$$

which leads, by setting $t = xN^2$, $x > 0$, to the limiting distribution result stated in Theorem 4.2.

4.3.4 Index of last hired candidate

Another quantity, where the results for $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ are of importance, is the index (i.e., the time) L_n of the last hired candidate. Namely, using simple reasoning, we can see that the probability that the m-th interviewed candidate is the last one hired satisfies:

$$\begin{aligned} \mathbb{P}\{L_n = m\} &= \mathbb{P}\{\text{We hire at position } m\} \cdot \mathbb{P}\{\text{No hirings from position } (m+1) \text{ till } n\} \\ &= \sum_{\ell=1}^{m} a^{[1]}_{m-1,\ell-1} \cdot \frac{\ell - 1}{m} \prod_{j=m}^{n-1} \left(1 - \frac{\ell}{j+1}\right) + \sum_{\ell=1}^{m-1} a^{[2]}_{m-1,\ell} \cdot \frac{\ell}{m} \prod_{j=m}^{n-1} \left(1 - \frac{\ell}{j+1}\right), \end{aligned} \tag{4.22}$$

for $1 < m \leq n$, whereas $\mathbb{P}\{L_n = 1\} = \frac{1}{n}$. Thus plugging the formulas (4.11) and (4.12) into (4.22) yields, after simple manipulations, the exact result stated in Theorem 4.3.

To characterize the limiting distribution of L_n we use the following asymptotic expansion that holds uniformly for $k, \ell = \mathcal{O}(n^{\frac{1}{2}+\epsilon})$, which is obtained by applying Stirling's formula:

$$\frac{\binom{n-k-1-\ell}{\ell-2}}{\binom{n}{\ell}} = \frac{\ell^2}{n^2}\, e^{-\frac{\ell(k+\ell)}{n}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell}\Big) + \mathcal{O}\Big(\frac{k+\ell}{n}\Big) + \mathcal{O}\Big(\frac{(k+\ell)^3}{n^2}\Big)\right). \tag{4.23}$$

Furthermore this expression is exponentially small for $\ell \geq n^{\frac{1}{2}+\epsilon}$ and arbitrary k. Thus, by setting $m = n - k$ in the exact formula for the probabilities $\mathbb{P}\{L_n = m\}$ given in Theorem 4.3 and using

(4.23), we get the following local approximation valid for $k = \mathcal{O}(n^{\frac{1}{2}+\epsilon})$:

$$\mathbb{P}\{L_n = n - k\} \sim 2 \sum_{\ell \geq 1} \frac{\ell^2}{n^2} e^{-\frac{\ell(k+\ell)}{n}}. \tag{4.24}$$

Setting $k = x\sqrt{n}$ and considering (4.24) as a Riemann sum of the corresponding integral we easily get the following asymptotic equivalent of the probabilities studied:

$$\mathbb{P}\{L_n = n - k\} \sim \frac{2}{\sqrt{n}} \int_0^\infty t^2 e^{-t(x+t)} dt, \quad \text{with } k = x\sqrt{n} \text{ and } x > 0. \tag{4.25}$$

The limiting distribution result stated in Theorem 4.3 immediately follows from (4.25).

To get an asymptotic expansion of the expectation $\mathbb{E}\{L_n\}$ it is advantageous to start with the explicit formula:

$$\begin{aligned}\mathbb{E}\{L_n\} &= \sum_{m=1}^{n} \frac{m}{\binom{n}{\ell}} \left(\sum_{\ell=1}^{m-1} \binom{m-1-\ell}{\ell-2} + \sum_{\ell=1}^{m} \frac{\ell-1}{\ell} \cdot \binom{m-\ell}{\ell-2} \right) \\ &= \sum_{\ell=1}^{n} \frac{(2\ell^2 - 3\ell + 1)n^2 - (3\ell^3 - 11\ell^2 + 10\ell - 3)n - 6\ell^3 + 12\ell^2 - 8\ell + 2}{\ell^2(n-\ell+1)} \cdot \frac{\binom{n-\ell+1}{\ell-1}}{\binom{n}{\ell}}, \end{aligned} \tag{4.26}$$

which can be obtained by first changing the order of summations, then applying standard combinatorial identities based on

$$\sum_{j=0}^{n} \binom{j}{\ell} = \binom{n+1}{\ell+1}. \tag{4.27}$$

Applying Stirling's formula to the summand in (4.26) yields

$$\mathbb{E}\{L_n\} = \sum_{\ell \geq 1} e^{-\frac{\ell^2}{n}} \cdot \left(2\ell - 3 + \frac{5\ell^2}{n} - \frac{2\ell^4}{n^2} \right) \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell^2}\Big) + \mathcal{O}\Big(\frac{\ell^6}{n^4}\Big) \right). \tag{4.28}$$

Then we can evaluate the occurring sums asymptotically, as $n \to \infty$, as follows:

$$\begin{aligned}\sum_{\ell \geq 1} \ell^j e^{-\frac{\ell^2}{n}} &= \frac{n^{j+1/2}}{2} \int_0^\infty e^{-t} t^{j-1/2} dt \\ &= \frac{1}{2} \Gamma\Big(\frac{j+1}{2}\Big) n^{\frac{j+1}{2}} + \mathcal{O}(1), \quad \text{for integers } j \geq 0, \end{aligned} \tag{4.29}$$

and

$$\begin{aligned}\sum_{\ell \geq 1} \frac{1}{\ell} e^{-\frac{\ell^2}{n}} &= \frac{1}{2} \int_{1/n}^{n} \frac{e^{-t}}{t} dt \\ &\sim \frac{1}{2} E_1\Big(\frac{1}{n}\Big), \quad \text{as } n \to \infty, \\ &= \frac{1}{2} \log n + \mathcal{O}(1), \end{aligned} \tag{4.30}$$

where E_1 is the exponential integral. Using (4.29) and (4.30) we obtain from (4.28), after collecting all contributions, the following asymptotic result for the expectation $\mathbb{E}\{L_n\}$ stated in Theorem 4.3:

$$\mathbb{E}\{L_n\} = n - \sqrt{\pi n} + \mathcal{O}(\log n).$$

4.3.5 Distance between the last two hirings

For the distance Δ_n between the last two hirings we can use a similar reasoning like we did in Subsection 4.3.4 for the index L_n of the last hired candidate. Considering the transition probabilities in the automaton given in Figure 4.2 and taking into account the "times" of the last two hirings yields then the desired description of the probabilities $\mathbb{P}\{\Delta_n = d\}$.

Assume that at time m the size of the hiring set is odd which happens with probability $a^{[1]}_{m,\ell}$, then the last two hirings in a sequence of n candidates have occurred at positions $m+1$ with probability $\frac{\ell}{m+1}$ (notice that now the size of the hiring set becomes even), and $m+d+1$ with probability $\frac{\ell+1}{m+d+1}$ (where we moved from odd to even size of the hiring set, then the rank ℓ has increased). In between those two positions all candidates got discarded, with probability $\prod_{j=m+1}^{m+d-1}\left(1-\frac{\ell+1}{j+1}\right)$, also after $m+d+1$ no more hirings until the last candidate, with probability $\prod_{j=m+d+1}^{n-1}\left(1-\frac{\ell+1}{j+1}\right)$; thus we have $\Delta_n = d$ (see Figure 4.3). Since the decisions of hiring and discarding candidates are

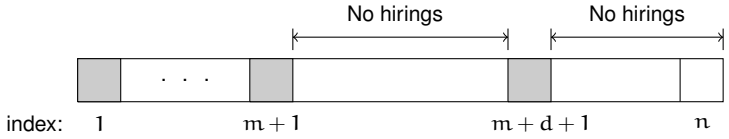


Figure 4.3: A plot shows the "times" of the last two consecutive hirings.

independent, then we can multiply all previously mentioned probabilities and summing over all positions from 1 to $n-d-1$ gives $\mathbb{P}\{\Delta_n = d\}$ in case of odd size of the hiring set.

Similar derivation follows for even size of the hiring set. It is left to consider the case when only the first candidate and the $(d+1)$-th coming candidate are hired, then we add the three quantities to obtain $\mathbb{P}\{\Delta_n = d\}$, for $1 \le d \le n-1$ and $n \ge 2$, with $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ given in Subsection 4.3.2, as follows:

$$\begin{aligned}\mathbb{P}\{\Delta_n = d\} = &\sum_{m=1}^{n-d-1}\sum_{\ell=1}^{m} a^{[1]}_{m,\ell}\cdot\frac{\ell}{m+1}\prod_{j=m+1}^{m+d-1}\left(1-\frac{\ell+1}{j+1}\right)\cdot\frac{\ell+1}{m+d+1}\prod_{j=m+d+1}^{n-1}\left(1-\frac{\ell+1}{j+1}\right)\\ &+\sum_{m=1}^{n-d-1}\sum_{\ell=1}^{m} a^{[2]}_{m,\ell}\cdot\frac{\ell}{m+1}\prod_{j=m+1}^{m+d-1}\left(1-\frac{\ell}{j+1}\right)\cdot\frac{\ell}{m+d+1}\prod_{j=m+d+1}^{n-1}\left(1-\frac{\ell+1}{j+1}\right)\\ &+a^{[1]}_{1,1}\cdot\prod_{j=1}^{d-1}\left(1-\frac{1}{j+1}\right)\cdot\frac{1}{d+1}\prod_{j=d+1}^{n-1}\left(1-\frac{2}{j+1}\right).\end{aligned} \tag{4.31}$$

Furthermore, we define $\Delta_n = 0$ if only one candidate is getting hired, which occurs if and only if the sequence starts with the maximum score and this, of course, happens with probability $\frac{1}{n}$ (this also covers the case when $n = 1$). Plugging the results (4.11) and (4.12) into (4.31) yields, after some simplifications, to the result stated in Theorem 4.4.

We will here only sketch very briefly the somewhat lengthy computations characterizing the limiting distribution of Δ_n. First, we remark that for asymptotic considerations it is advantageous to start with the following formula for the probabilities (valid for $1 \le d \le n-2$ and $n \ge 3$), which

can be deduced from Theorem 4.4 by applying basic combinatorial identities:

$$\mathbb{P}\{\Delta_n = d\} = \sum_{\ell=2}^{n-d-1} \frac{\binom{n-d-\ell}{\ell-1}}{\binom{n}{\ell+1}} \cdot \frac{\ell}{\ell+1} + \sum_{\ell=2}^{n-d-2} \frac{\binom{n-d-\ell-1}{\ell-1}}{\binom{n}{\ell+1}} \cdot \frac{\ell}{\ell-1} + R(n,d), \tag{4.32}$$

with

$$R(n,d) = \frac{2}{n(n-1)}\big(H_{n-2} - H_{d-1}\big) + \frac{1}{n(n-1)} - \sum_{m=3}^{n-d-1}\sum_{\ell=2}^{m-1} \frac{\ell d}{(m+d-\ell)(m-\ell)} \cdot \frac{\binom{m-\ell}{\ell-1}}{\binom{n}{\ell-1}}.$$

It turns out that, for $d = \mathcal{O}(n^{\frac{1}{2}+\epsilon})$, $R(n,d)$ is asymptotically negligible, i.e., $R(n,d) = \mathcal{O}\Big(\frac{d}{n^{\frac{3}{2}}}\Big)$, whereas the two sums of (4.32) contain the main contribution yielding the asymptotic equivalent

$$\mathbb{P}\{\Delta_n = d\} \sim \frac{2}{\sqrt{n}} \int_0^\infty t^2 e^{-t(x+t)} dt, \quad \text{with } d = x\sqrt{n} \text{ and } x > 0. \tag{4.33}$$

Thus, (4.33) characterizes the limiting distribution of $\frac{\Delta_n}{\sqrt{n}}$, which is the same one occurring in Theorem 4.3 for the parameter L_n.

4.3.6 Size of the hiring set conditioned on the score of first candidate

We study here the r.v. $h_{n,q}$ measuring the number of hired candidates conditioned on the event that the score U_n of the first candidate in the sequence of n candidates is q. To do this we introduce the numbers

$$a_{n,k,q} = \mathbb{P}\{h_n = k \text{ and } U_n = q\}, \tag{4.34}$$

which give the joint probability that the first candidate has score q and k candidates are hired in total. We mention that it is not difficult to extend the approach presented in Subsections 4.3.1-4.3.2 for computing the probabilities

$$a_{n,k} = \mathbb{P}\{h_n = k\},$$

by adapting the automaton of Figure 4.2 to obtain recurrences for the quantities

$$a^{[1]}_{n,\ell,q} = a_{n,2\ell-1,q} \quad \text{and} \quad a^{[2]}_{n,\ell,q} = a_{n,2\ell-2,q}.$$

However, it turns out that the task of determining the probabilities $a_{n,k,q}$ can be reduced by elementary considerations to the unconditioned probabilities $a_{n,k}$ computed already in Subsection 4.3.2, yielding thus the required results.

We do this in a two-step procedure. First we show that the probabilities $a_{n,k,1}$ (the first candidate is the one with smallest score) fully determines $a_{n,k,q}$, for general $q \leq n$. To do this, we introduce $\mathfrak{I}_{n,k,q}$, which denotes the set of n-permutations, where the first element is q and k elements are hired. It is immediate from the definition of the hiring strategy that, if the sequence starts with score q, then none of the lower $q-1$ scores in the sequence can get hired. Thus if we eliminate all these $q-1$ candidates from the original sequence and apply the hiring strategy to this "reduced" sequence the number of recruited candidates will be the same. In particular, if we eliminate in any permutation $\pi \in \mathfrak{I}_{n,k,q}$ all elements less than q we get, after relabelling $q, q+1, \ldots, n$ by $1, 2, \ldots, n-q+1$, a permutation $\pi' \in \mathfrak{I}_{n-q+1,k,1}$.

If we take into account the number of possibilities of eliminating subpermutations from the set $\mathfrak{I}_{n,k,q}$ to get the set $\mathfrak{I}_{n-q+1,k,1}$, then we can state the following useful relation:

$$|\mathfrak{I}_{n,k,q}| = |\mathfrak{I}_{n-q+1,k,1}| \cdot \binom{n-1}{q-1} \cdot (q-1)!. \tag{4.35}$$

Since

$$a_{n,k,q} = \frac{|\mathfrak{I}_{n,k,q}|}{n!},$$

then (4.35) yields

$$a_{n,k,q} = \frac{|\mathfrak{I}_{n-q+1,k,1}| \cdot \binom{n-1}{q-1} \cdot (q-1)!}{n!} = \frac{n-q+1}{n} \cdot a_{n-q+1,k,1}. \tag{4.36}$$

Second we show that the sequences $a_{n,k,1}$ and $a_{n,k}$ determine each other. We start with the obvious fact:

$$a_{n,k} = \sum_{q=1}^{n} a_{n,k,q},$$

which, by plugging (4.36) into it, yields

$$a_{n,k} = \frac{1}{n} \sum_{q=1}^{n} (n-q+1) a_{n-q+1,k,1} = \frac{1}{n} \sum_{q=1}^{n} q \cdot a_{q,k,1}. \tag{4.37}$$

Multiplying (4.37) by n and taking differences gives then

$$a_{n,k,1} = \frac{n a_{n,k} - (n-1) a_{n-1,k}}{n}. \tag{4.38}$$

By combining (4.36) and (4.38), we can link $a_{n,k,q}$ with $a_{n,k}$:

$$a_{n,k,q} = \frac{1}{n}\big((n-q+1) a_{n-q+1,k} - (n-q) a_{n-q,k}\big). \tag{4.39}$$

Of course, the distribution of $h_{n,q}$ can be obtained from (4.39) as follows:

$$\mathbb{P}\{h_{n,q} = k\} = \frac{\mathbb{P}\{h_n = k \text{ and } U_n = q\}}{\mathbb{P}\{U_n = q\}} = n a_{n,k,q} = (n-q+1) a_{n-q+1,k} - (n-q) a_{n-q,k}, \tag{4.40}$$

which, after plugging the exact formulas for $a_{n,k} = \mathbb{P}\{h_n = k\}$ obtained in Theorem 4.1 into (4.40), lead to the exact results stated in Theorem 4.5.

Due to the explicit nature of the exact results for the distribution of $h_{n,q}$, the limiting behaviour can be deduced from them quite easily: an application of Stirling's formula gives

$$\mathbb{P}\{h_{n,q} = k\} = \frac{k^3}{8(n-q)^2} e^{-\frac{k^2}{4(n-q)}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{k}\Big) + \mathcal{O}\Big(\frac{k}{n-q}\Big) + \mathcal{O}\Big(\frac{k^3}{(n-q)^2}\Big)\right), \tag{4.41}$$

uniformly for $k = \mathcal{O}\big((n-q)^{\frac{1}{2}+\epsilon}\big)$, whereas the probabilities are exponentially small for k larger. This characterizes the limiting distribution of $h_{n,q}$ as stated in Theorem 4.5, by considering the r.v. $\frac{h_{n,q}}{\sqrt{n-q}}$, and thus setting $x = \frac{k}{\sqrt{n-q}}$ in (4.41).

4.3.7 Score of last hired candidate

To show results for the score R_n of the last recruited candidate after n interviews, we study the r.v. $\tilde{R}_n = n + 1 - R_n$ (that is the rank of last hired candidate according to the ranking scheme: rank 1 is better than rank n). By considering the automaton given in Figure 4.2 and distinguishing according to the index of the last hired candidate, we obtain the following expression for the probability that the last recruited candidate has the r-th largest rank, with $1 \le r \le n$, amongst all n interviewed candidates:

$$\mathbb{P}\{\tilde{R}_n = r\} = \sum_{m=1}^{n-1} \sum_{\ell=r}^{m} a^{[1]}_{m,\ell} \cdot \frac{1}{m+1} \prod_{j=m+1}^{n-1} \left(1 - \frac{\ell+1}{j+1}\right) \tag{4.42}$$
$$+ \sum_{m=1}^{n-1} \sum_{\ell=r}^{m} a^{[2]}_{m,\ell} \cdot \frac{1}{m+1} \prod_{j=m+1}^{n-1} \left(1 - \frac{\ell}{j+1}\right) + a^{[1]}_{1,1} \cdot \prod_{j=1}^{n-1} \left(1 - \frac{1}{j+1}\right) \cdot [\![r = 1]\!].$$

Plugging the exact formulas (4.11) and (4.12) into (4.42) we get, after some simplifications by changing the order of summations and applying identity (4.27), the following result for the exact distribution of $\tilde{R}_n$:

$$\mathbb{P}\{\tilde{R}_n = r\} = \sum_{\ell=r}^{n-1} \frac{1}{\ell+1} \cdot \frac{\binom{n-\ell}{\ell}}{\binom{n}{\ell+1}} + \sum_{\ell=r}^{n-1} \frac{1}{\ell} \cdot \frac{\binom{n-\ell}{\ell-1}}{\binom{n}{\ell}}. \tag{4.43}$$

Of course, (4.43) also characterizes the distribution of $R_n = n + 1 - \tilde{R}_n$ as stated in Theorem 4.6.

The limiting distribution result given in Theorem 4.6 follows easily from (4.43) by applying Stirling's formula, which yields

$$\mathbb{P}\{n + 1 - R_n = r\} = \mathbb{P}\{\tilde{R}_n = r\} = \sum_{\ell=r}^{n^{\frac{1}{2}+\epsilon}} \frac{2}{n} e^{-\frac{\ell^2}{n}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell}\Big) + \mathcal{O}\Big(\frac{\ell}{n}\Big) + \mathcal{O}\Big(\frac{\ell^3}{n^2}\Big)\right).$$

Namely, setting $r = x\sqrt{n}$, shows that $\frac{n-R_n}{\sqrt{n}}$ converges in distribution to a r.v. Y, which has the density function

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

We further mention that the s-th integer moments of Y are given as follows:

$$\begin{aligned}\mathbb{E}(Y^s) &= \int_0^\infty x^s f_R(x) dx \\ &= 2\int_0^\infty e^{-t^2} \frac{t^{s+1}}{s+1} dt \\ &= \frac{2}{s+1} \Gamma\Big(\frac{s}{2} + 1\Big).\end{aligned}$$

4.3.8 Score of best discarded candidate

A direct recursive study of the r.v. M_n, which measures the score of the best discarded candidate after n interviews, 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

$$\hat{a}^{[1]}_{n,\ell,q} \quad \text{and} \quad \hat{a}^{[2]}_{n,\ell,q}, \quad \text{for } 0 \le q \le n-\ell, \tag{4.44}$$

which give the probabilities that, for n interviewed candidates, the threshold candidate 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 score r is simply given by the difference between the probability that all candidates with a score higher than r are recruited and the probability that all candidates with a score higher than $r-1$ are recruited, which yields the following relation eventually characterizing the distribution of M_n:

$$\mathbb{P}\{M_n = r\} = \sum_{\ell=1}^{n-r} \left(\hat{a}^{[1]}_{n,\ell,n-\ell-r} - \hat{a}^{[1]}_{n,\ell,n-\ell-r+1}\right) + \sum_{\ell=1}^{n-r} \left(\hat{a}^{[2]}_{n,\ell,n-\ell-r} - \hat{a}^{[2]}_{n,\ell,n-\ell-r+1}\right). \tag{4.45}$$

Of course, $\hat{a}^{[\cdot]}_{n,\ell,0} = a^{[\cdot]}_{n,\ell}$, where the latter numbers are defined in Subsection 4.3.1 and given by (4.11) and (4.12), since the ℓ highest ranked candidates are always hired if the threshold candidate has the ℓ-th largest score.
By an extension of the automaton for the transition probabilities as given in Figure 4.2 one gets that the quantities $\hat{a}^{[\cdot]}_{n,\ell,q}$ satisfy, for $n \ge 2$, $1 \le \ell \le n$ and $1 \le q \le n-\ell$, the following system of recurrences:

$$\hat{a}^{[1]}_{n,\ell,q} = \frac{\ell}{n} \cdot \hat{a}^{[2]}_{n-1,\ell,q-1} + \left(1 - \frac{\ell+q}{n}\right) \cdot \hat{a}^{[1]}_{n-1,\ell,q}, \tag{4.46a}$$

$$\hat{a}^{[2]}_{n,\ell,q} = \frac{\ell-1}{n} \cdot \hat{a}^{[1]}_{n-1,\ell-1,q} + \left(1 - \frac{\ell+q}{n}\right) \cdot \hat{a}^{[2]}_{n-1,\ell,q}. \tag{4.46b}$$

To reduce the system of PDEs for the corresponding generating functions we consider the normalized quantities

$$\hat{b}^{[\cdot]}_{n,\ell,q} = \frac{n!}{\ell!(n-q-\ell)!} \cdot \hat{a}^{[\cdot]}_{n,\ell,q}, \tag{4.47}$$

which yields, from (4.46a)-(4.46b), the following system of recurrences:

$$\hat{b}^{[1]}_{n,\ell,q} = \ell \cdot \hat{b}^{[2]}_{n-1,\ell,q-1} + \hat{b}^{[1]}_{n-1,\ell,q}, \tag{4.48a}$$

$$\ell \cdot \hat{b}^{[2]}_{n,\ell,q} = (\ell-1) \cdot \hat{b}^{[1]}_{n-1,\ell-1,q} + \ell \cdot \hat{b}^{[2]}_{n-1,\ell,q}. \tag{4.48b}$$

To solve this system of recurrences (4.48a)-(4.48b) we introduce the generating functions

$$\hat{B}^{[\cdot]}(z,u,v) = \sum_{n\ge 1} \sum_{1\le \ell\le n} \sum_{1\le q\le n-\ell} \hat{b}^{[\cdot]}_{n,\ell,q} \cdot z^n u^\ell v^q. \tag{4.49}$$

Due to the second recurrence (4.48b) we can express

$$\frac{\partial}{\partial u}\hat{B}^{[2]}(z,u,v) = \frac{zu}{1-z} \cdot \frac{\partial}{\partial u}\hat{B}^{[1]}(z,u,v)$$

and obtain eventually from (4.48a) the following PDE:

$$\frac{\partial}{\partial z}\hat{B}^{[1]}(z,u,v) - \frac{(1-z)^2}{z^2u^2v}\hat{B}^{[1]}(z,u,v) + \frac{z(1-z)}{(1-z-z^2u)^2} = 0. \tag{4.50}$$

It is not difficult to give an integral representation of the general solution of (4.50); however, it seems quite involved to extract coefficients from the particular solution of (4.50), which satisfies the initial condition $\widehat{B}^{[1]}_u(z,0,v) = 0$. Thus we choose another way to overcome this problem.
We introduce the functions

$$\overline{b}^{[1]}_\ell(z,v) = [u^\ell]\widehat{B}^{[1]}(z,u,v) = \sum_{n\geq \ell}\sum_{1\leq q\leq n-\ell} b^{[1]}_{n,\ell,q}\cdot z^n v^q \tag{4.51}$$

and extract coefficients $[u^\ell]$ from (4.50); this gives the following recurrence:

$$\overline{b}^{[1]}_\ell(z,v) = \frac{z^2 v}{(1-z)^2}\cdot(\ell-1)\cdot\overline{b}^{[1]}_{\ell-1}(z,v) + \frac{z^3 v}{(1-z)^3}\cdot(\ell-1)\cdot\left(\frac{z^2}{1-z}\right)^{\ell-2}. \tag{4.52}$$

The form of (4.52) suggests to introduce a further normalization via

$$\vartheta_\ell = \vartheta_\ell(z,v) = \frac{\overline{b}^{[1]}_\ell}{(\ell-1)!\left(\frac{z^2 v}{(1-z)^2}\right)^\ell}. \tag{4.53}$$

Recurrence (4.52) can then be written as follows:

$$\vartheta_\ell = \vartheta_{\ell-1} + \frac{1}{z(\ell-2)!}\left(\frac{1-z}{v}\right)^{\ell-1}, \tag{4.54}$$

and (4.54) can be solved easily by iterating it. Taking into account the initial value $\vartheta_0 = 0$ we get

$$\vartheta_\ell = \frac{1}{z}\sum_{j=1}^{\ell}\frac{1}{(j-2)!}\left(\frac{1-z}{v}\right)^{j-1}, \quad \ell\geq 0. \tag{4.55}$$

Combining (4.53) and (4.55) leads to the following formula for the functions $\overline{b}^{[1]}_\ell(z,v)$:

$$\overline{b}^{[1]}_\ell(z,v) = z^{2\ell-1}\sum_{j=2}^{\ell}\frac{v^{\ell-j+1}}{(1-z)^{2\ell-j+1}}\cdot\frac{(\ell-1)!}{(j-2)!}. \tag{4.56}$$

Extracting coefficients from (4.56) and taking into consideration (4.51) yields then the following exact formula for the numbers $\widehat{b}^{[1]}_{n,\ell,q}$:

$$\begin{aligned}\widehat{b}^{[1]}_{n,\ell,q} &= [z^n v^q]\overline{b}^{[1]}_\ell(z,v)\\ &= [z^n]z^{2\ell-1}\cdot\frac{(\ell-1)!}{(\ell-q-1)!}\cdot\frac{1}{(1-z)^{\ell+q}}\\ &= \frac{(\ell-1)!}{(\ell-q-1)!}\binom{n-\ell+q}{\ell+q-1},\end{aligned}$$

which, according to (4.47), leads to the following formula for $\widehat{a}^{[1]}_{n,\ell,q}$:

$$\widehat{a}^{[1]}_{n,\ell,q} = \frac{\binom{\ell-1}{q}\binom{n-\ell+q}{\ell+q-1}}{\binom{n}{q}\binom{n-q}{\ell}}, \quad 0\leq q\leq n-\ell. \tag{4.57}$$

An exact formula for $\hat{a}^{[2]}_{n,\ell,q}$ can be obtained from (4.57) via recurrence (4.46a):

$$\begin{aligned}\hat{a}^{[2]}_{n,\ell,q} &= \frac{n+1}{\ell}\left(\hat{a}^{[1]}_{n+1,\ell,q+1} - \left(1-\frac{\ell+q+1}{n+1}\right)\cdot \hat{a}^{[1]}_{n,\ell,q+1}\right)\\ &= \frac{\binom{\ell-1}{q+1}\binom{n-\ell+q+1}{\ell+q-1}}{\binom{n}{q+1}\binom{n-q-1}{\ell-1}}.\end{aligned} \tag{4.58}$$

Plugging (4.57) and (4.58) into (4.45) yields, after some manipulations, the exact formula for the probabilities $\mathbb{P}\{M_n = r\}$ stated in Theorem 4.7.

For the derivation of the asymptotic behaviour of M_n we consider the r.v. $\tilde{M}_n = n - M_n$. With the exact formula for the distribution of M_n given in Theorem 4.7 we obtain, for $1 \le r \le n-1$:

$$\begin{aligned}\mathbb{P}\{\tilde{M}_n = r\} = &\sum_{\ell=0}^{r-1}\left(1+\frac{(2\ell-r+1)(n+2\ell-r+1)}{r(n-r)}\right)\frac{\binom{r-\ell-1}{\ell}\binom{n+2\ell-r}{r-1}}{\binom{r}{\ell}\binom{n}{r}}\\ &+\sum_{\ell=0}^{r-1}\left(1+\frac{(2\ell-r+2)(n+2\ell-r+2)}{r(n-r)}\right)\frac{\binom{r-\ell-1}{\ell+1}\binom{n+2\ell-r+1}{r-1}}{\binom{r}{\ell+1}\binom{n}{r}}.\end{aligned}$$

For $r = 0$ (i.e., hiring everybody in the sequence): $\mathbb{P}\{M_n = 0\} = \mathbb{P}\{h_n = n\}$. An application of Stirling's formula gives

$$\mathbb{P}\{\tilde{M}_n = r\} = \sum_{\ell=0}^{r^{\frac{1}{2}+\epsilon}} \frac{4\ell}{n}\, e^{-\frac{\ell^2}{r}-\frac{r^2}{n}}\cdot\left(1+\mathcal{O}\Big(\frac{1}{\ell}\Big)+\mathcal{O}\Big(\frac{\ell}{r}\Big)+\mathcal{O}\Big(\frac{\ell r}{n}\Big)+\mathcal{O}\Big(\frac{r^3}{n^2}\Big)\right), \tag{4.59}$$

for $r = \mathcal{O}\big(n^{\frac{1}{2}+\epsilon}\big)$. Considering the sum occurring in (4.59) as a Riemann sum of the corresponding integral and setting $r = x\sqrt{n}$ we get that

$$\sqrt{n}\,\mathbb{P}\{\tilde{M}_n = x\} \sim 2xe^{-x^2}, \quad \text{for } x > 0, \tag{4.60}$$

and thus that $\frac{n-M_n}{\sqrt{n}} = \frac{\tilde{M}_n}{\sqrt{n}}$ converges in distribution to a Rayleigh distributed r.v. $\hat{M}$ with parameter $\sigma = \frac{1}{\sqrt{2}}$ as stated in Theorem 4.7.

4.3.9 Probability that a candidate with score q is getting hired

To compute the probability $p_{n,q}$ that the candidate with score q is getting hired in a sequence of n candidates, we introduce the numbers $\tilde{a}^{[1]}_{n,\ell,q}$ and $\tilde{a}^{[2]}_{n,\ell,q}$, which give the the probabilities that, for n interviewed candidates, the threshold candidate 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.

Note further that each candidate in the sequence with a score q larger or equal to the threshold candidate is getting hired anyway; thus in order to fully describe $p_{n,q}$ we further require the numbers $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ as defined in Subsection 4.3.1, since they give the probabilities that the threshold

candidate has the ℓ-th largest score in the sequence. It is immediate to see that the probabilities $p_{n,q}$ can then be obtained from these numbers via the following relation:

$$p_{n,q} = \sum_{\ell=1}^{n-q} \left(\tilde{a}^{[1]}_{n,\ell,n-\ell-q+1} + \tilde{a}^{[2]}_{n,\ell,n-\ell-q+1}\right) + \sum_{\ell=n-q+1}^{n} \left(a^{[1]}_{n,\ell} + a^{[2]}_{n,\ell}\right). \tag{4.61}$$

But the probabilities $\tilde{a}^{[1]}_{n,\ell,q}$ and $\tilde{a}^{[2]}_{n,\ell,q}$ satisfy, for $n \geq 2$, $1 \leq \ell \leq n$ and $1 \leq q \leq n-\ell$, the following system of recurrences, which can be obtained by an extension of the automaton for the transition probabilities given in Figure 4.2:

$$\tilde{a}^{[1]}_{n,\ell,q} = \frac{\ell}{n} \cdot \tilde{a}^{[2]}_{n-1,\ell,q-1} + \frac{q-1}{n} \cdot \tilde{a}^{[1]}_{n-1,\ell,q-1} + \left(1 - \frac{\ell+q}{n}\right) \cdot \tilde{a}^{[1]}_{n-1,\ell,q}, \tag{4.62a}$$

$$\tilde{a}^{[2]}_{n,\ell,q} = \frac{\ell-1}{n} \cdot \tilde{a}^{[1]}_{n-1,\ell-1,q} + \frac{q-1}{n} \cdot \tilde{a}^{[2]}_{n-1,\ell,q-1} + \left(1 - \frac{\ell+q}{n}\right) \cdot \tilde{a}^{[2]}_{n-1,\ell,q}, \tag{4.62b}$$

with initial values $\tilde{a}^{[1]}_{n,\ell,0} = \tilde{a}^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell,0} = a^{[2]}_{n,\ell}$.
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. Thus we introduce the numbers

$$b^{[\cdot]}_{n,\ell,q} = \frac{n!}{(\ell-1)!(q-1)!(n-\ell-q)!} \cdot \tilde{a}^{[\cdot]}_{n,\ell,q}, \tag{4.63}$$

leading, for $n \geq 2$, $1 \leq \ell \leq n$ and $1 \leq q \leq n-\ell$, to the recurrences

$$b^{[1]}_{n,\ell,q} = \frac{\ell}{q-1} \cdot b^{[2]}_{n-1,\ell,q-1} + b^{[1]}_{n-1,\ell,q-1} + b^{[1]}_{n-1,\ell,q}, \tag{4.64a}$$

$$b^{[2]}_{n,\ell,q} = b^{[1]}_{n-1,\ell-1,q} + b^{[2]}_{n-1,\ell,q-1} + b^{[2]}_{n-1,\ell,q}, \tag{4.64b}$$

with $b^{[\cdot]}_{n,\ell,q} = 0$, if $q = 0$ or $\ell = 0$. To treat the system of recurrences (4.64a)-(4.64b) we introduce the trivariate generating functions

$$B^{[\cdot]}(z,u,v) = \sum_{n\geq 1} \sum_{1\leq \ell \leq n} \sum_{1 \leq q \leq n-\ell} b^{[\cdot]}_{n,\ell,q} \cdot z^n u^\ell v^q. \tag{4.65}$$

From (4.64b) we can express $B^{[2]}(z,u,v) = \frac{zu}{1-z-zv} \cdot B^{[1]}(z,u,v)$ and eventually obtain from (4.64a) the PDE

$$v(1-z-zv)\frac{\partial}{\partial v}B^{[1]}(z,u,v) - \frac{z^2u^2v}{1-z-zv}\frac{\partial}{\partial u}B^{[1]}(z,u,v) - \left(1 - z + \frac{z^2uv}{1-z-zv}\right)B^{[1]}(z,u,v) = 0. \tag{4.66}$$

However, it seems to be rather involved to adapt the general solution of (4.66) to the boundary values, i.e., to find the proper solution to our problem; thus we proceed as follows.
We introduce

$$\hat{B}(z,u,w) = B^{[1]}\left(z, u, \frac{(1-z)w}{1+zw}\right), \tag{4.67}$$

so that the PDE (4.66) translates into the following one:

$$(1-z)w\frac{\partial}{\partial w}\hat{B}(z,u,w) - z^2u^2w\frac{\partial}{\partial u}\hat{B}(z,u,w) - (1-z+z^2uw)\hat{B}(z,u,w) = 0. \tag{4.68}$$

We introduce now the functions

$$\hat{b}_\ell(w) = \hat{b}_\ell(w,z) = [u^\ell]\hat{B}(z,u,w) \tag{4.69}$$

and extract coefficients $[u^\ell]$ from (4.68). This leads to the recurrence

$$w\frac{\partial}{\partial w}\hat{b}_\ell(w) - \hat{b}_\ell(w) = \frac{\ell z^2 w}{1-z}\hat{b}_{\ell-1}(w). \tag{4.70}$$

To solve (4.70) we introduce the numbers

$$\overline{b}_\ell(w) = \frac{\hat{b}_\ell(w)}{\ell!\left(\frac{z^2}{1-z}\right)^\ell}, \tag{4.71}$$

and the corresponding generating function

$$\overline{B}(z,u,w) = \sum_{\ell\geq 1}\overline{b}_\ell(w)u^\ell. \tag{4.72}$$

We get then from (4.70) the PDE

$$w\frac{\partial}{\partial w}\overline{B}(z,u,w) - (1+uw)\overline{B}(z,u,w) = 0, \tag{4.73}$$

whose general solution is given by

$$\overline{B}(z,u,w) = w\cdot C(z,u)\cdot e^{uw}, \tag{4.74}$$

with an arbitrary function $C(z,u)$. To characterize the unknown function $C(z,u)$ we have to compute $\frac{\partial}{\partial w}\hat{B}(z,u,0)$, since we will use (4.74) yielding

$$\frac{\partial}{\partial w}\overline{B}(z,u,w)\Big|_{w=0} = C(z,u). \tag{4.75}$$

According to (4.65), (4.67), (4.69), (4.71) and (4.72) we have

$$\overline{B}(z,u,w) = \sum_{n\geq 1}\sum_{1\leq\ell\leq n}\sum_{1\leq q\leq n-\ell}\frac{b^{[1]}_{n,\ell,q}}{\ell!}\cdot z^n\cdot\left(\frac{u(1-z)}{z^2}\right)^\ell\left(\frac{(1-z)w}{1+zw}\right)^q,$$

which gives

$$\frac{\partial}{\partial w}\overline{B}(z,u,w)\Big|_{w=0} = (1-z)\sum_{n\geq 1}\sum_{1\leq\ell\leq n}\frac{b^{[1]}_{n,\ell,1}}{\ell!}\cdot z^n\cdot\left(\frac{u(1-z)}{z^2}\right)^\ell, \tag{4.76}$$

where only the term of $q=1$ survives. But the numbers $\tilde{a}^{[1]}_{n,\ell,1}$ defined in the beginning of this subsection coincide with the numbers $\hat{a}^{[1]}_{n,\ell,1}$ defined in Subsection 4.3.8 during the computations

of the parameter M_n (as it is obvious from the corresponding definitions). Thus, using the explicit formula for $\hat{a}^{[1]}_{n,\ell,1}$ as given in (4.57) and taking into account (4.63), gives an explicit formula for $b^{[1]}_{n,\ell,1}$:

$$b^{[1]}_{n,\ell,1} = \ell(\ell-1)\binom{n-\ell+1}{\ell}. \tag{4.77}$$

Plugging (4.77) into (4.76) we obtain then

$$\frac{\partial}{\partial w}\overline{B}(z,u,w)\Big|_{w=0} = \sum_{n\geq 1}\sum_{1\leq \ell\leq n}\frac{1}{(\ell-2)!}\binom{n-\ell+1}{\ell}z^{n-2\ell}(1-z)^{\ell+1}u^{\ell}. \tag{4.78}$$

According to (4.75), equation (4.78) also characterizes the unknown functions $C(z,u)$ occurring in (4.74) and we obtain the following solution for $\overline{B}(z,u,w)$:

$$\overline{B}(z,u,w) = \left(\sum_{\ell\geq 2}\frac{1}{(\ell-2)!}(1-z)^{\ell+1}u^{\ell}\sum_{n\geq \ell}\binom{n-\ell+1}{\ell}z^{n-2\ell}\right)we^{uw}. \tag{4.79}$$

When going through all the transformations (4.65), (4.67), (4.69), (4.71) and (4.72) and simplifying the explicit formula (4.79), it is not difficult to see that

$$\begin{aligned}
\frac{b^{[1]}_{n,\ell,q}}{\ell!} &= [z^n u^\ell v^q]\overline{B}\left(z, \frac{uz^2}{1-z}, \frac{v}{1-z-zv}\right)\\
&= [z^n u^\ell v^q]\frac{z^3u^2v}{(1-z)^2(1-z-zv)}\cdot e^{\frac{uz^2(1+v)}{1-z-zv}}\\
&= \left[z^{n-3}v^{q-1}\right]\frac{1}{(1-z)^2(1-z-zv)(\ell-2)!}\cdot\left(\frac{z^2(1+v)}{1-z-zv}\right)^{\ell-2}\\
&= \sum_{k=0}^{\ell-2}\binom{\ell-2}{k}\left[z^{n-2\ell+1}v^{q-1-k}\right]\frac{1}{(1-z)^2(1-z-zv)^{\ell-1}}\\
&= \sum_{k=0}^{\ell-2}\binom{\ell-2}{k}\binom{q-1+k}{\ell-2}\cdot\left[z^{n-\ell-q-k}\right]\frac{1}{(1-z)^{q+k+2}}\\
&= \sum_{k=0}^{\ell-2}\binom{\ell-2}{k}\binom{q-1+k}{\ell-2}\binom{n-\ell+1}{q+k+1}.
\end{aligned} \tag{4.80}$$

Taking into account (4.63) and (4.80), we obtain the following explicit formula for the numbers $\tilde{a}^{[1]}_{n,\ell,q}$, valid for $1\leq \ell\leq n$ and $1\leq q\leq n-\ell$:

$$\tilde{a}^{[1]}_{n,\ell,q} = \frac{(\ell-1)}{n\binom{n-1}{\ell}\binom{n-\ell-1}{q-1}}\sum_{k=0}^{\ell-2}\binom{\ell-2}{k}\binom{q+k-1}{\ell-2}\binom{n-\ell+1}{q+k+1}. \tag{4.81}$$

Furthermore, by using recurrence (4.62a), we obtain from (4.81) an explicit formula for $\tilde{a}^{[2]}_{n,\ell,q}$, also valid for $1\leq \ell\leq n$ and $1\leq q\leq n-\ell$:

$$\tilde{a}^{[2]}_{n,\ell,q} = \frac{(\ell-1)}{n\binom{n-1}{\ell-1}\binom{n-\ell}{q}}\sum_{k=0}^{\ell-2}\binom{\ell-2}{k}\binom{q+k-1}{\ell-3}\binom{n-\ell+1}{q+k+1}. \tag{4.82}$$

Plugging (4.81) and (4.82) as well as (4.11) and (4.12) into (4.61) yields the exact formula for the probabilities $p_{n,q}$ as stated in Theorem 4.8.

4.3.10 Number of replacements

We study the number of replacements f_n when combining "hiring above the median" with the proposed replacement mechanism in Section 3.3.
To do this we express f_n as a sum of indicator r.v.'s $\mathcal{X}_j$, where $\mathcal{X}_j$ denotes the event that the j-th candidate in the sequence replaces the worst candidate hired so far:

$$f_n = \mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n.$$

It is not difficult to see that the probability of success of the r.v.'s $\mathcal{X}_j$ can be obtained as follows:

$$\mathbb{P}\{\mathcal{X}_j = 1\} = \sum_{\ell=1}^{j-1} \frac{\ell-1}{j} \cdot a^{[1]}_{j-1,\ell} + \sum_{\ell=1}^{j-1} \frac{\ell-2}{j} \cdot a^{[2]}_{j-1,\ell},$$

where the numbers $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ are given by (4.11) and (4.12). Hence, the expectation $\mathbb{E}\{f_n\}$ can be computed as follows:

$$\mathbb{E}\{f_n\} = \sum_{j=1}^{n} \mathbb{E}(\mathcal{X}_j) = \sum_{j=1}^{n} \sum_{\ell=1}^{j-1} \frac{\ell-1}{j} \cdot \frac{\binom{j-1-\ell}{\ell-1}}{\binom{j-1}{\ell}} + \sum_{j=1}^{n} \sum_{\ell=2}^{j-1} \frac{\ell-2}{j} \cdot \frac{\binom{j-1-\ell}{\ell-2}}{\binom{j-1}{\ell-1}}. \tag{4.83}$$

Applying Stirling's formula to (4.83) and carrying out the inner summation yields the following asymptotic expansion, which holds uniformly for $1 \le j \le n$:

$$\sum_{\ell=2}^{j-1} \left(\frac{\ell-1}{j} \cdot \frac{\binom{j-1-\ell}{\ell-1}}{\binom{j-1}{\ell}} + \frac{\ell-2}{j} \cdot \frac{\binom{j-1-\ell}{\ell-2}}{\binom{j-1}{\ell-1}} \right) = \frac{\sqrt{\pi}}{2\sqrt{j}} + \mathcal{O}\Big(\frac{1}{j}\Big). \tag{4.84}$$

Plugging (4.84) into (4.83) easily gives the asymptotic result for the expectation of f_n stated in Theorem 4.9.

4.4 Relationship with other on-line processes

In this section we consider the relationship between "hiring above the median" and two similar problems occurring in the literature. Subsection 4.4.1 explains the differences between our strategy and the "$\frac{1}{2}$-percentile rule" presented in Subsection 2.2.1. We get, using our approach, the explicit and limit distribution of the *the number of selected items* for the later process. Consequently, the distributional results for the *waiting time* are easily obtained for the $\frac{1}{2}$-percentile rule.

In Subsection 4.4.2, we similarly focus on the similarities (and differences) between the *number of hired candidates* under our strategy, and the *number of open tables* for the seating plan $(\frac{1}{2}, 1)$ of the CRP (Section 2.5). The results for the *waiting time* for this seating plan are also given.

4.4.1 The $\frac{1}{2}$-percentile rule

According to Definition 2.2, the p-percentile selection rule, with $0 < p \le 1$, selects the first candidate in the sequence, and then each new candidate is selected exactly if he has a better rank than the $\lceil pk \rceil$-th best quantile of the already selected candidates (with k denoting the number of

already selected candidates).
Let us consider now the special instance $p = \frac{1}{2}$. We can also formulate those selection rules by using the terminology of the hiring problem and assuming that better candidates have better scores. This hiring strategy, "$\frac{1}{2}$-percentile rule", reads then as follows:
" The first candidate is hired, and then a new candidate is hired if and only if his score is better than r_{k-j+1}, *with* $j = \lceil \frac{k}{2} \rceil$, *where* $r_1 < r_2 < \cdots < r_k$ *denote the (ordered) sequence of scores of the* k *already hired candidates."*

Thus, as explained before in Section 2.3, the $\frac{1}{2}$-percentile rule is very closely related to "hiring above the median"; the difference is simply that the latter uses the "lower" median as the threshold candidate, whereas the former uses the "upper" median (of course, a difference between both selection rules only appears when the size of the already recruited staff is *even*).
This can also be noticed by considering the quantity $X(\sigma)$ introduced in Section 2.4, which gives the number of choices to hire the next candidate right after σ. After k hirings, then, hiring above the median has $X_{\text{med}}(\sigma) = \lfloor \frac{k+2}{2} \rfloor = \lceil \frac{k+1}{2} \rceil$, while the $\frac{1}{2}$-percentile rule has $X_{[1/2]}(\sigma) = \lceil \frac{k}{2} \rceil$.

The $\frac{1}{2}$-percentile rule is thus *more restrictive* than "hiring above the median" and it will hold, for each sequence of candidates, that the number of selected candidates by applying this strategy is *not larger than* for hiring above the median.
In particular, when considering the r.v. $L_n^{[1/2]}$, which measures the *number of selected candidates* after n interviews when applying the $\frac{1}{2}$-percentile rule, it must hold $\mathbb{E}\left\{L_n^{[1/2]}\right\} \leq \mathbb{E}\{h_n\}$.

It turns out to be an interesting question to describe the influence of this small change (taking as threshold candidate the "upper" median instead of the "lower" median) to the behaviour of the number of selected candidates. As an addition to Theorem 2.1 by Krieger et al., we report new results concerning $L_n^{[1/2]}$. First we get $\mathbb{E}\left\{L_n^{[1/2]}\right\} \sim \frac{2}{3}\sqrt{\pi n}$, whereas $\mathbb{E}\{h_n\} \sim \sqrt{\pi n}$, thus taking the "upper" median instead of the "lower" median rules out about $\frac{1}{3}$ of the candidates on average. Second, it holds that the limiting behaviour of $L_n^{[1/2]}$ changes, i.e., the limiting distribution is no more a Rayleigh-distribution (as it is the case for h_n).

Moreover, we report the results for a new parameter for this selection rule, namely, the *waiting time*, and also complete some results given in Theorem 2.5 regarding the *average rank of the retained group* as shown next.

4.4.1.1 Results

Theorem 4.10 *For the* $\frac{1}{2}$*-percentile rule, let* $L_n^{[1/2]}$ *denote the number of selected candidates after* n *observations. Then the exact distribution of* $L_n^{[1/2]}$ *is given as follows:*

$$\mathbb{P}\left\{L_n^{[1/2]} = k\right\} = \begin{cases} \frac{\ell-1}{n}\left(H_{n-1} - H_{\ell-1}\right) + \frac{1}{n} - \sum_{j=1}^{\ell-1} \frac{1}{j} \cdot \frac{\binom{n-1-j}{\ell-2}}{\binom{n}{\ell-1}}, & \text{for } k = 2\ell - 1 \text{ odd}, \\ \frac{\ell}{n}\left(H_{n-1} - H_{\ell-1}\right) - \sum_{j=1}^{\ell-1} \frac{1}{j} \cdot \frac{\binom{n-1-j}{\ell-1}}{\binom{n}{\ell}}, & \text{for } k = 2\ell \text{ even}. \end{cases}$$

Asymptotically, as $n \to \infty$, $\frac{L_n^{[1/2]}}{\sqrt{n}} \xrightarrow{(d)} \Lambda$, *where* Λ *is characterized by its density function* $f(x)$*:*

$$f(x) = \frac{x}{2}\int_1^\infty \frac{e^{-\frac{tx^2}{4}}}{t}dt, \quad \textit{for } x > 0.$$

Furthermore, the expectation of $L_n^{[1/2]}$ *has the following asymptotic behaviour:*

$$\mathbb{E}\left\{L_n^{[1/2]}\right\} = \frac{2}{3}\sqrt{\pi n} + \mathcal{O}(\log n).$$

Thus $\frac{\mathbb{E}\left\{L_n^{[1/2]}\right\}}{\sqrt{n}} \to c_{\frac{1}{2}}$, *with* $c_{\frac{1}{2}} = \frac{2}{3}\sqrt{\pi}$.

Theorem 4.11 *For the* $\frac{1}{2}$*-percentile rule, let* $W_N^{[1/2]}$ *denote the "waiting time" (i.e., the number of candidates that have to be interviewed) until exactly* N *candidates are selected. Then asymptotically, as* $N \to \infty$, $\frac{W_N^{[1/2]}}{N^2} \xrightarrow{(d)} \hat{W}^{[1/2]}$, *where* $\hat{W}^{[1/2]}$ *has the following density function:*

$$g(x) = \frac{1}{4x^2}\cdot\int_1^\infty \frac{e^{-\frac{t}{4x}}}{t}dt, \quad \textit{for } x > 0.$$

(Note that the moments of $\hat{W}^{[1/2]}$ *do not exist.)*

Corollary 4.2 *For the* $\frac{1}{2}$*-percentile rule, let* $A_n^{[1/2]}$ *denote average rank of the retained group after* n *observations. Then we have* $\frac{\mathbb{E}\left\{A_n^{[1/2]}\right\}}{\sqrt{n}\log n} \to \frac{\sqrt{\pi}}{12}$.

4.4.1.2 Analysis

We consider the sequences $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$, which give now when applying the $\frac{1}{2}$-percentile rule, the probability that, after n interviews, the threshold candidate has *rank* ℓ and an *odd* number of candidates ($k = 2\ell - 1$) has been selected, whereas $a_{n,\ell}^{[2]}$ gives the probability that, after n interviews, the threshold candidate has *rank* ℓ and an *even* number of candidates ($k = 2\ell$) has been selected. Of course, the probabilities $\mathbb{P}\left\{L_n^{[1/2]} = k\right\}$ are then determined by

$$\mathbb{P}\left\{L_n^{[1/2]} = k\right\} = \begin{cases} a_{n,\frac{k+1}{2}}^{[1]}, & \text{for } k = 2\ell - 1 \text{ odd}, \\ a_{n,\frac{k}{2}}^{[2]}, & \text{for } k = 2\ell \text{ even}. \end{cases} \tag{4.85}$$

It is not difficult to set up the following recurrences for the numbers $a^{[1]}_{n,\ell}$ and $a^{[2]}_{n,\ell}$ (we omit here stating the "automaton" of the Markov chain and the transition probabilities since it is typical to that one in Figure 4.2 for "hiring above the median"):

$$a^{[1]}_{n,\ell} = \frac{\ell-1}{n}\cdot a^{[2]}_{n-1,\ell-1} + \left(1-\frac{\ell}{n}\right)\cdot a^{[1]}_{n-1,\ell}, \quad n \geq 2,\ 1 \leq \ell \leq n, \tag{4.86a}$$

$$a^{[2]}_{n,\ell} = \frac{\ell}{n}\cdot a^{[1]}_{n-1,\ell} + \left(1-\frac{\ell}{n}\right)\cdot a^{[2]}_{n-1,\ell}, \quad n \geq 2,\ 1 \leq \ell \leq n, \tag{4.86b}$$

with initial values $a^{[1]}_{1,1} = 1$ and $a^{[2]}_{1,1} = 0$. Outside the range $1 \leq \ell \leq n$ we define $a^{[1]}_{n,\ell} = a^{[2]}_{n,\ell} = 0$. To simplify this system of recurrences (4.86a)-(4.86b) we introduce the numbers

$$b^{[\cdot]}_{n,\ell} = \binom{n}{\ell} a^{[\cdot]}_{n,\ell}, \tag{4.87}$$

and obtain, after some manipulations eliminating $a^{[1]}_{n,\ell}$, the following recurrence for $b^{[2]}_{n,\ell}$:

$$(n-\ell)b^{[2]}_{n,\ell} = (n-\ell)b^{[2]}_{n-1,\ell} + [\![\ell = 1]\!] + \sum_{m=1}^{n-2}(\ell-1)b^{[2]}_{m,\ell-1}, \quad 1 \leq \ell \leq n-1. \tag{4.88}$$

Introducing the generating function

$$B(z,u) = \sum_{n\geq 2}\sum_{1\leq\ell\leq n-1} b^{[2]}_{n,\ell}\cdot z^n u^\ell, \tag{4.89}$$

we obtain from recurrence (4.88) the following first order linear PDE:

$$z(1-z)\frac{\partial}{\partial z}B(z,u) + \left(zu - u - \frac{u^2z^2}{1-z}\right)\frac{\partial}{\partial u}B(z,u) - zB(z,u) = \frac{z^2u}{1-z}. \tag{4.90}$$

Proceeding similar as we did for the numbers $a^{[\cdot]}_{n,\ell}$ in Subsection 4.3.2, i.e., adapting the (easily obtained) general solution of (4.90) to the initial conditions, leads to the following solution for the generating function $B(z,u)$:

$$B(z,u) = \frac{zu}{1-z-zu}\log\left(\frac{1-z-z^2u}{(1-z)^2}\right). \tag{4.91}$$

Extracting coefficients of (4.91) is quite a routine task and yields:

$$b^{[2]}_{n,\ell} = \binom{n-1}{\ell-1}(H_{n-1} - H_{\ell-1}) - \sum_{j=1}^{\ell-1}\frac{1}{j}\binom{n-j-1}{\ell-1}, \tag{4.92}$$

and due to (4.87) the following exact formula for $a^{[2]}_{n,\ell}$:

$$a^{[2]}_{n,\ell} = \frac{\ell}{n}\big(H_{n-1} - H_{\ell-1}\big) - \sum_{j=1}^{\ell-1}\frac{1}{j}\cdot\frac{\binom{n-1-j}{\ell-1}}{\binom{n}{\ell}}. \tag{4.93}$$

Plugging (4.93) into (4.86b) yields after some routine calculations also an exact formula for $a_{n,\ell}^{[1]}$:

$$\begin{aligned} a_{n,\ell}^{[1]} &= \frac{n+1}{\ell}\left(a_{n+1,\ell}^{[2]} - \frac{n+1-\ell}{n+1}\cdot a_{n,\ell}^{[2]}\right) \\ &= \frac{\ell-1}{n}\left(H_{n-1}-H_{\ell-1}\right) + \frac{1}{n} - \sum_{j=1}^{\ell-1}\frac{1}{j}\cdot\frac{\binom{n-1-j}{\ell-2}}{\binom{n}{\ell-1}}. \end{aligned} \tag{4.94}$$

Combining (4.93) and (4.94) yields, due to (4.85), the exact probability distribution of $L_n^{[1/2]}$ stated in Theorem 4.10.

To obtain the limiting distribution of $L_n^{[1/2]}$ we will use the exact formulas for $a_{n,\ell}^{[\cdot]}$ given in (4.93) and (4.94) and apply Stirling's formula; additionally we require the well-known asymptotic expansion of the harmonic numbers H_n:

$$H_n = \log n + \gamma + \mathcal{O}\Big(\frac{1}{n}\Big), \tag{4.95}$$

where γ denotes the Euler-Mascheroni constant. This yields, for $\ell = \mathcal{O}\big(n^{\frac{1}{2}+\epsilon}\big)$, the uniform asymptotic expansion

$$a_{n,\ell}^{[1]} \sim a_{n,\ell}^{[2]} = \frac{\ell}{n}\Big(-\gamma - \log\Big(\frac{\ell^2}{n}\Big) + \int_0^{\frac{\ell^2}{n}}\frac{1-e^{-t}}{t}dt\Big)\cdot\Big(1+\mathcal{O}\Big(\frac{1}{\ell}\Big)+\mathcal{O}\Big(\frac{\ell}{n}\Big)+\mathcal{O}\Big(\frac{\ell^3}{n^2}\Big)\Big). \tag{4.96}$$

After setting $\ell = x\sqrt{n}$, expansion (4.96) leads to

$$\sqrt{n}\,a_{n,\ell}^{[1]} \sim \sqrt{n}\,a_{n,\ell}^{[2]} \sim x\left(-\gamma - \log(x^2) + \int_0^{x^2}\frac{1-e^{-t}}{t}dt\right) = x\int_1^{\infty}\frac{e^{-tx^2}}{t}dt, \tag{4.97}$$

where the latter identity follows after partial integration and using the well-known integral evaluation

$$\int_0^{\infty}\log(t)\,e^{-t}dt = -\gamma. \tag{4.98}$$

The limiting distribution result stated in Theorem 4.10 follows then from (4.85) and (4.97).
An exact result for the expectation $\mathbb{E}\left\{L_n^{[1/2]}\right\}$ can be obtained by plugging (4.93) and (4.94) into

$$\mathbb{E}\left\{L_n^{[1/2]}\right\} = \sum_{\ell=1}^{n}\left((2\ell-1)\cdot a_{n,\ell}^{[1]} + 2\ell\cdot a_{n,\ell}^{[2]}\right). \tag{4.99}$$

We remark that, after applying identity (4.27) and some basic identities involving harmonic numbers, one obtains from (4.99) the following explicit formula:

$$\mathbb{E}\left\{L_n^{[1/2]}\right\} = \frac{4}{9}n^2 + \frac{7}{12}n + \frac{5}{36} - \frac{1}{6n} - \sum_{j=1}^{n-1} \frac{2j+1}{n-j} \cdot \frac{\binom{n-j}{j}}{\binom{n}{j}} \\ - \sum_{j=1}^{n-1} \frac{4j^4+13j^3+24j^2+12j+7+8j^2n+5jn+15+8n^2}{j(j+1)(j+2)(j+3)} \cdot \frac{\binom{n-1-j}{j}}{\binom{n}{j}}. \tag{4.100}$$

An asymptotic evaluation of the expression (4.100) can be carried out following the tracks used in Section 4.3, but requires (due to cancellations occurring) some care. First, by using partial fraction expansion, the asymptotic expansion (4.13) and evaluating sums asymptotically by the corresponding Riemann integral, one can easily rule out negligible terms yielding

$$\mathbb{E}\left\{L_n^{[1/2]}\right\} = \frac{4}{9}n^2 + \frac{7}{12}n - \sum_{j=1}^{n-1} \frac{8n^2}{j^{\overline{4}}} \frac{\binom{n-1-j}{j}}{\binom{n}{j}} - \sum_{j=1}^{n-1} \frac{(8j^2+5j+15)n}{j^{\overline{4}}} \frac{\binom{n-1-j}{j}}{\binom{n}{j}} \\ - 4\sum_{j=1}^{n-1} \frac{\binom{n-1-j}{j}}{\binom{n}{j}} + \mathcal{O}(\log n). \tag{4.101}$$

We comment on evaluating the first of the sums appearing in (4.101) asymptotically: basically we will applying Stirling's formula. However, to apply these asymptotic expansions one has to bring the summands into a suitable form. We do this by applying partial fraction expansion and shifting the range of summation to easily get

$$\sum_{j\geq 1} \frac{8}{j^{\overline{4}}} F(j) = \frac{4}{9} - \frac{8}{3(n+3)} + \sum_{j\geq 1} \frac{1}{j} \left(\frac{4}{3}F(j) - 4F(j-1) + 4F(j-2) - \frac{4}{3}F(j-3)\right), \tag{4.102}$$

with $F(j) = \frac{\binom{n-1-j}{j}}{\binom{n}{j}}$. This and expanding the summand obtained eventually yields

$$\sum_{j=1}^{n-1} \frac{8n^2}{j^{\overline{4}}} \frac{\binom{n-1-j}{j}}{\binom{n}{j}} = n^2\left(\frac{4}{9} - \frac{8}{3(n+3)}\right) \\ + \sum_{j\geq 1} \frac{\binom{n-j}{j}}{\binom{n}{j}} \left(\frac{8(2j-3)}{j} - \frac{8(4j^3-51j^2+116j-75)}{3jn} + \mathcal{O}\Big(\frac{j^3}{n^2}\Big)\right) \\ = \frac{4}{9}n^2 - \frac{8n}{3} + 16\sum_{j\geq 1} \frac{\binom{n-j}{j}}{\binom{n}{j}} - \frac{32}{3n}\sum_{j\geq 1} j^2 \frac{\binom{n-j}{j}}{\binom{n}{j}} + \mathcal{O}(\log n) \\ = \frac{4}{9}n^2 - \frac{8n}{3} + \frac{16}{3}\sqrt{\pi n} + \mathcal{O}(\log n). \tag{4.103}$$

The second sum of (4.101) can be treated in a similar way leading to

$$\sum_{j=1}^{n-1} \frac{(8j^2+5j+15)n}{j^{\overline{4}}} \frac{\binom{n-1-j}{j}}{\binom{n}{j}} = \frac{13}{4}n - 8\sqrt{\pi n} + \mathcal{O}(\log n), \tag{4.104}$$

whereas the asymptotic behaviour of the third sum is a routine calculation yielding

$$4\sum_{j=1}^{n-1}\frac{\binom{n-1-j}{j}}{\binom{n}{j}} = 2\sqrt{\pi n} + \mathcal{O}(1). \tag{4.105}$$

Combining (4.101), (4.103), (4.104) and (4.105) gives then the asymptotic result for $\mathbb{E}\left\{L_n^{[1/2]}\right\}$ stated in Theorem 4.10.

Waiting time. Of course, the exact results for the probabilities $\mathbb{P}\left\{L_n^{[1/2]} = k\right\}$ as given in Theorem 4.10 could be used to obtain the exact and asymptotic behaviour of further quantities of interest. We just state the result concerning the limiting distribution of the *waiting time* $W_N^{[1/2]}$ (i.e., the number of interviewed candidates required) to select N candidates, which follows easily as before (i.e., for h_n), we start with

$$\mathbb{P}\left\{W_N^{[1/2]} = t\right\} = \begin{cases} a_{t-1,\ell}^{[2]} \cdot \frac{\ell}{t}, & \text{for } N = 2\ell \text{ and } N \geq 2, \\ a_{t-1,\ell}^{[1]} \cdot \frac{\ell}{t}, & \text{for } N = 2\ell + 1 \text{ and } N \geq 3. \end{cases}$$

Then, we make use of the asymptotic expansion in (4.97) for $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$, to state the result in Theorem 4.11.

Average rank of the retained group. As a consequence of the characterization of the constant $c_{\frac{1}{2}}$ in Theorem 4.10, we can continue the work of Krieger et al. in Theorem 2.5, describing the asymptotic behaviour of the *average rank of the retained group* $A_n^{[1/2]}$ for the $\frac{1}{2}$-percentile rule. Recall from Theorem 2.5 that

$$\frac{\mathbb{E}\left\{A_n^{[1/2]}\right\}}{\sqrt{n}\log n} \to \frac{c_{\frac{1}{2}}}{8}, \quad \text{with} \quad c_{\frac{1}{2}} = \lim_{n\to\infty}\frac{\mathbb{E}\left\{L_n^{[1/2]}\right\}}{\sqrt{n}}.$$

Thus we obtain Corollary 4.2 from Theorems 4.10 and 2.5.

4.4.2 The seating plan $(\frac{1}{2}, 1)$

We have introduced the *Chinese restaurant process (CRP)* in Section 2.5, where probabilistic selection rules called seating plans (α, θ) are used to process a sequence of customers visiting the restaurant. As discussed before, the relationship between seating plans of the CRP and hiring strategies could be explored by considering the transition probabilities of two equivalent events: *opening new table* and *hiring new candidate* for seating plans and hiring strategies, respectively.

Let us consider the r.v. K_n, that denotes *number of occupied tables* after n customers have arrived in the restaurant. Assume $\theta = 1$ and $0 < \alpha < 1$ for the seating plan (α, θ); then the *number of hired candidates* h_n for the general strategy "hiring above the α-quantile" (see Definition 2.9) and K_n are both *Markov chains* with increments in $\{0, 1\}$.

Referring to Definition 2.10, we find that the seating plan $(\frac{1}{2}, 1)$ is very close to "hiring above the median", though they are not *equivalent*. This is clearly explained via the inhomogeneous

transition probabilities in the respective Markov chains as follows:

$$\mathbb{P}\left\{K_{n+1}^{(\frac{1}{2},1)} = k+1 | K_n^{(\frac{1}{2},1)} = k\right\} = \frac{\frac{k}{2}+1}{n+1} \quad \text{and} \quad \mathbb{P}\left\{K_{n+1}^{(\frac{1}{2},1)} = k | K_n^{(\frac{1}{2},1)} = k\right\} = \frac{n-\frac{k}{2}}{n+1},$$

$$\mathbb{P}\{h_{n+1} = k+1 | h_n = k\} = \frac{\lfloor \frac{k}{2}+1 \rfloor}{n+1} \quad \text{and} \quad \mathbb{P}\{h_{n+1} = k | h_n = k\} = \frac{\lceil n-\frac{k}{2} \rceil}{n+1}.$$

Thus, we obtain that, for $k = 2\ell - 2$ even, both r.v.'s have the same probability $\frac{\ell}{n+1}$ to be incremented by 1 at time $n+1$, whereas, for $k = 2\ell - 1$ odd, $K_{n+1}^{(\frac{1}{2},1)}$ has probability $\frac{\ell+\frac{1}{2}}{n+1}$, but h_{n+1} still has probability $\frac{\ell}{n+1}$ to be incremented by 1.

The behaviour of K_n for the two-parameter model has been studied precisely as given in Section 2.5. For the matter of comparison, we state here the corresponding results for $K_n^{(\frac{1}{2},1)}$, which are obtained by specializing $\theta = 1$ and $\alpha = \frac{1}{2}$ in the general formulas (2.2) and (2.3), and carrying out some simplifications.

Theorem 4.12 (Pitman) *For the seating plan $(\frac{1}{2}, 1)$, let $K_n^{(\frac{1}{2},1)}$ denote the number of occupied tables after n customers have arrived in the restaurant. Then the exact probability distribution of $K_n^{(\frac{1}{2},1)}$ is given as follows:*

$$\mathbb{P}\left\{K_n^{(\frac{1}{2},1)} = k\right\} = \frac{k(k+1)}{n\,2^{2n-k}}\binom{2n-k-1}{n-k}, \quad 1 \le k \le n.$$

Asymptotically, as $n \to \infty$, $\frac{K_n^{(\frac{1}{2},1)}}{\sqrt{n}} \xrightarrow{\text{(a.s.)}} K$, where K has the density function

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

Moreover, the expectation $\mathbb{E}\left\{K_n^{(\frac{1}{2},1)}\right\}$ is given by the following exact and asymptotic formulas:

$$\begin{aligned}\mathbb{E}\left\{K_n^{(\frac{1}{2},1)}\right\} &= \frac{2(2n+1)}{4^n}\binom{2n}{n} - 2 \\ &= \frac{4}{\sqrt{\pi}}\sqrt{n} + \mathcal{O}(1).\end{aligned}$$

It is further given in Theorem 2.23, that the limiting distribution K occurring is a "variant" of a Mittag-Leffler distribution, since K has the density function $f(x) = \frac{x^2}{2} \cdot g_{1/2}(x)$, where $g_{1/2}(x)$ is the Mittag-Leffler distribution with parameter $\frac{1}{2}$. We also observed that $K = \frac{K_n^{(\frac{1}{2},1)}}{\sqrt{n}}$ has a *Maxwell-Boltzmann* distribution with parameter $\sqrt{2}$.

Waiting time. A main difference in the behaviour of the CRP with seating plan $(\frac{1}{2}, 1)$ and the hiring process using "hiring above the median" occurs when studying the *waiting time* until N tables are occupied, i.e., the r.v. $T_N^{(\frac{1}{2},1)}$ which counts the *number of customers arrived in the restaurant until the N-th table is opened*, thus N tables are occupied for the first time. As a consequence of the exact result for the distribution of $K_n^{(\frac{1}{2},1)}$ stated in Theorem 4.12, it holds that

$$\mathbb{P}\left\{T_N^{(\frac{1}{2},1)} = t\right\} = \frac{N+1}{2t} \cdot \mathbb{P}\{K_{t-1} = N-1\}.$$

Then, it is not difficult to show the following exact and asymptotic behaviour of $T_N^{(\frac{1}{2},1)}$:

Theorem 4.13 *For the seating plan* $(\frac{1}{2},1)$*, let* $T_N^{(\frac{1}{2},1)}$ *denote the waiting time until* N *tables are occupied in the restaurant. Then the exact distribution of* $T_N^{(\frac{1}{2},1)}$ *is given as follows:*

$$\mathbb{P}\left\{T_N^{(\frac{1}{2},1)}=t\right\}=\frac{(N+1)N(N-1)}{t(t-1)\,2^{2t-N}}\binom{2t-N-2}{t-N},\quad t\geq N\geq 2,$$

and $\mathbb{P}\left\{T_1^{(\frac{1}{2},1)}=1\right\}=1$.

Asymptotically, as $N\to\infty$, $\frac{T_N^{(\frac{1}{2},1)}}{N^2}\xrightarrow{(d)}T$, *where* T *has the density function*

$$g(x)=\frac{1}{4\sqrt{\pi}x^{\frac{5}{2}}}\,e^{-\frac{1}{4x}},\quad \textit{for } x>0.$$

Moreover, the expectation of T_N *is given as follows:*

$$\mathbb{E}\left\{T_N^{(\frac{1}{2},1)}\right\}=\frac{N(N+1)}{2}.$$

4.5 Conclusions

We provided a rather detailed study of various hiring parameters related to the hiring process when applying the "hiring above the median" strategy. The analysis occurred in this chapter is based on a recursive approach, where we always took into account the rank of the threshold candidate. We have reported the results for "hiring above the median" in Section 4.2. We also gave a detailed explanation for our approach together with complete analysis in Section 4.3.
This approach proves useful again for the analysis of the "$\frac{1}{2}$-percentile rule", where we introduced new results for this selection rule in Theorems 4.10 and 4.11, and Corollary 4.2. The connections between "hiring above the median" and the seating plan $(\frac{1}{2},1)$ of the CRP has also been considered, indeed, new results for this seating plan are given in Theorem 4.13.

In the same way, it is also noticed that the seating plan $(\frac{1}{2},0)$ is very closely related to the $\frac{1}{2}$-percentile rule, where as usual we have to consider the transition probabilities of increment of the r.v.'s $K_n^{(\frac{1}{2},0)}$ and $L_n^{[1/2]}$ as follows:

$$\mathbb{P}\left\{K_{n+1}^{(\frac{1}{2},0)}=k+1|K_n^{(\frac{1}{2},0)}=k\right\}=\frac{k}{2n},$$
$$\mathbb{P}\left\{L_{n+1}^{[1/2]}=k+1|L_n^{[1/2]}=k\right\}=\frac{\lceil\frac{k}{2}\rceil}{n+1}.$$

Thus we can state the following relationship regarding the r.v. *number of selections* under these related selection rules: seating plan $(\frac{1}{2},0)$, $\frac{1}{2}$-percentile rule, hiring above the median, and seating plan $(\frac{1}{2},1)$, that is

$$K_n^{(\frac{1}{2},0)}\leq L_n^{[1/2]}\leq h_n\leq K_n^{(\frac{1}{2},1)},$$

which is obviously true since the size of the sample of selections increases (in probability) if we increase the probability of selecting elements.

As a consequence of Theorems 4.1, 4.12 and 4.10 we obtain again that a slight modification of the transition probabilities in the Markov chain yields a different limiting distribution as well as a different asymptotic behaviour of the expectation. This indicates once more that for such kind of problems a detailed analysis is required to precisely describe the asymptotic behaviour.

From another point of view, for the CRP, the expected waiting time until N tables are occupied is finite, whereas the corresponding quantity W_N for the hiring process, i.e., the number of candidates that have to be interviewed until N candidates are recruited, is infinity, for $N \geq 2$.

The present analysis approach will be useful again in the analysis of the class of "hiring above the α-quantile" for $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$; this is explained in detail in Chapter 5.

This study of "hiring above the median" gives new insights in the context of the CRP, as well as the p-percentile rules. For instance, the results for the *time of last occupied table* and the *time between opening the last two occupied tables* in the restaurant can be easily obtained for many classes of seating plans, i.e., $(\frac{1}{2}, 1)$, $(\frac{1}{2}, 0)$, $(\frac{1}{3}, 1)$, and others. Of course, all parameters studied for hiring above the α-quantile in general can be also defined for the p-percentile rules.

In particular, for the "$\frac{1}{2}$-percentile rule", since we already have the quantities $a_{n,\ell}^{[1]}$ and $a_{n,\ell}^{[2]}$, then the distributional results for the *index of last hired candidate* and the *distance between the last two hirings* follow directly with a little effort, while the limiting behaviour may be more involved. As the case of hiring above the median, other parameters for this percentile rule like the *rank of best discarded candidate* and others, will require considering auxiliary quantities. However, it is a matter of carrying out the computations, we leave them to a further work in the future.

Chapter 5

Hiring above the α-quantile

5.1 Introduction

We discuss in this chapter the general strategy, "hiring above the α-quantile of the hired staff", with $0 < \alpha < 1$, introduced first by Archibald and Martínez [5]. This strategy is a generalization of "hiring above the median" (with $\alpha = \frac{1}{2}$) discussed in Chapter 4. According to Definition 2.9, the α-quantile, with $0 < \alpha < 1$, of a sequence $x_1 < x_2 < \cdots < x_k$ of k elements is the element x_j with $j = \lceil \alpha k \rceil$. This strategy hires the first candidate, then any further candidate is hired if he ranks better than the α-quantile of the hired staff so far. We discuss here using the framework of Archibald and Martínez (refer to Section 2.4) to analyze hiring above the α-quantile. This approach can give "lower" and "upper" bounds on the studied parameters, giving us at least the order of growth of the expectation for those parameters, for general α. Since those bounds are not tight then we cannot deduce more than the order of growth.

We give first a summary of the main results for the bounds on three parameters: the *size of the hiring set,* h_n, the *gap of last hired candidate,* g_n and the *number of replacements,* f_n. The introduced theorems quantify precisely the order of growth of the expectation of the mentioned parameters, while similar results for other parameters are quite involved.

Moreover, we show that the framework of Archibald and Martínez can be used to analyze other probabilistic selection rules other than hiring strategies in a systematic analytic way. One example (given also in Chapter 4) is the seating plan $(\frac{1}{2}, 1)$ of the CRP (Section 2.5) which is exactly the upper bound of "hiring above the median", then we can obtain similar results for particular classes of seating plans like $(\alpha, 1)$ and $(\alpha, 0)$, $0 < \alpha < 1$.

After that, we show that a suitable extension of our recursive approach used to analyze hiring above the median (see Section 4.3), works well to obtain explicit results for h_n for "hiring above the α-quantile", when $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$.

The sequel of this chapter is organized as follows: Section 5.2 contains the main theorems of the lower and upper bounds on three parameters: h_n, g_n and f_n for the general case $0 < \alpha < 1$, followed by the proofs. Then, Section 5.3 gives the main result of the asymptotic distribution of h_n when $\alpha = \frac{1}{d}$ with $d \in \mathbb{N}$, together with the analysis. The results of this chapter appear in the technical report [49].

5.2 Lower and upper bounds

We follow here the framework introduced in Section 2.4 which uses directly the generating functions to perform the analysis. The first step is always to characterize the quantity $X(\sigma)$ of the strategy under study, which specifies how many candidates can be hired in the next step right after σ. This quantity is a unique value for each unique hiring strategy. For hiring above the α-quantile, $X(\sigma) = h(\sigma) - \lceil \alpha h(\sigma) \rceil + 1 = \lfloor (1-\alpha)h(\sigma) + 1 \rfloor$, with $0 < \alpha < 1$. For simplicity, we write $X(\sigma) = \lfloor a h(\sigma) + 1 \rfloor$, where $a = 1 - \alpha$. Rather than dealing with the ceilings, we shall consider the *lower* and *upper* bounds $X_\ell(\sigma) = a h_\ell(\sigma) + 1 - a$ and $X_u(\sigma) = a h_u(\sigma) + 1$, when $|\sigma| > 0$, and for the empty permutation, we set $X_\ell(\epsilon) = X_u(\epsilon) = 1$.

The lower and upper bounds X_ℓ and X_u will yield lower and upper bounds on several hiring parameters, i.e., $h_n^{(\alpha,\alpha)} \leq h_n \leq h_n^{(\alpha,1)}$ where $h_n^{(\alpha,\alpha)}$ and $h_n^{(\alpha,1)}$ represent the corresponding sizes of the hiring set under the strategies defined by $X_\ell(\sigma) = (1-\alpha)h_\ell(\sigma) + \alpha$ and $X_u(\sigma) = (1-\alpha)h_u(\sigma) + 1$, respectively, while h_n is the size of the hiring set under hiring above the α-quantile. We make use of the following proposition to establish such relationships for bounds on the parameters considered,

Proposition 5.1 *Let* A *and* B *be two pragmatic hiring strategies such that, for all* σ *with* $|\sigma| = n$, $X_A(\sigma) \leq X_B(\sigma)$. *Since both strategies are pragmatic, that means that if strategy* A *hires a candidate with score* j, *then the candidate will be also hired by strategy* B. *Then*

i) $h_n^{(A)} \leq_{st} h_n^{(B)}$.

ii) $W_N^{(A)} \geq_{st} W_N^{(B)}$.

iii) $L_n^{(A)} \leq_{st} L_n^{(B)}$.

iv) $g_n^{(A)} \leq_{st} g_n^{(B)}$.

v) $M_n^{(A)} \geq_{st} M_n^{(B)}$.

vi) $f_n^{(A)} \leq_{st} f_n^{(B)}$.

For any two positive random variables (r.v.'s) Y *and* Z, $Y \leq_{st} Z$ *(reads: "*Y *is stochastically smaller than or equal to* Z*") means that* $\mathbb{P}\{Y > t\} \leq \mathbb{P}\{Z > t\}$, *for all* $t \geq 0$. *Moreover,* $Y \leq_{st} Z$ *implies* $\mathbb{E}\{Y\} \leq \mathbb{E}\{Z\}$.

Proof:

i) It follows directly, since $X_A(\sigma) \leq X_B(\sigma)$, any candidate hired by A will be hired by B too.

ii) B has more choices than A to hire the next candidate, so B will have to wait less than A to hire another candidate.

iii) B hires at least the same candidates as A and possibly more.

iv) The reason is that B might hire a candidate worse than the last candidate hired by A, but not a candidate that is better; in that case A would hire that candidate too. Then, as $X(\sigma)$ increases the gap increases and vice-versa.

v) B cannot discard any candidate that A has hired, but A might discard a candidate that B would hire.

vi) We know that the number of choices to make a replacement right after processing a permutation σ is equal to $h(\sigma) - X(\sigma)$. Let say that $X_B(\sigma) = X_A(\sigma) + x(\sigma)$, assuming that $x(\sigma)$ is a monotone function. Since, for most pragmatic strategies, when $X(\sigma)$ is incremented by $\epsilon \leq 1$ (due to hiring a new candidate), then $h(\sigma)$ has been already incremented by 1. So that, during the hiring process, when $|\sigma| \geq |\sigma_o|$, $h(\sigma)$ should grow faster than $X(\sigma)$, hence

$$\begin{aligned} & h^{(B)} \geq h^{(A)} + x(\sigma) \quad \Rightarrow \quad h^{(B)} - h^{(A)} \geq x(\sigma) \quad \Rightarrow \quad h^{(B)} - h^{(A)} \geq X_B - X_A \\ & \Rightarrow \quad h^{(B)} - X_B \geq h^{(A)} - X_A \\ & \Rightarrow \quad f^{(A)} \leq f^{(B)}. \end{aligned}$$

■

It is important to clarify that both selection rules defined by X_ℓ and X_u are "pragmatic" but do not correspond to actual hiring strategies. Pragmaticity conditions (Definition 2.7) hold here but since we are dealing with rank-based strategies, then the function $X(\sigma)$ should give an *integer* value (as pointed out also by Krieger et al., Subsection 2.2.1) that is the number of choices to hire the next candidate. When $X(\sigma)$ is not always integer-valued then we cannot always specify the *threshold candidate* during hiring; in other terms the threshold candidate does not always exist in the hiring set.
For example, let us assume that there are k hired candidates so far by three strategies: hiring above the median with $X_{med} = \lfloor \frac{k+2}{2} \rfloor$, and the strategies defined by $X_\ell = \frac{1}{2}k + \frac{1}{2}$ and $X_u = \frac{1}{2}k + 1$. Then we can say that the two later strategies represent lower and upper bounds of hiring above the median, respectively. For odd $k = 2t - 1$, we have $X_{med} = t$, $X_\ell = t$ and $X_u = t + \frac{1}{2}$, while for even $k = 2t - 2$, we have $X_{med} = t$, $X_\ell = t - \frac{1}{2}$ and $X_u = t$.

In general, this can be proved by induction as follows: initially we have $X_A(\epsilon) = 1$ and $X_B(\epsilon) = 1$, then assuming that $X_A(\sigma) \leq X_B(\sigma)$ for all σ, $|\sigma| = n$ leads to get $h_n^{(A)} \leq h_n^{(A)}$. Now, since for almost all pragmatic strategies $X(\sigma) = f(h(\sigma))$ is a monotone function, then $h_n^{(A)} \leq h_n^{(A)}$ implies that $X_A(\sigma) \leq X_B(\sigma)$ and so on.

Notice that the strategy defined by X_u is *equivalent* to the seating plan $(\frac{1}{2}, 1)$—discussed in Subsection 4.4.2. The distributional and asymptotic results for the *number of selections* parameter for this rule have been characterized elsewhere, but it will be interesting to obtain similar results using the framework here.
We conclude that X_ℓ and X_u define two probabilistic selection rules in terms of the *probabilities of selection* (similar to the seating plans $(\alpha, 1)$ and $(\alpha, 0)$).

5.2.1 Results

Theorem 5.1 *For "hiring above the α-quantile", with $0 < \alpha < 1$, let h_n denote the size of the hiring set, then*

$$\mathbb{E}\left\{h_n^{(\alpha,\alpha)}\right\} \leq \mathbb{E}\{h_n\} \leq \mathbb{E}\left\{h_n^{(\alpha,1)}\right\},$$

where asymptotically as $n \to \infty$,

$$\mathbb{E}\left\{h_n^{(\alpha,\alpha)}\right\} = \frac{1}{(1-\alpha)(2-\alpha)} \cdot \frac{n^{1-\alpha}}{\Gamma(2-\alpha)} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right),$$
$$\mathbb{E}\left\{h_n^{(\alpha,1)}\right\} = \frac{1}{1-\alpha} \cdot \frac{n^{1-\alpha}}{\Gamma(2-\alpha)} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right).$$

Theorem 5.2 *For "hiring above the α-quantile", with* $0 < \alpha < 1$, *let* g_n *denote the gap of last hired candidate, then*

$$\mathbb{E}\left\{g_n^{(\alpha,\alpha)}\right\} \le \mathbb{E}\{g_n\} \le \mathbb{E}\left\{g_n^{(\alpha,1)}\right\},$$

where asymptotically as $n \to \infty$,

$$\mathbb{E}\left\{g_n^{(\alpha,\alpha)}\right\} = \frac{1}{2(2-\alpha)} \cdot \frac{n^{-\alpha}}{\Gamma(2-\alpha)} + \mathcal{O}\Big(\frac{1}{n}\Big),$$
$$\mathbb{E}\left\{g_n^{(\alpha,1)}\right\} = \frac{1}{2} \cdot \frac{n^{-\alpha}}{\Gamma(2-\alpha)} + \mathcal{O}\Big(\frac{1}{n}\Big).$$

Theorem 5.3 *For "hiring above the α-quantile", with* $0 < \alpha < 1$, *let* f_n *denote the number of replacements, then*

$$\mathbb{E}\left\{f_n^{(\alpha,\alpha)}\right\} \le \mathbb{E}\{f_n\} \le \mathbb{E}\left\{f_n^{(\alpha,1)}\right\},$$

where asymptotically as $n \to \infty$,

$$\mathbb{E}\left\{f_n^{(\alpha,\alpha)}\right\} = \frac{\alpha}{(1-\alpha)^2(2-\alpha)} \cdot \frac{n^{1-\alpha}}{\Gamma(2-\alpha)} - \frac{\alpha}{1-\alpha}\ln n + \mathcal{O}(1),$$
$$\mathbb{E}\left\{f_n^{(\alpha,1)}\right\} = \frac{\alpha}{(1-\alpha)^2} \cdot \frac{n^{1-\alpha}}{\Gamma(2-\alpha)} - \frac{1}{1-\alpha}\ln n + \mathcal{O}(1).$$

Theorem 5.4 *Let* $h_n^{(\frac{1}{2},\frac{1}{2})}$ *and* $h_n^{(\frac{1}{2},1)}$, *denote the sizes of the hiring sets for the selection rules defined by* $X_\ell(\sigma) = \frac{1}{2}h(\sigma) + \frac{1}{2}$ *and* $X_u(\sigma) = \frac{1}{2}h(\sigma) + 1$ *respectively, then these rules bound hiring above the median. The explicit distributions of* $h_n^{(\frac{1}{2},\frac{1}{2})}$ *and* $h_n^{(\frac{1}{2},1)}$ *are given as follows:*

$$\mathbb{P}\left\{h_n^{(\frac{1}{2},\frac{1}{2})} = k\right\} = \frac{k}{n}2^{k+1-2n}\sum_{j=0}^{n-k} 2^j \binom{2n-k-j-1}{n-1},$$
$$\mathbb{P}\left\{h_n^{(\frac{1}{2},1)} = k\right\} = (k+1)2^{k-2n}\frac{k}{n}\binom{2n-k-1}{n-1}.$$

Asymptotically as $n \to \infty$:
The normalized r.v. $\frac{h_n^{(\frac{1}{2},\frac{1}{2})}}{\sqrt{n}} \xrightarrow{(d)} Y$, *where* Y *has the probability density function:*

$$f(y) = \frac{y}{\sqrt{\pi}}\int_{t=y}^{\infty} e^{-\frac{t^2}{4}}\,dt,$$

similarly, $\frac{h_n^{(\frac{1}{2},1)}}{\sqrt{2n}} \xrightarrow{(d)} Z$, *where* Z *has a* Maxwell-Boltzmann *distribution with parameter* $\sqrt{2}$.

5.2.2 Analysis

5.2.2.1 Size of the hiring set

We have to apply Theorem 1.4 that describes the PDE of the size of hiring set. Also we need to define $X(\sigma)$ which is $ah(\sigma)+b$ as mentioned above. Remember that we use $a = 1-\alpha$ for simplicity. Thus, the PDE of $H(z,u)$ takes the following form

$$(1-z)\frac{\partial H_{a,b}}{\partial z} - au(u-1)\frac{\partial H_{a,b}}{\partial u} - (1+b(u-1))H_{a,b}(z,u) = (u-1)(1-b). \tag{5.1}$$

It is difficult to obtain a closed form of $H(z,u)$ from (5.1), so we will go directly to the next step by differentiating *w*.r.t. u and setting $u = 1$. Then

$$(1-z)\frac{\partial}{\partial z}h_{a,b}(z) - (1+a)h_{a,b}(z) - \frac{b}{1-z} = 1-b.$$

The solution turns out to be

$$h_{a,b}(z) = \frac{-1}{(1-z)^{1+a}}\left(\frac{(1-z)^a(az(b-1)+a+b)}{a(1+a)} + C\right),$$

with the initial condition $h(0) = 0$; we get thus $C = -\frac{b+a}{a(1+a)}$. From singularity analysis (review Section 1.4), we have then

$$[z^n]h_{a,b}(z) = \frac{n^a}{\Gamma(1+a)}\cdot\frac{b+a}{a(1+a)}\cdot\left(1+\mathcal{O}\Big(\frac{1}{n}\Big)\right), \tag{5.2}$$

where $\Gamma(\cdot)$ is the Gamma function. Then replacing b by $1-a$ in (5.2) we have a lower bound

$$[z^n]h_{a,a}(z) = \frac{1}{a(a+1)}\cdot\frac{n^a}{\Gamma(1+a)}\cdot\left(1+\mathcal{O}\Big(\frac{1}{n}\Big)\right). \tag{5.3}$$

Replacing b by 1 in (5.2), we have the upper bound

$$[z^n]h_{a,1}(z) = \frac{1}{a}\cdot\frac{n^a}{\Gamma(1+a)}\cdot\left(1+\mathcal{O}\Big(\frac{1}{n}\Big)\right). \tag{5.4}$$

We replace a by $1-\alpha$ in (5.3) and (5.4) to obtain the results for $\mathbb{E}\left\{h_n^{(\alpha,\alpha)}\right\}$ and $\mathbb{E}\left\{h_n^{(\alpha,1)}\right\}$, respectively, in Theorem 5.1. We get the same result as Theorem 2.20, for hiring above the α-quantile, we have $\mathbb{E}\{h_n\} = \Theta\big(n^{1-\alpha}\big)$.

5.2.2.2 Gap of last hired candidate

Following Theorem 2.17 for the gap. Let

$$X(z) = \sum_{\sigma\in\mathcal{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}.$$

Then for $X(\sigma) = ah(\sigma) + b$, we have

$$\begin{aligned}\sum_{\sigma\in\mathcal{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!} &= \sum_{\sigma\in\mathcal{P}} \big(ah(\sigma)+b\big)\frac{z^{|\sigma|}}{|\sigma|!}\\ &= a\cdot\sum_{\sigma\in\mathcal{P}} h(\sigma)\frac{z^{|\sigma|}}{|\sigma|!} + b\cdot\sum_{\sigma\in\mathcal{P}}\frac{z^{|\sigma|}}{|\sigma|!}\\ &= a\cdot h_{a,b}(z) + b.\end{aligned}$$

We have the value of $[z^n]h_{a,b}(z)$ from (5.2) hence,

$$\frac{1}{2}\big([z^n]X_{a,b}(z)-1\big) = \frac{b+a}{2(1+a)}\cdot\frac{n^{a-1}}{\Gamma(1+a)} + \mathcal{O}\Big(\frac{1}{n}\Big). \tag{5.5}$$

The lower bound is

$$\frac{1}{2(1+a)}\cdot\frac{n^{a-1}}{\Gamma(1+a)} + \mathcal{O}\Big(\frac{1}{n}\Big), \tag{5.6}$$

while the upper bound is

$$\frac{1}{2}\cdot\frac{n^{a-1}}{\Gamma(1+a)} + \mathcal{O}\Big(\frac{1}{n}\Big). \tag{5.7}$$

In general, as given before in Theorem 2.20, for hiring above the α-quantile, we get $\mathbb{E}\{g_n\} = \Theta(n^{-\alpha})$. Observe that for any $\alpha < 1$, $g_n \to 0$ as $n \to \infty$. We substitute $a = 1-\alpha$ in (5.6) and (5.7) to get the results stated in Theorem 5.2.

5.2.2.3 Number of replacements

We begin with the following trivariate generating function

$$F(z,u,v) = \sum_{\sigma\in\mathbb{P}} \frac{z^{|\sigma|}}{|\sigma|!}u^{f(\sigma)}v^{h(\sigma)}, \tag{5.8}$$

where $f(\sigma)$ is the number of replacements made to process the permutation σ. Again, we use the catalytic variable v to be able to proceed in the analysis of strategies that have $X(\sigma) = f(h(\sigma))$.

Referring to the discussion of hiring with replacements in Section 3.3, of the $|\sigma|+1$ possible rankings coming after σ, $|\sigma|+1-h(\sigma)$ will be discarded, $X(\sigma)$ will be hired without replacement and $h(\sigma)-X(\sigma)$ will be hired with replacement. So that the recurrence of $f(\sigma)$ will take the following form:

$$f(\sigma\circ j) = \begin{cases} f(\sigma), & \text{if } 1 \le j \le |\sigma|+1-h(\sigma) \text{ (j is discarded)},\\ f(\sigma)+1, & \text{if } |\sigma|+2-h(\sigma) \le j \le |\sigma|+1-X(\sigma) \text{ (replaces worst)},\\ f(\sigma), & \text{if } |\sigma|+2-X(\sigma) \le j \le |\sigma|+1 \text{ (j is hired)}.\end{cases}$$

This yields the next theorem, whose proof we omit as it closely follows that one of Theorem 1.4.

Theorem 5.5 *Let $F(z,u,v)$ be the generating function defined in* (5.8). *Let $X(\sigma)$ denote the number of ranks j, $1 \le j \le |\sigma|+1$, such that a candidate with score j will be hired without replacing anyone, if interviewed right after σ.*

Then

$$(1-z)\frac{\partial}{\partial z}F(z,u,v) - F(z,u,v) = v(u-1)\frac{\partial}{\partial v}F(z,u,v) + (v-u)\sum_{\sigma\in\mathbb{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}u^{f(\sigma)}v^{h(\sigma)}.$$

In order to compute the expected number of replacements for a random permutation of size n, we can differentiate $F(z,u,v)$ *w.*r.t. u and set $u = 1$. Then the differential equation given in Theorem 5.5 transforms into

$$\begin{aligned}(1-z)\frac{\partial}{\partial z}f(z,v) - f(z,v) = v\frac{\partial}{\partial v}F(z,1,v) + (v-1)\sum_{\sigma\in\mathbb{P}} X(\sigma)f(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}v^{h(\sigma)} \\ - \sum_{\sigma\in\mathbb{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}v^{h(\sigma)},\end{aligned} \tag{5.9}$$

with

$$f(z,v) = \frac{\partial}{\partial u}F(z,u,v)\bigg|_{u=1}.$$

Now, $F(z,1,v) = H(z,v)$, and we have to set $v = 1$ in (5.9) to get rid of the catalytic variable v and thus obtain the generating function $f(z)$ for the expected values f_n:

$$(1-z)\frac{d}{dz}f(z) - f(z) = h(z) - \sum_{\sigma\in\mathbb{P}} X(\sigma)\frac{z^{|\sigma|}}{|\sigma|!}, \tag{5.10}$$

where

$$h(z) = \frac{\partial}{\partial v}H(z,v)\bigg|_{v=1}$$

is the generating function for $\mathbb{E}\{h_n\}$, and the initial condition is $f(0) = 0$.
Now, we set $X(\sigma) = a\cdot h(\sigma) + b$ in (5.10), thus we have

$$(1-z)\frac{d}{dz}f_{a,b}(z) - f_{a,b}(z) = (1-a)h_{a,b}(z) - \frac{bz}{1-z},$$

where

$$h_{a,b}(z) = \frac{-1}{(1-z)^{a+1}}\left(\frac{(1-z)^a(az(b-1)+a+b)}{a(a+1)} - \frac{b+a}{a(1+a)}\right),$$

as explained above in this subsection. The solution for $f_{a,b}(z)$ is

$$f_{a,b}(z) = \frac{(a+b)(1-a)}{a^2(1+a)}\cdot\frac{1}{(1-z)^{1+a}} - \frac{b}{a}\cdot\frac{1}{1-z}\ln\Big(\frac{1}{1-z}\Big) + \frac{C+z(2b+a-1)}{1-z}, \tag{5.11}$$

where C is a constant, that can be computed using the initial condition $f(0) = 0$. However, the value of C is irrelevant, as the last term in (5.11) can be ignored as it is not dominant. Using again singularity analysis to obtain the asymptotic of the nth coefficient of both the lower bound ($b = 1 - a$) and upper bound ($b = 1$), we get

$$[z^n]f_{a,1-a}(z) = \frac{1-a}{a^2(1+a)} \cdot \frac{n^a}{\Gamma(1+a)} - \frac{1-a}{a} \ln n + \mathcal{O}(1), \tag{5.12}$$

$$[z^n]f_{a,1}(z) = \frac{(1-a)}{a^2} \cdot \frac{n^a}{\Gamma(1+a)} - \frac{1}{a} \ln n + \mathcal{O}(1). \tag{5.13}$$

Thus, for hiring above the α-quantile, $\mathbb{E}\{f_n\} = \Theta\big(n^{1-\alpha}\big)$. In particular, $\mathbb{E}\{f_n\} = \Theta\big(\mathbb{E}\{h_n\}\big)$. As usual, we replace a by $1 - \alpha$ to get the results in Theorem 5.3.

5.2.2.4 Bounds on hiring above the median

As we mentioned before, hiring above the median is a special case of hiring above the α-quantile when $\alpha = \frac{1}{2}$. So that we set $a = \frac{1}{2}$ in the general case equations to obtain the bounds of the parameters of this strategy.
Substituting $a = \frac{1}{2}$ in (5.3) and (5.4), we have the following bounds

$$\mathbb{E}\Big\{h_n^{(\frac{1}{2},\frac{1}{2})}\Big\} = \frac{8\sqrt{n}}{3\sqrt{\pi}}\left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right), \tag{5.14}$$

$$\mathbb{E}\Big\{h_n^{(\frac{1}{2},1)}\Big\} = \frac{4\sqrt{n}}{\sqrt{\pi}}\left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right). \tag{5.15}$$

This implies that, for hiring above the median, $\mathbb{E}\{h_n\} = \Theta\big(\sqrt{n}\big)$ as mentioned in Chapter 4.
The solution of the PDE for $H_{a,b}(z)$ is difficult in general, but not to when $a = \frac{1}{2}$ and $b \in \big\{\frac{1}{2}, 1\big\}$, which give us lower and upper bounds on the probability distribution of h_n. So, substituting $a = \frac{1}{2}$ and $b = \frac{1}{2}$ in (5.1) and using the initial condition $H(0, u) = 1$, we find the solution as

$$H_{\frac{1}{2},\frac{1}{2}}(z,u) = \frac{1-u^2}{(1-u)^2} - \frac{2u}{(1-u)(1-u+u\sqrt{1-z})} + \frac{2u}{(1-u)^2}\left(\frac{1}{2}\ln\left(\frac{1}{1-z}\right) - \ln\left(\frac{1}{1-u+u\sqrt{1-z}}\right)\right).$$

For the upper bound, we set $a = \frac{1}{2}$ and $b = 1$, then

$$H_{\frac{1}{2},1}(z,u) = \frac{1}{(1-u+u\sqrt{1-z})^2}.$$

Since we have these closed forms, then we can obtain the following useful information. First we can give the factorial moments of each r.v. there:

$$\begin{aligned}
\mathbb{E}\left\{h_n^{(\frac{1}{2},\frac{1}{2})\underline{r}}\right\} &= \Theta\left(n^{r/2}\right),\\
\mathbb{E}\left\{h_n^{(\frac{1}{2},1)\underline{r}}\right\} &= [z^n]\frac{(r+1)!\left(1-\sqrt{1-z}\right)^r}{(1-z)^{r/2+1}}\\
&= (-1)^r(r+1)!\sum_{k=0}^{r}(-1)^k\binom{r}{k}\binom{n+k/2}{n}\\
&= \Theta\left(n^{r/2}\right).
\end{aligned}$$

A more precise asymptotic estimation of $\mathbb{E}\left\{h_n^{(\frac{1}{2},1)\underline{r}}\right\}$ is given later in (5.17).

Now, extracting the coefficients of $[u^k z^n]H_{\frac{1}{2},1}(z,u)$ gives us the probability mass functions as follows

$$\begin{aligned}
\mathbb{P}\left\{h_n^{(\frac{1}{2},\frac{1}{2})}=k\right\} &= \frac{k}{n}2^{k+1-2n}\sum_{j=0}^{n-k}2^j\binom{2n-k-j-1}{n-1},\\
\mathbb{P}\left\{h_n^{(\frac{1}{2},1)}=k\right\} &= (k+1)2^{k-2n}\frac{k}{n}\binom{2n-k-1}{n-1}.
\end{aligned}\tag{5.16}$$

The result in (5.16) is exactly what we obtain before in Theorem 4.12 for the number of occupied tables in restaurant under the seating plan $(\frac{1}{2},1)$.
We can obtain the limiting distribution in both cases. However, it is known for the case of upper bound as given in Theorem 4.12, but it will be interesting also to obtain it here in a different way, namely, using the method of moments.
In case of lower bound, first we use the following absolute approximation given in [37]

$$\binom{2n-k-1}{n-1} \sim 2^{2n-k}\frac{n}{2n-k}\frac{e^{-k^2/4n}}{\sqrt{\pi n}}.$$

Then

$$\begin{aligned}
\mathbb{P}\left\{h_n^{(\frac{1}{2},\frac{1}{2})}=k\right\} &\sim \frac{k}{\sqrt{\pi}n^{3/2}}\sum_{j=0}^{n-k}e^{-\frac{(k+j)^2}{4n}}\\
&\sim \frac{k}{\sqrt{\pi}n}\left(\int_{t=\frac{k}{\sqrt{n}}}^{\sqrt{n}}e^{-\frac{t^2}{4}}dt+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)\\
&\sim \frac{k}{\sqrt{\pi}n}\int_{t=\frac{k}{\sqrt{n}}}^{\infty}e^{-\frac{t^4}{4}}dt, \quad \text{as } n\to\infty.
\end{aligned}$$

So that the normalized r.v. $\frac{h_n^{(\frac{1}{2},\frac{1}{2})}}{\sqrt{n}} \xrightarrow{(d)} Y$, where Y has the probability density function:

$$f(y) = \frac{y}{\sqrt{\pi}}\int_{t=y}^{\infty}e^{-\frac{t^2}{4}}dt$$

Moreover, we can compute the moments as follows:

$$\begin{aligned}\mathbb{E}\{Y^r\} &= \int_0^\infty y^r \cdot f(y)\,dy \\ &= \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{y^{r+2}}{r+2} e^{-\frac{y^2}{4}}\,dy \\ &= \frac{2^{r+2}}{\sqrt{\pi}(r+2)} \cdot \Gamma\Big(\frac{r+3}{2}\Big).\end{aligned}$$

For the upper bound, we start with following exact closed form of the moment generating function

$$\begin{aligned}\mathbb{E}\Big\{h_n^{(\frac{1}{2},1)\underline{r}}\Big\} &= [z^n] \frac{(r+1)!\Big(1-\sqrt{1-z}\Big)^r}{(1-z)^{r/2+1}} \\ &= [z^n] \frac{(r+1)!}{1-z}\Big(\frac{1}{\sqrt{1-z}} - 1\Big)^r.\end{aligned}$$

Then using the singularity analysis, one can write

$$\begin{aligned}\mathbb{E}\Big\{h_n^{(\frac{1}{2},1)\underline{r}}\Big\} &\sim [z^n] \frac{(r+1)!}{(1-z)^{1+r/2}} \\ &= (r+1)! \binom{n+r/2}{n} \\ &= (r+1)! \frac{(n+\frac{r}{2})(n+\frac{r}{2}-1)(n+\frac{r}{2}-2)\dots(n+1)}{(\frac{r}{2})!} \\ &= n^{r/2} \frac{(r+1)\cdot r!}{\Gamma(\frac{r}{2}+1)} \cdot \Big(1 + \mathcal{O}\Big(n^{r/2-1}\Big)\Big). \qquad (5.17)\end{aligned}$$

Now we can use the following property of Gamma function:

$$\Gamma\Big(\frac{r}{2}+\frac{1}{2}\Big) = \frac{\sqrt{\pi}\,2^{-r}\,r!}{\Gamma(\frac{r}{2}+1)},$$

thus

$$\mathbb{E}\Big\{h_n^{(\frac{1}{2},1)\underline{r}}\Big\} \sim n^{r/2}\,2^r\,\frac{(r+1)\Gamma(\frac{r}{2}+\frac{1}{2})}{\sqrt{\pi}}.$$

If we consider this normalized r.v. $h_n' = \frac{h_n^{(\frac{1}{2},1)}}{\sqrt{2n}}$, then

$$\mathbb{E}\Big\{h_n'^{\underline{r}}\Big\} \sim 2^{r/2}\,\frac{(r+1)\Gamma(\frac{r}{2}+\frac{1}{2})}{\sqrt{\pi}}, \qquad (5.18)$$

since the moment generating function of the Chi distribution with parameter $k = 3$ is given as

$$M_r = \frac{2^{r/2}\,\Gamma(\frac{r+3}{2})}{\Gamma(3/2)} = \frac{2^{r/2}\,(\frac{r+1}{2})\,\Gamma(\frac{r+1}{2})}{\frac{1}{2}\,\Gamma(1/2)} = \frac{2^{r/2}\,(r+1)\,\Gamma(\frac{r+1}{2})}{\sqrt{\pi}}. \qquad (5.19)$$

Using the method of moments, from (5.18) and (5.19) we can proof that asymptotically as $n \to \infty$, $\frac{h_n^{(\frac{1}{2},1)}}{\sqrt{2n}} \xrightarrow{(d)} Z$, where Z follows a Chi distribution with $k = 3$, namely Maxwell-Boltzmann distribution with parameter $\sqrt{2}$.

5.3 Hiring above the $\frac{1}{d}$-quantile

Our approach to the study of hiring above the median in Chapter 4 enables us to understand well the hiring process under such strategy, but we are not able to find a suitable combinatorial explanation for the probability distribution of h_n. This reflects that we have to investigate other special cases of α in order to generalize our results if possible. For example, hiring above the $\frac{1}{3}$-quantile: in this case we have three different states for the automaton according to the number of hired candidates k; where k is congruent to 0,1, or 2 modulo 3. We can write the recurrences of the quantities $a^{[1]}_{n,\ell}$, $a^{[2]}_{n,\ell}$ and $a^{[3]}_{n,\ell}$, which are the probabilities that after receiving n candidates, the threshold candidate has the ℓ-th largest score in the hiring set and k is 1(mod 3), 2(mod 3) and 0(mod 3) respectively. We give the main result for *hiring above the $\frac{1}{d}$-quantile*, then we discuss the analysis of the special case $\alpha = \frac{1}{3}$, followed by the proof of our theorem.

Theorem 5.6 *For "hiring above the $\frac{1}{d}$-quantile", $d \in \mathbb{N}$, let h_n denote the size of the hiring set after n interviews. Then the normalized r.v. $\frac{h_n}{n^{1-\frac{1}{d}}} \xrightarrow{(d)} X$, where X has the density function:*

$$f(x) = \frac{1}{d^{\frac{1}{d-1}}\left(\frac{1}{d-1}\right)!} \cdot x^{\frac{1}{d-1}} \cdot \exp\Big(-\frac{(d-1)^{d-1}}{d^d} \cdot x^d\Big), \quad x > 0.$$

Moreover,

$$\mathbb{E}\{X\} = \frac{d\left(\frac{1}{d-1} + \frac{1}{d} - 1\right)!}{(d-1)^{2-\frac{1}{d}}\left(\frac{1}{d-1}\right)!}.$$

5.3.1 Analysis

In general, the α-quantile of a sequence $r_1 < r_2 < \cdots < r_k$ of k elements is the element r_j with $j = \lceil \alpha k \rceil$, hence it is the $(k-j+1)$-th largest one. For $\alpha = \frac{1}{3}$ we can track the evolution of the threshold candidate which is the ℓ-th largest one in the hiring set via the following table:

k	1	2	3	4	5	6	7	$\cdots$
ℓ	1	2	3	3	4	5	5	$\cdots$

Notice that when we move from the state $a^{[3]}_{n,\ell}$ to $a^{[1]}_{n,\ell}$, ℓ is still the same while ℓ is incremented when moving from $a^{[1]}_{n,\ell}$ to $a^{[2]}_{n,\ell}$ or from $a^{[2]}_{n,\ell}$ to $a^{[3]}_{n,\ell}$. Thus for $n \geq 2$ and $1 \leq \ell \leq n$:

$$\begin{aligned}
a^{[1]}_{n,\ell} &= \Big(1 - \frac{\ell}{n}\Big) \cdot a^{[1]}_{n-1,\ell} + \frac{\ell}{n} \cdot a^{[3]}_{n-1,\ell} \\
a^{[2]}_{n,\ell} &= \Big(1 - \frac{\ell}{n}\Big) \cdot a^{[2]}_{n-1,\ell} + \frac{\ell-1}{n} \cdot a^{[1]}_{n-1,\ell-1} \\
a^{[3]}_{n,\ell} &= \Big(1 - \frac{\ell}{n}\Big) \cdot a^{[3]}_{n-1,\ell} + \frac{\ell-1}{n} \cdot a^{[2]}_{n-1,\ell-1},
\end{aligned}$$

with the initial conditions: $a^{[1]}_{1,1} = 1$, $a^{[2]}_{1,1} = a^{[3]}_{1,1} = 0$.
We simplify those recurrences by introducing a suitable normalization:

$$c^{[i]}_{n,\ell} = \frac{n!}{(n-\ell)! \cdot (\ell-1)!} \cdot a^{[i]}_{n,\ell}, \quad i = 1, 2, 3. \tag{5.20}$$

which leads us to the following relations:

$$\begin{aligned}(n-\ell)\cdot c^{[1]}_{n,\ell} &= (n-\ell)\cdot c^{[1]}_{n-1,\ell} + \ell\cdot c^{[3]}_{n-1,\ell}\\ c^{[2]}_{n,\ell} &= c^{[2]}_{n-1,\ell} + c^{[1]}_{n-1,\ell-1}\\ c^{[3]}_{n,\ell} &= c^{[3]}_{n-1,\ell} + c^{[2]}_{n-1,\ell-1},\end{aligned}$$

Now we introduce the following generating function:

$$C^{[i]}(z,u) = \sum_{n\geq 1}\sum_{1\leq\ell\leq n} c^{[i]}_{n,\ell}\cdot z^{n-\ell}u^{\ell}, \quad i = 1,2,3,$$

which when applied to the above system of recurrences gives:

$$\begin{aligned}\frac{\partial}{\partial z}\Big((1-z)C^{[1]}(z,u)\Big) &= u\cdot\frac{\partial}{\partial u}C^{[3]}(z,u)\\ C^{[2]}(z,u) &= \frac{u}{1-z}\cdot C^{[1]}(z,u)\\ C^{[3]}(z,u) &= \frac{u}{1-z}\cdot C^{[2]}(z,u),\end{aligned}$$

It is easy to infer that $C^{[3]}(z,u) = \frac{u^2}{(1-z)^2}\cdot C^{[1]}(z,u)$ to obtain the following PDE:

$$(1-z)^2\frac{\partial}{\partial z}\Big((1-z)C^{[1]}(z,u)\Big) = u\cdot\frac{\partial}{\partial u}\Big(u^2C^{[1]}(z,u)\Big),$$

whose explicit solution contains some arbitrary function $F(\cdot)$ as follows:

$$C^{[1]}(z,u) = \frac{1}{(1-z)u^2}\cdot F\left(\frac{u^2-(1-z)^2}{u^2(1-z)^2}\right).$$

But the initial condition $a^{[1]}_{1,1} = 1$ implies that $C^{[1]}(0,u) = \sum_{n\geq 1} c^{[1]}_{n,n}u^n = u$; this shows that:

$$F(x) = \frac{1}{(1-x)^{3/2}}.$$

Then we get:

$$C^{[1]}(z,u) = \frac{1}{(1-z)u^2}\frac{1}{\left(1-\frac{u^2-(1-z)^2}{u^2(1-z)^2}\right)^{3/2}} = \frac{u}{(1-z)\left(1-\left(\frac{1-(1-z)^2}{(1-z)^2}\right)u^2\right)^{3/2}}.$$

Now we can extract the coefficients of $c^{[1]}_{n,\ell}$ as follows:

$$\begin{aligned}
c^{[1]}_{n,\ell} &= [z^{n-\ell}u^\ell]C^{[1]}(z,u) = [z^{n-\ell}u^{\ell-1}]\frac{1}{(1-z)\left(1-\left(\frac{1-(1-z)^2}{(1-z)^2}\right)u^2\right)^{3/2}}\\
&= [z^{n-\ell}(u^2)^{(\ell-1)/2}]\frac{1}{(1-z)\left(1-\left(\frac{1-(1-z)^2}{(1-z)^2}\right)u^2\right)^{3/2}}\\
&= \binom{\frac{\ell}{2}}{\frac{\ell-1}{2}}\cdot[z^{n-\ell}]\frac{1}{1-z}\left(\frac{1-(1-z)^2}{(1-z)^2}\right)^{(\ell-1)/2}\\
&= \binom{\frac{\ell}{2}}{\frac{\ell-1}{2}}\cdot[z^{n-\frac{3\ell}{2}+\frac{1}{2}}]\frac{(1+(1-z))^{(\ell-1)/2}}{(1-z)^\ell}\\
&= \binom{\frac{\ell}{2}}{\frac{\ell-1}{2}}\cdot[z^{n-\frac{3\ell}{2}+\frac{1}{2}}]\sum_{j=0}^{\frac{\ell-1}{2}}\binom{\frac{\ell-1}{2}}{j}\cdot\frac{1}{(1-z)^{\ell-j}}\\
&= \binom{\frac{\ell}{2}}{\frac{\ell-1}{2}}\sum_{j=0}^{\frac{\ell-1}{2}}\binom{\frac{\ell-1}{2}}{j}\cdot\binom{n-\frac{\ell}{2}-j-\frac{1}{2}}{\ell-j-1}.
\end{aligned}$$

Remember the normalization used in (5.22), then we have:

$$a^{[1]}_{n,\ell} = \begin{cases} \frac{1}{\ell\binom{n}{\ell}}\cdot\binom{\frac{\ell}{2}}{\frac{\ell-1}{2}}\sum_{j=0}^{\frac{\ell-1}{2}}\binom{\frac{\ell-1}{2}}{j}\cdot\binom{n-\frac{\ell}{2}-j-\frac{1}{2}}{\ell-j-1}, & \text{if } k=1 \pmod 3.\\ 0, & \text{otherwise.}\end{cases} \tag{5.21}$$

For the other cases of $a^{[\cdot]}_{n,\ell}$, which turn out to be easier, we proceed as above: we obtain the corresponding generating function $C^{[\cdot]}(z,u)$ first, then we extract the coefficients.

5.3.1.1 Proof of Theorem 5.6

The approach used for the analysis still prove useful and as we have seen, the computations for $\alpha = \frac{1}{3}$ to obtain the fundamental quantities $a_{n,\ell}^{[\cdot]}$ were relatively easy and we got the desired results. This encourages us to investigate some generalization in this class of hiring strategies when $\alpha = \frac{1}{d}$; namely *hiring above the* $\frac{1}{d}$*-quantile*, $d \in \mathbb{N}$. We know that for $\alpha = \frac{1}{d}$, we have d recurrences that describe the relationships between the quantities $a_{n,\ell}^{[\cdot]}$, but the trick is always to find some suitable normalization to reduce the resulting system of differential equations into only one PDE in one function, after that we expect that the computations will be a routine task. We start writing the following recurrence relations for $n \geq 2$, $1 \leq \ell \leq n$ and $2 \leq i \leq d$:

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

We need again this normalization:

$$c_{n,\ell}^{[i]} = \frac{n!}{(n-\ell)! \cdot (\ell-1)!} \cdot a_{n,\ell}^{[i]}, \quad 1 \leq i \leq d, \tag{5.22}$$

together with the generating function:

$$C^{[i]}(z,u) = \sum_{n \geq 1} \sum_{1 \leq \ell \leq n} c_{n,\ell}^{[i]} \cdot z^{n-\ell} u^{\ell}, \quad 1 \leq i \leq d.$$

As before we obtain a system of PDEs:

$$\frac{\partial}{\partial z}\left((1-z)C^{[1]}(z,u)\right) = u \cdot \frac{\partial}{\partial u} C^{[d]}(z,u)$$
$$C^{[i]}(z,u) = \frac{u}{1-z} \cdot C^{[i-1]}(z,u).$$

Thus we have $C^{[d]}(z,u) = \frac{u^{d-1}}{(1-z)^{d-1}} \cdot C^{[1]}(z,u)$, and we can write the following PDE:

$$(1-z)^{d-1} \frac{\partial}{\partial z}\left((1-z)C^{[1]}(z,u)\right) = u \cdot \frac{\partial}{\partial u}\left(u^{d-1} C^{[1]}(z,u)\right). \tag{5.23}$$

A direct solution for this PDE will be too complicated, so we transform it into a simpler form. If we consider the function:

$$C_{\ell}^{[\cdot]}(z) = [u^{\ell}]C^{[\cdot]}(z,u) = \sum_{n \geq \ell} c_{n,\ell}^{[\cdot]} z^n,$$

Then treating (5.23) gives:

$$(1-z)^{d-1} \frac{\partial}{\partial z}\left((1-z)C_{\ell}^{[1]}(z)\right) = \ell \cdot C_{\ell-d+1}^{[1]}(z). \tag{5.24}$$

We need again to introduce another normalization in order to obtain the last PDE in a useful form. First, we have

$$\hat{C}_{\ell}^{[\cdot]}(z) = \frac{C_{\ell}^{[\cdot]}(z)}{\ell!^{(d-1)}},$$

where $X!^{(r)}$ denotes the multifactorial of X (1.8); it also has following alternative definition, which is suitable when $X = 1 (\mathrm{mod}\ r)$:

$$X!^{(r)} = r^{\frac{X-1}{r}} \frac{\Gamma(\frac{X}{r}+1)}{\Gamma(\frac{1}{r}+1)}. \tag{5.25}$$

Thus (5.24) becomes:

$$(1-z)^{d-1}\frac{\partial}{\partial z}\left((1-z)\hat{C}^{[1]}_{\ell}(z)\right) = \hat{C}^{[1]}_{\ell-d+1}(z),$$

multiplying both sides by u^{ℓ} and summing over $\ell \geq d-1$ yields:

$$(1-z)^{d-1}\frac{\partial}{\partial z}\left((1-z)\hat{C}^{[1]}(z,u)\right) = u^{d-1}\hat{C}^{[1]}(z,u),$$

whose solution is the following:

$$\hat{C}^{[1]}(z,u) = \frac{1}{1-z}\cdot F(u)\cdot \exp\left(\frac{u^{d-1}}{(d-1)(1-z)^{d-1}}\right),$$

using the initial condition $\hat{C}^{[1]}(0,u) = u$ gives us

$$F(u) = u\cdot \exp\Big(-\frac{u^{d-1}}{d-1}\Big).$$

Thus we get finally the solution

$$\hat{C}^{[1]}(z,u) = \frac{u}{1-z}\cdot \exp\left(\frac{u^{d-1}}{d-1}\left(\frac{1}{(1-z)^{d-1}}-1\right)\right).$$

Now extracting the coefficients is going as follows:

$$\begin{aligned}
\hat{c}^{[1]}_{n,\ell} &= [z^{n-\ell}u^{\ell}]\hat{C}^{[1]}(z,u)\\
&= [z^{n-\ell}(u^{d-1})^{\frac{\ell-1}{d-1}}]\frac{1}{1-z}\cdot\exp\left(\frac{u^{d-1}}{d-1}\left(\frac{1}{(1-z)^{d-1}}-1\right)\right)\\
&= [z^{n-\ell}]\frac{1}{1-z}\cdot\frac{1}{(d-1)^{\frac{\ell-1}{d-1}}\cdot(\frac{\ell-1}{d-1})!}\cdot\left(\frac{1}{(1-z)^{d-1}-1}\right)^{\frac{\ell-1}{d-1}}\\
&= \frac{1}{(d-1)^{\frac{\ell-1}{d-1}}\cdot(\frac{\ell-1}{d-1})!}\sum_{j=0}^{\frac{\ell-1}{d-1}}\binom{\frac{\ell-1}{d-1}}{j}(-1)^j\binom{n-j(d-1)-1}{\ell-j(d-1)-1}, \quad \text{for } \ell = 1(\mathrm{mod}\ (d-1)).
\end{aligned}$$

The result for the quantity which we are interested in follows easily,

$$\begin{aligned}
a^{[1]}_{n,\ell} &= \frac{\ell!^{(d-1)}}{\ell\cdot\binom{n}{\ell}}\cdot\hat{c}^{[1]}_{n,\ell}\\
&= \frac{\ell!^{(d-1)}}{\ell\binom{n}{\ell}(d-1)^{\frac{\ell-1}{d-1}}\cdot(\frac{\ell-1}{d-1})!}\sum_{j=0}^{\frac{\ell-1}{d-1}}\binom{\frac{\ell-1}{d-1}}{j}(-1)^j\binom{n-j(d-1)-1}{\ell-j(d-1)-1}, \quad \text{for } \ell = 1(\mathrm{mod}\ (d-1)).
\end{aligned}$$

Limit distribution. First, we have from (5.25) that

$$\ell!^{(d-1)} = \frac{(d-1)^{\frac{\ell-1}{d-1}} \cdot (\frac{\ell}{d-1})!}{(\frac{1}{d-1})!}, \quad \text{for } \ell = 1 (\mathrm{mod}\ (d-1)),$$

so that

$$\frac{\ell!^{(d-1)}}{(d-1)^{\frac{\ell-1}{d-1}} \cdot (\frac{\ell-1}{d-1})!} = \frac{(\frac{\ell}{d-1})!}{(\frac{\ell-1}{d-1})! \cdot (\frac{1}{d-1})!} = \frac{(\frac{\ell}{d-1})^{\frac{1}{d-1}}}{(\frac{1}{d-1})!} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{\ell}\Big)\right), \quad \text{as } d \text{ is fixed.}$$

Then we use Stirling's formula (1.2) as usual to do the asymptotic analysis. We have a sum of terms

$$T_j = \frac{1}{\binom{n-1}{\ell-1}} \cdot \binom{\frac{\ell-1}{d-1}}{j} \binom{n-j(d-1)-1}{\ell-j(d-1)-1},$$

so

$$\begin{aligned}\log(T_j) &= \log\left(\frac{(\frac{\ell-1}{d-1})!(n-j(d-1)-1)!(\ell-1)!}{(\frac{\ell-1}{d-1}-j)!(\ell-j(d-1)-1)!(n-1)!}\right) \\ &= -j \cdot \log(d-1) + d \cdot j \cdot \log \ell - j(d-1)\log n - \frac{j^2(d-1)}{2(\ell-1)} + \frac{j^2(d-1)^2}{2n} - \frac{j^2(d-1)^2}{2\ell} \\ &\quad + \mathcal{O}\Big(\frac{j}{\ell}\Big) + \mathcal{O}\Big(\frac{j^3}{\ell^2}\Big) + \mathcal{O}\Big(\frac{1}{\ell}\Big).\end{aligned}$$

After that we recover an asymptotic estimate for T_j:

$$T_j = \frac{1}{j!}\left(\frac{\ell^d}{(d-1)n^{d-1}}\right)^j \cdot \exp\left(-\frac{j^2(d-1)}{2(\ell-1)} + \frac{j^2(d-1)^2}{2n} - \frac{j^2(d-1)^2}{2\ell}\right) \cdot \left(1 + \mathcal{O}\Big(\frac{j}{\ell}\Big) + \mathcal{O}\Big(\frac{j^3}{\ell^2}\Big) + \mathcal{O}\Big(\frac{1}{\ell}\Big)\right).$$

Asymptotically as $n \to \infty$, $\frac{\ell^d}{n^{d-1}} = \mathcal{O}(1)$ and $\ell \gg d$, thus we have

$$T_j = \frac{(-1)^j}{j!}\left(\frac{\ell^d}{(d-1)n^{d-1}}\right)^j \cdot \left(1 + \mathcal{O}\Big(\frac{j^2}{\ell}\Big)\right),$$

and its summation can be approximated as follows:

$$\sum_{j=0}^{\frac{\ell-1}{d-1}} T_j = \sum_{j=0}^{\infty} T_j - \sum_{j=\frac{\ell-1}{d-1}}^{\infty} T_j.$$

Since

$$\sum_{j=\frac{\ell-1}{d-1}}^{\infty} T_j = \mathcal{O}\left(\exp\left(\frac{-(\ell-1)}{2(d-1)} \cdot \log\Big(\frac{\ell-1}{d-1}\Big)\right)\right),$$

$$\sum_{j=0}^{\infty} T_j \sim \exp\left(\frac{-\ell^d}{(d-1)n^{d-1}}\right),$$

we can show the limit distribution for the sequence $a_{n,\ell}^{[1]}$, for $\ell = \mathcal{O}\Big(n^{1-\frac{1}{d}}\Big)$, is asymptotically

$$a_{n,\ell}^{[1]} \sim \frac{1}{(d-1)^{1/(d-1)}(\frac{1}{d-1})!} \cdot \frac{\ell^{\frac{1}{d-1}}}{n} \cdot \exp\left(\frac{-\ell^d}{(d-1)n^{d-1}}\right).$$

Since $a_{n,\ell}^{[1]}$ represents the case when k is 1(mod d) then we can state the results for the size of the hiring set as follows:

$$\mathbb{P}\{h_n = k\} = a_{n,k\cdot\frac{d-1}{d}+\frac{1}{d}}^{[1]} \sim \frac{1}{d^{\frac{1}{d-1}}(\frac{1}{d-1})!}\cdot\frac{k^{\frac{1}{d-1}}}{n}\cdot\exp\left(-\frac{(d-1)^{d-1}}{d^d}\cdot\frac{k^d}{n^{d-1}}\right).$$

If we consider the normalized r.v. $\frac{h_n}{n^{1-\frac{1}{d}}}$, then we have

$$\mathbb{P}\left\{\frac{h_n}{n^{1-\frac{1}{d}}} = \frac{k}{n^{1-\frac{1}{d}}}\right\} \sim \frac{1}{n^{1-\frac{1}{d}}d^{\frac{1}{d-1}}(\frac{1}{d-1})!}\cdot\left(\frac{k}{n^{1-\frac{1}{d}}}\right)^{\frac{1}{d-1}}\cdot\exp\left(-\frac{(d-1)^{d-1}}{d^d}\cdot\left(\frac{k}{n^{1-\frac{1}{d}}}\right)^d\right).$$

Thus Theorem 5.6 follows easily.

5.4 Conclusions

We have seen in this chapter that the framework given by Archibald and Martínez can give useful information of many parameters related to the hiring process when "hiring above the α-quantile" is applied. The introduced theorems in Subsection 5.2.1 give the order of growth of the expectation of many hiring parameters via upper and lower "bounds". We clarify again that those bounds represent pragmatic selection rules but do not correspond to actual rank-based hiring strategies.

On the other hand, an extension of the recursive approach used to analyze "hiring above the median" was very helpful to get the distributional results of the size of the hiring set under hiring above the α-quantile, with $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$, as characterized in Theorem 5.6. As usual, since we have the quantities $a_{n,\ell}^{[i]}$, $i = 1, \ldots, d$, for the later strategy, then the results for other parameters like the index of last hired candidate, the distance between last two hirings, and others are possible, however the computations become more involved.

Moreover, the explicit and distributional results of the *number of selections* for the class of p-percentile rules, with $p = \frac{1}{d}$, $d \in \mathbb{N}$ are also in hand with a suitable extension of the recursive approach.

We think that for other particular cases of "hiring above the α-quantile", as well as the corresponding p-percentile rules, like $\alpha = \frac{2}{3}, \frac{3}{4}, \ldots$; our approach will do the job and we can obtain, at least, the distribution of the size of the hiring set.

For more general case of rational α, i.e., $\alpha = \frac{p}{q}$ where $\gcd(p, q) = 1$, there are a system of q recurrences and finding suitable normalization factors to treat the corresponding generating functions seems challenging. On the other hand this recursive approach will break down in case of irrational α.

Chapter 6

Hiring above the m-th best

6.1 Introduction

This chapter is devoted to study "hiring above the m-th best candidate strategy"; the rank-based hiring strategy introduced originally by Archibald and Martínez in [5]. Since the goal of hiring strategies is to hire many good candidates then it is reasonable that we select the extremes or the records of the sequence. This of course gives us a set of hired candidates with a very good quality but very few candidates (very slow hiring rate), i.e., the size of the hiring set is $\Theta(\log n)$, for a sequence of n candidates. But we can expand the base of selections by hiring any of the best m candidates seen so far in the sequence instead of only the best one.

According to Definition 2.8, our hiring strategy processes the sequence of candidates in two phases. In the initial phase, the first m interviewed candidates are hired regardless of their relative 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 is the *threshold* for this strategy and at any time step n there are m choices for hiring a new candidate which must have one of the ranks (or scores) $\{n, n-1, \ldots, n-m+1\}$.

In the sequel of this chapter, we use the subscript m in the notation of the studied parameters to stress their dependence on the *rigidity* m of this strategy, i.e., $h_{n,m}$, $L_{n,m}$, etc.

For example, let $m = 3$ and we have the following permutation of seven interviewed candidates, $\sigma^{(7)} = \underline{4}\,\underline{6}\,\underline{1}\,\underline{7}\,3\,\underline{5}\,2$. Then processing those candidates using hiring above the m-th best results in hiring set $\mathcal{H}(\sigma^{(7)}) = \{1, 2, 3, 4, 6\}$ where the underlined scores of $\sigma^{(7)}$ form the set of scores of hired candidates $\mathcal{Q}(\sigma^{(7)})$, whereas the ones with scores $\{3, 2\}$ are discarded. Thus we have the number of hired candidates $h_{7,3} = 5$, the gap of last hired candidate $g_{7,3} = 1 - \frac{5}{7} = \frac{2}{7}$, the index of last hired candidate $L_{7,3} = 6$, the distance between the last two hirings $\Delta_{7,3} = 2$, the score of best discarded candidate $M_{7,3} = 3$. If we apply the proposed hiring with replacement technique in Section 3.3, then we have the number of replacements $f_{7,3} = 1$ since the candidate with score 3 *replaces* the candidate with score 1, and $\mathcal{H}_R(\sigma^{(7)}) = \{1, 2, 4, 5, 6\}$ with $\mathcal{Q}_R(\sigma^{(7)}) = \{4, 6, 7, 3, 5\}$. Moreover, a candidate coming after σ_7 gets hired if he has a rank in the set $\{8, 7, 6\}$, whereas he gets discarded otherwise.

As a pragmatic hiring strategy, it holds for this strategy that, for any $n \geq m$, $\mathcal{Q}(\sigma)$ always contains the m best candidates seen so far (and maybe others), as stated in Theorem 2.16. To be more precise, $\mathcal{Q}(\sigma)$ under this strategy can be described as the set of left-to-right $(\leq m)$-maxima; of course,

the particular case $m = 1$ (*hiring above the best strategy*) coincides with the usual notion of records in a permutation (see Section 2.4). The connections between the hiring set (and the corresponding set of scores) under this hiring strategy and the two types of m-records are discussed in more detail in the next subsection.

6.1.1 Records

A good hiring strategy should hire good candidates, in particular those that stand out among the others. This is obviously related to the notion of *records* in a sequence or a permutation.

There is a vast literature on the subject of records in sequences and permutations, and several generalizations, e.g., for d-dimensional data. Here we discuss the two common definitions of m-records found in the literature (see, for instance, the book of Arnold et al. [6]), because of their close connection to $\mathcal{Q}(\sigma)$ under "hiring above the m-th best". We shall slightly adapt the general definition of the two types of records to stress the similarities.

Definition 6.1 *Given a permutation* $\sigma^{(n)} = \left(\sigma_1^{(n)}, \ldots, \sigma_n^{(n)}\right)$*, then we say that* $\sigma_i^{(n)}$ *(the* i*-th element in a permutation* σ *of size* n*) is a* `Type1` m*-record if* $\sigma_i^{(n)}$ *is the* m*-th largest element in* $\left\{\sigma_1^{(n)}, \ldots, \sigma_i^{(n)}\right\}$.

Quite clearly, the union of the sets of `Type1` i-records of σ, for $i = 1, 2, \ldots, m$ represents $\mathcal{Q}(\sigma)$ if we apply "hiring above the m-th best". If $r_{n,i}^{[1]}$ is the number of `Type1` i-records in a random permutation of size n and $h_{n,m}$ denotes the number of hired candidates then

$$h_{n,m} = r_{n,1}^{[1]} + r_{n,2}^{[1]} + \ldots + r_{n,m}^{[1]}.$$

`Type2` m-records are defined similarly as follows.

Definition 6.2 *Given a permutation* $\sigma^{(n)} = \left(\sigma_1^{(n)}, \ldots, \sigma_n^{(n)}\right)$*, then an element* $\sigma_i^{(n)}$ *is a* `Type2` m*-record if there exists* $j \geq i$ *such that* $\sigma_i^{(n)}$ *is the* m*-th largest in* $\left\{\sigma_1^{(n)}, \ldots, \sigma_j^{(n)}\right\}$.

Suppose that $\sigma_i^{(n)}$ is not any of the largest $m-1$ ranks, that is, $\sigma_i^{(n)} \notin \{n, n-1, \ldots, n-m+2\}$. Then if $\sigma_i^{(n)}$ is a `Type2` m-record then "hiring above the m-th best" hires the i-th candidate $\sigma_i^{(n)}$, because $\sigma_i^{(n)}$ ranked the m-th best or larger, with eventual later candidates making the rank of this i-th candidate drop to the m-th largest rank at some moment. And vice-versa: the rank of a hired candidate which is not in $\{n, n-1, \ldots, n-m+2\}$ is a `Type2` m-record. Thus, if we denote by $r_{n,m}^{[2]}$ the number of `Type2` m-records in a random permutation of size n then

$$h_{n,m} = r_{n,m}^{[2]} + m - 1, \tag{6.1}$$

because the $m-1$ candidates with ranks $\{n-m+2, \ldots, n\}$ get hired, but they are not `Type2` m-records. We give an example in Table 6.1, for $m = 1, 2$, and we have a sequence of length eight, to show the relationships explained in this subsection. As mentioned before, `Type2` m-records are often called m-records, for simplicity.

It is also useful to examine these notions of records and of "hiring above the m-th best" from an algorithmic point of view. To carry out the hiring process, we would setup a table T with m entries to contain the m largest ranks seen so far. The first m elements will fill the table, then for each subsequent element, if it is larger than the smallest in T then it enters into the table T and

		i	1	2	3	4	5	6	7	8
		s_i	1	2	2	4	4	4	7	7
		$\sigma^{(n)}$	1	3	2	6	5	4	8	7
$m = 1$	Type1		✓	✓		✓			✓	
	Type2		✓	✓		✓			✓	
	$\mathcal{Q}$		✓	✓		✓			✓	
$m = 2$	Type1				✓		✓			✓
	Type2		✓(2)	✓(4)	✓(3)	✓(7)	✓(5)			✓(8)
	$\mathcal{Q}$		✓	✓	✓	✓	✓		✓	✓

Table 6.1: An illustrative example of the two types of m-records, for $m = 1, 2$, and the corresponding set of scores of hired candidates, $\mathcal{Q}$, of "hiring above the m-th best". s_i represents the initial (relative) rank of the i-th element. $\sigma^{(n)}$ denotes the resulting permutation of eight elements ($n = 8$). For Type2 2-records, we put the number i next to the check mark to denote the "time" when the element $\sigma_i^{(n)}$ becomes a Type2 2-record.

the former smallest element in T is removed from T (we can keep it somewhere else if we need to collect the hired candidates). If the element is smaller than the minimum in T then it is discarded. Then the number of hired candidates $h_{n,m}$ is the number of times that the table T is updated. If we think in similar terms for the determination of m-records, we proceed as above, but an element is an m-record if and only if it occupies the m-th entry in T somewhen along the execution of the algorithm. Thus the number of m-records is the number of times that the m-th entry of T is updated. Once T already contains $m-1$ elements, each addition to T implies that its m-th entry will be updated. Either because the new element is the m-th largest seen so far (a Type1 m-record) or because the previous m-th leaves the table and the previous $(m-1)$-th becomes the m-th largest. Except for the first $m-1$ additions to fill up the first, second, ..., $(m-1)$-th entries in T, all hirings imply a new m-record. This algorithmic formulation (number of updates of the m-th entry) was used by Prodinger [83] in his investigation of m-records in sequences of i.i.d. geometric r.v.'s (as mentioned in Section 2.4).

The sequel of this chapter is organized as follows: Section 6.2 contains our results for "hiring above the m-th best". We give complete proofs of all presented theorems in Section 6.3. The relationship between this hiring strategy and the seating plan $(0, m)$ of the CRP is discussed, together with the results for a new parameter for the mentioned seating plan in Section 6.4. This chapter ends with the conclusions in Section 6.5. The results of this chapter appear in [48, 50].

6.2 Results

Theorem 6.1 *Let $h_{n,m}$ denote the size of the hiring set after n interviews. Then the exact distribution of $h_{n,m}$ is given as follows:*

$$\mathbb{P}\{h_{n,m} = j\} = \begin{cases} [\![n = j]\!], & \text{if } m > n, \\ \frac{m!\,m^{j-m}}{n!} \cdot \begin{bmatrix} n-m+1 \\ j-m+1 \end{bmatrix}, & \text{if } m \le j \le n. \end{cases}$$

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 $n \to \infty$:

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

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:

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

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

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

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

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

Theorem 6.2 *Let $W_{N,m}$ denote the waiting time (# interviews) for the strategy to hire N candidates. Then the exact distribution of $W_{N,m}$ is given as follows:*

$$\mathbb{P}\{W_{N,m} = t\} = \begin{cases} [\![N = t]\!], & \text{if } N \le m, \\ \frac{m!\,m^{N-m}}{t!} \cdot \left[\begin{smallmatrix} t-m \\ N-m \end{smallmatrix}\right], & \text{if } m < N \le t. \end{cases}$$

For $m \le N$ the expectation of $W_{N,m}$ is $\mathbb{E}\{W_{N,m}\} = m \cdot \left(\frac{m}{m-1}\right)^{N-m}$. Asymptotically as $m \to \infty$, $\mathbb{E}\{W_{N,m}\} \sim m \cdot e^{\frac{N}{m}-1}$.

Theorem 6.3 *Let $L_{n,m}$ denote the index of last hired candidate after n interviews. 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:

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

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

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

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

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

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

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

Theorem 6.4 *Let* $\Delta_{n,m}$ *denote the distance between the last two hirings after* n *interviews. Then the exact distribution of* $\Delta_{n,m}$ *is given as follows (for all other values of the parameters the probabilities are zero):*

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

ii) $m = 1 \le n$*:*

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

iii) $2 \le m \le n$*:*

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

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* $n \to \infty$*:*

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

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:*

i) 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*

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

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

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

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

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

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

Theorem 6.5 *Let* $M_{n,m}$ *denote the score of best discarded after* n *interviews. 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*

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

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:*

i) 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*

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

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

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

iii) $n - m \sim \alpha\sqrt{m}$*, 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.,*

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

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

Theorem 6.6 *Let* $f_{n,m}$ *denote the number of replacements done after processing* n *candidates using the mechanism "hiring with replacements". Then, for* $1 \le m \le n$*, the expectation of* $f_{n,m}$ *is given as follows:*

$$\mathbb{E}\{f_{n,m}\} = \frac{m}{2}\Big(H_n^2 - H_n^{(2)} + H_m^2 + H_m^{(2)}\Big) - m H_n H_m.$$

6.3 Analysis

We give here detailed analytical proofs of the theorems in Section 6.2. We show here the derivations of the explicit results characterizing the exact probability distributions of the considered hiring parameters. Then due to the explicit nature of these exact formulas, the asymptotic results follow from them essentially by applying Stirling's formula for the factorials (1.2), together with standard techniques, i.e. the asymptotic expansion of the logarithmic function for small values (1.3).

We have also checked our results for small values of m and n against the exact probabilities, and we have further tested them for large values of m and n by doing some experiments and they match very well [46].

6.3.1 Size of the hiring set

Probability distribution

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 in 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 is given as follows:

$$\begin{aligned} h_{n,m}(v) &= \sum_{\ell \geq 0} \mathbb{P}\{h_{n,m} = \ell\} v^\ell \\ &= v^m \prod_{j=m+1}^{n} \frac{mv + (j-m)}{j} \\ &= v^m \frac{(mv+n-m)! \cdot m!}{(mv)! \cdot n!} \\ &= v^m \frac{\binom{n+m(v-1)}{n}}{\binom{mv}{m}}. \end{aligned} \tag{6.2}$$

We point out that the corresponding probability generating function for `Type2` m-records [6] in permutations already appeared in [71] and later in [83], that is

$$G^{[2]}_{n,m}(v) = \prod_{j=m}^{n} \frac{mv + (j-m)}{j}, \tag{6.3}$$

it is noticed again from (6.2) and (6.3) that initially the hiring set under our strategy contains $(m-1)$ more candidates than the set of `Type2` m-records for the same input sequence. To get an

explicit result for the probabilities and thus the connection to unsigned Stirling numbers of first kind we introduce the generating function:

$$h_m(z,\nu) = \sum_{n\geq m} \binom{n}{m} h_{n,m}(\nu) z^n.$$

A simple computation shows then

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

Using the well-known generating function [57] of the Stirling numbers

$$\sum_{n,k} \begin{bmatrix} n \\ k \end{bmatrix} \frac{z^n}{n!} \nu^k = \frac{1}{(1-z)^\nu},$$

the explicit result for the distribution of $h_{n,m}$ easily follows.

We can also give a "combinatorial" explanation for the nice formula of the distribution of $h_{n,m}$ as follows: let us consider the number of n-permutations that have exactly j hirings; that is

$$\left|\mathcal{P}_{n,j}^{[m]}\right| = (m-1)! \cdot m^{j-m+1} \cdot \begin{bmatrix} n-m+1 \\ j-m+1 \end{bmatrix},$$

if we look at this class of permutations, then we say that the first $(m-1)$ candidates are always hired and there are $(m-1)!$ different arrangements of those $(m-1)$ starting candidates. Then the rest of the j hirings (in $(n-m+1)$ candidates) is counted by the unsigned Stirling numbers of the first kind. Finally if we consider the positions from the m-th coming candidate until the last one in the sequence, then at any position of the $j-m+1$ hiring ones there are m choices to make a hiring there; that gives a factor of m^{j-m+1}.
Formally speaking we consider permutations via their "rank-table", i.e., an n-permutation π is described uniquely via

$$\vec{r} = (r_1, r_2, \ldots, r_n), \quad 1 \leq r_i \leq i,$$

with r_i the relative rank of the element $\pi(i)$ amongst $\{\pi(1), \pi(2), \ldots, \pi(i)\}$. Then this set can be characterized as follows via their corresponding rank-tables:

$$\mathcal{P}_{n,j}^{[m]} = \left\{(r_1, r_2, \ldots, r_n) : 1 \leq r_i \leq i \text{ and } |\{i : r_i \leq m\}| = j\right\}.$$

Now we can give a bijection between the family of permutations $\mathcal{P}_{n,j}^{[m]}$ and all ordered triples of "rank-tables" $\vec{s}$, $\vec{t}$ and $\vec{r'}$, satisfying the following restrictions:

$$\begin{aligned} \vec{s} &= (s_1, s_2, \ldots, s_{m-1}), \text{ with } 1 \leq s_i \leq i, \\ \vec{t} &= (t_1, t_2, \ldots, t_{j-m+1}), \text{ with } 1 \leq t_i \leq m, \\ \vec{r'} &= (r'_1, r'_2, \ldots, r'_{n-m+1}), \text{ with } 1 \leq r'_i \leq i, \end{aligned}$$

i.e., we will show that

$$\mathcal{P}_{n,j}^{[m]} \cong \{(\vec{s}, \vec{t}, \vec{r'})\}.$$

Since the unsigned Stirling numbers of the first kind count permutations with respect to left-to-right minima, i.e.,

$$\begin{bmatrix} n \\ j \end{bmatrix} = \left|\{(r_1, r_2, \ldots, r_n) : \ 1 \le r_i \le i \text{ and } |\{i : r_i = 1\}| = j\}\right|,$$

we have then shown in a bijective way that $\left|\mathcal{P}_{n,j}^{[m]}\right| = (m-1)! \cdot m^{j-m+1} \cdot \begin{bmatrix} n-m+1 \\ j-m+1 \end{bmatrix}$.
But the tables $\vec{s}$, $\vec{t}$ and $\vec{r'}$ are related to the rank-table $\vec{r} = (r_1, r_2, \ldots, r_n)$ of a permutation in $\mathcal{P}_{n,j}^{[m]}$ in a straightforward way:

- $\vec{s} = (s_1, \ldots, s_{m-1})$: it is the rank-table of the first $m-1$ candidates (which are hired anyway):
$$s_i = r_i, \quad 1 \le i \le m-1.$$
- $\vec{t} = (t_1, t_2, \ldots, t_{j-m+1})$: it stores the ranks of the remaining hired candidates. Let us define the set H^* of indices of these $j-m+1$ hired candidates:
$$H^* = \{\ell \ge k : r_\ell \le m\} = \ell_1 < \ell_2 < \cdots < \ell_{j-m+1}.$$
It holds then:
$$t_i = r_{\ell_i}, \quad 1 \le i \le j-m+1.$$
- $\vec{r'} = (r'_1, r'_2, \ldots, r'_{n-m+1})$: it is the "reduced" rank table (of the m-th, $(m+1)$-th, ..., n-th candidate):
$$r'_i = \begin{cases} 1, & \text{if } r_{i+m-1} \le m, \\ r_{i+m-1} - m + 1, & \text{if } r_{i+m-1} > m. \end{cases}$$

So it is easy to see that this indeed gives a "one-to-one"-correspondence of these objects by showing that each permutation of $\mathcal{P}_{n,j}^{[m]}$ yields a different triple $(\vec{s}, \vec{t}, \vec{r'})$ and all such triples indeed occur.

Expectation and variance

The explicit result for $h_m(z, v)$ easily gives, via differentiating r times with respect to v, evaluating at $v = 1$ and extracting coefficients, 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 Theorem [**?**]. The corresponding asymptotic results follow from the asymptotic expansion of the first and second order harmonic numbers, $H_n = \log n + \gamma + \mathcal{O}\left(\frac{1}{n}\right)$ and $H_n^{(2)} = \frac{\pi^2}{6} - n^{-1} + \mathcal{O}\left(\frac{1}{n^2}\right)$.

Limiting distribution

To show the limiting distribution results as $n \to \infty$, we compute the moment generating function (m.g.f.) $\mathbb{E}\left\{e^{h^*_{n,m} s}\right\}$ of a suitably normalized version $h^*_{n,m}$ of $h_{n,m}$ (depending on the region of interest) which converges pointwise for each real s to the m.g.f. $\mathbb{E}\left\{e^{Xs}\right\}$ of a certain r.v. X. Then, an application of the theorem of Curtiss (1.1) shows the weak convergence of $h^*_{n,m}$ to X.

We start with the closed form of the probability generating function $h_{n,m}(v)$ in (6.2), then using Stirling's formula gives us the following useful asymptotic expansion:

$$\log h_{n,m}(v) = m(v-1)(\log n - \log m) + \big(n + m(v-1)\big) \log\left(1 + \frac{m(v-1)}{n}\right) + \mathcal{O}(1-v) + \mathcal{O}\left(\frac{1}{m}\right) \quad (6.4)$$

i) The main region $n - m \gg \sqrt{n}$:
We consider here 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 $\sigma^2 = \sigma^2_{n,m} = m\big(\log n - \log m - 1 + \frac{m}{n}\big)$, yielding thus the m.g.f.

$$\mathbb{E}\left\{e^{h^*_{n,m}s}\right\} = e^{-\frac{\mu}{\sigma}s} \cdot h_{n,m}\left(e^{\frac{s}{\sigma}}\right).$$

Since for m is fixed, the central limit theorem has been shown already in [5], then we consider here only $m \to \infty$. Now we substitute in (6.4), doing some computations, to get the following asymptotic expansion (which holds for any fixed real s):

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

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 yields the stated central limit theorem. We can simplify $\sigma_{n,m}$ by neglecting the term $\frac{m}{n} - 1$ as $n \to \infty$ and $n \gg m$.

ii) $n - m = \mathcal{O}\big(\sqrt{n}\big)$:
We consider the r.v. $\overline{h}_{n,m} = n - h_{n,m}$, then there are two ways to show the convergence to Poisson distribution:
(1) *The method of moments*: the moment generating function of the r.v. $\overline{h}_{n,m}$ is

$$\mathbb{E}\left\{e^{\overline{h}^s_{n,m}}\right\} = e^{ns} \cdot h_{n,m}(e^{-s}).$$

Hence we substitute in (6.4) to get the expansion

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

Since $e^{\lambda(e^s-1)}$ is the m.g.f. of a Poisson distributed r.v. with parameter $\lambda = \frac{(n-m)^2}{2n}$, then the limiting distribution result for $n - m \sim \alpha\sqrt{n}$ follows.

(2) *From the explicit form of the probability distribution*: let us recall it again, for $m \le j \le n$:

$$\mathbb{P}\{h_{n,m} = j\} = \frac{m!\,m^{j-m}}{n!} \cdot \begin{bmatrix} n-m+1 \\ j-m+1 \end{bmatrix}.$$

For simplicity we set $k = n - m$ and $\ell = n - j$, so that

$$\mathbb{P}\{n - h_{n,m} = n - j\} = \mathbb{P}\{\overline{h}_{n,m} = \ell\} = \frac{(n-k)!(n-k)^{k-\ell}}{n!} \cdot \begin{bmatrix} k+1 \\ k-\ell+1 \end{bmatrix}. \tag{6.5}$$

First we have

$$\begin{bmatrix} k+1 \\ k-\ell+1 \end{bmatrix} = \frac{k^{2\ell}}{\ell!2^\ell} \cdot \Big(1 + \mathcal{O}\Big(\frac{1}{k}\Big)\Big), \quad \text{as } k \to \infty \text{ and fixed } \ell, \tag{6.6}$$

where identity (1.6) for the unsigned Stirling numbers of the first kind tells us that

$$\begin{bmatrix} k \\ \ell \end{bmatrix} = [z^\ell]\, z^{\overline{k}} = [z^\ell] z(z+1)\dots(z+k-1).$$

Thus

$$\begin{aligned} \begin{bmatrix} k \\ k-\ell \end{bmatrix} &= [z^\ell] \left(\frac{1}{z}\right)^{\overline{k}} = [z^\ell] \prod_{i=1}^{k} \big(1+(i-1)z\big) \\ &= [z^\ell] \frac{z^k \Gamma\left(k+\frac{1}{z}\right)}{\Gamma\left(\frac{1}{z}\right)}. \end{aligned}$$

Applying Stirling's formula yields:

$$\begin{bmatrix} k \\ k-\ell \end{bmatrix} \sim [z^\ell] e^{-k} (1+kz)^{k+\frac{1}{z}+\frac{1}{2}},$$

then expanding this form using some computer algebra system shows that

$$\begin{aligned} (1+kz)^{k+\frac{1}{z}+\frac{1}{2}} = e^k \Big(1 &+ \frac{k^2}{2}\big(1+\mathcal{O}(k^{-1})\big)z + \frac{k^4}{8}\big(1+\mathcal{O}(k^{-1})\big)z^2 \\ &+ \frac{k^6}{48}\big(1+\mathcal{O}(k^{-1})\big)z^3 + \frac{k^8}{384}\big(1+\mathcal{O}(k^{-1})\big)z^4 + \dots\Big). \end{aligned}$$

The main term is $\frac{z^\ell k^{2\ell}}{c_\ell}$, where $c_\ell = \ell! 2^\ell$. Thus, we get the asymptotic approximation in (6.6). The asymptotic approximation of the other multiplied factor in (6.5) is as follows:

$$\frac{(n-k)!(n-k)^{k-\ell}}{n!} \sim n^{-\ell} e^{\frac{-k^2}{2n}}, \tag{6.7}$$

after applying Stirling's formula as usual. Now, considering (6.6) and (6.7) gives us the asymptotic expansion for (6.5), which is valid for $k = \mathcal{O}\big(\sqrt{n}\,\big)$:

$$\mathbb{P}\{\overline{h}_{n,m} = \ell\} \sim \frac{1}{\ell!}\Big(\frac{k^2}{2n}\Big)^\ell e^{-\frac{k^2}{2n}},$$

that is the probability density function of a Poisson distribution with parameter $\lambda = \frac{k^2}{2n} = \frac{(n-m)^2}{2n}$.

iii) $n - m = o\big(\sqrt{n}\,\big)$:
The strategy is expected to hire many candidates then we consider the r.v. $\tilde{h}_{n,m} = n - h_{n,m}$. It is easy to see that the m.g.f. of $\tilde{h}_{n,m}$ converges to 0, or $\mathbb{P}\{h_{n,m} = n\} = 1 - o(1)$, which shows the stated theorem for this region also.

6.3.2 Waiting time

The result for this parameter follows easily from the explicit form of the distribution of $h_{n,m}$. When we consider the moment that the size of the hiring set is exactly N, we see

$$\mathbb{P}\{W_{N,m} = t\} = \frac{m}{t} \cdot \mathbb{P}\{h_{t-1,m} = N-1\}, \quad \text{for } m < N \le t.$$

Then using Theorem 6.1 we get directly the result. For the expectation, we make use of the following interesting identity given by Kuba and Prodinger [62] for $s, d \in \mathbb{N}$:

$$\sum_{j\geq 1} \frac{\left[\begin{smallmatrix} j \\ d \end{smallmatrix}\right]}{j!\binom{s+j}{j}} = \frac{1}{s^d},$$

together with a useful identity for the unsigned Stirling numbers of the first kind (1.5), thus we have:

$$\begin{aligned}
\mathbb{E}\{W_{N,m}\} &= \sum_{t\geq N} t \cdot \frac{m! m^{N-m}}{t!} \begin{bmatrix} t-m \\ N-m \end{bmatrix} \\
&= m! m^{N-m} \left(\sum_{t\geq N-m} \frac{\left[\begin{smallmatrix} t \\ N-m \end{smallmatrix}\right]}{(t+m-1)!} - \sum_{t\geq N-m-1} \frac{\left[\begin{smallmatrix} t \\ N-m-1 \end{smallmatrix}\right]}{(t+m)!} \right) \\
&= m \cdot \left(\frac{m}{m-1}\right)^{N-m}.
\end{aligned}$$

Asymptotically as m becomes large,

$$\mathbb{E}\{W_{N,m}\} = m \lim_{m\to\infty} \left(1 - \frac{1}{m}\right)^{m(1-\frac{N}{m})} \sim m \cdot e^{\frac{N}{m}-1},$$

which is fully consistent with $\mathbb{E}\{h_{n,m}\}$.

6.3.3 Index of last hired candidate

Trivially, for $n > m$ one gets $\mathbb{P}\{L_{n,m} = n\} = 1$, thus we only have to consider the range $1 \leq m \leq n$. It is immediate from the definition of "hiring above the m-th best" (see also Subsect. 6.3.1) 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 $L_{n,m}$:

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

The results for the expectation and the variance can be obtained easily as follows:

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

$$\begin{aligned}
\mathbb{V}\{L_{n,m}\} &= \frac{1}{\binom{n}{m}} \sum_{j=m}^{n} j^2 \binom{j-1}{m-1} - \left(\frac{m(n+1)}{m+1}\right)^2 \\
&= \frac{m}{(m+2)(m+1)^2} n^2 + \frac{2m^2+m}{(m+2)(m+1)} n + \frac{m^2}{(m+2)(m+1)}.
\end{aligned}$$

The limiting distribution results, for $n \to \infty$, can be obtained by applying Stirling's formula to the exact formula for the probabilities.

i) The main region: $m \to \infty$ but $m = o(n)$ and let $j = n - k$ where $k = o(n)$ then

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

from which one immediately gets that $\frac{m}{n}(n - L_{n,m}) \xrightarrow{(d)} L$, L has density function $f(x) = e^x$.

ii) m is fixed, i.e., $m = \Theta(1)$:

$$\mathbb{P}\{L_{n,m} = j\} = \frac{mj!(n-m)!}{j(j-m)!n!} = \frac{m}{n}\left(\frac{j}{n}\right)^{m-1} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{j}\Big)\right),$$

so that $\frac{L_{n,m}}{n} \xrightarrow{(d)} L$, where L has density function $f(x) = m \cdot x^{m-1}$, $0 < x < 1$, which is Beta distribution with parameters $\alpha = m$ and $\beta = 1$.

iii) $m \sim \alpha n$ with $0 < \alpha < 1$: let $j = n - k$ and $k = o(n)$,

$$\begin{aligned}
\mathbb{P}\{L_{n,m} = n - k\} &= \frac{m(n-m)!(n-k-1)!}{n!(n-k-m)!} \\
&= \frac{m(n-m)^k \left(1 + \mathcal{O}\Big(\frac{1}{n-m}\Big)\right)}{(n-k)n^k \left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right)} \\
&= \frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)^k \cdot \left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right),
\end{aligned}$$

for $k \in \mathbb{N}$ we can say that asymptotically $(n - L_{n,m})$ is geometrically distributed with success probability $p = \alpha$.

iv) $n - m = o(n)$: it is very likely that the strategy recruits almost everybody, so that

$$\mathbb{P}\{L_{n,m} = n\} = \frac{m}{n} = 1 + o\Big(\frac{1}{n}\Big), \quad \text{then } (n - L_{n,m}) \xrightarrow{(d)} 0.$$

6.3.4 Distance between the last two hirings

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:

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = d\} &= \sum_{j=m+d}^{n} \Big[\mathbb{P}\{\text{We hire at position } (j-d)\} \cdot \mathbb{P}\{\text{No hirings from position } (j-d+1) \text{ till } (j-1)\} \\ &\qquad \cdot \mathbb{P}\{\text{We hire at position } j\} \cdot \mathbb{P}\{\text{No hirings from position } (j+1) \text{ till } n\}\Big] \\ &= \sum_{j=m+d}^{n} \frac{m}{j-d} \cdot \prod_{\ell=j-d}^{j-1} \left(1-\frac{m}{\ell}\right) \cdot \frac{m}{j} \cdot \prod_{\ell=j+1}^{n} \left(1-\frac{m}{\ell}\right) = \frac{m}{\binom{n}{m}} \sum_{j=m+d}^{n} \frac{1}{j-m}\binom{j-d-1}{m-1}.\end{aligned} \tag{6.8}$$

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

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = 1\} &= \frac{m}{\binom{n}{m}} \sum_{j=m+1}^{n} \frac{1}{j-m}\binom{j-2}{m-1} + \mathbb{P}\{L_{n,m} = m\} \\ &= \frac{m}{\binom{n}{m}} \sum_{j=m+1}^{n} \frac{1}{j-m}\binom{j-2}{m-1} + \frac{1}{\binom{n}{m}} \\ &= \frac{1}{m-1}\left(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\right),\end{aligned}$$

where the last simplification follows from a summation formula. Finally, for the instance $m = 1$ the formula (6.8) 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 $\mathbb{P}\{\Delta_{n,1} = 0\} = \frac{1}{n}$.
The expectation and variance are computed as follows for $2 \le m \le n$:

$$\begin{aligned}\mathbb{E}\{\Delta_{n,m}\} &= \frac{m}{\binom{n}{m}} \sum_{d=2}^{n-m} d \sum_{j=m+d}^{n} \frac{1}{j-m}\binom{j-d-1}{m-1} + \frac{1}{m-1}\left(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\right) \\ &= \frac{m}{\binom{n}{m}} \sum_{j=m+2}^{n} \frac{1}{j-m} \sum_{d=2}^{j-m} d\binom{j-d-1}{m-1} + \frac{1}{m-1}\left(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\right) \\ &= \frac{m(n+1)}{(m+1)^2} + \frac{1}{(m+1)\binom{n}{m}},\end{aligned}$$

$$\begin{aligned}\mathbb{V}\{\Delta_{n,m}\} &= \frac{m}{\binom{n}{m}} \sum_{d=2}^{n-m} d^2 \sum_{j=m+d}^{n} \frac{1}{j-m}\binom{j-d-1}{m-1} + \frac{1}{m-1}\left(\frac{m^2}{n} - \frac{1}{\binom{n}{m}}\right) - \mathbb{E}\{\Delta_n\}^2 \\ &= \frac{m^4+2m^3+2m^2+2m}{(m+2)^2(m+1)^4} \cdot n^2\left(1+\mathcal{O}\Big(\frac{1}{n}\Big)\right), \quad \text{for } m = o(n).\end{aligned}$$

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 $2 \leq m \leq n$ and $n \to \infty$:

i) The main region $m \to \infty$, but $m = o(n)$: doing some simplifications to the formula (6.8) gives us

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = d\} &= m^2 \sum_{j=d}^{n-m} \frac{1}{(n-m-j+d)} \cdot \frac{(n-j-1)!(n-m)!}{(n-j-m)!n!} \\ &= m^2 \sum_{j=d}^{n-m} \frac{1}{(n-m-j+d)} \cdot \frac{1}{n} e^{\frac{-jm}{n}} \cdot \left(1 + \mathcal{O}\Big(\frac{j^2 m}{n^2}\Big) + \mathcal{O}\Big(\frac{jm^2}{n^2}\Big)\right) \\ &\sim \frac{m}{n} \cdot \int_{\frac{dm}{n}}^{\infty} e^{-t} dt = \frac{m}{n} \cdot e^{\frac{-md}{n}},\end{aligned}$$

where the main contribution for this local approximation is for $d = \mathcal{O}\left(\frac{n}{m}\right)$. Thus the random variable $\frac{m}{n}\Delta_{n,m} \xrightarrow{(d)} X$, where X has density function $f(x) = e^x$, $x > 0$.

ii) m is fixed: starting with formula (6.8),

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = d\} &= \frac{m \cdot m!}{n^m \cdot \left(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\right)} \sum_{j=m+d}^{n} \frac{1}{j-m} \cdot \frac{(j-d)^{m-1}}{(m-1)!} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{j-d}\Big)\right) \\ &\sim \frac{m^2}{n} \cdot \int_{\frac{d}{n}}^{1} \frac{1}{t} \cdot \Big(t - \frac{d}{n}\Big)^{m-1} dt,\end{aligned}$$

which implies that $\frac{\Delta_{n,m}}{n} \xrightarrow{(d)} X$, where X has the following density function,

$$f_m(x) = m^2 \int_x^1 \frac{1}{t}(t-x)^{m-1} dt, \quad 0 < x < 1.$$

We can express $f_m(x)$ also in the more explicit form stated in the theorem as follows:

$$\begin{aligned}f_m(x) &= m^2 \int_1^x \frac{1}{t} \sum_{l=0}^{m-1} \binom{m-1}{l} (-1)^l x^l t^{m-1-l} dt \\ &= m^2 \left(\sum_{l=0}^{m-2} \binom{m-1}{l} \frac{(-1)^l x^l}{m-1-l} (1 - x^{m-1-l}) + (-1)^m x^{m-1} \log(x) \right) \\ &= m^2 \left(\sum_{j=0}^{m-2} \frac{(-1)^l x^l}{m-1-l} \sum_{l=0}^{j} \binom{m-2-l}{m-2-j} (-1)^j x^j + (-1)^m x^{m-1} \log(x) \right) \\ &= m^2 \left(\sum_{j=0}^{m-2} \frac{x^j (-1)^j}{m-1-l} \binom{m-1}{j} + \sum_{l=0}^{m-2} \frac{(-1)^{m-1} x^{m-1}}{m-1-l} + (-1)^m x^{m-1} \log(x) \right),\end{aligned}$$

A couple of simple additional computations yield the result.

iii) $m \sim \alpha n$:

$$\begin{aligned}\mathbb{P}\{\Delta_{n,m} = d\} &= m^2 \sum_{j=d}^{n-m} \frac{1}{(n-m-j+d)} \cdot \frac{(n-j-1)!(n-m)!}{(n-j-m)!n!} \\ &= m^2 \sum_{j \geq d} \frac{1}{n-m} \frac{(n-m)^j}{n^{j+1}} \cdot \left(1 + \mathcal{O}\Big(\frac{1}{n-m}\Big)\right), \quad \text{for fixed } j, \\ &\sim \frac{m}{n}\Big(1 - \frac{m}{n}\Big)^{d-1},\end{aligned}$$

which says that asymptotically $(\Delta_{n,m} - 1)$ is geometrically distributed with probability of success $p = \alpha$.

iv) $n - m = o(n)$: we expect that the strategy will recruit many candidates, then $\Delta_{n,m}$ takes small values, so that for $d = 1$ we have

$$\mathbb{P}\{\Delta_{n,m} = 1\} = \frac{m^2}{(m-1)n} + \mathcal{O}\Big(\frac{1}{n}\Big) = 1 + o(1),$$

that is enough to show that $\Delta_{n,m}$ converges to 1 as $n - m = o(n)$.

6.3.5 Score of best discarded candidate

To show the explicit result for the exact distribution of $M_{n,m}$ on the case $1 \leq m \leq n$ and $1 \leq b \leq n - m$, we proceed as we did in Subsection 4.3.8. We have to consider an auxiliary quantity, namely the probability $a_{n,m,j}$, with $0 \leq j \leq n - m$, that *all of the* $m + j$ *highest ranked candidates are hired (and maybe others)*. Thus the probability that the best discarded candidate has rank $1 \leq b \leq n - m$ is simply given by the difference between the probability that all candidates with a rank higher than b are recruited and the probability that all candidates with a rank higher than $b - 1$ are recruited, i.e.,

$$\mathbb{P}\{M_{n,m} = b\} = a_{n,m,n-m-b} - a_{n,m,n+1-m-b}. \tag{6.9}$$

The corresponding recurrence of the sequence $a_{n,m,j}$ can be stated as follows for $1 \leq j \leq n - m$:

$$a_{n,m,j} = \frac{m}{n} \cdot a_{n-1,m,j-1} + \left(1 - \frac{j+m}{n}\right) \cdot a_{n-1,m,j}, \tag{6.10}$$

with $a_{n,m,0} = 1$ where the best m candidates are always hired as we know.
To solve this recurrence let us introduce the normalization:

$$b_{n,m,j} = \frac{n!}{m!(n-m-j)!} \cdot a_{n,m,j}.$$

Hence the recurrence equation (6.10) becomes,

$$b_{n,m,j} = m \cdot b_{n-1,m,j-1} + \cdot b_{n-1,m,j}, \tag{6.11}$$

Now we can introduce the generating function

$$B(z,u,v) = \sum_{n \geq 1} \sum_{1 \leq m \leq n} \sum_{1 \leq j \leq n-m} b_{n,m,j} z^n u^m v^j.$$

Applying the generating function to (6.11) yields the following PDE:

$$(1-z)B(z,u,v) = zuvB_u(z,u,v) + \frac{z^2uv}{(1-z(1+u))^2}.$$

A simple trick to extract the required coefficients without explicitly solving the PDE is to deal with $[u^m]B(z,u,v) = b_m(z,v)$, so that the PDE is reduced to

$$(1-z)b_m(z,v) = mzvb_m(z,v) + m\frac{z^2v}{(1-z)^2}\left(\frac{z}{1-z}\right)^{m-1},$$

which gives directly

$$b_m(z,v) = \frac{mvz^{m+1}}{(1-z-mzv)(1+z)^{m+1}}.$$

Now it is easy to extract the coefficients to get the result for our quantity:

$$\begin{aligned} a_{n,m,j} &= \frac{m!(n-m-j)!}{n!}\cdot b_{n,m,j} \\ &= \frac{m!(n-m-j)!}{n!}\cdot m^j\binom{n}{m+j} \\ &= \frac{m!m^j}{(m+j)!}, \quad \text{for } 0 \le j \le n-m \end{aligned} \tag{6.12}$$

If we plus this answer in (6.9) then we get the result stated in the theorem. We can explain the result obtained in (6.12) as follows: 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

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

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 whole distribution of $M_{n,m}$.

The expectation of $M_{n,m}$ is computed as follows:

$$\begin{aligned}\mathbb{E}\{M_{n,m}\} &= \sum_{b=1}^{n-m} b \cdot \frac{m!(n-m-b+1)m^{n-m-b}}{(n-b+1)!} \\ &= n-m+1-m!m^{-m}\sum_{j=m}^{n}\frac{m^j}{j!},\end{aligned}$$

and the term $S_m = m!m^{-m}\sum_{j=m}^{n}\frac{m^j}{j!}$ is exactly the following function introduced by Knuth in [56]:

$$\begin{aligned}R(m) &= \sum_{j\geq 0}\frac{m!m^j}{(m+j)!} \\ &= \sqrt{\frac{\pi m}{2}}+\frac{1}{3}+\frac{1}{12}\sqrt{\frac{\pi}{2m}}+\frac{4}{135m}+\frac{1}{288}\sqrt{\frac{\pi}{2m^3}}+\mathcal{O}\left(\frac{1}{m^2}\right).\end{aligned}$$

Thus asymptotically as $n\to\infty$ and uniformly for $1\leq m\leq n$,

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

The variance of $M_{n,m}$ can be obtained via similar computations as follows

$$\begin{aligned}\mathbb{V}\{M_{n,m}\} &= \sum_{b=1}^{n-m} b^2\cdot\frac{m!(n-m-b+1)m^{n-m-b}}{(n-b+1)!} - \mathbb{E}\{M_{n,m}\}^2 \\ &= 2m - S_m^2 + S_m \\ &= \left(2-\frac{\pi}{2}\right)\cdot m + \frac{\sqrt{2\pi m}}{6}+\mathcal{O}(1), \quad \text{as } n\to\infty.\end{aligned}$$

The limiting behaviour of $M_{n,m}$ is characterized depending on the size relation between n and m as previous parameters. For fixed m it is trivially derived from the exact formula of the distribution. We give here the details for other regions of interest:

i) The main region $m\to\infty$, but $n-m\gg\sqrt{m}$: we get the following local expansion

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

which immediately entails that $\frac{n-m-M_{n,m}}{\sqrt{m}}\xrightarrow{(d)}Y$, where Y has the density function $f(y)=ye^{-\frac{y^2}{2}}$, $y>0$, thus Y is Rayleigh distributed r.v. with parameter 1. We notice that the asymptotic result for the expectation follows also from the local expansion obtained here.

ii) $n-m=o\big(\sqrt{m}\,\big)$: that means that m is very close to n so that almost all candidates are hired and $M_{n,m}$ tends to 0,

$$\mathbb{P}\{M_{n,m}=0\} = \frac{m!m^{n-m}}{n!}\sim 1+o(1).$$

iii) $n-m \sim \alpha\sqrt{m}$, $\alpha > 0$: we consider the r.v. $\tilde{M}_{n,m} = n-m-M_{n,m}$ and compute the distribution function $\mathbb{P}\{\tilde{M}_{n,m} \leq \ell\}$. For $0 \leq \ell \leq n-m-1$ we get

$$\begin{aligned}\mathbb{P}\{\tilde{M}_{n,m} \leq \ell\} &= \sum_{j=0}^{\ell} \mathbb{P}\{\tilde{M}_{n,m} = j\} \\ &= \sum_{j=0}^{\ell} \frac{(j+1)m!m^j}{(m+j+1)!} \\ &= 1 - \frac{m!m^{\ell+1}}{(m+\ell+1)!},\end{aligned}$$

since the sum telescopes. Together with $\mathbb{P}\{\tilde{M}_{n,m} \leq n-m\} = 1$ we get

$$\mathbb{P}\{\tilde{M}_{n,m} \leq \ell\} = \begin{cases} 1 - \frac{m!m^{\ell+1}}{(m+\ell+1)!}, & 0 \leq \ell \leq n-m-1, \\ 1, & \ell \geq n-m. \end{cases}$$

We consider now the region $n-m \sim \alpha\sqrt{m}$, $\alpha > 0$, we set $\ell = x\sqrt{m}$.
Since

$$\frac{m!m^{\ell+1}}{(m+\ell+1)!} = e^{-\frac{\ell^2}{2m}} \cdot \left(1 + \mathcal{O}\Big(\frac{\ell}{m}\Big) + \mathcal{O}\Big(\frac{\ell^3}{m^2}\Big)\right),$$

we get

$$\mathbb{P}\left\{\frac{\tilde{M}_{n,m}}{\sqrt{m}} \leq x\right\} \to 1 - e^{-\frac{x^2}{2}}, \quad \text{for } 0 \leq x < \alpha.$$

Moreover, it holds $\mathbb{P}\left\{\frac{\tilde{M}_{n,m}}{\sqrt{m}} \leq x\right\} \to 1$, for $x \geq \alpha$. This shows that

$$\frac{\tilde{M}_{n,m}}{\sqrt{m}} \xrightarrow{(d)} Y_\alpha,$$

where the r.v. Y_α has the cumulative density function

$$F_\alpha(x) = \begin{cases} 1 - e^{-\frac{x^2}{2}}, & 0 \leq x < \alpha, \\ 1, & x \geq \alpha. \end{cases}$$

Note that $F(x) = 1 - e^{-\frac{x^2}{2}}$, $x \geq 0$ is the cumulative function of a Rayleigh(1) distributed r.v. Thus $F_\alpha(x)$ is the distribution function of $\min\big(\text{Rayleigh}(1), \alpha\big)$ as stated in the theorem.

6.3.6 Number of replacements

Let us consider the indicator r.v. $Y_{j,m}$ which takes the value 1 if a replacement occurs at step j (after receiving j candidates) and 0 otherwise. Since the number of replacements $f_{n,m}$ depends on the size of the hiring set $h_{n,m}$ but the contrary is not true (where the size of the hiring set is not affected by using the replacement mechanism), then the conditional probability of $Y_{j,m}$ is given as follows:

$$\mathbb{P}\{Y_{j,m} = 1 | h_{j-1,m} = k\} = \begin{cases} 0, & \text{if } j \leq m, \\ \frac{k-m}{j}, & \text{if } j > m \text{ and } k \leq j-1, \end{cases}$$

where we have to condition on the size of the hiring set in the previous step, $j-1$. Let the size of the hiring set be k after $j-1$ interviews. There are m possible scores which would be directly hired at step j and those candidates who are worse than the m-th best hired candidate at step $j-1$ **AND** better than the worst hired one, get hired substituting the worst hired. So that there are $k-(m-1)-1 = k-m$ possibilities at step j to replace the worst candidate in the hiring set with a better one.

Since we have the exact distribution of $h_{n,m}$ (Theorem 6.1) then we can proceed in the computations as follows:

$$\begin{aligned}\mathbb{P}\{Y_{j,m}=1\} &= \sum_{k=m}^{j-1} \mathbb{P}\{Y_{j,m}=1 | h_{j-1,m}=k\}\cdot \mathbb{P}\{h_{j-1,m}=k\}\\ &= \frac{(m-1)!}{j!}\sum_{t=1}^{j-m}(t-1)\begin{bmatrix} j-m \\ t \end{bmatrix} m^t\\ &= \frac{m}{j}\big(H_{j-1}-H_m\big).\end{aligned}$$

Now we can compute the expectation easily:

$$\begin{aligned}\mathbb{E}\{f_{n,m}\} &= \sum_{j=m+1}^{n}\mathbb{P}\{Y_{j,m}=1\}\\ &= \frac{m}{2}\left(H_n^2 - H_n^{(2)} + H_m^2 + H_m^{(2)}\right) - mH_nH_m\\ &= \frac{m}{2}\cdot\log^2\left(\frac{n}{m}\right) + \mathcal{O}(m), \quad \text{uniformly for } m\le n.\end{aligned}$$

6.4 The seating plan $(0, m)$

We have already discussed, in Subsection 4.4.2, the connections between the hiring process when "hiring above the median" is used to process the sequence of candidates, and the Chinese Restaurant Process (CRP) (Section 2.5) when the seating plan $(\frac{1}{2}, 1)$ is applied for a bunch of customers.

We go here similarly, where we consider the quantity *number of occupied tables after receiving* n *customers*, call it K_n. Thus, as pointed out before, both K_n and *the size of the hiring set* under any natural hiring strategy represent the same Markov chain with increments in $\{0,1\}$ and inhomogeneous transition probabilities.

In particular, let m be positive integer and $1 \le m \le n$ then the seating plan $(0, m)$ processes the sequence of customers exactly like "hiring above the m-th best" during the selection phase (review Section 6.1). The major difference is that the initial phase (where candidates get hired with probability 1) for our hiring strategy which takes $(m-1)$ time intervals more than the seating plan $(0, m)$. Let $K_n^{(0,m)}$ denote the number of occupied tables in the restaurant if the seating plan $(0, m)$ is applied for n customers. Then the following table shows the probabilities of increment for $K_n^{(0,m)}$ and $h_{n,m}$:

n	1	2	3	...	m	$m+1$	$m+2$	...
$K_n^{(0,m)}$	1	$\frac{m}{m+1}$	$\frac{m}{m+2}$	...	$\frac{m}{2m-1}$	$\frac{m}{2m}$	$\frac{m}{2m+1}$	...
$h_{n,m}$	1	1	1	...	1	$\frac{m}{m+1}$	$\frac{m}{m+2}$	...

So that for $1 \leq m \leq n$ we can state the following relationship:

$$\mathbb{P}\{h_{n,m} = k\} = \mathbb{P}\left\{K_{n-m+1}^{(0,m)} = k - m + 1\right\}. \tag{6.13}$$

A simple derivation, after specializing $\theta = m$ and $\alpha = 0$ in the general formulas (2.2) and (2.3), shows the following theorem for the distribution and expectation of $K_n^{(0,m)}$:

Theorem 6.7 (Pitman [77]) *For the seating plan $(0, m)$, let $K_n^{(0,m)}$ denote the number of occupied tables after n customers have arrived in the restaurant. Then the exact probability distribution of $K_n^{(0,m)}$ is given as follows:*

$$\mathbb{P}\left\{K_n^{(0,m)} = k\right\} = \frac{m^k \Gamma(m)}{\Gamma(n+m)} \begin{bmatrix} n \\ k \end{bmatrix}, \tag{6.14}$$

and the expectation is $\mathbb{E}\left\{K_n^{(0,m)}\right\} = m\big(\log(n+m-1) - \log(m-1)\big) + \mathcal{O}(1)$.

This result of the expectation coincides with Theorem 6.1 according to (6.13). As a consequence, for the special case $m = 1$: $h_{n,1}$ for *hiring above the best strategy* and $K_n^{(0,1)}$ for the seating plan $(0, 1)$ are identical.

Moreover, some hiring parameters studied here for "hiring above the m-th best" can make sense also for CRP under the seating plan $(0, m)$. We mean that results for new quantities associated to seating plan $(0, m)$ like the *time of last occupied table*, $L_n^{(0,m)}$, the *time between opening the last two occupied tables*, $\Delta_n^{(0,m)}$, and the *waiting time until N tables are occupied* in the restaurant, $T_N^{(0,m)}$, are in hand, depending on the corresponding introduced results in Theorems 6.3, 6.4 and 6.2, respectively.

Thus, the following relationships (similar to (6.13)) hold for $1 \leq m \leq n$:

$$\mathbb{P}\left\{L_n^{(0,m)} = j\right\} = \mathbb{P}\{L_{n+m-1,m} = j + m - 1\}, \tag{6.15}$$

$$\mathbb{P}\left\{\Delta_n^{(0,m)} = d\right\} = \mathbb{P}\{\Delta_{n+m-1,m} = d + m - 1\}, \tag{6.16}$$

$$\mathbb{P}\left\{T_N^{(0,m)} = t\right\} = \mathbb{P}\{W_{N+m-1,m} = t + m - 1\}, \quad m \leq N \leq t. \tag{6.17}$$

For instance, we mention here only the results for the *waiting time*.

Theorem 6.8 (CRP) *For the seating plan $(0, m)$, let $T_N^{(0,m)}$ denote the waiting time until N tables are occupied in the restaurant. Then the distribution of $T_N^{(0,m)}$ is given as follow*

$$\mathbb{P}\left\{T_N^{(0,m)} = t\right\} = \frac{m^N \Gamma(m)}{\Gamma(t+m)} \begin{bmatrix} t-1 \\ N-1 \end{bmatrix},$$

and the expectation is $\mathbb{E}\left\{T_N^{(0,m)}\right\} = m\left(\frac{m}{m-1}\right)^{N-1} - m + 1$.

A small check shows that the result for $\mathbb{E}\left\{T_N^{(0,m)}\right\}$ matches exactly the results for $\mathbb{E}\left\{K_n^{(0,m)}\right\}$ in Theorem 6.7 as expected. The proof of the last theorem is omitted since it is similar to that one of Theorem 6.2 in Subsection 6.3.2.

6.5 Conclusions

We have presented various theorems, in Section 6.2 that describe the properties of the hiring process when applying "hiring above the m-th best candidate". These results provide a very detailed picture of this natural hiring strategy.

The connections between the strategy and the two types of m-records have been studied in Subsection 6.1.1, in particular, showing that the hired candidates are those $\texttt{Type2}$ m-records in the sequence together with the best $m-1$ ones.
Moreover, the results for new statistics related to m-records can be deduced directly from the results of the corresponding hiring parameters. As already explained in Subsection 6.1.1, after the first $m-1$ elements in the sequence, whenever we hire a new candidate, a new m-record is encountered. Then, the *index of last* m*-record*, that is the "time" when the last m-record has been encountered, is exactly $L_{n,m}$, Theorem 6.3.
Also, the *distance between the last two* m*-records*, i.e., the number of elements (which are non m-records) between the last two m-records in the sequence, is exactly $\Delta_{n,m}$, Theorem 6.4.
Finally, the *last* m*-record*, also the *largest* m*-record*, is, trivially, that element with rank $n-m+1$ in a sequence of n elements. Studying the *number of replacements* makes little sense in the context of m-records.

We have also discussed the relationship between this hiring strategy and the seating plan $(0, m)$ of the CRP, showing that both are *equivalent* (after a time shift). Moreover, new interesting quantities in the context of the CRP have been introduced, where the results for the seating plan $(0, m)$ related to the *index of last open table* and the *time between opening the last two occupied tables* can be easily obtained from (6.15) and (6.16) respectively. The explicit results for the *waiting time* are also given in Theorem 6.8.

It is obvious from Theorem 2.19 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 last hiring satisfying $\frac{L_{n,m}}{n} < 1$ (Theorem 6.3). Also Theorem 6.2 for the waiting time $W_{N,m}$ shows that it is expected for the strategy to take an exponential number of interviews to hire N candidates.

Chapter 7

Applications to data streaming algorithms

7.1 Introduction

The sequential selection processes have proved very useful in diverse fields. For example, the Chinese restaurant process (Section 2.5) was the main tool in a model-based Bayesian approach used for clustering microarray gene expression data [84], and the asymptotic results of some Polya's urn models were used to obtain an estimation of the Computer memory requirements of the 2-3 trees, a well-known Computer data structure for storage organization [8]. We show in this chapter a new application of the hiring problem in the analysis of data streaming algorithms.

The problem addressed here is the so-called "cardinality estimation problem". We are given a very large multiset $S = (s_1, s_2, \ldots)$ which has a total number of elements $N = |S|$ and may contain repetitions. Then we are interested in extracting the number of distinct elements n that represents the *cardinality* of the underlying set of S. Data streaming algorithms typically require processing a huge data set *sequentially*, very quickly in a single pass (or few passes at most) over all elements, using a limited memory and giving answers of the queries within a small percentage of error is sufficient. Due to these restrictions, a deterministic solution (imagine that N is far beyond the RAM capacity) becomes rather too expensive, and does not meet the requirements. The idea of designing probabilistic algorithms to *estimate* the cardinality of a huge data set become then a reasonable and feasible alternative.

This problem is one of the first to be framed in terms of streaming algorithms. It has arisen in the early 1980s in the database community when IBM's researchers were trying to optimize some intermediate algorithmic operations on data bases. After the pioneering work of Flajolet and Martin [36] in 1985, the problem has received a lot of attention in many other fields like data mining and network security. For example, estimating the distinct number of flows (sequence of packets identified by a source address and a destination address) is of interest and has many applications in network monitoring and network security (see [27, 39]).

7.1.1 Prior work on cardinality estimation problem

Since the first algorithm solving the problem, PROBABILISTIC COUNTING by Flajolet and Martin, new proposals have appeared bearing new algorithms that improve the efficiency or introduce new point of views to attack the problem. Flajolet himself payed a special interest to this particular problem and has visited it in a score of publications, i.e., [25, 31, 32, 33, 34]. Almost all available approximate counting algorithms apply a hash function to every element in the sequence S, in order to get rid of repetitions (since any distinct element and its copies will be hashed into the same hash value). Furthermore, the hashed values can be seen as uniformly distributed random variables (r.v.'s) in $(0, 1)$. Then there are two main points of view for handling the hash values (see [34, 66]):

- **Bit-pattern model**: the hash values are considered as binary strings. Then observing the longest run of zeros in the hash values is used to set an estimator. Techniques using this principle were investigated by Flajolet in several papers, i.e., [25, 34, 36].
- **Order statistic model**: here, the hash values are considered as real numbers in $(0, 1)$. Then the k-th order statistic of the set, i.e., the k-th smallest value seen in the sequence of data is used to set an estimator. Cardinality estimators proposed by Bar-Yossef et al. [9], Giroire [44], Lumbroso [66] and others are build upon this idea.

In this context, we have to prove that the proposed estimator is unbiased or at least asymptotically unbiased (as $n \to \infty$): its expectation (asymptotically) be n. Next, we need to characterize the standard error of the estimator; that is the common accuracy measure. Providing the limiting distribution of the estimator completes the picture but this is often too complicated and only few limit distributions are known for such algorithms; one of them is given by Lumbroso in [66].

From the practical point of view, the best obtained algorithm is the HYPERLOGLOG [34] by Flajolet et al.: it is very fast, simple to implement and it has an accuracy $1.03/\sqrt{m}$ using m words of 5 bytes to estimate cardinalities up to 2^{40}.

Notation. In Chapter 6 we were using the letter m to refer to the strategy "hiring above the m-th best". Here, we are using k instead of m, to refer to the mentioned strategy, because when speaking about data streaming algorithms we find that m is often used to denote the number of substreams m in the so-called "stochastic averaging" technique. We have seen in Subsection 6.1.1 that there are two types of k-records but when mentioning "k-records" here alone we mean `Type2` k-records.

7.1.2 Data streams as random permutations

Here we introduce a different technique to process the hash values that represent the input data stream. If we take into consideration only the hash values of the first occurrence of each distinct element then these constitute a *random permutation* as long as we assume that the (distinct) hash values are i.i.d. random numbers in $(0, 1)$. Hence considering the input data stream as a random permutation opens the door to make use of the properties of random permutations. It is easy to notice that k-*records* are insensitive to repetitions in the input sequence. Hence, we can adapt the explicit distributional results for the hiring parameter *number of hired candidates* to design our first unbiased cardinality estimator, which we call RECORDINALITY. The mathematical computations are very simple and elegant. The probability distribution of RECORDINALITY and its limit are also

easy to compute.

To implement the strategy "hiring above the k-th best", the data sequence is processed in a simple way: it uses only k memory units plus one counter to report the number of k-records, requires few operations per element and of course needs only one single pass over all elements. Hiring above the k-th best considers only the relative ranks of the hash values (not the values themselves) as explained in Chapter 6. Thus, we argue that RECORDINALITY is a practical estimator that can be implemented without hash functions, but rather works assuming that the data streams satisfies the random-order model.

We introduce another estimator, called DISCARDINALITY, based on the hiring parameter *best discarded candidate* that is the largest non k-record after processing the input data sequence using "hiring above the k-th best". From a practical point of view, implementing DISCARDINALITY requires more memory than RECORDINALITY, takes more processing time for the same input stream, and using hash functions is essential.

There were previous proposals to use records to estimate the number of distinct elements in a sample by Moreno-Rebollo et al. [70] in 1996. Later in 2000, Moreno-Rebollo et al. [71] revisited the problem and presented cardinality estimators based on k-records (both types of k-records). In particular, they obtained similar results to our estimator RECORDINALITY using a purely probabilistic approach. The work of Moreno-Rebollo et al. [70, 71] was not cast in the framework of data stream analysis nor did it touch the most algorithmic aspects. Thus, it went largely unnoticed in the data stream community. In particular, we become a ware of [70, 71] after [47] was published. The main results of Moreno-Rebollo et al. and ours are equivalent (thought different techniques were used); in other aspects, they complement and give alternative viewpoints.

The sequel of this chapter is organized as follows: we review the main results of Moreno-Rebollo et al. in Section 7.2. Section 7.3 introduces our main results, analysis and experimental work for RECORDINALITY. The extensions and discussion of RECORDINALITY are presented in Section 7.4. The second estimator DISCARDINALITY is introduced in Section 7.5, where we give the main results, analysis and experimental work. In Section 7.6, we discuss using the largest non k-record for another problem: similarity index estimation. The chapter ends with the conclusions of the presented work and a discussion about the preliminary ideas left for future.

7.2 Related work on the random-order model

In their first work [70], Moreno-Rebollo, Blázquez, Chamorro and Acosta have considered the *number of lower records* (i.e., left-to-right minima), or just *records* since upper records have identical distribution as lower records, in a random sample of unknonw size n from a continuous distribution, to estimate n. They introduced two unbiased estimators as characterized in the following theorems.

Theorem 7.1 (Moreno-Rebollo et al., 1996) *Let r_1 be the observed number of records (1-records) in a sample with unknown size n, then*

$$\mathcal{T}_1 = 2^{r_1} - 1$$

is an unbiased estimator of n *(i.e.,* $\mathbb{E}\{\mathcal{T}\} = n$*) and*

$$\mathbb{V}\{\mathcal{T}_1\} = \frac{(n+3)(n+2)(n+1)}{6} - (n+1)^2.$$

Using the probability generating function of r_1 (that is, $h_{n,1}(v)$ in (6.2)) directly proves the unbiasedness of $\mathcal{T}_1$.

Theorem 7.2 (Moreno-Rebollo et al., 1996) *Let* r_1 *be the observed number of records in a sample with unknown size* n*, then the maximum likelihood estimator (m.l.e.) of* n *is given by*

$$\mathcal{T}^{(\text{m.l.e.})} = \min\{n : \texttt{Mo}(N_n) = r_1\},$$

where N_n is the r.v. that represents the number of records of a sample of size n, and `Mo` is the mode of N_n. Moreno-Rebollo, Chamorro, Blázquez and Gómez have proposed also estimators suggested by the method of moments approach, that was the motivation of their work in [71]. Recall Definitions 6.1 and 6.2 for `Type1` and `Type2` k-records, then the following theorems show the results of the estimators given in [71]:

Theorem 7.3 (Moreno-Rebollo et al., 2000) *Let* $r_k^{[1]}$ *denote the observed number of* `Type1` k*-records in a sample with unknown size* n*, then*

$$\mathcal{T}_{k,\lambda}^{(1)} = k2^{r_k^{[1]}} - 1 + \lambda(1-k)^{r_k^{[1]}}, \quad \lambda \in \mathbb{R},$$

is an unbiased estimator of n *for any* λ*, and*

$$\mathbb{V}\{\mathcal{T}_{k,\lambda}^{(1)}\} = k^2 G_{n,k}^{[1]}(4) - (n+1)^2 + \lambda^2 G_{n,k}^{[1]}((1-k)^2) + 2\lambda k G_{n,k}^{[1]}(2(1-k)),$$

where $G_{n,k}^{[1]}(k)$ *is the probability generating function of* `Type1` k*-records.*

Theorem 7.4 (Moreno-Rebollo et al., 2000) *Let* $r_k^{[2]}$ *denote the observed number of* `Type2` k*-records in a sample with unknown size* n*, then the unique unbiased estimator of* n*, based on* $r_k^{[2]}$ *is*

$$\mathcal{T}_k^{(2)} = k\left(\frac{k+1}{k}\right)^{r_k^{[2]}} - 1,$$

and

$$\mathbb{V}\{\mathcal{T}_k^{(2)}\} = k^2 G_{n,k}^{[2]}\left(\left(\frac{r+1}{r}\right)^2\right) - (n+1)^2,$$

where $G_{n,k}^{[2]}(k)$ *is the probability generating function of* `Type2` k*-records (given in* (6.3)*).*

Moreover, Cramer [21] has studied the asymptotic properties of those estimators; in particular he established bounds for the m.l.e. given in Theorem 7.2.

7.3 RECORDINALITY: a cardinality estimator based on k-records

In the hiring problem, we assume that we can rank all candidates from best to worst without ties which leads to the random permutation model—as mentioned in Subsection 3.1. Then, the considered k-records (refers always to `Type2` k-records) over permutations are strict. According to Definition6.2, even if the sequence of data contains repetitions then only the first occurrence of a value may be considered a k-record. Thus the statistic *number of* k*-records* is insensitive to repetitions while other statistics like number of descents and number of inversions are affected by repetitions. Other hiring parameters like index of last hired candidate and distance between the last two hirings cannot be used to design a useful estimator because they are also sensitive to repetitions.

Moreover, as already mentioned before, we use a hash function that map every element in the sequence into a *random* value in $(0, 1)$; then replications of the same element are mapped into same hash value. This enables us to consider the data stream as a *random permutation*[1].

We give a pseudo-code for RECORDINALITY and the main theorems in Subsection 7.3.1. In Subsection 7.3.2 we show the analysis of our theorems. Preliminary results for the limiting distribution of RECORDINALITY is given in Subsection 7.3.3. We report many experimental results that validate our theoretical findings and compare RECORDINALITY with other existing algorithms in Subsection 7.3.4.

7.3.1 Results

RECORDINALITY simply implements the strategy "hiring above the k-th best" and reports *the number of selected values*, call it r_k, over the hash values that represent the input data stream.

In Algorithm 1, at every step, RECORDINALITY keeps in H the k largest values seen so far on the generated sequence of hash values. Once H is full with the first k distinct values, the second phase of the algorithm starts. If the hash value y of the current element is smaller than the k-th largest value seen so far, we are sure that y is a non-k-record, whether this is its first apparition in the stream or not. Otherwise, we check if y is already in H —then y would be simply discarded because it is a repetition, and if not, we remove the smallest in H and add y to H, incrementing the counter r_k of distinct "hired" or selected elements.

It might happen that there are less than k distinct elements in the stream, thus the number r_k in this case is the cardinality of the stream. When the input stream S is exhausted, RECORDINALITY produces an estimation of the unknown cardinality n according to the following theorem:

Theorem 7.5 *Let* $r_k \in \mathbb{N}$ *be the empirically observed number of selections of the strategy "hiring above the* k*-th best" after processing a sequence with unknown the number of distinct elements* n*; then the estimator* $\mathcal{R}_k$*, is defined by*

$$\mathcal{R}_k = \begin{cases} k \cdot \left(1 + \frac{1}{k}\right)^{r_k - k + 1} - 1, & \text{if } r_k \geq k, \\ r_k, & \text{if } r_k < k. \end{cases}$$

is an unbiased *estimator of* n*, in the sense that* $\mathbb{E}\{\mathcal{R}_k\} = n$.

[1] We can disregard collusions of the hash values; by a proper choice of the hash function we can make the probability of collisions (different elements with the same hash value) virtually 0.

Algorithm 1 Using the number of k-records to estimate cardinality

procedure RECORDINALITY(S,k)
▷ $S = s_1, \ldots, s_N$; $N \gg k$

▷ H: the k largest values seen so far in S
▷ r_k: the number of k-records in S plus $k - 1$

$H \leftarrow (\emptyset, \ldots, \emptyset)$; $r_k \leftarrow 0$;
$i \leftarrow 1$;
while $i \leqslant N$ **do**
$y \leftarrow \text{HASH}(s_i)$;
if $|H| < k$ **then**
if $y \notin H$ **then**
$H \leftarrow H \cup \{y\}$; $r_k \leftarrow r_k + 1$;
end if
else if $y \geqslant \min(H) \wedge y \notin H$ **then**
$H \leftarrow H - \min(H) \cup \{y\}$; $r_k \leftarrow r_k + 1$;
end if
$i \leftarrow i + 1$;
end while
if $r_k \leqslant k$ **then**
return r_k;
else
return $k \cdot \left(1 + \frac{1}{k}\right)^{r_k - k + 1} - 1$;
end if
end procedure

Moreover, the standard error, defined as $\frac{\sqrt{\mathbb{V}\{\mathcal{R}_k\}}}{n}$, gives the accuracy of $\mathcal{R}_k$ according to the following theorem:

Theorem 7.6 *The exact accuracy of the estimator $\mathcal{R}_k$, expressed in terms of standard error,*

$$\mathbb{SE}\{\mathcal{R}_k\} = \sqrt{\frac{k\Gamma(k+1)}{\Gamma(k+2+\frac{1}{k})} \cdot \frac{\Gamma(n+3+\frac{1}{k})}{n^2 \cdot \Gamma(n+1)} - 1 - \frac{2}{n} - \frac{1}{n^2}},$$

Asymptotically as $n \to \infty$ and k is large, the accuracy satisfies

$$\mathbb{SE}\{\mathcal{R}_k\} \sim \sqrt{\left(\frac{n}{ke}\right)^{\frac{1}{k}} - 1}.$$

Theorem 7.6 tells us that this estimator loses accuracy as n becomes very big. For all practical applications, a memory of 1KB plus one counter is enough to estimate the cardinality of a huge data stream that contains several millions of distinct elements, with accuracy less that 10%.

7.3.2 Analysis

Our starting point is the exact formula of the probability distribution of the number of updates of the counter r_k, that is given in Theorem 6.1, we recall it here

$$\mathbb{P}\{r_k = j\} = \begin{cases} [\![n = j]\!], & \text{if } k > n, \\ \\ \frac{k! \cdot k^{j-k}}{n!} \cdot \left[\begin{smallmatrix} n-k+1 \\ j-k+1 \end{smallmatrix}\right], & \text{if } k \le j \le n. \end{cases}$$

Since the expectation of r_k holds uniformly for $1 \le k \le n$ and $n \to \infty$; that is

$$\mathbb{E}\{r_k\} = k(\log n - \log k + 1) + \mathcal{O}(1),$$

which suggests that the following $\mathcal{Z}_k$ is a rough estimator of the unknown n, where

$$\mathcal{Z}_k = \exp(\alpha_k \cdot r_k),$$

with α_k some corrective factor to be determined later. For $k < n$, the expected value of $\mathcal{Z}_k$ can be computed as follows,

$$\begin{aligned} \mathbb{E}\{\mathcal{Z}_k\} &= \sum_{j=k}^{n} \exp(\alpha_k \cdot j) \cdot \frac{k! \cdot k^{j-k}}{n!} \begin{bmatrix} n-k+1 \\ j-k+1 \end{bmatrix} \\ &= \frac{k! \exp((k-1)\alpha_k)}{kn!} \sum_{j=1}^{n-k+1} \begin{bmatrix} n-k+1 \\ j \end{bmatrix} (k \exp(\alpha_k))^j. \end{aligned}$$

By applying identity (1.6) for the unsigned Stirling numbers of the first kind, we get

$$\mathbb{E}\{\mathcal{Z}_k\} = \frac{k! \exp((k-1)\alpha_k)}{kn!} \cdot \left(k \exp(\alpha_k)\right)^{\overline{n-k+1}}. \tag{7.1}$$

Since we have $n!$ in the denominator, then we need to choose α_k such that $n!$ is canceled but some linear factor of n remains. Thus taking $\alpha_k = \log(1 + \frac{1}{k})$ leads to

$$\Big(k\exp(\alpha_k)\Big)^{\overline{n-k+1}} = \frac{(n+1)!}{k!}.$$

Substituting in 7.1 gives us:

$$\begin{aligned}\mathbb{E}\{\mathcal{Z}_k\} &= \frac{k!\exp((k-1)\alpha_k)}{kn!}\Big|_{\alpha_k=\log(1+1/k)} \cdot \frac{(n+1)!}{k!} \\ &= (n+1)\cdot\frac{1}{k}\cdot\Big(1+\frac{1}{k}\Big)^{k-1}.\end{aligned}$$

Now we can set the following unbiased estimator,

$$\begin{aligned}\mathcal{R}_k &= k\Big(1+\frac{1}{k}\Big)^{1-k}\cdot\mathcal{Z}_k\Big|_{\alpha_k=\log(1+1/k)} - 1 \\ &= k\cdot\Big(1+\frac{1}{k}\Big)^{r_k-k+1} - 1.\end{aligned}$$

Indeed,

$$\begin{aligned}\mathbb{E}\{\mathcal{R}_k\} &= k(1+\frac{1}{k})^{-k+1}\sum_{j=k}^{n}(1+\frac{1}{k})^j\cdot\frac{k!k^{j-k}}{n!}\begin{bmatrix}n-k+1\\ j-k+1\end{bmatrix} - 1 \\ &= \frac{k!}{n!}\sum_{j=1}^{n-k+1}\begin{bmatrix}n-k+1\\ j\end{bmatrix}(k+1)^j \\ &= n.\end{aligned}$$

Thus RECORDINALITY is an *unbiased estimator* of the unknown number of distinct elements n. In particular, RECORDINALITY works for the whole range of cardinalities $n > k$, and is not only *asymptotically unbiased*, as it is the case for many other existing estimators.

Standard error computations: As we know, the standard error of an estimator $\mathcal{Z}$ is defined as follows:

$$\mathbb{SE}\{\mathcal{Z}\} = \frac{1}{n}\sqrt{\mathbb{V}\{\mathcal{Z}\}}.$$

To compute the variance of RECORDINALITY, first we compute the second moment:

$$\begin{aligned}\mathbb{E}\{\mathcal{R}_k^2\} &= \mathbb{E}\left\{\left(k\cdot\left(1+\frac{1}{k}\right)^{r_k-k+1}-1\right)^2\right\}\\ &= \phi_k^2\cdot\mathbb{E}\{\lambda_k^{2\cdot r_k}\}-2n-1,\end{aligned}$$

with $\phi_k = k\cdot\left(1+\frac{1}{k}\right)^{-k+1}$ and $\lambda_k = 1+\frac{1}{k}$. Using the explicit formula of the distribution of r_k again to get,

$$\begin{aligned}\mathbb{E}\{\mathcal{R}_k^2\} &= \phi_k^2\sum_{j=k}^{n}\lambda_k^{2j}\cdot\frac{k!k^{j-k}}{n!}\begin{bmatrix}n-k+1\\ j-k+1\end{bmatrix}\\ &= \frac{\phi_k^2\lambda_k^{2(k-1)}k!k^{-1}}{n!}\sum_{j=1}^{n-k+1}\begin{bmatrix}n-k+1\\ j\end{bmatrix}(k\lambda_k^2)^j\\ &= f_k\cdot\frac{\Gamma(n-k+k\lambda_k^2+1)}{\Gamma(n+1)},\end{aligned}$$

with

$$f_k = \frac{\phi_k^2\lambda_k^{2(k-1)}k!k^{-1}}{(k\lambda_k^2-1)!}.$$

After doing a couple of simplifications, we get

$$\begin{aligned}f_k &= \frac{k\Gamma(k+1)}{\Gamma(k+2+\frac{1}{k})},\\ k\lambda_k^2-k &= 2+\frac{1}{k}.\end{aligned}$$

Finally, we can obtain the *exact* formula for the standard error of RECORDINALITY:

$$\begin{aligned}\mathbb{SE}\{\mathcal{R}_k\} &= \frac{1}{n}\sqrt{f_k\cdot\frac{\Gamma(n+3+\frac{1}{k})}{\Gamma(n+1)}-2n-1-n^2}\\ &= \sqrt{f_k\cdot\frac{\Gamma(n+3+\frac{1}{k})}{n^2\cdot\Gamma(n+1)}-1-\frac{2}{n}-\frac{1}{n^2}}.\end{aligned}$$

Asymptotically as $n\to\infty$, we can use Stirling's formula (1.2) to get:

$$\mathbb{SE}\{\mathcal{R}_k\} = \sqrt{f_k\cdot n^{\frac{1}{k}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)-1}.$$

We can obtain a more simpler formula for larger k; we apply Stirling's formula again to f_k. Then

$$\begin{aligned}\mathbb{SE}\{\mathcal{R}_k\} &= \sqrt{\left(\frac{1}{ke}\right)^{\frac{1}{k}}\left(1+\mathcal{O}\left(\frac{1}{k^2}\right)\right)\cdot n^{\frac{1}{k}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)-1}\\ &\sim \sqrt{\left(\frac{n}{ke}\right)^{\frac{1}{k}}-1}.\end{aligned}$$

7.3.3 Limit distribution

As Theorem 6.1 tells us, the r.v. r_k converges to a *Normal* distribution. Since the estimator $\mathcal{R}_k$ is based on r_k, then it is natural to expect $\mathcal{R}_k$ to have a Log-normal distribution according to the following theorem:

Theorem 7.7 *For any* $1 \leq k \leq n$, *the estimator* $\mathcal{R}_k$ *satisfies*

$$\frac{1}{\sigma} \cdot \log_{1+1/k}\left(\frac{\mathcal{R}_k + 1}{k+1}\right) - \sigma \xrightarrow{(d)} \mathcal{N}(0,1),$$

where

$$\sigma = \sqrt{k\,(\log n - \log k)}.$$

Proof: we do a direct application of the *continuous mapping theorem* (1.1) to get this rough result (while it still requires more enhancements). We start with the mathematical formula of the estimator

$$\mathcal{R}_k = \left(1 + \frac{1}{k}\right)^{r_k - k + 1} - 1,$$

thus

$$\log_{1+1/k}\left(\frac{\mathcal{R}_k}{k}\right) + k - 1 = r_k,$$

but Theorem 6.1 tells us that

$$\frac{r_k - \mu}{\sigma} \xrightarrow{(d)} \mathcal{N}(0,1),$$

with

$$\mu = k\big(\log n - \log k + 1\big), \quad \sigma = \sqrt{k\big(\log n - \log k\big)}.$$

Then we can write

$$\frac{1}{\sigma}\left(\log_{1+1/k}\left(\frac{\mathcal{R}_k}{k}\right) + k - 1 - \mu\right) \xrightarrow{(d)} \mathcal{N}(0,1),$$

doing some simplifications leads to the result stated in Theorem 7.7. ■

This distributional result allows us to provide this corollary which, while not rigorous, restates the properties of the distribution in more immediate terms.

Corollary 7.1 *Let* σ *be the standard deviation of the distribution for some large* k*; the algorithm* RECORDINALITY *is expected to provide estimates within* σ, 2σ, 3σ *of the exact count in respectively* at least *68%, 95% and 99% of all cases.*

When k *is smaller, the estimates may be significantly more concentrated. For instance, for* $k = 10$, *the estimates are within* σ, 2σ, 3σ *of the exact count in respectively 91%, 96% and 99% of all cases.*

7.3.4 Experimental results

We focus here on two aims: first, to show that our theoretical results are validated by practical simulations; second, to give some idea of how RECORDINALITY compares against similar cardinality estimation algorithms.

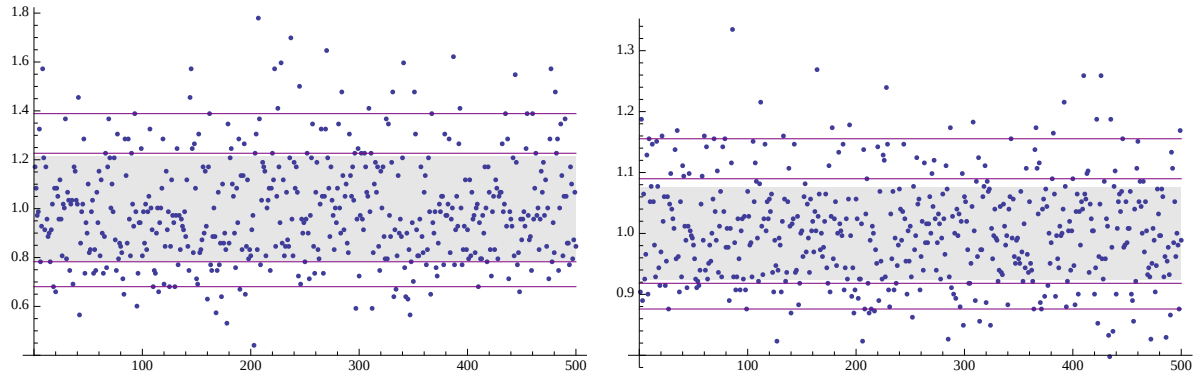


Figure 7.1: Two plots showing the accuracy of 500 estimates of the number of distinct elements contained in Shakespeare's *A Midsummer Night's Dream*, on the left for $k = 64$ and on the right for $k = 256$. Above the top and below the bottom lines lie 5% of the estimates; contained in between the two centermost lines is 70% of the estimates. As a reference, the gray rectangle delimits the area within one standard deviation from the mean.

Unbiasedness and standard error. In both plots of Figure 7.1, we have plotted the accuracy of 500 estimates of the number of distinct elements contained in Shakespeare's *A Midsummer Night's Dream*, each made with a new random hash function[2]; the accuracy is expressed on the y-axis as the ratio of the estimate to the actual number of distinct elements, which is $n = 3031$.

As stated by Theorem 7.5, RECORDINALITY is an unbiased estimator, which can be seen from the fact that the points on the plot are concentrated around 1.0. The gray rectangle delimits the area where estimates are within one standard deviation from the mean, as stated by Theorem 7.6; this mostly coincides with the area between the second and third (empirically placed) level lines containing 70% of the estimates, in accordance with Corollary 7.1.

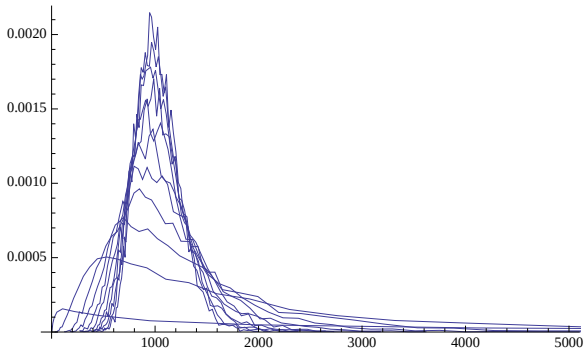


Figure 7.2: Each curve is the empirical distribution of 10 000 estimates made by RECORDINALITY of some random text containing $n = 1000$ distinct elements, for $k = 1, 5, 10, \ldots, 50$. This validates theoretical calculations showing that estimates made by the algorithm are log-normally distributed.

[2]We use an affine transformation of a base hash function, using large prime parameters.

Distribution. Figure 7.2 displays the empirical distribution made from 10 000 estimates on some random text using RECORDINALITY, for values of $k = 1, 5, 10, \ldots, 50$. It convincingly shows that the estimates are log-normally distributed.

Comparative performances. Tables 7.1, 7.2 and 7.3 compare the outcome of several thousand simulations using four different algorithms, on various types of data. For Shakespeare's play, where $n = 3031$, and where we required hashing, we would proceed as follow: for each simulation, we would draw a random hash function, apply it to the words of the stream, then feed the same hashed stream to each of the four algorithms. For the two random streams, containing respectively $n = 6000$ and $n = 50\,000$ distinct elements, we directly generated random uniform variables and used them as input for the four algorithms.

The algorithms we chose to compare RECORDINALITY to, were picked for the following reasons: *Adaptive Sampling* [31] is the only other cardinality estimation algorithm which, in addition, provides a random sample of the underlying set of the stream (see Subsection 7.4.4), so it seemed pertinent to compare its performance with that of our algorithm; the algorithm based on the k-th order statistic [9] functions maintaining the same data structure as RECORDINALITY and practically the same information; finally, HYPERLOGLOG [34] is the optimal algorithm used in practice.

7.4 Extensions and discussion

We have presented RECORDINALITY in the light of a new approach of treating the cardinality estimation problem. Besides the simple and elegant mathematical analysis of RECORDINALITY, we believe that RECORDINALITY (and our approach in general) still have other specific interesting features. In this section we discuss some promising ideas related to RECORDINALITY involving switching totally toward the random-order model, combining RECORDINALITY with other estimators and using RECORDINALITY as a sampling algorithm. We also discuss using the stochastic averaging technique with RECORDINALITY.

7.4.1 RECORDINALITY without hash functions

Most cardinality estimation algorithms are based on hash functions. As previously expounded, these have a number of beneficial features, but in this subsection we will only focus on two: first, hash functions chop up and mix the data until it looks *quasi-random*; second, hash functions reduce arbitrary data into *computable values*—either random $(0, 1)$ reals, integers, or random bits. Estimators are then typically just functions which take these values, and output an estimate of the unknown number of distinct elements.

It is noteworthy that these algorithms do not just rely on hash functions for pseudo-randomness, but also for converting the input data to a useful form. Hash functions are thus central, and indeed it is a well-known fact (see for instance [34]) that, in these applications, they generally account for roughly 80% of the run time[3].

On the other hand, the algorithms we describe here, such as RECORDINALITY, never use the values provided by hash functions; instead, because only the *relative ranks* of the hash values are taken into account, the hash functions serve merely to, in a sense, *randomly permute* the input data. But what if the data is *already* a random permutation?

[3]Though in theory hash functions can be calculated extremely fast using hardware, this is, in practice, seldom done.

k	RECORDINALITY		*Adaptive Sampling*		k-th Order Statistic		HYPERLOGLOG	
	Avg.	Error	Avg.	Error	Avg.	Error	Avg.	Error
4	2737	1.04	3047	0.70	4050	0.89	2926	0.61
8	2811	0.73	3014	0.41	3495	0.44	3147	0.42
16	3040	0.54	3012	0.31	3219	0.28	2981	0.26
32	3010	0.34	3078	0.20	3159	0.18	3001	0.18
64	3020	0.22	3020	0.15	3071	0.12	3011	0.13
128	3042	0.14	3032	0.11	3070	0.10	3031	0.09
256	3044	0.08	3027	0.07	3037	0.06	3025	0.06
512	3043	0.04	3043	0.05	3046	0.04	2975	0.08

Table 7.1: Estimating the number of distinct elements in Shakespeare's *A Midsummer Night's Dream* ($n = 3031$). "Avg." represents the average estimate and "Error" is the empirical standard deviation divided by n, both calculated over 10 000 simulations.

k	RECORDINALITY		*Adaptive Sampling*		k-th Order Statistic		HYPERLOGLOG	
	Avg.	Error	Avg.	Error	Avg.	Error	Avg.	Error
4	5569	1.35	5826	0.67	7715	0.86	6148	0.70
8	6162	1.06	5899	0.42	6677	0.43	5938	0.40
16	6278	0.64	6008	0.31	6381	0.28	6131	0.31
32	6172	0.39	5930	0.21	6172	0.19	6058	0.19
64	6009	0.23	5974	0.15	6104	0.13	5949	0.13
128	5993	0.14	5974	0.10	6050	0.09	5996	0.09

Table 7.2: Similar experiments for a random stream containing $n = 6000$ distinct elements.

k	RECORDINALITY		*Adaptive Sampling*		k-th Order Statistic		HYPERLOGLOG	
	Avg.	Error	Avg.	Error	Avg.	Error	Avg.	Error
4	43658	1.19	59474	0.94	81724	1.30	44302	0.42
8	35230	0.52	47432	0.38	57028	0.41	52905	0.39
16	57723	0.98	49889	0.29	52990	0.23	51522	0.27
32	48686	0.45	49480	0.23	50556	0.18	48009	0.16
64	47617	0.34	50524	0.14	51146	0.13	49345	0.14
128	50097	0.17	50452	0.09	50947	0.08	51531	0.10
256	51742	0.11	50857	0.06	50348	0.06	49287	0.06
512	49496	0.09	49920	0.06	50084	0.04	49016	0.04

Table 7.3: Experiments for a random stream containing $n = 50\,000$ distinct elements—here 25 000 simulations were run.

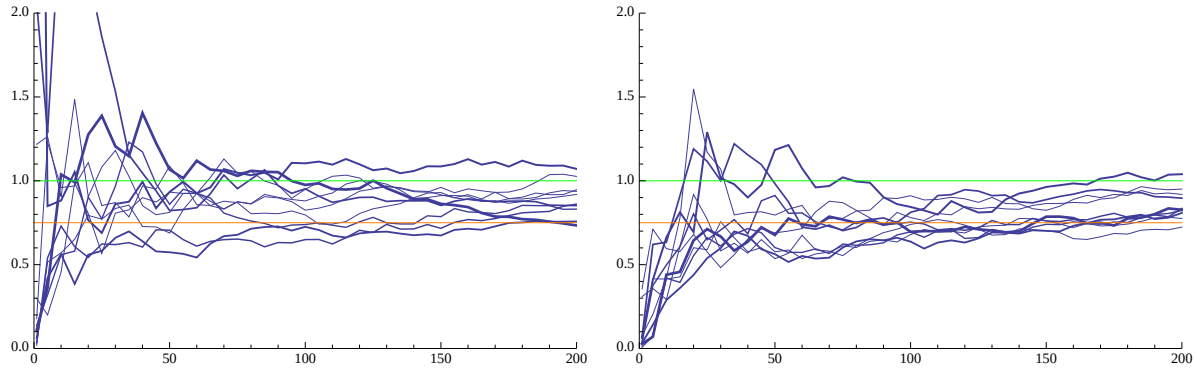


Figure 7.3: Here, we determine the accuracy of estimating the number of distinct words (expressed as the ratio of the estimate to the actual number of distinct words) using a straightforward version of RECORDINALITY which forsakes the use of hash function, as a function of k the memory usage. (a) *Left:* each curve corresponds to one of Shakespeare's tragedies, unmodified and processed *as is*; the *thickness* of the curve is proportional to the number of distinct elements contained in the text, which ranges from 2884 to 4725. (b) *Right:* as a control experiment, we also tried using random permutations of the texts.

Random-order model. In 1980, Munro and Paterson [73], in the context of selecting the median from a list of unsorted elements, were the first to consider the advantages of assuming all orderings of the stream to be equally likely—in other words, to consider that the input stream is a random permutation.

Initially, the idea did not gain much traction, presumably because this assumption seems even more theoretically unjustifiable than considering hashed values as nearly uniform r.v.'s. In 2007, McGregor, in his PhD thesis [68] and related articles, provided the first extensive discussion on the topic and some arguments as to when data streams may be considered random (enough), see [68, §3.1.2]. This allows us to state the following theorem.

Theorem 7.8 *In the random-order model, the algorithm derived from* RECORDINALITY *by removing the hash function and ranking input elements in lexicographical order, is an unbiased estimator of the number of distinct elements, exhibiting the previously demonstrated properties.*

Initial simulations. Determining the practical applicability of this result would require some serious experimentation and, in all likelihood, would certainly result in algorithms which are not universal: for instance, perhaps the generic algorithm given in Theorem 7.8 can be tweaked to give accurate estimation for literary English texts, or for some other class of input streams.

Out of sheer curiosity, we ran this version of RECORDINALITY on Shakespeare's tragedies for varying values of k. The results of these simulations are plotted in Figure 7.3.

Several points are striking. First of all, given the fact that Shakespeare's plays cannot reasonably be considered to be random-order streams, it is quite remarkable that the estimates are so accurate: starting around $k = 50$, most estimates fall within a 25% accuracy. As a matter of fact, the algorithm seems to consistently underestimate the number of distinct elements, to such an extent that it might be possible to compensate for this and obtain a better accuracy. The smaller cardinalities (indicated by the thinner lines) might be better estimated than larger cardinalities,

but this is far from obvious (this behaviour is perhaps according to the standard error formula of RECORDINALITY. For same k, small cardinalities are better estimated than larger ones). Finally, it seems notable that accuracy increases proportionally to the memory used, up until $k = 50$, past which any additional memory seems to be wasted.

A second plot, on the right, gives a second set of experimental results, when instead of the original texts, we use a random permutations of these. Interestingly, the initial gross over-estimations of the previous simulations (which are truncated on the plot) do not seem to occur; while the cut-off in accuracy still seems to be around $k = 50$.

It remains of course—related to experimental validation—to see whether the overhead introduced by comparing strings (instead of the integers computed by the hash function) compensates any gain in speed from avoiding hash functions—one possible avenue would be to store the largest elements in a *ternary search tree*.

7.4.2 Stochastic averaging and RECORDINALITY

Stochastic averaging technique has been introduced by Flajolet and Martin [36], which simulates taking many different estimates of a same stream at a fraction of the computational cost. Using this technique, the values of the data stream are uniformly split into m substreams, estimations for n/m are made separately in each substream, then averaged and scaled up: the expected gain in accuracy is of order $\sqrt{m}$ for a memory usage that is multiplied by m.

This technique is not pertinent to algorithms based on the k-th order statistic, such as [9] or our algorithm RECORDINALITY, because taking the k-th order statistic (as opposed to the minimum, which is the first order statistic) is essentially the same as averaging k minima. In other terms, coupling stochastic averaging with an algorithm based the k-th order statistic is *redundant*.

Even so, we can calculate the accuracy of RECORDINALITY in which we split the stream in m substreams using stochastic averaging (see any of [25, 34, 36, 66] for details), and combined them using an arithmetic mean.

Consider the repartition vector $\overline{n} = [n_1, \ldots, n_m]$ with $n_1 + \ldots + n_m = n$, where n_i is the number of distinct elements which are hashed into the i-th substream, and let $X_{(i)}$ be the r.v. for the estimated number of distinct elements in corresponding substream. We make no assumptions here on the estimator which is used, other than it is unbiased, i.e. $\mathbb{E}\{X_{(i)}\} = n_i$.

Conditioned on some fixed repartition $\overline{n}$, the $X_{(i)}$ are independent. Thus, if $X = X_{(1)} + \ldots + X_{(m)}$ is the sum of the m independent r.v., then,

$$\mathbb{E}\{X\} = \mathbb{E}\{X_{(1)}\} + \ldots + \mathbb{E}\{X_{(m)}\} = n_1 + \ldots + n_m = n.$$

Using the multinomial theorem, we show that the actual expected value (no longer conditioned on a particular repartition) is equal to n. But more importantly,

$$\mathbb{V}\{X\} = \sum_{n_1,\ldots,n_m} \frac{1}{m^n} \binom{n}{n_1, \ldots, n_m} \left(\mathbb{V}\{X_{(1)}\} + \ldots + \mathbb{V}\{X_{(m)}\}\right),$$

in particular, when all variances are equal

$$\mathbb{V}\{X\} = m\mathbb{V}\{X_{(i)}\} \sum_{n_1,\ldots,n_m} \frac{1}{m^n} \binom{n}{n_1, \ldots, n_m} = m\mathbb{V}\{X_{(i)}\}.$$

Hence, we can establish the following theorem:

m	k	Avg.	$sd./n$	$\mathbb{SE}\{\mathcal{R}_k\}$
1	500	0.013	8.85	9.72
2	250	0.029	9.04	9.74
5	100	0.196	9.90	9.80
10	50	0.702	9.67	9.91
20	25	0.855	9.80	10.13
50	10	2.352	10.70	10.82
100	5	2.820	12.88	12.12
250	2	3.418	17.62	17.29
500	1	5.928	31.79	31.62

Table 7.4: This table summarizes the empirical results of running 100 trials of RECORDINALITY on a data set `Random` which contains 150 000 pseudo-randomly drawn reals in $(0, 1)$ and using total memory $M = mk = 500$. The algorithm behaves exactly the same way on each trial; only the choice of the hash function will change from one run to the other. The columns show the variation of several parameters as we change the number of substreams m (or for that matter the value of k). "Avg." gives the relative error of the corresponding average estimate ($|n - \text{avg. estimate}|/n$), expressed in percents. "$sd./n$" reports the normalized sample standard deviation, also in percents. "$\mathbb{SE}$" represents the exact mathematical standard error.

Theorem 7.9 *When using stochastic averaging to split the stream into m substreams, the accuracy of the estimator $\mathcal{R}_k$, expressed in terms of standard error, asymptotically as $n \to \infty$, satisfies*

$$\mathbb{SE}\{\mathcal{R}_k\} \sim \frac{1}{\sqrt{m}} \sqrt{\left(\frac{n}{mke}\right)^{\frac{1}{k}} - 1}.$$

Theorem 7.9 makes it clear that, if memory is to be spent, it is always better to increase k than m (and since the total memory usage is km, there is no reason not to favor the former).

This conclusion is strongly supported also by a small experiment which we report its results in Table 7.4.

7.4.3 Hybrid estimators

It is often the case that estimators for the number of distinct elements are *asymptotically* unbiased. This usually means that they get more accurate as the cardinality to estimate is large; but conversely, they suffer from *nonlinear distortions* for smaller cardinalities.

For instance, using stochastic averaging delays the asymptotic regime of an algorithm for well-understood reasons. Several methods have been devised to circumvent this issue. In [66], Lumbroso characterized the initial distortion introduced by stochastic averaging using a Poisson model, and was able to reverse it. In the LOGLOG family of articles [25, 34], among several others (with an extensive discussion in [24]), the idea has been to switch to an auxiliary algorithm, *Linear Counting* [92], when it is detected that the number of distinct elements is small.

This second solution is algorithmically pertinent: stochastic averaging uses an array of buckets, and is distorted when too many buckets are empty; *Linear Counting* uses the number of empty buckets to make its estimate. The data structure and computations are shared by both estimators, and thus no overhead is required to make use of them both.

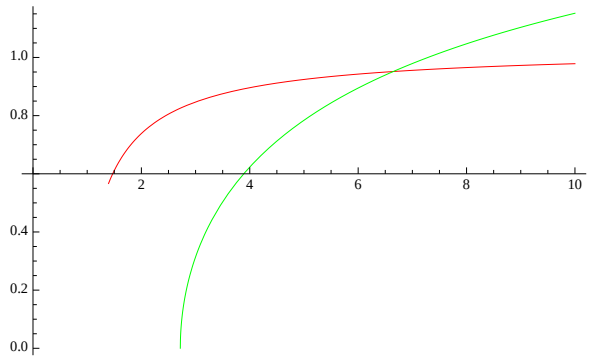


Figure 7.4: This plot compares the *theoretical* dispersion of the estimator based on the k-th order statistic (in red) against that of RECORDINALITY using k-records (in green), as a function of the cardinality expressed as a multiple of k. For cardinalities up to about $n = 6k$, RECORDINALITY is less dispersed (and thus leads to more accurate estimates). The y-axis is rescaled to be independent of k.

Algorithms based on order statistics. But combining an algorithm with *Linear Counting* does not make sense in all cases. Consider Bar-Yossef *et al.*'s first algorithm in [9]: it keeps track of m_k, the k-th minimum of the hashed values, and then estimates that $n \sim (k-1)/m_k$.

The data structure maintained for this algorithm is some kind of a balanced binary tree. *Linear Counting* would in addition require maintaining a separate array. Although this would not change the order of the space and time complexities, it would significantly increase the storage used, doubling it.

On the other hand, Bar-Yossef *et al.*'s order statistic algorithm uses the exact same data structure as RECORDINALITY, and would only require a small extra $\mathcal{O}(\log\log n)$-sized counter to track the number of times the k minima are changed. Figure 7.4 suggests it would be useful to use RECORDINALITY in conjunction with the estimator based on the k-th order statistics, and to switch to the former for cardinalities n that are smaller than $6k$ (where k is the number of minima stored).

Hybridization. A complementary idea comes from the observation that the *values* of the k smallest elements of a stream are independent from their *position* in the stream and thus, from the number of times the data structure maintaining the current k smallest elements is updated while scanning the stream.

In other words, the statistic behind Bar-Yossef *et al.*'s first algorithm (the k-th order statistic) and that behind RECORDINALITY (essentially, the number of k-records) are *independent*. What if, instead of using one or the either, we used *both*?

We expect the accuracy to improve, if for no other reason than because the standard deviation is sub-additive. If we note Bar-Yossef *et al.*'s first algorithm and RECORDINALITY respectively as $\mathcal{O}_k$ and $\mathcal{R}_k$, we define the hybrid algorithm $\mathcal{H}_{\lambda,k}$ as

$$\mathcal{H}_{\lambda,k} = \lambda\mathcal{O}_k + (1-\lambda)\mathcal{R}_k \qquad \text{with} \quad \lambda \in (0,1).$$

Because of the aforementioned independence,

$$\mathbb{V}\{\mathcal{H}_{\lambda,k}\} = \sqrt{\lambda^2\mathbb{V}\{\mathcal{O}_k\} + (1-\lambda)^2\mathbb{V}\{\mathcal{R}_k\}}. \tag{7.2}$$

A suitable value for parameter λ is to be determined, either empirically through simulations, or by plugging in the theoretical variances of the algorithms in (7.2) and maximizing the resulting function (or some approximation of it).

Finally it is worthwhile to mention that RECORDINALITY can be thus hybridized with any existing cardinality estimation algorithm as, to the best of our knowledge, the statistic it uses is likewise independent with all those previously considered (although, as we have said, it does not always make algorithmic sense).

7.4.4 Distinct sampling

We discuss a different usage of the algorithm RECORDINALITY; that is producing a *random sample* from the underlying set of the input data stream. Of course, in such application hash functions are essential for RECORDINALITY. It is well known that sampling has a rich history in different fields like Statistics and Computer Science. If we restrict ourselves in the context related to data streaming, we find that the challenges facing sampling algorithms are not far from cardinality estimation algorithms (as discussed in this chapter). There are also many sampling algorithms; we mention for example *Adaptive Sampling* which is introduced by Mark Wegman and analyzed by Flajolet [31, 32].

This algorithm was subsequently rediscovered by several authors: by Bar-Yossef *et al.* in [9], but also, and perhaps most famously, by Gibbons [42], who most pertinently renamed it *Distinct Sampling*. More recently, Monemizadeh and Woodruff [69] introduced some generalization of its basic idea.

The crucial feature of Adaptive Sampling is that at any point during the course of its execution, the algorithm (parameterized to use m words of memory) stores a uniform sample of between $m/2$ and m distinct elements from the set underlying the stream. Hence in these algorithms, the *size* of the sample is a random variable.

While for RECORDINALITY, the table of the k largest elements maintained is a uniform sample of exactly k elements of the set underlying the stream. This is easy to see: the presence of an element in RECORDINALITY's cache depends only on its hashed value, which is considered uniformly random (in other words, without hash functions, the elements in the cache of RECORDINALITY no longer are a random sample).

Moreover, we think of developing algorithms that can generate distinct samples with tunable size, where the sample size is not constant but rather depends on the unknown number of distinct elements in the stream n. For instance, the set of selected values by hiring above the k-th best has on average $\mathcal{O}(k \log n)$ out of n distinct elements. But the behaviour of this strategy shows some bias in the selection process which makes the selected sample is not totally random. That is because the first k elements in the stream are always contained in the set of selections. After that, the position of an element is crucial where if this element is locally among the best k values seen so far (i.e., if at most there are $k-1$ larger values preceded this element in the stream), then it is selected. In order to compensate this situation, we have to combine the strategy with the replacement mechanism introduced in Section 3.3. With applying hiring with replacement, neither of the first k elements entering the sample nor next selected elements will remain there except it is among the best h_n elements, where the basis strategy would select h_n elements, as Theorem 3.1 states. Again, the largest $\mathcal{O}(k \log n)$ elements are position-free, and thus the generated sample is totally random and in the same time its size is a r.v. which depends on the unknown cardinality

n.

Also other strategies, combined with the replacement mechanism, like hiring above the α-quantile (Chapter 5), with $0 < \alpha < 1$, and p-percentile rules (Subsection 2.2.1), with $0 < p \leq 1$, can be used for the same purpose. The sample of selected values in both cases will have on average $\mathcal{O}(n^{1-\alpha})$ and $\mathcal{O}(n^p)$, respectively. That is quite good because we know the leading factors in case of p-percentile rules for general p, as well as for hiring above the α-quantile for $\alpha = \frac{1}{d}$, $d \in \mathbb{N}$.

Of course, while using any strategy to do the job, we have to keep a pair of the hash value and the actual element in order to recover those elements after processing the whole sequence.

7.5 DISCARDINALITY: a cardinality estimator based on largest non-k-record

Here, we introduce another cardinality estimator; called DISCARDINALITY which also exploits the strategy "hiring above the k-th best" and retrieves the *largest discarded value* from the generated hash values of the input data stream to report the estimation. Remember that after scanning the input sequence using hiring above the k-th best, then there are the set of selections represented by the hiring set and its related set of scores—as explained in Chapter 6, and the discarded elements because any of those discarded has arrived after at least k larger elements than it. So that the largest discarded value is indeed the *largest non-k-record* in the particular permutation of the underlying set of the generated hash values.

By definition, the largest k values (among n distinct ones) are always contained in the set of selections and may be others, but any of the values with ranks (scores) $\{n-k, n-k-1, \ldots, 1\}$ can be the largest non-k-record. Thus DISCARDINALITY depends on the *value* of the *largest non-k-record*. Therefore it is similar to estimators based on order statistics but it will depend on the order (the position) of the element in the data stream unlike the order statistics estimators (see for example, [9]).

We give in Subsection 7.5.1 a pseudo-code for DISCARDINALITY and the main theorems. Subsection 7.5.2 contains the details of the analysis. In Subsection 7.5.3, we discuss using stochastic averaging with DISCARDINALITY. We report many experimental results for DISCARDINALITY in Subsection 7.5.4.

7.5.1 Results

Algorithm 2 introduces a pseudo-code for DISCARDINALITY. It has similar behaviour like RECORDINALITY but we need now to keep the k-records that are smaller than those in H but larger than the largest non-k-record, call it d_k. When a hash value y is added to the k largest values in H, the minimum of H is moved to G. If $y < \min(H)$ then y might be a non-k-record or a repetition of a previously seen k-record. If it is larger than d_k and it is not a repetition of some previous k-record then we update d_k accordingly. While RECORDINALITY needs to keep k elements plus one counter, DISCARDINALITY might need more space, because of the auxiliary space for G. Since $\mathbb{E}\{|G|\} = \mathcal{O}(\sqrt{k})$ (see Theorem 6.5), then the expected memory space for our second algorithm is $k + \mathcal{O}(\sqrt{k})$ elements plus one memory location for d_k.

As explained before, DISCARDINALITY takes into consideration the order of the first occurrence of

each distinct hash value, in contras to all other estimators based on order statistics that are independent of the order.

Moreover, it might happen that d_k is never being updated (still zero) during scanning the stream using DISCARDINALITY; however the probability of this event is very small (almost negligible); it is $\frac{k!k^{n-k}}{n!}$ for n distinct elements in the stream, but in this case DISCARDINALITY can report the *exact* number of distinct elements n that is the number of k-records plus k-1 also. Referring to Subsection 6.3.5 that gives the analysis of the parameter score of best discarded candidate, then this parameter takes the value zero in case of hiring everybody from the input sequence. That means that the *number of selections* r_k, in this case, is exactly the permutation size n. Thus, if we let DISCARDINALITY count r_k, and we do check if $d_k = 0$ (assuming that $0 < d_k < 1$) at the end of the algorithm then, in this case we can decide that the cardinality of the stream is the value of the counter r_k with probability 1.

Thus DISCARDINALITY uses d_k to give an estimation of the unknown cardinality n according to the following theorem:

Theorem 7.10 *Let* $d_k \in (0,1)$ *be the empirically observed largest value which is a non-*k*-record and assume* $d_k > 0$*, after processing a sequence with unknown number of distinct elements* n*, using "hiring above the* k*-th best". Then the estimator* $\mathcal{D}_k$ *defined as*

$$\mathcal{D}_k = \frac{\gamma_k}{1 - d_k},$$

is an asymptotically unbiased estimator *of* n*, in the sense that* $\mathbb{E}\{\mathcal{D}_k\} \sim n$*, where* γ_k *is a corrective factor given as follows:*

$$\gamma_k = \left(\sum_{j=1}^{n-k} \frac{1}{n-j} \mathbb{P}\{d_k = j\} \right)^{-1} \sim k + \sqrt{\frac{\pi k}{2}} + \frac{\pi}{4} - \frac{8}{3}, \text{ as } n \to \infty.$$

We give next the standard error of the estimator $\mathcal{D}_k$.

Theorem 7.11 *The accuracy of the estimator* $\mathcal{D}_k$*, expressed in terms of standard error, asymptotically as* $n \to \infty$*, satisfies*

$$\mathbb{SE}\{\mathcal{D}_k\} \sim \frac{1.19}{\sqrt{k}} \left(1 + \mathcal{O}\Big(\frac{1}{\sqrt{k}}\Big) \right).$$

7.5.2 Analysis

Let us recall the distribution of the parameter best discarded candidate, $M_{n,k}$, under the strategy hiring above the k-th best (Theorem 6.5):

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

Algorithm 2 Using the largest non-k-record to estimate cardinality

procedure DISCARDINALITY(S,k)
▷ $S = s_1, \ldots, s_N$; $N \gg k$

▷ H: the k largest values seen so far in S
▷ d_k: the largest non-k-record in S
▷ G: $H \cup G$ are the values seen so far larger than d_k
▷ r_k: number of k-records in S plus $k - 1$

$H \leftarrow (\emptyset, \ldots, \emptyset)$; $G \leftarrow (\emptyset, \ldots, \emptyset)$;
$d_k \leftarrow 0$; $r_k \leftarrow 0$;
$i \leftarrow 1$;
while $i \leqslant N$ **do**
  $y \leftarrow$ HASH(s_i);
  if $|H| < k$ **then**
    if $y \notin H$ **then**
      $H \leftarrow H \cup \{y\}$; $r_k \leftarrow r_k + 1$;
    end if
  else if $y \geqslant \min(H)$ **then**
    if $y \notin H$ **then**
      $G \leftarrow G \cup \min(H)$; $H \leftarrow H - \min(H) \cup \{y\}$; $r_k \leftarrow r_k + 1$;
    end if
  else
    ▷ $|H| \geqslant k \wedge y < \min(H)$
    if $y \geqslant d_k \wedge y \notin G$ **then**
      $d_k \leftarrow y$; $G \leftarrow G \setminus \{x \in G \mid x < d_k\}$;
    end if
  end if
  $i \leftarrow i + 1$;
end while
if $d_k = 0$ **then**
  return r_k;
else
  return $\frac{k+\sqrt{\frac{\pi k}{2}}+\frac{\pi}{4}-\frac{8}{3}}{1-d_k}$;
end if
end procedure

And the expectation holds uniformly for $1 \leq k \leq n$:

$$\mathbb{E}\{M_{n,k}\} = n - \left(k + \mathcal{O}\left(\sqrt{k}\right)\right).$$

Since the input to DISCARDINALITY is a sequence of real numbers which can be considered i.i.d. r.v.'s from $\mathtt{Unif}(0,1)$ distribution, then after processing the whole stream, the largest non-k-record value d_k has the following distribution:

$$f_{d_k}(x) = \sum_{j=1}^{n-k} f_j(x) \cdot \mathbb{P}\{M_{n,k} = j\},$$

where the density function of the j-th order statistic of n i.i.d. r.v.'s from $\mathtt{Unif}(0,1)$ is

$$f_j(x) = j\binom{n}{j} x^{j-1}(1-x)^{n-j}.$$

The expectation of d_k is then

$$\begin{aligned}\mathbb{E}\{d_k\} &= \int_0^1 x \cdot \sum_{j=1}^{n-k} f_j(x) \cdot \mathbb{P}\{M_{n,k} = j\}dx \\ &= \frac{1}{n+1}\sum_{j=1}^{n-k} j \cdot \mathbb{P}\{M_{n,k} = j\} = \frac{\mathbb{E}\{M_{n,k}\}}{n+1}.\end{aligned}$$

This suggests the following estimator for the number of distinct elements n:

$$\mathcal{I} = \frac{\gamma_k}{1-d_k},$$

where γ_k is some correcting factor to be determined:

$$\begin{aligned}\mathbb{E}\{\mathcal{I}\} &= \mathbb{E}\left\{\frac{\gamma_k}{1-d_k}\right\} \\ &= \gamma_k \int_0^1 \frac{1}{1-x} f_{d_k}(x)dx \\ &= \gamma_k \cdot n \cdot \sum_{j=1}^{n-k} \frac{1}{n-j}\mathbb{P}\{M_{n,k} = j\}.\end{aligned}$$

Thus it is clear that we should choose

$$\gamma_k = \left(\sum_{j=1}^{n-k} \frac{1}{n-j}\mathbb{P}\{M_{n,k} = j\}\right)^{-1},$$

in order to obtain an unbiased estimator, that is $\mathbb{E}\{\mathcal{I}\} = n$.
Now we are looking for some asymptotic approximation for γ_k in the main region $n \to \infty$ and $n \gg k$, that is the case here as we are processing very large sequences with finite memory. First let us find an asymptotic approximation for the summation:

$$\begin{aligned} S_k &= \sum_{j=1}^{n-k} \frac{1}{n-j} \mathbb{P}\{M_{n,k} = j\} \\ &= \sum_{b=1}^{n-k} \frac{k! k^{n-k-b}}{(n-b+1)!} - \sum_{b=1}^{n-k} \frac{k!(k-1)k^{n-k-b}}{(n-b)(n-b+1)!} \\ &= k \sum_{b=1}^{n-k} \frac{k! k^{n-k-b}}{(n-b+1)!} - (k-1) \sum_{b=1}^{n-k} \frac{k! k^{n-k-b}}{(n-b)(n-b)!} \\ &= T_1(k) - T_2(k), \end{aligned}$$

As $n \to \infty$, $T_1(k)$ can be expressed in terms of $R(k)$ below whose asymptotic approximation is given by Knuth [56].

$$\begin{aligned} R(k) &= \sum_{j \geq 0} \frac{k! k^j}{(k+j)!} \\ &= \sqrt{\frac{\pi k}{2}} + \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2k}} + \frac{4}{135k} + \frac{1}{288}\sqrt{\frac{\pi}{2k^3}} + \mathcal{O}\Big(\frac{1}{k^2}\Big). \end{aligned} \tag{7.3}$$

Thus

$$T_1(k) = R(k) - 1 + \mathcal{O}\left(\frac{k! k^{n+1-k}}{(n+1)!}\right), \quad \text{as } n \to \infty \text{ and } n \gg k.$$

$T_2(k)$ can be approximated as follows:

$$\begin{aligned} T_2(k) &= (k-1) \sum_{j \geq 0} \frac{k! k^j}{(j+k)(j+k)!} + \mathcal{O}\left(\frac{k! k^{n-k}}{n \cdot n!}\right), \quad \text{as } n \to \infty \text{ and } n \gg k. \\ &\sim (k-1) \sum_{j \geq 0} \frac{k! k^j}{(j+k)!} \cdot \left(\frac{1}{k} - \frac{j}{k^2} + \frac{j^2}{k^3} - \frac{j^3}{k^4} + \frac{j^4}{k^5} - \frac{j^5}{k^6} + \dots\right). \end{aligned}$$

Now, considering the first few terms of the expansion of $\frac{1}{j+k}$, then evaluating the sum is enough to obtain the significant main order terms of $T_2(k)$, in terms of $R(k)$ too. We have as $n \to \infty$ and $n \gg k$,

$$T_2(k) = \left(1 + \frac{1}{k^2} + \frac{4}{k^3} + \mathcal{O}\Big(\frac{1}{k^4}\Big)\right) R(k) - 1 - \frac{1}{k} - \frac{3}{k^2} - \frac{12}{k^3} + \mathcal{O}\Big(\frac{1}{k^4}\Big).$$

Doing the necessary simplifications gives us:

$$S_k = \frac{1}{k} - \frac{\sqrt{\pi/2}}{k\sqrt{k}} + \frac{8}{3k^2} - \frac{\sqrt{\pi/2}}{12k^2\sqrt{k}} + \mathcal{O}\Big(\frac{1}{k^3}\Big). \tag{7.4}$$

As $\gamma_k = S_k^{-1}$, then we do the following trick to find the asymptotic estimate for γ_k:

$$\begin{aligned}\gamma_k &= \frac{k}{1-\sqrt{\frac{\pi}{2k}}+\frac{8}{3k}+\mathcal{O}\Big(\frac{1}{k^{3/2}}\Big)}\\ &\sim k\cdot\exp\left(-\log\Big(1-\sqrt{\frac{\pi}{2k}}+\frac{8}{3k}\Big)\right)\\ &= k+\sqrt{\frac{\pi k}{2}}+\frac{\pi}{4}-\frac{8}{3}+\mathcal{O}\Big(\frac{1}{\sqrt{k}}\Big).\end{aligned}$$

Finally, considering the main order terms, we take $\gamma_k \sim k+\sqrt{\frac{\pi k}{2}}+\frac{\pi}{4}-\frac{8}{3}$, as $n\to\infty$ and $n \gg k$; we can derive more lower order terms if necessary.

Standard error computations: As we did before, we start with computing the second moment of DISCARDINALITY,

$$\begin{aligned}\mathbb{E}\{\mathcal{D}_k^2\} &= \frac{1}{S_k^2}\cdot\mathbb{E}\left\{\Big(\frac{1}{1-d_k}\Big)^2\right\}\\ &= \frac{1}{S_k^2}\int_0^1\frac{1}{(1-x)^2}\sum_{j=1}^{n-k}f_j(x)\mathbb{P}\{M_{n,k}=j\}dx\\ &= n(n-1)\frac{1}{S_k^2}\sum_{j=0}^{n-k}\frac{jk!k^{j-1}}{(j+k-2)(j+k-1)(j+k)!},\end{aligned}$$

and thus we have to compute an asymptotic expansion of the new summation. We will treat it as $T_2(k)$ where we expand the two multiplied factors in the denominator, then the summation can be expressed in terms of $R(k)$ in (7.3). Finally we obtain the following approximation , as $n\to\infty$ and k is large enough but $n \gg k$:

$$\begin{aligned}\sum_{j=0}^{n-k}\frac{jk!k^{j-1}}{(j+k-2)(j+k-1)(j+k)!} &= \sum_{j\geq 0}\frac{jk!k^{j-1}}{(j+k-2)(j+k-1)(j+k)!}+\mathcal{O}\left(\frac{(n-k+1)k!k^{n-k}}{n(n-1)(n+1)!}\right)\\ &= \frac{1}{k^2}-\frac{\sqrt{2\pi}}{k^{5/2}}+\frac{25}{3k^3}-\frac{109\sqrt{2\pi}}{12k^{7/2}}+\mathcal{O}\Big(\frac{1}{k^4}\Big).\end{aligned}$$

Using the asymptotic expansions of S_k which is obtained in (7.4), after carrying out the necessary calculations, we get (as $n\to\infty$):

$$\begin{aligned}\mathbb{SE}\{\mathcal{D}_k\} &= \frac{1}{n}\sqrt{\mathbb{E}\{\mathcal{D}_k^2\}-n^2}\\ &= \sqrt{\frac{\frac{1}{k^2}-\frac{\sqrt{2\pi}}{k^{5/2}}+\frac{25}{3k^3}-\frac{109\sqrt{2\pi}}{12k^{7/2}}+\mathcal{O}\Big(\frac{1}{k^4}\Big)}{\frac{1}{k^2}-\frac{\sqrt{2\pi}}{k^{5/2}}+\big(\frac{16}{3}+\frac{\pi}{2}\big)\frac{1}{k^3}-\frac{8\sqrt{2\pi}}{3k^{7/2}}+\mathcal{O}\Big(\frac{1}{k^4}\Big)}-1}\\ &= \frac{1.19}{\sqrt{k}}\left(1+\mathcal{O}\Big(\frac{1}{\sqrt{k}}\Big)\right).\end{aligned}$$

7.5.3 Stochastic averaging and DISCARDINALITY

In this case, stochastic averaging has a double effect on the accuracy and speed-up of DISCARDINALITY. If we combine the results for d_k of m substreams using the arithmetic mean (as we did in Subsection 7.4.2), then the standard error of DISCARDINALITY is given according to the following theorem:

Theorem 7.12 *When using stochastic averaging to split the stream into m substreams, the accuracy of the estimator $\mathcal{D}_k$, expressed in terms of standard error, asymptotically satisfies*

$$\mathbb{SE}\{\mathcal{D}_k\} = \frac{1.19}{\sqrt{mk}}\left(1 + \mathcal{O}\Big(\frac{1}{\sqrt{k}}\Big)\right).$$

The proof of this theorems is omitted as it is analogous of that for Theorem 7.9.

7.5.4 Experimental results

We study here the practical behaviour of DISCARDINALITY combined with stochastic averaging, in order to check our analytic results. Another interesting goal is to report the amount of additional memory used by DISCARDINALITY to keep those values that are larger than the largest non k-record in the input hashes; hence we can make a good conjecture for the exact amount, since theoretically, there are, on average, $\mathcal{O}(\sqrt{k}\,)$ elements larger than the largest non k-record for hiring above the k-th best. So, the total memory necessary for DISCARDINALITY using m substreams is $M \approx m\Big(k + \mathcal{O}(\sqrt{k}\,)\Big)$, with two locations for d_r and r_k.

We report the results of two experiments; in Table 7.5 we give the empirical results for the standard error of DISCARDINALITY when processing three different data sets: 1) the celebrated novel `Hamlet` by Shakespeare, with $N \approx 33\,000$ total words and a vocabulary of $n = 5\,316$ distinct words; 2) an English text book, `Book`, with $N \approx 200\,000$ and $n = 19\,918$; and 3) a multiset `Random` which is a sequence of 150 000 pseudo-randomly drawn reals in $(0, 1)$ (and of which the number of distinct elements has been measured as well). For each data set, we have repeated the estimation 100 times, using a new random hash function. These results show clearly the trade-off between memory and accuracy.

In Table 7.6 we test different combinations, for the memory available, between k and m. We report the average (over the 100 trials) number of additional elements per substream: recall that the algorithm needs to store, on average, $\mathcal{O}(\sqrt{k}\,)$ additional elements per substream. Indeed, the values given in that last column of Table 7.6 agree with the theoretical analysis, growing roughly as $3.5\sqrt{k}$.

Since a larger number of substreams yields a better execution time (more substreams means smaller tables to cope with), the best choice is probably to combine a moderate number of substreams (e.g., $m = 4$) with a relatively large k. The tables also show that large values of k lead to better, more accurate estimates, in accordance with the theoretical analysis in Theorem 7.12.

7.6 Other applications

We discuss here a preliminary idea of another algorithmic application of the strategy "hiring above the k-th best". The addressed problem in this section is what called "similarity index estimation"

m	k	Hamlet	Book	Random
5	10	16.09	15.07	14.94
5	20	11.81	10.91	10.15
10	50	4.52	4.43	4.49

Table 7.5: Empirical sample standard deviation divided by n (in percents) for 100 estimates of the cardinality for different data sets using DISCARDINALITY, and with different memory sizes. The data sets are Shakespeare's `Hamlet` ($N \approx 33\,000$ and $n = 5\,316$), an English text book, `Book` ($N \approx 200\,000$ and $n = 19\,918$), and a multiset `Random` ($N \approx 150\,000$ and n is measured).

m	k	Avg.	sd./n	Aux./m
1	128	0.14	10.47	39.59
2	64	0.29	10.59	27.64
4	32	0.40	10.32	19.05
8	16	0.46	9.73	12.98
16	8	0.56	9.87	8.78
32	4	0.86	10.18	5.93
64	2	1.27	12.06	3.93
128	1	1.96	17.54	2.53

Table 7.6: Empirical results for 100 estimates of the cardinality of a data set, with $N \approx 100\,000$ and n is measured, using DISCARDINALITY. The main memory $M = mk = 128$. "Avg." gives the relative error of the corresponding average estimate ($|n - \text{avg. estimate}|/n$), in percents. "sd./n" reports the normalized sample standard deviation, also in percents. The mathematical standard error is approximately 10%. "Aux./m" gives the average (over the 100 trials) number of additional elements—rather than the table of size k—per substream.

or "Jaccard similarity" estimation (see for example [14]) of two data sets; that is another interesting problem in the context of data streaming algorithms and has diverse applications in Networks, Databases and others.

Similarity index estimation

Due to the huge increment of the dimensions of static or on-line databases, it becomes very likely that some document is almost identical to another one. Then many applications related to database clustering and management require computing the similarity index between two huge data sets. In many real-life situations, the storage and computational requirements for computing exactly the similarity between two datasets are so expensive, consume a lot of time or prohibitive totally. This motivates us to move toward approximate techniques to find feasible solutions (with accepted accuracy to be useful). In stead of processing the whole document, a representative sample which is called *a sketch* of the document is randomly chosen then only those sketches are processed to obtain the required information; that is an index $0 \leq R \leq 1$. If R exceeds some predetermined threshold value then the two processed documents are similar or belong to the same cluster; otherwise they do not.

There are various techniques to attack this problem; the major aspect of each is the way of using *hash functions*. For example, minwise hashing by Broder in [14], Locality sensitive hashing by Andoni and Indyk [4], b-bit minwise hashing introduced by Li and König in [65] and many others.

In the *similarity index estimation* problem: each document is associated with a set of *shingles* which is a string of w consecutive words in the document, i.e. $w = 5$ in several studies (see [65] and references therein). Then given two data sets: $\mathcal{A}$ with cardinality $n_{\mathcal{A}}$ and $\mathcal{B}$ with cardinality $n_{\mathcal{B}}$, then the similarity index can be defined as follows:

$$\rho = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} = \frac{n_{\mathcal{A}} + n_{\mathcal{B}} - n_{\mathcal{A}\cup\mathcal{B}}}{n_{\mathcal{A}\cup\mathcal{B}}} = \frac{n_{\mathcal{A}} + n_{\mathcal{B}}}{n_{\mathcal{A}\cup\mathcal{B}}} - 1,$$

where

$$\rho = \begin{cases} 1, & \text{if } \mathcal{A} \text{ and } \mathcal{B} \text{ are identical } (n_{\mathcal{A}\circ\mathcal{B}} = n_{\mathcal{A}} = n_{\mathcal{B}}), \\ 0, & \text{if } \mathcal{A} \text{ and } \mathcal{B} \text{ are totally distinct } (n_{\mathcal{A}\circ\mathcal{B}} = n_{\mathcal{A}} + n_{\mathcal{B}}), \\ R,\ 0 \le R \le 1, & \text{if } \mathcal{A} \text{ and } \mathcal{B} \text{ have some similarity.} \end{cases}$$

Our proposed approach. We can cast the problem of similarity index estimation in the frame of cardinality estimation. Notice that set $\mathcal{A}$ contains $n_{\mathcal{A}}$ distinct shingles; same thing for $\mathcal{B}$, but if we merge both sets into one set $\mathcal{C}$ then the number of distinct shingles in $\mathcal{C}$ is $n_{\mathcal{C}} = n_{\mathcal{A}} + n_{\mathcal{B}} - j$ where $j = 0$ means that $\mathcal{A}$ and $\mathcal{B}$ are *disjoint* (all shingles in $\mathcal{C}$ are distinct), $j = n_{\mathcal{B}}$ means that $\mathcal{A}$ and $\mathcal{B}$ are *identical* (all shingles of second part of $\mathcal{C}$ are repetitions of first one), while $1 \le j < n_{\mathcal{B}}$ reflects the amount of duplication between $\mathcal{B}$ and $\mathcal{A}$.

The common tool that helps here is of course *hash functions*. We apply one random hash function to $\mathcal{A}$ and $\mathcal{B}$ (assuming that the probability of collisions is negligible), then we try to compute the cardinality of $\mathcal{A} \circ \mathcal{B}$ where "$\circ$" refers to the concatenation of $\mathcal{A}$ followed by $\mathcal{B}$.

It is quite clear that deterministic solutions are out of scope here, imagine that to represent one document of 10^5 distinct English words, we need total number of 5-shingles equals $(10^5)^5 = \mathcal{O}(10^{25})$ [65]; that means that we have the same restrictions as cardinality estimation problem.

Using the score of best discarded. To estimate ρ we have to compute $n_{\mathcal{A}\cup\mathcal{B}}$, assuming that $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ are given. To do that in an efficient way, we do not have to merge the two sets $\mathcal{A}$ and $\mathcal{B}$ but rather we can apply *hiring above the* k*-th best* to the concatenation of two smaller sets. First, $\mathcal{A}$ is processed using the strategy and we extract the set $\mathcal{G}\{\mathcal{A}\}$ that is the set of all values larger than or equal to the largest non-k-record in $\mathcal{A}$ (i.e., $\mathcal{G} = H \cup G$ in Algorithm 2), called $M_{\mathcal{A}}$. We do the same for $\mathcal{B}$ to extract the set $\mathcal{G}\{\mathcal{B}\}$. Then we process the set $\mathcal{G}\{\mathcal{A}\} \circ \mathcal{G}\{\mathcal{B}\}$, where sets $\mathcal{G}\{\mathcal{A}\}$ and $\mathcal{G}\{\mathcal{B}\}$ contain all necessary values to compute the largest non-k-record $M_{\mathcal{A}\circ\mathcal{B}}$ of the set $\mathcal{A} \circ \mathcal{B}$.

Notice that the k-th largest value in $\mathcal{G}\{\mathcal{A}\}$ is the initial threshold when we start applying the strategy to $\mathcal{G}\{\mathcal{A}\} \circ \mathcal{G}\{\mathcal{B}\}$ and $M_{\mathcal{A}\circ\mathcal{B}}$ is initially set to $M_{\mathcal{A}}$. Thus, at the end, either $M_{\mathcal{A}\circ\mathcal{B}} = M_{\mathcal{A}}$ which means $\mathcal{B}$ is *identical* to $\mathcal{A}$, then the estimation based on $M_{\mathcal{A}\circ\mathcal{B}}$ shows that $n_{\mathcal{A}\circ\mathcal{B}} = n_{\mathcal{A}}$, or $M_{\mathcal{A}\circ\mathcal{B}} \neq M_{\mathcal{A}}$ then if the cardinality estimation shows that $n_{\mathcal{A}\circ\mathcal{B}} = n_{\mathcal{A}} + n_{\mathcal{B}}$, that means $\mathcal{B}$ and $\mathcal{A}$ are *disjoint*, otherwise $n_{\mathcal{A}} < n_{\mathcal{A}\circ\mathcal{B}} < n_{\mathcal{A}} + n_{\mathcal{B}}$; in this case $\mathcal{B}$ has some duplicated shingles in $\mathcal{A}$.

The sketch of each document is $\mathcal{G}$ whose size is $k + \mathcal{O}(\sqrt{k})$—on average as previously shown in Subsection 7.5.4. From Algorithm 2, $\mathcal{G}$ is computed very fast using only one single pass over the shingles of the processed document.

We have pointed out in Subsection 7.5.1 that the parameter score of best discarded is sensitive to the "position" of elements in the stream. Such property is not provided by any other estimator—as far as we know. Given two documents that have same set of shingles but one of them is a permutation of the other, then other estimators will decide that the similarity is 100% which does not mean much more than one document is an arbitrary permutation of the other (also pointed out by Broder [14]), while this estimator, in most cases, will report the similarity as less than one, which is more accurate.

We argue that $n_{\mathcal{A}\circ\mathcal{B}}$ gives us the value of $n_{\mathcal{A}\cup\mathcal{B}}$. Our next mission is to use $M_{\mathcal{A}\circ\mathcal{B}}$ to estimate the cardinality $n_{\mathcal{A}\circ\mathcal{B}}$, then to calculate the accuracy of the estimate; this should be relatively easy after the experience we got from the analysis of DISCARDINALITY.

The empirical work is very helpful here. We have to develop some experiments that will help us to investigate the properties of the proposed algorithm, other than comparing to other existing algorithms that use same resources (i.e., hash functions, memory, etc.). Also it will be interesting to investigate this property of the algorithm regarding the order of the vocabularies in the document.

Using the size of the hiring set. We can still make use of considering a data stream as a random permutation of its underlying set. Hence, there is possibility to avoid using hash functions (as explained before in Subsection 7.4.1) especially as the processed documents are often textual. Again we use "hiring above the k-th best", then computing the *size of the hiring set* can help to estimate the set similarity index. Given two sets $\mathcal{A}$ and $\mathcal{B}$ (representing two documents), then we apply the strategy to both sets separately to draw a sketch of $\mathcal{A}$ that is $\mathcal{E}_{\mathcal{A}}$ the largest k values in $\mathcal{A}$, and for $\mathcal{B}$ to obtain $\mathcal{H}_{\mathcal{B}}$ which is the hiring set (notice that $\mathcal{E}_{\mathcal{B}} \supset \mathcal{H}_{\mathcal{B}}$). We also keep the size of the hiring set in both sets $r_{\mathcal{A}}$ and $r_{\mathcal{B}}$. Now we have to compute the size of the hiring set of $\mathcal{A} \circ \mathcal{B}$ which goes according to one of three cases. Let $\max(\mathcal{A})$ and $k\text{-th}(\mathcal{A})$ represent the largest and k-th largest elements in $\mathcal{A}$, respectively, then

i) If $k\text{-th}(\mathcal{A}) \geq \max(\mathcal{B})$, then $r_{\mathcal{A}\circ\mathcal{B}} = r_{\mathcal{A}}$, where no more k-records will be encountered in $\mathcal{B}$.

ii) If $k\text{-th}(\mathcal{B}) < k\text{-th}(\mathcal{A}) < \max(\mathcal{B})$, i.e., there are x, $1 \leq x < k$, distinct values in $\mathcal{E}_{\mathcal{B}}$ larger than $k\text{-th}(\mathcal{A})$, then $r_{\mathcal{A}\circ\mathcal{B}} = r_{\mathcal{A}} + x$.

iii) If $k\text{-th}(\mathcal{A}) \leq k\text{-th}(\mathcal{B})$, then we have to process the set $\mathcal{E}_{\mathcal{A}} \circ \mathcal{H}_{\mathcal{B}}$ to obtain $r_{\mathcal{A}\circ\mathcal{B}}$.

After that, $r_{\mathcal{A}\circ\mathcal{B}}$ is used to estimate $n_{\mathcal{A}\circ\mathcal{B}}$ which we argue that gives the value of $n_{\mathcal{A}\cup\mathcal{B}}$. Of course the size of the sketch in this case is $\mathcal{O}(k \log n)$ which consumes more memory than all other estimators (in most cases only k memory units are necessary), but we gain two important advantageous: first, hashing is not necessary any more, that saves a lot of the processing time, other than saving the effort to implement very good independent hash functions. Second, the order of shingles in the document is taken into consideration as the size of hiring set parameter is sensitive to elements positions in the sequence.

It is left to perform the necessary calculations to obtain an unbiased estimator, which is expected to be doable. Moreover, we have to check imperially the gain in the processing time against the consumed memory, and comparing this algorithm with other known ones that depend essentially on hashing.

Another common related task is **clustering a huge set of documents** into smaller subsets of *similar* documents. In this case, we generate the sketch of the first document, call it the "kernel" then we process on-line (without generating a sketch) the shingles of the next document using the sketch of the kernel and compute the similarity index ρ. If ρ exceeds some predetermined threshold (i.e., $\rho \geq 0.5$) then the later document will be classified into the same cluster of the kernel. The same steps are repeated if there are other clusters, otherwise, we have to generate a sketch of the new document as it starts a new cluster. This procedure is also efficient, fast and it will save a lot of memory as we need not to generate a sketch for each processed document.

7.7 Conclusions

We have seen in this chapter how can we investment some results obtained in the hiring problem in order to develop useful applications in the data streaming field. Two estimators of the number of distinct elements n in a data stream that may contain repetitions were established. First one is RECORDINALITY that uses the number of k-records in the stream to give an estimate of n. While our results related to RECORDINALITY are independent, this estimator has been introduced once before as given in Theorem 7.4.

We gave a pseudo-code for RECORDINALITY in Algorithm 1, the mathematical formula in Theorem 7.5, and characterized its accuracy in Theorem 7.6, as well as a preliminary result of the limiting distribution in Theorem 7.7. We were able to give full analysis of our results with the help of the rich distributional and asymptotic results obtained for the parameter *number of hired candidate* under the strategy "hiring above the k-th best". Moreover, we have shown many advantageous and useful extensions such as this estimator is unbiased (not only asymptotically unbiased, as the case for most known estimators), combining it with other estimators to improve the estimates since small cardinalities are well estimated by RECORDINALITY while other estimators are doing better for large cardinalities, it can work in the random-order model (see Theorem 7.8) and avoid using hash functions, with few modifications, by depending only on the Lexicographical order among elements especially when processing text documents, and it could be used to generate random samples (of the underlying stream) whose size depends on n in an efficient way. We have also discussed combining RECORDINALITY with the stochastic averaging technique (see Theorem 7.9), showing that this will not improve the accuracy as required; it is always useful to dedicate the whole available memory to one RECORDINALITY's cash. Thus, there are many promising ideas related to RECORDINALITY, which are worth to be investigated in the future.

Our second estimator is DISCARDINALITY, that uses the value of the largest non-k-record in the hash values of the input stream to give an estimate of n. We gave for a pseudo-code for DISCARDINALITY in Algorithm 2, the mathematical formula in Theorem 7.10, and reported the accuracy in Theorem 7.11. We were able to give the analysis of this estimator depending one the distributional results obtained for the parameter *score of best discarded* under the strategy "hiring above the k-th best", while the limiting distribution seems to be complicated. We have shown that using stochastic averaging with DISCARDINALITY is very useful as the accuracy improves as stated in Theorem 7.12.

In practice, DISCARDINALITY is not as efficient as RECORDINALITY because it consumes a lot of memory compared to other estimators. However, the algorithmic idea of both looks very useful for studying another interesting problem, that is estimating the similarity between two data sets, and its associated task: clustering a huge set of documents into smaller subsets with similar

documents. We have introduced two proposed algorithms to study the later problem, pinpointing the positive features of this approach, and this is also left as a future work.

Part IV

Conclusions and Future Work

IV.1 Overview

We have shown in this thesis several results for various hiring parameters and their usefulness in particular applications. Those parameters give us a very precise picture on the behaviour of two hiring strategies, namely, "hiring above the median" and "hiring above the m-th best". We have also discussed the relationship between those strategies and other selection rules like the p-percentile rules (Subsection 2.2.1) and the seating plans of the Chinese restaurant process (Section 2.5). Besides the conclusions given by the end of each chapter, we introduce in this section a recap of the obtained results in the thesis for the studied hiring strategies and related selection rules. We will highlight again the open problems and the work left for the future.

The class of "pragmatic rank-based selection rules" has two main categories according to the way in which decisions are taken. Let us talk using the terminology of the "hiring problem", then we can easily switch to other related problems. In a direct way, the *number of choices to hire the next candidate* might be *fixed* along the hiring process, such as "hiring above the m-th best", or it depends on a random variable that is the *number of hired candidates* so far, such as "hiring above the α-quantile". In Tables IV.7 and IV.8, we show the obtained results (with references), so far, for many quantities of interest under different selection rules.

	parameter	Hiring above the α-quantile		p-percentile rules		seating plan (α, θ)	
dynamics	# selections	h_n	$0 < \alpha < 1$: [5], Ch. 5	L_n	$0 < p \leq 1$: [40, 59]	K_n	(α, θ): [77]
			$\alpha = \frac{1}{d}$: Ch. 5		$p = \frac{1}{2}$: Ch. 4		$(\alpha, 1)$: Ch. 5
			$\alpha = \frac{1}{2}$: Ch. 4				$(\frac{1}{2}, 1)$: Ch. 4
	waiting time	W_N	$\alpha = \frac{1}{2}$: [15], Ch. 4	T_N	$p = \frac{1}{2}$: Ch. 4	T_N	$(\frac{1}{2}, 1)$: Ch. 4
	index	L_n	$\alpha = \frac{1}{2}$: Ch. 4		open		open
	distance	Δ_n	$\alpha = \frac{1}{2}$: Ch. 4		open		open
quality	last score	R_n	$\alpha = \frac{1}{2}$: Ch. 4		open		ND
	gap	g_n	$0 < \alpha < 1$: [5], Ch. 5		open		ND
	avg. score		open	A_n	$0 < p \leq 1$: [59]		ND
	best discarded	M_n	$\alpha = \frac{1}{2}$: Ch. 4		open		ND
	# replacements	f_n	$0 < \alpha < 1$: Ch. 5		open		ND
			$\alpha = \frac{1}{2}$: Ch. 4				

Table IV.7: Results of "hiring above the α-quantile", "hiring above the median" ($\alpha = \frac{1}{2}$) and related problems. The parameters shown here are: first, the *dynamics indicators*: "# selections" the *number of selections*, "waiting time" the *number of observations until* N *items are selected*, "index" the *index of last selected item*, and "distance" the *distance between the last two selections*. Second, the *quality indicators*: "last score" the *score (rank) of last selected item*, "gap" the *gap of last selected item*, "avg. score" the *average score (rank) of selections*, "best discarded" the *score (rank) of best discarded item*, and "# replacements" the *number of replacements*. "ND" means that this parameter is not defined in that context. "open" refers to that no results for the parameter are known yet.

	parameter	Hiring above the m-th best		m-records	seating plan $(0, m)$	
dynamics	# selections	$h_{n,m}$	[5], Ch. 6	[83], Ch. 6 $m = 1$: [6]	K_n	[77], Ch. 6
	waiting time	$W_{N,m}$	Ch. 6	Ch. 6	T_N	Ch. 6
	index	$L_{n,m}$	Ch. 6	Ch. 6 $m = 1$: [6]		Ch. 6
	distance	$\Delta_{n,m}$	Ch. 6	Ch. 6 $m = 1$: [6]		Ch. 6
quality	gap	$g_{n,m}$	[5], Ch. 6	trivial		ND
	best discarded	$M_{n,m}$	Ch. 6	Ch. 6		ND
	# replacements	$f_{n,m}$	Ch. 6	ND		ND

Table IV.8: Results of "hiring above the m-th best" ($m = \Theta(1)$ or $m = f(n)$) and related problems. We use same conventions as in Table IV.7.

We point out that the results for some parameters (put as "open" in Table IV.7) can be obtained easily, i.e., the *index of last selected item* and *distance between the last two selections* for the "$\frac{1}{2}$-percentile rule" and particular instances of the seating plans such as $(\frac{1}{2}, 1)$ and $(\frac{1}{2}, 0)$. More results about the *rank of last selected item* and *rank of best discarded item* are also in hand for the $\frac{1}{2}$-percentile rule, while such parameters make a little sense in the context of the CRP. Also, the expectation of the *number of replacements* for the $\frac{1}{2}$-percentile rule can be obtained similarly as we did for hiring above the median.

Moreover, for the p-percentile rules, with $p = \frac{1}{d}$, $d \in \mathbb{N}$, the distributional and asymptotic results for the *number of selected items* can be obtained similarly as done in Chapter 5 for hiring above the α-quantile, with $\alpha = \frac{1}{d}$.

We argue that the expectations of the parameters, *number of selected items*, *gap of last selected item* and *number of replacements* for the p-percentile rules, with $0 < p \leq 1$, have the same order of growth like the corresponding parameters for hiring above the α-quantile (in Chapter 5). Of course, this can be formally proved using the framework of Archibald and Martínez (Section 2.4).

The parameter *average score (rank) of selections*, which has been studied for the p-percentile rules, can also be analyzed for hiring above the α-quantile, following the probabilistic analysis of Krieger et al. in [59]. However, we think that this parameter is not natural for rank-based strategies. On the other hand, such parameter is more informative for score-based strategies (as already discussed by Krieger et al. in [60, 61] for different distributions) as it depends directly on the distribution of the scores.

For the *number of replacements*, f_n, computing the probability distribution is a challenging problem. The difficulty stems from the dependency on the *number of hired candidates*, h_n. Thus the starting point here is to compute the joint probability of f_n and h_n; that seems extremely complicated. Besides its own interest, knowing the probability distribution of f_n would be helpful to develop new sampling algorithms. We hope that we can go further on the knowledge of f_n under different hiring strategies.

Moreover, we discuss here some ideas and open problems that we would like to investigate in

the future. One important question is how to compare two hiring strategies, which requires a suitable definition of the notion of "optimality".

As already pointed out by Broder et al. [15] and Archibald and Martínez [5], non-degenerate hiring strategies[4] always exhibit trade-offs between the quality of the hired staff and the rate at which they hire. "Hiring above the m-th best" provides an excellent example. By playing around with the value of m, 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 $\Delta_{n,m}$ decreases (better hiring rate), but the gap of last hired candidate $g_{n,m}$ 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 of best discarded candidate $M_{n,m}$.
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. For instance, one can say that 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.

We discuss in Section IV.2 *"probabilistic hiring strategies"*. In Section IV.3 we introduce *"multicriteria hiring"* as a practical extension of the hiring problem. We have followed the framework of Archibald and Martínez and obtained a generic PDE for the *number of hired candidates* for general r, where r is the number of attributes that each candidate has. We also propose other variants of the hiring problem in Section IV.4 like *"batch hiring"*, *"hiring with sliding-window"* and *"hybrid hiring"*.

IV.2 Probabilistic hiring strategies

Another approach in getting further insight into the "hiring above the α-quantile" strategies (Chapter 5), 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 threshold candidate.
- Each time a new candidate is "examined" his rank will be compared with the rank or score of the threshold candidate.
  - If the new candidate does not have a larger rank, then he will not be hired and the threshold candidate remains the same.
  - If the new candidate has a rank larger than the threshold candidate, then he will be hired and furthermore with a certain probability $1-p$ (which might depend on certain quantities) the threshold candidate remains the same, but with probability p the threshold candidate changes to the recruited candidate with the lowest score larger than the actual threshold candidate, i.e., the "next better candidate" will be the new threshold candidate.

"Hiring above the median" (generalized for "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

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

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.

IV.3 Multicriteria hiring problem

"Multicriteria hiring" is a practical extension of the standard hiring problem. In multicriteria hiring, the preference between candidates is based on r attributes. Every candidate is given r quality ranks relative to the candidates seen so far. As a simple real-life example, these attributes might correspond to some characteristics of candidates for a job, such as education, work experience, international skills, etc. The sequence of ranks of incoming candidates is now modeled by r random permutations; every permutation represents the sequence of ranks for one attribute.

We follow the framework of Archibald and Martínez (review Section 2.4) to do the analysis here. We have presented also an example of using the symbolic method in Section 1.3 to derive a general PDE for the *number of hired candidate*. Let us consider that the attributes are uncorrelated or independent. That means that, at any stage n we have a tuple of $\vec{\sigma} = (\sigma_1, \ldots, \sigma_r)$ of random permutations each of size n with probability of occurrence $1/n!^r$. So, the generating function of the size of hiring set will be

$$H_r(z, u) = \sum_{\vec{\sigma} \in \mathcal{P}_r} \frac{z^{|\vec{\sigma}|}}{|\vec{\sigma}|!^r} u^{h(\vec{\sigma})}, \tag{IV.5}$$

where $\mathcal{P}_r = \{\vec{\sigma} = (\sigma_1, \ldots, \sigma_r) | \sigma_i \in \mathcal{P},\ |\vec{\sigma}| = |\sigma_1| = \ldots = |\sigma_r|\}$. The subscript r in $H_r(z, u)$ corresponds to multiattribute or multicriteria hiring, with $H_1(z, u) \equiv H(z, u)$ corresponding the standard hiring problem. The recurrence of the size of hiring set is the same as before where

$$h(\vec{\sigma} \circ \vec{j}) = h(\vec{\sigma}) + X_{\vec{j}}(\vec{\sigma}),$$

with $h(\vec{\sigma}) = 0$ if $|\vec{\sigma}| = 0$ and $\vec{j} = (j_1, j_2, \ldots, j_r)$, $1 \le j_i \le |\vec{\sigma}| + 1$; $\vec{j}$ is the vector of ranks of the incoming candidate. The indicator r.v. $X_{\vec{j}}(\vec{\sigma})$ is defined as follows

$$X_{\vec{j}}(\vec{\sigma}) = \begin{cases} 1, & \text{if a candidate with a vector of ranks } \vec{j} \text{ is hired right after } \vec{\sigma}, \\ 0, & \text{otherwise.} \end{cases}$$

And the crucial quantity $X(\vec{\sigma})$ has a similar definition as before,

$$X(\vec{\sigma}) = \sum_{\vec{j} \in (1 \ldots |\vec{\sigma}|+1)^r} X_{\vec{j}}(\vec{\sigma}),$$

which tells us how many "vectors" of ranks among the $(|\vec{\sigma}| + 1)^r$ possible ones, could be hired after $\vec{\sigma}$, under the applied strategy. Doing a simple derivation, we have the following theorem

Theorem IV.13 *Let $H_r(z, u)$ be the generating function defined in* (IV.5)*. Let $X(\vec{\sigma})$ denote the number of vectors $\vec{j}$, $(1, \ldots, 1) \le \vec{j} \le (|\vec{\sigma}| + 1, \ldots, |\vec{\sigma}| + 1)$, such that a candidate with vector of ranks $\vec{j}$ will be hired if interviewed right after $\vec{\sigma}$, that is, $X(\vec{\sigma})$ is the number of vectors $\vec{j}$ such that $h_r(\vec{\sigma} \circ \vec{j}) = h_r(\vec{\sigma}) + 1$. Then*

$$\sum_{j=1}^{r+1} z^{j-1} \left\{ {r+1 \atop j} \right\} \frac{\partial^{j-1}}{\partial z^{j-1}} H_r(z, u) = (1 - u) \sum_{\vec{\sigma} \in \mathcal{P}_r} X(\vec{\sigma}) \frac{z^{|\vec{\sigma}|}}{|\vec{\sigma}|!^r} u^{h(\vec{\sigma})}.$$

The derivation of this theorem is straightforward, with the help of identity (1.7) for Stirling numbers of the second kind, and very similar to that one of Theorem 1.4.

One simple strategy in this class is "hiring above the best in any attribute". For example, let $r = 2$, then

$$X(\vec{\sigma}) = 2 \cdot |\vec{\sigma}| + 1.$$

Here, the *threshold level* is defined by the set of candidates with maximum scores in all attributes, or it is just one candidate who is the best in all attributes. So that further candidates should rank better than the threshold level to get hired.

Another proposed strategy is "hiring above the Pareto optima", where Pareto optima is the set of candidates which are not *dominated* by any other candidate. In this case, determining $X(\vec{\sigma})$ seems to be more tricky.
For example, let $r = 2$ and the following table shows the scores of eight interviewed candidates, each one has two scores r_1 and r_2,

n	1	2	3	4	5	6	7	8
r_1	2	1	6	5	7	3	4	8
r_2	6	5	2	4	1	8	7	3

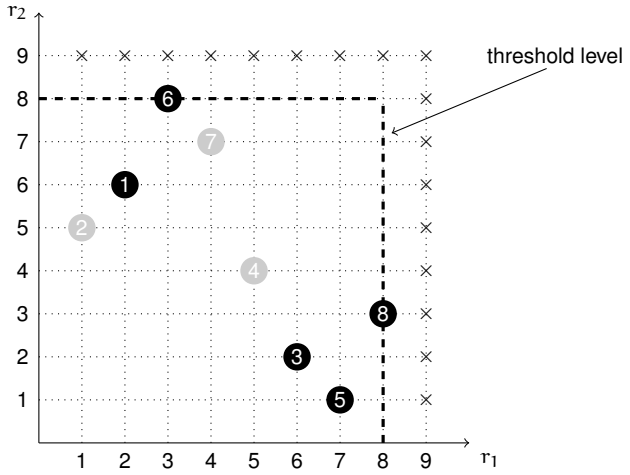
The situation after processing the last candidate using "hiring above the best in any attribute" and "hiring above the Pareto optima" is as explained in Figure IV.1. By definition, the set of Pareto optima contains the best candidate in any attribute, together with candidates who are not dominated by others. Then, it is true that the hiring set under "hiring above the best in any attribute" is a subset of the corresponding hiring set of "hiring above the Pareto optima" for the same sequence of candidates.

IV.4 Other variants of the hiring problem

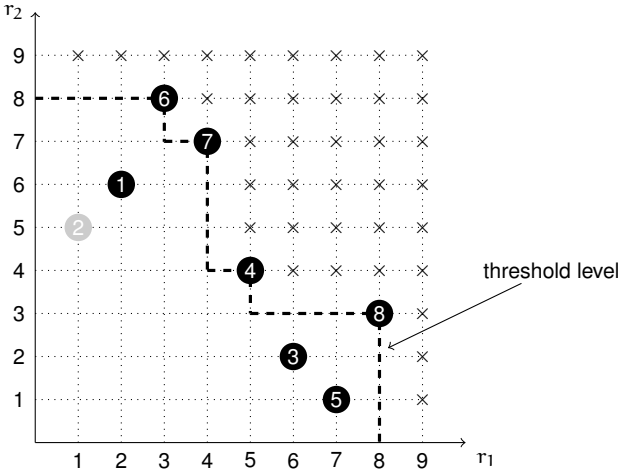
"Batch hiring". Candidates come in blocks or batches of size b, and decisions for all the candidates in the block are simultaneously taken.

"Sliding-window". We can change our mind and hire some candidate who we already interviewed and provisionally discarded if we have not interviewed more than $w - 1$ candidates afterwards. In other words, for a window size w we take a decision for the n-th candidate at time $n + w - 1$; the decision is made with knowledge of the scores or ranks of the $n + w - 1$ candidates seen so far.

"Hybrid hiring". In order to improve the quality of the hired staff, we might try to use a mixture of hiring strategies. For instance, we could start applying *hiring above the best* for the initial K candidates, then apply *hiring above the median* for the rest of the sequence. We conjecture that the hiring set may shrink compared to using hiring above the median alone, but the quality of the hiring set will be improved.



(a) "Hiring above the best in any attribute". r_1 and r_2 axes represent the scores of candidates according to the first and the second attribute after receiving eight candidates, respectively. The number inside the circle denotes the arrival time of that candidate. Hired candidates are in black, while discarded ones are in gray. If the 9-th coming candidate ranks above the *threshold level*, i.e., the incoming candidate attributes correspond to one of the coordinates marked as $\times$, then gets hired, and discarded otherwise.



(b) "Hiring above the Pareto optima". Same conventions are used as in the previous subfigure. If the 9-th coming candidate ranks above the *threshold level* (in any of the coordinates marked as $\times$), then he is hired, and discarded otherwise. Notice that the 4-th and 7-th candidates are not dominated by any other candidate, hence they belong to the current Pareto optima level.

Figure IV.1: Example of multicriteria hiring.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Books on Mathematics. Dover, New York, 1972.

[2] M. Ajtai, N. Megiddo, and O. Waarts. Improved algorithms and analysis for secretary problems and generalizations. *SIAM Journal on Discrete Mathematics*, 14:1–27, 2001.

[3] E. A. Amin. K-th upper record values and their moments. *International Mathematical Forum*, 6(61):3013 – 3021, 2011.

[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, 2008.

[5] M. Archibald and C. Martínez. The hiring problem and permutations. In *DMTCS Proceedings, 21st International Colloquium on Formal Power Series and Algebraic Combinatorics (FPSAC)*, volume AK, pages 63–76, 2009.

[6] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *Records*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York, 1998.

[7] M. Babaioff, N. Immorlica, and D. Kempe. A knapsack secretary problem with applications. *APPROX-RANDOM*, pages 16–28, 2007.

[8] A. Bagchi and A. Pal. Asymptotic normality in the generalized Pólya−Eggenberger urn model, with an application to computer data structures. *SIAM Journal on Algebraic Discrete Methods*, 6(3):394–405, 1985.

[9] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting Distinct Elements in a Data Stream. In J. D. P. Rolim and S. P. Vadhan, editors, *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques (RANDOM'02)*, pages 1–10. Springer, London, UK, 2002.

[10] M. Bateni, M. Hajiaghayi, and M. Zadimoghaddam. Submodular secretary problem and extensions. In *Proceedings of the 13th International Conference on Approximation, and 14th International Conference on Randomization, and Combinatorial Optimization: Algorithms and Techniques (APPROX/RANDOM'10)*, pages 39–52. Springer-Verlag, Berlin, Heidelberg, 2010.

[11] J. N. Bearden and R. O. Murphy. On generalized secretary problems. In M. Abdellaoui, R. D. Luce, M. J. Machina, and B. Munier, editors, *Uncertainty and risk: mental, formal, experimental representations*, volume 41 of *Theory and Decision Library C*, pages 187–205. Springer, Berlin, Heidelberg, 2007.

[12] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2009.

[13] G. Blom, D. Thorburn, and T. A. Vessey. The distribution of the record position and its applications. *The American Statistician*, 44(2):151–153, 1990.

[14] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society, Washington, DC, USA, 1997.

[15] 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 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'08)*, pages 1184–1193. SIAM, Philadelphia, PA, USA, 2008.

[16] T. Bruss. Sum the odds to one and stop. *Annals of Probability*, 28(3):1384–1391, 2000.

[17] T. Bruss, M. Drmota, and G. Louchard. The complete solution of the competitive rank selection problem. *Algorithmica*, 22(4):413–447, 1998.

[18] N. Buchbinder, K. Jain, and M. Singh. Secretary problems via linear programming. In *Proceedings of the 14th International Conference on Integer Programming and Combinatorial Optimization (IPCO'10)*, pages 163–176. Springer-Verlag, Berlin, Heidelberg, 2010.

[19] R. W. Chen, V. N. Nair, A. M. Odlyzko, and Y. Vardi. Optimal sequential selection of n random variables under a constraint. *Journal of Applied Probability*, 21:537–547, 1984.

[20] F. Chung, S. Handjani, and D. Jungreis. Generalizations of Polya's urn problem. *Annals of Combinatorics*, 7:141–153, 2003.

[21] E. Cramer. Asymptotic properties of estimators of the sample size in a record model. *Statistical Papers*, 41(2):159–171, 2000.

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

[23] C. Dietz, D. van der Laan, and A. Ridder. Approximate results for a generalized secretary problem. Tinbergen Institute Discussion Papers 10-092/4, Tinbergen Institute, 2010.

[24] M. Durand. *Combinatoire analytique et algorithmique des ensembles de données*. PhD thesis, École Polytechnique, France, 2004.

[25] M. Durand and P. Flajolet. LogLog Counting of Large Cardinalities. In G. Di Battista and U. Zwick, editors, *Proceedings of the 11th European Symposium on Algorithms*, volume 2832 of *LNCS*, pages 605–617, 2003.

[26] W. Dziubdziela and A. Tomicka-Stisz. Stochastic ordering of random k-th record values. *Applicationes Mathematicae*, 26(3):293–298, 1999.

[27] C. Estan, G. Varghese, and M. Fisk. Bitmap algorithms for counting active flows on high-speed links. *IEEE/ACM Trans. Netw.*, 14(5):925–937, 2006.

[28] W. Feller. *An introduction to probability theory and its applications*, volume 2. Wiley, 2nd edition, 1966.

[29] W. Feller. *An introduction to probability theory and its applications*, volume 1. Wiley, 3rd edition, 1968.

[30] T. Ferguson. Who solved the secretary problem? *Statistical Science*, 4(3):282–296, 1989.

[31] P. Flajolet. On adaptive sampling. *Computing*, 34:391–400, 1990.

[32] P. Flajolet. Adaptive Sampling. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*, volume Supplement I, page 28. Kluwer Academic Publishers, 1997.

[33] P. Flajolet. Counting by coin tossings. In M. J. Maher, editor, *Proceedings of the 9th Asian Computing Science Conference*, volume 3321, pages 1–12, 2004.

[34] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier. HyperLoglog: the analysis of a near-optimal cardinality estimation algorithm. In P. Jacquet, editor, *Proceedings of Analysis of Algorithms 2007*, pages 127–146, 2007.

[35] P. Flajolet and G. N. Martin. Probabilistic Counting. In *Proceedings of the 24th Annual IEEE Symposium on Foundations of Computer Science (FOCS'83)*, pages 76–82, 1983.

[36] P. Flajolet and G. N. Martin. Probabilistic Counting Algorithms for Data Base Applications. *Journal of Computer and System Sciences*, 31(2):182–209, 1985.

[37] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge Univ. Press, 2008.

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

[39] É. Fusy and F. Giroire. Estimating the number of active flows in a data stream over a sliding window. In D. Appelgate, editor, *Proceedings of the 9th Workshop on Algorithm Engineering and Experiments and the 4th Workshop on Analytic Algorithmics and Combinatorics*, pages 223–231. SIAM press, 2007.

[40] J. Gaither and M. D. Ward. Analytic methods for select sets. *Probability in the Engineering and Informational Sciences*, 26:561–568, 2012.

[41] M. Gardner. Mathematical games. *Scientific American*, pages 150–153, 1960.

[42] P. B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *Proceedings of the International Conference on Very Large Data Bases*, pages 541–550, 2001.

[43] J. P. Gilbert and F. Mosteller. Recognizing the maximum of a sequence. *Journal of the American Statistical Association*, 61(313):35–73, 1966.

[44] F. Giroire. Order statistics and estimating cardinalities of massive data sets. *Discrete Applied Mathematics*, pages 406–427, 2000.

[45] K. S. Glasser, R. Holzsager, and A. Barron. The d-choice secretary problem. *Defense Technical Information Center*, 1979.

[46] A. Helmi. The hiring problem: An analytic and experimental study. Master's thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 2010.

[47] A. Helmi, J. Lumbroso, C. Martínez, and A. Viola. Data streams as random permutations: the distinct element problem. In *DMTCS Proceedings, the 23rd International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, number 1, 2012.

[48] A. Helmi, C. Martínez, and A. Panholzer. Analysis of "hiring above the m-th best candidate strategy", 2012. Submitted to *Algorithmica*.

[49] A. Helmi, C. Martínez, and A. Panholzer. Analysis of the "hiring above the α-quantile" strategy. Technical Report LSI-12-15-R, 2012.

[50] A. Helmi, C. Martínez, and A. Panholzer. Hiring above the m-th best candidate: A generalization of records in permutations. In D. Fernández-Baca, editor, *Proceedings of the 10th Latin American Symposium on Theoretical Informatics (LATIN'12)*, volume 7256 of *LNCS*, pages 470–481. Springer, Berlin, Heidelberg, 2012.

[51] A. Helmi and A. Panholzer. Analysis of "hiring above the median": a "Lake Wobegon" strategy for the hiring problem. In *Proceedings of the ACM-SIAM Meeting on Analytic Algorithmics and Combinatorics (ANALCO'12)*, pages 75–83, 2012.

[52] A. Helmi and A. Panholzer. Analysis of the hiring above the median selection strategy for the hiring problem. *Algorithmica*, pages 1–42, 2012.

[53] H.-K. Hwang. On convergence rates in the central limit theorems for combinatorial structures. *European Journal of Combinatorics*, 19(3):329–343, 1998.

[54] H.-K. Hwang, M. Kuba, and A. Panholzer. Analysis of some exactly solvable diminishing urn models. In *Proccedings, the 19th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC)*, 2007.

[55] R. Kleinberg. A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the 16th annual ACM-SIAM symposium on Discrete algorithms (SODA'05)*, pages 630–631. SIAM, Philadelphia, PA, USA, 2005.

[56] D. E. Knuth. *The Art of Computer Programming : Fundamental Algorithms*, volume 1. Addison Wesley, 3rd edition, 1997.

[57] D. E. Knuth, R. L. Graham, and O. Patashnik. *Concrete Mathematics*. Addison Wesley, 1994.

[58] H. Kösters. A note on multiple stopping rules. *Journal of Mathematical Programming and Operations Research*, 53(1):69–75, 2004.

[59] A. M. Krieger, M. Pollak, and E. Samuel-Cahn. Select sets: rank and file. *Annals of Applied Probability*, 17:360–385, 2007.

[60] A. M. Krieger, M. Pollak, and E. Samuel-Cahn. Beat the mean: sequential selection by better than average rules. *Journal of Applied Probability*, 45:244–259, 2008.

[61] A. M. Krieger, M. Pollak, and E. Samuel-Cahn. Extreme(ly) mean(ingful): sequential formation of a quality group. *Annals of Applied Probability*, 20:2261–2294, 2010.

[62] M. Kuba and H. Prodinger. A Note on Stirling Series. *Integers - Electronic journal of Combinatorial number theory*, 10:A34, 393–406, 2010.

[63] R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer series in statistics. Springer-Verlag, 1983.

[64] M. D. Lee, T. A. Gregory, and M. B. Welsh. Decision-making on the full information secretary problem. In K. Forbus, D. Gentner, and T. Reiger, editors, *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 819–824. Lawrence Erlbaum, 2005.

[65] P. Li and A. C. König. Theory and applications of b-bit minwise hashing. *Commun. ACM*, 54(8):101–109, 2011.

[66] J. Lumbroso. An optimal cardinality estimation algorithm based on order statistics and its full analysis. In *DMTCS Proceedings, 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, pages 489–504, 2010.

[67] H. Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC, 1st edition, 2008.

[68] A. McGregor. *Processing Data Streams*. PhD thesis, University of Pennsylvania, 2007.

[69] M. Monemizadeh and D. P. Woodruff. 1-pass relative-error L_p-sampling with applications. In *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA'10)*, pages 1143–1160. SIAM, Philadelphia, PA, USA, 2010.

[70] J. L. Moreno-Rebollo, I. Barranco-Chamorro, F. López-Blázquez, and T. Gómez Gómez. On the estimating the unknown sample size from the number of records. *Statistics and Probability Letters*, 31(1):7–12, 1996.

[71] J. L. Moreno-Rebollo, F. López-Blázquez, I. Barranco-Chamorro, and A. Pascual-Acosta. Estimating the unknown sample size. *Journal of Statistical Planning and Inference*, (83):311–318, 2000.

[72] T. F. Móri. The random secretary problem with multiple choice. *Annales Universitatis Scientiarum Budapestinensis de Rolando Etvs Nominatae. Sectio Computatorica*, 5:91–102, 1984.

[73] J. I. Munro and M. S. Paterson. Selection and sorting with limited storage. *Theoretical Computer Science*, 12(3):315–323, 1980.

[74] V. B. Nevzorov. Distribution of k-th record values in the discrete case. *Journal of Mathematical Sciences*, 43:2830–2833, 1988.

[75] M. L. Nikolaev and G. Y. Sofronov. Multiple optimal stopping rules for the sum of independent random variables. *Diskretnaya Matematika*, 19:42–51, 2007.

[76] P. Norvig. Hiring: The Lake Wobegon strategy. Google Research Blog, March 11, 2006. `http://googleresearch.blogspot.com.es/2006/03/hiring-lake-wobegon-strategy.html`.

[77] J. Pitman. *Combinatorial Stochastic Processes*. Berlin: Springer-Verlag, 2006. Available at: http://works.bepress.com/jim_pitman/1 and via SpringerLink.

[78] E. Platen. About secretary problems, Mathematical Statistics. *Banach Center Publications PWN-Polish Scientific Publications, Warsaw*, 6:257266, 1980.

[79] M. R. R. W. Poeth. Decision making based on sequences. Master's thesis, Master Operations Research. Maastricht University, Faculty of Humanities and Sciences, Department of Knowledge Engineering, 2009.

[80] J. Preater. A multiple stopping problem,. *Probability in the Engineering and Informational Sciences*, 8:169–177, 1994.

[81] J. Preater. On multiple choice secretary problems. *Mathematics of Operations Research*, 19:597–602, 1994.

[82] J. Preater. Sequential selection with a better-than-average rule. *Statistics and Probability Letters*, 50:187–191, 2000.

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

[84] Z. S. Qin. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22(16):1988–1997, 2006.

[85] S. I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Applied Probability. Springer, New York, 1987.

[86] D. A. Seale and A. Rapoport. Sequential decision making with relative ranks: An experimental investigation of the "secretary problem". *Organizational Behavior and Human Decision Processes*, 69:221–236, 1997.

[87] R. Sedgewick and P. Flajolet. *An introduction to the analysis of algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1996.

[88] J. Sonnemans. Decisions and strategies in a sequential search experiment. *Journal of Economic Psychology*, 21(1):91 – 102, 2000.

[89] W. E. Stein, D. A. Seale, and A. Rapoport. Analysis of heuristic solutions to the best choice problem. *European Journal of Operational Research*, 151:140–152, 2003.

[90] M. E. Taylor. *Partial differential equations. Basic theory*. Texts in Applied Mathematics. Springer, New York, 1996.

[91] N. M. Temme. Uniform asymptotic expansions of the incomplete Gamma functions and the incomplete Beta function. *Mathematics of Computation*, 29(132):1109–1114, 1975.

[92] K.-Y. Whang, B. T. Vander-Zanden, and H. M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Transactions on Database Systems*, 15(2):208–229, 1990.