MINIMEAN OPTIMAL KEY ARRANGEMENTS IN HASH TABLES

Andrew Chi-Chih Yao

CS-TR-145-88

March 1988
Minimean Optimal Key Arrangements in Hash Tables

Andrew Chi-Chih Yao

Department of Computer Science
Princeton University
Princeton, New Jersey 08544

Abstract

For an open-address hash function $h$ and a set $A$ of $n$ keys, let $C_h(A)$ be the expected retrieval cost when the keys are arranged to minimize the expected retrieval cost in a full table. It is shown that, asymptotically for large $n$, when $h$ satisfies a certain doubly dispersive property, as is the case for uniform hashing or double hashing, $C_h(A) = O(1)$ with probability $1 - o(1)$ for a random $A$.

\footnote{This research was supported in part by the National Science Foundation under Grant No. DCR-8308109.}
1 Introduction

Hashing techniques are commonly employed in information storage and retrieval (see e.g. Knuth [Kn]). In Gonnet and Munro [GM] and Rivest [R], the question of optimally arranging a set of keys in a static hash table was studied. In particular, it was shown in Rivest [R] that, asymptotically for large \( n \), when uniform hashing is used, one can, with probability \( 1 - o(1) \), arrange \( n \) keys in a full table such that the worst-case retrieval cost is \( O(\log n) \). A similar result for double hashing was later proved in Yao [Y]. For the optimal static hash table that minimizes the expected retrieval cost, it was suggested in Gonnet and Munro [GM] that an \( O(1) \) expected retrieval cost can be achieved even for full tables, when either uniform hashing or double hashing is used. In this paper we will give a proof of this conjecture.

Let \( A = (a_{ij}) \in \mathcal{A}_n \), where \( \mathcal{A}_n \) is the set of all \( n \times n \) matrices of real numbers. For any permutation \( \sigma \) of \( (1, 2, \ldots, n) \), let \( C(A, \sigma) = \frac{1}{n} \sum_{1 \leq i \leq n} a_{i, \sigma(i)} \). Define the cost of \( A \) as \( C(A) = \min_{\sigma} C(A, \sigma) \).

We are interested in the typical value of \( C(A) \), when \( A \) is randomly generated according to certain distributions. Let \( \Sigma_n \) be the set of all permutations of \( (1, 2, \ldots, n) \). For any \( \sigma \in \Sigma_n \), let \( b(\sigma) = (b_1, b_2, \ldots, b_n) \) be an \( n \)-tuple of integers defined by \( b_{\sigma(i)} = i \) for \( 1 \leq i \leq n \). (Informally, any \( \sigma \) specifies the hash sequence \( (\sigma(1), \sigma(2), \ldots, \sigma(n)) \) for a key \( K \), with \( \sigma(i) \) being the \( i \)-th location to be probed when \( K \) is being retrieved; thus, \( b_j \) is the cost for retrieving \( K \) if \( K \) is stored in location \( j \) in the table.) A hash function \( h \) (for table size \( n \)) is specified by a probability distribution \( p_h \) over \( \Sigma_n \); let \( \mathcal{H}_n \) be the family of all hash functions for a table of size \( n \). Generate a random \( A = (a_{ij}) \in \mathcal{A}_n \) by picking independently, for each \( 1 \leq i \leq n \), a random permutation \( \rho^{(i)} \in \Sigma_n \) distributed according to \( p_h \), and let \( (a_{i1}, a_{i2}, \ldots, a_{in}) = b(\rho^{(i)}) \); let \( q_h \) denotes the probability distribution on \( \mathcal{A}_n \) corresponding to such a random \( A \).

For any hash function \( h \), let \( \Lambda_h(i, j, k, \ell) \) denote the set \( \sigma \in \Sigma_n \) such that \( \sigma(i) = k \) and \( \sigma(j) = \ell \); let \( \lambda_h(i, j, k, \ell) = \sum_{\sigma \in \Lambda} p_h(\sigma) \), where \( \Lambda = \Lambda_h(i, j, k, \ell) \). We say that \( h \) is doubly dispersive if \( \lambda_h(i, j, k, \ell) = 1/n(n-1) \) for all \( 1 \leq i \neq j \leq n \) and \( 1 \leq k \neq \ell \leq n \). For example, the uniform hashing function \( h_0 \) and the double hashing function \( h_1 \) (only for prime integers \( n \)) are both doubly-dispersive, where \( h_0 \) is the uniform distribution over \( \Sigma_n \), and \( h_1(\sigma) = 1/n(n-1) \) if \( \sigma(1), \sigma(2), \ldots, \sigma(n) \) forms an arithmetic progression, i.e. \( \sigma_j \equiv \sigma_1 + (j-1)(\sigma_2 - \sigma_1) \pmod{n} \) for \( 1 \leq j \leq n \), and \( h_1(\sigma) = 0 \) otherwise. Our main result is the next theorem.

Theorem 1 There exist positive constants \( c_1, c_2, c_3 \) such that the following is true: if \( A \) is a random matrix distributed according to \( q_h \), where \( h \in \mathcal{H}_n \) is doubly-dispersive, then \( C(A) \leq c_1 \) with probability \( \geq 1 - \frac{c_2}{n^{c_3}} \).

Corollary There exists a positive constant \( c_4 \) such that, if \( A \) is a random matrix distributed according to \( q_h \), where \( h \in \mathcal{H}_n \) is doubly-dispersive, then \( E(C(A)) \leq c_4 \).
Thus, if \( A \) is generated by either uniform hashing function or double hashing function, then \( C(A) = O(1) \) with probability \( 1 - o(1) \) as \( n \to \infty \).

We now demonstrate that Theorem 1 gives the \( O(1) \) expected retrieval time about hash functions. Given a random set of keys \( K = \{K_1, K_2, \ldots, K_n\} \) with \( \rho^{(i)} \) being the hash sequence for key \( K_i \), any permutation \( \sigma \in \Sigma_n \) defines an arrangement \( R_\sigma \) of the keys in a table of size \( n \), i.e. \( K_i \) in location \( \sigma(i) \) for \( 1 \leq i \leq n \). Let \( A = (a_{ij}) \) with \( (a_{i1}, a_{i2}, \ldots, a_{in}) = b(\rho^{(i)}) \), then the cost of retrieving \( K_i \) is \( a_{i,\sigma(i)} \); if we assume that all keys are equally likely to be retrieved, the expected retrieval cost for \( K \) under \( R_\sigma \) is \( \frac{1}{n} \sum_{1 \leq i \leq n} a_{i,\sigma(i)} \), which is \( C(A, \sigma) \). Thus, \( C(A) \) is the optimal expected retrieval cost for \( K \). Theorem 1 states that, if we use any doubly-dispersive hash function \( h \), then a random set \( K \) of \( n \) keys can almost always be arranged in a full hash table such that the expected retrieval cost is \( O(1) \).

As observed in Gonnet and Munro [GM] and Rivest [R], the optimal key arrangements problem is directed related to the classical minimum assignment problem. When viewed from this perspective, Theorem 1 is about the probable behavior of the optimum cost of certain random assignment problems. There are several well known results in the literature on this topic. In Lazarus [L], it was proved that, for a random \( n \times n \) matrix \( A = (a_{ij}) \) with each \( a_{ij} \) being an independent uniform random variable over \([0, 1] \), \( E(C(A)) \geq 1 + 1/e + O(1/n) \); Walkup [W] showed that, for all \( n \), \( E(C(A)) < 3 \). In Karp [Ka1], with the same probability distribution, it was shown that with probability \( 1 - o(1) \), \( \frac{1}{3} < C(A) < 3 \) for a random \( n \times n \) matrix \( A \); more recently, Karp [Ka2] showed that, for all \( n \), \( E(C(A)) < 2 \). In our result, some dependency relation among the entries has been introduced into the model.

## 2 Main Line of Arguments

In this section we first state without proof two propositions, and then employ them to prove Theorem 1. The proof of the two propositions will be left to Sections 3 and 4. In Section 5, a proof of the corollary to Theorem 1 will be given. We remark that results from [Y] will be needed in the proof of Lemma 7 and in Section 5.

Let \( N \leq n \) be any positive integer. Generate a random \( N \times n \) matrix \( D = (d_{ij}) \) by picking independently, for each \( 1 \leq i \leq N \), a random \( \rho^{(i)} \in \Sigma_n \) distributed according to \( p_h \) and let \( (d_{i1}, d_{i2}, \ldots, d_{in}) = b(\rho^{(i)}) \); let \( q_{h,N} \) be the probability distribution for such a random \( D \). Clearly, \( q_{h,n} \) is just \( q_h \).

For any \( S \subseteq \{1, 2, \ldots, n\} \) with \( 0 < |S| \leq n \), let \( \Delta_S \) denote the set of all injective functions \( \omega: S \to \{1, 2, \ldots, N\} \). For any \( N \times n \) matrix \( D = (d_{ij}) \) and \( \omega \in \Delta_S \), define \( \alpha(D, S, \omega) = \frac{1}{|S|} \sum_{j \in S} d_{\omega(j),j} \). Let \( \alpha(D, S) = \min_{\omega \in \Delta_S} \alpha(D, S, \omega) \).
Let $\lambda, \mu$ be any fixed numbers with $0 < \lambda < 1$ and $0 < \mu < 10^{-4}\lambda^4$. Let $c_5 = 1/(1 - e^{-\lambda/4})$ and $\epsilon = e^{-\lambda\sqrt{2/8}}$; clearly, $0 < \epsilon < 1$. Suppose $|\lambda n| \leq N < n$. Take a random $D$ distributed according to $q_{h,N}$, and let $Z_I$ denote the event that $\alpha(D, S) < n$ for all $S \subseteq \{1, 2, \ldots, n\}$ with $|S| \leq \mu n$.

**Proposition I** $\Pr\{Z_I\} \geq 1 - c_5 \epsilon^n$ for all $n \geq 1/\mu$.

**Proof.** See Section 3. □

Let $A = (a_{ij}) \in A_n$ for which no row contains repeated entries. For any integers $1 \leq i, k \leq n$, let $I_k(i, A)$ denote the set of all integers $j$, $1 \leq j \leq n$, for which $a_{ij}$ are among the $k$ smallest elements in the $i$-th row of $A$. That is, $\{a_{ij} : j \in I_k(i, A)\}$ consists of the $k$ smallest elements of $a_{i1}, a_{i2}, \ldots, a_{in}$. Let $J_k(A) = \{(i, j) : 1 \leq i \leq n, j \in I_k(i, A)\}$. For any $T \subseteq \{1, 2, \ldots, n\}$, define $V_k(A, T)$ as the set $\{i : \exists j \in T \text{ with } (i, j) \in J_k(A)\}$. Thus, $J_k(A)$ is the set of locations in $A$ that contain all the smallest $k$ elements in every row, and $V_k(A, T)$ is the set of every row with at least one of its $k$ smallest elements occurring in some column of $T$.

Let $0 < \gamma < 1/10$ be any fixed number, and $k = \lceil \gamma^{-c} \rceil$ where $c = (32/e)^{10}$. Take a random $A \in A_n$ distributed according to $q_h$. Let $Z_{II}$ denote the event that $|V_k(A, T)| \geq 2|T|$ for all $T \subseteq \{1, 2, \ldots, n\}$ with $\gamma n \leq |T| \leq 2\gamma n$. Let $Z_{III}$ denote the event that $|V_k(A, T)| \geq |T|$ for all $T \subseteq \{1, 2, \ldots, n\}$ with $|T| > 2\gamma n$. Let $\epsilon' = 2^{(\ln 2)/10}$.

**Proposition II** There exists a constant $N_\gamma$ such that $\Pr\{Z_{II} \land Z_{III}\} \geq 1 - 2/n^{\epsilon'}$ for all $n \geq N_\gamma$.

**Proof.** See Section 4. □

We proceed to prove Theorem 1. Let $\gamma = 10^{-6}$ and $k = \lceil \gamma^{-c} \rceil$. Let $A \in A_n$ be any matrix for which no row contains repeated entries, and $S \subseteq \{1, 2, \ldots, n\}$ with $|S| \leq \gamma n$. We will say that $S$ is a virtuous column set for $A$ if the following is true: For all $T \subseteq \{1, 2, \ldots, n\} - S$ with $|T| \leq \gamma n$, we have $|V_k(A, T)| \geq 2|T|$.

Take a random $A \in A_n$ distributed according to $q_h$. Let $Z_{IV}$ be the event that there exists a virtuous column set $S$.

**Lemma 1** $\Pr\{Z_{IV}\} \geq 1 - 2/n^{\epsilon'}$ for all sufficiently large $n$.

**Proof.** Initially, set $S \leftarrow \emptyset$ and $W \leftarrow \{1, 2, \ldots, n\}$. Repeat the following process: as long as there exists a $T \subseteq W$ with $|T| \leq \gamma n$ and $|V_k(A, T)| < 2|T|$, choose lexicographically the smallest such $T$, set $S \leftarrow S \cup T$ and $W \leftarrow W - T$; stop when either $|S| > \gamma n$ or no such $T$ can be found. Let $S_A$ denote the set $S$ when the process stops. As can be readily verified by induction, $S \cap W = \emptyset$ and $W = \{1, 2, \ldots, n\} - S$ at any time. Furthermore, $|V_k(A, S)| < 2|S|$ at any time.

Take a random $A \in A_n$ distributed according to $q_h$. Let $Z_1$ denote the event $\neg Z_{II}$ and let $Z_2$ denote the event that $|S_A| > \gamma n$. From the halting condition, it is clear that $\neg Z_2$ implies that $S_A$
is a virtuous column set. Also, by Proposition II, \( \Pr\{Z_1\} \leq 2/n^\ell \). If we can prove that \( Z_2 \) implies \( Z_1 \), then \( \Pr\{-Z_2\} = 1 - \Pr\{Z_2\} \geq 1 - \Pr\{Z_1\} \geq 1 - 2/n^\ell \); Lemma 1 will thus be proved. We now show that \( Z_2 \) implies \( Z_1 \). If \( Z_2 \) is true, then \( |S_A| > \gamma n \). Let \( S_A = S_1 \cup T_1 \) where \( S_1, T_1 \) are the last values of \( S \) and \( T \) before \( S \) becomes \( S_A \). Then \( |S_1| \leq \gamma n, \ |T_1| \leq \gamma n \), and hence \( |S_A| \leq 2\gamma n \). Now, \( |V_k(A, S_A)| < 2|S_A| \) as noted previously. Thus, \( S_A \) is a witness for \( \neg Z_II \). That is, \( Z_1 \) is true. \( \square \)

Take a random \( A = (a_{ij}) \in A_n \) distributed according to \( q_h \). Let \( A_1 \) be the \([N/2] \times n\) matrix obtained from the top \([N/2]\) rows of \( A \), and \( A_2 \) be the \([N/2] \times n\) matrix obtained from the bottom \([N/2]\) rows of \( A \). Let \( Z_3 \) be the event \( Z_{II} \land Z_{III} \land Z_{IV} \). Let \( \lambda = 1/2 \) and \( \mu = 10^{-6} \). Thus, \( \mu = \gamma \). Define \( Z_4 \) to be the event \( Z_f \) in Proposition I, in which \( D \) is defined as \( A_1 \). Similarly, define \( Z_5 \) to be the event \( Z_I \) in Proposition I, in which \( D \) is defined as \( A_2 \). By Propositions I, II and Lemma 1, \( \Pr\{Z_3 \land Z_4 \land Z_5\} = 1 - O(1/n^\ell) \). We will now show that, when \( Z_3 \land Z_4 \land Z_5 \) is true, there exists a set \( F \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\} \) such that

\[
\sum_{(i,j) \in F} a_{ij} \leq \left( \binom{k+1}{2} + 2 \right) n,
\]

and

\[
|\{i: \exists j \in T, (i,j) \in F\}| \geq |T| \text{ for all } T \subseteq \{1, 2, \ldots, n\}.
\]

This would imply Theorem 1, since (2) guarantees, by Hall's Theorem [H] on matching, the existence of a permutation \( \sigma \in \Sigma_n \) such that \((i, \sigma(i)) \in F\) for all \( 1 \leq i \leq n \), and (1) then guarantees that for this \( \sigma \), \( \sum_{1 \leq i \leq n} a_{i\sigma(i)} \leq (\binom{k+1}{2} + 2)n \).

Suppose \( Z_3 \land Z_4 \land Z_5 \) is true. Let \( S \subseteq \{1, 2, \ldots, n\} \) be a virtuous column set, which must exist since \( Z_3 \) is true. Then, by definition, \( |S| \leq \gamma n = \mu n \). As \( Z_4 \) is true, there exists an injective function \( \omega_1: S \rightarrow \{1, 2, \ldots, \lfloor n/2 \rfloor\} \) such that \( \sum_{j \in S} a_{\omega_1(j), j} < n \). As \( Z_5 \) is true, there exists an injective function \( \omega_2: S \rightarrow \{\lfloor n/2 \rfloor + 1, \ldots, n - 1, n\} \) such that \( \sum_{j \in S} a_{\omega_2(j), j} < n \). Let \( F_0 = \{(\omega_1(j), j), (\omega_2(j), j) : j \in S\} \), and \( F = F_0 \cup J_k(A) \).

Now \( \sum_{(i,j) \in F_0} a_{ij} = \sum_{j \in S} a_{\omega_1(j), j} + \sum_{j \in S} a_{\omega_2(j), j} < 2n \) and \( \sum_{(i,j) \in J_k(A)} a_{ij} = \binom{k+1}{2} n \). Clearly, inequality (1) is satisfied. It remains to prove (2).

Lemma 2 For every \( T \subseteq S \), the set \( Y_T \), defined by \( \{i: \exists j \in T, (i,j) \in F_0\} \), satisfies \( |Y_T| = 2|T| \).

Proof. \( Y_T = \{\omega_1(j), \omega_2(j) : j \in T\} \). \( \square \)

For any \( T \subseteq \{1, 2, \ldots, n\} \), let \( Y'_T = \{i: \exists j \in T, (i,j) \in F\} \). We need to prove \( |Y'_T| \geq |T| \).

CASE 1. If \( |T| > \gamma n \), then as \( Z_{II} \) and \( Z_{III} \) are true, we have \( |V_k(A, T)| \geq |T| \). This implies \( |Y'_T| \geq |T| \).

CASE 2. If \( |T| \leq \gamma n \), let \( T_1 = T \cap S \) and \( T_2 = T \cap (\{1, 2, \ldots, n\} - S) \). By Lemma 2, \( |Y_{T_1}| = 2|T_1| \). Also \( |V_k(A, T_2)| \geq 2|T_2| \) since \( S \) is virtuous. It follows that \( |Y'_T| \geq |Y_{T_1} \cup V_k(A, T_2)| \geq \max\{|Y_{T_1}|, |V_k(A, T_2)|\} \geq 2 \max\{|T_1|, |T_2|\} \geq |T_1| + |T_2| = |T| \).
This completes the proof of (2), and hence, Theorem 1.

3 Proof of Proposition I

The following simple inequality will be proved in the Appendix:

$$x > \frac{8}{\lambda} (1 + 2\ln x) \quad \text{for all} \quad x \geq \frac{1}{\sqrt{\mu}} . \quad (3)$$

Let $r = e^{-\lambda/2}$. We consider an infinite sequence $Y$ of independent identically distributed random variables $Y_1, Y_2, Y_3, \ldots$ with $\Pr\{Y_i = k\} = r^{k-1}(1-r)$ for integers $k \geq 1$.

Lemma 3 For all $n \geq 1/\mu$,

$$\Pr\{Y_1 + Y_2 + \ldots + Y_m \geq \sqrt{\mu} n\} \leq \frac{e^{-\lambda\sqrt{\mu} n/4}}{1 - e^{-\lambda/4}},$$

where $m = \lfloor \mu n \rfloor$.

Proof. Let $Y^{(m)} = \sum_{1 \leq i \leq m} Y_i$. Consider the generating function $g_m(x) = \sum_{k \geq 1} \Pr\{Y^{(m)} = k\} x^k$. A standard calculation gives

$$g_m(x) = \left( \sum_{k \geq 1} \Pr\{Y_1 = k\} x^k \right)^m$$

$$= \left( \sum_{k \geq 1} r^{k-1}(1-r)x^k \right)^m$$

$$= (1-r)^m x^m (1-rx)^{-m}.$$

It follows that, for $k \geq m$,

$$\Pr\{Y^{(m)} = k\} = (1-r)^m \left( \frac{-m}{k-m} \right) r^{k-m}$$

$$= (1-r)^m \frac{m(m+1) \ldots (k-1)}{(k-m)!} r^{k-m}$$

$$= \left( \frac{1-r}{r} \right)^m r^k \frac{k-1}{m-1}$$

$$\leq (e^{1/2} - 1)^m r^k \frac{k^m}{m!}$$

$$\leq r^k \left( \frac{ek}{m} \right)^m$$

$$= e^{-k \ln(1/r) - m - m \ln(k/m)}. \quad (4)$$
For $k \geq \sqrt{\mu} n$, we have $k/m \geq 1/\sqrt{\mu}$, and thus by (3),

$$
\frac{1}{2} k \ln \frac{1}{r} - m - m \ln \frac{k}{m} = \frac{m \lambda}{4} \left( \frac{k}{m} - \frac{4}{\lambda} \left( 1 + \ln \frac{k}{m} \right) \right) > 0.
$$

(5)

From (4) and (5), we have, for all $k \geq \sqrt{\mu} n$,

$$
\Pr\{Y^{(m)} = k\} \leq e^{-(k \ln(1/r))/2} = e^{-\lambda k/4}.
$$

Thus,

$$
\Pr\{Y^{(m)} \geq \sqrt{\mu} n\} \leq \sum_{k \geq \lceil \sqrt{\mu} n \rceil} e^{-\lambda k/4} \leq e^{-\lambda \sqrt{\mu} n/4} \left( 1 - e^{-\lambda/4} \right).
$$

This proves the lemma. □

We now turn to the proof of Proposition I. Let $D = (d_{ij})$ be an $N \times n$ matrix of real numbers. For any $S \subseteq \{1, 2, \ldots, n\}$ with $|S| = m = \lceil \mu n \rceil$, define an injective function $\omega_{S,D}: S \to \{1, 2, \ldots, N\}$ to be described below. Write $S = \{j_1, j_2, \ldots, j_m\}$, where $j_1 < j_2 < \ldots < j_m$. The following procedure clearly defines an injective function $\omega_{S,D}$:

**Procedure ASSIGN** $(D, \omega_{S,D})$;

begin $W \leftarrow D$;

for $i = 1$ to $m$ do

begin

find in column $j_i$ of $W$ a smallest entry $d_{ij_i}$

(in case of ties, pick the smallest qualified $i$);

set $\omega_{S,D}(j_i) \leftarrow i$;

set to $\infty$ all entries in row $i$ of $W$;

end

end ASSIGN.

Take a random $D$ distributed according to $q_{h,N}$, and let $a_n(S) = \Pr\{\alpha(D, S, \omega_{D,S}) \geq \sqrt{\mu} n\}$. We will prove that for any $S$ with $|S| = m$,

$$
a_n(S) \leq e^{-\lambda \sqrt{\mu} n/4} \left( 1 - e^{-\lambda/4} \right).
$$

(6)

To prove (6), we analyze procedure ASSIGN in the next two lemmas. Fix $S$. For a random $D$, let $I_i$ denote the random variable corresponding to the value $i$ assigned to $\omega_{S,D}(j_i)$, and let $X_i$ be the random variable for $d_{ij_i}$, $1 \leq i \leq m$. Thus, $\alpha(D, S, \omega_{D,S}) = \sum_{1 \leq i \leq m} X_i$. For each
0 \leq \ell < m$, let $M_\ell$ be the set of all $\bar{x} = (x_1, x_2, \ldots, x_\ell)$ where $x_i$ are positive integers satisfying $\sum_{1 \leq j \leq \ell} x_i \leq \sqrt{\mu \ n}$; let $L_\ell$ be the set of all $\bar{i} = (i_1, i_2, \ldots, i_\ell)$ where $i_t \in \{1, 2, \ldots, N\}$ are distinct integers. For any $\bar{x} = (x_1, x_2, \ldots, x_\ell) \in M_\ell$, $\bar{i} = (i_1, i_2, \ldots, i_\ell) \in L_\ell$, and integer $k \geq 1$, let $\delta_{\ell}(\bar{x}, \bar{i}, k) = \Pr\{X_{\ell+1} \geq k | X_{\ell} = x_t, I_{\ell} = i_t \text{ for } 1 \leq t \leq \ell\}.

Lemma 4 $\delta_{\ell}(\bar{x}, \bar{i}, k) \leq e^{-\lambda(k-1)/2}$.

**Proof.** For each $1 \leq s \leq N$, let $B_s$ be the set of $(a_1, a_2, \ldots, a_n) \in \Sigma_n$ such that, for all $1 \leq t \leq \ell$, the following is true: $a_{j_t} > x_t$ if $s < i_t$, $a_{j_t} = x_t$ if $s = i_t$, and $a_{j_t} \geq x_t$ if $s > i_t$. We further partition each $B_s$ into $B_{s,1} \cup B_{s,2} \cup \ldots \cup B_{s,n}$, where $B_{s,k'}$ consists of those $(a_1, a_2, \ldots, a_n) \in B_s$ with $a_{j_{t+1}} = k'$. It is easily verified that an $N \times n$ matrix $D = (d_{ij})$ satisfies $X_t = x_t$, $I_t = i_t$ for $1 \leq t \leq \ell$ if and only if $\bar{d}_s \in B_s$ for all $1 \leq s \leq N$, where $\bar{d}_s = (d_{s,1}, d_{s,2}, \ldots, d_{s,n})$; also $D$ satisfies $X_t = x_t$, $I_t = i_t$ for $1 \leq t \leq \ell$ and $X_{\ell+1} \geq k$ if and only if $\bar{d}_i \in B_i$ for $1 \leq t \leq \ell$ and $\bar{d}_s \in \bigcup_{k' \leq k} B_{s,k'}$ for all $s \neq i_1, i_2, \ldots, i_\ell$. For a random $D$ distributed according to $q_{h,N}$, all rows $\bar{d}_s$ are independently generated, and thus,

$$
\delta_{\ell}(\bar{x}, \bar{i}, k) = \prod_{s \neq i_1, i_2, \ldots, i_\ell} \Pr\{\bar{d}_s \in \bigcup_{k' \leq k} B_{s,k'} | \bar{d}_s \in B_s\}.
$$

$$
= \prod_{s \neq i_1, i_2, \ldots, i_\ell} \left(1 - \Pr\{\bar{d}_s \in \bigcup_{1 \leq k' < k} B_{s,k'} | \bar{d}_s \in B_s\}\right)\nonumber
$$

$$
\leq \prod_{s \neq i_1, i_2, \ldots, i_\ell} \left(1 - \Pr\{\bar{d}_s \in \bigcup_{1 \leq k' < k} B_{s,k'}\}\right)\nonumber
$$

$$
= \prod_{s \neq i_1, i_2, \ldots, i_\ell} \left(1 - \sum_{1 \leq k' < k} \Pr\{\bar{d}_s \in B_{s,k'}\}\right).\quad (7)
$$

Now,

$$
\Pr\{\bar{d}_s \in B_{s,k'}\} \geq \Pr\{d_{s,j_{t+1}} = k'\} - \sum_{1 \leq t \leq \ell} \sum_{1 \leq z \leq x_t \atop z \neq k'} \Pr\{(d_{s,j_{t+1}} = k') \land (d_{s,j_t} = z)\}.\quad (8)
$$

Since $h$ is a doubly-dispersive hash function, we have, for $k' \neq z$,

$$
\Pr\{(d_{s,j_{t+1}} = k') \land (d_{s,j_t} = z)\} = \frac{1}{n(n-1)}.\quad (9)
$$

Let $v \in \{1, 2, \ldots, n\} - \{j_{t+1}\}$, then we have

$$
\Pr\{d_{s,j_{t+1}} = k'\} = \sum_{1 \leq k' \leq n \atop k' \neq k} \Pr\{d_{s,j_{t+1}} = k' \land (d_{s,v} = k)\} = \frac{n-1}{n(n-1)} = \frac{1}{n}.\quad (10)
$$
As $\tilde{x} \in M_{\ell}$, it follows from (8), (9) and (10) that
\[
\Pr\{d_s \in B_{s,k}\} \geq \frac{1}{n} - \frac{\sum_{1 \leq t \leq \ell} x_t}{n(n-1)} \\
\geq \frac{1}{n} - \frac{\sqrt{\mu}}{n-1}.
\] (11)

From (7) and (11) we obtain
\[
\delta_{\ell}(\tilde{x}, \bar{t}, k) \leq (1 - (k-1)\left(\frac{1}{n} - \frac{\sqrt{\mu}}{n-1}\right))^{N-\ell} \\
\leq e^{-k(l-1)(\frac{1}{n} - \frac{\sqrt{\mu}}{n-1})(N-\ell)}.
\] (12)

As $N - \ell \geq \lambda n - \mu n - 1 \geq \frac{3}{4} \lambda n$, we obtain from (12) $\delta_{\ell}(\tilde{x}, \bar{t}, k) \leq e^{-\lambda(k-1)/2}$. This proves Lemma 4. \(\square\)

Now consider both the sequence of random variables $X_1, X_2, \ldots, X_m$ under discussion and the infinite sequence $Y$ of random variables $Y_1, Y_2, Y_3, \ldots$ defined at the beginning of this section. It is clear that, for all $i, k \geq 1$, $\Pr\{Y_i \geq k\} = r^{k-1} = e^{-\lambda(k-1)/2}$. It follows from Lemma 4 that, for any $0 \leq \ell < m$, $k \geq 1$, and $\tilde{x} = (x_1, x_2, \ldots, x_\ell) \in M_{\ell}$,
\[
\Pr\{X_{\ell+1} \geq k | X_t = x_t, 1 \leq t \leq \ell\} \leq \Pr\{Y_{\ell+1} \geq k\}.
\] (13)

**Lemma 5** For any integer $s \leq \mu n$,
\[
\Pr\left\{\sum_{1 \leq t \leq m} X_t \geq s\right\} \leq \Pr\left\{\sum_{1 \leq t \leq m} Y_t \geq s\right\}.
\]

**Proof.** We will prove the following more general statement: for any integers $j, t, s$, where $0 < t \leq j$ and $s \leq \mu n$,
\[
\Pr\left\{\sum_{1 \leq i \leq t} X_i + \sum_{t < i \leq j} Y_i \geq s\right\} \leq \Pr\left\{\sum_{1 \leq i \leq t-1} X_i + \sum_{t \leq i \leq j} Y_i \geq s\right\}.
\] (14)

We prove (14) by induction on $j \geq 1$.

If $j = 1$, then (14) follows from (13). Now, let $j_0 > 1$, and assume that we have proved (14) for all $j < j_0$; we need to prove it for $j = j_0$. It $t < j_0$, then using the induction hypothesis, we have
\[
\Pr\{X_1 + \ldots + X_t + Y_{t+1} + \ldots + Y_{j_0} \geq s\} \\
= \sum_{k \geq 1} \Pr\{Y_{t+1} + \ldots + Y_{j_0} = k\} \cdot \Pr\{X_1 + \ldots + X_t \geq s - k\} \\
\leq \sum_{k \geq 1} \Pr\{Y_{t+1} + \ldots + Y_{j_0} = k\} \cdot \Pr\{X_1 + \ldots + X_{t-1} + Y_t \geq s - k\} \\
= \Pr\{X_1 + \ldots + X_{t-1} + Y_t + \ldots + Y_{j_0} \geq s\}.
\]

9
Thus the inequality (14) is true for $j = j_0$ in this case.

If $t = j_0$, then

$$\Pr\{X_1 + \ldots + X_{j_0} \geq s\} = \sum_{k \geq 1} \Pr\{X_1 + \ldots + X_{j_0-1} = k\} \cdot \Pr\{X_{j_0} \geq s - k \mid X_1 + \ldots + X_{j_0-1} = k\}$$

(15)

Now, let $M_k$ denote the set of $(x_1, x_2, \ldots, x_{j_0-1})$ with all integers $x_i > 0$ and $\sum_{1 \leq i \leq j_0-1} x_i = k'.

Using (13), we have

$$\Pr\{X_{j_0} \geq s - k \mid \sum_{1 \leq t < j_0} X_t = k\} = \sum_{(x_1, \ldots, x_{j_0-1}) \in M_k} \Pr\{\land_{1 \leq t < j_0} (X_t = x_t) \mid \sum_{1 \leq t < j_0} X_t = k\} \cdot \Pr\{X_{j_0} \geq s - k \mid \land_{1 \leq t < j_0} (X_t = x_t)\}$$

$$\leq \sum_{(x_1, \ldots, x_{j_0-1}) \in M_k} \Pr\{\land_{1 \leq t < j_0} (X_t = x_t) \mid \sum_{1 \leq t < j_0} X_t = k\} \cdot \Pr\{Y_{j_0} \geq s - k\}$$

$$= \Pr\{Y_{j_0} \geq s - k\} .$$

(16)

From (15) and (16) we obtain

$$\Pr\{X_1 + \ldots + X_{j_0} \geq s\} \leq \Pr\{X_1 + \ldots + X_{j_0-1} + Y_{j_0} \geq s\} .$$

This completes the inductive proof of (14). We have proved Lemma 5. □

From Lemma 3 and Lemma 5 we obtain, for $m = \lfloor \mu n \rfloor$,

$$\Pr\left\{ \sum_{1 \leq t \leq m} X_t \geq \sqrt{\mu n} \right\} \leq \frac{e^{-\lambda \sqrt{n}/4}}{1 - e^{-\lambda/4}} .$$

(17)

This immediately gives (6), as $\alpha(D, S, \omega_{D, S}) = \sum_{1 \leq t \leq m} X_t$.

We will now use (6) to complete the proof of Proposition 1. Take a random $D$ distributed according to $q_{\lambda, \nu}$. Let $\nu$ denote the probability that there exists an $S \subseteq \{1, 2, \ldots, n\}$ such that $|S| = m$ and $\alpha(D, S) \geq \sqrt{\mu n}$. We infer from (6) that, for each $S$ with $|S| = m$,

$$\Pr\{\alpha(D, S) \geq \sqrt{\mu n}\} \leq \frac{e^{-\lambda \sqrt{n}/4}}{1 - e^{-\lambda/4}} .$$

(18)

It follows that

$$\nu \leq \sum_{S \mid |S| = m} \Pr\{\alpha(D, S) \geq \sqrt{\mu n}\}$$

$$\leq \left( \begin{array}{c} n \\ m \end{array} \right) \frac{e^{-\lambda \sqrt{n}/4}}{1 - e^{-\lambda/4}}$$

$$\leq \left( \frac{e}{m} \right)^m \frac{e^{-\lambda \sqrt{n}/4}}{1 - e^{-\lambda/4}}$$

$$= e^{m(1 + \ln(n/m))} \frac{e^{-\lambda \sqrt{n}/4}}{1 - e^{-\lambda/4}} .$$

(19)
Using (3) with \( x = (n/m)^{1/2} \), it is elementary to check that
\[
-\frac{\lambda \sqrt{\mu} n}{8} + m \left( 1 + \ln \frac{n}{m} \right) < 0 .
\]
Thus, we have from (19)
\[
\nu \leq \frac{e^{-\lambda \sqrt{\mu} n/8}}{1 - e^{-\lambda/4}} = c_5 e^n.
\]
This implies Proposition I, since it is clear that \( \Pr\{\neg Z_I\} \leq \nu \).

4 Proof of Proposition II

We will prove
\[
\Pr\{\neg Z_{II}\} \leq n \left( \frac{1}{2} \right)^{n/2} ,
\]
and
\[
\Pr\{\neg Z_{III}\} \leq \frac{n}{2^n} + \frac{1}{n^2} + \frac{\ln n}{n^{4/5}} ,
\]
from which Proposition II follows immediately. The techniques used in this section involve adaptations of the methods employed in [Y].

For any \( 1 \leq m \leq n \), let \( T_m \) be the family of all \( T \subseteq \{1, 2, \ldots, n\} \) with \( |T| = m \). Let \( T \in T_m \). Take a random \( A \in A_n \) distributed according to \( q_h \). Define random variables \( Z_{T,i} \), \( 1 \leq i \leq n \), such that \( Z_{T,i} = 1 \) if \( I_k(i, A) \cap T \neq \emptyset \), and \( Z_{T,i} = 0 \) otherwise. Then \( \sum_{1 \leq i \leq n} Z_{T,i} \) takes on the value \( |V_k(A, T)| \). Let \( \beta_T = \Pr\{Z_{T,1} = 0\} \). Clearly, one has then \( \beta_T = \Pr\{Z_{T,i} = 0\} \) for all \( 1 \leq i \leq n \).

Lemma 6 Suppose \( 2 \leq m \leq n \) and \( T \in T_m \). Then \( \beta_T \leq 4n/(mk) \).

Proof. Take a random permutation \( \sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n)) \in \Sigma_n \) distributed according to \( p_h \). Then \( \beta_T \) is equal to the probability that \( \{\sigma(1), \sigma(2), \ldots, \sigma(k)\} \cap T = \emptyset \). Define random variables \( F_j \), \( 1 \leq j \leq n \), such that \( F_j = 1 \) if \( j \in \{\sigma(1), \sigma(2), \ldots, \sigma(k)\} \) and \( F_j = 0 \) otherwise; let \( F = \sum_{j \in T} F_j \). We have
\[
\beta_T = \Pr\{F = 0\} .
\]

Now, as \( h \) is doubly dispersive, we have for all \( 1 \leq i < j \leq n \),
\[
E(F_i F_j) = \Pr\{(\sigma(s) = i) \land (\sigma(t) = j) \text{ for some } 1 \leq s \neq t \leq k\}
\]
\[
= \sum_{1 \leq s \neq t \leq k} \Pr\{(\sigma(s) = i) \land (\sigma(t) = j)\}
\]
\[
= \frac{k(k-1)}{n(n-1)} ,
\]

11
and, letting \( \ell \) be an arbitrary element of \( \{1, 2, \ldots, n\} - \{i\} \), we have

\[
E(F_i) = \Pr \left\{ (\sigma(s) = i) \land (\sigma(t) = \ell) \text{ for some } 1 \leq s \leq k, 1 \leq t \leq n, t \neq s \right\}
= \sum_{1 \leq s \leq k} \sum_{1 \leq t \leq n, t \neq s} \Pr \left\{ (\sigma(s) = i) \land (\sigma(t) = \ell) \right\}
= \frac{k(n-1)}{n(n-1)}
= \frac{k}{n}.
\]

(24)

From (23) and (24), we obtain

\[
E(F) = \sum_{j \in T} E(F_j)
= \frac{mk}{n},
\]

(25)

and, noting that \( F_j^2 = F_j \),

\[
\text{Var}(F) = \sum_{j \in T} E(F_j^2) - \sum_{j \in T} (E(F_j))^2 + 2 \sum_{i < j \atop i, j \in T} \left( E(F_i F_j) - E(F_i) E(F_j) \right)
= \sum_{j \in T} E(F_j) - \sum_{j \in T} (E(F_j))^2 + 2 \binom{m}{2} \left( \frac{k(k-1)}{n(n-1)} - \frac{k^2}{n^2} \right)
= \frac{mk}{n} \left( 1 - \frac{k}{n} \right) \left( 1 - \frac{m-1}{n-1} \right)
\leq \frac{mk}{n}.
\]

(26)

Chebycheff's Inequality then gives

\[
\Pr \{ F = 0 \} \leq \Pr \left\{ \left| \frac{F - \frac{mk}{n}}{\frac{mk}{n}} \right| > \frac{1}{2} \frac{mk}{n} \right\}
\leq \frac{\frac{(mk/n)}{\left(\frac{mk}{2n}\right)^2}}{4n}
= \frac{4n}{mk}.
\]

(27)

Lemma 6 follows from (22) and (27) immediately. \( \Box \)

We remark that Lemma 6 is valid for all \( 1 \leq k \leq n \). We now prove (20). Let \( k = \lceil \gamma^{-c} \rceil \). Let \( T \in T_m \). As \( Z_{T,i} \), \( 1 \leq i \leq n \), are independent random variables, we obtain, with the help of Lemma 6,

\[
\Pr \left\{ \left| V_k(A, T) \right| < 2|T| \right\} = \Pr \left\{ \sum_{1 \leq i \leq n} Z_{T,i} < 2m \right\}
\]
\[ \leq \binom{n}{n - 2m} \beta_T^{n - 2m} \]
\[ \leq \left( \frac{n}{2m} \right) \left( \frac{4n}{mk} \right)^{n - 2m}. \]  

(28)

It follows that

\[
\begin{align*}
\Pr\{\neg Z_{II} \} & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \sum_{T \in T_m} \Pr\{|V_k(A, T)| < 2|T|\} \\ & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \binom{n}{m} \left( \frac{n}{2m} \right) \left( \frac{4n}{mk} \right)^{n - 2m} \\ & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \frac{n^m}{m!} \frac{n^{2m}}{(2m)!} \left( \frac{4n}{mk} \right)^{n - 2m} \\ & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \left( \frac{ne}{m} \right)^m \left( \frac{ne}{2m} \right)^{2m} \left( \frac{4n}{mk} \right)^{n - 2m} \\ & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \left( \frac{n}{m} \right)^{n+m} \left( \frac{4e}{k} \right)^{n - 2m} \\ & \leq \sum_{\gamma n \leq m \leq 2\gamma n} \frac{1}{\gamma^{n+m}} \left( \frac{4e}{k} \right)^{n - 2m} \\ & \leq n \left( \frac{4e}{k} \right)^{n - 2m} \\ & \leq n \left( \frac{1}{2} \right)^{n/2}.
\end{align*}
\]

This proves (20).

We now turn to the proof of (21). Define \( n_1 = 2\gamma n, n_2 = \left( 1 - \frac{1}{(32e)^8} \right) n + 1, n_3 = n - \frac{1}{10} \ln n, \) and \( n_4 = n. \) For \( 1 \leq i \leq 3, \) let \( T^{(i)} = \bigcup_{n_i \leq m \leq n_{i+1}} T_m. \) Take a random \( A \in A_n \) distributed according to \( q_n, \) let \( G_i \) be the event that there exists a \( T \in T^{(i)} \) with \( |V_k(A, T)| < |T|. \) As \( \neg Z_{III} = G_1 \lor G_2 \lor G_3, \) we need only prove the following equations:

\[
\begin{align*}
\Pr\{G_1\} & \leq \frac{n}{2^n}, \\
\Pr\{G_2\} & \leq \frac{1}{n^e}, \\
\Pr\{G_3\} & \leq \frac{\ln n}{n^{4/5}}.
\end{align*}
\]

(29)  (30)  (31)

Similar to (28), we have from Lemma 6 that, for \( T \in T_m, \)

\[
\Pr\{|V_k(A, T)| < T\} = \Pr\left\{ \sum_{1 \leq i \leq n} \gamma_{T,i} < m \right\} \leq \left( \frac{n}{n - m + 1} \right) \beta_T^{n-m+1}
\]

13
\[
\Pr\{G_1\} \leq \sum_{n_1 < m \leq n_2} \sum_{T \in T_m} \Pr\{|V_k(A,T)| < T\} \\
\leq \sum_{n_1 < m \leq n_2} \binom{n}{m} \binom{n}{m-1} \left(\frac{4n}{mk}\right)^{n-m+1} \\
\leq \sum_{n_1 < m \leq n_2} \left(\frac{ne}{m}\right)^n \left(\frac{ne}{m-1}\right)^{m-1} \left(\frac{4n}{mk}\right)^{n-m+1} \\
\leq \sum_{n_1 < m \leq n_2} \left(\frac{2ne}{n_1}\right)^{2n_2} \left(\frac{4n}{kn_1}\right)^{n-n_2} \\
= \sum_{n_1 < m \leq n_2} \left(\frac{e}{\gamma}\right)^{2n_2} \left(\frac{2}{k\gamma}\right)^{n/(32e)^8} \\
\leq \sum_{n_1 < m \leq n_2} \left(\frac{e}{\gamma}\right)^{2n} \left(\frac{2\gamma}{\gamma}\right)^{n/(32e)^8} \\
\leq \sum_{n_1 < m \leq n_2} \frac{e^{2n} \gamma(32e)^2 n}{\gamma^{2n}} \\
\leq \sum_{n_1 < m \leq n_2} \frac{1}{2^n} \\
\leq \frac{n}{2^n}.
\]

Thus, \(\Pr\{G_1\} \leq \frac{n}{2^n}\).

This proves (29).

To prepare for the proof of (30), we take a random \(A \in \mathcal{A}_n\) distributed according to \(q_h\), and let \(D_s\) be the event that there exists \(s\) integers \(1 \leq i_1 < i_2 < \ldots < i_s \leq n\) such that \(\bigcup_{1 \leq \ell \leq s} I_3(i_\ell, A)\) \(< s\).

**Lemma 7** For \(1 \leq s \leq n/(32e)^8\), \(\Pr\{D_s\} \leq 1/2^s\).

**Proof.** This result was derived for double hashing in [Y, equation (12)]; the proof extends straightforwardly to any hash function \(h\) that is doubly dispersive. \(\square\)

**Lemma 8** Let \(2 \leq m < n, 1 \leq t \leq n\). If there exists \(T \in T_m\) with \(|V_t(A,T)| < |T|\), then there exist \(n - m + 1\) integers \(1 \leq i_1 < i_2 < \ldots < i_{n-m+1} \leq n\) such that \(\bigcup_{1 \leq \ell \leq n-m+1} I_t(i_\ell, A)\) \(< n - m + 1\).

**Proof.** Suppose \(|V_t(A,T)| < |T|\). Let \(W = \{1,2,\ldots,n\} - V_t(A,T)\). Then \(|W| > n - m\). Let \(\{i_1, i_2, \ldots, i_{n-m+1}\} \subseteq W\) with \(i_1 < i_2 < \ldots < i_{n-m+1}\). Then \(\bigcup_{1 \leq \ell \leq n-m+1} I_t(i_\ell, A)\) \(\cap T = \emptyset\). Hence, \(\bigcup_{1 \leq \ell \leq n-m+1} I_t(i_\ell, A)\) \(< n - |T| = n - m\). \(\square\)
We now prove (30). Using Lemmas 7 and 8, we obtain
\[
\Pr\{G_2\} \leq \sum_{n_2 < m \leq n_3} \Pr\{\exists T \in \mathcal{T}_m \text{ with } |V_k(A, T)| < |T|\}
\leq \sum_{n_2 < m \leq n_3} \Pr\{\exists T \in \mathcal{T}_m \text{ with } |V_3(A, T)| < |T|\}
\leq \sum_{n_2 < m \leq n_3} \Pr\{D_{n-m+1}\}
\leq \sum_{n-n_3+1 \leq s < n-n_3+1} \Pr\{D_s\}
= \sum_{1 + \frac{1}{10} \ln n \leq s < \frac{1}{2(32e)^{\beta}} n} \Pr\{D_s\}
\leq \sum_{s \geq 1 + \frac{1}{10} \ln n} \frac{1}{2^s}
\leq \frac{1}{2^{10} \ln n}
= \frac{1}{n^{\beta}}.
\]

This proves (30).

We now turn to the proof of (31). Let \( n_3 < m < n \), and \( T \in \mathcal{T}_m \). Take a random \( \sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n)) \in \Sigma_m \) distributed according to \( p_h \). Then,
\[
\beta_T = \Pr\{\{\sigma(1), \sigma(2), \ldots, \sigma(k)\} \cap T = \emptyset\}
\leq \Pr\{\{\sigma(1), \sigma(2)\} \cap T = \emptyset\}
= \Pr\{\{\sigma(1), \sigma(2)\} \subseteq \{1, 2, \ldots, n\} - T\}
= \sum_{i,j \neq T \atop i \neq j} \Pr\{ (\sigma(1) = i) \land (\sigma(2) = j) \}.
\]
Since \( h \) is doubly dispersive, we have then
\[
\beta_T \leq \frac{(n - m)(n - m - 1)}{n(n - 1)}
\leq \left( \frac{n - m}{n} \right)^2.
\]
Thus, writing \( s = n - m \), we obtain
\[
\Pr\{|V_k(A, T)| < |T|\} = \Pr\left\{ \sum_{1 \leq i \leq n} Z_{T,i} < m \right\}
\leq \left( \frac{n}{n - m + 1} \right) \beta_T^{n-m+1}
\leq \left( \frac{n}{s + 1} \right) \left( \frac{s}{n} \right)^{2(s+1)}.
\]

(34)
Clearly, (34) is also valid for $m = n$ and $T = \{1, 2, \ldots, n\}$. It follows that

$$
\Pr\{G_3\} \leq \sum_{n_3 < m \leq n} \sum_{T \in \mathcal{T}_m} \Pr\{|V_k(A, T)| < |T|\}
$$

$$
\leq \sum_{0 \leq s < \frac{1}{10} \ln n} \sum_{T \in \mathcal{T}_{n-s}} \binom{n}{s+1} \left(\frac{s}{n}\right)^2 (s+1) (\frac{s}{n})^{2(s+1)}
$$

$$
= \sum_{0 \leq s < \frac{1}{10} \ln n} \binom{n}{s} \binom{n}{s+1} \left(\frac{s}{n}\right)^2 (s+1) (\frac{s}{n})^{2(s+1)}
$$

$$
\leq \sum_{1 \leq s < \frac{1}{10} \ln n} \left(\frac{ne}{s}\right)^2 \frac{n}{s+1} \left(\frac{s}{n}\right)^2 s^{2s+2} n^{2s+2}
$$

$$
\leq \sum_{1 \leq s < \frac{1}{10} \ln n} s e^{2s} \frac{n}{n^{2s+2}}
$$

$$
\leq \frac{\ln n}{n^{4/5}}.
$$

This proves (31).

We have completed the proof of Proposition II.

5 Derivation of Corollary

In Section 4, we proved two equalities (23) and (24) which we summarize below. Let $k$ be any integer satisfying $1 \leq k \leq n$. Take a random permutation $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n)) \in \Sigma_n$ distributed according to $p_h$, where $h \in \mathcal{H}_n$ is doubly dispersive. Define random variables $F_j$, $1 \leq j \leq n$, such that $F_j = 1$ if $j \in \{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$ and $F_j = 0$ otherwise. Then, for all $1 \leq i \leq n$,

$$
E(F_i) = \frac{k}{n},
$$

and, for all $1 \leq i \neq j \leq n$,

$$
E(F_i F_j) = \frac{k(k-1)}{n(n-1)}.
$$

In [Y], it was proved that, for the double hashing function $h$, if we take a random $A = (a_{ij}) \in \mathcal{A}_n$ distributed according to $q_h$, then with probability $1 - \frac{c_6}{n^5}$, there exists a $\sigma \in \Sigma_n$ satisfying $\max_i a_{i,\sigma(i)} \leq \lambda_1 \ln n$, where $c_6$ and $\lambda_1$ are positive constants. Clearly, when such $\sigma$
exists, \( C(A) \leq \lambda_1 \ln n \). The proof in [Y] in fact holds for any hash function \( h \) satisfying the two equalities (35) and (36), and hence for any doubly-dispersive function \( h \).

Let \( c_4 = c_1 + c_6 + c_2 \lambda_1 \max_{n \geq 1} (\ln n / n^s) \). The above discussion and Theorem 1 immediately give

\[
E(C(A)) \leq \Pr\{C(A) \leq c_1\} c_1 + \Pr\{\lambda_1 \ln n \geq C(A) > c_1\} \lambda_1 \ln n + \Pr\{C(A) > \lambda_1 \ln n\} n
\leq c_1 + \frac{c_2}{n^s} \lambda_1 \ln n + \frac{c_6}{n^5} n
\leq c_4.
\]

This proves the corollary to Theorem 1.

6 Remarks

One motivation for this work is to investigate how good double hashing is, as a substitute for uniform hashing. Guibas and Szemerédi [GS] showed that double hashing has a performance that is virtually indistinguishable from uniform hashing, when hashing is used in the standard way to maintain a dynamic hash table, at least up to a certain load factor. In Yao [Y], it was shown that double hashing has asymptotically, up to a multiplicative constant, the same worst case retrieval time as uniform hashing, when hashing is employed to build a static dictionary. In the present paper, we have proved that this is also the case, when the average retrieval time of the static dictionary is adopted as the performance measure. From an application viewpoint, our result is not conclusive, since the constants involved in Theorem 1 and its corollary are very large. A challenging open problem is to derive tight bounds on \( E(C(A)) \) for uniform hashing and double hashing, so that their performance can be compared satisfactorily. For example, can one prove that \( E(C(A)) < 10 \) for double hashing? Simulation results in Gonnet and Munro [GM] indicate that \( E(C(A)) \) is close to the value 3. For uniform hashing, it is possible to prove reasonable upper bounds on \( E(C(A)) \) using ideas from Walkup [W], but an accurate determination of \( E(C(A)) \), say within 20\%, seems to be an interesting but difficult open problem.

Appendix: Proof of an Inequality

In this Appendix, we will prove Inequality (3) in Section 3 of this paper. Let \( \lambda, \mu \) be constants such that \( 0 < \mu < 10^{-4} \lambda^4 < \lambda < 1 \). We will prove that, for all \( x \geq 1/\sqrt{\mu} \),

\[
x > \frac{8}{\lambda} (1 + 2 \ln x).
\]  

(A1)
Let \( f(x) = x - \frac{8}{\lambda}(1 + 2\ln x) \). To prove (A1), it suffices to show that

\[
f\left(\frac{100}{\lambda^2}\right) > 0 ,
\]

(A2)

and for all \( x \geq \frac{100}{\lambda^2} \),

\[
f'(x) \geq 0 .
\]

(A3)

Now, \( f'(x) = 1 - \frac{16}{\lambda^2} \), and (A3) clearly holds. To prove (A2), observe that the function

\[
f\left(\frac{100}{\lambda^2}\right) = \frac{4}{\lambda} g(\lambda),
\]

where \( g(\lambda) \) is defined as \( \frac{25}{\lambda} - 2 - 8\ln \frac{10}{\lambda} \); \( g(\lambda) \) satisfies \( g(1) > 0 \) and, for all \( 0 < \lambda \leq 1 \), \( g'(\lambda) = -\frac{25}{\lambda^2} + \frac{8}{\lambda} = \frac{8}{\lambda}(1 - \frac{25}{8\lambda}) < 0 \). It follows that \( g(\lambda) > 0 \) for all \( 0 < \lambda \leq 1 \), and hence \( f\left(\frac{100}{\lambda^2}\right) > 0 \). This completes the proof of (A1).

References


