Discrete Applied Mathematics 289 (2021) 374–382

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

New bounds for perfect k-hashing

Simone Costa [a,*], Marco Dalai [b]

[a] DICATAM - Sez. Matematica, Università degli Studi di Brescia, Via Branze 43, I-25123 Brescia, Italy
[b] DII, Università degli Studi di Brescia, Via Branze 38, I-25123 Brescia, Italy

ARTICLE INFO

Article history:
Received 25 February 2020
Received in revised form 22 October 2020
Accepted 2 November 2020
Available online 20 November 2020

Keywords:
Trifference
Perfect k-hashing

ABSTRACT

Let $C \subseteq \{1, \ldots, k\}^n$ be such that for any k distinct elements of C there exists a coordinate where they all differ simultaneously. Fredman and Komlós studied upper and lower bounds on the largest cardinality of such a set C, in particular proving that as $n \to \infty$, $|C| \leq \exp(nk!/k^{k-1} + o(n))$. Improvements over this result were first derived by different authors for $k = 4$. More recently, Guruswami and Riazanov showed that the coefficient $k!/k^{k-1}$ is certainly not tight for any $k > 3$ and provided explicit improvements for $k = 5, 6$, which are immediately extendable to $k > 6$ modulo a conjecture on the maxima of certain polynomials.

In this paper, we first prove their conjecture, completing the derivation of their new bound for any k. Then, we develop a different method which gives further substantial improvements for $k = 5, 6$.

1. Introduction

For positive integers $k \geq 2$ and $n \geq 1$, consider a subset $C \subset \{1, \ldots, k\}^n$ with the property that for any k distinct elements of C there exists a coordinate where they all differ. We call such a set a *perfect k-hash code* of length n, or simply k-hash for brevity. The name is motivated by the idea that if each coordinate of C is interpreted as a k-hash function on a set U of cardinality $|C|$, then any k elements of U are hashed onto $\{1, 2, \ldots, k\}$ by at least one function.

Determining the largest possible cardinality of such a set C as a function of k and n is a classic combinatorial problem in theoretical computer science. One standard formulation is to study, for fixed k, the growth of the largest possible $|C|$ as n goes to infinity. It is known that $|C|$ grows exponentially in n. Then one usually defines the *rate* of the code as[1]

$$R = \frac{\log |C|}{n} \tag{1}$$

and asks for bounds on the rate of codes of maximal cardinality as $n \to \infty$, that is

$$R_k = \limsup R,$$

where the limsup is over all k-hash codes as n goes to infinity. This formulation of the problem can also be cast as a problem, in information theory, of determining the zero-error capacity under list decoding for certain channels. Few lower bounds on R_k are known. First results in this sense were given by [8,7] and a better bound was derived by [13] for $k = 3$. More recently, new lower bounds were derived in [17] for many other values of k (actually infinitely many).

* Corresponding author.
E-mail addresses: simone.costa@unibs.it (S. Costa), marco.dalai@unibs.it (M. Dalai).

[1] Here and in the whole paper $\log x$ is understood to be in base two.

https://doi.org/10.1016/j.dam.2020.11.001
0166-218X/

In this paper we consider upper bounds on R_k. A simple packing argument (see [7]) shows that for all $k \geq 2$ one has $R_k \leq \log(k/(k-1))$. For $k = 3$, the simplest non-trivial case, this evaluates to $\log(3/2) \approx 0.5850$ and is still the best known upper bound to date (the best lower bound is $1/4 \log(9/5) \approx 0.212$). For $k \geq 4$, the first important result was derived by Fredman and Komlós [8], who proved that

$$R_k \leq k!/k^{k-1} . \tag{2}$$

We also refer to [11,13] and [12] where the Fredman–Komlós bound (and some generalizations to hypergraphs) has been cast using the language of graph entropy and to [15] where a simple probabilistic proof has been presented. Improvements were obtained for $k = 4$ in [1,2] and more recently in [5,6]. The most recent progress we are aware of was obtained in [9] where the Fredman–Komlós bound is proved to be non-tight for any $k \geq 5$. Explicit computations of new better values were provided there for $k = 5, 6$, with a technique which extends to larger k modulo a conjecture on the maxima of certain polynomials. Other recent papers on this topic that deserve to be recalled are [3], where a related problem is studied in a different asymptotic regime, and [4] where the authors attempt to use the polynomial method to upperbound R_3 and they state some limitations of this method.

In this paper we make some progress on this problem. We first prove the conjecture formulated in [9] and thus complete their new explicit upper bounds on R_k which beat the Fredman–Komlós bound for all $k \geq 5$. Our main contribution is then to expand on the idea used in [6] to derive a substantial improvement for $k = 5, 6$. In Section 2 we give a brief summary of the approaches used in [6] and in [9], upon which we build our contribution. In Section 3 we prove the conjecture stated in [9] and give a numerical evaluation of the ensuing bound for $k > 6$. In Section 4 we present our improvement for $k = 5, 6$.

2. Background

The bounds presented in [8,2,6] and [9] can all be derived by starting with the following Lemma on graph covering (see [16]).

Lemma 1 (*Hansel [10]*)**.** *Let K_r be a complete graph on r vertices and let $G_1, \ldots, G_m$ be bipartite graphs on those same vertices such that $\cup_i G_i = K_r$. Let finally $\tau(G_i)$ represent the number of non-isolated vertices in G_i. Then*

$$\sum_{i=1}^{m} \tau(G_i) \geq \log r . \tag{3}$$

The connection with k-hashing comes from the following application. Given a k-hash code C, fix any $(k-2)$-elements subset $\{x_1, x_2, \ldots, x_{k-2}\}$ in C. For any coordinate i let $G_i^{x_1,\ldots,x_{k-2}}$ be the bipartite graph with vertex set $G \setminus \{x_1, x_2, \ldots, x_{k-2}\}$ and edge set

$$E = \left\{(v, w) : x_{1,i}, x_{2,i}, \ldots, x_{k-2,i}, v, w \text{ are all distinct}\right\} . \tag{4}$$

Then, since C is a k-hash code, we note that $\bigcup_i G_i^{x_1,\ldots,x_{k-2}}$ is the complete graph on $G \setminus \{x_1, x_2, \ldots, x_{k-2}\}$ and so

$$\sum_{i=1}^{n} \tau(G_i^{x_1,\ldots,x_{k-2}}) \geq \log(|C| - k + 2) . \tag{5}$$

This inequality can be used to prove upper bounds on $|C|$. Since it holds for any choice of $x_1, x_2, \ldots, x_{k-2}$, one can show that the right hand side is small by proving that left hand side cannot be too large for all possible choices of $x_1, x_2, \ldots, x_{k-2}$. One can either use it for some specific choice or take expectation over any random selection.

Let f_i be probability distribution of the ith coordinate of C, that is, $f_{i,a}$ is the fraction of elements of C whose ith coordinate is a. Note that the graph in (4) is empty if the $x_{1,i}, x_{2,i}, \ldots, x_{k-2,i}$ are not all distinct. We will say in this case that $x_1, x_2, \ldots, x_{k-2}$ *collide* in coordinate i. Then, we have

$$\tau(G_i^{x_1,\ldots,x_{k-2}}) = \begin{cases} 0 & x_1, \ldots, x_{k-2} \text{ collide in coordinate } i \\ \left(\frac{|C|}{|C|-k+2}\right)\left(1 - \sum_{j=1}^{k-2} f_{i,x_{ji}}\right) & \text{otherwise} \end{cases} \tag{6}$$

So, one can make the left hand side in (5) small by either taking a set $x_1, \ldots, x_{k-2}$ which collide in many coordinates, so forcing the corresponding τ's to zero, or by taking a set which uses "popular" values in many coordinates.

The Fredman–Komlós bound is obtained by taking expectation in (5) over a uniform random extraction of $x_1, x_2, \ldots, x_{k-2}$. By linearity of expectation the computation can be performed over each single coordinate. Denoting with $\mathbb{E}$ the expectation, for large n and $|C|$

$$\begin{aligned} &\mathbb{E}[\tau(G_i^{x_1,\ldots,x_{k-2}})] \\ &\quad = (1 + o(1)) \sum_{\substack{\text{distinct} \\ a_1,\ldots,a_{k-2}}} f_{i,a_1} f_{i,a_2} \cdots f_{i,a_{k-2}} (1 - f_{i,a_1} \cdots - f_{i,a_{k-2}}) \end{aligned}$$

where the coefficient $o(1)$ is due to sampling without replacement. One can show that the worst-case f_i is the uniform distribution, which gives

$$\mathbb{E}[\tau(G_i^{x_1,\dots,x_{k-2}})] \le \frac{k!}{k^{k-1}}\,(1+o(1))\ . \tag{7}$$

The procedures used in [6] and [9] are based on the idea that one can also take $x_1, x_2, \dots, x_{k-2}$ uniformly from a subset $C' \subset C$ which ensures they collide in all coordinates i in some subset $T \subset \{1, 2 \dots, n\}$. Then, if $g_{i,a}$ is the frequency of symbol a in the coordinate $i \notin T$ of C', one has

$$\begin{aligned}&\mathbb{E}[\tau(G_i^{x_1,\dots,x_{k-2}})]\\ &\quad = (1+o(1)) \sum_{\substack{\text{distinct}\\ a_1,\dots,a_{k-2}}} g_{i,a_1} g_{i,a_2} \cdots g_{i,a_{k-2}} (1 - f_{i,a_1} \cdots - f_{i,a_{k-2}})\end{aligned} \tag{8}$$

The worst case g and f here, if taken independently, give in general a value which exceeds the $k!/k^{k-1}$ of (7). In [6], for $k = 4$, it was shown that one can deal with this by also taking C' randomly from a partition of C (based on the values in positions $i \in T$), thus adding an additional (outer) expectation. In that case g_i is also random and constrained to satisfy $\mathbb{E}[g_i] = f_i$. Using some concavity argument it was shown that under this random selection the bound (7) still holds for $i \notin T$, thus gaining on average compared to [8]. However, for $k > 4$ that approach seems infeasible. The idea used in [9] is to suppress the random selection of C' and show that one can carefully choose C' so that $x_1, \dots, x_{k-2}$ collide in a portion of the coordinates large enough to more than compensate the increase in $\mathbb{E}[\tau(G_i^{x_1,\dots,x_{k-2}})]$ for $i \notin T$ obtained in (8) with respect to (7). This leads to a proof that (2) is not tight for all $k > 4$. They also state a conjecture on the maxima of some polynomials (proved there for $k = 5, 6$) whose validity allows to extend their derivation of new numerical bounds to any given $k > 6$.

In the next two sections we present our contribution. First we prove the conjecture formulated by the authors in [9], thus completing their computation of new bounds on R_k for all k. Then, we prove stronger results for $k = 5, 6$. Our idea is based on a symmetrization of (8) which allows us to resurrect the random selection of C' in an effective way, replacing the concavity argument of [6] with new bounds on the maxima of some polynomials.

3. Guruswami–Riazanov bounds

A crucial role in all bounds discussed in this paper is played by the sum appearing in Eq. (8). We simplify the notation and set, for general probability vectors $g = (g_1, \dots, g_k)$ and $f = (f_1, \dots, f_k)$,

$$\psi(g, f) = \sum_{\sigma \in S_k} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)} f_{\sigma(k-1)}\,, \tag{9}$$

observing that Eq. (8) can be rewritten as

$$\mathbb{E}[\tau(G_i^{x_1,\dots,x_{k-2}})] = (1+o(1))\,\psi(g_i, f_i) \tag{10}$$

We can now prove the conjecture stated in [9].

Proposition 1 (*Conjecture 1 [9]*). *Under the constraints $f_i \ge \gamma, \forall i$, $\psi(g, f)$ attains a maximum in a point (g, f) with vector f of the form $f = (\gamma, \dots, \gamma, 1 - (k-1)\gamma)$.*

Proof. Since $\psi(g, f)$ is invariant under (identical) permutations on g and f, we can study maxima for which g_k is the minimum among the values $g_1, g_2, \dots, g_k$ and show that for those points $f = (\gamma, \dots, \gamma, 1 - (k-1)\gamma)$. We prove this by considering the components of f one by one. Assume on the contrary that $f_1 > \gamma$. Given $\epsilon \le f_1 - \gamma$, set $\tilde{f} = (f_1 - \epsilon, f_2 \dots, f_{k-1}, f_k + \epsilon)$. Then

$$\begin{aligned}\psi(g, \tilde{f}) = &\sum_{\sigma:\sigma(k-1)\neq 1,k} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)} f_{\sigma(k-1)}\\ &+ \sum_{\sigma:\sigma(k-1)=1} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)} (f_1 - \epsilon)\\ &+ \sum_{\sigma:\sigma(k-1)=k} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)} (f_k + \epsilon)\\ = \psi(g, f) &- \epsilon \cdot \sum_{\sigma:\sigma(k-1)=1} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)}\\ &+ \epsilon \cdot \sum_{\sigma:\sigma(k-1)=k} g_{\sigma(1)} g_{\sigma(2)} \cdots g_{\sigma(k-2)}\,.\end{aligned}$$

Table 1
Numerical values for the bounds on R_k from [8] and from [9] in light of Proposition 1. All numbers are rounded upwards.

k	5	6	7	8
Bound from [8]	0.19200	0.092593	0.04284	0.019227
Bound from [9]	0.19079	0.092279	0.04279	0.019213

Since we assumed $g_1 \geq g_k$,

$$\sum_{\sigma:\sigma(k-1)=1} g_{\sigma(1)}g_{\sigma(2)}\cdots g_{\sigma(k-2)} \leq \sum_{\sigma:\sigma(k-1)=k} g_{\sigma(1)}g_{\sigma(2)}\cdots g_{\sigma(k-2)}. \tag{11}$$

and hence $\psi(g,\tilde{f}) \geq \psi(g,f)$. By repeating the above procedure for $f_2, f_3, \ldots, f_{k-1}$, we find that indeed $f = (\gamma, \ldots, \gamma, 1-(k-1)\gamma)$ maximizes $\psi(g,f)$ under the considered constraints whenever g_k is the minimum among $g_1, g_2, \ldots, g_k$, and in particular for the optimal g sorted in this way. □

It terms of g, it was already shown in [9] that assuming the above result one could show that the maximum value of $\psi(g,f)$, under the constraint that $f_i \geq \gamma$, $\forall i$, is attained at a point (g,f) with g of the form $(\beta, \beta, \ldots, 1-(k-1)\beta)$. Assuming this, the method developed in [9] for their explicit bounds can be extended to all k. Table 1 gives numerical results[2] for the first values of k.

4. Better bounds for small k

In this section we combine insights from both the approaches of [6] and [9]. Instead of looking at one subcode C', as done in [9], we follow the idea in [6]. We consider a partition $\{C_\omega : \omega \in \Omega\}$ of our k-hash code C and randomly select a subcode C_ω. Then we randomly extract codewords $x_1, \ldots, x_{k-2}$ from C_ω and bound the expected value in (8) over both random code and codewords. At this point, we replace the concavity argument of [6] with a symmetrization trick combined with new bounds on the maxima of certain polynomials. This procedure leads to the following nontrivial improvement on the rates R_5 and R_6.

Theorem 1. *For $k = 5, 6$ the following bounds hold*

- $R_5 \leq 0.1697$;
- $R_6 \leq 0.0875$.

4.1. Proof of Theorem 1

Here our goal is to find a family of subcodes such that any $k-2$ codewords $x_1, x_2, \ldots, x_{k-2}$ of a given subcode C_ω collide in all coordinates of $T = [1, \ell]$ for a carefully chosen value of ℓ, that is, for any coordinate $t \in T$ there exist i, j such that $x_{i,t} = x_{j,t}$. This will ensure that the coordinates from T contribute 0 to the LHS of (5). To do this, we cover all the possible prefixes of length ℓ; the following lemma can be seen as a special case of the known results on the fractional clique covering number (see [14]).

Lemma 2. *For any positive ϵ, for ℓ large enough, there exists a partition Ω of $\{1, 2, \ldots, k\}^\ell$ such that:*

1. $|\Omega| \leq \left\lfloor \left(\frac{k}{k-3} + \epsilon\right)^\ell \right\rfloor$.
2. *For all $\omega \in \Omega$ and $i = 1, \ldots, \ell$, the ith projection of ω has cardinality at most $k-3$.*

In particular, for any $\omega \in \Omega$, any $k-2$ sequences in ω collide in all coordinates $i = 1, \ldots, \ell$.

Proof. For any $i \in [1, k]$, consider the set $A_i = \{i, i+1, \ldots, i+(k-4)\}$, where the sums are performed modulo k in $[1, k]$. To a string $s = (i_1, \ldots, i_\ell)$ in $[1, k]^\ell$ we associate a set $\omega_s = A_{i_1} \times A_{i_2} \times \cdots \times A_{i_\ell} \subset [1, k]^\ell$. Fix a word $x \in [1, k]^\ell$, and choose uniformly at random the string s; the probability that $x \notin \omega_s$ is $1 - \left(\frac{k-3}{k}\right)^\ell$. Therefore, if we choose randomly h strings $s_1, \ldots, s_h$, the probability that $x \notin (\omega_{s_1} \cup \cdots \cup \omega_{s_h})$ is $\left(1 - \left(\frac{k-3}{k}\right)^\ell\right)^h$. Hence, the expected number of words $x \in [1, k]^\ell$ that do not belong to any of the $\omega_{s_1}, \ldots, \omega_{s_h}$ is

$$\mathbb{E}(|\{x \in [1, k]^\ell : \ x \notin \omega_{s_1} \cup \cdots \cup \omega_{s_h}\}|) = k^\ell \left(1 - \left(\frac{k-3}{k}\right)^\ell\right)^h.$$

[2] We believe that due to a minor error in the computation, the bound given for R_5 in [9] is not really the best possible using their method. We report here the optimal.

If this value is smaller than 1, then there exists a choice of $s_1, \dots, s_h$ such that the family $\{\omega_{s_1}, \dots, \omega_{s_h}\}$ covers the whole set $[1, k]^\ell$. This happens whenever

$$k^\ell \left(1 - \left(\frac{k-3}{k}\right)^\ell\right)^h < 1$$

or equivalently

$$h > \frac{-\ell \log k}{\log\left(1 - \left(\frac{k-3}{k}\right)^\ell\right)},$$

which holds for

$$h > \ell \left(\frac{k}{k-3}\right)^\ell \frac{\log k}{\log e}.$$

For ℓ large enough, setting $h = \left\lfloor \left(\frac{k}{k-3} + \epsilon\right)^\ell \right\rfloor$ we have the desired inequality.

Removing possible intersections between the sets ω_s we obtain a partition of $[1, k]^\ell$ with the desired properties, since condition 2) is satisfied by construction. □

Let $\Omega = \{\omega_1, \dots, \omega_h\}$ be a partition of $[1, k]^\ell$ as derived from Lemma 2 and consider the family of subcodes $C_{\omega_1}, \dots, C_{\omega_h}$ of C defined by

$$C_\omega = \{x \in C : (x_1, x_2, \dots, x_\ell) \in \omega\}.$$

Clearly, any $k-2$ codewords $x_1, x_2, \dots, x_{k-2}$ of a given subcode C_ω collide in all coordinates of $T = [1, \ell]$. As in [6], define a subcode C_ω to be *heavy* if $|C_\omega| > n$ and to be *light* otherwise. We can show that, if ℓ is not too large, most of the codewords are contained in heavy subcodes. Indeed, if we consider ℓ such that $\left(\frac{k}{k-3} + \epsilon\right)^\ell \le 2^{nR - 2\log n}$, that is $\ell \le \frac{nR - 2\log n}{\log\left(\frac{k}{k-3} + \epsilon\right)}$, we have that

$$\left| \bigcup_{C_\omega \text{ is light}} C_\omega \right| \le n \left(\frac{k}{k-3} + \epsilon\right)^\ell \le n 2^{nR - 2\log n} = \frac{|C|}{n}.$$

This means that at least a fraction $(1 - 1/n)$ of the codewords are in heavy subcodes. If we remove from C the light codes, the rate changes by an amount $\frac{1}{n}\log(1 - 1/n)$, which vanishes as n grows. So, in the following we can assume, without loss of generality, that all the subcodes are heavy.

We are finally ready to describe our strategy to pick the codewords $x_1, \dots, x_{k-2}$: first we choose a subcode C_ω with probability $\lambda_\omega = |C_\omega|/|C|$ and then we pick uniformly at random (and without replacement) $x_1, \dots, x_{k-2}$ from C_ω. Since those codewords collide in all the coordinates from the set $T = [1, \ell]$, we obtain in (5):

$$\log(|C| - k + 2) \le \mathbb{E}_{\omega \in \Omega}(\mathbb{E}[\sum_{i \in [\ell+1, n]} \tau(G_i^{x_1, x_2, \dots, x_{k-2}})]) \tag{12}$$

$$= \sum_{i \in [\ell+1, n]} \mathbb{E}_{\omega \in \Omega}(\mathbb{E}[\tau(G_i^{x_1, x_2, \dots, x_{k-2}})]). \tag{13}$$

Let again f_i be probability distribution of the ith coordinate of C, and let $f_{i|\omega}$ be the distribution of the subcode C_ω. Invoking (10) for the expectation over the random choice of $x_1, \dots, x_{k-2}$, we can write for $i \in [\ell + 1, n]$

$$\mathbb{E}_{\omega \in \Omega}(\mathbb{E}[\tau(G_i^{x_1, x_2, \dots, x_{k-2}})]) = (1 + o(1)) \sum_{\omega \in \Omega} \lambda_\omega \psi(f_{i|\omega}, f_i).$$

Since $f_i = \sum_{\mu \in \Omega} \lambda_\mu f_{i|\mu}$ and ψ is linear in its second variable, we have that

$$\mathbb{E}_{\omega \in \Omega}(\mathbb{E}[\tau(G_i^{x_1, x_2, \dots, x_{k-2}})]) = (1 + o(1)) \sum_{\omega, \mu \in \Omega} \lambda_\omega \lambda_\mu \psi(f_{i|\omega}, f_{i|\mu}).$$

We exploit now a simple yet effective trick. Since the sum above is symmetric in ω and μ, we can write

$$\mathbb{E}_{\omega \in \Omega}(\mathbb{E}[\tau(G_i^{x_1, x_2, \dots, x_{k-2}})])$$

$$= (1 + o(1)) \frac{1}{2} \sum_{\omega, \mu \in \Omega} \lambda_\omega \lambda_\mu [\psi(f_{i|\omega}, f_{i|\mu}) + \psi(f_{i|\mu}, f_{i|\omega})]. \tag{14}$$

Here, we note that $f_{i|\omega}$ has no relation with $f_{i|\mu}$. Therefore we can just consider the following polynomial function over two generic probability vectors $p = (p_1, p_2, \dots, p_k)$ and $q = (q_1, q_2, \dots, q_k)$

$$\begin{aligned}\Psi(p; q) &:= \psi(p, q) + \psi(q, p)\\ &= \sum_{\sigma\in S_k} p_{\sigma(1)}p_{\sigma(2)}\cdots p_{\sigma(k-2)}q_{\sigma(k-1)} + q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)}.\end{aligned} \tag{15}$$

Because of (14), if M_k is the maximum of Ψ over probabilistic vectors p and q, Eq. (13) says that

$$\begin{aligned}\log|C| &\le (1 + o(1))\frac{1}{2}(n-\ell)\sum_{\omega,\mu\in\Omega}\lambda_\omega\lambda_\mu M_k\\ &= (1 + o(1))\frac{1}{2}(n-\ell)M_k.\end{aligned}$$

Recalling that $|C| = 2^{nR}$ and taking $\ell = \left\lfloor \frac{nR - 2\log n}{\log\left(\frac{k}{k-3}+\epsilon\right)} \right\rfloor$, we obtain

$$R \le (1 + o(1))\left[1 - \frac{R - 2\log(n)/n}{\log\left(\frac{k}{k-3}+\epsilon\right)}\right]\frac{M_k}{2}.$$

Rearranging the terms, taking $n \to \infty$ first and then $\epsilon \to 0$, we deduce the following proposition.

Proposition 2. *Let M_k be the maximum of Ψ over probabilistic vectors $p = (p_1, p_2, \dots, p_k)$ and $q = (q_1, q_2, \dots, q_k)$. Then we have the following upperbound on R_k*

$$R_k \le \left(\frac{2}{M_k} + \frac{1}{\log(k/(k-3))}\right)^{-1}.$$

In the next subsection we will prove that $M_5 = \frac{15(48+\sqrt{5})}{1936} \approx 0.389226$ and $M_6 = 24/125$. This implies Theorem 1.

4.2. Bounds on Ψ

The goal of this subsection is to find the maximum of the function Ψ as defined in (15). For this purpose we first introduce two lemmas that provide some restrictions on this maximum.

Lemma 3. *Let $\bar p = (\bar p_1, \dots, \bar p_k)$ and $\bar q = (\bar q_1, \dots, \bar q_k)$ be two probabilistic vectors. If $(\bar p; \bar q)$ is a maximum for Ψ such that $\bar p_1, \bar p_2, \bar q_1, \bar q_2$ are nonzero, then also $(\frac{\bar p_1+\bar p_2}{2}, \frac{\bar p_1+\bar p_2}{2}, \bar p_3, \dots, \bar p_k; \frac{\bar q_1+\bar q_2}{2}, \frac{\bar q_1+\bar q_2}{2}, \bar q_3, \dots, \bar q_k)$ is a maximum for Ψ.*

Proof. If $\bar P = (\bar p; \bar q)$ is a maximum for $\Psi(p; q)$ under the constraints $p_1 + p_2 + \cdots + p_k = 1$ and $q_1 + q_2 + \cdots + q_k = 1$, then it is a maximum also under the stronger constraints $p_1 + p_2 = c_1$, $q_1 + q_2 = c_2$ where $c_1 = \bar p_1 + \bar p_2$, $c_2 = \bar q_1 + \bar q_2$, and $p_i = \bar p_i$, $q_i = \bar q_i$ for $i \in \{3, 4, \dots, k\}$. Because of the Lagrange multiplier method this means that:

$$\frac{\partial\Psi}{\partial p_1}\Big|_{\bar P} = \frac{\partial\Psi}{\partial p_2}\Big|_{\bar P}$$

and

$$\frac{\partial\Psi}{\partial q_1}\Big|_{\bar P} = \frac{\partial\Psi}{\partial q_2}\Big|_{\bar P}.$$

After simple algebraic manipulations, it follows that

$$(\bar p_1 - \bar p_2)a + (\bar q_1 - \bar q_2)b = 0$$

and

$$(\bar q_1 - \bar q_2)d + (\bar p_1 - \bar p_2)c = 0$$

where $a = \frac{\partial^2\Psi}{\partial p_1\partial p_2}\big|_{\bar P}$, $b = \frac{\partial^2\Psi}{\partial p_1\partial q_2}\big|_{\bar P} = \frac{\partial^2\Psi}{\partial q_1\partial p_2}\big|_{\bar P} = c$ and $d = \frac{\partial^2\Psi}{\partial q_1\partial q_2}\big|_{\bar P}$. If we set $\bar p_1 - \bar p_2 = x$, $\bar q_1 - \bar q_2 = y$, the previous equations became:

$$\begin{cases} ax + by = 0;\\ cx + dy = 0.\end{cases}$$

In the case $ad - bc \neq 0$ the previous system admits only the solution $x = y = 0$ that means $\bar p_1 = \bar p_2$ and $\bar q_1 = \bar q_2$. It is clear that here we have $\bar p_1 = \frac{\bar p_1+\bar p_2}{2} = \bar p_2$, $\bar q_1 = \frac{\bar q_1+\bar q_2}{2} = \bar q_2$ and hence the thesis is satisfied.

Let us assume $ad - bc = 0$. Then there exists a line L of points $P(t)$ such that $P(1) = \bar{P}$, $P(0) = (\frac{\bar{p}_1+\bar{p}_2}{2}, \frac{\bar{p}_1+\bar{p}_2}{2}, \bar{p}_3, \ldots, \bar{p}_k; \frac{\bar{q}_1+\bar{q}_2}{2}, \frac{\bar{q}_1+\bar{q}_2}{2}, \bar{q}_3, \ldots, \bar{q}_k)$ and

$$\frac{\partial \Psi}{\partial p_1}\Big|_{P(t)} - \frac{\partial \Psi}{\partial p_2}\Big|_{P(t)} = \frac{\partial \Psi}{\partial q_1}\Big|_{P(t)} - \frac{\partial \Psi}{\partial q_2}\Big|_{P(t)} = 0.$$

It follows that $\Psi(P(t))$ is constantly equal to the value of Ψ in $\bar{P} = P(1)$. Since $(\frac{\bar{p}_1+\bar{p}_2}{2}, \frac{\bar{p}_1+\bar{p}_2}{2}, \bar{p}_3, \ldots, \bar{p}_k, \frac{\bar{q}_1+\bar{q}_2}{2}, \frac{\bar{q}_1+\bar{q}_2}{2}, \bar{q}_3, \ldots, \bar{q}_k)$ belongs to the line L, this point is also a maximum for Ψ. □

With essentially the same proof we also obtain the following result.

Lemma 4. *Let $\bar{p} = (\bar{p}_1, \ldots, \bar{p}_k)$ and $\bar{q} = (\bar{q}_1, \ldots, \bar{q}_k)$ be two probabilistic vectors. If $(\bar{p}; \bar{q})$ is a maximum for Ψ such that $\bar{p}_1, \bar{p}_2$ are nonzero while $\bar{q}_1 = \bar{q}_2 = 0$ then also $(\frac{\bar{p}_1+\bar{p}_2}{2}, \frac{\bar{p}_1+\bar{p}_2}{2}, \bar{p}_3, \ldots, \bar{p}_k; 0, 0, \bar{q}_3, \ldots, \bar{q}_k)$ is a maximum for Ψ.*

In the next two lemmas, we will provide some further restrictions on the maximum of Ψ using just some combinatorial arguments.

Lemma 5. *We have that:*

$$\Psi(0, p_2, \ldots, p_k; 0, q_2, \ldots, q_k) \leq \Psi(0, p_2, \ldots, p_k; q_2, 0, q_3, \ldots, q_k).$$

Proof. Because of the definition, we have that $\Psi(0, p_2, \ldots, p_k; 0, q_2, \ldots, q_k)$ evaluates as

$$\sum_{\sigma:\ \sigma(k)=1} p_{\sigma(1)}p_{\sigma(2)}\cdots p_{\sigma(k-2)}q_{\sigma(k-1)} + q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)}.$$

The crucial observation is that when we expand the expression of the quantity $\Psi(0, p_2, \ldots, p_k; q_2, 0, q_3, \ldots, q_k)$, all terms in the previous expression will still appear, with the addition of other terms including the factor p_2q_2. In particular, one can write

$$\Psi(0, p_2, \ldots, p_k; q_2, 0, q_3, \ldots, q_k) =$$

$$\sum_{\sigma:\ \sigma(k)=1} p_{\sigma(1)}p_{\sigma(2)}\cdots p_{\sigma(k-2)}q_{\sigma(k-1)} + q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)} +$$

$$(k-2)p_2q_2\left(\sum_{\sigma\in Sym(3,\ldots,k)} p_{\sigma(3)}\cdots p_{\sigma(k-1)} + q_{\sigma(3)}\cdots q_{\sigma(k-1)}\right) =$$

$$\Psi(0, p_2, \ldots, p_k; 0, q_2, \ldots, q_k) +$$

$$(k-2)p_2q_2\left(\sum_{\sigma\in Sym(3,\ldots,k)} p_{\sigma(3)}\cdots p_{\sigma(k-1)} + q_{\sigma(3)}\cdots q_{\sigma(k-1)}\right).$$

The claim follows since each term of the last sum is non negative. □

The following Lemma is in the same spirit of Proposition 1.

Lemma 6. *We have that:*

$$\Psi(p_1, \ldots, p_{k-3}, 0, 0, 0; q_1, q_2, \ldots, q_k) \leq \Psi\left(1, 0\ldots, 0; 0, \frac{1}{(k-1)}, \ldots, \frac{1}{(k-1)}\right).$$

Proof. We suppose, without loss of generality that q_1 is the minimum among the values $q_1, q_2, \ldots, q_{k-3}$. Setting $p = (p_1, \ldots, p_{k-3}, 0, 0, 0)$ and $q = (q_1, \ldots, q_k)$, we have

$$\begin{aligned}\Psi(p; q) = &\sum_{\sigma:\ \sigma(k-1)\notin\{1,2\}} q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)}\\ &+ \frac{p_1+p_2}{2}\sum_{\sigma:\ \{\sigma(k-1),\sigma(k)\}=\{1,2\}} q_{\sigma(1)}\cdots q_{\sigma(k-2)}\\ &+ (p_1q_2+q_1p_2)(k-2)\sum_{\sigma\in Sym(3,\ldots,k)} q_{\sigma(3)}\cdots q_{\sigma(k-1)}.\end{aligned}$$

Similarly, setting $p' = (p_1 + p_2, 0, p_3, \ldots, p_{k-3}, 0, 0, 0)$, we have that:

$$\begin{aligned}\Psi(p'; q) = &\sum_{\sigma:\ \sigma(k-1)\notin\{1,2\}} q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)}\\ &+ \frac{p_1+p_2}{2}\sum_{\sigma:\ \{\sigma(k-1),\sigma(k)\}=\{1,2\}} q_{\sigma(1)}\cdots q_{\sigma(k-2)}\\ &+ (p_1+p_2)q_2(k-2)\sum_{\sigma\in Sym(3,\ldots,k)} q_{\sigma(3)}\cdots q_{\sigma(k-1)}.\end{aligned}$$

Since $q_1 \le q_2$ we have that

$$\Psi(p; q) \le \Psi(p'; q).$$

Reiterating the previous procedure, since q_1 is the minimum among the values $q_1, \ldots, q_{k-3}$, we obtain

$$\Psi(p_1, \ldots, p_{k-3}, 0, 0, 0; q_1, q_2, \ldots, q_k) \le \Psi(1, 0, \ldots, 0, 0; q_1, q_2, \ldots, q_k). \tag{16}$$

Since q_1 does not appear in the value of $\Psi(1, 0, \ldots, 0, 0; q_1, q_2, \ldots, q_k)$, this is certainly maximized for $q_1 = 0$. Finally, due to the Muirhead's inequality, we obtain that the RHS of (16) is maximized for $q_2 = q_3 = \cdots = q_k = \frac{1}{k-1}$. □

As a consequence of the previous lemmas, Ψ attains a maximum in a point of one of the following types:

(a) $\left(1, 0 \ldots, 0; 0, \frac{1}{(k-1)}, \ldots, \frac{1}{(k-1)}\right)$;
(b) $(1/k, \ldots, 1/k; 1/k, \ldots, 1/k)$;
(c) $(0, 0, \alpha, \ldots, \alpha, \beta, \beta; \gamma, \gamma, \delta, \ldots, \delta, 0, 0)$
where $(k-4)\alpha + 2\beta = 1$ and $2\gamma + (k-4)\delta = 1$;
(d) $(0, 0, \alpha, \ldots, \alpha, \beta; \gamma, \gamma, \delta, \ldots, \delta, 0)$
where $(k-3)\alpha + \beta = 1$ and $2\gamma + (k-3)\delta = 1$;
(e) $(0, 0, 1/(k-2), \ldots, 1/(k-2); \gamma, \gamma, \delta, \ldots, \delta)$
where $2\gamma + (k-2)\delta = 1$;
(f) $(0, \alpha, \ldots, \alpha, \beta; \gamma, \delta, \ldots, \delta, 0)$
where $(k-2)\alpha + \beta = 1$ and $\gamma + (k-2)\delta = 1$;
(g) $(0, 1/(k-1), \ldots, 1/(k-1); \gamma, \delta, \ldots, \delta)$
where $\gamma + (k-1)\delta = 1$.

In particular, because of Lemma 6, a maximum with three or more p-coordinates (resp. q-coordinates) equal to zero is also attained in a point of the form (a). Otherwise, there are at most two zero coordinates both for the vector p and for the vector q. Due to Lemma 5, we can then assume those zeros are in different positions and finally, using Lemmas 3 and 4, we obtain the required characterization of the maximum.

For $k = 5, 6$, we have inspected using Mathematica all cases listed above and determined the maximum explicitly.

Theorem 2. *The following hold:*

- *for $k = 5$, the global maximum of Ψ is $\frac{15(48+\sqrt{5})}{1936} \approx 0.389226$ and is obtained in case (g) with $\delta = 1/44(4+\sqrt{5})$ and $\gamma = 1 - 4\delta$;*
- *for $k = 6$, the global maximum of Ψ is $24/125 = 0.192$, obtained in case (a).*

Theorem 1 follows immediately from Theorem 2 and Proposition 2.

Remark 1. For $k > 6$, the value obtained for p and q as in case (a), which we conjecture to be the true maximum, is too big to improve the known upper bounds on R_k.

Acknowledgments

This research was partially supported by Italian Ministry of Education under Grant PRIN 2015 D72F16000790001. Helpful discussions with Jaikumar Radhakrishnan and Venkatesan Guruswami are gratefully acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dam.2020.11.001.

References

[1] E. Arikan, A bound on the zero-error list coding capacity, in: Proceedings. IEEE International Symposium on Information Theory, 1993, pp. 152–152.
[2] E. Arikan, An upper bound on the zero-error list-coding capacity, IEEE Trans. Inform. Theory 40 (1994) 1237–1240.
[3] S. Bhandari, J. Radhakrishnan, Bounds on the zero-error list-decoding capacity of the q/(q-1) channel, IEEE International Symposium on Information Theory, ISIT, 2018, pp. 906–910.
[4] S. Costa, M. Dalai, A gap in the slice rank of *k*-tensors, J. Combin. Theory Ser. A 177 (2021) 105335.
[5] M. Dalai, V. Guruswami, J. Radhakrishnan, An improved bound on the zero-error listdecoding capacity of the 4/3 channel, in: IEEE International Symposium on Information Theory, ISIT, 2017, pp. 1658–1662.
[6] M. Dalai, V. Guruswami, J. Radhakrishnan, An improved bound on the zero-error listdecoding capacity of the 4/3 channel, IEEE Trans. Inform. Theory 66 (2) (2020) 749–756.
[7] P. Elias, Zero error capacity under list decoding, IEEE Trans. Inform. Theory 34 (1988) 1070–1074.
[8] Michael L. Fredman, János Komlós, On the size of separating systems and families of perfect hash functions, SIAM J. Algebr. Discrete Methods 5 (1984) 61–68.
[9] V. Guruswami, A. Riazanov, Beating Fredman-Komlos for perfect *k*-hashing, in: Leibniz International Proceedings in Informatics, 2019.
[10] G. Hansel, Nombre minimal de contacts de fermature nécessaires pour réaliser une fonction booléenne symétrique de *n* variables, C. R. Acad. Sci., Paris (1964) 6037–6040.
[11] J. Korner, Coding of an information source having ambiguous alphabet and the entropy of graphs, in: 6th Prague Conference on Information Theory, 1973, pp. 411–425.
[12] J. Korner, Fredman–Komlós bounds and information theory, SIAM J. Algebr. Discrete Methods 7 (4) (1986) 560–570.
[13] J. Korner, K. Marton, New bounds for perfect hashing via information theory, European J. Combin. 9 (1988) 523–530.
[14] R.J. McEliece, E.C. Posner, Hide and seek, data storage, and entropy, Ann. Math. Stat. 42 (5) (1971) 1706–1716.
[15] A. Nilli, Perfect hashing and probability, Combin. Probab. Comput. 3 (1994) 407–409.
[16] J. Radhakrishnan, Entropy and Counting, available at: http://www.tcs.tifr.res.in/jaikumar/Papers/EntropyAndCounting.pdf.
[17] C. Xing, C. Yuan, Beating the probabilistic lower bound on perfect hashing, 2019, arXiv preprint arXiv:1908.08792.