Improved Bounds for (b, k)-hashing

Stefano Della Fiore[*], Simone Costa[†], Marco Dalai[*]

Abstract

For fixed integers n and $b \geq k$, let $A(b, k, n)$ the largest size of a subset of $\{1, 2, \ldots, b\}^n$ such that, for any k distinct elements in the set, there is a coordinate where they all differ. Bounding $A(b, k, n)$ is a problem of relevant interest in information theory and in computer science, relating to the zero-error capacity with list decoding and with the study of (b, k)-hash families of functions. It is known that, for fixed b and k, $A(b, k, n)$ grows exponentially in n. In this paper, we determine new exponential upper bounds for different values of b and k.

A first bound on $A(b, k, n)$ for general b and k was derived by Fredman and Komlós in the '80s and improved for certain $b \neq k$ by Körner and Marton and by Arikan. Only very recently better bounds were derived for general b and k by Guruswami and Riazanov, while stronger results for small values of $b = k$ were obtained by Arikan, by Dalai, Guruswami and Radhakrishnan, and by Costa and Dalai. In this paper, we strengthen the bounds for some specific values of b and k. Our contribution is a new computational method for obtaining upper bounds on the values of a quadratic form defined over discrete probability distributions in arbitrary dimensions, which emerged as a central ingredient in recent works. The proposed method reduces an infinite-dimensional problem to a finite one, which we manage to further simplify by means of a series of optimality conditions.

Keywords: perfect hashing, list decoding, zero-error capacity, extremal combinatorics
MSC: 68R05

1 Introduction

The problem considered in this paper has a twofold history that connects it naturally with combinatorial aspects of computer science and information theory. Let b, k, and n be integers and let C be a subset of $\{1, 2, \ldots, b\}^n$ with the property that for any k distinct elements we can find a coordinate where they all differ. Such a set can be interpreted, by looking at it coordinate-wise, as a family of n hashing functions on some universe of size $|C|$. The required property then says that the family is a perfect hash family, that is, any k elements in the universe are k-partitioned by at least one function. Alternatively, C can be interpreted as a code of rate $\log(|C|)/n$ for communication over a channel with b inputs. Assume that the channel is a $b/(k-1)$ channel, meaning that any $k-1$ of the b inputs share one output but no k distinct inputs do (see Figure 1). The required property for C is what is needed for the receiver to always be able to produce a list of $k-1$ codewords of C which must necessarily include the one that was sent; that is, zero-error communication with $(k-1)$-list decoding is possible. Indeed, the condition implies that any k codewords use, in at least one coordinate, k different symbols, and one of them will not be compatible with the received symbol in that coordinate. We refer the reader to [8], [9], [13], [14] for an overview of the more general context of this problem. Some recent important results in a different asymptotic setting can be found in [4].

[*]DII, University of Brescia, Via Branze 38, 25123 Brescia, Italy. Emails: s.dellafiore001@unibs.it, marco.dalai@unibs.it.

[†]DICATAM, University of Brescia, Via Branze 43, 25123 Brescia, Italy. Email: simone.costa@unibs.it.
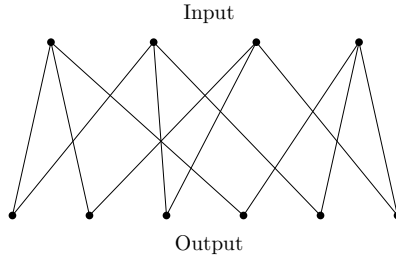
Input

Output

Figure 1: A 4/2 channel. Edges represent positive probabilities. Here, zero-error communication is possible when decoding with list-size equal to 2.

We will call any subset C of $\{1, 2, \ldots, b\}^n$ with the described property a (b, k)-hash code. For the reasons mentioned above, bounding the size of (b, k)-hash codes is a combinatorial problem that has been of interest in both computer science and information theory. Let $A(b, k, n)$ be the largest size of such a code. It is known that for fixed b and k, $A(b, k, n)$ grows exponentially in n, and a challenging problem consists in bounding the exponent. We will thus study the quantity

$$R_{(b,k)} = \limsup_{n \to \infty} \frac{1}{n} \log A(b, k, n). \tag{1}$$

Note that, throughout the paper, all logarithms are to base 2.

Few lower bounds on $R_{(b,k)}$ are known. First results in this sense were given by [9], [8] and a better bound was derived by [12] for $b = k = 3$. More recently, new lower bounds were derived in [16] for infinitely many other values of k. The first, landmark result concerning the upper bounds was obtained by Fredman-Komlós [9], who showed that

$$R_{(b,k)} \leq \frac{b^{\underline{k-1}}}{b^{k-1}} \log(b - k + 2), \tag{2}$$

where $b^{\underline{k-1}} = b(b-1)\cdots(b-k+2)$. Progress has since been rare. A generalization of the bound given in equation (2) was derived by Körner and Marton [12] in the form

$$R_{(b,k)} \leq \min_{2 \leq j \leq k-2} \frac{b^{\underline{j+1}}}{b^{j+1}} \log \frac{b-j}{k-j-1}. \tag{3}$$

Nilli [14] provided an elementary proof of (3) without considerations of graph entropy or hypergraph entropy. This bound was further improved for different values of b and k by Arikan [3]. In the case $b = k$, an improvement was first obtained for $k = 4$ in [2] and then in [6], [7]. The latter only focuses on $b = k = 4$, but the procedure can be extended to general b and k. As shown in the next sections, it leads to the following bound.

Lemma 1. *For general b and k, we have*

$$R_{(b,k)} \leq \left(\frac{1}{\log b} + \frac{b^2}{(b^2 - 3b + 2) \log \frac{b-2}{k-3}} \right)^{-1}. \tag{4}$$

In [10], the authors prove that the Fredman-Komlós bound is not tight for any $b \geq k > 3$; explicit better values were given there for $b = k = 5, 6$, and for larger $b = k$ modulo a conjecture which is proved in [5], where further improvements are also obtained for $b = k = 5, 6$. The case of $b \neq k$ is not described in detail in [10] but, as the authors mention, it is straightforward. We do not write here the bound since it has a complicated expression.

2

In this paper, we attack some of the cases which appear not to be optimally handled by those methods. In particular, we build on the results obtained in [5] and add an improvement that leads to better bounds for many pairs of (b, k) values. The results of [5] for $b = k$ were derived following an approach common to many recent works by introducing a symmetrization which reduces to the problem of bounding a quadratic form on probability distributions. We give a more general exposition for the general b, k case, anticipating here the key lemma whose proof we give for completeness in the next section. Fix an integer j in the range $2, \ldots, k - 2$ and define, for probability vectors $p, q \in \mathbb{R}^b$, the function

$$\Psi_j(p; q) = \frac{1}{(b - j - 1)!} \sum_{\sigma} p_{\sigma(1)} p_{\sigma(2)} \cdots p_{\sigma(j)} q_{\sigma(j+1)} + q_{\sigma(1)} q_{\sigma(2)} \cdots q_{\sigma(j)} p_{\sigma(j+1)}, \quad (5)$$

where σ ranges over all permutations of $\{1, 2, \ldots, b\}$. Define then

$$\mathbf{M}_j = \sup_{\lambda} \sum_{p, q} \lambda_p \lambda_q \Psi_j(p; q) \quad (6)$$

where λ ranges over all probability distributions on finite sets of probability vectors in $\mathbb{R}^b$, so that λ_p is the probability associated to the probability vector p. Then, the following bound holds.

Lemma 2. *For* $j = 2, \ldots, k - 2$,

$$R_{(b,k)} \leq \left(\frac{2}{\mathbf{M}_j \log \frac{b-j}{k-j-1}} + \frac{1}{\log \left(\frac{b}{j-1} \right)} \right)^{-1}. \quad (7)$$

The results in [5] were obtained using in (7), for $b = k$ and $j = k - 2$, the upper bound

$$\mathbf{M}_j \leq \max_{p, q} \Psi_j(p; q). \quad (8)$$

A weakness in this bound comes from the fact that distributions p and q that maximize $\Psi_j(p; q)$ exhibit in many cases some opposing asymmetries, in the sense that they give higher probabilities to different symbols. When used as a replacement for *each* of the pairs of p and q in (6), we have a rather conservative bound, because pairs (p, q) which give high values for $\Psi_j(p; q)$ will give low values for $\Psi_j(p; p)$ and $\Psi_j(q; q)$, and equation (6) contains a weighted contribution from all pairings of p and q. In this paper, we present a computational method for obtaining more refined bounds on $\mathbf{M}_j$ for general b, k values which lead to improvements on the best-known bounds on $R_{(b,k)}$ for many b, k pairs.

In Table 1 we give a comparison between bounds (4) and (3), the bounds of [3] and [10] and our new bounds for different values of b and k. In Table 2 we show that for some (b, k)-cases the bound (4) is the best bound among all the current known bounds, in particular when b is much larger than k. Finally, in Table 3 we provide some (b, k)-cases where the bound of [10] is the current best known bound, in particular when b and k are large and nearly equal. Clearly, the cases reported in Tables 2 and 3 are not exhaustive, but they have been properly selected to point out that our method does not always provide the best bounds. The integers in the parentheses for bounds [10], [3] and [12] in Table 2 represent the optimal value of a parameter which has the same role as j in (3). When its value is not reported, as well as in Tables 1 and 3, it is equal to $k - 2$ for our bounds and for bounds of [10], [3] and [12]. Instead, for bound (4) it is always equal to 2.

The paper is structured as follows. In Section 2 we give some background proving Lemmas 1 and 2. In Section 3 we present the first part of our computational method to bound $\mathbf{M}_j$ by partitioning the domain of possible p and q distributions and then working on the subdomains. The second part is presented in Section 4, where we derive optimality

Table 1: Upper bounds on $R_{(b,k)}$. All numbers are rounded upwards.

(b,k)	Our method	[3]	[10]	[12]	(b,k)	Our method	[3]	[10]	[12]
$(5,5)$	**0.16894**[1]	0.23560	0.19079	0.19200	$(6,5)$	**0.34512**[1]	0.44149	0.43207	0.44027
$(6,6)$	**0.08475**[1]	0.15484	0.09228	0.09260	$(7,6)$	**0.19897**[2]	0.30554	0.23524	0.23765
$(8,6)$	**0.31799**[2]	0.44888	0.40330	0.41016	$(9,6)$	**0.43237**[2]	0.58303	0.58486	0.59455
$(10,6)$	**0.53909**[2]	0.73304	0.76977	0.78170	$(11,6)$	**0.63766**[2]	0.87038	0.95285	0.96640
$(12,6)$	**0.72848**[2]	0.99588	1.13118	1.14584	$(13,6)$	**0.81227**[2]	1.11084	1.30322	1.31855
$(14,6)$	**0.88978**[2]	1.21657	1.46822	1.48388	$(7,7)$	**0.04090**[1]	0.09747	0.04279	0.04284
$(8,7)$	**0.10865**[2]	0.20340	0.12134	0.12189	$(9,7)$	**0.19054**[2]	0.31204	0.22547	0.22761
$(10,7)$	**0.27741**[2]	0.41982	0.34615	0.35108	$(11,7)$	**0.36424**[2]	0.52472	0.47856	0.48538
$(12,7)$	**0.44850**[2]	0.65160	0.61698	0.62549	$(13,7)$	**0.52902**[2]	0.77148	0.75796	0.76792
$(14,7)$	**0.60538**[2]	0.88384	0.89915	0.91027	$(8,8)$	**0.01889**[1]	0.05769	0.01922	0.01923
$(9,8)$	**0.05616**[1]	0.12874	0.06001	0.06013	$(10,8)$	**0.10791**[2]	0.20754	0.12048	0.12096
$(11,8)$	**0.16878**[2]	0.29023	0.19680	0.19818	$(12,8)$	**0.23451**[2]	0.37434	0.28470	0.28797
$(13,8)$	**0.30214**[2]	0.45827	0.38245	0.38694	$(14,8)$	**0.36974**[2]	0.56612	0.48658	0.49227
$(10,9)$	**0.02773**[1]	0.07668	0.02874	0.02876	$(11,9)$	**0.05796**[2]	0.13098	0.06197	0.06208
$(12,9)$	**0.09730**[2]	0.19157	0.10746	0.10778	$(13,9)$	**0.14332**[2]	0.25611	0.16368	0.16444
$(14,9)$	**0.19382**[2]	0.32294	0.22865	0.23033	$(11,10)$	**0.01321**[1]	0.04289	0.01342	0.01343
$(12,10)$	**0.02978**[1]	0.07806	0.03093	0.03095	$(13,10)$	**0.05342**[2]	0.12009	0.05674	0.05681
$(14,10)$	**0.08332**[2]	0.16726	0.09071	0.09090	$(13,11)$	**0.01476**[1]	0.04400	0.01506	0.01506
$(14,11)$	**0.02815**[2]	0.07141	0.02915	0.02916	$(14,12)$	**0.00712**[1]	0.02361	0.00718	0.00718
$(15,13)$	**0.00335**[1]	0.01218	0.00336	0.00336					

[1] Bounds obtained with the procedure of Section 3, strictly improving also the generalization of [5] to the (b,k)-case.
[2] Bounds where the procedure of Section 3 reduces to the same solution obtained by generalization of [5].

Table 2: Upper bounds on $R_{(b,k)}$. All numbers are rounded upwards.

(b,k)	[6]*	[5]*	[3]	[10]	[12]
$(5,4)$	**0.57303**	0.66126	0.61142	0.74834	0.73697(0)
$(6,4)$	**0.77709**	0.87963	0.83904	1.09604	1.00000(0)
$(7,4)$	**0.94372**	1.03711	1.02931	1.40593	1.22239(0)
$(100,6)$	**2.81342**	—	3.61848(2)	4.87959(2)	4.32193(0)
$(100,7)$	**2.67473**	—	3.41158(2)	4.47696(2)	4.05889(0)

Missing values indicate impossibility to compute the bound due to high computational complexity.
*The generalized bound for the (b,k) case.

Table 3: Upper bounds on $R_{(b,k)}$. All numbers are rounded upwards.

(b,k)	[10]	[5]*	[3]	[12]
$(9,9)$	**$8.4288 \cdot 10^{-3}$**	0.00946	0.03182	$8.4300 \cdot 10^{-3}$
$(10,10)$	**$3.6287 \cdot 10^{-3}$**	0.00419	0.01642	$3.6288 \cdot 10^{-3}$
$(11,11)$	**$1.53895 \cdot 10^{-3}$**	0.00181	0.00803	$1.53897 \cdot 10^{-3}$
$(12,11)$	**$6.13036 \cdot 10^{-3}$**	0.00664	0.02266	$6.13075 \cdot 10^{-3}$
$(12,12)$	**$6.44678 \cdot 10^{-4}$**	0.00077	0.00377	$6.44679 \cdot 10^{-4}$
$(13,12)$	**$2.75350 \cdot 10^{-3}$**	0.00305	0.01143	$2.75355 \cdot 10^{-3}$
$(13,13)$	**$2.672760 \cdot 10^{-4}$**	0.00033	0.00172	$2.672761 \cdot 10^{-4}$
$(14,13)$	**$1.218595 \cdot 10^{-3}$**	0.00138	0.00556	$1.218599 \cdot 10^{-3}$

*The generalized bound for the (b,k) case.

conditions on p and q over such subdomains, which allow us to reduce the problem to a manageable one that can be solved exactly. Finally, in Section 5 we show that at least some of the bounds that we obtain are not tight, although a quantitative improvement is not explicitly derived.

2 Background

The best upper bounds on $R_{(b,k)}$ available in the literature can all be seen as different applications of a central idea, which is the study of (b,k)-hashing by comparison with a combination of binary partitions. This mainline of approach to the problem comes from the original work of Fredman and Komlós [9]. A clear and productive formulation of the idea was given by Radhakrishnan in terms of Hansel's lemma [15], which remained the main tool used in all recent results [7], [10] and [5].

A *hypergraph* $\mathcal{H}$ is a family E of subsets of a finite set V where the subsets in E are called edges and the elements of V are called vertices. If all the edges have size d then we say that $\mathcal{H}$ is a d-uniform hypergraph. We state the Hansel's Lemma here for the reader's convenience.

Lemma 3 (Hansel for Hypergraphs [11], [14])**.** *Let K_r^d be the complete d-uniform hypergraph on r vertices and let $G_1, \ldots, G_m$ be c-partite d-uniform hypergraphs on those vertices such that $\cup_i G_i = K_r^d$. Let $\tau(G_i)$ be the number of non-isolated vertices in G_i. Then*

$$\log \frac{c}{d-1} \sum_{i=1}^{m} \tau(G_i) \geq \log \frac{r}{d-1} \,. \tag{9}$$

Using this main ingredient, we provide here a proof of Lemma 2, which extends the bound used in [5] to general b and k. We refer the reader to [5] for a more detailed discussion on connections with other previous bounds in the literature.

Proof of Lemma 2. Given a (b,k)-hash code C of rate R, fix any j elements $x_1, x_2, \ldots, x_j$ in C, with j in the range $2, \ldots, k-2$. For any coordinate i let $G_i^{x_1, \ldots, x_j}$ be the $(b-j)$-partite $(k-j)$-uniform hypergraph with vertex set $C \setminus \{x_1, x_2, \ldots, x_j\}$ and edge set

$$E = \big\{ \{y_1, \ldots, y_{k-j}\} : x_{1,i}, \ldots, x_{j,i}, y_{1,i}, \ldots, y_{k-j,i} \text{ are all distinct} \big\} \,. \tag{10}$$

Since C is a (b,k)-hash code, then $\bigcup_i G_i^{x_1, \ldots, x_j}$ is the complete $(k-j)$-uniform hypergraph on $C \setminus \{x_1, x_2, \ldots, x_j\}$ and so

$$\log \frac{b-j}{k-j-1} \sum_{i=1}^{n} \tau(G_i^{x_1, \ldots, x_j}) \geq \log \frac{|C|-j}{k-j-1} \,. \tag{11}$$

Inequality (11) holds for any choice of $x_1, x_2, \ldots, x_j$, so the main goal is proving that the left hand side is not too large for all possible choices of $x_1, x_2, \ldots, x_j$. The choice can be deterministic or we can take the expectation over any random selection.

First note that if the $x_{1,i}, x_{2,i}, \ldots, x_{j,i}$ are not all distinct (let us say that they "collide") then the hypergraph defined by (10) is empty, that is the corresponding τ in the left hand side of (11) is zero. Otherwise, $\tau(G_i^{x_1, \ldots, x_j})$ depends on the frequency of different symbols in the i-th coordinate of the code. Let f_i be their distribution, meaning that $f_{i,a}$ is the fraction of elements of C whose i-th coordinate is a. Then, we have

$$\tau(G_i^{x_1, \ldots, x_j}) = \begin{cases} 0 & x_1, \ldots, x_j \text{ collide in coordinate } i \\ \left(\frac{|C|}{|C|-j} \right) \left(1 - \sum_{h=1}^{j} f_{i,x_{h,i}} \right) & \text{otherwise} \end{cases} \,. \tag{12}$$

We partition the code C into subcodes C_ω, $\omega \in \Omega$ in such a way that each subcode has a size which grows unbounded with n and uses in any of its first ℓ coordinates only $j-1$ symbols, where ℓ denotes the length of the prefix. It can be shown, by an easy extension of the method used for the case $b = k$ and $j = k-2$ in [5], that if the original code has rate R, then for any $\epsilon > 0$ one can do this with a choice of $\ell = n(R-\epsilon)/\log\left(\frac{b}{j-1}\right)$ for n large enough. Given such a partition of our code, if we select codewords $x_1, \ldots, x_j$ within the same

subcode C_ω, they will collide in the first ℓ coordinates and the corresponding contribution to the left-hand side of (11) will be zero. The next step is to add randomization. Pick randomly one of the subcodes C_ω and randomly select the codewords $x_1, \ldots, x_j$ within C_ω. Then an upper bound on $|C|$ is obtained by taking an expectation on the left-hand side of (11)

$$\log \frac{|C| - j}{k - j - 1} \leq \log \frac{b - j}{k - j - 1} \mathbb{E}_\omega \left(\mathbb{E} \left[\sum_{i \in [\ell+1,n]} \tau(G_i^{x_1,x_2,\ldots,x_j})|\omega \right] \right)$$

$$= \log \frac{b - j}{k - j - 1} \sum_{i \in [\ell+1,n]} \mathbb{E}_\omega(\mathbb{E}[\tau(G_i^{x_1,x_2,\ldots,x_j})|\omega]). \tag{13}$$

Here, each subcode C_ω is taken with probability $\lambda_\omega = |C_\omega|/|C|$, and $x_1, \ldots, x_j$ are taken uniformly at random (without repetitions) from C_ω.

Let now $f_{i|\omega}$ be the distribution of the i-th coordinate of the subcode C_ω (with components, say, $f_{i,a|\omega}$) . Then, for $i > \ell$, we can write

$$\mathbb{E}[\tau(G_i^{x_1,\ldots,x_j})|\omega] = (1 + o(1)) \sum_{\substack{\text{distinct} \\ a_1,\ldots,a_j}} f_{i,a_1|\omega} f_{i,a_2|\omega} \cdots f_{i,a_j|\omega} (1 - f_{i,a_1} - \ldots - f_{i,a_j}) \tag{14}$$

where the $o(1)$ is meant as $n \to \infty$ and is due, under the assumption that C_ω grows unbounded with n, to sampling without replacement within C_ω. Now, since $\lambda_\omega = |C_\omega|/|C|$, f_i is actually the expectation of $f_{i|\omega}$ over ω, that is, using a different dummy variable μ to index the subcodes for convenience,

$$f_i = \sum_\mu \lambda_\mu f_{i|\mu} \, .$$

Using this in (14), one notices that when taking a further expectation over ω it is possible to operate a symmetrization in ω and μ. The expectation of (14) over ω can then be written as

$$\mathbb{E}_\omega[\tau(G_i^{x_1,x_2,\ldots,x_j})] = (1 + o(1)) \frac{1}{2} \sum_{\omega,\mu \in \Omega} \lambda_\omega \lambda_\mu \Psi_j(f_{i|\omega}, f_{i|\mu}) \, , \tag{15}$$

so that

$$\mathbb{E}_\omega[\tau(G_i^{x_1,x_2,\ldots,x_j})] \leq (1 + o(1)) \frac{1}{2} \mathbf{M}_j \, . \tag{16}$$

This leads to

$$\log |C| \leq (1 + o(1)) \frac{1}{2} (n - \ell) \mathbf{M}_j \log \frac{b - j}{k - j - 1} \, , \tag{17}$$

from which, using the value of ℓ described above, one deduces

$$R \leq (1 + o(1)) \frac{1}{2} \left[1 - \frac{R}{\log \left(\frac{b}{j-1} \right)} \right] \mathbf{M}_j \log \frac{b - j}{k - j - 1}.$$

Explicitating in R we conclude the proof of the Lemma. $\qquad \square$

The first part of the above derivation follows the same method used in [6]. In particular, the proof of Lemma 1 can be obtained using $j = 2$ and looking at (14) as a quadratic form in $f_{i|\omega}$ with kernel of elements $(1 - f_{i,a_1} - f_{i,a_2})$. The procedure used in [6] can then be applied also for $b \geq k$ with some simple variations.

Proof of Lemma 1. Set $j = 2$ in (14). Proceeding as in [6], it can be shown that the right hand side, as a quadratic form in $f_{i|\omega}$, is a concave function on the simplex of probability distributions if all the values $f_{i,a}$ are not larger than $1/2$. Assume first that this holds for all $i \in [\ell+1, n]$. The expectation over ω is then bounded by the value obtained by replacing both $f_{i|\omega}$ and f_i with a uniform distribution, which is easily evaluated to be $(b^2 - 3b + 2)/b^2$. When used in (13) this gives the bound of Lemma 1. It remains to show that we can assume without loss of generality that $f_{i,a} \leq 1/2$ for all i and a. Again the procedure is a generalization of what was done in [6]. Suppose that there exists a coordinate $i \in \{1, 2, \ldots, n\}$ for which (rename the symbols if needed) $f_{i,1} \geq f_{i,2} \geq \ldots \geq f_{i,b}$ with $f_{i,1} > 1/2$. Note that we must then have $f_{i,1} + f_{i,2} + \ldots + f_{i,k-1} \geq (b + k - 3)/(2b - 2)$. We can build another (b, k)-hash code C' by removing all the codewords in C for which the symbol in the i-th coordinate is in $\{k, k+1, \ldots, b\}$ and by deleting this coordinate in the remaining codewords. Clearly C' has length $n - 1$ and cardinality $|C'| \geq |C| \cdot (b + k - 3)/(2b - 2)$. This process can be iterated, say t times, in order to get a code $\tilde{C}$ of length $n - t$ in which $f_{i,a} \leq 1/2$ for all $i \in \{1, 2, \ldots, n - t\}$ and for all $a \in \{1, 2, \ldots, b\}$ and such that

$$|\tilde{C}| \geq |C| \left(\frac{b + k - 3}{2b - 2} \right)^t . \tag{18}$$

Let $B(b, k)$ be the right hand side of (4). We can apply the previous part of the proof to $\tilde{C}$ and bound the rate R of C as

$$\frac{1}{n} \log |C| \leq \frac{1}{n} \log |\tilde{C}| + \frac{t}{n} \log \left(\frac{2b - 2}{b + k - 3} \right)$$

$$\leq \frac{n - t}{n} B(b, k) + \frac{t}{n} \log \left(\frac{2b - 2}{b + k - 3} \right) + o(1)$$

$$\leq B(b, k) - \frac{t}{n} \left[B(b, k) - \log \left(\frac{2b - 2}{b + k - 3} \right) \right] + o(1) .$$

The proof of the Lemma is concluded if we prove that $B(b, k) > \log \frac{2b-2}{b+k-3}$ for $b \geq k \geq 4$. We verify this inequality considering the following three different ranges of b and k:

1. Suppose that $12 \leq k \leq b \leq (k - 3)^2$. Then

$$B(b, k) \stackrel{(i)}{>} \frac{2}{3} \cdot \frac{(b^2 - 3b + 2) \log b \log \left(\frac{b-2}{k-3} \right)}{b^2 \log(b)} > \frac{2}{3}(1 - 3/b) \log \left(\frac{b - 2}{k - 3} \right)$$

$$\stackrel{(ii)}{\geq} \frac{1}{2} \log \left(\frac{b - 2}{k - 3} \right) ,$$

   where (i) is true since $\log \left(\frac{b-2}{k-3} \right) \leq 1/2 \log b$ for $b \leq (k - 3)^2$, while (ii) since $b \geq 12$. Then, it can be verified that for $b \geq k \geq 12$ we have that

$$\frac{1}{2} \log \left(\frac{b - 2}{k - 3} \right) > \log \left(\frac{2b - 2}{b + k - 3} \right) .$$

2. Suppose that $b \geq 8k - 22$ and $k \geq 4$. Then

$$B(b, k) > \frac{(b^2 - 3b + 2) \log b \log \left(\frac{b-2}{k-3} \right)}{2b^2 \log(b)} > \frac{1}{2}(1 - 3/b) \log \left(\frac{b - 2}{k - 3} \right)$$

$$\stackrel{(i)}{>} \frac{1}{3} \log \left(\frac{b - 2}{k - 3} \right) ,$$

7

where (i) is true since $b > 9$. Then, it can be easily verified that for $b \geq 8k - 22$ we have that

$$\frac{1}{3} \log \left(\frac{b-2}{k-3} \right) \geq 1 > \log \left(\frac{2b-2}{b+k-3} \right) .$$

3. All the cases $b \geq k = 4, 5, \ldots, 11$ can be verified manually or by using a symbolic computation software.

Finally, we see that the ranges of b, as functions of k, in the first two cases intersect because

$$(k-3)^2 \geq 8k - 22$$

is verified for every $k \geq 12$. Therefore the thesis of the lemma follows. $\qquad\square$

3 Bounding the quadratic form

We now enter the problem of determining better upper bounds on the value of $\mathbf{M}_j$ defined in (6). We consider partitions of $\mathcal{P}_b$, the set of probability distributions on b elements, into disjoint subsets to find upper bounds on the quadratic form (6) in terms of simpler ones. If we have a partition $\{\mathcal{P}_b^0, \mathcal{P}_b^1, \ldots, \mathcal{P}_b^r\}$ of $\mathcal{P}_b$ and we define

$$m_{i,h} = \sup_{p \in \mathcal{P}_b^i, q \in \mathcal{P}_b^h} \Psi_j(p, q), \qquad \eta_i = \sum_{p \in \mathcal{P}_b^i} \lambda_p ,$$

then clearly

$$\sum_{p,q} \lambda_p \lambda_q \Psi_j(p, q) \leq \sum_{i,h} \sum_{p \in \mathcal{P}_b^i} \sum_{q \in \mathcal{P}_b^h} \lambda_p \lambda_q m_{i,h} \leq \sum_{i,h} \eta_i \eta_h m_{i,h} . \tag{19}$$

This is a convenient simplification since we have now an r-dimensional problem which we might be able to deal with in some computationally feasible way. We will use this procedure with two different partitions in terms of how balanced or unbalanced the distributions are. We take $b+1$ subsets with some symmetry which allows us to further reduce the complexity.

Partition based on maximum value. We first consider a partition of $\mathcal{P}_b$ in terms of the largest probability value which appears in a distribution. We use a parameter $\epsilon \leq 1/(j+1)$; all quantities will depend on ϵ but we do not write this to avoid cluttering the notation. We define b sets of unbalanced distributions

$$\check{\mathcal{P}}_b^i = \{p \in \mathcal{P}_b : p_i > 1 - \epsilon\}$$

for every $1 \leq i \leq b$, and correspondingly a set of balanced distributions

$$\check{\mathcal{P}}_b^0 = \{p \in \mathcal{P}_b : p_i \leq 1 - \epsilon \; \forall i\} .$$

Note that these are all disjoint sets since $\epsilon < 1/2$ when $j \geq 2$. Following the scheme mentioned above, we can consider the values $m_{i,h}$ and η_i for this specific partition. However, due to symmetry, the values $m_{i,h}$ can be reduced to only four cases, depending on whether p and q are both balanced, one balanced and one unbalanced, or both unbalanced, either on the same coordinate or on different coordinates.

Assuming $1 \leq i, h \leq b$ with $i \neq h$, the following quantities are then well defined and independent of the specific values chosen for i and h

$$\begin{aligned}
\widetilde{M}_1 &= \sup_{p,q \in \check{\mathcal{P}}_b^0} \Psi_j(p; q) & \widetilde{M}_2 &= \sup_{p \in \check{\mathcal{P}}_b^0, q \in \check{\mathcal{P}}_b^i} \Psi_j(p; q) \\
\widetilde{M}_3 &= \sup_{p,q \in \check{\mathcal{P}}_b^i} \Psi_j(p; q) & \widetilde{M}_4 &= \sup_{p \in \check{\mathcal{P}}_b^i, q \in \check{\mathcal{P}}_b^h} \Psi_j(p; q)
\end{aligned} \tag{20}$$

8

These values can then be used in (19) in place of the values $m_{i,h}$.

Partition based on the minimum value. We also consider a partition of $\mathcal{P}_b$ using constraints from below. Again we use a parameter ϵ which will be then tuned. We assume here $\epsilon < 1/b$. Consider now the following disjoint sets of unbalanced distributions

$$\widehat{\mathcal{P}}_b^i = \{p \in \mathcal{P}_b : p_i < \epsilon \,, p_h \geq p_i \ \forall h \,, p_h > p_i \ \forall h < i\}$$

for $1 \leq i \leq b$, that is, distributions in $\widehat{\mathcal{P}}_b^i$ have a minimum component in the i-th coordinate, which is smaller than ϵ, and strictly smaller than any of the preceding components (unless of course $i = 1$). Correspondingly, define a set of balanced distributions as

$$\widehat{\mathcal{P}}_b^0 = \{p \in \mathcal{P}_b : p_i \geq \epsilon \ \forall i\} \,.$$

The symmetry argument mentioned before also applies in this case and we can continue in analogy replacing the $m_{i,h}$ of (19) with the following quantities

$$
\begin{aligned}
\widehat{M_1} &= \sup_{p,q \in \widehat{\mathcal{P}}_b^0} \Psi_j(p;q) & \widehat{M_2} &= \sup_{p \in \widehat{\mathcal{P}}_b^0, q \in \widehat{\mathcal{P}}_b^i} \Psi_j(p;q) \\
\widehat{M_3} &= \sup_{p,q \in \widehat{\mathcal{P}}_b^i} \Psi_j(p;q) & \widehat{M_4} &= \sup_{p \in \widehat{\mathcal{P}}_b^i, q \in \widehat{\mathcal{P}}_b^h} \Psi_j(p;q)
\end{aligned}
\tag{21}
$$

where again $1 \leq i, h \leq b$ with $i \neq h$.

Quadratic form. Applying the above scheme with the symmetric partitions we just defined, we can now rewrite the upper bound of equation (19) in the form

$$\sum_{p,q} \lambda_p \lambda_q \Psi_j(p;q) \leq \eta_0^2 M_1 + 2\eta_0 \sum_{i=1}^b \eta_i M_2 + \sum_{i=1}^b \eta_i^2 M_3 + 2 \sum_{i<h} \eta_i \eta_h M_4 \,, \tag{22}$$

where either the $\widehat{M_i}$'s or the $\widecheck{M_i}$'s can be used in place of the M_i's.

Call M the maximum value achieved by the right hand side of (22) over all possible probability distributions $\eta = (\eta_0, \eta_1, \ldots, \eta_b)$. We show that under assumptions that are verified in our setting, the value of M can be determined explicitly.

Lemma 4. *Let M_1, M_2, M_3 and M_4 be positive numbers such that $M_4 > M_3$ and, for a probability distribution $\eta = (\eta_0, \eta_1, \ldots, \eta_b)$, define the function*

$$f(\eta) = \eta_0^2 M_1 + 2\eta_0 \sum_{i=1}^b \eta_i M_2 + \sum_{i=1}^b \eta_i^2 M_3 + 2 \sum_{i<h} \eta_i \eta_h M_4 \,.$$

Then

$$M = \max_\eta f(\eta) \tag{23}$$

is attained at $\eta_1 = \eta_2 = \ldots = \eta_b$ and

$$\eta_0 = \begin{cases} \dfrac{M_2 - \frac{1}{b} M_3 - \frac{b-1}{b} M_4}{2M_2 - M_1 - \frac{1}{b} M_3 - \frac{b-1}{b} M_4}, & \text{if } M_2 > M_1, M_3, M_4 \\ 0 \text{ or } 1, & \text{otherwise} \end{cases} \,.$$

Proof. Since $\sum_{i=1}^b \eta_i = (1 - \eta_0)$, f can be written as

$$\eta_0^2 M_1 + 2(1 - \eta_0)\eta_0 M_2 + \sum_{i=1}^b \eta_i^2 M_3 + 2 \sum_{i<h} \eta_i \eta_h M_4.$$

Note that

$$\sum_{i=1}^b \eta_i^2 M_3 + 2 \sum_{i<h} \eta_i \eta_h M_4 = \sum_{i=1}^b \eta_i^2 (M_3 - M_4) + (1 - \eta_0)^2 M_4.$$

Since $M_3 < M_4$ and $\sum_{i=1}^{b} \eta_i = 1 - \eta_0$, this sum is maximized when $\eta_1 = \eta_2 = \ldots = \eta_b = (1 - \eta_0)/b$. Therefore we have to maximize the quantity

$$\eta_0^2 M_1 + 2(1 - \eta_0)\eta_0 M_2 + \frac{1}{b}(1 - \eta_0)^2(M_3 - M_4) + (1 - \eta_0)^2 M_4 \,,$$

which is just a quadratic in η_0 that achieves its maximum in $[0, 1]$ at the point described in the statement of the Lemma. $\qquad\square$

We will describe in the next Section our procedure to determine, or upper bound the values $\widehat{M}_i$, $\widetilde{M}_i$. Using these bounds in equation (22) we thus obtain an upper bound on $\mathbf{M}_j$ defined in (6). Applying Lemma 2 we obtain our main result.

Theorem 1. *The bounds of Table 1 hold.*

Remark 1. *The bounds on $R_{(7,7)}$, $R_{(8,8)}$, $R_{(9,8)}$, $R_{(10,9)}$, $R_{(11,10)}$, $R_{(12,10)}$, $R_{(13,11)}$, $R_{(14,12)}$ and $R_{(15,13)}$ are obtained using the partition based on the maximum value $\{\breve{\mathcal{P}}_b^i\}_{i=0,\ldots,b}$. The bounds on $R_{(5,5)}$, $R_{(6,5)}$ and $R_{(6,6)}$ are obtained using the partition based on the minimum value $\{\widehat{\mathcal{P}}_b^i\}_{i=0,\ldots,b}$.*

All other cases, those underlined in Table 1, are obtained computing, as done in [5], the global maximum of Ψ_{k-2}, which is attained for uniform distributions. Therefore, the partitioning process in these particular cases cannot make any improvements.

Based on the result in [7], or its generalization given in equation (4) and on Theorem 1 for $(b, k) = (6, 6)$, we are led to formulate the following conjecture.

Conjecture 1. *For $b \geq k > 3$,*

$$R_{(b,k)} \leq \min_{2 \leq j \leq k-2} \left(\frac{1}{\log \frac{b}{j-1}} + \frac{b^{j+1}}{b^{j+1}\log \frac{b-j}{k-j-1}} \right)^{-1} \,.$$

Note that the conjectured expression can be seen as a modification of the Körner-Marton bound in (3) which takes into account the effects of prefix-based partitions.

4 Computation of M in (23)

In light of Lemma 4, the main problem for the computation of M is determining the $\widetilde{M}_i$'s and $\widehat{M}_i$'s defined in equations (20) and (21). This requires determining the maximum values taken by $\Psi_j(p; q)$ for p and q constrained to specific subsets $\breve{\mathcal{P}}_b^i$ or $\widehat{\mathcal{P}}_b^i$. Following a procedure similar to that of [5], here we prove that, under certain conditions, the distributions p and q achieving those maxima have many equal components. This, together with other simplifications that will be presented later, allows us to greatly reduce the complexity in the search for the maxima (see Remarks 2 and 3 below). For this purpose we first present three Lemmas, which generalize Lemmas 3, 4 and 5 of [5].

Lemma 5 (Extension of Lemma 3 in [5]). *Let ℓ be an integer in $[2, b]$ and, for $i \in [1, \ell]$, consider the nonempty intervals $I_i = [a_i, b_i]$ and $J_i = [c_i, d_i]$. Set $D_p = I_1 \times I_2 \times \cdots \times I_\ell \times \overline{p_{\ell+1}} \times \cdots \times \overline{p_b}$ and $D_q = J_1 \times J_2 \times \cdots \times J_\ell \times \overline{q_{\ell+1}} \times \cdots \times \overline{q_b}$. Consider the set D of pairs of probability vectors (p, q) such that p belongs to D_p and q belongs to D_q. Then if $(\overline{p}; \overline{q})$ is a maximum point for Ψ_j in D then either $\overline{p_i} = \overline{p_h}$ and $\overline{q_i} = \overline{q_h}$ for any $i, h \in [1, \ell]$ or there is a maximum for Ψ_j on the boundary of D (as projected on the first ℓ coordinates).*

Note that, in particular, in the latter case, we have a maximum point $(\overline{p}; \overline{q})$ for Ψ_j with at least one index $i \in [1, \ell]$ such that either $\overline{p_i} \in \{a_i, b_i\}$ or $\overline{q_i} \in \{c_i, d_i\}$.

Proof. Let us assume that $\overline{P} = (\overline{p}; \overline{q})$ is a maximum point for Ψ_j in D and $\overline{p_1}, \overline{p_2}, \ldots, \overline{p_\ell}$ or $\overline{q_1}, \overline{q_2}, \ldots, \overline{q_\ell}$ are not all equal. By symmetry, assume without loss of generality that $\overline{p_1} \neq \overline{p_2}$. Now, if $\overline{P}$ is a maximum for Ψ_j not on the boundary D, then it is a maximum also under the stronger constraints $p_1 + p_2 = c_1$, $q_1 + q_2 = c_2$ where $c_1 = \overline{p_1} + \overline{p_1}$, $c_2 = \overline{q_1} + \overline{q_2}$, and $p_i = \overline{p_i}, q_i = \overline{q_i}$ for $i \in \{3, 4, \ldots, \ell\}$. Then, let us consider the line L of points $P(t)$ such that

$$P(t) = P(0) + t\left(\frac{\overline{p_1} - \overline{p_2}}{2}, \frac{-\overline{p_1} + \overline{p_2}}{2}, 0, \ldots, 0; \frac{\overline{q_1} - \overline{q_2}}{2}, \frac{-\overline{q_1} + \overline{q_2}}{2}, 0, \ldots, 0\right),$$

where $P(0) = (\frac{\overline{p_1} + \overline{p_2}}{2}, \frac{\overline{p_1} + \overline{p_2}}{2}, \overline{p_3}, \ldots, \overline{p_b}; \frac{\overline{q_1} + \overline{q_2}}{2}, \frac{\overline{q_1} + \overline{q_2}}{2}, \overline{q_3}, \ldots, \overline{q_b})$, so that $P(1) = \bar{P}$.

It is easy to see that $\Psi_j(P(t))$ is of degree 2 and, if $\overline{P}$ is not on the boundary of D, then $t = 1$ must be a stationary point for $\Psi_j(P(t))$. Moreover $\Psi_j(P(t))$ is an even function because:

$$\Psi_j(P(-t)) = P(0) - t\left(\frac{\overline{p_1} - \overline{p_2}}{2}, \frac{-\overline{p_1} + \overline{p_2}}{2}, 0, \ldots, 0; \frac{\overline{q_1} - \overline{q_2}}{2}, \frac{-\overline{q_1} + \overline{q_2}}{2}, 0, \ldots, 0\right)$$

$$= P(0) + t\left(\frac{\overline{p_2} - \overline{p_1}}{2}, \frac{-\overline{p_2} + \overline{p_1}}{2}, 0, \ldots, 0; \frac{\overline{q_2} - \overline{q_1}}{2}, \frac{-\overline{q_2} + \overline{q_1}}{2}, 0, \ldots, 0\right)$$

$$= \Psi_j(P(t)).$$

This means that $\Psi_j(P(t)) = \alpha t^2 + \beta$ for some α and β in $\mathbb{R}$. Therefore $t = 0$ would be another stationary point for $\Psi_j(P(t))$ but this is possible only if $\alpha = 0$ that is $\Psi_j(P(t))$ is a constant.

The thesis follows because, in this case, the maximum is also attained on the boundary of D. □

With essentially the same proof we obtain

Lemma 6 (Extension of Lemma 4 in [5]). *Let ℓ be an integer in $[2, b]$ and, for $i \in [1, \ell]$, consider the nonempty intervals $I_i = [a_i, b_i]$. Set $D_p = I_1 \times I_2 \times \cdots \times I_\ell \times \overline{p_{\ell+1}} \times \cdots \times \overline{p_b}$ and $D_q = \overline{q_1} \times \overline{q_2} \times \cdots \times \overline{q_\ell} \times \overline{q_{\ell+1}} \times \cdots \times \overline{q_b}$ where $\overline{q_i} = \overline{q_h}$ for any $i, h \in [1, \ell]$. Consider the set D of pairs of probability vectors (p, q) such that p belongs to D_p and q belongs to D_q. Then if $(\overline{p}; \overline{q})$ is a maximum point for Ψ_j in D then either $\overline{p_i} = \overline{p_h}$ for any $i \in [1, \ell]$ or there is a maximum for Ψ_j on the boundary of D.*

Note that, in particular, in the latter case, we have a maximum point $(\overline{p}; \overline{q})$ for Ψ_j with at least one index $i \in [1, \ell]$ such that $\overline{p_i} \in \{a_i, b_i\}$.

Now we present a Lemma that allows us to assume that the coordinates of p and q are properly rearranged depending on their values.

Lemma 7 (Extension of Lemma 5 in [5]). *If $p_1 \leq p_2$, and $q_1 \leq q_2$, then*

$$\Psi_j(p_1, p_2, p_3, \ldots, p_b; q_1, q_2, q_3, \ldots, q_b) \leq \Psi_j(p_1, p_2, p_3, \ldots, p_b; q_2, q_1, q_3, \ldots, q_b). \quad (24)$$

Proof. Using the definition of Ψ_j in eq. (5), inequality (24) can be restated by only considering the terms in the summation which differ in the two sides, that is, those corresponding to permutations σ such that $1 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j+1) = 2$ and $2 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j+1) = 1$. Hence inequality (24) becomes

$$(p_1 q_2 + p_2 q_1) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$

$$\leq (p_1 q_1 + p_2 q_2) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$

which can be restated as

$$(p_2 - p_1)(q_2 - q_1) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} \geq 0$$

This is always true since $p_1 \leq p_2$ and $q_1 \leq q_2$. □

Using the above lemmas, we are able to isolate a relatively small set of possible configurations for the p and q which give the value $\widetilde{M}_1$.

Proposition 1. $\widetilde{M}_1$ *is attained in one of the following points:*

1) *for $(p;q)$ of the form*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2-2},\gamma,1-\epsilon;\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2-2},1-\epsilon,\zeta)$$

where $\alpha,\delta > 0,\quad \beta,\eta,\gamma,\zeta \geq 0$ and

$$l_2\alpha + (b-l_1-l_2-2)\beta + \gamma + (1-\epsilon) = 1 = l_1\delta + (b-l_1-l_2-2)\eta + (1-\epsilon) + \zeta;$$

2) *for $(p;q)$ of the form*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2-1},\gamma;\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2-1},1-\epsilon)$$

where $\alpha,\delta > 0,\quad \beta,\eta,\gamma \geq 0$ and

$$l_2\alpha + (b-l_1-l_2-1)\beta + \gamma = 1 = l_1\delta + (b-l_1-l_2-1)\eta + (1-\epsilon);$$

3) *for $(p;q)$ of the form*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2};\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2})$$

where $\alpha,\delta > 0,\quad \beta,\eta \geq 0$ and

$$l_2\alpha + (b-l_1-l_2)\beta = 1 = l_1\delta + (b-l_1-l_2)\eta.$$

Proof. Remember that the value $\widetilde{M}_1$ is the maximum of Ψ_j over pairs (p,q) with p and q in $\check{\mathcal{P}}_b^0$. Moreover, due to Lemma 7, we have that p and q do not have a value $1-\epsilon$ in the same coordinate. Similarly, again because of Lemma 7, either the zeros of p and q are in different positions (i.e. if $p_i = 0$ then $q_i \neq 0$) or for any i at least one between p_i and q_i is zero.

According to the positions where values $1-\epsilon$ and zero can appear as coordinates of p and q, we have that $\widetilde{M}_1$ is attained in one of the following points:

1A) p and q have respectively l_1 and l_2 zeros in different positions, both have a coordinate with value $1-\epsilon$ and those are in different positions:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{\beta_1,\ldots,\beta_{b-l_1-l_2-2}}^{b-l_1-l_2-2},\gamma,1-\epsilon;\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta_1,\ldots,\eta_{b-l_1-l_2-2}}^{b-l_1-l_2-2},1-\epsilon,\zeta);$$

1B) p and q have respectively l_1 and l_2 zeros in different positions, additional $b-l_1-l_2-2$ zeros in the same positions, both have a coordinate with value $1-\epsilon$ and those are in different positions:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2-2},0,1-\epsilon;\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2-2},1-\epsilon,0);$$

2A) p and q have respectively l_1 and l_2 zeros in different positions, p has no coordinate of value $1 - \epsilon$ but q has:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{\beta_1,\ldots,\beta_{b-l_1-l_2-1}}^{b-l_1-l_2-1},\gamma;\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta_1,\ldots,\eta_{b-l_1-l_2-1}}^{b-l_1-l_2-1},1-\epsilon);$$

2B) p and q have respectively l_1 and l_2 zeros in different positions, additional $b - l_1 - l_2 - 1$ zeros in the same positions, p has no coordinate of value $1 - \epsilon$ but q has:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2-1},0;\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2-1},1-\epsilon);$$

3A) p and q have respectively l_1 and l_2 zeros in different positions and both have no coordinates with value $1 - \epsilon$:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{\beta_1,\ldots,\beta_{b-l_1-l_2}}^{b-l_1-l_2};\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta_1,\ldots,\eta_{b-l_1-l_2}}^{b-l_1-l_2});$$

3B) p and q have respectively l_1 and l_2 zeros in different positions, additional $b - l_1 - l_2$ zeros in the same positions and neither has a coordiante of value $1 - \epsilon$:

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha_1,\ldots,\alpha_{l_2}}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2};\overbrace{\delta_1,\ldots,\delta_{l_1}}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{0,\ldots,0}^{b-l_1-l_2}).$$

Moreover, in all those cases, the allowed domains for p and q satisfy either the hypothesis of Lemma 5 or those of Lemma 6. This means that we can average the α's (i.e. we can assume that all the α's are equal), the β's, the δ's, and the η's. The thesis follows allowing β and η to possibly be zero and noting that the case $1B$ becomes a subcase of $1A$, $2B$ becomes a subcase of $2A$ and $3B$ becomes a subcase of $3A$. $\qquad\square$

Remark 2. *As seen in Proposition 1, Lemmas 5, 6 and 7 reduce the maxima candidates to a finite set of possible configurations. Still, the number of such configurations increases with b, and the ensuing optimization problems depend on 4 free variables in the case 1. The direct evaluation of the maxima of Ψ_j on those configurations can in principle be performed by symbolic computation software, but the resources needed are excessive. In the following lemmas, we provide additional simplifications to obtain the exact evaluations of the maxima.*

Due to the following lemma, whose proof can be found in the appendix, we can assume that the number of zeros that appear in p (resp. in q) is either $b - 2$ or at most $b - j$. Note that this simplification does not decrease the number of free variables but it reduces the total number of cases.

Lemma 8 (Extension of Lemma 6 in [5]). *Suppose that $q_1 \leq q_2 \leq \ldots \leq q_{j-1}$. If all the p_i are less than or equal to $1 - \alpha$ where $0 \leq \alpha < 1$, then*

$$\Psi_j(p_1, p_2, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b)$$
$$\leq \Psi_j(1 - \alpha, \alpha, 0, \ldots, 0; q_1, q_2, \ldots, q_b). \qquad (25)$$

The following lemma, whose proof can be found in the appendix, takes care of the cases when there is at least one element greater or equal to $1 - \epsilon$ in p or q vector. If this element is p_1, because of Lemma 7 we can assume q_1 is the minimum among the q-values if we are maximizing Ψ_j. For the evaluation of $\widetilde{M}_1$, this implies that $q_1 = 0$ whenever $p_1 = 1 - \epsilon$ and vice-versa.

Lemma 9. *Assume that $\epsilon \leq \frac{1}{j+1}$, $p_1 \geq 1 - \epsilon$ and $q_1 \leq q_2 \leq \ldots \leq q_b$. Then*

$$\Psi_j(p_1, p_2, \ldots, p_b; q_1, q_2, \ldots, q_b) \leq \Psi_j(p_1, p_2, \ldots, p_b; 0, q_1 + q_2, q_3, \ldots, q_b). \qquad (26)$$

Thanks to Lemmas 8 and 9, we obtain the following proposition.

Proposition 2. $\widetilde{M}_1$ *is attained in one of the following points:*

1) *for $(p; q)$ of the form*

$$(\overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2 - 2}, 0, 1 - \epsilon; \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2 - 2}, 1 - \epsilon, 0)$$

where $\alpha, \beta, \delta, \eta \geq 0$ and

$$l_2 \alpha + (b - l_1 - l_2 - 2)\beta + (1 - \epsilon) = 1 = l_1 \delta + (b - l_1 - l_2 - 2)\eta + (1 - \epsilon);$$

2) *for $(p; q)$ of the form*

$$(\overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2 - 1}, 0; \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2 - 1}, 1 - \epsilon)$$

where $\alpha, \beta, \delta, \eta \geq 0$ and

$$l_2 \alpha + (b - l_1 - l_2 - 1)\beta = 1 = l_1 \delta + (b - l_1 - l_2 - 1)\eta + (1 - \epsilon);$$

3) *for $(p; q)$ of the form*

$$(\overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2}; \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2})$$

where $\alpha, \beta, \delta, \eta \geq 0$ and

$$l_2 \alpha + (b - l_1 - l_2)\beta = 1 = l_1 \delta + (b - l_1 - l_2)\eta.$$

Moreover, we can assume that the number of zeros that appear in p and in q is either $b - 2$ or at most $b - j$.

Proof. We consider the finite list of cases provided by Proposition 1 and we relax the domains of p and q allowing α and δ to be 0. Here, due to Lemma 9, a maximum with exactly one element equal to $1 - \epsilon$ in p (resp. q) implies a zero in the same coordinate of q (resp. p). Finally, because of Lemma 8, a maximum with $b - j + 1$ or more coordinates in p (resp. q) equal to zero is also attained in a point of the form $p = (1 - \epsilon, \epsilon, 0, \ldots, 0)$ ($q = (1 - \epsilon, \epsilon, 0, \ldots, 0)$). $\square$

14

Proposition 3. $\widecheck{M}_2$ *is upper bounded by the global maximum of Ψ_j which is attained in a point $(p;q)$ of the following form:*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2};\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2})$$

where $\alpha,\beta,\delta,\eta \geq 0$ and

$$l_2\alpha + (b-l_1-l_2)\beta = 1 = l_1\delta + (b-l_1-l_2)\eta.$$

Moreover, we can assume that the number of zeros that appear in p and in q is either $b-1$ or at most $b-j$.

Proof. In order to find the global maximum of Ψ_j we need no restriction on the pairs (p,q), i.e., $p \in [0,1]^b$ and $q \in [0,1]^b$. Using Lemmas 5, 6, 7 and 8, we can easily derived the desired points. $\qquad\square$

Now, to provide a list of possible maxima also for the other $\widecheck{M}_i$ and $\widehat{M}_i$, we need also the following additional lemma.

Lemma 10. *Assume that $\epsilon < \frac{1}{2}$, $q_1 \geq 1-\epsilon$ and $0 < \delta \leq \epsilon$, then*

$$\Psi_j(1-\epsilon+\delta,p_2,p_3,\ldots,p_b;q_1,q_2,\ldots,q_b) < \Psi_j(1-\epsilon,p_2+\delta,p_3,\ldots,p_b;q_1,q_2,\ldots,q_b). \quad (27)$$

Thanks to Lemma 10, whose proof can be found in the appendix, we obtain the following proposition.

Proposition 4. $\widecheck{M}_3$ *is attained in a point $(p;q)$ of the following form:*

$$(1-\epsilon,\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2-1};1-\epsilon,\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2-1})$$

where $\alpha,\beta,\delta,\eta \geq 0$ and

$$l_2\alpha + (b-l_1-l_2-1)\beta = \epsilon = l_1\delta + (b-l_1-l_2-1)\eta.$$

Moreover, we can assume that the number of zeros that appear in p and in q is either $b-2$ or at most $b-j$.

Proof. In order to find the values $\widecheck{M}_3$ we need to restrict the function Ψ_j to the pairs (p,q) such that p and q belong to $\widecheck{\mathcal{P}}_b^1$ (by symmetry we can fix an arbitrary coordinate).

Using Lemmas 5, 6, 7 and 8, we see that $\widecheck{M}_3$ is attained in a point of the following form:

$$(\gamma,\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2-1};\zeta,\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2-1})$$

where $\alpha,\beta,\delta,\eta \geq 0,\quad \gamma,\zeta \geq 1-\epsilon$ *and*

$$\gamma + l_2\alpha + (b-l_1-l_2-1)\beta = 1 = \zeta + l_1\delta + (b-l_1-l_2-1)\eta.$$

Finally, because of Lemma 10 a maximum with $\gamma,\zeta \geq 1-\epsilon$ is also attained in a point with $\gamma = \zeta = 1-\epsilon$. $\qquad\square$

Proposition 5. $\widecheck{M}_4$ *is attained in one of the following points:*

1) for $(p; q)$ of the form

$$(\gamma, \alpha, \ldots, \alpha, 0; 0, \delta, \ldots, \delta, \zeta)$$

where $\alpha, \delta \geq 0$, $\gamma, \zeta \geq 1 - \epsilon$, and

$$\gamma + (b - 2)\alpha = 1 = (b - 2)\delta + \zeta.$$

2) for $(p; q)$ of the form

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{b - l_1 - 2}, 0; 0, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{\eta, \ldots, \eta}_{b - l_1 - 2}, \zeta)$$

where $\alpha, \delta, \eta \geq 0$, $\gamma \geq 1 - \epsilon$, $\zeta \in \{1 - \epsilon, 1\}$, and

$$(b - l_1 - 2)\alpha + \gamma = 1 = l_1 \delta + (b - l_1 - 2)\eta + \zeta.$$

3) for $(p; q)$ of the form

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2 - 2}, 0; 0, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2 - 2}, \zeta)$$

where $\alpha, \beta, \delta, \eta \geq 0$, $\gamma, \zeta \in \{1 - \epsilon, 1\}$, and

$$l_2 \alpha + (b - l_1 - l_2 - 2)\beta + \gamma = 1 = l_1 \delta + (b - l_1 - l_2 - 2)\eta + \zeta.$$

Moreover, we can assume that the number of zeros that appear in p and in q is either $b - 1$ or at most $b - j$.

Proof. In order to find the values $\widetilde{M_4}$ we need to restrict the function Ψ_j to the pairs $(p; q)$ such that p belongs to $\check{\mathcal{P}}_b^1$ and q belongs to $\check{\mathcal{P}}_b^b$ (by symmetry we can choose, arbitrarily, two different coordinates).

Using Lemmas 5, 6, 7, 8 and 9 we see that $\widetilde{M_4}$ is attained in a point of the following form:

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2 - 2}, 0; 0, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2 - 2}, \zeta)$$

where $\alpha, \beta, \delta, \eta \geq 0$, $\gamma, \zeta \geq 1 - \epsilon$, and

$$l_2 \alpha + (b - l_1 - l_2 - 2)\beta + \gamma = 1 = l_1 \delta + (b - l_1 - l_2 - 2)\eta + \zeta.$$

Finally, we can split this case into three cases. The first one is for $l_1 = l_2 = 0$, the second one for $l_1 > 0, l_2 = 0$ and the third one for $l_1, l_2 > 0$. By symmetry the case $l_1 = 0, l_2 > 0$ is included in the second case. For the second case, by Lemma 6 it is easy to see that δ or ζ must be on the boundary in order to be a valid point for $\widetilde{M_4}$. The same argument can be carried out for the third case which implies that $\gamma, \zeta \in \{1 - \epsilon, 1\}$. $\qquad\square$

Proposition 6. $\widehat{M_1}$ *is attained in a point $(p; q)$ of the following form:*

$$(\overbrace{\epsilon, \ldots, \epsilon}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b - l_1 - l_2}; \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{\epsilon, \ldots, \epsilon}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b - l_1 - l_2})$$

where $\alpha, \beta, \delta, \eta \geq \epsilon$ and

$$l_1 \epsilon + \alpha + (b - l_1 - l_2)\beta = 1 = l_2 \epsilon + l_1 \delta + (b - l_1 - l_2)\eta.$$

Proof. In order to find the values $\widehat{M}_1$ we need to restrict the function Ψ_j to the pairs (p,q) such that p and q belong to $\widehat{\mathcal{P}}_b^0$. Using Lemmas 5, 6 and 7 we obtain the thesis. $\qquad\square$

Proposition 7. $\widehat{M}_2$ *is attained in one of the following points:*

1) *for $(p;q)$ of the form*

$$(\overbrace{\alpha,\ldots,\alpha}^{l_1},\underbrace{\beta,\ldots,\beta}_{b-l_1};\overbrace{\eta,\ldots,\eta}^{l_1},\underbrace{\epsilon,\ldots,\epsilon}_{b-l_1})$$

where $0 \le \alpha \le \epsilon, \quad \beta \ge 0, \quad \eta \ge \epsilon,$ *and*

$$l_1\alpha + (b-l_1)\beta = 1 = l_1\eta + (b-l_1)\epsilon.$$

2) *for $(p;q)$ of the form*

$$(\epsilon,\overbrace{\alpha,\ldots,\alpha}^{l_1},\underbrace{\beta,\ldots,\beta}_{b-l_1-1};\zeta,\overbrace{\eta,\ldots,\eta}^{l_1},\underbrace{\epsilon,\ldots,\epsilon}_{b-l_1-1})$$

where $\alpha, \beta \ge 0, \quad \zeta, \eta \ge \epsilon,$ *and*

$$\epsilon + l_1\alpha + (b-l_1-1)\beta = 1 = \zeta + l_1\eta + (b-l_1-1)\epsilon.$$

3) *for $(p;q)$ of the form*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2};\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{\eta,\ldots,\eta}_{l_2},\overbrace{\epsilon,\ldots,\epsilon}^{b-l_1-l_2})$$

where $\alpha, \beta \ge 0, \quad \delta, \eta \ge \epsilon$ *and*

$$l_2\alpha + (b-l_1-l_2)\beta = 1 = l_1\delta + (b-l_1-l_2)\eta.$$

Moreover, we can assume that the number of zeros that appear in p is either $b-1$ or at most $b-j$.

Proof. In order to find the values $\widehat{M}_2$ we need to restrict the function Ψ_j to the pairs (p,q) such that p belongs to $\widehat{\mathcal{P}}_b^1$ and q belongs to $\widehat{\mathcal{P}}_b^0$. In addition, we relax the domain of p by removing the constraint on p_1 to be a minimum coordinate., i.e., $p \in [0,\epsilon] \times [0,1]^{b-1}$. However, this implies that p belongs to $\widehat{\mathcal{P}}_b^i$ for some $i \in [1,b]$. Therefore, by symmetry, we are still considering valid candidates for $\widehat{M}_2$ under this domain.

Using Lemmas 5, 6, 7 and 8, we see that $\widehat{M}_2$ is attained in a point of the following form:

$$(\gamma,\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{b-l_1-l_2-1},\overbrace{\beta,\ldots,\beta}^{l_2};\zeta,\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{\eta,\ldots,\eta}_{b-l_1-l_2-1},\overbrace{\epsilon,\ldots,\epsilon}^{l_2})$$

where $\alpha, \beta \ge 0, \quad 0 \le \gamma \le \epsilon, \quad \zeta, \delta, \eta \ge \epsilon,$ and

$$\gamma + (b-l_1-l_2-1)\alpha + l_2\beta = 1 = \zeta + l_1\delta + (b-l_1-l_2-1)\eta + l_2\epsilon.$$

Finally, we can split this case into three cases. The first one is when $l_1 = 0$ and the average between γ and the α-components is less than or equal to ϵ, the second one for $\gamma = \epsilon$ and $l_1 = 0$ while the third one for $\gamma = 0$ and $l_1 \ge 0$. We have not considered the case $\gamma = \epsilon$ and $l_1 > 0$ since it is a subcase of the third one. $\qquad\square$

Proposition 8. *An upper bound on $\widehat{M}_3$ is obtained by computing the maximum of Ψ_j over points of the following form:*

1) for $(p;q)$ of the form

$$(\beta, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b-l_1-l_2-1}; \eta, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b-l_1-l_2-1})$$

where $\alpha, \delta \geq 0$, $\quad 0 \leq \beta, \eta \leq \epsilon$ and

$$l_2\alpha + (b - l_1 - l_2)\beta = 1 = l_1\delta + (b - l_1 - l_2)\eta.$$

2) for $(p;q)$ of the form

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b-l_1-l_2-1}; 0, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b-l_1-l_2-1})$$

where $\alpha, \beta, \delta, \eta \geq 0$, $\quad 0 \leq \gamma \leq \epsilon$ and

$$\gamma + l_2\alpha + (b - l_1 - l_2 - 1)\beta = 1 = l_1\delta + (b - l_1 - l_2 - 1)\eta.$$

3) for $(p;q)$ of the form

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b-l_1-l_2-1}; \epsilon, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b-l_1-l_2-1})$$

where $\alpha, \beta, \delta, \eta \geq 0$, $\quad 0 \leq \gamma \leq \epsilon$ and

$$\gamma + l_2\alpha + (b - l_1 - l_2 - 1)\beta = 1 = \epsilon + l_1\delta + (b - l_1 - l_2 - 1)\eta.$$

Moreover, we can assume that the number of zeros that appear in p and in q is either $b - 1$ or at most $b - j$.

Proof. In order to find and upper on the values $\widehat{M}_3$ we need to restrict the function Ψ_j to the pairs (p, q) such that p and q belong to $\widehat{\mathcal{P}}_b^1$ (by symmetry we can fix an arbitrary coordinate). In addition, we relax the domains of p and q by removing the constraints on p_1 and q_1 to be minimum components, i.e., $p, q \in [0, \epsilon] \times [0, 1]^{b-1}$.

Using Lemmas 5, 6, 7 and 8, we see that under this extended domain an upper bound on $\widehat{M}_3$ is attained in a point of the following form:

$$(\gamma, \overbrace{0, \ldots, 0}^{l_1}, \underbrace{\alpha, \ldots, \alpha}_{l_2}, \overbrace{\beta, \ldots, \beta}^{b-l_1-l_2-1}; \zeta, \overbrace{\delta, \ldots, \delta}^{l_1}, \underbrace{0, \ldots, 0}_{l_2}, \overbrace{\eta, \ldots, \eta}^{b-l_1-l_2-1})$$

where $\alpha, \beta, \delta, \eta \geq 0$, $\quad 0 \leq \gamma, \zeta \leq \epsilon$ and

$$\gamma + l_2\alpha + (b - l_1 - l_2 - 1)\beta = 1 = \zeta + l_1\delta + (b - l_1 - l_2 - 1)\eta.$$

Finally, we can split this case into three cases. The first one is when the averages between γ and the β-components and between ζ and the η-components are less than or equal to ϵ, the second one for $0 \leq \gamma \leq \epsilon$ and $\zeta = 0$, and the third one for $0 \leq \gamma \leq \epsilon$ and $\zeta = \epsilon$. By symmetry, the cases in which $\gamma = 0$ and $0 \leq \zeta \leq \epsilon$ or $\gamma = \epsilon$ and $0 \leq \zeta \leq \epsilon$ are included in the second and third cases. $\qquad \square$

Proposition 9. $\widehat{M}_4$ *is upper bounded by the global maximum of* Ψ_j *which is attained in a point* $(p; q)$ *of the following form:*

$$(\overbrace{0,\ldots,0}^{l_1},\underbrace{\alpha,\ldots,\alpha}_{l_2},\overbrace{\beta,\ldots,\beta}^{b-l_1-l_2};\overbrace{\delta,\ldots,\delta}^{l_1},\underbrace{0,\ldots,0}_{l_2},\overbrace{\eta,\ldots,\eta}^{b-l_1-l_2})$$

where $\alpha,\beta,\delta,\eta \geq 0$ *and*

$$l_2\alpha + (b - l_1 - l_2)\beta = 1 = l_1\delta + (b - l_1 - l_2)\eta.$$

Moreover, we can assume that the number of zeros that appear in p *and in* q *is either* $b-1$ *or at most* $b-j$.

Proof. Analogous to the proof of Proposition 3. □

Remark 3. *Each configuration that appears in the list of possible maxima in the previous propositions leads to an optimization problem that depends on at most 3 free variables. Therefore, for given* b *and* k, *we can analytically determine, using Mathematica, those maxima.*

The previous propositions allow us to determine a finite list of maxima candidates for each $\widetilde{M}_i$ and $\widehat{M}_i$. We have analytically determined and inspected using Mathematica all the possible maximum points. We have restricted our attention to (b, k)-cases for small b and k in order to avoid excessive computational complexity. It is important to note that for the (b, k)-cases that we have considered (see Propositions 10 and 11) the global maximum of Ψ_j, for $j = k - 2$, satisfy the domains of $\widetilde{M}_2$ and $\widehat{M}_4$. Therefore for these particular cases, we are not upper bounding the values of $\widetilde{M}_2$ and $\widehat{M}_4$ but, instead, we are computing the exact values. Based on the results of computations, we choose the values of j and ϵ for each (b, k)-case to improve the current best-known bounds on $R_{(b,k)}$. A more careful choice of these parameters could lead to better bounds except for the case $b = k = 6$ (see Remark 5).

Proposition 10. *For* $j = k - 2$, *for the values of* ϵ *shown, the* $\widetilde{M}_i$*'s are as shown in the following table*

(b, k)	ϵ	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_4$
$(7, 7)$	$9/100$	0.085679	0.092593	0.000006	0.000107
$(8, 8)$	$3/25$	0.038453	0.042840	0.000002	0.000022
$(9, 8)$	$1/10$	0.075870	0.076905	0.000001	0.000015
$(10, 9)$	$1/15$	0.036289	0.037935	$3.4 \cdot 10^{-9}$	$8.5 \cdot 10^{-8}$
$(11, 10)$	$1/11$	0.016928	0.018144	$1.4 \cdot 10^{-9}$	$2.7 \cdot 10^{-8}$
$(12, 10)$	$1/20$	0.030945	0.031036	$2.1 \cdot 10^{-11}$	$7.0 \cdot 10^{-9}$
$(13, 11)$	$1/25$	0.015057	0.015473	$7.8 \cdot 10^{-14}$	$3.5 \cdot 10^{-12}$
$(14, 12)$	$1/13$	0.007176	0.007529	$1.2 \cdot 10^{-12}$	$2.6 \cdot 10^{-11}$
$(15, 13)$	$1/12$	0.003360	0.003588	$1.1 \cdot 10^{-13}$	$2.3 \cdot 10^{-12}$

$\widetilde{M}_1$ is attained at $(\frac{1}{b},\ldots,\frac{1}{b};\frac{1}{b},\ldots,\frac{1}{b})$

$\widetilde{M}_2$ is attained at $(1, 0,\ldots,0; 0, \frac{1}{b-1},\ldots,\frac{1}{b-1})$

$\widetilde{M}_3$ is attained at $(1 - \epsilon, \frac{\epsilon}{b-1},\ldots,\frac{\epsilon}{b-1}; 1 - \epsilon, \frac{\epsilon}{b-1},\ldots,\frac{\epsilon}{b-1})$

$\widetilde{M}_4$ is attained at $(1 - \epsilon, \frac{\epsilon}{b-2},\ldots,\frac{\epsilon}{b-2}, 0; 0, \frac{\epsilon}{b-2},\ldots,\frac{\epsilon}{b-2}, 1 - \epsilon)$

Proposition 11. *For $j = 3$, $(b, k) = (5, 5)$ and $\epsilon = \frac{1}{44}(4 + \sqrt{5})$, the $\widehat{M}_i$'s are as shown in the following table*

$\widehat{M}_i$	Attained at point $(p; q)$	Value $\approx$
$\widehat{M}_1$	$(\epsilon, \frac{1-\epsilon}{b-1}, \ldots, \frac{1-\epsilon}{b-1}; \gamma, \delta, \ldots, \delta), \delta \approx 0.185275$	0.384033
$\widehat{M}_2$	$(0, \frac{1}{b-1}, \ldots, \frac{1}{b-1}; \gamma, \delta, \ldots, \delta), \delta = \epsilon$	0.389226
$\widehat{M}_3$	$(\epsilon, \frac{1-\epsilon}{b-2}, \ldots, \frac{1-\epsilon}{b-2}, 0; \epsilon, \alpha, \ldots, \alpha, \beta), \beta \approx 0.454183$	0.374759
$\widehat{M}_4$	$(0, \frac{1}{b-1}, \ldots, \frac{1}{b-1}; \gamma, \delta, \ldots, \delta), \delta = \epsilon$	0.389226

For $j = 3$, $(b, k) = (6, 5)$ and $\epsilon = \frac{1}{10}$, the $\widehat{M}_i$'s are as shown in the following table

$\widehat{M}_i$	Attained at point $(p; q)$	Value $\approx$
$\widehat{M}_1$	$(\epsilon, \frac{1-\epsilon}{b-1}, \ldots, \frac{1-\epsilon}{b-1}; \gamma, \delta, \ldots, \delta), \delta \approx 0.153159$	0.555625
$\widehat{M}_2$	$(0, \frac{1}{b-1}, \ldots, \frac{1}{b-1}; \gamma, \delta, \ldots, \delta), \delta \approx 0.130217$	0.558467
$\widehat{M}_3$	$(\epsilon, \frac{1-\epsilon}{b-2}, \ldots, \frac{1-\epsilon}{b-2}, 0; \epsilon, \alpha, \ldots, \alpha, \beta), \beta \approx 0.376930$	0.535106
$\widehat{M}_4$	$(0, \frac{1}{b-1}, \ldots, \frac{1}{b-1}; \gamma, \delta, \ldots, \delta), \delta \approx 0.130217$	0.558467

For $j = 4$, $(b, k) = (6, 6)$ and $\epsilon = \frac{1}{20}$, the $\widehat{M}_i$'s are as shown in the following table

$\widehat{M}_i$	Attained at point $(p; q)$	Value $\approx$
$\widehat{M}_1$	$(\frac{1}{b}, \ldots, \frac{1}{b}; \frac{1}{b}, \ldots, \frac{1}{b})$	0.185185
$\widehat{M}_2$	$(\epsilon, \frac{1-\epsilon}{b-1}, \ldots, \frac{1-\epsilon}{b-1}; \gamma, \delta, \ldots, \delta), \delta \approx 0.147757$	0.178857
$\widehat{M}_3$	$(\epsilon, 0, \frac{1-\epsilon}{b-2}, \ldots, \frac{1-\epsilon}{b-2}; 0, 1, 0, \ldots, 0)$	0.140664
$\widehat{M}_4$	$(1, 0, \ldots, 0; 0, \frac{1}{b-1}, \ldots, \frac{1}{b-1})$	0.192000

Remark 4. *The values reported for $\widehat{M}_3$ are not approximate values of the exact values of $\widehat{M}_3$ but upper bounds. We point out that the value $\widehat{M}_1$ for $b = k = 6$ is only attained for uniform distributions. This will be important for a qualitative analysis of our bound on $R_{(b,k)}$ for different values of b and k, see Section 5.*

As a consequence of Propositions 10, 11 and equation (22) we are able to evaluate the values of M for both the partitions $\{\check{P}_b^i\}_{i=0,\ldots,b}$ and $\{\widehat{P}_b^i\}_{i=0,\ldots,b}$. Then we state the following theorem

Theorem 2. *Using the partition $\{\check{P}_b^i\}_{i=0,\ldots,b}$ we have*

(b, k)	$(7, 7)$	$(8, 8)$	$(9, 8)$	$(10, 9)$	$(11, 10)$
M	≈ 0.0861594	≈ 0.0388599	≈ 0.0758830	≈ 0.0363565	≈ 0.0170049

(b, k)	$(12, 10)$	$(13, 11)$	$(14, 12)$	$(15, 13)$	
M	≈ 0.0309448	≈ 0.0150674	≈ 0.0071917	≈ 0.0033733	

Using the partition $\{\widehat{P}_b^i\}_{i=0,\ldots,b}$ we have

(b,k)	$(5,5)$	$(6,5)$	$(6,6)$
M	≈ 0.3873676	≈ 0.5567010	$\frac{5}{27} \approx 0.185185$

Remark 5. *For the underlined (b,k)-cases reported in Table 1, it is interesting to note that the maximum in (8) is only achieved for uniform distributions. This means that, for these particular cases, any new upper bounds that can be found on the quadratic form in equation (15) cannot further improve those bounds. Note that, for $(b,4)$-cases when $b \geq 4$, the maximum of the quadratic form in (15) is only achieved for uniform distributions if we suppose that the frequency of each symbol in all the coordinates of the code is less than or equal to $1/2$. For the special case $b = k = 6$, the values we obtained for the M_i constants are such that the resulting η_0 in the statement of Lemma 4 is equal to 1. That is, the constant M is actually M_1, which means that in our bound the worst-case scenario is given by the balanced subcodes. The resulting value $M_1 = 5/27$ is actually the value attained by $\Psi_j(p,q)$ for uniform p and q. Roughly speaking, this should be interpreted as saying that our procedure is unable to give for $R_{(6,6)}$ a bound smaller than $5/59$ because such a rate might in principle be achieved if all subcodes have a uniform distribution on each coordinate. However, for such globally balanced codes, one can use a different argument based on the minimum distance of the code to get even stronger upper bounds on $R_{(b,k)}$. In the next section, we combine the two procedures to deduce a rigorous proof that indeed the bounds shown in Table 1 are not sharp for different values of b and k.*

5 A qualitative analysis on $R_{(b,k)}$

In this section we show that, at least for the underlined (b,k)-cases in Table 1 and for case $(b,k) = (6,6)$, the bound in equation (7), for $j = k-2$ with $\mathbf{M}_{k-2} = \Psi_{k-2}(1/b,\ldots,1/b; 1/b,\ldots,1/b)$, is not sharp. We also show that, the bound given in equation (4) is not sharp for every $(b,4)$-cases with $b \geq 5$ and $j = 2$. In this discussion, we use continuity arguments whose quantitative analysis would require long and complicated computations. For this reason, we do not provide explicit numerical improvements on $R_{(b,k)}$ and only show that the bounds on $R_{(b,k)}$ can be improved.

To prove our statement, we invoke an upper bound from [1] on the minimum hamming distance $d_H(C)$ of a b-ary code C with a given rate R. It suffices here to mention that, set $\delta := d_H(C)/n$, this bound is of the form $\delta \leq F(R)$ for a suitable decreasing continuous function F. Due to the monotonicity of F there exists a maximum value of R for which the inequality $R \leq \frac{(b-2)!}{(b-k+1)!b^{k-3}}F(R)$ is satisfied.

Using Mathematica on the specific bound in [1], one finds that

$$R \leq \frac{(b-2)^{\underline{k-3}}}{b^{k-3}}F(R) \implies R < U_{(b,k)}$$

where $(b-2)^{\underline{k-3}} = (b-2)\cdots(b-k+2)$ and $U_{(b,k)}$ takes the values shown in Table 4 for some (b,k) pairs. Note that most of these pairs are actually those underlined in Table 1.

Because of the continuity of F, this implies that there exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that

$$R \leq \left(\frac{(b-2)^{\underline{k-3}}}{b^{k-3}} + \epsilon_1\right)F(R) + \epsilon_2 \implies R < U_{(b,k)} + 10^{-5}$$

We note that, if $p_1 = p_2 = \cdots = p_b = 1/b$, given $i \neq h \in [1,b]$ and chosen at random $x_1,\ldots,x_{k-4},z \in [1,b]$ according to the distribution p, the probability that $i,h,x_1,\ldots,x_{k-4}$, z are all different is $(b-2)^{\underline{k-3}}/b^{k-3}$. Therefore, by continuity, there exists $\epsilon_3 > 0$ such that given $i \neq h \in [1,b]$ and chosen at random $x_1,\ldots,x_{k-4},z \in [1,b]$ according to the distribution p' where $p'_1, p'_2, \ldots, p'_b \in [1/b - \epsilon_3, 1/b + \epsilon_3]$, the probability that $i,h,x_1,\ldots,x_{k-4},z$ are all

21

Table 4: $U_{(b,k)}$ values

$U_{(6,6)} = 0.08469$	$U_{(7,6)} = 0.13440$	$U_{(8,6)} = 0.18125$	$U_{(9,6)} = 0.22405$
$U_{(10,6)} = 0.26268$	$U_{(11,6)} = 0.29744$	$U_{(12,6)} = 0.32874$	$U_{(13,6)} = 0.35699$
$U_{(14,6)} = 0.38258$	$U_{(8,7)} = 0.07200$	$U_{(9,7)} = 0.10510$	$U_{(10,7)} = 0.13822$
$U_{(11,7)} = 0.17025$	$U_{(12,7)} = 0.20068$	$U_{(13,7)} = 0.22930$	$U_{(14,7)} = 0.25609$
$U_{(10,8)} = 0.05749$	$U_{(11,8)} = 0.08043$	$U_{(12,8)} = 0.10419$	$U_{(13,8)} = 0.12808$
$U_{(14,8)} = 0.15163$	$U_{(11,9)} = 0.03006$	$U_{(12,9)} = 0.04465$	$U_{(13,9)} = 0.06081$
$U_{(14,9)} = 0.07799$	$U_{(13,10)} = 0.02386$	$U_{(14,10)} = 0.03412$	$U_{(14,11)} = 0.01236$

different is less than $(b-2)\frac{k-3}{}/b^{k-3} + \epsilon_1$. Now we divide the coordinates $i \in [1, n]$ into two sets according to whether the distribution f_i has all its values in $[1/b - \epsilon_3, 1/b + \epsilon_3]$ or not. More precisely, we define:

$$U_{\epsilon_3} := \{i \in [1, n] : f_{i,h} \in [1/b - \epsilon_3, 1/b + \epsilon_3], \ \forall h \in [1, b]\}.$$

We can assume, up to reordering the coordinates, that $U_{\epsilon_3} = [1, t]$ for some value t. Here we divide the discussion into two cases, according to the size of t, and we show that in both cases a better bound on $R_{(b,k)}$ can be obtained.

A) Let us assume that $t < n(1 - \epsilon_2)$

As a consequence of Hansel's Lemma, we have the following

$$\log(|C|) \leq (1 + o(1)) \frac{1}{2} \sum_{i \in [\ell+1, n]} \sum_{\omega, \mu \in \Omega} \lambda_\omega \lambda_\mu \Psi_{k-2}(f_{i|\omega}, f_{i|\mu})$$

$$\leq (1 + o(1)) \frac{1}{2} \left[\sum_{i \in [\ell+1, t]} M + \sum_{i \in [t+1, n]} \sum_{\omega, \mu \in \Omega} \lambda_\omega \lambda_\mu \Psi_{k-2}(f_{i|\omega}, f_{i|\mu}) \right]$$

where M is the upperbound of equation (22) given in Theorem 2. Due to the following lemma, we are able to provide a better upper bound to the second term of the sum.

Lemma 11. *Assume that $f_{i,h} \notin [1/b - \epsilon_3, 1/b + \epsilon_3]$ for some $h \in [1, b]$. Then there exists $M' < M$ such that:*

$$\sum_{\omega, \mu \in \Omega} \lambda_\omega \lambda_\mu \Psi_{k-2}(f_{i|\omega}, f_{i|\mu}) \leq M'.$$

Proof. Consider first for simplicity the case when $f_{i,h} < 1/b - \epsilon_3$. Let $\Omega' \subseteq \Omega$ be the subset of the ω for which $f_{i,h|\omega} \geq 1/b - \epsilon_3/2$. Then, since

$$f_{i,h} = \sum_{\omega \in \Omega} \lambda_\omega f_{i,h|\omega} \geq \sum_{\omega \in \Omega'} \lambda_\omega f_{i,h|\omega} \geq (1/b - \epsilon_3/2) \sum_{\omega \in \Omega'} \lambda_\omega,$$

we deduce that

$$\sum_{\omega \in \Omega'} \lambda_\omega \leq \frac{1/b - \epsilon_3}{1/b - \epsilon_3/2}.$$

From Remarks 1, 4 and 5, we know that the maximum of the quadratic form in (15) is only achieved for uniform distributions. This means that $M = \Psi_{k-2}(1/b, \ldots, 1/b; 1/b, \ldots, 1/b)$

for the (b, k)-cases under consideration. Therefore there is some constant $M_{\epsilon_3} < M$ such that if either $f_{i,h|\omega}$ or $f_{i,h|\mu}$ are not Ω', then $\Psi_{k-2}(f_{i|\omega}, f_{i|\mu}) \leq M_{\epsilon_3}$. This implies

$$\sum_{\omega,\mu\in\Omega} \lambda_\omega \lambda_\mu \Psi_{k-2}(f_{i|\omega}, f_{i|\mu}) \leq \left(\frac{1/b - \epsilon_3}{1/b - \epsilon_3/2}\right)^2 M + \left(1 - \left(\frac{1/b - \epsilon_3}{1/b - \epsilon_3/2}\right)^2\right) M_{\epsilon_3}$$

and hence the statement of the lemma for the case $f_{i,h} < 1/b - \epsilon_3$. A similar proof holds for $f_{i,h} > 1/b + \epsilon_3$. $\qquad\square$

In the case $\ell \geq t$, we immediately obtain that $\log(|C|) \leq (n - \ell)\frac{M'}{2}$ which leads to the upperbound $R < \frac{M'}{2+M'}$ that is better than the one shown in Table 1. So we can assume $\ell < t$ and therefore:

$$\log(|C|) \leq (t - \ell)\frac{M}{2} + (n - t)\frac{M'}{2} \leq (n - n\epsilon_2 - \ell)\frac{M}{2} + n\epsilon_2\frac{M'}{2}.$$

Since $\ell = \left\lfloor \frac{nR - 2\log n}{\log(2+\bar{\epsilon})} \right\rfloor = \lfloor nR - 2\log n \rfloor (1 + o(1))$, dividing by n we get:

$$R \leq \frac{1}{2}\left[\frac{M(1 - \epsilon_2 - R + 2\frac{\log n}{n})}{1 - R + 2\frac{\log n}{n}} + \frac{M'\epsilon_2}{1 - R + 2\frac{\log n}{n}}\right](1 - R + 2\frac{\log n}{n})(1 + o(1)).$$

Set $M'' = \frac{M(1 - \epsilon_2 - R)}{1 - R} + \frac{M'\epsilon_2}{1 - R}$ we have that $M'' < M$ and, taking $n \to \infty$, we obtain:

$$R \leq \frac{M''}{2}(1 - R)(1 + o(1)).$$

It means that $R < \frac{M''}{2+M''}$ and since $M'' < M$, it follows that the bound is not sharp under the assumption of the case A.

B) Let us assume that $t \geq n(1 - \epsilon_2)$

Let us fix two words $u, u' \in C$ at minimum hamming distance and let us choose at random x, y. Because of Hansel Lemma we have that:

$$\log(|C|) \leq \sum_{i=1}^{n} \mathbb{E}[\tau(G_i^{u,u',x_1,\ldots,x_{k-4}})].$$

We recall that $0 \leq \tau(G_i^{u,u',x_1,\ldots,x_{k-4}}) \leq 1$ and, if $u_i \neq u'_i$, $\tau(G_i^{u,u',x_1,\ldots,x_{k-4}})$ is the probability that given $z \notin \{u, u', x_1, \ldots, x_{k-4}\}$ we have that $u_i, u'_i, x_{1i}, x_{(k-4)i}, z_i$ are all different. Since we have chosen at random also $x_1, \ldots, x_{k-4}$, $\mathbb{E}[\tau(G_i^{u,u',x_1,\ldots,x_{k-4}})]$ coincides with the probability that given $x_1, \ldots, x_{k-4}, z \notin \{u, u'\}$ we have that $u_i, u'_i, x_{1i}, x_{(k-4)i}, z_i$ are all different. Therefore $\mathbb{E}[\tau(G_i^{u,u',x_1,\ldots,x_{k-4}})] \leq (b-2)^{\underline{k-3}}/b^{k-3} + \epsilon_1$ for any $i \in [1, t]$ when $u_i \neq u'_i$, otherwise if $u_i = u'_i$ then the expected value is 0. This means that

$$\log(|C|) \leq \left(\frac{(b-2)^{\underline{k-3}}}{b^{k-3}} + \epsilon_1\right)\sum_{i=1}^{t} \mathbb{1}_{u_i \neq u'_i} + \sum_{i=t+1}^{n} 1$$

$$\leq \left(\frac{(b-2)^{\underline{k-3}}}{b^{k-3}} + \epsilon_1\right)\sum_{i=1}^{n} \mathbb{1}_{u_i \neq u'_i} + \sum_{i=n(1-\epsilon_2)}^{n} 1$$

and hence

$$\log(|C|) \leq \left(\frac{(b-2)^{\underline{k-3}}}{b^{k-3}} + \epsilon_1\right) d_H(u, u') + n\epsilon_2.$$

Dividing by n and remembering that u, u' are at minimal hamming distance, we obtain that:

$$R \leq \left(\frac{(b-2)^{k-3}}{b^{k-3}} + \epsilon_1 \right) \delta + \epsilon_2 \leq \left(\frac{(b-2)^{k-3}}{b^{k-3}} + \epsilon_1 \right) F(R) + \epsilon_2.$$

But, because of the definition of ϵ_1 and ϵ_2, this implies that $R < U_{(b,k)} + 10^{-5}$. It can be easily checked that the bound in Theorem 3 is strictly greater than $U_{(b,k)} + 10^{-5}$ for every (b,k)-cases under consideration and therefore:

Theorem 3.

$$R_{(b,k)} < \left(\frac{1}{\log \frac{b}{k-3}} + \frac{b^{k-1}}{b^{\underline{k-1}} \log(b-k+2)} \right)^{-1}$$

for the (b,k)-cases shown in Table 4.

For cases $(b,4)$ when $b \geq 4$ we know thanks to [6] that the maximum of (14), under the constraint that $f_{i,a} \leq \frac{1}{2}$ for every $i = 1, \ldots, n$ and every $a = 1, \ldots, b$, is only achieved for uniform distributions. Therefore we can use the Plotkin bound instead of the Aaltonen bound in order to prove that bound (4) is not sharp when $k = 4$ and $b \geq 5$.

Let C be a $(b,4)$-hash code with rate R and suppose that the frequency of the symbols in all the coordinates of C is uniform. Then by Hansel we get

$$R \leq \frac{b-2}{b} \cdot \delta, \tag{28}$$

where $\delta = d_H(C)/n$. The Plotkin bound for q-ary codes with $\delta \leq (b-1)/b$ is the following

$$R \leq \log b \left(1 - \delta \cdot \frac{b}{b-1} \right). \tag{29}$$

Since equation (28) is increasing in δ while (29) is decreasing then we can combine the two bounds to get

$$R \leq \frac{b(b-1) \log b}{(b-1)(b-2) + b^2 \log(b)}. \tag{30}$$

It is easy to see that the bound given in (4) for $k = 4$ is always strictly greater than (30) for every $b > 4$. Then, by a continuity argument (as done previously) one can show that bound (4) for $k = 4$ is not sharp for every $b \geq 5$. Therefore we have the following theorem.

Theorem 4. *For every integer $b > 4$*

$$R_{(b,4)} < \left(\frac{1}{\log b} + \frac{b^2}{(b^2 - 3b + 2) \log(b-2)} \right)^{-1}.$$

References

[1] M. Aaltonen. *A new upper bound on nonbinary block codes, Discrete Math.* 83 (1990), 139–160.

[2] E. Arikan, An upper bound on the zero-error list-coding capacity, *IEEE Transactions on Information Theory* 40 (1994), 1237–1240.

[3] E. Arikan, An improved graph-entropy bound for perfect hashing, *IEEE International Symposium on Information Theory* (1994).

[4] S. Bhandari and J. Radhakrishnan, Bounds on the Zero-Error List-Decoding Capacity of the q/(q-1) Channel, *IEEE Transactions on Information Theory* 289 (2022), 238–247.

[5] S. Costa, M. Dalai, New bounds for perfect k-hashing, *Discrete Applied Mathematics* 289 (2021), 374–382.

[6] M. Dalai, V. Guruswami, and J. Radhakrishnan, An improved bound on the zero-error listdecoding capacity of the 4/3 channel, *IEEE International Symposium on Information Theory* (2017).

[7] M. Dalai, V. Guruswami, and J. Radhakrishnan, An improved bound on the zero-error listdecoding capacity of the 4/3 channel, in *IEEE Transactions on Information Theory*, 66 (2020), 749–756.

[8] P. Elias, Zero error capacity under list decoding, *IEEE Transactions on Information Theory* 34 (1988), 1070–1074.

[9] Michael L. Fredman and János Komlós, On the Size of Separating Systems and Families of Perfect Hash Functions, *SIAM Journal on Algebraic Discrete Methods* 5 (1984), 61–68.

[10] V. Guruswami, A. Riazanov, Beating Fredman-Komlós for perfect k-hashing, *Journal of Combinatorial Theory, Series A* 188 (2022).

[11] 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.

[12] J. Körner and K. Marton, New Bounds for Perfect Hashing via Information Theory, *European Journal of Combinatorics* 9 (1988), 523–530.

[13] J. Körner, Fredman–Komlós bounds and information theory, *SIAM Journal on Algebraic Discrete Methods* 7 (1986), 560–570.

[14] A. Nilli, "Perfect hashing and probability," *Combinatorics, Probability and Computing*, 3 (1994), 407–409.

[15] J. Radhakrishnan, Entropy and Counting, available at: http://www.tcs.tifr.res.in/~jaikumar/Papers/EntropyAndCounting.pdf.

[16] C. Xing and C. Yuan, Beating the probabilistic lower bound on perfect hashing, *arXiv preprint arXiv:1908.08792* (2019).

Appendix

Here we provide the proofs of Lemmas 8, 9, and 10 stated in Section 4.

Proof of Lemma 8. Let $0 \le \delta \le p_2$. We first prove that

$$\Psi_j(p_1, p_2, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b)$$
$$\le \Psi_j(p_1 + \delta, p_2 - \delta, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b). \quad (31)$$

Using the definition of Ψ_j in eq. (5), inequality (31) can be restated by only considering the terms in the summation which differ in the two sides, that is, those corresponding to permutations σ such that $1 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j + 1) = 2$ and $2 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j + 1) = 1$. This gives

$$(p_1 q_2 + p_2 q_1) \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$
$$\le ((p_1 + \delta)q_2 + (p_2 - \delta)q_1) \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j+1)} \cdot$$

Rearranging the terms we have

$$\delta(q_2 - q_1) \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j+1)} \ge 0 \,.$$

Therefore, inequality (31) is thus satisfied since $q_1 \le q_2$ and $\delta \ge 0$. Moreover, given $h > i$ and δ such that $0 \le \delta \le p_h$, with the same argument we have

$$\Psi_j(p_1, \ldots, p_i, \ldots, p_h, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b)$$
$$\le \Psi_j(p_1, \ldots, p_i + \delta, \ldots, p_h - \delta, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b). \quad (32)$$

Using multiple times inequality (32) we get the following chain of inequalities

$$\Psi_j(p_1, p_2, \ldots, p_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b)$$
$$\le \Psi_j(p_1, \alpha, p'_3, \ldots, p'_{j-1}, 0, \ldots, 0; q_1, q_2, \ldots, q_b)$$
$$\le \Psi_j(1 - \alpha, \alpha, 0, \ldots, 0; q_1, q_2, \ldots, q_b) \,,$$

where $p'_3 + \ldots + p'_{j-1} = 1 - p_1 - \alpha$ and $p'_i \in [0, 1 - \alpha]$ for $i = 3, \ldots, j - 1$. $\qquad \square$

Proof of Lemma 9. Using the definition of Ψ_j in eq. (5), inequality (26) can be restated by only considering the terms in the summation which differ in the two sides, that is, those corresponding to permutations σ such that $1 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j + 1) = 2$ and $2 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j + 1) = 1$ and $\{1, 2\} \subseteq \{\sigma(1), \ldots, \sigma(j)\}$. Hence inequality (26) becomes:

$$(p_1 q_2 + p_2 q_1) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$
$$(j - 1)q_2 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j)} p_{\sigma(j+1)}$$
$$\le p_1(q_1 + q_2) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} \cdot$$

That is

$$p_2 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)q_2 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j)} p_{\sigma(j+1)}$$

$$\leq p_1 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)}.$$

We have

$$p_2 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)q_2 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j)} p_{\sigma(j+1)}$$

$$\overset{(i)}{\leq} q_1 \epsilon \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)q_1 \epsilon \sum_{\sigma \in Sym(3,\ldots,b)} q_2 q_{\sigma(3)} \cdots q_{\sigma(j)}$$

$$\overset{(ii)}{\leq} q_1 \epsilon \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + j q_1 \epsilon \sum_{\sigma \in Sym(3,\ldots,b)} q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$

$$\overset{(iii)}{\leq} (1-\epsilon)q_1 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$

$$\overset{(iiii)}{\leq} p_1 q_1 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)}$$

Inequality (i) holds because $p_2, p_3, \ldots, p_b \leq \epsilon$, inequality (ii) because $q_2 \leq q_3 \leq \ldots \leq q_b$, inequality (iii) due to the assumption $\epsilon \leq \frac{1}{j+1}$ and inequality $(iiii)$ since $p_1 \geq 1 - \epsilon$. $\qquad \square$

Proof of Lemma 10. Using the definition of Ψ_j in eq. (5), inequality (27) can be restated by only considering the terms in the summation which differ in the two sides, that is, those corresponding to permutations σ such that $1 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j+1) = 2$ and $2 \in \{\sigma(1), \ldots, \sigma(j)\}$, $\sigma(j+1) = 1$ and $\{1,2\} \subseteq \{\sigma(1), \ldots, \sigma(j)\}$. Therefore we have that

$$((1-\epsilon+\delta)q_2 + q_1 p_2) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)(1-\epsilon+\delta)p_2 \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j)} q_{\sigma(j+1)}$$

$$< ((1-\epsilon)q_2 + q_1(p_2+\delta)) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)(1-\epsilon)(p_2+\delta) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j)} q_{\sigma(j+1)}.$$

That is

$$\delta(q_1 - q_2) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j+1)} + q_{\sigma(3)} \cdots q_{\sigma(j+1)} +$$

$$(j-1)\delta(1-\epsilon-p_2) \sum_{\sigma \in Sym(3,\ldots,b)} p_{\sigma(3)} \cdots p_{\sigma(j)} q_{\sigma(j+1)} > 0.$$

Which is satisfied because $q_2 < q_1$, $p_2 < 1 - \epsilon$ and $\delta > 0$. $\qquad \square$