Beating Fredman-Komlós for Perfect k-Hashing

Venkatesan Guruswami
Computer Science Department, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, USA, 15213
venkatg@cs.cmu.edu

Andrii Riazanov
Computer Science Department, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, USA, 15213
riazanov@cs.cmu.edu

Abstract

We say a subset $C \subseteq \{1, 2, \dots, k\}^n$ is a k-hash code (also called k-separated) if for every subset of k codewords from C, there exists a coordinate where all these codewords have distinct values. Understanding the largest possible rate (in bits), defined as $(\log_2 |C|)/n$, of a k-hash code is a classical problem. It arises in two equivalent contexts: (i) the smallest size possible for a perfect hash family that maps a universe of N elements into $\{1, 2, \dots, k\}$, and (ii) the zero-error capacity for decoding with lists of size less than k for a certain combinatorial channel.

A general upper bound of $k!/k^{k-1}$ on the rate of a k-hash code (in the limit of large n) was obtained by Fredman and Komlós in 1984 for any $k \geq 4$. While better bounds have been obtained for $k = 4$, their original bound has remained the best known for each $k \geq 5$. In this work, we present a method to obtain the first improvement to the Fredman-Komlós bound for every $k \geq 5$, and we apply this method to give explicit numerical bounds for $k = 5, 6$.

2012 ACM Subject Classification Theory of computation → Error-correcting codes; Mathematics of computing → Coding theory

Keywords and phrases Coding theory, perfect hashing, hash family, graph entropy, zero-error information theory

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.92

Category Track A: Algorithms, Complexity and Games

Related Version A full version is available at `https://eccc.weizmann.ac.il/report/2018/096/`.

Funding *Venkatesan Guruswami*: Research supported in part by NSF grants CCF-1422045 and CCF-1563742.

Acknowledgements Some of this work was done when the first author was visiting the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore and the Center of Mathematical Sciences and Applications, Harvard University.

1 Introduction

A code of length n over an alphabet of size k is a subset $C \subseteq \{1, 2, \dots, k\}^n$. We say such a code C is a *k-hash code* (also called k-separated in the literature), if for every subset of k distinct codewords $\{c^{(1)}, c^{(2)}, \dots, c^{(k)}\}$ from C, there exists a coordinate j such that all these codewords differ in this coordinate, i.e. $\{c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(k)}\} = \{1, 2, \dots, k\}$. The rate (in bits) of the code is defined as $R = \frac{\log_2 |C|}{n}$. Then for each fixed integer k, let R_k be the limit superior ($\limsup$), as $n \to \infty$, of the rate of the largest k-hash code of length n.

The study of the quantity R_k is a fundamental problem in combinatorics, information theory, and computer science. As the name suggests, k-hash codes have strong connections to the hashing problem. A family of functions mapping a universe of size N to the set

© Venkatesan Guruswami and Andrii Riazanov;
licensed under Creative Commons License CC-BY
46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 92; pp. 92:1–92:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$\{1, 2, \ldots, k\}$ is called a perfect k-hash family if any k elements of the universe are mapped in one-to-one fashion by at least one hash function from this family. If C is a k-hash code, then a perfect k-hash family for universe C with n functions is just the family of coordinate projections. Therefore, R_k gives the growth rate of the size of universes for which perfect k-hash families of a given size exist. Equivalently, an upper bound on R_k is equivalent to a lower bound on the size of a perfect k-hash family as a function of the universe size.

An equivalent information-theoretic context in which k-hash codes arise concerns zero-error list decoding on certain channels. A channel can be thought of as a bipartite graph (V, W, E), where V is the set of channel inputs, W is the set of channel outputs, and $(v, w) \in E$ if on input v the channel can output w. The $k/(k-1)$ channel then is the channel with $V = W = \{1, 2, \ldots, k\}$, and $(v, w) \in E$ iff $v \neq w$. In this context, R_k is the largest asymptotic rate at which one can communicate using n repeated uses of the channel (as n grows), if we want to ensure that the receiver can identify a subset of at most $k-1$ sequences that is guaranteed to contain the transmitted sequence. See [4, 3] for more details.

Studying the rates of the codes and hashing family sizes in the above settings is a longstanding problem. A probabilistic argument shows the existence of k-hash codes with rate at least $\frac{1}{k-1} \log \frac{1}{1-k!/k^k} - o(1)$ [5, 10], and better bounds are known for some small values of k. Our focus here is on *upper bounds* on R_k, that is limitations on the size of k-hash codes. Here the best-known general upper bound on the rate R_k dates all the way back to the 1984 paper of Fredman and Komlós [5]:

$$R_k \leq \frac{k!}{k^{k-1}} =: \alpha_k \ . \tag{1}$$

For large k the multiplicative discrepancy between the probabilistic lower bound on R_k and the above Fredman-Komlós upper bound (1) grows approximately as k^2, so the current bounds on the rate require tightening to obtain better estimations of R_k. There is another trivial upper bound, $R_k \leq \log_2 \left(\frac{k}{k-1}\right)$, that follows from a simple double-counting or first moment method. The above bound (1) is much better than this bound for $k \geq 4$. For $k = 3$ (which is called the trifference problem by Körner), however, $R_3 \leq \log_2(3/2) \approx 0.585$ remains the best upper bound, and improving it (or showing it can be achieved!) is a major combinatorial challenge. For the case $k = 4$, the bound (1) which states $R_4 \leq 0.375$ has been improved, first by Arikan to 0.3512 [1], and recently by Dalai, Guruswami, and Radhakrishnan [3] to $6/19 \leq 0.3158$.

However, the above quantity α_k remained the best known upper bound on R_k for each $k > 4$. Our main result gives the first improvement to the Fredman-Komlós bound (1) for $k \geq 5$, proving that R_k is strictly smaller than α_k for every k.

▶ **Theorem 1.** *For all $k \geq 4$ there exists β_k such that $R_k \leq \beta_k < \alpha_k$. For $k = 5, 6$, we have the explicit upper bounds $R_5 < \beta_5 = 0.190826 < 0.192 = \alpha_5$, and $R_6 < \beta_6 = 0.0922789 < 0.0\overline{925} = \frac{5}{54} = \alpha_6$.*

Our approach provides a method to compute the explicit bound β_k for any $k \geq 5$ by finding a root of a degree-$O(k)$ polynomial, which lies within a specific interval. Moreover, we present a technical conjecture on the optimum values of certain polynomial optimization problems over the simplex, assuming which even stronger upper bounds on R_k can be obtained. Using the tools of numerical optimization in *Mathematica*, we verify this conjecture for the cases $k = 5, 6$, which gives us better values of β_5 and β_6, claimed in the Theorem above.

Our approach is also applicable to the (b, k)-hashing problem for $b \geq k$, where one considers codes $C \subseteq \{1, 2, \ldots, b\}^n$ with the property that for any k distinct codewords $\{c^{(1)}, c^{(2)}, \ldots, c^{(k)}\}$ from C there exists a coordinate j such that all these codewords differ in this coordinate. Using exactly the same arguments, we obtain an improvement on the Körner-Marton upper bound [9] on the rate of such codes. When $b = k$, this latter bound is identical to the Fredman-Komlós bound, but can be better than the corresponding bound in [5] when $b > k$. For some small pairs of values (b, k) with $b > k$, the Körner-Marton bound was further improved by Arikan [1]. In those cases, the bounds we get are probably weaker than Arikan's. For this reason and for the sake of simplicity, in this paper we analyze only the case $b = k$, which corresponds to k-hashing, but all our proofs generalize in a straightforward way for (b, k)-hashing as well.

2 Background and approach

The previous general upper bounds on the rates of k-hash codes by Fredman and Komlós [5], Körner and Marton [9], and Arikan [1] are all based on information-theoretic inequalities for graph covering, related to the Hansel lemma [6]. Körner [10] cast the Fredman-Komlós proof in the language of graph entropy, which he had introduced in [8] (see [12] for a nice survey on graph entropy). Körner and Marton [9] generalized this approach to the hypergraph case, which led to improvements to the Fredman-Komlós bound for the (b, k)-hashing problem in certain cases when $b > k$, but not for R_k. In this paper we use the following version of the Hansel lemma, which is also proved in [11] via a simple probabilistic argument:

▶ **Lemma 2** (Hansel). *Let K_m be a complete graph on m vertices. Let also $G_1, G_2, \dots, G_t$ be bipartite graphs, such that $E(K_m) = \bigcup_{i=1}^{t} E(G_i)$. Denote by $\tau(G_i)$ the fraction of non-isolated vertices in G_i. Then the following holds:*

$$\log m \le \sum_{i=1}^{t} \tau(G_i). \tag{2}$$

To relate this lemma to the context of the paper, consider a k-hash code $C \subseteq [k]^n$. Take a subset of this code $\{x_1, x_2, \dots, x_{k-2}\} \subseteq C$, and define bipartite graphs $G_i^{x_1,\dots,x_{k-2}}$, for $i \in [n]$, as follows:

$$V(G_i^{x_1,\dots,x_{k-2}}) = C \setminus \{x_1, x_2, \dots, x_{k-2}\},$$
$$E(G_i^{x_1,\dots,x_{k-2}}) = \Big\{\{y_1, y_2\} \;:\; (y_1)_i, (y_2)_i, (x_1)_i, (x_2)_i, \dots, (x_{k-2})_i \text{ are distinct}\Big\}.$$

Note that since C is a k-hash code, for any pair $\{y_1, y_2\} \subseteq C \setminus \{x_1, x_2, \dots, x_{k-2}\}$, there exists some coordinate i, such that all the k codewords $y_1, y_2, x_1, x_2, \dots, x_{k-2}$ differ in the i^{th} coordinate. In other words, $\{y_1, y_2\} \in E(G_i^{x_1,\dots,x_{k-2}})$ for this i. Therefore, $E\left(K_{|C|-(k-2)}\right) = \bigcup_{i=1}^{n} E(G_i^{x_1,\dots,x_{k-2}})$. Then Hansel lemma 2 applies directly, and denoting $\tau_i(x_1, x_2, \dots, x_{k-2}) = \tau\left(G_i^{x_1,\dots,x_{k-2}}\right)$, we obtain

$$\log\left(|C| - k + 2\right) \le \sum_{i=1}^{n} \tau_i(x_1, x_2, \dots, x_{k-2}). \tag{3}$$

Taking the expectation over the choice of $x_1, x_2, \dots, x_{k-2}$, we get

$$\log\left(|C| - k + 2\right) \le \sum_{i=1}^{n} \mathbb{E}[\tau_i(x_1, x_2, \dots, x_{k-2})]. \tag{4}$$

By bounding the RHS of the above inequality one might obtain an upper bound on $\log |C|$, and thus on the rate of this code. Different strategies to pick the codewords $\{x_1, x_2, \dots, x_{k-2}\}$ from C lead to different approaches to bound the RHS of (4). Here we briefly present the ideas underlying the previous works and then outline our approach.

In the original bound by Fredman and Komlós [5] the codewords $x_1, x_2, \dots, x_{k-2}$ are picked independently at random from the code C. Then one can use symmetry arguments (or Muirhead's inequality) to bound the RHS of (4), which leads to the inequality

$$R_k \leq \frac{k!}{k^{k-1}}. \tag{5}$$

Due to the symmetry arguments involved, this bound is actually tight only in the case when the frequencies of the symbols of the code C in each coordinate are exactly uniform.

Arikan [1, 2] used rate versus distance results (the Plotkin bound) from coding theory to ensure that it is possible to pick $x_1, x_2, \dots, x_{k-2}$ which agree on many coordinates. Note that this already guarantees that many terms in the RHS of (3) equal 0. Together with an argument which allows to modify the code so that it doesn't have any coordinate where the symbols have an overly skewed (far from uniform) frequency, Arikan was able to improve the bound (5) for $k = 4$. However, no improvement was gained for larger k.

Dalai, Guruswami, and Radhakrishnan [3] combine aspects of the above two approaches for the case $k = 4$. As in Arikan's work, they pick x_1, x_2 to agree on the first several coordinates. However, instead of a fixed such choice, they pick such a pair at random from a rich subcode of C with a common prefix. Considering such subcodes with common prefixes is a standard approach that leads to the Plotkin bound. The technical crux of the argument in [3] is a concavity claim for some quadratic form which says that despite conditioning on a common prefix, which might greatly alter the frequency vector of symbols in any coordinate in the suffix, the Fredman-Komlós bound for completely random x_1, x_2 is still valid on those coordinates. (In some sense, only the average frequency vector over all prefixes matters, not the individual ones.) Actually this holds modulo a technical condition that there are no coordinates with very skewed symbol distribution, which can be ensured by some pre-processing of the code similar to [2]. Thus some terms in (4) are equal to 0 and the other are bounded by 3/8, and balancing these appropriately, a bound of $R_4 \leq \frac{6}{19}$ is obtained in [3].

In this work, we follow the strategy of [3] for general k by picking $x_1, x_2, \dots, x_{k-2}$ randomly so that they all lie in a rich subcode of C. However, rather than taking Plotkin-type subcode with a common prefix, we consider a subcode C which takes at most $(k-3)$ values on each coordinate from some large set T. This again implies that the coordinates from T contribute 0 to the RHS of (4). In this case, however, the analogous concavity claim seems out of reach, as one has to argue about degree-$(k-2)$ polynomials rather than quadratics. We instead take a different approach that works directly with the arbitrary symbol frequencies that may arise upon conditioning within a subcode, avoiding the averaging or concavity step. (This leads to worse bounds, but still allows to beat the Fredman-Komlós bound for $k > 4$.) However, another problem arises in that the constraint on the code to have non-skewed frequencies in each coordinate cannot be dealt with using Arikan's argument for large k. To cope with this issue, we differentiate two separate cases: (i) where C has only a few coordinates with skewed distributions of symbols, and (ii) where there are a lot of such coordinates.

- In the first case, we pick the coordinates T (where $x_1, x_2, \dots, x_{k-2}$ are chosen to collide) to include all these skewed coordinates. Note that this is unlike [1, 3] where any choice of T of prescribed size works. Our choice of T ensures that in the remaining coordinates the frequency vector is not too far from uniform, and we apply the approach of [3] to get an improvement upon the Fredman-Komlós bound.

- In the second case, we use the original random strategy of picking $x_1, x_2, \ldots, x_{k-2}$ as in [5]. The idea here is that the bound (5) is tight only when all the frequencies of symbols are exactly uniform. Then, in the case when there are a lot of far-from-uniform frequencies, it is possible to improve the bound (5).

By picking the correct way to differentiate between skewed and non-skewed distributions, we then obtain an improvement on the Fredman-Komlós bound (5) for every $k \geq 4$ in section 3.3. As mentioned earlier, this is the *first* such improvement for $k \geq 5$. For $k = 5$ and $k = 6$ we also use numerical optimization tools to provide slightly stronger explicit bounds on R_k.

3 Upper bound on the rate of k-hash codes

Let $\Sigma = \{1, 2, \ldots, k\} = [k]$, and let $C \subseteq \Sigma^n$ be a k-hash code with rate $R = \frac{\log |C|}{n}$ (all logarithm are to the base 2). Let $f_i \in \mathbb{R}^k$ be the frequency vector of symbols of the code for each coordinate $i \in [n]$, namely:

$$f_i[a] = \frac{1}{|C|} |\{x \in C : x_i = a\}|.$$

Throughout the analysis, we will be interested in two cases: when for most of the coordinates the distribution of codeword symbols is close to uniform (non-skewed), or when this doesn't hold. To define the term "close to uniform" formally, we consider a *threshold* γ, that satisfies $\frac{1}{2k-3} \leq \gamma \leq \frac{1}{k}$, and say that $f \in \mathbb{R}^k$ is close to uniform when $f[a] \geq \gamma$ for all $a \in [k]$. Denote then $P_\gamma = \{i \in [n] : \min\limits_{a \in \Sigma} f_i[a] \geq \gamma\}$ – the set of all the coordinates for which the distribution of codeword symbols is close to uniform. Denote also $\ell := \left\lfloor \frac{nR - \log n}{\log\left(\frac{k}{k-3}\right)} \right\rfloor$. We then consider two cases:

1. *Unbalanced:* $|P_\gamma| < n - \ell$, so there is a decent fraction of coordinates where the distribution of codeword symbols is skewed. For this case, we apply a random strategy to pick $x_1, x_2, \ldots, x_{k-2}$ in (4).
2. *Almost balanced:* $|P_\gamma| \geq n - \ell$, so for almost all coordinates, the distribution of codeword symbols is close to uniform. Then we follow the approach from [3] to pick $x_1, x_2, \ldots, x_{k-2}$ which collide on many coordinates.

For both of these cases, we will obtain some bounds on the rate of C, which depend on the threshold γ. It then will remain to choose γ in a manner ensuring that both these bounds beat (5). Then, since for any code C exactly one of the cases holds, we can obtain a general upper bound on the rate.

Before we continue with studying the two cases separately, let's look at how we can estimate $\tau_i(x_1, x_2, \ldots, x_{k-2})$. Clearly, the codeword $y \in C$ appears non-isolated in the graph $G_i^{x_1, x_2, \ldots, x_{k-2}}$ only if all the codewords $x_1, x_2, \ldots, x_{k-2}$ and y differ in the i^{th} coordinate. Therefore, the fraction of non-isolated vertices in $G_i^{x_1, x_2, \ldots, x_{k-2}}$ satisfies

$$\begin{aligned} \tau_i(x_1, \ldots, x_{k-2}) \leq \left(\frac{|C|}{|C| - (k-2)}\right) \Big(1 - f_i[(x_1)_i] - f_i[(x_2)_i] - \cdots - f_i[(x_{k-2})_i]\Big) \times \\ \times \mathbf{1}\Big[(x_1)_i, (x_2)_i, \ldots, (x_{k-2})_i \text{ distinct}\Big], \end{aligned} \tag{6}$$

where $\mathbf{1}[E]$ is the indicator variable for an event/condition E.

3.1 Unbalanced case

We will pick $x_1, x_2, \ldots, x_{k-2}$ uniformly at random without replacement from C to obtain an upper bound on the rate of C from (4). Taking the expectations of the both sides in (6) gives

$$\begin{aligned}
&\mathbb{E}[\tau_i(x_1, \ldots, x_{k-2})] \\
&\quad \leq \frac{|C|}{|C|-k+2} \sum_{\substack{a_1,\ldots,a_{k-2} \,\in\, \Sigma \\ \{a_s\} \text{ distinct}}} \left(1 - \sum_{s=1}^{k-2} f_i[a_s]\right) \cdot \mathbb{P}\big[(x_s)_i = a_s,\ s = 1, \ldots, (k-2)\big] \\
&\quad = \frac{|C|}{|C|-k+2} \prod_{j=0}^{k-3} \frac{|C|}{|C|-j} \sum_{\substack{a_1,a_2,\ldots,a_{k-2} \,\in\, \Sigma \\ \{a_s\} \text{ distinct}}} \left(1 - \sum_{s=1}^{k-2} f_i[a_s]\right) \cdot f_i[a_1] f_i[a_2] \ldots f_i[a_{k-2}],
\end{aligned} \tag{7}$$

where the coefficients $\frac{|C|}{|C|-j}$, $j = 0, 1, \ldots, k-3$ appear because we pick elements from C without replacement. Define the following function of two probability vectors $g, f \in \mathbb{R}^k$:

$$\phi_k(g, f) := \sum_{\substack{a_1,a_2,\ldots,a_{k-2} \in \Sigma \\ \{a_s\} \text{ distinct}}} \prod_{s=1}^{k-2} g[a_s] \Big(1 - \sum_{s=1}^{k-2} f[a_s]\Big). \tag{8}$$

Using this notation, we derive from (7):

$$\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})] \leq \phi_k(f_i, f_i)\big(1 + o(1)\big). \tag{9}$$

Since $\sum_{a \in \Sigma} f_i[a] = 1$, it is easy to see that $\phi_k(f_i, f_i)$ is a symmetric expression in $f_i[a]$ for all $a \in \Sigma$. Denote by $S_h^t(g)$ the h-th elementary symmetric sum of the first t coordinates of the vector $g \in \mathbb{R}^k$, i.e. the sum of all products of h distinct elements from $\{g[1], g[2], \ldots, g[t]\}$. Then we can write

$$\phi_k(f_i, f_i) = (k-2)! \cdot \binom{k-1}{k-2} S_{k-1}^k(f_i) = (k-1)! \cdot S_{k-1}^k(f_i)$$

It is not hard to show that $S_h^k(g)$ for g being a probability vector in $\mathbb{R}^k$ is maximized when g is uniform. Indeed, if there are two non-equal coordinates $g[a] \neq g[b]$, then substituting the values in these coordinates by their arithmetic average strictly increases the value of $S_h^k(g)$. Then let us denote by u the uniform distribution on k elements, i.e. $u[a] = 1/k$ for all $a \in [k]$, and so $S_h^k(g) \leq S_h^k(u)$. Then in (9)

$$\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})] \leq (k-1)! \cdot S_{k-1}^k(f_i) \cdot \big(1 + o(1)\big) \leq (k-1)! \cdot S_{k-1}^k(u) \cdot \big(1 + o(1)\big),$$

where we compute $S_{k-1}^k(u) = \binom{k}{k-1} \cdot \left(\frac{1}{k}\right)^{k-1} = \left(\frac{1}{k}\right)^{k-2}$. Therefore, we retrieve

$$\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})] \leq \frac{(k-1)!}{k^{k-2}} \cdot \big(1 + o(1)\big) = \frac{k!}{k^{k-1}} \cdot \big(1 + o(1)\big). \tag{10}$$

Substituting this inequality into (4), notice that we derive exactly the Fredman-Komlós bound (5). Denote then

$$\alpha_k := \frac{k!}{k^{k-1}},$$

the Fredman-Komlós upper bound on the rate R_k.

Now recall that we are considering the unbalanced case, in which there are a lot of coordinates with frequencies of codeword symbols being far from uniform. Take i to be any of such coordinates, and let for convenience $f = f_i$, so $\min_{a\in\Sigma} f[a] < \gamma$. Without loss of generality, say $f[k] < \gamma$. Notice the following trivial property of symmetric sums:

$$\phi_k(f,f) = (k-1)! \cdot S^k_{k-1}(f) = (k-1)!\Big(S^{k-1}_{k-1}(f) + f[k] \cdot S^{k-1}_{k-2}(f)\Big).$$

The above expression is symmetric in the first $(k-1)$ coordinates of f. Let's then fix $f[k]$, and do the same averaging operations with all the remaining coordinates of f, making in the end $f'[1] = f'[2] = \cdots = f'[k-1] = \frac{1-f[k]}{k-1}$. The value of $\phi_k(f,f)$ only increases after such operations, so

$$\phi_k(f,f) \le (k-1)!\Bigg(f'[1]f'[2]\cdots f'[k-1] + f[k] \cdot S^{k-1}_{k-2}(f')\Bigg).$$

Let $y = \frac{1-f[k]}{k-1}$, so $f[k] = 1-(k-1)y$. Since $0 \le f[k] < \gamma$ by the assumption above, it holds $\frac{1-\gamma}{k-1} \le y \le \frac{1}{k-1}$. Recall that we took the threshold $\gamma \le \frac{1}{k}$, thus $y \ge \frac{1-\gamma}{k-1} \ge \frac{1}{k}$. Then

$$\phi_k(f,f) \le (k-1)!\Bigg(y^{k-1} + \big(1-(k-1)y\big)\cdot(k-1)y^{k-2}\Bigg) = (k-1)!y^{k-2}\Bigg((k-1)-(k^2-2k)y\Bigg).$$

Denote $G_k(y) = (k-1)!y^{k-2}\Bigg((k-1)-(k^2-2k)y\Bigg)$, so $\phi_k(f,f) \le G_k(y)$. We have

$$(G_k(y))' = (k-1)!(k-1)(k-2)y^{k-3}\big(1-ky\big), \tag{11}$$

so the derivative of G_k is negative on the interval $\frac{1}{k} \le \frac{1-\gamma}{k-1} < y \le \frac{1}{k-1}$, and it is zero at $y = \frac{1}{k}$. Therefore, we finally obtain for any such f:

$$\phi_k(f,f) \le \max_{y\in\left[\frac{1-\gamma}{k-1},\frac{1}{k-1}\right]} G_k(y) = G_k\left(\frac{1-\gamma}{k-1}\right). \tag{12}$$

Note that $G_k\left(\frac{1-\gamma}{k-1}\right) \le G_k\left(\frac{1}{k}\right) = \alpha_k$ for any $\gamma \le \frac{1}{k}$, and the strict inequality $G_k\left(\frac{1-\gamma}{k-1}\right) < G_k\left(\frac{1}{k}\right) = \alpha_k$ holds when $\gamma < \frac{1}{k}$.

So if $\min_{a\in[k]} f_i[a] < \gamma$ for some coordinate i, we retrieved the bound

$$\mathbb{E}[\tau_i(x_1,x_2,\dots,x_{k-2})] \le G_k\left(\frac{1-\gamma}{k-1}\right)\big(1+o(1)\big). \tag{13}$$

For now we obtained two bounds for the summands in the RHS of (4): (i) the bound (10) holds for all the coordinates, and (ii) the bound (13) holds for the coordinates with codeword symbol frequencies far from uniform. As we noted above, the second bound is strictly stronger than the first bound when we take the threshold $\gamma < \frac{1}{k}$. Also recall that in the unbalanced case which we now consider, there are a lot of coordinates of the second type, so essentially the bound (13) applies many times. Let's now formalize this argument to obtain an improvement on the Fredman-Komlós bound for the unbalanced case.

Denote $\xi_k(\gamma) = G_k\left(\frac{1-\gamma}{k-1}\right)$, then

$$\xi_k(\gamma) = (k-1)!\frac{(1-\gamma)^{k-2}}{(k-1)^{k-2}}\frac{(k-1)^2-(k^2-2k)(1-\gamma)}{k-1} = \frac{(k-2)!(1-\gamma)^{k-2}\big((k^2-2k)\gamma+1\big)}{(k-1)^{k-2}}, \tag{14}$$

and note that $\xi_k(\gamma) \le \alpha_k$ for $\gamma \le \frac{1}{k}$. Recall that we denoted by P_γ the set of coordinates i for which $\min_{a\in\Sigma} f_i[a] \ge \gamma$. For such $i \in P_\gamma$ we directly apply the bound (10). For all the other coordinates $i \in [n] \setminus P_\gamma$ we use the inequality (13). In the unbalanced case $|P_\gamma| < n - \ell$, thus $n - |P_\gamma| > \ell$. Applying all these arguments to (4), we obtain

$$\begin{aligned}\log(|C| - k + 2) &\le \Big(|P_\gamma|\alpha_k + \big(n - |P_\gamma|\big)\xi_k(\gamma)\Big)(1 + o(1))\\ &< \Big(n\alpha_k - \ell\left(\alpha_k - \xi_k(\gamma)\right)\Big)(1 + o(1))\\ &\le \left(n\alpha_k - \frac{nR}{\log\left(\frac{k}{k-3}\right)}\left(\alpha_k - \xi_k(\gamma)\right) + o(n)\right)(1 + o(1)),\end{aligned}$$

where $\ell = \left\lfloor \frac{nR - \log n}{\log\left(\frac{k}{k-3}\right)} \right\rfloor$. Since $|C| = 2^{Rn}$, the above implies for $n \to \infty$:

$$R \le \alpha_k - \frac{R\left(\alpha_k - \xi_k(\gamma)\right)}{\log\left(\frac{k}{k-3}\right)} + o(1),$$

$$\boxed{R_k^{\text{unbal}}(\gamma) \le \frac{\alpha_k}{1 + \frac{\alpha_k - \xi_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}}.} \tag{15}$$

Note that for $\gamma = \frac{1}{k}$ the above bound becomes equal to α_k, since $\xi_k(1/k) = G\left(\frac{1-1/k}{k-1}\right) = G\left(\frac{1}{k}\right) = \alpha_k$. Moreover, the previous analysis (11) of the function $G_k(\cdot)$ implies that the RHS of the above bound is strictly increasing as a function of γ. Thus the bound (15) is strictly better than the Fredman-Komlós bound for the unbalanced case for any threshold $\gamma < \dfrac{1}{k}$.

3.2 Almost balanced case

For this case we extend the approach used in [3] for 4-hashing. Namely, in [3] the authors considered a Plotkin-type subcode of C containing the words with the common prefix, and then picked x_1, x_2 from this subcode. For our purposes, we will consider a rich subcode of codewords which can take a restricted set of symbols on some fixed set of coordinates, and choose $x_1, x_2, \ldots, x_{k-2}$ randomly from the subcode. In the almost balanced case, we are able to ensure that the distributions of codeword symbols in all non-fixed coordinates are close to uniform, which will allow us to use some continuity argument to bound the RHS of (4).

In the almost balanced case we assume $|P_\gamma| \ge n - \ell$, so there are at most ℓ coordinates where the distribution of codeword symbols is skewed. The set of such coordinates is $\overline{P_\gamma} = [n] \setminus P_\gamma$, $|\overline{P_\gamma}| \le \ell$. Then take any subset $T \subset [n]$, such that $\overline{P_\gamma} \subseteq T$ and $|T| = \ell$, and denote $S = [n] \setminus T$.

Our goal is to find a subcode of C of sufficient size, such that any $(k-2)$ codewords $x_1, x_2, \ldots, x_{k-2}$ from this subcode collide in all the coordinates from T. In other words, for any coordinate $t \in T$ there should exist i, j such that $(x_i)_t = (x_j)_t$. This will ensure that the coordinates from T contribute 0 to the RHS of (4), which will allow us to prove a better bound on the rate of the code C. We will now define the subcodes which satisfy this property.

First, denote by $\binom{\Sigma}{p}$ the family of p-element subsets of the alphabet $\Sigma = \{1, 2, \ldots, k\}$. Then define $\Omega := \underbrace{\binom{\Sigma}{k-3} \times \binom{\Sigma}{k-3} \times \cdots \times \binom{\Sigma}{k-3}}_{\ell}$.

Now, for any $\omega \in \Omega$ and any string $s \in \Sigma^\ell$, denote $s \vdash \omega$ if $s_1 \in \omega_1, s_2 \in \omega_2, \ldots, s_\ell \in \omega_\ell$. Then, for any $\omega \in \Omega$, we define:

$$C_\omega := \{x \in C \,:\, x_{\{T\}} \vdash \omega\},$$

where $x_{\{T\}}$ is the projection of the codeword x on the set of coordinates T. Notice that C_ω has the property we discussed above. Indeed, for any pick $x_1, x_2, \ldots, x_{k-2} \in C_\omega$ and any $t \in T$, it holds $(x_1)_t, (x_2)_t, \ldots, (x_{k-2})_t \in \omega_t$, but $|\omega_t| = k-3$, and therefore $(x_1)_t, (x_2)_t, \ldots, (x_{k-2})_t$ are not all distinct.

Denote then $M_\omega = |C_\omega|$. Note that for each $x \in C$ there are exactly $\binom{k-1}{k-4}^\ell$ different elements $\omega \in \Omega$ such that $x_{\{T\}} \vdash \omega$. Therefore

$$\sum_{\omega\in\Omega} M_\omega = |C| \cdot \binom{k-1}{k-4}^\ell.$$

It suffices to prove that there exists at least one $\omega \in \Omega$ such that $M_\omega \geq n$ for our arguments further. For the sake of contradiction, suppose $M_\omega < n$ for all $\omega \in \Omega$. But then

$$2^{nR} = |C| = \sum_{\omega\in\Omega} M_\omega \frac{1}{\binom{k-1}{k-4}^\ell} < \frac{\binom{k}{k-3}^\ell}{\binom{k-1}{k-4}^\ell} \cdot n = \left(\frac{k}{k-3}\right)^\ell n = 2^{\ell\cdot\log\frac{k}{k-3} + \log n} \leq 2^{nR},$$

where $\ell = \left\lfloor \frac{nR - \log n}{\log\left(\frac{k}{k-3}\right)} \right\rfloor$. The above is a contradiction, so there exists such $\omega \in \Omega$ that $M_\omega \geq n$.

We are finally ready to describe the strategy to pick the codewords $x_1, x_2, \ldots, x_{k-2}$ in the almost balanced case. We do the following: first, deterministically choose some $\omega \in \Omega$ such that $M_\omega \geq n$, and then pick $x_1, x_2, \ldots, x_{k-2}$ uniformly at random (without replacement) from C_ω. Since all the codewords collide on the coordinates from the set T, we obtain in (4):

$$\log(|C| - k + 2) \leq \sum_{m\in[n]} \mathbb{E}[\tau_m(x_1, x_2, \ldots, x_{k-2})] = \sum_{m\in S} \mathbb{E}[\tau_m(x_1, x_2, \ldots, x_{k-2})]. \tag{16}$$

Now fix some $m \in S$, and let $f_{m|\omega}$ be the frequency vector of the m^{th} coordinate in the subcode C_ω. Taking expectation over the choice of $x_1, x_2, \ldots, x_{k-2}$ in (6) with respect to the the random strategy described above, we have

$$\begin{aligned}&\mathbb{E}[\tau_m(x_1, x_2, \ldots, x_{k-2})]\\ &= \frac{|C|}{|C| - k + 2} \prod_{j=0}^{k-3} \frac{|C_\omega|}{|C_\omega| - j} \sum_{\substack{a_1,\ldots,a_{k-2} \in \Sigma \\ \{a_s\} \text{ distinct}}} \left(1 - \sum_{s=1}^{k-2} f_m[a_s]\right) \cdot f_{m|\omega}[a_1] f_{m|\omega}[a_2] \ldots f_{m|\omega}[a_{k-2}],\end{aligned}$$

where the coefficients $\frac{|C_\omega|}{|C_\omega| - j}$, $j = 0, 1, \ldots, (k-3)$ appear because we pick $(k-2)$ elements from C_ω without replacement. Since we took ω such that $|C_\omega| \geq n$, it follows that $\frac{|C_\omega|}{|C_\omega|-j} \leq \frac{n}{n-j}$.

Using the function $\phi_k(g, f)$ which was defined in (8), we can rewrite the above as

$$\mathbb{E}[\tau_m(x_1, x_2, \ldots, x_{k-2})] \leq \prod_{j=0}^{k-2} \left(\frac{n}{n-j}\right) \phi_k(f_{m|\omega}, f_m) = \phi_k(f_{m|\omega}, f_m) \cdot \big(1 + o(1)\big). \tag{17}$$

Consider the following definition:

$$\boxed{\theta_k(\gamma) := \max_{g,f}\{\phi_k(g, f) : f, g \in \mathbb{R}^k \text{ are probability vectors, } \min_{a\in\Sigma} f[a] \geq \gamma\}.} \tag{18}$$

Let's first consider the bound we obtain using this definition, and then analyze $\theta_k(\gamma)$.

Since $\min_{a\in\Sigma} f_m[a] \geq \gamma$ by construction of the set S, we have $\phi_k(f_{m|\omega}, f_m) \leq \theta_k(\gamma)$ for any $m \in S$, so substituting it into (17) gives

$$\mathbb{E}[\tau_m(x_1, x_2, \ldots, x_{k-2})] \leq \theta_k(\gamma) \cdot \big(1 + o(1)\big).$$

Therefore, in (16) we derive

$$\begin{aligned}\log(|C| - k + 2) &\leq |S| \cdot \theta_k(\gamma)\big(1 + o(1)\big) = (n - \ell) \cdot \theta_k(\gamma)\big(1 + o(1)\big) \\ &\leq \left(n - \frac{nR}{\log\left(\frac{k}{k-3}\right)} + \frac{\log n}{\log\left(\frac{k}{k-3}\right)} + 1\right) \theta_k(\gamma)\big(1 + o(1)\big).\end{aligned}$$

Recall that $|C| = 2^{nR}$, thus for $n \to \infty$

$$R \leq \left(1 - \frac{R}{\log\left(\frac{k}{k-3}\right)}\right) \theta_k(\gamma) + o(1),$$

$$\boxed{R_k^{\mathrm{bal}}(\gamma) \leq \frac{\theta_k(\gamma)}{1 + \frac{\theta_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}}}. \tag{19}$$

It now remains to understand how $\theta_k(\gamma)$, defined in (18), behaves as a function of γ.

Upper bound for $\theta_k(\gamma)$

First, note that for $\gamma = \frac{1}{k}$ the only probability vector f with $\min_{a\in\Sigma} f[a] \geq \gamma$ is the uniform vector u. Then $\phi_k(g, u)$ is an elementary symmetric sum of all the coordinates of g, and therefore we obtain $\phi_k(g, u) \leq \phi_k(u, u) = \alpha_k$, and so $\theta_k(1/k) = \alpha_k$.

Now take any $\gamma \leq \frac{1}{k}$, and let g, f be probability vectors in $\mathbb{R}^k$ such that $f[a] \geq \gamma$ for $a \in \Sigma$. We will further use "f_a" to refer to the a^{th} coordinate of vector f rather than "$f[a]$".

Let $t = \binom{k}{2} = k(k-1)/2$ and let $P_1, P_2, \ldots, P_t$ be an enumeration of all $(k-2)$-element subsets of $\Sigma = \{1, 2, \ldots, k\}$. Then we have from (8)

$$\phi_k(g, f) = \sum_{\substack{a_1,\ldots,a_{k-2}\in\Sigma \\ \{a_i\}\ \text{distinct}}} \prod_{i=1}^{k-2} g_{a_i}\left(1 - \sum_{i=1}^{k-2} f_{a_i}\right) = (k-2)! \sum_{j=1}^{t} \left[\prod_{a\in P_j} g_a \cdot \left(1 - \sum_{a\in P_j} f_a\right)\right]. \tag{20}$$

Denote $d_j := \prod_{a\in P_j} g_a$, and let $d_{(i)}$ be the i^{th} order statistic of the set $\{d_1, d_2, \ldots, d_t\}$, i.e. $\{d_{(1)}, d_{(2)}, \ldots, d_{(t)}\} = \{d_1, d_2, \ldots, d_t\}$ and $d_{(1)} \geq d_{(2)} \geq \cdots \geq d_{(t)}$. The proof of the following claim can be found in the full version of the paper.

▷ Claim 3.

$$\phi_k(g, f) \leq (k-2)! \left[(1 - k\gamma) \sum_{j=1}^{k-1} d_{(j)} + 2\gamma \sum_{j=1}^{t} d_{(j)}\right]. \tag{21}$$

Now, $\sum_{j=1}^{t} d_{(j)} = \sum_{j=1}^{t} d_j$ is just an elementary symmetric sum of degree $k-2$ for the probability vector g, so this expression is maximized for the uniform vector, and then we get $\sum_{j=1}^{t} d_{(j)} \leq \binom{k}{2}\left(\frac{1}{k}\right)^{k-2}$.

Finally, we use $d_{(j)} \leq \left(\frac{1}{k-2}\right)^{k-2}$, because each $d_{(j)}$ is a product of $(k-2)$ coordinates of some probability vector. Thus we obtain

$$\phi_k(g,f) \leq (k-2)!\left[(1-k\gamma)\frac{(k-1)}{(k-2)^{k-2}} + \gamma\frac{(k-1)}{k^{k-3}}\right] = (k-1)!\left[\frac{(1-k\gamma)}{(k-2)^{k-2}} + \frac{\gamma}{k^{k-3}}\right]. \quad (22)$$

Since this holds for any g and any f such that $\min_{a\in\Sigma} f[a] \geq \gamma$, we obtain an upper bound

$$\theta_k(\gamma) \leq (k-1)!\left[\frac{(1-k\gamma)}{(k-2)^{k-2}} + \frac{\gamma}{k^{k-3}}\right] =: \rho_k(\gamma). \quad (23)$$

Note that $\rho_k(\gamma)$ is linear in γ for a fixed k, and that $\rho_k(1/k) = \alpha_k = \theta_k(1/k)$, so this upper bound is tight for $\gamma = \frac{1}{k}$.

3.2.1 Conjecture on the exact value of $\theta_k(\gamma)$

We now describe the conjecture we make on the exact value of $\theta_k(\gamma)$. Consider the upper bound (21) on $\phi_k(g,f)$. Even for some fixed ordering of the coordinates of g, say (without loss of generality) $g_1 \geq g_2 \geq \cdots \geq g_k \geq 0$, there might be different cases of orderings within the set $\{d_1, d_2, \ldots, d_t\}$. Say there are q_k different ways to take the first $(k-1)$ order statistics within the set $\{d_1, d_2, \ldots, d_t\}$ (with this fixed ordering), then there would be q_k different functionals of d_i's, and thus of g_i's, in the RHS of (21), call them $\Theta_k^{(1)}(g,\gamma), \Theta_k^{(2)}(g,\gamma), \ldots, \Theta_k^{(q_k)}(g,\gamma)$. Since exactly one ordering within $\{d_1, \ldots, d_t\}$ is correct for any particular vector g, we obtain

$$\phi_k(g,f) \leq \max_{i=1,2,\ldots,q_k} \Theta_k^{(i)}(g,\gamma).$$

Then define

$$\theta_k^{(i)}(\gamma) := \max_x \left\{\Theta_k^{(i)}(x,\gamma) : \sum_{j=1}^{k} x_j = 1, x \geq 0\right\}, \qquad \text{for } i = 1,2,\ldots,q_k, \quad (24)$$

and so the quantity $\theta_k(\gamma)$ defined in (18) satisfies

$$\theta_k(\gamma) \leq \max_{i=1,2,\ldots,q_k} \theta_k^{(i)}(\gamma).$$

So to find an upper bound on $\theta_k(\gamma)$ it suffices to find the maximum among $\theta_k^{(i)}(\gamma)$ for $i = 1,2,\ldots,q_k$. Unfortunately, q_k grows exponentially as k increases, so it is not clear how to do this efficiently. We introduce a conjecture below, which suggests that we can determine which of the values $\theta_k^{(i)}(\gamma)$, $i = 1,2,\ldots,q_k$, is the greatest for any k.

Specifically, the conjecture is stated as follows: we assume that the maximum among all the values $\theta_k^{(i)}(\gamma)$, $i = 1,2,\ldots,q_k$, is the greatest for the functional $\Theta_k^{(i)}(x,\gamma)$ corresponding to the case, when the first $(k-1)$ order statistics of the set $\{d_1, d_2, \ldots, d_t\}$ form the set $\left\{\frac{\prod_{i=1}^{k-1} g_i}{g_a}\right\}_{a\in[k-1]}$. In other words, $\{d_{(1)}, d_{(2)}, \ldots, d_{(k-1)}\}$ correspond to the sets P_j that contain all their $(k-2)$ elements from $\{1,2,\ldots,k-1\}$ (recall $d_j = \prod_{a\in P_j} g_j$). So the first $(k-1)$ order statistics are formed as the products of only the first $(k-1)$ coordinates of g, ignoring the coordinate g_k. Our intuition behind the assumption that this functional will have the greatest maximum is based on symmetry arguments.

Recall that we denote by $S_h^t(g)$ the h-th elementary symmetric sum of the first t coordinates of g. Then the above conjecture can be formalized as follows:

▶ **Conjecture 4.**

$$\theta_k(\gamma) = \max_x \left\{ (k-2)! \left[\left(1-(k-2)\gamma\right) S_{k-2}^{k-1}(x) + 2\gamma \cdot x_k \cdot S_{k-3}^{k-1}(x) \right] : \quad \sum_{i=1}^{k} x_i = 1, x \geq 0 \right\}. \quad (25)$$

Indeed, the function $\Theta_k^\gamma(g) = (k-2)! \left[\left(1 - (k-2)\gamma\right) S_{k-2}^{k-1}(g) + 2\gamma \cdot g_k \cdot S_{k-3}^{k-1}(g) \right]$ just corresponds to the functional in the RHS of (21) in the case we discussed above.

Simple computation of the RHS of (25) gives us another formulation of this conjecture:

$$\theta_k(\gamma) = \frac{(k-1)!(k-3)^{k-3}\gamma^{k-2}}{\left((k^2-2k)\gamma - 1\right)^{k-3}}. \quad (26)$$

3.3 Improvement of the Fredman-Komlós bound

In this section we show that it is possible to choose such a threshold γ that both bounds (15) and (19) are stronger than the Fredman-Komlós bound.

Using (23) in the bound (19), we obtain

$$R_k^{\text{bal}}(\gamma) \leq \frac{\theta_k(\gamma)}{1 + \frac{\theta_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}} \leq \frac{\rho_k(\gamma)}{1 + \frac{\rho_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}}. \quad (27)$$

Since for any k-hash code C either the unbalanced or the almost balanced case holds, and we get to choose the threshold γ to differentiate between these cases, the above, combined with (15), gives us the following upper bound for the rate in the general case:

$$R_k \leq \min_{\gamma \in \left(\frac{1}{2k-3}, \frac{1}{k}\right)} \max \left\{ \frac{\rho_k(\gamma)}{1 + \frac{\rho_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}}, \frac{\alpha_k}{1 + \frac{\alpha_k - \xi_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}} \right\}. \quad (28)$$

The optimal threshold γ is such that the bounds (27) and (15) are equal, since the first bound becomes stronger as γ increases, while the second bound becomes weaker. Therefore, the optimal threshold is the solution of the following equation:

$$\frac{\rho_k(\gamma)}{1 + \frac{\rho_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}} = \frac{\alpha_k}{1 + \frac{\alpha_k - \xi_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}} \quad (29)$$

where $\alpha_k = \dfrac{k!}{k^{k-1}}$ is the Fredman-Komlós bound, $\rho_k(\gamma)$ can be found using expression (23), and $\xi_k(\gamma)$ is found via (14). Note that $\rho_k(\gamma)$ is a linear function and $\xi_k(\gamma)$ is a rational functions with degree $O(k)$, and therefore the above equation is equivalent to finding a root of a polynomial of degree $O(k)$ in variable γ, which lies in the interval $\left(\frac{1}{2k-3}, \frac{1}{k}\right)$. Such a solution certainly exists, because at $\gamma = \frac{1}{k}$ the LHS is less than α_k, while the RHS is equal to α_k, however at $\gamma = \frac{1}{2k-3}$ the LHS is greater than α_k while the RHS is less than α_k. Therefore, there exists a point $\gamma^* \in \left(\frac{1}{2k-3}, \frac{1}{k}\right)$ where these bounds are equal, since these functions are continuous. The values of the bounds for $\gamma = \frac{1}{k}$ guarantee that both bounds will be less than α_k when we take the optimal threshold γ^*. Therefore, for each k this optimal threshold γ^*, substituted into (28), gives a new upper bound on the rate of k-hash codes, which is stronger than the Fredman-Komlós bound (5).

Assuming the conjecture 4 holds, we obtain a stronger bound by using the exact value of $\theta_k(\gamma)$ instead of its upper bound $\rho_k(\gamma)$. The optimal threshold in this case is the solution of

$$\frac{\theta_k(\gamma)}{1+\frac{\theta_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}} = \frac{\alpha_k}{1+\frac{\alpha_k-\xi_k(\gamma)}{\log\left(\frac{k}{k-3}\right)}}, \tag{30}$$

where the value for $\theta_k(\gamma)$ is taken from (26). The value attained by the above expressions at the optimal threshold is the new upper bound on R_k.

New bounds for $k = 5$ and $k = 6$

Currently, the conjecture 4 is formulated in such a way that for any γ (from an appropriate range) one particular functional, parametrized with γ, has a maximum value on a simplex larger than some other set of functionals, also parametrized with γ. However, for our purposes of obtaining an upper bound on R_k, we only need the conjecture to hold specifically for the value of the optimal threshold $\gamma = \gamma^*$, found via (30). This is because if (26) holds for this γ^*, we then know that the upper bounds (15) and (19) for unbalanced and balanced cases are equal by the choice of γ^* and thus give a general upper bound on the rate R_k.

So to use the conjecture for a given k it just suffices to solve all optimization problems (24) for this value γ^*, and check if the conjecture indeed holds (namely, that the maximum of $\Theta_k^{(i)}(g,\gamma)$ is the greatest for the functional $\Theta_k^{\gamma}(g)$ described in the conjecture).

Applying (30) for $k = 5$ gives us the optimal threshold $\gamma_5^* \approx 0.136163$. We then use the numberical optimization tools in *Wolfram Mathematica* [7] to verify that the conjecture 4 indeed holds, which in this case reduces to optimizing two degree-3 multilinear polynomials with 5 variables over the simplex. After verifying the conjecture, we obtain the new general bound for 5-hashing:

$$R_5 < 0.190826 < 0.192 = \frac{24}{125} = \alpha_5.$$

For $k = 6$, the above approach gives us:

$$R_6 < 0.0922789 < 0.0\overline{925} = \frac{5}{54} = \alpha_6.$$

References

1. E. Arikan. A Bound on the Zero-Error List Coding Capacity. In *Proceedings. IEEE International Symposium on Information Theory*, pages 152–152, 1993. doi:10.1109/ISIT.1993.748467.
2. E. Arikan. An upper bound on the zero-error list-coding capacity. *IEEE Transactions on Information Theory*, 40(4):1237–1240, 1994. doi:10.1109/18.335947.
3. M. Dalai, V. Guruswami, and J. Radhakrishnan. An improved bound on the zero-error list-decoding capacity of the 4/3 channel. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1658–1662, 2017. doi:10.1109/ISIT.2017.8006811.
4. Peter Elias. Zero error capacity under list decoding. *IEEE Trans. Information Theory*, 34(5):1070–1074, 1988. doi:10.1109/18.21233.
5. 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(1):61–68, 1984. doi:10.1137/0605009.
6. G. Hansel. Nombre minimal de contacts de fermature nécessaires pour réaliser une fonction booléenne symétrique de n variables. *C. R. Acad. Sci. Paris*, pages 6037–6040, 1964.
7. Wolfram Research, Inc. Mathematica, Version 11.3. Champaign, IL, 2018.

8 J. Körner. Coding of an information source having ambiguous alphabet and the entropy of graphs. *6th Prague Conference on Information Theory*, pages 411–425, 1973.

9 J. Körner and K. Marton. New Bounds for Perfect Hashing via Information Theory. *European Journal of Combinatorics*, 9(6):523–530, 1988. doi:10.1016/s0195-6698(88)80048-9.

10 János Körner. Fredman–Komlós bounds and information theory. *SIAM Journal on Algebraic Discrete Methods*, 7(4):560–570, 1986. doi:10.1137/0607062.

11 A. Nilli. Perfect Hashing and Probability. *Combinatorics, Probability and Computing*, 3(03):407–409, 1994. doi:10.1017/s0963548300001280.

12 Jaikumar Radhakrishnan. Entropy and Counting, 2001. URL: http://www.tcs.tifr.res.in/~jaikumar/Papers/EntropyAndCounting.pdf.