Electronic Colloquium on Computational Complexity, Report No. 108 (2019)

Beating the probabilistic lower bound on perfect hashing

Chaoping Xing* Chen Yuan†

Abstract

For an interger $q \geqslant 2$, a perfect q-hash code C is a block code over $\mathbb{Z}_q := \mathbb{Z}/q\mathbb{Z}$ of length n in which every subset $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\}$ of q elements is separated, i.e., there exists $i \in [n]$ such that $\{\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_q)\} = \mathbb{Z}_q$, where $\mathrm{proj}_i(\mathbf{c}_j)$ denotes the ith position of $\mathbf{c}_j$. Finding the maximum size $M(n,q)$ of perfect q-hash codes of length n, for given q and n, is a fundamental problem in combinatorics, information theory, and computer science. In this paper, we are interested in asymptotical behavior of this problem. More precisely speaking, we will focus on the quantity $R_q := \limsup_{n\to\infty} \frac{\log_2 M(n,q)}{n}$.

A well-known probabilistic argument shows an existence lower bound on R_q, namely $R_q \geqslant \frac{1}{q-1}\log_2\left(\frac{1}{1-q!/q^q}\right)$ [8, 10]. This is still the best-known lower bound till now except for the case $q = 3$ for which Körner and Matron [11] found that the concatenation technique could lead to a perfect 3-hash code beating this the probabilistic lower bound. The improvement on the lower bound on R_3 was discovered in 1988 and there has been no any progress on lower bound on R_q for more than 30 years despite of some work on upper bounds on R_q. In this paper we show that this probabilistic lower bound can be improved for $q = 4, 8$ and all odd integers between 3 and 25, and *every sufficiently large odd* q. Our idea is based on a modified concatenation which is different from the classical concatenation for which both the inner and outer codes are separated. However, for our concatenation we do not require that the inner code is a perfect q-hash code. This gives a more flexible choice of inner codes and hence we are able to beat the probabilistic existence lower bound on R_q.

*School of Electronics, Information and Electric Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Email:⟨xingcp@sjtu.edu.cn⟩

†Centrum Wiskunde & Informatica, Amsterdam, the Netherlands. Email:⟨Chen.Yuan@cwi.nl⟩

ISSN 1433-8092

1 Introduction

One of the most powerful tools to derive lower bounds in theoretical computer science and extremal combinatorics is probabilistic method [1]. Roughly speaking, to prove the existence of an object of a given size satisfying certain conditions, one shows that a random object of this size (maybe after being slightly modified) has a positive probability to satisfy these conditions. In many problems the lower bound given by this method is conjectured exact, at least asymptotically, and sometimes one can prove it is indeed so. This means that optimal solutions to such problems are rather common. On the other hand, when the probabilistic lower bound is not asymptotically exact, optimal solutions tend to be rare and have some particular structure. So, from a theoretical point of view, it is of great importance to know whether a problem belongs to one or the other of these two classes. Some exceptional examples where the probabilistic lower bounds are not asymptotically exact include the Gilbert-Varshamov bound in coding theory [14] and the probabilistic lower bound on perfect hash codes [11]. In this paper, we study lower bounds on perfect hash codes and compare them with the probabilistic lower bound.

A q-ary code $C \subseteq \mathbb{Z}_q^n$ is said to be a perfect q-hash code if for every subset of C of q elements, there exists an coordinate where the q codewords in this subset have distinct values. The rate of a perfect q-hash code is defined as $R_C = \frac{\log_2 |C|}{n}$.

The existence of perfect q-hash code gives rise to a perfect q-hash family. To see this, let C be the whole universe and the projection of each coordinate be a hash function. Then, for any q elements of this universe, there exists a hash function mapping them to distinct value. Other application of perfect q-hash code includes the zero-error list decoding on certain channel. 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 $(w, v) \in E$ if on input v, the channel can output w. The $q/(q-1)$ channel then is the channel with $V = W = \{0, 1, \ldots, q-1\}$, and $(v, w) \in E$ if and only if $v \neq w$. If we want to ensure that the receiver can identify a subset of at most $q-1$ sequences that is guaranteed to contain the transmitted sequence, one can communicate via n repeated uses of the channel. See [7, 6] for more details.

In this paper, we only consider the asymptotic behavior of rates of perfect q-hash codes, namely, we focus on the quantity $R_q := \limsup_{n\to\infty} \frac{\log_2 M(n,q)}{n}$, where $M(n, q)$ stands for the maximum size of perfect q-hash codes of length n.

The study of R_q could be dated back to 80s. There are a few works dedicated to the upper bound on R_q. Fredman and Komlós [8] showed a general upper bound: $R_q \leqslant \frac{q!}{q^{q-1}}$ for all $q \geqslant 2$. Arikan [2] improved this bound for $q = 4$, and then Dalai, Guruswami and Radhakrishnan [6] further improved the upper bound on R_4. Recently, Guruswami and Riazanov [9] discovered a stronger bound for every $q \geq 4$. Although there are some works towards tightening the upper bound on R_q. There are a little work dedicated to lower bounds on R_q. A probabilistic argument shows the existence of perfect q-hash code with rate $R_q \geqslant \frac{1}{q-1} \log_2 \left(\frac{1}{1-q!/q^q}\right)$ [8, 10]. This is still the best-known lower bound till now except for the case $q = 3$ for which Körner and Matron [11] found that the concatenation technique could lead to a perfect 3-hash code beating this the probabilistic lower bound. The improvement on the lower bound on R_3 was discovered in 1988 and there has been no any progress on lower bounds on R_q for more than 30 years. Körner and Matron's idea is to concatenate an outer code, an 9-ary 3-hash code and an inner code, a perfect 3-hash code with size 9. They further posed an open problem whether there exist perfect q-hash code beating the

random argument for every q. In this paper, we provide a partial and affirmative answer to this open problem. We show that there exists perfect q-hash code beating the random argument for every sufficiently large odd q. To complement this result, we also prove the existence of perfect q-hash code that could beat random result for small q such as $q = 4, 8$ and all odd integers q between 3 and 25 (in fact for many other odd integers between 27 and 155 (see Remark 4)). Our computer search result together with asymptotical result suggests that our construction might beat the the probabilistic lower bound for every odd q.

The main technique of this paper is a modified concatenation. Unlike Körner and Matron's concatenation for which it is required that both inner and outer codes be separated, we abandon this separated requirement on the inner code at a cost of an even stricter requirement on the outer code. By relaxing the condition that the inner code is q-perfect hash code, we have more freedom to construct the inner code. As a result, we are able to improve the lower bound on R_q.

Before explaining our technique in detail, let us recall the concatenation technique by Körner and Matron. One first relies on the probabilistic argument to show the existence of an m-ary outer code C_1 of length n_1 that is q-separated with $q \leqslant m$, i.e., for every q-element subset of C_1 (a q-element set is a set of cardinality q), there exists $i \in \{1, 2, \ldots, n_1\}$ such that elements of this q-subset are pairwise distinct at position i. Then they construct a perfect q-hash code C_2 of length n_2 as an inner code. By concatenating C_1 and C_2 (this concatenation is slightly different form the classical concatenation in coding theory (see Lemma 2.3 for detail)), one obtains a perfect q-hash code of length $n_1 n_2$.

In our concatenation, we make a trade-off between inner code and outer code by relaxing the condition on the inner code and imposing a stricter condition on the outer code. We first take a set $\mathcal{A}$ consisting of some q-element subsets of $\mathbb{Z}_m$. Then we use the probabilistic argument to show existence of an m-ary outer code C_1 such that, for every q-element subset $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\}$ of C_1, $\{\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_{n_1})\} \in \mathcal{A}$, where $\mathrm{proj}_i(\mathbf{c}_j)$ stands for the ith coordinate of $\mathbf{c}_j$. Note that Körner and Matron's concatenation requires that $\{\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_{n_1})\}$ are pairwise distinct. Thus, we have a stricter condition on C_1. If we could find a q-ary inner code C_2 such that there are at least $|\mathcal{A}|$ q-element subsets of C_1 that are separated, concatenating these two codes could be shown to be a perfect q-hash code. Now, it remains to look for suitable inner code C_2. One good candidate for the inner code is an MDS code. In this paper, we choose an q-ary MDS code of length 3 and dimension 2 to be inner code C_2. To show improvement, we have to estimate the number of separated q-element subsets of C_2. We further reduce determining the number of separated q-element subsets of C_2 to determining the number of q-element subsets of C_2 in which all three positions are separated. It turns out that the latter problem is equivalent to the following well-known combinatorial problem: determine the number s_q of pairs (π_1, π_2) of bijections $\mathbb{Z}_q \to \mathbb{Z}_q$ such that $\pi_1 + \pi_2$ is a bijection as well. In literature, there is an asymptotic result on s_q for odd number q [13] which can be used to estimate the number of separated q-element subsets of C_2. As a result, we are able to improve R_q for large odd q. When it comes to the case where q is small, our computer search reveals that other codes could give a better lower bound although a $[3, 2]$-MDS code still leads to an improved lower bound on R_q.

Our main result is a new lower bound valid for every odd integer q.

Theorem 1.1. *For every odd integer q, there exists a perfect q-hash code over $\mathbb{Z}_q$ with rate R satisfying*

$$R \geqslant -\frac{1}{3(q-1)} \log_2 \left(1 - 3\frac{q!}{q^q} + 3\frac{(q!)^2}{q^{2q}} - \left(\frac{1}{\sqrt{e}} + o(1)\right) \frac{(q!)^3}{q^{3q-1}}\right).$$

This rate outperforms the probabilistic lower bound, $R_q \geqslant -\frac{1}{(q-1)}\log_2(1-\frac{q!}{q^q})$, for all sufficiently large odd q.

The paper is organized as follows. In Section 2, we propose a new concatenation technique and derive a lower bound on R_q in terms of the number of separated q-element subsets of the inner code. In Section 3, we provide several candidates for the inner code of our concatenation technique and estimate the number of separated q-element subsets for these candidates. By plugging this number into the lower bound in Section 2, we manage to prove that the probabilistic lower bound on R_q can be improved in many cases.

2 $\mathcal{A}$-friendly codes and concatenation

2.1 Hash code

Denote by $[n]$ the set $\{1, 2, \ldots, n\}$. A set containing q elements is called a q-element set. Define $\mathbb{Z}_m := \mathbb{Z}/m\mathbb{Z}$ be a congruence class of integers modulo m. Assume that $m \geqslant q$, then a q-element subset $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\}$ of $\mathbb{Z}_m^N$ is called separated if there exists $i \in [N]$ such that $\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_q)$ are pairwise distinct. Denote by $\mathbb{F}_q$ the finite field with q elements.

A subset C of $\mathbb{Z}_m^N$ is called an m-ary code of length N. For an integer $q \leqslant m$, an m-ary code C of length N is called an m-ary q-hash code if every q-element subset of C is separated. In particular, we say that C is a perfect q-hash code if $m = q$.

Let us generalize the notion of an m-ary q-hash code. Let $\binom{\mathbb{Z}_m}{q}$ denote the collection of all q-element subsets of $\mathbb{Z}_m = \{0, \ldots, m-1\}$. Let $\mathcal{A}$ be a subset of $\binom{\mathbb{Z}_m}{q}$ and let C be a code in $\mathbb{Z}_m^N$. We say that a q-element subset $\{\mathbf{c}_1, \ldots, \mathbf{c}_q\}$ of $\mathbb{Z}_m^N$ is $\mathcal{A}$-friendly if there exists $i \in [N]$ such that $\{\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_q)\} \in \mathcal{A}$. Otherwise, we call $\{\mathbf{c}_1, \ldots, \mathbf{c}_q\}$ an $\mathcal{A}$-unfriendly subset. If every q-element subset of C is $\mathcal{A}$-friendly, we say that C is an $\mathcal{A}$-friendly code. In particular, this definition coincides with an m-ary q-hash code when $\mathcal{A} = \binom{\mathbb{Z}_m}{q}$.

2.2 Random $\mathcal{A}$-friendly codes

In this subsection, by applying a probabilistic argument, we show existence of $\mathcal{A}$-friendly codes. This is a generalization for random argument showing existence of perfect hash codes.

Lemma 2.1. *Let $\mathcal{A}$ be a nonempty subset of $\binom{\mathbb{Z}_m}{q}$. Then there exists an m-ary $\mathcal{A}$-friendly code C of length N and size at least $\lceil \frac{M}{3} \rceil$ as long as*

$$\binom{M}{q}\left(1 - \frac{q!|\mathcal{A}|}{m^q}\right)^N \leq \frac{M}{2q} \tag{1}$$

Proof. We sample M codewords $\mathbf{c}_1, \ldots, \mathbf{c}_M$ uniformly at random in $\mathbb{Z}_m^N$ with replacement. The number of collisions is negligible compared to M. To see this, let $X_{i,j}$ be the $0,1$-random variable such that $X_{i,j} = 1$ if $\mathbf{c}_i = \mathbf{c}_j$ and $X_{i,j} = 0$ otherwise. It is clear $P[X_{i,j} = 1] = m^{-N}$. It follows that $E[\sum_{1 \leq i < j \leq M} X_{i,j}] = \binom{M}{2} m^{-N} = o(M)$ due to the fact that $M = o(m^N)$. Next, we bound the number of $\mathcal{A}$-friendly q-element sets from these M codewords. Let us consider the q-element set

$\{\mathbf{c}_1, \ldots, \mathbf{c}_q\}$ with $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,N})$. For any $j \in [n]$, the probability that $\{c_{1,j}, \ldots, c_{q,j}\} \in \mathcal{A}$ is $\frac{q!|\mathcal{A}|}{m^q}$ as $c_{i,j}$ is picked uniformly at random in $\mathbb{Z}_m$. It follows that the probability that $\{\mathbf{c}_1, \ldots, \mathbf{c}_q\}$ is not $\mathcal{A}$-friendly is $(1 - \frac{q!|A|}{m^q})^N$. There are at most $\binom{M}{q}$ q-element sets from $\{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$. By union bound, the expected number of $\mathcal{A}$-unfriendly q-element sets is at most $\binom{M}{q}\left(1 - \frac{q!|\mathcal{A}|}{m^q}\right)^N \leq \frac{M}{2q}$. Remove all the codewords that lie in any of these $\mathcal{A}$-unfriendly q-element sets. Then, we remove at most $q \times \frac{M}{2q} = \frac{M}{2}$ codewords. According to our previous argument, there are $o(M)$ collisions among these M codewords. Remove these $o(M)$ codewords and we obtain an $\mathcal{A}$-friendly code of size at least $\frac{M}{3}$. The desired result follows. □

Remark 1. Note that in [11], the set $\mathcal{A}$ is the collection of all q-element subsets of $\mathbb{Z}_m$. Thus, our random argument can be viewed as a generalization of the argument in [11]. This generalization allows us to relax the constraint on our inner code C_1, i.e., C_1 is not necessary a perfect q-hash code, although we propose a stricter constraint on the outer code. Instead of requiring that C_1 is a perfect q-hash code, we only require that $|\mathcal{A}|/\binom{m}{q}$ fraction of q-element sets of C_1 are separated.

If we choose $m = q$ in Lemma 2.1, then $|\mathcal{A}| = 1$. Thus, we obtain a random construction of perfect q-hash codes.

Corollary 2.2. *Let $q \geqslant 2$. Then there exists q-hash code of length N and size at least $\left\lceil \frac{M}{3} \right\rceil$ as long as*

$$\binom{M}{q}\left(1 - \frac{q!}{q^q}\right)^N \leq \frac{M}{2q}. \tag{2}$$

In particular, we have a random q-hash code with rate

$$R = \frac{\log_2 M}{N} = -\frac{1}{q-1}\log_2\left(1 - \frac{q!}{q^q}\right) + \frac{O(1)}{N}. \tag{3}$$

Hence, we have a probabilistic lower bound

$$R_q \geqslant \frac{1}{q-1}\log_2\left(\frac{1}{1 - q!/q^q}\right). \tag{4}$$

Proof. As $\binom{M}{q} \leqslant \frac{M^q}{q!}$, the following inequality

$$\frac{M^q}{q!}\left(1 - \frac{q!}{q^q}\right)^N \leq \frac{M}{2q} \tag{5}$$

implies the inequality (2). Choose M to be the largest integer satisfying the inequality (5) and consider the limit $\lim_{N\to\infty} \frac{\log_2 M}{N}$. The desired equality (3) follows. □

2.3 A concatenation technique

Let C be a q-ary code of length n and size m. Denote by $\mathcal{S}(C)$ the collection of all q-element subsets of C that are separated.

Lemma 2.3. *Let C be a q-ary code of length n and size m. Then one has*

$$R_q \geqslant -\frac{1}{(q-1)n}\log_2\left(1 - \frac{q!|\mathcal{S}(C)|}{m^q}\right). \tag{6}$$

Proof. Denote C by C_2 and put $n_2 = n$. Let π be any bijection from C_2 to $\mathbb{Z}_m$. Define $\mathcal{A} := \bigcup_{\{\mathbf{c}_1,\ldots,\mathbf{c}_q\}\in\mathcal{S}(C)}\{\pi(\mathbf{c}_1),\ldots,\pi(\mathbf{c}_q)\}$. It is clear that $\mathcal{A} \subseteq \binom{\mathbb{Z}_m}{q}$ and $|\mathcal{A}| = |\mathcal{S}(C)|$. Lemma 2.1 tells us that there exists an m-ary $\mathcal{A}$-friendly code C_1 of length n_1 with rate

$$R = -\frac{1}{(q-1)}\log_2\left(1 - \frac{q!|\mathcal{A}|}{m^q}\right) + \frac{O(1)}{n_1}.$$

Let C be the concatenation of C_1 with C_2, i.e.,

$$C := \{\pi^{-1}(\mathbf{c}) = (\pi^{-1}(c_1), \pi^{-1}(c_2), \ldots, \pi^{-1}(c_{n_1})) : \ \mathbf{c} = (c_1, c_2, \ldots, c_{n_1}) \in C_1\}.$$

Clearly, the rate of C is $R = -\frac{1}{n(q-1)}\log_2(1 - \frac{q!|\mathcal{A}|}{m^q}) + \frac{O(1)}{n_1 n_2}$. It remains to show that C is a perfect q-hash code.

Choose a q-element subset $\{\pi^{-1}(\mathbf{c}_1), \pi^{-1}(\mathbf{c}_2), \ldots, \pi^{-1}(\mathbf{c}_q)\}$ from C with $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\}$ being a q-element subset of C_1. Since C_1 is $\mathcal{A}$-friendly, there exists $i \in [N]$ such that $\{\mathrm{proj}_i(\mathbf{c}_1), \mathrm{proj}_i(\mathbf{c}_2), \ldots, \mathrm{proj}_i(\mathbf{c}_q)\} \in \mathcal{A}$. This implies that $\{\pi^{-1}(\mathrm{proj}_i(\mathbf{c}_1)), \ldots, \pi^{-1}(\mathrm{proj}_i(\mathbf{c}_q))\} \in \mathcal{S}(C)$ and thus $\{\pi^{-1}(\mathbf{c}_1), \pi^{-1}(\mathbf{c}_2), \ldots, \pi^{-1}(\mathbf{c}_q)\}$ is separated. The desired result follows from the definition of perfect q-hash codes. □

Remark 2. Given a q-ary inner code C_2 of length n, Lemma 2.3 tells us there must exist an outer code whose concatenation with this inner code gives a perfect q-hash code with rate $-\frac{1}{n(q-1)}\log_2(1 - \frac{q!|\mathcal{S}(C_2)|}{m^q})$. That means we only need to focus on finding good inner codes C_2 with large subset $\mathcal{S}(C_2)$. In what follows, when we talk about concatenation, we only specify the inner code. The outer code is always given by Lemma 2.3.

3 Beating probabilistic lower bound

By Lemma 2.3, to have a good lower bound on R_q, one needs a find to a q-ary inner code C of length n such that $\mathcal{S}(C)$ has large size for fixed q, n and size $|C|$. However, determining (or even estimating) the size of $\mathcal{S}(C)$ for a given inner code C with dimension at least 2 seems very difficult. In this section, we estimate the size of $\mathcal{S}(C)$ for some classes of codes and show that these inner codes give lower bounds on R_q better than the probabilistic lower bound (4)

3.1 Lower bounds from linear codes

To overcome the problem of estimating the size of $\mathcal{S}(C)$, we need to resort to linearity and dual distance of linear codes and narrow our target to linear codes with simple structure. In this subsection, we investigate a promising candidate for the inner code.

Let us recall some facts on linear codes. Let q be a prime power and let C be a q-ary $[n,k]$-linear code. A subset I of $[n]$ of size k is called an information set of C if every codeword $\mathbf{c} \in C$ is uniquely determined by $\mathbf{c}_I$, where $\mathbf{c}_I$ is the projection of $\mathbf{c}$ at I. In other words, let G be a generator of C, then a subset I of $[n]$ of size k is an information set of C if and only if G_I is a $k \times k$ invertible matrix, where G_I is the submatrix of G consisting of those columns of G indexed by $i \in I$.

Lemma 3.1. *Let C be a q-ary $[n,k]$-linear code with dual distance $d^\perp$. Then for any subset J of $[n]$ with $|J| \leqslant d^\perp - 1$, there exists an information set I such that $J \subseteq I$.*

Proof. Let G be a generator of C. As C has dual distance $d^\perp$, any $d^\perp - 1$ columns of G are linearly independent. Thus, the submatrix G_J has rank $|J|$. Hence, one can find a subset I of $[n]$ of size k such that $J \subseteq I$ and G_I has rank k. The proof is completed. □

Let q be a prime power and assume that there is a q-ary $[n,k]$-linear code C with dual distance $d^\perp$. For each $i \in [n]$, define the set

$$\mathcal{A}_i = \{\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\} \subseteq C : \{\mathrm{proj}_i(\mathbf{c}_1)), \ldots, (\mathrm{proj}_i(\mathbf{c}_q))\} = \{0, \ldots, q-1\}\}.$$

Thus, we have $\mathcal{S}(C) = \cup_{i=1}^n \mathcal{A}_i$.

For any subset T of $[n]$, we denote by $\mathcal{A}_T$ the set $\cap_{i \in T} \mathcal{A}_i$. Let A_i denote the number

$$A_i = \sum_{T \subseteq [n], |T| = i} |\mathcal{A}_T|. \tag{7}$$

Lemma 3.2. *Let C be a q-ary $[n,k]$-linear code with dual distance $d^\perp$. Then*

$$|\mathcal{S}(C)| = \sum_{i=1}^{d^\perp - 1} (-1)^{i-1} \binom{n}{i} q^{q(k-i)} (q!)^{i-1} + \sum_{i=d^\perp}^{n} (-1)^i A_i. \tag{8}$$

Proof. First we claim that for any $j \in [d^\perp - 1]$ and subset J of $[n]$ with $|J| = j$, we have $|\mathcal{A}_J| = q^{q(k-j-1)}(q!)^{j-1}$.

By Lemma 3.1, we can choose an information set $I \subseteq [n]$ that includes J. For any matrix M in $\mathbb{F}_q^{q \times k}$, by the definition of the information set, there is a unique q-tuple $(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q)$ such that

$$M = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_q \end{pmatrix}_I. \tag{9}$$

It is clear that $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\} \in \mathcal{A}_J$ if and only every column of M_J is a permutation of $(0, \ldots, q-1)$. There are $(q!)^{|J|} = (q!)^j$ ways to pick M_J and $q^{q(|I|-|J|)} = q^{q(k-j)}$ ways to pick M_{I-J}. This gives $(q!)^j q^{q(k-j)}$ different q-tuples $(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q)$ with $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\} \in \mathcal{A}_T$. It follows that the number of q-element sets in $\mathcal{A}_T$ is $(q!)^{j-1} q^{q(k-j)}$.

By the inclusion-exclusion principle, we have

$$|\mathcal{S}(C)| = |\bigcup_{i=1}^n \mathcal{A}_i| = \sum_{i=1}^{d^\perp - 1} (-1)^{i-1} \binom{n}{i} q^{q(k-i)} (q!)^{i-1} + \sum_{i=d^\perp}^{n} (-1)^{i-1} A_i.$$

□

By the equality (8), we have

$$\begin{aligned}
|\mathcal{S}(C)| &= \sum_{i=1}^{n}(-1)^{i-1}\binom{n}{i}q^{q(k-i)}(q!)^{i-1} - \sum_{i=d^\perp}^{n}(-1)^{i-1}\binom{n}{i}q^{q(k-i)}(q!)^{i-1} + \sum_{i=d^\perp}^{n}(-1)^{i-1}A_i \\
&= \frac{-q^{qk}}{q!q^{qn}}\left(-q^{qn} + (q^q - q!)^n\right) - \sum_{i=d^\perp}^{n}(-1)^{i-1}\binom{n}{i}q^{q(k-i)}(q!)^{i-1} + \sum_{i=d^\perp}^{n}(-1)^{i-1}A_i \\
&= \frac{q^{qk}}{q!}\left(1 - \left(1 - \frac{q!}{q^q}\right)^n\right) - \sum_{i=d^\perp}^{n}(-1)^{i-1}\binom{n}{i}q^{q(k-i)}(q!)^{i-1} + \sum_{i=d^\perp}^{n}(-1)^{i-1}A_i.
\end{aligned}$$

Thus, we have

$$1 - \frac{q!|\mathcal{S}(C)|}{q^{qk}} = \left(1 - \frac{q!}{q^q}\right)^n + \sum_{i=d^\perp}^{n}(-1)^{i-1}\binom{n}{i}\left(\frac{q!}{q^q}\right)^i - \frac{q!}{q^{qk}}\sum_{i=d^\perp}^{n}(-1)^{i-1}A_i.$$

Hence, in order to beat the probabilistic lower bound, we need to verify the following inequality for an inner code $C = [n,k]_q$.

$$\sum_{i=d^\perp}^{n}(-1)^{i-1}\binom{n}{i}\left(\frac{q!}{q^q}\right)^i < \frac{q!}{q^{qk}}\sum_{i=d^\perp}^{n}(-1)^{i-1}A_i \tag{10}$$

Lemma 3.2 shows that computing $|\mathcal{S}(C)|$ is reduced to computing A_i for $i = d^\perp, d^\perp+1, \ldots, n$. However, if $d^\perp$ is too far from n, we have to compute many A_i and this is rather difficult. The simplest case is $d^\perp = n$ where we need to compute only A_n. In this case the dimension k is at least $n-1$. Therefore, let us consider $[n, n-1]$ MDS codes. On the other hand, when C has dimension $n-1$, we do not require that q is a prime power. Precisely speaking, we have the following result.

Lemma 3.3. *Let $q \geqslant 2$ be an integer and consider the q-ary code $C = \{(x_1, \ldots, x_{n-1}, \sum_{i=1}^{n-1} x_i) : x_1, \ldots, x_{n-1} \in \mathbb{Z}_q\}$. Let A_n denote the cardinality of the set*

$$\{\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q\} \subseteq C : \ \{\mathrm{proj}_i(\mathbf{c}_1)), \ldots, (\mathrm{proj}_i(\mathbf{c}_q))\} = \{0, \ldots, q-1\}\} \text{ for any } i \in [n]\}.$$

Then $|\mathcal{S}(C)| = \frac{q^{q(n-1)}}{q!}\left(1 - \left(1 - \frac{q!}{q^q}\right)^n\right) - (-1)^{n-1}q^{-q}(q!)^{n-1} + (-1)^{n-1}A_n$.

Proof. One can show that in this case, we have $A_i = q^{q(n-i-1)}(q!)^{i-1}$ for any $1 \leqslant i \leqslant n-1$. The desired result follows from the same arguments as in Lemma 3.2. □

Corollary 3.4. *Let $q \geqslant 2$ be an integer. Let A_n be the number given in Lemma 3.3. If*

$$(-1)^{n-1}A_n > (-1)^{n-1}\frac{(q!)^{n-1}}{q^q}, \tag{11}$$

Then there exist families of perfect q-hash code over $\mathbb{F}_q$ with rate better than the probabilistic lower bound (4).

Proof. Let C be the q-ary code defined in Lemma 3.3. Then

$$1 - \frac{q!|\mathcal{S}(C)|}{q^{q(n-1)}} = \left(1 - \frac{q!}{q^q}\right)^n + (-1)^{n-1}\left(\frac{q!}{q^q}\right)^n - (-1)^{n-1}\frac{q!}{q^{q(n-1)}}A_n < \left(1 - \frac{q!}{q^q}\right)^n.$$

The desired result follows. □

If C is the code of length 3 given in Lemma 3.3, i.e, $C = \{(x, y, x + y) : \ x, y \in \mathbb{Z}_q\}$, then determining A_3 given in Lemma 3.3 is actually reduced to the following well-known combinatorial problem: determine the number s_q of pairs (π_1, π_2) of bijections $\mathbb{Z}_q \to \mathbb{Z}_q$ such that $\pi_1 + \pi_2$ is a bijection as well. The relation between A_3 and s_q is $A_3 = \frac{s_q}{q!}$.

The number s_q has been studied somewhat extensively, but under a different guise [3, 5, 4, 15, 13]. It is in general very difficult to determine the exact value of s_q unless q is an even number for which $s_q = 0$. It has been conjectured in [15] that there exists two positive constant c_1 and c_2 such that $c_1^q(q!)^2 < s_q < c_2^q(q!)^2$ for all odd q. Various upper bounds are given [13]. To beat the probabilistic lower bound on R_q, we want to show $s_p > (\frac{q!}{q^q})^2$. That means, we are only interested in the lower bounds on s_q. A generic lower bound is $s_q \geqslant 3.246^n \times n!$ for all odd n. However, there is still a very big gap between this lower bound and the aforementioned conjecture. On the other hands, there are various algorithms to numerically approximate s_n [12].

By taking exact value of s_q for all odd q between 3 and 25 from [12], we obtain the following result.

Corollary 3.5. *There exists a family of perfect q-hash codes over $\mathbb{Z}_q$ with rate better than the probabilistic lower bound* (4) *for all odd q between* 3 *and* 25*.*

Proof. By Corollary 3.4, it is sufficient to verify the inequality

$$\frac{s_q}{q!} > \frac{(q!)^2}{q^q} \tag{12}$$

for all odd q between 3 and 25. Taking the values of s_q from Table I of [12] gives the desired claim. □

Remark 3. For completeness, we list the values of $A_3 = \frac{s_q}{q!}$ and $\frac{(q!)^2}{q^q}$ for odd $q \in [3, 25]$ in Table 3. We observe that the ratio A_3 over $\frac{(q!)^2}{q^q}$ grows slowly but monotonically. In fact, we will see that this ratio is asymptotically equal to $\frac{q}{\sqrt{e}}$ in our following discussion.

Remark 4. In literatures, various algorithms were proposed to compute s_q for large odd q. Using these algorithms, for many odd q in the interval $[27, 155]$, estimation on s_q is given in [12]. One can verify from these estimation that the probabilistic lower bound (4) is improved for all odd integers q for which available values of s_q are given in [12].

For even q, we have $s_q = 0$. Therefore, we cannot use the codes defined in Lemma 3.3. Instead, we can replace $\mathbb{Z}_q$ by $\mathbb{F}_q$ if q is a prime power.

Corollary 3.6. *There exists a family of perfect q-hash code over $\mathbb{F}_q$ with rate better than the probabilistic lower bound* (4) *for $q = 4, 8, 9$. Furthermore, the lower bound on R_9 given this corollary is better than that in Corollary 3.5 and the probabilistic lower bound.*

$\mathbb{Z}_q$	$\mathbb{Z}_5$	$\mathbb{Z}_7$	$\mathbb{Z}_9$	$\mathbb{Z}_{11}$	$\mathbb{Z}_{13}$	$\mathbb{Z}_{15}$
A_3	15	133	2025	37851	1.03×10^6	3.63×10^7
$\frac{(q!)^2}{q^q}$	4.6	30.8	339.9	5584.6	1.28×10^5	3.90×10^6
Ratio	3.26	4.32	5.96	6.78	8.04	9.30
$\mathbb{Z}_q$	$\mathbb{Z}_{17}$	$\mathbb{Z}_{19}$	$\mathbb{Z}_{21}$	$\mathbb{Z}_{23}$	$\mathbb{Z}_{25}$	
A_3	1.60×10^9	8.76×10^{10}	5.77×10^{12}	4.52×10^{14}	4.16×10^{16}	
$\frac{(q!)^2}{q^q}$	$1.52 * 10^8$	7.47×10^9	4.47×10^{11}	3.2×10^{13}	2.70×10^{15}	
Ratio	10.53	11.71	12.93	14.12	15.4	

Table 1: The comparison between A_3 and $\frac{(q!)^2}{q^q}$ for small odd q.

Proof. Let C be a code with the form

$$C = \{(x, y, x + y) : x, y \in \mathbb{F}_q\}.$$

Let A_3 be defined in (7). With the help of computer search, we get the values A_3 of C: 8 for code over $\mathbb{F}_4$, 384 for code over $\mathbb{F}_8$ and 2241 for code over $\mathbb{F}_9$, respectively. We note the fact that A_3 from code over $\mathbb{F}_9$ is 2241, while A_3 from code over $\mathbb{Z}_9$ is 2025.

It is straightforward to verify that the inequality $A_3 > \frac{(q!)^2}{q^q}$ is satisfied for $q = 4, 8$ and 9.

□

Remark 5. The lower bound on R_3 given in [11] is $R_3 \geqslant \frac{1}{4} \log_2 \frac{9}{5}$. Let C be a ternary $[4, 3]$-MDS code. The computer search shows that $|\mathcal{S}(C)| = 84$. By Lemma 2.3, we also obtain the same lower bound $R_3 \geqslant \frac{1}{4} \log_2 \frac{9}{5}$.

This remark indicates that q-ary MDS codes of larger length sometimes leads to a better lower bound on R_q than q-ary MDS codes of length 3 and dimension 2. This is further confirmed by the following example for $q = 4$.

Corollary 3.7. *There exists a family of perfect* 4*-hash code over* $\mathbb{F}_4$ *with rate at least* 0.049586*. This is better than both the lower bound given in Corollary 3.6 and the probabilistic lower bound.*

Proof. Assume $\mathbb{F}_4 = \{0, 1, \alpha, \alpha + 1\}$. Consider a $[5, 2]$-MDS code:

$$C = \{(a, b, a + b, a\alpha + b, a(\alpha + 1) + b) : a, b \in \mathbb{F}_4\}.$$

By computer search, we find that there are 1100 out of $\binom{32}{4}$ 4-element subsets of C that are separated. Plugging it parameters into Lemma 2.3, we obtain perfect 4-hash code with rate 0.049586. □

For some odd integers q large than 25, there are also some lower bounds on s_q [12]. By these lower bounds, we can verify that the probabilistic lower bound on R_q are improved for odd integers between 27 and 155. The computer search can only help to solve the small case. To lower bound s_q for large q, we have to look for a lower bound with rigorous mathematical proof. Fortunately, a recent progress on asymptotic behavior of s_q is given in [13]. Recall that there is a conjecture

saying that, for all odd q, the number s_q lies in between $c_1^n n!^2$ and $c_2^n n!^2$ for some constants c_1, c_2. This conjecture is recently confirmed in [13]. Moreover, they even close the gap by showing $c_1 = c_2 = \frac{1}{\sqrt{e}} + o(1)$.

Proposition 3.8 ([13]). *Let q be an odd integer. Then, the number s_q is $(\frac{1}{\sqrt{e}} + o(1))\frac{q!^3}{q^{q-1}}$, and hence A_3 defined in Lemma 3.3 is $(\frac{1}{\sqrt{e}} + o(1))\frac{q!^2}{q^{q-1}}$.*

Plugging A_3 in Proposition 3.8 into (8) and (6) gives the following theorem.

Theorem 3.9. *For odd integer q, there exists perfect q-hash code over $\mathbb{Z}_q$ with rate at least*

$$R = -\frac{1}{3(q-1)} \log_2 \left(1 - 3\frac{q!}{q^q} + 3\frac{(q!)^2}{q^{2q}} - \left(\frac{1}{\sqrt{e}} + o(1)\right)\frac{(q!)^3}{q^{3q-1}}\right).$$

Moreover, for every sufficiently large odd q this rate is bigger than that given by the probabilistic lower bound.

Proof. It remains to compare this rate with (3). It suffices to show that $A_3 > \frac{(q!)^2}{q^q}$. For large q, this inequality is reduced to prove $\left(\frac{1}{\sqrt{e}} + o(1)\right)\frac{(q!)^3}{q^{3q-1}} > \frac{(q!)^3}{q^{3q}}$. This holds as $\frac{1}{\sqrt{e}} + o(1) > \frac{1}{q}$ for sufficiently large q. □

3.2 Lower bounds on R_5 and R_7

In previous subsection, we make use of linear codes C and estimate the size $|\mathcal{S}(C)|$ either numerically or asymptotically. However, linear codes do not always give the best lower bound on R_q. In this subsection we present a class of nonlinear inner code C_2 where many q-element subsets are separated.

Lemma 3.10. *Assume q is a prime. There exists a code C over $\mathbb{Z}_q$ with length q and size $2q$ such that $|\mathcal{S}(C)| = 2^q q - 2(q-1)$.*

Proof. Let $C_1 = \{\mathbf{c}_1 = (0, 1 \ldots, q-1), \mathbf{c}_2 = (1, 2 \ldots q-1, 0), \ldots, \mathbf{c}_q = (q-1, 0, \ldots, q-2)\}$, i.e., C_1 consists of the codeword $(0, 1 \ldots, q-1)$ and its ith shifts for $i = 1, \ldots, q-1$. Let $C_2 = \{i \cdot \mathbf{1} : \ 0 \leqslant i \leqslant q-1\}$, where $\mathbf{1}$ stands for all-one vector of length q. Let $C = C_1 \cup C_2$. Obviously, C has length q and size $2q$. It remains to show that $|\mathcal{S}(C)| = 2^q q - 2(q-1)$.

We pick any $0 < i < q$ codewords $\mathbf{c}_1, \ldots, \mathbf{c}_i$ from C_1. Denote by $\mathbf{c}_j = (c_{j,1}, \ldots, c_{j,q})$ for $j \in [q]$. For coordinate $t \in [q]$, let $B_t := \{c_{1,t}, c_{2,t}, \ldots, c_{i,t}\}$ be the collection of t-th components of $\mathbf{c}_1, \ldots, \mathbf{c}_i$. It is clear that $|B_t| = i$ by observing that all codewords in C_1 have distinct values on each coordinate. Moreover, we will prove that $B_1, \ldots, B_q$ are distinct if $0 < i < q$. Assume not and we have $B_1 = B_a$ for some $a \in [q]$. The structure of code C_1 tells us that $c_{j,a} = c_{j,1} + a - 1$ for $j = 1, \ldots, i$. This coupled with $B_1 = B_a$ implies that $c_{1,1}, c_{1,1} + a - 1$ both belong to B_1. Continue this argument and we finally arrive at $\{c_{1,1}, c_{1,1} + a - 1, \ldots, c_{1,1} + (q-1)(a-1)\} \subseteq B_1$. It is clear that $c_{1,1}, c_{1,1} + a - 1, \ldots, c_{1,1} + (q-1)(a-1)$ are distinct which contradicts our assumption that $|B_t| = i < q$.

Now, we know that $B_1, \ldots, B_q$ are distinct. For each set $B_t = \{c_{1,t}, c_{2,t}, \ldots, c_{i,t}\}$, we choose a $(q-i)$-element set A_t $\{\mathbf{i} : i \notin B_t\} \subseteq C_2$. It is clear that $\mathbf{c}_1, \ldots, \mathbf{c}_i$ and the codewords in B have distinct symbols on i-th coordinate. Moreover, for each value t, the set A_t is distinct due to the fact

that $B_1, \ldots, B_q$ are distinct. That means, for any $0 < i < q$-element set of C_1, we could obtain q distinct q-element sets of C that are separated. If $i = 0$ or $i = q$, it is clear that the only q-element set that are separated is C_1 or C_2. Thus, the total number of q-element sets of C that are separated is $\sum_{i=1}^{q-1} q\binom{q}{i} + 2 = 2^q q - 2(q-1)$. □

Combined this construction with Lemma 2.3 gives following lower bounds on R_q for $q = 5$ and 7.

Corollary 3.11. *One has $R_5 \geqslant 0.01452$ and $R_7 \geqslant 0.001483$. Furthermore, the lower bounds on R_5 and R_7 given in this corollary are better than those in Corollary 3.5 and the probabilistic lower bound.*

Proof. Take the inner code to be the code in Lemma 3.10 for $q = 5$ and 7, respectively. The desired result follows from Lemma 3.10 and 2.3. □

Let us end this section by tabulating our best lower bound, denoted by R_{new}, obtained in this paper and the probabilistic lower bound denoted by R_{ran} for some small q. We omit cases for $q \geq 13$.

q	4	5	7	8	9	11
R_{new}	0.0495	0.01452	0.001483	4.95909×10^{-4}	1.689931×10^{-4}	$2.01855746 \times 10^{-5}$
R_{ran}	0.0473	0.01412	0.001476	4.95905×10^{-4}	1.689929×10^{-4}	$2.01855739 \times 10^{-5}$

Table 2: Lower bounds

References

[1] Noga Alon and Joel H. Spencer. *The Probabilistic Method, Third Edition*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2008.

[2] Erdal Arikan. Upper bound on the zero-error list-coding capacity. *Information Theory, IEEE Transactions on*, 40:1237 – 1240, 08 1994.

[3] Miklos Bona. *Handbook of enumerative combinatorics*. Discrete Mathematics and Its Applications. CRC Press, Hoboken, NJ, 2015.

[4] C. Cooper. A lower bound for the number of good permutations. *Nat. Acad. Sci. Ukraine*, 213:15–25, 2000.

[5] C. Cooper, R. Gilchrist, I. N. Kovalenko, and D. Novakovic. Estimation of the number of "good" permutatio with applications to cryptography. *Cybernetics and Systems Analysis*, 35(5):688–693, Sep 1999.

[6] M. Dalai, V. G. Carnegie, 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, June 2017.

[7] P. Elias. Zero error capacity under list decoding. *IEEE Transactions on Information Theory*, 34(5):1070–1074, Sep. 1988.

[8] M. Fredman and J. Komlós. On the size of separating systems and families of perfect hash functions. *SIAM Journal on Algebraic Discrete Methods*, 5(1):61–68, 1984.

[9] Venkatesan Guruswami and Andrii Riazanov. Beating Fredman-Komlós for Perfect k-Hashing. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132, pages 92:1–92:14, Dagstuhl, Germany, 2019.

[10] J. Körner. Fredman-komlós bounds and information theory. *SIAM Journal on Algebraic Discrete Methods*, pages 560–570, 1986.

[11] J. Körner and K. Marton. New bounds for perfect hashing via information theory. *European Journal of Combinatorics*, 9(6):523–530, 1988.

[12] N Kuznetsov. Applying fast simulation to find the number of good permutations. *Cybernetics and Systems Analysis - CYBERN SYST ANAL-ENGL TR*, 43:830–837, 11 2007.

[13] F. Manners S. Eberhard and R. Mrazović. Additive triples of bijections, or the toroidal semiqueens problem. *Journal of the European Mathematical Society*, 21(2):441–463, 2019.

[14] M. A. Tsfasman, S. G. Vlădutx, and Th. Zink. Modular curves, shimura curves, and goppa codes, better than varshamov-gilbert bound. *Mathematische Nachrichten*, 109(1):21–28, 1982.

[15] Ilan Vardi. *Computational recreations in Mathematica*. Addison Wesley, 1991.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il