Zero error list-decoding capacity of the $q/(q-1)$ channel

Sourav Chakraborty[1], Jaikumar Radhakrishnan[2,3], Nandakumar Raghunathan [⋆4], and Prashant Sasatte[⋆⋆5]

[1] Department of Computer Science, University of Chicago, Chicago, USA
[2] Toyota Technological Institute at Chicago, USA
[3] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India
[4] Microsoft, Seattle, USA
[5] University of Waterloo, Waterloo, Canada

Abstract. Let m, q, ℓ be positive integers such that $m \geq \ell \geq q$. A family $\mathcal{H}$ of functions from $[m]$ to $[q]$ is said to be an (m, q, ℓ)-family if for every subset S of $[m]$ with ℓ elements, there is an $h \in \mathcal{H}$ such that $h(S) = [q]$. Let, $N(m, q, \ell)$ be the size of the smallest (m, q, ℓ)-family. We show that for all q, $\ell \leq 1.58q$ and all sufficiently large m, we have

$$N(m, q, \ell) = \exp(\Omega(q)) \log m.$$

Special cases of this follow from results shown earlier in the context of perfect hashing: a theorem of Fredman & Komlós (1984) implies that $N(m, q, q) = \exp(\Omega(q)) \log m$, and a theorem of Körner (1986) shows that $N(m, q, q+1) = \exp(\Omega(q)) \log m$. We conjecture that $N(m, q, \ell) = \exp(\Omega(q)) \log m$ if $\ell = O(q)$. A standard probabilistic construction shows that for all q, $\ell \geq q$ and all sufficiently large m,

$$N(m, q, \ell) = \exp(O(q)) \log m.$$

Our motivation for studying this problem arises from its close connection to a problem in coding theory, namely, the problem of determining the *zero error* list-decoding capacity for a certain channel studied by Elias [IEEE Transactions on Information Theory, Vol. 34, No. 5, 1070–1074, 1988]. Our result implies that for the so called $q/(q-1)$ channel, the capacity is exponentially small in q, even if the list size is allowed to be as big as $1.58q$. The earlier results of Fredman & Komlós and Körner cited above imply that the capacity is exponentially small if the list size is at most $q + 1$.

1 Introduction

Shannon [S56] studied the zero error capacity of discrete finite memoryless noisy channels. Such a channel can be modeled as a bipartite graph (V, W, E),

[⋆] Work done while the author was at the University of Chicago.
[⋆⋆] Work done while the author was at TIFR, Mumbai.

where $(v, w) \in E$ iff the letter w can be received when the letter v is transmitted. The goal then, is to encode messages as strings of letters from the input alphabet V and recover it from the received message. The goal naturally is to use as few input letters as possible and still recover the intended message perfectly. Shannon [S56] and Lovász [L79] determined the best rate of transmission achievable under this model for several specific channels.

We are interested in the list-decoding version of this problem, studied by Elias [E88]. For example, consider the channel shown in Figure 1. It is not hard to see that for this channel, no matter how many letters are used in the encoding, it is impossible to recover an input message uniquely (assuming there are at least two possibilities for the input message). However, it is not hard to see that one can always encode messages using strings of letters such that based on the received message one can narrow down the possibilities to just two, that is, we cannot decode exactly but we can list-decode with a list of size two. This motivates the following definition.

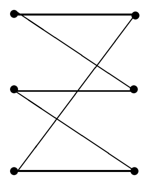

Fig. 1. The 3/2 channel

Definition 1 (Code, Rate). *Consider a channel $C = ([q], [q'], E)$ with q input letters and q' output letters. We say that a sequence $\sigma \in [q]^n$ is compatible with $\sigma' \in [q']^n$ if for $i = 1, 2, \ldots, n$, we have $(\sigma[i], \sigma'[i]) \in E$. A subset $S \subseteq [q]^n$ is said to be a zero error ℓ-list-decoding code for the channel C if for all σ' in $[q']^n$,*

$$|\{\sigma \in S : \sigma \text{ and } \sigma' \text{ are compatible}\}| \leq \ell.$$

Let $n(m, C, \ell)$ be the minimum n such that there is a zero error ℓ-list-decoding code S for the channel C, such that $S \geq m$. The zero error list-of-ℓ rate of the code S is

$$R_{\mathcal{C},\ell}(S) = \frac{1}{n} \log \left(\frac{m}{\ell} \right),$$

and the zero error capacity of the channel $\mathcal{C}$ is the least upper bound of the attainable zero error list-of-ℓ rates of all codes.

For the $3/2$ channel Elias [E88] proved that the zero error capacity when $\ell = 2$ is lower bounded by $\log(3) - 1.5 \approx 0.08$ and upper bounded (see (2) below) by $\log(3) - 1 \approx 0.58$. In this paper, we study generalization's of the $3/2$ channel. The $q/(q-1)$ channel corresponds to the complete bipartite graph $K_{q,q}$ minus a perfect matching. Thus, the transmission of any letter can result in all but one of the letters being received. It is easy to see that that it is not possible to design a code where one can always recover the original message exactly. However, it is possible to design codes that perform list-decoding with lists of size $q-1$. In fact a routine probabilistic argument shows the following.

Proposition 1. $n(m, q, q-1) = \exp(O(q)) \log m.$

The $q/(q-1)$ channel is thus a natural and simple channel where exact decoding is not possible, but list-decoding with moderate size lists is possible. The main point of interest for us is that the rate of the code promised by Proposition 1 is exponentially small as a function of q. Is this exponentially small rate the best we can hope for if the list size is restricted to be $q-1$? Yes, and this follows from a lower bound on the size of families of perfect hash functions shown by Fredman and Komlós [FK84]. A generalization of the result of Fredman and Komlós obtained by Körner [K86], implies that the rate is exponentially small even if we allow the decoder to produce lists of size q. For what list size, then, can we expect list-decoding codes with constant or inverse polynomial rate?

Proposition 2. *For all q we have, $n(m, q, \lceil q \ln q \rceil) = O(q \log m)$.*

On the other hand, it can be shown that the rate cannot be better than $\frac{1}{q}$ unless the list size is allowed to depend on m.

Proposition 3. *All functions $\ell : \mathbb{Z}^+ \to \mathbb{Z}^+$ and all q, for all large enough m, $n(m, q, \ell(q)) \geq q \log m$.*

Thus, we know that the rate is exponentially small when the list size is required to be exactly q, and it is an inverse polynomial when the list size is $\theta(q \ln q)$. These observations, however, do not completely determine the dependence of the rate on the list size, or even the smallest list size (as a function of q) for which there are codes with rate significantly better than an inverse exponential. We conjecture the following.

Conjecture 1 *The conjecture has two parts.*

1. *For all constants $c > 0$, there is a constant ϵ, such that for all large m, we have*
$$n(m, q, cq) \geq \exp(\epsilon q) \log m$$

2. *For all function $\ell(q) = o(q \log q)$ and for all large m we have*

$$n(m, q, \ell(q)) \geq q^{\omega(1)} \log m$$

In this paper, we make progress towards the first part of the conjecture.

Theorem 1 (Main result). *For $\epsilon > 0$, there is a $delta > 0$ such that for all large q and for all large enough m, we have $n(m, q, (\gamma - \epsilon)q)) \geq \exp(\delta q) \log m$, where $\gamma = e/(e-1) \approx 1.58$.*

2 Techniques

As stated above the inverse exponential upper bounds on the rate when the list size is $q-1$ or q follow from results proved earlier in connection with hashing. In this section, we formally state this connection, review the previous techniques, and then outline the argument we use to obtain our result.

2.1 Connection to hashing

Definition 2. *Let q, ℓ, m be integers such that $1 \leq q \leq \ell \leq m$. A family $\mathcal{H}$ of functions from $[m]$ to $[q]$ is said to be an (m, q, ℓ)-family of hash functions if for all ℓ-sized subsets S of $[m]$, there is a function $h \in \mathcal{H}$ such that $h(S) = [q]$. $N(m, q, \ell)$ is the size of the smallest (m, q, ℓ)-family of hash functions. For convenience, we will allow m, q and ℓ to positive real numbers, and use $N(m, q, \ell)$ to mean the size of the smallest (m', q', ℓ')-family where m', q' and ℓ' are integers such that $m' \geq m$, $q' \geq q$ and $\ell' \leq \ell$.*

The connection between the family of hash functions and zero error list-decoding codes for the $q/(q-1)$ channel is straightforward. Suppose we have an (m, q, ℓ)-code $\mathcal{C} \subseteq [q]^n$. Such a code naturally gives rise to n functions $h_1, h_2, \ldots, h_n$ from $[m]$ to $[q]$: where $h_i(j) = k$ iff the i-th letter of the j-th codeword is $k \in [q]$. It is then straightforward to verify that for every set $S \subseteq [m]$ of size $\ell + 1$, we have $h_i(S) = [q]$ for some $i \in \{1, 2, \ldots, n\}$. This translations works in the other direction as well: if there is an (m, q, ℓ)-family of hash functions of size n, then there is an $(m, q, \ell - 1)$-code $\mathcal{C} \subseteq [q]^n$.

Proposition 4. *For all m, q, and ℓ, we have $n(m, q, \ell) = N(m, q, \ell + 1)$.*

In light of the above, we will concentrate on showing lower bounds for $N(m, q, \ell)$. Our main result can then be reformulated as follows.

Theorem 2. *For all $\epsilon > 0$, there is a $\delta > 0$, such that for all large q, $\ell \leq \left(\frac{e}{e-1} - \epsilon\right) q$, and all large enough m,*

$$N(m, q, \ell) \geq \exp(\delta q) \log m.$$

It is easy to see that Theorem 1 follows immediately from this. In Section 4.2 we will formally prove this theorem. We now present an overview.

2.2 The lower bound argument

It will be helpful to review the proof of the lower bound shown by Fredman and Komlós [FK84].

Theorem 3. *For all large q and all large enough m, $N(m,q,q) \geq O(q^{-1/2}\exp(q))\log m$.*

First, we note two simple lower bounds on $N(m,q,q)$. First, any one hash function can perfectly hash at most $\left(\frac{m}{q}\right)^q$ sets of size q. So,

$$N(m,q,\ell) \geq \binom{m}{q}\left(\frac{m}{q}\right)^{-q} \approx \frac{1}{\sqrt{2\pi q}}\exp(q). \tag{1}$$

This bound has the required exponential dependence on q but not the logarithmic dependence on m. A different argument gives us a logarithmic dependence on m. If we restrict attention to all elements of $[m]$ that are mapped by the first hash function to some $q-1$ of the $[q]$ elements of $[m]$, then clearly every q-sized subset of these elements must be perfectly hashed by at least one of the remaining hash functions. From this, we conclude that $N(m,q,q) \geq N\left((\frac{q-1}{q})m,q,q\right)$ (provided $m \geq q$), which implies

$$N(m,q,q) \geq \log_{\frac{q}{q-1}}\left(\frac{m}{q-1}\right) \geq q\log\left(\frac{m}{q-1}\right). \tag{2}$$

[A similar calculation can be used to justify Proposition 3.] This bound, gives us the required logarithmic dependence on m but not the exponential dependence on q. Fredman and Komlós devised an ingenious argument that combined the merits of (1) and (2). Consider a set T of size $q-2$. Clearly, if a function maps two of the elements of T to the same value in $[q]$, then this hash function is incapable of perfectly hashing any q-element superset $T' \supseteq T$. An averaging argument shows that for if T is chosen uniformly at random then all but an exponentially small fraction of the original family do map some two elements of T to the same element. Furthermore, for every two elements of $[m]-T$ one of the remaining hash functions (that are on-to-one on T) must map these two elements differently. By (2), the number of hash functions remaining must be at least $\log(m-q+2)$. Thus, the size of original family must be at $\exp(\Omega(q))\log(m-q+2)$. (The arguments used by Fredman and Komlós and Körner are more sophisticated and yield slightly better bounds.)

Our argument is similar. Suppose we have an (m, q, ℓ) family of hash functions where $\ell = 1 + \frac{1}{e} - \epsilon$ (for some $\epsilon > 0$). As in the argument above, we pick a set T of size $q-2$. The main observation now is that for any fixed function $h : [m] \to [q]$, the expected size $h(T)$ (as T is chosen at random) is about $q\left(1 - \frac{1}{e}\right)$, and there is a sharp concentration of measure near this mean value. Thus, only for an exponentially small fraction of the hash functions in the family is the image of T at least $q\left(1 - \frac{1}{e} + \epsilon\right)$. The majority of the functions already suffer so many collisions on T, that they cannot cover all of $[q]$ when an additional $\left(\frac{1}{e} - \epsilon\right) q$ elements are added to T. Using an argument similar to the one used to show (2), we conclude that an exponentially small fraction of the original family must be at least $\log(m - q + 2)$. The lower bound will follow from this. This argument is presented in detail in Section 4.1. It however is somewhat weaker than the bound claimed in Theorem 2. The stronger result is obtained by applying this idea recursively. The formal proof proceeds by induction, and is presented in Section 4.2.

3 Preliminaries

In this section we develop the tools that will be necessary in the proof of Theorem 2 in Section 4.1 and 4.2.

Definition 3 (Derived function). *Let m, q , q' be integers such that $m \geq q > q' \geq 1$ and let $T \subseteq [m]$ Let $h : [m] \to [q]$ be a hash function such that $|h(T)| \leq q - q'$. Then, the function $h_{T,q'}$ is defined as follows. Let $j_1, j_2, \ldots, j_{q'}$ be the smallest q' elements of $[q] - h(T)$. Then, for all $i \in [m]$, let*

$$h_{T,q'}(i) = \begin{cases} k \text{ if } h(i) = j_k \\ 1 \text{ otherwise} \end{cases} .$$

The following proposition follows immediately from our definition.

Proposition 5. *Let $h : [m] \to [q]$. If $T, T' \subseteq [m]$ are such that $|h(T)| \leq q - q'$ and $h(T \cup T') = [q]$, then $h_{T,q'}(T') = [q']$.*

Lemma 1. *If $\mathcal{H}$ is a family of hash functions from $[m]$ to $[2]$. Then, there is a subset $U \subseteq [m]$ of size at least $m/2^{|\mathcal{H}|}$ such that $|h(U)| = 1$, for all $h \in \mathcal{H}$.*

Proof. Consider the map from $[m]$ to $\{1, 2\}^{|\mathcal{H}|}$ defined by $i \mapsto \langle h(i) : h \in \mathcal{H} \rangle$. The range of this map has size exactly $2^{|\mathcal{H}|}$. It follows that there are at least $m/2^{|\mathcal{H}|}$ elements of the domain $[m]$ that map to the same element. $\square$

Definition 4 (Dangerous function). *We say that the function $h : [m] \to [q]$ is ϵ-dangerous for the set $T \subseteq [m]$ if $|h(T)| \geq q\left(1 - \frac{1}{e} + \epsilon\right)$.*

Lemma 2. *Let $h : [m] \to [q]$. Let T be a random subset of $[m]$ chosen uniformly from among all subsets of $[m]$ of size $q-2$. Then, if $m \gg q$,*

$$\Pr_T[h \text{ is } \epsilon\text{-dangerous for } T] \leq 2\exp(-2\epsilon^2 q).$$

To prove Lemma 2, we will need the following concentration result due to McDiarmid.

Lemma 3 (see McDiarmid [M89]). *Let $X_1, X_2, \ldots, X_n$ be independent random variables with each X_k taking values in a finite set A and let $f : A^n \to \mathbb{R}$. For all k, let f change by at most c_k if only the value of X_k is changed, that is, $max_{x \in A^k} |f(x) - f(y)| \leq c_k$, when x and y differ only in the kth coordinate. If $Y = f(X_1, X_2, \ldots, X_n)$ is the random variable with expectation $\mathbf{E}[Y]$, then for any $t \geq 0$,*

$$\Pr[Y - \mathbf{E}[Y] \geq t] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_k^2}\right).$$

Proof (of Lemma 2). Pick $q-2$ elements from $[m]$ with replacement, let the resulting set be T. With probability more than $\left(1 - \frac{\binom{q-2}{2}}{m}\right)$ we have that $|T| = q-2$. Now, fix an $h \in \mathcal{H}$. For $j \in [q]$, the probability that $j \notin h(T)$ is exactly, $\left(1 - \frac{|h^{-1}(j)|}{m}\right)^{q-2}$. Thus, by linearity of expectation, we have

$$\begin{aligned}
\mathbf{E}[|[q] - h(T)|] &= \sum_{j=1}^{q} \left(1 - \frac{|h^{-1}(j)|}{m}\right)^{q-2} \\
&\geq q\left(1 - \frac{1}{qm}\sum_{j=1}^{q} |h^{-1}(j)|\right)^{q-2} \\
&= q\left(1 - \frac{1}{q}\right)^{q-2} \\
&= q\left(1 + \frac{1}{q-1}\right)^{-(q-2)} \\
&\geq q\exp\left(-\frac{q-2}{q-1}\right) \\
&\geq \frac{q}{e}.
\end{aligned}$$

The second inequality follows from Holder's Inequality. Thus, $\mathbf{E}[|h(T)|] \leq q(1 - \frac{1}{e})$. We think of $|h(T)|$ as a function of $f(X_1, X_2, \ldots, X_q)$, of $q-2$

independent random variables $X_1, X_2, \ldots, X_{q-2}$ (each distributed uniformly over the set $[m]$).

Note that f changes at most by 1 if the value of any of the the variables is changed (leaving the rest unchanged). We may thus conclude from Lemma 3 that

$$\Pr\left[|h(T)| \geq q\left(1 - \frac{1}{e} + \epsilon\right)\right] \leq \exp\left(-2\frac{\epsilon^2 q^2}{q-2}\right) \leq \exp\left(-2\epsilon^2 q\right).$$

This implies that if T is chosen to be a random set of size $q-2$, then the probability that h is dangerous for T is at most

$$\left(1 - \frac{\binom{q-2}{2}}{m}\right)^{-1} \exp\left(-2\epsilon^2 q\right) \leq 2\exp\left(-2\epsilon^2 q\right).$$

□

Corollary 1. *Let $\mathcal{H}$ be a family of hash functions from $[m]$ to $[q]$. Then, there is a set $T \subseteq [m]$ of size $q-2$ such that at most $2\exp(-2\epsilon^2 q)|\mathcal{H}|$ hash functions in $\mathcal{H}$ are ϵ-dangerous for T.*

Proof. Pick T at random. By Lemma 2, the expected number of ϵ-dangerous hash functions for T is at most $2\exp(-2\epsilon^2 q)|\mathcal{H}|$. There must be a at least one choice for T with this property. □

4 Proof of Theorem 2

4.1 A weaker bound

Our goal in this section is to show the following weaker form of the main theorem, which will serve as the basis for the inductive argument, when we present the proof of the main result.

Theorem 4. *For $\epsilon > 0$, large q and all large enough m, we have $N(m, q, (\gamma - \epsilon)q) \geq \exp(\delta q)\log m$, where $\gamma = 1 + \frac{1}{e} \approx 1.37$.*

Proof. Let $\mathcal{H}$ be an (m, q, ℓ)-family with $\ell \leq (\gamma - \epsilon)q$. By Corollary 1, we have a set T of size $q-2$ such that the number of functions that are ϵ-dangerous for T is at most $2\exp(-2\epsilon^2 q)|\mathcal{H}|$. Fix such a T and consider the derived family

$$\mathcal{H}' = \{h_{T,2} : h \in H \text{ is } \epsilon\text{-dangerous for } T\}.$$

By Lemma 1, there is a set $U \subseteq [m]$ of size $m/2^{|\mathcal{H}'|}$ such that $|h'(U)| = 1$ for all $h' \in \mathcal{H}'$. We claim that $|U| < \left\lceil q\left(\frac{1}{e} - \epsilon\right)\right\rceil$. For, otherwise let T' be a subset of U of size $q' = \left\lceil q\left(\frac{1}{e} - \epsilon\right)\right\rceil$, and consider the set $T \cup T'$. If $h \in \mathcal{H}$ is not dangerous then $|h(T)| < q - q'$ and, therefore, $|h(T \cup T')| < q$. On the other hand if h is ϵ-dangerous for T, then our definition of U ensures that $|h(T')| = 1$ and, therefore, $|h(T \cup T')| \leq |T| + 1 < q$. Thus,

$$\frac{m}{2^{|\mathcal{H}'|}} < \left\lceil q\left(\frac{1}{e} - \epsilon\right)\right\rceil .$$

This, together with $|\mathcal{H}'| \leq 2\exp(-2\epsilon^2 q)|\mathcal{H}|$ implies our claim. □

4.2 The general bound

In this section, we will prove the Theorem 2. It will be convenient to restate it in a form suitable for an inductive proof. For $k \geq 1$ and $\epsilon > 0$, let

$$\ell_k(q, \epsilon) \;=\; q\left(1 + \frac{1}{e} + \frac{1}{e^2} + \cdots + \frac{1}{e^k} - \epsilon\right) - 2k.$$

Theorem 5 (Version of main theorem). *For all $k \geq 1$, $\epsilon > 0$, $q \geq 2k$ and all large enough m,*

$$N(m, q, \ell_k(q, \epsilon)) \geq \frac{1}{4k}\exp\left(\frac{2\epsilon^2 q}{e^{2k}}\right)\log m.$$

Proof. We will use induction on k. The base case $k = 1$, follows from the Theorem 4 proved in the previous section.

Induction step: Suppose the claim is false, that is, there is an (m_k, q_k, ℓ_k)-family $\mathcal{H}_k$ such that $\ell_k \leq \ell_k(q_k, \epsilon)$ and

$$|\mathcal{H}_k| < \frac{1}{4k}\exp\left(\frac{2\epsilon^2 q_k}{e^{2k}}\right)\log m. \tag{3}$$

[We use k in the subscript for parameters associated with the $\mathcal{H}_k$ to make emphasize the correspondence with the parameter k used in the induction.] From $\mathcal{H}_k$ we will derive an $(m_{k-1}, q_{k-1}, \ell_{k-1})$-family $\mathcal{H}_{k-1}$ such that

$$|\mathcal{H}_{k-1}| \leq |\mathcal{H}_k|; \tag{4}$$

$$m_{k-1} \geq m_k^{1-\frac{1}{k}}; \tag{5}$$

$$q_{k-1} \geq q_k\left(\frac{1}{e} - \frac{\epsilon}{4}\right) \geq \frac{q_k}{\epsilon^2}; \tag{6}$$

$$\ell_{k-1} \leq \ell_{k-1}(q_{k-1}, \epsilon). \tag{7}$$

Then, using the induction hypothesis, we obtain

$$
\begin{aligned}
|\mathcal{H}_k| \geq |\mathcal{H}_{k-1}| &\geq \frac{1}{4(k-1)} \exp\left(\frac{2\epsilon^2 q_{k-1}}{e^{2(k-1)}}\right) \log m_{k-1} \quad \text{(by the induction hypothesis)} \\
&\geq \frac{1}{4(k-1)} \exp\left(\frac{2\epsilon^2 q_k}{e^{2k}}\right) \left(1 - \frac{1}{k}\right) \log m_k; \\
&= \frac{1}{4k} \exp\left(\frac{2\epsilon^2 q_k}{e^{2k}}\right) \log m_k,
\end{aligned}
$$

contradicting (3).

It remains to describe how $\mathcal{H}_{k-1}$ is obtained from $\mathcal{H}_k$. The idea, as outlined in the introduction, is this. In the hope of using induction, we will first pick a subset T of size $q_k - 2$, which most hash functions map into a small number of elements. These functions can now be viewed as mapping $[m_k] - T$ to $[q_k]$, so that the problem reduces to one of covering a large subset of $[q_k]$ with $\ell_k - q_k + 2$. However, not all functions are guaranteed to be so well-behaved. For the few functions that do perform well on T, we need to take evasive action, by restricting attention to a subset of the universe on which these functions are guaranteed to be fail.

This idea is implemented as follows. Using Corollary 1, we first obtain a set $T \subseteq [m_k]$ of size $q_k - 2$ such that at most

$$
2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 q_k\right)|\mathcal{H}_k| \leq 2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 q_k\right)\cdot\frac{1}{4k}\exp\left(\frac{2\epsilon^2 q_k}{e^{2k}}\right)\log m \leq \frac{1}{2k}\log m_k
$$

hash functions in $\mathcal{H}_k$ are $\left(\frac{\epsilon}{4}\right)$-dangerous for T. Now consider the family of derived functions

$$
\mathcal{H}' = \{h_{T,2} : h \in \mathcal{H}_k \text{ is } \left(\tfrac{\epsilon}{4}\right)\text{-dangerous for } T\}.
$$

Using Lemma 1, we obtain a set $U \subseteq [m_k]$ of size at least $(m_k - q_k + 2)m_k^{-\frac{1}{2k}}$ such that $|h(U)| = 1$ for all $h \in \mathcal{H}'$. Our family $\mathcal{H}_{k-1}$ will be the following set of hash functions from U to $[q_{k-1}]$ (where $q_{k-1} = \lceil q_k(\frac{1}{e} - \frac{\epsilon}{4})\rceil$).

$$
\mathcal{H}_{k-1} = \{h_{T,q_{k-1}} : h \in \mathcal{H}_k \text{ is } \underline{\text{not}} \left(\tfrac{\epsilon}{4}\right)\text{-dangerous for } T\}.
$$

We claim that for all $T' \subseteq U$ of size $\ell_k - (q_k - 2)$, there is a function $h \in \mathcal{H}_{k-1}$ such that $h(T') = [q_{k-1}]$. For, consider the set $T \cup T'$ of size ℓ_k. By the definition of $\mathcal{H}_k$ there is an $h \in \mathcal{H}_k$ such that $h(T \cup T') = [q_k]$. Such an h is not $\left(\frac{\epsilon}{4}\right)$-dangerous for T because our definition of U ensures that $|h(T \cup T')| < q_k$. So for such an h we have

$$
|h(T)| < q_k\left(1 - \frac{1}{e} + \frac{\epsilon}{4}\right) \leq q_k - \left\lceil q_k(\frac{1}{e} - \frac{\epsilon}{4})\right\rceil = q_k - q_{k-1}.
$$

Hence, for such an h, by Proposition 5, we have $h_{T,q_{k-1}}(T') = [q_{k-1}]$.

Thus, $\mathcal{H}_{k-1}$ is an $(m_{k-1}, q_{k-1}, \ell_{k-1})$-family for $\ell_{k-1} = \ell_k - (q_k - 2)$. In particular $l_{k-1} \geq q_{k-1}$. We need to verify (4)–(7). The definition of $|\mathcal{H}_{k-1}|$ immediately implies (4). To verify (5) note that for $m_k \gg q_k$,

$$|U| \geq (m_k - q_k + 2)m_k^{-\frac{1}{2k}} \geq m_k^{1-\frac{1}{k}}.$$

Since $q_{k-1} = \lceil q_k(\frac{1}{e} - \frac{\epsilon}{4}) \rceil$, (6) holds. Finally, to justify (7), note that

$$\begin{aligned}
\ell_{k-1}(q_{k-1}, \epsilon) &\geq q_k \left(\frac{1}{e} - \frac{\epsilon}{4}\right)\left(1 + \frac{1}{e} + \cdots + \frac{1}{e^{k-1}} - \epsilon\right) - 2(k-1) \\
&\geq q_k \left[\frac{1}{e} + \frac{1}{e^2} + \cdots + \frac{1}{e^k} - \frac{\epsilon}{e} - \epsilon\left(1 + \frac{1}{e} + \cdots + \frac{1}{e^{k-1}}\right)\right] - 2(k-1) \\
&\geq q_k \left(\frac{1}{e} + \frac{1}{e^2} + \cdots + \frac{1}{e^k} - \epsilon\right) - 2(k-1) \\
&\geq q_k \left(1 + \frac{1}{e} + \cdots + \frac{1}{e^{k-1}} - \epsilon\right) - 2k - q_k + 2 \\
&\geq \ell_k(q_k, \epsilon) - (q_k - 2) \\
&\geq \ell_k - (q_k - 2) \\
&= \ell_{k-1}.
\end{aligned}$$

□

Acknowledgement

We thank Venkat Guruswami for introducing us to this problem.

References

E88. P. Elias. Zero Error Capacity Under List Decoding, IEEE Transactions on Information Theory, vol. 34, No. 5, (1988): 1070-1074.

FK84. M. Fredman, J. Komlós. On the Size of Separating Systems and Families of Perfect Hash Functions, SIAM J. Alg. and Disc. Meth., Vol. 5, No. 1 (1984): 61-68.

K86. J. Körner. Fredman-Komlós bounds and information theory, SIAM J. Algebraic and Discrete Methods, 7 (1986): 560-570.

L79. L. Lovász. On the Shannon Capacity of a Graph, IEEE Trans. Inform. Theory, Vol. IT-25 (1979): 1-7.

M89. C. McDiarmid. On the method of bounded differences, Surveys in Combinatorics, vol 141 LMS Lecture Notes Series (1989): 148–188.

S56. C.E. Shannon. The zero error capacity of a noisy channel, IEEE Trans. Inform. Theory, Vol. IT-2, no. 3, (1956): 8-19. (Reprinted in D. Slepian, Ed., Key Papers in the Development of Information Theory. New York: IEEE Press (1974): 112-123)