Note on the size of binary Armstrong codes

Aart Blokhuis*, Andries Brouwer, Attila Sali†

April 19, 2012

Abstract

We show for binary Armstrong codes $\mathrm{Arm}(2,k,n)$ that asymptotically $n/k \le 1.224$, while such a code is shown to exist whenever $n/k \le 1.12$. We also construct an $\mathrm{Arm}(2,n-2,n)$ and $\mathrm{Arm}(2,n-3,n)$ for all admissible n.

Keywords: coding theory; databases; Armstrong codes
AMS subject classification: 94B60; 94B65.

1 Introduction

An Armstrong code $\mathrm{Arm}(q,k,n)$ is a code of length n over an alphabet of size q with minimum Hamming distance $d = n-k+1$ and the additional property that for every subset of size $k-1 = n-d$ of the coordinate positions there are two codewords that agree there (so the minimum distance occurs 'in all directions').‡ For example, the code consisting of the rows of an n by n identity matrix is an $\mathrm{Arm}(q,n-1,n)$ and the code of the $n+1$ vectors $\mathbf{c}_i = (1,\dots,1,0,\dots,0)$ with i ones followed by $n-i$ zeroes is an $\mathrm{Arm}(q,n,n)$ for all q.

Armstrong codes have their origin in Database Theory, see for instance [8]. The main questions of this note were introduced in [6] and investigated in the papers [1, 7].

In this note we take $q = 2$, and give necessary and sufficient conditions for the existence of an $\mathrm{Arm}(2,k,n)$.

2 Armstrong codes $\mathrm{Arm}(2,k,n)$ for $k \ge n-3$

We have seen above that an $\mathrm{Arm}(2,n,n)$ and $\mathrm{Arm}(2,n-1,n)$ exists for all $n > 0$.

Proposition 1 *An* $\mathrm{Arm}(2,n-2,n)$ *exists if and only if* $n \ge 9$. *An* $\mathrm{Arm}(2,n-3,n)$ *exists if and only if* $n \ge 10$.

*Partially supported by ERC Advanced Research Grant no 267165 (DISCONV)

†Partially supported by F.I.S.T. Marie Curie Host Fellowship for the Transfer of Knowledge, and OTKA grant NK 78439

‡The referee asks us to stress that the code is not necessarily linear, and that k does not denote the dimension of the code. Instead, k has the stated meaning.

Proof. By deleting one coordinate position in an $\mathrm{Arm}(q,k,n)$, one obtains an $\mathrm{Arm}(q,k,n-1)$. Consequently, the existence of an $\mathrm{Arm}(2,n-2,n)$ for $n \geq 9$ follows from that of an $\mathrm{Arm}(2,n-3,n)$ for $n \geq 10$.

A. Keszler showed in her diploma thesis [2] using computer that no $\mathrm{Arm}(2,n-2,n)$ exists for $n \leq 8$. It follows that no $\mathrm{Arm}(2,n-3,n)$ exists for $n \leq 9$.

An $\mathrm{Arm}(2,7,10)$ can be constructed by taking a Steiner system $S(3,4,10)$[§] and adding the all-0 vector.

It remains to construct an $\mathrm{Arm}(2,n-3,n)$ for $n \geq 11$. Let an (m,M,d)-code be a binary code of word length m, size M, and minimum distance (at least) d.

First assume that $n \geq 23$. We construct an $(n,n,12)$-code. From a Hadamard matrix of order $4t$, where $n \leq 4t \leq 2n-22$, we obtain a $(4t,8t,2t)$-code. If $n = 4t$, then discard $4t$ words. If $n < 4t$, then shorten the code once (this yields a $(4t-1,4t,2t)$-code), discard $4t-n$ code words, and $4t-1-n$ coordinate positions. In both cases we find an $(n,n,12)$-code.

Partition the quadruples from an n-set into n collections such that two quadruples in the same collection intersect in at most 2 elements by putting quadruple $\{p,q,r,s\}$ in collection $\mathcal{T}_i$ if $p+q+r+s \equiv i \pmod{n}$. Let $C = \{\mathbf{c}_0,\ldots,\mathbf{c}_{n-1}\}$ be an $(n,n,12)$-code. Construct an $\mathrm{Arm}(2,n-3,n)$ by taking the code words in C together with the words $\mathbf{c}_i + \mathbf{t}$ for every $T \in \mathcal{T}_i$, where $\mathbf{t}$ is the characteristic vector of T.

For $14 \leq n \leq 16$, look at the 2165 extended perfect $(16,2048,4)$-codes (classified in [5]). One finds that five of these (numbers 2099, 2108, 2121, 2122 and 2124) are Armstrong. Appropriate shortenings give Armstrong codes for $n = 15$ and $n = 14$ (but not for $n = 13$).

For $14 \leq n \leq 22$ Armstrong codes can be obtained by computer, using a greedy procedure: Start by putting the zero word in the code. Then enumerate all binary words in lexicographic order, adding a word to the code obtained so far when it has the required minimum distance, and it provides at least one difference that did not occur earlier. For $n = 11, 12, 13$, a randomized version of this greedy procedure works. □

3 A lower bound

For general k we have the following. Recall that $d = n-k+1$.

Theorem 2 ([1], Theorem 2.2) *An* $\mathrm{Arm}(2,k,n)$ *exists when* $n \geq 9.09d$. *An* $\mathrm{Arm}(2,k,n)$ *exists when* $n \leq 1.12k$.

Proof. The second claim follows from the first one. Katona et al. [1] show (in formula (9)) that $\mathrm{Arm}(2,k,n)$ exists when $d\binom{n}{d}^2 \leq 2^{n-2}$. And this holds when $d \geq 1$ and $n \geq ad$ with $a \geq 9.08861$. □

4 Upper bounds

In [1], Theorem 3.3, it is shown that if an $\mathrm{Arm}(2,k,n)$ exists, and $k \geq 7$, then $n \leq 2(k-1)$ (that is, $n \geq 2d$). Here we asymptotically improve the constant 2 to $\frac{5}{4}$ (so that $n \geq 5d$ when d is large).

[§](also known as a Steiner quadruple system $SQS(10)$)

Proposition 3 *Let $A(n,d)$ and $A(n,d,w)$ denote the maximum size of a binary code of word length n, minimum distance d (and constant weight w). Suppose an* $\mathrm{Arm}(2,k,n)$ *exists. Then* $2\binom{n}{d} \le A(n,d)A(n,d,d)$.

Proof. If $\mathcal{C}$ is an $\mathrm{Arm}(2,k,n)$ and we look at all spheres of radius d around code words, then we see each difference at least twice. □

Write $L(x) = x\log_2(x)$. Below we will use the following standard estimate for binomial coefficients. It follows from Stirling's theorem, and is valid for m sufficiently large, β, γ and $\beta-\gamma$ bounded away from zero, small compared to m, but not necessarily constant. $\frac{1}{m}\log_2\binom{\beta m}{\gamma m} \approx L(\beta) - L(\gamma) - L(\beta-\gamma)$. With the binary entropy function $H_2(x) = -L(x) - L(1-x)$, we have $\frac{1}{n}\log_2\binom{n}{\alpha n} \approx H_2(\alpha)$.

Let $d = \delta n$. Let $\kappa_0 = \kappa_0(\delta)$ be such that a code of length n with constant weight d and minimum distance d has size at most $2^{\kappa_0 n}$. Let $\kappa_1 = \kappa_1(\delta)$ be such that an arbitrary code with length n and minimum distance d has size at most $2^{\kappa_1 n}$.

Proposition 3 says that if an $\mathrm{Arm}(2,k,n)$ exists, then $2\binom{n}{d} \le 2^{(\kappa_0+\kappa_1)n}$. Hence $H_2(\delta) \le \kappa_0(\delta) + \kappa_1(\delta)$. Various bounds on $\kappa_0(\delta)$ and $\kappa_1(\delta)$ now give upper bounds for n/k for Armstrong codes.

Theorem 4 *If an Armstrong code* $\mathrm{Arm}(2,k,n)$ *exists, then we have asymptotically* $n \le 1.224k$.

Proof. The sphere packing bound (really, ball packing bound) gives an upper bound $\kappa_1 = 1 - H_2(\delta/2)$. Let C be a code of word length n, constant weight d, and minimum distance d. Let $m = \lfloor d/2 \rfloor$. Then $|C| \le \binom{n}{m+1}/\binom{d}{m+1}$, because every $(m+1)$-set of coordinates is covered by a code word from C at most once. It follows that we can take $\kappa_0 = L(\frac{1}{2}\delta) - L(\delta) - L(1-\frac{1}{2}\delta)$. Solving $H_2(\delta) \le \kappa_0(\delta) + \kappa_1(\delta)$ yields $\delta \le 0.2271$, so that $n \le 1.294k$.

The Elias-Bassalygo bound gives $\kappa_1 = 1 - H_2((1-\sqrt{1-2\delta})/2)$, better than the sphere packing bound. This time we find $\delta \le 0.212$, so that $n \le 1.27k$.

A weak form of the McEliece-Rodemich-Rumsey-Welch bound ([4], (1.5)) allows us to take $\kappa_1 = H_2(\frac{1}{2} - \sqrt{\delta(1-\delta)})$. This is better again (for $\delta > 0.15$), and yields $\delta \le 0.205$, so that $n \le 1.258k$.

An improved value for κ_0 (see [3], p. 643) is

$$\kappa_0 = H_2\left(\frac{1}{2} - \sqrt{\frac{1}{4} - \left(\sqrt{\delta(1-\delta) - \frac{\delta}{2}(1-\frac{\delta}{2})} - \frac{\delta}{2}\right)^2}\right).$$

Using it yields $\delta \le 0.18506$ and hence $n \le 1.2271k$.

A stronger form of the McEliece-Rodemich-Rumsey-Welch bound ([4], (1.4)) has $\kappa_1 = \min\{1 + g(u^2) - g(u^2 + 2\delta u + 2\delta) \mid 0 \le u \le 1 - 2\delta\}$, where $g(x) = H_2((1-\sqrt{1-x})/2)$. With $u = 0.25$ this says $\kappa_1 = 1 + g(\frac{1}{16}) - g(\frac{1}{16} + \frac{5\delta}{2})$. This yields $\delta \le 0.183$ and hence $n \le 1.224k$. □

References

[1] Katona, G. O. H., Sali, A., and Schewe, K.-D., Codes that attain minimum distance in all possible directions. *Central European J. of Math. 6*, 1–11 (2008).

[2] Keszler, A., Adatbázisok Extremális Kombinatorikai Problémái, *Diploma thesis Budapest University of Technology and Economics*, (2008) .

[3] Levenshtein, V. I., Universal bounds for codes and designs, pp. 499–648 in: Handbook of Coding Theory, V. S. Pless and W. C. Huffman, eds., Elsevier, Amsterdam, 1998.

[4] McEliece, R. J., Rodemich, E. R., Rumsey Jr., H. C., and Welch, L. R., New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities, *IEEE Trans. Inf. Th. 23*, 157–166 (1977).

[5] Östergård, P. R. J., and Pottonen, O., The perfect binary one-error-correcting codes of length 15: Part I—Classification. `arXiv:0806.2513`, Dec 2009.

[6] Sali, A., and Schewe, K.-D., Keys and Armstrong databases in trees with restructuring. *Acta Cybernetica 18*, 529–556 (2008).

[7] Sali, A., and Székely, L., On the Existence of Armstrong Instances with Bounded Domains, *Lecture Notes in Computer Science 4932*, 151–157 (2008).

[8] Sali, A., Coding Theory Motivated by Relational Databases. *Lecture Notes in Computer Science 6834*, 96–113 (2011).

Address of the authors:

A. Blokhuis and A. E. Brouwer,
Department of Mathematics and Comp. Science,
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven,
The Netherlands.

A. Sali,
Alfréd Rényi Institute of Mathematics
Hungarian Academy of Sciences
Budapest, P.O. Box 127, H-1364 Hungary

e-mail: `aartb@win.tue.nl`, `aeb@cwi.nl`, `sali.attila@renyi.mta.hu`