A trellis-based bound on $(2,1)$-separating codes

Hans Georg Schaathun[1]* and Gérard D. Cohen[2]

[1] Dept. Informatics,
University of Bergen
Pb. 7800
N-5020 Bergen
Norway
⟨georg@ii.uib.no⟩

[2] Dept. Informatique et Reseaux,
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault,
F-75634 Paris Cedex 13
France
⟨cohen@enst.fr⟩

Abstract We explore some links between higher weights of binary codes based on entropy/length profiles and the asymptotic rate of $(2,1)$-separating codes. These codes find applications in digital fingerprinting and broadcast encryption for example. We conjecture some bounds on the higher weights, whose proof would considerably strengthen the upper bound on the rate of $(2,1)$-separating codes.

Keywords: *Trellis, coding bounds, separating codes.*

1 The problem

The concept of (t,u)-separating codes has been studied for about 35 years in the literature, with applications including fault-tolerant systems, automata synthesis, and construction of hash functions. For a survey one may read [15]. The concept has been revived by the study of digital fingerprinting [3]; a $(t,1)$-separating code is the same as a t-frameproof code, on which we elaborate now.

In broadcast encryption, a company distributes a unique decoder to each user. Users may collude and combine their decoders to forge a new one. The company wants to limit this or trace back illegal decoders to the offending users. Among the forbidden moves: framing an innocent user. This goal can be achieved with frameproof codes. One can consult e.g. [2,16] for more.

In this paper we study binary $(2,1)$-separating codes ($(2,1)$-SS). An (n,M) code is a subset of size M from the set of binary vectors of length n. The code or a set of codewords will often be regarded as matrices, with the codewords forming the rows.

* Research was supported by the Norwegian Research Council under Grant Number 146874/420 and the AURORA program.

	Linear		Nonlinear	
	Rate	Ref.	Rate	Ref
Known construction	0.156	[6]	0.1845	[8]
New construction			0.2033	Theorem 2
Existence	0.2075	Well-known e.g. [6]	0.2075	Well-known e.g. [10]
Upper bound	0.28	Well-known e.g. [6]	0.5	[10]

Table 1. Rate bounds for $(2,1)$-SS.

Definition 1 *Let* $\mathbf{a},\mathbf{b},\mathbf{c}$ *be three vectors. We say that* $\mathbf{a}$ *is separated from* $(\mathbf{b},\mathbf{c})$ *if there is at least one postion* i *such that* $a_i \neq b_i$ *and* $a_i \neq c_i$*.*

An (n,M) *code is* $(2,1)$*-separating if for every ordered triplet* $(\mathbf{a},\mathbf{b},\mathbf{c})$*,* $\mathbf{a}$ *is separated from* $(\mathbf{b},\mathbf{c})$*.*

Many interesting mathematical problems are equivalent to that of $(2,1)$-SS. An overview of this is found in [10]. A linear $(2,1)$-separating code is equivalent to an intersecting code [6], and some results have been proved independently for intersecting and for separating codes. For further details on intersecting codes, see [5] and references therein.

The rate of the code is

$$R = \frac{\log M}{n}.$$

For an asymptotic family of codes (n_i, M_i) codes $(C_1, C_2, \ldots)$ where $M_i > M_{i-1}$ for all i, the rate is defined as

$$R = \limsup_{i\to\infty} \frac{\log M_i}{n_i}.$$

The known bounds on asymptotic families of $(2,1)$-SS are shown in Table 1. By abuse of language, an asymptotic family of codes will also be called an asymptotic code.

We observe a huge gap between the upper and lower bounds for non-linear codes. Our goal is to reduce this gap. The references in the table are given primarily for easy access and are not necessarily the first occurrences of the results, which are sometimes folklore.

Section 2 gives a minor result, namely a new construction slightly improving the lower bound. In Section 3, we make some observations about the trellis of a $(2,1)$-SS. In Section 4, we discuss higher weights of arbitrary codes and we introduce the 'tistance' of a code and make some conjecture. Section 5, we prove bounds for $(2,1)$-SS in terms of the 'tistance' and show how the conjectures would give major improvements of the upper bounds if proved.

2 A new asymptotic construction

Theorem 2 *The codes obtained by concatenating an arbitrary subcode of 121 words from the* $(15,2^7)$ *shortened Kerdock code* $K'(4)$ *with codes as described* *by Xing [18] ($t = 2$) over* $\mathsf{GF}(11^2)$*, is a family of* $(2,1)$*-SS of asymptotic rate* $R = 0.2033$.

Proof. It is well known that the concatenation of two $(2,1)$-SS is a $(2,1)$-SS. The shortened Kerdock code $K'(4)$ was proved to be a $(2,1)$-SS in [11]. The Xing codes were proved to be $(2,1)$-SS in [18]. Let us recall for convenience their parameters.

Suppose that $q = p^{2r}$ with p prime, and that t is an integer such that $2 \leq t \leq \sqrt{q} - 1$. Then there is an assymptotic family of $(t,1)$-separating codes with rate

$$R = \frac{1}{t} - \frac{1}{\sqrt{q}-1} + \frac{1 - 2\log_q t}{t(\sqrt{q}-1)}.$$

We take $K'(4)$ which is a $(15, 2^7)$ $(2,1)$-SS, wherefrom we pick 11^2 arbitrary codewords. This code can be concatenated with a Xing code over $\mathsf{GF}(11^2)$, for which a rate of approximately 0.4355 and minimum distance more than 0.5 is obtainable. This gives a concatenated code which is $(2,1)$-separating with the stated rate.

It is a bit unclear how easily we can construct the sequences of curves on which the Xing codes are based; we have found no explicit construction in the literature, but it is hinted that the construction should be feasible. The alternative is to use the random construction of [6,10] for a rate of 0.2075, but that is certainly computationally intractable even for moderate code sizes.

3 Trellises for $(2,1)$-SS

We know that a code can always be described as a trellis, and trellises have been studied a lot as devices for decoding. We will not rule out the possibility that someone will want to use trellis decoding of separating codes at some point, but that *is not* our concern. We want to derive bounds on the rate of $(2,1)$-separating codes, and it appears that such bounds may be derived by studying the trellis.

A trellis is a graph where the vertices are divided into $(n+1)$ classes called *times*. Every edge goes from a vertex at time i to a vertex at time $i+1$, and is labeled with an element from some alphabet Q. The vertices of a trellis are called *states*. A trellis also have the property that time 0 and time n each has exactly one vertex, called respectively the *initial* and the *final* states. There is at least one path from the initial state to each vertex of the graph, and at least one path to the final state from each vertex of the graph.

A binary (n, M) code corresponds to a trellis with label alphabet $\{0, 1\}$. Every path from time 0 to time n defines a codeword by the labels of the edges. Every codeword is defined by at least one such path.

A trellis is most often considered as an undirected graph, but we will nevertheless say that an edge between a state v at time $i-1$ to some state w at time i goes from v to w.

If each codeword corresponds to exactly one trellis path, then we say that the trellis is *one-to-one*. A *proper* trellis is one where two edges from the same

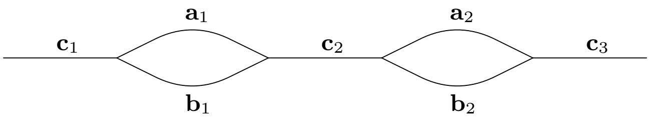

Figure 1. The impossible subtrellis in Proposition 3.

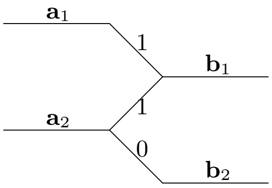

Figure 2. The impossible subtrellis in Proposition 5.

vertex never have the same label. If, in addition, no two edges into the same vertex have the same label, then we say that the trellis is *biproper*. It is known that every block code corresponds to a unique minimal proper trellis, i.e. the proper trellis with the minimal number of edges.

Proposition 3 *In a trellis corresponding to a* $(2,1)$*-separating code, if two distinct paths join in some vertex* v*, the joint path cannot rebranch at any later time.*

Proof. If the trellis were to contain two paths which first join and later rebranch, it would mean a sub-trellis as given in Figure 1. If so, we can consider the three vectors

$$\begin{aligned}\mathbf{v}_1 &= \mathbf{c}_1||\mathbf{a}_1||\mathbf{c}_2||\mathbf{a}_2||\mathbf{c}_3\\ \mathbf{v}_2 &= \mathbf{c}_1||\mathbf{a}_1||\mathbf{c}_2||\mathbf{b}_2||\mathbf{c}_3\\ \mathbf{v}_3 &= \mathbf{c}_1||\mathbf{b}_1||\mathbf{c}_2||\mathbf{a}_2||\mathbf{c}_3\end{aligned}$$

Now $\mathbf{v}_1$ is not separated from $(\mathbf{v}_2,\mathbf{v}_3)$, so the trellis cannot be $(2,1)$-separating.

Corrollary 4 *The trellis of a* $(2,1)$*-separating code cannot contain a vertex with both two incoming and two outgoing edges (often called a butterfly vertex).*

Proposition 5 *Every* $(2,1)$*-separating code has a biproper trellis.*

Proof. We consider the minimal proper trellis of a $(2,1)$-separating code. Let v be a vertex with two incoming edges with the same label, say 1. This must mean that we have a subtrellis like the one drawn in Figure 2, where $\mathbf{a}_1||0||\mathbf{b}_2$ is not a codeword. Observe the three codewords

$$\begin{aligned}\mathbf{v}_1 &= \mathbf{a}_1||1||\mathbf{b}_1\\ \mathbf{v}_2 &= \mathbf{a}_2||1||\mathbf{b}_1\\ \mathbf{v}_3 &= \mathbf{a}_2||0||\mathbf{b}_2.\end{aligned}$$

Here, $\mathbf{v}_2$ is not separated from $(\mathbf{v}_1, \mathbf{v}_3)$.

4 Entropy/Length Profiles and higher weights

This section deals with general codes, not necessarily separating. Higher weights, or generalised Hamming weights, have been studied for linear codes since 1977 [9] and have received considerable interest with the definition of the weight hierarchy in 1991 [17]. For non-linear codes, different definitions have been suggested [4,1,14]. We will primarily use the entropy/length profiles (ELP) from [13]. The ELP was used to define the weight hierarchy in [14].

Let X be a stochastic variable, representing a codeword drawn uniformly at random from some code C. Write $[n] = \{1, 2, \dots, n\}$. For any subset $I \subseteq [n]$, let X_I be the vector $(X_i : i \in I)$, where $X = (X_i : i \in [n])$. Clearly X_I is also a stochastic variable, but not necessarily uniformly distributed.

Definition 6 (Entropy) *The (binary) entropy of a discrete stochastic variable X drawn from a set $\mathcal{X}$ is defined as*

$$H(X) = -\sum_{x\in\mathcal{X}} P(X = x) \log P(X = x).$$

The conditional entropy of X with respect to another discrete stochastic variable Y from $\mathcal{Y}$ is

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} P(Y = y) \sum_{x\in\mathcal{X}} P(X = x|Y = y) \log P(X = x|Y = y).$$

We define the (unordered) conditional ELP to be the sequence $(h_i : i \in [n])$ where

$$h_i = \max_{\#I=i} H(X_I|X_{[n]\setminus I}),$$

we also have the ordered conditional ELP $(g_i : i \in [n])$ where

$$g_i = H(X_{[i]}|X_{\{i+1,\dots,n\}}).$$

Evidently, g_i depends on the coordinate ordering and may thus be different for two equivalent codes. On the other hand, h_i is the maximum of g_i for all equivalent codes, and thus invariant throughout an equivalence class.

The weight hierarchy as defined in [14] is $\{i|h_i > h_{i-1}\}$. It is an interesting point that the weight hierarchy of a linear code always has k elements, while there is no way to predict the number of elements in the weight hierarchy of a non-linear code, no matter which definition is used.

In this paper we use the parameters

$$t_j := \min\{i : g_i \geq j\}, \quad \text{where } j = 0, \dots, \lfloor \log M \rfloor .$$

We have particular interest in the first parameter, $t := t_1$, which we are going to call the 'tistance' of the code. For all codes $t \geq d$, and for linear codes we have $t = d$. We also define a normalised measure $\tau := t/n$.

Lemma 7 [14, Lemma 2] *For any $(n, M)_q$ code C, we have $h_l(C) \geq r$ where $r = l + log_q M - n$.*

Proposition 8 (Tistance Singleton bound) *For any $(n, M)_q$ code with tistance t, we have $t < n - log_q M + 2$.*

The proposition follows directly from Lemma 7, by setting $l = t$ and noting that $h_t < 2$. Note that if $M = q^k$ for some integer k, then $t \leq n - k + 1$, which is the more common Singleton form of the bound.

Corrollary 9 *For an asymptotic class of codes, we have $\tau \leq 1 - R$.*

Conjecture 10 (Plotkin-type bound) *For all asymptotic codes, we have $R \leq R_P(\tau) := 1 - 2\tau$.*

The regular asymptotic Plotkin bound states that $R \leq 1 - 2\delta$. Since $\tau \geq \delta$, the conjecture is stronger than this.

Conjecture 11 *Let $R_{LP}(\delta)$ be the MRRW bound [12]. For an asymptotical non-linear code, it holds that $R \leq R_{LP}(\tau)$.*

Obviously, Conjecture 11 implies Conjecture 10, because the MRRW bound is stronger than the Plotkin bound. We state the conjectures separately to encourage work on a Plotkin-type bound in terms of t. The usual Plotkin bound has a cleaner expression and a simpler proof than the MRRW bound, and thus Conjecture 11 may well be considerably harder to prove.

5 Trellis bounds on $(2, 1)$-SS

At time i, let $\Sigma_i = \{\sigma_1, \dots, \sigma_a\}$ be the set of states with more than one incoming path. For any state σ, let $P(\sigma)$ be the number of distinct incoming paths respectively. Remember from Proposition 3 that any state $\sigma \in \Sigma_i$ has only one outgoing path. We get that

$$g_i = \sum_{\sigma \in \Sigma_i} \frac{P(\sigma)}{M} h(P(\sigma)) = M^{-1} \sum_{\sigma \in \Sigma_t} P(\sigma) \log P(\sigma), \tag{1}$$

or equivalently that

$$M = g_i^{-1} \sum_{\sigma \in \Sigma_t} P(\sigma) \log P(\sigma). \tag{2}$$

Setting $i = t$, we get that

$$M \leq \sum_{\sigma \in \Sigma_t} P(\sigma) \log P(\sigma). \tag{3}$$

Theorem 12 *A* $(2,1)$*-separating code has size* $M \leq 2^t$.

Proof. Let Σ_i be the set of states at time i with multiple incoming paths as before, and let Σ_i' be the set of states at time $i \leq t$ with a unique incoming path, and a path leading to a state in Σ_t. Obviously, we have $\#\Sigma_i + \#\Sigma_i' \leq 2^i$. Also note that a state in Σ_i (for $i < t$) must have a (unique) outgoing path leading to a state in Σ_t. Observe that $\Sigma_t' = \emptyset$.

We will prove that for $i = 0, \ldots, t$, we have

$$M \leq 2^i \left(\sum_{\sigma \in \Sigma_{t-i}} P(\sigma) \log P(\sigma) + \#\Sigma_{t-i}' \right). \tag{4}$$

This holds for $i = 0$ by (3), so it is sufficient to show that

$$\sum_{\sigma \in \Sigma_i} P(\sigma) \log P(\sigma) + \#\Sigma_i' \leq 2 \left(\sum_{\sigma \in \Sigma_{i-1}} P(\sigma) \log P(\sigma) + \#\Sigma_{i-1}' \right). \tag{5}$$

for $0 < i \leq t$. Since $\Sigma_0 = \emptyset$ and Σ_0' is the singleton set containing the initial state, (4) implies $M \leq 2^t$ by inserting $i = t$, which will prove the theorem.

Each state σ in Σ_i must have paths from one or two states $\Sigma_{i-1} \cup \Sigma_{i-1}'$. If there is only one such state σ', then we have $P(\sigma) = P(\sigma')$ and $\sigma' \in \Sigma_{i-1}$.

If there are two such states σ_1 and σ_2, we get that $P(\sigma_1) + P(\sigma_2) = P(\sigma)$. If $P(\sigma_j) = 1$, we have $\sigma_j \in \Sigma_{i-1}'$, otherwise $\sigma_j \in \Sigma_{i-1}$. Observe that

$$P(\sigma) \log P(\sigma) \leq 2(P(\sigma_1) \log P(\sigma_1) + P(\sigma_2) \log P(\sigma_2)), \quad \text{if } \sigma_1, \sigma_2 \in \Sigma_{i-1}, \tag{6}$$

$$P(\sigma) \log P(\sigma) \leq 2P(\sigma_1) \log P(\sigma_1) + 1, \quad \text{if } \sigma_1 \in \Sigma_{i-1}, \sigma_2 \in \Sigma_{i-1}', \tag{7}$$

$$P(\sigma) \log P(\sigma) = 2, \quad \text{if } \sigma_1, \sigma_2 \in \Sigma_{i-1}'. \tag{8}$$

Each of these three equations describes one type of state $\sigma \in \Sigma_i$. Recall that $\sigma_j \in \Sigma_{i-1}$ can have but one outgoing edge. For any state $\sigma \in \Sigma_i'$ there is one state $\sigma' \in \Sigma_{i-1}'$, and each such state σ' has a path to one or two states in $\Sigma_i \cup \Sigma_i'$.

We note that each $\sigma \in \Sigma_i$ in (5) contributes to the right hand side with the maximum amount from the bounds (6) to (8). The term $\#\Sigma'_{i-1}$ is multiplied by two to reflect the fact that each $\sigma' \in \Sigma'_{i-1}$ can have an edge to two different states in $\Sigma_i \cup \Sigma'_i$. This proves the bound.

Proposition 13 *If Conjecture 10 is true, then any asymptotical $(2,1)$-SS has rate $R \leq 1/3$. Similarly, if Conjecture 11 is true, then any asymptotical $(2,1)$-SS has rate $R \leq 0.28$.*

The proof is is similar to the ones used to prove upper bounds on linear $(2,1)$-SS in past, see e.g. [15,6].

Proof. From the Plotkin-type bound on τ, we get $\tau \leq \frac{1}{2}(1-R)$, and from Theorem 12 we thus get $R \leq \frac{1}{2}(1-R)$ which proves the result. The proof of the second sentence is similar, replacing the Plotkin bound by the MRRW bound.

Remark 14 *Theorem 12 combined with the tistance Singleton bound, $R \leq 1-\tau$, implies that $R \leq 0.5$ for any $(2,1)$-SS by the proof above, providing a new proof for the old bound. Any stronger bound on R in terms of τ for non-linear codes, will improve the rate bound for $(2,1)$-separating codes.*

Remark 15 *By using (2), we get for any i that*

$$M \leq h_i^{-1} 2^i,$$

by a proof similar to that of Theorem 12.

6 Balance

From Table 1, we know that an asymptotic upper bound of the rate of a $(1,2)$-separating code is $R \leq 1/2$. Starting with an asymptotic family with rate close to $1/2$, we construct a family with the same rate and only codewords with weight close to $n/2$. Let C be a $(1,2)$-SS of rate $R = 1/2-\alpha$, where $\alpha > 0$ is a sufficiently small constant.

Consider a partition (P_1, P_2) of the coordinates with $|P_1| = \lceil (1/2 + 1.1\alpha)n \rceil =: n_1$. Let $U_i \subseteq C$ be the set of codewords matching no other codeword on P_i. It is easy to check that $C \subset U_1 \cup U_2$. (Otherwise, some codeword would be matched by at most two others on P_1 and P_2, thus not separated). Since $|U_2| \leq 2^{|P_2|} = o(|C|)$, we get $|U_1| = (1-o(1))|C|$.

Projecting C on P_1 gives a code $C_1(n_1, 2^{(1/2-\alpha)n}(1-o(1)))$ of rate $R_1 \approx (1/2-\alpha)/(1/2+1.1\alpha) \approx 1-4.2\alpha$. Thus, the relative dominating weight ω_1 in C_1 must be close to $1/2$.

Now, we expurgate by keeping only codewords of C which get relative weight ω_1 when projected on P_1. Thus we get a code C' with rate asymptotically equal to that of C.

We repeat the procedure with a new partition (P'_1, P'_2), almost disjoint from the previous one (i.e., we take $|P_1 \cap P'_1| = \lceil 2.2\alpha n \rceil$). The code C'' obtained after the second expurgation retains both $(1,2)$-separation and rate $\approx 1/2$. Its codewords, being balanced on P_1 and P_1', are 'almost' balanced, as the following theorem states.

Theorem 16 *For all $c'' \in C''$, we have $|w(c'')/n - 1/2| = o(1)$.*

Remark 17 *This result generalises easily to $(1,t)$-separation. Any such code with rate close to the optimal rate of $1/t$ is almost balanced.*

We have translated the old combinatorial question of separating codes into the language of trellises. This has enabled us to shed new light on the matter, by putting to use concepts like entropy and higher weights.

References

1. L. A. Bassalygo. Supports of a code. In G. Cohen, M. Giusti, and T. Mora, editors, *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, volume 948 of *Springer Lecture Notes in Computer Science*. Springer-Verlag, 1995.
2. Simon R. Blackburn. Frameproof codes. *SIAM J. Discrete Math.*, 16(3):499–510, 2003.
3. Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. Inform. Theory*, 44(5):1897–1905, 1998. Presented in part at CRYPTO'95.
4. Gérard Cohen, Simon Litsyn, and Gilles Zémor. Upper bounds on generalized distances. *IEEE Trans. Inform. Theory*, 40(6):2090–2092, 1994.
5. Gérard Cohen and Gilles Zémor. Intersecting codes and independent families. *IEEE Trans. Inform. Theory*, 40:1872–1881, 1994.
6. Gérard D. Cohen, Sylvia B. Encheva, Simon Litsyn, and Hans Georg Schaathun. Intersecting codes and separating codes. *Discrete Applied Mathematics*, 128(1):75–83, 2003.
7. Gérard D. Cohen, Sylvia B. Encheva, and Hans Georg Schaathun. More on $(2,2)$-separating codes. *IEEE Trans. Inform. Theory*, 48(9):2606–2609, September 2002.
8. Gérard D. Cohen and Hans Georg Schaathun. Asymptotic overview on separating codes. Technical Report 248, Dept. of Informatics, University of Bergen, May 2003. Available at http://www.ii.uib.no/publikasjoner/texrap/index.shtml.
9. Tor Helleseth, Torleiv Kløve, and Johannes Mykkeltveit. The weight distribution of irreducible cyclic codes with block lengths $n_1((q^l-1)/n)$. *Discrete Math.*, 18:179–211, 1977.
10. János Körner. On the extremal combinatorics of the Hamming space. *J. Combin. Theory Ser. A*, 71(1):112–126, 1995.
11. A. Krasnopeev and Yu. L. Sagalovich. The Kerdock codes and separating systems. In *Eight International Workshop on Algebraic and Combinatorial Coding Theory*, 2002.
12. Robert J. McEliece, Eugene R. Rodemich, Howard Rumsey, Jr., and Lloyd R. Welch. New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities. *IEEE Trans. Inform. Theory*, IT-23(2):157–166, 1977.
13. Ilan Reuven and Yair Be'ery. Entropy/length profiles, bounds on the minimal covering of bipartite graphs, and the trellis complexity of nonlinear codes. *IEEE Trans. Inform. Theory*, 44(2):580–598, March 1998.

14. Ilan Reuven and Yair Be'ery. Generalized Hamming weights of nonlinear codes and the relation to the Z_4-linear representation. *IEEE Trans. Inform. Theory*, 45(2):713–720, March 1999.
15. Yu. L. Sagalovich. Separating systems. *Problems of Information Transmission*, 30(2):105–123, 1994.
16. Jessica N. Staddon, Douglas R. Stinson, and Ruizhong Wei. Combinatorial properties of frameproof and traceability codes. *IEEE Trans. Inform. Theory*, 47(3):1042–1049, 2001.
17. Victor K. Wei. Generalized Hamming weights for linear codes. *IEEE Trans. Inform. Theory*, 37(5):1412–1418, 1991.
18. Chaoping Xing. Asymptotic bounds on frameproof codes. *IEEE Trans. Inform. Theory*, 40(11):2991–2995, November 2002.