A quadratic alphabet threshold for non-overlapping codes

Peiwei Wu Qian Peng*
The University of Melbourne

hexagon:2611.01321v1

Abstract

Let $q_0(n)$ be the least integer such that every alphabet of size $q \geq q_0(n)$ admits a partition $I \sqcup J$ for which $I^{n-1}J$ is a maximum non-overlapping code of length n. Chee proved that $q_0(n)$ exists. We prove $q_0(n) = n^2/2 + O(n)$ and the finite bound $q_0(n) \leq 4n(n-1)$. The upper bound compares a uniform loss in the real partition relaxation with the cost of rounding a single alphabet split. Moving one two-letter word between parts produces an infinite family of exceptions and disproves a conjecture of Stanovnik, Moškon and Mraz. We determine the maximum for every alphabet size at lengths five and six. Their thresholds are 4 and 10, respectively. At length six the exceptional alphabet sizes are $2, 3, 4, 9$. On nine symbols the maximum is 33,872, exceeding the best I^5J code by 258.

1 The alphabet threshold

A code $C \subseteq \Sigma^n$ is *non-overlapping* if a nonempty proper prefix of a word in C never equals a suffix of another word in C or of the same word. Let $S_n(q)$ denote the largest size of such a code when $|\Sigma| = q$. For any partition $\Sigma = I \sqcup J$ into nonempty parts, $I^{n-1}J$ is non-overlapping. The best size given by these Blackburn codes is

$$B_n(q) = \max_{1 \leq s < q} s^{n-1}(q - s).$$

Blackburn conjectured that $S_n(q) = B_n(q)$ for each fixed $n \geq 2$ and all sufficiently large q [1]. Chee proved this conjecture and asked for the least integer $q_0(n) \geq 2$ such that equality holds for every $q \geq q_0(n)$ [2, Theorem I.1 and Section IV].

Theorem 1. *For all $n \geq 3$, $q_0(n) \leq 4n(n-1)$. For every $n \geq 12$,*

$$q_0(n) > n\big(\lfloor n/2 \rfloor - 2\big) + \lceil n/2 \rceil.$$

For $n \geq 564$, $q_0(n) \leq \lfloor n^2/2 + 143n \rfloor + 1$. In particular, $q_0(n) = n^2/2 + O(n)$.

Stanovnik, Moškon and Mraz conjectured that some maximum partition code has $R_i = \varnothing$ for all $i > 1 + n/(q+1)$, in the notation of (1). For $q \geq n$ this forces every right part above the first level to be empty and gives $S_n(q) = B_n(q)$ [4, Conjecture 1 and Remark 3]. An example of length six already contradicts that prediction.

At lengths two, three and four the maximum equals $B_n(q)$ over every alphabet [1, 4]. The next two lengths admit complete classifications.

Theorem 2. *$S_5(2) = 2$, $S_5(3) = 17$, and $S_5(q) = B_5(q)$ for all $q \geq 4$. Thus $q_0(5) = 4$.*

*Corresponding author.

Theorem 3. *For $q \geq 2$, one has $S_6(q) = B_6(q)$ except at four alphabet sizes, whose maxima are given below.*

q	2	3	4	9
$S_6(q)$	3	41	251	33,872
$B_6(q)$	1	32	243	33,614

In particular, $q_0(6) = 10$.

Table 1 of [4] already gives the maxima for $2 \leq q \leq 6$ at both lengths, including the first three exceptional values in Theorem 3. The accompanying data give an optimal ternary partition at length five with a single-word split at level two, up to reversal [5]. Theorems 2 and 3 extend these tables to every alphabet size.

2 An exceptional alphabet for every large length

We use the construction of Fimmel et al. [3] in the form analyzed by Stanovnik, Moškon and Mraz [4, Construction 6]. Partition $W_1 = \Sigma$ into L_1, R_1. For each $i \geq 2$, form

$$W_i = \bigcup_{j=1}^{i-1} L_j R_{i-j}, \qquad W_i = L_i \sqcup R_i \quad (i < n). \tag{1}$$

The union over j is disjoint. Every W_n constructed in this way is non-overlapping, and every maximum code has such a representation [4, Theorems 3 and 6, Proposition 7]. Reversing the words interchanges the two sides. Write l_i, r_i, w_i for the respective cardinalities.

Set $l_1 = a$, $r_1 = b$ and move a single word of $W_2 = L_1R_1$ into L_2. Put every other two-letter word and every subsequent level on the right. Thus $r_1 = b$, $r_2 = ab - 1$, and the resulting code has size r_n, where

$$r_j = ar_{j-1} + r_{j-2} \qquad (3 \leq j \leq n).$$

We use this same family for all lengths in the lower bound.

Proof of the lower bound in Theorem 1. Let $k = \lfloor n/2 \rfloor - 2$, $d = \lceil n/2 \rceil$, $q = nk + d$, $b = k + 1$ and $a = q - b$. Then $(n-2)b - a = 1$ and $a > n - 2$ for $n \geq 12$. The continuous maximum of $f(s) = s^{n-1}(q - s)$ occurs at $s_* = q(n-1)/n$, with $s_* - a = 1 - d/n \in (0, 1/2]$. In fact a is an integer maximizer. The identity $f'(s) = ns^{n-2}(s_* - s)$ shows that for equal distances from s_* the loss to the right is at least the loss to the left. Since $a + 1$ is at least as far from the peak as a, $f(a + 1) \leq f(a)$, and unimodality handles the other integers.

Write $v_j = r_j/a^{j-1}$. Then $v_1 = b$, $v_2 = b - 1/a$ and $v_j = v_{j-1} + v_{j-2}/a^2$. All terms are at least $b - 1/a$. Summing the increments from $j = 3$ to n yields

$$v_n \geq b - \frac{1}{a} + \frac{(n-2)(b - 1/a)}{a^2} = b + \frac{1}{a^2} - \frac{n-2}{a^3} > b.$$

It follows that $r_n > ba^{n-1} = B_n(q)$, so $q_0(n) > q$. Dividing this choice of q by n^2 gives $\liminf_{n\to\infty} q_0(n)/n^2 \geq 1/2$. □

3 A uniform loss for moderate partitions

Normalize each level cardinality in (1) by dividing it by q^i and retain the notation l_i, r_i, w_i for the normalized quantities. Write $a = l_1$, $b = r_1$, $M = w_n$; thus $a + b = 1$. Put

$$c = \frac{n-1}{n}, \qquad m = \frac{1}{n-1}, \qquad \delta = \frac{1}{2n(n-1)}, \qquad \beta(x) = x^{n-1}(1 - x).$$

If $\max(a,b) \geq (n-2)/(n-1)$, Chee's skew-partition theorem gives $M \leq B_n(q)/q^n$ [2, Theorem III.2]. For the remaining partitions, both a and b lie in $[m, c-2\delta]$. We call these partitions moderate; the following estimate also holds at the endpoints.

Let $L(z) = \sum_{i=1}^{n-1} l_i z^i$, let $\rho > 0$ satisfy $L(\rho) = 1$, and put $t = 1/\rho$. Chee's root identity [2, Lemma II.5] reads

$$M = \beta(t) - D, \qquad D = \sum_{\substack{1 \leq i,j < n \\ i+j > n}} l_i r_j \rho^{i+j-n}. \tag{2}$$

For a nonempty code $M > 0$, so $0 < t < 1$ and $\rho > 1$. The identity is valid in either orientation of the partition, using the corresponding L and ρ.

Lemma 4. *If $q \geq 2n$ and a moderate partition attains $S_n(q)$, then*

$$\beta(c) - M \;\geq\; \frac{1}{96n(n-1)^2}.$$

Proof. First, choose $s = \lceil qc \rceil$. Then $s/q \geq c$ and $1 - s/q \geq 1/(2n)$. Since $c^{n-1} \geq 1/3$, the Blackburn code of this split has normalized size at least $1/(6n)$, so $M \geq 1/(6n)$. We use $(1+1/j)^j \leq e < 3$ for positive integers j.

Suppose the endpoint terms $ar_{n-1} + bl_{n-1}$ in w_n account for at least $M/2$. Since $a+b=1$, one of r_{n-1}, l_{n-1} is at least $M/2$. Reverse the partition if needed to arrange $r_{n-1} \geq M/2$, then take the root in (2). The terms with $j = n-1$ give

$$D \geq r_{n-1} \sum_{i=2}^{n-1} l_i \rho^{i-1} = r_{n-1}(t-a).$$

If $t \geq c - \delta$, then $t - a \geq \delta$ and $\beta(c) - M \geq M\delta/2$. If $t < c - \delta$, integrate $\beta'(x) = nx^{n-2}(c-x)$ from $c-\delta$ to c. Since $(c-\delta)^{n-2} \geq ((n-2)/(n-1))^{n-2} \geq 1/3$, monotonicity on $[0,c]$ gives $\beta(c) - M \geq n\delta^2/6$.

Otherwise the terms $l_i r_j$ with $i + j = n$ and $i, j \geq 2$ sum to at least $M/2$. For each such pair, the recurrence yields $w_{i+1} \geq m l_i$ and $w_{j+1} \geq m r_j$. If $l_{i+1} \geq m l_i/2$, select the tail term $l_{i+1} r_j$. Otherwise, if $r_{j+1} \geq m r_j/2$, select $l_i r_{j+1}$. In the remaining case $r_{i+1} \geq m l_i/2$ and $l_{j+1} \geq m r_j/2$, so select $l_{j+1} r_{i+1}$. In every case the selected term is at least $m^2 l_i r_j/4$. Each term of total degree $n+1$ can be selected at most twice, by decrementing its first or its second index. Each term of degree $n+2$ can be selected at most once, since the ordered pair $(j+1, i+1)$ determines (i,j). The weights in D are at least one, so $D \geq m^2 M/16$.

The three lower bounds $M\delta/2$, $n\delta^2/6$ and $m^2 M/16$ are each at least $1/(96n(n-1)^2)$. □

Proof of the upper bound in Theorem 1. Assume $q \geq 4n(n-1)$ and round qc to a nearest integer s, which lies between 1 and $q-1$. Since $|\beta'(x)| \leq n|x-c|$ on $[0,1]$,

$$\beta(c) - \beta(s/q) \leq \frac{n}{8q^2} < \frac{1}{96n(n-1)^2}.$$

The Blackburn code with split s thus beats every putative moderate maximum, by Lemma 4. A maximum must have a skew partition, for which Chee's theorem gives the asserted equality. □

4 The linear error term

Proof of the asymptotic formula in Theorem 1. The alphabet sizes in the lower bound differ from $n^2/2$ by $O(n)$. We prove that any exceptional size $q \geq n^2/2$ also differs from $n^2/2$ by

at most $O(n)$, with absolute constants throughout. Choose a maximum partition for such an alphabet. Since $S_n(q) > B_n(q)$, Chee's skew-partition theorem makes it moderate. Put

$$\Delta = \beta(c) - M, \qquad P = c^{n-2}, \qquad F = \sum_{i=2}^{n-2} l_i r_{n-i}.$$

The root identity gives $0 \le D \le \Delta$. Let s be a nearest integer to qc. The rounding estimate of Section 3 gives $\Delta \le n/(8q^2) \le 1/(2n^3)$.

For $i+j = n$, $i, j \ge 2$, use $w_{i+1} \ge bl_i$ and $w_{j+1} \ge ar_j$ in the interior argument of Lemma 4. Select $l_{i+1}r_j$ if $l_{i+1} \ge bl_i/2$, and otherwise select $l_i r_{j+1}$ if $r_{j+1} \ge ar_j/2$. In the remaining case $l_{j+1}r_{i+1} \ge abl_i r_j/4$. Each choice is at least $abl_i r_j/4$, with multiplicity at most two. Hence $D \ge abF/8$. Moderateness gives $ab \ge m(1-m) \ge 1/[2(n-1)]$, so $F \le 16(n-1)\Delta = O(n^{-2})$.

The endpoint terms sum to $M - F = ar_{n-1} + bl_{n-1}$. Reverse the partition if necessary so that $r_{n-1} \ge M - F$, and use the root ρ of this orientation. Reversal preserves M and F. The root equation gives $t \ge a$, and the terms with $j = n-1$ in (2) give $D \ge (M-F)(t-a)$.

Write $u = n^2(c-t)$. For $n \ge 4$ we have $|u| \le 3$. Indeed, if $t < c - 1/n$, monotonicity and integration of $\beta'(x) = nx^{n-2}(c-x)$ on $[c-1/n, c]$ give

$$\Delta \ge \beta(c) - \beta(t) \ge \frac{(1-2/n)^{n-2}}{2n} \ge \frac{1}{18n},$$

since $(1 + 2/(n-2))^{n-2} \le e^2 < 9$. This contradicts $\Delta \le 1/(2n^3)$. For $t \ge c - 1/n$, every x between t and c satisfies $x^{n-2} \ge 1/9$. Integration on that interval gives $n(c-t)^2/18 \le \Delta$, proving the claim.

Both t and s/q lie within $3/n^2$ of c. For every x between c and either point, the mean value theorem on $[0,1]$ gives

$$|x^{n-2} - P| \le (n-2)|x - c| \le 3/n.$$

Since $P \ge c^{n-1} \ge 1/3$ and $|u| \le 3$, integration gives

$$\frac{n^3}{P}\big(\beta(c) - \beta(t)\big) \ge \left(1 - \frac{9}{n}\right)\frac{u^2}{2}, \tag{3}$$

$$\frac{n^3\Delta}{P} \le \frac{n^4}{8q^2}\left(1 + \frac{3}{n}\right). \tag{4}$$

In the grid estimate the integration interval has length at most $1/(2q) \le 1/n^2$, giving $|x^{n-2} - P| \le 1/n$ there.

Set $w = n(M-F)/P$. The identities $1 - c = 1/n$ and $nM/P = c - n\Delta/P$ give

$$1 - w = \frac{1}{n} + \frac{n(\Delta + F)}{P} \le \frac{25}{n},$$

since $\Delta + F \le 16n\Delta$ and $\Delta \le 1/(2n^3)$. In particular $0 \le w \le 1$. Moderateness gives $a \le c - 1/[n(n-1)]$, so $n^2(t-a) \ge n/(n-1) - u \ge 1 - u$. The endpoint estimate now gives $n^3 D/P \ge w(1-u) \ge 1 - u - 100/n$, since $|1-u| \le 4$. Adding (3) yields

$$\frac{n^3\Delta}{P} \ge \frac{u^2}{2} + 1 - u - \frac{141}{n} = \frac{1 + (u-1)^2}{2} - \frac{141}{n}.$$

If $n \ge 564$, comparison with (4) gives

$$\frac{2q}{n^2} \le \sqrt{1 + \frac{285}{n - 282}} \le 1 + \frac{285}{2(n-282)} \le 1 + \frac{285}{n}.$$

In particular, $q \le n^2/2 + 143n$. All larger alphabets admit a maximum Blackburn code, proving the finite bound and the asymptotic formula. □

5 Exact optimization at length five

Proof of Theorem 2. Return to integer counts and use reversal to assume $a = l_1 \geq b = r_1 > 0$, $q = a + b$. Write $x = l_2$, $y = r_2 = ab - x$, and $T = ay + bx = w_3$. With $u = l_3$ and $s = l_4$ the last two levels obey

$$w_4 = a(T - u) + xy + bu, \qquad w_5 = a(w_4 - s) + x(T - u) + uy + bs.$$

At fixed x, u the coefficient of s is $b - a \leq 0$, so $s = 0$ maximizes w_5. The remaining expression is affine in u. Its two endpoint values, at $u = 0, T$ respectively, are

$$F_0(x) = a^4 b + a^2(3b - a)x - (2a - b)x^2,$$
$$F_1(x) = 2a^3b^2 - 2ab(a - b)x - bx^2.$$

The maximum for the split is therefore $\max_{0 \leq x \leq ab} \max(F_0(x), F_1(x))$, with x integral.

When $a \geq 3b$, both polynomials are at most $a^4 b$. When $b \leq a \leq 3b$, both are at most $81q^5/1024$. For F_1 this follows by bounding it by $2a^3b^2$, whose maximum for real $a + b = q$ is $216q^5/3125$. For F_0, complete the square to obtain

$$F_0(x) \leq \frac{a^4((a + b)^2 + 4b^2)}{4(2a - b)}.$$

To compare this with $81q^5/1024$, multiply through by the positive denominators. The required nonnegative difference factors as

$$81(2a - b)(a + b)^5 - 256a^4((a + b)^2 + 4b^2)$$
$$= (3b - a)\big(94a^5 + 65a^4b + 260a^3b^2 - 30a^2b^3 - 90ab^4 - 27b^5\big).$$

The three subtracted terms have sum at most $147a^3b^2$.

If $[3q/4, 4q/5]$ contains an integer s, then $s^4(q - s) \geq 81q^5/1024$ by monotonicity up to $4q/5$. This condition holds whenever $q = 4h + 5j$ with $h, j \geq 0$, by taking $s = 3h + 4j$. It covers $q = 4, 5, 8, 9, 10$ and every $q \geq 12$ (use $12, 13, 14, 15$ and add fours). For $q = 6, 11$ use $s = 5, 9$ respectively; direct substitution gives the same lower bound. The remaining small splits are displayed below. Each F_0 maximum is obtained from the integer or integers nearest its vertex in $[0, ab]$, and F_1 is maximized at zero.

q	a	$\max F_0$	$\max F_1$
2	1	2	2
3	2	17	16
7	4	1088	1152
7	5	1268	1000
7	6	1296	432

For $q = 7$ the maximum is $6^4 = B_5(7)$. Every endpoint partition used in the calculation is feasible in (1), which gives equality in the two exceptional cases as well. □

6 Exact optimization at length six

Proof of Theorem 3. We first realize the exceptional sizes. For $q = 2$ use the words 001011, 001101 and 001111. Their one-letter prefixes and suffixes differ. At lengths two through five, each prefix starts with 00, which occurs only at the start of each word. For $q = 3, 4$ use the single marked pair of Section 2 with $a = q - 1$ and $b = 1$. Then $r_2 = a - 1$ and $r_j = ar_{j-1} + r_{j-2}$, giving $r_6 = 41,251$, respectively.

For $q = 9$, let $I = \{0, \dots, 6\}$ and $J = \{7, 8\}$. Take all words starting in I and ending in J such that every interior letter from J is a 7 immediately preceded by 0 or 1. Exclude the two endings 07 and 17. Any proper prefix ending in J ends with one of these pairs, so it cannot equal a suffix. The first two counts are $r_1 = 2$ and $r_2 = 12$. Removing either one initial letter from I or one of the two marked pairs gives $r_j = 7r_{j-1} + 2r_{j-2}$. Thus $r_6 = 33{,}872$. With just the pair 07 the recurrence in Section 2 gives 33,851, which already exceeds $B_6(9)$ and refutes Conjecture 1 of [4].

For the upper bound it suffices by Theorem 1 to treat $2 \le q < 120$. Reverse the partition if necessary to ensure $a = l_1 \ge b = r_1 > 0$. Put $x = l_2$, $y = ab - x$, $T = ay + bx = w_3$ and $u = l_3$. Then

$$w_4 = a(T - u) + xy + bu.$$

At level five we may place all words on the right, since the coefficient of l_5 in w_6 is $b - a \le 0$. For a fixed u, a word assigned to R_4 has coefficient $a^2 + x$ in w_6, whereas one assigned to L_4 has coefficient $ab + y$. Hence with $D = \max(a^2 + x, ab + y)$ the largest objective for these three initial splits is

$$\begin{aligned} Dw_4 + a\big(x(T - u) + uy\big) + u(T - u) &= A + Ku - u^2, \\ A &= D(aT + xy) + axT, \\ K &= D(b - a) + a(y - x) + T. \end{aligned}$$

For nonnegative integral u this is at most $A + \lfloor \max(K, 0)^2/4 \rfloor$. The following integer calculation evaluates that upper bound at every remaining q, a, x and records just the strict excesses over $B_6(q)$.

```
exceptions = {}
for q in range(2, 120):
    B = max(s**5 * (q-s) for s in range(1, q))
    U = 0
    for a in range((q+1)//2, q):
        b = q-a
        for x in range(a*b+1):
            y = a*b-x
            T = a*y+b*x
            D = max(a*a+x, a*b+y)
            A = D*(a*T+x*y)+a*x*T
            K = D*(b-a)+a*(y-x)+T
            U = max(U, A+max(K,0)**2//4)
    if U > B:
        exceptions[q] = U
```

The four entries are $(2, 3), (3, 41), (4, 251), (9, 33872)$. The constructions above attain all four bounds. For every other q a Blackburn split attains the upper bound. □

Formal verification. The accompanying Lean 4 development defines a code as a finite set of words, constructs the Blackburn codes and the exceptional family of Section 2, and verifies the bounds and limiting ratio for the least threshold $q_0(n)$. It also certifies the maximum at lengths five and six for every alphabet size. For the finite calculation at length six, it bounds the quadratic envelope on intervals of x, subdividing any interval until its bound is at most the claimed maximum.

For the upper bound the formal proof constructs integer layer counts that dominate each actual code C. Begin with each codeword split into single letters. Label a block left if it is

a prefix of a word of C and right otherwise. While at least two blocks remain, the first is left and the last is right by non-overlap. Merge an adjacent left–right pair and repeat. This parses every codeword. Counting possible pairs bounds the number of parsed words of length i by the convolution of the shorter left and right counts. Keep the left counts and increase the right counts successively until equality holds in the recurrence. Induction then bounds $|C|$ by the final convolution. This argument permits multiple parses of a word. Lean applies the verified root identity, skew estimate and moderate-partition bounds to the resulting integer profile.

The Lean source is available in the proof repository, version 1.1.0, and archived with the earlier preprint at doi:10.5281/zenodo.24590143.

Author contributions and use of AI. Peiwei Wu and Qian Peng contributed to the research ideas, discussed the mathematical arguments, and reviewed the manuscript and its results. They used OpenAI Codex to develop the mathematical derivations, draft the manuscript, and write the accompanying Lean formalization.

References

[1] S. R. Blackburn, Non-overlapping codes, *IEEE Trans. Inform. Theory* **61** (2015), 4890–4894. doi:10.1109/TIT.2015.2456634.

[2] Y. M. Chee, A proof of Blackburn's conjecture on non-overlapping codes, 2026. arXiv:2611.28776v2.

[3] E. Fimmel, C. J. Michel, F. Pirot, J.-S. Sereni and L. Strüngmann, Diletter and triletter comma-free codes over finite alphabets, *Australas. J. Combin.* **86** (2023), 233–270.

[4] L. Stanovnik, M. Moškon and M. Mraz, In search of maximum non-overlapping codes, *Des. Codes Cryptogr.* **92** (2024), 1299–1326. doi:10.1007/s10623-023-01344-z.

[5] L. Stanovnik, M. Moškon and M. Mraz, Computational data for *In search of maximum non-overlapping codes*, dataset accompanying [4].