A quadratic lower bound for intervals in the Hales–Jewett theorem

Peter Graham
https://orcid.org/0009-0007-8493-1505

November 5, 2026

Zenodo archive (Lean development in `lean`, `checks`): https://doi.org/10.5281/zenodo.23972681

hexagon:2611.00837v1

Abstract

A combinatorial line in $[m]^n$ is q-fold if its wildcard set, the set of coordinates that vary along the line, is a union of at most q intervals of consecutive coordinates. Let $I(m,r)$ be the least q such that, for n large, every r-coloring of $[m]^n$ has a monochromatic q-fold line. We show that $I(m,2) > (m-3)^2/128$ for every $m \geq 3$, so $I(m,r) \geq I(m,2)$ grows at least quadratically in m for every $r \geq 2$. This answers Question 3.1 of Conlon and Kamčev, who asked whether some $m \geq 4$, $r \geq 2$ and $c > 1$ admit r-colorings of $[m]^n$ with no monochromatic line whose wildcard set is a union of at most cr intervals: $m = 18$, $r = 2$ and any $1 < c < 3/2$ do, for every n. For $m = 3$ the value is known, r or $r-1$, by work of Conlon and Kamčev, Leader and Räty, and Kamčev and Spiegel; Kamčev and Spiegel noted that $r-1$ is the best explicit lower bound the known methods give, and Shelah's proof of the Hales–Jewett theorem gives the upper bound $\mathrm{HJ}(m-1,r)$. To rule out lines with at most Q intervals, we color x by the parity of $\lfloor S(x)/(32Q) \rfloor$, where $S(x)$ is the sum of the squared differences of consecutive letters of the word $0x$. Along such a line, S is a quadratic in the free letter whose values rise through a whole interval of length $32Q$ in smaller steps, so the line meets both colors. This coloring theorem is proved in Lean from the definitions; the step to $I(m,r)$ is a short argument on paper.

This work has not yet been reviewed by experts. The proofs, the Lean formalization and the text were produced by AI models (from OpenAI and Anthropic), working under the author's direction.

2020 Mathematics Subject Classification. 05D10.
Keywords. Hales–Jewett theorem, combinatorial lines, Ramsey theory, colorings.

1 Introduction

For integers $m, n \geq 1$ let $[m]^n$ be the set of words of length n over the alphabet $[m]$. We take $[m] = \{0, 1, \ldots, m-1\}$; relabeling the letters does not change the quantity $I(m,r)$ defined below, while the coloring we construct uses the letters as integers. A *root* is a word $w \in ([m] \cup \{*\})^n$ with at least one $*$, its *wildcard set* is $W = \{i : w_i = *\}$, and for a letter t the word $w(t)$ replaces every $*$ by t. The *combinatorial line* of w is $\{w(0), \ldots, w(m-1)\}$. For example, the root $w = *2{*}*$ over $[5]$ has wildcard set $\{1,3,4\}$, and $w(3) = 3233$. The Hales–Jewett theorem [3] says that for all m and r, if n is large then every r-coloring of $[m]^n$ has a monochromatic combinatorial line; the least such n is $\mathrm{HJ}(m,r)$.

A *run* of a root is a maximal interval of positions contained in W. The line of w is *q-fold* if w has at most q runs, that is, if its wildcard set is a union of at most q intervals; the root $*2{*}*$ above has two runs, so its line is q-fold for every $q \geq 2$. Following Kamčev and Spiegel [2, Section 3], let $I(m,r)$ be the least q such that, for all sufficiently large n, every r-coloring of $[m]^n$ contains

a monochromatic q-fold line. Since wildcard sets are nonempty, $I(m,r) \geq 1$. For $x \in [m]^n$ put $x_0 = 0$ and

$$S(x) = \sum_{i=1}^{n} (x_i - x_{i-1})^2, \qquad \chi_B(x) = \lfloor S(x)/B \rfloor \bmod 2.$$

So $S(x)$ is the sum of the squared jumps between consecutive letters of x, starting from a letter 0 placed before position 1; for example, $S(3233) = 3^2 + 1^2 + 1^2 + 0^2 = 11$. The nonnegative integers split into *blocks* $[kB, (k+1)B)$ of length B, and χ_B colors x by the parity of the block containing $S(x)$, so consecutive blocks have opposite colors.

Shelah's proof of the Hales–Jewett theorem [5] finds monochromatic lines with few runs: as Conlon and Kamčev record [1, Section 3], it gives $I(m,r) \leq \mathrm{HJ}(m-1,r)$. For alphabets of size at least four they asked, as Question 3.1 of the arXiv version of [1]:

> do there exist $m \geq 4$, $r \geq 2$ and $c > 1$ such that there are r-colorings of $[m]^n$ "containing no monochromatic combinatorial line whose wildcard set is the union of at most cr intervals"?

1.1 Main results

Theorem 1. *Let $Q \geq 1$ and $K \geq 0$ be integers with $K^2 \geq 32Q$ (for example $K = \lceil \sqrt{32Q} \rceil$), and let $m \geq 2K + 2$. For every $n \geq 1$, the 2-coloring χ_{32Q} of $[m]^n$ has no monochromatic Q-fold combinatorial line.*

Corollary 2. *Let $Q \geq 1$ be an integer. If $m \geq 2\lceil \sqrt{32Q} \rceil + 2$ then $I(m,2) \geq Q + 1$. Consequently $I(m,2) > (m-3)^2/128$ for every $m \geq 3$, and $I(m,r) \geq I(m,2)$ for every $r \geq 2$, so $I(m,r) \to \infty$ as $m \to \infty$ for each fixed r.*

Corollary 3. *For every n, the coloring χ_{64} of $[18]^n$ has no monochromatic 2-fold combinatorial line. Hence $m = 18$, $r = 2$ and any c with $1 < c < 3/2$ answer Question 3.1 positively. More generally, for every $r \geq 2$ and $c > 1$, every $m \geq 2\lceil \sqrt{32\lfloor cr \rfloor} \rceil + 2$ has, for every n, an r-coloring of $[m]^n$ with no monochromatic $\lfloor cr \rfloor$-fold line, that is, none whose wildcard set is a union of at most cr intervals.*

Theorem 1, and the statement that for every Q some alphabet size m admits, for every n, a 2-coloring of $[m]^n$ with no monochromatic Q-fold line, are proved in Lean from the definitions (Section 3). The step from them to $I(m,r)$ is the short argument proving Corollary 2.

1.2 Background and prior work

Conlon and Kamčev's Theorem 1.1 states that for any n and any odd $r > 1$ there is an r-coloring of $[3]^n$ "containing no monochromatic combinatorial line whose wildcard set is the union of fewer than r intervals" [1]; with Shelah's bound this gives $I(3,r) = r$ for odd r. Leader and Räty [4, Theorem 2] proved that for n sufficiently large every 2-coloring of $[3]^n$ has a monochromatic line whose wildcard set (their "active coordinate set") is an interval, disproving a conjecture of Conlon and Kamčev; that is, $I(3,2) = 1$. Kamčev and Spiegel [2, Theorem 2] proved that for even r and n large every r-coloring of $[3]^n$ has a monochromatic line whose wildcard set "is a union of at most $r-1$ intervals," which with Conlon and Kamčev's coloring gives $I(3,r) = r - 1$ for even r. Conlon and Kamčev added that "even a marginal improvement on either the upper or lower bound would be interesting" [1, Section 3].

Kamčev and Spiegel [2, Section 3] asked, as Question 19, whether one can improve on the bounds $\max\{I(m-1,r), I(m,r-1)\} \leq I(m,r) \leq \mathrm{HJ}(m-1,r)$, the first of which gives $I(m,r) \geq I(3,r)$ for

$m \geq 3$, adding: "In particular, $r-1$ is the best explicit lower bound that the present methods can give on $I(m,r)$, which is probably far from the truth." In 2026 Mouhib restated the question for alphabets of size at least four as open [6, Question 66] (also [7, Question 8.9] and [8, Question 4.21]), noting that "any lower bound requires genuinely asymmetric constructions": by [6, Proposition 51], colorings invariant under permuting coordinates cannot give one. The coloring χ_B is not of that kind, since S depends on adjacent pairs of letters.

Corollary 2 gives $I(m,r) \geq \max\{I(3,r), \lfloor (m-3)^2/128 \rfloor + 1\}$, a lower bound growing quadratically in m that exceeds the previous one whenever $(m-3)^2 \geq 128r$, and Corollary 3 answers Question 3.1 as stated. The natural stronger form of Question 3.1, whether $I(m,r) \geq cr$ for one fixed m and all r, remains open, as do Question 3.2 of [1], the upper-bound half of Question 19, and the true order of $I(m,2)$.

1.3 Proof overview

A root with q runs touches at most $2q$ fixed letters across a run boundary, and only those boundary terms of S depend on the free letter t. So $S(w(t)) = C + \sum_{c \in N}(t-c)^2$ for a constant C and a multiset N of between q and $2q$ letters: a parabola in t whose leading coefficient is at most $2Q$. On one side of its vertex, the alphabet offers $K+1$ consecutive letters along which this parabola rises by at least $K^2 \geq 32Q$ in total, and up to the first letter at which it has risen by $32Q$ every step is smaller than $32Q$ (below $22Q$). A sequence that climbs by at least one block length in steps of at most one block length must enter the next block of χ_{32Q}, which has the other color. The hard step is to find a coloring whose values along every line are this rigid. Colorings invariant under permuting coordinates cannot work [6, Proposition 51]; because S depends on adjacent pairs of letters, only a line's few run boundaries see the free letter, and this makes its values a parabola.

2 Proof of Theorem 1

Lemma 1 (values along a line)**.** *Let w be a root with $q \geq 1$ runs. There are an integer C and a multiset N of letters in $[m]$, of size A with $q \leq A \leq 2q$, such that $S(w(t)) = C + \sum_{c \in N}(t-c)^2$ for every $t \in [m]$.*

Proof. Put $w_0 = 0$ and consider the term $(x_i - x_{i-1})^2$ of $S(w(t))$ for $1 \leq i \leq n$. If neither i nor $i-1$ is in W, the term does not depend on t; this includes $i = 1 \notin W$, since position 0 is the letter 0. If both are in W, the term is $(t-t)^2 = 0$. If $i \in W$ and $i-1 \notin W$, then i is the first position of a run and the term is $(t - w_{i-1})^2$. If $i \notin W$ and $i-1 \in W$, then $i-1$ is the last position of a run and the term is $(t-w_i)^2$. Every run has one first position and at most one following position, so the boundary letters form a multiset N of size between q and $2q$, and C is the sum of the constant terms. □

Lemma 2 (crossing a block)**.** *Let $B \geq 1$ and let $v_0 \leq v_1 \leq \cdots \leq v_J$ be nonnegative integers with $v_{i+1} - v_i \leq B$ for every i and $v_J - v_0 \geq B$. Then $\lfloor v_0/B \rfloor$ and $\lfloor v_i/B \rfloor$ have different parities for some i.*

Proof. Let $k = \lfloor v_0/B \rfloor$. Since $v_J \geq v_0 + B \geq (k+1)B$, there is a least $i \geq 1$ with $v_i \geq (k+1)B$. Then $v_{i-1} < (k+1)B$, so $v_i \leq v_{i-1} + B < (k+2)B$ and $\lfloor v_i/B \rfloor = k+1$. □

Lemma 3 (a parabola leaves every block)**.** *Let $Q \geq 1$ and $K \geq 0$ be integers with $K^2 \geq B$, where $B = 32Q$, and let $m \geq 2K+2$. Let N be a multiset of A letters in $[m]$ with $1 \leq A \leq 2Q$, and let $f(t) = \sum_{c \in N}(t-c)^2$. Then there are distinct letters $t_0, \ldots, t_J \in [m]$ with $J \geq 1$ such that $f(t_0) \leq f(t_1) \leq \cdots \leq f(t_J)$, each increment $f(t_{i+1}) - f(t_i)$ is at most B, and $f(t_J) - f(t_0) \geq B$.*

Proof. Write $s = \sum_{c \in N} c$, so $f(t) = At^2 - 2st + \sum_{c \in N} c^2$ with vertex s/A and $0 \le s \le A(m-1)$. Since $f(m-1-t)$ is the same kind of polynomial for the multiset $\{m-1-c : c \in N\}$, we may assume $2s \le A(m-1)$, so that $s/A \le (m-1)/2$. Let $t_0 = \lceil s/A \rceil$ and $t_i = t_0 + i$. Then $t_0 \le \lceil (m-1)/2 \rceil$, so $t_i \le m-1$ for $i \le \lfloor (m-1)/2 \rfloor$, and $\lfloor (m-1)/2 \rfloor \ge K$ because $m \ge 2K+2$. With $d_i = t_i - s/A$, which satisfies $i \le d_i < i+1$,

$$f(t_{i+1}) - f(t_i) = A((d_i+1)^2 - d_i^2) = A(2d_i+1) \in [A(2i+1),\, A(2i+3)).$$

Summing the lower bound gives $f(t_j) - f(t_0) \ge Aj^2$, so $f(t_K) - f(t_0) \ge AK^2 \ge B$. Let $J \ge 1$ be least with $f(t_J) - f(t_0) \ge B$. By minimality $A(J-1)^2 \le f(t_{J-1}) - f(t_0) < B$, so $J - 1 < \sqrt{B/A}$. Each increment up to t_J is then less than

$$A(2(J-1)+3) < 2\sqrt{AB} + 3A \le 2\sqrt{2Q \cdot 32Q} + 6Q = 22Q < B. \qquad \square$$

Proof of Theorem 1. Let w be a root with q runs, $1 \le q \le Q$. By Lemma 1, $S(w(t)) = C + f(t)$, where f is as in Lemma 3 with $1 \le A \le 2q \le 2Q$. Lemma 3 gives letters $t_0, \ldots, t_J$ along which $v_i = S(w(t_i)) = C + f(t_i)$ satisfies the hypotheses of Lemma 2 with $B = 32Q$. Hence $\chi_{32Q}(w(t_0)) \ne \chi_{32Q}(w(t_i))$ for some i, and the line of w is not monochromatic. $\square$

Proof of Corollary 2. With $K = \lceil \sqrt{32Q} \rceil$, Theorem 1 gives, for every n, a 2-coloring of $[m]^n$ with no monochromatic Q-fold line, and hence none that is q-fold for any $q \le Q$, since a q-fold line is Q-fold. So no $q \le Q$ meets the definition of $I(m,2)$, and $I(m,2) \ge Q+1$. For $m \ge 3$ let $k = \lfloor (m-2)/2 \rfloor \ge (m-3)/2$ and $Q = \lfloor k^2/32 \rfloor$. If $Q \ge 1$, then $\lceil \sqrt{32Q} \rceil \le k$ and $m \ge 2k+2$, so $I(m,2) \ge Q+1 > k^2/32 \ge (m-3)^2/128$. If $Q = 0$, then $k^2 < 32$, so $(m-3)^2 \le 4k^2 < 128$ and $I(m,2) \ge 1 > (m-3)^2/128$, since wildcard sets are nonempty. Finally, a 2-coloring is an r-coloring, so $I(m,r) \ge I(m,2)$. $\square$

Proof of Corollary 3. Take $Q = 2$ and $K = 8$ in Theorem 1: $K^2 = 64 = 32Q$ and $2K+2 = 18$. If $1 < c < 3/2$ then $2 < cr < 3$, so a union of at most cr intervals is a union of at most two. For general r and c, take $Q = \lfloor cr \rfloor \ge 2$ and $K = \lceil \sqrt{32Q} \rceil$, and use the 2-coloring χ_{32Q} as an r-coloring. $\square$

Remark. The constants are not optimized.

3 Formal verification

Theorem 1 is proved in Lean 4 with Mathlib as `CK18HJ.main : MainStatement`, and the statement that for every Q some m admits, for every n, a 2-coloring of $[m]^n$ with no monochromatic Q-fold line as `CK18HJ.unbounded : UnboundedStatement`. The two statements read

```
def MainStatement : Prop :=
  ∀ Q K m n : ℕ, 1 ≤ Q → 32 * Q ≤ K * K → 2 * K + 2 ≤ m →
    ∀ w : Fin n → Option (Fin m), (∃ i, w i = none) → runs w ≤ Q →
      ∃ s t : Fin m, colour (32 * Q) (point w s) ≠ colour (32 * Q) (point w t)

def UnboundedStatement : Prop :=
  ∀ Q : ℕ, ∃ m : ℕ, ∀ n : ℕ, ∃ c : (Fin n → Fin m) → Bool,
    ∀ w : Fin n → Option (Fin m), (∃ i, w i = none) → runs w ≤ Q →
      ∃ s t : Fin m, c (point w s) ≠ c (point w t)
```

They match the theorem as follows. In `MainStatement` all four parameters are natural numbers, so $K \geq 0$ as in Theorem 1. Letters are `Fin m`, that is $\{0, \ldots, m-1\}$, and positions are `Fin n`, numbered from 0 rather than 1. A root is a map `Fin n → Option (Fin m)` with `none` for $*$, and the hypothesis $\exists$ `i, w i = none` says that the wildcard set is nonempty. `point w t` is $w(t)$. `runs w` counts the wildcard positions that are first or follow a letter, which is the number of runs. `S x` is $\sum_i (x_i - x_{i-1})^2$ over the integers, with the letter 0 before the first position, and `colour B x` is `decide ((S x / B) % 2 = 1)`; since $S(x) \geq 0$, integer division is the floor. The conclusion, two letters of the line with different colors, says that the line is not monochromatic. The step from these statements to the quantity $I(m, r)$, which is not defined in Lean, is the one-line argument of Corollary 2.

The build also restates both theorems with their full types, checks the examples of the introduction by evaluation in the kernel, and checks that both theorems depend only on the axioms `propext`, `Classical.choice` and `Quot.sound`. The proof uses no compiled evaluation. The Lean files accompany this paper. They pin Lean 4 v4.32.2 and a fixed Mathlib revision; with the Lean toolchain manager `elan` installed, running `lake exe cache get` and then `lake build` in the project's `lean` folder downloads Mathlib's compiled library and checks the proof, the restated statements, the examples and the axioms. The build itself takes under a minute and about 2 GB of memory.

Acknowledgments

The question answered here was posed by David Conlon and Nina Kamčev, and the quantity $I(m, r)$ and the remark on explicit lower bounds are from Nina Kamčev and Christoph Spiegel. Imre Leader and Eero Räty's theorem shaped the question for larger alphabets, and Younes Mouhib's 2026 papers and thesis record the question as open and show that symmetric colorings cannot answer it.

References

[1] D. Conlon and N. Kamčev, *Intervals in the Hales–Jewett theorem*, European J. Combin. 132 (2026), 104258, doi:10.1016/j.ejc.2025.104258; quotations and the numbering of Question 3.1 follow arXiv:1801.08919v2 (2018). The journal version was revised after that version (its references add [2]), and its wording may differ.

[2] N. Kamčev and C. Spiegel, *Another note on intervals in the Hales–Jewett theorem*, Electron. J. Combin. 29(1) (2022), #P1.62, doi:10.37236/9400; quotations and numbering follow this version, while arXiv:1811.04628v1 numbers these results Theorem 1.2 and Question 3.1 and reads "r is the best explicit lower bound".

[3] A. W. Hales and R. I. Jewett, *Regularity and positional games*, Trans. Amer. Math. Soc. 106 (1963), 222–229, doi:10.1090/S0002-9947-1963-0143712-1.

[4] I. Leader and E. Räty, *A note on intervals in the Hales–Jewett theorem*, Electron. J. Combin. 25(3) (2018), #P3.15, doi:10.37236/7664.

[5] S. Shelah, *Primitive recursive bounds for van der Waerden numbers*, J. Amer. Math. Soc. 1 (1988), 683–697, doi:10.1090/S0894-0347-1988-0929498-X.

[6] Y. Mouhib, *Lower bounds for the Hales–Jewett numbers via symmetric and one-weight colorings*, arXiv:2606.22155v2 (2026).

[7] Y. Mouhib, *One-weight colorings, the symmetric class, and lower bounds for Hales–Jewett numbers*, arXiv:2607.02226v1 (2026).

[8] Y. Mouhib, *Improving bounds on Hales–Jewett numbers: symmetric colorings, SAT solvers, line-family variants, and forcing structures*, master's thesis, ETH Zürich (2026), ys-mouhib.github.io/Mouhib_Master_Thesis.pdf.