

Four diagonals suffice: completeness of Parikh matrices when letters occur at most twice

Christopher Yamba

October 2026

Abstract

The Parikh matrix of a word over an ordered alphabet records, diagonal by diagonal, the numbers of occurrences of the scattered subwords made of consecutive letters. For permutations the pair counts on the third diagonal determine the whole matrix (the completeness theorem of the author’s 2006 thesis, and a special case of Salomaa’s γ -property theorem of 2005), while Salomaa showed that in general no fixed number of diagonals suffices. We prove that if every letter other than the first and the last letter of the alphabet occurs at most twice, the first four diagonals determine the Parikh matrix, and we give the long counts explicitly: the count of $j(j+1)\cdots l$ factorises across every edge e at which w avoids the scattered subword $e(e+1)e(e+1)$, and along a chain of interleaved pairs it is a bilinear expression in four counts of length 2 and 3. This proves, for such words, a statement that Salomaa reports was sometimes conjectured in general, and it contains Simpson’s result for palindromes in which every letter occurs twice; in particular, the Parikh classes of words in which each letter occurs twice are the sets of words with given third and fourth diagonals. Salomaa’s construction shows that counts of length at most n do not suffice when the inner letters occur n times; we show that one copy of the first letter is enough, giving for every $n \geq 2$ two words of content $(1, n, \dots, n)$ whose Parikh matrices differ only in the corner. Their window counts are ballot numbers, among them the Catalan number C_n , by a reflection argument. We conjecture that counts of length at most $n+1$ always suffice when the inner letters occur at most n times, and support it by exhaustive computation for $n = 3$ and $n = 4$. The main theorem, the rectangle rule, the crossing chain and the corner statement of the sharpness family for $n = 3, 4$ are formalised in Lean 4 at the level of words.

Keywords: Parikh matrix; scattered subword; M-ambiguity; Cauchy inequality for words; ballot numbers; formal verification.

Mathematics Subject Classification (2020): 68R15, 05A05, 05A15.

1 Introduction

Let w be a word over the ordered alphabet $1 < 2 < \cdots < k$. Its *Parikh matrix* $\Psi(w)$, introduced by Mateescu, Salomaa, Salomaa and Yu [2], is the upper unitriangular $(k+1) \times (k+1)$ matrix whose entry $(j, l+1)$ is the number $|w|_{j..l}$ of occurrences of $j(j+1)\cdots l$ in w as a scattered subword. Its second diagonal is the classical Parikh vector, the third holds the pair counts $|w|_{i(i+1)}$, the fourth the triple counts $|w|_{i(i+1)(i+2)}$, and so on: the d -th diagonal holds the counts of length $d-1$.

The author’s 2006 thesis [11] read the Parikh matrix of a permutation diagonal by diagonal (“it seems natural to continue this procedure, investigating the entries by going through the diagonals in increasing order” [11, printed p. 8, PDF p. 12]) and proved the *Parikh matrix completeness theorem*: the third diagonal of the Parikh matrix of a permutation determines

the whole matrix [11, Theorem 3.5, printed pp. 19–20, PDF pp. 23–24]. For words with repeated letters the third diagonal no longer suffices [7, p. 226], and [12, Theorem A] says exactly when it does. Salomaa’s γ -property [6], in the form quoted in [7, Theorem 3, p. 226], gives a sufficient condition for the second and third diagonals to determine the matrix; it appeared the year before the thesis and already contains the permutation case, since the condition is empty when every letter occurs once; its condition fails as soon as two adjacent letters occur at least twice each and have a pair count other than 0, 1, $n_i n_{i+1} - 1$ and $n_i n_{i+1}$. Salomaa then showed that *no* fixed number of diagonals suffices in general: for every entry there are two Parikh matrices that differ only there [7, Theorem 4, p. 228]. He adds that it is “not so easy to give examples of different Parikh matrices whose second, third and fourth diagonals coincide. (Indeed, it was sometimes conjectured that such examples do not exist.)” [7, p. 228]. His examples use many repeated letters: in his witnesses for a window of $n + 1$ letters, the inner and last letters occur n times and the first letter 2^{n-1} times (and $2^{n-1}t$ times when the two corners are to differ by t) [7, Lemma 3, p. 227, and proof of Theorem 4].

This note asks what happens when the multiplicities are bounded, and answers the first case. For $j < l$ the *window* $[j, l]$ is the set of letters $j, \dots, l$, its *corner* is the count $|w|_{j..l}$, and $j + 1, \dots, l - 1$ are its *inner letters*.

Theorem A. *Let $j < l$ with $l - j \geq 3$, and let w be a word in which every inner letter of $[j, l]$ occurs at most twice. Then $|w|_{j..l}$ is determined by the counts $|w|_{j'..l'}$ with $j \leq j' \leq l' \leq l$ and $l' - j' < l - j$: every word w' with the same such counts has $|w'|_{j..l} = |w|_{j..l}$. The value is given explicitly by Theorem 4.2.*

By induction on the length of the window we obtain the completeness theorem for doubled letters.

Corollary 1.1. *If every letter other than 1 and k occurs at most twice in w , then the first four diagonals of $\Psi(w)$ determine $\Psi(w)$: every word with the same first four diagonals has the same Parikh matrix.*

For permutations it is weaker by one diagonal than the thesis’s theorem, and it contains Simpson’s result for palindromes in which every letter occurs exactly twice [10, Corollary 4, p. 331], on which the third diagonal is a function of the second [10, Theorem 3(b)]. It confirms, for these words, the statement reported in [7, p. 228]. The number of diagonals cannot be lowered, nor the multiplicity bound raised: the first three diagonals do not suffice (Example 4.5), and “twice” cannot be replaced by “three times”, by Salomaa’s construction [7, Lemma 3 and proof of Theorem 4] with $n = 3$, or by Theorem B (Example 4.5). The proof (Sections 3 and 4) rests on two rules that extend the thesis’s product rule. The *rectangle rule* (Lemma 3.1) factorises a long count across any edge whose two letters meet as a rectangle; it gives a sufficient condition for equality in the Cauchy inequality for words [3, 5]. The *crossing chain* (Proposition 4.1) computes a long count along interleaved pairs $i(i + 1)i(i + 1)$. In particular the Parikh classes of the words in which each of k letters occurs twice are the sets of words with given third and fourth diagonals (Corollary 4.6).

In the other direction, the following family shows that the first letter of Salomaa’s witnesses can occur just once.

Theorem B. *For every $n \geq 2$ there are two words A_n, B_n of content $(1, n, n, \dots, n)$, over the letters $1, \dots, n + 1$, whose Parikh matrices differ only in the corner entry: $|A_n|_{1..(n+1)} = 1$ and $|B_n|_{1..(n+1)} = 0$. For $2 \leq i < j \leq n + 1$ and $m = j - i$, the counts $|A_n|_{i..j} = |B_n|_{i..j}$ are the ballot numbers $\binom{n+m}{m+1} - \binom{n+m}{m-1}$; in particular $|A_n|_{2..(n+1)}$ is the Catalan number C_n .*

So, as Salomaa’s construction already shows, when the inner letters occur n times the counts of length at most n do not determine the corner; Theorem B shows that this needs only one copy of the first letter. We conjecture that this is the only obstruction.

Conjecture C. Let $n \geq 1$. If every inner letter of $[j, l]$ occurs at most n times in w and $l - j \geq n + 1$, then $|w|_{j..l}$ is determined by the shorter counts of the window. Hence, if every letter other than 1 and k occurs at most n times, the first $n + 2$ diagonals determine the Parikh matrix.

The case $n = 1$ is the cut rule of Lemma 2.2, and $n = 2$ is Theorem A. Section 6 reports exhaustive evidence for $n = 3$ and $n = 4$; if true, the conjecture says that Salomaa's construction is optimal in its inner multiplicities. Section 7 describes a formalisation in Lean 4 of Theorem A, Corollary 1.1, the rectangle rule, the crossing chain and the corner statement of Theorem B for $n = 3, 4$, together with the computer checks of every numerical statement; the supplementary files are on the article's web page and in the repository <https://github.com/cyamba/yamba-ai-math/tree/main/research/parikh>. Our companion articles [13, 12] study the largest Parikh classes and the values of the fourth diagonal for words with repeated letters, and [12] determines exactly when the first three diagonals suffice [12, Theorem A]; the present note takes up the question of the further diagonals raised at the end of [12]. Its preliminaries, the product formula (2.1), the realisation of projections and the cut rule (Lemma 2.2) are repeated from [12, Section 2 and Lemma 3.1] to keep the note self-contained; no proof here depends on [13, 12].

2 Preliminaries

A *content* is a vector $\mathbf{n} = (n_1, \dots, n_k)$ of positive integers, and $\mathcal{W}(\mathbf{n})$ is the set of words over $1 < \dots < k$ with n_i letters i . Subwords are scattered: $|w|_u$ counts the position sequences $p_1 < \dots < p_\ell$ with $w_{p_1} \dots w_{p_\ell} = u$. We write $|w|_{j..l} = |w|_{j(j+1)\dots l}$ for $j \leq l$, so that $|w|_{m..m} = n_m$, and call $l - j + 1$ the *length* of the count. The Parikh matrix $\Psi(w)$ has entry $|w|_{j..l}$ in position $(j, l + 1)$, and is the product, from left to right, of the matrices $I + E_{m, m+1}$ over the letters m of w [2, Definition 1.1, Theorem 2.1]; in the thesis it is $\psi_{M_k}(w)$ [11, Definition 2.8, Theorem 2.10, printed pp. 12–13, PDF pp. 16–17]. *Edge* i joins the letters i and $i + 1$, and $x_i = |w|_{i(i+1)}$ is its *area*. In a window $[j, l]$ the *inner edges* are the edges e with $j < e$ and $e + 1 < l$. A *Parikh class* is the set of words with a given Parikh matrix.

The *projection* of w on edge i , its subword on the letters i and $i + 1$, is encoded by $A^{(i)}(w) = (A_1^{(i)}, \dots, A_{n_{i+1}}^{(i)})$, where $A_s^{(i)}$ is the number of letters i before the s -th letter $i + 1$; it is nondecreasing, with values in $\{0, \dots, n_i\}$. Let $\mathbf{B}_i$ be the $n_i \times n_{i+1}$ matrix with entry 1 in position (r, s) if the r -th letter i precedes the s -th letter $i + 1$, that is, if $r \leq A_s^{(i)}$, and 0 otherwise. It is a *Ferrers matrix*: its s -th column is the indicator of the rows $1, \dots, A_s^{(i)}$, the *heights*. Let $\mathbf{1}$ be the all-ones column vector. Expanding the product below, the term indexed by $(r_j, \dots, r_l)$ is 1 exactly when the r_j -th letter j , $\dots$, the r_l -th letter l form an occurrence of $j(j + 1) \dots l$; hence

$$|w|_{j..l} = \mathbf{1}^\top \mathbf{B}_j \mathbf{B}_{j+1} \dots \mathbf{B}_{l-1} \mathbf{1} \quad (j < l). \quad (2.1)$$

Conversely, every tuple of nondecreasing sequences $A^{(i)} \in \{0, \dots, n_i\}^{n_{i+1}}$ is the tuple of projections of a word: start from 1^{n_1} and insert, for $i = 1, \dots, k - 1$ in turn, the s -th letter $i + 1$ just before the $(A_s^{(i)} + 1)$ -th letter i , or at the end if $A_s^{(i)} = n_i$ (so every area $0 \leq x_i \leq n_i n_{i+1}$ occurs, which is [3, Lemma 8, p. 19]).

Lemma 2.1. *Let $\mathbf{B}_e$ be the Ferrers matrix of edge e in w .*

- (i) *The rank of $\mathbf{B}_e$ is the number of distinct nonzero heights.*
- (ii) *$\mathbf{B}_e$ has rank at most 1 if and only if w has no subword $e(e + 1)e(e + 1)$.*
- (iii) *If $n_e, n_{e+1} \leq 2$, then $\mathbf{B}_e$ has rank 2 if and only if $n_e = n_{e+1} = 2$ and $x_e = 3$, that is, if and only if the letters $e, e + 1$ appear in the order $e(e + 1)e(e + 1)$.*

Proof. (i) Equal heights give equal columns and zero heights zero columns, and the indicators of initial segments of distinct lengths are linearly independent. (ii) An occurrence of $e(e+1)e(e+1)$ uses the r_1 -th and r_2 -th letters e and the s_1 -th and s_2 -th letters $e+1$ with $r_1 \leq A_{s_1}^{(e)} < r_2 \leq A_{s_2}^{(e)}$, so two distinct nonzero heights occur; conversely, from nonzero heights $A_{s_1}^{(e)} < A_{s_2}^{(e)}$ take $r_1 = 1$ and $r_2 = A_{s_2}^{(e)}$. (iii) A Ferrers matrix with at most two rows and two columns has two distinct nonzero heights only if it is 2×2 with heights $(1, 2)$, which is the only 2×2 Ferrers matrix of area 3. $\square$

Lemma 2.2 (Cut rule). (i) If $j \leq e < e+1 \leq l$ and $x_e = 0$, then $|w|_{j..l} = 0$.

(ii) If $j < m < l$ and $n_m = 1$, then $|w|_{j..l} = |w|_{j..m} \cdot |w|_{m..l}$.

Proof. (i) $\mathbf{B}_e = 0$ in (2.1). (ii) The row vector $\mathbf{1}^\top \mathbf{B}_j \cdots \mathbf{B}_{m-1}$ and the column vector $\mathbf{B}_m \cdots \mathbf{B}_{l-1} \mathbf{1}$ have one entry each, equal to $|w|_{j..m}$ and $|w|_{m..l}$. $\square$

For permutations, (ii) gives $|w|_{j..l} = \prod_{i=j}^{l-1} |w|_{i(i+1)}$, the product form of the rule behind the thesis's completeness theorem, that a count is 1 exactly when the two shorter overlapping counts are 1 [11, Proposition 3.4, Theorem 3.5, printed pp. 17–20, PDF pp. 21–24]; with $l = j + 2$, (ii) is the case $|w|_b = 1$ of the equality condition in [3, Lemma 2, p. 12]. So Conjecture C holds for $n = 1$: a window with $l - j \geq 2$ and a single copy of each inner letter is determined by (ii).

3 Rectangles

Lemma 3.1 (Rectangle rule). If w has no subword $e(e+1)e(e+1)$, then for all $j \leq e < e+1 \leq l$,

$$|w|_{j..l} \cdot |w|_{e(e+1)} = |w|_{j..e+1} \cdot |w|_{e..l}.$$

Proof. By Lemma 2.1(ii), $\mathbf{B}_e = uv^\top$ for column vectors u, v . With $P = \mathbf{1}^\top \mathbf{B}_j \cdots \mathbf{B}_{e-1} u$ and $Q = v^\top \mathbf{B}_{e+1} \cdots \mathbf{B}_{l-1} \mathbf{1}$, (2.1) gives $|w|_{j..l} = PQ$, $|w|_{j..e+1} = P(v^\top \mathbf{1})$, $|w|_{e..l} = (\mathbf{1}^\top u)Q$ and $|w|_{e(e+1)} = (\mathbf{1}^\top u)(v^\top \mathbf{1})$. $\square$

In words, w avoids $e(e+1)e(e+1)$ exactly when its letters e and $e+1$ appear in the order $(e+1)^{p_1} e^{q_1} (e+1)^{r_1} e^{s_1}$: the pairs of a letter e before a letter $e+1$ form a rectangle, the first q_1 letters e times the middle r_1 letters $e+1$, and every occurrence of $j \cdots l$ through edge e is an occurrence of $j \cdots e$ ending in the rectangle followed by an occurrence of $(e+1) \cdots l$ starting in it. When $n_e = 1$ or $n_{e+1} = 1$, edge e is always a rectangle, and Lemma 3.1 follows from Lemma 2.2(ii) with $m = e$ or $m = e+1$, and is equivalent to it when $x_e \neq 0$.

Remark 3.2. Lemma 3.1 gives a sufficient condition for equality in the Cauchy inequality $|w|_y |w|_{xyz} \leq |w|_{xy} |w|_{yz}$ [3, Corollary 4, p. 12], [5, Theorem 4, p. 193], with $y = e(e+1)$, $x = j \cdots (e-1)$ and $z = (e+2) \cdots l$; the condition is not necessary, as $w = 1112323444$ with $(j, e, l) = (1, 2, 4)$ shows. For y a single letter the equality case is known in full [3, Lemma 2, p. 12]. Salomaa notes that “no general theory exists concerning the cases when the Cauchy inequality actually reduces to an equality” [5, p. 193]. For two letters a, b the analogous statement is known: $|w|_{ab} |w|_{a^m b^n} \leq |w|_{a^m b} |w|_{ab^n}$ for $m, n \geq 2$, strict exactly when w contains $abab$ and the right side is nonzero [5, Theorem 5, p. 195], the case $m = n = 2$ being [3, Lemma 3, p. 12]. We regard Lemma 3.1 as the counterpart of that statement for consecutive letters, not as new in substance. We have not been able to consult [8], whose title and abstract announce a study of equality cases.

4 Crossings, and the proof of Theorem A

Call edge e a *crossing* if $n_e = n_{e+1} = 2$ and $x_e = 3$; by Lemma 2.1(iii) its Ferrers matrix is $U_2 = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}$ and the letters $e, e+1$ appear in the order $e(e+1)e(e+1)$.

Proposition 4.1 (Crossing chain). *Let $j < l$ with $l - j \geq 3$, let every inner letter of $[j, l]$ occur twice and every inner edge be a crossing, and put $p = l - j - 2$. Let a_s be the number of letters j before the s -th letter $j+1$, and d_s the number of letters l after the s -th letter $l-1$ ($s = 1, 2$). Then*

$$|w|_{j..l} = a_1 d_1 + p a_1 d_2 + a_2 d_2, \\ |w|_{j(j+1)} = a_1 + a_2, \quad |w|_{j..j+2} = 2a_1 + a_2, \quad |w|_{(l-1)l} = d_1 + d_2, \quad |w|_{l-2..l} = d_1 + 2d_2.$$

Proof. By (2.1), $\mathbf{1}^\top \mathbf{B}_j = (a_1, a_2)$ and $\mathbf{B}_{l-1} \mathbf{1} = (d_1, d_2)^\top$, and the p inner edges contribute $U_2^p = \begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix}$. So $|w|_{j..l} = (a_1, a_2) U_2^p (d_1, d_2)^\top$. Edge $j+1$ is inner, so $|w|_{j..j+2} = (a_1, a_2) U_2 \mathbf{1} = 2a_1 + a_2$; symmetrically for $|w|_{l-2..l}$. $\square$

Combinatorially, an occurrence of $(j+1) \cdots (l-1)$ chooses the first or the second copy of each inner letter, and a crossing forbids the second copy of a letter to be followed by the first copy of the next; so the occurrence switches from first to second copies at most once. One occurrence uses only first copies, one only second copies, and p switch, at one of the p inner edges; this is the matrix U_2^p .

Theorem 4.2. *Let $j < l$ with $l - j \geq 3$, and suppose every inner letter of $[j, l]$ occurs at most twice in w .*

- (i) *If some inner edge e has $x_e = 0$, then $|w|_{j..l} = 0$.*
- (ii) *Otherwise, if some inner edge e is not a crossing, then $|w|_{j..l} = |w|_{j..e+1} \cdot |w|_{e..l} / x_e$.*
- (iii) *Otherwise every inner edge is a crossing, and with $X = |w|_{j(j+1)}$, $Y = |w|_{j..j+2}$, $X' = |w|_{(l-1)l}$, $Y' = |w|_{l-2..l}$ and $p = l - j - 2$,*

$$|w|_{j..l} = (Y - X)(2X' - Y') + p(Y - X)(Y' - X') + (2X - Y)(Y' - X').$$

In each case the right side involves only counts $|w|_{j'..l'}$ with $j \leq j' \leq l' \leq l$ and $l' - j' < l - j$, and which case applies is decided by the multiplicities of the inner letters and the areas of the inner edges. The formula in (ii) holds for every inner edge e that is not a crossing; we fix e to be the least one.

Proof. Inner edges exist since $l - j \geq 3$. (i) is Lemma 2.2(i). In (ii), the letters e and $e+1$ occur at most twice and edge e is not a crossing, so $\mathbf{B}_e$, which has at most two rows, has rank at most 1 by Lemma 2.1(iii) and avoids $e(e+1)e(e+1)$ by Lemma 2.1(ii); Lemma 3.1 applies with $x_e \neq 0$, and the two factors are shorter because $j < e$ and $e+1 < l$. In (iii) every inner letter is an end of an inner edge, so it occurs twice, and Proposition 4.1 applies; solve its last four equations for a_1, a_2, d_1, d_2 . The case distinction uses only n_e, n_{e+1} and x_e for the inner edges e . $\square$

Proof of Theorem A. Let w' have the same shorter counts as w in the window. It has the same multiplicities of the inner letters, hence at most two copies of each, and the same areas, so the same case of Theorem 4.2 applies to w and w' , with the same right side (in case (ii) with the same edge e , the least inner edge that is not a crossing, which is determined by these counts). $\square$

Proof of Corollary 1.1. Let w' have the same first four diagonals as w . We show $|w'|_{j..l} = |w|_{j..l}$ for $1 \leq j \leq l \leq k$ by strong induction on $l - j$. For $l - j \leq 2$ these are entries of the first four diagonals. For $l - j \geq 3$ the inner letters of $[j, l]$ lie in $\{2, \dots, k - 1\}$, so they occur at most twice in w , and by induction w' has the same shorter counts in the window; Theorem A applies. $\square$

Remark 4.3. In case (iii), the matrix form in the proof of Proposition 4.1 gives $|w|_{j+1..l-1} = p + 2$, $|w|_{j..l-1} = (p + 1)a_1 + a_2$ and $|w|_{j+1..l} = d_1 + (p + 1)d_2$, so

$$|w|_{j..l-1} |w|_{j+1..l} - |w|_{j..l} |w|_{j+1..l-1} = (a_2 - a_1)(d_1 - d_2),$$

the product of the number of letters j between the two letters $j + 1$ and the number of letters l between the two letters $l - 1$. This 2×2 minor of $\Psi(w)$ is nonnegative, as every minor of a Parikh matrix is [3, Theorem 6, p. 12], [5, Theorem 6, p. 196].

Remark 4.4. Two special cases of Corollary 1.1 are known, and a third result overlaps it. For permutations the second and third diagonals already determine the matrix, and more generally whenever every pair count has Salomaa's γ -property, whatever the multiplicities [7, Theorem 3, p. 226]. Under the hypothesis of Corollary 1.1 the γ -property fails exactly at the edges e with $n_e, n_{e+1} \geq 2$ and $x_e \notin \{0, 1, n_e n_{e+1} - 1, n_e n_{e+1}\}$: at an inner edge this means area 2, where the projection may be $e(e+1)(e+1)e$ or $(e+1)ee(e+1)$, and at an end edge with $n_1 = m \geq 2$ and $n_2 = 2$ (or $n_k = m \geq 2$ and $n_{k-1} = 2$) an area other than $0, 1, 2m - 1, 2m$; for instance 43211321 and 14323211, of content $(3, 2, 2, 1)$, agree on the first three diagonals and have $|w|_{123} = 0$ and 1. For palindromes in which every letter occurs exactly twice, every pair count is 2 and, by (2.1), every longer count is 0 or 2, so the γ -property fails at every edge and crossings do not occur, and the alphabet and the fourth diagonal determine the matrix [10, Theorem 3, Corollary 4, p. 331]; on these words the third diagonal is a function of the second [10, Theorem 3(b)], so Corollary 1.1 contains this result. The general statement and the formulas of Theorem 4.2 were not located in the literature we consulted.

Example 4.5. The words 21332 and 32123 of content $(1, 2, 2)$ have the same counts of length at most 2 but $|w|_{123} = 0$ and 1; so the fourth diagonal cannot be omitted in Corollary 1.1. For a three-letter window (a, b, c) the corner is determined, for every word of that content, by the counts of length at most 2 if and only if $b = 1$ or $a = c = 1$; this is the case $k = 3$ of [12, Theorem A(iii)]. Sufficiency is also the γ -property [7, Theorem 3]; directly, the corner is $x_1 x_2$ if $b = 1$ (Lemma 2.2(ii)) and $\max(0, x_1 + x_2 - b)$ if $a = c = 1$, and that the corner is not determined in general is noted in [7, p. 226]. For necessity, when $b \geq 2$ and $c \geq 2$ put 2^{b-2} and then 1^{a-1} in front of both words 21332, 32123 and 3^{c-2} after them; this keeps the shorter counts equal and the corners different. Use the reverse-complement when $a \geq 2$ and $c = 1$. Raising “twice” to “three times” fails as well: Salomaa's construction [7, Lemma 3 and proof of Theorem 4] with $n = 3$ gives two words of content $(4, 3, 3, 3)$, and Theorem B with $n = 3$ gives the words 3221443234 and 2343212443 of content $(1, 3, 3, 3)$, which agree on every count of length at most 3 and have $|w|_{1234} = 1$ and 0.

For the words in which every letter occurs twice, the content (2^k) , Corollary 1.1 identifies the Parikh classes, the sets of words with a common Parikh matrix, whose sizes are Salomaa's degrees of ambiguity [9, Definition 6, p. 109].

Corollary 4.6. *On the content (2^k) the Parikh classes are exactly the sets of words with given pair counts $(|w|_{i(i+1)})_i$ and triple counts $(|w|_{i(i+1)(i+2)})_i$. For $k = 3, \dots, 8$ there are 30, 173, 993, 5702, 32745 and 188045 such classes, and of the sets of words with given pair counts, 5, 39, 261, 1577, 9049 and 50181 split into several of them.*

Proof. The first statement is Corollary 1.1; the numbers are computed (Section 7). $\square$

5 A sharpness family

Let $n \geq 2$, let U be the $n \times n$ Ferrers matrix with heights $(1, 2, \dots, n)$, so $U_{r,s} = [r \leq s]$, and for $1 \leq t \leq n$ let $L_t = U - \mathbf{e}_t \mathbf{e}_t^\top$, the Ferrers matrix with heights $(1, \dots, t-1, t-1, t+1, \dots, n)$. As a projection, U is the crossing chain $pqpq \cdots pq$ of two letters, and L_t is the same word with its t -th pair pq inverted.

Lemma 5.1. *For $1 \leq m \leq n$ and $1 \leq a \leq n - m + 1$,*

$$\mathbf{1}^\top L_a L_{a+1} \cdots L_{a+m-1} \mathbf{1} = \binom{n+m}{m+1} - \binom{n+m}{m-1}.$$

Proof. The left side counts the sequences $1 \leq x_0 \leq x_1 \leq \cdots \leq x_m \leq n$ in which no $h \in [m]$ has $x_{h-1} = x_h = a+h-1$. Put $y_h = x_h + h$: these are the $(m+1)$ -subsets Y of $[n+m]$, and the forbidden event is $y_{h-1} = a+2h-2, y_h = a+2h-1$. Encode Y as a walk $S_0 = 0, S_1, \dots, S_{n+m}$ with step $+1$ at the elements of Y and -1 elsewhere; it ends at $S_{n+m} = m - n + 2$. Since $\#\{y \in Y : y < a+2h-2\} = h-1$ exactly when $S_{a+2h-3} = 1-a$, the forbidden event says that the walk crosses the level $\ell = 2-a$ upwards, from $\ell-1$ through ℓ to $\ell+1$, at times $a+2h-3, \dots, a+2h-1$; conversely every upward crossing of ℓ has this form, by parity and because $S_t \geq -t$ (parity gives $t-1 = a+2h-3$ for an integer h , $S_{t-1} \geq -(t-1)$ gives $h \geq 1$, and $h \leq m$ because Y then has $h+1$ elements up to $a+2h-1$). So the bad walks are those that visit $\ell-1$ and later $\ell+1$, since a walk with steps ± 1 that does so crosses ℓ upwards between the two visits. Since $S_0 = 0 \geq \ell-1$, reflecting the part before the first visit to $\ell-1$ maps them bijectively onto the walks from $2\ell-2$ to $m-n+2$ that visit $\ell+1$; as $m-n+2 \leq \ell+1$ (that is, $a \leq n-m+1$), reflecting the part after the first visit to $\ell+1$ maps these onto all walks from $2\ell-2$ to $2\ell+2-(m-n+2)$. Their displacement $n-m+2$ needs $n+1$ up-steps among $n+m$, so there are $\binom{n+m}{n+1} = \binom{n+m}{m-1}$ bad walks, out of $\binom{n+m}{m+1}$. $\square$

The argument is the reflection principle applied twice; for $a = 1$ the lemma is the classical ballot count of the lattice paths that stay weakly below a diagonal [1, Theorem 10.3.1].

For $m = n-1$ the value is $\binom{2n-1}{n} - \binom{2n-1}{n+1} = C_n$. In general it is the ballot number $\binom{n+m}{m+1} - \binom{n+m}{m-1} = \frac{n-m}{n+1} \binom{n+m+1}{m+1}$, an entry of Catalan's triangle [4, A009766]; it does not depend on a . We have not found Lemma 5.1 in matrix form, or the words A_n, B_n , in the literature we consulted.

Proof of Theorem B. Over the letters $1, \dots, n+1$ with content $(1, n, \dots, n)$, take the projection $(0, \dots, 0, 1)$ on edge 1 (the letter 1 stands just before the last letter 2), and on edge $i \in \{2, \dots, n\}$ the heights of L_{i-1} for A_n and of L_i for B_n ; the words exist by the realisation remark of Section 2. For $2 \leq i < j \leq n+1$, (2.1) gives $|A_n|_{i..j} = \mathbf{1}^\top L_{i-1} \cdots L_{j-2} \mathbf{1}$ and $|B_n|_{i..j} = \mathbf{1}^\top L_i \cdots L_{j-1} \mathbf{1}$, both equal to the ballot number of Lemma 5.1 with $m = j-i$. For the counts starting at the letter 1, $\mathbf{1}^\top \mathbf{B}_1 = \mathbf{e}_n^\top$, and $\mathbf{e}_n^\top L_t = \mathbf{e}_n^\top$ for $t < n$ while $\mathbf{e}_n^\top L_n = 0$. Hence $|A_n|_{1..j} = 1$ for $2 \leq j \leq n+1$, and $|B_n|_{1..j} = 1$ for $j \leq n$ but 0 for $j = n+1$. $\square$

For $n = 2, 3, 4$ the words are $A_2 = 32123, B_2 = 21332; A_3 = 3221443234, B_3 = 2343212443$; and $A_4 = 32245432132554345, B_4 = 23432215454323554$. In Salomaa's construction [7, Theorem 4] the window has the same length and the inner letters the same multiplicity n , but the first letter occurs 2^{n-1} times. In A_n and B_n the last letter occurs n times; we do not know the least possible end multiplicities in general (Section 6).

6 The conjecture for higher multiplicities

By Theorem B, Conjecture C would be sharp in its bound $l-j \geq n+1$. Its case $n = 3$ cannot be proved as Theorem A was: an edge with three letters of each kind can have area

6 with heights $(1, 2, 3)$, of rank 3, or $(0, 3, 3)$, of rank 1, so the area no longer tells which rule applies. The conjecture is a statement about finitely many objects for each window content, since the counts depend only on the projections, and every tuple of projections is realised. We call a window content *independent* if two tuples of projections agree on every count of the window except the corner and differ there, and *determined* otherwise. Table 1 lists the window contents that we checked over all tuples of projections, by hashing the tuples according to their counts other than the corner; up to reversal of the content, which preserves the question by the reverse-complement symmetry (reverse the word and replace each letter i by $k + 1 - i$, which permutes the counts), each listed content was found determined or, where so marked, independent.

n	window contents	k	result
2	every content with $n_i \leq 2$ and $\sum n_i \leq 11$; (2^k) , $k \leq 8$	4–8	determined (Theorem A)
3	(a, x, y, z, c) with $x, y, z \leq 3$, $\max(x, y, z) = 3$, $a, c \leq 6$ (357 contents)	5	determined
3	$(a, 3, 3, 3, 3, c)$ with $a, c \leq 4$; (a, x, y, z, t, c) with $x, y, z, t \in \{2, 3\}$, $a, c \leq 2$	6	determined
4	$(1, 4, 4, 4, 4, 1)$, $(1, 4, 4, 4, 4, 2)$, $(2, 4, 4, 4, 4, 2)$, $(1, 4, 4, 4, 4, 4)$, $(2, 4, 4, 4, 4, 3)$	6	determined
4	$(a, 4, 4, 4, c)$ with $a, c \leq 4$ and $a + c \geq 5$	5	independent
4	$(a, 4, 4, 4, c)$ with $a, c \leq 4$ and $a + c \leq 4$	5	determined

Table 1: Exhaustive tests of Conjecture C. A six-letter window with $n = 4$ is in its range; the contents $(1, 4, 4, 4, 4, 4)$ and $(2, 4, 4, 4, 4, 3)$ are sharp tests, as their five-letter subwindows are independent. The independent five-letter windows are consistent with it, being one letter short.

The end letters do not matter for the determination in Theorem A, but they do matter for independence below the bound: among the windows $(a, 3, 3, c)$ with $a, c \leq 6$ the corner is independent exactly when $\max(a, c) \geq 3$, and among $(a, 4, 4, 4, c)$ with $a, c \leq 4$ exactly when $a + c \geq 5$. We have no conjecture for the least end multiplicities.

7 Formal and computer verification

Lean. The file `ParikhCompleteness.lean` (supplementary) formalises the statements of Sections 3 and 4 at the level of words, in Lean 4.34.0 with its core library and no other dependency. Words are lists of natural numbers, $|w|_{j..l}$ is defined by the usual recursion for scattered subwords, and no matrix model is used: the proofs are inductions on the word, following the morphism property of the Parikh matrix one letter at a time, together with splitting lemmas at the first copy of a letter. The formalised theorems are the rectangle rule (Lemma 3.1), the crossing chain, in a parametrised form from which Theorem 4.2(iii) follows by solving two 2×2 systems, the consequence of Lemma 2.1(ii)–(iii) used in the proof (with at most two copies of e and $e + 1$, an edge that is not a crossing avoids $e(e + 1)e(e + 1)$), Theorem A and Corollary 1.1, and, by evaluation in the kernel, Example 4.5 and the corner statement of Theorem B for $n = 3, 4$. They depend only on the standard axioms (propositional extensionality, choice and quotient soundness), and the rectangle rule on propositional extensionality and quotient soundness only, and the evaluated cases on none; there is no `sorry`. Lemma 5.1 and Theorem B for general n are proved on paper only.

Computation. The script `check_article3.py` (supplementary, standard-library Python) re-checks: Lemma 3.1 and Theorem 4.2 on every word of every content with $3 \leq k \leq 5$, $n_i \leq 3$ and at most 9 letters; Example 4.5; Lemma 5.1 for $n \leq 30$ and all m, a , and Theorem B

for $n \leq 30$ by building the words and counting; the class numbers of Corollary 4.6; the first row and the five-letter rows of Table 1, and the windows $(a, 3, 3, c)$ of Section 6. The larger entries of Table 1 were computed by the supplementary C program `corner.c`, and most of them were reproduced by an independent Python program written for the review of this note; neither is part of any proof.

Acknowledgements and disclosure

Generative AI tools, Anthropic’s Claude (through Claude Code), assisted with the exploration, the drafting of proofs and of the Lean formalisation, the computational checks, the literature search and the drafting of this article. Mathematical responsibility and authorship rest with Christopher Yamba. The use of these tools is not an independent human peer review.

References

- [1] C. Krattenthaler, *Lattice path enumeration*, Handbook of Enumerative Combinatorics (M. Bóna, ed.), CRC Press, Boca Raton, FL, 2015, <https://doi.org/10.1201/b18255-16>, arXiv:1503.05930.
- [2] A. Mateescu, A. Salomaa, K. Salomaa, and S. Yu, *A sharpening of the Parikh mapping*, RAIRO Theor. Inform. Appl. **35** (2001), no. 6, 551–564, <https://doi.org/10.1051/ita:2001131>.
- [3] A. Mateescu, A. Salomaa, and S. Yu, *Subword histories and Parikh matrices*, J. Comput. System Sci. **68** (2004), no. 1, 1–21, <https://doi.org/10.1016/j.jcss.2003.04.001>.
- [4] OEIS Foundation Inc., *The On-Line Encyclopedia of Integer Sequences*, <https://oeis.org>, accessed 23 October 2026.
- [5] A. Salomaa, *Connections between subwords and certain matrix mappings*, Theoret. Comput. Sci. **340** (2005), no. 2, 188–203, <https://doi.org/10.1016/j.tcs.2005.03.024>.
- [6] ———, *On the injectivity of Parikh matrix mappings*, Fund. Inform. **64** (2005), no. 1–4, 391–404, Zbl 1102.68072.
- [7] ———, *Independence of certain quantities indicating subword occurrences*, Theoret. Comput. Sci. **362** (2006), no. 1–3, 222–231, <https://doi.org/10.1016/j.tcs.2006.06.025>.
- [8] ———, *On a Cauchy-type inequality for subword occurrences*, An. Univ. Bucureşti Inform. **62** (2015), no. 2, 101–112, Bibliographic record: https://tucs.fi/publications/view/?pub_id=jSalomaa_Arto15b.
- [9] ———, *Parikh matrices: subword indicators and degrees of ambiguity*, Adventures Between Lower Bounds and Higher Altitudes (H.-J. Böckenhauer, D. Komm, and W. Unger, eds.), Lecture Notes in Comput. Sci., vol. 11011, Springer, Cham, 2018, pp. 100–112, https://doi.org/10.1007/978-3-319-98355-4_7.
- [10] J. Simpson, *Short witnesses for Parikh-friendly permutations*, Australas. J. Combin. **78** (2020), no. 2, 329–334, <https://ajc-new.maths.uq.edu.au/v78.p329>.
- [11] C. Yamba, *Parikh matrices and permutation statistics*, Examensarbeten i matematik 2006:8, Matematiska institutionen, Stockholms universitet, 2006, https://kurser.math.su.se/pluginfile.php/16103/mod_folder/content/0/2006/2006_08_report.pdf.
- [12] ———, *From Euler numbers to the golden ratio: rigidity, ranges and the most ambiguous Parikh matrices for words with repeated letters*, Research preprint, version 1, 23 October 2026, <https://www.christopheryamba.com/math/parikh-golden-ratio/>.
- [13] ———, *Partition stabilization and semigroup gaps for ternary Parikh matrices*, Research preprint, version 1, 19 October 2026, <https://www.christopheryamba.com/math/ternary-parikh/>.