THE FINITENESS CONJECTURE HOLDS IN $(\mathrm{SL}_2 \mathbb{Z}_{\geq 0})^2$

GIOVANNI PANTI AND DAVIDE SCLOSA

ABSTRACT. Let A, B be matrices in $\mathrm{SL}_2 \mathbb{R}$ having trace greater than or equal to 2. Assume the pair A, B is coherently oriented, that is, can be conjugated to a pair having nonnegative entries. Assume also that either A, B^{-1} is coherently oriented as well, or A, B have integer entries. Then the Lagarias-Wang finiteness conjecture holds for the set $\{A, B\}$, with optimal product in $\{A, B, AB, A^2B, AB^2\}$. In particular, it holds for every pair of 2×2 matrices with nonnegative integer entries and determinant 1.

1. INTRODUCTION

Given a finite set Σ of square matrices of the same dimension and with real entries, the *joint spectral radius* of Σ is

$$\tilde{\rho}(\Sigma) = \lim_{n\to\infty} \max\{\|C\|^{1/n} : C \in \Sigma^n\},$$

where Σ^n is the set of all products of n matrices from Σ, repetitions allowed, and $\|-\|$ is the operator norm induced from some vector norm, whose choice is irrelevant. In short, $\tilde{\rho}(\Sigma)$ measures the maximal exponential growth rate of vectors under the action of Σ. Introduced by Rota and Strang in the sixties, the joint spectral radius remained unnoticed for almost thirty years. Research burst after the publication of a series of paper by Daubechies and Lagarias in the nineties, and connections with fields such as wavelet regularity, optimization, control theory, combinatorics, Lyapunov exponents and ergodic theory rapidly emerged. We refer to [17], [12], [15], [4], [21] and references therein for a broad panorama. In this paper we are concerned with a further connection, namely to dynamics on the hyperbolic plane.

Despite the simple definition, the computation of the joint spectral radius is a notoriously difficult problem (indeed it is NP-hard [27]), even in the restricted form of just determining whether it is nonzero. By the Berger-Wang theorem [2] we have the equivalent characterization

$$\tilde{\rho}(\Sigma) = \sup_n \max\{\rho(C)^{1/n} : C \in \Sigma^n\}, \tag{1.1}$$

2020 Math. Subj. Class.: 05A05; 15A60; 37F32.

The first author is partially supported by the MIUR Grant E83C18000100006 *Regular and stochastic behaviour in dynamical systems*.

This is an author-created, un-copyedited version of an article which appeared in *Nonlinearity*, n. 34, pp. 5234–5260, 2021. The publisher is not responsible for any errors or omissions in this version of the manuscript or any version derived from it. The Version of Record is available online at `https://iopscience.iop.org/article/10.1088/1361-6544/ac0484`.

where $\rho(C)$ is the spectral radius of C, and in [22, p. 19] Lagarias and Wang put forward the *finiteness conjecture*, namely the possibility that the supremum in (1.1) is always a maximum. Although in its full generality the conjecture was refuted in [5], counterexamples are difficult to construct, and are widely believed to be rare. The complexity of the matter already appears in the simplest setting, namely sets Σ containing just two 2×2 matrices. Indeed, such sets appear in the literature both as finiteness counterexamples [3], [13], [15], [24], as well as families of finiteness examples [18], [8], [20].

In this paper we adopt a geometric viewpoint and deal with sets $\Sigma = \{A, B\}$ of matrices in $\mathrm{SL}_2\,\mathbb{R}$, the group of 2×2 matrices with real entries and determinant one. Such matrices act on the hyperbolic plane $\mathcal{H} = \{z \in \mathbb{C} : \operatorname{im} z > 0\}$ via Möbius isometries $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) * z = (az + b)/(cz + d)$, and whenever the group Γ generated by Σ is fuchsian (i.e., acts on $\mathcal{H}$ in a properly discontinuous way) the quotient $X = \Gamma\backslash\mathcal{H}$ is a complete hyperbolic surface. In this case, asking about the joint spectral radius of Σ amounts to asking about the supremum of *mean free paths* along closed geodesics on X, namely about the supremum of mean time intervals between successive crossings of fixed cuts of X (corresponding to the generators A, B of Γ) that can be realized among closed geodesics; see [25, §10] for the case of billiards in hyperbolic polygons with ideal vertices. This geometric point of view appears also in [7, §6], where it is discussed the case of two hyperbolic translations coherently oriented and with disjoint axes (corresponding to X being a pair of pants, provided the two axes are sufficiently far apart). It also appears in [10], although the authors are concerned there with the limiting distribution of mean free paths (which turns out to be gaussian), rather than with their maximal value.

We summarize our results as follows, referring to the following sections for detailed statements. Fix $\Sigma = \{A, B\} \subset \mathrm{SL}_2\,\mathbb{R}$ with $\operatorname{tr}(A), \operatorname{tr}(B) \geq 2$. We say that $C \in \Sigma^n$ is an *optimal product* if $\tilde{\rho}(\Sigma) = \rho(C)^{1/n}$ and for no $1 \leq k < n$ and $D \in \Sigma^k$ we have $\tilde{\rho}(\Sigma) = \rho(D)^{1/k}$. The existence of optimal products amounts to the validity of the finiteness conjecture; their uniqueness —which may or may not hold— is intended up to conjugation. We assume that A, B are coherently oriented; this is a geometric condition on the translation axes of A and B (see Definition 2.1) that turns out to be equivalent to the fact that A, B are simultaneously conjugate to a pair of nonnegative matrices. Discarding the trivial case in which A and B commute and hence are simultaneously diagonalizable (or triangularizable, if parabolic), we will prove the following results.

(I) If A and B are hyperbolic with intersecting axes, then the one with larger trace is the unique optimal product. If they have the same trace, they are both optimal.

(II) If A and B are hyperbolic with asymptotically parallel axes, then the one with larger trace is the unique optimal product; this also holds if one of the two is parabolic with fixed point equal to one of the two fixed points of the other. If they have the same trace, then:

(II.1) If the attracting fixed point of one of the two is repelling for the other, then A and B are both optimal, and no other product is optimal;
(II.2) Otherwise, every product which is not a power is optimal.

(III) If A and B have the same trace and the pair A, B^{-1} is *not* coherently oriented, then AB is the unique optimal product.

(IV) The above statements leave uncovered the cases in which A, B are either both hyperbolic with different traces and ultraparallel axes, or one of the two is parabolic with fixed point distinct from the two fixed points of the other. Assume we are in one of these cases with $2 \leq \mathrm{tr}(A) < \mathrm{tr}(B)$, and further assume that A and B have integer entries.
(IV.1) If $\mathrm{tr}(AB) < \mathrm{tr}(B^2)$, then B is the unique optimal product;
(IV.2) If $\mathrm{tr}(AB) = \mathrm{tr}(B^2)$, then AB^2 is the unique optimal product;
(IV.3) If $\mathrm{tr}(AB) > \mathrm{tr}(B^2)$, then $\mathrm{tr}((AB)^3)$ and $\mathrm{tr}((AB^2)^2)$ differ at least by 2; if the former is larger, then AB is the unique optimal product, otherwise so is AB^2;
(IV.4) The statements (IV.1), (IV.2), (IV.3) are false if the assumption about integer entries is dropped.

Putting together the above statements, we obtain the result stated in the abstract.

Remark 1.1. The result stated in the title is strictly weaker than that stated in (I)–(IV). In particular, we are proving the finiteness conjecture for all pairs $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ which are simultaneously conjugate to a pair $A', B' \in \mathrm{SL}_2\,\mathbb{R}$ with nonnegative entries (this implies that both A and B have traces greater than or equal to 2). It may well occur that no conjugation yields A', B' with nonnegative *integer* entries; the pair C, D in Example 2.3 is such a specimen. The key point here is that the combinatorial proofs in §5–7 only require that *the traces* of arbitrary products of A, B are integers, and this property is conjugation-invariant.

Remark 1.2. Our results leave open the case of integer matrices of trace ≥ 2 and not coherently oriented. It is possible that the finiteness conjecture still holds in these circumstances, or even holds for arbitrary finite subsets of $\mathrm{SL}_2\,\mathbb{Z}$; this is a substantial open problem. We only remark here that it is not surprising that maps moving points "in opposite directions" are inherently more difficult to analyze than maps working in agreement; a classical example of this phenomenon arises with linked twist maps [26]. In our context, the difference of the two settings is well exemplified by the following not coherently oriented pair (zero entries are replaced by spaces):

$$A = \begin{pmatrix} 1 & \\ 2 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -3 \\ & 1 \end{pmatrix}.$$

One can show —for example, by constructing an appropriate extremal polytope norm [12]— that $\{A, B\}$ has the finiteness property. However, the resulting optimal product is A^2B^2, which is essentially different from the optimal products arising from the coherently ordered pairs treated in our work, that are always Christoffel words (i.e., induce periodic sturmian sequences).

Our paper is organized as follows: in §2 we give the definition of coherent orientation for pairs of nonelliptic matrices in $\mathrm{SL}_2\,\mathbb{R}$ and prove the equivalence alluded to above. We then establish in Theorem 2.5 inequalities relating the translation length of a matrix product with the sum of the translation lengths of the factors. In §3 we recast the finiteness property in terms of the existence of maximal elements for a certain preorder defined in the free semigroup on two generators; this interpretation allows us to replace optimal matrix products with better behaved optimal words. We prove statements (I), (II), (III) above in Theorems 3.4, 3.5, and 3.6. In §4 we provide counterexamples and settle (IV.4). In §5 we restrict attention to integer matrices and move from geometric arguments to combinatorial ones, establishing (IV.1) in Theorem 5.7. The statements (IV.2) and (IV.3) are more involved, requiring a section each, and are established in Theorems 6.7 and 7.5.

2. Coherently oriented nonelliptic matrices

The Möbius action of $\mathrm{SL}_2\,\mathbb{R}$ cited in §1 extends naturally to the euclidean boundary of the hyperbolic plane, namely the real projective line $\mathrm{P}^1\,\mathbb{R} = \partial\mathcal{H}$. A nonidentity matrix $A \in \mathrm{SL}_2\,\mathbb{R}$ is then *elliptic*, *parabolic*, or *hyperbolic* according to the number of fixed points —either zero, one, or two— it has in $\partial\mathcal{H}$; equivalently, according to the absolute value of its trace being less than, equal to, or greater than 2. Note that replacing A with $-A$ does not change the action and is irrelevant with respect to anything related to spectral radii. If A is hyperbolic, one of its fixed points is attracting and we will be denoted by α^+, the repelling one being denoted α^-; similar conventions hold for other letters $B, C, \ldots$. If A is parabolic, we agree that $\alpha^+ = \alpha^-$ is the only fixed point of A.

Let A be a nonidentity matrix in $\mathrm{SL}_2\,\mathbb{R}$ of trace ≥ 2, and let d denote hyperbolic distance (see [1] or [19] for basics of hyperbolic geometry). The *translation length* of A is

$$\ell(A) = \inf\{d(z, A * z) : z \in \mathcal{H}\}.$$

It has value 0 if and only if the infimum is not realized by any z, if and only if A is parabolic. If A is hyperbolic, then the set of points z realizing the infimum are precisely those points that lie on the *translation axis* of A, namely the unique geodesic of ideal endpoints α^+ and α^-. For A as above, spectral radius, trace, and translation length have neat relationships, namely

$$\begin{aligned} \rho &= \mathrm{tr}/2 + \sqrt{(\mathrm{tr}/2)^2 - 1} = \exp(\ell/2), \\ \mathrm{tr} &= \rho + \rho^{-1} = 2\cosh(\ell/2), \\ \ell &= 2\,\mathrm{arccosh}(\mathrm{tr}/2) = 2\log\rho. \end{aligned} \tag{2.1}$$

Since the functions involved in (2.1) are order-preserving bijections between the intervals $[1, \infty)$ (for spectral radius), $[2, \infty)$ (for trace), and $[0, \infty)$ (for translation length), comparing nonelliptic matrices with respect to one of these characteristics is the same as comparing them with respect to any other. Moreover, for every $A, B \in \mathrm{SL}_2\,\mathbb{R}$ with trace ≥ 2, we have $\rho(A) < \rho(B)$ if and only if $\rho(A^n) < \rho(B^n)$ for some (equivalently, for all) $n \geq 1$, and the same statement holds for trace and for translation length.

We look at the ideal boundary $\partial\mathcal{H}$ as a topological circle, cyclically ordered by the ternary betweenness relation $\alpha \prec \beta \prec \gamma$, which reads "$\alpha, \beta, \gamma$ are pairwise distinct, and traveling from α to γ counterclockwise we meet β". Every pair of distinct points α, β determines then two closed intervals, namely $[\alpha, \beta] = \{\alpha, \beta\} \cup \{x : \alpha \prec x \prec \beta\}$ and $[\beta, \alpha] = \{\beta, \alpha\} \cup \{x : \beta \prec x \prec \alpha\}$.

Definition 2.1. Let A, B be noncommuting matrices in $\mathrm{SL}_2\,\mathbb{R}$, both with trace greater than or equal to 2. If $\alpha^+ = \beta^+$, let $I^+ = \{\alpha^+\}$. If $\alpha^+ \neq \beta^+$, let I^+ be the one, of the two intervals $[\alpha^+, \beta^+]$ and $[\beta^+, \alpha^+]$, which is mapped into itself by both A and B, if such an interval exists (if it does then it is unique, since $AB \neq BA$ implies $\{\alpha^+, \alpha^-\} \neq \{\beta^+, \beta^-\}$). If such an interval does not exists, leave I^+ undefined. Replace in the above lines A, B with A^{-1}, B^{-1}, and α^+, β^+ with α^-, β^-, obtaining the definition of I^-. If both of I^+ and I^- are defined, then we say that the pair A, B is *coherently oriented*. If A, B are coherently oriented, but A, B^{-1} are not, then we say that A, B are *well oriented*.

It is clear that A, B are coherently oriented if and only if so are A^{-1}, B^{-1}. Coherently oriented hyperbolic pairs with ultraparallel axes are necessarily well oriented; see Example 2.3 and Figure 1 for taxonomy.

Lemma 2.2. *Let A, B be noncommuting matrices in $\mathrm{SL}_2\,\mathbb{R}$, both with trace ≥ 2. Then they are coherently oriented if and only if there exists a (not necessarily unique) matrix $C \in \mathrm{SL}_2\,\mathbb{R}$ such that CAC^{-1} and CBC^{-1} have nonnegative entries.*

Proof. Assume A, B are coherently oriented with I^+, I^- as in Definition 2.1. Since $\{\alpha^+, \alpha^-\} \neq \{\beta^+, \beta^-\}$, at least one of I^+, I^- is not a singleton, say I^+. Let K be the closure of the complement of I^+. Then $\alpha^-, \beta^- \in K$; indeed if, say, α^- were in the interior of I^+ we would have $A * \beta^+ \notin I^+$, which is impossible. This fact implies that $A^{-1}[K] \cup B^{-1}[K] \subseteq K$. Let now C be any matrix in $\mathrm{SL}_2\,\mathbb{R}$ such that $C[I^+] = [0, \infty]$. Setting $D = CAC^{-1}$ and $E = CBC^{-1}$ we obtain $D[0, \infty] \cup E[0, \infty] \subseteq [0, \infty]$ and $D^{-1}[\infty, 0] \cup E^{-1}[\infty, 0] \subseteq [\infty, 0]$. Write $D = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$; we want to prove that $a, b, c, d \geq 0$. Since $D[0, \infty] \subseteq [0, \infty]$, we have that a and c have the same sign, and so do b and d. The involution $S = \left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right)$ exchanges $[0, \infty]$ with $[\infty, 0]$. As a consequence, $SD^{-1}S^{-1} = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ maps $[0, \infty]$ into itself, which implies that a and b have the same sign, and so do c and d. We conclude that all of a, b, c, d have the same sign, which must be positive, since $\mathrm{tr}(D) \geq 2$; the same argument works for E.

Conversely, let A, B have nonnegative entries; then $\alpha^+, \beta^+ \in [0, \infty]$ and $\alpha^-, \beta^- \in [\infty, 0]$. Taking I^+ to be the interval of endpoints α^+, β^+ which is contained in $[0, \infty]$, and analogously for I^-, we see that A, B satisfy the conditions of coherent orientation, which are plainly preserved under conjugation. □

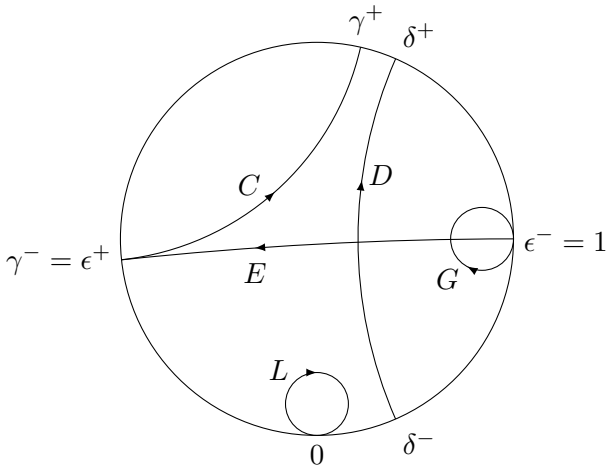

FIGURE 1. Examples of coherently oriented pairs

Example 2.3. Consider the following matrices:

$$C = \begin{pmatrix} 9 & 8 \\ 1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 5 & -1 \\ 1 & \end{pmatrix}, \quad E = \frac{1}{10}\begin{pmatrix} 17-2\sqrt{6} & -12+2\sqrt{6} \\ -3-2\sqrt{6} & 8+2\sqrt{6} \end{pmatrix},$$
$$G = \begin{pmatrix} 5 & -4 \\ 4 & -3 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & \\ 1 & 1 \end{pmatrix}.$$

We draw in Figure 1 the oriented translation axes of the hyperbolic C, D, E, as well as oriented horocycles corresponding to the parabolic G, L; note that although we work in the upper-plane model $\mathcal{H}$, we draw pictures in the Poincaré disk model.

Direct checking shows that coherently oriented pairs can be classified in six subcases as follows, three of them being well oriented.

- The parabolic-parabolic case, which is necessarily well oriented. The pair G^{-1}, L above is an example (with $I^+ = [0,1]$ and $I^- = [1,0]$); note that G, L are not coherently oriented. This case is covered by Theorem 3.6.
- The parabolic-hyperbolic case, which splits in two. A pair may be well oriented (e.g., C, G with $I^+ = [1, \gamma^+]$, $I^- = [\gamma^-, 1]$), or coherently oriented but not well oriented (e.g., E, G). The first subcase is covered by Theorems 5.7, 6.7, 7.5, and the second by Theorem 3.5.
- The hyperbolic-hyperbolic case. This splits in three, the subcases of intersecting (such as $D^{\pm 1}, E^{\pm 1}$) or asymptotically parallel axes (such as $C^{\pm 1}, E^{\pm 1}$) being not well oriented. The remaining case, exemplified by C, D, is of course well oriented. The first subcase is covered by Theorem 3.4, the second by Theorem 3.5, and the third by Theorems 5.7, 6.7, 7.5.

Lemma 2.4. *Let A, B have trace ≥ 2, and assume they are coherently oriented*; *let C be a product of A and B.*

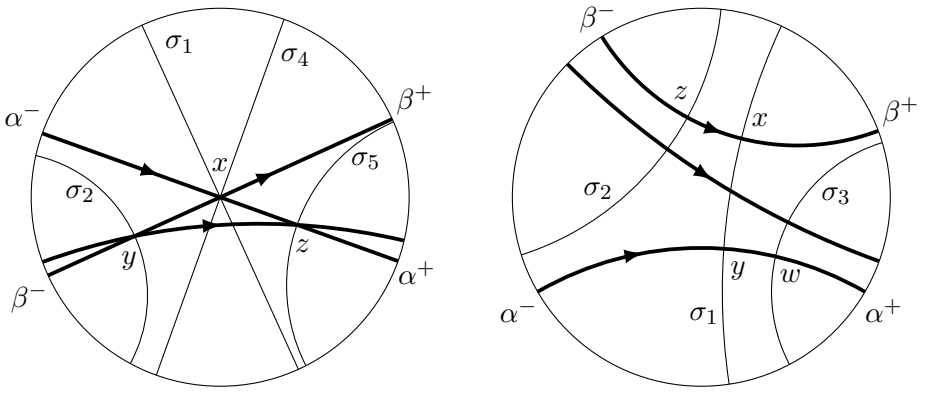

FIGURE 2. Coherently oriented geodesics, intersecting case left, nonintersecting right

(1) *We have* $\mathrm{tr}(C) \geq 2$, $\gamma^+ \in I^+$, $\gamma^- \in I^-$; *in particular, if* $I^+ \cap I^- = \emptyset$ *then* C *is hyperbolic.*
(2) *If* $\alpha^+ \neq \beta^+$ *and* C *is not a power of* B, *then* $\gamma^+ \neq \beta^+$; *an analogous statement holds for repelling fixed points.*

Proof. (1) Surely $\mathrm{tr}(C) > 0$ by Lemma 2.2. Since $A[I^+] \cup B[I^+] \subseteq I^+$ we have $C[I^+] \subseteq I^+$ and a descending chain $I^+ \supseteq C[I^+] \supseteq C^2[I^+] \supseteq \cdots$ that shrinks to $\gamma^+ \in I^+$; thus $\mathrm{tr}(C) \geq 2$. Inverting both A and B we get $\gamma^- \in I^-$.

(2) We have $C = DAB^k$, for some $k \geq 0$ and some product D of A and B. Then β^+ is an endpoint of the interval $B^k[I^+]$, and does not belong to $AB^k[I^+]$. Any further application of A and B to $AB^k[I^+]$ leaves β^+ outside, and therefore $\beta^+ \notin C[I^+]$, which implies $\gamma^+ \neq \beta^+$. □

Theorem 2.5. *Let* A, B *be hyperbolic and coherently oriented with* $I^+ \cap I^- = \emptyset$. *Then* $\ell(AB)$ *is less than, equal to, or greater than* $\ell(A)+\ell(B)$ *if and only if the axes of* A *and* B *are intersecting, asymptotically parallel, or ultraparallel, respectively.*

Proof. Assume that the axes are asymptotically parallel. Then, possibly replacing A, B with A^{-1}, B^{-1}, we may conjugate and assume $\alpha^+ = \beta^+ = \infty$ (since I^+ and I^- do not intersect, $\alpha^+ = \beta^-$ is excluded). We then have

$$A = \begin{pmatrix} r & t \\ & r^{-1} \end{pmatrix}, \quad B = \begin{pmatrix} s & u \\ & s^{-1} \end{pmatrix},$$

with $r, s > 1$. Since $\ell(A) = 2\log r$, and analogously for B and AB, we obtain $\ell(AB) = \ell(A) + \ell(B)$.

We assume that the axes intersect and apply a classical construction. Denote by x the intersection point, by y the point at distance $\ell(B)/2$ from x moving towards β^-, and by z the point at distance $\ell(A)/2$ from x moving towards α^+. We sketch the situation in Figure 2 left. Let σ_1 and σ_2 be the geodesics perpendicular to the axis σ_3 of B and passing through x and y, respectively. Also, let σ_4 and σ_5 be the geodesics perpendicular to the axis σ_6 of A and passing through x and z,

respectively. For each $i = 1, \dots, 6$, the reflection S_i through σ_i is an isometric involution of $\mathcal{H}$, and we have $A = S_5 S_4$ and $B = S_1 S_2$. The composition $S_4 S_1$ is a rotation about x, and equals the composition $S_6 S_3$, because the pair (σ_4, σ_1) is mapped to (σ_6, σ_3) by a rotation of $\pi/2$ about x. Summing up, we obtain

$$AB = (S_5 S_4)(S_1 S_2) = S_5(S_4 S_1)S_2 = S_5(S_6 S_3)S_2 = (S_5 S_6)(S_3 S_2),$$

which is the composition of a rotation of π about y, followed by a rotation of π about z. These two rotations leave the geodesic through y and z invariant, and thus this geodesic is the axis of AB; moreover,

$$\ell(AB) = d(y, AB * y) = d(y, S_5 S_6 * y) = 2d(y, z).$$

By the triangle inequality we conclude

$$\frac{\ell(AB)}{2} = d(y, z) < d(y, x) + d(x, z) = \frac{\ell(A)}{2} + \frac{\ell(B)}{2},$$

as desired.

Assume now that the axes of A and B are ultraparallel; see Figure 2 right. Then they determine a unique common perpendicular, denoted by σ_1, which intersects the axis of B in x and the axis of A in y. Let z be the point at distance $\ell(B)/2$ from x moving towards β^-, and w the point at distance $\ell(A)/2$ from y towards α^+. Draw the perpendicular σ_2 at z to the axis of B, and the perpendicular σ_3 at w to the axis of A. Defining as above S_i to be the reflection of mirror σ_i, we have $A = S_3 S_1$, $B = S_1 S_2$, and $AB = S_3 S_2$, because S_1 cancels. Ultraparallel geodesics have a well-defined hyperbolic distance, still denoted by d, and we have $d(\sigma_1, \sigma_2) = \ell(B)/2$, $d(\sigma_1, \sigma_3) = \ell(A)/2$, and $d(\sigma_2, \sigma_3) = \ell(AB)/2$.

Denote by $\overline{\sigma}_i$ the euclidean circle (possibly a straight line) in $\mathbb{C}$ of which σ_i is an arc. Then the $\overline{\sigma}_i$s are pairwise nonintersecting (because so are the σ_is, and an intersection outside $\mathcal{H}$ would produce an intersection inside, by Möbius inversion through $\partial\mathcal{H}$), and $\overline{\sigma}_1$ separates $\overline{\sigma}_2$ from $\overline{\sigma}_3$, meaning that any circle intersecting $\overline{\sigma}_2$ and $\overline{\sigma}_3$ intersects $\overline{\sigma}_1$ too. The circles $\partial\mathcal{H}$ and $\overline{\mathrm{axis}(AB)}$ are distinct and perpendicular to both $\overline{\sigma}_2$ and $\overline{\sigma}_3$; thus the set of circles perpendicular to both $\partial\mathcal{H}$ and $\overline{\mathrm{axis}(AB)}$ constitute the *coaxial pencil* determined by the pair $\overline{\sigma}_2, \overline{\sigma}_3$ [9, §4]. The key observation here is that $\overline{\sigma}_1$ does not belong to this pencil, since it is not perpendicular to $\mathrm{axis}(AB)$ (because common perpendiculars to pairs of ultraparallel geodesics are unique, and A, B, AB have distinct axes by Lemma 2.4(2)).

Now, while points in the hyperbolic plane obey the triangle inequality, ultraparallel geodesics obey the *non-triangle inequality* [9, §6], according to which the distance between ultraparallel geodesics is strictly greater than the sum of distances between the two given geodesics and a third one, separating the two but not coaxial to them. In our case we get

$$\frac{\ell(AB)}{2} = d(\sigma_2, \sigma_3) > d(\sigma_2, \sigma_1) + d(\sigma_1, \sigma_3) = \frac{\ell(B)}{2} + \frac{\ell(A)}{2},$$

again as desired. □

3. Words

As anticipated in §1, some caution is required in defining the length of matrix products. The problem is, of course, that a given pair $A, B \in \mathrm{SL}_2\,\mathbb{R}$ (even with nonnegative entries) may fail to generate not only a free group —a tolerable fault— but even a free semigroup. For example, the matrices

$$A = \begin{pmatrix} 1 & 1/\sqrt{6} \\ & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & \\ 1/\sqrt{6} & 1 \end{pmatrix},$$

satisfy the nontrivial identity

$$A^2B^3A^2 = BA^6B = \begin{pmatrix} 2 & \sqrt{6} \\ \sqrt{6}/2 & 2 \end{pmatrix};$$

see [6] for other examples. We deal with the issue by working with free semigroups of words, as follows.

Let $\{a, b\}$ be a two-letter alphabet, F_2^+ the free semigroup of words w of length $|w| \geq 1$, and F_2 the enveloping free group. Once a pair $A, B \in \mathrm{SL}_2\,\mathbb{R}$ has been fixed, we consider the group homomorphism $\phi : F_2 \to \mathrm{SL}_2\,\mathbb{R}$ that maps a to A and b to B, and the induced character $[-] : F_2 \to \mathbb{R}$ defined by $[w] = \mathrm{tr}(\phi(w))$.

Lemma 3.1. *The following statements are true.*

1. *Let ϕ', $[-]'$ be induced by another matrix choice $A', B' \in \mathrm{SL}_2\,\mathbb{R}$. If $[a] = [a]'$, $[b] = [b]'$, $[ab] = [ab]'$, then $[-] = [-]'$.*
2. $[w][u] = [wu] + [wu^{-1}]$.
3. *Given w, let u be either w^{-1}, or a rotation of w, or the reversal of w (that is, w written backwards). Then $[u] = [w]$.*
4. $[wuv] = [wv][u] - [wu^{-1}v]$, *and thus* $[wu^2v] = [wuv][u] - [wv]$.

Proof. (1) follows from the fact [14, Theorem 3.1] that, for a fixed w, there exists a polynomial $f_w \in \mathbb{Z}[x, y, z]$ such that, for varying ϕ, we have $\mathrm{tr}(\phi(w)) = f_w\big(\mathrm{tr}(\phi(a)), \mathrm{tr}(\phi(b)), \mathrm{tr}(\phi(ab))\big)$. (2) and the identity $[w] = [w^{-1}]$ are well known, and the invariance of $[-]$ under word rotation follows from the invariance of trace under conjugation. Define $\phi'(a) = A^{-1}$, $\phi'(b) = B^{-1}$; then $[-]' = [-]$ by (1) and invariance under rotation and group inversion. Letting u be the reversal of w, we obtain $[u] = [w^{-1}]' = [w^{-1}] = [w]$, which proves (3). Finally, (4) follows from (2) and rotation invariance. □

A key feature of our formalism is that not only word length in F_2^+ is better behaved than matrix product length, but the implicit comparison of spectral radii in (1.1) becomes an explicit preorder on words, as follows.

Definition 3.2. Let $A, B \in \mathrm{SL}_2\,\mathbb{R}$ have trace greater than or equal to 2, and assume that they are coherently oriented. Define ϕ, $[-]$ as above; by Lemma 2.4(1), the restriction of $[-]$ to F_2^+ takes values in $\mathbb{R}_{\geq 0}$. We define a binary relation $\preceq_{A,B}$ on F_2^+ by

$$w \preceq_{A,B} u \quad \text{if and only if} \quad \big[w^{|u|}\big] \leq \big[u^{|w|}\big].$$

By saying that a word is *maximal* we mean maximal with respect to $\preceq$ (for simplicity's sake we are dropping in the notation the dependence from A and B). A

complete set of optimal words is a possibly infinite subset $\{v_1, v_2, \ldots\}$ of F_2^+ such that:

- every v_i is maximal, and is a *Lyndon word*, i.e., is strictly smaller in the lexicographic order than any of its proper rotations (in particular, it is not a power);
- $v_i \neq v_j$ for $i \neq j$;
- every maximal w is a power of a rotation of some (necessarily unique) v_i.

A word that belongs to a complete set of optimal words is an *optimal* word.

Lemma 3.3. (1) *We have $w \preceq u$ if and only if $\big[w^{m/|w|}\big] \leq \big[u^{m/|u|}\big]$, where m is any common multiple of $|w|$ and $|u|$.*
(2) *The relation $\preceq$ on F_2^+ is a preorder, and every two words are comparable.*
(3) *If a complete set of optimal words exists, then it is unique.*
(4) *The finiteness conjecture holds for $\Sigma = \{A, B\}$ precisely when F_2^+ contains maximal —equivalently, optimal— words.*

Proof. Let $W = \phi(w)$, and analogously for u and v (later on we will apply this uppercase/lowercase convention without further notice). By the remarks following Equations (2.1), we have $w \preceq u$ if and only if $\rho\big(W^{|u|}\big) \leq \rho\big(U^{|w|}\big)$ if and only if $\rho\big(W^{m/|w|}\big) \leq \rho\big(U^{m/|u|}\big)$ if and only if $\big[w^{m/|w|}\big] \leq \big[u^{m/|u|}\big]$. It is clear that $\preceq$ is reflexive and every two words are comparable. Assuming $w \preceq u \preceq v$, and letting m be a common multiple of $|w|, |u|, |v|$, we obtain $\big[w^{m/|w|}\big] \leq \big[u^{m/|u|}\big] \leq \big[v^{m/|v|}\big]$, and thus $w \preceq v$. Let S and S' be two complete sets of optimal words, and let $v \in S$. Since v is maximal, it is a power of a rotation of some $v' \in S'$; by the elementary properties of Lyndon words, $v = v'$. The remaining assertions follow straight from the definitions; note that every word is both greater and less than any of its powers. In particular, if the maximal word w is a power of u, then u is maximal as well. □

We can now make precise and prove (I), (II) and (III) in §1. We stipulate for the rest of this paper, and without further repetitions, that A, B are noncommuting matrices in $\mathrm{SL}_2\,\mathbb{R}$, of trace greater than or equal to 2, and coherently oriented. The following theorem settles (I).

Theorem 3.4. *Let A, B be both hyperbolic, and assume that the translation axes intersect. If $[a] \leq [b]$ then b is an optimal word, and so is a provided $[a] = [b]$. There are no other optimal words.*

Proof. We show by induction that for every word w of length $n \geq 1$ we have $w \preceq b$. For $n = 1$ or $w \in \{a^n, b^n\}$ this is true. Let $n > 1$ and $w = au$ without loss of generality, with u not a power of a. By Lemma 2.4, $v^+ \in I^+ \setminus \{\alpha^+\}$ and $v^- \in I^- \setminus \{\alpha^-\}$. Therefore the axes of A and of U intersect, and by Theorem 2.5 and inductive hypothesis we have $\ell(W) < \ell(A) + \ell(U) \leq n\ell(B)$. Since $\ell(A^n) = n\ell(B)$ if and only if $[a] = [b]$, this also shows uniqueness. □

Theorem 3.5. *Let B be hyperbolic.*

(1) *If A is hyperbolic as well and the translation axes are asymptotically parallel, then*:

(1.1) *If $[a] \neq [b]$, then the only optimal word is the one among a and b that corresponds to the matrix with larger trace.*

(1.2) *If $[a] = [b]$ and $I^+ \cap I^- = \emptyset$, then every word which is not a power is optimal. Conversely, if I^+ and I^- intersect (necessarily in a singleton), then both a and b are optimal, and there are no other optimal words.*

(2) *If A is parabolic with $\alpha^+ = \alpha^- \in \{\beta^+, \beta^-\}$, then b is the only optimal word.*

Proof. (1) By conjugating, possibly exchanging A with B and inverting both of them, we may assume $\alpha^+ = \beta^+ = \infty$ when $I^+ \cap I^- = \emptyset$ and $\alpha^+ = \beta^- = \infty$ when $I^+ \cap I^- \neq \emptyset$. Let $r = \rho(A)$ and $s = \rho(B)$. In the first case, after a further conjugation by a parabolic matrix fixing ∞, and by a diagonal matrix, we obtain

$$A = \begin{pmatrix} r & 1 \\ & r^{-1} \end{pmatrix}, \quad B = \begin{pmatrix} s & \\ & s^{-1} \end{pmatrix}.$$

In the second case we similarly obtain

$$A = \begin{pmatrix} r & 1 \\ & r^{-1} \end{pmatrix}, \quad B = \begin{pmatrix} s^{-1} & \\ & s \end{pmatrix}.$$

It remains to check our claims (1.1) and (1.2) on these two pairs, which is easily done by direct inspection.

(2) is obvious: up to a conjugation we have

$$A = \begin{pmatrix} 1 & r \\ & 1 \end{pmatrix}, \quad B = \begin{pmatrix} s & \\ & s^{-1} \end{pmatrix},$$

for some $r \in \mathbb{R} \setminus \{0\}$ and $s \in \mathbb{R}_{>0} \setminus \{1\}$. □

Theorem 3.6. *Let $\mathrm{tr}(A) = \mathrm{tr}(B)$ and assume that the pair A, B is well oriented. Then the only optimal word is ab.*

Proof. By possibly exchanging A with B, and after an appropriate conjugation, we reduce our matrices to the standard form

$$A = \begin{pmatrix} 1 & \\ r & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & r \\ & 1 \end{pmatrix},$$

for some $r > 0$ in the parabolic case, or to the form

$$A = D^{-1}HD, \quad B = DHD^{-1},$$

in the hyperbolic one. In this second case we set

$$H = \begin{pmatrix} \cosh(\ell/2) & \sinh(\ell/2) \\ \sinh(\ell/2) & \cosh(\ell/2) \end{pmatrix}, \quad D = \begin{pmatrix} \exp(d/4) & \\ & \exp(-d/4) \end{pmatrix},$$

with $\ell = \ell(A) = \ell(B) > 0$ and $d = d(\mathrm{axis}(A), \mathrm{axis}(B)) > 0$. We will establish the result by showing that, for every word u which is not a power of ab or or ba, we have $u \prec ab$.

Fix such a u of length n, and let n_a, n_b be the number of occurrences of a —respectively b— in it. If one of n_a, n_b is zero, our claim is true: this is clear in

the parabolic case (because AB is hyperbolic), and follows from Theorem 2.5 in the hyperbolic one.

Claim. Let $\mathcal{W}(n_a, n_b)$ be the set of words containing n_a occurrences of a and n_b of b, and assume without loss of generality $n_a \geq n_b$. Let $w \in \mathcal{W}(n_a, n_b)$ with $[w]$ maximal among words in $\mathcal{W}(n_a, n_b)$. Then every occurrence of b in w is isolated, that is, is preceded and followed, in the cyclic order, by occurrences of a.

Proof of Claim. In the hyperbolic case this is the content of [16, Lemma 5-3]. The same statement holds in the parabolic case as well. Indeed, the proof of [16, Lemma 5-3] works by repeatedly applying the identity in Lemma 3.1(2), while making use of the following facts (references being relative to the quoted paper).

(1) Equation (5.1), namely

$$[a^p b^q a^t b^s] - [a^{p+t} b^{q+s}] = pqts\big([aba^{-1}b^{-1}] - 2\big), \tag{3.1}$$

(in the parabolic case, the Chebychev polynomials α_k, β_k of [16, §2] are both equal to k). By explicit computation, in our case we have

$$\begin{aligned} [a^p b^q a^t b^s] &= 2 + (p+t)(q+s)r^2 + pqtsr^4, \\ [a^{p+t} b^{q+s}] &= 2 + (p+t)(q+s)r^2, \\ [aba^{-1}b^{-1}] &= r^4 + 2, \end{aligned}$$

and (3.1) remains true.

(2) Lemma 5-2, which carries through.

(3) Lemma 4-3, which is only used through the inequality $[a^p b^q] > [a^{p-1} b^{q-1}]$; by direct computation one easily checks that this inequality still holds.

Having proved our claim, we may safely assume that all appearances of b in u are isolated. Since by assumption u is not a power of ab or of ba, not all occurrences of a are isolated; therefore, up to a rotation, we have

$$u = a^{k_1} b a^{k_2} \cdots b a^{k_t},$$

for some $t \geq 2$ and $k_1, \dots, k_t \geq 1$. We shall show that υ^+ is in the interior of $I = [0, 1]$ and υ^- in the interior of $[-1, 0]$. Let $k \geq 1$; it is clear that $A^k[I] \subset I$ both in the parabolic and in the hyperbolic case. We also have $A^k B[I] \subset I$; indeed, it suffice to consider $k = 1$. In the parabolic case one easily computes

$$AB[I] = \left[\frac{r}{r^2+1}, \frac{r+1}{r^2+r+1}\right] \subset I.$$

The hyperbolic case reduces to the computation of $AB * 1$, since $0 < AB * 0 < AB * 1$ anyway. Let $AB\binom{1}{1} = \binom{s}{t} \in \mathbb{R}^2_{>0}$. Then a little help from SageMath establishes that

$$\begin{aligned} t - s &= \sinh(-d) + \frac{1}{2}\sinh(d+l) + \frac{1}{2}\sinh(d-l) \\ &= \sinh(-d) + \sinh(d)\cosh(l) \\ &= \sinh(d)\big(\cosh(l) - 1\big) > 0, \end{aligned}$$

and thus $AB * 1 < 1$. As in the proof of Lemma 2.4 we obtain $A^{k_1} B A^{k_2} \cdots B A^{k_t}[I] \subset I$, and we conclude that υ^+ is in the interior of I.

The same argument, applied to the reversal v of u, shows that the attracting fixed point of $\phi(v)$ is in the interior of I as well. Letting $J = \left(\begin{smallmatrix} -1 & \\ 1 & \end{smallmatrix}\right)$, we see that J conjugates A with A^{-1} and B with B^{-1}, so that $\phi(v) = J\phi(u^{-1})J$. This implies that v^- is the J-image of this attracting fixed point, and thus is in the interior of $[-1, 0]$.

We now exchange a with b in u, obtaining u'. This corresponds to conjugating A and B by $F = \left(\begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix}\right)$; in particular, the attracting fixed point of $U' = \phi(u')$ is in the interior of $[1, \infty]$, and the repelling one in $[\infty, -1]$. As a consequence, the translation axes of U and of U' are ultraparallel.

By Lemma 3.1(1) we have $\ell(U) = \ell(U')$, and by Theorem 2.5 $\ell(UU') > 2\ell(U) = \ell(U^2)$. We have thus found a word, namely uu', that strictly dominates u in the $\preceq$ preorder. Since uu' contains the same number $n = n_a + n_b$ of occurrences of a and of b, we apply again our Claim above and infer $uu' \preceq ab$. This yields $u \prec ab$, as required. □

4. Counterexamples

We have thus proved (I), (II), (III) in §1, covering all cases in which A, B are coherently oriented but not well oriented. From now on we restrict attention to well oriented matrices with integer entries, and prove (IV); our tools, and the overall tone of our paper, will perceptibly move from geometry to combinatorics. This is unavoidable, since finiteness counterexample do indeed exist for well oriented pairs in $\mathrm{SL}_2\,\mathbb{R}$, wildly popping out as the pair varies smoothly in certain 1-parameters families of perfectly tame well oriented translations; this is the case, e.g., of the Morris example in [24, §2.2]. In order to make this tone shift more palatable to the reader, we begin by providing counterexamples, that is, by discussing (IV.4).

Let us first note that any triple $(r, s, t) \in \mathbb{R}^3_{\geq 2}$ determines uniquely up to conjugation a pair $A, B \in \mathrm{SL}_2\,\mathbb{R}$ such that $\mathrm{tr}(A) = r$, $\mathrm{tr}(B) = s$, $\mathrm{tr}(AB) = t$; indeed such triples give coordinates for the Teichmüller space of hyperbolic pair of pants. Let us fix $\mathrm{tr}(A) = 101/50$, and vary $\mathrm{tr}(B) = x$ in the interval $[101/50, 113/50]$. Adjusting the distance between the axes we may impose that the difference

$$\Delta = \mathrm{tr}(AB) - \mathrm{tr}(B^2)$$

be constant, in particular equal to 0 or to any small positive or negative number; once Δ is fixed we can compare words in F_2^+ in the $\preceq$ order.

We fix $\Delta = 0$, so that $[ab] = [b^2] = x^2 - 2$, and compare ab^2 with ab^3. We must compute $\big[(ab^2)^4\big]$ and $\big[(ab^3)^3\big]$. We have $[ab^2] = [ab][b] - [a] = (x^2 - 2)x - 101/50$ and $\big[(ab^2)^4\big] = T_4\big([ab^2]\big)$, where $T_k(y)$ is the degree k polynomial defined recursively by $T_0(y) = 2, T_1(y) = y, T_k(y) = yT_{k-1}(y) - T_{k-2}(y)$. Thus $\big[(ab^2)^4\big]$ is a polynomial in x of degree 12, and so is $\big[(ab^3)^3\big]$. Explicit computation gives

$$\begin{aligned}\big[(ab^2)^4\big] - \big[(ab^3)^3\big] = x^{10} - 101/50x^9 - 9x^8 + 303/25x^7 \\ + 98103/2500x^6 - 909/50x^5 - 46103/500x^4 - 2080903/125000x^3 \\ + 105559/1250x^2 + 1618727/31250x + 2050401/6250000,\end{aligned}$$

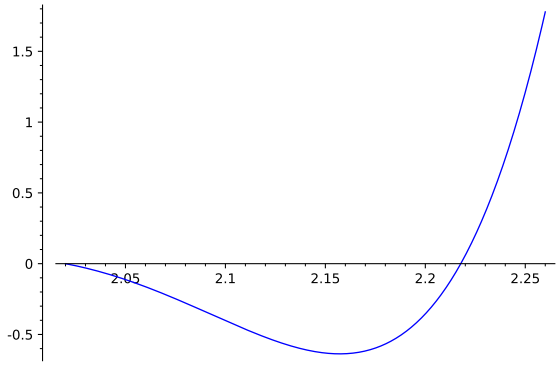

FIGURE 3. Graph of $\big[(ab^2)^4\big] - \big[(ab^3)^3\big]$ as a function of $[b]$.

whose graph appears in Figure 3. Therefore, for x ranging in an appropriate interval, we have $ab^2 \prec ab^3$ and the word ab^2 is not maximal, contrary to (IV.2).

Fix now $\Delta = -1/50$, so that $\operatorname{tr}(AB)$ is slightly less than $\operatorname{tr}(B^2)$. Then $[ab^2] = (x^2 - 2 - 1/50)x - 101/50$, while $[b^3] = T_3(x) = x^3 - 3x$. Thus, for large x, the word ab^2 dominates b, contrary to (IV.1).

Finally, let $\Delta = 1/50$; by analogous computations we obtain

$$\begin{aligned}\big[(ab)^3\big] &= x^6 - 297/50x^4 + 21903/2500x^2 - 227799/125000,\\ \big[(ab^2)^2\big] &= x^6 - 99/25x^4 - 101/25x^3 + 9801/2500x^2\\ &\quad + 9999/1250x + 5201/2500,\end{aligned}$$

which have the same value at $x_0 \sim 2.0255364739899213\ldots$. We compare the two words ab and ab^2 with their concatenation $abab^2$ by computing the differences $\big[(ab)^5\big] - \big[(abab^2)^2\big]$ and $\big[(ab^2)^5\big] - \big[(abab^2)^3\big]$, which are polynomials in x of degree 8 and 13, respectively. These polynomials are negative at x_0, so both ab and ab^2 are strictly dominated by $abab^2$, and (IV.3) fails.

5. INTEGER MATRICES AND CASE (IV.1)

Since in Theorems 3.4, 3.5, 3.6 we covered the case in which A and B have equal trace, we assume from now on, without loss of generality, that $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ are well oriented and satisfy $2 \le \operatorname{tr}(A) < \operatorname{tr}(B)$. For completeness's sake we provide one specimen for each of the cases (IV.1)–(IV.3) in §1; let L be as in Example 2.3

and $N = \left(\begin{smallmatrix} 1 & 1 \\ 1 & \end{smallmatrix}\right)$. Then we have the following examples:

$$\begin{aligned}
A = L, B = LN \qquad & \operatorname{tr}(AB) = 4 < 7 = \operatorname{tr}(B^2), \\
A = LNL, B = NLN^3 \qquad & \operatorname{tr}(AB) = 34 = \operatorname{tr}(B^2), \\
A = L^3N, B = N^2LN^2 \qquad & \operatorname{tr}(AB) = 40 > 34 = \operatorname{tr}(B^2) \\
& \operatorname{tr}((AB)^3) = 63880 > 55223 = \operatorname{tr}((AB^2)^2), \\
A = L^{11}, B = LNL \qquad & \operatorname{tr}(AB) = 15 > 14 = \operatorname{tr}(B^2) \\
& \operatorname{tr}((AB)^3) = 3330 < 3362 = \operatorname{tr}((AB^2)^2).
\end{aligned}$$

Definition 5.1. A *subword* of the word $w \in F_2^+$ is a possibly empty word obtained from w by deleting one or more not necessarily contiguous letters.

Since we are now working with matrices having integer entries, the range of $[-]$ on F_2^+ is $\mathbb{Z}_{\geq 2}$. The remark (1) in the following lemma is thus trivial, but key in our proofs.

Lemma 5.2. *Let w, u be words in F_2^+.*

(1) *$[w] < [u]$ if and only if $[w] \leq [u] - 1$.*
(2) *If u is a subword of w then $[u] < [w]$, exception being made for the case in which A is parabolic and w a power of a.*

Proof. In order to prove (2) we assume that we are not in the exceptional case, in which $[u] = [w] = 2$. It suffices to consider the removal of a single letter c, which by rotation invariance we may assume being the first one; let then $w = cu$. If both of C and U are hyperbolic then Theorem 2.5 applies, yielding $\ell(W) \geq \ell(C) + \ell(U)$. Thus $\ell(W) > \ell(U)$ and $[w] > [u]$ by the remarks following Equation (2.1).

Suppose $C = A$ is parabolic; then, since we are not in the exceptional case, U contains B as a factor and is hyperbolic. Moreover, by Lemma 2.4(2), neither of υ^+, υ^- equals the fixed point of C. By Lemma 2.2 we may assume $C, U \in \mathrm{SL}_2 \mathbb{R}$ have nonnegative entries, and a further conjugation —if needed— by the matrix F in the proof of Theorem 3.6 reduces us to the case

$$C = \begin{pmatrix} 1 & \\ r & 1 \end{pmatrix}, \quad U = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with $r, a, d > 0$ and $b, c \geq 0$. Now, c can be 0 —it is so precisely when one of υ^+, υ^- equals ∞— but b cannot, because otherwise one of υ^+, υ^- would equal the fixed point 0 of C. We thus obtain $[w] = a + rb + d > a + d = [u]$. An analogous proof applies if U is parabolic. □

Remark 5.3. A cofinite fuchsian group Γ has *bounded gaps* if any two distinct traces of elements of Γ differ at least by a fixed positive constant, depending on Γ; Lemma 5.2(1), that we will use repeatedly in our computations, is just the obvious remark that $\mathrm{SL}_2 \mathbb{Z}$ has bounded gaps. It is an intriguing speculation —but much wilder than the one in Remark 1.2— that the class of fuchsian groups derived from quaternion algebras (which is included in the class of bounded gaps groups, and conjecturally exhausts it [23, p. 423], [11, Conjecture 1.4]) is a natural candidate for a class of matrix groups that do not contain finiteness counterexamples.

Lemma 5.4. *If* $[ab] < [b^2]$ *then* $[ab^k] < [b^{k+1}]$*, for every* $k \geq 1$*.*

Proof. We work by induction. The case $k = 1$ is by hypothesis, and for $k = 2$ we have

$$[ab^2] = [ab][b] - [a] \leq \big([b^2] - 1\big)[b] - [a] < [b^3] - 1.$$

Let $k > 2$; repeatedly applying Lemma 3.1(2) to the left side we obtain

$$\begin{aligned}
[ab^k] &= [ab][b^{k-1}] - [ab^{2-k}] \\
&= [ab][b^{k-1}] - [a][b^{k-2}] + [ab^{k-2}] \\
&< \big([b^2] - 1\big)[b^{k-1}] - [b^{k-2}] + [b^{k-1}] \\
&= [b^{k+1}] + [b^{k-3}] - [b^{k-1}] - [b^{k-2}] + [b^{k-1}] \\
&= [b^{k+1}] + [b^{k-3}] - [b^{k-2}] \\
&\leq [b^{k+1}] - 1.
\end{aligned}$$

Here the third line follows by induction hypothesis, and the last one from $[b^{k-3}] < [b^{k-2}]$, which is valid for $k > 2$. □

Lemma 5.5. *Among all words of fixed length, the trace-maximizing ones do not contain the factor* a^2*.*

Proof. Since $[a] < [b]$, no such word w can be a power of a. Assume that w contains a^2. Then, up to a rotation, $w = bua^2$, and it is enough to prove $[bua^2] < [buab]$. The right side equals $[bua][b] - [ua]$, and the other side $[bua][a] - [bu]$. The difference is then greater than $[bua] - [ua] + [bu]$, which is strictly positive by Lemma 5.2(2). □

Lemma 5.6. *If* $[ab] < [b^2]$ *then, for every* $s, k_1, \cdots, k_s \geq 1$*, we have*

$$[ab^{k_1} \cdots ab^{k_s}] < [b^{k_1 + \cdots + k_s + s}].$$

Proof. We work by induction on s, the case $s = 1$ having been proved in Lemma 5.4. Let $s \geq 2$. We can suppose $k_{s-1} \geq k_s$, which ensures that

$$[ab^{k_1} \cdots ab^{k_{s-1}}(ab^{k_s})^{-1}] = [b^{k_1} \cdots ab^{k_{s-1} - k_s}]$$

is positive. We thus obtain

$$\begin{aligned}
[ab^{k_1} \cdots ab^{k_s}] &< [ab^{k_1} \cdots ab^{k_{s-1}}][ab^{k_s}] \\
&\leq ([b^{k_1 + \cdots + k_{s-1} + s - 1}] - 1)([b^{k_s + 1}] - 1) \\
&= [b^{k_1 + \cdots + k_{s-1} + k_s + s}] + [b^{k_1 + \cdots + k_{s-1} - k_s + s - 2}] \\
&\qquad - [b^{k_1 + \cdots + k_{s-1} + s - 1}] - ([b^{k_s + 1}] - 1) \\
&< [b^{k_1 + \cdots + k_{s-1} + k_s + s}] - ([b^{k_s + 1}] - 1) \\
&< [b^{k_1 + \cdots + k_{s-1} + k_s + s}].
\end{aligned}$$

□

We can now prove §1(IV.1).

Theorem 5.7. *Let $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ be well oriented, and assume $2 \le \mathrm{tr}(A) < \mathrm{tr}(B)$ and $\mathrm{tr}(AB) < \mathrm{tr}(B^2)$. Then b is the only optimal word.*

Proof. Let w be a word of length n, trace-maximizing among all words of the same length; we must prove that w does not contain any a. By Lemma 5.5, w does not contain a^2 as a factor. After a rotation we may apply Lemma 5.6, and conclude that $w = b^n$. □

6. Case (IV.2)

The remaining cases (IV.2) and (IV.3) are more involved and require a section each. The standing assumptions in this section for the well oriented pair $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ are $2 \le \mathrm{tr}(A) < \mathrm{tr}(B)$ and $\mathrm{tr}(AB) = \mathrm{tr}(B^2)$. They yield

$$[ab] - [a][b] = [b^2] - [a][b] \ge [b^2] - ([b]-1)[b] = [b] - 2 \ge [a] - 1 > 0,$$

or, equivalently, $[ab^{-1}] < 0$. This will be useful several times.

Lemma 6.1. *Fix a word w and assume $s \ge 0$. Then we have*:

(1) $[wab(ab^2)^s ab^3] < [wab^2(ab^2)^s ab^2]$;
(2) $[wab^3(ab^2)^s ab] < [wab^2(ab^2)^s ab^2]$, *if w is empty or begins with a.*

Proof. We prove (1). We have

$$\begin{aligned}
[wab(ab^2)^s ab^3] &= [wab(ab^2)^s ab][b^2] - [wab(ab^2)^s ab^{-1}] \\
&= [wab(ab^2)^s ab][b]^2 - 2[wab(ab^2)^s ab] \\
&\quad - [wab(ab^2)^s a][b] + [wab(ab^2)^s ab] \\
&= [wab(ab^2)^s ab][b]^2 - [wab(ab^2)^s ab] - [wab(ab^2)^s a][b],
\end{aligned}$$

and, by Lemma 3.1(4),

$$\begin{aligned}
[wab^2(ab^2)^s ab^2] &= [wab(ab^2)^s ab^2][b] - [wa(ab^2)^s ab^2] \\
&= [wab(ab^2)^s ab][b]^2 - [wab(ab^2)^s a][b] - [wa(ab^2)^s ab^2].
\end{aligned}$$

Subtracting the first end result from the second we get

$$\begin{aligned}
[wab(ab^2)^s ab] - [wa(ab^2)^s ab^2] &= [w(ab^2)^s ab][ab] - [w(ab)^{-1}(ab^2)^s ab] \\
&\quad - [wa(ab^2)^s ab][b] + [wa(ab^2)^s a] \\
&= [w(ab^2)^s ab]\big([ab] - [a][b]\big) - [wb(ab^2)^{s-1} ab] \\
&\quad + [wb^2(ab^2)^{s-1} ab][b] + [wa(ab^2)^s a].
\end{aligned}$$

If $s \ge 1$ this is strictly positive by the observation preceding the lemma and Lemma 5.2(2). This also holds when $s = 0$, since the sum of the two middle terms becomes $-[w] + [wb][b] > 0$.

The proof of (2) is similar, except that in the second expansion we work on the last ab^2. We have

$$\begin{aligned}[wab^3(ab^2)^sab] &= [wab(ab^2)^sab][b^2] - [wab^{-1}(ab^2)^sab] \\ &= [wab(ab^2)^sab][b]^2 - 2[wab(ab^2)^sab] \\ &\quad - [wa(ab^2)^sab][b] + [wab(ab^2)^sab] \\ &= [wab(ab^2)^sab][b]^2 - [wab(ab^2)^sab] - [wa(ab^2)^sab][b],\end{aligned}$$

and

$$\begin{aligned}[wab^2(ab^2)^sab^2] &= [wab^2(ab^2)^sab][b] - [wab^2(ab^2)^sa] \\ &= [wab(ab^2)^sab][b]^2 - [wa(ab^2)^sab][b] - [wab^2(ab^2)^sa].\end{aligned}$$

As above, subtracting the two end results we get

$$\begin{aligned}&[wab(ab^2)^sab] - [wab^2(ab^2)^sa] \\ &\quad = [wab(ab^2)^sab] - [wab(ab^2)^sa][b] + [wa(ab^2)^sa] \\ &\quad = [wab(ab^2)^sa][b] - [wab(ab^2)^sab^{-1}] - [wab(ab^2)^sa][b] + [wa(ab^2)^sa] \\ &\quad = [wa(ab^2)^sa] - [wab(ab^2)^sab^{-1}] \\ &\quad = [wa(ab^2)^sa] - [wab(ab^2)^s][ab^{-1}] + [wab(ab^2)^sba^{-1}].\end{aligned}$$

Since $[ab^{-1}] < 0$ and w is empty or beginning with a, this is positive. □

Lemma 6.2. *Under the same hypotheses of Lemma* 6.1 *we have*:

(1) $[wab(ab^2)^sab^4] < [wab^2(ab^2)^sab^3]$,
(2) $[wab^4(ab^2)^sab] < [wab^3(ab^2)^sab^2]$, *if* w *is empty or begins with* a.

Proof. (1) follows by Lemma 6.1(1), applied to the word bw. (2) Write $\widetilde{w}$ for the reversal of w. Then by Lemma 3.1(3) we obtain

$$\begin{aligned}[wab^4(ab^2)^sab] &= [ba(b^2a)^sb^4a\widetilde{w}] \\ &= [a\widetilde{w}b(ab^2)^sab^4] \\ &< [a\widetilde{w}b^2(ab^2)^sab^3] \qquad \text{(by (1))} \\ &= [wab^3a(b^2a)^sb^2] \\ &= [wab^3(ab^2)^sab^2].\end{aligned}$$

Note that the use of (1) in the third line is legitimate, since $\widetilde{w}$ ends with a, or is empty. □

Lemma 6.3. *Let* w *be a word that is empty or ends with* b, *and let* $k, h \geq 0$. *Then we have*

$$[ab^2wab(ab^2)^kab(ab^2)^hab] < [ab^2w(ab^2)^{k+h+2}].$$

Proof. On the left hand side we have

$$\begin{aligned}
&[ab^2wab(ab^2)^kab(ab^2)^hab] \\
&\quad = [bwab(ab^2)^kab(ab^2)^hab][ab] - [bwab(ab^2)^kab(ab^2)^h] \\
&\quad = [bwab(ab^2)^k(ab^2)^hab][ab]^2 - [bwab(ab^2)^kb(ab^2)^{h-1}ab][ab] \\
&\qquad - [bwab(ab^2)^kab(ab^2)^h] \\
&\quad = [bw(ab^2)^k(ab^2)^hab][ab]^3 - [bwb(ab^2)^{k+h-1}ab][ab]^2 \\
&\qquad - [bwab(ab^2)^kb(ab^2)^{h-1}ab][ab] - [bwab(ab^2)^kab(ab^2)^h] \\
&\quad = [w(ab^2)^{k+h+1}][ab]^3 - [wb(ab^2)^{k+h}][ab]^2 \\
&\qquad - [wab(ab^2)^kb(ab^2)^h][ab] - [bwab(ab^2)^kab(ab^2)^h].
\end{aligned} \tag{6.1}$$

On the other side we have

$$\begin{aligned}
[ab^2w(ab^2)^{k+h+2}] &= [ab^2w(ab^2)^{k+h}][(ab^2)^2] - [ab^2w(ab^2)^{k+h-2}] \\
&= [w(ab^2)^{k+h+1}][(ab^2)^2] - [w(ab^2)^{k+h-1}].
\end{aligned}$$

It is enough to show that $[w(ab^2)^{k+h-1}] < [wb(ab^2)^{k+h}][ab]^2$ and that $[ab]^3 < [(ab^2)^2]$. If $k+h \geq 1$ the first inequality is clear. If $k = h = 0$, it amounts to

$$[wb^{-1}(ab)^{-1}] < [wb][ab]^2,$$

or equivalently

$$[wb^{-1}][ab] < [wb][ab]^2 + [wb^{-1}ab],$$

which holds, since w is empty or ends with b.

We now show $[ab]^3 < [(ab^2)^2]$. We have

$$[ab]^3 = [b^2]^3 = [b^2]\big([b^4]+2\big) = [b^6] + 3[b^2].$$

Let $[b] - [a] = \Delta \geq 1$; then

$$\begin{aligned}
[(ab^2)^2] &= [ab^2]^2 - 2 = \big([ab][b] - [a]\big)^2 - 2 = \big([b^2][b] - [a]\big)^2 - 2 \\
&= \big([b^3] + \Delta\big)^2 - 2 = [b^3]^2 - 2 + 2[b^3]\Delta + \Delta^2 = [b^6] + 2[b^3]\Delta + \Delta^2.
\end{aligned}$$

We thus have to show $2[b^3]\Delta + \Delta^2 > 3[b^2]$, and it is enough to prove the case $\Delta = 1$, namely $2[b^3] - 3[b^2] + 1 > 0$. We compute

$$\begin{aligned}
2[b^3] - 3[b^2] + 1 &= 2T_3\big([b]\big) - 3T_2\big([b]\big) + 1 \\
&= 2[b]^3 - 3[b]^2 - 6[b] + 7.
\end{aligned}$$

Since the polynomial $2x^3 - 3x^2 - 6x + 7$ has three real roots, all of them strictly less than 3, and the trace of B is at least 3, the desired inequality follows. □

Lemma 6.4. *Let w be a word that is empty or begins with a, and let $s \geq 0$. Then we have*:

(1) $[wab^4(ab^2)^sab^3] < [w(ab^2)^{s+3}]$;
(2) $[wab^3(ab^2)^sab^4] < [w(ab^2)^{s+3}]$.

Proof. We prove (1). On the left side we have

$$\begin{aligned}
&[wab^4(ab^2)^s ab^3] \\
&\quad = [wab^2(ab^2)^s ab^3][b^2] - [wa(ab^2)^s ab^3] \\
&\quad = [wab^2(ab^2)^s ab^2][b][b^2] - [wab^2(ab^2)^s ab][b^2] - [wa(ab^2)^s ab^3] \\
&\quad = [w(ab^2)^{s+2}][b][ab] - [w(ab^2)^{s+1}ab][b^2] - [wa(ab^2)^{s+1}b] \\
&\quad = [w(ab^2)^{s+2}][ab^2] + [w(ab^2)^{s+2}][a] \\
&\qquad - [w(ab^2)^{s+1}ab][b^2] - [wa(ab^2)^{s+1}b],
\end{aligned}$$

while on the other side we have

$$[w(ab^2)^{s+3}] = [w(ab^2)^{s+2}][ab^2] - [w(ab^2)^{s+1}].$$

We obtain $[w(ab^2)^{s+1}] < [wa(ab^2)^{s+1}b]$ from Lemma 5.2(2). We complete the proof by computing

$$\begin{aligned}
&[w(ab^2)^{s+1}ab][b^2] - [w(ab^2)^{s+2}][a] \\
&\quad = [w(ab^2)^{s+1}ab][ab] - [w(ab^2)^{s+1}abab] - [w(ab^2)^{s+1}aba^{-1}b] \\
&\quad = [w(ab^2)^{s+1}abab] + [w(ab^2)^{s+1}] - [w(ab^2)^{s+1}abab] - [w(ab^2)^{s+1}aba^{-1}b] \\
&\quad = [w(ab^2)^{s+1}] - [w(ab^2)^{s+1}aba^{-1}b] \\
&\quad = [w(ab^2)^{s+1}] - [w(ab^2)^{s+1}ab][a^{-1}b] + [w(ab^2)^{s+1}a^2],
\end{aligned}$$

whose end result is positive since $[a^{-1}b] < 0$.

(2) can be obtained from (1) as in the proof of Lemma 6.2. □

Lemma 6.5. *Let w be a word that is empty or ends with b, and let $k, h \geq 0$. Then we have*

$$[wab^4(ab^2)^k ab^4(ab^2)^h ab^4] < [w(ab^2)^{k+h+5}].$$

Proof. On the left side we have

$$\begin{aligned}
&[wab^4(ab^2)^k ab^4(ab^2)^h ab^4] \\
&= [wab^4(ab^2)^k ab^4(ab^2)^h ab^2][b^2] - [wab^4(ab^2)^k ab^4(ab^2)^h a] \\
&= [wab^4(ab^2)^k ab^2(ab^2)^h ab^2][b^2]^2 - [wab^4(ab^2)^k a(ab^2)^h ab^2][b^2] \\
&\qquad - [wab^4(ab^2)^k ab^4(ab^2)^h a] \\
&= [wab^2(ab^2)^k ab^2(ab^2)^h ab^2][b^2]^3 - [wa(ab^2)^k ab^2(ab^2)^h ab^2][b^2]^2 \\
&\qquad - [wab^4(ab^2)^k a(ab^2)^h ab^2][b^2] - [wab^4(ab^2)^k ab^4(ab^2)^h a] \\
&= [w(ab^2)^{k+h+3}][b^2]^3 - [wa(ab^2)^{k+h+2}][b^2]^2 \\
&\qquad - [wab^4(ab^2)^k a(ab^2)^{h+1}][b^2] - [wab^4(ab^2)^k ab^4(ab^2)^h a].
\end{aligned}$$

On the other side we have

$$[w(ab^2)^{k+h+5}] = [w(ab^2)^{k+h+3}][(ab^2)^2] - [w(ab^2)^{k+h+1}].$$

The second end result is strictly greater than the first, because $[w(ab^2)^{k+h+1}] < [wa(ab^2)^{k+h+2}][b^2]^2$ by Lemma 5.2(2), and $[b^2]^3 = [ab]^3 < [(ab^2)^2]$ by the proof of Lemma 6.3. □

Lemma 6.6. *Let w be a word that begins with ab or is empty, and let $k, h \geq 0$. Then*

$$[wab^3(ab^2)^k ab^3(ab^2)^h ab^3] < [w(ab^2)^{k+h+4}].$$

Proof. This time we work on the term on the right hand side; the step in the middle of the following identity chain results from $[(ab)b^2] + [(ab)b^{-2}] = [ab][b^2] = [ab]^2$.

$$\begin{aligned}
&[w(ab^2)^{k+h+4}] \\
&\quad = [wab(ba)bb(ab^2)^{k+h+2}] \\
&\quad = [wab^3(ab^2)^{k+h+2}][ba] - [waba^{-1}b(ab^2)^{k+h+2}] \\
&\quad = [wab^3(ab^2)^k ab^2(ab)b(ab^2)^h][ab] - [waba^{-1}b(ab^2)^{k+h+2}] \\
&\quad = [wab^3(ab^2)^k ab^3(ab^2)^h][ab]^2 - [wab^3(ab^2)^k aba^{-1}b(ab^2)^h][ab] \\
&\qquad - [waba^{-1}b(ab^2)^{k+h+2}] \\
&\quad = [wab^3(ab^2)^k ab^3(ab^2)^h]\big([ab^3] + [ab^{-1}]\big) \\
&\qquad - [wab^3(ab^2)^k aba^{-1}b(ab^2)^h][ab] - [waba^{-1}b(ab^2)^{k+h+2}] \\
&\quad = [wab^3(ab^2)^k ab^3(ab^2)^h ab^3] + [wab^3(ab^2)^k ab^3(ab^2)^{h-1}ab^{-1}a^{-1}] \\
&\qquad + [wab^3(ab^2)^k ab^3(ab^2)^h][ab^{-1}] \\
&\qquad - [wab^3(ab^2)^k aba^{-1}b(ab^2)^h][ab] - [waba^{-1}b(ab^2)^{k+h+2}].
\end{aligned}$$

Now, the first summand of the end result is the left side of the desired inequality, and the second is positive due to our hypotheses on w. We develop the fourth summand, in order to make the third appear:

$$\begin{aligned}
&[wab^3(ab^2)^k aba^{-1}b(ab^2)^h][ab] \\
&\quad = [wab^3(ab^2)^k ab(ab^2)^h][ab][a^{-1}b] - [wab^3(ab^2)^k a^2(ab^2)^h][ab] \\
&\quad = [wab^3(ab^2)^k ab(ab^2)^h][b^2][a^{-1}b] - [wab^3(ab^2)^k a^2(ab^2)^h][ab] \\
&\quad = [wab^3(ab^2)^k ab^3(ab^2)^h][a^{-1}b] + [wab^3(ab^2)^k ab^{-1}(ab^2)^h][a^{-1}b] \\
&\qquad - [wab^3(ab^2)^k a^2(ab^2)^h][ab] \\
&\quad = [wab^3(ab^2)^k ab^3(ab^2)^h][ab^{-1}] \\
&\qquad + [wab^3(ab^2)^k a(ab^2)^h][a^{-1}b][b] - [wab^3(ab^2)^k ab(ab^2)^h][a^{-1}b] \\
&\qquad - [wab^3(ab^2)^k a^2(ab^2)^h][ab].
\end{aligned}$$

We are then left with proving that the sum

$$\begin{aligned}
&- [wab^3(ab^2)^k a(ab^2)^h][a^{-1}b][b] + [wab^3(ab^2)^k ab(ab^2)^h][a^{-1}b] \\
&\quad + [wab^3(ab^2)^k a^2(ab^2)^h][ab] - [wab(ab^2)^{k+h+2}][a^{-1}b] + [wa^2(ab^2)^{k+h+2}]
\end{aligned}$$

is positive. The first, third, and last summand surely are, and so is the sum or the second with the fourth, because

$$[wab^3(ab^2)^k ab(ab^2)^h] < [w(ab^2)^{k+h+2}] < [wab(ab^2)^{k+h+2}],$$

by Lemma 6.1(2) and Lemma 5.2(2). □

We finally arrive at §1(IV.2).

Theorem 6.7. *Let $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ be well oriented, and assume $2 \le \mathrm{tr}(A) < \mathrm{tr}(B)$ and $\mathrm{tr}(AB) = \mathrm{tr}(B^2)$. The ab^2 is the only optimal word.*

Proof. Let u be a word which is trace-maximizing among words of the same length; by the remarks following Equation (2.1) we may assume, possibly replacing u with its cube, that this length is a multiple of 3. We have to prove that u is a power of a rotation of ab^2. By Lemma 5.5, u does not contain a^2 as a factor, up to rotations. We have

$$\begin{aligned}
[wab^5] =& [wab^2][b^3] - [wab^{-1}] \\
&= [wab^2]\big([bb][b] - [b]\big) - [wab^{-1}] \\
&= [wab^2]\big([ab][b] - [b]\big) - [wab^{-1}] \\
&\le [wab^2]\big([ab][b] - [a] - 1\big) - [wab^{-1}] \\
&= [wab^2]\big([ab^2] - 1\big) - [wab^{-1}] \\
&= [w(ab^2)^2] + [w] - [wab^2] - [wab^{-1}] \\
&= [w(ab^2)^2] + [w] - [wab^2] - [wa][b] + [wab] \\
&< [w(ab^2)^2],
\end{aligned}$$

and therefore b^5 is also excluded. Moreover

$$[b^3] = [bb][b] - [b] = [ab][b] - [b] = [ab^2] + [a] - [b] < [ab^2]$$

shows that u is not a power of b. Summing up, u uniquely factorizes as a product of ab, ab^2, ab^3, and ab^4; we refer to this as the *syllabic decomposition* of u.

Suppose that ab occurs as a syllable (occurrences are always intended up to rotations). Since the length is a multiple of 3, at least one of the following cases must hold:

- one of ab^3 and ab^4 occurs as well,
- ab occurs at least thrice.

These occurrences will be separated by zero or more occurrences of ab^2. In any case, Lemmas 6.1, 6.2, and 6.3 apply, and u is not trace-maximizing. Thus the syllable ab does not occur in u.

Suppose ab^4 occurs. Then

- either ab^3 occurs as well,
- or ab^4 occurs at least thrice.

Lemmas 6.4 and 6.5 treat these cases, and thus exclude ab^4.

We have established that the only syllables occurring in u are ab^2 and ab^3. If the latter occurs, it must do so at least thrice, and Lemma 6.6 applies. We are then left with occurrences of ab^2 only, and the proof is complete. □

7. Case (IV.3)

The standing assumptions in this final section are that the well oriented pair $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ satisfies $2 \le \mathrm{tr}(A) < \mathrm{tr}(B)$ and $\mathrm{tr}(AB) > \mathrm{tr}(B^2)$. First, a dichotomy.

Lemma 7.1. *Under the above assumptions the numbers* $[(ab)^3]$ *and* $[(ab^2)^2]$ *differ by at least* 2*.*

Proof. We work modulo $[ab]$. We have

$$[(ab)^3] \equiv [(ab)^2][ab] - [ab] \equiv 0,$$

and also

$$[(ab(ba)b^2] \equiv [ab^3][ba] - [aba^{-1}b] \equiv -[ab][a^{-1}b] + [a^2] \equiv [a^2].$$

Taking the difference we get

$$[(ab^2)^2] - [(ab)^3] \equiv [a^2].$$

Now, the coset $[a^2] + \mathbb{Z}[ab]$ does not intersect the interval $\{-1, 0, 1\}$. Indeed, its points closest to zero are $[a^2] \ge 2$ and $[a^2] - [ab] \le [a^2] - [b^2] - 1 \le -2$. □

The case $[(ab)^3] > [(ab^2)^2]$ of the dichotomy follows familiar patterns.

Lemma 7.2. *Assume* $[(ab^2)^2] \le [(ab)^3] - 2$*. Fix a word* w*, and let* $k \ge 1$ *be such that the length of* $wab^2(ab)^k ab^2$ *is a multiple of* 6*. If* w *is empty, or is a product of* ab *and* ab^2*, then we have*

$$[wab^2(ab)^k ab^2] < [w(ab)^{k+3}].$$

Proof. On the left side we have

$$\begin{aligned}
&[wab^2(ab)^k ab^2] \\
&= [wab^2(ab)^k][ab^2] - [wab^2(ab)^{k-1}ab^{-1}a^{-1}] \\
&= [w(ab)^k][ab^2]^2 - [wb^{-1}(ab)^{k-1}][ab^2] - [wab^2(ab)^{k-1}ab^{-1}a^{-1}] \\
&= [w(ab)^k]\big([(ab^2)^2] + 2\big) - [wb^{-1}(ab)^{k-1}][ab^2] - [wab^2(ab)^{k-1}ab^{-1}a^{-1}] \\
&\le [w(ab)^k][(ab)^3] - [wb^{-1}(ab)^{k-1}][ab^2] - [wab^2(ab)^{k-1}ab^{-1}a^{-1}] \\
&= [w(ab)^{k+3}] + [w(ab)^{k-3}] - [wb^{-1}(ab)^{k-1}][ab^2] - [wab^2(ab)^{k-1}ab^{-1}a^{-1}].
\end{aligned}$$

We then have to prove that the sum of all but the first summands of the above end result is negative. We will prove this fact by showing the following two inequalities:

$$\begin{aligned}
[w(ab)^{k-3}] &\le [wb^{-1}(ab)^{k-1}][ab^2], \\
[wab^2(ab)^{k-1}ab^{-1}a^{-1}] &> 0.
\end{aligned}$$

Suppose $k \geq 3$. If w is non empty, it begins with ab and ends with b. Thus all the inverted letters simplify and the inequalities follow from Lemma 5.2(2). This also holds when w is the empty word, with different simplifications.

Suppose $k < 3$. Then w cannot be empty, for reasons of length, and the second inequality causes no problems. If $k = 2$ the first inequality becomes $[wb^{-1}a^{-1}] \leq [wb^{-1}ab][ab^2]$, clearly true. If $k = 1$ it becomes $[w(ab)^{-2}] \leq [wb^{-1}][ab^2]$, which is also true since $[w(ab)^{-2}] = [wb^{-1}a^{-1}][ab] - [w]$. □

The other case $[(ab)^3] < [(ab^2)^2]$ will be treated via the following lemma.

Lemma 7.3. *Assume* $[(ab)^3] \leq [(ab^2)^2] - 2$. *Fix a word* w, *and let* $k, h \geq 0$ *be such that the length of* $ab^2wab(ab^2)^kab(ab^2)^hab$ *is a multiple of* 6. *If* w *is empty, or* w *is a product of* ab *and* ab^2, *then we have*

$$[ab^2wab(ab^2)^kab(ab^2)^hab] < [ab^2w(ab^2)^{k+h+2}].$$

Unfortunately, although the end inequality is the same, the proof of Lemma 7.3 is harder than that of its twin Lemma 6.3. This is due to the fact that the final step in the proof of Lemma 7.3, namely the inequality $[ab]^3 \leq [(ab^2)^2]$, may fail; for example, it fails for the matrices $A = L^{11}$, $B = LNL$ cited in §5. We have thus to make do with the weaker $[ab]^3 \leq [(ab^2)^2] + [ab]$, that can still be quite narrow: for the above matrices we have

$$[ab]^3 = 3375 \leq 3377 = 3362 + 15 = [(ab^2)^2] + [ab].$$

The following lemma establishes that weaker inequality, as well as the fact that the minimal difference $\mathrm{tr}(AB) = 15 > 14 = \mathrm{tr}(B^2)$ for the above matrix pair is no coincidence.

Lemma 7.4. *Assume* $[(ab^2)^2] \geq [(ab)^3] + 2$. *Then the following formulas hold*:

$$[ab] = [b^2] + 1, \tag{7.1}$$

$$2[a] \leq [b], \tag{7.2}$$

$$[ab]^3 \leq [(ab^2)^2] + [ab]. \tag{7.3}$$

Proof. Let $\alpha = [a]$, $\beta = [b]$, $x = [ab]$; then we have $[(ab)^3] = T_3(x) = x^3 - 3x$ and $[(ab^2)^2] = T_2(\beta x - \alpha) = (\beta x - \alpha)^2 - 2$.

Assuming the negation of (7.1), we have $x \geq [b^2] + 2 = \beta^2$ and thus

$$\begin{aligned}[(ab)^3] - [(ab^2)^2] &= x^3 - \beta^2 x^2 + (4\beta - 3)x - 2 \\ &\geq (4\beta - 3)x - 2.\end{aligned}$$

Since β and x are both at least 3, the last term is positive, which is a contradiction; this establishes (7.1).

Applying (7.1) and its equivalent form $[b]^2 = [ab] + 1$ several times, we compute

$$\begin{aligned}
[(ab^2)^2] &= [ab^2ab][b] - [a^2b^2] \\
&= [abab][b]^2 - [a^2b][b] - [a^2b][b] + [a^2] \\
&= [(ab)^2]\big([ab] + 1\big) - 2\big([a][ab][b] - [b]^2\big) + [a]^2 - 2 \\
&= [(ab)^3] + [ab] + [(ab)^2] - 2[a][ab][b] + 2[b]^2 + [a]^2 - 2 \\
&= [(ab)^3] + \beta^2 - 1 + (\beta^2 - 1)^2 - 2 - 2\alpha\beta(\beta^2 - 1) + 2\beta^2 + \alpha^2 - 2 \\
&= [(ab)^3] + \beta^4 - 2\alpha\beta^3 + \beta^2 + 2\alpha\beta + \alpha^2 - 4;
\end{aligned}$$

therefore our hypothesis yield

$$2\alpha\beta^3 - \beta^4 \leq \beta^2 + 2\alpha\beta + \alpha^2 - 6.$$

We change variables by setting $\alpha = \lambda\beta$, and obtain

$$\begin{aligned}
(2\lambda - 1)\beta^4 &\leq (1 + 2\lambda + \lambda^2)\beta^2 - 6, \\
(2\lambda - 1)\beta &< \frac{(1+\lambda)^2}{\beta}, \\
2\lambda\beta &< \beta + \frac{(1+\lambda)^2}{\beta}, \\
2\alpha &\leq \beta + \left\lfloor \frac{(1+\lambda)^2}{\beta} \right\rfloor,
\end{aligned}$$

the last step justified by the fact that $2\lambda\beta = 2\alpha$ is an integer. If $\beta = 3$ then $\alpha = 2$ and $\lambda = 2/3$, leading to $4 \leq 3$, a contradiction. Therefore $\beta \geq 4$ and the floor part is zero since $1 + \lambda < 2$; this proves (7.2).

We previously computed that

$$[(ab^2)^2] - [(ab)^3] = \beta^4 - 2\alpha\beta^3 + \beta^2 + 2\alpha\beta + \alpha^2 - 4.$$

Since $[ab]^3 = [(ab)^3] + 3[ab]$ and $[ab] = \beta^2 - 1$ we obtain

$$\begin{aligned}
[ab]^3 &= [(ab^2)^2] - \beta^4 + 2\alpha\beta^3 - \beta^2 - 2\alpha\beta - \alpha^2 + 4 + 3\beta^2 - 3 \\
&= [(ab^2)^2] - \beta^4 + 2\alpha\beta^3 + 2\beta^2 - 2\alpha\beta - \alpha^2 + 1 \\
&= [(ab^2)^2] + (2\alpha - \beta)(\beta^3 - \beta) + \beta^2 - \alpha^2 + 1 \\
&\leq [(ab^2)^2] + \beta^2 - \alpha^2 + 1 \\
&\leq [(ab^2)^2] + [ab],
\end{aligned}$$

thus settling (7.3). □

Proof of Lemma 7.3. Continuing from (6.1) (whose proof does not depend on the relative values of $[a]$, $[b]$, $[ab]$) and applying (7.3), we obtain

$$\begin{aligned}
[ab^2wab(ab^2)^kab(ab^2)^hab] &= [w(ab^2)^{k+h+1}][ab]^3 - [wb(ab^2)^{k+h}][ab]^2 \\
&\quad - [wab(ab^2)^kb(ab^2)^h][ab] - [bwab(ab^2)^kab(ab^2)^h] \\
&\leq [w(ab^2)^{k+h+1}][(ab^2)^2] \\
&\quad + [w(ab^2)^{k+h+1}][ab] - [wb(ab^2)^{k+h}][ab]^2 \\
&\quad - [wab(ab^2)^kb(ab^2)^h][ab] - [bwab(ab^2)^kab(ab^2)^h] \\
&= [w(ab^2)^{k+h+3}] + [w(ab^2)^{k+h-1}] \\
&\quad + [w(ab^2)^{k+h+1}][ab] - [wb(ab^2)^{k+h}][ab]^2 \\
&\quad - [wab(ab^2)^kb(ab^2)^h][ab] - [bwab(ab^2)^kab(ab^2)^h].
\end{aligned}$$

Thus we need to prove that

$$\begin{aligned}
&[w(ab^2)^{k+h-1}] + [w(ab^2)^{k+h+1}][ab] \\
&< [wb(ab^2)^{k+h}][ab]^2 + [wab(ab^2)^kb(ab^2)^h][ab] + [bwab(ab^2)^kab(ab^2)^h].
\end{aligned} \tag{7.4}$$

We will need to switch the positions of some factors, and this will be accomplished by the formula

$$[xyzw] = [xzyw] + [xz^{-1}yw] - [xyz^{-1}w], \tag{7.5}$$

which moves z to the left; of course, an analogous formula holds for moving to the right. Formula (7.5) follows from Lemma 3.1(2), since both $[xyzw] + [xyz^{-1}w]$ and $[xzyw] + [xz^{-1}yw]$ equal $[xyw][z]$.

In order to obtain (7.4), we work on the second summand of the second line. Moving b to the left we get

$$\begin{aligned}
[wab(ab^2)^kb(ab^2)^h][ab] &= [w(ab^2)^{k+h+1}][ab] + [wa(ab^2)^{k+h}][ab] \\
&\quad - [wab(ab^2)^{k-1}ab(ab^2)^h][ab] \\
&= [w(ab^2)^{k+h+1}][ab] + [wa(ab^2)^{k+h}][ab] \\
&\quad - [wab(ab^2)^{k-1}abab(ab^2)^h] - [wab(ab^2)^{k+h-1}].
\end{aligned}$$

Substituting this back into (7.4), the summand $[w(ab^2)^{k+h+1}][ab]$ simplifies. The summand $[wab(ab^2)^{k-1}(ab)^2(ab^2)^h]$ is, by Lemma 5.2(2), always dominated by the last term of (7.4), even when $k = h = 0$. Removing both of them we remain with the inequality

$$\begin{aligned}
&[w(ab^2)^{k+h-1}] + [wab(ab^2)^{k+h-1}] \\
&\qquad < [wb(ab^2)^{k+h}][ab][ab] + [wa(ab^2)^{k+h}][ab],
\end{aligned}$$

which holds by Lemma 5.2(2). The case $k = h = 0$ must be checked apart, but causes no problems. □

Theorem 7.5. *Let $A, B \in \mathrm{SL}_2\,\mathbb{Z}$ be well oriented, and assume $2 \leq \mathrm{tr}(A) < \mathrm{tr}(B)$ and $\mathrm{tr}(AB) > \mathrm{tr}(B^2)$. If $\mathrm{tr}((AB)^3) > \mathrm{tr}((AB^2)^2)$, then ab is the only optimal word; otherwise, so is ab^2*

Proof. As in the proof of Theorem 6.7, let u be a word of length a multiple of 6 which is trace-maximizing among all words of the same length. The statement will result by proving that u is a power of ab (in case $[(ab)^3] > [(ab^2)^2]$), or of ab^2 (in case $[(ab)^3] < [(ab^2)^2]$, no equality being possible by Lemma 7.1).

Since $[b^2] < [ab]$, at least one a appears in u, but a^2 does not by Lemma 5.5. We claim that the factor b^3 is also excluded. Indeed, for every w we have

$$\begin{aligned}[wab^3] = [wab][b^2] - [wab^{-1}] &\leq [wab][ab] - [wab] - [wab^{-1}] \\ &= [w(ab)^2] + [w] - [wa][b] < [w(ab)^2].\end{aligned}$$

Therefore, u factors uniquely as a product of the syllables ab and ab^2.

Suppose that both syllables appear; since $|u|$ is a multiple of 6, ab^2 must appear at least two times, and ab at least three times. If $ab^2 \prec ab$ then Lemma 7.2 contradicts the maximality of u, and the same does Lemma 7.3 if $ab \prec ab^2$. Thus only one syllable appears in u, and the proof is complete. □

References

[1] A. F. Beardon. *The geometry of discrete groups*. Springer, 1995.

[2] M. A. Berger and Y. Wang. Bounded semigroups of matrices. *Linear Algebra Appl.*, 166:21–27, 1992.

[3] V. D. Blondel, J. Theys, and A. A. Vladimirov. An elementary counterexample to the finiteness conjecture. *SIAM J. Matrix Anal. Appl.*, 24(4):963–970, 2003.

[4] J. Bochi. Ergodic optimization of Birkhoff averages and Lyapunov exponents. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. III. Invited lectures*, pages 1825–1846. World Sci. Publ., 2018.

[5] T. Bousch and J. Mairesse. Asymptotic height optimization for topical IFS, Tetris heaps, and the finiteness conjecture. *J. Amer. Math. Soc.*, 15(1):77–111, 2002.

[6] J. L. Brenner and A. Charnow. Free semigroups of 2×2 matrices. *Pacific J. Math.*, 77(1):57–69, 1978.

[7] E. Breuillard and C. Sert. The joint spectrum. *J. London Math. Soc.*, 103(3):943–990, 2021.

[8] A. Cicone, N. Guglielmi, S. Serra-Capizzano, and M. Zennaro. Finiteness property of pairs of 2×2 sign-matrices via real extremal polytope norms. *Linear Algebra Appl.*, 432(2-3):796–816, 2010.

[9] H. S. M. Coxeter. Inversive geometry. *Educational Studies in Mathematics*, 3(3/4):310–321, 1971.

[10] I. Gekhtman, S. J. Taylor, and G. Tiozzo. A central limit theorem for random closed geodesics: proof of the Chas-Li-Maskit conjecture. *Adv. Math.*, 358:106852, 18 pp., 2019.

[11] S. Geninska and E. Leuzinger. A geometric characterization of arithmetic Fuchsian groups. *Duke Math. J.*, 142(1):111–125, 2008.

[12] N. Guglielmi and M. Zennaro. Stability of linear problems: joint spectral radius of sets of matrices. In *Current challenges in stability issues for numerical differential equations*, volume 2082 of *Lecture Notes in Math.*, pages 265–313. Springer, 2014.

[13] K. G. Hare, I. D. Morris, N. Sidorov, and J. Theys. An explicit counterexample to the Lagarias-Wang finiteness conjecture. *Adv. Math.*, 226(6):4667–4701, 2011.

[14] R. D. Horowitz. Characters of free groups represented in the two-dimensional special linear group. *Comm. Pure Appl. Math.*, 25:635–649, 1972.

[15] O. Jenkinson and M. Pollicott. Joint spectral radius, Sturmian measures and the finiteness conjecture. *Ergodic Theory Dynam. Systems*, 38(8):3062–3100, 2018.
[16] T. Jørgensen and K. Smith. On certain semigroups of hyperbolic isometries. *Duke Math. J.*, 61(1):1–10, 1990.
[17] R. Jungers. *The joint spectral radius*, volume 385 of *Lecture Notes in Control and Information Sciences*. Springer, 2009.
[18] R. M. Jungers and V. D. Blondel. On the finiteness property for rational matrices. *Linear Algebra Appl.*, 428(10):2283–2295, 2008.
[19] S. Katok. *Fuchsian groups*. University of Chicago Press, 1992.
[20] V. Kozyakin. Hourglass alternative and the finiteness conjecture for the spectral characteristics of sets of non-negative matrices. *Linear Algebra Appl.*, 489:167–185, 2016.
[21] V. Kozyakin. Minimax joint spectral radius and stabilizability of discrete-time linear switching control systems. *Discrete Contin. Dyn. Syst. Ser. B*, 24(8):3537–3556, 2019.
[22] J. C. Lagarias and Y. Wang. The finiteness conjecture for the generalized spectral radius of a set of matrices. *Linear Algebra Appl.*, 214:17–42, 1995.
[23] W. Luo and P. Sarnak. Number variance for arithmetic hyperbolic surfaces. *Comm. Math. Phys.*, 161(2):419–432, 1994.
[24] E. Oregón-Reyes. Properties of sets of isometries of Gromov hyperbolic spaces. *Groups Geom. Dyn.*, 12(3):889–910, 2018.
[25] G. Panti. Billiards on pythagorean triples and their Minkowski functions. *Discrete Contin. Dyn. Syst.*, 40(7):4341–4378, 2020.
[26] F. Przytycki. Ergodicity of toral linked twist mappings. *Ann. Sci. École Norm. Sup. (4)*, 16(3):345–354, 1983.
[27] J. N. Tsitsiklis and V. D. Blondel. The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. *Math. Control Signals Systems*, 10(1):31–40, 1997.

University of Udine, Department of Mathematics, Computer Science and Physics, via delle Scienze 206, 33100 Udine
Email address: giovanni.panti@uniud.it

Vrije Universiteit Amsterdam, Faculteit der Bètawetenschappen, De Boelelaan 1111, 1081HV Amsterdam
Email address: davide.sclosa@gmail.com