

REPRESENTING THE SPECIAL LINEAR GROUP WITH BLOCK UNITRIANGULAR MATRICES

JOHN URSCHHEL

ABSTRACT. We prove that every element of the special linear group can be represented as the product of at most six block unitriangular matrices, and that there exist matrices for which six products are necessary, independent of indexing. We present an analogous result for the general linear group. These results serve as general statements regarding the representational power of alternating linear updates. The factorizations and lower bounds of this work immediately imply tight estimates on the expressive power of linear affine coupling blocks in machine learning.

INTRODUCTION

Let \mathbb{F} be an arbitrary field and $M_{m,n}(\mathbb{F})$ denote the set of $m \times n$ matrices with coefficients in \mathbb{F} . When $m = n$, we simply write $M_n(\mathbb{F})$. Let $GL_n(\mathbb{F})$ denote the group of $n \times n$ non-singular matrices with coefficients in \mathbb{F} , and $SL_n(\mathbb{F})$ denote the subgroup of $GL_n(\mathbb{F})$ consisting of matrices with determinant one. Let I denote the identity matrix and 0 denote the zero matrix; the dimension of each is always clear from context. Let S_n denote the symmetric group and P_π denote the permutation $\pi \in S_n$ of the columns of I . For $X \subset M_n(\mathbb{F})$, let $X^k \subset M_n(\mathbb{F})$ denote the set of k -fold products of elements of X . Let $BL_{m,n}(\mathbb{F})$ and $BU_{m,n}(\mathbb{F})$ denote the subgroups of $SL_{m+n}(\mathbb{F})$ consisting of block lower and upper unitriangular matrices, respectively, with block partition $\{1, \dots, m\}$ and $\{m+1, \dots, m+n\}$:

$$BL_{m,n}(\mathbb{F}) = \left\{ \left[\begin{array}{c|c} I & 0 \\ A & I \end{array} \right] \mid A \in M_{n,m}(\mathbb{F}) \right\},$$

$$BU_{m,n}(\mathbb{F}) = \left\{ \left[\begin{array}{c|c} I & A \\ 0 & I \end{array} \right] \mid A \in M_{m,n}(\mathbb{F}) \right\}.$$

In this work, we prove the existence of the following block unitriangular factorization of the special linear group.

Theorem 1. *For every $M \in SL_{2n}(\mathbb{F})$, there exists $A_1, \dots, A_6 \in M_n(\mathbb{F})$ such that*

$$M = \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix} \begin{bmatrix} I & A_6 \\ 0 & I \end{bmatrix}.$$

For every $m+n > 3$, there exists some $M \in SL_{m+n}(\mathbb{F})$ such that $M \notin [BL_{m,n}(\mathbb{F}) \cup BU_{m,n}(\mathbb{F})]^5$. Furthermore, if \mathbb{F} has at least four elements, then the lower bound holds

SOCIETY OF FELLOWS, HARVARD UNIVERSITY, CAMBRIDGE, MA
DEPARTMENT OF MATHEMATICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MA

E-mail address: urschel@mit.edu.

independent of indexing: there exists $M \in \text{SL}_{m+n}(\mathbb{F})$ such that $P_\pi M P_{\pi^{-1}} \notin [\text{BL}_{m,n}(\mathbb{F}) \cup \text{BU}_{m,n}(\mathbb{F})]^5$ for all permutations $\pi \in \text{S}_{m+n}$.

We have an analogous theorem for the general linear group. Let $\text{T}_{m,n}(\mathbb{F})$ denote the set of matrices of the form $\begin{bmatrix} B & 0 \\ A & C \end{bmatrix}$ or $\begin{bmatrix} B & A \\ 0 & C \end{bmatrix}$, with $B \in \text{M}_m(\mathbb{F})$ and $C \in \text{M}_n(\mathbb{F})$ both diagonal. We prove the following result.

Theorem 2. *For every $M \in \text{GL}_{2n}(\mathbb{F})$, there exists $A_1, \dots, A_6 \in \text{M}_n(\mathbb{F})$ such that*

$$M = \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix} \begin{bmatrix} D & A_6 \\ 0 & I \end{bmatrix}$$

for any diagonal $D \in \text{GL}_n(\mathbb{F})$ with $\det(D) = \det(M)$. Furthermore, for every $m+n > 3$, there exists $M \in \text{SL}_{m+n}(\mathbb{F})$ such that $M \notin [\text{T}_{m+n}(\mathbb{F})]^5$.

The factorizations of Theorems 1 and 2 are efficiently computable. Our construction is fairly unique among block matrix factorizations.¹ It can be viewed as a generalized version of a block LU factorization. One benefit of our construction is that such a factorization always exists, whereas the existence of a block LU factorization relies on the invertibility of the upper left block. Moreover, the above theorems serve as general results regarding the representational power of alternating linear updates.

In fact, Theorem 2 immediately solves an open problem regarding affine coupling networks in machine learning. An affine coupling block is a function $f : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+n}$ of the form $f(x_u, x_v) = (x_u, x_v \odot \exp(s(x_u)) + t(x_u))$, where $x_u \in \mathbb{R}^m$, $x_v \in \mathbb{R}^n$ (typically, $m \approx n$), $s, t : \mathbb{R}^m \rightarrow \mathbb{R}^n$, and \odot is the entry-wise product. A version of these functions (with $s = 0$) was originally introduced by Dinh, Krueger, and Y. Bengio in their NICE deep learning model [2]. Dinh, Sohl-Dickstein, and S. Bengio expanded that work to real NVP (non-volume preserving) transformations (e.g., the general formulation above) [3]. These papers led, in part, to the popularization of normalizing flows in machine learning, a general class of diffeomorphisms that map some standard distribution (say, a standard Gaussian vector) to a more complex one. We refer the reader to [6, 8] for two excellent surveys of this emerging field. As discussed in [8], the “question of foremost importance” is the expressive power of such models, even when restricted to simple inputs. Recently, Koehler, Mehta, and Risteski studied the expressive power of linear affine couplings, i.e., matrices of the form $\begin{bmatrix} I & 0 \\ A & D \end{bmatrix}$ and $\begin{bmatrix} D & A \\ 0 & I \end{bmatrix}$ for $A \in \text{M}_n(\mathbb{R})$ and diagonal $D \in \text{GL}_n(\mathbb{R})$ with strictly positive entries [7]. They posed the following question: how many linear affine coupling layers are needed to represent an arbitrary orientation-preserving matrix? They produced a 47-layer construction and showed that at least five layers are necessary. Theorem 2 answers their question exactly.

Corollary 3. *Every matrix $M \in \text{GL}_{2n}(\mathbb{R})$ with $\det(M) > 0$ can be represented by a depth-six linear affine coupling network. In addition, for every $n > 1$, there exists $M \in \text{SL}_{2n}(\mathbb{R})$ for which six layers are necessary, i.e., M cannot be exactly represented by a depth-five linear affine coupling network.*

The lower bound of Corollary 3 immediately implies one for the non-linear setting, by considering the function Mx , and applying a Jacobian argument; see [7, Corollary

¹One somewhat similar construction is the $LU\tilde{L}$ factorization proposed by Serre and Püschel, where L, \tilde{L} are block lower unitriangular and U is block upper triangular; the authors studied how close (in a rank sense) such a factorization can be to block diagonal [9].

6] for details. Theorem 1 gives an analogous result for the aforementioned NICE model (where $D = I$), with an additional lower bound independent of indexing, e.g., a learned partition cannot do uniformly better than an arbitrary one. The improvement in construction from depth 47 to depth six leads to a significant practical difference in terms of architecture. For example, since permutation matrices can be represented with six layers, the choice of partition may be of limited importance. Furthermore, the improved construction has consequences for maximum likelihood estimation, as Corollary 3 implies that the distributions representable as the application of a six-layer linear affine coupling network to $N(0, I)$ are exactly the set of $N(0, \Sigma)$ with Σ invertible; see [7, Appendix A.2] for details.

PROOF OF THEOREMS 1 AND 2

In what follows, we make use of the theory of commutators, i.e., elements of a group G of the form $[g, h] := g^{-1}h^{-1}gh$ for some $g, h \in G$. We recall the following consequence of a combination of results of R.C. Thompson.

Lemma 4 ([10, 11]). *If $\mathrm{SL}_n(\mathbb{F}) \neq \mathrm{SL}_2(\mathrm{GF}(2))$, then every element is a commutator of $\mathrm{GL}_n(\mathbb{F})$.*

Furthermore, given $A \in \mathrm{SL}_n(\mathbb{F}) \neq \mathrm{SL}_2(\mathrm{GF}(2))$, $X, Y \in \mathrm{GL}_n(\mathbb{F})$ satisfying $A = [X, Y]$ are efficiently computable; see [10, 11] for details. Using Lemma 4 and well-chosen block unitriangular matrices, we produce a five-layer decomposition for matrices with a non-singular upper right block.

Lemma 5. *Let $M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \in \mathrm{SL}_{2n}(\mathbb{F}) \neq \mathrm{SL}_4(\mathrm{GF}(2))$ and $M_2 \in \mathrm{GL}_n(\mathbb{F})$. Then*

$$M = \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix},$$

where

$$\begin{aligned} A_1 &= M_4 M_2^{-1} + M_2^{-1} X^{-1} Y^{-1} (I - X) - M_2^{-1} X^{-1}, \\ A_2 &= X M_2, \\ A_3 &= M_2^{-1} X^{-1} (Y - I), \\ A_4 &= Y^{-1} (I - X) M_2, \\ A_5 &= M_2^{-1} (M_1 - Y), \end{aligned}$$

and $X, Y \in \mathrm{GL}_n(\mathbb{F})$ satisfy $[X, Y] = M_2(M_4 M_2^{-1} M_1 - M_3)$.

Proof. $\det [M_2(M_4 M_2^{-1} M_1 - M_3)] = \det M$ [5, Sec. 0.8.5], and so, by Lemma 4, there exists $X, Y \in \mathrm{GL}_n(\mathbb{F})$ with $[X, Y] = M_2(M_4 M_2^{-1} M_1 - M_3)$. The result follows from a short computation:

$$\begin{aligned} \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} &= \begin{bmatrix} Y & X M_2 \\ M_4 M_2^{-1} Y - M_2^{-1} [X, Y] & (M_4 M_2^{-1} X + M_2^{-1} [X, Y] Y^{-1} (I - X)) M_2 \end{bmatrix}, \\ \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix} &= \begin{bmatrix} Y^{-1} (M_1 - X (M_1 - Y)) & Y^{-1} (I - X) M_2 \\ M_2^{-1} (M_1 - Y) & I \end{bmatrix}, \end{aligned}$$

and, given $[X, Y] = M_2(M_4 M_2^{-1} M_1 - M_3)$, their product equals M . \square

$\mathrm{SL}_4(\mathrm{GF}(2))$ cannot be treated using Lemma 5, as the matrices $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ are not commutators of $\mathrm{SL}_2(\mathrm{GF}(2))$. Despite this, elements of $\mathrm{SL}_4(\mathrm{GF}(2))$ with non-singular upper right block can still be represented as the product of five block unitriangular matrices, which, given the small group size, is easily verified by exhaustive search.²

Lemma 6 ([12]). *Let $M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \in \mathrm{SL}_4(\mathrm{GF}(2))$ and $M_2 \in \mathrm{SL}_2(\mathrm{GF}(2))$. Then there exists $A_1, \dots, A_5 \in \mathrm{M}_2(\mathrm{GF}(2))$ such that $M = \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix}$.*

The desired factorizations of Theorems 1 and 2 follow from the application of Lemmas 5 and 6 to the product $M \begin{bmatrix} B & A \\ 0 & I \end{bmatrix}$ for some diagonal $B \in \mathrm{GL}_n(\mathbb{F})$ satisfying $\det(B) = \det(M)^{-1}$ and $A \in \mathrm{M}_n(\mathbb{F})$ satisfying $M_1 A + M_2 \in \mathrm{GL}_n(\mathbb{F})$. That such a matrix A exists is a consequence of the following simple lemma, as $M \in \mathrm{GL}_{2n}(\mathbb{F})$ implies $\mathrm{coker}(M_1) \cap \mathrm{coker}(M_2) = 0$.

Lemma 7. *For any $A, B \in \mathrm{M}_n(\mathbb{F})$, there exists $C \in \mathrm{M}_n(\mathbb{F})$ such that $CA + B \in \mathrm{GL}_n(\mathbb{F})$ if and only if $\ker(A) \cap \ker(B) = 0$.*

Proof. $\ker(A) \cap \ker(B) = 0$ is clearly necessary, as $\ker(A) \cap \ker(B) \subset \ker(CA + B)$. The converse also follows quickly. Simply choose C to be any matrix for which $\mathrm{im}(C)$ is a complement of $\mathrm{im}(B)$ and $\ker(C) \cap \{Ax \mid x \in \ker(B)\} = 0$. Such a matrix always exists, as $\ker(A) \cap \ker(B) = 0$ and the rank-nullity theorem together imply $\dim(\{Ax \mid x \in \ker(B)\}) = \dim(\mathrm{im}(C))$. \square

We now consider the lower bounds of Theorems 1 and 2. We have the following lemma regarding the representation of block diagonal matrices.

Lemma 8. *If $\begin{bmatrix} M_1 & 0 \\ 0 & M_4 \end{bmatrix} \in [\mathrm{T}_{m,n}(\mathbb{F})]^5$, $M_1 \in \mathrm{GL}_m(\mathbb{F})$, $M_4 \in \mathrm{GL}_n(\mathbb{F})$, then there exist diagonal matrices $D \in \mathrm{GL}_m(\mathbb{F})$ and $\tilde{D} \in \mathrm{GL}_n(\mathbb{F})$ such that*

$$m \cdot 1 + \mathrm{trace}(M_4 \tilde{D}) = n \cdot 1 + \mathrm{trace}(M_1^{-1} D).$$

Proof. It suffices to consider an $LULUL$ factorization, as the lemma statement is invariant under transpose. Suppose

$$\begin{bmatrix} M_1 & 0 \\ 0 & M_4 \end{bmatrix} = \begin{bmatrix} B_1 & 0 \\ A_1 & C_1 \end{bmatrix} \begin{bmatrix} B_2 & A_2 \\ 0 & C_2 \end{bmatrix} \begin{bmatrix} B_3 & 0 \\ A_3 & C_3 \end{bmatrix} \begin{bmatrix} B_4 & A_4 \\ 0 & C_4 \end{bmatrix} \begin{bmatrix} B_5 & 0 \\ A_5 & C_5 \end{bmatrix}$$

for some $A_1, A_3, A_5 \in \mathrm{M}_{n,m}(\mathbb{F})$, $A_2, A_4 \in \mathrm{M}_{m,n}(\mathbb{F})$, and diagonal matrices $B_1, \dots, B_5 \in \mathrm{GL}_m(\mathbb{F})$ and $C_1, \dots, C_5 \in \mathrm{GL}_n(\mathbb{F})$. We have

$$\begin{bmatrix} B_1 & 0 \\ A_1 & C_1 \end{bmatrix} \begin{bmatrix} B_2 & A_2 \\ 0 & C_2 \end{bmatrix} \begin{bmatrix} B_3 & 0 \\ A_3 & C_3 \end{bmatrix} = \begin{bmatrix} B_1 B_2 B_3 + B_1 A_2 A_3 & B_1 A_2 C_3 \\ A_1 B_2 B_3 + C_1 C_2 A_3 + A_1 A_2 A_3 & C_1 C_2 C_3 + A_1 A_2 C_3 \end{bmatrix}$$

and

$$\begin{aligned} \begin{bmatrix} M_1 & 0 \\ 0 & M_4 \end{bmatrix} \begin{bmatrix} B_5 & 0 \\ A_5 & C_5 \end{bmatrix}^{-1} \begin{bmatrix} B_4 & A_4 \\ 0 & C_4 \end{bmatrix}^{-1} &= \begin{bmatrix} M_1 & 0 \\ 0 & M_4 \end{bmatrix} \begin{bmatrix} B_5^{-1} & 0 \\ -C_5^{-1} A_5 B_5^{-1} & C_5^{-1} \end{bmatrix} \begin{bmatrix} B_4^{-1} & -B_4^{-1} A_4 C_4^{-1} \\ 0 & C_4^{-1} \end{bmatrix} \\ &= \begin{bmatrix} M_1 B_5^{-1} B_4^{-1} & -M_1 B_5^{-1} B_4^{-1} A_4 C_4^{-1} \\ -M_4 C_5^{-1} A_5 B_5^{-1} B_4^{-1} & M_4 C_5^{-1} (I + A_5 B_5^{-1} B_4^{-1} A_4) C_4^{-1} \end{bmatrix}. \end{aligned}$$

²See repository [12] for a short computer-assisted proof (using the Julia programming language [1]); the program terminates in under a second on a personal computer. It is also possible to prove Lemma 6 via an involved case analysis. The details are left to the interested reader.

By setting these matrices equal and inspecting the upper left and right blocks, we deduce that $A_2A_3 = B_1^{-1}M_1B_5^{-1}B_4^{-1} - B_2B_3$ and $A_4C_4^{-1} = -B_4B_5M_1^{-1}B_1A_2C_3$. Using the former equality applied to the lower left block,

$$-M_4C_5^{-1}A_5B_5^{-1}B_4^{-1} = A_1B_1^{-1}M_1B_5^{-1}B_4^{-1} + C_1C_2A_3,$$

which, all together, implies (using the lower right block)

$$\begin{aligned} M_4C_5^{-1}C_4^{-1} &= (C_1C_2 + A_1A_2)C_3 - M_4C_5^{-1}A_5B_5^{-1}B_4^{-1}A_4C_4^{-1} \\ &= (C_1C_2 + A_1A_2)C_3 + (A_1B_1^{-1}M_1B_5^{-1}B_4^{-1} + C_1C_2A_3)(-B_4B_5M_1^{-1}B_1A_2C_3) \\ &= C_1C_2C_3 - C_1C_2A_3B_4B_5M_1^{-1}B_1A_2C_3. \end{aligned}$$

Therefore,

$$A_2(A_3B_4B_5M_1^{-1}B_1) = I - B_2B_3B_4B_5M_1^{-1}B_1$$

and

$$(A_3B_4B_5M_1^{-1}B_1)A_2 = I - C_2^{-1}C_1^{-1}M_4C_5^{-1}C_4^{-1}C_3^{-1}.$$

Taking the trace of each gives our desired result, as the product of two matrices has a fixed trace, independent of the order of operands. \square

Consider the matrices $X \in \text{GL}_m(\mathbb{F})$ and $Y \in \text{GL}_n(\mathbb{F})$, $m, n > 1$, defined as follows:

$$X(i, j) = \begin{cases} 1 & \text{if } j - i = 1 \pmod{m} \\ 0 & \text{otherwise} \end{cases},$$

$$Y(i, j) = \begin{cases} \delta(m \cdot 1 - n \cdot 1) & \text{if } i = j = 1 \\ (-1)^{m+n} & \text{if } i = n, j = 1 \\ 1 & \text{if } j - i = 1 \\ 0 & \text{otherwise} \end{cases},$$

where $\delta(\cdot)$ is the Kronecker delta function. We have $\det(X) = \det(Y) = (-1)^{m+1}$, $\text{trace}(XD) = 0$ for all diagonal $D \in \text{GL}_m(\mathbb{F})$, and, for every diagonal $\tilde{D} \in \text{GL}_n(\mathbb{F})$, $\text{trace}(Y\tilde{D}) \neq 0$ if and only if $m \cdot 1 = n \cdot 1$. Therefore, by Lemma 8, $\begin{bmatrix} X^{-1} & 0 \\ 0 & Y \end{bmatrix} \notin [\text{T}_{m,n}(\mathbb{F})]^5$. To complete our desired lower bound, we must briefly analyze the case when either m or n is equal to one. If, say, $n = 1$ and $m > 2$, let us keep X as above and set $Y = (-1)^{m+1}$, so that $\begin{bmatrix} X^{-1} & 0 \\ 0 & Y \end{bmatrix} \in \text{SL}_{m+1}(\mathbb{F})$. By the analysis in the proof of Lemma 8, if $\begin{bmatrix} X^{-1} & 0 \\ 0 & Y \end{bmatrix} \in [\text{T}_{m,n}(\mathbb{F})]^5$, then $I - \hat{D}XD$ is a rank one matrix for some diagonal \hat{D} , $D \in \text{GL}_m(\mathbb{F})$. However, this is not possible, as $[I - \hat{D}XD](1, 1) = [I - \hat{D}XD](2, 2) = 1$ and $[I - \hat{D}XD](2, 1) = 0$. This completes the proof of Theorem 2. When \mathbb{F} has at least four elements, the lower bound for $\text{SL}_n(\mathbb{F})$ holds independent of indexing. The following lemma completes the proof of Theorem 1.

Lemma 9. *If \mathbb{F} has at least four elements, then, for every $m + n > 3$, there exists $M \in \text{SL}_{m+n}(\mathbb{F})$ such that $P_\pi M P_{\pi^{-1}} \notin [\text{BL}_{m,n}(\mathbb{F}) \cup \text{BU}_{m,n}(\mathbb{F})]^5$ for all permutations $\pi \in \mathcal{S}_{m+n}$.*

Proof. Let M be diagonal, with diagonal elements $g, h, (gh)^{-1}$ (not necessarily distinct), and $2n - 3$ copies of 1, for some $g, h \neq 1$ satisfying $gh \neq 1$. Such $g, h \in \mathbb{F}$ always exists

when \mathbb{F} has at least four elements (take any $g_1, g_2 \neq 0, 1$ distinct; either $g_1^2 \neq 1$ or $g_1 g_2 \neq 1$). Now suppose

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_4 \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_1 & I \end{bmatrix} \begin{bmatrix} I & A_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_3 & I \end{bmatrix} \begin{bmatrix} I & A_4 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ A_5 & I \end{bmatrix}$$

for some $A_1, A_3, A_5 \in M_{n,m}(\mathbb{F})$ and $A_2, A_4 \in M_{m,n}(\mathbb{F})$. Repeating the same analysis as in the proof of Lemma 8, we find that

$$A_2(A_3 M_1^{-1}) = I - M_1^{-1} \quad \text{and} \quad (A_3 M_1^{-1})A_2 = I - M_4.$$

The product of two matrices has a fixed set of non-zero characteristic roots, independent of the order of operands [4, Theorem 1]. However, in total, exactly three elements of $I - M_1^{-1}$ and $I - M_4$ are non-zero. Therefore, there is no ordering and bipartition of the diagonal elements such that the non-zero characteristic roots, taken with multiplicity, of $I - M_1^{-1}$ and $I - M_4$ are the same, a contradiction. \square

Acknowledgements. The author thanks Louisa Thomas for improving the style of presentation.

REFERENCES

- [1] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.
- [2] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- [3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016.
- [4] Harley Flanders. Elementary divisors of AB and BA. *Proceedings of the American Mathematical Society*, 2(6):871–874, 1951.
- [5] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [6] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.
- [7] Frederic Koehler, Viraj Mehta, and Andrej Risteski. Representational aspects of depth and conditioning in normalizing flows. In *International Conference on Machine Learning*, pages 5628–5636. PMLR, 2021.
- [8] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.
- [9] François Serre and Markus Püschel. Generalizing block LU factorization: A lower–upper–lower block triangular decomposition with minimal off-diagonal ranks. *Linear Algebra and its Applications*, 509:114–142, 2016.
- [10] RC Thompson. *Commutators in the special and general linear groups*. PhD thesis, California Institute of Technology, 1960.
- [11] RC Thompson. Commutators in the special and general linear groups. *Transactions of the American Mathematical Society*, 101(1):16–33, 1961.
- [12] John Urschel. GitHub repository proving Lemma 6. <https://github.com/johnurschel/SL2GF4>, 2023.