Isometry groups of regular and context-free languages for the Levenshtein distance

Vladimir Yankovskiy

Abstract

This article is a partial answer to the question of which groups can be represented as isometry groups of languages of a certain type. Namely, from two arbitrary finite groups $G$ and $H$, a method is indicated for constructing a regular language whose isometry group is isomorphic to $G \times H^n$, and it is proved that context-free isometry groups embed into $\Pi_{n=1}^\infty S_n$, and there exists a regular language with isometry group $\Pi_{n=1}^\infty S_{2n}$.

MSC 2020: 05E18, 20B25, 20H15
ACM: F.4.3

1 Introduction

This article is a partial answer to the question of which groups can be represented as isometry groups of regular and context-free languages. Namely, the following three theorems are proved:

**Theorem 1.** Let $G$ and $H$ be arbitrary finite groups. Then there is a regular language $L \subset \{0; 1\}^*$ such that $\text{Isom}(L) \cong G \times H^n$.

**Theorem 2.** Let $G$ be the isometry group of some context-free language $L$ for the Levenshtein distance. Then it is a subgroup of $\Pi_{n=1}^\infty S_n$.

**Theorem 3.** There is a regular language $L \subset \{0; 1\}^*$ such that $\text{Isom}(L) \cong \Pi_{n=1}^\infty S_{2n}$.

Here and below, $S_n$ denotes a symmetric group on $n$ elements, $C_n$ denotes a cyclic group of order $n$, $\Pi$ denotes the Cartesian product of groups.

The isometry groups of finite languages have been studied before. For example, in [3] it is proved that for an arbitrary finite alphabet $A$, the isometry group of $A^n$ for the Hamming distance is isomorphic to $S_{|A|}^n$. Here we consider isometry groups for the Levenshtein distance, first considered in [4], which allows one to find the distance between words of different lengths.

The article consists of 8 sections (including the introduction):

- Section 2 formulates the definition of the Levenshtein distance, and also classifies isometry groups of one-character languages.
• Sections 3 and 4 carry out the preparatory work necessary for the proof of Theorem 1.
• Section 5 proves Theorem 1.
• In Section 6, the preparatory work necessary for the proof of Theorem 2 is carried out.
• Section 7 proves Theorem 2.
• Section 8 proves Theorem 3.

2 Levenshtein distance and language isometries

Definition 1. Let $A$ be a finite alphabet, $\alpha, \beta \in A^*$. Then the Levenshtein distance $d(\alpha, \beta)$ is the minimum number of one-character operations (inserts, deletions, substitutions) required to turn $\alpha$ into $\beta$.

Obviously, $d$ is a metric on $A^*$.

Also, there is an alternative way to specify the Levenshtein distance by the recursive formula:

Proposition 1.\textsuperscript{(5)}

$$d(a, b) = \begin{cases} 
|a| & |b| = 0 \\
|b| & |a| = 0 \\
d(a.\text{tail}, b.\text{tail}) & a.\text{head} = b.\text{head} \\
1 + \min(d(a.\text{tail}, b.\text{tail}), d(a.\text{tail}, b), d(b.\text{tail}, a)) & a.\text{head} \neq b.\text{head}
\end{cases}$$

where $a.\text{tail}$ is the suffix of $a$ containing all of its characters except the first one, and $a.\text{head}$ is its first element.

In particular, from this formula and the invariance of the Levenshtein distance under the "reflection" of words, it follows that $d(u.xv, uyv) = d(x,y)$.

Definition 2. Let $L_1 \in A^*$ and $L_2 \in B^*$ be formal languages. Then the bijection $\phi : L_1 \to L_2$ is an isometry of the languages $L_1$ and $L_2$ iff $\forall u, v \in L_1$, $d(\phi(u), \phi(v)) = d(u, v)$

The set $\text{Isom}(L)$ of all isometries of the language $L$ into itself forms a composition group.

The class of isometry groups of languages over a one-element alphabet is rather small.

Proposition 2. Let $L \subset \{a\}^*$. Then $\text{Isom}(L) \in \{E, C_2\}$.

Proof. $l : a^n \mapsto n$ is an isometry of $\{a\}^*$ onto the metric space $(\mathbb{N}, d(m, n) = |n - m|)$. From this we can conclude that $\text{Isom}(l) \cong \text{Isom}(l(L))$.

Suppose now that $N \subset \mathbb{N}$, $|N| \geq 2$ (everything is obvious for $|N| = 1$). Let
\( n_0, n_1, \ldots \) be elements of \( N \) sorted in ascending order, \( \phi \) be an isometry from \( N \).

Let \( \phi(n_0) \in \{ n_0, n_{|N|} \} \). Otherwise, if \( \phi(n_0) = n_i \neq n_{|N|} \), then \( |n_{i+1} - n_{i-1}| = |\phi^{-1}(n_{i+1}) - \phi^{-1}(n_{i-1})| < |\phi^{-1}(n_{i+1}) - n_0| + |n_0 - \phi^{-1}(n_{i-1})| = |(n_{i+1} - n_{i-1})| + |n_i - n_{i-1}| = |n_{i+1} - n_{i-1}| \), which is impossible.

Let \( \phi(n_0) = n_0 \). Let us prove by induction that \( \phi \) is trivial:
The step is already complete.

**Base:** suppose \( \phi(n_i) = n_i \) for all \( i < k \). Then \( n_k \) is the only nearest element to \( n_{k-1} \) not contained in \( \{ n_0, \ldots, n_{k-1} \} \). This means that \( \phi(n_k) \) is the only closest element to \( \phi(n_{k-1}) = n_{k-1} \) not contained in \( \phi(\{ n_0, \ldots, n_{k-1} \}) = \{ n_0, \ldots, n_{k-1} \} \), i.e. \( \phi(n_k) = n_k \).

Let now \( \phi(n_0) = n_{|N|} \). Let us prove by induction that \( \phi : n_k \mapsto n_{|N|}-k \).
The step is already complete.

**Base:** suppose \( \phi(n_i) = n_{|N|}-i \) for all \( i < k \). Then \( n_k \) is the only nearest element to \( n_{k-1} \) not contained in \( \{ n_0, \ldots, n_{k-1} \} \). This means that \( \phi(n_k) \) is the only closest element to \( \phi(n_{k-1}) = n_{|N|}-k+1 \) not contained in \( \phi(\{ n_0, \ldots, n_{k-1} \}) = \{ n_{|N|}, \ldots, n_{|N|}-k+1 \} \), i.e. \( \phi(n_k) = n_{|N|}-k \).

Thus, there can be no other isometries, except for these two (and the second one is far from always realized).

Hence, \( \text{Isom}(L) \in \{ E, C_2 \} \)

However, for the two-element alphabet this is no longer the case.

### 3 Finite groups as isometry groups of uniform languages

**Definition 3.** A uniform language is a language in which the lengths of all words are equal.

All uniform languages are finite.

**Lemma 1.** Let \( \Gamma(V,E) \) be a finite simple cubic graph. Then there exists a language \( \{ w_v \}_{v \in V} \subset \{ 0; 1 \}^{3|E|} \) such that

\[
    d(w_u, w_v) = \begin{cases} 
        4 & (u, v) \in E \\
        6 & (u, v) \not\in E 
    \end{cases}
\]

**Proof.** Let \( E = \{ e_0, \ldots, e_m \} \). Let’s define \( i : V \times E \to \{ 0, 1 \}^* \) as follows:

\[
    i(v, e) = \begin{cases} 
        1 & v \in e \\
        0 & v \not\in e 
    \end{cases}
\]

Now we define \( w_v = i(v, e_0)11i(v,e_1)11\ldots11i(v,e_{|E|})11 \). For all \( v \in V \) \( w_v \) contains exactly \( |E| - 3 \) zeros and \( 2|E| + 3 \) ones. In this case, all zeros are in the positions of dividing 3 (if you count from the left).

Suppose that \( (u, v) \in E \). Let us show that \( d(w_u, w_v) = 4 \).
In this case, \( u \) and \( v \) share the same edge, and the two edges are different. That is, \( w_u \) can be converted to \( w_v \) using four character substitutions. Let us show that less than four is impossible.

All possible transformations of less than four operations that convert the original word into a valid one — this is the replacement of two characters or the removal of a character and then adding the same one, or the replacement of \( 1 \rightarrow 0 \) (\( 0 \rightarrow 1 \)), adding a new character 1 (0) and removing the character 0(1). Moreover, between the added symbol and the removed one (in all cases) there can only be ones (since zeros cannot be in positions that do not divide 3). In each case, no more than one edge will change, but two must be changed.

Now consider the case \((u,v) \notin E\). Let us show that \( d(w_u, w_v) = 6 \).

In this case, \( u \) and \( v \) have all three distinct edges. That is, \( w_u \) can be converted to \( w_v \) with six character substitutions. In this case (by analogy with the previous reasoning), it will not be possible to change three edges in less than six operations, i.e. \( d(w_u, w_v) = 6 \).

Thus, the language we need has been constructed.

\[ \square \]

**Theorem 4.** (2) A group is finite if and only if it is isomorphic to the automorphism group of some cubic graph.

**Corollary 1.** A group is finite if and only if it is isomorphic to the isometry group of some uniform language.

**Proof.** The isometry group of the language constructed in Lemma 1 coincides with the automorphism group of the corresponding cubic graph by construction.

\[ \square \]

### 4 Isometry of languages and countable Cartesian powers of finite groups

**Definition 4.** An \( n \)-ary de Bruijn word over a finite alphabet \( A \) is a word of length \(|A|^n + n - 1\) containing all words of length \( n \) as subwords.

De Bruijn words exist for any \( A \) and \( n \). Their explicit construction is described in [1].

**Lemma 2.** Let \( L \subset A^n \), \( w \) be an \( n \)-ary de Bruijn word over the alphabet \( A \). Then \( Isom(Lw^*) \cong Isom(L)^\mathbb{N} \).

**Proof.** Let \( u, v \in L \) and \( p, q \in \mathbb{N} \). Let us show that

\[
   d(uw^p, vw^q) = \begin{cases} 
   |pq|(|A|^n + n - 1) & p \neq q \\
   d(u,v) & p = q 
   \end{cases}
\]

Indeed, in the first case, a longer word is obtained from a shorter one by removing everything superfluous. In the second case, it is enough to perform
transformation only on different prefixes. From this, in particular, it follows that any transformation of the form \( \phi : uw^k \mapsto \psi(k)(u)w^k \), where \( \psi \) is an arbitrary function \( \mathbb{N} \to \text{Isom}(L) \), is an isometry of \( Lw^* \). To verify that there are no other isometries, note that \( uw^k \) has exactly \( 2|L| \) "neighbors" at distance \( t(|A|^n + n - 1) \), for \( t < k \) and exactly \( |L| \) for \( t > k \).

\[ \square \]

5 Proof of Theorem 1

Using Corollary 1, we can construct uniform languages \( L_1 \subset A^n \) and \( L_2 \subset A^m \) such that \( \text{Isom}(L_1) \cong G \) and \( \text{Isom}(L_2) \cong H \). Now let \( w_1 \) be an \( n \)-ary de Bruijn word, and \( w_2 \) be an \( m \)-ary de Bruijn word. Then by Lemma 2 \( \text{Isom}(L_2w_2^n) \cong H^N \).

Now consider the language \( L_1 \cup w_1^n L_2w_2^n \), where \( t = \lceil \frac{2^n + m + n - 1}{2} \rceil + 1 \). It is easy to see that for all \( u \in L_1, v \in L_2, k \in \mathbb{N} \) \( d(u, w_1^nvw_2^k) = t2^n + k2^m + (t - 1)n + (k + t)(m - 1) \).

Hence, for every \( \phi \in \text{Isom}(L_1) \) and \( \psi \in \text{Isom}(L_2w_2^n) \), the map \( \chi(\phi, \psi) \) taking all \( u \in L_1 \) to \( \phi(u) \), and all \( w_1^n \) to \( \psi(v) \), is an isometry of \( L_1 \cup w_1^n L_2w_2^n \). At the same time, there can be no other isometries, since the word \( v \) from the new language belongs to \( L_1 \) if and only if it has no "neighbors" at a distance greater than \( n \) but less than \( t(2^n + n + 1) \). So \( \text{Isom}(L_1 \cup w_1^n L_2w_2^n) \cong G \times H^N \). Moreover, \( L_1 \cup w_1 L_2w_2^n \) is regular by construction.

6 Length Conservation Lemma

**Lemma 3.** If the language \( L \) is context-free, then there exists an integer \( p \geq 1 \), called the growth length, such that any word \( s \in L \) of length at least \( p \) can be written as \( s = uwxxy \), where \( |vx| \geq 1 \), \( |vw| \leq p \) and \( \forall n \in \mathbb{N} \) \( uv^nxw^n \) \( y \in L \).

**Lemma 4.** Let \( L \) be a context-free language with growth length \( p \). Then for any \( s \in L, \phi \in \text{Isom}(L) \) \( ||s| - |\phi(s)|| \leq p \).

**Proof.** We will prove by contradiction. Let \( \exists s \in L \), incl. \( ||s| - |\phi(s)|| > p \). Let us assume without loss of generality that \( |s| > |\phi(s)| \). Then \( s - p > |\phi(s)| \).

Moreover, since by Lemma 3 any word \( s' \in L \) of length at least \( p \) can be written as \( s' = uv'v'w'x'y' \), where \( |v'x'| \geq 1 \) and \( u'w'y' \in L \), language \( L \) contains a word \( s_i \) of length at most \( p \), which can be obtained from \( s \) by performing only deletion operations.

On the other hand, by Lemma 3 \( s \) can be written as \( s = uvwxy \), where \( |vx| \geq 1, |vw| \leq p \) and \( \forall n \in \mathbb{N} \) \( uv^nwx^n \) \( y \in L \). Let \( t = \lceil \frac{|s| - |s_i|}{|vx|} \rceil + 1, s_r = uv^twx^ty \in L \). Then, \( d(s, s_r) = d(s, s_i), d(s, s_r) \geq |s| - p > |\phi(s)| \) and \( d(s_i, s_r) = d(s_i) + d(s, s_r) \).
So \( |\phi(s)| + d(s, s_r) < d(s, s_t) + d(s, s_r) = d(s_t, s_r) = d(\phi(s_t), \phi(s_r)) \leq \max(|\phi(s)| + d(s, s_t), |\phi(s)| + d(s, s_r)) = \max(|\phi(s)| + d(s, s_t), |\phi(s)| + d(s, s_r)) = |\phi(s)| + d(s, s_r). \) Contradiction. 

7 Proof of Theorem 2

Since by Lemma 4 for any \( s \in L, \phi \in Isom(L) \) \(|s| - |\phi(s)|| \leq p,\) we can conclude that all orbits of the natural action of Isom\((L)\) on \( L \) are finite. Let \( n_1, n_2, \ldots \) be the orders of these orbits. Then, since Isom\((L)\) acts freely on \( L \), it embeds in \( \Pi_{i=1}^\infty S_{n_i} \leq \Pi_{n=1}^\infty S_n. \)

8 Proof of Theorem 3

Consider the language \( L = (010)^*(110)(010)^* \cap \{0,1\}^6)^* \). Let us show that for arbitrary two words \( u, v \in L \) such that \(|u| - |v| = 6, v \) can be obtained from \( u \) with exactly 6 deletions. Note that at positions dividing 3, each word can contain only one 1. Let these special units in the words \( u \) and \( v \) be at positions \( 3i \) and \( 3j \), respectively. Then, if \( i = j \), it suffices to remove from \( u \) the postfix of length 6 (of the form 010010). Otherwise, you must first remove the subword \( u_{3j}u_{3j+1}u_{3j+2} = 110 \) from \( u \), and then from the subword \( u_{3j}u_{3j+1}u_{3j+2}u_{3j+3} = 0100 \) (numbering is now according to the new order after the previous deletion) remove \( u_{3j} \) and \( u_{3j+2}u_{3j+3} \). Thus we get \( v \).

From the above it follows by induction that for \( u, v \in L \)

\[ d(u, v) = \max(||u| - |v||, 2) \]

. Thus we see that the distance between words does not depend on anything other than their length. This means that for any sequence of permutations \( \sigma_i \) of elements \( L \cap A^{6i} \) the map \( \phi : u \mapsto \sigma_{\phi(u)}(u) \) is an isometry .

At the same time, the absence of other isometries follows from the fact that each word \( u \in L \) has exactly \( 2|u| \) neighbors at a distance of 6.

That is, since \(|L \cap A^{6i}| = 2i, Isom(L) \cong \Pi_{n=1}^\infty S_{2n}. \)

Bibliography

[1] N.G. De Bruijn. “A combinatorial problem”. In: Koninklijke Nederlandse Akademie v. Wetenschappen 49 (1946), pp. 758–764.

[2] R. Frucht. “Graphs of degree three with a given abstract group”. In: Canadian Journal of Mathematics 1 (1949), pp. 365–378.

[3] A.A. Markov. “On transformations, not spreading distortions”. 1956.

[4] Y. Bar-Hillel, M. Perles, and E. Shamir. “On formal properties of simple phrase-structure grammars”. In: Zeitschrift für Phonetik, Sprachwissenschaft, und Kommunikationsforschung 14 (1961), pp. 143–172.
[5] V.I. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals.” In: Soviet Physics Doklady 10 (1965), pp. 707–710.