On the Complexity of Computing the Profinite Closure of a Rational Language

Pierre-Cyrille Heam

On the Complexity of Computing the Profinite Closure of a Rational Language

P.-C. Héam

November 16, 2011

Projet INRIA-CASSIS
Laboratoire d'Informatique de l'Université de Franche-Comté
16 route de Gray, 25030 Besançon Cedex. FRANCE.

Abstract

The profinite topology is used in rational languages classification. In particular, several important decidability problems, related to the Malcev product, reduce to the computation of the closure of a rational language in the profinite topology. It is known that given a rational language by a deterministic automaton, computing a deterministic automaton accepting its profinite closure can be done with an exponential upper bound. This paper is dedicated the study of a lower bound for this problem: we prove that in some cases, if the alphabet contains at least three letters, it requires an exponential time.

1 Preliminaries

For more informations on automata and languages theory we refer the reader to [1, 4, 9]. For a general reference on profinite topologies see [5, 13].

1.1 Introduction

The profinite topology is used to characterise certain classes of rational languages: the languages of level 1/2 in the group hierarchy and the languages recognisable by reversible automata [11, 14]. Moreover profinite topologies on the free group or on the free monoid play a crucial role in the theory of finite semigroups [2, 7, 12, 3]. In particular, several important decidability problems, related to the Malcev product, reduce to the computation of the closure of a rational language in the profinite topology.

It is known that the profinite closure of a rational language is rational too [16, 8]. The first algorithm was given in [15] for languages given by rational expressions, while [17, 10] provide algorithms on finite automata. In this paper we are interesting in the following problem:

Profinite Closure
Input: A finite deterministic n-states automaton $\mathcal{B}$ on the alphabet A.
Output: A finite deterministic automaton accepting the profinite closure of $L(\mathcal{B})$.

A solution to this problem is known to be computable in time $O(2^n)$ [10]. We prove in this paper that it can not be done, in some cases, faster than in exponential time (if the alphabet contains at least three letters).

In the first part of this paper, we introduce useful notations and definitions. In the second one, we recall an algorithm [17, 10] to answer the above problem. The last section of this paper is dedicated to the main result of the paper: we will prove there exists a family or rational languages K_n such that:

(1) The minimal automaton of K_n has $3n$ states,

(2) The minimal automaton of the profinite closure of K_n has $\Omega(4^n/\sqrt{n})$ states.

Notice that topological notions related to this paper are technical and required a wide mathematical background. However, the proved result can be easily understood using only automata theoretic arguments. In order to not overload the reader, we do not expose the mathematical background in this article. The interested reader is referred to [5, 13] for more information on profinite topologies. Particular topological properties of rational languages are studied in [6]. We just provide short definitions in the next section.

1.2 Background and notations

Let A be a finite alphabet and let $\overline{A} = \{\overline{a} \mid a \in A\}$ be a copy of A. Finally, let $\tilde{A}$ be the disjoint union of A and $\overline{A}$. The map $a \mapsto \overline{a}$ from A onto $\overline{A}$ can be extended to a one-to-one function from $\tilde{A}$ into itself by setting $\overline{\overline{a}} = a$. A word of A^* is said to be reduced if it does not contain any factor of the form $a\overline{a}$ with $a \in \tilde{A}$. We denote by $\equiv$ the monoid congruence generated by the relations $a\overline{a} \equiv 1$ for all $a \in \tilde{A}$. The set $\tilde{A}/_{\equiv}$ is a group for the quotient law, called the free group over A. Let π be the projection from $\tilde{A}$ into this group, which is a monoid morphism. We denote by $D(\tilde{A})$ the set $\pi^{-1}(\varepsilon)$, i.e. the set of words of $\tilde{A}^*$ that can be rewritten into ε using the rewriting rules $a\overline{a} \to \varepsilon$, with $a \in \tilde{A}$.

The family of normal subgroups of the free group with finite index forms a basis of open sets for the profinite topology on the free group. Similarly, the class of group languages (regular languages whose syntactic monoids are finite groups) forms a basis of open sets for the profinite topology on A^*. Note that the profinite closure in A^* of a language L is the intersection of A^* with its profinite closure in the free group. Throughout the end of this paper, the considered profinite topology is the one on A^*.

Recall that a *finite automaton* is a 5-tuple $\mathcal{A} = (Q, B, E, I, F)$ where Q is a finite set of *states*, B is the alphabet, $E \subseteq Q \times B \times Q$ is the set of *edges* (or *transitions*), $I \subseteq Q$ is the set of *initial states* and $F \subseteq Q$ is the set of *final states*. A *path* in $\mathcal{A}$ is a finite sequence of consecutive edges:

$$p = (q_0, a_0, q_1), (q_1, a_1, q_2), \cdots, (q_{n-1}, a_n, q_n)$$

The *label* of the path p is the word $a_1 a_2 \cdots a_n$, its *origin* is q_0 and its *end* is q_n. A word is accepted by $\mathcal{A}$ if it is the label of a path in $\mathcal{A}$ having its origin in I and its end in F. Such a path is said to be *successful*. The set of words accepted by $\mathcal{A}$ is denoted by $L(\mathcal{A})$.

For every state q and language K we denote by $q \cdot_{\mathcal{A}} K$ (or $q \cdot K$ if there is no ambiguity on $\mathcal{A}$), the subset of Q of all the states which are the end of a path having its origin in q and its label in K. An automaton is said to be *trim* if for each state q there exists a path from an initial state to q and a path from q to a final state. An automaton is *deterministic* if it has a unique initial state and does not contain any pair of edges of the form (q, a, q_1) and (q, a, q_2) with $q_1 \neq q_2$. An important result of automata theory states that for any automaton $\mathcal{A}$ there exists exactly one deterministic automaton (up to isomorphism) with a minimal number of states which accepts the same language. It is called the *minimal automaton* of $L(\mathcal{A})$. Two states p and q of an automaton are *Nerode-equivalent* if for every word u, $p \cdot u$ is final if and only if $q \cdot u$ is final too. It is well known that a trim deterministic automaton is minimal if and only if all classes of the Nerode equivalence are singletons.

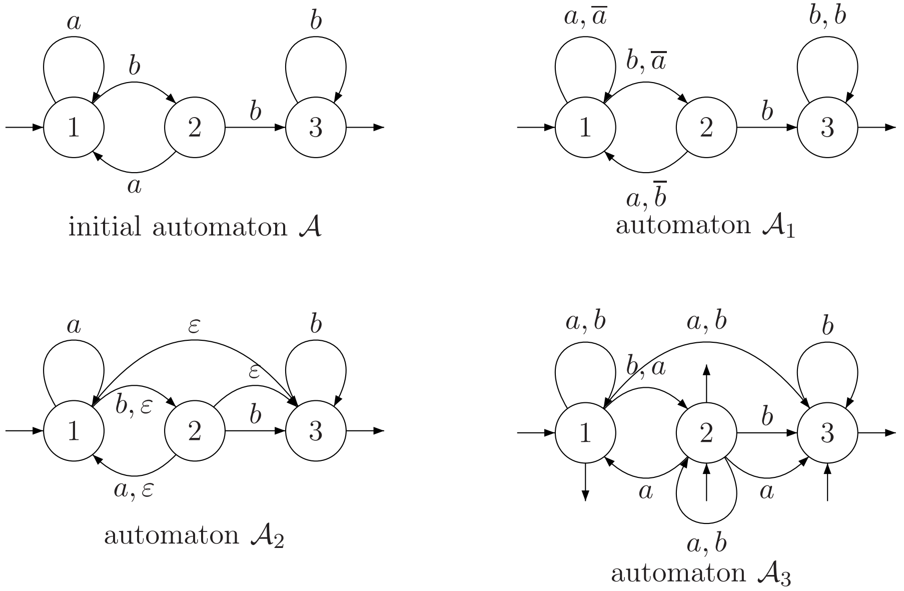
Let $\mathcal{A}$ be an automaton with set of states Q and set of transitions E. A subset P of Q is said to be *strongly connected* if, for each pair p and q of states in P, there exist a path from p to q and a path from q to p. A strongly connected component of $\mathcal{A}$ is a maximal (for the inclusion) set of states which is strongly connected. The strongly connected components of $\mathcal{A}$ form a partition of Q. A transition (p, a, q) of $\mathcal{A}$ is *internal to a strongly connected component* if p and q belongs to the same strongly connected component. It is said *internal* if it is internal to some strongly connected component and *external* otherwise.

The class of *rational* languages of A^* is the smallest class of languages closed under product, finite union and star operation. It is well known that a language of A^* is rational if and only if it can be accepted by a finite automaton.

1.3 Profinite Closure of a Rational Language

It is known that the profinite closure of a rational language is rational too [16].

In this direction, we use the following algorithm [17, 10], called Profinite-Closure, working on a finite trim automaton $\mathcal{A} = (Q, A, E, I, F)$ in order to compute the profinite closure of $L(\mathcal{A})$.



initial automaton $\mathcal{A}$

automaton $\mathcal{A}_1$

automaton $\mathcal{A}_2$

automaton $\mathcal{A}_3$

Figure 1: Algorithm ProfiniteClosure on an example

1. Compute the strongly connected components of $\mathcal{A}$.

2. Compute the set T of external transitions.

3. Compute $E_1 = \{(q, a, p) \mid (p, \overline{a}, q) \in E \setminus T\}$. Let $\mathcal{A}_1 = (Q, \tilde{A}, E \cup E_1, I, F)$.

4. Compute $E_2 = \{(p, \varepsilon, q) \mid p \neq q,\ q \in p \cdot_{\mathcal{A}_1} u,\ u \in D(\tilde{A})\}$. Let $\mathcal{A}_2 = (Q, A, E \cup E_2, I, F)$.

5. Return $\mathcal{A}_3$, the automaton obtained for $(Q, A, E \cup E_2, I, F)$ by a classical ε-transitions elimination.

In order to obtain a resulting deterministic automaton, one can use the standard determinization algorithm, which is known to be exponential in the worst case [9]. We illustrate how this algorithm is working on a graphical example on Fig. 1.

2 Main result

We will prove in this section there exists a family of rational languages K_n such that:

(1) The minimal automaton of K_n has $3n$ states,

(2) The minimal automaton of the profinite closure of K_n has $\Omega(4^n/\sqrt{n})$ states.

We first need the following technical lemma.

Lemma 1 *Let a and b be the two following permutations of $Q_n = \{1, \dots, 2n\}$ (with $n \geq 2$):*

$$a = (2, 3, \cdots, 2n, 1) \quad b = (1, 2).$$

Let $E_n = \{(i, a, a(i)) \mid i \in Q_n\} \cup \{(i, b, b(i)) \mid i \in Q_n\}$ and

$$\mathcal{A}_n = (Q_n, \{a, b\}, E_n, \{1, \dots, n\}, 2n).$$

The minimal automaton of $L(\mathcal{A}_n)$ has $(2n)!/(n!)^2$ states.

Proof. Notice first that a and b generate the symmetric group of $\{1, \dots, 2n\}$. For every word $u \in \{a, b\}^*$, we define σ_u by $\sigma_{ua} = a \circ \sigma_u$ and $\sigma_{ub} = b \circ \sigma_u$. Therefore, for every permutation σ of A there exists a word u such that $\sigma = \sigma_u$.

Consider the following automata $\mathcal{A}'_n$ obtained by the classical determinization algorithm:

- The set of states of $\mathcal{A}'_n$ is composed by all the images of $\{1, \dots, n\}$ by all permutations of Q_n. We obtain all the subsets of A with n elements. Therefore, since Q_n has $2n$ elements, $\mathcal{A}'_n$ has $(2n)!/(n!)^2$ states .

- We have a transition between two parts R and S labelled by a [resp. par b], if and only if $a(R) = S$ [resp. $b(R) = S$].

- Initial state of $\mathcal{A}'_n$ is $\{1, \dots, n\}$.

- Final states of $\mathcal{A}'_n$ are all states containing $2n$.

By construction $L(\mathcal{A}'_n) = L(\mathcal{A}_n)$ (see [9]).
Now we claim that $\mathcal{A}'_n$ is minimal. Consider the two distinct states $R = \{a_1, \cdots, a_n\}$ and $S = \{b_1, \cdots, b_n\}$ of $\mathcal{A}'_n$. We will prove that R and S are not Nerode-equivalent. Following cases arise:

- If R is final and S is not final, then $R \cdot \varepsilon$ is final but not $S \cdot \varepsilon$. Thus R and S are not Nerode-equivalent. The same argument holds if S is final and R is not final.
- If R and S are either both final or both non final, then there exists $a_k \in R$ such that $a_k \neq 2n$ and $a_k \notin S$. Consider a word u such that $\sigma_u = (a_k, 2n)$. The unique state of $R \cdot u$ contains $2n$; it is final. However the unique state of $S \cdot u$ doesn't contain $2n$ since $a_k \notin S$.

Since $\mathcal{A}'_n$ is trim, we proved that the minimal automaton of $L(\mathcal{A}_n)$ has $(2n)!/(n!)^2$ states. □

With previous lemma notations, we consider now the following family of automata denoted $(\mathcal{B}_n)_{n \in \mathbb{N}}$ and illustrated on Fig. 2.

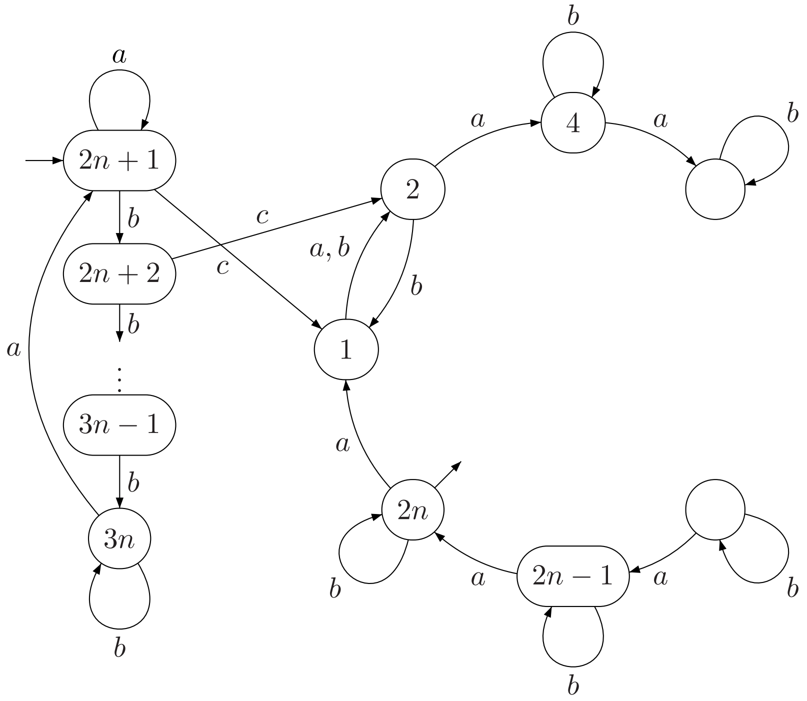


Figure 2: Automaton $\mathcal{B}_n$

- Set of states of $\mathcal{B}_n$ is $\{1, \ldots, 3n\}$.
- Alphabet of $\mathcal{B}_n$ is $\{a, b, c\}$.
- Transitions of $\mathcal{B}_n$ are the following tuples:
  - $(2n+k, b, 2n+k+1)$ for $1 \leq k \leq n-1$,
  - $(2n+k, c, k)$ for $1 \leq k \leq n-1$,
  - $(k, a, a(k))$ and $(k, b, b(k))$ for $1 \leq k \leq 2n$,
  - $(3n, b, 3n)$, $(2n+1, a, 2n+1)$ and $(3n, a, 2n+1)$,
- The unique initial state of $\mathcal{B}_n$ is $2n+1$,
- The unique final state of $\mathcal{B}_n$ is $2n$.

Theorem 2 *With above notations, $\mathcal{B}_n$ is minimal, has $3n$ states and the minimal automaton of the profinite closure of $L(\mathcal{B}_n)$ has $\Omega(4^n/\sqrt{n})$ states.*

Proof. In this proof $A = \{a, b, c\}$. We first prove that $\mathcal{B}_n$ is minimal. Indeed consider two distinct states k_1 and k_2 of $\mathcal{B}_n$. The following cases arise:

- If $1 \leq k_1 \leq 2n$ and $1 \leq k_2 \leq 2n$, then $k_1 \cdot a^{2n-k_1}$ is final. Without loss of generality, we may assume that $k_2 < k_1$. Therefore $k_2 \cdot a^{2n-k_1}$ is $2n - (k1 - k2)$ which is not final. Thus the two states are not Nerode-equivalent.

- If $1 \leq k_1 \leq 2n$ and $2n+1 \leq k_2 \leq 3n$, then one can reach a final states from k_2 using a word u containing the letter c. By construction, $k_1 \cdot u$ is not final. Thus k_1 and k_2 are not Nerode-equivalent.

- If $1 \leq k_2 \leq 2n$ and $2n+1 \leq k_1 \leq 3n$, we conclude similarly.

- If $2n+1 \leq k_2 \leq 3n$ and $2n+1 \leq k_1 \leq 3n$, then $k_1 \cdot ca^{4n-k_1}$ is final and $k_2 \cdot ca^{4n-k_1}$ is not final. Thus the two states are not Nerode-equivalent.

Since $\mathcal{B}_n$ is trim, it follows that $\mathcal{B}_n$ is a $3n$ states minimal automaton.
We will now apply the algorithm $\mathsf{ProfiniteClosure}$ on $\mathcal{B}_n$.

1. The strongly connected components of $\mathcal{B}_n$ are $\{1, \dots, 2n\}$ on one hand and $\{2n+1, \dots, 3n\}$ on the other hand.

2. The set of external transitions is reduced to all transitions labelled by c.

3. We add the reverse transitions labelled in $\overline{A}$ of internal transitions. Notice we only add transitions labelled by $\overline{a}$ and $\overline{b}$. We have the automaton represented on Fig. 3.

4. In the strongly connected component $\{2n+1, \dots, 3n\}$, one has a path from k_1 to k_2 labelled by $b^{3n-k_1}(\overline{b})^{n-k_1+k_2-1}a\overline{a}b^{k_2-2n-1}$. Therefore there is an ε-transition between all pairs of distinct states of $\{2n+1, \dots, 3n\}$.

   Since every word labelling a path from a state of $\{2n+1, \dots, 3n\}$ to a state of $\{1, \dots, 2n\}$ contains an occurrence of c and no occurrence of $\overline{c}$, there is no ε-transition between the two strongly connected components.

   Now for every state p of $\{1, \dots, 2n\}$, $p \cdot a\overline{a} = p \cdot \overline{a}a = p \cdot b\overline{b} = p \cdot \overline{b}b = \{p\}$, thus we add no ε-transition in $\{1, \dots, 2n\}$;

5. It follows that the profinite closure of $L(\mathcal{B}_n)$ is the language $\{a, b\}^* cL(\mathcal{A}_n)$ whose minimal automaton is given on Fig. 4.

Observe that, using lemma 1, this automaton is minimal. Consequently, the profinite closure of $L(\mathcal{B}_n)$ has $1 + \frac{(2n)!}{(n!)^2}$ states.

Thus, using Stirling Formula, the minimal automaton of the profinite closure of $L(\mathcal{B}_n)$ has $\Omega(4^n/\sqrt{n})$ states. □

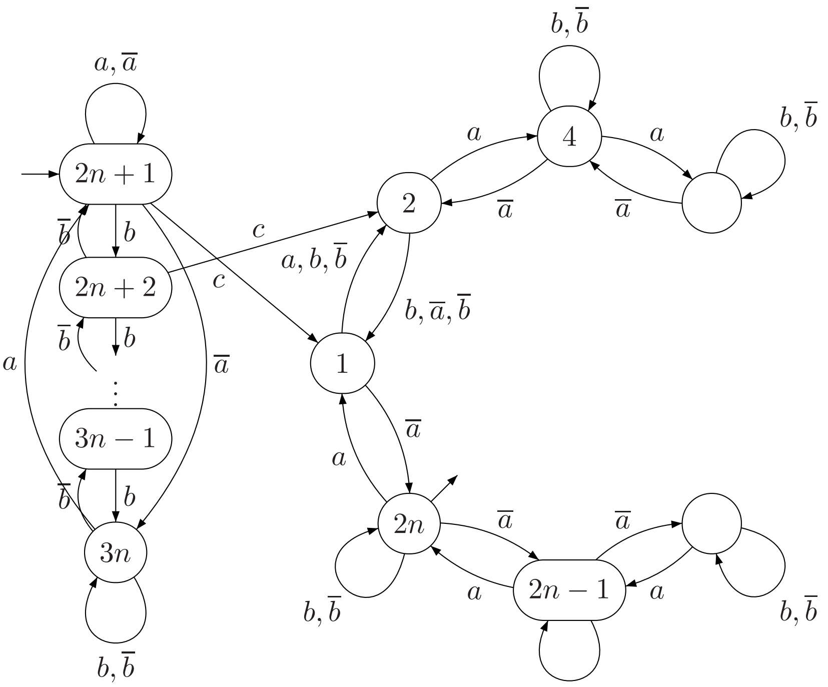


Figure 3: Automaton $\mathcal{B}_n$

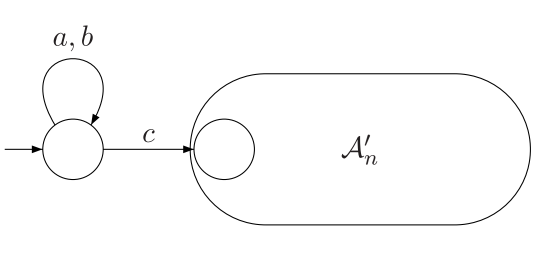


Figure 4: Minimal automaton of $\{a, b\}^* cL(\mathcal{A}_n)$

We proved that the **Profinite Closure** problem can not be solved, in some cases, faster than in exponential time (if the alphabet contains at least three letters).

3 Conclusion

In this paper we studied the complexity of computing the profinite closure of a rational language when the chosen representation for rational languages is finite deterministic automata. We proved this computation requires an exponential time in the worst case (if the alphabet contains at least three letters). One can easily verify that the problem is polynomial for unary alphabet. However, as far as we know, the problem is still open for two letters alphabets.

The author would like to thank J.-E. Pin and P. Weil for helpful discussions. This work was partially granted by the french project ANR-2010-BLAN-0202-02-FREC.

References

[1] J. Berstel. *Transductions and Context-Free-Languages*. B.G. Teubner, Stuttgart, 1979.

[2] M. Delgado. Abelian pointlikes of a monoid. *Semigroup forum*, 56:339–361, 1998.

[3] M. Delgado and P.-C. Heam. A polynomial time algorithm to compute the abelian kernel of a finite monoid. *Semigroup Forum*, 67:97–110, 2003.

[4] S. Eilenberg. *Automata, Languages and Machines*, volume C. Academic Press, New York, 1978.

[5] M. Hall. A topology for free groups and related groups. *Ann. of Maths*, 52, 1950.

[6] P.-C. Héam. Some topological properties of rational sets. *Journal of Automata, Languages and Combinatorics*, 6(3):275–290, 2001.

[7] K. Henckell, S. Margolis, J. Pin, and J. Rhodes. Ash's type II theorem, profinite topology and Malcev products. part I. *Int. J. Algebra and Comput.*, 1:411–436, 1991.

[8] B. Herwig and D. Lascar. Extending partial automorphism and the profinite topology on the free groups. *Transactions of the American Mathematical Society*, 352:1985–2021, 2000.

[9] J. Hopcroft and J. Ullman. *Introduction to automata theory, languages, and computation.* Addison-Wesley, 1980.

[10] P.-C. Héam. Automata for pro-**v** topologies. In *4th International Conference on Implementation Implementation and Application of Automata (CIAA)*, volume 2088 of *Lecture Notes in Computer Science*, pages 135–142, 2001.

[11] J.-E. Pin. On the language accepted by finite reversible automata. In T. Ottmann, editor, *Automata, Languages and Programming, 14th International Colloquium*, volume 267 of *Lecture Notes in Computer Science*, pages 237–249, Karlsruhe, Germany, 13–17 July 1987. Springer-Verlag.

[12] J.-E. Pin. A topological approach to a conjecture of Rhodes. *Bull. Austral. Math. Soc.*, 38:421–431, 1988.

[13] J.-E. Pin. Topologies for the free monoid. *Journal of Algebra*, 137(2):297–337, 1991.

[14] J.-E. Pin. Polynomial closure of group languages and open sets of the Hall topology. *Lecture Notes in Computer Science*, 820:424–432, 1994.

[15] J.-E. Pin and C. Reutenauer. A conjecture on the Hall topology on a free group. *Bull. London Math. Soc.*, 25:356–362, 1991.

[16] L. Ribes and P. Zalesskii. On the profinite topology on a free group. *Bull. London Math. Soc.*, 25:37–43, 1993.

[17] B. Steinberg. Finite state automata: A geometric approach. *Transactions of the American Mathematical Society*, 353(9):3409–3464, 2001.