dr hab. Jan Otop, prof. UWr
Instytut Informatyki
Uniwersytet Wrocławski

Wrocław, 17 lipca 2023

Recenzja rozprawy doktorskiej
Commutative images of languages over infinite alphabets
magistra Mohnisha Pattathurajana

Tematyka rozprawy

W przedstawionej rozprawie doktorskiej badane są języki formalne nad alfabetami nieskończonymi — w szczególności odpowiedniki języków regularnych oraz bezkontekstowych. Języki regularne (odpowiednio bezkontekstowe) nad alfabetami nieskończonymi mogą być zdefiniowane na kilka sposobów, w szczególności poprzez dodanie rejestrów przechowujących litery lub poprzez zmianę pojęcia „skończoności" przy zachowaniu klasycznych definicji. W drugim przypadku, dopuszcza się aby np. zbiór stanów oraz relacja przejścia w automacie niedeterministycznym były jedynie "orbitowo skończone" czyli skończone z dokładnością do permutacji alfabetu. Przykładowo, nad alfabetem złożonym z liczb naturalnych, rozważając jedynie relację równości pomiędzy liczbami, zbiór par różnych liczb naturalnych jest nieskończony, ale jest orbitowo skończony. Zarówno niedeterministyczne automaty z rejestrami jak i niedeterministyczne automaty orbitowo skończone definiują tę samą klasę języków: języki regularnych nad alfabetami nieskończonymi. Przykładem takiego języka regularnego jest zbiór skończonych ciągów liczb naturalnych, w których pewna liczba się powtarza, natomiast dopełnienie tego języka, złożone ze skończonych ciągów unikalnych liczb naturalnych nie jest językiem regularnym. Zatem języki regularne nad alfabetami nieskończonymi nie są domknięte ze względu na dopełnienie, co pokazuje, że pomimo zastosowania klasycznej definicji, automaty orbitowo skończone istotnie różnią się od automatów skończonych.

Wyniki przedstawione w Rozprawie dotyczą rozszerzenia twierdzenia Parikha o obrazach przemiennych na języki nad alfabetami nieskończonymi. Obraz przemienny (lub obraz Parikha) słowa to wektor krotności liter z tego słowa a obraz przemienny języka, to zbiór obrazów (wektorów) słów z danego języka. Można myśleć, że obraz przemienny otrzymuje się z języka przez wymuszenie przemienności liter, przez co słowa są jednoznacznie definiowane przez krotności występowania poszczególnych liter. Twierdzenie Parikha mówi, że klasy obrazów przemiennych zarówno języków regularnych jak i bezkontekstowych pokrywają się i są równe klasie zbiorów semiliniowych (skończone sumy zbiorów liniowych). To w szczególności oznacza, że różnice pomiędzy językami regularnymi oraz bezkontestowymy wynikają wyłącznie z porządku liter w słowach a nie krotności liter; każdy język bezkontekstowy zdefiniowany jedynie na podstawie krotności liter jest językiem regularnym.

O ile języki regularne oraz języki bezkontekstowe są reprezentowane przez odpowiednio automaty z rejestrami oraz gramatyki bezkontekstowe z rejestrami, które bezpośrednio rozszerzają klasyczne definicje, zbiory semiliniowe nie mają odpowiednika pozwalającego na bezpośrednie uogólnienie twierdzenia Parikha. Wyniki przedstawione w Rozprawie rozpoczynają się od dyskusji, jakie zbiory (nieskończonych) wektorów powinny być odpowiednikiem zbiorów semiliniowych z twierdzenia Parikha, bo już ten fakt nie jest oczywisty. To w szczególności pokazuje trudność podjętego zadania.

Jan Otop

Ocena rozprawy

Wyniki. W Rozprawie łączone są wyżej wymienione dwa podejścia do definiowania obiektów nad alfabetami nieskończonymi. Języki regularne oraz bezkontekstowe nad alfabetami nieskończonymi są reprezentowane przez, odpowiednio, niedeterministyczne automaty z rejestrami oraz gramatyki bezkontekstowe z rejestrami. Zbiory wektorów nad alfabetami nieskończonymi, w szczególności odpowiedniki zbiorów semiliniowych, są zdefiniowane przy użyciu pojęcia orbitowej skończoności.

Pierwszym wkładem Rozprawy **(W1)** jest definicja dwóch klas zbiorów wektorów nad nieskończonym alfabetem: (odpowiednika) zbiorów semiliniowych oraz zbiorów regularnych (obrazów wyrażeń regularnych), które są otrzymane przez zmianę w klasycznych definicjach sum skończonych na sumy orbitowo skończone. Tak zdefiniowane zbiory semiliniowe są istotnym podzbiorem zbiorów regularnych (Twierdzenie 5.3.1). W szczególności, jest wskazany język rozpoznawany przez niedeterministyczny automat z jednym rejestrem, którego obraz przemienny jest zbiorem regularnym ale nie jest zbiorem semiliniowym (nad alfabetem nieskończonym). Dowód, że obraz przemienny wskazanego języka nie jest semilinowy jest dość techniczny i niestety nie towarzyszą mu intuicje.

Twierdzenie 5.3.1 pokazuje, że twierdzenie Parikha nie ma dokładnego przeniesienia na alfabety nieskończone. Co więcej, to zbiory regularne a nie semiliniowe są odpowiednim kandydatem do reprezentowania obrazów przemiennych języków regularnych (nad alfabetem nieskończonym). W dalszej części rozprawy Doktorant skupia się na zbiorach regularnych oraz automatach i gramatykach z pojedynczym rejestrem.

Drugim wynikiem **(W2)** jest Twierdzenie 6.1.1 mówiące, że obrazy przemienne języków rozpoznawanych przez niedeterministyczne automaty z jednym rejestrem są zbiorami regularnymi. Dowód tego twierdzenia jest najbardziej złożonym wynikiem z Rozprawy, leczy dzięki intuicjom czyta się go lepiej niż dowód Twierdzenia 5.3.1. Sam dowód składa się z ciągu redukcji, t.j., stwierdzeń, że obraz przemienny danego języka jest zbiorem regularnym, jeżeli obraz przemienny innego języka jest regularny. Dla ostatniego języka z ciągu redukcji, języka anty-cykli nieograniczonego rzędu, jest pokazane wprost, że jego obraz przemienny jest regularny. Dowód regularności jest pomysłowy; opiera się na zdefiniowaniu grafu odpowiadającego literom i pokazanie, że każde słowo z języka odpowiada cyklowi Hamiltona w tym grafie a każdy cykl Hamiltona definiuje słowo z języka.

Trzeci wynik **(W3)** to twierdzenie, że obrazy przemienne języków bezkontekstowych definiowanych przez gramatyki z pojedynczym rejestrem jest regularny. Dowód jest poprzez sprytną redukcję do wyniku **(W2)**. To pokazuje regularność obrazów przemiennych zarówno języków rozpoznawanych przez (niedeterministyczne) automaty z jednym rejestrem jak i języków generowanych przez gramatyki bezkontekstowe z jednym rejestrem.

O ile wyniki **(W2)** i **(W3)** mówiły o zawieraniu się pewnych obrazów przemiennych w zbiorach regularnych, czwarty wynik **(W4)** charakteryzuje również zbiory regularne. W ramach tego wyniku, zdefiniowane są hierarchiczne automaty z rejestrami, które są podklasą niedeterministycznych automatów z rejestrami oraz dwa twierdzenia. Pierwsze twierdzenie (Wniosek 7.6.5) mówi, że obrazy przemienne języków rozpoznawanych przez hierarchiczne automaty z rejestrami to dokładnie zbiory regularne. To w szczególności oznacza, że każdy zbiór regularny jest obrazem pewnego niedeterministycznego automaty z rejestrami (który jest dodatkowo hierarchiczny). Aby pokazać, że te zbiory są regularne Doktorant pokazuje redukcję do wyniku **(W2)**; dowód jest poprzez indukcję względem liczby rejestrów a Twierdzenie 6.1.1 jest bazą indukcji oraz jest użyte w kroku indukcyjnym.

W ramach **(W4)** podane jest również twierdzenie mówiące, że hierarchiczne automaty z rejestrami mają istotnie większą siłę wyrazu niż automaty z jednym rejestrem, ale istotnie mniejszą niż automaty z rejestrami. Niestety, w Rozprawie tylko drugie stwierdzenie jest pokazane (Twierdzenie 7.4.1).

Podsumowując, uważam że wyniki przedstawione w Rozprawie są ciekawe i oryginalne oraz stanowią punkt wyjściowy dla dalszych badań.

Znaczenie. Problem rozważany w rozprawie jest naturalny, ważny i ciekawy. Rozważane obiekty, czyli języki nad alfabetami nieskończonymi mają zastosowania w automatycznej weryfikacji opro-

Jan Otop

gramowania, analizie statycznej czy językach zapytań w grafowych bazach danych. Twierdzenie Parikha jest jednym z fundamentalnych twierdzeń w teorii języków formalnych pozwalającym zrozumieć relację pomiędzy językami regularnymi i bezkontekstowymi. Zatem pytanie, czy twierdzenie Parikha pozostaje prawdziwe przy nieskończonym alfabecie jest naturalne oraz ważne.

Wyniki **(W1)**, **(W2)** oraz **(W3)** zostały opublikowane w materiałach bardzo dobrej konferencji LICS (2021), a wynik **(W4)** w materiałach dobrej konferencji FSTTCS (2021). Praca z LICS ma 4 autorów, więc nie jest jasne jaki jest wkład Doktoranta w przedstawione wyniki.

Techniki dowodowe. W Rozprawie jest przedstawiony krótki przegląd teorii zbiorów z atomami w celu wprowadzenia pojęcia orbitowej skończoności oraz powiązanych pojęć. Jest to w zasadzie jedyne zaawansowane technicznie pojęcie używane w Rozprawie — dowody w Rozprawie są najczęściej elementarne i nie polegają na zewnętrznych wynikach.

Dowody przedstawione w pracy są pomysłowe i nietrywialne. Najbardziej skomplikowany dowód Twierdzenia 6.1.1 (o obrazach przemiennych automatów z pojedynczym rejestrem) jest pomysłowy i dość dobrze opisany. Bardzo podobało mi się wykorzystanie tego twierdzenia do pokazania wyników **(W3)** oraz **(W4)**.

Jakość redakcyjna. Rozprawa jest napisana dość dobrze, Doktorant obrazuje przedstawiane pojęcia przykładami a dowody są podzielone na logiczne kroki. Mam jednak pewne zastrzeżenia. Po pierwsze, brakuje mi przy bardziej skomplikowanych argumentach podania intuicji, które by wyjaśniały pojawiające się lematy. Brak intuicji utrudnia czytanie pracy, gdy nie widać związku pomiędzy lematami a dowodzonym twierdzeniem dopóki się nie przeczyta całego dowodu. Na szczęście nie jest to zasada, bo dowód głównego Twierdzenia 6.1.1 jest poprzedzony planem dowodu. Po drugie, pojawiają się niejasne stwierdzenie oraz literówki utrudniające czytanie, które jednak wyjaśniają się przy dalszej lekturze Rozprawy. Listę poprawek zamieszczam w suplemencie do recenzji.

Konkluzja

Uważam, że wyniki przedstawione w pracy są ważne i oryginalne a Doktorant wykazał się umiejętnością rozwiązywania nietrywialnych problemów. Sama rozprawa jest również dobrze zredagowana (lista sugerowanych poprawek jest w suplemencie). Dlatego uważam, że Rozprawa magistra Mohnisha Pattathurajana **spełnia wymagania ustawowe stawiane rozprawom doktorskim** i wnoszę o dopuszczenie go do dalszych etapów postępowania o nadanie stopnia doktora.

Jan Otop

UNIWERSYTET WARSZAWSKI
BIURO RAD NAUKOWYCH
2023 -07- 21
WPŁYNĘŁO
L.dz. 1469 ....Podpis...........

dr hab. Jan Otop, prof. UWr
Instytut Informatyki
Uniwersytet Wrocławski

Wrocław, 14 lipca 2023

Suplement do recenzji rozprawy
Commutative images of languages over infinite alphabets.

There are two lists below: a list of high-level suggestions and a list of low-level comments, which are mostly typos.

The list of high-level suggestions:

- page 13: "Furthermore, as a corollary of the above two results one derives one-register context-free grammars are Parikh-equivalent to register automata, but not necessarily to one-register automata (see also Theorem 8.6.1 below)". This is confusing. You did not prove that Parikh images of 1-NRA do not cover all rational sets over an orbit-finite alphabet. Did I miss something?
- page 14: Figure 1.2. In the transition from the red to the green state I think we need "$x_1 \neq x_2$". As it is now, a_2 and a_3 can be the same. You implicitly assume that registers store different values (page 31). Please make it more explicit or drop this assumption.
- page 15 and the whole paper: I am not a native speaker, but it seems strange that there are not articles before "Hamiltonian cycle" or "Parikh image". Are you sure this is correct?
- page 31: "Implicitely, the input alphabet of (...)" What do you mean by "implicitly"?
- page 34, Example 3.2.11. I think that we need $x \neq x'$ in the second transition.
- page 35: Example 3.2.15. What is the difference between labeling transitions with y instead of no label ($\top$)?
- page 45: Definition 4.1.1. is a bit sloppy, especially in contrast to Definition 5.1.1, which formally addresses orbit-finite unions. It would be perhaps acceptable if a careful definition of rational sets was followed with more informal definition of semilinear sets. However, in my opinion the best solution is to improve Definition 4.1.1 to the standard of Definition 5.1.1.
- page 50: "containing words, where all maximal constant infixes have even length" I would suggest to change "constant" to "same-letter". The current statement is confusing.
- page 55: It would really help, if you could give some intuitions first. Alternatively, you could state first some properties, which are contradictory, so a reader would know what is going on.
- page 64: "In the sequel we assume that the constraint in every transition rule of 1-NRA defines exactly one orbit in A^3" You could have explain in one sentence what does it mean. It would save a reader some effort.
- page 72: The definition of Σ. I would suggest to rewrite it to $\Sigma = \{\langle a, \{b, c\}\rangle \mid a \neq b, a \neq c\}$. Currently, it is unclear whether b, c are some selected constants or the elements from α. I was a bit puzzled.

- page 73. "By the very definition (...)" It is not clear at all! Please comment of this.
- page 75: "directed graphs without self-loops". I see that you use this assumption only once in the proof of Lemma 6.10.7. It should have been repeated. Or better yet, remove this assumption and consider all nodes to have self-loops (as they should).
- page 86 Example 7.2.2. Is it supposed to be the example separating 2-HRA from 1-NRA? If yes, then I would contrast it with Example 3.2.1, where a very similar language is recognized by a 1-NRA (and I think it can be adopted to recognized L'). If not, why do you claim that HRA are strictly more expressible than 1-NRA?
- page 88: "(which is forcedly different than the new value of the second register, but could be equal to the previous value of the first register)" Do you assume that different registers store different values? Could you drop this assumption and state all the constraints explicitly?
- page 89. The proof of Lemma 7.4.3. You could have stated that h' is separated.
- page 95: The definition of Δ': I believe you should drop x_1, x_1' in the definition of the rule from Δ' as you want to have an HRA with $k-1$ registers.
- page 98: "As a corollary, we deduce that the language of every 1-CFG is Parikh-equivalent to the language of an NRA. The number of registers of the NRA may be larger than one." Please be more explicit. The question whether every 1-CFG is Parikh-equivalent to some 1-NRA is open (right?), but it is Parikh-equivalent to some HRA, which is an NRA.
- It would help, if you stated more conjectures in the paper (e.g. as in the above comment), so a reader would know what has been shown and what is unknown.
- page 98: In Section 1, you explain what is the outline of the chapter, rather than what would the a high-lever plan of the proof. It is difficult to keep focus on notions (say compatible pair) if one has no idea what is it used for.

Detailed comments (which is not comprehensive):

- page 10: "no satisfactory extension of Kleene theorem to infinite alphabet" →"(...) to infinite alphabets"
- page 11: $(a_i, a_i + 1) \to (a_i, a_{i+1})$; Unline →Unlike.
- page 15: "Everry"
- page 16 "such tree" →"such a tree"; "the later" →"the latter".
- page 19: "all set" →"all sets".
- page 30: "is is".
- page 35: "several closure," →"several closure properties,"
- page 39: "register value may be testes only once" →"register value may be tested only once"
- page 40: "where last" →"where the last".
- page 43: "mullary rules" →"nullary rules"
- page 48: "the smallest class" →"the least class"?
- page 49: "of L_2" →"of L_1".
- page 50: Example 4.2.2 →Example 4.2.3.
- page 64: "a automaton" →"an automaton", "and end at" →"and ends at".

- page 73: "anti-chain" →"anti-cycle".
- page 77: "OUT_DEG$(d) \geq 3$" →"OUT_DEG$(d) \geq n-3$".
- page 80: "Therefore we will complete the proof of rationality of anti-cycles of bounded order" →"(...) unbounded (...)"
- page 83: "and a also" →"and also".
- page 88: "tools that is is used" →"tools that are is used".
- page 94: "Claim 7.6.2" →"Lemma 7.6.2"; "all these transitions are all at level 1" →"all these transitions are at level 1".

UNIWERSYTET WARSZAWSKI
BIURO RAD NAUKOWYCH
2023 -07- 21
WPŁYNĘŁO
L.dz................Podpis...........