dr hab. Małgorzata Bednarska-Bzdęga
Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

2-05-2022

# Recenzja rozprawy doktorskiej mgra Dominika Bojko
## *Selected topics on randomized algorithms*

Rozprawa ma dwa aspekty: teoretyczny i związany z algorytmami stosowanymi w praktyce. Niniejsza recenzja dotyczy tylko tego pierwszego, ponieważ ocena, na ile uzyskane wyniki są istotne w zastosowaniach algorytmicznych, przekracza moje kompetencje. Recenzja dotyczy rozprawy w wersji drukowanej (która nie jest tożsama z plikiem pdf, który również otrzymałam).

Z bogatej kolekcji wyników uzyskanych przez autora poniżej przedstawiam te, które najbardziej wpłynęły na moją ocenę merytoryczną rozprawy. Opisuję je, nie trzymając się dokładnie sformułowań autora, używając języka probabilistycznego (i pomijając pewne subtelności), by ułatwić zrozumienie czytelnikowi nieobeznanemu z licznymi oznaczeniami z rozprawy. Omawiane wyniki pochodzą z rozdziałów 2, 3 i 4, stanowiących najważniejszą część rozprawy.

### Rozdział 2

Załóżmy, że generujemy losowo i niezależnie $n \geqslant 2$ liczb ze zbioru $[L] = \{1, 2, \ldots, L\}$ zgodnie z rozkładem prawdopodobieństwa $\bar{p}^{(L)}$ na tym zbiorze. Niech $\Pr[\bar{p}^{(L)}, n]$ oznacza prawdopodobieństwo zdarzenia, że wśród wygenerowanych $n$ liczb jest dokładnie jeden element maksymalny. Autor rozważa problem w kilku wariantach.

(1) Dla ustalonych $n$ i $L$ chcemy tak dobrać rozkład $\bar{p}^{(L)}$ (który może zależeć od $n$), by prawdopodobieństwo $\Pr[\bar{p}^{(L)}, n]$ było jak największe. Rozkład $\bar{p}^{(L)}$, dla którego $\Pr[\bar{p}^{(L)}, n]$ jest największe, nazwijmy optymalnym. Autor wykazuje, że taki rozkład optymalny jest dokładnie jeden dla każdego $n \geqslant 2$ i $L \in \mathbb{N}$ oraz wyznacza go w postaci rekurencyjnej (twierdzenie 2.5.1). Autor podaje też inną zależność rekurencyjną, charakteryzującą rozkład optymalny, związaną z prawdopodobieństwami $q_L(n)$, że w danym kroku wygenerujemy liczbę mniejszą od $L$ (twierdzenia 2.5.2 i 2.5.3). Problem jest również rozważany w wersji asymptotycznej.

(2) Dla ustalonego $L$ zakładamy, że rozkład $\bar{p}^{(L)}$ jest jednostajny. Autor zauważa, że wtedy $\Pr[\bar{p}^{(L)}, n]$ jest funkcją malejącą zmiennej $n$ (twierdzenie 2.6.3).

(3) Załóżmy, że rozkład $\bar{p}^{(L)}$ jest jednostajny. Dla ustalonych $\varepsilon \in (0, 1)$ i $N \in \mathbb{N}$ szukamy najmniejszego $K \in N$ o tej własności, że dla wszystkich $n \leqslant N$ i $L \leqslant 2^K$ zachodzi $\Pr[\bar{p}^{(L)}, n] \geqslant 1 - \varepsilon$. Autor znajduje najmniejsze takie $K$ (jego postać przedstawiona jest w twierdzeniu 2.6.4).

(4) Ustalmy $N \in \mathbb{N}$ i $\varepsilon \in (0, 2/27)$ i załóżmy, że rozkład $\bar{p}^{(L)}$ jest ograniczonym do $[L]$ rozkładem geometrycznym, z prawdopodobieństwem sukcesu $p$. Badamy, jakie warunki dla $L$ i $p$ – zależnych od $N$ i $\varepsilon$ – gwarantują, że dla wszystkich $n \leqslant N$ zachodzi $\Pr[\bar{p}^{(L)}, n] \geqslant 1 - \varepsilon$.

Rozważmy teraz inny eksperyment losowy, składający się z dwóch etapów. W pierwszym etapie generujemy losowo i niezależnie $n \geqslant 2$ liczb ze zbioru $[L_1]$, zgodnie z ograniczonym do $[L_1]$ rozkładem geometrycznym, z prawdopodobieństwem sukcesu $p$. Następnie sprawdzamy, ile jest liczb największych w tak otrzymanym ciągu – załóżmy że jest ich $X$. W drugim etapie $X$-krotnie losujemy w sposób jednostajny liczbę ze zbioru $[L_2]$. Niech $\Pr[Sc_n(p, L_1, L_2)]$ oznacza prawdopodobieństwo zdarzenia, że wśród wygenerowanych w drugim etapie liczb jest dokładnie jeden element maksymalny. Również w tym wariancie autor rozważa kilka problemów.

(5) Czy $\Pr[Sc_n(p, L_1, L_2)]$ jest funkcją malejącą względem $n$?

(6) Ustalmy $\varepsilon \in (0, 0.6)$ i $N \in \mathbb{N}$. Badamy, jakie warunki dla $p, L_1, L_2$ (jako funkcji $N$ i $\varepsilon$)) gwarantują, że dla wszystkich $n \leqslant N$ zachodzi $\Pr[Sc_n(p, L_1, L_2)] \geqslant 1 - \varepsilon$.

Rozdział 3

Problem I. Dana jest liczba naturalna $n$ i rozkład prawdopodobieństwa $\bar{\varrho} = (\varrho_i)_{i\in[n]}$. Załóżmy, że istnieje łańcuch Markowa $(X_i)$ na zbiorze stanów $[n]$, zwany diabelskimi schodami (zdefiniowany w podrozdziale 3.2.4), o tej własności, że dla wszystkich $i \geqslant n$ oraz $j \in [n]$ zachodzi $\Pr[X_i = j] = \varrho_j$. Dla szerokiej klasy rozkładów $\bar{\varrho}$ autor konstruuje algorytm implementujący wspomniany łańcuch Markowa w taki sposób, by algorytm był oszczędny w sensie pamięci i liczby wywołań generatora liczb losowych. Autor analizuje własności tego algorytmu, między innymi wyznacza za pomocą dystrybuanty rozkładu $\bar{\varrho}$ wartość oczekiwaną i dystrybuantę dla liczby kroków algorytmu, pomiędzy dwoma kolejnymi wywołaniami generatora liczb losowych. Klasa rozkładów $\bar{\varrho}$, dla których algorytm ten można stosować, nie obejmuje np. rozkładu jednostajnego, więc w celu rozszerzenia rozważanej klasy rozkładów autor konstruuje algorytm ogólniejszy, którego sercem jest implementacja kilku łańcuchów Markowa, będących diabelskimi schodami.

Problem II. Dany jest ciąg $\alpha_1, \alpha_2, \ldots$ liczb z przedziału $(0, 1)$. Przypuśćmy, że wykonujemy kolejno (nieskończenie wiele razy) próby Bernoulliego, przy czym w próbie $j$ prawdopodobieństwo sukcesu wynosi $\alpha_j$, dla $j = 1, 2, \ldots$. Po $n$-tej próbie przez $K_n$ oznaczamy liczbę prób wykonanych od ostatniego sukcesu (jeśli $n$-ta próba zakończyła się sukcesem, wtedy $K_n = 1$). Autor studiuje asymptotyczne własności ciągu zmiennych losowych $K_n$ (przy $n \to \infty$), dla ciągów $(\alpha_n)$ postaci $\alpha_n = g/n^a$, gdzie $g > 0$ i $a \geqslant 0$. Dla ustalonych parametrów $g$ i $a$ badana jest wartość oczekiwana $E(K_n)$ oraz zbieżność $K_n$ (niekiedy po przeskalowaniu) względem dystrybuanty. Jak widać w podsumowaniu wyników w podrozdziale 3.3.2, wartości $\alpha$ można podzielić na 6 przedziałów, w których asymptotyka jest bardzo odmienna.

Rozdział 4

Rozważmy łańcuch Markowa $(M_n)$ o zbiorze stanów $\mathbb{N}_0$, liczbę nieujemną $\delta$ i ciąg liczb $(\varepsilon_n)$. Mówimy, że $(M_n)$ spełnia warunek $((\varepsilon_n), \delta)$-DP, jeżeli dla wszystkich $n, m$ różniących się o 1 i dla każdego $S \subseteq \mathbb{N}_0$ zachodzi $\Pr[M_n \in S] \leqslant e^{\varepsilon_n} \Pr[M_m \in S] + \delta$. Autor wyznacza $(\varepsilon_n)$ i $\delta$ wystarczające do tego, by $(M_n)$ spełniał warunek $((\varepsilon_n), \delta)$-DP, w przypadku dwóch procesów Markowa: dla tak zwanego licznika Morrisa (twierdzenie 4.5.1) i dla łańcucha, w którym $M_n$ jest maksimum $n$ niezależnych zmiennych losowych o (przesuniętym) rozkładzie geometrycznym $Geo(1/2)$ (twierdzenie 4.5.5).

**Ocena merytoryczna**

Poniżej przez (1)–(6) oznaczam opisane wyżej zagadnienia. W rozdziale 2, w części dotyczącej zagadnień (1)–(4), autor postawił problem naturalny na gruncie rachunku prawdopodobieństwa i przeanalizował go pod różnymi kątami, metodami wymagającymi solidnego warsztatu matematycznego. Dodatkowym atutem jest fakt, że uzyskane wyniki mają zastosowanie w teorii algorytmów, w problemach wyboru lidera, które mają solidne umocowanie w literaturze. Do uzyskania tych wyników autor posłużył się wprawdzie standardowymi metodami, ale wymagały one sporo pracy i biegłości w używaniu różnorodnych narzędzi analitycznych i rachunku prawdopodobieństwa. Wyniki teoretyczne uzupełnił nietrywialnymi symulacjami, co wymagało dobrego warsztatu informatycznego.

Autor poruszył bardzo ciekawy problem monotoniczności względem $n$ prawdopodobieństwa, że wśród wygenerowanych wg danej procedury losowej $n$ liczb jest dokładnie jeden element maksymalny. Dla procesu losowego (2), gdy rozkład $\bar{p}^{(L)}$ jest jednostajny, ta monotoniczność zachodzi. Z kolei – jak zauważa autor w uwagach kończących dodatek B – gdyby rozkład $\bar{p}^{(L)}$ był ograniczonym rozkładem geometrycznym, monotoniczności w ogólnym przypadku by nie było. Tym ciekawszym wynikiem byłoby twierdzenie 2.9.1 o monotoniczności w przypadku procesu losowego opisanego przed (5), a w rozprawie nazywanego algorytmem GULE. Niestety dowód tego twierdzenia budzi moje wątpliwości i nie wiem, dla jakich parametrów $K_1, K_2$ twierdzenie jest prawdziwe. Na pewno w takiej ogólności, jak jest w pracy, twierdzenie nie zachodzi, bowiem autor dopuszcza przypadek, że $L_2 = 2^{K_2} - 1 = 1$ i wtedy z twierdzenia 2.9.1 wynikałaby monotoniczność dla wspomnianego wyżej ograniczonego rozkładu geometrycznego – wiemy zaś, że wtedy monotoniczności nie ma. W dowodzie twierdzenia 2.9.1 niepoprawne wg mnie są szacowania prawdopodobieństw w części dotyczącej scenariusza 3. Rozważane tam są pewne prawdopodobieństwa warunkowe (tak podejrzewam, choć autor tego nie napisał), przy założeniu, że maksimum pewnych zmiennych losowych wynosi $m$. Fakt, że są to prawdopodobieństwa warunkowe nie został uwzględniony w niektórych szacowaniach. Nie widzę też, jak uzasadnić oszacowanie dolne na $\Pr[B_n]$, a uzasadnienie autora jest niewystarczające. Bez

twierdzenia 2.9.1 główne twierdzenie podrozdziału 2.9, czyli twierdzenie 2.9.3, zachodziłoby dla $n = N$, nie dla $n \leqslant N$, niemniej dowód twierdzenia 2.9.3 pozostałby bez zmian.

Jednym z głównych twierdzeń rozdziału 2 jest twierdzenie 2.8.1. Jest to drugie (i ostatnie) twierdzenie, na którego dowodzie utknęłam w martwym punkcie. Dowód jest skomplikowany technicznie, zajmuje 13 stron dodatku B, a miejsce sprawiające mi problem jest tylko jedno: Skąd wzięła się pierwsza równość w (B.11)? Jest dla mnie jasne, że $S_{n,Geo(p,L)}$ jest iloczynem dwóch opisanych przez autora zdarzeń, ale nie jest dla mnie jasne, że te dwa zdarzenia są niezależne. Jeśli są zależne, to wszystko byłoby dobrze, gdyby znak $=$ w (B.11) zastąpić przez $\geqslant$, ale taką nierówność należałoby uzasadnić (nie wiem, jak to zrobić).

W rozdziale 3 podoba mi się sposób, w jaki autor prowadzi analizę problemu I. Nie ma tu fajerwerków w dowodach, natomiast budowanie teorii jest ładne: Najpierw autor rozważa klasę funkcji dyskretnie wklęsłych, uzasadnia dlaczego akurat ta klasa jest dobrym wyborem, następnie rozwija swój argument (oparty na diabelskich schodach), by klasę rozszerzyć o inne funkcje. Wg mnie świadczy to o dojrzałości matematycznej autora. Problem II (w pracy nazywany PLU) dotyczy zmiennych losowych $K_n$, bardzo naturalnych z probabilistycznego punktu widzenia. Równie naturalne jest badanie asymptotyki dla $E(K_n)$ i zbieżności wg dystrybuanty. Dowody są nieco żmudne rachunkowo, za to wskazują na dobre umiejętności techniczne autora.

Rozdział 4 wzbudził najmniej mojego entuzjazmu. Wg mnie odbiega od pozostałych pod względem chociażby precyzji językowej i staranności w definiowaniu oznaczeń i pojęć. Autor analizuje dwa znane w literaturze algorytmy a analiza jest dość komplikowana rachunkowo. W niektórych miejscach dowodów uzasadnienia były dla mnie zbyt skąpe, o czym piszę niżej, wśród uwag do dowodów, ale nie było to istotne dla poprawności samych stwierdzeń.

Pewien niedosyt budzi fakt, że liczne wyniki zawarte w rozprawie nie przełożyły się na choćby kilka publikacji.

## Pomniejsze uwagi i pytania do dowodów i twierdzeń

(a) s.69, pierwszy akapit podrozdziału 2.5.6, „The definition of $\bar{p}(n)^{(L)}$ and Lemma 6 show that ... $\leqslant \Pr[\bar{p}^{(L)}(n), n]$":
Nie widzę, w jaki sposób ta nierówność wynika z definicji $\bar{p}(n)^{(L)}$ i lematu 6.

(b) s.69, początek podrozdziału 2.5.6:
Dlaczego $\Pr[\bar{p}^{(L)}(n), n-1] \leqslant \Pr[\bar{p}^{(L)}(n), n]$?

(c) s.70, dowód twierdzenia, 2.5.7, ostatni akapit:
Dlaczego $\Pr[\bar{p}^{(L)}(n), n] \leqslant 1 - \frac{1}{L}$?

(d) s.102, założenia lematu 9:
Powinno być $\ln(n-1)$ zamiast $\ln n$.

(e) s.102, treść lematu 10:
W mianowniku powinno być $2L_2$ zamiast $L_2$.

(f) s.146, treść wniosku 7:
W mianowniku powinno być $n^2$ zamiast $n$.

(g) s.159, twierdzenie 4.5.1:
Pierwszy wyeksponowany wzór tego twierdzenia jest nieco mylący. Nie taka nierówność jest dowodzona i nie z niej wynika postulowana własność DP.

(h) s.161, po stwierdzeniu 2, „This claim emerges from lemmas 27 and 28 from Appendix E.":
Nie widzę tego wynikania. (Niemniej stwierdzenie 2 jest prawdziwe.)

(i) s.161, akapit po stwierdzeniu 2, „we are going to use Claim 1 from Appendix E":
Nie znalazłam Claim 1 w dodatku E.

(j) s.164, założenia twierdzenia 4.5.5 (i później s.169, zdanie po przykładzie 12):
W mianowniku powinno być $e^{\varepsilon}$ zamiast $e^{-\varepsilon}$.

Uwaga: Bardzo dokładnie sprawdziłam tylko wybrane dowody. Sprawdzenie wszystkich kroków rozumowań w tej liczącej 229 stron pracy zajęłoby mi dwu- lub trzykrotnie więcej czasu niż wymagane ustawowo 2 miesiące.

## Uwagi redakcyjne

Praca jest napisana w języku angielskim. Czytelne i starannie opracowane są rysunki i tabele. Autor dołożył starań, by wyjaśniać to, o czym pisze, matematycznie i na poziomie intuicyjnym. W zdecydowanej większości tekstu dobrze mu się to udało, choć małe zastrzeżenia do niektórych fragmentów rozprawy mam. Kilka z poniższych uwag ma charakter bardzo subiektywny. Żadne nie wpływają na moją ogólną, pozytywną ocenę pracy.

- Praca zyskałaby na korekcie językowej. Powtarzające się (drobne i na szczęście nieczęste) błędy językowe nieco psują przyjemność czytania.
- Za dużo jest w pracy przypisów. Utrudniały mi czytanie i rozpraszały. Co gorsza, odnośniki do przypisów pojawiają się przy formułach i symbolach matematycznych. Jednym z przykładów jest $D^3$ na s. 115. Trójka jest tu odnośnikiem.
- Szkoda, że w pracy nie ma spisu oznaczeń. Jest spis pseudokodów, ilustracji i tabel, ale nie miałam z nich potrzeby korzystać, w przeciwieństwie do spisu oznaczeń.
- Osobna numeracja dla: twierdzeń, lematów, faktów, obserwacji itp. bardzo utrudniła mi śledzenie rozumowań. Przy tak długiej pracy szukanie twierdzenia o danym numerze, gdy po drodze mija się numery na przemian mniejsze i większe, jest męczące.
- Niektóre odnośniki do pozycji w bibliografii są niewłaściwe, np. na s.78 jest [62], a powinno być [64]; cytowanych [97] i [98] w bibliografii nie ma. Ten problem być może występuje tylko w wersji drukowanej rozprawy, bo w pliku pdf bibliografia i odnośniki do niej są inne. Niestety zauważyłam to już po zrecenzowaniu pracy.
- podrozdział 2.9.1: Niejasne jest, na jakim zbiorze rozważany jest rozkład jednostajny. Z fragmentu o tytule „A description of GULE" wynika, że chodzi o zbiór $[0 : L_2 - 2]$, potem w „Arbitrary simulations" rozkład jest na zbiorze $[1 : L_2]$ i $L_2 = 2^{K_2}$, następnie w dowodzie twierdzenia 2.9.1 mamy $L_2 = 2^{K_2} - 1$ i nie podano, na jakim zbiorze rozpatrywany jest rozkład jednostajny.
- Pomysł na przedstawienie rozumowania w podrozdziale 2.9.5 nie był dla mnie, jako czytelniczki, dobry. Rozumiem, że intencją autora było pokazać, jak rodziły się oszacowania parametrów w twierdzeniu 2.9.3, lecz mi ta lektura raczej zaciemniła obraz, niż pomogła. W takim „rozumowaniu od tyłu" pojawia się np. problem dwuznaczności wyrażeń typu „X must hold": Czy chodzi o implikację i jeśli X nie zajdzie, to algorytm nie zadziała, czy raczej o życzenie, by X zaszło, dzięki czemu nierówności będą takie, jak nam pasuje. W wielu miejscach nie wiedziałam, czy autor coś zakłada, czy chce, by zaszło. Chyba jednak wolałabym, by podane zostało najpierw twierdzenie 2.9.5 a potem zwięzłe objaśnienie intuicji stojącej za wyborem parametrów. Z drugiej strony, zdaję sobie sprawę, że takie zwięzłe, intuicyjne wyjaśnienie byłoby tu wyzwaniem.
- s.86, drugi wyeksponowany wzór: $Y_1$ i $Y_2$ nie zostały wcześniej zdefiniowane.
- podrozdział 3.3.2: Pojawia się zmienna losowa $L_n$, zdefiniowana dopiero w następnym podrozdziale.
- s.139, Lemma 11: Lemat zaczyna zwrot "Monotonicity Let".
- Nie jestem przekonana, że dobrym pomysłem jest używanie dwóch oznaczeń na kwantyfikator generalny, tzn. zarówno $\bigwedge$, jak i $\forall$.
- W podrozdziale 4.2 mylące dla mnie jest używanie raz $\mathbb{N}^{card(\mathcal{X})}$ a w innych miejscach $\mathbb{N}_0^{card(\mathcal{X})}$.
- s.153, definicja 4.2.2: Po tej kluczowej definicji spodziewałabym się większej precyzji. Mamy tu niezdefiniowane wcześniej pojęcie Range($\mathcal{M}$), mamy też definicję pojęcia $(\varepsilon, \delta)$-differentialy private algorithm, podczas gdy później to pojęcie stosowane jest do ciągu zmiennych losowych (czy ciąg zmiennych losowych jest algorytmem?), zaś nierówność definiująca to pojęcie używana jest nie tylko dla ustalonego $\varepsilon$, ale też ciągu $(\varepsilon_n)$. Wprawdzie w tekście pojawia się później informacja, że możemy uzależnić $\varepsilon$ od $n$, lecz nierówność będąca odpowiednikiem tej z definicji 4.2.2 powinna być zapisana, a nie pozostawiona domyślności czytelnika.
- s.153, po definicji 4.2.2: Oznaczenie $(\varepsilon, \delta)$-DP nie zostało zdefiniowane.
- s.154, akapit przed faktem 4.3.1: Czym są $x$ i $y$? Co to jest $M(x)$?

- s.157, przedostatni akapit: Brakuje „bits" przed „are required".
- s.159, pierwszy wyeksponowany wzór: Brakuje indeksu $k$ przy HyperLogLog.
- s.164, twierdzenie 4.5.5: Zbiór $S$ nie został zdefiniowany.
- s.171: Pojawia się niezdefiniowane oznaczenie PRNG.
- s.177, opis tabeli A.4: Powinno być $L = 8$ zamiast $n = 8$.
- s.182, obserwacja 4: Podejrzewam, że zmienna losowa $W_{n,p}$ ma rozkład WMGeo$(n, p)$, ale nie zostało to napisane.
- s.184, wniosek 5: Założenie Asm.3 jest tu niepotrzebne i trochę mylące, bo później autor używa wniosku 5, by uzasadnić (B.16), gdzie warunek Asm.3 jest tezą.
- W podziękowaniach autor dziękuje swoim współautorom. To nieco zaskakujące, bo bibliografia nie zawiera publikacji autora.

**Konkluzja**

Stwierdzam, że rozprawa mgra Dominika Bojko spełnia zwyczajowe i ustawowe wymagania dla prac doktorskich. Uważam, że pan Bojko w pełni zasługuje na nadanie mu stopnia doktora.

Małgorzata Bednarska-Bzdęga