

Milé čtenářky, milí čtenáři,

máte v rukou nové dvojčíslo časopisu Kvaternion. Časopisu, který se už od roku 2012 snaží – s důrazem na čtivost a srozumitelnost – přinášet mix odborných a přehledových matematických článků, dále příspěvků založených na vybraných pracích studentů oboru Matematického inženýrství na FSI VUT v Brně a případných dalších typů textů (jako např. informací o matematických soutěžích, či rozborů zajímavých úloh).

Věříme, že i letošní dvojčíslo přináší Vám, čtenářům, pestrou nabídku témat, která odpovídá našim původním záměrům. Posuďte sami. Sekci odborných a přehledových článků otevírá studie Jiřího Klašky věnovaná Jakóbczykově hypotéze o Fermatových číslech; na tomto místě lze doporučit i související autorův článek o Mersennových číslech, který vyšel v Kvaternionu v r. 2021. Následuje příspěvek Radovana Potůčka prezentující elegantní způsoby, jak sečíst jisté redukované harmonické řady, tedy harmonické řady, kde jsou některé členy vynechány. Autorem článku o kombinatorických hrách je Václav Vopravil. Konkrétně je uvažována hra Hackenstring, pro kterou jsou analyzovány vítězné strategie pro určité pozice. Čtenář jistě ocení detailní rozpracování textu. Prvním ze „studentských“ článků je příspěvek Alžběty Kočendové, který jednak popisuje matematický aparát Hausdorffovy dimenze, dále se věnuje detekci blesku v obraze a též poskytuje výpočet Hausdorffovy dimenze blesku. Dalším textem založeným na bakalářské práci je příspěvek Jana Michálka zabývající se analýzou, která odhaluje výhody použití geometrických algeber pro aplikace v kvantovém počítání. Dušan Oberta ve svém článku přibližuje čtenářům tzv. regresní stromy a náhodné lesy, což jsou užitečné statistické nástroje. Stojí za zmínku, že řada autorových teoretických úvah je v literatuře nová. Posledním článkem tohoto dvojčísla je příspěvek Lubomíra Pažourka. Je rovněž založen na bakalářské práci a zabývá se matematickou podstatou některých rankingových metod, kde jejich jednotícím prvkem je Perronova-Frobeniova věta.

Zábavné a inspirativní čtení Vám jménem redakčního kolektivu přeje

Pavel Řehák

JAKÓBCZYKOVA HYPOTÉZA O FERMATOVÝCH ČÍSLECH

JIŘÍ KLAŠKA

Věnováno profesoru Michalu Krížkovi

ABSTRAKT. Článek je volným pokračováním předchozí autorovy studie [15], která byla věnována hypotéze polského matematika Franciszka Jakóbczyka (1905-1992) o Mersennových číslech. Neméně zajímavá je Jakóbczykova hypotéza o Fermatových číslech, která tvrdí, že každé Fermatovo číslo je buď prvočíslo, nebo je součinem různých prvočísel. Hypotéza byla poprvé publikována v roce 1951 a její důkaz nebyl do dnešní doby nalezen.

1. FERMATOVA ČÍSLA

Fermatova čísla jsou přirozená čísla tvaru

$$F_n = 2^{2^n} + 1, \text{ kde } n \in \mathbb{N} \cup \{0\}. \quad (1.1)$$

Na základě definice (1.1) můžeme snadno určit hodnoty prvních pěti Fermatových čísel:

$$F_0 = 3, F_1 = 5, F_2 = 17, F_3 = 257, F_4 = 65537. \quad (1.2)$$

Čísla F_n poprvé podrobně studoval francouzský matematik Pierre de Fermat (1601–1665) v souvislosti s otázkou, která z čísel tvaru $2^m + 1$, $m \in \mathbb{N} \cup \{0\}$, mohou být prvočísla. Tento problém byl přirozeným rozšířením Fermatových úvah o prvočíselnosti Mersennových čísel, to je čísel $M_m = 2^m - 1$, kde $m \in \mathbb{N} \cup \{0\}$. Připomeňme, že pro Mersennova čísla platí následující zajímavá implikace:

$$\text{Je-li } 2^m - 1 \text{ prvočíslo, pak } m \text{ je prvočíslo.} \quad (1.3)$$

Opačná implikace k implikaci (1.3) ale neplatí, protože například $2^{11} - 1 = 23 \cdot 89$. Ve skutečnosti je do roku 2023 známo pouze 51 Mersennových prvočísel a další se intenzivně hledají [13]. Více podrobností o Mersennových číslech lze nalézt v knihách [7] a [20] nebo též v autorově článku [15].

Pro čísla $2^m + 1$ platí následující tvrzení:

$$\text{Je-li } 2^m + 1 \text{ prvočíslo a } m > 1, \text{ pak } m = 2^n \text{ pro nějaké } n \in \mathbb{N}. \quad (1.4)$$

Je zřejmé, že právě objev implikace (1.4) vedl Fermata ke studiu čísel (1.1), která dnes nazýváme jeho jménem. Další částí příběhu Fermatových čísel bylo zjištění, že

2020 MSC. Primární 11A41; Sekundární 11A07, 11-02.

Klíčová slova. Fermatova čísla, Franciszek Jakóbczyk, Wieferichova prvočísla.

všechna čísla uvedená v seznamu (1.2) jsou prvočísla. To vedlo Fermata k formulaci Hypotézy 1.1.

Hypotéza 1.1 (Fermat, 1640). *Každé Fermatovo číslo $F_n = 2^{2^n} + 1$ je prvočísl.*

I když byl Fermat o správnosti Hypotézy 1.1 zcela přesvědčen, její důkaz nepředložil. Naznačoval však, že Hypotézu 1.1 bude možné dokázat metodou nekonečného sestupu [8, str. 375]. Na základě dochované korespondence se rovněž zdá, že další významný francouzský matematik Bernard Frénicle de Bessy (1604–1674) s Fermatovým tvrzením souhlasil. Historické podrobnosti o korespondenci mezi Fermatem a Fréniclem z roku 1640 lze nalézt v Fletcherově článku [11]. Navíc Marin Mersenne (1588–1648) v knize *Novarum Physico-Matematicarum* z roku 1647 uvedl, že každé Fermatovo číslo je prvočísl. Uvedené skutečnosti byly příčinou, že Hypotéza 1.1 byla považována za dokázané tvrzení a téměř sto následujících let se problémem nikdo nezabýval. Teprve v roce 1732 švýcarský matematik Leonhard Euler (1707–1783) pomocí protipříkladu dokázal, že tvrzení Hypotézy 1.1 není pravdivé. Euler našel rozklad čísla

$$F_5 = 4294967297 = 641 \cdot 6700417.$$

Tento objev oživil zájem matematiků o studium Fermatových čísel a jejich vlastností. Eulerovým objevem také vznikla nová otázka, totiž zda Fermatových prvočísel je konečně nebo nekonečně mnoho.

Mezi nejkrásnější doposud dokázaná tvrzení o Fermatových číslech patří věta, kterou dokázal švýcarský matematik Christian Goldbach (1690–1764) a kterou dnes nazýváme Goldbachova věta.

Věta 1.2 (Goldbach, 1730). *Žádná dvě různá Fermatova čísla nejsou dělitelná stejným prvočíslem (tj. nemají společného dělitele většího než jedna).*

Snadným důsledkem Goldbachovy věty je fundamentální matematické tvrzení, totiž že množina všech prvočísel je nekonečná. Další krásné a zcela mimořádné tvrzení objevil a dokázal německý matematik Carl Friedrich Gauss (1777–1855). Toto tvrzení nečekaným způsobem propojilo problematiku Fermatových prvočísel s geometrií.

Věta 1.3 (Gauss, 1796). *Pravidelný mnohoúhelník je eukleidovsky konstruovatelný (tj. pouze pomocí pravítka a kružítka) právě tehdy, když počet jeho vrcholů je roven číslu $n = 2^r \cdot p_1 \cdots p_s$, kde $n \in \mathbb{N}$, $n \geq 3$, $r, s \in \mathbb{N} \cup \{0\}$ a $p_1, \dots, p_s$ jsou navzájem různá Fermatova prvočísla.*

Protože doposud nevíme, zda seznam (1.2) obsahuje všechna Fermatova prvočísla, nevíme ani kolik existuje pravidelných mnohoúhelníků, které lze eukleidovsky zkonstruovat. Na základě Gaussovy věty ale víme, že existuje aspoň 31 pravidelných mnohoúhelníků s lichým počtem vrcholů, které lze zkonstruovat pomocí pravítka a kružítka.

V roce 1854 našel dánský matematik Thomas Clausen (1801–1885) rozklad čísla

$$F_6 = 18446744073709551617 = 274177 \cdot 67280421310721.$$

Clausen o svém výsledku informoval Gausse v korespondenci ze dne 1. ledna 1855, tedy krátce před Gaussovou smrtí. Gauss zemřel 23. února 1855. Clausenův důležitý objev nebyl nikdy publikován, až v roce 1964 Kurtem Reinhardem Biermannem v historické studii [1, str. 185]. Tato skutečnost vysvětluje, proč jako objevitel rozkladu čísla F_6 je ve většině článků [4, str. 431], [6, str. 175] i knih [7, str. 172], [8, str. 377] uváděn francouzský matematik Fortuné Landry (1799–1895), který rozklad našel v roce 1880 a publikoval v *Comptes rendus* [21]. Analýzou Landryho výsledku se podrobně věnoval H. C. Williams v článku [39]. Mezi literární zdroje, které uvádí informaci o prvním objeviteli rozkladu Fermatova čísla F_6 správně, jsou například velmi obsažné a kvalitní knihy [20, 32]. To, že se informace o Biermannově článku šíří matematickou komunitou jen pomalu, svědčí poznámka v knize Paula Ribenboima [32, str. 85] z roku 1996, kde sděluje, že na Biermannovu studii ho upozornil jeho kolega A. Hinz.

Důležitou větu, která popisuje tvar prvočíselných dělitelů Fermatových čísel publikoval v roce 1878 [24] francouzský matematik François Édouard Anatole Lucas (1842–1891).

Věta 1.4 (Lucas, 1878). *Nechť $m \in \mathbb{N}$, $m > 1$ a nechť p je libovolné prvočíslo takové, že $p \mid F_m$. Pak $p \equiv 1 \pmod{2^{m+2}}$. To znamená, že p je tvaru $p = k \cdot 2^n + 1$, kde $k, n \in \mathbb{N}$ a $n \geq m + 2$.*

V roce 1903 Alfred Edward Western [6] úspěšně použil Lucasovu větu k nalezení prvočíselných dělitelů několika Fermatových čísel:

$$37 \cdot 2^{16} + 1 \mid F_9, \quad 397 \cdot 2^{16} + 1 \mid F_{12}, \quad 7 \cdot 139 \cdot 2^{16} + 1 \mid F_{12}, \quad 13 \cdot 2^{20} + 1 \mid F_{18}. \quad (1.5)$$

Do druhé světové války pak bylo nalezeno několik dalších dělitelů Fermatových čísel, ale žádný výrazný pokrok ve studiu problematiky Fermatových čísel nenastal. S nástupem počítačů bylo logické očekávat, že nové výpočetní možnosti přinesou nové výsledky a další kompletní rozklady Fermatových čísel. To nastalo až v roce 1970, kdy Michael A. Morrison a John Brillhart [26] objevili rozklad Fermatova čísla F_7 :

$$F_7 = 59649589127497217 \cdot 5704689200685129054721.$$

Pro další výklad bude vhodné zavést, v souladu s označením [4, str. 429], následující konvenci. Pro dané $n \in \mathbb{N}$ nechť P_n označuje prvočíslo mající v dekadické soustavě právě n cifer a C_n nechť označuje složené přirozené číslo mající právě n cifer. Bude-li v dalším textu třeba rozlišit dvě různá prvočísla se stejným počtem cifer, použijeme u jednoho z čísel jako horní index symbol hvězdička. Zavedenou konvenci použijeme již při formulaci následujícího objevu. V roce 1980 našli Richard P. Brent a John M. Pollard [3] kompletní rozklad Fermatova čísla F_8 :

$$F_8 = 1238926361552897 \cdot P_{62},$$

kde

$$P_{62} = 93461639715357977769163558199606896585051237541638188580280321.$$

V roce 1990 A. K. Lestra, H. W. Lestra, M. S. Manasse a J. M. Pollard [23] za asistence mnoha spolupracovníků a využití přibližně různých 700 pracovišť dokončili rozklad Fermatova čísla F_9 :

$$F_9 = 2424833 \cdot 7455602825647884208337395736200454918783366342657 \cdot P_{99},$$

kde

$$P_{99} = 74164006262753080152478714190193747405994078109751 \\ 9023905821316144415759504705008092818711693940737.$$

Připomeňme, že faktor $2424833 = 37 \cdot 2^{16} + 1$ objevil již Western v roce 1903, viz (1.5).

V roce 1995 našel Richard P. Brand [4] kompletní rozklad čísla F_{10} :

$$F_{10} = 45592577 \cdot 6487031809 \cdot 4659775785220018543264560743076778192897 \cdot P_{252},$$

kde

$$P_{252} = 130439874405488189727484768796509903946608530841611892186895295 \\ 776832416251471863574140227977573104895898783928842923844831149 \\ 032913798729088601617946094119449010595906710130531906171018354 \\ 491609619193912488538116080712299672322806217820753127014424577.$$

Pro úplnost poznamenejme, že faktor 45592577 objevil již v roce 1953 Selfridge a faktor 6487031809 našel Brillhart v roce 1962.

Je jistě pozoruhodné, že rozklad Fermatova čísla F_{11} byl znám dříve než kompletní rozklady čísel F_9 a F_{10} . Rozklad čísla F_{11} byl nalezen již v roce 1888 a má tvar

$$F_{11} = 319489 \cdot 974849 \cdot 167988556341760475137 \cdot 3560841906445833920513 \cdot P_{564}.$$

Faktory 319489 a 974849 čísla F_{11} objevil již v roce 1899 Cunningham [6, str. 175]. Hlavní, zbývající část rozkladu čísla F_{11} našel Richard P. Brand [2]. Přesnou hodnotu velkého prvočísla P_{564} může čtenář nalézt například v knize [20, str. 209].

Kompletní rozklad čísla F_{12} a ani žádného dalšího Fermatova čísla není do roku 2023 znám. Částečný rozklad čísla F_{12} má tvar

$$F_{12} = 114689 \cdot 26017793 \cdot 63766529 \cdot 190274191361 \cdot 1256132134125569 \cdot$$

$$568630647535356955169033410940867804839360742060818433 \cdot C_{1133}.$$

Faktor 114689 objevili v roce 1877 Lucas a Pervouchine, faktory 26017793 a 63766529 našel v roce 1903 Western, faktor 190274191361 objevili v roce 1974 Hallyburton a Brillhart a faktor 1256132134125569 našel v roce 1986 Baillie. Poslední pokrok v rozkladu čísla F_{12} nastal v roce 2010, kdy Michael Vang objevil faktor

$$P_{54} = 568630647535356955169033410940867804839360742060818433.$$

Problém určit všechny prvočíselné dělitele složeného čísla C_{1133} je jednou z velkých výzev před kterými výzkumníci Fermatových čísel v současné době stojí. Pokud je

číslo C_{1133} rovno součinu dvou přibližně stejně velkých prvočísel, pak může nalezení jeho rozkladu trvat ještě velmi dlouhou dobu. Na druhé straně nelze vyloučit, že pokrok v konstrukci kvantových počítačů spolu s objevy nových postupů faktorizace přirozených čísel umožní rozložit toto číslo mnohem rychleji, než si dnes dokážeme představit. Sledovat další vývoj faktorizace čísla C_{1133} může být proto velmi zajímavé a poučné. Aktuální informace o nově nalezených dělitelech Fermatových čísel F_n je možné sledovat například na internetové stránce *Distributed Search for Fermat Number Divisors* [9] nebo na stránce *Prime factors $k \cdot 2^n + 1$ of Fermat numbers F_m and complete factoring status* [29]. Internetové stránky [9] a [29] nejsou jedinými stránkami, které se Fermatovými čísly zabývají. Seznam dalších stránek lze nalézt v [20, str. 243].

Vědomosti dosažené o prvočíselných rozkladech Fermatových čísel $F_5, \dots, F_{12}$ za 383 let jejich výzkumu (1640–2023) lze shrnout do následujícího chronologického přehledu.

$$\begin{aligned}
F_5 &= P_3 \cdot P_7, & (1732) \\
F_6 &= P_6 \cdot P_{14}, & (1854) \\
F_7 &= P_{17} \cdot P_{22}, & (1970) \\
F_8 &= P_{16} \cdot P_{62}, & (1980) \\
F_9 &= P_7 \cdot P_{49} \cdot P_{99}, & (1990) \\
F_{10} &= P_8 \cdot P_{10} \cdot P_{40} \cdot P_{252}, & (1995) \\
F_{11} &= P_6 \cdot P_6^* \cdot P_{21} \cdot P_{22} \cdot P_{564}, & (1988) \\
F_{12} &= P_6 \cdot P_8 \cdot P_8^* \cdot P_{12} \cdot P_{16} \cdot P_{54} \cdot C_{1133}. & (2023)
\end{aligned} \tag{1.6}$$

Připomeňme také, že již v roce 1925 byl zahájen obecněji zaměřený projekt, který se zabývá prvočíselnými rozklady čísel tvaru $b^n \pm 1$, která se nazývají Cunninghamova čísla. Tento název byl zaveden na počest Allana J. C. Cunninghama (1842–1928), který spolu s Herbertem J. Woodallem tento projekt založili. Výsledkem jejich snažení je, že v dnešní době existují již tři tištěné verze tabulek rozkladů čísel $b^n \pm 1$, z nichž nejnovější je z roku 2002. Největší přínos projektů [9, 13, 29] a jim podobným, spočívá ve vytváření nových efektivních metod pro testování prvočíselnosti a pro hledání kompletních prvočíselných rozkladů přirozených čísel. Tyto metody mají v dnešní době velký strategický význam.

Konečně je vhodné připomenout, že o problematice Fermatových čísel byly napsány stovky publikací. Čtenáře lze odkázat například na patnáctou kapitolu prvního svazku známé Dicksonovy *Historie teorie čísel* [8], ve které jsou shrnuty nejdůležitější objevy o Fermatových číslech do roku 1919. Dále pak na knihy [7, 20], nebo na články našich autorů Eduarda Fuchse [12, str. 70–74], Michala Křížka [18, 19], Karla Lepky [22, str. 148] a Štefana Porubského [28].

2. JAKÓBCZYKOVA HYPOTÉZA

V roce 1951 publikoval polský kněz a matematik Franciszek Jakóbczyk (1905–1992) dva zajímavé a obtížné problémy týkající Fermatových a Mersennových čísel. Původní formulace hypotéz byla uveřejněna ve francouzštině [14, str. 127]:

Hypothèse. *Les nombres de Fermat et de Mersenne sont des nombres impairs du type $n = p_1 \cdot p_2 \cdots p_k$ où $k \geq 1$.*

O Jakóbczykově hypotéze, která se týká Mersennových čísel se zmiňuje známý polský matematik Waclaw Sierpiński (1882–1969) v knize *Co wiemy a czego nie wiemy o liczbach pierwszych* [33, str. 70], která byla publikována ve Varšavě v roce 1961. Tuto hypotézu lze rovněž nalézt v Sierpińského knize *A Selection of Problems in the Theory of Numbers* [34, str. 92], která vyšla v New Yorku v roce 1964. Ve skutečnosti druhá část knihy [34, str. 25–97] obsahuje anglický překlad polského originálu [33]. Je zajímavé, že Sierpiński do svých knih nezařadil také formulaci Jakóbczykovy hypotézy o Fermatových číslech. Tato skutečnost je možná jednou z příčin, proč dodnes není Jakóbczykovo jméno v souvislosti s jeho hypotézou o Fermatových číslech uváděno. Podrobnosti o životě a díle Franciszka Jakóbczyka může čtenář nalézt v článku [27].

Připomeňme, že Jakóbczykově hypotéze týkající se případu Mersennových čísel je věnována samostatná přehledová studie [15]. V souvislosti s Mersennovými čísly je také vhodné upozornit na dvě nedávné autorovy publikace [16] a [17]. Článek [16] je věnován alternativnímu důkazu hlavní Skulovy věty prezentované v [35]. Viz též [15, str. 55]. V [17] je pak Skulova věta zobecněna pro případ Cunninghamových čísel $b^n \pm 1$.

Zaměříme nyní pozornost na Jakóbczykovu hypotézu o Fermatových číslech.

Hypotéza 2.1 (Jakóbczyk, 1951). *Každé Fermatovo číslo $F_n = 2^{2^n} + 1$ je buď prvočíslo nebo je součinem různých prvočísel.*

Je zřejmé, že Hypotézu 2.1 lze formulovat následujícím ekvivalentním způsobem:

Žádné Fermatovo číslo není dělitelné druhou mocninou prvočísla.

I když se Hypotéza 2.1 objevuje ve známých knihách [7, str. 192], [20, str. 160] a [32, str. 88], z nichž [7] a [20] jsou speciálně věnovány Fermatovým číslům, neobsahují tyto publikace žádné informace o jejím autoru, ani o době jejího vzniku.

Pro formulaci hlavních výsledků dosažených o Hypotéze 2.1 bude vhodné připomenout některé základní definice. Necht $a, m \in \mathbb{N}$, $m \geq 2$ a necht $\gcd(a, m) = 1$. Pak číslo

$$q_m(a) = \frac{a^{\varphi(m)} - 1}{m} \quad (2.1)$$

se nazývá Eulerův kvocient čísla m se základem a . Číslo $\varphi(m)$ vyskytující se v definičním vztahu (2.1) je rovno počtu všech $k \in \mathbb{N}$, kde $1 \leq k \leq m$, která jsou nesoudělná s m . Funkce φ se nazývá Eulerova funkce. Podle Eulerovy věty (viz například [7, str. 14], [20, str. 20]) platí, že $a^{\varphi(m)} \equiv 1 \pmod{m}$. Odtud plyne, že $q_m(a)$ je celé číslo. Je-li m prvočíslo, pak $\varphi(m) = m - 1$ a kvocient (2.1) se nazývá Fermatův kvocient. Podrobnější informace může čtenář nalézt v [15, str. 51–52]. Dále, nejmenší číslo $k \in \mathbb{N}$ takové, že $a^k \equiv 1 \pmod{m}$ se nazývá multiplikativní řád čísla a modulo m a označuje se $\text{ord}_m(a)$. Základní vlastnosti multiplikativního řádu lze nalézt například v [7, str. 30] nebo v [17, str. 3–4].

V roce 1909 dokázal německý matematik Arthur Wieferich (1884–1954) pozoruhodné tvrzení, týkající se prvního případu velké Fermatovy věty [38].

Věta 2.2 (Wieferich, 1909). *Nechť p je liché prvočíslo a nechť x, y, z jsou celá čísla nedělitelná p , vyhovující rovnici $x^p + y^p = z^p$. Pak $2^{p-1} \equiv 1 \pmod{p^2}$.*

Na počest Wieferichova objevu jsou prvočísla p splňující kongruenci $2^{p-1} \equiv 1 \pmod{p^2}$ nazývána Wieferichova prvočísla. Na základě výše uvedených definic není obtížné dokázat, že podmínky (i)–(iii) jsou ekvivalentní:

$$\begin{aligned} \text{(i)} \quad & 2^{p-1} \equiv 1 \pmod{p^2}, \\ \text{(ii)} \quad & q_p(2) \equiv 0 \pmod{p}, \\ \text{(iii)} \quad & \text{ord}_{p^2}(2) = \text{ord}_p(2). \end{aligned} \tag{2.2}$$

Je jistě zajímavé, že v roce 1909, kdy Arthur Wieferich svoje slavné tvrzení dokázal, nebylo žádné Wieferichovo prvočíslo ještě známo. Do dnešní doby byla objevena pouze dvě prvočísla splňující (2.2). První, $w_1 = 1093$, objevil v roce 1913 Waldemar Meissner [25] a druhé, $w_2 = 3511$, objevil v roce 1922 Nicolas Beeger [5]. Podrobná historická fakta týkající se Wieferichových prvočísel jsou uvedena v článku [15, str. 52–54]. Připomeňme, že nedávný projekt *PrimeGrid* [30] zaměřený na hledání Wieferichových prvočísel prokázal, že třetí Wieferichovo prvočíslo, pokud existuje, musí být větší než

$$2^{64} = 18446744073709551616 \doteq 1,8 \cdot 10^{19}. \tag{2.3}$$

Projekt *PrimeGrid* byl ukončen v prosinci 2022. Tedy ani po sto letech od Beegerova objevu nebylo přes velkou snahu mnoha matematiků další Wieferichovo prvočíslo nalezeno. Sám Beeger v roce 1939 vyslovil domněnku, že žádné další Wieferichovo prvočíslo již neexistuje.

První důležité tvrzení, které propojilo Jakóbczykovu hypotézu o Fermatových číslech s Wieferichovými prvočíslly dokázali v roce 1967 Le Roy J. Warren a Henry G. Bray [37].

Věta 2.3 (Warren, Bray, 1967). *Nechť $n \in \mathbb{N}$, $n > 1$ a nechť p je libovolné liché prvočíslo. Jestliže $p \mid F_n$, pak*

$$2^{\frac{p-1}{2}} \equiv 1 \pmod{F_n}.$$

Z Věty 2.3 snadno plyne následující důsledek.

Důsledek 2.4. *Nechť $n \in \mathbb{N}$, $n > 1$ a nechť p je libovolné liché prvočíslo. Jestliže $p^2 \mid F_n$, pak p je Wieferichovo prvočíslo, tj. $2^{p-1} \equiv 1 \pmod{p^2}$.*

Protože v článku [37] není důkaz Důsledku 2.4 uveden, uvedeme důkaz nyní.

Důkaz. Předpokládejme, že $p^2 \mid F_n$. Pak $p \mid F_n$ a podle Věty 2.3 platí $2^{(p-1)/2} \equiv 1 \pmod{F_n}$. To však znamená, že $F_n \mid (2^{(p-1)/2} - 1)(2^{(p-1)/2} + 1) = 2^{p-1} - 1$. Protože relace dělitelnosti $\mid$ je tranzitivní, konjunkce podmínek $p^2 \mid F_n$ a $F_n \mid 2^{p-1} - 1$ implikuje $p^2 \mid 2^{p-1} - 1$. Tato relace ale podle části (i) vztahu (2.2) znamená, že p je Wieferichovo prvočíslo. $\square$

V roce 1979 Paulo Ribenboim [31, str. 86–88] publikoval následující dvě tvrzení.

Věta 2.5 (Ribenoim, 1979). *Nechť $n \in \mathbb{N}$, $n \geq 2$ a nechť p je libovolné prvočíslo. Jestliže $p^2 \mid F_n$, pak $p = k \cdot 2^r + 1$, kde $k, r \in \mathbb{N}$, $r \geq n + 2$, k je liché a*

$$q_p(k) = \frac{k^{p-1} - 1}{p} \equiv 1 \pmod{p}.$$

Věta 2.6 (Ribenoim, 1979). *Nechť $a, n \in \mathbb{N}$, $a \geq 2$ a nechť p je libovolné liché prvočíslo. Jestliže $p \mid F_{a,n} = a^{a^n} + 1$ a $q_p(a) \equiv 0 \pmod{p}$, pak $p^2 \mid F_{a,n}$.*

Speciálně, je-li $a = 2$, pak Věta 2.6, společně s částí (ii) vztahu (2.2), zaručuje platnost následující implikace:

$$\text{Jestliže } p \mid F_n \text{ a } 2^{p-1} \equiv 1 \pmod{p^2}, \text{ pak } p^2 \mid F_n. \quad (2.4)$$

Důsledek 2.4 spolu s implikací (2.4) pak dokazuje platnost Věty 2.7.

Věta 2.7. *Nechť $n \in \mathbb{N} \cup \{0\}$, $a \geq 2$ a nechť p je libovolné liché prvočíslo. Jestliže $p \mid F_n$, pak $p^2 \mid F_n$ právě tehdy, když $2^{p-1} \equiv 1 \pmod{p^2}$.*

Kompletní důkaz Věty 2.7 lze nalézt v [20, str. 68]. Z důvodu historické korektnosti poznamenejme, že autorství implikace (2.4) není zcela zřejmé. Michal Křížek v článku [18, str. 252] nepřímou uvádí jako autora Věty 2.7 Wilfrida Kellera a odkazuje zde čtenáře na nepublikovaný Kellerův rukopis. Také odkazy uvedené v [20, str. 68] původ implikace (2.4) neobjasňují. Kromě toho, Paulo Ribenoim v [31, str. 87] uvádí, že jeho Věta 2.6 je přeformulováním výsledků uvedených v článku [10], jehož autorkou je Jeanne Ferentinou–Nicolacopoulou.

V souvislosti s Větou 2.7 je rovněž vhodné zmínit, že v roce 2023 autor tohoto článku publikoval tvrzení (viz [17, str. 19]), které zjednodušuje logickou strukturu Věty 2.7 následujícím způsobem:

Věta 2.8 (Klaška, 2023). *Nechť $n \in \mathbb{N} \cup \{0\}$ a nechť p je libovolné liché prvočíslo. Pak $p^2 \mid F_n$ právě tehdy, když p je Wieferichovo prvočíslo a $\text{ord}_p(2) = 2^{n+1}$.*

Není obtížné ověřit, že pro jediná dvě známá Wieferichova prvočísla platí

$$\begin{aligned} \text{ord}_{1093}(2) &= \text{ord}_{1093^2}(2) = 364 = 2^2 \cdot 7 \cdot 13, \\ \text{ord}_{3511}(2) &= \text{ord}_{3511^2}(2) = 1755 = 3^3 \cdot 5 \cdot 13. \end{aligned}$$

Odtud a z (2.3) nyní plyne, že žádné Fermatovo číslo není dělitelné druhou mocninou žádného prvočísla p , kde $p < 1,8 \cdot 10^{19}$. Zkoumání četnosti výskytu prvočísel p splňujících podmínku $\text{ord}_p(2) = 2^k$ pro nějaké $k \in \mathbb{N}$ vede rovněž k zajímavým závěrům. Například je známo, že pro $p < 4,18 \cdot 10^{18}$ existuje pouze padesát prvočísel splňujících podmínku $\text{ord}_p(2) = 2^k$ pro nějaké $k \in \mathbb{N}$. Seznam těchto prvočísel je následující:

3, 5, 17, 257, 641, 65537, 114689, 274177, 319489, 974849, 2424833, 6700417, 13631489, 26017793, 45592577, 63766529, 167772161, 825753601, 1214251009, 6487031809, 70525124609, 190274191361, 646730219521, 2710954639361, 2748779069441, 4485296422913, 6597069766657, 25409026523137, 25991531462657, 31065037602817, 46179488366593, 67280421310721,

76861124116481, 151413703311361, 640126220763137, 1095981164658689,
 1238926361552897, 1256132134125569, 2327042503868417, 2405286912458753,
 2917004348489729, 59649589127497217, 204393464266227713,
 231292694251438081, 1033434552359452673, 1529992420282859521,
 2170072644496392193, 2663848877152141313, 3603109844542291969,
 4179340454199820289.

Dále je možné dokázat [20, str. 37], že prvočísla uvedená v tomto seznamu jsou právě ta prvočísla, která dělí některé z Fermatových čísel. Přesněji formulováno, platí Věta 2.9.

Věta 2.9. *Nechť $n \in \mathbb{N} \cup \{0\}$ a necht p je libovolné liché prvočíslo. Pak $p \mid F_n$ právě tehdy, když $\text{ord}_p(2) = 2^{n+1}$.*

Více informací o posloupnosti prvočíselných dělitelů Fermatových čísel lze nalézt na webové stránce *The On-Line Encyclopedia of Integer Sequences* [36, A023394] a také v knize [20, str. 37].

Článek zakončíme formulací Problému 2.10, který byl inspirován přehledem známých rozkladů Fermatových čísel uvedených v (1.6), viz [20, str. 159] a [18, str. 249].

Problém 2.10 (Křížek, 1995). *Dokažte nebo vyvratte následující tvrzení. Necht $\delta(F_n)$ označuje počet všech prvočíselných dělitelů Fermatova čísla F_n . Pak pro každé $n \in \mathbb{N} \cup \{0\}$ platí $\delta(F_n) \leq \delta(F_{n+1})$.*

Z přehledu (1.6) okamžitě vidíme, že

$$(\delta(F_n))_{n=0}^{\infty} = (1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 5, \dots).$$

Je zřejmé, že pokud je posloupnost $(\delta(F_n))_{n=0}^{\infty}$ neklesající, pak seznam (1.2) obsahuje všechna Fermatova prvočísla. Speciální část Problému 2.10, která se týká výskytu prvočísel v posloupnosti $(F_n)_{n=0}^{\infty}$, lze formulovat jako samostatnou hypotézu, viz [20, str. 163].

Hypotéza 2.11 (Křížek, 1995). *$\delta(F_n) = 1$ právě tehdy, když $n \leq 4$.*

Na základě našich současných vědomostí můžeme předpokládat, že řešení problémů formulovaných Michalem Křížkem a nalezení důkazu Jakóbczykovy hypotézy bude velmi obtížné. Nelze ani vyloučit, že řešení těchto jednoduše formulovaných problémů leží zcela mimo hranice lidských možností.

REFERENCE

- [1] K. R. Biermann: *Thomas Clausen, Mathematiker und Astronom*, Journal für die Reine und Angewandte Mathematik **216** (1964), 159–198.
- [2] R. P. Brent: *Factorization of the eleventh Fermat number*, Abstracts Amer. Math. Soc. **10** (1989), 176–177.
- [3] R. P. Brent, J. M. Pollard: *Factorization of the eighth Fermat number*, Mathematics of Computation **36** (1981), 627–630.

- [4] R. P. Brent: *Factorization of the tenth Fermat number*, Mathematics of Computation **68** (1999), 429–451.
- [5] N. Beeger: *On a new case of the congruence $2^{p-1} \equiv 1 \pmod{p^2}$* , Messenger of Mathematics **51** (1922), 149–150.
- [6] A. J. Cunningham, A. E. Western: *On Fermats numbers*, Proc. London Math. Soc. **(2)1** (1904), 175.
- [7] E. Deza: *Mersenne Numbers and Fermat Numbers, Selected Chapters of Number Theory: Special Numbers*, World Scientific, 2021.
- [8] L. E. Dickson: *History of the Theory of Numbers*, Vol. I, Dover Publications, 2005.
- [9] Distributed Search for Fermat Number Divisors, online <http://www.fermatsearch.org>.
- [10] J. Ferentinou–Nicolacopoulou: *Une propriété des diviseurs du nombre $r^{r^m} + 1$. Applications au dernier théorème de Fermat*, Bulletin of the Greek Mathematical Society **4** (1963), 121–126.
- [11] C. R. Fletcher: *A reconstruction Frenicle–Fermat correspondence of 1640*, Historia Mathematica **18.4** (1991), 344–351.
- [12] E. Fuchs: *Co ještě nevíme o přirozených číslech (2) aneb Od dokonalých čísel k Fermatovým prvočísłům*, Učitel matematiky **7.2** (1999), 65–74.
- [13] Great Internet Mersenne Prime Search, online <https://www.mersenne.org>.
- [14] F. Jakóbczyk: *Les applications de la fonction $\lambda_g(n)$ à l'étude des fractions périodiques et de la congruence chinoise $2^n - 2 \equiv 0 \pmod{n}$* , Annales Universitatis Mariae Curie-Skłodowska **5** (1951), 97–138.
- [15] J. Klaška: *Hypotéza Franciszka Jakóbczyka o Mersennových číslech*, Kvaternion (2021), 47–57.
- [16] J. Klaška: *A simple proof of Skula's theorem on prime power divisors of Mersenne numbers*, Journal of Integer Sequences **25** (2022), Article 22.4.3.
- [17] J. Klaška: *Jakóbczyk's hypothesis on Mersenne numbers and generalizations of Skula's theorem*, Journal of Integer Sequences **26** (2023), Article 23.3.8.
- [18] M. Křížek: *O Fermatových číslech*, Pokroky matematiky, fyziky a astronomie **40.5** (1995), 243–253.
- [19] M. Křížek: *Od Fermatových prvočísel ke geometrii*, Cahiers du CEFRES, (2002), 131–161.
- [20] M. Křížek, F. Luca, L. Somer: *17 Lectures on Fermat Numbers, From Number Theory to Geometry*, Canadian Mathematical Society, Springer, 2001.
- [21] F. Landry: *Note sur la décomposition du nombre $2^{64} + 1$* , Comptes rendus de l'Académie des Sciences A **91** (1880), 138.
- [22] K. Lepka: *Malá Fermatova věta*, Učitel matematiky **5.3** (1997), 143–150.
- [23] A. K. Lenstra, H. W. Lenstra, M. S. Manasse, J. M. Pollard: *The factorization of the ninth Fermat number*, Mathematics of Computation **61** (1993), 319–349.
- [24] E. Lucas: *Théorèmes d'arithmétique*, Atti della Reale Accademia delle Scienze di Torino **13** (1878), 271–284.
- [25] W. Meissner: *Über die Teilbarkeit von $2^p - 2$ durch das Quadrat der Primzahl $p = 1093$* , Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, Berlin, 1913, 663–667.
- [26] M. A. Morrison, J. Brillhart: *The factorization of F_7* , Bull. Amer. Math. Soc. **77** (1971), 264.
- [27] H. Piersa: *Śp. Ksiądz dr Franciszek Jakóbczyk (9 X 1905 – 3 VI 1992)*, Roczniki Filozoficzne **39–40**, spis 3, (1991–1992), 5–7.
- [28] Š. Porubský: *Fermat a teorie čísel aneb problematika dělitelů a dokonalá čísla*, Cahiers du CEFRES, 2002, 49–86.
- [29] Prime factors $k \cdot 2^n + 1$ of Fermat numbers F_m and complete factoring status (compiled by Wilfrid Keller), online <http://www.prothsearch.com/fermat.html>.
- [30] PrimeGrid, online <https://www.primegrid.com>.
- [31] P. Ribenboim: *On the square factors of the numbers of Fermat and Ferentinou–Nicolacopoulou*, Bulletin of the Greek Mathematical Society **20** (1979), 81–92.

- [32] P. Ribenboim: *The new Book of Prime Number Records*, Springer-Verlag, New York, 1996.
- [33] W. Sierpiński: *Co wiemy, a czego nie wiemy o liczbach pierwszych*, Warszawa, 1961.
- [34] W. Sierpiński: *A Selection of Problems in the Theory of Numbers*, New York, 1964.
- [35] L. Skula: *Prime power divisors of Mersenne numbers and Wieferich primes of higher order*, Integers, Electronic Journal of Combinatorial Number Theory **19** (2019).
- [36] The On-Line Encyclopedia of Integer Sequences (OEIS), online <https://oeis.org>.
- [37] L. R. Warren, H. G. Bray: *On the square-freeness of Fermat and Mersenne numbers*, Pacific Journal of Mathematics **22** (1967), 563–564.
- [38] A. Wieferich: *Zum letzten Fermat'schen Theorem*, Journal für die Reine und Angewandte Mathematik **136** (1909), 293–302.
- [39] H. C. Williams: *How was F_6 factored?*, Mathematics of Computation **61** (1993), 463–474.

Jiří Klaška, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně,
Technická 2, 61669 Brno, Česká republika,
e-mail: klaska@fme.vutbr.cz

SOUČTY ŘAD PŘEVŘÁCENÝCH HODNOT VŠECH SOUČINŮ GENEROVANÝCH DVĚMA A VÍCE PŘIROZENÝMI ČÍSLY

RADOVAN POTŮČEK

ABSTRAKT. V tomto článku je prezentován jeden typ nekonečných číselných řad, které jsou generovány převrácenými hodnotami ke všem součinům dvou a více přirozených čísel. Tyto řady tak představují příklad redukované harmonické řady, tedy harmonické řady s některými vynechanými členy. Jsou uvedeny dva způsoby, jak tyto řady sečíst. Jedním způsobem je zápis těchto řad ve tvaru součtu dílčích řad a určení příslušných dílčích součtů postupným užitím vzorce pro součet nekonečné geometrické řady. Druhým, výrazně jednodušším, způsobem je zápis těchto řad ve tvaru součinu a užití vzorce pro součet jednotlivých nekonečných geometrických řad. Získané výsledky jsou zobecněny pro n generujících přirozených čísel a numericky ověřeny pro dvě a tři jednociferná čísla užitím systému počítačové algebry Maple.

1. ÚVOD

Článek je inspirován úlohou 5.4 z roku 2002 na určení součtu speciální nekonečné řady obsaženou ve studijním materiálu [5]. Úlohy z různých oblastí matematiky jsou každý týden ve formě studijních materiálů předkládány studentům a dalším zájemcům o hlubší studium matematiky sdružených do skupiny *the Berkeley Math Circle* při Kalifornské univerzitě v Berkeley. Touto úlohou bylo určit součet všech zlomků, které mají v čitateli jedničku a ve jmenovateli součiny generované všemi prvočíselnými děliteli čísla 2002, tedy prvočísky 2, 7, 11 a 13. Jednalo se tak o úlohu určit součet nekonečné číselné řady

$$\begin{aligned} \frac{1}{2} + \frac{1}{2^2} + \frac{1}{7} + \frac{1}{2^3} + \frac{1}{11} + \frac{1}{13} + \frac{1}{2 \cdot 7} + \frac{1}{2^4} + \frac{1}{2 \cdot 11} + \frac{1}{2 \cdot 13} + \frac{1}{2 \cdot 2 \cdot 7} + \cdots \\ = \frac{1}{2} + \frac{1}{4} + \frac{1}{7} + \frac{1}{8} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{1}{16} + \frac{1}{22} + \frac{1}{26} + \frac{1}{28} + \cdots \end{aligned}$$

V tomto článku se budeme zabývat jistým zobecněním této úlohy spočívajícím v tom, že budeme určovat součet všech zlomků, které mají v čitateli jedničku a ve jmenovateli součiny generované všemi 2-prvkovými i víceprvkovými podmnožinami jednociferných čísel s výjimkou čísla 1, tedy přirozenými čísly 2, 3, 4, 5, 6, 7, 8, 9.

2020 MSC. Primární 40A02.

Klíčová slova. Nekonečná řada, geometrická řada, harmonická řada, redukovaná harmonická řada, systém počítačové algebry Maple.

Práce prezentovaná v tomto článku je podporována Projektem pro rozvoj organizace „DZRO Vojenské autonomní a robotické systémy“.

Číslo 1 neuvažujeme proto, že budeme hledat součty konvergentních řad a součet zlomků majících ve jmenovateli mocniny čísla 1, tj. dílčí součet

$$\frac{1}{1} + \frac{1}{1^2} + \frac{1}{1^3} + \frac{1}{1^4} + \frac{1}{1^5} + \dots$$

obsažený v jakékoliv řadě generované i dalšími jednocifernými čísly, tvoří divergentní řadu se součtem ∞ .

Protože výše zmíněná úloha i dále řešené úlohy se týkají redukovaných harmonických řad, uvedeme nejprve stručný přehled několika základních a dalších potřebných pojmů z teorie nekonečných číselných řad.

2. ZÁKLADNÍ POJMY

Definice 2.1. Pro libovolnou číselnou posloupnost $\{a_k\}$ je odpovídající **nekonečná řada**, dále stručněji **řada**, definována jako součet

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \dots$$

Definice 2.2. Posloupnost **částečných součtů** $\{s_n\}$ řady $\sum_{k=1}^{\infty} a_k$ je definována pro každé $n \in \mathbb{N}$ jako posloupnost konečných součtů tvaru

$$s_n = \sum_{k=1}^n a_k = a_1 + a_2 + \dots + a_n.$$

Definice 2.3. Jestliže posloupnost částečných součtů $\{s_n\}$ má limitu s , tj. $\lim_{n \rightarrow \infty} s_n = s$, potom je řada $\sum_{k=1}^{\infty} a_k$ **konvergentní**. Pokud tato limita s neexistuje nebo je nevlastní, tedy $s = \pm\infty$, je řada $\sum_{k=1}^{\infty} a_k$ **divergentní**. V případě konvergentní řady říkáme, že řada $\sum_{k=1}^{\infty} a_k$ má **součet** s a píšeme $\sum_{k=1}^{\infty} a_k = s$.

Definice 2.4. Jestliže konverguje řada $\sum_{k=1}^{\infty} a_k$ i řada $\sum_{k=1}^{\infty} |a_k|$, pak říkáme, že řada $\sum_{k=1}^{\infty} a_k$ je **absolutně konvergentní**.

Poznamenejme, že pro konvergentní řady platí asociativní a distributivní zákon. Připomeňme, že přerovnání členů absolutně konvergentní řady nemá vliv ani na její konvergenci, ani na její součet. **Součinem** dvou absolutně konvergentních řad $\sum_{k=1}^{\infty} a_k$ a $\sum_{k=1}^{\infty} b_k$ se součty s_a a s_b je opět absolutně konvergentní řada, která má součet $s_a s_b$. Součin dvou absolutně konvergentních řad je přitom dán vztahem

$$\left(\sum_{k=1}^{\infty} a_k \right) \left(\sum_{k=1}^{\infty} b_k \right) = \sum_{k=2}^{\infty} \sum_{\ell=1}^{k-1} a_{\ell} b_{k-\ell} =$$

$$= a_1b_1 + a_1b_2 + a_2b_1 + a_1b_3 + \\ + a_2b_2 + a_3b_1 + a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1 + \dots$$

Definice 2.5. Geometrická řada je definována jako součet nekonečně mnoha členů geometrické posloupnosti, tedy posloupnosti s prvním členem a , v níž je podíl q - dvou po sobě jdoucích členů konstantní. Geometrická řada je tak řada tvaru

$$a + aq + aq^2 + \dots + aq^{n-1} + \dots = \sum_{k=0}^{\infty} aq^k.$$

Připomeňme, že pro součet s konvergentní geometrické řady s prvním členem a a kvocientem q , kde $|q| < 1$, platí známý vzorec

$$s = \frac{a}{1-q}. \quad (2.1)$$

Příklad 2.6. Součtem řady převrácených hodnot jistých přirozených čísel rozumíme součet příslušných zlomků s jedničkou v čitateli. Příkladem takového součtu je:

a) Součet řady převrácených hodnot druhých mocnin přirozených čísel (tato úloha pocházející z roku 1650 je nazývána **Basilejský problém**)

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \dots = \frac{\pi^2}{6} \doteq 1,644934.$$

b) Součet řady převrácených hodnot třetích mocnin přirozených čísel, jehož výsledkem je tzv. **Apéryho konstanta**

$$\sum_{k=1}^{\infty} \frac{1}{k^3} = \frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{4^3} + \dots \doteq 1,202057.$$

Obecně představuje součet $\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$ takzvanou **Riemannovu funkci zeta** komplexní proměnné s . Je známo, že tato řada konverguje, pokud je reálná část čísla s větší než jedna. Tedy $\zeta(2) = \pi^2/6$ a $\zeta(3)$ představuje Apéryho konstantu.

Definice 2.7. Harmonická řada je definována jako řada převrácených hodnot ke všem přirozeným číslům, takže je tvaru

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$$

Název řady, o níž lze např. integrálním kritériem konvergence snadno dokázat, že diverguje, je dán tím, že každý člen této řady kromě prvního je harmonickým průměrem jeho dvou sousedních členů. Takže např.

$$\frac{2}{1/(1/3) + 1/(1/5)} = \frac{2}{3+5} = \frac{1}{4}.$$

Definice 2.8. Redukovanou harmonickou řadu definujeme jako neúplnou harmonickou řadu, z níž byly vynechány některé její členy.

Příklad 2.9. K důležitým příkladům redukováných harmonických řad patří:

a) Řada převrácených hodnot k prvočísłům a k číslu 1

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \frac{1}{13} + \dots,$$

jejíž divergenci dokázal již Leonhard Euler (viz např. knihu [2]).

b) Řada převrácených hodnot k faktoriálům nezáporných celých čísel

$$\sum_{k=0}^{\infty} \frac{1}{k!} = \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \dots,$$

která konverguje a jejíž součet je Eulerovo číslo $e = 2,718281828\dots$

c) **Kempnerova řada** K_c , která vznikne z harmonické řady vynecháním všech členů $1/n$, kde n obsahuje v desítkové soustavě číslici c . Tedy např. vynecháním všech členů obsahujících ve jmenovateli číslici 9 obdržíme řadu

$$K_9 = \frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{8} + \frac{1}{10} + \frac{1}{11} + \dots + \frac{1}{18} + \frac{1}{20} + \frac{1}{21} + \dots + \frac{1}{88} + \frac{1}{100} + \frac{1}{101} + \dots,$$

která byla poprvé studována v roce 1914 v článku [4]. Touto řadou se zabýval britský matematik Aubrey John Kempner (1880–1972), který v článku dokázal, že řada K_9 překvapivě konverguje.

Lze dokázat, že všech dalších devět Kempnerových řad $K_0, K_1, \dots, K_8$ konverguje. Kempnerovým řadám je mj. věnován článek [1] nebo část 3. kapitoly v knize [3].

Následující tabulka udává přibližné hodnoty součtů $S_0, S_1, \dots, S_9$ těchto řad zaokrouhlené na tři desetinná místa.

c	0	1	2	3	4	5	6	7	8	9
S_c	23,103	16,177	19,257	20,570	21,327	21,835	22,206	22,493	22,726	22,921

Tabulka 1. Přibližné hodnoty součtů Kempnerových řad.

Poznamenejme, že v článku [6] je uveden algoritmus, který zobecňuje problém chybějících číslic na chybějící řetězec číslic. Např. součet zobecněné Kempnerovy řady s vynechaným číselným řetězcem 42 je v tomto článku vyčíslen přibližně hodnotou 228,446.

3. DVA ZPŮSOBY VÝPOČTU SOUČTŮ ŘAD PŘEVŘÁCENÝCH HODNOT VŠECH SOUČINŮ GENEROVANÝCH DVĚMA A TŘEMI PŘIROZENÝMI ČÍSLY

Oba dále uvedené způsoby výpočtu součtu těchto řad nejprve ukážeme na dvou konkrétních příkladech součtu řad generovaných dvěma a třemi přirozenými čísly. Druhý, a výrazně kratší, způsob následně zobecníme na určení součtů řad generovaných libovolným počtem přirozených čísel.

Příklad 3.1. Určete součet řady převrácených hodnot všech součinů generovaných čísly 2 a 3, tj. součet řady

$$\begin{aligned} & \frac{1}{2} + \frac{1}{3} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{2 \cdot 3} + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{2^2 \cdot 3} + \frac{1}{3^2 \cdot 2} \\ & \quad + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{2^3 \cdot 3} + \frac{1}{3^3 \cdot 2} + \frac{1}{2^2 \cdot 3^2} + \dots \\ & = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{9} + \frac{1}{6} + \frac{1}{8} + \frac{1}{27} + \frac{1}{12} \\ & \quad + \frac{1}{18} + \frac{1}{16} + \frac{1}{81} + \frac{1}{24} + \frac{1}{54} + \frac{1}{36} + \dots \end{aligned}$$

Řešení:

a) Srovnávacím kritériem lze ukázat, že daná řada konverguje, což zaručuje platnost asociativního zákona. Uvažujme majorantní řadu, která má oproti dané řadě ve jmenovatelích zlomků pouze základy 2:

$$\frac{2}{2^1} + \frac{3}{2^2} + \frac{4}{2^3} + \frac{5}{2^4} + \dots + \frac{k+1}{2^k} + \dots$$

Konvergenci této majorantní řady lze ověřit limitním podílovým kritériem, kdy

$$\lim_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = \lim_{k \rightarrow \infty} \frac{k+2}{2^{k+1}} \frac{2^k}{k+1} = \frac{1}{2} \lim_{k \rightarrow \infty} \frac{k+2}{k+1} = \frac{1}{2} < 1.$$

Daná řada, kterou budeme dále značit T , tak má konečný součet S . Protože všechny její členy jsou kladné, konverguje řada T absolutně, takže její členy lze přerovnat. Členy řady T nyní uzavřeme podle součtu exponentů. Dostaneme tak řadu tvaru

$$\begin{aligned} T &= \frac{1}{2} + \frac{1}{3} + \left(\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{2 \cdot 3} \right) + \left(\frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{2^2 \cdot 3} + \frac{1}{3^2 \cdot 2} \right) \\ & \quad + \left(\frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{2^3 \cdot 3} + \frac{1}{3^3 \cdot 2} + \frac{1}{2^2 \cdot 3^2} \right) + \dots \end{aligned} \quad (3.1)$$

Pro určení součtu S řady T vhodně přerovnáme, a to na tvar

$$\begin{aligned} T &= \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \dots \right) + \left(\frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^4} + \dots \right) \\ & \quad + \left(\frac{1}{2 \cdot 3} + \frac{1}{2^2 \cdot 3} + \frac{1}{3^2 \cdot 2} + \frac{1}{2^3 \cdot 3} + \frac{1}{3^3 \cdot 2} + \frac{1}{2^2 \cdot 3^2} + \dots \right). \end{aligned}$$

Takto přerovnanou řadu T nyní rozdělíme na tři dílčí řady T_2 , T_3 a $T_{2 \cdot 3}$, kde

$$T_2 = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \dots = \frac{1}{2} \left(1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \dots \right), \quad (3.2)$$

$$T_3 = \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^4} + \dots = \frac{1}{3} \left(1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \dots \right), \quad (3.3)$$

$$T_{2 \cdot 3} = \frac{1}{2 \cdot 3} + \frac{1}{2^2 \cdot 3} + \frac{1}{3^2 \cdot 2} + \frac{1}{2^3 \cdot 3} + \frac{1}{3^3 \cdot 2} + \frac{1}{2^2 \cdot 3^2} + \dots,$$

tj.

$$T_{2,3} = \frac{1}{2 \cdot 3} \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{2 \cdot 3} + \cdots \right). \quad (3.4)$$

Součet S určíme pomocí součtů dílčích řad (3.2)–(3.4). Užitím vzorce (2.1) pro součet konvergentní geometrické řady získáme součty S_2 a S_3 řad (3.2) a (3.3):

$$S_2 = \frac{1}{2} \frac{1}{1 - 1/2} = \frac{1}{2} \frac{2}{2 - 1} = 1, \quad (3.5)$$

$$S_3 = \frac{1}{3} \frac{1}{1 - 1/3} = \frac{1}{3} \frac{3}{3 - 1} = \frac{1}{2}. \quad (3.6)$$

Je zřejmé, že součet $S_{2,3}$ dílčí řady (3.4) můžeme zapsat ve tvaru

$$S_{2,3} = \frac{1}{2 \cdot 3} (1 + S) = \frac{1}{6} (1 + S). \quad (3.7)$$

Na základě platnosti absolutní konvergence řady (3.1)) můžeme její součet S zapsat ve tvaru

$$S = S_2 + S_3 + S_{2,3}.$$

Užitím vztahů (3.5)–(3.7) obdržíme rovnici

$$S = 1 + \frac{1}{2} + \frac{1}{6} (1 + S).$$

Vynásobením rovnice číslem 6 dostaneme rovnici

$$6S = 6 + 3 + 1 + S,$$

tj. $5S = 10$, odkud

$$S = 2.$$

b) Uvažujme dvě geometrické řady U a V s prvním členem 1 a s kvocienty $2^{-1} = \frac{1}{2}$ a $3^{-1} = \frac{1}{3}$, tedy geometrické řady tvaru

$$U = 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \cdots,$$

$$V = 1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^4} + \cdots$$

se součty S_U a S_V , které podle vzorce (2.1) jsou

$$S_U = \frac{1}{1 - 1/2} = \frac{2}{2 - 1} = 2,$$

$$S_V = \frac{1}{1 - 1/3} = \frac{3}{3 - 1} = \frac{3}{2}.$$

Protože jsou řady U a V absolutně konvergentní, je jejich součinem opět absolutně konvergentní řada

$$U \cdot V = \left(1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \cdots \right) \left(1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^4} + \cdots \right).$$

Je zřejmé, že roznásobením členů obou řad obdržíme řadu tvaru

$$U \cdot V = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{2 \cdot 3} + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{2^2 \cdot 3} + \frac{1}{3^2 \cdot 2} + \cdots,$$

tedy řadu $1 + T$ se součtem $1 + S$. Z rovnice

$$1 + S = S_U S_V$$

dostáváme

$$S = S_U S_V - 1,$$

tj. $S = 2 \frac{3}{2} - 1$, odkud

$$S = 2.$$

Oběma způsoby výpočtu jsme tak obdrželi stejný výsledek: řada T má součet $S = 2$.

Příklad 3.2. Určete součet řady převrácených hodnot všech součinů generovaných čísly 4, 5 a 6, tj. součet řady

$$\begin{aligned} \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{4^2} + \frac{1}{4 \cdot 5} + \frac{1}{4 \cdot 6} + \frac{1}{5^2} + \frac{1}{5 \cdot 6} + \frac{1}{6^2} + \frac{1}{4^3} + \frac{1}{4^2 \cdot 5} + \dots \\ = \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{16} + \frac{1}{20} + \frac{1}{24} + \frac{1}{25} + \frac{1}{30} + \frac{1}{36} + \frac{1}{64} + \frac{1}{80} + \dots \end{aligned}$$

Řešení:

a) Srovnávacím kritériem lze ukázat, že daná řada konverguje, což zaručuje platnost asociativního zákona. Uvažujme majorantní řadu, která má oproti dané řadě ve jmenovatelích zlomků pouze základy 4:

$$\frac{3}{4^1} + \frac{6}{4^2} + \frac{10}{4^3} + \frac{15}{4^4} + \dots + \frac{\binom{k+2}{2}}{2 \cdot 4^k} + \dots,$$

tedy řadu tvaru

$$\sum_{k=1}^{\infty} \frac{(k+2)(k+1)}{2 \cdot 4^k}.$$

Konvergenzi této majorantní řady lze opět ověřit limitním podílovým kritériem:

$$\lim_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = \lim_{k \rightarrow \infty} \frac{(k+3)(k+2)}{2 \cdot 4^{k+1}} \frac{2 \cdot 4^k}{(k+2)(k+1)} = \frac{1}{4} \lim_{k \rightarrow \infty} \frac{k+3}{k+1} = \frac{1}{4} < 1.$$

Daná řada, kterou budeme stejně jako v příkladu 3.1 značit T , má tedy konečný součet S . Protože všechny její členy jsou kladné, konverguje řada T absolutně, takže její členy můžeme přerovnat. Členy řady T nyní uzavorkujeme podle součtu exponentů. Dostaneme tak řadu tvaru

$$\begin{aligned} T = \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \left(\frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{4 \cdot 5} + \frac{1}{4 \cdot 6} + \frac{1}{5 \cdot 6} \right) \\ + \left(\frac{1}{4^3} + \frac{1}{5^3} + \frac{1}{6^3} + \frac{1}{4^2 \cdot 5} + \frac{1}{4^2 \cdot 6} + \frac{1}{5^2 \cdot 4} + \frac{1}{5^2 \cdot 6} \right. \\ \left. + \frac{1}{6^2 \cdot 4} + \frac{1}{6^2 \cdot 5} + \frac{1}{4 \cdot 5 \cdot 6} \right) \\ + \left(\frac{1}{4^4} + \frac{1}{5^4} + \frac{1}{6^4} + \frac{1}{4^3 \cdot 5} + \frac{1}{4^3 \cdot 6} + \frac{1}{5^3 \cdot 4} + \frac{1}{5^3 \cdot 6} + \frac{1}{6^3 \cdot 4} + \frac{1}{6^3 \cdot 5} \right) \end{aligned}$$

$$+ \frac{1}{4^2 \cdot 5 \cdot 6} + \frac{1}{5^2 \cdot 4 \cdot 6} + \frac{1}{6^2 \cdot 4 \cdot 5} + \frac{1}{4^2 \cdot 5^2} + \frac{1}{4^2 \cdot 6^2} + \frac{1}{5^2 \cdot 6^2} \Big) + \dots$$

Pro určení součtu S řadu T vhodně přerovnáme, a to na tvar

$$\begin{aligned} T = & \left(\frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \dots \right) + \left(\frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \frac{1}{5^4} + \dots \right) \\ & + \left(\frac{1}{6} + \frac{1}{6^2} + \frac{1}{6^3} + \frac{1}{6^4} + \dots \right) \\ & + \left(\frac{1}{4 \cdot 5} + \frac{1}{4^2 \cdot 5} + \frac{1}{5^2 \cdot 4} + \frac{1}{4 \cdot 5 \cdot 6} + \frac{1}{4^3 \cdot 5} + \frac{1}{5^3 \cdot 4} \right. \\ & \quad \left. + \frac{1}{4^2 \cdot 5 \cdot 6} + \frac{1}{5^2 \cdot 4 \cdot 6} + \frac{1}{6^2 \cdot 4 \cdot 5} + \frac{1}{4^2 \cdot 5^2} + \dots \right) \\ & + \left(\frac{1}{4 \cdot 6} + \frac{1}{4^2 \cdot 6} + \frac{1}{6^2 \cdot 4} + \frac{1}{4^3 \cdot 6} + \frac{1}{6^3 \cdot 4} + \frac{1}{4^2 \cdot 6^2} + \dots \right) \\ & + \left(\frac{1}{5 \cdot 6} + \frac{1}{5^2 \cdot 6} + \frac{1}{6^2 \cdot 5} + \frac{1}{5^3 \cdot 6} + \frac{1}{6^3 \cdot 5} + \frac{1}{5^2 \cdot 6^2} + \dots \right). \end{aligned} \quad (3.8)$$

Takto přerovnanou řadu T nyní rozdělíme na šest dílčích řad T_4 , T_5 , T_6 , $T_{4 \cdot 5}$, $T_{4 \cdot 6}$ a $T_{5 \cdot 6}$, kde

$$T_4 = \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \dots = \frac{1}{4} \left(1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \dots \right), \quad (3.9)$$

$$T_5 = \frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \frac{1}{5^4} + \dots = \frac{1}{5} \left(1 + \frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \dots \right), \quad (3.10)$$

$$T_6 = \frac{1}{6} + \frac{1}{6^2} + \frac{1}{6^3} + \dots = \frac{1}{6} \left(1 + \frac{1}{6} + \frac{1}{6^2} + \frac{1}{6^3} + \dots \right), \quad (3.11)$$

$$\begin{aligned} T_{4 \cdot 5} = & \frac{1}{4 \cdot 5} + \frac{1}{4^2 \cdot 5} + \frac{1}{5^2 \cdot 4} + \frac{1}{4 \cdot 5 \cdot 6} + \frac{1}{4^3 \cdot 5} + \frac{1}{5^3 \cdot 4} \\ & + \frac{1}{4^2 \cdot 5 \cdot 6} + \frac{1}{5^2 \cdot 4 \cdot 6} + \frac{1}{6^2 \cdot 4 \cdot 5} + \frac{1}{4^2 \cdot 5^2} + \dots, \end{aligned}$$

tj.

$$\begin{aligned} T_{4 \cdot 5} = & \frac{1}{4 \cdot 5} \left(1 + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} \right. \\ & \quad \left. + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{4 \cdot 5} + \frac{1}{4 \cdot 6} + \frac{1}{5 \cdot 6} + \dots \right), \end{aligned} \quad (3.12)$$

$$T_{4 \cdot 6} = \frac{1}{4 \cdot 6} + \frac{1}{4^2 \cdot 6} + \frac{1}{6^2 \cdot 4} + \frac{1}{4^3 \cdot 6} + \frac{1}{6^3 \cdot 4} + \frac{1}{4^2 \cdot 6^2} + \dots,$$

tj.

$$T_{4 \cdot 6} = \frac{1}{4 \cdot 6} \left(1 + \frac{1}{4} + \frac{1}{6} + \frac{1}{4^2} + \frac{1}{6^2} + \frac{1}{4 \cdot 6} + \dots \right), \quad (3.13)$$

$$T_{5 \cdot 6} = \frac{1}{5 \cdot 6} + \frac{1}{5^2 \cdot 6} + \frac{1}{6^2 \cdot 5} + \frac{1}{5^3 \cdot 6} + \frac{1}{6^3 \cdot 5} + \frac{1}{5^2 \cdot 6^2} + \dots,$$

tj.

$$T_{5 \cdot 6} = \frac{1}{5 \cdot 6} \left(1 + \frac{1}{5} + \frac{1}{6} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{5 \cdot 6} + \dots \right). \quad (3.14)$$

Hledaný součet S určíme pomocí součtů dílčích řad (3.9)–(3.14). Užitím vzorce (2.1) pro součet konvergentní geometrické řady získáme součty S_4 , S_5 a S_6 řad (3.9), (3.10) a (3.11):

$$S_4 = \frac{1}{4} \frac{1}{1 - 1/4} = \frac{1}{4} \frac{4}{4 - 1} = \frac{1}{3}, \quad (3.15)$$

$$S_5 = \frac{1}{5} \frac{1}{1 - 1/5} = \frac{1}{5} \frac{5}{5 - 1} = \frac{1}{4}, \quad (3.16)$$

$$S_6 = \frac{1}{6} \frac{1}{1 - 1/6} = \frac{1}{6} \frac{6}{6 - 1} = \frac{1}{5}. \quad (3.17)$$

Je zřejmé, že součet $S_{4 \cdot 5}$ dílčí řady (3.12) můžeme zapsat ve tvaru

$$S_{4 \cdot 5} = \frac{1}{4 \cdot 5} (1 + S) = \frac{1}{20} (1 + S). \quad (3.18)$$

Součty $S_{4 \cdot 6}$ a $S_{5 \cdot 6}$ řad (3.13) a (3.14) jsou součty všech zlomků s jedničkou v čitateli a ve jmenovateli se všemi mocninami součinů čísel 4, 6 a 5, 6.

Výraz v závorkách ve vyjádření součtu $S_{4 \cdot 6}$ řady (3.13) přerovnáme a zapíšeme jako součet čísla 1 a tří dílčích řad T_4 , T_6 a $T_{4 \cdot 6}$ se součty $S_4 = 1/3$, $S_6 = 1/5$ a $S_{4 \cdot 6}$. Dostáváme tak rovnici

$$S_{4 \cdot 6} = \frac{1}{4 \cdot 6} \left(1 + \frac{1}{3} + \frac{1}{5} + S_{4 \cdot 6} \right).$$

Vynásobením číslem 360 obdržíme rovnici $360S_{4 \cdot 6} = 15 + 5 + 3 + 15S_{4 \cdot 6}$, odkud

$$S_{4 \cdot 6} = \frac{23}{345} = \frac{1}{15}. \quad (3.19)$$

Analogicky obdržíme pro součet $S_{5 \cdot 6}$ rovnici

$$S_{5 \cdot 6} = \frac{1}{5 \cdot 6} \left(1 + \frac{1}{4} + \frac{1}{5} + S_{5 \cdot 6} \right).$$

Z této rovnice obdržíme úpravou rovnici $600S_{5 \cdot 6} = 20 + 5 + 4 + 20S_{5 \cdot 6}$, odkud

$$S_{5 \cdot 6} = \frac{29}{580} = \frac{1}{20}. \quad (3.20)$$

Na základě platnosti absolutní konvergence řady T můžeme vzhledem k (3.8) její součet S zapsat ve tvaru

$$S = S_4 + S_5 + S_6 + S_{4 \cdot 5} + S_{4 \cdot 6} + S_{5 \cdot 6}.$$

Užitím vztahů (3.15)–(3.20) obdržíme rovnici

$$S = \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{20} (1 + S) + \frac{1}{15} + \frac{1}{20}.$$

Vynásobením rovnice číslem 60 dostaneme rovnici

$$60S = 20 + 15 + 12 + 3 + 3S + 4 + 3,$$

tj. $57S = 57$, odkud

$$S = 1.$$

b) Uvažujme tři geometrické řady U , V a W s prvním členem 1 a s kvocienty $4^{-1} = \frac{1}{4}$, $5^{-1} = \frac{1}{5}$ a $6^{-1} = \frac{1}{6}$, tedy geometrické řady tvaru

$$\begin{aligned} U &= 1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \cdots, \\ V &= 1 + \frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \frac{1}{5^4} + \cdots, \\ W &= 1 + \frac{1}{6} + \frac{1}{6^2} + \frac{1}{6^3} + \frac{1}{6^4} + \cdots \end{aligned}$$

se součty S_U , S_V a S_W , které podle vzorce (2.1) jsou

$$\begin{aligned} S_U &= \frac{1}{1 - 1/4} = \frac{4}{4 - 1} = \frac{4}{3}, \\ S_V &= \frac{1}{1 - 1/5} = \frac{5}{5 - 1} = \frac{5}{4}, \\ S_W &= \frac{1}{1 - 1/6} = \frac{6}{6 - 1} = \frac{6}{5}. \end{aligned}$$

Protože jsou řady U , V a W absolutně konvergentní, je jejich součinem opět absolutně konvergentní řada

$$\begin{aligned} U \cdot V \cdot W &= \left(1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \cdots\right) \left(1 + \frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \frac{1}{5^4} + \cdots\right) \times \\ &\quad \times \left(1 + \frac{1}{6} + \frac{1}{6^2} + \frac{1}{6^3} + \frac{1}{6^4} + \cdots\right) \end{aligned}$$

Je zřejmé, že roznásobením členů těchto tří řad obdržíme řadu tvaru

$$\begin{aligned} U \cdot V \cdot W &= 1 + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{4 \cdot 5} + \frac{1}{4 \cdot 6} + \frac{1}{5 \cdot 6} \\ &\quad + \frac{1}{4^3} + \frac{1}{5^3} + \frac{1}{6^3} + \frac{1}{4^2 \cdot 5} + \frac{1}{4^2 \cdot 6} + \frac{1}{5^2 \cdot 4} + \frac{1}{5^2 \cdot 6} \\ &\quad + \frac{1}{6^2 \cdot 4} + \frac{1}{6^2 \cdot 5} + \frac{1}{4 \cdot 5 \cdot 6} + \frac{1}{4^4} + \frac{1}{5^4} + \frac{1}{6^4} + \cdots, \end{aligned}$$

tedy řadu $1 + T$ se součtem $1 + S$. Z rovnice

$$1 + S = S_U S_V S_W$$

dostáváme

$$S = S_U S_V S_W - 1,$$

tj. $S = \frac{4}{3} \cdot \frac{5}{4} \cdot \frac{6}{5} - 1$, odkud

$$S = 1.$$

Oběma způsoby výpočtu jsme tak obdrželi opět stejný výsledek: Daná řada T má součet $S = 1$.

Je zřejmé, že druhý způsob výpočtu se zápisem dané řady ve tvaru součinu dílčích konvergentních geometrických řad je výrazně jednodušší a pro více než dvě přirozená čísla generující tyto řady i značně přehlednější. Proto nyní uvedeme hlavní výsledek týkající se druhého způsobu určení součtu uvažovaných řad.

4. VZOREC PRO VÝPOČET SOUČTU ŘADY PŘEVŘÁCENÝCH HODNOT VŠECH SOUČINŮ GENEROVANÝCH n PŘIROZENÝMI ČÍSLY

Zobecněním druhého způsobu řešení výše uvedených příkladů 3.1 a 3.2 dostáváme následující tvrzení.

Věta 4.1. *Součet $S(a_1, a_2, \dots, a_n)$ řady převrácených hodnot všech součinů generovaných n různými jednocifernými čísly $a_1, a_2, \dots, a_n$ většími než 1, kde $2 \leq n \leq 8$, je dán vztahem*

$$S(a_1, a_2, \dots, a_n) = \prod_{k=1}^n \frac{1}{1 - 1/a_k} - 1,$$

tedy vztahem

$$S(a_1, a_2, \dots, a_n) = \prod_{k=1}^n \frac{a_k}{a_k - 1} - 1. \quad (4.1)$$

Poznámka 4.2. Ve speciálním případě, kdy čísla $a_1, a_2, \dots, a_n$ tvoří posloupnost n po sobě jdoucích přirozených čísel $(m, m+1, \dots, m+n-1)$, se v součinu $\prod_{k=1}^n \frac{a_k}{a_k-1}$ ve vztahu (4.1) zřejmě vykrátí všechny čitatele a jmenovatele, až na poslední čísel a první jmenovatel, takže pro n po sobě jdoucích přirozených čísel začínajících číslem m platí jednodušší vztah

$$S(m, m+1, \dots, m+n-1) = \frac{m+n-1}{m-1} - 1,$$

tedy vztah

$$S(m, m+1, \dots, m+n-1) = \frac{n}{m-1}, \quad (4.2)$$

speciálně pak vztah

$$S(2, 3, \dots, n+1) = n.$$

Takže například $S(2, 3) = \frac{2}{2-1} = 2$ a $S(4, 5, 6) = \frac{3}{4-1} = 1$, jak jsme vypočítali výše v příkladech 3.1, a 3.2.

Poznámka 4.3. Poznamenejme, že výše uvedený vzorec (4.1) samozřejmě platí nejen pro jednociferná čísla, ale pro libovolná přirozená čísla s výjimkou jedničky. Takže například v úvodu uvedená motivační úloha na určení součtu všech zlomků, které mají v čitateli jedničku a ve jmenovateli součiny generované všemi prvočíselnými děliteli čísla 2002, tedy prvočísla 2, 7, 11 a 13, má součet

$$S(2, 7, 11, 13) = \frac{2}{2-1} \frac{7}{7-1} \frac{11}{11-1} \frac{13}{13-1} - 1 = \frac{2 \cdot 7 \cdot 11 \cdot 13}{1 \cdot 6 \cdot 10 \cdot 12} - 1 = 1,708\bar{5}.$$

5. NUMERICKÉ OVĚŘENÍ VZORCE

Vzorec (4.1) ověříme numericky alespoň pro dvě a tři jednociferná generující čísla užitím základního programovacího jazyka systému počítačové algebry Maple 2022. Užijeme přitom následující procedury `sum2` a `sum3`, které porovnávají hodnoty součtů vypočtené vzorcem (4.1) s přibližnými hodnotami součtů členů těchto řad, pro které má součet p všech exponentů n jednociferných činitelů ve jmenovateli hodnotu $p = 100$, pro kterou jsou délky trvání příslušných výpočtů ještě docela krátké. Procedurou `sum2` bylo vypočteno $\binom{8}{2} = 28$ součtů řad během 6 sekund a procedurou `sum3` bylo vypočteno $\binom{8}{3} = 56$ součtů řad během 86 sekund.

Pokud bychom výpočty prováděli pro všechna generující jednociferná čísla s výjimkou jedničky, bylo by takových dílčích vypočtených součtů celkem

$$\binom{8}{2} + \binom{8}{3} + \binom{8}{4} + \binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8} = 28 + 56 + 70 + 56 + 28 + 8 + 1 = 247.$$

Obě níže uvedené procedury `sum2` a `sum3` byly následně užity v rámci dvou, resp. tří, vnořených cyklů.

```
sum2:=proc(a,b,p)
    local i,j,n,re,s,sf;
    s:=0; sf:=(a*b)/((a-1)*(b-1))-1;
    for n from 1 to p do
        for i from 0 to n do
            for j from 0 to n-i do
                if i+j = n then
                    s:=s+1/(a^i*b^j);
                end if;
            end do;
        end do;
    end do;
    re:=abs(s-sf)/s;
    print("sum for the most power",p,"for",a,b,"is",evalf[16](s),
        "sf is",sf,"relative error is",evalf[10](re));
end proc;

for a from 2 to 8 do
    for b from a+1 to 9 while b <> a do
        sum2(a,b,100);
    end do;
end do;

sum3:=proc(a,b,c,p)
    local i,j,k,n,re,s,sf;
    s:=0; sf:=(a*b*c)/((a-1)*(b-1)*(c-1))-1;
    for n from 1 to p do
        for i from 0 to n do
            for j from 0 to n-i do
```

```

        for k from 0 to n-i-j do
            if i+j+k = n then
                s:=s+1/(a^i*b^j*c^k);
            end if;
        end do;
    end do;
end do;
end do;
re:=abs(s-sf)/s;
print("sum for the most power",p,"for",a,b,c,"is",evalf[16](s),
      "sf is",sf,"relative error is",evalf[10](re));
end proc;

for a from 2 to 8 do
    for b from a+1 to 9 while b <> a do
        for c from b+1 to 9 while c <> b do
            sum3(a,b,c,100);
        end do;
    end do;
end do;

```

Pro součet exponentů $p = 100$ byly součty řad generovaných dvěma čísly vypočteny s relativními chybami v rozpětí řádově 10^{-30} až 10^{-90} a součty řad generovaných třemi čísly s relativními chybami v rozpětí řádově 10^{-30} až 10^{-84} .

Tabulka 2 obsahuje součty řad generovaných dvěma jednocifernými čísly většími než jedna určené vzorcem (4.1) a ověřené procedurou `sum2`.

Celočíselných hodnot přitom nabývají pouze dva součty: $s(2, 3) = 2$ (příklad 3.1) a $s(3, 4) = 1$.

$a \setminus b$	3	4	5	6	7	8	9
2	2	$5/3$	$3/2$	$7/5$	$4/3$	$9/7$	$5/4$
3	×	1	$7/8$	$4/5$	$3/4$	$5/7$	$11/16$
4	×	×	$2/3$	$3/5$	$5/9$	$11/21$	$1/2$
5	×	×	×	$1/2$	$11/24$	$3/7$	$13/32$
6	×	×	×	×	$2/5$	$13/35$	$7/20$
7	×	×	×	×	×	$1/3$	$5/16$
8	×	×	×	×	×	×	$2/3$

Tabulka 2. Součty řad generovaných dvěma jednocifernými čísly.

Tabulka 3, která je uvedena na následující straně, obsahuje součty řad generovaných třemi jednocifernými čísly většími než jedna určené vzorcem (4.1) a ověřené procedurou `sum3`.

Celočíselných hodnot nabývá jen těchto pět součtů: $s(2, 3, 4) = 3$, $s(2, 4, 9) = 2$, $s(2, 5, 6) = 2$, $s(3, 7, 8) = 1$ a $s(4, 5, 6) = 1$ (příklad 3.2).

$a = 2$	$c = 4$	$c = 5$	$c = 6$	$c = 7$	$c = 8$	$c = 9$
$b = 3$	3	11/4	13/5	5/2	17/7	19/8
$b = 4$	×	7/3	11/5	19/9	43/21	2
$b = 5$	×	×	2	23/12	13/7	29/16
$b = 6$	×	×	×	9/5	61/35	17/10
$b = 7$	×	×	×	×	5/3	13/8
$b = 8$	×	×	×	×	×	11/7
$a = 3$	×	$c = 5$	$c = 6$	$c = 7$	$c = 8$	$c = 9$
$b = 4$	×	3/2	7/5	4/3	9/7	5/4
$b = 5$	×	×	5/4	19/16	8/7	71/64
$b = 6$	×	×	×	11/10	37/35	41/40
$b = 7$	×	×	×	×	1	31/32
$b = 8$	×	×	×	×	×	13/14
$a = 4$	×	×	$c = 6$	$c = 7$	$c = 8$	$c = 9$
$b = 5$	×	×	1	17/18	19/21	7/8
$b = 6$	×	×	×	13/15	29/35	4/5
$b = 7$	×	×	×	×	7/9	3/4
$b = 8$	×	×	×	×	×	5/7
$a = 5$	×	×	×	$c = 7$	$c = 8$	$c = 9$
$b = 6$	×	×	×	3/4	5/7	11/16
$b = 7$	×	×	×	×	2/3	41/64
$b = 8$	×	×	×	×	×	17/28
$a = 6$	×	×	×	×	$c = 8$	$c = 9$
$b = 7$	×	×	×	×	3/5	23/40
$b = 8$	×	×	×	×	×	19/35
$a = 7$	×	×	×	×	×	$c = 9$
$b = 8$	×	×	×	×	×	1/2

Tabulka 3. Součty řad generovaných třemi jednocifernými čísly.

6. ZÁVĚR

Článek se zabývá jedním typem nekonečných číselných řad, které jsou generovány převrácenými hodnotami ke všem součinům dvou a více přirozených čísel. Tyto řady představují speciální případ redukované harmonické řady

Na konkrétních příkladech jsou prezentovány dva způsoby, jak tyto řady sečíst. Jednak užitím jejich zápisu ve tvaru součtu dílčích řad, které představují konvergentní nekonečné geometrické řady, a jednak značně jednodušším způsobem, při kterém řady zapíšeme ve tvaru součinu dílčích konvergentních nekonečných geometrických řad.

Na základě konkrétních příkladů těchto řad se dvěma a třemi generátory je uvedena věta obsahující vzorec pro libovolných n generujících přirozených čísel

$$S(a_1, a_2, \dots, a_n) = \prod_{k=1}^n \frac{a_k}{a_k - 1} - 1$$

a rovněž její důsledek pro n generátorů, které tvoří posloupnost po sobě jdoucích přirozených čísel. Získané výsledky jsou numericky ověřeny pro dvě a tři jednocíferné generátory užitím systému počítačové algebry Maple.

Lze tedy konstatovat, že řady, které jsou generovány převrácenými hodnotami ke všem součinům dvou a více přirozených čísel, představují speciální případ konvergentních nekonečných řad, podobně jako geometrické řady a řady s teleskopickou vlastností, jejichž součet je určen relativně jednoduchým analytickým vzorcem.

REFERENCE

- [1] R. Baillie: *Summing the curious series of Kempner and Irwin*, 2008, online <http://arxiv.org/pdf/0806.4410.pdf> [Cit. 2023-08-31].
- [2] G. H. Hardy, E. M. Wright: *An Introduction to the Theory of Numbers*, 4th Edition, Oxford University Press, London, 1975, online <https://1library.net/document/zz5r2woq-hardy-wright-introduction-theory-numbers-pdf.html> [Cit. 2023-08-31].
- [3] J. Havil: *Gamma: Exploring Euler's Constant*, Princeton University Press, New Jersey, 2003, online https://www.academia.edu/8413794/Gamma_Exploring_Eulers_Constant_Havil [Cit. 2023-08-31].
- [4] A. J. Kempner: *A Curious Convergent Series*, Amer. Math. Monthly **21**(1) (1914), 48–50, online <https://www.jstor.org/stable/2972074> [Cit. 2023-08-31].
- [5] T. Rike: *Infinite Series (Berkeley seminary)*, The study material of the Berkeley Math Circle, March 24, 2002, online <https://www.docin.com/p-107850630.html> [Cit. 2023-08-31].
- [6] T. Schmelzer, R. Baillie: *Summing a Curious, Slowly Convergent Series*, Amer. Math. Monthly **115**(6) (1979), 525–540, online https://www.researchgate.net/publication/233542102_Summing_a_Curious_Slowly_Convergent_Series [Cit. 2023-08-31].

Radovan Potůček, Katedra matematiky a fyziky, Fakulta vojenských technologií, Univerzita obrany, Kounicova 65, 662 10 Brno, Česká republika,
e-mail: Radovan.Potucek@unob.cz

HRA HACKENSTRING A DYADICKÁ ČÍSLA

VÁCLAV VOPRAVIL

ABSTRAKT. V příspěvku se zabýváme kombinatorickými hrami pro dva hráče s úplnou informací a bez náhody. Ukážeme, že krátká hra HACKENSTRING má hodnoty pouze dyadická racionální čísla.

1. ÚVOD

Jedním z cílů teorie kombinatorických her je najít vítězné strategie. Jiným cílem je určit (Conwayovu, viz [4] a [2]) hodnotu hry a tím korigovat své tahy. Tedy odpovědět na otázky, kdo vyhraje a o kolik. Dalším cílem teorie je určit první vítězný tah. V obecném případě se nám to ne vždy podaří. V tomto článku uvádíme vítězné strategie pro určité pozice hry BÍLOČERNÝ HACKENSTRING. BÍLOČERNÝ HACKENSTRING je graf bílých a černých hran, které jsou modelovány kameny v řadách a spojeny vlevo s vrcholem nazývaným země (základna). Základnu budeme znázorňovat jako svislou černou čáru. Hru hrají dva hráči: Levý a pRavý. Na tahu Levý musí vybrat bílý kámen (hrana grafu), který odstraní. Všechny hrany, které již nejsou spojeny se základnou, jsou také smazány. Tah pRavého hráče je podobný, ale musí si vybrat černý kámen, který chce odstranit (smazat). První hráč, který nemůže táhnout, prohrává. Každé pozici hry BÍLOČERNÝ HACKENSTRING lze přiřadit dyadické racionální číslo. Tato hodnota zcela určuje, kdo hru vyhraje. Dobrou představou je vláček s bílými a černými kameny, vlevo je lokomotiva. Po odebrání svého vagonku hráč odebírá i všechny vagonky, které nesouvisí se základnou. Hra se hraje i s více vláčky srovnanými do řádků pod sebe.

Poznamenejme ještě, že hra HACKENSTRING se studuje také v monografii [1], v níž lze nalézt srozumitelně podaný úvod do teorie kombinatorických her. Článek je založen na [9, 10]. V českém jazyce lze doporučit [3, 11, 12]. Pro první seznámení s tématem pak první díl [2] a [5].

2. BÍLOČERNÝ HACKENSTRING

V této části je naším cílem seznámit čtenáře s kombinatorickou teorií her s důrazem na analýzu pozic ve hře BÍLOČERNÝ HACKENSTRING. Nyní by měl čtenář předpokládat, že všechny hry se hrají podle pravidel obvyklých pro BÍLOČERNÝ HACKENSTRING. Slova hra a pozice jsou zaměnitelná. Téměř veškerý materiál v této sekci lze nalézt v [9, 10] a [5].

2020 MSC. Primární 91A46.

Klíčová slova. teorie kombinatorických her, hra hackenstring, dyadická čísla.

2.1. BíLoČeRný Hackenstring

Hra BíLoČeRný HACKENSTRING je hra, která se hraje na grafu bíLých a čeRných hran. Existuje jedinečný vrchol, který označíme jako základnu. Je vhodné nakreslit vrchol země jako svislou černou čáru. Hru hrají dva hráči, Levý a pRavý, v tazích se hráči střídají. Na tahu musí hráč odstranit hranu své barvy. Levý může smazat bíLé a pRavý může smazat čeRné hrany. Po smazání hrany všechny ostatní hrany, které nejsou spojeny se základnou cestou, jsou také smazány. První hráč, který nemůže táhnout, prohrává.

BíLé hrany budeme modelovat bíLými kroužky (kameny) a čeRné budeme modelovat čeRnými kameny (kroužky). Odebírá se zprava. Příklad takové hry je na obrázku 2.1. V něm i ve zbytku článku budeme bílé hrany znázorňovat bílými ka-



Obrázek 2.1. Příklad hry BíLoČeRný HACKENSTRING.

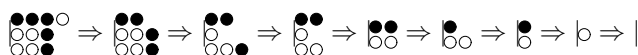
meny a čeRné hrany čeRnými kameny. Vrátime-li se k výše uvedenému příkladu, můžeme provést hrubou silou pomocí tužky a papíru. Vítězem této hry je Levý hráč. Levý může tuto hru vyhrát bez ohledu na to, zda ve hře začínal, či ne.

Další příklad je uveden na obrázku 2.2. Ten je zvláštní tím, že má přirozenou



Obrázek 2.2. Příklad hry Tweedle Dee Tweedle Dum.

strategii pro druhého hráče. Druhý hráč může napodobovat tahy prvního hráče. Tímto způsobem druhému hráči nikdy nedojdou tahy dříve než prvnímu hráči. Strategie jednoho hráče, který napodobuje tahy druhého hráče, se nazývá strategie Tweedle Dee Tweedle Dum (TDTD). Výše uvedená hra je vítěznou pro druhého hráče (ten, který ve hře nezačínal).



Obrázek 2.3. Ukázka hry, pRavý začíná.

Zde začíná čeRný hráč a v poslední pozici, protože tam není žádný čeRný kámen, prohrává, nemá žádný možný tah.



Obrázek 2.4. Několik počátečních pozic.

2.2. Základní definice

Nechť G je hra. Píšeme

$$G = \{\mathcal{G}^L \mid \mathcal{G}^R\},$$

kde $\mathcal{G}^L = \{\text{množina dílčích pozic } G, \text{ kam Levý může zahrát jedním tahem}\}$ a dále $\mathcal{G}^R = \{\text{množina dílčích pozic } G, \text{ kam může pravý zahrát prvním tahem v } G\}$.

Definujeme prvky $\mathcal{G}^L$ jako levé možnosti G . Levou možnost označujeme také jako G^L . Definujeme analogické termíny pro pravou variantu (možnost). V tomto článku budeme psát

$$G = \{G^L \mid G^R\},$$

kde G^L zahrnuje všechny levé možnosti a G^R zahrnuje všechny možné pravé možnosti.

Hra G je krátká hra, pokud splňuje následující dvě vlastnosti:

- G má konečný počet dílčích pozic.
- V G neexistuje nekonečná posloupnost tahů, tj. hra nakonec skončí.

Všechny hry v této kapitole jsou krátké hry. Hrajeme podle normální konvence, že první hráč, který není schopen tahu, prohrává. Předpokládáme také, že každá hra BÍLOČERNÝ HACKENSTRING má konečný počet hran.

Věta 2.1. *Každý BÍLOČERNÝ HACKENSTRING G je krátká hra.*

Důkaz. Nechť G je hra BÍLOČERNÝ HACKENSTRING. Na G se budeme dívat jako na graf bílých a černých hran. Podle předpokladu má G n hran, kde n nějaké nezáporné celé číslo. Každý tah na G sníží počet hran alespoň o jednu. V nejvýše n tazích bude tedy G prázdný. Z toho vyplývá, že na G neexistuje nekonečná posloupnost tahů. G má konečný počet podgrafů. Každá podpozice G je podgrafem G . Můžeme tedy shrnout, že G je krátká. $\square$

2.3. Krátké hry

V této části zkonstruujeme všechny krátké hry. Budeme definovat určitý vztah mezi krátkými hrami a uvidíme, že množina tříd ekvivalence spolu s určitým doplněním tvoří částečně uspořádanou abelovskou grupu.

Definujeme $0 = \{|\}$ a $\tilde{\mathbb{G}}_0 = \{0\}$. Krátké hry narozené v den n rekurzivně definujeme takto

$$\tilde{\mathbb{G}}_n = \{\{G^L \mid G^R\}; G^L, G^R \in \tilde{\mathbb{G}}_{n-1}\}.$$

Definujeme

$$\tilde{\mathbb{G}} = \bigcup_{n \in \mathbb{N}} \tilde{\mathbb{G}}_n$$

jako množinu krátkých her.

Definujeme disjunktí součet na hrách jako

$$G + H = \{G^L + H, G + H^L \mid G^R + H, G + H^R\}.$$

Věta 2.2. *Součet na hrách vytváří abelovskou pologrupu.*

Důkaz. Nejprve indukci dokážeme, že sčítání je komutativní. Podívejte se, $G + H = \{G^L + H, G + H^L \mid \dots\} = \{H + G^L, H^L + G \mid \dots\} = H + G$. Důkaz pro asociativitu je podobný. $\square$

Třidu výsledků hry, označovanou $v(G)$, definujeme následujícím způsobem:

- $v(G) = \mathcal{L} \Leftrightarrow$ levá strana může vyhrát, když táhne první nebo druhá,
- $v(G) = \mathcal{N} \Leftrightarrow$ vyhrává hráč, který je na tahu jako první,
- $v(G) = \mathcal{P} \Leftrightarrow$ vyhrává druhý hráč na tahu,
- $v(G) = \mathcal{R} \Leftrightarrow$ pravá strana může vyhrát, když táhne první nebo druhá.

Hry, patřící do $\mathcal{L}$, se nazývají kladné, patřící do $\mathcal{R}$ záporné, patřící do $\mathcal{P}$ nulové a patřící do $\mathcal{N}$ označíme jako fuzzy.

Věta 2.3 (Základní věta teorie kombinatorických her). *Pro libovolnou krátkou G může vyhrát buď levá strana, která hraje první, nebo pravá strana, která hraje druhá, ale ne obě současně.*

Důkaz. Necht G je krátká hra a G^L je libovolná levá možnost G . Potom indukci a pomocí symetrie může pravá strana vyhrát G^L hrající jako první, nebo může vyhrát levá strana G^L hrající jako druhá, ale ne obojí. Jestliže pro všechny levé možnosti G^L , může vyhrát pRavý hrající G^L jako první, pak může pRavý vyhrát G hrající jako druhý. Na druhou stranu, pokud existuje levá možnost G^L taková, že levá strana může vyhrát hrající jako druhá, pak může levá strana vyhrát G přechodem na takovou variantu G^L . Je jasné, že nastane přesně jeden případ z těchto dvou možností. To dokazuje tvrzení. $\square$

Důsledek 2.4. *Každá krátká hra G patří přesně do jedné třídy výsledků.*

Důkaz. Podle věty 2.3 patří každá krátká hra G do $\mathcal{L}$, $\mathcal{R}$, $\mathcal{N}$ nebo $\mathcal{P}$. Pokud hra G patřila do více než jedné výsledné třídy, pak bychom měli protipříklad k větě 2.3. $\square$

Lemma 2.5. *Jestliže X patří do $\mathcal{P}$, pak $v(G + X) = v(G)$.*

Důkaz. Je třeba uvažovat dva případy:

Případ 1: Předpokládejme, že levá strana může vyhrát G , když hraje jako druhá. Předpokládejme, že pravá strana táhne jako první na $G + X$. Pak má levá strana zaručenou odpověď na té složce, ve které pravá strana táhla. Levá strana může pokračovat v této strategii, dokud pravé straně nedojdou tahy.

Případ 2: Předpokládejme, že Levý může vyhrát G hrající jako první. Pak má levá strana vítězný tah na G , který označíme jako G^L . Všimněte si, že Levý může vyhrát G^L , hraje-li jako druhý. Z toho plyne, že levá strana má tah na $G^L + X$, což je vítězný tah podle strategie v případě 1. $\square$

Lemma 2.6. *Jestliže G a H patří do $\mathcal{L}$ nebo $\mathcal{P}$, pak $G + H$ patří také do $\mathcal{L}$ nebo $\mathcal{P}$.*

Důkaz. Stačí ukázat indukci, že pRavý nemůže vyhrát, když ve hře začíná. Pokud pravá strana táhne první, pak podle indukční hypotézy musí mít Levý odpověď ve stejné komponentě. Levý tak může vždy najít správnou odpověď v komponentě, kde zahrál pRavý hráč. Proto také levé straně nedojdou tahy dříve než pravé straně, a protože tyto hry jsou krátké, pravé straně musí nakonec dojít tahy. $\square$

Definice 2.7. Definujeme opačnou hodnotu krátké hry jako

$$-G = \{-G^R \mid -G^L\}.$$

Definice 2.8 (Kvaziuspořádání na $\tilde{\mathbb{G}}$). Relaci $\geq$ na $\tilde{\mathbb{G}}$ definujeme tak, že říkáme, že $G \geq H$, jestliže a pouze tehdy, když $v(G + (-H)) = \mathcal{L}$ nebo $\mathcal{P}$.

Věta 2.9. $\geq$ je kvaziuspořádání na $\tilde{\mathbb{G}}$.

Důkaz. (Reflexivita): Uvažujme hru $G + (-G)$. Pak každý tah prvního hráče může být napodoben druhým hráčem ve druhé složce. Z toho vyplývá, že druhému hráči nikdy nedojdou tahy dříve než prvnímu hráči. Protože obě hry jsou krátké, prvnímu hráči musí nakonec dojít tahy. Proto $G + (-G)$ patří do $\mathcal{P}$ a $G \geq G$.

(Tranzitivita): Předpokládejme, že $G \geq H$ a $H \geq J$. Pak $G + (-H)$ a $H + (-J)$ patří do $\mathcal{L}$ nebo $\mathcal{P}$. Podle lemmatu 2.6 patří $G + (-H) + H + (-J)$ do $\mathcal{L}$ nebo $\mathcal{P}$. Podle lemmatu 2.5 a vzhledem k tomu, že $H + (-H)$ patří do $\mathcal{P}$, máme $G + (-J)$ patří do $\mathcal{L}$ nebo $\mathcal{P}$. $\square$

Lemma 2.10. Necht G, H jsou krátké hry. Pak $-(G + H) = (-G) + (-H)$.

Důkaz. Podle definice 2.7 a indukci dostaneme

$$\begin{aligned} -(G + H) &= -\{G^L + H, G + H^L \mid \dots\} = \{-(G^R + H), -(G + H)^R \mid \dots\} \\ &= \{-G^R - H, -G - H^R \mid \dots\} = (-G) + (-H). \end{aligned}$$

$\square$

Věta 2.11. Pro všechny krátké hry G, H, X platí, že $G \geq H$ implikuje $G + X \geq H + X$.

Důkaz. S použitím věty 2.2 a lemmatu 2.10, $(G + X) + (-(H + X)) = G + (X + (-X)) - H$. Protože $X + (-X)$ patří do $\mathcal{P}$, díky lemma 2.5 dostaneme $v(G - H) = v(G + X + (-(H + X)))$. Protože $G \geq H$, máme $G + X \geq H + X$. $\square$

Ukázali jsme, že $(\tilde{\mathbb{G}}, +)$ je kvaziuspořádaná abelovská pologrupa. Dále definujeme relaci na $\tilde{\mathbb{G}}$ jako $G = H$ tehdy a jen tehdy, když $G \geq H$ a $H \geq G$.

Věta 2.12. Výše definovaná relace $=$ je relace ekvivalence na $\tilde{\mathbb{G}}$.

Důkaz. Necht $G, H, J \in \tilde{\mathbb{G}}$.

(Reflexivita): Ekvivalentně $G \geq G$, což ukazuje, že $G + (-G)$ patří do $\mathcal{P}$. $G = G$.

(Symetričnost): Předpokládejme, že $G = H$. Pak $G \geq H$ a $H \geq G$. Ekvivalentně platí, že $H \geq G$ a $G \geq H$. Tedy $H = G$.

(Tranzitivita): Předpokládejme, že $G = H$ a $H = J$. Pak $G \geq H$ a $H \geq J$. Protože $\geq$ je tranzitivní, máme $G \geq J$. Podobně $J \geq G$. Takže $G = J$. $\square$

Definujeme $\mathbb{G}$ jako množinu tříd ekvivalence tvořenou $\tilde{\mathbb{G}}$ a výše uvedenou relací. Prvky $\mathbb{G}$ označujeme jako herní hodnoty. Sčítání v $\mathbb{G}$ definujeme jako $[G] + [H] = [G + H]$. Dále definujeme relaci $\geq$ na $\mathbb{G}$ pomocí $[G] \geq [H]$ tehdy a jen tehdy, když $G' \geq H'$ pro nějaké $G' \in [G]$ a nějaké $H' \in [H]$. Říkáme, že $[G] \models [H]$ tehdy a jen tehdy, když $[G] \not\geq [H]$. Ukážeme, že $(\mathbb{G}, +)$ je částečně uspořádaná abelovská grupa.

Věta 2.13. $[G] \geq [H]$ platí tehdy a jen tehdy, když $G \geq H$ pro nějaké $G \in [G]$ a nějaké $H \in [H]$ je dobře definovaná.

Důkaz. Necht $[G], [H]$ jsou herní hodnoty a $G, G' \in [G]$ a $H, H' \in [H]$. Předpokládejme, že $G \geq H$. Chceme ukázat, že $G' \geq H'$. Všimněte si, že $H = H'$ znamená, že $-H = -H'$. Z toho vyplývá, že $-H' \geq -H$. Máme také $G' \geq G$, $G' + (-H') \geq G + (-H')$ (Podle věty 2.11) $G' + (-H') \geq G + (-H)$ (Podle věty 2.11) $G' + (-H') \geq G + (-H)$ (Protože $\geq$ v $\tilde{\mathbb{G}}$ je tranzitivní). Znovu použijeme tranzitivitu a dostaneme $G' + (-H') \geq 0$. Z toho plyne, že $G' + (-H')$ patří do $\mathcal{L}$ nebo $\mathcal{P}$, z čehož vyplývá, že $G' \geq H'$. $\square$

Věta 2.14. $\geq$ je částečné uspořádání na $\mathbb{G}$.

Důkaz. Necht $[G], [H], [J]$ jsou herní hodnoty.

(Reflexivita): Vidíme, že $G \geq G$, protože $G + (-G)$ patří do $\mathcal{P}$. Proto $[G] \geq [G]$.

(Tranzitivita): Předpokládejme, že $[G] \geq [H]$ a $[H] \geq [J]$. Pak $(G + (-H)) + (H + (-J))$ je součet dvou her, které patří do $\mathcal{L}$ nebo $\mathcal{P}$. Součet tedy také patří do $\mathcal{L}$ nebo $\mathcal{P}$. Součet můžeme také přepsat jako $G + (-H + H) + (-J)$. Protože $-H + H$ patří do $\mathcal{P}$, můžeme uzavřít, že výsledek $G + (-J)$ je stejný jako výsledek $G + (-H + H) + (-J)$. Z toho vyplývá, že $G \geq J$. To znamená, že $[G] \geq [J]$.

(Antisymetričnost): Předpokládejme, že $[G] \geq [H]$ a $[H] \geq [G]$. Pak $G \geq H$ a $H \geq G$. Podle naší definice = máme $G = H$, z čehož vyplývá $[G] = [H]$. $\square$

Věta 2.15. Výše definované sčítání je dobře definováno.

Důkaz. Necht $[G], [G'], [H], [H'] \in \mathbb{G}$ a předpokládejme, že $[G] = [G']$ a $[H] = [H']$. Chceme ukázat $[G + H] = [G' + H']$. Vidíme, že $G \geq G'$ a $H \geq H'$. Takže $G - G'$ a $H - H'$ jsou v $\mathcal{L}$ nebo $\mathcal{P}$. Pak hra $(G + H) - (G' + H') = (G - G') + (H - H')$ patří do $\mathcal{L}$ nebo $\mathcal{P}$. Ekvivalentně platí, že $G + H \geq G' + H'$. S použitím symetrického argumentu platí, že $G' + H' \geq G + H$. Toto ukazuje, že $[G + H] = [G' + H']$ a tedy nezáleží na reprezentantech. $\square$

Věta 2.16. $(\mathbb{G}, +)$ je komutativní plogrupa.

Důkaz. Tvrzení platí, protože $(\tilde{\mathbb{G}}, +)$ je komutativní plogrupa. $\square$

Věta 2.17. $[0]$ je identita v $\mathbb{G}$.

Důkaz. Všimněte si, že $[G] + [0] = [G + 0] = [G]$ pro všechna $[G] \in \mathbb{G}$. $\square$

Lemma 2.18. Jestliže $[G] \geq [H]$, pak $[G] + [J] \geq [H] + [J]$.

Důkaz. Nejprve ukážeme, že pokud $G \geq H$, pak $G + J \geq H + J$. Všimněte si, že $(G + J) + (-(H + J)) = (G + (-H)) + (J + (-J))$ je součet her patřících do $\mathcal{L}$ nebo $\mathcal{P}$. Proto $G + J + (-(H + J))$ také patří do $\mathcal{L}$ nebo $\mathcal{P}$. To ukazuje, že $G + J \geq H + J$. Konkrétně, jestliže $[G] \geq [H]$, pak $G \geq H$. Ukázali jsme, že z toho vyplývá $G + J \geq H + J$. Z toho plyne také, že $[G] + [J] \geq [H] + [J]$. $\square$

Věta 2.19. *Jestliže $[G] = [H]$, pak $[G] + [J] = [H] + [J]$.*

Důkaz. Předpokládejme, že $G = H$. Pak $G + J \geq H + J$ a $G + J \leq H + J$. Z definice vyplývá, že platí i rovnost, což je důkazem věty. $\square$

Lemma 2.20. *$v(G) = \mathcal{P}$ tehdy a jen tehdy, když $[G] = [0]$.*

Důkaz. $(\Rightarrow)$: Předpokládejme, že $v(G) = \mathcal{P}$. Pak G a $-G$ patří do $\mathcal{P}$. Proto $G \geq 0$ a $0 \geq G$. Ekvivalentně tedy $[G] = [0]$.

$(\Leftarrow)$: Předpokládejme, že $[G] = 0$. Pak G a $-G$ patří do $\mathcal{L}$ nebo $\mathcal{P}$. Pokud jedna z nich, řekněme G , patří do $\mathcal{L}$, pak musíme mít: $-G$ patří do $\mathcal{R}$, což je spor. Proto G a $-G$ patří do $\mathcal{P}$. Konkrétně G patří do $\mathcal{P}$. $\square$

Věta 2.21. *Pro libovolné $[G] \in \mathbb{G}$ máme $[G] + [-G] = [0]$.*

Důkaz. Vidíme, že v $G + (-G)$ je vítězem druhý hráč pomocí TDTD argumentu.¹ Podle předchozího lematu 2.20 je $[G] + [-G] = [G + (-G)] = [0]$. $\square$

Z toho můžeme vyvodit, že $(\mathbb{G}, +)$ je částečně uspořádaná abelovská grupa. Na tomto místě upustíme od zápisu v závorkách.

Důsledek 2.22. *Nechť G, H jsou herní hodnoty. Pak $G + (-H) = 0$ tehdy a jen tehdy, když $G = H$.*

Důkaz. Máme $G + (-H) = 0 \Leftrightarrow G + (-H) + H = 0 + H \Leftrightarrow G + 0 = H \Leftrightarrow G = H$. $\square$

V dalších kapitolách využijeme výše uvedený důsledek k důkazu některých pravidel.

2.4. Kanonické formy

Intuitivně jsou některé tahy lepší než jiné. V těchto případech se zdá být intuitivní ignorovat tahy, které nejsou optimální. Co máme na mysli optimálními tahy, bude předmětem této části.

Definice 2.23. Nechť G je hra a G^{L_1} a G^{L_2} jsou dvě levé možnosti hry G . Říkáme, že G^{L_2} je dominantní G^{L_1} , jestliže $G^{L_1} \geq G^{L_2}$.

Věta 2.24. *Nechť G je hra. Předpokládejme, že $G^{L'}$ je dominantní levá možnost. Uvažujme $G = \{G^{L'}, G^L \mid G^R\}$ a $H = \{G^L \mid G^R\}$, kde G^L zahrnuje všechny levé možnosti kromě $G^{L'}$. Pak $G = H$.*

¹Strategie TDTD se také nazývá metodou kradení strategie, metodou kopírování tahů nebo strategy stealing.

Důkaz. Chceme dokázat, že $G - H = 0$, což je ekvivalentní tomu, že v $G - H$ je vítězem druhý hráč. Všimněte si, že $G - H = \{G^{L'}, G^L \mid G^R\} + \{-G^R \mid -G^L\}$. Předpokládejme, že levá strana se nejprve přesune do $G^{L'} - H$. Podle předpokladu existuje G^L taková, že $G^L \geq G^{L'}$. Proto se pRavý hráč může přesunout do $G^{L'} - G^L \leq 0$. PRavý si vynutil výhru. Každý další tah jednoho z hráčů má odpovídající tah druhého hráče na 0. Zejména pokud Levý táhne na $G^L - H$, pak pravá strana může odpovědět na $G^L - G^L = 0$. A pokud Levý táhne na $G - G^R$, pak pRavá strana může reagovat na $G^R - G^R = 0$. Podobný argument platí i pro zahajovací tahy pRavého. Můžeme dojít k závěru, že v $G - H$ vítězí druhý hráč, tj. že $G = H$. $\square$

Lemma 2.25. *Pro krátké hry G, H, J platí, že pokud $G \triangleright H$ a $H \geq J$, pak $G \triangleright J$ (tj. levá strana má vítězný tah na $G - J$).*

Důkaz. Poznamenejme, že $G - J = (G - H) + (H - J)$. Při hře $G - H$ máme dáno, že Levý má vyhrávající tah, řekněme $(G - H)^L$. Označíme tento tah jako $(G - H)^L$. Pak $(G - H)^L \geq 0$. Součet her, které jsou nulové nebo kladné, je hra, která je nulová nebo kladná. Protože $(G - H)^L \geq 0$ a $H - J \geq 0$, máme $(G - H)^L + (H - J) \geq 0$. To ukazuje, že $(G - H)^L + (H - J)$ je vítězný tah pro levou stranu, což dokazuje lemma. $\square$

Lemma 2.26. *Jestliže $J \triangleleft H$, pak $G + J \triangleleft G + H$.*

Důkaz. Uvažujme rozdílovou hru $(G + H) - (G + J)$. Tu můžeme přepsat jako $(G - G) + (H - J)$. Rádi bychom ukázali, že $0 \not\geq (G - G) + (H - J)$, tj. že levá strana má výhru, začíná-li. Protože $J \triangleleft H$, Levý má vítězný tah na $H - J$. Označíme takovou možnost $(H - J)^L$. Uvažujme levou možnost $(G - G) + (H - J)^L$. To je součet her, které jsou ≥ 0 . Proto $(G - G) + (H - J)^L \geq 0$. To ukazuje, že levá strana má vítězný tah na $(G + H) - (G + J)$. $\square$

Definice 2.27. Necht G je hra a G^L je její levá možnost. Předpokládejme, že existuje pravá možnost G^{LR} taková, že $G^{LR} \leq G$. Potom říkáme, že G^L je reverzibilní přes G^{LR} .

Věta 2.28. *Necht $G = \{G^{L'}, G^L \mid G^R\}$ je hra. Předpokládejme, že $G^{L'}$ je reverzibilní přes $G^{L'R}$. Necht $H = \{G^{L'RL}, G^L \mid G^R\}$, kde $G^{L'RL}$ zahrnuje všechny levé možnosti $G^{L'R}$. Pak $G = H$.*

Důkaz. Necht G a H jsou definovány, jak je uvedeno výše. Chceme ukázat, že $G - H = 0$ (tj. vítězství druhého hráče). Všimněte si, že $G - H = \{G^{L'}, G^L \mid G^R\} + \{-G^R \mid -G^{L'RL}, G^L\}$. Uvažujme levou stranu a úvodní tah do $G^{L'} - H$. Pravá strana může táhnout na $G^{L'RL} - H$. Levá strana má dvě možnosti. Pokud levá strana se přesune do $G^{L'RL} - H$, pak má pravá strana vítězný tah na H do 0. V opačném případě levá strana táhne na $G^{L'R} - G^R$. Máme $G^{L'R} \leq G \triangleleft G^R$. Podle lemmatu 2.25 pRavý zvítězí tahem na $G^{L'R} - G^R$.

Na druhou stranu předpokládejme, že se pRavý nejprve přesune na $G - G^{L'RL}$. Uvažujme $G^{L'RL} \triangleleft G^{L'R}$. Z toho vyplývá, že $G - G^{L'RL} \triangleright G - G^{L'R} \geq 0$. Podle lemmatu 2.25 platí, že $G - G^{L'RL} \triangleright 0$, z čehož vyplývá, že Levý má vítěznou

odpověď. Všechny ostatní první tahy mají odpověď Tweedle Dee Tweedle Dum na 0. Můžeme uzavřít $G = H$. $\square$

Definice 2.29. Krátká hra G je v kanonické formě (tvaru), jestliže pro libovolnou dílčí pozici H hry G (včetně samotné G), nemá H žádné dominované možnosti a žádné reverzibilní možnosti.

Věta 2.30. *Nechť G je krátká hra. Existuje krátká hra K v kanonické formě taková, že $G = K$.*

Důkaz. Uvažujeme hru $G = \{G^L \mid G^R\}$. Indukcí můžeme předpokládat, že všechny dílčí pozice G jsou v kanonickém tvaru. Pro uvedení G do kanonické formy máme následující metodu:

- 1) Nahradíme všechny reverzibilní možnosti G^L příslušnými G^{LRL} . Totéž provedeme pro reverzibilní možnosti G^R .
- 2) Odstraníme všechny dominantní možnosti.
- 3) Pokud výsledná hra nemá žádné reverzibilní možnosti, STOP. V opačném případě se vrátíme ke kroku 1).

Výše uvedený postup musí skončit v konečném počtu kroků, protože G je krátká. Definujte výslednou hru jako G' . Pak $G' = G$ a G' je v kanonickém tvaru. $\square$

Definice 2.31. Nechť G, H jsou hry. Říkáme, že $G \cong H$ tehdy a jen tehdy, když jejich herní stromy jsou identické (tj. jsou si rovny v $\tilde{\mathbb{G}}$).²

Věta 2.32. *Jestliže $H = K$ a H, K jsou v kanonické formě, pak $H \cong K$.*

Důkaz. Nechť H^L je levá možnost H . Pak $H^L - K \triangleleft 0$. Jestliže má pravá strana vítězný tah na H^L , pak $H^{LR} - K \leq 0$. To však znamená, že $H^{LR} \leq K = H$, tedy H^L je reverzibilní přes H^{LR} . To je v rozporu s naším předpokladem, že H je kanonická. Z toho vyplývá, že vítězný tah pRavého musí mít tvar $H^L - K^L \leq 0$. Z toho vyplývá, že $H^L \leq K^L$. K^L na chvíli uvažujeme pevně.

Předpokládejme, že se pRavý přesune nejprve do $H - K^L$. Pak má levá strana vítězný tah. Pokud je to na $H - K^{LR}$, pak $H - K^{LR} \geq 0$, tzn. že $K \geq K^{LR}$. Z toho vyplývá, že K^L je reverzibilní prostřednictvím K^L , což odporuje našemu předpokladu, že K je kanonický. Z toho vyplývá, že Levého vítězný tah musí mít tvar $H^{L'} - K^L$. Z toho vyplývá, že $H^L \leq K^L \leq H^{L'}$. Ale H nemá žádné dominantní tahy. Z toho vyplývá, že $H^L = K^L = H^{L'}$.

Pro všechny H^L , existuje K^L taková, že $H^L = K^L$ a naopak. Totéž můžeme udělat i s pravými možnostmi.

Dále ukážeme, že $H \cong K$. Z definice kanonického tvaru je každá dílčí pozice H a K kanonická. Konkrétně pro každou H^L a K^L takové, že $H^L = K^L$, máme $H^L \cong K^L$ indukci. Z toho vyplývá, že $H \cong K$. $\square$

²Stromy a grafy jsou studovány v [12].

2.5. Jednoduchost

Našimi hlavními nástroji pro studium HACKENSTRINGOVÝCH HER jsou věta o jednoduchosti a pravidlo jednoduchosti. Ukážeme, že každá pozice hry BÍLOČERNÝ HACKENSTRING je číslo (definice 2.33). Podle věty o jednoduchosti lze každou pozici hry BÍLOČERNÝ HACKENSTRING identifikovat jako určité dyadické racionální číslo. Pravidlo jednoduchosti nám dává způsob, jak toto dyadické racionální číslo nalézt. Všimněte si, že korespondence mezi čísly a dyadickými racionálními čísly respektuje částečné uspořádání. Zejména sčítání čísel (definice 2.33) respektuje běžnou aritmetiku. Proto například pokud lze pozici hry BÍLOČERNÝ HACKENSTRING $\circ \bullet$ identifikovat jako $\frac{1}{2}$, pak její výsledná třída je $\mathcal{L}$, tj. kladná.

Definice 2.33. Říkáme, že G je číslo, jestliže $H^L < H^R$ platí pro všechny dílčí pozice H v G (včetně samotného G).

Lemma 2.34. Je-li G číslo, pak $G^L < G < G^R$ pro všechny levé a pravé možnosti G .

Důkaz. Necht G^L je levá možnost G . Ukážeme, že v rozdílu $G - G^L$ zvítězí Levý. Levá strana může vyhrát, když bude začínat, přesunem do $G^L - G^L$.

Předpokládejme, že pravá strana táhne jako první. Pokud pravá strana táhne na G , máme $G^R - G^L > 0$ podle definice čísla. Z toho vyplývá, že levá strana zvítězí. V opačném případě pokračuje pravý na $-G^L$. Výsledná hra je $G + (-G^L)$. Levá strana odpovídá na $G^L + (-G^L)$. Tato hra je kladné číslo a levý má vítězný tah indukci. Můžeme uzavřít, že $G^L < G$ pro všechny levé možnosti G . Podobným argumentem, $G < G^R$ platí pro všechny pravé možnosti G . $\square$

Věta 2.35. Je-li G číslo, pak $-G$ je číslo.

Důkaz. Předpokládejme, že G je číslo. Pak $G^L < G^R$ pro všechny levé a pravé možnosti G . Nerovnost můžeme přepsat jako $-G^R < -G^L$. Levé možnosti $-G$ však vypadají jako $-G^R$ a pravé možnosti $-G$ vypadají jako $-G^L$. Z toho vyplývá, že $(-G)^L < (-G)^R$ pro všechny levé a pravé možnosti G . Z toho můžeme vyvodit, že $-G$ je číslo. $\square$

Věta 2.36. Jsou-li G a H čísla, pak $G + H$ je číslo.

Důkaz. Z definice, že G a H jsou čísla, vyplývá, že $G^L < G^R$ a $H^L < H^R$. Z toho vyplývá, že $G^L + H < G^R + H$ a $G + H^L < G + H^R$. Proto stačí ukázat $G^L + H < G + H^R$ a $G + H^L < G^R + H$. První nerovnost lze přepsat jako $0 < (G - G^L) + (H^R - H)$. Tato nerovnost platí podle lematu 2.34. Druhou nerovnost lze přepsat jako $0 < (G^R - G) + (H - H^L)$. Tato nerovnost platí také podle stejného lematu 2.34. $\square$

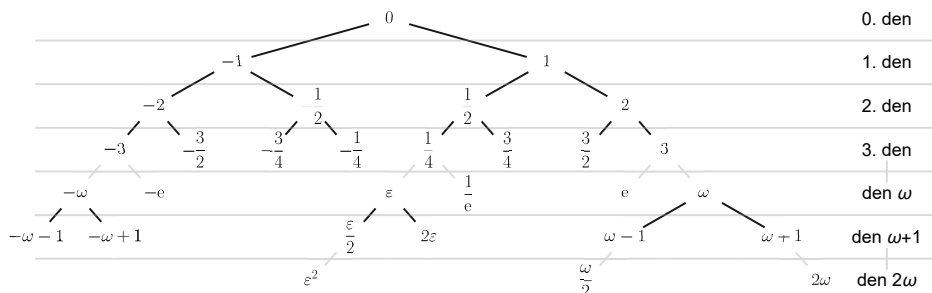
Množina čísel je neprázdná podmnožina $\mathbb{G}$ uzavřená při sčítání a opačného čísla. Můžeme tedy uzavřít, že množina čísel tvoří podgrupu $\mathbb{G}$.

Definice 2.37. Definujeme $\mathbb{Z}[\frac{1}{2}] = \{\frac{a}{2^b}; a, b \in \mathbb{Z}\}$ a tuto množinu nazýváme množinou dyadických racionálních čísel.

Všimněte si, že $(\mathbb{Z}[\frac{1}{2}], +)$ je podgrupa $(\mathbb{Q}, +)$. Naším dalším cílem je ukázat, že existuje grupový izomorfismus z podgrupy čísel do grupy dyadických racionálních čísel. Následovat bude věta o jednoduchosti a pravidlo jednoduchosti krátce poté, přičemž obě jsou pro studium hry BÍLOČERNÝ HACKENSTRING klíčové.

Uvažujme nyní jeden bílý kámen ve hře BÍLOČERNÝ HACKENSTRING $\circ$. V této hře může zahrát Levý do 0 a pravý nemá žádný pravidly povolený tah. Tedy $\circ \equiv \{0 \mid \}$. Tato hra se označí 1, protože Levý má výhodu jednoho tahu. Protože $1 > 0$, generuje podgrupu grupy všech her izomorfní se $\mathbb{Z}$. Tedy máme $2 = 1 + 1 = \{1 \mid \}$, $3 = 2 + 1 = \{2 \mid \}$, $\dots$ a obecně $n + 1 = \{n \mid \}$ a $-(n + 1) = \{ \mid -n \}$.

Další příklady mohou být $\circ \bullet = \{0 \mid 1\} = \frac{1}{2}$, $\circ \bullet \bullet = \frac{1}{2} + \frac{1}{2} = 1$, $\circ \bullet \bullet \bullet = \frac{1}{4}$ nebo $\circ \bullet \bullet \bullet \bullet = \frac{1}{32}$. Rovnost $\frac{1}{2} + \frac{1}{2} = 1$ lze jednoduše dokázat pomocí rozdílu $\frac{1}{2} + \frac{1}{2} - 1$, ve kterém vyhraje druhý hráč. Čísla s větším jmenovatelem dostaneme podobnou konstrukcí: $\frac{1}{2^{n+1}} = \{0 \mid \frac{1}{2^n}\}$ a taková čísla generují podgrupu her izomorfní s grupou $\mathbb{Z}[\frac{1}{2}]$, tj. grupou dyadických racionálních čísel: $\mathbb{Z}[\frac{1}{2}] = \{q \in \mathbb{Q}; 2^n q \in \mathbb{Z} \text{ pro nějaké } n \geq 0\}$. Kanonická forma $m/2^n$ (zkrácený tvar) je $\frac{m}{2^n} = \{\frac{m-1}{2^n} \mid \frac{m+1}{2^n}\}$. Induktivní struktura čísel se nejlépe zrcadlí na obrázku 2.5. Pro každé n je právě 2^n čísel vzniklých v den n . Například $\circ \bullet \bullet = \{0, \circ \mid \circ \bullet\} = \{0, 1 \mid 2\} = \frac{3}{2}$.



Obrázek 2.5. Geneze nadreálných čísel, viz [11] a [7].

Je-li x číslo, potom žádnému hráči se nechce v x zahrát jako první, protože ztrácí výhodu (tzv. studená hra), tedy $x^L < x < x^R$ pro každé x^L a x^R .

Začneme definicí následujících krátkých her: Připomeňme, že $0 \cong \{ \mid \}$. Dále definujeme $[1] = \{0 \mid \}$. Můžeme si představit $[1]$ jako hru, ve které má levá strana výhodu v hodnotě 1 tahu. Při hraní rozdílové hry vidíme, že $[1] = \{0 \mid [1]\} + \{0 \mid [1]\}$. Z toho vyplývá, že definujeme $[\frac{1}{2}] = \{0 \mid [1]\}$. Podobně vidíme, že $[\frac{1}{2}] = \{0 \mid [\frac{1}{2}]\} + \{0 \mid [\frac{1}{2}]\}$. Proto definujeme $[\frac{1}{4}] = \{0 \mid [\frac{1}{2}]\}$. To naznačuje obecnější definici:

Definice 2.38. Pro libovolné nezáporné celé číslo n definujeme $[\frac{1}{2^n}]$ vztahem $[\frac{1}{2^n}] = \{0 \mid [\frac{1}{2^{n-1}}]\}$.

Všimněme si, že $[\frac{1}{2^n}]$ je ostře větší než 0, a to indukci.

Lemma 2.39. $\left\lfloor \frac{1}{2^n} \right\rfloor + \left\lfloor \frac{1}{2^n} \right\rfloor = \left\lfloor \frac{1}{2^{n-1}} \right\rfloor$ pro libovolné celé číslo $n \geq 1$.

Důkaz. Ukážeme, že ve hře $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor$ vyhraje druhý hráč. Předpokládejme, že Levý táhne jako první. Pokud Levý táhne dále $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor$, pak se pravá strana nutně pohybuje na jednom z $-\left\lfloor \frac{1}{2^n} \right\rfloor$, která je záporná. Na druhou stranu levá strana může otevřít přesunem do polohy $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor = -\left\lfloor \frac{1}{2^n} \right\rfloor$, která je záporná. Z toho můžeme vyvodit, že Levý nemůže vyhrát, pokud začínal. Dále ukážeme, že pRavý nemůže vyhrát, když táhne jako první. Pokud pravá strana postupuje dále $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor$, pak výsledná hra je $\left\lfloor \frac{1}{2^{n-2}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor$. Indukcí, $\left\lfloor \frac{1}{2^{n-2}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor = \left\lfloor \frac{1}{2^{n-1}} \right\rfloor + \left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor = \left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor + \left\lfloor \frac{1}{2^{n-1}} \right\rfloor - \left\lfloor \frac{1}{2^n} \right\rfloor > 0$. Na druhou stranu pRavý může hru otevřít tím, že se přesune na $-\left\lfloor \frac{1}{2^n} \right\rfloor$. Výsledná hra je $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor + 0 - \left\lfloor \frac{1}{2^n} \right\rfloor$. Levá strana vyhrává přesunem do $\left\lfloor \frac{1}{2^{n-1}} \right\rfloor + 0 - \left\lfloor \frac{1}{2^{n-1}} \right\rfloor = 0$. Z toho vyplývá, že nemůže zvítězit, pokud táhne jako první. Z toho můžeme vyvodit požadovanou rovnost. $\square$

Lemma 2.40. Necht A, B jsou posety. Je-li $f: A \rightarrow B$ zobrazení, které respektuje částečné uspořádání, pak je f injektivní.

Důkaz. Předpokládejme, že $f(x) = f(y)$. Pak $f(x) \geq f(y)$ a $f(x) \leq f(y)$. Protože f respektuje částečné uspořádání, máme $x \geq y$ a $x \leq y$. Ekvivalentně $x = y$. To ukazuje, že f je injektivní. $\square$

Věta 2.41. Existuje injektivní grupový homomorfismus z $\mathbb{Z}\left[\frac{1}{2}\right]$ do $\mathbb{G}$.

Důkaz. Pro liché a definujeme zobrazení f pomocí $\frac{a}{2^b} \rightarrow a\left[\frac{1}{2^b}\right]$, kde $a\left[\frac{1}{2^b}\right] = \underbrace{\left[\frac{1}{2^b}\right] + \dots + \left[\frac{1}{2^b}\right]}_{a \text{ krát}}$. Necht $\frac{a}{2^b}, \frac{m}{2^n} \in \mathbb{Z}\left[\frac{1}{2}\right]$. Pak $\frac{a}{2^b} + \frac{m}{2^n}$ se zobrazí na $(a2^{n-b} + m)\left[\frac{1}{2^n}\right] = a2^{n-b}\left[\frac{1}{2^n}\right] + m\left[\frac{1}{2^n}\right] = a\left[\frac{1}{2^b}\right] + m\left[\frac{1}{2^n}\right]$. Poslední rovnost je pravdivá, protože opakovaná aplikace lematu 2.39 ukazuje $2^{n-b}\left[\frac{1}{2^n}\right] = \left[\frac{1}{2^b}\right]$. Z toho můžeme vyvodit, že f je grupový homomorfismus. Navíc toto zobrazení respektuje částečné uspořádání G . Proto je f také injektivní. $\square$

Nyní můžeme ztotožnit hru $a\left[\frac{1}{2^b}\right]$ s dyadickou racionální hrou $\frac{a}{2^b}$. Z tohoto důvodu nyní upustíme od zápisu v závorkách.

Definice 2.42. Necht $I \subset \mathbb{Z}\left[\frac{1}{2}\right]$. Pak I je interval, jestliže pro libovolné $x, y \in I$ s $x > y$ máme $z \in I$ pro všechna z taková, že $x > z > y$.

Definice 2.43. Necht G je krátká hra. Definujme narozeniny G , označené $\mathbf{b}(G)$, jako nejmenší celé číslo n takové, že $G \in \mathbb{G}_n$. Pro úplnost definujeme také formální narozeniny G , označené $\tilde{\mathbf{b}}(G)$, jako nejmenší celé číslo n takové, že $G \in \tilde{\mathbb{G}}_n$.

Všimněte si, že $\mathbf{b}(G)$ se zajímá pouze o herní hodnotu G , zatímco $\tilde{\mathbf{b}}(G)$ se zajímá o strukturu G , tj. o její možnosti.

Lemma 2.44. Necht $\frac{a}{2^b} \in \mathbb{Z}\left[\frac{1}{2}\right]$, kde a je liché. Pak $\frac{a}{2^b} = \left\{ \frac{a-1}{2^b} \mid \frac{a+1}{2^b} \right\}$ je kanonický tvar $\frac{a}{2^b}$.

Důkaz. Napíšeme $\frac{a}{2^b} = \underbrace{\frac{1}{2^b} + \frac{1}{2^b} + \cdots + \frac{1}{2^b}}_{a \text{ krát}}$. Připomeňme, že $\frac{1}{2^b} = \{0 \mid \frac{1}{2^{b-1}}\}$.

Existuje tedy přesně jedna levá možnost, a to je $0 + \frac{1}{2^b} + \cdots + \frac{1}{2^b} = \frac{a-1}{2^b}$. Podobně existuje přesně jedna pravá možnost. To je $\frac{1}{2^{b-1}} + \frac{1}{2^b} + \cdots + \frac{1}{2^b} = \frac{2}{2^b} + \frac{1}{2^b} + \cdots + \frac{1}{2^b} = \frac{a+1}{2^b}$. Můžeme dojít k závěru, že $\frac{a}{2^b} = \{\frac{a-1}{2^b} \mid \frac{a+1}{2^b}\}$. Všimněte si, že $\frac{a}{2^b}$ nemá žádné dominantní možnosti, protože by to vyžadovalo alespoň dvě levé strany nebo alespoň dvě pravé možnosti. Dále ukážeme $\frac{a}{2^b}$ nemá žádné reverzibilní možnosti. Pro přehlednost zápisu definujeme $G = \frac{a}{2^b}$. Všimněte si, že $\frac{a-1}{2^b} = \frac{a'}{2^{b'}}$ je v základním tvaru, kde a' je liché a $b' < b$. Víme, že $\frac{a'}{2^{b'}} = \left\{ \frac{a-1}{2^b} - \frac{1}{2^{b'}} \mid \frac{a-1}{2^b} + \frac{1}{2^{b'}} \right\}$. Z toho vyplývá, že $G^{LR} = \frac{a-1}{2^b} + \frac{1}{2^{b'}}$, kde $b' < b$. Proto platí $G^{LR} \geq \frac{a-1}{2^b} + \frac{1}{2^{b-1}} = \frac{a+1}{2^b} > G$. Z toho vyplývá, že G^L není reverzibilní přes G^{LR} . Z toho můžeme vyvodit, že G nemá reverzibilní levé možnosti. Podobným argumentem můžeme dokázat, že G nemá žádné reverzibilní pravé možnosti. Tím je dokázáno lemma. $\square$

Lemma 2.45. *Je-li G v kanonické formě, pak $\tilde{\mathbf{b}}(G) = \mathbf{b}(G)$.*

Důkaz. Necht $\tilde{\mathbf{b}}(G) = n$. Je jisté, že herní hodnota G se vyskytuje ve formálním tvaru nebo před ním. $\tilde{\mathbf{b}}(G) \geq \mathbf{b}(G)$. Předpokládejme sporem, že $\tilde{\mathbf{b}}(G) > \mathbf{b}(G)$. Pak existuje krátká hra H taková, že $\mathbf{b}(G) = \tilde{\mathbf{b}}(H)$. Ale kanonická forma H , nazvěme ji K , se jistě narodila v době, kdy se H narodila. Takže $\mathbf{b}(G) = \tilde{\mathbf{b}}(H) \geq \tilde{\mathbf{b}}(K)$. Ale $G \cong K$, protože kanonické formy jsou jedinečné, proto máme $\mathbf{b}(G) \geq \tilde{\mathbf{b}}(G)$. Podle našeho předpokladu to dává $\tilde{\mathbf{b}}(G) > \mathbf{b}(G) \geq \tilde{\mathbf{b}}(G)$, což je spor. Můžeme tedy uzavřít, že $\tilde{\mathbf{b}}(G) = \mathbf{b}(G)$. $\square$

Lemma 2.46. *Je-li G v kanonické formě, pak $\mathbf{b}(G^L) < \mathbf{b}(G)$ a $\mathbf{b}(G^R) < \mathbf{b}(G)$ pro všechny G^L, G^R .*

Důkaz. Necht G^L a G^R jsou možnosti v G . Máme za to, že G je v kanonické formě. Z definice kanonické formy vyplývá, že všechny možnosti G jsou kanonické. Z toho vyplývá, že $\mathbf{b}(G) = \tilde{\mathbf{b}}(G) > \tilde{\mathbf{b}}(G^L) = \mathbf{b}(G^L)$, kde prostřední nerovnost vyplývá z naší rekursivní konstrukce krátkých her. Podobný argument ukazuje, že $\mathbf{b}(G^R) < \mathbf{b}(G)$. $\square$

Lemma 2.47. *Necht $I \in \mathbb{Z}[\frac{1}{2}]$ je neprázdný interval. Pak existuje jedinečné $x \in I$ z minimálních narozenin.*

Důkaz. Předpokládejme, že $x, y \in I$ mají stejné narozeniny, řekněme n , a jsou v kanonické formě. Můžeme předpokládat, že $x > y$. Pak $x - y > 0$. Levý má tedy vyhrávající tah tvaru $x^L - y \geq 0$ nebo $x - y^R \geq 0$. V prvním případě máme $x > x^L \geq y$, z čehož vyplývá, že $x^L \in I$. V druhém případě, máme $x \geq y^R > y$, z čehož vyplývá, že $y^R \in I$. Poznamenejme, že x^L a y^R mají narozeniny ostře menší než n . Našli jsme tedy prvek s narozeninami ostře menšími než n . Opakováním tohoto postupu n -krát získáme jedinečný prvek s minimálními narozeninami. $\square$

Věta 2.48 (Věta o jednoduchosti). *Nechť G je krátká hra a definujme $I(G) = \{x \in \mathbb{Z}[\frac{1}{2}]; \forall G^L, G^R, G^L \triangleleft x \triangleleft G^R\}$. Pokud $I(G) \neq \emptyset$, pak $G = x$, kde x je jedinečný tvar, prvek v I s minimálními narozeninami.*

Důkaz. Předpokládejme, že $I(G) \neq \emptyset$, a necht x je jedinečný prvek v $I(G)$ s minimálním datem narození. Ukážeme, že $G - x = 0$. Stačí ukázat, že v $G - x$ vyhraje druhý hráč. Předpokládejme, že Levý začíná. Levý má dvě možnosti: $G^L - x$, nebo $G - x^R$. Ale $G^L - x \triangleleft 0$ podle naší definice $I(G)$, takže tento tah je pro Levého prohrávající. Na druhou stranu, $x^R \notin I(G)$, protože má ostře menší datum narození než x . Proto existuje levá možnost G taková, že $G^L \geq x^R$ nebo existuje pravá možnost G taková, že $G^R \leq x^R$. Pokud platí druhá možnost, pak má pRavý vítěznou variantu, totiž odpověď na $G - x^R$. Pokud pRavý začíná, pak $G^L \geq x^R > x$, což je v rozporu s tím, že $x \in I(G)$. Z toho můžeme vyvodit, že Levý nemůže vyhrát, když táhne jako první. Pomocí podobného argumentu můžeme usoudit že pRavý nemůže vyhrát, když táhne jako první. Z toho vyplývá, že $G = x$. $\square$

Věta 2.49 (Pravidlo jednoduchosti). *Nechť G je krátká hra a definujme $I(G) = \{x \in \mathbb{Z}[\frac{1}{2}]; \forall G^L, G^R, G^L \triangleleft x \triangleleft G^R\}$. Předpokládejme, že $I(G)$ není prázdný. Pak jedinečný prvek v $I(G)$ s minimálními narozeninami lze najít takto:*

- 1) Pokud $I(G)$ obsahuje celé číslo, pak x je celé číslo nejmenší velikosti.
- 2) V opačném případě je x jedinečný prvek minimálního jmenovatele.

Důkaz. Nejprve předpokládejme, že $I(G)$ obsahuje kladné a záporné číslo. Pak $I(G)$ obsahuje 0 a jedinečným prvkem minimálního jmenovatele je 0. Proto můžeme předpokládat, že $I(G)$ obsahuje pouze kladná čísla. Předpokládejme, že $I(G)$ obsahuje celé číslo. Napíšeme celá čísla $I(G)$ ve vzestupném pořadí jako $n_1 < n_2 < n_3 < \dots$. Máme $n_1 = \{n_1 - 1\}$. Ukážeme, že $G - n_1 = 0$. Předpokládejme, že první je Levý hráč. Má jedinou možnost: $G^L - n_1 - 1$. Protože $n_1 \in I(G)$, máme $G^L - n_1 \triangleleft 0$. pRavý si tedy může vynutit výhru. Předpokládejme, že pravá strana táhne jako první. Má dvě možnosti: $G^R - n_1$ a $G - (n_1 - 1)$. V prvním případě máme $G^R - n_1 \triangleright 0$, a tak si Levý může vynutit výhru. V druhém případě máme $n_1 - 1 \notin I(G)$. Proto můžeme najít takovou levou možnost, že $G^L \geq n_1 - 1$, nebo můžeme najít pravou variantu takovou, že $G^R \leq n_1 - 1$. Druhý případ nemůže nastat, protože by to znamenalo, že $G^R \leq n - 1 < n$. Musíme tedy mít $G^L - (n_1 - 1) \geq 0$. To ukazuje, že Levý má vítěznou odpověď na $G - (n_1 - 1)$. Z toho můžeme vyvodit závěr $G = n_1$.

Nyní předpokládejme, že $I(G)$ neobsahuje žádná celá čísla. Necht $x = \frac{a}{2^b}$ je jedinečný prvek minimálního jmenovatele. Nejprve dokážeme, že takový prvek existuje. Za tímto účelem ukážeme, že pokud existují dva různé prvky se stejným jmenovatelem v $I(G)$, pak můžeme najít prvek v $I(G)$ s ostře menším jmenovatelem. Předpokládejme, že $\frac{m}{2^n}$ a $\frac{m+k}{2^n}$ jsou prvky v $I(G)$, v základním tvaru. Je-li $k \geq 2$, pak $(m+1)/2^n \in I(G)$. Na druhé straně, je-li $k = 1$, pak buď $m/2^n$ nebo $(m+k)/2^n$ není v základním tvaru. To dokazuje tvrzení.

Nyní ukážeme $G - x = 0$. Předpokládejme, že Levý táhne první. Má dvě možnosti: $G^L - x$ a $G - x^R$. Protože $G^L \triangleleft x$, máme, že pRavý má vítěznou odpověď

na $G^L - x$. Co se týče $G - x^R$, protože x^R je číslo ostře menšího jmenovatele než x , máme $x^R \notin I$. Pak můžeme najít takovou variantu G , že $G^L \geq x^R$ nebo $G^R \leq x^R$. První z nich však implikuje $G^L \geq x^R > x$, což je spor. Takže musíme mít $G^R - x^R \leq 0$, což ukazuje, že pRavý má vítěznou odpověď na $G - x^R$. Podobně se ukazuje, že pRavý nemůže vyhrát, když táhne jako první. Z toho můžeme vyvodit, že v $G - x$ vítězí druhý hráč, tj. $G = x$. $\square$

Podle předcházející věty dostaneme: $\{-2 \mid 5\} = 0$, $\{\frac{1}{2} \mid 3\} = 1$, $\{\frac{1}{8} \mid \frac{5}{8}\} = \frac{1}{2}$ a $\{\frac{1}{2} \mid \frac{7}{8}\} = \frac{3}{4}$.

Příklad 2.50. Budeme hledat hodnotu hry $G = \{\frac{1}{4} \mid 2\}$. Prvním kandidátem hodnoty je číslo mezi levou a pravou částí G , může to být aritmetický průměr $\frac{x+y}{2}$. Ptejme se tedy, zda $G - \frac{9}{8} = 0$. Tuto hypotézu vyvrátíme: Podle definice dyadického zlomku platí $\frac{9}{8} = \{\frac{8}{8} \mid \frac{10}{8}\} = \{1 \mid \frac{5}{4}\}$. Dokázat (nebo vyvrátit) v naší teorii hypotézu znamená zahrát si hru. Jak se asi bude hrát $\{\frac{1}{4} \mid 2\} + \{-\frac{5}{4} \mid -1\}$? Začne-li pRavý hráč, asi nezahraje do hry 2, ale spíše zahraje do hry -1 . Hra se posune do postavení $\{\frac{1}{4} \mid 2\} + \{ \mid 0\}$. Levý hráč je donucen zahrát do $\frac{1}{4}$. Ve hře $\frac{1}{4} - 1 < 0$ existuje vítězná strategie pro pRavého hráče a tedy hra $G - \frac{9}{8}$ není nulová. Musíme tedy vybrat jiného kandidáta pro hodnotu hry G . Nabízí se vybrat krajní body hry $\frac{5}{4}$. Vezmeme jednodušší číslo 1 a budeme vyšetřovat, zda $G - 1 = 0$, tj. zda ve hře $G - 1$ existuje vítězná strategie pro druhého hráče. Zahrajme si tedy hru $\{\frac{1}{4} \mid 2\} + \{ \mid 0\}$. Začne-li Levý hráč, nemůže zahrát ve druhé hře a bude hrát do hry $\frac{1}{4}$. Hra se posune do pozice $\frac{1}{4} - 1 < 0$, kde existuje vítězná strategie pro pRavého hráče. Začne-li ve hře $G - 1$ naopak pRavý hráč, může zahrát do nuly (není výhodné, Levý odpoví do $\frac{1}{4}$), nebo do 2, a hra bude pokračovat hrou $2 - 1 = 1 > 0$, ve které opět existuje vítězná strategie pro Levého hráče. Tedy ve hře $G - 1$ existuje vítězná strategie pro druhého hráče. Odtud získáváme $G = 1$.

Věta 2.51. *Je-li G číslo, pak $I(G)$ není prázdné. Zejména je-li G číslo, pak G je dyadické racionální číslo.*

Důkaz. Předpokládejme, že G je číslo. Pak každá možnost G je číslo. Indukcí zjistíme, že můžeme předpokládat, že $I(G^L)$ a $I(G^R)$ jsou neprázdné pro každou možnost G . Podle jednoduchosti je každé G^L a G^R dyadické racionální číslo. Protože G je krátká, musí mít konečně možností. Proto můžeme najít maximum G^L a minimum G^R . Označme je jako $G^L = \frac{a}{2^b}$ a $G^R = \frac{c}{2^d}$, kde a, c jsou liché. Pak $x = \frac{G^L + G^R}{2}$ je dyadické racionální číslo mezi nimi. Konkrétně $G^L < x < G^R$ pro všechny možnosti G . Můžeme uzavřít, že $I(G)$ je neprázdná. $\square$

Ukázali jsme, že pokud je G krátká hra a číslo, pak je G dyadické racionální číslo. Obecně však může být $I(G)$ prázdná. V takovém případě G není číslo. Některé příklady ne-čísla jsou $\{56 \mid -56\}$, $\{0 \mid 0\}$ a $\{0 \mid \{0 \mid 0\}\}$. Poslední dvě hry se obvykle označují jako $*$ a $\uparrow$.

3. DYADICKÁ RACIONÁLNÍ ČÍSLA

Po nezbytném teoretickém stručném úvodu CGT (teorie kombinatorických her) se budeme dále věnovat především dyadickým racionálním číslům a jejich aplikaci na hru HACKENSTRING. Naše dyadická racionální čísla se také označují $\mathbb{Z}[\frac{1}{2}]$. Jedná se o obor integrity celých čísel s adjungovaným prvkem $\frac{1}{2}$. Je to nejmenší podobor integrity tělesa racionálních čísel $\mathbb{Q}$ (ve smyslu inkluze) obsahující současně všechna celá čísla a číslo $\frac{1}{2}$. Díky konstrukci dyadických čísel je jich spočetně. Z druhé strany jedná se o podmnožinu spočetné množiny všech racionálních čísel $\mathbb{Q}$, tedy $\mathbb{Z}[\frac{1}{2}]$ je spočetná.

Poznámka 3.1. Uvažujme $\mathbb{Z} \times \mathbb{N}$ a zavedeme relaci $\cong$ takto:

$$1) (a, b) \cong (a', b') \Leftrightarrow a \cdot 2^{b'} = a' \cdot 2^b.$$

Relace $\cong$ je na množině $\mathbb{Z} \times \mathbb{N}$ ekvivalence, tj. relací reflexivní, symetrickou a tranzitivní. Tato relace vytváří rozklad $(\mathbb{Z} \times \mathbb{N})/\cong$ na třídy. Typická třída rozkladu je $T_{(a,b)} = \{(a', b') \mid (a, b) \cong (a', b')\}$, druhý obor relace $\cong$ příslušný k prvku (a, b) . Na třídách rozkladu definujeme dvě binární operace:

$$2) T_{(a,b)} \oplus T_{(a',b')} = T_{(a \cdot 2^{b'} + a' \cdot 2^b, b+b')},$$

$$3) T_{(a,b)} \otimes T_{(a',b')} = T_{(aa', b+b')}.$$

Operace $\oplus, \otimes$ jsou dobře definovány, nezáleží na reprezentantech tříd. Označme $\mathbb{Z}[\frac{1}{2}] = (\mathbb{Z} \times \mathbb{N})/\cong$. Zobrazení $f: \mathbb{Z} \rightarrow (\mathbb{Z} \times \mathbb{N})/\cong$ definujeme takto: $f(a) = T_{(a,0)}$. Toto zobrazení je homomorfismus. Tradičně píšeme $T_{(a,b)} = \frac{a}{2^b}$.

$$4) \text{ Pomocí struktury } (\mathbb{Z}, <) \text{ definujeme } \prec \text{ na třídách takto: } T_{(a,b)} \prec T_{(a',b')} \Leftrightarrow a \cdot 2^{b'} < a' \cdot 2^b.$$

Strukturu $\mathbb{Z}[\frac{1}{2}] = (\mathbb{Z} \times \mathbb{N})/\cong$ nazýváme oborem dyadických racionálních čísel, podobně vznikají i p -adická čísla, speciálně desetinná (racionální) čísla.

Absolutní hodnotu získáme takto: $|T_{(a,b)}| = T_{(|a|,b)}$.

Příklad 3.2. Platí následující tvrzení: *Dyadická racionální čísla jsou hustá v reálných číslech.*

Důkaz. Předpokládejme, že $a < b$, a a b jsou reálná čísla. Díky Archimédovu axiómu existuje $n \in \mathbb{N}$ takové, že

$$0 < \frac{1}{n} < b - a,$$

což znamená, že

$$0 < \frac{1}{2^n} < \frac{1}{n} < b - a.$$

Platí

$$1 < 2^n b - 2^n a$$

a tedy existuje celé číslo m takové, že

$$2^n a < m < 2^n b,$$

tj.

$$a < \frac{m}{2^n} < b.$$

Tedy dyadická racionální čísla jsou hustá v $\mathbb{R}$ (mezi každými dvěma různými reálnými čísly leží alespoň jedno dyadické racionální číslo). $\square$

Pozorování 3.3. *Nechť $x > 0$; $x \in \mathbb{Z}[\frac{1}{2}]$, $x \notin \mathbb{Z}$ (kladné necelé dyadické racionální číslo x). Potom x se může zapsat jednoznačně ve tvaru*

$$x = 1 + 1 + \cdots + 1 - \frac{1}{2} \pm \frac{1}{4} \pm \frac{1}{8} \pm \cdots \pm \frac{1}{2^n}.$$

Důkaz. Uvažujme $\frac{a}{2^n} > 0$ v základním tvaru, tedy $a > 0$, a je liché, $n > 0$. Potom $a - 1$ a $a + 1$ jsou sudá. Jedno z nich je dokonce dělitelné čtyřmi. Bez újmy na obecnosti, necht $4|(a - 1)$. Pro $n = 1$ dostaneme $\frac{a}{2}$ a je jasné, že je pouze jeden způsob zápisu x v požadovaném tvaru.

Nechť nyní $n > 1$. Jistě platí $4 \nmid (a + 1)$. Potom

$$\frac{a + 1}{2^n} = \frac{\frac{a+1}{2}}{2^{n-1}}.$$

Podle indukční hypotézy lze to napsat jednoznačně v odpovídajícím tvaru,

$$\frac{a + 1}{2^n} = \frac{\frac{a+1}{2}}{2^{n-1}} = x = 1 + 1 + \cdots + 1 - \frac{1}{2} \pm \frac{1}{4} \pm \frac{1}{8} \pm \cdots \pm \frac{1}{2^{n-1}}.$$

Potom ale

$$x = \frac{a}{2^n} = 1 + 1 + \cdots + 1 - \frac{1}{2} \pm \frac{1}{4} \pm \frac{1}{8} \pm \cdots \pm \frac{1}{2^{n-1}} - \frac{1}{2^n}.$$

Na konci musíme odečíst, protože po předání bychom dostali $\frac{a-1}{2^n}$ jako výraz s $\frac{1}{2^{n-1}}$, což by bylo v rozporu s tvrzením $4|(a - 1)$. Z konstrukce je zřejmé, že vyjádření je jednoznačné. $\square$

Nyní můžeme pro každé dyadické racionální číslo definovat odpovídající hru HACKENSTRING. Pro každé $x \in \mathbb{Z}[\frac{1}{2}]$. Pro $x > 0$ nasčítáme 1, které odpovídají počátečnímu segmentu hry HACKENSTRING, první čeRný kámen označíme $-\frac{1}{2}$ a dále píšeme $\frac{1}{2^i}$ se znaménkem $-$ a $+$, které odpovídají následujícím čeRným a bíLým kamenům. Analogicky postupujeme i pro $x < 0$.

Příklad 3.4. $2\frac{3}{4} = 2 + \frac{3}{4} = 1 + 1 + 1 - \frac{1}{2} + \frac{1}{4}$, odpovídající hra HACKENSTRING je $\circ\circ\circ\bullet\circ$.

Jiné metody jsou uvedeny na straně 59 a 58.

Pozorování 3.5. *Zobrazení $x \rightarrow [x]$ zachovává uspořádání.*

Zlomky můžeme také psát ve tvaru

$$\left[\frac{a}{2^n} \right] = \left\{ \left[\frac{a-1}{2^n} \right] \mid \left[\frac{a+1}{2^n} \right] \right\}.$$

To je také nová definice této hry a navíc v kanonickém tvaru.

Věta 3.6. *Pro $x, y \in \mathbb{Z}[\frac{1}{2}]$ platí $[x + y] = [x] + [y]$.*

Důkaz. Případ 1 pro $x, y \in \mathbb{Z}$ jsme již ověřili.

Případ 2. Uvažujme $x \in \mathbb{Z}$, $y \notin \mathbb{Z}$. Bez újmy na obecnosti předpokládejme, že x je kladné (pro $x = 0$ je tvrzení triviální). Potom:

$$\begin{aligned} [x] + [y] &= \{[x-1] \mid\} + \left\{ \left[\frac{a-1}{2^n} \right] \mid \left[\frac{a+1}{2^n} \right] \right\} \\ &= \left\{ [x-1] + [y], [x] + \left[\frac{a-1}{2^n} \right] \mid [x] + \left[\frac{a+1}{2^n} \right] \right\} \\ &= \left\{ [x+y-1], \left[x+y - \frac{1}{2^n} \right] \mid \left[x+y + \frac{1}{2^n} \right] \right\} \\ &= \left\{ \left[x+y - \frac{1}{2^n} \right] \mid \left[x+y + \frac{1}{2^n} \right] \right\} = [x+y]. \end{aligned}$$

Podobně pro $y \in \mathbb{Z}$.

Případ 3. Necht x, y nejsou celá čísla. Označme $x = \frac{a}{2^n}$ a $y = \frac{b}{2^m}$ a nejdříve předpokládejme, že $n > m$. Pak

$$\begin{aligned} [x] + [y] &= \left\{ \left[x - \frac{1}{2^n} \right] + [y], [x] + \left[y - \frac{1}{2^m} \right] \mid \dots \right\} \\ &= \left\{ \left[x+y - \frac{1}{2^n} \right], \left[x+y - \frac{1}{2^m} \right] \mid \dots \right\} \\ &= \left\{ \left[x+y - \frac{1}{2^n} \right] \mid \left[x+y + \frac{1}{2^n} \right] \right\} \\ &= [x+y]. \end{aligned}$$

Konečně uvažme také případ $n = m$. Ukážeme, že platí

$$[x] + [y] = \left\{ \left[x+y - \frac{1}{2^n} \right] \mid \left[x+y + \frac{1}{2^n} \right] \right\}.$$

Tento případ vyžaduje více kontrol, protože toto není kanonický tvar $x+y$. Číslo $x+y$ má menší jmenovatel než 2^n . Víme, že

$$z = \left\{ \left[x+y - \frac{1}{2^n} \right] \mid \left[x+y + \frac{1}{2^n} \right] \right\}$$

je číslo, které je nejjednodušší a leží mezi (levou a pravou možností). Číslo $x+y$ samozřejmě splňuje vhodné nerovnosti, ale předpokládejme, že z je nyní opravdu nejjednodušší. Potom z získáme pro výraz $x+y$, kde $|z - (x+z)| \geq \frac{1}{2^k}$, kde $\frac{1}{2^k}$ je jmenovatelem $x+y$ a tedy $\frac{1}{2^k} \geq \frac{1}{2^{n-1}}$. Všimněte si, že z a $x+y$ se snaží ležet v otevřeném intervalu, jehož délka je přesně $\frac{1}{2^{n-1}}$, což není možné. Proto $z = x+y$. $\square$

Závěr 3.7. Vytvořili jsme izomorfismus uspořádané Abelovy grupy $\mathbb{Z}[\frac{1}{2}]$ do čísel.

3.1. Dyadická reprezentace racionálních čísel

Nechť a je nenulové přirozené číslo. Díky algoritmu opakovaného dělení se zbytkem lze a reprezentovat ve tvaru

$$a = \sum_{j=0}^l q_j 2^j, \quad (3.1)$$

s přirozenými čísly $0 \leq q_j \leq 1$ pro $j = 0, \dots, l$ a $q_l \neq 0$. Pro takový součet (3.1) zavádíme dyadickou (dvojkovou) reprezentaci

$$a = q_l q_{l-1} \dots q_1 q_0.$$

Dyadický zápis lze jednoduše rozšířit pro celá čísla. Je-li celé číslo a záporné, potom $a = -|a|$. Tedy dyadická reprezentace celého čísla $|a|$ bude dyadická reprezentace a tvaru

$$a = -q_l q_{l-1} \dots q_1 q_0.$$

Naše pozorování nyní rozšíříme na všechna racionální čísla a jejich dyadickou reprezentaci (tvar). Takže budeme předpokládat, že $\frac{a}{b}$ reprezentuje racionální číslo; $a, b \in \mathbb{Z}$ a $b \neq 0$. Bez ohledu na obecnost budeme dále předpokládat, že $b > 0$. Použijeme-li algoritmus dělení se zbytkem na celá čísla, pro a, b najdeme celá čísla q, r taková, že $0 \leq r < b$ tak, že

$$a = q \cdot b + r \Leftrightarrow \frac{a}{b} = q + \frac{r}{b}.$$

Pro kladné celé číslo q získáme dyadickou reprezentaci

$$q = \sum_{j=0}^l q_j 2^j = q_l q_{l-1} \dots q_1 q_0$$

a pro záporné $-q_l q_{l-1} \dots q_1 q_0$. Tedy

$$q = \pm q_l q_{l-1} \dots q_1 q_0.$$

Nyní předpokládejme dyadickou reprezentaci racionálního čísla $0 \leq \frac{r}{b} < 1$. Také můžeme zapsat jako

$$\frac{r}{b} = \frac{1}{2} \cdot \frac{2 \cdot r}{b}. \quad (3.2)$$

Zaměříme se na podíl $\frac{2 \cdot r}{b}$ se zbytkem. Získáme přirozená čísla q_{-1}, r_{-1} , taková, že $0 \leq r_{-1} < b$ a platí

$$2 \cdot r = q_{-1} \cdot b + r_{-1} \Leftrightarrow \frac{2 \cdot r}{b} = q_{-1} + \frac{r_{-1}}{b}. \quad (3.3)$$

Z nerovnosti $\frac{r}{b} < 1$ získáme:

$$0 \leq q_{-1} = \frac{2 \cdot r}{b} - \frac{r_{-1}}{b} < \frac{2 \cdot r}{b} < 2,$$

tak, že $0 \leq q_{-1} \leq 1$. Dosadíme-li (3.3) do (3.2), dostaneme

$$\frac{r}{b} = \frac{1}{2} \cdot \frac{2 \cdot r}{b} = \frac{1}{2} \left(q_{-1} + \frac{r_{-1}}{b} \right) = \frac{q_{-1}}{2} + \frac{1}{2} \cdot \frac{r_{-1}}{b} = \frac{q_{-1}}{2} + \frac{1}{2^2} \cdot \frac{2 \cdot r_{-1}}{b}.$$

Je-li $r_{-1} \neq 0$, vydělíme (opět) $2 \cdot r_{-1}$ se zbytkem b a získáme přirozená čísla q_{-2}, r_{-2} , taková, že $0 \leq r_{-2} < b$ taková, že

$$2 \cdot r_{-1} = q_{-2} \cdot b + r_{-2} \Leftrightarrow \frac{2 \cdot r_{-1}}{b} = q_{-2} + \frac{r_{-2}}{b}.$$

Tak jako dříve, $0 \leq q_{-2} \leq 1$ a spojíme-li naše výsledky, dostaneme

$$\frac{r}{b} = \frac{q_{-1}}{2} + \frac{1}{2^2} \cdot \frac{2 \cdot r_{-1}}{b} = \frac{q_{-1}}{2} + \frac{1}{2^2} \left(q_{-2} + \frac{r_{-2}}{b} \right) = \frac{q_{-1}}{2} + \frac{q_{-2}}{2^2} + \frac{1}{2^3} \cdot \frac{2 \cdot r_{-2}}{b}.$$

Budeme-li postupovat dál, získáme přirozená čísla q_{-3}, r_{-3} , taková, že $0 \leq q_{-3} \leq 1$ a $0 \leq r_{-3} < b$ a platí

$$\frac{r}{b} = \frac{q_{-1}}{2} + \frac{q_{-2}}{2^2} + \frac{q_{-3}}{2^3} + \frac{1}{2^4} \cdot \frac{2 \cdot r_{-3}}{b}.$$

Po k krocích získáme přirozená čísla q_{-k}, r_{-k} taková, že $0 \leq q_{-k} \leq 1$ a $0 \leq r_{-k} < b$ a platí

$$\frac{r}{b} = \sum_{j=1}^k \frac{q_{-j}}{2^j} + \frac{1}{2^{k+1}} \cdot \frac{2 \cdot r_{-k}}{b}.$$

Tímto postupem získáme dvě možnosti. Buď existuje $k \in \mathbb{N}$, $k > 0$ tak, že $r_{-k} = 0$, nebo zbytek r_{-j} je vždy nenulový pro všechna $j = 1, 2, 3, \dots$

Definice 3.8. S označením viz výše definujeme pro $a, b \in \mathbb{Z}$ a $b \neq 0$ dyadickou reprezentaci následovně:

- 1) Je-li $r = 0$ nebo existuje $k \in \mathbb{N}$, $k > 0$ a $r_{-k} = 0$ položíme

$$\pm q_l \dots q_0, q_{-1} \dots q_{-k} := \pm \sum_{j=-l}^k \frac{q_{-j}}{2^j}$$

a $\pm q_l \dots q_0, q_{-1} \dots q_{-k}$ nazýváme dyadickou (binární, dvojkovou) reprezentaci nebo dyadickým rozvojem racionálního čísla $\frac{a}{b}$.

- 2) Pokud všechny zbytky r_{-j} jsou nenulové, položíme formálně

$$\pm q_l \dots q_0, q_{-1} \dots q_{-k} \dots := \pm \sum_{j=-l}^{\infty} \frac{q_{-j}}{2^j}$$

a $\pm q_l \dots q_0, q_{-1} \dots q_{-k} \dots$ stejně nazýváme dyadickou (binární, dvojkovou) reprezentaci nebo dyadickým rozvojem racionálního čísla $\frac{a}{b}$.

Definice 3.9. Nekonečný dyadický rozvoj $\pm q_l \dots q_0, q_{-1} \dots q_{-k} \dots$ nazýváme periodickým, pokud existují přirozená čísla $v \geq 0$ a $p > 0$ taková, že $q_{-(v+j)} = q_{-(v+j+p)} = q_{-(v+j+2p)} = \dots$ pro $j = 1, 2, \dots, p$. Píšeme zjednodušeně pruh nahore pro periodu:

$$\pm q_l \dots q_0, q_{-1} \dots q_{-v} \overline{q_{-(v+1)} \dots q_{-(v+p)}}.$$

Nejmenší takové p se nazývá perioda dyadického rozvoje racionálního čísla $\frac{a}{b}$.

Nekonečná dyadická reprezentace racionálního čísla je periodická.

Věta 3.10. Necht $a, b \in \mathbb{Z}$, $b \neq 0$. Je-li dyadický rozvoj racionálního čísla $\frac{a}{b}$ nekončící, potom je periodický.

Důkaz. Předpokládejme, že racionální číslo $\frac{a}{b}$ nemá konečný dyadický rozvoj. Podle konstrukce dyadického rozvoje čísla $\frac{a}{b}$ získáváme nekonečnou posloupnost zbytků $r_0 := r, r_{-1}, r_{-2}, r_{-3}, \dots$, ovšem všechny jsou podmnožinou $\{0, \dots, b-1\}$. Odtud jistě musí existovat dva zbytky r_{-j_1} a r_{-j_2} tak, že jsou identické. Bez újmy na obecnosti můžeme předpokládat, že $j_2 > j_1 \geq 0$ a pro pevně zvolené j_1 vybereme minimální rozdíl $p := j_2 - j_1$. Díky algoritmu získání dyadického rozvoje čísla $\frac{a}{b}$, získáme

$$\begin{aligned} r_{-j_1} &= r_{-(j_1+p)} = r_{-(j_1+2p)} = \dots, \\ r_{-(j_1+1)} &= r_{-(j_1+1+p)} = r_{-(j_1+1+2p)} = \dots, \\ &\vdots \\ r_{-(j_1+p-1)} &= r_{-(j_1+2p-1)} = r_{-(j_1+3p-1)} = \dots. \end{aligned}$$

Díky tomu, že jsme vybrali j_1 minimální, dostáváme tvrzení věty. $\square$

3.2. Další vlastnosti dyadických racionálních čísel

V dalším kroku prozkoumáme hru

$$G = \circ \bullet \bullet \bullet.$$

Když napíšeme možnosti hry G , dostaneme $G = \{0 \mid 1/2, 1\}$. Podle pravidla o zjednodušení můžeme odstranit 1, takže $G = \{0 \mid 1/2\}$. Trocha práce by vás měla přesvědčit, že $G + G + (-1/2) = 0$ (tj. je to výhra druhého hráče). Definujme tedy $1/4 = \{0 \mid 1/2\}$.

Obecněji uvažujme následující hru HACKENSTRING:

$$\circ \bullet \bullet \bullet \bullet \dots \bullet$$

Rádi bychom označili $G_n = 1/2^n$. K tomu potřebujeme ukázat, že $G_n + G_n = G_{n-1}$ pro $n = 1, 2, 3, \dots$. Důkaz provedeme indukcí.

Důkaz. Předpokládejme, že již víme, že $G_n + G_n = G_{n-1}$ pro $n = 1, 2, \dots, t-1$. Pak pro $n = t$ máme

$$G_t = \{0 \mid G_0, G_1, G_2, \dots, G_{t-1}\}.$$

Je snadné vidět, že $G_0 > G_1 > G_2 > \dots$ (není to tak těžké, takže to necháme jako cvičení), takže hra se zjednoduší na $G_t = \{0 \mid G_{t-1}\}$. Zbývá tedy ukázat, že

$$\{0 \mid G_{t-1}\} + \{0 \mid G_{t-1}\} + (-G_{t-1}) = 0.$$

Všimněte si, že $-G_{t-1} = \{-G_{t-2} \mid 0\}$. Důkaz, že ve výše uvedené hře zvítězí druhý hráč:

- Předpokládejme, že začíná Levý hráč.
 - Pokud Levý vezme z $\{0 \mid G_{t-1}\}$ 0, pak pRavý může vzít z $\{0 \mid G_{t-1}\}$ G_{t-1} , čímž ponechá součet 0 a vyhraje.
 - Pokud Levý vezme z $(-G_{t-1})$ $(-G_{t-2})$, pak pRavý vezme z $\{0 \mid G_{t-1}\}$ G_{t-1} . Podle indukční hypotézy $G_{t-1} - G_{t-2} = -G_{t-2}$, takže hra je nyní $G_t - G_{t-2} < 0$. pRavý vyhrává.
- Předpokládejme, že začíná pRavý.

- Pokud pRavý vezme z $\{0 \mid G_{t-1}\}$ G_{t-1} , Levý vezme ze hry $\{0 \mid G_{t-1}\}$ 0, tedy ponechá 0 a vyhraje.
- Pokud pRavý vezme $(-G_{t-1})$ 0, pak Levý vezme z $\{0 \mid G_{t-1}\}$ 0. To dává $\{0 \mid G_{t-1}\} = G_t$, která je jednoznačně kladná. Levý tedy vyhrává.

Tím je důkaz dokončen. $\square$

Na závěr můžeme označit $G_n = 1/2^n$. Formálně dostaneme následující rekursivní definici:

$$1/2 = \{0 \mid 1\}, \quad 1/4 = \{0 \mid 1/2\}, \quad 1/8 = \{0 \mid 1/4\}, \quad 1/16 = \{0 \mid 1/8\}, \dots$$

Nyní máme tendenci si naivně myslet, že můžeme prostě brát průměry $\{a \mid b\} = \frac{a+b}{2}$. Bohužel to neplatí obecně, jak ukazují následující příklady.

3.2.1. Příklady.

- 1) Uvažujme hru $G = \{-5 \mid 2\}$. Pokud začíná Levý, nechává záporné číslo a vyhrává pRavý. Pokud začne pRavý, nechá kladné číslo a Levý vyhrává. Vyhrává tedy druhý hráč. To znamená, že $G = 0$.
- 2) Uvažujme $G = \{1/4 \mid 1\}$. Tvrdíme, že $G = 1/2$. Ve skutečnosti stačí ukázat, že v $G - 1/2 = G + \{-1 \mid 0\}$ vyhraje 2. hráč. To je ale snadné, důkaz přenecháme čtenáři.

I přes tento drobný neúspěch můžeme v některých případech brát průměry.

Je-li p celé číslo a m nezáporné celé číslo, pak:

$$\left\{ \frac{p}{2^m} \mid \frac{p+1}{2^m} \right\} = \frac{2p+1}{2^{m+1}}.$$

Důkaz. Nejprve uvažujme případ $2p+1 > 0$. Zapišme pravou stranu jako součet $(2p+1)$ členů, z nichž každý je roven $1/2^{m+1} = \{0 \mid 1/2^m\}$. Pokud tedy v tomto herním součtu začíná Levý, nemá jinou možnost než hrát $\{0 \mid 1/2^m\}$. Stejně tak, pokud začíná pRavý, je jeho jediným tahem z $\{0 \mid 1/2^m\}$ do $1/2^m$. Uvážíme-li tyto možnosti, dostaneme hru $\left\{ \frac{2p}{2^{m+1}} \mid \frac{2p}{2^{m+1}} + \frac{1}{2^m} \right\}$, což je přesně levá strana. Případ, kdy $2p+1 < 0$, je podobný a je ponechán na čtenáři. $\square$

3.2.2. Analýza. Půjde nám především o to, abychom zjistili, pro kterého hráče je daná hra výhodná. Koncová pozice pozice je výhodná pro druhého hráče v normální variantě hry. Začínající hráč nemůže táhnout. Množina tahů je tedy $\emptyset$, což také zapisujeme $\{\}$, kde vlevo píšeme tahy bílého a vpravo tahy černého, a označujeme 0. Je-li v jednom řádku jeden bílý kámen $\circ$, černý hráč nemůže táhnout a bílý hráč má výhodu jednoho tahu. Hra je tedy výhodná pro bílého hráče. Takže možné tahy jsou $\{\circ \mid\} = 1$. Bude-li počáteční pozice tvořena dvěma bílými kameny, třeba $\circ\circ$, bílý může zahrát do 0 nebo do 1, zatímco černý hráč nemůže táhnout. Tuto situaci píšeme $\{0, 1 \mid\}$ a její hodnotu označujeme 2 (dva tahy k dobru pro bílého hráče). Indukcí pro n bílých kamenů získáme

$$n = \{0, 1, 2, \dots, (n-1) \mid\}.$$

Stejně jsme mohli postupovat pro černé kameny a získali bychom

$$-n = \{\mid 0, -1, -2, \dots, -(n-1)\}.$$

Nyní přidáme k n bílým kamenům jeden černý $\circ \bullet \dots \circ \bullet$. Možné tahy jsou $\{0, 1, 2, \dots, (n-1) \mid n\} = n - 1/2$. Přidáme k n bílým kamenům dva černé $\circ \bullet \dots \circ \bullet \bullet$, dostaneme tyto možnosti $\{0, 1, 2, \dots, (n-1) \mid n, n-1/2\} = n - 1/2 - 1/4$. Zleva můžeme vynechat všechny nevýhodné tahy a nechat tah s největší hodnotou, zprava obdobně vynecháme slabé tahy a necháme pouze nejmenší hodnotu. Tak dostaneme celkem přehledný zápis $\{n-1 \mid n-1/2\} = n - 1/2 - 1/4$. Přidáme-li vpravo jeden bílý kámen, hodnota bude $n - 1/2 - 1/4 + 1/8$, atd.

3.2.3. Čísla $1/2^n$. Ve hře HACKENSTRING můžeme pozicím přiřadit jejich dyadická racionální čísla takto: $\circ = 1$, $\circ \bullet = 1/2$, $\circ \bullet \bullet = 1/4$, $\circ \bullet \bullet \bullet = 1/8$, $\dots$, $\circ \bullet \dots \bullet = 1/2^n$. Toto přiřazení je celkem přirozené (odhalíte pravidlo?).

3.2.4. Analýza $1/2$. Dyadickému racionálnímu číslu $1/2$ jsme přiřadili hodnotu hry $\circ \bullet$. Pokud ukážeme, že $\circ \bullet + \circ \bullet + (-1) = 0$, tj. $\circ \bullet + \circ \bullet + \bullet = 0$, tedy ve hře $\circ \bullet + \circ \bullet + \bullet$ existuje vítězná strategie pro druhého hráče ($\rightsquigarrow 2$). Jak se bude asi hrát?

Začne-li pRavý hráč: Hráč nezahraje v poslední hře (tento tah může udělat i později) a zahraje v první (druhé) hře do postavení $\circ \bullet \bullet$. Ze stejného důvodu Levý hráč, hráč na tahu, nezahraje svůj tah v první hře a zahraje ve druhé hře do postavení $\circ \bullet$ (odebere celý druhý řádek). V této pozici je na tahu pRavý hráč a prohraje. Tedy $R \rightsquigarrow L$.

Začne-li Levý z postavení $\circ \bullet \bullet$ bude analýza podobná. Svým tahem se dostane hra do postavení $\circ \bullet \bullet$, kde hraje pRavý. Hráč jistě neodebere celou poslední hromádku (proč?), ale zahraje do postavení $\circ \bullet$ a vyhraje. Tedy ve hře $\circ \bullet \bullet \rightsquigarrow 2$. Analogicky se postupuje i v případě $\circ \bullet \bullet \bullet$ ($\rightsquigarrow 2$), atd.

Příklad 3.11. Víme, že $\frac{1}{2} + \frac{1}{2} = 1$. Uvažujme nyní hru HACKENSTRING $\circ \bullet \bullet \bullet$ a porovnejme ji s hrou $\frac{1}{2} = \circ \bullet$. Tedy budeme hrát hru $\circ \bullet \bullet \bullet + \circ \bullet$.

Zahraje-li Levý do $\circ \bullet \bullet \bullet$, rychle prohraje, protože pRavý zahraje do nulové hry $\circ \bullet$. Tedy Levý zahraje jinak a po jeho tahu hry vypadá postavení tato: $\circ \bullet \bullet \bullet$. Výhodnější pro pRavého hráče je odebrání nejpravější kámen, tedy zahraje do $\circ \bullet \bullet$, kde Levý má méně tahů a také prohraje.

Obráceně: Začne-li pRavý, neodebere kámen z posledního řádku, což může udělat i později, ale dostane se do postavení $\circ \bullet \bullet \bullet$, kde okamžitě vyhraje Levý, protože první a třetí řádek je opačný (nulová hra) a druhý řádek začíná bílým kamenem.

Tedy hra $\circ \bullet \bullet \bullet + \circ \bullet$ je nulová a $\circ \bullet \bullet \bullet = -\circ \bullet = \circ \bullet = \frac{1}{2}$. Má-li $\circ \bullet \bullet$ číselnou hodnotu, nutně $\circ \bullet \bullet = \frac{1}{4}$.

Příklad 3.12. Ukážeme, že

$$\underbrace{\circ \bullet \bullet \dots \bullet}_n + \underbrace{\circ \bullet \bullet \dots \bullet}_n = \underbrace{\circ \bullet \bullet \dots \bullet}_{n-1}$$

přímo, indukcí a strategicky.

Je potřeba dokázat, že ve hře

$$\begin{array}{c} | \bullet \bullet \cdots \bullet + | \bullet \bullet \cdots \bullet + | \circ \circ \cdots \circ \\ \underbrace{\hspace{1.5cm}}_n \quad \underbrace{\hspace{1.5cm}}_n \quad \underbrace{\hspace{1.5cm}}_{n-1} \end{array}$$

existuje vítězná strategie pro druhého hráče.

R: Začne pRavý. Nezačne určitě v poslední hře, tento tah může udělat kdykoliv později. Zahraje svůj optimální tah v první (druhé) hře a odebere svůj nejpravější kámen. Hra je tedy v postavení

$$\begin{array}{c} | \bullet \bullet \cdots \bullet + | \bullet \bullet \cdots \bullet + | \circ \circ \cdots \circ \\ \underbrace{\hspace{1.5cm}}_{n-1} \quad \underbrace{\hspace{1.5cm}}_n \quad \underbrace{\hspace{1.5cm}}_{n-1} \end{array}$$

Zde první a poslední hra je opačnou ($\rightsquigarrow 2$) a tedy Levý zahraje v druhé hře a vyhraje ($R \rightsquigarrow L$).

L: Levý hráč nezahraje v prvních dvou hrách (zde jistě vyhraje) a své tahy může udělat kdykoliv později (naopak, zahraje-li, prohraje). Zahraje tedy v poslední hře. Zde zahraje svůj nejlepší tah, tj. nejpravější a hra se posune do postavení:

$$\begin{array}{c} | \bullet \bullet \cdots \bullet + | \bullet \bullet \cdots \bullet + | \circ \circ \cdots \circ \\ \underbrace{\hspace{1.5cm}}_{n-1} \quad \underbrace{\hspace{1.5cm}}_n \quad \underbrace{\hspace{1.5cm}}_{n-2} \end{array}$$

a na tahu je pRavý hráč. Tento hráč nesebere svůj první kámen v poslední hře (prohrál by) a ve hrách

$$\begin{array}{c} | \bullet \bullet \cdots \bullet + | \bullet \bullet \cdots \bullet \\ \underbrace{\hspace{1.5cm}}_{n-1} \quad \underbrace{\hspace{1.5cm}}_n \end{array}$$

si vybere svůj lepší tah.³ Zahraje do pozice

$$\begin{array}{c} | \bullet \bullet \cdots \bullet + | \bullet \bullet \cdots \bullet \\ \underbrace{\hspace{1.5cm}}_{n-1} \quad \underbrace{\hspace{1.5cm}}_{n-1} \end{array}$$

Tato pozice je díky indukčnímu předpokladu rovna hře

$$\begin{array}{c} | \bullet \bullet \cdots \bullet \\ \underbrace{\hspace{1.5cm}}_{n-2} \end{array}$$

a prohraje.

Platí tedy věta:

Věta 3.13. Pro $n \in \mathbb{N}$ platí

$$\frac{1}{2^n} + \frac{1}{2^n} = \frac{1}{2^{n-1}}.$$

³ $\begin{array}{c} | \bullet \bullet \cdots \bullet \\ \underbrace{\hspace{1.5cm}}_n \end{array} = \{ | \begin{array}{c} | \\ | \bullet \\ | \bullet \bullet \\ | \bullet \bullet \bullet \\ \vdots \\ | \bullet \bullet \cdots \bullet \end{array} \} = \{ 0 \mid 1, \frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^{n-1}} \} =$
 $\{ 0 \mid \frac{1}{2^{n-1}} \} = \frac{1}{2^n}.$

Příklad 3.14. Vezměme hru (half) $h = \{0 \mid 1\}$ je číslo mezi 0 a 1. Ukážeme, že h je $1/2$. Spočítáme

$$\begin{aligned} 1 - h &= 1 + (-h) = 1 + (-\{0 \mid 1\}) = 1 + \{-1 \mid 0\} \\ &= \{0 - h, 1 + (-1) \mid 1 + 0\} = \{-h, 0 \mid 1\} \\ &\stackrel{=}{=} \{0 \mid 1\} = h, \end{aligned}$$

h je kladná

odtud $1 - h = h$ a tedy $h = 1/2$. Rovnice $1 = h + h$ má i netriviální řešení.

3.2.5. Hráč má výhodu $1/2^n$ tahu. Pomocí Conwayova formalismu (sendvičový princip) označujeme:

$$\begin{aligned} \circ &\equiv \{0 \mid\} \equiv 1 & \circ\bullet &\equiv \{0 \mid 1/2, 1\} = \{0 \mid 1/2\} \equiv 1/4 \\ \circ\bullet &\equiv \{0 \mid 1\} \equiv 1/2 & \circ\bullet\bullet &\equiv \{0 \mid 1/4, 1/2, 1\} = \{0 \mid 1/4\} \equiv 1/8 \end{aligned}$$

atd.

V kanonickém tvaru dostaneme $\circ \equiv 1$, $\circ\bullet \equiv 1/2$, $\circ\bullet\bullet \equiv 1/4$ a $\circ\bullet\bullet\bullet \equiv 1/8$. Obecně $\underbrace{\circ\bullet\bullet\cdots\bullet}_{n \times} = 1/2^n$.

Můžeme ukázat, že platí následující pravidlo.

Věta 3.15. Platí $1 > 1/2 > 1/4 > \cdots > 1/2^n > \cdots$.

Důkaz. Případ $1 > 1/2$ přenecháme čtenáři. Ukážeme, že z platnosti vztahu $1/2^k > 1/2^{k+1}$ plyne pro $k \geq 1$ platnost $1/2^{k+1} > 1/2^{k+2}$. Jinými slovy $d_k = \{0 \mid 1/2^{k-1}\} + \{-1/2^k \mid 0\} > 0$ implikuje

$$d_{k+1} = \{0 \mid 1/2^k\} + \{-1/2^{k+1} \mid 0\} > 0,$$

pro $k \geq 1$. Pokud ve hře d_{k+1} začíná Levý hráč, zahraje ve druhé komponentě a získá $1/2^{k+1} - 1/2^{k+1} = 0$ a tedy vyhraje. Pokud začíná pRavý a zahraje ve druhé komponentě do 0, odpoví Levý v první komponentě do 0 a vyhraje. Pokud by pRavý šel svým prvním tahem v d_{k+1} do $1/2^k - 1/2^{k+2}$, Levý odpoví do $1/2^k - 1/2^{k+1}$, která je kladná ($\rightsquigarrow L$) díky indukční hypotéze. Tím je důkaz hotov. $\square$

Věta 3.16. Pro každé celé $n \geq 1$ je zlomek $1/2^n$ ekvivalentní číslu $\{0 \mid 1/2^{n-1}\}$.

Důkaz. Necht $x = 1/2^{n-1}$ a $y \equiv \{0 \mid x\}$. Díky indukci je $x = \{0 \mid 2x\}$. Protože $0 < y < x < x + y < 2x$, je x mezi y a $x + y$, ale pro x^L a x^R tato vlastnost neplatí. Podle věty o zjednodušení je také $x = \{y \mid x + y\}$, které je ekvivalentní $2y$ podle definice sčítání. $\square$

Nyní máme k dispozici všechny mocniny $1/2$. Z toho také plyne, že každé celé číslo můžeme vydělit mocninou $1/2$. Z definice je vidět, že ve všech hrách $1/2^n \rightsquigarrow L$, zatímco ve hrách $-1/2^n \rightsquigarrow R$. Ve hře HACKENSTRING získáme opačné pozice záměnou barev.

Čísla tvaru $m/2^n$ jsou definována prostřednictvím součtu her.

Definice 3.17. Pro každé celé číslo $m > 0$, číslo $m/2^n$ je

$$\underbrace{1/2^n + 1/2^n + \cdots + 1/2^n}_{m \times}.$$

Lemma 3.18. Pro každé celé číslo $n > 0$ platí $2/2^n = 1/2^{n-1}$.

Podotkněme, že $1/2 + 1/2 = 1$ a tedy toto lemma je zobecněním této rovnosti. Důkaz indukci přenecháme čtenáři.

Věta 3.19. Pro celá čísla $0 \leq m \leq n$ platí $2^m/2^n = 1/2^{n-m}$.

Důkaz. Protože $2/2^n = 1/2^{n-1}$, je $2/2^n + 2/2^n = 1/2^{n-1} + 1/2^{n-1}$ nebo $4/2^n = 2/2^{n-1}$. Aplikujeme-li předcházející větu, dostaneme: $2/2^{n-1} = 1/2^{n-2}$. Tedy $2^2/2^n = 1/2^{n-2}$. Budeme-li pokračovat naznačeným způsobem získáme tvrzení věty (indukcí). $\square$

Důsledkem předcházející věty je tvrzení, že každé dyadické racionální číslo lze zapsat ve tvaru, kde v čitateli je liché celé číslo. Zabývejme se nyní optimálními tahy.

Věta 3.20. Pro každé nezáporné celé číslo $n \geq 0$ a libovolné liché celé číslo k platí

$$\frac{k}{2^n} = \left\{ \frac{k-1}{2^n} \mid \frac{k+1}{2^n} \right\}.$$

Důkaz.

$$\frac{k}{2^n} = \underbrace{\frac{1}{2^n} + \frac{1}{2^n} + \cdots + \frac{1}{2^n}}_{k \text{ krát}} = \left\{ \frac{k-1}{2^n} + 0 \mid \frac{k-1}{2^n} + \frac{1}{2^{n-1}} \right\} = \left\{ \frac{k-1}{2^n} \mid \frac{k+1}{2^n} \right\}.$$

$\square$

Nabízí se otázka, co lze vytěžit z optimálního tahu ve hře $1/2 + 1/2 - 1$. Obecně hrajeme-li ve hrách typu

$$(2q+1)/2^n + (2q+1)/2^m \quad \text{nebo} \quad (2q+1)/2^n - (2q+1)/2^m$$

pro $n > m \geq 0$. Následující věta shrne naše pozorování.

Věta 3.21. Pro $n > m \geq 0$ optimální tah ve hrách $1/2^n + 1/2^m$ nebo $1/2^n - 1/2^m$ je začít ve hře $1/2^n$ (s vyšší mocninou dvojky).

Důkaz. Označme $s = 1/2^n + 1/2^m = \{0 \mid 1/2^{n-1}\} + \{0 \mid 1/2^{m-1}\}$.

Předpokládejme, že Levý začíná ve hře s . Pokud Levý zahraje, dostane buď $1/2^n$ nebo $1/2^m$. Lepším tahem je zahrát v první hře (hra je delší), protože $1/2^n > 1/2^m$ (podle předpokladu) a tedy vyhraje.

Nechť dále ve hře s začne pravý hráč. Zahraje-li v první hře, získá postavení $1/2^{n-1} + 1/2^m$ nebo $1/2^n + 1/2^{m-1}$. Zde lepší tah je zahrát v první hře, protože $1/2^{n-1} + 1/2^m < 1/2^n + 1/2^{m-1} \Leftrightarrow 2/2^n + 1/2^m < 1/2^n + 2/2^m \Leftrightarrow 1/2^m > 1/2^n$.

Zbytek důkazu, tj. $t = 1/2^n - 1/2^m$, kde je lepší zahrát pro $n > m \geq 0$, přenecháme čtenáři. $\square$

Důsledek 3.22. Pro $n > m \geq 0$ optimální tah ve hrách

$$(2p+1)/2^n + (2q+1)/2^m \quad \text{nebo} \quad (2p+1)/2^n - (2q+1)/2^m$$

je zahrát ve hře $(2p+1)/2^n$ (s vyšší mocninou dvojky).

Toto zobecnění a jeho důkaz přenecháme čtenáři, plyne z předcházející věty.

3.2.6. Dyadická čísla obecně. Víme, že $3/8 = \begin{smallmatrix} \circ & \bullet & \bullet & \bullet \\ \circ & \bullet & \bullet & \bullet \end{smallmatrix}$, $1/8 + 1/8 + 1/8 = 3/8$. Číslo $3/8$ lze ale zapsat i jinak.

Věta 3.23. Necht $r \geq 0$ a $n \geq 1$ jsou celá čísla. Potom dyadické racionální číslo má herní vyjádření

$$\frac{2r+1}{2^n} = \left\{ \frac{r}{2^{n-1}} \mid \frac{r+1}{2^{n-1}} \right\}.$$

Důkaz. $2r+1$ kopií $1/2^n$ je $\frac{2r+1}{2^n}$. Začne-li Levý hráč, odebere celou jednu větev a hru zanechá s hodnotou (v postavení) $\frac{2r}{2^n} = \frac{r}{2^{n-1}}$. Z druhé strany, začne-li pRavý, odebere jeden svůj nejpravější kámen a hru zanechá v postavení s hodnotou

$$\frac{2r}{2^n} + \frac{1}{2^{n-1}} = \frac{r}{2^{n-1}} + \frac{1}{2^{n-1}} = \frac{r+1}{2^{n-1}}.$$

Tedy

$$\frac{2r+1}{2^n} \equiv \left\{ \frac{r}{2^{n-1}} \mid \frac{r+1}{2^{n-1}} \right\}.$$

□

Alternativní zápis hry je

$$\circ \bullet \bullet \bullet \equiv \{\circ \bullet \bullet, \mid \circ \bullet, \circ\} = \{1/4, 0 \mid 1/2, 1\} = \{1/4 \mid 1/2\} = 3/8$$

(viz předcházející věta pro $r = 1$ a $n = 3$).

3.2.7. Algoritmy výpočtů. Nyní na chvíli zapomeneme na zápis čísel pomocí dvou závorek $\{\cdot \mid \cdot\}$ a budeme se věnovat jednořádkovému zápisu možných pozic ve hře HACKENSTRING. S minimem znalostí tak budeme schopni vypočítat hodnoty takových her HACKENSTRING pomocí dvojkové soustavy. Tato metoda je extrémně užitečná i pro analýzu některých nekonečných her, kde se nám podaří nalézt pravidlo a umožní nám sestavit herní hodnoty všech racionálních čísel. Pomocí binární reprezentace a této metody bude užitečné nejen pro vyšetřování racionálních čísel $\mathbb{Q}$, ale umožní nám počítat i některá neracionální reálná čísla (reálné hodnoty těchto her). Pomocí tohoto pravidla budeme schopni výpočtu hodnot hry HACKENSTRINGU na jednom řádku. Pro HACKENSTRING vytvoříme jednořádkový zápis pozice. Hodnoty vyjádříme pomocí zlomků nebo jako decimální hodnoty s určením opakovací sekvence (pravidla pokračování).

Nyní začneme s tím, že budeme vypočítávat hodnoty daných jednořádkových her.

Pravidlo: Pravidlo popíšeme pomocí několika kroků takto:

- 1) Mějme nějaký jednořádkový HACKENSTRING a jeho pozici. Jednotlivé kameny budeme po řadě kódovat zleva do prava pomocí písmen L a R . Každý bílý kámen zakódujeme L a černý kámen R . Například hodnotu hry $1\frac{3}{8}$ zakódujeme jako $[LLRRL]$, protože hra je v postavení $\circ\circ\bullet\bullet\circ$.
- 2) První změnu L, R označíme (desetinnou) čárkou, tj. např. $[LLRRL] = [L, RL]$ nebo jiný příklad $[RRRLRLR] = [RR, RLR]$.
- 3) Dále za (desetinnou) čárkou zaměníme každé R nulou 0 a každé L jedničkou 1 . Např. $[L, RL] = [L, 01]$ nebo jiný příklad $[RR, RLR] = [RR, 010]$.
- 4) Na konec naší posloupnosti zapíšeme ještě jednu jedničku, takže dostaneme $L, 011$ nebo $RR, 0101$.
- 5) Počet počátečních (stejnorodých) znaků odpovídá celému číslu, např. $1 + 0,011$ nebo $(-2) + 0,101$.
- 6) Před (desetinnou) čárkou máme celé číslo v desítkové soustavě a za čárkou je číslo ve dvojkové soustavě.

Tato metoda se nazývá Berlekampovo pravidlo (1974). Užitečné je také obrátit toto pravidlo, tj. k dané hodnotě sestavit konkrétní pozici hry HACKENSTRING. Tato metoda se ukáže jako zvlášť zajímavá při vyšetřování nekonečných her.

Následující pravidlo je přirozenější. Pravidlo získávání hodnot her je vcelku jednoduché (T. van Roode (2002), viz [8]). Začíná-li hra bílým kamenem, označíme ho $+1$, naopak -1 . Dále budeme předpokládat bez újmy na obecnost, že řádek začíná bílým kamenem (pro černý kámen je situace opačná). Nasčítáme hodnoty bílých kamenů od začátku bez přeskočení nějakého černého kamene (dostaneme celé kladné číslo). Bude-li následovat černý kámen, připočteme -2^{-n} , v opačném případě připočteme 2^{-n} . V našem prvním příkladě pozice $\circ\bullet\bullet\circ\bullet\bullet\circ$ má hodnotu

$$+1 - 1/2 - 1/4 + 1/8 - 1/16 - 1/32 - 1/64 + 1/128.$$

Má-li hra více řádků, sečteme hodnoty všech řádků. Pokud výsledná hodnota je kladná vyhraje bílý hráč, v opačném případě černý. Je-li hodnota nulová, první hráč nemůže zahrát dobře (bez chyb druhého) a nemůže si zajistit vítězství. Pro nulovou hodnotu existuje vítězná strategie pro druhého hráče.

Dále se budeme zabývat také „zlomky“ s nedvojkovým jmenovatelem.

Pravidlo: Metodu popíšeme v několika krocích:

- 1) Pro libovolné racionální číslo n oddělíme celou část k a desetinnou část q . Např. pro $3\frac{11}{32}$ dostaneme $k = 3$ a $q = \frac{11}{32}$.
- 2) Celé číslo prozatím necháme a desetinnou část převedeme do dvojkové soustavy. Pro zlomek q dostaneme

$$q = \sum_{i=0}^{\infty} a_i \cdot \left(\frac{1}{2}\right)^i, \quad a_i \in \{0, 1\}.$$

Např. je-li $q = \frac{11}{32}$, potom

$$q = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{16} + 1 \cdot \frac{1}{32}.$$

Posloupnost a_i zaznamenáme do závorek, tj. v našem příkladě získáme $[01011]$

- 3) K této posloupnosti přidáme celé číslo k , tj. $[k, a_1 a_2 \dots]$. V našem příkladu $3, 01011$.
- 4) Nyní zakódujeme celou část k pomocí L (pokud je $k \geq 0$) nebo R (je-li $k < 0$). V našem případě je to $n = [LLL, 01011]$.
- 5) Poslední jedničku vynecháme a dvojkovou (necelou) část opět zakódujeme, místo 0 píšeme R , místo 1 píšeme L .
- 6) Znak $,$ zaměníme za LR .
- 7) Konečně nakreslíme herní situaci ve hře HACKENSTRING, zaměníme-li L bílým kamenem a R černým, tedy

$$| \circ \circ \circ \circ \bullet \bullet \bullet \circ \circ = 3 \frac{11}{32}.$$

Ilustrujte tuto metodu ještě na příkladu $2\frac{3}{4}$.

Nyní tuto metodu použijeme pro komplikovanější racionální čísla. Budeme hledat pozici pro hru s hodnotou $1/3$. Jak už víme, pozice bude nekonečná.

Příklad 3.24. Necht $n = 1/3$, $k = 0$ a $q = 1/3$. Číslo $1/3$ vyjádříme ve dvojkové soustavě:

$$1/3 = \sum_{i=0}^{\infty} a_i \cdot \left(\frac{1}{2}\right)^i = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{16} + 0 \cdot \frac{1}{32} + 1 \cdot \frac{1}{64} + \dots$$

Vidíme, že dostáváme posloupnost $[LR\overline{RL}]$. Pro $1/5$ dostaneme $[LR\overline{RRLL}]$ nebo pro $20/7 = 2\frac{6}{7} = 2, \overline{110} = [LLL\overline{RLRL}]$ atd.

3.2.8. Rekurzivní formule. Rekurzivní formule nám pomůže spočítat hodnotu libovolného HACKENSTRINGU, viz [3]. Pro jednoduchost uvažujme $\underbrace{| \circ G}_{\text{napojení}}$, tj. na-

pojení hry G za hru $\circ$. Toto napojování budeme nyní označovat $(1 : G)$. Levý hráč na tahu v pozici $(1 : G)$ může zahrát do pozice $(1 : G^L)$, tj. svůj tah udělá někde ve hře G , nebo do hry 0 . Pravý hráč může táhnout pouze ve hře G a svým tahem se dostane do pozice $(1 : G^R)$.

Tedy hodnota hry je

$$(1 : G) = \{0, (1 : G^L) \mid (1 : G^R)\}.$$

Pro napojování $(-1 : G)$ platí analogická formule.

Spočítáme několik prvních hodnot, pro celé číslo n dostaneme:

$$\begin{array}{cccccccccccccccc} n = & \dots & -5 & -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 & \dots, \\ (1 : n) = & \dots & \frac{1}{32} & \frac{1}{16} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & 1 & 2 & 3 & 4 & 5 & 6 & \dots, \end{array}$$

tj. pro $n = 0, 1, \dots$ platí

$$(1 : n) = n + 1$$

a pro $n = 1, 2, \dots$ máme

$$(1 : -n) = \frac{1}{2^n}. \quad (3.4)$$

Hodnota hry $\underbrace{|\bullet\bullet\cdots\bullet|}_{n\times}$ je $-n$. Hodnota hry $\underbrace{|\circ\circ\cdots\circ|}_{n\times}$ je $(1 : n)$ a hodnota $\underbrace{|\circ\bullet\cdots\bullet|}_{n\times}$ je $(1 : -n)$. Ve hře $(1 : -n)$ má Levý hráč jedinou možnost odebrat první související bílý kámen $\circ$, zatímco pravý hráč může odebírat své černé kameny.

Tedy hodnota bude

$$(1 : -n) = \{0 \mid (1 : 0), (1 : -1), \dots, (1 : -(n-1))\}.$$

Indukcí předpokládejme, že $(1 : -k) = \frac{1}{2^k}$ for $k = 0, 1, \dots, n-1$, potom dostaneme:

$$(1 : -n) = \left\{0 \mid 1, \frac{1}{2}, \dots, \frac{1}{2^{n-1}}\right\} = \left\{0 \mid \frac{1}{2^{n-1}}\right\} = \frac{1}{2^n},$$

viz (3.4).

Funkce $f(x) = (1 : x)$ je po částech lineární. Pro dyadická čísla x dostaneme

$$\text{pro } n \leq x \leq n+1, \quad (1 : x) = f(x) = (f(n+1) - f(n)) \cdot (x - n) + f(n).$$

Samozřejmě situace pro $(-1 : x)$ je analogická. Několik prvních hodnot $(-1 : x)$ pro celá čísla x jsou

$$\begin{array}{ccccccccccccccccc} x = & \dots & -5 & -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 & \dots \\ -1 : x = & \dots & -6 & -5 & -4 & -3 & -2 & -1 & -\frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & -\frac{1}{16} & -\frac{1}{32} & \dots \end{array}$$

Poznamenejme, že všechny pozice HACKENSTRINGU jsou čísla.

Příklad 3.25. Nalezené zákonitosti nám pomohou efektivně vypočítávat hodnoty černobílých řad. Například určíme hodnotu $|\circ\bullet\bullet\bullet|$.

- 1) Nejdříve pomocí rekurzivní formule získáme postupně tyto hodnoty:
 - a) $|\bullet| = -1$,
 - b) $|\circ\bullet| = (1 : -1) = 1/2$,
 - c) $|\bullet\bullet| = (-1 : 1/2) = -3/4$,
 - d) $|\circ\bullet\bullet| = (1 : -3/4) = 5/8$,
 - e) $|\circ\bullet\bullet\bullet| = (1 : 5/8) = 13/8$.
- 2) Výpočet přímo, viz [6]: Idea je skákání po číselné ose, pro bílé na začátku budeme skákat po jedničce, zbylé skoky vždy zkracujeme o polovinu, pro černý kámen se vracíme. Takže hodnota $|\circ\bullet\bullet\bullet| = 1 + 1 - 1/2 + 1/4 - 1/8 = 13/8$.
- 3) Je jednoduché odvodit také $(x : 1) = \{x^L, x \mid x^R\} = \{x \mid x^R\}$ a $(x : -1) = \{x^L \mid x\}$. (Napojování shora, zprava.) Postupně dostaneme:
 - a) $|\circ| = 1$,
 - b) $|\circ\circ| = (1 : 1) = 2$,
 - c) $|\circ\bullet| = (2 : -1) = (\{1 \mid\} : -1) = \{1 \mid 2\} = 3/2$,
 - d) $|\circ\bullet\circ| = (3/2 : 1) = (\{1 \mid 2\} : 1) = \{3/2 \mid 2\} = 7/4$,
 - e) $|\circ\bullet\bullet\bullet| = (7/4 : -1) = (\{3/2 \mid 2\} : -1) = \{3/2 \mid 7/4\} = 13/8$.
- 4) (Berlekampovo pravidlo)
 - a) Okódujeme řadu kamenů tak, že místo bílého kamene $\circ$ píšeme 1 a černého kamene 0. Pro náš příklad tedy 11010.

- b) Vyznačíme první změnu cifer závorkou: $1(10)10$.
 c) Sečteme čísla před závorkou, místo závorky píšeme desetinnou čárku a další cifry reprezentují dvojkový zápis. Nakonec přidáme ještě jednu jedničku. Pro náš příklad je tedy $(1, 101)_2 = 1 + 1/2 + 1/8 = 13/8$.

Příklad 3.26. Nalezneme hodnotu HACKENSTRINGU $\circ\bullet\bullet\circ$. Nalezneme pozici HACKENSTRINGU s hodnotou $7/8$.

Řešení:

- 1) Hodnota je $1 - 1/2 - 1/4 + 1/8 = (8 - 4 - 2 + 1)/8 = 3/8$.
- 2) Hodnota je $(0, 011)_2 = 1/4 + 1/8 = 3/8$ (Berlekampovo pravidlo).
- 3) Hodnota je $\{\circ\bullet\bullet, 0 \mid \circ\bullet, \circ\} = \{1/4 \mid 1/2\} = 3/8$.

Protože $7/8 = 1 - 1/8$, je také $\bullet\circ\circ\circ$. Číslo $7/8$ je také $7 \cdot 1/8$. Jednořádková verze je:

- 1) $7/8 = 1 - 1/2 + 1/4 + 1/8 = \circ\bullet\circ\circ$,
- 2) $7/8 = (0, 111)_2 = \bullet\circ\circ\circ$ (Berlekampovo pravidlo),
- 3) $7/8 = \{3/4 \mid 1\} = \{\bullet\circ\circ \mid \circ\} = \bullet\circ\circ\circ$.

PODĚKOVÁNÍ

Autor vřele děkuje anonymnímu recenzentovi, šéfredктору Kvaternionu Miroslavu Kurešovi, Haně Rajdlové, Jakubu Šolcovi, Aleně Šolcové a Jiřímu Šremrovi za velmi užitečné komentáře a návrhy na vylepšení textu.

REFERENCE

- [1] M. Albert, R. Nowakowski, D. Wolfe: *Lessons in Play: An Introduction to Combinatorial Game Theory*, A K Peters, Ltd./CRC Press, Natick, MA, 2ed., 2019.
- [2] E. R. Berlekamp, J. H. Conway, R. K. Guy: *Winning Ways for your Mathematical Plays*; 2ed., vol. 1–4, A. K. Peters Ltd., 2001–2004.
- [3] J. Cihlář, V. Vopravil: *Hry a čísla* (On Games and Numbers), PF UJEP Ústí nad Labem, 125 str., 1983, 1995.
- [4] J. H. Conway: *On Numbers and Games*, Academic Press, 2ed., 2001.
- [5] L. R. Haff, W. J. Garner: *An Introduction to Combinatorial Game Theory*, Lulu.com, 2019.
- [6] A. Kirilov, I. Klumova, A. Sosinskij: *Syurrealnye chisla*, Kvant **11** (1979), rusky.
- [7] D. Knuth: *Nadreálná čísla*, Pokroky matematiky, fyziky a astronomie, **23** (1978), 66–76, 130–139, 187–196, 246–261 (překlad: Helena Nešetřilová).
- [8] T. van Roode: *Partizan Forms of Hackenbush Combinatorial Games*, Master's thesis, University of Calgary, Calgary, Alberta, July 2002.
- [9] D. Schleicher, M. Stoll: *An Introduction to Conway's Games and Numbers*, Moscow Math Journal 6:359 (2006).
- [10] A. Siegel: *Combinatorial Games Theory*, Graduate Studies in Mathematics, Vol. **146**, AMS 2013.
- [11] P. Stehlík, V. Vopravil: *John Horton Conway (1937–2020)*. Pokroky matematiky, fyziky a astronomie, roč. 65, č. **3** (2020), 125–148.
- [12] V. Vopravil, J. Porkert: *Hry a strategie*, Rozhledy matematicko-fyzikální, roč. **70** (1992), 52–57.

Václav Vopravil, Praha, Česká republika,
 e-mail: gs@wopravil.cz

HAUSDORFFOVA DIMENZE BLESKU

ALŽBĚTA KOČENDOVÁ

ABSTRAKT. V tomto článku je popsána teorie související s Hausdorffovou dimenzí. Hausdorffova dimenze je využívána k popisu míry členitosti fraktálu. Mezi fraktály patří spousta přírodních objektů a jevů. Jedním z takových jevů je právě blesk, na který se tento článek zaměřuje. Součástí článku je popis tvorby programu v MATLABu na detekci blesku v obraze a následný výpočet Hausdorffovy dimenze tohoto blesku.

ÚVOD

Tento článek se zabývá matematickou teorií Hausdorffovy dimenze a následnou aplikací této teorie na obrázky blesků. Hausdorffova dimenze má velký význam ve fraktální geometrii. Udává míru členitosti jednotlivých fraktálů. Mezi fraktály nepatří jen uměle vytvořené, na pohled velmi zajímavé útvary, ale i přírodní objekty, či jevy. Například blesk lze chápat také jako fraktál. Blesky jsou fascinujícím a zároveň nebezpečným přírodním jevem, který se vyskytuje po celém světě. Studium Hausdorffovy dimenze blesku může pomoci porozumět jeho struktuře a vývoji. Součástí této práce je program v MATLABu na detekci blesku v obraze a následný výpočet jeho Hausdorffovy dimenze.

1. TEORIE K HAUSDORFFOVĚ DIMENZI

V této kapitole si vysvětlíme matematické pojmy, které jsou nezbytné k definování Hausdorffovy dimenze. Zaměříme se na metriky a metrické prostory, následně se dostaneme k definicím Hausdorffovy míry a Hausdorffovy dimenze. Poté si vysvětlíme pojem fraktál. Definice a věty, které se vyskytují v této kapitole, jsou převzaty z [2, 4, 5, 6, 7].

1.1. Motivační příklad

Představme si, že máme úsečku, čtverec a krychli vytvořenou z nějaké hmoty. U všech těchto útvarů zmenšíme rozměry na $\frac{1}{2}$ původních rozměrů, viz obrázek 1.1. Podívejme se, kolik hmoty je potřeba na takto vzniklé útvary ve srovnání

2020 MSC. Primární 68U10; Sekundární 28A78.

Klíčová slova. Hausdorffova míra, Hausdorffova dimenze, fraktál, detekce, blesk, metoda box-counting, aproximace.

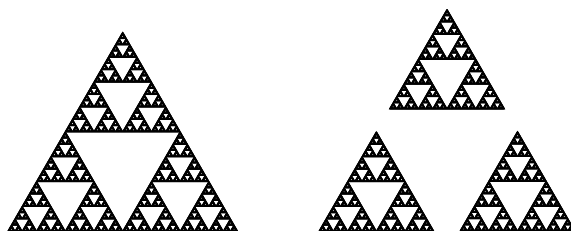
Článek vznikl na základě bakalářské práce autorky v oboru Matematické inženýrství na FSI VUT v Brně. Vedoucí práce byla Jana Hoderová z Ústavu matematiky FSI VUT v Brně.

s původními útvary. Na zmenšenou úsečku potřebujeme $\frac{1}{2}$ množství hmoty původní úsečky, u čtverce je to $\frac{1}{4}$ množství hmoty původního čtverce a u krychle $\frac{1}{8}$ množství hmoty původní krychle. Relativní množství hmoty nových útvarů vyjádříme jako mocninu koeficientu zvětšení, který je v našem případě $\frac{1}{2}$. Pro úsečku získáme $\frac{1}{2} = \left(\frac{1}{2}\right)^1$, pro čtverec $\frac{1}{4} = \left(\frac{1}{2}\right)^2$ a pro krychli $\frac{1}{8} = \left(\frac{1}{2}\right)^3$. Všimněme si, že exponent je souhlasný s dimenzí útvaru. Úsečka je jednorozměrný útvar, čtverec dvourozměrný a krychle trojrozměrná.



Obrázek 1.1. Znázornění zmenšení úsečky, čtverce a krychle

Stejný postup aplikujeme na Sierpiňského trojúhelník, který je na obrázku 1.2 vlevo. Tento trojúhelník se skládá ze tří trojúhelníků o polovičních rozměrech zob-



Obrázek 1.2. Sierpiňského trojúhelník

razených na obrázku 1.2 vpravo. Všechny tři menší trojúhelníky jsou v podstatě stejné jako celý trojúhelník, pokud nebereme ohled na měřítko. Tato vlastnost se nazývá soběpodobnost. Vidíme tedy, že zmenšení rozměrů na $\frac{1}{2}$ způsobí zmenšení potřebné hmoty na $\frac{1}{3}$ a podle předchozích úvah hledáme takové k , pro které platí

$$\left(\frac{1}{2}\right)^k = \frac{1}{3}. \quad (1.1)$$

Řešením rovnice (1.1) získáme k , které má význam fraktální dimenze Sierpiňského trojúhelníku. Tu určíme úpravami

$$\ln \left(\frac{1}{2}\right)^k = \ln \frac{1}{3}, \quad k \ln \frac{1}{2} = \ln \frac{1}{3}, \quad k = \frac{\ln 3}{\ln 2} \approx 1,58496.$$

Dimenze Sierpiňského trojúhelníku nám vyšla neceločíselná, což je právě typická vlastnost fraktálů.

1.2. Metrický prostor a metrika

Definice 1.1 (metrického prostoru a metriky). Necht $M \neq \emptyset$ je libovolná množina a $\rho: M \times M \rightarrow \langle 0, \infty \rangle$ zobrazení splňující pro všechna $x, y, z \in M$

- (m1) $\rho(x, y) = 0 \iff x = y$, (axiom totožnosti)
- (m2) $\rho(x, y) = \rho(y, x)$, (axiom symetrie)
- (m3) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$. (trojúhelníková nerovnost)

Potom zobrazení ρ nazýváme *metrika na M* , dvojici (M, ρ) nazýváme *metrický prostor*. Prvky množiny M se nazývají *body metrického prostoru* (M, ρ) a číslo $\rho(x, y)$ se nazývá *vzdálenost bodů x, y v prostoru (M, ρ)* .

Definice 1.2 (průměru množiny). Necht (M, ρ) je metrický prostor a $A \subseteq M$ je neprázdná množina. Je-li množina $\{\rho(x, y) : x, y \in A\}$ shora ohraničená, definujeme *průměr množiny A* jako

$$d(A) := \sup\{\rho(x, y) : x, y \in A\}.$$

Jestliže množina $\{\rho(x, y) : x, y \in A\}$ není shora ohraničená, tak klademe $d(A) := \infty$. Množina $A \subseteq M$ se nazývá *omezená*, jestliže $d(A) < \infty$.

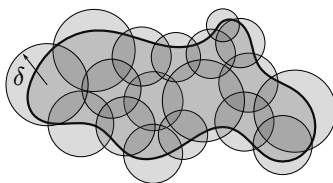
Definice 1.3 (vzdálenosti dvou množin). Necht (M, ρ) je metrický prostor a $A, B \subseteq M$ jsou neprázdné množiny. Potom definujeme *vzdálenost množin A a B* vztahem

$$\rho(A, B) := \inf\{\rho(a, b) : a \in A, b \in B\}.$$

Vzdálenost bodu x od množiny A definujeme jako

$$\rho(x, A) := \rho(\{x\}, A).$$

Definice 1.4 (δ -pokrytí množiny). Necht (M, ρ) je metrický prostor a $A \subset M$. Konečný nebo spočetný systém množin $U_i \subseteq M$ takový, že $d(U_i) \leq \delta$ pro každé i , se nazývá *δ -pokrytí množiny A* , pokud platí $A \subseteq \bigcup_i U_i$.



Obrázek 1.3. δ -pokrytí dvourozměrné množiny

1.3. Míra

Definice 1.5 (σ -algebry a měřitelného prostoru). Necht X je libovolná neprázdná množina. Systém S jejích podmnožin se nazývá *σ -algebra*, jestliže platí

- (s1) $X \in S$,
- (s2) $A \in S \implies X \setminus A \in S$,

$$(s3) \quad A_n \in S, n \in \mathbb{N} \implies \bigcup_{n=i}^{\infty} A_n \in S;$$

tedy systém S je uzavřený vzhledem k operaci doplňku a operaci spočetného sjednocení. Dvojice (X, S) se nazývá *měřitelný prostor*.

Definice 1.6 (míry na σ -algebře). Necht (X, S) je měřitelný prostor. Funkci $\mu: S \rightarrow \mathbb{R}$ nazveme *mírou na měřitelném prostoru* (X, S) , jestliže platí

$$(p1) \quad \mu(\emptyset) = 0,$$

$$(p2) \quad \mu(A) \geq 0 \text{ pro každou množinu } A \in S,$$

$$(p3) \quad A_n \in S, n \in \mathbb{N}, \text{ jsou po dvou disjunktní množiny} \implies \mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Vlastnost (p3) z definice 1.6 se nazývá σ -aditivita. Trojice (X, S, μ) se nazývá *měřitelný prostor s mírou* μ .

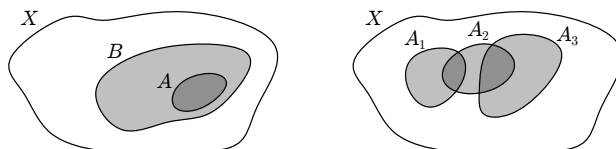
Definice 1.7 (vnější míry na množině). Necht X je neprázdná množina. Funkce $\mu^*: 2^X \rightarrow \langle 0, \infty \rangle$ se nazývá *vnější míra na X* , jestliže platí

$$(v1) \quad \mu^*(\emptyset) = 0,$$

$$(v2) \quad A, B \subseteq X, A \subseteq B \implies \mu^*(A) \leq \mu^*(B),$$

$$(v3) \quad A_n \subset X, n \in \mathbb{N} \implies \mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu^*(A_n).$$

Vlastnost (v3) z definice 1.7 se nazývá *subaditivita* vnější míry, což je slabší vlastnost než σ -aditivita míry. Na obrázku 1.4 vidíme znázornění vlastností (v2) a (v3) z definice 1.7. Po vnější míře nechceme žádné zvláštní vlastnosti. Levá část



Obrázek 1.4. Vlastnosti vnější míry

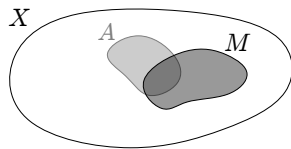
obrázku znázorňuje fakt, že libovolná podmnožina množiny bude mít menší míru než tato podmnožina. Na pravé části obrázku můžeme vidět, že sjednocení množin bude mít menší míru než součet měr těchto množin.

Definice 1.8 (μ^* -měřitelné množiny). Necht μ^* je vnější míra na X . Množina $M \subseteq X$ se nazývá μ^* -*měřitelná* právě tehdy, když pro libovolnou množinu $A \subset X$ platí

$$\mu^*(A) = \mu^*(A \cap M) + \mu^*(A \setminus M).$$

K představě této vlastnosti μ^* -měřitelné množiny M pomůže názorný obrázek 1.5.

Věta 1.9 ([6]). Necht μ^* je vnější míra na X . Systém $\mathcal{M}$ μ^* -měřitelných množin je σ -algebra a zúžení μ^* na $\mathcal{M}$ je míra.

Obrázek 1.5. μ^* -měřitelná množina M

Definice 1.10 (vnější metrické míry). Nechť (M, ρ) je metrický prostor a μ^* je vnější míra na M . Pokud pro všechny množiny $E, F \subseteq M$ splňující $\rho(E, F) > 0$ platí

$$\mu^*(E \cup F) = \mu^*(E) + \mu^*(F),$$

pak vnější míru μ^* nazýváme *vnější metrická míra*.

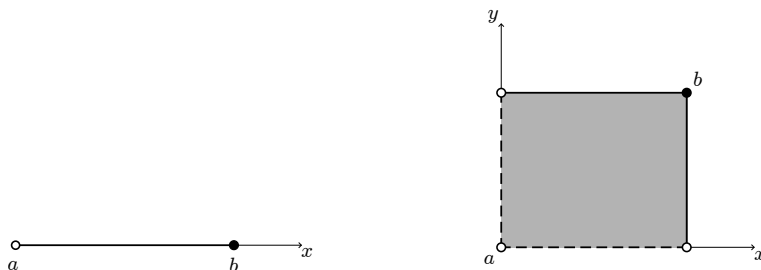
Dále se budeme pohybovat v prostoru $\mathbb{R}^d$. Body tohoto prostoru $a \in \mathbb{R}^d$ budeme zapisovat ve tvaru $a := (a_1, \dots, a_d)$.

Definice 1.11 (polouzavřeného intervalu). *Polouzavřeným intervalem* rozumíme množinu

$$(a, b) := \{x \in \mathbb{R}^d : a_j < x_j \leq b_j, j \in \{1, \dots, d\}\}.$$

Systém všech polouzavřených intervalů značíme J^d .

Na obrázku 1.6 vidíme příklady polouzavřených intervalů v prostorech $\mathbb{R}$ a $\mathbb{R}^2$. V prostoru $\mathbb{R}$ se jedná o intuitivní představu polouzavřeného intervalu neboli

Obrázek 1.6. Polouzavřený interval v $\mathbb{R}$ a $\mathbb{R}^2$

úsečky bez jednoho bodu. V prostoru $\mathbb{R}^2$ tuto představu pouze rozšíříme o jednu dimenzi a získáme tak obdélník bez dvou stran, tedy i bez třech vrcholů.

Definice 1.12 (objemu intervalu). Nechť $I \in J^d$. Potom *objem* $\lambda_d(I)$ intervalu $I := (a, b)$ definujeme jako $\lambda_d(I) = \prod_{j=1}^d (b_j - a_j)$.

Objem intervalu lze interpretovat pro jednorozměrný interval $I_1 = (a, b)$ jako délku tohoto intervalu $\lambda_1(I_1) = b - a$, pro dvourozměrný interval $I_2 = (a_1, b_1) \times (a_2, b_2)$ jako obsah obdélníku $\lambda_2(I_2) = (b_1 - a_1)(b_2 - a_2)$, pro trojrozměrný interval by se jednalo o objem kvádra a podobně bychom mohli pokračovat do vyšších dimenzí.

Definice 1.13 (vnější Lebesgueovy míry). Necht $A \subset \mathbb{R}^d$. Definujeme *vnější d -rozměrnou Lebesgueovu míru* množiny A jako

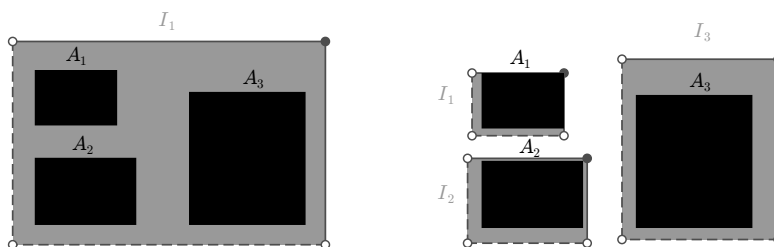
$$\lambda_d^*(A) := \inf \left\{ \sum_{n=1}^{\infty} \lambda_d(I_n) : I_n \in \mathcal{J}^d, A \subset \bigcup_{n \in \mathbb{N}} I_n \right\}.$$

Příklad 1.14. Zaměřme se nejprve na jednorozměrnou vnější Lebesgueovu míru množiny A . Nejdříve sestrojíme systém polouzavřených intervalů I_n takových, že jejich sjednocením pokryjeme celou množinu A . Poté sečteme objemy všech těchto intervalů. Pokrytí množiny polouzavřenými intervaly můžeme ovšem provést mnoha způsoby a ze všech těchto způsobů hledáme infimum součtu objemů intervalů. Na obrázku 1.7 vidíme dva možné způsoby pokrytí množiny $A = A_1 \cup A_2$ polouzavřenými intervaly I_n .



Obrázek 1.7. Pokrytí jednorozměrné množiny polouzavřenými intervaly

Analogicky si můžeme představit dvourozměrnou vnější Lebesgueovu míru. Příklad dvourozměrné množiny a její možné pokrytí vidíme na obrázku 1.8. Množina $A = A_1 \cup A_2 \cup A_3$ je složená ze tří obdélníků. V prvním případě ji pokryjeme jedním intervalem, ve druhém případě třemi intervaly, které mají zajisté menší součet objemů. Snadno si tedy představíme, že k infimu tohoto součtu dospějeme zmenšováním intervalů co nejtěsněji k jednotlivým množinám A_n .



Obrázek 1.8. Pokrytí dvourozměrné množiny polouzavřenými intervaly

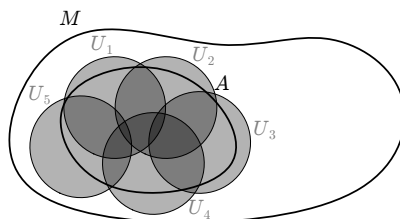
1.4. Hausdorffova míra

Věta 1.15 ([6]). Necht (M, ρ) je metrický prostor, $\delta > 0$ a $s > 0$. Potom zobrazení $H_\delta^{*s} : 2^M \rightarrow \langle 0, \infty \rangle$ definované vztahem

$$H_\delta^{*s}(A) = \inf \left\{ \sum_{i=1}^{\infty} (d(U_i))^s : \{U_i\}_{i=1}^{\infty} \text{ je } \delta\text{-pokrytí množiny } A \right\} \quad \text{pro } A \subseteq M$$

je vnější míra na množině M .

Infimum v předchozí větě bereme přes všechna možná δ -pokrytí $\{U_i\}_{i=1}^{\infty}$ množiny $A \subseteq M$, viz obrázek 1.9.



Obrázek 1.9. Pokrytí množiny A v (M, ρ)

Definice 1.16 (Hausdorffovy s -rozměrné vnější míry). Necht (M, ρ) je metrický prostor a $s > 0$. Definujeme zobrazení $H^{*s}: 2^M \rightarrow \langle 0, \infty \rangle$ vztahem

$$H^{*s}(A) = \lim_{\delta \rightarrow 0^+} H_{\delta}^{*s}(A) \quad \text{pro } A \subseteq M,$$

kde zobrazení H_{δ}^{*s} je definováno ve větě 1.15. Zobrazení H^{*s} nazveme *Hausdorffova s -rozměrná vnější míra*.

Definice 1.17. Necht (M, ρ) je metrický prostor a $s > 0$ je libovolné reálné číslo. Zúžení Hausdorffovy s -rozměrná vnější míry H^{*s} na σ -algebru všech H^{*s} -měřitelných množin je míra na množině M , kterou nazýváme *Hausdorffova s -rozměrná míra H^s* .

Věta 1.18 ([6]). Necht (M, ρ) je metrický prostor a $s > 0$ je libovolné reálné číslo. Hausdorffova s -rozměrná vnější míra H^{*s} je vnější metrickou mírou na množině M .

Věta 1.19 ([6]). Necht (M, ρ) je metrický prostor a $A \subset M$ je H^s -měřitelná množina. Potom existuje právě jedno reálné číslo $t > 0$ takové, že

- (h1) $0 < H^t(A) < \infty$,
- (h2) $H^r(A) = \infty$ pro každé $r \in \mathbb{R}$ splňující $0 < r < t$,
- (h3) $H^r(A) = 0$ pro každé $r \in \mathbb{R}$ splňující $r > t$.

Definice Hausdorffovy dimenze je založena na této vlastnosti Hausdorffovy vnější míry. Číslo t , pro které platí vlastnosti z věty 1.19, je Hausdorffova dimenze množiny $A \subset M$.

Poznámka. Hausdorffovu vnější míru na (M, ρ) lze definovat jako

$$H^{*s}(A) = \lim_{n \rightarrow \infty} H_n^{*s}(A) \quad \text{pro } A \subseteq M,$$

kde pro $n \in \mathbb{N}$ je

$$H_n^{*s}(A) = \inf \left\{ \sum_{i=1}^{\infty} (d(A_{n,i}))^s : A_{n,i} \subseteq M, d(A_{n,i}) \leq \frac{1}{n}, A \subseteq \bigcup_{i \in \mathbb{N}} A_{n,i} \right\}.$$

Když nerovnost zaměníme za rovnost, dostaneme tzv. mřížkovou vnější míru.

Definice 1.20 (mřížkové vnější míry). Necht (M, ρ) je metrický prostor a $s > 0$. Zobrazení $G^{*s} : 2^M \rightarrow \langle 0, \infty \rangle$ definované vztahem

$$G^{*s}(A) = \lim_{n \rightarrow \infty} G_n^{*s}(A) \quad \text{pro } A \subseteq M,$$

kde pro $n \in \mathbb{N}$ je

$$G_n^{*s}(A) = \inf \left\{ \sum_{i=1}^{\infty} (d(A_{n,i}))^s : A_{n,i} \subseteq M, d(A_{n,i}) = \frac{1}{n}, A \subseteq \bigcup_{i \in \mathbb{N}} A_{n,i} \right\},$$

nazýváme *mřížková vnější míra na M* .

Lze ukázat, že G^{*s} je vnější míra na M a že její zúžení na množinu všech G^{*s} -měřitelných množin je míra na M . Tuto míru značíme G^s a nazýváme ji *s-rozměrnou mřížkovou mírou*. Pro mřížkovou míru platí analogie věty 1.19. Dále platí, že je-li množina A G^s -měřitelná, pak je také H^s -měřitelná a $G^s(A) = H^s(A)$.

1.5. Fraktál

Definice 1.21 (Hausdorffovy dimenze). *Hausdorffovou dimenzí* množiny $A \subset M$, kde (M, ρ) je metrický prostor, rozumíme číslo

$$D = \sup \{s > 0 : H^{*s}(A) = \infty\}.$$

Důležitá vlastnost Hausdorffovy dimenze je, že na rozdíl od dimenze topologické nemusí být nutně celočíselná. Topologickou dimenzí geometrického útvaru U se obvykle rozumí nejmenší $n \in \mathbb{N} \cup \{0\}$ takové, že pro každé $\delta > 0$ existuje δ -pokrytí tohoto útvaru otevřenými množinami tak, že každý bod útvaru U je pokryt nejvýše $n + 1$ okolími.

Definice 1.22 (mřížkové dimenze). *Mřížkovou dimenzí* množiny $A \subset M$, kde (M, ρ) je metrický prostor, rozumíme číslo

$$D_G = \sup \{s > 0 : G^{*s}(A) = \infty\}.$$

Jestliže A je G^s -měřitelná množina, pak její mřížková dimenze je rovna Hausdorffově dimenzi.

Poznámka. *Fraktální dimenzí* rozumíme takovou dimenzi, která připouští ne-celočíselné hodnoty.

Věta 1.23. *Necht $A \subset \mathbb{R}^n$ je H^s -měřitelná množina. Potom $0 < s \leq n$. Tedy Hausdorffova dimenze množiny A může být nejvýše rovna dimenzi prostoru $\mathbb{R}^n$.*

Definice 1.24 (fraktálu). *Fraktálem* nazveme takovou množinu, která je H^s -měřitelná a jejíž Hausdorffova dimenze je ostře větší než topologická dimenze.

2. TEORIE K DETEKCI BLESKU

V této kapitole si vysvětlíme matematický aparát, který budeme následně potřebovat pro detekci blesku v obraze. Jedná se zejména o diskrétní konvoluci, díky které budeme schopni aproximovat gradient obrazu. Samotný gradient je také náplní této kapitoly a budeme ho potřebovat pro detekci hran. Definice v této kapitole je převzata z [3].

2.1. Konvoluce

Definice 2.1 (konvoluce funkcí dvou proměnných). Předpokládejme, že pro funkce $f, g: \mathbb{R}^2 \rightarrow \mathbb{C}$ integrály

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(x, y)| dx dy, \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(x, y)| dx dy$$

existují a jsou konečné. Potom *konvoluci funkcí f a g* značíme $f * g$ a definujeme jako

$$(f * g)(t, s) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) g(t - x, s - y) dx dy \quad \text{pro } (t, s) \in \mathbb{R}^2.$$

Funkce g se nazývá *konvoluční jádro*.

Definice 2.1 popisuje konvoluci dvou spojitých funkcí. Při zpracování obrazu je obvykle potřeba konvoluce diskrétních funkcí dvou proměnných, protože obraz je popsán pomocí jednotlivých pixelů v rovině. Pro obraz $f: \{0, 1, \dots, m-1\} \times \{0, 1, \dots, n-1\} \rightarrow \mathbb{R}$ o rozměrech $m \times n$ použijeme vztah

$$(f * g)(t, s) = \sum_{x=0}^{k-1} \sum_{y=0}^{\ell-1} f(t+x, s+y) g(x, y) \\ \text{pro } t \in \{0, \dots, m-k\}, s \in \{0, \dots, n-\ell\},$$

kde konvoluční jádro $g: \{0, 1, \dots, k-1\} \times \{0, 1, \dots, \ell-1\} \rightarrow \mathbb{R}$ má rozměry $k \times \ell$.

Příklad 2.2. Diskrétní funkci dvou proměnných lze reprezentovat pomocí matice. Hodnoty x, y jsou souřadnice prvku matice a funkční hodnota $f(x, y)$ je hodnota na této pozici matice. Uvedme tedy konkrétní příklad diskrétní konvoluce v maticovém tvaru

$$\begin{pmatrix} 0 & 3 & 1 & 6 & 4 & 1 & 0 \\ 5 & 1 & 8 & 8 & 1 & 3 & 5 \\ 5 & 0 & 8 & 5 & 5 & 9 & 4 \\ 2 & 0 & 4 & 4 & 5 & 2 & 8 \\ 7 & 0 & 1 & 1 & 0 & 3 & 7 \\ 4 & 1 & 1 & 0 & 7 & 4 & 8 \end{pmatrix} * \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 22 & -14 & -11 & 3 \\ 11 & 21 & -12 & 1 & 5 \\ 1 & 14 & -2 & 2 & 12 \\ -13 & 5 & 5 & 6 & 18 \end{pmatrix}.$$

zdrojový obraz konvoluční jádro výsledek konvoluce

2.2. Gradient

Gradient se využívá v oblasti zpracování obrazu pro detekci hran. Norma gradientu funkce dvou proměnných udává, jak moc se v daných místech mění funkční hodnota. Hrany se v obraze vyznačují velkými změnami hodnot, takže vysoké hodnoty normy gradientu značí velkou pravděpodobnost, že se jedná o hranu. My budeme ale potřebovat diskrétní aproximaci gradientu, protože máme obraz složený z pixelů. Jednou z takových aproximací je diskrétní konvoluce obrazu se Sobelovým

operátorem. Tak jako gradient funkce dvou proměnných má dvě složky, tak i Sobelův operátor budeme muset aplikovat ve dvou směrech pro aproximaci derivace ve směru osy x a ve směru osy y . Pro získání x -ové složky aproximace gradientu provedeme konvoluci s konvolučním jádrem

$$\mathbf{g}_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad (2.1)$$

a pro získání y -ové složky aproximace gradientu provedeme konvoluci s konvolučním jádrem

$$\mathbf{g}_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}. \quad (2.2)$$

Provedením těchto operací získáme dvě matice, jedna obsahuje hodnoty x -ové složky aproximace gradientu, tu označíme $\mathbf{G}_x$ a druhá hodnoty y -ové složky aproximace gradientu, tu označíme $\mathbf{G}_y$. Aproximovaný gradient označíme $\mathbf{G} = (\mathbf{G}_x, \mathbf{G}_y)$. Nás zajímá norma gradientu $\mathbf{G}$, takže po prvcích provedeme výpočet normy aproximovaného gradientu

$$\|\mathbf{G}\| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}. \quad (2.3)$$

Poté pro detekci hran najdeme maximum této normy a hledáme všechny pixely, které mají hodnotu dostatečně vysokou vůči tomuto maximu.

3. TEORIE K VÝPOČTU HAUSDORFFOVY DIMENZE

V této kapitole se zaměříme na teorii potřebnou k výpočtu Hausdorffovy dimenze. Vysvětlíme si metodu nejmenších čtverců, kterou budeme potřebovat k prokládání dat přímkou. Poté si popíšeme metodu box-counting, kterou použijeme pro výpočet Hausdorffovy dimenze a která využívá zmiňovanou metodu nejmenších čtverců. V této kapitole je čerpáno z [1].

3.1. Metoda nejmenších čtverců

Při výpočtu Hausdorffovy dimenze potřebujeme proložit přímkou danými body a následně zjistit její směrnici. K tomu využijeme numerickou metodu zvanou metoda nejmenších čtverců (MNČ). Tato metoda zvolí přímku tak, aby součet druhých mocnin vzdálenosti všech bodů od přímky ve směru osy y byl minimální.

3.1.1. Matematická formulace MNČ. Necht t je nezávisle proměnná a $y(t)$ je neznámá funkce proměnné t , kterou chceme aproximovat. Provedli jsme m měření a přibližně jsme tedy zjistili hodnoty y pro určité různé t , takže $y_i \approx y(t_i)$, pro $i = 1, 2, \dots, m$. Chceme modelovat y lineární kombinací n bázeových funkcí φ_j pro nějaké $n \leq m$, tedy

$$y(t) \approx x_1\varphi_1(t) + x_2\varphi_2(t) + \dots + x_n\varphi_n(t) =: R_n(t).$$

Funkce R_n se ve statistice nazývá *lineární regresní funkce*. Bázeové funkce φ_j navrhujeme podle očekávaného průběhu funkce y . Určíme parametry $x_1, x_2, \dots, x_n$

tak, aby

$$y_i \approx R_n(t_i), \quad i = 1, 2, \dots, m, \quad \text{neboli} \quad \mathbf{y} \approx \mathbf{A}\mathbf{x},$$

kde $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$ jsou naměřené hodnoty, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ je vektor neznámých parametrů a

$$\mathbf{A} = \begin{pmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \dots & \varphi_n(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) & \dots & \varphi_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(t_m) & \varphi_2(t_m) & \dots & \varphi_n(t_m) \end{pmatrix} =: (\varphi_1, \varphi_2, \dots, \varphi_n).$$

Matice $\mathbf{A}$ se nazývá *návrhová matice*. Vektor $\varphi_j = (\varphi_j(t_1), \varphi_j(t_2), \dots, \varphi_j(t_m))^T$, $j = 1, 2, \dots, n$, je j -tý sloupec matice $\mathbf{A}$. Rozdíly mezi naměřenými hodnotami y_i a aproximovanými hodnotami $R_n(t_i)$ se nazývají *rezidua* a platí pro ně

$$r_i := y_i - R_n(t_i) = y_i - \sum_{j=1}^n \varphi_j(t_i)x_j = y_i - \sum_{j=1}^n a_{ij}x_j, \quad i = 1, 2, \dots, m,$$

kde $a_{ij} = \varphi_j(t_i)$. Pro reziduum tedy platí $\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}$. Naším cílem je určit parametry x_i tak, aby rezidua byla co nejmenší. Minimalizujeme tedy součet druhých mocnin reziduí, jinými slovy

$$\|\mathbf{r}\|_2^2 = \sum_{i=1}^m r_i^2 \rightarrow \min. \quad (3.1)$$

3.1.2. Normální soustava rovnic. Řešení minimalizační úlohy (3.1) musí splňovat nutnou podmínku pro extrém

$$\frac{\partial \|\mathbf{r}\|_2^2}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^m \left(y_i - \sum_{j=1}^n a_{ij}x_j \right)^2 = 0, \quad k = 1, 2, \dots, n.$$

Provedme naznačenou derivaci:

$$\frac{\partial \|\mathbf{r}\|_2^2}{\partial x_k} = 2 \sum_{i=1}^m \left(y_i - \sum_{j=1}^n a_{ij}x_j \right) (-a_{ik}) = 0, \quad k = 1, 2, \dots, n.$$

Odtud úpravou získáme vztah

$$\sum_{i=1}^n \left(\sum_{j=1}^m a_{ik}a_{ij} \right) x_j = \sum_{i=1}^m a_{ik}y_i, \quad k = 1, 2, \dots, n,$$

který je možné zapsat jako

$$\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{y}. \quad (3.2)$$

Tato soustava lineárních rovnic (3.2) se nazývá *normální soustava rovnic*.

3.1.3. Prokládání dat přímkou. V případě výpočtu Hausdorffovy dimenze potřebujeme aproximovat data přímkou, neboli polynomem prvního stupně. Bázové funkce tedy budou pouze dvě a to $\varphi_1(t) = t$ a $\varphi_2(t) = 1$. Takže máme dva neznámé parametry. Vektor těchto parametrů označíme $(a, b)^T$, aby nedocházelo ke kolizi s rovnicí přímkou. Změříme hodnoty y_i pro t_i , kde $i = 1, 2, \dots, m$. Pro matici $\mathbf{A}$ platí

$$\mathbf{A} = \begin{pmatrix} \varphi_1(t_1) & \varphi_2(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) \\ \vdots & \vdots \\ \varphi_1(t_m) & \varphi_2(t_m) \end{pmatrix} = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix}.$$

Dosazením do vztahu (3.2) získáme

$$\begin{pmatrix} \sum_{i=1}^m t_i^2 & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & m \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m t_i y_i \\ \sum_{i=1}^m y_i \end{pmatrix}. \quad (3.3)$$

Z této soustavy dopočítáme parametry a a b . Tím získáme předpis přímkou $y = at + b$.

3.2. Metoda box-counting

K výpočtu Hausdorffovy dimenze lze použít metodu box-counting. Jedná se o metodu odvozenou přímo z definice Hausdorffovy míry. Hausdorffova míra množiny A je definována jako

$$H^s(A) = \lim_{\delta \rightarrow 0^+} H_\delta^{*s}(A), \quad (3.4)$$

kde

$$H_\delta^{*s}(A) = \inf \left\{ \sum_{i=1}^{\infty} (d(U_i))^s : \{U_i\}_{i=1}^{\infty} \text{ je } \delta\text{-pokrytí množiny } A \right\},$$

viz definici 1.17. Dále budeme uvažovat metrický prostor $(\mathbb{R}^2, \rho_\infty)$, kde ρ_∞ je maximální metrika definovaná na množině $\mathbb{R}^2$ jako

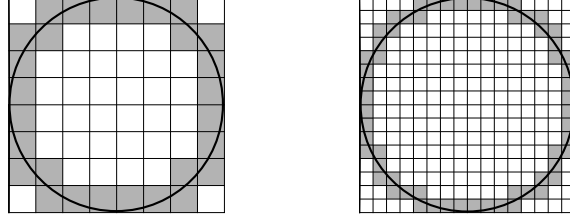
$$\rho_\infty(x, y) = \max \{|x_1 - y_1|, |x_2 - y_2|\}.$$

Jednotková koule, neboli množina bodů se vzdáleností menší nebo rovnou 1 od nějakého bodu a , by v této metrice vypadala jako čtverec. Položme $\delta = \frac{1}{k}$. Množinu A pokryjeme čtverci s průměrem δ , které mezi sebou po dvojicích mají průnik nejvýše jednu stranu. Počet čtverců obsahujících část množiny A označíme $N(k)$. Dosazením těchto parametrů do (3.4) dostaneme

$$H^s(A) = \lim_{k \rightarrow \infty} N(k) \left(\frac{1}{k} \right)^s.$$

Znázornění takových pokrytí můžeme vidět na obrázku 3.1, kde k zvolíme tak, že průměr čtverců je $\frac{1}{k}$ a potom $N(k)$ je počet šedých čtverců.

Při výpočtu Hausdorffovy dimenze z obrázku jsme omezeni počtem pixelů. Hodnotu k nemůžeme zvětšovat do nekonečna, protože tím zmenšujeme průměr čtverců



Obrázek 3.1. Pokrytí kružnice čtverci

$\frac{1}{k}$ k nule a ten nemůže být menší než průměr pixelu. Limitu tedy neuvažujeme a dostáváme pouze aproximaci

$$H^s(A) \approx N(k) \left(\frac{1}{k} \right)^s.$$

Odtud vyjádříme následujícími úpravami Hausdorffovu dimenzi s :

$$\begin{aligned} \ln H^s(A) &\approx \ln \left(N(k) \left(\frac{1}{k} \right)^s \right), \\ \ln H^s(A) &\approx \ln N(k) + s \cdot \ln \left(\frac{1}{k} \right), \\ -\ln N(k) &\approx s \cdot \ln \left(\frac{1}{k} \right) - \ln H^s(A), \\ \ln N(k) &\approx s \cdot \ln k + \ln H^s(A). \end{aligned}$$

Vztah $\ln N(k) \approx s \ln k + \ln H^s(A)$ připomíná rovnici přímky $y = at + b$ v logaritmických souřadnicích, kde $a = s$, $b = \ln H^s(A)$, $t = \ln k$, $y = \ln N(k)$. Koeficient $b = \ln H^s(A)$ je logaritmus Hausdorffovy míry množiny A , takže b je vskutku konstanta. Pro výpočet Hausdorffovy dimenze počítáme $N(k)$ pro různá k , jinými slovy počítáme y pro různá t . Tím získáváme body, které by měly přibližně ležet na přímce. Můžeme tedy použít metodu nejmenších čtverců k proložení přímky těmito body a pomocí směrnice a této přímky přibližně určit Hausdorffovu dimenzi s . Uvažujme tedy m přibližných hodnot $y_i \approx y(t_i)$. Dosazením za a , b , t a y do (3.3) dostaneme

$$\begin{pmatrix} \sum_{i=1}^m (\ln k_i)^2 & \sum_{i=1}^m \ln k_i \\ \sum_{i=1}^m \ln k_i & m \end{pmatrix} \begin{pmatrix} s \\ \ln H^s(A) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m \ln k_i \ln N(k_i) \\ \sum_{i=1}^m \ln N(k_i) \end{pmatrix}.$$

Z této soustavy dopočítáme Hausdorffovu dimenzi s . Nejdříve vynásobíme matici a tím získáme soustavu dvou rovnic o dvou neznámých s a $\ln H^s(A)$, tedy

$$s \sum_{i=1}^m (\ln k_i)^2 + \ln H^s(A) \sum_{i=1}^m \ln k_i = \sum_{i=1}^m \ln k_i \ln N(k_i), \quad (3.5)$$

$$s \sum_{i=1}^m \ln k_i + m \ln H^s(A) = \sum_{i=1}^m \ln N(k_i). \quad (3.6)$$

Z rovnice (3.6) vyjádříme

$$\ln H^s(A) = \frac{\sum_{i=1}^m \ln N(k_i) - s \sum_{i=1}^m \ln k_i}{m} \quad (3.7)$$

a dosadíme do rovnice (3.5):

$$s \sum_{i=1}^m (\ln k_i)^2 + \frac{\sum_{i=1}^m \ln N(k_i) - s \sum_{i=1}^m \ln k_i}{m} \sum_{i=1}^m \ln k_i = \sum_{i=1}^m \ln k_i \ln N(k_i).$$

Odtud už snadno vyjádříme s :

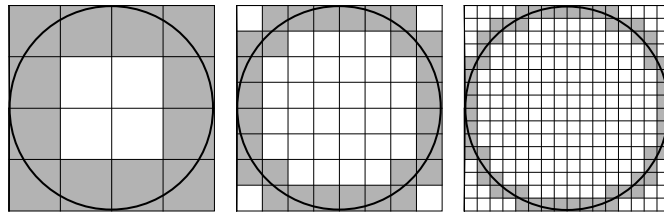
$$sm \sum_{i=1}^m (\ln k_i)^2 + \sum_{i=1}^m \ln N(k_i) \sum_{i=1}^m \ln k_i - s \left(\sum_{i=1}^m \ln k_i \right)^2 = m \sum_{i=1}^m \ln k_i \ln N(k_i),$$

$$s = \frac{m \sum_{i=1}^m \ln k_i \ln N(k_i) - \sum_{i=1}^m \ln N(k_i) \sum_{i=1}^m \ln k_i}{m \sum_{i=1}^m (\ln k_i)^2 - \left(\sum_{i=1}^m \ln k_i \right)^2}.$$

Získali jsme vztah pro Hausdorffovu dimenzi, do kterého stačí už jen dosadit různá k_i a k nim příslušná $N(k_i)$.

K implementaci této metody je potřeba zavést pojem *škálovací faktor*. Mějme při první iteraci pokrytí množiny čtverci o průměru $\frac{1}{k}$, v ostatních iteracích zmenšujeme průměr čtverců na polovinu průměru čtverců předchozí iterace. Škálovacím faktorem pro i -tou iteraci rozumíme hodnotu zmenšení průměru čtverců oproti první iteraci. První iterace má tedy škálovací faktor 1, protože je $\frac{1}{1}$ násobkem průměru $\frac{1}{k}$. Druhá iterace má škálovací faktor 2, protože je $\frac{1}{2}$ násobkem průměru $\frac{1}{k}$. Podobně i -tá iterace má škálovací faktor 2^{i-1} .

Příklad 3.1. Postup při výpočtu Hausdorffovy dimenze metodou box-counting si ukážeme na kružnici. Na obrázku 3.2 vidíme tři iterace pokrytí kružnice čtverci. Počty těchto čtverců a k nim příslušné hodnoty škálovacího faktoru jsou uvedeny



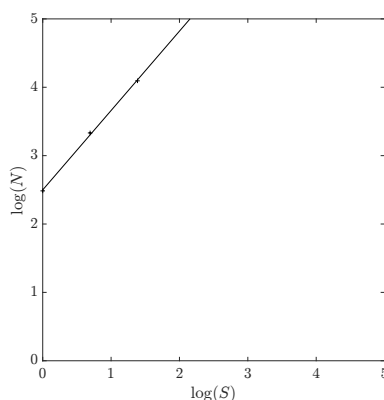
Obrázek 3.2. Pokrytí kružnice čtverci

v tabulce 3.1. Zároveň zde vidíme i přirozené logaritmy těchto hodnot, které budeme potřebovat k prokládání přímky daty. Hodnoty $\log S$ a $\log N$ znázorníme

iterace	škálovací faktor S	počet čtverců N	$\log S$	$\log N$
první	1	12	0	2,4849
druhá	2	28	0,6931	3,3322
třetí	4	60	1,3863	4,0943

Tabulka 3.1. Hodnoty škálovacího faktoru a počty pokrývajících čtverců

graficky jako body a pomocí metody nejmenších čtverců těmito body proložíme přímkou. Body a proloženou přímkou vidíme na obrázku 3.3. Směrnice této přímky



Obrázek 3.3. Proložení dat přímkou

je 1,1610. Hausdorffova dimenze kružnice nám tedy vyšla 1,1610. Pro přesnější výsledky bychom mohli provést více iterací, ale pro vysvětlení postupu nám tyto tři iterace stačí.

4. TVORBA PROGRAMU

V této kapitole si popíšeme postup při tvorbě programu na detekci blesku v obraze, jeho zjednodušení na jednopixelovou křivku a následný výpočet Hausdorffovy dimenze tohoto blesku. Nejprve převedeme obrázek na černobílý, následně najdeme v obrázku blesk a zjednodušíme ho na jednopixelovou křivku. Nakonec spočítáme Hausdorffovu dimenzi takto zjednodušeného blesku.

4.1. Příprava obrázku

Nejdříve načteme barevný obrázek o rozměrech, které si označíme $m \times n$. Tento obrázek lze tedy chápat jako matici $m \times n$, kterou si označíme $\mathbf{O}$. V MATLABu

má tato matice prvky, kterými jsou jednorozměrná pole o třech prvcích, kde jsou uloženy hodnoty jednotlivých složek RGB, což si lze představit také jako tři matice, kde každá obsahuje informace pouze o jedné ze tří složek. Obrázek následně převedeme na černobílou verzi tak, aby se co nejvíce eliminoval šum. K tomu použijeme součet poměru složek RGB. Červenou složku násobíme koeficientem 0,25, zelenou 0,5 a modrou 0,25. Označme matici černobílého obrázku $\mathbf{BW}$ a matice barevných složek $\mathbf{R}, \mathbf{G}, \mathbf{B}$. Potom platí $\mathbf{BW} = 0,25\mathbf{R} + 0,5\mathbf{G} + 0,25\mathbf{B}$.

4.2. Detekce blesku

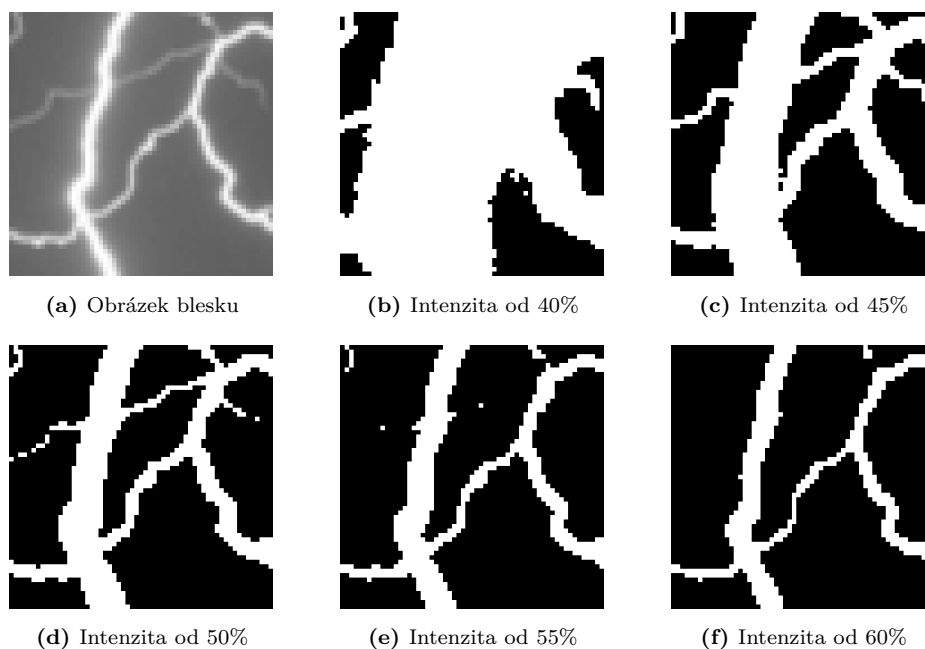
Po převodu obrázku na černobílý je nutno v něm najít blesk. Hledání blesku provedeme na základě myšlenky, že blesk by měl být nejsvětlejší část obrázku a zároveň by hrany blesku měly být velmi výrazné. Vytvoříme si tedy dvě matice. Jedna bude obsahovat informace o tom, kde se nachází pixely s dostatečně velkou intenzitou a druhá matice bude obsahovat výsledek po detekci hran. Obě tyto matice obsahují hodnoty pouze 0 a 1, kde 1 znamená, že pixel je součástí blesku a 0, že nikoliv. U takových matic lze tedy provádět logické operace. Hledané pixely blesku mají vysokou intenzitu nebo se v jejich okolí prudce mění intenzita. Matici nesoucí informaci o blesku tím pádem získáme jako logický součet těchto dvou matic.

4.2.1. Hledání dostatečné intenzity. Projdeme tedy všechny prvky matice $\mathbf{BW}$ a najdeme jejich maximální hodnotu, jinými slovy projdeme všechny pixely obrázku a najdeme hodnotu nejsvětlejšího pixelu. Tuto hodnotu si uložíme a vytvoříme si nulovou matici $\mathbf{V}$. Zvolíme si hodnotu parametru p tak, že $0 \leq p \leq 1$, který značí minimální hodnotu intenzity vůči maximu, kterou musí blesk mít. Znovu projdeme všechny prvky matice $\mathbf{BW}$ a těm, které mají hodnotu alespoň p , přiřadíme v matici $\mathbf{V}$ hodnotu 1. Ostatní pixely této matice mají hodnotu 0.

Příklad 4.1. Hledání dostatečné intenzity si předvedeme na obrázku 4.1a, který už je převeden na černobílou verzi. Najdeme maximální hodnotu mezi všemi pixely a následně volíme různé hodnoty parametru p . Hledáme pixely s hodnotou větší nebo rovnou p -násobku nalezeného maxima a těm přiřadíme hodnotu 1, neboli bílou barvu. Ostatní pixely mají hodnotu 0 a jsou černé. Výsledky lze vidět na obrázcích 4.1b až 4.1f.

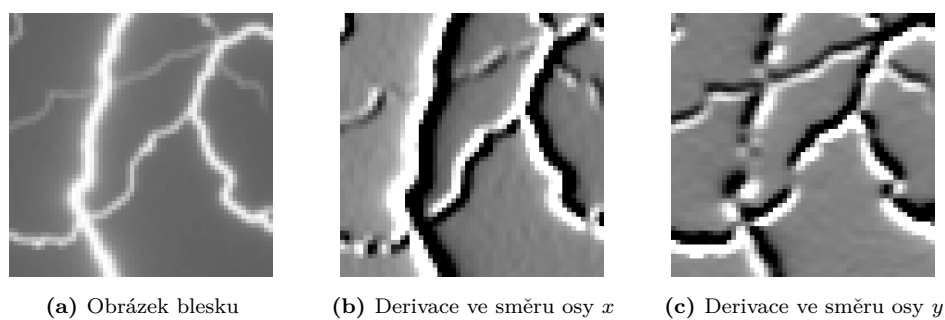
Příklad 4.1 je realizován na obrázku o malém rozlišení, abychom mohli sledovat, co se děje na pixelové úrovni a tím získali představu, jak metoda funguje.

4.2.2. Detekce hran. Detekci hran realizujeme pomocí Sobelova operátoru, provedeme tedy konvoluci černobílého obrazu $\mathbf{BW}$ s konvolučním jádrem $\mathbf{g}_x$ ze vztahu (2.1), tím získáme matici $\mathbf{G}_x$ a poté i s konvolučním jádrem $\mathbf{g}_y$ ze vztahu (2.2) a tím získáme matici $\mathbf{G}_y$. Následně vypočítáme normu aproximovaného gradientu $\mathbf{G}$ podle vztahu (2.3). Projdeme všechny prvky matice s hodnotami této normy a najdeme mezi nimi maximum. Poté se budeme řídit hodnotou parametru g , který volíme tak, že $0 \leq g \leq 1$. Projdeme znovu všechny prvky a pokud mají hodnotu alespoň g krát nalezené maximum, nahradíme je hodnotou 1, v opačném případě je nahradíme hodnotou 0. Tím získáme matici nesoucí informaci o tom, které prvky považujeme za hrany a které ne.



Obrázek 4.1. Obrázek blesku a pixely s určitou minimální intenzitou

Příklad 4.2. Pro názornost provedeme konvoluci obrázku 4.2a se Sobelovým operátorem. Na obrázku 4.2b vidíme výsledek konvoluce s konvolučním jádrem $\mathbf{g}_x$ a na obrázku 4.2c výsledek konvoluce s konvolučním jádrem $\mathbf{g}_y$. Známe prvky



Obrázek 4.2. Obrázek blesku a směrové derivace

matic $\mathbf{G}_x$ a $\mathbf{G}_y$, takže spočítáme normu aproximovaného gradientu $\mathbf{G} = (\mathbf{G}_x, \mathbf{G}_y)$ jako

$$\|\mathbf{G}\| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}.$$

Tato norma je zobrazena na obrázku 4.3. Vidíme, že černé pixely reprezentující záporné hodnoty se zde už nevyskytují, to je způsobeno druhou mocninou matic $\mathbf{G}_x$ a $\mathbf{G}_y$.



Obrázek 4.3. Norma aproximovaného gradientu

Následně najdeme maximální hodnotu v matici normy gradientu a zvolíme si různé parametry g , které nám udávají minimální požadovanou hodnotu relativně k hodnotě tohoto maxima. Pro tyto hodnoty g nalezneme všechny pixely splňující tuto podmínku. Výsledky tohoto postupu provedeného pro různá g vidíme na obrázcích 4.4b až 4.4f.



(a) Obrázek blesku



(b) Norma gradientu od 5%



(c) Norma gradientu od 10%



(d) Norma gradientu od 15%



(e) Norma gradientu od 20%

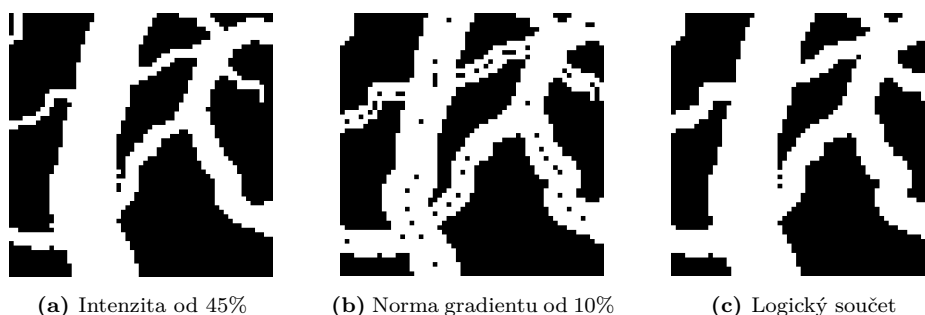


(f) Norma gradientu od 25%

Obrázek 4.4. Obrázek blesku a pixely s určitou minimální hodnotou normy gradientu

4.2.3. Konstrukce blesku. Provedli jsme hledání dostatečné intenzity obrázku a detekci hran. Teď se vrátíme k úvaze, že blesk je logický součet těchto dvou výsledků.

Příklad 4.3. Na výsledcích z příkladů 4.1 a 4.2 si ukážeme provedení logického součtu. Máme dvě matice, jejichž prvky nabývají pouze hodnot 0 a 1. Logický součet těchto matic nám dá matici, kde bude na prvku hodnota 1, pokud je hodnota 1 na alespoň jedné ze sčítaných matic, jinak je zde 0. Realizaci vidíme na obrázku 4.5. Tento výsledek už lze považovat za detekovaný blesk.



Obrázek 4.5. Znáznornění logického součtu

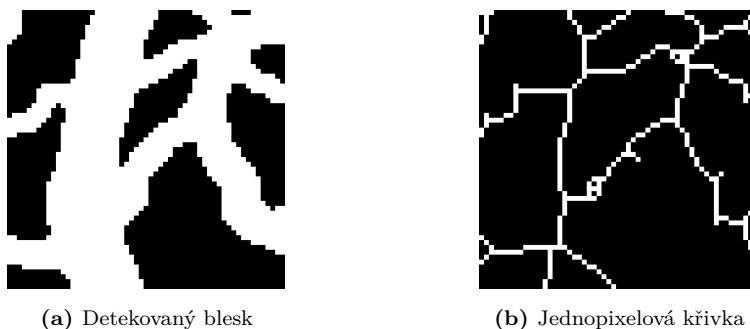
4.3. Aproximace jednopixelovou křivkou

K aproximaci blesku jednopixelovou křivkou, tedy křivkou, která má tloušťku pouze velikosti jednoho pixelu, využijeme funkci `bwmorph`, která slouží k provádění morfologických operací na binárních obrazech. Vstupy této funkce jsou obraz, název prováděné operace a případně další specifikace. V našem případě potřebujeme provést skeletonizaci, která má jakožto vstupní parametr funkce `bwmorph`, v MATLABu název `'skel'`. Výstupem této funkce je „kostra“ vstupního obrazu, neboli jednopixelová křivka.

Příklad 4.4. Příklad si ukážeme na obrázku 4.5, který jsme už v předchozí kapitole připravili k aproximaci jednopixelovou křivkou. Na obrázku 4.6 vidíme srovnání detekovaného blesku s výslednou aproximací.

4.4. Výpočet Hausdorffovy dimenze

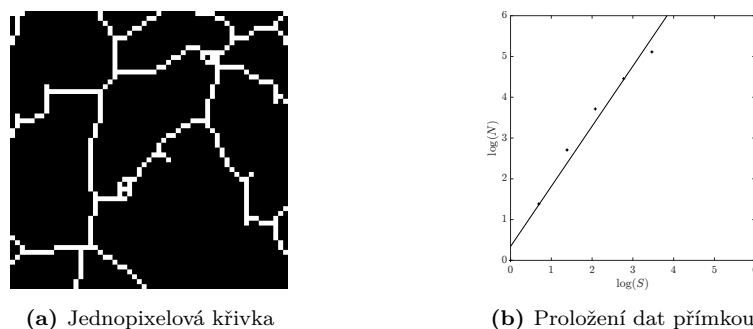
K výpočtu Hausdorffovy dimenze použijeme metodu box-counting. Tato metoda vyžaduje dělení obrazu na stejně velké čtverce, kde se mění velikost čtverců v každé iteraci. Pro usnadnění takového dělení si rozšíříme obraz o černé pixely tak, aby byl obraz čtvercový o délce strany nějaké mocniny dvojky. Tuto mocninu dvojky volíme nejmenší možnou. Po této úpravě se bude obraz snadno dělit na čtverce. Zvolíme první dělení a následné dělení získáme rovnoměrným rozdělením jednoho čtverce na čtyři menší čtverce. Upravené rozměry nám zaručují, že tímto způsobem můžeme pokračovat v dělení až do iterace, ve které mají čtverce velikost jednoho pixelu.



Obrázek 4.6. Aproximace blesku jednapixelovou křivkou

Vytvoříme si tedy dvě pole. První pole S bude sloužit k zapisování hodnot škálovacího faktoru. Do druhého pole N budeme zapisovat počet čtverců, které obsahují alespoň jeden bílý pixel. V každé iteraci i zapíšeme hodnoty $S(i)$ a $N(i)$. Máme tedy naměřené hodnoty škálovacího faktoru a k nim příslušné počty čtverců. Těmito daty potřebujeme po jejich zlogaritmování proložit přímkou. K tomu využijeme metodu nejmenších čtverců. Jakmile máme proloženou přímku, stačí zjistit jaká je její směrnice. Ta je totiž rovna vypočítané Hausdorffově dimenzi.

Příklad 4.5. V předchozí kapitole jsme si připravili jednapixelovou křivku aproximující blesk. Na ní si teď ukážeme výsledky po výpočtu Hausdorffovy dimenze. Na obrázku 4.7 vidíme jednapixelovou křivku a proložení získaných bodů přímkou. Směrnice přímky, tedy i Hausdorffova dimenze křivky, vychází v tomto případě 1,45722.



Obrázek 4.7. Výpočet Hausdorffovy dimenze

5. HAUSDORFFOVA DIMENZE BLESKU

V této kapitole se konečně dostáváme k výpočtu Hausdorffovy dimenze blesku. Pomocí programu vypočítáme Hausdorffovu dimenzi několika obrázků blesků, které

nejdříve vhodně upravíme. Výsledky výpočtu Hausdorffovy dimenze vidíme v tabulce 5.1. Hausdorffova dimenze se u různých blesků liší. Ve zkoumaných případech této kapitoly se pohybuje mezi 1,156 a 1,501.

Hausdorffova dimenze obrázků blesků							
1,501	1,333	1,271	1,270	1,156	1,373	1,204	1,235

Tabulka 5.1. Vypočítané hodnoty Hausdorffovy dimenze

ZÁVĚR

V tomto článku byla popsána teorie související s Hausdorffovou dimenzí. Následně byl vytvořen program na detekci blesku v obraze, aproximaci blesku jednopixelovou křivkou a následný výpočet Hausdorffovy dimenze. Tento program byl použit na výpočet Hausdorffovy dimenze blesků. Tato dimenze se u různých blesků liší podle míry zmenšení napětí mezi mraky nebo mezi mrakem a zemí. Rozdílné hodnoty Hausdorffovy dimenze jsou způsobené také tím, že blesk je trojrozměrný objekt, ale my detekujeme jeho tvar z průmětu do roviny.

REFERENCE

- [1] L. Čermák, R. Hlavička: *Numerické metody*, Akademické nakladatelství CERM, Brno, 2016.
- [2] Z. Došlá: *Metrické prostory: teorie a příklady*, Masarykova univerzita, Brno, 2016.
- [3] G. B. Folland: *Fourier analysis and its applications*, Pure and applied undergraduate texts 4, American Mathematical Society, Providence, 1992.
- [4] D. Martišek: *Matematické principy grafických systémů*, Littera, Brno, 2002.
- [5] I. Netuka: *Teorie míry a integrálu*, Studijní text MFF UK, 2010.
- [6] J. Tomáš: *Měření Hausdorffovy dimenze reálných objektů*, Disertační práce, Vysoké učení technické v Brně, Fakulta strojního inženýrství, Brno, 2009.
- [7] J. Vostal: *Numerické metody měření Hausdorffovy dimenze*, Bakalářská práce, Vysoké učení technické v Brně, Fakulta strojního inženýrství, Brno, 2015.

Alžběta Kočendová, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 61669 Brno, Česká republika,
e-mail: 226835@vutbr.cz

APLIKACE GEOMETRICKÝCH ALGEBER V KVANTOVÉM POČÍTÁNÍ

JAN MICHÁLEK

ABSTRAKT. Tato práce se zabývá využitím geometrických algeber v kvantovém počítání. Začíná definicí obecné Cliffordovy algebry a poté odvozuje specifickou komplexní geometrickou algebru, která je vhodná pro reprezentaci kvantových výpočtů. Tento přístup je porovnáván s tradiční metodou použití klasické maticové reprezentace. Analýzou a porovnáním těchto dvou metod si tato práce klade za cíl poskytnout poznatky o potenciálních výhodách použití geometrických algeber pro aplikace v kvantovém počítání.

ÚVOD

Kvantové počítání je relativně novým, i když velmi dynamickým oborem informatiky. Pro další rozvoj je nezbytné solidní matematické ukotvení. V současné době je nejběžnějším způsobem reprezentace kvantových stavů a operací takzvaný Diracův formalismus. Jedná se o elegantní nástroj, který nám pomáhá zvládnout všechny komplikace, které nám kvantové počítání přináší. Vede k maticové reprezentaci, kde jsou qubity chápány jako vektory a kvantové brány jako matice. Dimenze matic však s vyšším počtem vstupních qubitů roste a stává se výpočetně náročnější.

Tato práce se pokusí poskytnout alternativu. Bude ukázáno, že aparát geometrických algeber může být pro kvantové výpočty stejně vhodný a v některých konkrétních aplikacích dokonce výhodnější. Bude zavedena obecná Cliffordova algebra a odvozena speciální geometrická algebra. Kvantové stavy pak budou reprezentovány prvky geometrické algebry. Kvantové brány jsou také asociovány s prvky této algebry a to umožňuje přímou aplikaci a výpočty v rámci jedné algebraické struktury. Tento přístup bude porovnán s maticovou reprezentací na jednoduchých příkladech.

1. ALGEBRAICKÝ ZÁKLAD

V této kapitole jsou uvedeny a vysvětleny klíčové pojmy potřebné k pochopení prezentovaných konceptů.

2020 *MSC*. Primární 68Q12; Sekundární 15A66.

Klíčová slova. Cliffordova algebra, geometrická algebra, kvantové počítání.

Článek vznikl z základě bakalářské práce autora v oboru Matematické inženýrství na FSI VUT v Brně. Vedoucí práce byl Petr Vašík z Ústavu matematiky FSI VUT v Brně.

1.1. Tenzorové prostory

Tenzory jsou abstraktní objekty, které lze definovat několika způsoby. Nejprve představíme univerzální vlastnost, která nám dává omezení, která musí být vždy splněna. Tato definice však není konstruktivní. Proto uvedeme také alternativní definici tenzoru.

Definice 1.1 (Univerzální vlastnost). Tenzorový součin dvou vektorových prostorů V a W , které jsou nad stejným polem K , je vektorový prostor $V \otimes W$ spolu s bilineárním zobrazením $\otimes : (\mathbf{v}, \mathbf{w}) \mapsto \mathbf{v} \otimes \mathbf{w}$, $\otimes : V \times W \rightarrow V \otimes W$, takovým, že pro každé bilineární zobrazení $h : V \times W \rightarrow Z$ existuje jednoznačné lineární zobrazení $\bar{h} : V \otimes W \rightarrow Z$, které splňuje $h = \bar{h} \circ \otimes$ (následující diagram komutuje).

$$\begin{array}{ccc} V \times W & \xrightarrow{\otimes} & V \otimes W \\ & \searrow h & \downarrow \bar{h} \\ & & Z \end{array}$$

Univerzální vlastnost by nám měla sloužit jako ověření, zda je daná konstrukce tenzorového součinu platná. Pokud platí univerzální vlastnost, pak je zkonstruovaný tenzorový součin správný.

Definice 1.2 (Tenzorový součin). Mějme dva vektorové prostory V a W nad stejným polem K . Tenzorový součin $V \otimes W$ je vektorový prostor spolu s příslušným bilineárním zobrazením $V \times W \rightarrow V \otimes W$, které vezme prvky $\mathbf{v} \in V$ a $\mathbf{w} \in W$ a zobrazí je na prvek vektorového prostoru $V \otimes W$, $(\mathbf{v}, \mathbf{w}) \mapsto \mathbf{v} \otimes \mathbf{w}$.

Definice 1.3 (Dimenze tenzorového prostoru). Mějme dva vektorové prostory V a W s $\dim V = n$ a $\dim W = m$. Pak tenzorový prostor $V \otimes W$ má dimenzi $\dim V \otimes W = mn$.

Jedním zvláště zajímavým příkladem tenzorového součinu je *Kroneckerův součin*. Je široce využíván v kvantovém počítání. Kroneckerův součin bere jako vstupní prvky dvě matice a jako výstup vytváří jinou matici s vyšší dimenzí.

Definice 1.4 (Kroneckerův součin). Vezměme dvě matice A, B ve tvaru

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pq} \end{pmatrix}.$$

Pak jejich Kroneckerův součin, značený $C = A \otimes B$, má tvar

$$C = A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}. \quad (1.1)$$

1.2. Tenzorová algebra

Po zavedení tenzorových prostorů lze nyní zkonstruovat tenzorovou algebru.

Definice 1.5 (Tenzorová mocnina). Nechť V je vektorový prostor nad polem K . Pro každé $k \in \mathbb{N}$ je definována k -tá tenzorová mocnina V jako tenzorový součin V k -krát sám se sebou:

$$T^k V = V^{\otimes k} = V \otimes V \otimes \cdots \otimes V.$$

Definice 1.6 (Tenzorová algebra). Tenzorová algebra $T(V)$ je konstruována jako přímá suma $T^k V$ pro $k = 0, 1, 2, \dots$, tj.

$$T(V) = \bigoplus_{k=0}^{\infty} T^k V = K \oplus V \oplus (V \otimes V) \oplus (V \otimes V \otimes V) \oplus \cdots. \quad (1.2)$$

Pod $T^0 V$ je rozuměno základní pole K .

Definice 1.7 (Gradovaná algebra). Algebra A nad polem K je nazývána gradovanou, pokud může být zapsána jako přímá suma $A = \bigoplus_{k=0}^{\infty} A^k$ vektorových prostorů nad K , takových že zobrazení násobení splňuje $A^k \times A^l \rightarrow A^{k+l}$.

Z definice tenzorové algebry je zřejmé, že tato algebra je gradovaná. Například skalár ($T^0 V$) je tenzor řádu 0 a má stupeň 0, vektor ($T^1 V$) je tenzor řádu 1 a má stupeň 1 a tak dále. Tato vlastnost tenzorové algebry umožňuje manipulaci s objekty různých stupňů v rámci jedné algebraické struktury.

1.3. Cliffordova algebra

Cliffordova algebra je značena $Cl(V, Q)$, kde V je vektorový prostor a Q je kvadratická forma definovaná na vektorovém prostoru V . Přesná definice pak může být zapsána následovně.

Definice 1.8 (Cliffordova algebra, [6]). Nechť V je n -rozměrný vektorový prostor nad základním polem K a Q je kvadratická forma na V . Cliffordova algebra je pak definována jako

$$T(V)/I(Q) = T(V)/(\mathbf{v} \otimes \mathbf{v} - Q(\mathbf{v})\mathbb{1})$$

kde $T(V)$ je tenzorová algebra vektorového prostoru V , $I(Q)$ je oboustranný ideál generovaný kvadratickou formou Q a $\mathbb{1}$ je multiplikativní identita ve V .

Budeme se soustředit zejména na případy, kde základní pole K je $\mathbb{R}$ nebo $\mathbb{C}$. Cliffordova algebra má také některé zajímavé vlastnosti. Tato algebra je volná, unitární, asociativní a také gradovaná.

Alternativním způsobem popisu Cliffordovy algebry je *univerzální vlastnost*. Tento přístup není konstruktivní, ale poskytuje dobrý kontext pro hlubší porozumění.

Definice 1.9 (Univerzální vlastnost Cliffordovy algebry). Nechť V je vektorový prostor nad polem K , nechť A je unitární asociativní algebra s operací $\odot$. Nechť $Cl(V, Q)$ je Cliffordova algebra s operací $*$. Dále nechť i je lineární zobrazení $i :$

$V \rightarrow Cl(V, Q)$ splňující $i(\mathbf{v}) * i(\mathbf{v}) = Q(\mathbf{v})\mathbb{1}$ pro všechna $\mathbf{v} \in V$. Toto zobrazení je definováno pomocí **univerzální vlastnosti**: Pro libovolnou unitární asociativní algebru A nad K a libovolné lineární zobrazení $j : V \rightarrow A$, které splňuje $j(\mathbf{v}) \odot j(\mathbf{v}) = Q(\mathbf{v})\mathbb{1}_A$, kde $\mathbb{1}_A$ značí multiplikativní jednotku v A , pro všechna $\mathbf{v} \in V$, existuje jednoznačný algebraický homomorfismus $f : Cl(V, Q) \rightarrow A$, pro který platí, že následující diagram komutuje.

$$\begin{array}{ccc} V & \xrightarrow{i} & Cl(V, Q) \\ & \searrow j & \downarrow f \\ & & A \end{array}$$

Komutováním je myšleno, $f \circ i = j$.

Vektorový prostor V je vybaven kvadratickou formou Q , a proto může být vytvořena ortogonální báze. S využitím symetrické bilineární formy asociované kvadratické formě Q můžeme psát: $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$ pro $i \neq j$ a $\langle \mathbf{e}_i, \mathbf{e}_i \rangle = Q(\mathbf{e}_i)$.

Často se setkáváme s Cliffordovými algebry zapsanými ve formě $Cl_{p,q}(V, Q)$. Ty jsou nazývány p, q Cliffordovými algebry. Nyní ukážeme obecný příklad vysvětlující význam koeficientů p, q , správně označovaných jako *signatura* Cliffordovy algebry.

Nechť $K = \mathbb{R}$ je základní pole pro vektorový prostor $V = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ vybavený kvadratickou formou $Q(\alpha_1 \mathbf{v}_1 + \dots + \alpha_n \mathbf{v}_n)$, kde $\alpha \in \mathbb{R}$. Formu můžeme zapsat následujícím způsobem:

$$(\alpha_1 \quad \dots \quad \alpha_n) A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix},$$

kde uvažujeme α_i jako koeficienty vektoru $\mathbf{v}$ zapsaného v bázi, a A značí matici odpovídající dané kvadratické formě.

Protože víme, že matice A je symetrická, $A = A^T$, z definice kvadratické formy plyne, že A je *ortogonálně diagonalizovatelná*:

$$U^T A U = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Nyní je báze změněna vzhledem k diagonalizaci, $V = \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Kvadratická forma je nyní přepsána do tvaru

$$Q(\beta_1 \mathbf{x}_1 + \dots + \beta_n \mathbf{x}_n) = \lambda_1 \beta_1^2 + \dots + \lambda_n \beta_n^2, \quad \beta_i \in \mathbb{R}.$$

To znamená, že můžeme jednoduše psát $Q(\mathbf{x}_i) = \lambda_i$, odkud odvodíme $\mathbf{x}_i^2 = \lambda_i$.

Podívejme se nyní detailně čemu může být λ_i rovno. Možnosti se rozpadnou do tří neizomorfních případů:

$$\lambda_i = \begin{cases} 1, \\ 0, \\ -1. \end{cases}$$

Lze ukázat, že jiné případy, například 2, jsou izomorfní buď s 1 nebo -1 .

Shrnutí všech poznatků a informací je následující. Necht A je Cliffordova algebra $A \cong Cl(V, Q)$ s n generátory, pak vektorový prostor V a kvadratická forma Q mohou být zapsány následovně (kvadratická forma je již ortogonalizována s nenulovými prvky na diagonále):

$$V = \text{span} \{ \mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{y}_1, \dots, \mathbf{y}_q, \mathbf{z}_1, \dots, \mathbf{z}_r \}, \quad p + q + r = n,$$

$$Q(\alpha_1 \mathbf{x}_1 + \dots + \alpha_p \mathbf{x}_p + \beta_1 \mathbf{y}_1 + \dots + \beta_q \mathbf{y}_q + \gamma_1 \mathbf{z}_1 + \dots + \gamma_r \mathbf{z}_r), \quad \alpha_i, \beta_i, \gamma_i \in \mathbb{R}.$$

Pro každé $\mathbf{x}, \mathbf{y}, \mathbf{z}$ nyní platí, již dříve popsany, speciální případ:

$$\mathbf{x}_i^2 = 1, \quad \mathbf{y}_i^2 = -1, \quad \mathbf{z}_i^2 = 0.$$

V této práci se budeme zajímat o algebry, kde $r = 0$. Koeficienty p, q lze nyní snadno vysvětlit. Počet prvků v bázi, jejichž druhá mocnina je 1, je koeficient p , a počet prvků v bázi, jejichž druhá mocnina je -1 , je koeficient q . Konkrétní algebra je potom zapsána se svou signaturou:

$$Cl_{p,q}(V, Q).$$

1.4. Vnější algebra

Vnější algebra, známá také jako Grassmannova algebra, se používá zejména v geometrii při studiu ploch, objemů a objektů vyšší dimenze. Využívá vnější součin.

Definice 1.10 (Vnější algebra). Necht V je n -rozměrný vektorový prostor nad základním polem K a I je ideál generovaný $(\mathbf{u} \otimes \mathbf{u})$. Vnější algebrou pak rozumíme:

$$T(V)/(\mathbf{u} \otimes \mathbf{u}),$$

což se obvykle značí jako $\bigwedge(V)$.

Na této algebře je definován tzv. vnější součin, v angličtině známý jako *wedge product*, viz [6].

Definice 1.11 (Vnější součin). Necht V je vektorový prostor nad základním polem K . Potom zobrazení $\bigwedge : V \times V \rightarrow V \otimes V$ se nazývá *vnější součin*, jestliže má následující vlastnosti:

(A1) *Antisymetrie*:

$$\mathbf{u} \wedge \mathbf{v} = -\mathbf{v} \wedge \mathbf{u},$$

(A2) *Škálování*:

$$(\alpha \mathbf{u}) \wedge \mathbf{v} = \alpha(\mathbf{u} \wedge \mathbf{v}),$$

(A3) *Distributivita*:

$$\mathbf{u} \wedge (\mathbf{v} + \mathbf{w}) = (\mathbf{u} \wedge \mathbf{v}) + (\mathbf{u} \wedge \mathbf{w})$$

pro všechna $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ a $\alpha \in K$.

Tato definice může být rozšířena na $\bigwedge : V \times \cdots \times V \rightarrow V^{\otimes k}$, čímž se přidá asociativita. Lineární kombinace součinů dvou vektorů $\mathbf{u} \wedge \mathbf{v}$ se nazývá *bivektor*, podobně lineární kombinace součinů tří vektorů se nazývá *trivektor* a lineární kombinace součinů k vektorů se nazývá *k-vektor*. Lineární kombinace smíšených k -vektorů se nazývá *multivektor*. Budeme také používat termín *blade*. Blade k -třídy se skládá z vnějšího součinu k vektorů. Bivektor lze chápat jako orientovanou rovinu a trivektor jako orientovaný objem.

Definice 1.12 (Vnější mocnina). k -tá vnější mocnina, $\bigwedge^k(V)$, je definována jako vektorový podprostor $\bigwedge(V)$ generovaný prvky ve tvaru $\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \cdots \wedge \mathbf{x}_k$, kde $\mathbf{x}_i \in V$.

1.5. Bra-Ket notace

Při práci s vektory bude použita bra-ket notace. Tato notace je často využívána v kvantovém počítání k popisu kvantových stavů. Poskytuje způsob, jak popsat vektor $\mathbf{v}$ v komplexním vektorovém prostoru. *Ket* je označován jako $|a\rangle$, a *bra* je označován jako $\langle a|$.

Použití bra-ket notace je efektivnější než zápis vektorů do sloupců či řádků a umožňuje snadnější manipulaci s vektory. *Bra* reprezentuje sloupcový vektor a *ket* vznikne jeho transpozicí a komplexním sdružením:

$$|a\rangle = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \quad \langle a| = (a_1^* \quad a_2^* \quad \dots \quad a_n^*).$$

Obecně lze psát: $\langle a|^\dagger = |a\rangle$, $|a\rangle^\dagger = \langle a|$, kde symbol $\dagger$ je používán pro označení konjugované transpozice.

Vnitřní součin dvou vektorů je pak zapsán ve tvaru $\langle a||b\rangle$, což se zkracuje na $\langle a|b\rangle$.

2. KVANTOVÉ POČÍTÁNÍ

2.1. Qubit

Qubit je základní jednotkou v kvantové informatice a liší se od klasického bitu, který má pouze dva možné stavy - 0 nebo 1. Qubit totiž může existovat v superpozici obou stavů. To znamená, že qubit může uchovávat nejen 0 nebo 1, ale všechny možné kombinace těchto stavů současně. To otevírá široké spektrum možností pro zpracování a ukládání informací způsoby, které klasické bity neumožňují. Tato zvýšená flexibilita má však svou protiváhu. Hlavní výzvou v kvantovém počítání je najít způsoby, jak pomocí matematického aparátu reprezentovat a manipulovat s těmito superpozicemi. Další výzvou je navrhnout a sestavit hardware, které by tyto výpočty mohl efektivně implementovat.

Pro reprezentaci qubitů budou definovány dva ortonormální báze stavy. Budou značeny $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ a $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Tyto dva stavy jsou většinou nazývány výpočetní bázi.

Obecný kvantový stav je ztotožněn s normalizovaným vektorem, generovaným z ortonormální výpočetní báze v dvoudimenzionálním prostoru. Jak jsme již dříve zmínili, qubit je kombinací, v našem případě lineární kombinací, základních stavů a může být napsán následujícím způsobem:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad (2.1)$$

kde koeficienty α a β jsou komplexní čísla.

Koeficienty α a β jsou správně nazývány *pravděpodobnostní amplitudy*. Pokud změříme daný qubit, pravděpodobnosti získání výsledků $|0\rangle$ nebo $|1\rangle$ jsou $|\alpha|^2$ případně $|\beta|^2$. Druhý axiom pravděpodobnosti nám dává omezení $|\alpha|^2 + |\beta|^2 = 1$, které musí vždy platit. Koeficienty α a β jsou komplexní čísla a právě kvůli tomu má qubit čtyři stupně volnosti. Existuje více způsobů, jak se vyrovnat s tím, že pracujeme s objektem, který má čtyři stupně volnosti. Dají se také nalézt způsoby, jak qubit vhodně vizualizovat.

2.2. Blochova sféra

Elegantním způsobem reprezentace qubitů je Blochova sféra.

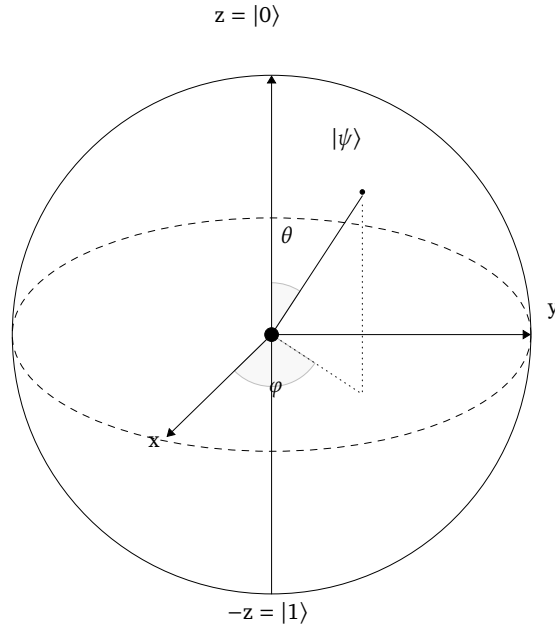
Vraťme se ke vztahu (2.1), můžeme vidět, že k popsání jednoho qubitů je potřeba kombinace dvou základních stavů. Také víme, že koeficienty α a β jsou komplexní čísla. Každý daný stav tak může být popsán čtyřmi reálnými čísly. Takový systém by bylo obtížné vizualizovat.

Ovšem vzhledem k tomu, že jsme identifikovali kvantové stavy s komplexními vektory, kde koeficienty jsou pravděpodobnostní amplitudy, můžeme nyní zavést *fázi*. Fáze je úhel mezi reálnou a imaginární složkou komplexního vektoru. Ve skutečnosti existují dva druhy fází. *Relativní fáze* se týká fázového rozdílu mezi dvěma pravděpodobnostními amplitudami. Je to míra toho, jak se amplitudy vzájemně ovlivňují. *Globální fáze* se týká celkové fáze kvantového stavu. Je to společný fázový faktor, který ovlivňuje všechny pravděpodobnostní amplitudy ve stavu. Zdůrazněme však, že globální fáze neovlivňuje pravděpodobnosti měření konkrétního výsledku. Navíc globální fáze nemá fyzikální význam, protože se zruší při výpočtu pravděpodobností. To znamená, že jeden stupeň volnosti lze odstranit. Pokud vezmeme v úvahu omezení dané druhým axiomem pravděpodobnosti, zbavíme se dalšího stupně volnosti. To nás vede k systému se dvěma stupni volnosti.

Jak již název napovídá, k reprezentaci qubitů bude použita sféra. A protože máme dva stupně volnosti, lze efektivně využít sférických souřadnic. Obecný qubit lze pak zapsat následujícím způsobem:

$$|\psi\rangle = \cos(\theta/2)|0\rangle + e^{i\phi} \sin(\theta/2)|1\rangle = \cos(\theta/2)|0\rangle + (\cos\phi + i\sin\phi) \sin(\phi/2)|1\rangle,$$

kde $0 \leq \theta \leq \pi$ a $0 \leq \phi \leq 2\pi$.



Obrázek 2.1. Blochova sféra.

Při vhodném výběru parametrů θ a ϕ lze každý stav jednoznačně popsat, viz Obrázek 2.1. Přesněji řečeno platí jednoznačná korespondence mezi kvantovými stavy (reprezentovanými qubity) a body na sféře. Je důležité si uvědomit, že qubit má podobu dvourozměrného vektoru. Volba báze je zcela libovolná. V této práci jsme od začátku pracovali s $|0\rangle$ a $|1\rangle$. Tyto stavy lze ztotožnit se severním a jižním pólem naší sféry. Nicméně každý pár dvou protilehlých bodů je ortogonální, jejich vnitřní součin je nulový. Mohou tedy sloužit jako ortogonální báze. Z praktických důvodů se však budeme držet volby $|0\rangle$ a $|1\rangle$.

Stavy $|0\rangle$ a $|1\rangle$ jsou umístěny na z -ové ose. Nyní se podívejme blíže na x -ovou a y -ovou osu. Mají svůj vlastní pár protilehlých bodů. Na x -ové ose obvykle hovoříme o $|+\rangle$, $|-\rangle$, a na y -ové ose o $|i\rangle$, $|-i\rangle$. Tyto páry také mohou tvořit bázi, jak jsme již dříve uvedli. Jelikož byla zvolena báze $|0\rangle$, $|1\rangle$, vyjádříme nyní tyto páry bodů v této bázi a dostaneme

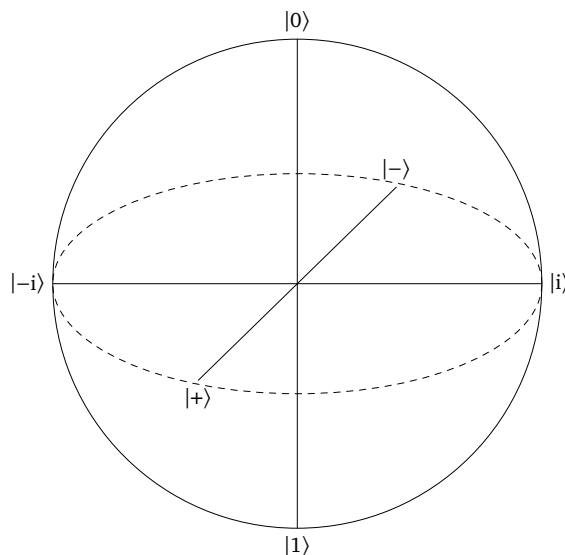
$$|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \quad |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$$

a

$$|i\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle), \quad |-i\rangle = \frac{1}{\sqrt{2}}(|0\rangle - i|1\rangle).$$

Tyto body jsou poté vizualizovány na sféře, viz Obrázek 2.2.

Pokusíme se pochopit klasický bit z tohoto nového pohledu. Bylo řečeno, že bit může nabývat pouze dvou hodnot, 0 a 1. Může tedy být buď na severním pólu nebo



Obrázek 2.2. Blochova sféra – osy.

na jižním pólu naší sféry a nikde jinde. Qubit může zaujmout libovolnou pozici na sféře, zatímco klasický bit si musí vybrat mezi dvěma protilehlými body, které jsou zvoleny jako báze. Zde je jasně vidět obrovská variabilita, kterou poskytuje použití qubitů pro výpočty.

2.3. Multi qubity

Qubity byly spojeny se stavovými vektory a stavový vektor o dimenzi 2 byl použit pro reprezentaci jednoho qubitu. K reprezentaci multi qubitů stačí zvýšit dimenzi vektoru. Jeden qubit je reprezentován 2-dimenzionálním vektorem, dva qubity jsou reprezentovány 4-dimenzionálním vektorem a n -qubit je reprezentován 2^n -dimenzionálním vektorem. Dimenze také určuje, kolik bázových vektorů je potřeba. Podívejme se na příklad dvou qubitů a jak je zde stavový vektor konstruován.

$$|a\rangle = a_{00}|00\rangle + a_{01}|01\rangle + a_{10}|10\rangle + a_{11}|11\rangle = \begin{pmatrix} a_{00} \\ a_{01} \\ a_{10} \\ a_{11} \end{pmatrix}.$$

Připomeňme si, že koeficienty a_{ij} jsou *pravděpodobnostní amplitudy*. Kvadrát koeficientu $|a_{ij}|^2$ je pravděpodobnost změření stavu $|ij\rangle$.

Druhý axiom pravděpodobnosti musí platit i zde a proto dospíváme k následující normalizační podmínce:

$$|a_{00}|^2 + |a_{01}|^2 + |a_{10}|^2 + |a_{11}|^2 = 1.$$

Stejná konstrukce bude použita pro n -qubity. Normalizační podmínka musí vždy platit.

Pro dva oddělené qubity definujeme *kolektivní stav* pomocí Kroneckerova součinu, viz vztah (1.1). K reprezentaci kolektivního stavu také použijeme stavový vektor. Dimenze stavového vektoru je 2^n , kde n je počet qubitů, které popisujeme. Podívejme se na příklad, jak je kolektivní stav dvou qubitů získán

$$|a\rangle = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \quad |b\rangle = \begin{pmatrix} b_0 \\ a_1 \end{pmatrix},$$

$$|ba\rangle = |b\rangle \otimes |a\rangle = \begin{pmatrix} b_0 \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \\ b_1 \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} b_0 a_0 \\ b_0 a_1 \\ b_1 a_0 \\ b_1 a_1 \end{pmatrix}.$$

Vstupem byly dva qubity a použili jsme Kroneckerův součin k sestavení stavového vektoru reprezentujícího kolektivní stav. Konečný stavový vektor má dimenzi 4. Bylo řečeno, že stavový vektor popisující kolektivní stav má dimenzi 2^n . Vzali jsme dva qubity, tedy 2^2 , což je vskutku 4.

2.4. Měření

Uvedli jsme, že qubit může získat nejen hodnoty 0 a 1, ale všechny jejich kombinace. Avšak po provedení měření je qubit změřen buď na pozici odpovídající 0 nebo na pozici odpovídající 1. Je důležité objasnit, v jaké bázi se měření provádí. Stejný qubit můžeme měřit v různých bazích a získat různé výsledky. Měření v bázi $|0\rangle, |1\rangle$ dává možné výsledky $|0\rangle, |1\rangle$. Ale pokud změníme bázi například na $|+\rangle, |-\rangle$, pak jsou možné výsledky $|+\rangle$ a $|-\rangle$. Volba báze je klíčová pro správný výklad měření. Budeme se držet báze $|0\rangle, |1\rangle$.

Ve skutečnosti je poměrně jednoduché spočítat pravděpodobnost výsledků pro libovolný qubit. Výpočet vnitřního součinu mezi bázevým stavem a stavovým vektorem reprezentujícím qubit dává pravděpodobnost, že qubit skončí v konkrétním bázevém stavu. Přesněji řečeno, druhá mocnina této hodnoty je požadovaná pravděpodobnost. Vezměme si například obecný vektor $|\psi\rangle$ a podívejme se, jak získat pravděpodobnosti změřením v bázevých stavech $|0\rangle, |1\rangle$,

$$p(|0\rangle) = |\langle 0|\psi\rangle|^2, \quad p(|1\rangle) = |\langle 1|\psi\rangle|^2.$$

3. KVANTOVÉ BRÁNY

Identifikovali jsme qubit se stavovým vektorem a nyní ho můžeme transformovat. Transformace qubitu je ve své podstatě pouze rotací na Blochově sféře. V klasické informatice pracují výpočty s logickými bránami. Tyto brány berou jako vstup jeden nebo více bitů, provádí operace definované matematickou logikou a jako výstup vrací opět bit nebo více bitů.

Pro smysluplné kvantové počítání je nutné zavést tyto logické brány. Blochova sféra je velmi užitečná pro vizualizaci toho, jak fungují brány na konkrétním qubitu. Pro brány na více qubitech bude použito tensorového součinu.

3.1. Brány na qubitu

Brány na qubitu berou jako vstup jeden qubit reprezentovaný stavovým vektorem, aplikují operátor a dávají jako výstup jiný stavový vektor

$$M|\psi\rangle \rightarrow |\psi'\rangle. \quad (3.1)$$

Můžeme vidět, že operátor M změní náš stavový vektor na jiný. Z lineární algebry je dobře známo, že matice se používají k transformaci vektorů. Matice tedy budou reprezentovat kvantové brány.

Tyto matice musí být unitární, protože norma vektoru musí být zachována.

Definice 3.1 (Unitární matice). Invertibilní čtvercová matice se nazývá *unitární*, když její komplexně sdružená transpozice A^* je také její inverzní matice, $AA^* = A^*A = I$, kde I je jednotková matice.

Existuje nekonečně mnoho způsobů, jak napsat obecný tvar operátoru M . Jeden z možných tvarů je následující:

$$M(\phi, \theta, \lambda) = \begin{pmatrix} \cos(\theta/2) & -e^{i\lambda} \sin(\theta/2) \\ e^{i\phi} \sin(\theta/2) & e^{i(\phi+\lambda)} \cos(\theta/2) \end{pmatrix}.$$

Tento tvar je však nepraktický a zbytečně složitý. Proto bude definována množina základních matic a požadovaný operátor bude poté získán kombinací těchto matic. S relativně malou množinou bázových matic lze získat libovolnou požadovanou matici. Navíc součin dvou unitárních matic je také unitární matice. Zvolené matice budou násobeny, vytvoří novou matici, která bude reprezentovat operátor. Stavový vektor je pak násoben konečnou maticí a tím se provede transformace.

Protože jsme zavedli Blochovu sféru pro reprezentaci qubitů, můžeme si představit transformace jako rotace na Blochově sféře. Pokud je splněna podmínka unitarity matice, zůstává stavový vektor na sféře a jen se kolem ní otáčí.

3.1.1. NOT brány. Základní logickou bránou v klasické informatice je NOT brána. Přepisuje hodnotu bitu na opačnou možnost. Bit s hodnotou 0 resp. 1 je změněn na 1 resp. 0. Situace u qubitů je trochu složitější. Jelikož qubit může nabývat libovolné hodnoty na Blochově sféře, nemůže být NOT brána tak snadno definována.

Začneme nejjednodušším příkladem. Pokud je qubit na severním pólu Blochovy sféry, což reprezentuje hodnotu $|0\rangle$, pak by NOT brána měla změnit jeho hodnotu na $|1\rangle$. Toho lze dosáhnout rotací stavového vektoru kolem x -ové nebo y -ové osy. Když je NOT brána aplikována na stavový vektor, jeho hodnota je změněna na hodnotu portilehlého bodu. Matice, které tuto transformaci provádějí, se ne vždy dají snadno nalézt. Prozatím bude stačit definovat rotace kolem osy x , y a z :

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

3.1.2. Hadamardova brána. Možná nejdůležitější bránou v kvantovém počítání je takzvaná Hadamardova brána. Zaujímá zvláštní pozici v množině elementárních matic. Na první pohled nemusí být zřejmé, proč je tato matice tak zásadní. Její důležitost bude vysvětlena později, nyní bude poskytnuta pouze definice. Když Hadamardova brána působí na stavový vektor, otáčí vektor kolem osy $(x - z)$. Forma matice reprezentující Hadamardovu bránu je následující:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

3.1.3. Fázová brána. Fázová brána slouží k otáčení stavového vektoru kolem osy z . Jako parametr přijímá reálné číslo ϕ , které určuje úhel otočení. Forma matice je následující:

$$P = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\phi} \end{pmatrix}.$$

Parametr ϕ můžeme měnit a získáme tak tzv. rodinu fázových bran.

3.2. Brány pro multi qubity

Brány pro multi qubity jsou klíčovým prvkem kvantového počítání, protože umožňují provádět operace na více qubitech současně. Stejně jako u bran pro jednotlivé qubity i u bran pro multi qubity se spoléhá na matematické operace s maticemi pro transformaci stavových vektorů. Nicméně matice používané pro multi qubitové brány jsou mnohem větší, s rozměry závislými na počtu vstupních qubitů. Rozměr matice brány pro multi qubit je roven $2^n \times 2^n$, kde n je počet vstupních qubitů. Tyto matice musí splňovat stejné vlastnosti jako matice pro jednotlivé qubity a jejich konstrukce vyžaduje pečlivé zvážení toho, jak brána působí na báze stavy.

Pro ilustraci tohoto konceptu můžeme zmínit některé běžné příklady bran pro multi qubity. Například existují 2-qubitové brány jako je CNOT (Controlled NOT) brána, SWAP brána a CZ brána (controlled phase gate). Kromě toho existují 3-qubitové brány, jako je Toffoliho brána (také známá jako CCNOT brána). Je důležité poznamenat, že tyto brány jsou jen několika příklady z mnoha možných bran pro multi qubity a volba konkrétní brány bude záviset na požadovaném účelu daného výpočetního systému. Nicméně porozumění základům toho, jak tyto brány fungují a jak ovlivňují stavové vektory, je klíčovým krokem při návrhu efektivních kvantových obvodů.

3.2.1. CNOT brána. Controlled-NOT brána je nejpoužívanější bránou pro multi qubity. Přijímá dva qubity jako vstup, zkontroluje hodnotu prvního qubitu nazývaného *controlled qubit*, a pokud je 1, obrátí hodnotu druhého qubitu nazývaného *target qubit*. Podívejme se, jak brána působí na báze stavy:

$$\text{CNOT}|00\rangle = |00\rangle, \quad \text{CNOT}|01\rangle = |01\rangle, \quad \text{CNOT}|10\rangle = |11\rangle, \quad \text{CNOT}|11\rangle = |10\rangle.$$

Matice reprezentující CNOT bránu má následující formu:

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

První dva řádky odpovídají případům, kdy je controlled qubit ve stavu $|0\rangle$ a proto se nic nemění. Třetí a čtvrtý řádek odpovídá situaci, kdy je controlled qubit ve stavu $|1\rangle$ a tedy se hodnota target qubitů obrátí.

3.2.2. SWAP brána. SWAP brána je velmi jednoduchá. Jednoduše prohodí stavy dvou qubitů. Nemá vliv na stavy $|00\rangle$, $|11\rangle$, ale ovlivňuje stavy $|10\rangle$, $|01\rangle$. Matice, která ji reprezentuje je následující:

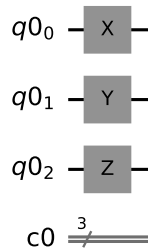
$$\text{SWAP} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

3.2.3. CZ brána. Také nazývaná Controlled-Z brána provádí rotaci kolem osy z , pokud je první qubit (controlled qubit) roven 1. Má například vliv na stav $|11\rangle$, který je poté změněn na $-|11\rangle$. Obecná forma je následující:

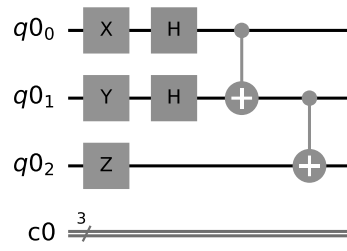
$$\text{CZ} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

3.3. Kvantový obvod

Nyní můžeme začít s konstrukcemi kvantových obvodů. Budeme používat nástroj Qiskit [11]. Podívejme se na příklad na Obrázku 3.1.



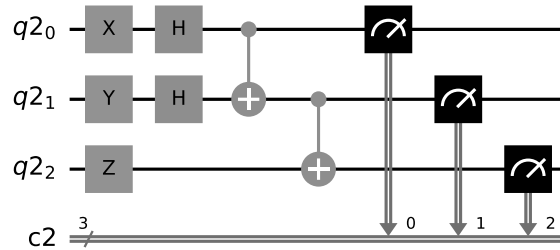
Obrázek 3.1. Základní brány.



Obrázek 3.2. Hadamardova brána a C-NOT brána.

Tento obvod bere jako vstup tři qubity. Každý z nich je uložen na samostatném vlákne. Tyto vlákna budou označeny jako $q0_i$ a budou použity k uložení qubitů. Zajímá nás pouze dolní index i a nula ve výrazech pro qubity pro nás nemá žádný

zvláštní význam. Klasické bity jsou také použity v obvodech. Používají se k uložení hodnoty získané z měření qubitu. Jsou označeny jako $c0$ a spojeny do jednoho vlákna, aby se ušetřilo místo a obvody byly srozumitelnější. Můžeme vidět, že na kvantových drátech jsou již kvantové brány, konkrétně brány X , Y a Z . Tyto brány působí na konkrétní qubity. Nyní mohou být přidány další brány, například Hadamardova brána a C-NOT brána, viz Obrázek 3.2. Všimněme si, že C-NOT brána je brána 2-qubitová, což znamená, že má jako vstup dva qubity. Posledním krokem k funkčnímu obvodu je přidání měření. Zde jsou klasické bity užitečné. Měření je umístěno na vlákno s qubitem a výsledek je uložen do klasického bitu. Protože existují pouze dvě možnosti, jak měření může skončit, je možné tuto hodnotu uložit do klasického bitu. Na Obrázku 3.3 je vidět, jak je měření prováděno a kde jsou výsledky uloženy.



Obrázek 3.3. Měření.

Tento obvod nemá žádný konkrétní účel, ale může sloužit jako příklad. V následující kapitole bude ukázáno, jak můžeme tyto obvody reprezentovat pomocí geometrických algeber.

4. GEOMETRICKÉ ALGEBRY V KVANTOVÉM POČÍTÁNÍ

Myšlenka implementace geometrických algeber (GA) v kvantovém počítání (QC) spočívá ve speciálním způsobu reprezentace qubitů. Jednotlivé qubity jsou asociovány s prvky GA a stejně tak i kvantové brány jsou asociovány s prvky GA. Díky tomu se nám otevírá cesta pro přímou implementaci. Jednotlivé transformace qubitů pomocí kvantových bran se dají reprezentovat jako součin prvků, které reprezentují kvantový stav a kvantovou bránu.

4.1. Geometrická algebra

Nyní definujeme vhodný aparát GA pro účely QC. V první kapitole byly diskutovány a obecně definovány základní algebraické pojmy. Tyto definice budou nyní mírně upraveny pro konkrétní účely. Začneme s případem reálné GA. Abychom mohli správně definovat GA, musí být zaveden nový pojem nazývaný *geometrický součin*, viz [3, 4]. Geometrický součin je kombinací vnitřního a vnějšího součinu. Mějme reálný vektorový prostor $\mathbb{R}^n$. Vektorový prostor má ortonormální bázi $(\mathbf{e}_1, \dots, \mathbf{e}_n)$. Je také vybaven bilineárním formou B se signaturou (p, q) .

$$B(\mathbf{e}_i, \mathbf{e}_j) = \begin{cases} 1, & \text{je-li } i = j = 1, \dots, p, \\ -1, & \text{je-li } i = j = p + 1, \dots, m, \\ 0, & \text{je-li } i \neq j. \end{cases} \quad (4.1)$$

Všimněme si, že bilineární forma B je symetrická. Proto lze definovat k ní asociovanou kvadratickou formu Q . Vztah je následující:

$$B(\mathbf{a}, \mathbf{b}) = \frac{1}{2}(Q(\mathbf{a} + \mathbf{b}) - Q(\mathbf{a}) - Q(\mathbf{b})). \quad (4.2)$$

Definice 4.1. V případě vektorového prostoru $\mathbb{R}^n$ s bilineární formou B je vnitřní součin definován jako

$$\mathbf{e}_i \cdot \mathbf{e}_j = B(\mathbf{e}_i, \mathbf{e}_j).$$

Pro definici vnějšího součinu použijeme definici z vnější algebry, Definice 1.11. Sečtení těchto dvou součinů vede k tzv. geometrickému součinu, viz [8].

Definice 4.2 (Geometrický součin). Geometrický součin, označovaný jako $\mathbf{e}_i \mathbf{e}_j$, dvou bázových vektorů je kombinací vnitřního a vnějšího součinu

$$\mathbf{e}_i \mathbf{e}_j = \mathbf{e}_i \cdot \mathbf{e}_j + \mathbf{e}_i \wedge \mathbf{e}_j.$$

Je důležité si všimnout všech užitečných vlastností, které geometrický součin má. Tyto vlastnosti budou široce využívány, protože nám umožňují zjednodušit výpočty.

Poznámka. Vzhledem k tomu, že vnitřní součin dvou ortonormálních prvků je roven nule, viz (4.1), redukuje se geometrický součin dvou ortonormálních prvků na jejich vnější součin,

$$\mathbf{e}_i \mathbf{e}_j = \mathbf{e}_i \wedge \mathbf{e}_j.$$

Podobné chování má také vnější součin. V případě bázových prvků víme, že vnější součin prvku se sebou samým je roven nule. Proto můžeme odvodit, že geometrický součin dvou ortonormálních bázových prvků se redukuje na vnitřní součin, který je buď plus nebo mínus jedna,

$$\mathbf{e}_i \mathbf{e}_i = \mathbf{e}_i \cdot \mathbf{e}_i = \begin{cases} 1, & \text{pokud } i = j = 1, \dots, p, \\ -1, & \text{pokud } i = j = p + 1, \dots, m. \end{cases}$$

Tyto definice jsou nyní rozšířeny na obecný stupeň r bladu. Pro vnitřní součin dostaneme tvar

$$\mathbf{e}_j \cdot \mathbf{e}_A = \sum_{k=1}^r (-1)^k B(\mathbf{e}_j, \mathbf{e}_{i_k}) \mathbf{e}_{A \setminus \{i_k\}}.$$

Pro vnější součin získáme

$$\mathbf{e}_j \wedge \mathbf{e}_A = \begin{cases} \mathbf{e}_j \wedge \mathbf{e}_{i_1} \wedge \dots \wedge \mathbf{e}_{i_r}, & \text{pokud } j \notin A, \\ 0 & \text{pokud } j \in A. \end{cases}$$

A pro geometrický součin jednoduše definujeme

$$\mathbf{e}_i \mathbf{e}_A = \mathbf{e}_j \cdot \mathbf{e}_A + \mathbf{e}_j \wedge \mathbf{e}_A.$$

Zkonstruovaná algebra se nazývá reálná geometrická algebra s příslušným geometrickým součinem, který se také nazývá Cliffordův součin. Obvykle je značena jako $\mathbb{G}_{p,q}$. Tato algebra je ve skutečnosti dobře známou Cliffordovou algebrou, viz Definice 1.8. Kvadratická forma, která generuje ideál, je ta asociovaná s bilineární formou (4.2). S vědomím toho, že zkonstruovaná geometrická algebra je ve skutečnosti Cliffordova algebra, můžeme využít všechny vlastnosti, které Cliffordova algebra má. Tato algebra je asociativní a také volná, což znamená, že ji lze generovat z báze prvků pomocí geometrického součinu.

Vnitřní a vnější součin lze považovat za symetrický a antisymetrický součin, $\mathbf{a} \cdot \mathbf{b} = \frac{1}{2}(\mathbf{ab} + \mathbf{ba})$ a $\mathbf{a} \wedge \mathbf{b} = \frac{1}{2}(\mathbf{ab} - \mathbf{ba})$. Tyto vlastnosti lze odvodit přímo z definice těchto operací a poté jsou zapsány ve formě geometrického součinu.

4.2. Komplexifikace

Tato kapitola využívá mnoho definic z [5]. Rozšiřuje tyto koncepty, propojuje je s již definovanými a poskytuje příklady pro lepší porozumění.

Abychom mohli reprezentovat komplexní povahu qubitů, je nutné zavést komplexní čísla. Nastavení základního pole na $\mathbb{C}$ ale samo o sobě nestačí, protože potřebujeme komplexní chování přenést přímo do algebry. Toho lze dosáhnout definicí speciální ortogonální lineární transformace a také nastavením báze vektorového prostoru na sudý. Uvažujme lineární ortogonální transformaci $J : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ takovou, že $J^2 = -1$, kde 1 je identita. Akce J na ortonormální bázi $(\mathbf{e}_1, \dots, \mathbf{e}_{2n})$ je následující, viz [5]:

$$J(\mathbf{e}_j) = -\mathbf{e}_{j+n}, \quad J(\mathbf{e}_{j+n}) = \mathbf{e}_j,$$

kde $j = 1, \dots, n$. S použitím J nyní představíme novou transformovanou bázi, která se v kontextu QC ukazuje být velmi praktickou, viz [5].

Definice 4.3. Wittova báze je tvaru

$$\begin{aligned} f_j &= \frac{1}{2}(1 + iJ)(\mathbf{e}_j) = \frac{1}{2}(\mathbf{e}_j - i\mathbf{e}_{j+n}), \quad j = 1, \dots, n, \\ f_j^\dagger &= \frac{1}{2}(1 - iJ)(\mathbf{e}_j) = \frac{1}{2}(\mathbf{e}_j + i\mathbf{e}_{j+n}), \quad j = 1, \dots, n. \end{aligned}$$

Tato definice je zvláště elegantní, protože můžeme ověřit, že pro každé $j = 1, \dots, n$ platí $f_j^2 = 0$ a $f_j^{\dagger 2} = 0$. Skutečně,

$$\begin{aligned} f_j^2 &= \frac{1}{2}(\mathbf{e}_j - i\mathbf{e}_{j+n})\frac{1}{2}(\mathbf{e}_j - i\mathbf{e}_{j+n}) \\ &= \frac{1}{4}(\mathbf{e}_j^2 - i\mathbf{e}_j\mathbf{e}_{j+n} - i\mathbf{e}_{j+n}\mathbf{e}_j + i^2\mathbf{e}_{j+n}^2) \\ &= \frac{1}{4}(1 - i\mathbf{e}_j\mathbf{e}_{j+n} + i\mathbf{e}_j\mathbf{e}_{j+n} - 1) = \frac{1}{4}(0) = 0. \end{aligned} \tag{4.3}$$

Volba tohoto konkrétního $f, f^\dagger$ je také motivována skutečností, že jsou splněny jak Grassmanovy tak dualní identity:

$$\begin{aligned} f f^\dagger + f^\dagger f &= 1, \\ f_j f_k + f_k f_j &= f_j^\dagger f_k^\dagger + f_k^\dagger f_j^\dagger, \quad j, k = 1, \dots, n, \\ f_j f_k^\dagger + f_k^\dagger f_j &= \delta_{jk}, \quad j, k = 1, \dots, n, \end{aligned} \quad (4.4)$$

kde symbol δ_{jk} je Kroneckerovo delta.

Jednoduchým způsobem, jak zkonstruovat bázi celé Cliffordovy algebry $\mathbb{C}_{2n}$, je použití geometrického součinu. Můžeme použít Grassmannovy blady Wittových prvků. Konstrukce je následující;

$$\begin{aligned} f_j f_k &= f_j \cdot f_k + f_j \wedge f_k = f_j \wedge f_k \\ f_j^\dagger f_k^\dagger &= f_j^\dagger \cdot f_k^\dagger + f_j^\dagger \wedge f_k^\dagger = f_j^\dagger \wedge f_k^\dagger \\ f_j f_k^\dagger &= f_j \cdot f_k^\dagger + f_j \wedge f_k^\dagger = \frac{1}{2} \delta_{jk} + f_j \wedge f_k^\dagger \end{aligned} \quad (4.5)$$

Pro rozšíření Wittovy báze na celou algebru $\mathbb{C}_{2n}$ stačí vzít 2^{2n} možných geometrických součinů

$$(f_1)^{i_1} (f_1^\dagger)^{j_1} \dots (f_n)^{i_n} (f_n^\dagger)^{j_n}, \quad i_k, j_k \in \{0, 1\} \text{ pro } k = 1, \dots, n.$$

4.3. Qubit v GA

Jak jsme již dříve zmínili, qubit bude reprezentován prvkem ležícím v komplexní Cliffordově algebře $\mathbb{C}_2$. Můžeme použít Wittovy bazové vektory $(f, f^\dagger)$ k vytvoření báze pro komplexní Cliffordovu algebru. Bázi lze následně zapsat ve tvaru $(1, f, f^\dagger, f f^\dagger)$. Předtím, než identifikujeme qubit s konkrétním prvkem, zmiňme podrobněji vlastnosti této algebry. Volba báze nás vede k dvěma primitivním idempotentům v této algebře, konkrétně $I = f f^\dagger$ a $K = f^\dagger f$. Konvence v oblasti fyziky, do kterých zde nebudeme do detailu zabíhat, nám pro použití určují idempotent I . S touto volbou lze dospět k následující reprezentaci stavů $|0\rangle$ a $|1\rangle$ u qubitů:

$$|0\rangle = I = f f^\dagger, \quad |1\rangle = f^\dagger I = f^\dagger f f^\dagger = (1 - f f^\dagger) f^\dagger = f^\dagger.$$

Nyní můžeme přejít k definici qubitů, viz vztah (2.1), a použít naši reprezentaci stavů $|0\rangle$ a $|1\rangle$:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \alpha f f^\dagger + \beta f^\dagger = (\alpha + \beta f^\dagger) I.$$

4.4. Brány v GA

Po zbytek této práce budeme držet tuto reprezentaci stavů $|0\rangle$ a $|1\rangle$. Nyní můžeme přistoupit ke konstrukci bran. Základní brány lze snadno získat s ohledem na jejich účinek na bazové stavy. K odlišení mezi maticovou reprezentací a GA reprezentací budeme používat různou notaci, viz [5]:

$$\begin{aligned} \text{X-brána: } \lambda_X &= f^\dagger + f, \\ \text{Y-brána: } \lambda_Y &= i f^\dagger - i f, \\ \text{Z-brána: } \lambda_Z &= f f^\dagger - f^\dagger f, \end{aligned}$$

Uvedli jsme formu Z-brány v GA. Víme však, že Z-brána je pouze speciálním případem obecnější fázové brány. V jazyce GA má tato brána následující tvar, viz [5]:

$$\text{Fázová brána: } \lambda_P = ff^\dagger - e^{i\phi} f^\dagger f.$$

Konkrétní fázové brány, o kterých jsme dříve mluvili, se získávají obdobným způsobem. Parametr ϕ se mění a vznikají různé brány. A při volbě $\phi = \pi$ dostáváme Z-bránu. Poslední důležitou branou, která nebyla zmíněna, je Hadamardova brána. Její konstrukce není intuitivní, proto přecházíme rovnou ke konečnému tvaru, viz [5]:

$$\lambda_H = \frac{1}{\sqrt{2}}(ff^\dagger - f^\dagger f + f + f^\dagger).$$

Všechny základní brány jsou zdefinovány a lze dojít k závěru, že každou bránu lze zapsat jako

$$\lambda = aff^\dagger + bf + cf^\dagger + df^\dagger f,$$

kde $a, b, c, d \in \mathbb{C}$ tak, že platí $a^2 + c^2 = b^2 + d^2 = 1$ a $\bar{b}a + \bar{d}c = 0$.

Tyto omezení zřetelně připomínají podmínky, které jsme specifikovali při definování obecné matice pro reprezentaci kvantové brány při maticové reprezentaci. Není to náhoda a také to ukazuje ekvivalenci mezi těmito dvěma typy reprezentace. Můžeme dokonce vyjádřit vztah

$$aff^\dagger + bf + cf^\dagger + df^\dagger f \longleftrightarrow \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

4.5. Multi qubity a jejich brány v GA

Připomeňme si nejprve, že pro bázi komplexní algebry $\mathbb{C}_{2^n}$ používáme prvky $(f_1, f_1^\dagger, \dots, f_n, f_n^\dagger)$ a s nimi konstruujeme 2^{2^n} kombinací jejich geometrického součinu. To vede k dvěma primitivním idempotentům I a K . Idempotent I bude opět použit

$$I = I_1 \dots I_n = f_1 f_1^\dagger \dots f_n f_n^\dagger,$$

kde $I_i = f_i f_i^\dagger$.

Nyní bude zavedena definice obecného n -qubitu. Realizace je provedena v prostoru dimenze $N = 2^{2^n}$ s ortonormální bází

$$|i_1 \dots i_n\rangle = (f_1^\dagger)^{i_1} \dots (f_n^\dagger)^{i_n} I,$$

kde $i_1, \dots, i_n \in \{0, 1\}$.

Následujícím krokem je zkonstruování bran pro multi qubity. Základní brány budou zapsány ve formě matice. Brány reprezentující matice jsou obvykle řídké. Proto není obtížné zapsat matici ve formě sumy, ve které má každá matice pouze jednu nenulovou hodnotu:

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Každá matice v sumě je nyní zapsána ve formě vnějšího součinu. A vnější součin lze velmi jednoduše přepsat do GA. Také víme, že stav $\langle i_1 i_2 |$ je pouze hermitovskou konjugací stavu $|i_1 i_2\rangle$. Proto můžeme odvodit

$$\begin{aligned} & |00\rangle\langle 00| + |01\rangle\langle 01| + |11\rangle\langle 10| + |10\rangle\langle 11| \\ &= II^\dagger + f_2^\dagger II^\dagger f_2 + f_1^\dagger f_2^\dagger II f_1 + f_1^\dagger II^\dagger f_2 f_1 \\ &= I + f_2^\dagger I f_2 + f_1^\dagger f_2^\dagger I f_1 + f_1^\dagger I f_2 f_1 \\ &= f_1 f_1^\dagger f_2 f_2^\dagger + f_1 f_1^\dagger f_2^\dagger f_2 - f_1^\dagger f_1 f_2^\dagger - f_1^\dagger f_1 f_2 \\ &= f_1 f_1^\dagger - f_1^\dagger f_1 (f_2^\dagger + f_2) \end{aligned}$$

Forma ostatních bran pro více qubitů je získána stejným způsobem. Tedy

$$\begin{aligned} \lambda_{CNOT} &= f_1 f_1^\dagger - f_1^\dagger f_1 (f_2^\dagger + f_2), \\ \lambda_{CZ} &= f_1 f_1^\dagger + f_1^\dagger f_1 (f_2 f_2^\dagger - f_2^\dagger f_2), \\ \lambda_{SWAP} &= f_1 f_1^\dagger f_2 f_2^\dagger + f_1^\dagger f_1 f_2^\dagger f_2 + f_1^\dagger f_2 - f_1 f_2^\dagger. \end{aligned} \tag{4.6}$$

Tato konstrukce lze použít pro libovolnou bránu. Maticová reprezentace je zapsána ve formě vnějšího součinu a pak převedena do GA.

4.6. Paralelní brány

Podobně jako v maticové reprezentaci i v GA reprezentaci je třeba vyřešit, jak reprezentovat paralelní brány působící na multi qubity. Řešením bude použití tenzorového součinu.

Definice 4.4 (Tensorový součin bran). Tensorový součin $\lambda_1 \otimes \dots \otimes \lambda_n$, kde $\lambda_k \in \{f_k f_k^\dagger, f_k^\dagger f_k, f_k, f_k^\dagger\}$ pro každé $k = 1, \dots, n$, je reprezentován geometrickým součinem $(-1)^s \lambda_1 \dots \lambda_n$, kde znaménko je určeno kardinalitou množin S_i , tak že $s = \sum_i |S_i|$, kde

$$S_i = \left\{ l < i : \lambda_l = f_l \text{ nebo } \lambda_l = f_l^\dagger f_l \right\} \text{ v případě, že } \lambda_i = f_i \text{ nebo } \lambda_i = f_i^\dagger.$$

4.7. Měření

Poslední věcí, kterou je třeba zavést, je měření. Pro výpočet pravděpodobnosti určitého výsledku v maticové reprezentaci jsme pouze provedli vnitřní součin určitého výsledku s transformovaným qubitem. A druhá mocnina tohoto vnitřního součinu byla konkrétní pravděpodobnost. V GA se použije takzvaná skalární projekce, označovaná jako $[\cdot]_0$. Skalární projekce vezme pouze skalární část multivektoru. Můžeme pro ní použít následující vzorec

$$\langle \rho | \psi \rangle = 2^n [\rho^\dagger \psi]_0,$$

kde n je počet qubitů. Tento faktor je třeba zavést kvůli našemu výběru báze a kvůli spinorové povaze reprezentace, kterou zde nebudeme diskutovat podrobněji. Podívejme se na příklad, jak tato skalární projekce funguje:

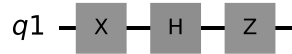
$$\langle 0|1\rangle = 2^1 [ff^\dagger f^\dagger]_0 = 0, \quad \langle 0|0\rangle = 2^1 [ff^\dagger]_0 = 1$$

Výsledky jsou dle očekávání. Připomeňme pouze, že $ff^\dagger = \frac{1}{2} + f \wedge f^\dagger$ a tedy $[ff^\dagger]_0 = \frac{1}{2}$.

4.8. Příklad obvodu

V této kapitole si ukážeme některé jednoduché příklady, jak reprezentovat kvantové obvody v GA. Začneme příkladem sériového sestavení bran, kde se nesetkáme s tenzorovým součinem. Vždy budeme porovnávat maticovou reprezentaci a GA reprezentaci.

Příklad 4.5 (Sériové sestavení bran). První příklad ukazuje obvod, kde jsou všechny brány umístěny na jedno vlákno, jak lze vidět na Obrázku 4.1.



Obrázek 4.1. Sériové sestavení bran.

Začneme s maticovou reprezentací. Všechny matice se vynásobí a získáme matici reprezentující obvod

$$M = ZHX = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}. \quad (4.7)$$

Pro libovolný qubit nyní můžeme spočítat pravděpodobnost, že qubit skončí v konkrétním bázeovém stavu. Qubit je vynásoben maticí obvodu, a poté se provede vnitřní součin s konkrétním bázeovým stavem. Budeme používat bázeové qubity $|\psi\rangle = |0\rangle$ a $|\rho\rangle = |1\rangle$. Spočítáme transformovaný qubit a pak provedeme vnitřní součin s bázeovým stavem pro získání pravděpodobností:

$$\begin{aligned} |\psi'\rangle &= M|\psi\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \\ |\rho'\rangle &= M|\rho\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\ p(|0\rangle) &= |\langle 0|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right|^2 = \left| -\frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}, \\ p(|1\rangle) &= |\langle 1|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}, \\ p(|0\rangle) &= |\langle 0|\rho'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}, \end{aligned}$$

$$p(|1\rangle) = |\langle 1|\rho'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}.$$

Pro oba vektory je součet pravděpodobností roven 1. Omezení dáno druhým axiomem pravděpodobnosti je splněno, a můžeme dospět k závěru, že naše výpočty jsou správné:

$$p(|0\rangle) + p(|1\rangle) = \frac{1}{2} + \frac{1}{2} = 1.$$

Nyní si ukážme alternativu. Obvod bude reprezentován pomocí aparátu GA. Pak

$$\begin{aligned} \lambda_M &= \lambda_Z \lambda_H \lambda_X = (ff^\dagger - f^\dagger f) \frac{1}{\sqrt{2}} (ff^\dagger - f^\dagger f + f + f^\dagger) (f^\dagger + f) \\ &= \frac{1}{\sqrt{2}} (ff^\dagger ff^\dagger - ff^\dagger f^\dagger f + ff^\dagger f + ff^\dagger f^\dagger) (f^\dagger + f) \\ &= \frac{1}{\sqrt{2}} (ff^\dagger + f + f^\dagger f - f^\dagger) (f^\dagger + f) \\ &= \frac{1}{\sqrt{2}} (ff^\dagger f^\dagger + ff^\dagger f + ff^\dagger + ff = f^\dagger ff^\dagger + f^\dagger ff - f^\dagger f^\dagger - f^\dagger f) \\ &= \frac{1}{\sqrt{2}} (f^\dagger - f^\dagger f + f + ff^\dagger). \end{aligned}$$

Tento výraz může být považován za ekvivalent k finální matici získané v maticové reprezentaci. Nyní budou spočteny transformované qubity:

$$\begin{aligned} |\psi'\rangle &= \lambda_M |\psi\rangle = \frac{1}{\sqrt{2}} (f^\dagger - f^\dagger f + f + ff^\dagger) (ff^\dagger) \\ &= \frac{1}{\sqrt{2}} (f^\dagger ff^\dagger - f^\dagger fff^\dagger + fff^\dagger + ff^\dagger ff^\dagger) \\ &= \frac{1}{\sqrt{2}} (f^\dagger + fff^\dagger), \\ |\rho'\rangle &= \lambda_M |\rho\rangle = \frac{1}{\sqrt{2}} (f^\dagger - f^\dagger f + f + ff^\dagger) f^\dagger \\ &= \frac{1}{\sqrt{2}} (f^\dagger f^\dagger - f^\dagger fff^\dagger + ff^\dagger + ff^\dagger f^\dagger) \\ &= \frac{1}{\sqrt{2}} (ff^\dagger - f^\dagger). \end{aligned}$$

Můžeme přistoupit k výpočtu pravděpodobnosti změření qubitů v konkrétních báзовých stavech:

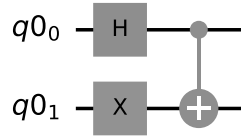
$$\begin{aligned} p(|0\rangle) &= |\langle 0|\psi'\rangle|^2 = \left| 2 \left[ff^\dagger \frac{1}{\sqrt{2}} (f^\dagger + fff^\dagger) \right]_0 \right|^2 = \left| \frac{2}{\sqrt{2}} [ff^\dagger f^\dagger + ff^\dagger fff^\dagger]_0 \right|^2 \\ &= \left| \frac{2}{\sqrt{2}} [ff^\dagger]_0 \right|^2 = \left| \frac{2}{\sqrt{2}} \frac{1}{2} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}. \end{aligned}$$

Ostatní pravděpodobnosti jsou spočteny stejným způsobem a proto je ukázána pouze zkrácená verze:

$$\begin{aligned} p(|1\rangle) &= |\langle 1|\psi'\rangle|^2 = \left| 2 \left[f \frac{1}{\sqrt{2}} (f^\dagger + f f^\dagger) \right]_0 \right|^2 = \frac{1}{2}, \\ p(|0\rangle) &= |\langle 0|\rho'\rangle|^2 = \left| 2 \left[f f^\dagger \frac{1}{\sqrt{2}} (f f^\dagger - f^\dagger) \right]_0 \right|^2 = \frac{1}{2}, \\ p(|1\rangle) &= |\langle 1|\rho'\rangle|^2 = \left| 2 \left[f \frac{1}{\sqrt{2}} (f f^\dagger - f^\dagger) \right]_0 \right|^2 = \frac{1}{2}. \end{aligned}$$

Nyní můžeme ověřit, že součet pravděpodobností je pro oba vektory 1. Pravděpodobnosti spočítané pomocí GA reprezentace jsou stejné jako pravděpodobnosti spočítané pomocí maticové reprezentace. GA reprezentace se ukazuje jako korektní.

Příklad 4.6 (Paralelní sestavení bran). V tomto příkladu je zkoumáno chování paralelního sestavení bran. Na jednom vlákne je Hadamardova brána a na druhém X-brána. Poté je použita CNOT brána, přičemž první qubit je controlled qubitem, viz Obrázek 4.2.



Obrázek 4.2. Obvod s CNOT bránou.

Nejprve bude ukázána maticová reprezentace. Bude proveden tenzorový součin Hadamardovy brány a X-brány. Forma CNOT brány je známá, viz rovnice 4.6:

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

a

$$\begin{aligned} H \otimes X &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & 1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ 1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & -1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix}. \end{aligned}$$

Finální matice je získána součinem CNOT brány s $H \otimes X$, tj.

$$\begin{aligned} M = \text{CNOT}(H \otimes X) &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \end{aligned}$$

Ukážeme si účinek obvodu na dvou příkladech qubitů, a to konkrétně na $\psi = |00\rangle$ a $\rho = |01\rangle$:

$$\begin{aligned} |\psi'\rangle = M|\psi\rangle &= \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \\ |\rho'\rangle = M|\rho\rangle &= \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Získali jsme transformované qubity a můžeme provést měření:

$$\begin{aligned} p(|00\rangle) &= |\langle 00|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \right|^2 = 0, \\ p(|01\rangle) &= |\langle 01|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}, \\ p(|10\rangle) &= |\langle 10|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2}, \\ p(|11\rangle) &= |\langle 11|\psi'\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \right|^2 = 0. \end{aligned}$$

Měření druhého qubitu bude zkráceno.

$$\begin{aligned} p(|00\rangle) &= |\langle 00|\rho'\rangle|^2 = \frac{1}{2}, & p(|01\rangle) &= |\langle 01|\rho'\rangle|^2 = 0, \\ p(|10\rangle) &= |\langle 10|\rho'\rangle|^2 = 0, & p(|11\rangle) &= |\langle 11|\rho'\rangle|^2 = \frac{1}{2}. \end{aligned}$$

Jednoduše lze ověřit že součet pravděpodobností je 1 pro oba případy.

Tentýž obvod je nyní realizován ve formě GA. Je proveden tenzorový součin Hadamardovy brány a X-brány. Forma CNOT hradla je známa, viz (4.6). Proto

$$\begin{aligned} \lambda_{HX} &= \lambda_H \otimes \lambda_X = \frac{1}{\sqrt{2}}(f_1 f_1^\dagger - f_1^\dagger f_1 + f_1 + f_1^\dagger)(f_2^\dagger + f_2), \\ &= \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1 f_1^\dagger f_2 + f_1^\dagger f_1 f_2^\dagger + f_1^\dagger f_1 f_2 - f_1 f_2^\dagger - f_1 f_2 + f_1^\dagger f_2^\dagger + f_1^\dagger f_2) \end{aligned}$$

$$\lambda_{CNOT} = f_1 f_1^\dagger - f_1^\dagger f_1 f_2^\dagger - f_1^\dagger f_1 f_2.$$

Konečná reprezentace obvodu je získána jednoduše vynásobením výrazů λ_{HX} a λ_{CNOT} . S využitím pravidel uvedených v (4.3), (4.4) a (4.5) můžeme zjednodušit na finální tvar

$$\begin{aligned} \lambda_M &= \lambda_{CNOT} \lambda_{HX} \\ &= \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1 f_1^\dagger f_2 - f_1 f_2^\dagger - f_1 f_2 - f_1^\dagger f_1 f_2^\dagger f_2 + f_1^\dagger f_2^\dagger f_2 \\ &\quad + f_1^\dagger f_1 f_2 f_2^\dagger + f_1^\dagger f_2 f_2^\dagger). \end{aligned}$$

Všimněme si, že v maticové reprezentaci má finální matice 8 nenulových prvků. Výraz v GA reprezentaci má také 8 prvků. Pro ověření správnosti vypočítáme pravděpodobnosti výsledků měření pro qubity $\psi = |00\rangle$ a $\rho = |01\rangle$:

$$\begin{aligned} |\psi'\rangle &= \lambda_M |\psi\rangle = \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1 f_1^\dagger f_2 - f_1 f_2^\dagger - f_1 f_2 - f_1^\dagger f_1 f_2^\dagger f_2 + f_1^\dagger f_2^\dagger f_2 \\ &\quad + f_1^\dagger f_1 f_2 f_2^\dagger + f_1^\dagger f_2 f_2^\dagger)(f_1 f_1^\dagger f_2 f_2^\dagger) = \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1^\dagger f_2 f_2^\dagger), \\ |\rho'\rangle &= \lambda_M |\rho\rangle = \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1 f_1^\dagger f_2 - f_1 f_2^\dagger - f_1 f_2 - f_1^\dagger f_1 f_2^\dagger f_2 + f_1^\dagger f_2^\dagger f_2 \\ &\quad + f_1^\dagger f_1 f_2 f_2^\dagger + f_1^\dagger f_2 f_2^\dagger)(f_1 f_1^\dagger f_2 f_2^\dagger) = \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2 f_2^\dagger + f_1^\dagger f_2^\dagger). \end{aligned}$$

K výpočtu pravděpodobností jednotlivých výsledků použijeme skalární projekci. Je důležité změnit člen před skalární projekcí. Tento příklad pracuje se dvěma qubity, proto změníma na $2^n = 2^2 = 4$ a dostaneme

$$\begin{aligned} p(|00\rangle) &= |\langle 00|\psi'\rangle|^2 = \left| 4 \left[f_1 f_1^\dagger f_2 f_2^\dagger \frac{1}{\sqrt{2}}(f_1 f_1^\dagger f_2^\dagger + f_1^\dagger f_2 f_2^\dagger) \right]_0 \right|^2 \\ &= \left| \frac{4}{\sqrt{2}} [0]_0 \right|^2 = 0, \end{aligned}$$

$$\begin{aligned}
p(|01\rangle) &= |\langle 00|\psi'\rangle|^2 = \left| 4 \left[f_1 f_1^\dagger f_2 \frac{1}{\sqrt{2}} (f_1 f_1^\dagger f_2^\dagger + f_1^\dagger f_2 f_2^\dagger) \right]_0 \right|^2 \\
&= \left| \frac{4}{\sqrt{2}} \left[f_1 f_1^\dagger f_2 f_2^\dagger \right]_0 \right|^2 = \left| \frac{4}{\sqrt{2}} \frac{1}{4} \right|^2 = \left| \frac{1}{\sqrt{2}} \right|^2 = \frac{1}{2},
\end{aligned}$$

$$p(|10\rangle) = |\langle 10|\psi'\rangle|^2 = \frac{1}{2},$$

$$p(|11\rangle) = |\langle 11|\psi'\rangle|^2 = 0.$$

Součet všech pravděpodobností je 1, což je správně. Nyní můžeme porovnat výsledky s těmi získanými pomocí maticové reprezentace. Všechny pravděpodobnosti jsou stejné. Pro druhý multi qubit je postup analogický a ukázány jsou pouze konečné pravděpodobnosti:

$$\begin{aligned}
p(|00\rangle) &= |\langle 00|\rho'\rangle|^2 = \frac{1}{2}, & p(|01\rangle) &= |\langle 01|\rho'\rangle|^2 = 0, \\
p(|10\rangle) &= |\langle 10|\rho'\rangle|^2 = 0, & p(|11\rangle) &= |\langle 11|\rho'\rangle|^2 = \frac{1}{2}.
\end{aligned}$$

V případě druhého multi qubitu je součet pravděpodobností také 1. Jsou stejné jako v maticové reprezentaci, a lze tedy dospět k závěru, že reprezentace pomocí GA je ve všech směrech správná a funkční.

Je zřejmé, že s vyšším počtem vstupních qubitů se rozměr matic exponenciálně zvětšuje. Výpočet se stává velmi náročným, protože pracujeme s obrovskými maticemi. Reprezentace v GA je velmi intuitivní a umožňuje mnoho zjednodušení během výpočtu. Počet prvků reprezentujících kvantové stavy nebo kvantové brány se také zvětšuje, ale obvykle se výraz během výpočtu zjednoduší a mnoho prvků zcela vypadne. Je také třeba poznamenat, že geometrický součin některých prvků v GA je roven 0. Toto vlastnost zjednodušuje výpočet a také šetří spoustu paměti.

Největší problém s GA reprezentací je ten, že neexistuje žádný software pro přímou implementaci. GAALOPWeb nabízí některé možnosti, ale s mnoha omezeními. Při vhodné implementaci lze pomocí GA realizovat QC velmi efektivně. Ekvivalence maticové a GA reprezentace také umožňuje použití vhodnější reprezentace pro konkrétní problém.

5. SHRNUÍ

V první kapitole se definují klíčové pojmy z algebry a jsou dány do patřičného kontextu. Obecně je popsán tenzorový součin a tenzorové prostory. Je odvozena Cliffordova algebra a diskutovány všechny její důležité vlastnosti, [6].

Druhá kapitola začíná definicí qubitu. Zdůrazněn je rozdíl mezi klasickým bitem a qubitem. Abychom mohli řešit problém s vizualizací qubitu, snížíme počet stupňů volnosti na dva a poté přepíšeme definici qubitu do sférických souřadnic. To umožní definovat tzv. Blochovu sféru, která se používá k vizualizaci různých qubitů a jejich transformací. Definice qubitu je rozšířena na multi qubity pomocí Kroneckerova součinu. V jednoduchosti je také popsáno, jak se provádí měření. Tato kapitola je založena na znalostech získaných z [2, 7, 13].

Další kapitola ukazuje, jak odvodit obecnou kvantovou bránu a jaké vlastnosti musí splňovat. Jsou definovány základní brány a demonstrován jejich účinek na konkrétní qubity. Definice brány pro jeden qubit je rozšířena na bránu pro multi qubity pomocí Kroneckerova součinu.

Hlavní část práce aplikuje znalosti a konstruuje aparát geometrické algebry (GA) vhodné pro kvantové výpočty (QC). Začínáme definicí GA nad reálnými čísly. Zvláštní pozornost je věnována geometrickému součinu, kterým je tato algebra vybavena. Tento součin hraje klíčovou roli po celou práci, a proto je pečlivě odvozen z vnitřního a vnějšího součinu. Jsou zdůrazněny důležité vlastnosti geometrického součinu. Abychom však mohli správně reprezentovat qubity v GA, musíme do naší definice zahrnout i komplexní čísla. Nastavení základního pole na komplexní čísla nestačí, protože komplexní chování musí být přeneseno přímo do algebry. Proto je představena ortogonální lineární transformace. Báze zkonstruovaná pomocí této transformace se nazývá Wittova báze a hraje velmi důležitou roli. Jsou zkoumány důležité vlastnosti Wittovy báze pro účely dalších výpočtů. Další užitečné informace k této konstrukci GA lze nalézt v [5].

S vhodným aparátem GA můžeme přejít k definici qubitu. Klíčovým konceptem je identifikovat qubity s prvky GA. Je popsána obecná definice qubitu a poté rozšířena na multi qubity. Kvantové brány jsou také identifikovány s prvky GA, což umožňuje výpočet v rámci jedné algebraické struktury. Je prozkoumán vztah mezi maticovou a GA reprezentací. Poté je pečlivě odvozena forma kvantových bran zapsaných v GA. Je ukázán univerzální způsob, jak sestavit libovolnou bránu pro multi qubity. Stručně diskutujeme také to, jak provádět měření v GA.

Povedlo se sestavit aparát GA, který umožňuje intuitivní a přímočarý výpočet. Všechny získané výsledky potvrdily správnost tohoto přístupu. Elegantní definice báze umožňuje mnoho zjednodušení během výpočtu, což ho dělá rychlejší a potenciálně i méně náročným na paměť. Maticová reprezentace má problém se zvyšujícím se rozměrem matic. Tento růst je exponenciální a i pro malý počet qubitů může být výpočet velmi náročný. Rozměr výrazů reprezentujících kvantové brány v GA také roste, ale není tak dramatický a často se dá zjednodušit. Největším problémem je samotná implementace GA. Komplexní čísla jsou obtížně reprezentovatelná a v současné době neexistuje žádný software vhodný pro výpočty v GA. GAALOPWeb nabízí alternativu, ale pracuje s GA nad reálnými čísly a používá trochu odlišný přístup.

REFERENCE

- [1] M. E. Browne: *Schaum's outlines physics for engineering and science*, McGraw-Hill, New York, 2010.
- [2] F. De Lima Marquezino, et al.: *A primer on quantum computing*, SpringerBriefs in Computer Science, 2019.
- [3] L. Dorst, D. H. F. Fontijne, S. Mann: *Geometric algebra for computer science: an object-oriented approach to geometry*, Morgan Kaufmann Publishers, 2007.
- [4] D. Hildenbrand: *Introduction to geometric algebra computing*, Chapman and Hall/CRC, 2018.
- [5] J. Hrdina, A. Návrat, P. Vašík: *Quantum computing based on complex Clifford algebras*, 2022, online <https://doi.org/10.1007/s1128-022-03648-w>.

- [6] M. S. Lane, G. Birkhoff: *Algebra*, American Mathematical Society, Providence, 1999.
- [7] M. A. Nielsen, I. L. CHUANG: *Quantum computation and quantum information*, Cambridge University Press, Cambridge, 2010.
- [8] C. Perwass: *Geometric algebra with applications in engineering*, Springer, 2009.
- [9] L. Procházka: *Inverzní kinematika robotického ramene s předepsanou trajektorií efektoru pomocí geometrické algebry*, Bakalářská práce, Vysoké učení technické v Brně, Fakulta strojního inženýrství, Brno, 2022.
- [10] E. M. Purcell, D. Morin: *Electricity and magnetism*, Cambridge University Press, Cambridge, 2018.
- [11] Open-source quantum development, online <https://qiskit.org/>, accessed Apr 28, 2023.
- [12] J. Rue, S. Xambo: *Mathematical essentials of quantum computing*, Lecture notes UPC, online <https://web.mat.upc.edu/sebastia.xambo/QC/qc.pdf>.
- [13] *Theory of computing*, ACM Press, New York, 1993.

Jan Michálek, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 61669 Brno, Česká republika,
e-mail: 228579@vutbr.cz

REGRESSION TREES AND RANDOM FORESTS FOR PREDICTIONS

DUŠAN OBERTA

ABSTRAKT. Regression trees are widely used in statistics to capture, not always trivial, relationships between predictors (i.e. independent variables) and a response variable (i.e. dependent variable). They can be used in a variety of situations where other statistical tools are not suitable, even in situations where the number of predictors is greater than the number of observations in the set of training data. Random forests generalize the concept of regression trees to reduce variance and improve stability of simple regression trees. Apart from the classical regression trees based on the least squares method, the concept of maximum likelihood with the assumption of gamma distribution of the response variable is described and derived by the author. Compared to literature found, slightly different proofs of theorems regarding pruning of regression trees are offered, as well as a thorough derivation of confidence intervals for the expected value of the response variable is offered as own work of the author. Introduction to the concept of random forests is covered in the last part of the article.

1. INTRODUCTION

Nowadays, we are surrounded by huge amounts of data. Considering the fact that more and more data are obtained every day from numerous human activities, it is crucial to know how to process the data and model the relationship between, usually multiple, predictor variables and a response variable properly. Predictor variable (i.e. predictor), also known as an explanatory variable, is an independent variable that can be changed by an observer in order to obtain an outcome (i.e. response variable). Predictors can be either real-valued (called continuous predictors), or they can take only the values from a finite set (in this case, they are called categorical predictors). A response variable is a dependent variable, which is influenced by predictors. Our goal is to create a statistical model which best

2020 *MSC*. Primární 62F40; Sekundární 62F25, 62H30, 68T05.

Klíčová slova. Regression trees, least squares, maximum likelihood estimation, confidence intervals, k-fold cross-validation, bagging, bootstrapping, random forests.

Článok vznikol na základe bakalárskej práce autora v obore Matematické inžinýrství na FSI VUT v Brne. Vedúcou práce bola Zuzana Hübnerová z Ústavu matematiky FSI VUT v Brně.

Ďakujem doc. Hübnerovej za ochotu, pomoc a priateľský prístup počas písania bakalárskej práce, v ktorej sa pojednáva (nielen) o problematike regresných stromov a náhodných lesov. Taktiež by som sa chcel poďakovať svojej rodine za podporu počas štúdia aj mimo neho.

captures the relationship, based on some observations made (i.e. a set of training data).

Imagine we know that a response variable Y is explained by an explanatory variable X by the relationship $Y = \alpha + \beta X + \varepsilon$. Based on a set of n observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$ we have done, we want to compute the estimations $\hat{\alpha}, \hat{\beta}$ of the real values of parameters α, β . In real world, each of the observations satisfies the equation $y_i = \alpha + \beta x_i + \epsilon_i$ (for $i = 1, \dots, n$), where ϵ_i is a random variable known as a random error, which is the difference between the observed and the real value of an experiment. Throughout this article we assume that $\epsilon_1, \dots, \epsilon_n$ are independent, meaning that a random error which occurs while making one observation does not affect any other observations made. This is quite a reasonable assumption, anyway it is not always fulfilled in real world applications. The case where the random errors are not independent is not discussed in this article, and more information regarding this topic can be found in [5]. Also, when deriving confidence intervals for the expected value of the response variable for *regression trees* (see Section 2.6), we assume that $\epsilon_1, \dots, \epsilon_n$ are normally distributed with zero mean and constant variance. Moreover, it is definitely a reasonable requirement that the computed estimations $\hat{\alpha}, \hat{\beta}$ are the “best” (in some sense), whatever the “best” means. Depending on the “best” criterion used, we obtain different statistical models. The two “best” criteria used for *regression trees* in this article are a well-known least squares criterion (see Section 2.3) and a maximum likelihood criterion (see Section 2.7) with the assumption of gamma distribution of the response variable (for simplicity, the words “assumption of gamma distribution” will be sometimes omitted, since in this article we will only consider maximum likelihood estimation with the assumption of gamma distribution of the response variable).

Probably the best-known tool for making predictions is *linear regression*. Being the best-known comes with certain limitations, such as the relationship between predictors and the expected value of the response variable must be linear. In order to derive interval estimates of unknown parameters using *linear regression*, normal distribution of the response variable is assumed. Generalization of *linear regression* is offered by *generalized linear models*, where the distribution of the response variable can be any distribution from the so called *exponential family*, and the expected value of the response variable is connected to the predictors by a link function, which is a strictly monotone, differentiable function. Nevertheless, there are still some limitations concerning this statistical models (e.g. the number of observations must be at least equal to the number of unknown parameters). More detailed information regarding *linear regression* can be found in [1] or [4], and regarding the *generalized linear models*, in [2].

Regression trees, described in Section 2, have no limitations concerning the relationship between predictors and the response variable. Also, as mentioned in the abstract, there are no limitations regarding the number of unknown parameters and the number of observations, which is not true for both *linear regression* and

generalized linear models. They are also known for their simplicity and interpretability. On the other hand, they might be unstable, meaning that a slightly different data set might result in quite different predictions. Two different approaches of growing the *trees* are described, least squares (see Section 2.3) and maximum likelihood estimation (see Section 2.7), with the latter being own derivation of the author. Confidence intervals for the expected value of the response variable (assuming a fixed partition of the sample space) are derived thoroughly in Section 2.6. Also, *k-fold cross-validation* is shown as a tool for finding optimal values of tuning parameters (not only) for *regression trees*. More information regarding *regression trees* can be found in [4].

As mentioned in the previous paragraph, *regression trees* tend to be unstable. Section 3 provides an introduction to the concept of *random forests*, which are built using the basic concepts of bootstrapping, bagging and *regression trees* in order to reduce variance of the model when compared to single *regression trees* and improve its stability. As in the previous section, confidence intervals for the expected value of the response variable (assuming a fixed partition of the sample space) are stated, with a sketch of their derivation. Also specific algorithm for the practical implementation of *random forests* is described. More detailed information regarding *random forests* can be found in [4] and [8], and regarding the confidence intervals in [3].

2. REGRESSION TREES

In this chapter, we will introduce some basic concepts regarding *regression trees*. Proper mathematical definition of a *tree* (see Definition 2.1) requires defining some terms from graph theory. However, for purposes of this article, more intuitive and vague definitions of the terms and concepts (e.g. Definition 2.2) will be sufficient.

Regression trees are used to model a relationship between predictors and a response variable. They are, for example, suitable to use in situations where either the relationship is too complicated to capture by a simple model, or the number of independent variables is relatively large when compared to the number of observations.

2.1. Basic Concepts

Definition 2.1. A *tree* is an *undirected graph*, in which any two *vertices* are connected by exactly one *path*.

Definition 2.2. A *binary tree* T consists of a non-empty set of *nodes*, such that each *node* (called *parent node*) contains either no *subnodes*, or precisely two *subnodes* (called *child nodes*), and that there is exactly one *node* (called the *root*), which is not a *child node* of any other *node*. *Nodes* containing no *child nodes* are called the *leaves* or *terminal nodes*.

Remark. Since we are interested only in *binary trees*, for simplification, instead of the term *binary tree*, we will be using only the term *tree* (omitting the word “binary”).

2.2. Introduction to Regression Trees

The main idea behind *regression trees* is that by starting with all the data at the *root*, we move downwards until a *leaf* is reached. At each *parent node*, we move towards one of its two *child nodes* depending on the values of predictors – we select a predictor and its value, which provide the “best” partition (according to some criterion), and then each data sample, which was originally in the *parent node* is moved into the corresponding *child node* (depending on the selected “best” predictor and its value). The selected predictor is called a *split predictor*, its value is called a *split value*, and together they are called a *split*. We stop the splitting process when some minimal *node size* (i.e., the minimal number of observations in a *node*) is reached.

Let $X_1, \dots, X_l$ be continuous predictors and $X_{l+1}, \dots, X_k$ be categorical predictors with $m_{l+1}, \dots, m_k$ categories (i.e., $X_i \in \mathbb{R}$ and $X_j \in G_j := \{g_j^1, \dots, g_j^{m_j}\}$, for $i = 1, \dots, l$ and $j = l+1, \dots, k$). Denote the sample space

$$\mathcal{D} := \mathbb{R}^l \times \prod_{i=l+1}^k G_i.$$

Let $\{R_i\}_{i=1}^m$ be a finite partition of $\mathcal{D}$, such that for $j = 1, \dots, m$, each region R_j is of the form

$$R_j = \prod_{i=1}^l (a_i; b_i] \times \prod_{i=l+1}^k S_i,$$

where S_i is a non-empty subset of G_i and $a_i, b_i \in \mathbb{R} \cup \{-\infty, \infty\}$ (in case $b_i = \infty$, the corresponding interval is considered to be open from the right side).

Let $Y_1, \dots, Y_n$ be random variables, $\mathbf{X} \in \mathcal{D}^n$ an $n \times k$ matrix (each of the n rows of $\mathbf{X}$ being an element of $\mathcal{D}$), $\beta_1, \dots, \beta_m$ unknown parameters and $\varepsilon_1, \dots, \varepsilon_n$ random variables such that $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$, for $i = 1, \dots, n$. Denote $\mathbf{x}_i := (x_{i1}, \dots, x_{ik})^T \in \mathcal{D}$ and $\mathbf{X}_j := (x_{1j}, \dots, x_{nj})^T$. We can write $\mathbf{X}$ in the form

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ & \ddots & \\ x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = (\mathbf{X}_1 \quad \dots \quad \mathbf{X}_k).$$

Now assume a regression model, where only a constant function is fitted in each of the partition regions, i.e. the corresponding model is of the form

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i = \sum_{j=1}^m \beta_j \chi_{R_j}(\mathbf{x}_i) + \varepsilon_i; \quad i = 1, \dots, n, \quad (2.1)$$

where χ_{R_j} is an indicator function of subset R_j of set $\mathcal{D}$, i.e.

$$\chi_{R_j}(\mathbf{x}) := \begin{cases} 1, & \text{if } \mathbf{x} \in R_j, \\ 0, & \text{else.} \end{cases}$$

Remark. The model defined in (2.1) is typically used for *regression trees*, when the partition is done according to the least squares criterion (see Section 2.3).

When considering the maximum likelihood criterion with the assumption of gamma distribution of the response variable (see Section 2.7), this model is no longer valid.

Remark. The partition $\{R_i\}_{i=1}^m$ described in this section simply means, that the $\mathbb{R}^l$ part of the sample space is divided into l -dimensional rectangles for continuous predictors, and the finite sets corresponding to categorical predictors are divided into non-empty subsets.

2.3. Growing a Regression Tree

2.3.1. Parameters Estimation. Consider a set of training data

$$\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\},$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T$, for $i = 1, \dots, n$. Notice that there are no assumptions regarding the relationship between n and k . Suppose, in this section, that we already have a partition of $\mathcal{D}$ into m regions $R_1, \dots, R_m$. Assuming that the model is of the form (2.1), we want to estimate the values of parameters $\beta_1, \dots, \beta_m$. Denote $f_i := f(\mathbf{x}_i)$ and consider the criterion of minimizing the residual sum of squares, i.e. minimizing the expression

$$RSS := RSS(\beta_1, \dots, \beta_m) = \sum_{i=1}^n (y_i - f_i)^2 = \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n y_i f_i + \sum_{i=1}^n f_i^2. \quad (2.2)$$

Differentiating both sides of (2.2) with respect to β_u and setting it to 0, we obtain

$$\frac{\partial RSS}{\partial \beta_u} = \dots = -2 \sum_{\{i | \mathbf{x}_i \in R_u\}} y_i + 2 \text{card}(\{i | \mathbf{x}_i \in R_u\}) \beta_u = 0, \quad (2.3)$$

where $\text{card}(A)$ denotes the cardinality of set A . Denote $\mathcal{C}_u := \text{card}(\{i | \mathbf{x}_i \in R_u\})$. Assuming that $\{i | \mathbf{x}_i \in R_u\}$ is non-empty (i.e., each of the partition regions contains at least one element), (2.3) has a unique solution

$$\hat{\beta}_u = \frac{1}{\mathcal{C}_u} \sum_{\{i | \mathbf{x}_i \in R_u\}} y_i. \quad (2.4)$$

Hence $\hat{\beta}_u$ is just the average of $y_i | \mathbf{x}_i \in R_u$. It is an easy exercise to verify that the *Hessian* matrix of (2.2) is positive definite, thus the expression (2.4) minimizes (2.2).

Remark. Details of (2.3), as well as verifying the positive definiteness of the *Hessian* matrix of (2.2) can be found in [7].

2.3.2. Finding an Optimal Partition. Now, assuming (2.4) is the least squares estimator of β_u , we need to find an optimal partition of the sample space $\mathcal{D}$.

For a categorical predictor X_j ($j \in \{l+1, \dots, k\}$), define the set of all possible non-empty *binary splits* as

$$\mathcal{B}_j := \{\{S; G_j \setminus S\} | S \subset G_j \wedge S \neq \emptyset; G_j\}.$$

Denote J_j the number of all possible non-empty *binary splits* for a categorical predictor X_j

$$J_j := \text{card}(\mathcal{B}_j) = \frac{\text{card}(\mathcal{P}(G_j) \setminus \{\emptyset; G_j\})}{2} = \frac{2^{\text{card}(G_j)} - 2}{2} = 2^{\text{card}(G_j)-1} - 1,$$

where $\mathcal{P}(A)$ denotes the power set of A . Thus we can write $\mathcal{B}_j$ as

$$\mathcal{B}_j = \{B_j^i\}_{i=1}^{J_j} = \left\{ \left\{ B_j^{i;1}; B_j^{i;2} \right\} \right\}_{i=1}^{J_j},$$

where B_j^i denotes the i -th element of $\mathcal{B}_j$.

Remark. Remember that G_j denotes the set of all values a categorical predictor X_j can take. Then $\mathcal{B}_j$ is a set of unordered tuples, where each of the tuples contains a non-empty subset of G_j and its complement.

Finding a best binary partition in terms of minimal residual sum of squares is generally computationally infeasible. Instead, we use a greedy algorithm. Starting with all the data samples, consider a *split predictor* X_j ($j \in \{1, \dots, k\}$). For a *split predictor* X_j and *split point* s define $R_1(j, s)$ and $R_2(j, s)$ as

$$\begin{aligned} R_1(j, s) &:= \begin{cases} \{\mathbf{x} \mid x_j \leq s\}, & \text{if } X_j \text{ is a continuous predictor,} \\ \{\mathbf{x} \mid x_j \in B_j^{s;1}\}, & \text{if } X_j \text{ is a categorical predictor,} \end{cases} \\ R_2(j, s) &:= \begin{cases} \{\mathbf{x} \mid x_j > s\}, & \text{if } X_j \text{ is a continuous predictor,} \\ \{\mathbf{x} \mid x_j \in B_j^{s;2}\}, & \text{if } X_j \text{ is a categorical predictor,} \end{cases} \end{aligned} \quad (2.5)$$

where $s \in \mathbb{R}$ for continuous variable and $s \in \{1, \dots, J_j\}$ for categorical variable and x_j denotes the j -th element of $\mathbf{x}$.

It is obvious that for every pair (j, s) , $\{R_1(j, s); R_2(j, s)\}$ form a partition of $\mathcal{D}$. For classical *regression trees*, the index j of *split predictor* X_j and *split point* s are chosen as the solution of the optimization problem

$$\min_{j; s} \left(\min_{\beta_1, \dots, \beta_m} \sum_{i=1}^n (y_i - f_i)^2 \right). \quad (2.6)$$

For every pair (j, s) , the inner minimization of (2.6) is solved by (2.4). If for continuous predictors we restrict ourselves to the set of values taken by training data, the outer minimization can be simply solved by iterating through all the predictors and possible values of s . Finding the best split, we divide the data into two regions R_1 and R_2 and repeat the splitting process on both the regions. The whole process is then repeated again and again on all of the resulting regions, until the minimal *node size* (i.e. the number of observations in a *node*).

Remark. Note that we have found *an* optimal partition, not *the* optimal partition. Due to the fact that we have a finite number of observations, for the partition obtained in the previous paragraph, there is an (actually uncountable) infinite number of partitions minimizing (2.5), if for continuous predictors we do not restrict ourselves just to the set of values taken by our training data.

2.4. Pruning a Regression Tree

A very large *tree* might overfit the data, while a *tree* which is too small might not capture the important structure of the data. In this section, we will describe how to grow an optimal sized *tree*. Theorems stated in this section are stated and proved in [8], nevertheless the proofs of Theorem 2.10 and Theorem 2.11 were modified and more clarified by the author.

Definition 2.3. A *subtree* of a *tree* T is a *tree* T_S with *root* a *node* of T , such that each *node* of T_S is also a *node* of T . It is denoted as: $T_S \subseteq T$. Then T_S is called a *rooted subtree* of a *tree* T , if its *root* is the *root* of T .

Remark. If not specified otherwise or being clear from the context, the term *subtree* means the maximum possible *subtree* (in terms of the number of *nodes*).

Definition 2.4. A *branch* T_B at a *non-terminal node* t of a *tree* T is the *subtree* rooted at one of its *child nodes*.

Definition 2.5. The number of *leaves* of a *tree* T is called the *size* of a *tree*, i.e.

$$\text{size}(T) := \text{card}(\{t \mid t \text{ is a leaf of } T\}).$$

Instead of simply finding a *tree* with the minimal residual sum of squares, we need to consider also the *size* of the *tree*. For a given *tree* T and real number α , define the cost complexity criterion as

$$R_\alpha(T) := R(T) + \alpha \text{size}(T), \quad (2.7)$$

where $R(T)$ is the residual sum of squares as defined in (2.2).

Consider growing a large *tree* T_0 , stopping the splitting process only when some minimum *node size* (sample size at the specific *node*) is reached. The idea is to find for a given α , a *subtree* $T(\alpha) \subseteq T_0$ minimizing (2.7). Note that for $\alpha \leq 0$, the solution is the *full tree* T_0 . We will show that for a given *tree* T_0 , there is a nested sequence of *subtrees* $\{T_k\}_{k=0}^q$ (i.e. $T_q \subseteq \dots \subseteq T_0$) and an increasing sequence of real numbers $\{\alpha_k\}_{k=1}^q$ such that for $k = 1, \dots, q-1$, T_k is an *optimal tree* for $\alpha \in [\alpha_k; \alpha_{k+1})$, T_q is an *optimal tree* for $\alpha \geq \alpha_q$ and T_0 is an *optimal tree* for $\alpha < \alpha_1$. We will also provide an algorithm on how to construct the nested sequence $\{T_k\}_{k=0}^q$.

Remark. Consider a *node* t of a *tree* T . Values $R_\alpha(t)$, $R(t)$ and $\text{size}(t)$ are defined in the sense of $R_\alpha(T_{t_0})$, $R(T_{t_0})$ and $\text{size}(T_{t_0})$, where T_{t_0} is a *trivial subtree* (consisting only of the *root*) rooted at the *node* t . It is easy to see that $\text{size}(t) = 1$.

Remark. For a *node* t of a *tree* T , denote T_t the *subtree* of T rooted at t .

Definition 2.6. Consider a *tree* T . The reduction function $g(t, T)$ for a *non-terminal node* t of a *tree* T is defined as

$$g(t, T) := \frac{R(t) - R(T_t)}{\text{size}(T_t) - \text{size}(t)}. \quad (2.8)$$

Definition 2.7. *Pruning* of a *tree* T at a *non-terminal node* t is to replace T_t by t .

Lemma 2.8. *Consider a tree T and a non-terminal node t . Then for a given $\alpha \in \mathbb{R}$*

$$g(t, T) > \alpha \text{ if and only if } R_\alpha(t) > R_\alpha(T_t).$$

Proof. The proof follows directly from definitions (2.7) and (2.8). $\square$

Remark. For a non-terminal node t of a tree T it follows that:

$$R_\alpha(T_t) = \sum_{\{T_B | T_B \text{ is a branch at } t\}} R_\alpha(T_B).$$

Theorem 2.9. *Consider a tree T and $\alpha \in \mathbb{R}$. Suppose that we visit all the nodes of T in the bottom-up order (i.e. starting from the leaves, finishing with the root, and visiting each node before its parent) and prune at a non-terminal node t only if*

$$R_\alpha(t) \leq R_\alpha(T'_t), \quad (2.9)$$

for the current tree T' . Then the resulting tree is $T(\alpha)$.

Proof. The proof will be done using mathematical induction. Suppose that when a node t is considered, all the branches at t are optimally pruned. It is obvious that this is true for the leaves. At node t , we either prune it with the value $R_\alpha(t)$, or not with the value $R_\alpha(T'_t)$ if this is strictly smaller. If there is a subtree T''_t rooted at t with a smaller value of R_α , it must be non-trivial and there must be a branch T_B with $R_\alpha(T''_B) < R_\alpha(T'_B)$ and so T'_B is not optimally pruned, which is a contradiction. Thus after a node t is considered, T'_t is optimally pruned. After the root is considered, the resulting tree is optimally pruned, hence it is $T(\alpha)$. $\square$

Remark. Theorem 2.9 gives us an algorithm to find $T(\alpha)$ for a single $\alpha \in \mathbb{R}$.

Theorem 2.10. *Consider a tree T . Denote*

$$\alpha_1 := \min_{\{t | t \text{ is a non-terminal node of } T\}} g(t, T). \quad (2.10)$$

The optimally pruned tree is T for $\alpha < \alpha_1$ and $T_1 := T(\alpha_1)$ is obtained by pruning at all the nodes t with $g(t, T) = \alpha_1$. Further, $g(t, T_1) > \alpha_1$ for all non-terminal nodes t of T_1 .

Proof. Assume $\alpha < \alpha_1$. From (2.10) it follows that $g(t, T) \geq \alpha_1 > \alpha$, for all non-terminal nodes t , thus $R_\alpha(t) > R_\alpha(T_t)$ (see Lemma 2.8). The optimality of T is then a direct consequence of Theorem 2.9. Now consider $\alpha = \alpha_1$ and pruning by Theorem 2.9. Since $g(t, T) \geq \alpha_1$, then also $R_\alpha(t) \geq R_\alpha(T_t)$ for all non-terminal nodes t , hence pruning is only applied at all nodes t with $g(t, T) = \alpha_1$ (see condition (2.9)). Thus the second part of the theorem is proved. As a consequence, whenever the tree is pruned, $R_\alpha(T'_s)$ is unchanged for all the nodes s of the new tree. Since there were no nodes with $R_{\alpha_1}(t) < R_{\alpha_1}(T_t)$ and all the nodes with $R_{\alpha_1}(t) = R_{\alpha_1}(T_t)$ were pruned, all the remaining non-terminal nodes satisfy $R_{\alpha_1}(t) > R_{\alpha_1}(T_{1t})$, which, according to Lemma 2.8 is equivalent to $g(t, T_1) > \alpha_1$ for all non-terminal nodes t of T_1 . $\square$

Theorem 2.11. *For $\beta > \alpha$, $T(\beta)$ is a subtree of $T(\alpha)$ and is the result of β -pruning of $T(\alpha)$.*

Proof. To show that $T(\beta)$ is a subtree of $T(\alpha)$, it is sufficient to show that $T_t^{\prime\beta}$ is a subtree of $T_t^{\prime\alpha}$ for every node t during β -pruning and α -pruning, respectively. It is obviously true at the leaves. At node t we compare $R_\alpha(t)$ to $R_\alpha(T_t^{\prime\alpha})$ ($R_\beta(t)$ to $R_\beta(T_t^{\prime\beta})$) and prune the tree if the first is weakly smaller. We must show that if $R_\alpha(t) \leq R_\alpha(T_t^{\prime\alpha})$ then $R_\beta(t) \leq R_\beta(T_t^{\prime\beta})$. Suppose that $R_\alpha(t) \leq R_\alpha(T_t^{\prime\alpha})$. During α -pruning, the current tree $T_t^{\prime\alpha}$ is optimal at each step, thus also $R_\alpha(T_t^{\prime\alpha}) \leq R_\alpha(T_t^{\prime\beta})$ at each step. Then

$$\begin{aligned}
R_\beta(t) &\stackrel{(2.7)}{=} R(t) + \beta \text{size}(t) \stackrel{(2.7)}{=} R_\alpha(t) + (\beta - \alpha) \text{size}(t) \\
&\leq R_\alpha(T_t^{\prime\alpha}) + (\beta - \alpha) \text{size}(t) \leq R_\alpha(T_t^{\prime\beta}) + (\beta - \alpha) \text{size}(t) \\
&\stackrel{(2.7)}{=} R(T_t^{\prime\beta}) + \alpha \text{size}(T_t^{\prime\beta}) + (\beta - \alpha) \text{size}(t) \\
&\stackrel{(2.7)}{=} R_\beta(T_t^{\prime\beta}) + (\alpha - \beta) \text{size}(T_t^{\prime\beta}) + (\beta - \alpha) \text{size}(t) \\
&= R_\beta(T_t^{\prime\beta}) - (\beta - \alpha) \left(\text{size}(T_t^{\prime\beta}) - \text{size}(t) \right) \\
&\stackrel{\beta > \alpha \ \& \ \text{size}(T_t^{\prime\beta}) \geq \text{size}(t)}{\leq} R_\beta(T_{-t}^{\prime\beta}).
\end{aligned}$$

Thus $T(\beta)$ is a subtree of $T(\alpha)$. Since $T(\beta)$ minimizes $R_\beta(T')$ over all the rooted subtrees T' of T and is a subtree of $T(\alpha)$, it also minimizes $R_\beta(T')$ over all the rooted subtrees T' of $T(\alpha)$. $\square$

The algorithm from Theorem 2.10 can be applied to the new tree $T_1 := T(\alpha_1)$ to find $\alpha_2 > \alpha_1$ and $T_2 := T(\alpha_2)$ and so on until T_q is the trivial tree (i.e. the root of $T_0 := T$). From Theorem 2.10 and Theorem 2.11 it follows that for $\alpha_1 \leq \alpha < \alpha_2$, $T(\alpha) = T_1$ and $T(\alpha_2) = T_2$. Repeating the process, $\alpha_1 < \alpha_2 < \dots < \alpha_q$ and $T_0 \supset T_1 \supset \dots \supset T_q$ are obtained such that $T(\alpha) = T_i$ for $\alpha_i \leq \alpha < \alpha_{i+1}$ ($i = 1, \dots, q-1$). Finally, the full tree T_0 is optimal for $\alpha < \alpha_1$ (see Theorem 2.10) and the trivial tree T_q is optimal for $\alpha \geq \alpha_q$, which follows from Theorem 2.11.

Algorithm 2.12. The following algorithm can be used to find for a given tree T , the tree sequence $\{T_k\}_{k=0}^q$ and $\{\alpha_k\}_{k=1}^q$ as described above:

1. Set $k := 0$ and $T_0 := T$.
2. Visit all the non-terminal nodes t in bottom-up order, compute $g(t, T_k)$ using (2.8) and set α_{k+1} as

$$\alpha_{k+1} := \min_{\{t | t \text{ is a non-terminal node of } T_k\}} g(t, T_k).$$

3. Visit all the nodes in top-down order and prune whenever $g(t, T_k) = \alpha_{k+1}$.
4. Set $T_{k+1} := T_k'$, where T_k' is the tree obtained by α_{k+1} -pruning of T_k .
5. If T_{k+1} is a non-trivial tree, set $k := k + 1$ and go to Step 2.

2.5. k-Fold Cross-validation

Assume using Algorithm 2.12 to obtain a sequence of nested *trees* $\{T_k\}_{k=0}^q$, each *tree* being optimal for some $\alpha \in I_\alpha \subseteq \mathbb{R}$. Our goal is to choose the best *subtree* for our data. We already know that T_0 might overfit the data, but the *trivial tree* T_q is probably not the right model either. For choosing the best *subtree*, *k-fold cross-validation* can be used (more information regarding this topic can be found in [4]).

Consider a problem as described in Section 2.3. Denote $\mathcal{A}$ the set of training data, i.e.

$$\mathcal{A} := \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}.$$

Consider splitting $\mathcal{A}$ into k “roughly equal-sized” non-empty subsets $A_1, \dots, A_k$ such that

$$\bigcup_{i=1}^k A_i = \mathcal{A} \wedge A_i \cap A_j = \emptyset; \quad i \neq j \wedge \max_{i,j \in \{1, \dots, k\}} \{|\text{card}(A_i) - \text{card}(A_j)|\} \leq 1.$$

We define an indexing function κ as

$$\begin{aligned} \kappa : \{1, \dots, n\} &\rightarrow \{1, \dots, k\}, \\ i &\mapsto j \mid \mathbf{x}_i \in A_j. \end{aligned}$$

Denote by $\hat{f}^j(x)$ the fitted model using data from the set: $\mathcal{A} \setminus A_j$. The *cross-validation* estimate of prediction error is defined as

$$CV(\hat{f}) := \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}^{\kappa(i)}(\mathbf{x}_i)),$$

where $L(y, \hat{f}(\mathbf{x}))$ is the loss function measuring errors between the observed value y and the predicted value $\hat{f}(\mathbf{x})$. Typical choices for the loss function are *squared error*

$$L(y, \hat{f}(\mathbf{x})) := (y - \hat{f}(\mathbf{x}))^2,$$

and an *absolute error*

$$L(y, \hat{f}(\mathbf{x})) := |y - \hat{f}(\mathbf{x})|.$$

The case $k = n$ is called the *leave-one-out cross-validation*. In this case κ is an identity function (i.e. $\kappa(i) = i$), and for the i -th observation, the fit is computed using all the data samples except the i -th.

Consider a set of models $f(\mathbf{x}, \alpha)$ indexed by a tuning parameter α . Denote by $\hat{f}^p(\mathbf{x}, \alpha)$ the α -th model fit on sample $\mathcal{A} \setminus A_p$. Define the *cross-validation estimate* of prediction error for this set of models as

$$CV(\hat{f}, \alpha) := \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}^{\kappa(i)}(\mathbf{x}_i, \alpha)). \quad (2.11)$$

Our goal is to find parameter $\hat{\alpha}$ minimizing (2.11). Our final chosen model is then $f(\mathbf{x}, \hat{\alpha})$, which is fitted to all the data.

2.6. Confidence Intervals for the Expected Value

Consider a problem as described in Section 2.2. Consider a model of the form (2.1). Assume that $\varepsilon_1, \dots, \varepsilon_n$ are independent. Then also $Y_1, \dots, Y_n$, each Y_i being a function of ε_i , are independent. Defining an indexing function τ as

$$\begin{aligned} \tau: \mathcal{D} &\rightarrow \{1, \dots, m\}, \\ \mathbf{x} &\mapsto j \mid \mathbf{x} \in R_j, \end{aligned} \quad (2.12)$$

we can write (2.1) as

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i = \beta_{\tau(\mathbf{x}_i)} + \varepsilon_i. \quad (2.13)$$

Theorem 2.13. For $\hat{\beta}_j$ (where $j = 1, \dots, m$) as defined in (2.4), it follows that

$$\mathbb{E}(\hat{\beta}_j) = \beta_j, \quad (2.14)$$

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{C_j}. \quad (2.15)$$

Proof. Firstly we need to compute the expected value and variance of Y_i (where $i = 1, \dots, n$)

$$\mathbb{E}(Y_i) \stackrel{(2.13)}{=} \mathbb{E}(\beta_{\tau(\mathbf{x}_i)} + \varepsilon_i) = \mathbb{E}(\beta_{\tau(\mathbf{x}_i)}) + \mathbb{E}(\varepsilon_i) \stackrel{\mathbb{E}(\varepsilon_i)=0}{=} \beta_{\tau(\mathbf{x}_i)}, \quad (2.16)$$

$$\text{var}(Y_i) \stackrel{(2.13)}{=} \text{var}(\beta_{\tau(\mathbf{x}_i)} + \varepsilon_i) = \text{var}(\varepsilon_i) \stackrel{\text{var}(\varepsilon_i)=\sigma^2}{=} \sigma^2. \quad (2.17)$$

Using (2.4), both the desired equalities can be computed directly

$$\begin{aligned} \mathbb{E}(\hat{\beta}_j) &= \mathbb{E}\left(\frac{1}{C_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} Y_i\right) = \frac{1}{C_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} \mathbb{E}(Y_i) \stackrel{(2.16)}{=} \frac{1}{C_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} \beta_{\tau(\mathbf{x}_i)} \\ &\stackrel{(2.12)}{=} \frac{1}{C_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} \beta_j = \frac{1}{C_j} C_j \beta_j = \beta_j, \\ \text{var}(\hat{\beta}_j) &= \text{var}\left(\frac{1}{C_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} Y_i\right) = \frac{1}{C_j^2} \text{var}\left(\sum_{\{i|\mathbf{x}_i \in R_j\}} Y_i\right) \\ &\stackrel{Y_1, \dots, Y_n \text{ are independent}}{=} \frac{1}{C_j^2} \sum_{\{i|\mathbf{x}_i \in R_j\}} \text{var}(Y_i) \stackrel{(2.17)}{=} \frac{1}{C_j^2} \sum_{\{i|\mathbf{x}_i \in R_j\}} \sigma^2 \\ &= \frac{1}{C_j^2} C_j \sigma^2 = \frac{\sigma^2}{C_j}. \end{aligned}$$

□

Remark. According to (2.14), $\hat{\beta}_j$ is an *unbiased* estimator of β_j .

Remark. Denote the residual sum of squares as

$$S_e := \sum_{i=1}^n (Y_i - \hat{\beta}_{\tau(\mathbf{x}_i)})^2 = \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} (Y_i - \hat{\beta}_j)^2. \quad (2.18)$$

Remark. Further on, define $\mathbf{1} := (1, \dots, 1)^T$, $\mathbf{0} := (0, \dots, 0)^T$ and denote $\mathbf{I}$ as the identity matrix. If the size of $\mathbf{I}$ ($\mathbf{1}$ or $\mathbf{0}$) is not clear from the context, we will use $\mathbf{I}_q$ ($\mathbf{1}_q$ or $\mathbf{0}_q$), where the lower index q indicates the matrix (vector) of size q .

Denote $\mathbf{Y}_j := (Y_{j_1}, \dots, Y_{j_{c_j}})^T$ the vector of all Y_i 's, for which $\mathbf{x}_i \in R_j$. From (2.16), (2.17) and independence of Y_i 's, it follows that

$$\mathbb{E}(\mathbf{Y}_j) = \beta_j \mathbf{1}, \quad (2.19)$$

$$\text{var}(\mathbf{Y}_j) = \sigma^2 \mathbf{I}. \quad (2.20)$$

Using (2.4), it is easy to see that

$$\hat{\beta}_j = \frac{1}{C_j} \mathbf{1}^T \mathbf{Y}_j. \quad (2.21)$$

Remark. Notice that

$$\sum_{j=1}^m C_j = n. \quad (2.22)$$

Lemma 2.14. *Define $\mathbf{M}_j$ as*

$$\mathbf{M}_j := \mathbf{I}_{C_j} - \frac{1}{C_j} \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T. \quad (2.23)$$

Then $\mathbf{M}_j$ is symmetric, idempotent and it follows that

$$\mathbf{1}_{C_j}^T \mathbf{M}_j = \mathbf{0}^T. \quad (2.24)$$

Proof. The proof of symmetry is trivial. Idempotence follows from

$$\begin{aligned} \mathbf{M}_j \mathbf{M}_j &= \mathbf{I}_{C_j} - \frac{2}{C_j} \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T + \frac{1}{C_j^2} \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T \\ &\stackrel{\mathbf{1}_{C_j}^T \mathbf{1}_{C_j} = C_j}{=} \mathbf{I}_{C_j} - \frac{2}{C_j} \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T + \frac{1}{C_j^2} \mathbf{1}_{C_j} C_j \mathbf{1}_{C_j}^T = \mathbf{M}_j. \end{aligned}$$

Equation (2.24) can be obtained simply by direct computation. Indeed

$$\mathbf{1}_{C_j}^T \mathbf{M}_j = \mathbf{1}_{C_j}^T - \frac{1}{C_j} \mathbf{1}_{C_j}^T \mathbf{1}_{C_j} \mathbf{1}_{C_j}^T \stackrel{\mathbf{1}_{C_j}^T \mathbf{1}_{C_j} = C_j}{=} \mathbf{1}_{C_j}^T - \frac{1}{C_j} C_j \mathbf{1}_{C_j}^T = \mathbf{0}^T.$$

□

Theorem 2.15. *Define random variable s^2 as*

$$s^2 := \frac{S_e}{n - m}, \quad (2.25)$$

where S_e was defined in (2.18). Then s^2 is an unbiased estimate of σ^2 , i.e.

$$\mathbb{E}(s^2) = \sigma^2. \quad (2.26)$$

Proof. Firstly we compute $E(S_e)$ as follows

$$\begin{aligned}
 E(S_e) &\stackrel{(2.18)}{=} E\left(\sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} (Y_i - \hat{\beta}_j)^2\right) \\
 &\stackrel{(2.21)}{=} E\left(\sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} \left(Y_i - \frac{1}{\mathcal{C}_j} \mathbf{1}^T \mathbf{Y}_j\right)^2\right) \\
 &= \sum_{j=1}^m E\left(\left(\mathbf{Y}_j - \frac{1}{\mathcal{C}_j} \mathbf{1} \mathbf{1}^T \mathbf{Y}_j\right)^T \left(\mathbf{Y}_j - \frac{1}{\mathcal{C}_j} \mathbf{1} \mathbf{1}^T \mathbf{Y}_j\right)\right) \\
 &\stackrel{(2.23)}{=} \sum_{j=1}^m E(\mathbf{Y}_j^T \mathbf{M}_j^T \mathbf{M}_j \mathbf{Y}_j) \stackrel{\text{Lemma 2.14}}{=} \sum_{j=1}^m E(\mathbf{Y}_j^T \mathbf{M}_j \mathbf{Y}_j) \\
 &\stackrel{\text{Theorem 4.18 in [1]}}{=} \sum_{j=1}^m \left(\text{Tr}(\mathbf{M}_j \text{var}(\mathbf{Y}_j)) + (E(\mathbf{Y}_j))^T \mathbf{M}_j E(\mathbf{Y}_j) \right) \\
 &\stackrel{(2.19) \& (2.20)}{=} \sum_{j=1}^m \left(\text{Tr}(\mathbf{M}_j \sigma^2 \mathbf{I}) + \beta_j \mathbf{1}^T \mathbf{M}_j \beta_j \mathbf{1} \right) \\
 &\stackrel{(2.24)}{=} \sum_{j=1}^m \sigma^2 \text{Tr}(\mathbf{M}_j) \stackrel{(2.23)}{=} \sigma^2 \sum_{j=1}^m \left(\text{Tr}\left(\mathbf{I}_{\mathcal{C}_j} - \frac{1}{\mathcal{C}_j} \mathbf{1}_{\mathcal{C}_j} \mathbf{1}_{\mathcal{C}_j}^T\right) \right) \\
 &= \sigma^2 \sum_{j=1}^m (\mathcal{C}_j - 1) \stackrel{(2.22)}{=} \sigma^2 (n - m),
 \end{aligned} \tag{2.27}$$

Dividing (2.27) by $(n - m)$, (2.26) is obtained. $\square$

Further on in this section, assume that for $i = 1, \dots, n$, ε_i are normally distributed (i.e., $\varepsilon_i \sim N(0, \sigma^2)$), where the expected value and variance of ε_i was defined in Section 2.2). Then also Y_i are normally distributed, for $i = 1, \dots, n$.

Lemma 2.16. *Matrix $\mathbf{M}_j$ (see (2.23)) is positive semi-definite.*

Proof. Multiplying $\mathbf{M}_j$ by vector $\mathbf{c}$ (of length $\mathcal{C}_j$) from left and right, we obtain

$$\begin{aligned}
 \mathbf{c}^T \mathbf{M}_j \mathbf{c} &\stackrel{(2.23)}{=} \mathbf{c}^T \mathbf{c} - \frac{1}{\mathcal{C}_j} \mathbf{c}^T \mathbf{1}_{\mathcal{C}_j} \mathbf{1}_{\mathcal{C}_j}^T \mathbf{c} \\
 &\stackrel{\mathbf{1}_{\mathcal{C}_j}^T \mathbf{1}_{\mathcal{C}_j} = \mathcal{C}_j}{=} \frac{1}{\mathcal{C}_j} \left(\mathbf{c}^T \mathbf{c} \cdot \mathbf{1}_{\mathcal{C}_j}^T \mathbf{1}_{\mathcal{C}_j} - (\mathbf{c}^T \mathbf{1}_{\mathcal{C}_j})^2 \right) \geq 0,
 \end{aligned}$$

where the last inequality is a well-known *Cauchy-Schwarz* inequality for $\mathbb{R}^{\mathcal{C}_j}$

$$|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle; \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{C}_j},$$

where $\langle \mathbf{u}, \mathbf{v} \rangle$ denotes the inner product of vectors $\mathbf{u}, \mathbf{v}$ (in this case, $\mathbf{u} = \mathbf{c}$ and $\mathbf{v} = \mathbf{1}_{\mathcal{C}_j}$). $\square$

Theorem 2.17. Consider s^2 as defined in (2.25). Assume that $\sigma^2 > 0$ and $\mathcal{C}_j \geq 2$, for $j = 1, \dots, m$. Then

$$\frac{(n-m)s^2}{\sigma^2} \sim \chi_{n-m}^2.$$

Proof. Define random variable Z_i as

$$Z_i := \frac{Y_i - \beta_{\tau(\mathbf{x}_i)}}{\sigma}. \quad (2.28)$$

It is easy to see that

$$\frac{1}{\mathcal{C}_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} Z_i = \frac{1}{\mathcal{C}_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} \frac{(Y_i - \beta_j)}{\sigma} \stackrel{(2.4)}{=} \frac{\hat{\beta}_j - \beta_j}{\sigma}. \quad (2.29)$$

Denote $\mathbf{Z}_j := (Z_{j1}, \dots, Z_{j\mathcal{C}_j})^T$ the vector of all Z_i 's, for which $\mathbf{x}_i \in R_j$. From (2.16), (2.17) and Z_i being a linear combination of Y_i , which is normally distributed, it follows that $Z_i \sim N(0, 1)$ (for $i = 1, \dots, n$). Finally, since Y_i 's are independent, Z_i 's are also independent, thus $\mathbf{Z}_j \sim N(\mathbf{0}_{\mathcal{C}_j}, \mathbf{I}_{\mathcal{C}_j})$. Furthermore, not only Z_i 's are independent, but $\mathbf{Z}_p$ and $\mathbf{Z}_q$ are also independent for $p \neq q$. Then

$$\begin{aligned} \frac{(n-m)s^2}{\sigma^2} &\stackrel{(2.25) \& (2.18)}{=} \frac{1}{\sigma^2} \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} (Y_i - \hat{\beta}_j)^2 \\ &= \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} \left(\frac{(Y_i - \beta_j) - (\hat{\beta}_j - \beta_j)}{\sigma} \right)^2 \\ &\stackrel{(2.28) \& (2.29)}{=} \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} \left(Z_i - \frac{1}{\mathcal{C}_j} \sum_{\{k|\mathbf{x}_k \in R_j\}} Z_k \right)^2 \\ &= \sum_{j=1}^m \left(\mathbf{Z}_j - \frac{1}{\mathcal{C}_j} \mathbf{1} \mathbf{1}^T \mathbf{Z}_j \right)^T \left(\mathbf{Z}_j - \frac{1}{\mathcal{C}_j} \mathbf{1} \mathbf{1}^T \mathbf{Z}_j \right) \\ &\stackrel{\text{Lemma 2.14}}{=} \sum_{j=1}^m \mathbf{Z}_j^T \mathbf{M}_j \mathbf{Z}_j. \end{aligned} \quad (2.30)$$

From Lemma 2.14 and Lemma 2.16, it follows that $\mathbf{M}_j$ is symmetric, idempotent, non-zero and positive semi-definite, hence by Theorem 4.16 in [1] we obtain

$$\mathbf{Z}_j^T \mathbf{M}_j \mathbf{Z}_j \sim \chi_{\text{Tr}(\mathbf{M}_j \mathbf{I})}^2,$$

which yields

$$\mathbf{Z}_j^T \mathbf{M}_j \mathbf{Z}_j \sim \chi_{\mathcal{C}_j-1}^2. \quad (2.31)$$

Finally, using Theorem 4.13 in [1], independence of $\mathbf{Z}_i$'s, (2.22), (2.30) and (2.31), we obtain

$$\frac{(n-m)s^2}{\sigma^2} = \sum_{j=1}^m \mathbf{Z}_j^T \mathbf{M}_j \mathbf{Z}_j \sim \chi_{n-m}^2.$$

□

Theorem 2.18. Assume that $\sigma^2 > 0$ and $\mathcal{C}_j \geq 2$, for $j = 1, \dots, m$. Then $\hat{\beta}_j$ and s^2 are independent.

Proof. Using (2.30), we can write s^2 as

$$s^2 = \frac{\sigma^2}{n-m} \sum_{j=1}^m \mathbf{Z}_j^T \mathbf{M}_j \mathbf{Z}_j \stackrel{(2.28)}{=} \sum_{j=1}^m (\mathbf{Y}_j - \beta_j \mathbf{1})^T \frac{1}{n-m} \mathbf{M}_j (\mathbf{Y}_j - \beta_j \mathbf{1}).$$

From (2.19), (2.20) and normality of $\mathbf{Y}_i$'s it follows that $\mathbf{Y}_j \sim N(\beta_j \mathbf{1}, \sigma^2 \mathbf{I})$. Denote $\mathbf{B} := \frac{1}{\mathcal{C}_j} \mathbf{1}_{\mathcal{C}_j}^T$. Then

$$\mathbf{B} \text{var}(\mathbf{Y}_j) \left(\frac{1}{n-m} \mathbf{M}_j \right) = \frac{\sigma^2}{\mathcal{C}_j(n-m)} \mathbf{1}^T \mathbf{I} \mathbf{M}_j \stackrel{(2.24)}{=} \mathbf{0}^T. \quad (2.32)$$

Since $\frac{1}{n-m} \mathbf{M}_j$ is positive semi-definite (see Lemma 2.16) and using (2.32), we obtain from Theorem 4.19 in [1] that $(\mathbf{Y}_j - \beta_j \mathbf{1})^T \frac{1}{n-m} \mathbf{M}_j (\mathbf{Y}_j - \beta_j \mathbf{1})$ and $\hat{\beta}_j = \mathbf{B} \mathbf{Y}_j$ are independent. Moreover, $(\mathbf{Y}_i - \beta_i \mathbf{1})^T \frac{1}{n-m} \mathbf{M}_i (\mathbf{Y}_i - \beta_i \mathbf{1})$ and $\hat{\beta}_j$ are independent for $i \neq j$, since $\mathbf{Y}_i$'s are independent. Finally, since s^2 is a sum of m independent random variables, where each of them is independent with $\hat{\beta}_j$, then s^2 and $\hat{\beta}_j$ are also independent. $\square$

Theorem 2.19. Consider $\tilde{x} \in \mathcal{D}$. Then

$$\frac{\hat{\beta}_{\tau(\tilde{x})} - \beta_{\tau(\tilde{x})}}{s} \sqrt{\mathcal{C}_{\tau(\tilde{x})}} \sim t_{n-m},$$

where t_{n-m} is a Student's t -distribution with $n-m$ degrees of freedom.

Proof. It follows from Theorem 2.13 and $\hat{\beta}_j$ being normally distributed, that

$$\hat{\beta}_{\tau(\tilde{x})} \sim N\left(\beta_{\tau(\tilde{x})}, \frac{\sigma^2}{\mathcal{C}_{\tau(\tilde{x})}}\right),$$

thus

$$\frac{\hat{\beta}_{\tau(\tilde{x})} - \beta_{\tau(\tilde{x})}}{\sigma} \sqrt{\mathcal{C}_{\tau(\tilde{x})}} \sim N(0, 1). \quad (2.33)$$

Using Theorem 4.22 in [1], (2.33), Theorem 2.17 and Theorem 2.18, we obtain

$$\frac{\frac{\hat{\beta}_{\tau(\tilde{x})} - \beta_{\tau(\tilde{x})}}{\sigma} \sqrt{\mathcal{C}_{\tau(\tilde{x})}}}{\sqrt{\frac{(n-m)s^2}{\sigma^2}}} \sqrt{n-m} = \frac{\hat{\beta}_{\tau(\tilde{x})} - \beta_{\tau(\tilde{x})}}{s} \sqrt{\mathcal{C}_{\tau(\tilde{x})}} \sim t_{n-m}.$$

$\square$

Remark. Using Theorem 2.19, we can construct the $(1 - \alpha)$ -confidence interval for $E(\tilde{Y})$, where $\tilde{Y}$ is an independent future observation associated with $\tilde{\mathbf{x}}$

$$\left(\hat{\beta}_{\tau(\tilde{\mathbf{x}})} - t_{n-m}(\alpha) \frac{s}{\sqrt{\mathcal{C}_{\tau(\tilde{\mathbf{x}})}}}; \hat{\beta}_{\tau(\tilde{\mathbf{x}})} + t_{n-m}(\alpha) \frac{s}{\sqrt{\mathcal{C}_{\tau(\tilde{\mathbf{x}})}}} \right),$$

where for a random variable $T \sim t_p$, value $t_p(\alpha)$ is defined as $(1 - \frac{\alpha}{2})$ -quantile, i.e.

$$P\{|T| \geq t_p(\alpha)\} = \alpha. \quad (2.34)$$

Remark. In Theorem 2.17 and Theorem 2.18, we assumed that $\mathcal{C}_j \geq 2$, which is quite reasonable, since we do not want to grow *trees* with *leaves* containing only one observation. Such *trees* would be probably overfitted and we would need to use e.g. *k-fold cross-validation* (see Section 2.5) in order to obtain an optimal *subtree*.

2.7. Maximum Likelihood Estimation for Gamma Distribution of the Response Variable

In Section 2.3, for a given partition of $\mathcal{D}$, the values of $\beta_1, \dots, \beta_m$ were estimated by minimizing the residual sum of squares (see (2.2)). If the response variable has gamma distribution, we might use another approach, maximum likelihood estimation. Maximum likelihood is described in detail in [2] in the context of *generalized linear models*. The crucial difference between least squares and maximum likelihood is that instead of minimizing a loss function, we estimate the values of unknown parameters in such way, that amongst all the possible considered models, the probability of making such observations as we have done, is the greatest.

Let $Y_i \sim \Gamma(\alpha, \beta_i)$, for $i = 1, \dots, n$, where, $\beta_1, \dots, \beta_n$ are the parameters of interest and $\alpha > 0$ is regarded as a nuisance parameter. Consider k predictors and denote $\mathbf{x}_i := (x_{i1}, \dots, x_{ik})^T \in \mathcal{D}$, for $i = 1, \dots, n$. Consider a partition of $\mathcal{D}$ into m regions $R_1, \dots, R_m$. We are not interested in parameters $\beta_1, \dots, \beta_n$ directly, but instead we estimate parameters $\mu_1, \dots, \mu_m$, where for $j = 1, \dots, m$ it follows that

$$E(Y_i) = \mu_j; \quad \forall i \mid \mathbf{x}_i \in R_j. \quad (2.35)$$

According to [1], for $Y_i \sim \Gamma(\alpha, \beta_i)$ it follows that

$$E(Y_i) = \frac{\alpha}{\beta_i}. \quad (2.36)$$

Combining (2.35) and (2.36), we obtain

$$\beta_i = \frac{\alpha}{\mu_j}; \quad \forall i \mid \mathbf{x}_i \in R_j; \quad j = 1, \dots, m. \quad (2.37)$$

Using this notation, our model of interest (compare with (2.1)) is now of the form

$$Y_i = \sum_{j=1}^m \mu_j \cdot \chi_{R_j}(\mathbf{x}_i) + \varepsilon_i; \quad i = 1, \dots, n,$$

The probability density function $f_i := f_i(y_i; \beta_i)$ of Y_i (see [1]) is of the form

$$f_i = \exp(-y_i \beta_i + \alpha \ln \beta_i + (\alpha - 1) \ln y_i - \ln \Gamma(\alpha)), \quad (2.38)$$

where $\Gamma(\alpha)$ is the gamma function.

Remark. Do not confuse the notation f_i with the notation used in Section 2.3. In this subsection, f_i denotes the probability density function of Y_i .

Denote $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^T$. Assuming that $Y_1, \dots, Y_n$ are independent, we can write the joint log-likelihood function as

$$\begin{aligned}
 \ell &:= \ell(\mathbf{y}; \boldsymbol{\mu}) = \ln f(\mathbf{y}; \boldsymbol{\mu}) \stackrel{Y_1, \dots, Y_n \text{ are independent}}{=} \ln \left(\prod_{i=1}^n f_i \right) = \sum_{i=1}^n \ln f_i \\
 &\stackrel{(2.38)}{=} \sum_{i=1}^n \ln \exp \left(-y_i \beta_i + \alpha \ln \beta_i + (\alpha - 1) \ln y_i - \ln \Gamma(\alpha) \right) \\
 &= \sum_{i=1}^n \left(-y_i \beta_i + \alpha \ln \beta_i + (\alpha - 1) \ln y_i - \ln \Gamma(\alpha) \right) \\
 &= -n \ln \Gamma(\alpha) + \sum_{j=1}^m \sum_{\{i | \mathbf{x}_i \in R_j\}} (-y_i \beta_i + \alpha \ln \beta_i + (\alpha - 1) \ln y_i) \\
 &\stackrel{(2.37)}{=} -n \ln \Gamma(\alpha) + \sum_{j=1}^m \sum_{\{i | \mathbf{x}_i \in R_j\}} \left(-y_i \frac{\alpha}{\mu_j} + \alpha \ln \frac{\alpha}{\mu_j} + (\alpha - 1) \ln y_i \right).
 \end{aligned} \tag{2.39}$$

Our goal is to find the values of $\mu_1, \dots, \mu_m$ maximizing the expression (2.39). Differentiating both sides of (2.39) with respect to μ_u and setting it to 0, we obtain

$$\begin{aligned}
 \frac{\partial \ell}{\partial \mu_u} &= \sum_{\{i | \mathbf{x}_i \in R_u\}} \frac{\partial}{\partial \mu_u} \left(-y_i \frac{\alpha}{\mu_u} + \alpha \ln \frac{\alpha}{\mu_u} + (\alpha - 1) \ln y_i \right) \\
 &= \sum_{\{i | \mathbf{x}_i \in R_u\}} \left(y_i \frac{\alpha}{\mu_u^2} + \alpha \frac{\mu_u}{\alpha} \left(-\frac{\alpha}{\mu_u^2} \right) \right) = \frac{\alpha}{\mu_u^2} \sum_{\{i | \mathbf{x}_i \in R_u\}} (y_i - \mu_u) \\
 &= \frac{\alpha}{\mu_u^2} \left(\sum_{\{i | \mathbf{x}_i \in R_u\}} y_i - \mathcal{C}_u \mu_u \right) = 0,
 \end{aligned} \tag{2.40}$$

where $\mathcal{C}_u$ has the same meaning as in Section 2.3. Assuming that $\{i | \mathbf{x}_i \in R_u\}$ is non-empty, we obtain a unique solution of (2.40)

$$\hat{\mu}_u = \frac{1}{\mathcal{C}_u} \sum_{\{i | \mathbf{x}_i \in R_u\}} y_i. \tag{2.41}$$

Similarly as in Section 2.3.1, it is an easy exercise to verify that (2.41) maximizes (2.39), by verifying that the Hessian matrix of (2.39) is negative definite. Details of these computations can be found in [7].

Notice that (2.41) is formally the same as (2.4). For a given partition of $\mathcal{D}$ and for a future independent observation $\tilde{Y}$ (associated with $\tilde{\mathbf{x}}$), the predicted value, $\mu_j | \tilde{\mathbf{x}} \in R_j$, is the same as it was for the least squares estimator derived in Section 2.3.

The difference between the least squares minimization approach and maximum likelihood estimation is how the *split predictor* X_j and *split point* s are obtained. Instead of solving (2.6), different optimization problem is considered

$$\max_{j; s} \left(\max_{\mu_1, \dots, \mu_m} \ell(\mathbf{y}; \boldsymbol{\mu}) \right), \tag{2.42}$$

where the meaning of j and s was explained in Section 2.3.2 and $\boldsymbol{\mu} := (\mu_1, \dots, \mu_m)$. For every pair (j, s) , the inner maximization of (2.42) is solved by (2.41). Similarly as in Section 2.3.2, the outer maximization can be simply computed by iterating through all the predictors and possible values of s and choosing the pair with the greatest value of (2.39) evaluated at the point $\hat{\boldsymbol{\mu}} := (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$ computed according to (2.41). If we assume $\alpha > 0$ to be fixed, it is obvious that maximizing the value of (2.39) evaluated at the point $\hat{\boldsymbol{\mu}}$ for different partitions $\{R_1(j_q, s_q); R_2(j_q, s_q)\}$ (for q being from some finite index set) is the same as maximizing the expression

$$\begin{aligned}
& \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} \left(-y_i \frac{\alpha}{\mu_j} + \alpha \ln \frac{\alpha}{\mu_j} + (\alpha - 1) \ln y_i \right) \Big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}} \\
&= (\alpha - 1) \sum_{j=1}^m \sum_{\{i|\mathbf{x}_i \in R_j\}} \ln y_i + n\alpha \ln \alpha \\
&\quad - \alpha \sum_{j=1}^m \left(\frac{1}{\hat{\mu}_j} \sum_{\{i|\mathbf{x}_i \in R_j\}} y_i + \mathcal{C}_j \ln \hat{\mu}_j \right) \\
&\stackrel{(2.41) \ \& \ (2.22)}{=} (\alpha - 1) \sum_{i=1}^n \ln y_i + n\alpha \ln \alpha - n\alpha - \alpha \sum_{j=1}^m \mathcal{C}_j \ln \hat{\mu}_j.
\end{aligned} \tag{2.43}$$

We can see that the first three terms of (2.43) do not depend on the partition of $\mathcal{D}$, hence for fixed $\alpha > 0$, the outer maximization of (2.42) is solved by computing and comparing the values of

$$\ell^* := - \sum_{j=1}^m \mathcal{C}_j \ln \hat{\mu}_j \tag{2.44}$$

for all the pairs (j, s) and selecting the pair with the greatest value of (2.44).

Remark. The outer maximization of (2.42) can be replaced by minimization of so called *deviance*, which provides exactly the same results. Details of this approach can be found in [7].

3. RANDOM FORESTS

Regression trees as described in Section 2 have some disadvantages, e.g. instability or the chance of overfitting the data easily. These issues can be overcome by *random forests*, which improve the predictions in terms of variance, by simply growing multiple *regression trees*, each of them on a slightly different dataset (as we will see in this chapter). Firstly, we will briefly introduce the basic concepts of bootstrap and bagging, which will be important for growing *random forests*.

3.1. Bootstrapping and Bagging

This section provides only a brief introduction into the bootstrapping and bagging. More detailed information regarding this topic can be found in [4].

Consider a set of training data $\mathcal{A} := \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. Denote $S(\mathcal{A})$ some quantity computed from data $\mathcal{A}$ (e.g. a prediction at some input point based on some model). The basic idea of bootstrapping is to randomly draw datasets with replacement from the set of training data, each sample the same size as the original one. This is done B times, producing B bootstrapped datasets. Then the quantity $S(\cdot)$ is computed from each of the B bootstrapped datasets. According to [4], using bootstrap sampling, we can compute the estimation $\widehat{\text{var}}(S(\mathcal{A}))$ of the variance of $S(\mathcal{A})$ as

$$\widehat{\text{var}}(S(\mathcal{A})) = \frac{1}{B-1} \sum_{b=1}^B (S(\mathcal{A}^b) - \bar{S})^2, \quad (3.1)$$

where $\mathcal{A}^b$ denotes the b -th bootstrapped dataset from $\mathcal{A}$ and $\bar{S}$ is defined as

$$\bar{S} := \frac{1}{B} \sum_{b=1}^B S(\mathcal{A}^b).$$

Now we will show how to use bootstrap to improve the prediction itself. Suppose we fit a model $\hat{f}$ to our training data $\mathcal{A}$ and obtain a prediction $\hat{f}(\tilde{\mathbf{x}})$ at input $\tilde{\mathbf{x}}$. Bagging averages this prediction over B bootstrapped samples. For b -th bootstrapped sample $\mathcal{A}^b$, we fit our model to obtain prediction $\hat{f}^b(\tilde{\mathbf{x}})$. The bagging estimate is defined as

$$\hat{f}_{\text{bag}}(\tilde{\mathbf{x}}) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(\tilde{\mathbf{x}}).$$

When bagging is applied on *regression trees*, each bootstrapped *tree* might contain different *splits* and might have a different number of *leaves*. The bagged estimate is the average prediction at $\tilde{\mathbf{x}}$ over all the B *trees*.

Bagging can significantly reduce the variance of unstable procedures like *regression trees* (see [4]). Applying bagging on *regression trees*, any simple structure is lost. Hence interpretability is lost as a result of improving the stability of the model. Instability of a *regression tree* means that a small change in the training data can result in a very different series of *splits*, hence providing quite different results.

The key difference between bootstrapping and bagging is that bootstrapping is a random sampling with replacement, and bagging is performing the bootstrap multiple times and training an estimator for each of the bootstrapped data and then aggregating the predictions to make a final prediction. So bootstrapping is a sampling technique, whereas bagging, also known as bootstrap aggregation, is a machine learning ensemble technique designed to improve accuracy of statistical models used for prediction and classification.

3.2. Introduction to Random Forests

In Section 3.1, basic idea of bagging was introduced, where we fit a *regression tree* many times to bootstrapped samples of the training data and average the

results. *Random forests* technique is a modification of bagging, that builds a large collection of *trees*, while reducing correlation amongst them at the same time.

An important feature of *trees* obtained by bagging is that they are identically distributed (i.d.), since each *tree* is built on a data sample randomly drawn with replacement from the same population. As a result, the expectation and variance of each *tree* is the same, so also the bias of bagged *trees* (i.e., squared distance between predicted and real values) is unchanged in comparison to the individual *trees*. Hence the only improvement can be through variance reduction of bagged *trees* when compared to individual *trees*.

Theorem 3.1. *Let $X_1, \dots, X_n$ be i.d. random variables (i.e., $\text{var}(X_i) = \sigma^2$, for $i = 1, \dots, n$). Assume that $\text{cor}(X_i, X_j) = \rho \geq 0$, for $i \neq j$ and $i, j = 1, \dots, n$. Denote $\bar{X}_n$ the average of $\mathbf{X} := (X_1, \dots, X_n)^T$, i.e.*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \mathbf{1}^T \mathbf{X}. \quad (3.2)$$

Then variance of $\bar{X}_n$ is of the form

$$\text{var}(\bar{X}_n) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2. \quad (3.3)$$

Proof. The variance-covariance matrix of $\mathbf{X}$ is of the form

$$\text{var}(\mathbf{X}) = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & & \ddots & \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{pmatrix} = \rho\sigma^2 \mathbf{1}\mathbf{1}^T + \sigma^2(1-\rho)\mathbf{I}. \quad (3.4)$$

Using basic properties of variance, we obtain

$$\begin{aligned} \text{var}(\bar{X}_n) &\stackrel{(3.2)}{=} \text{var}\left(\frac{1}{n} \mathbf{1}^T \mathbf{X}\right) = \frac{1}{n^2} \text{var}(\mathbf{1}^T \mathbf{X}) \\ &\stackrel{\text{var}(\mathbf{A}\mathbf{X}) = \mathbf{A} \text{var}(\mathbf{X}) \mathbf{A}^T}{=} \frac{1}{n^2} \mathbf{1}^T \text{var}(\mathbf{X}) \mathbf{1} \\ &\stackrel{(3.4)}{=} \frac{\sigma^2}{n^2} \mathbf{1}^T (\rho \mathbf{1}\mathbf{1}^T + (1-\rho)\mathbf{I}) \mathbf{1} \\ &= \frac{\sigma^2}{n^2} (\rho \mathbf{1}^T \mathbf{1} \mathbf{1}^T \mathbf{1} + (1-\rho) \mathbf{1}^T \mathbf{1}) \\ &\stackrel{\mathbf{1}^T \mathbf{1} = n}{=} \frac{\sigma^2}{n^2} (n^2 \rho + n(1-\rho)) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2. \end{aligned}$$

□

Remark. Notice that if X_i 's are independent (i.e., $\rho = 0$) in Theorem 3.1, we obtain $\text{var}(\bar{X}_n) = \frac{1}{n}\sigma^2$, which is a well-known property of variance of a mean of independent, identically distributed (i.i.d.) random variables.

Remark. An important result obtained from Theorem 3.1 is

$$\lim_{n \rightarrow \infty} \text{var}(\bar{X}_n) \stackrel{(3.3)}{=} \rho\sigma^2 \leq \sigma^2 = \text{var}(X_i); \quad i \in \mathbb{N},$$

since $0 \leq \rho \leq 1$.

Knowing that *regression trees* obtained by bagging are i.d. and assuming that correlation between the *trees* is constant and less than 1, we can decrease the variance of a single *regression tree* model by growing a sufficiently large number of *trees*. The basic idea of *random forests* is to improve the variance reduction obtained by bagging by reducing the correlation between *trees* without increasing their variance too much. This is achieved by selecting a random subset of $m \leq k$ predictors before each *split* when growing a *tree*.

Algorithm 3.2. Consider a set of training data $\mathcal{A} := \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T$, for $i = 1 \dots, n$. Each of the k predictors is either continuous or categorical. The following algorithm can be used to grow a *random forest*.

1. For $b = 1$ to B :
 - (a) Draw a bootstrap data sample $\mathcal{A}^b$ of size n from original data $\mathcal{A}$.
 - (b) Grow a *regression tree* T_b on bootstrapped data $\mathcal{A}^b$, by recursively repeating the following steps for each *leaf* of the *tree*, until the minimum *node size* is reached for all the *leaves* of the current *tree*:
 - (i) Select a random subset of $m \leq k$ predictors from k original predictors.
 - (ii) Select the best *split predictor* and *split point* (see Section 2.3 and Section 2.7, respectively) from the selected m predictors.
 - (iii) Split the *node* into two *nodes* only if the resulting *nodes* satisfy the condition of minimum *node size*, and perform Steps 1(b)(i)–1(b)(iii) on both of the resulting *nodes*.
 - (c) If needed, perform pruning of the *tree* T_b (see Section 2.4) and use the *k-fold cross-validation* (see Section 2.5) to select the best *subtree* T'_b , and set $T_b := T'_b$.
2. Output the ensemble of *regression trees* $\{T_b\}_{b=1}^B$.
3. Prediction at a point $\tilde{\mathbf{x}}$ is computed as

$$\bar{T}(\tilde{\mathbf{x}}) := \frac{1}{B} \sum_{b=1}^B T_b(\tilde{\mathbf{x}}),$$

where $T_b(\tilde{\mathbf{x}})$ is a prediction at $\tilde{\mathbf{x}}$ obtained from b -th *tree*.

According to [4], recommended minimum *node size* is 5, and the recommended value for m in Step 1(b)(i) is $\lfloor \frac{k}{3} \rfloor$, where $\lfloor x \rfloor$ is the floor function defined as

$$\lfloor x \rfloor := \max \{m \in \mathbb{Z} \mid m \leq x\}.$$

In practice, the best values for both these parameters depend on the particular problem and can be considered as tuning parameters.

When growing a *regression tree*, only the most relevant predictors are selected for *splits*, whereas predictors that do not affect Y_i 's very much are not selected. When the number of predictors k is large, but the number of relevant predictors is relatively very small when compared to k , *random forests* are not likely to perform

good for small values of m , since there is small chance of a relevant predictor being selected at each *split*. Before each *split*, the probability of a relevant predictor being selected follows a hypergeometric distribution.

3.3. Confidence Intervals for the Expected Value

Consider a random sample $\mathcal{A}$ from an unknown distribution F . Consider a parameter of interest $\theta = t(F)$ and its estimate $\hat{\theta} = s(\mathcal{A})$. For $b = 1, \dots, B$, consider a bootstrap dataset $\mathcal{A}^b$ from $\mathcal{A}$ and define a bootstrap replication of $\hat{\theta}$ as $\hat{\theta}^b := s(\mathcal{A}^b)$. Then according to [3], the bootstrap estimate $\hat{s}$ of the *standard error* of a *statistic* $\hat{\theta}$ can be computed as

$$\hat{s} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^b - \bar{\hat{\theta}})^2}, \quad (3.5)$$

where $\bar{\hat{\theta}}$ is defined as

$$\bar{\hat{\theta}} := \frac{1}{B} \sum_{b=1}^B \hat{\theta}^b.$$

Remark. Notice the similarity between (3.1) and (3.5). Indeed, the *standard error* of a *statistic* is the estimation of the square root of variance of its sampling distribution.

Now, consider an independent future observation $\tilde{Y}$ associated with $\tilde{\mathbf{x}}$. The value $T_b(\tilde{\mathbf{x}})$ (defined in Step 3 of Algorithm 3.2) can be considered as a bootstrap replication of $E(\tilde{Y})$, and from (3.5), the estimate $\hat{s}$ of the *standard error* of $E(\tilde{Y})$ can be computed as

$$\hat{s} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (T_b(\tilde{\mathbf{x}}) - \bar{T}(\tilde{\mathbf{x}}))^2},$$

where $\bar{T}(\tilde{\mathbf{x}})$ was defined also in Step 3.

Using bootstrap, we can obtain confidence intervals without making any assumptions about data. For each of the B bootstrapped data samples we compute

$$Z_b := \frac{T_b(\tilde{\mathbf{x}}) - \bar{T}(\tilde{\mathbf{x}})}{s_b},$$

where s_b is the estimated *standard error* for the b -th *tree* and can be computed as (also compare with (2.15))

$$s_b = \sqrt{\frac{s^2}{\mathcal{C}_{\tau(\tilde{\mathbf{x}})}}},$$

where s^2 was defined in (2.25). Consider a random variable Z

$$Z = \frac{\bar{T}(\tilde{\mathbf{x}}) - E(\tilde{Y})}{\hat{s}}.$$

According to [3], the $(1 - \alpha)$ -quantile of Z can be estimated by the value $\hat{t}(\alpha)$, which satisfies the following condition

$$\frac{\text{card}(\{b \mid Z_b \leq \hat{t}(\alpha)\})}{B} = \alpha. \quad (3.6)$$

If $B\alpha$ is not an integer, then e.g. interpolation can be used to compute $\hat{t}(\alpha)$. Finally, using (3.6), the bootstrap- t $(1 - \alpha)$ -confidence interval for $E(\tilde{Y})$ can be computed as

$$\left(\bar{T}(\tilde{\mathbf{x}}) - \hat{s} \hat{t} \left(1 - \frac{\alpha}{2} \right); \bar{T}(\tilde{\mathbf{x}}) - \hat{s} \hat{t} \left(\frac{\alpha}{2} \right) \right).$$

Another approach to compute confidence intervals is using the empirical distribution function. Define the empirical $(1 - \alpha)$ -quantile $\hat{t}_e(\alpha)$ as the value satisfying

$$\frac{\text{card}(\{b \mid T_b(\tilde{\mathbf{x}}) \leq \hat{t}_e(\alpha)\})}{B} = \alpha. \quad (3.7)$$

From (3.7), the empirical $(1 - \alpha)$ -confidence interval for $E(\tilde{Y})$ can be computed simply as

$$\left(\hat{t}_e \left(\frac{\alpha}{2} \right); \hat{t}_e \left(1 - \frac{\alpha}{2} \right) \right).$$

Remark. Notice a slight difference in notations of different quantiles defined in this article. In (2.34), the value $t_p(\alpha)$ denotes $(1 - \frac{\alpha}{2})$ -quantile, whilst in (3.6) and (3.7), values $F_{m,n}(\alpha)$, $\hat{t}(\alpha)$ and $\hat{t}_e(\alpha)$ denote $(1 - \alpha)$ -quantiles. This is because of the symmetry of *Student's t*-distribution. Since reader might be used to different notation, we'd rather emphasize the notation used in this article.

4. CONCLUSION

Section 2 provides an introduction to the concept of *regression trees*. Two principles, least squares and maximum likelihood estimation, of growing the *trees* are described and confidence intervals for the expected value were also derived. Introduction to *regression trees* can be found in [4] and more detailed information can be found in [8], according to which Section 2.4 was written. Proofs of theorems in Section 2.4 can also be found in [8], but were slightly modified (especially the proof of Theorem 2.10), since the author was not completely satisfied by the proofs provided in [8]. Since we could not find literature covering the topic of confidence intervals from Section 2.6 to the desired extent, the results in this section were derived by the author, following the derivation of confidence intervals for the expected value for *linear regression* from [1], since the process was quite similar. Also, for the same reason, the maximum likelihood estimation with the assumption of gamma distribution of the response variable of regression trees from Section 2.7 was derived by the author, inspired by the process of derivation of *generalized linear models* described in [2].

Section 3 deals with *random forests*. Although this chapter might seem to be short when compared to the previous one, the opposite is true. Most of the required apparatus for *random forests* has been already derived in Section 2, since *random forests* use *regression trees* together with bootstrapping and bagging to reduce

the variance of single *regression trees* and provide even more accurate predictions. Confidence intervals for the expected value were also derived for *random forests*. The introduction to *random forests* (Section 3.1 and Section 3.2 can be found in [4]). Confidence intervals from Section 3.3 were application of confidence intervals for bootstrap described in [3].

Application of described statistical models (i.e. *regression trees* and *random forests*) on real data can be found in [7], as well as comparison of those two models with *linear regression* and *generalized linear models*. Due to skewness of the distribution of our data, we decided to derive a maximum likelihood estimation for regression trees (see Section 2.7), assuming a response variable with gamma distribution. This gamma distribution reflects skewness of the distribution of our data in [7], which was also reflected in the final results, since maximum likelihood method for *regression trees* provided better results when compared to ordinary least squares (see Section 2.3) (and also the best results amongst all the studied statistical models).

REFERENCE

- [1] J. Anděl: *Základy matematické statistiky*, MatfyzPress, Praha, 2011.
- [2] A. J. Dobson, A. G. Barnett: *An Introduction to Generalized Linear Models*, Chapman and Hall/CRC, New York, 2008.
- [3] B. Efron, R. Tibshirani: *An Introduction to the Bootstrap*, Chapman and Hall/CRC, New York, 1993.
- [4] T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2016.
- [5] P. Hebák, J. Hustopecký I. Malá: *Vícerozměrné statistické metody [2]* Informatorium, Praha, 2005.
- [6] P. McCullagh, J. A. Nelder: *Generalized Linear Models*, Chapman and Hall/CRC, New York, 1989.
- [7] D. Oberta: *Statistical Models for Prediction of Project Duration*, Bachelor's Thesis; Faculty of Mechanical Engineering, Brno University of Technology, Brno, 2023.
- [8] B. D. Ripley: *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.

Dušan Oberta, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 61669 Brno, Česká republika,
e-mail: 228584@vutbr.cz

MATEMATICKÉ METODY V NĚKTERÝCH RANKINGOVÝCH MODELECH

LUBOMÍR PAŽOUREK

ABSTRAKT. Tento článek, napsán na základě stejnojmenné bakalářské práce autora, se zabývá matematickou podstatou některých rankingových metod. Jejich jednotlívým prvkem je tzv. Perronova-Frobeniova věta pro nezáporné a ireducibilní matice, která formuluje podmínky pro existenci kladného vlastního čísla a kladného vlastního vektoru dané matice. Cílem článku je uvést přehled potřebných teoretických výsledků, vysvětlit jejich aplikaci v rámci některých rankingových metod a provést simulace při vyhodnocení některých soutěží. Následně porovnat uvedené metody s některými, které se často používají v šachových turnajích.

1. ÚVOD

Už od pradávna lidé mezi sebou navzájem soupeřili a snažili se být těmi nejlepšími. Pro respektování dosažených výsledků a jednotlivých umístění v co největší míře musela být nastavena pravidla, podle kterých se různé věci bodovaly a podle kterých se určovalo průběžné i závěrečné pořadí. Ať už to bylo ve sportovních soutěžích týmů, které budeme rozebírat právě my, nebo v libovolných jiných, kde je cílem sestavit nějaký žebříček podle výkonnosti či úspěšnosti.

Otázka rankingu – tedy sestavení pořadí na základě vybraných faktorů – je někdy velmi jednoduchá, někdy zase velmi obtížná. Existují například atletické disciplíny, ve kterých rozhoduje o pořadí pouze výkonnost. Ve skoku vysokém pořadí atleta či atletky určuje nejvyšší zdolaná výška, v běhu na 1000 metrů pouze čas. Klasifikace je z tohoto hlediska objektivní a nezavdává příčinu k případným stížnostem.

V případě hodnocení týmových soutěží, především fotbalu a hokeje v Česku nebo Evropě, o pořadí rozhoduje celkový bodový zisk, který je určen na základě výsledků jednotlivých zápasů a předem stanovených kritérií. V české fotbalové lize se každý tým s ostatními celky potká dvakrát, v české extralize ledního hokeje

2020 MSC. Primární 15A18.

Klíčová slova. hodnotící, turnajová a ireducibilní matice, vlastní číslo, vlastní vektor, Perronova-Frobeniova věta, rankingové metody.

Článek vznikl na základě bakalářské práce autora v oboru Matematické inženýrství na FSI VUT v Brně. Vedoucí práce byl Jan Čermák z Ústavu matematiky FSI VUT v Brně.

Velmi bych chtěl poděkovat prof. RNDr. Janu Čermákovi, CSc. za velmi vstřícný přístup, veškerou pomoc, cenné rady a připomínky při psaní závěrečné práce i tohoto článku.

každý s každým čtyřikrát. Týmy se mezi sebou potkávají tedy stejnou měrou, takže v tomto ohledu nemůže nikdo namítat, že konečné pořadí není spravedlivé.

Existují ale i soutěže, ve kterých týmy mezi sebou nehrají stejný počet zápasů, nebo se dokonce některé nestřetnou vůbec. Poté pořadí sestavené na konci celého ročníku pouhým sečtením bodů může vyvolat velkou nevoli mezi příznivci jednotlivých týmů, protože se mohou domnívat, že zrovna jejich tým měl těžší los a díky tomu skončil v pořadí níže než měl, a než fanoušci doufali.

Pokusíme se tedy najít takové rankingové systémy, které nebudou zohledňovat jen bodové zisky z utkání, ale i kvalitu jednotlivých protivníků. V tomto článku se zaměříme zejména na ty, které využívají různých, zajímavých matematických metod a teoretických výsledků z lineární algebry. Následně se pokusíme zhodnotit, jestli by bylo lepší používat k sestavování pořadí právě tyto rankingové systémy, v čem by byly jejich benefity, a v čem naopak případná úskalí.

Článek je rozčleněn do sedmi kapitol. V druhé kapitole zavedeme a okomentujeme důležité související matematické pojmy. V třetí kapitole se detailněji podíváme na znění Perronovy-Frobeniovy věty, která je matematickým základem některých významných rankingových metod. Ve čtvrté kapitole ukážeme, jak reprezentovat tabulky s body z jednotlivých utkání maticovým zápisem, a tyto speciální matice a některé jejich vlastnosti podrobněji prozkoumáme. Následně uvedeme využití kladných vlastních vektorů garantových Perronovou-Frobeniovou větou v uvažovaných rankingových systémech. Prozkoumáme také různé varianty bodových ohodnocení zápasů. V další kapitole si budeme ilustrovat použití navržených rankingových systémů na třech příkladech z reálného světa sportu, konkrétně na fotbalové Premier League, hokejové NHL a lize amerického fotbalu NFL. A v předposlední kapitole si námi popisované metody porovnáme s metodami, které jsou používány v šachových turnajích.

2. PŘEHLED MATEMATICKÉHO APARÁTU

Nejprve si připomeneme některé základní pojmy teorie matic spolu se základními komentáři. S těmito pojmy další část tohoto textu pracuje. Při zavedení těchto pojmů budeme předpokládat, že A je čtvercová matice řádu n (tuto skutečnost nebudeme opakovat).

2.1. Vlastní čísla a vlastní vektory

Začneme s pojmem vlastní číslo a vlastní vektor, které v tomto článku hrají velmi významnou roli.

Definice 2.1. Necht I je jednotková matice řádu n . Pak algebraickou rovnici

$$\det(A - \lambda I) = 0$$

s neznámou λ nazýváme *charakteristickou rovnici* matice A . Kořeny této charakteristické rovnice pak nazveme *vlastní čísla* matice A . Je-li λ vlastní číslo A , pak existuje nenulové řešení $\mathbf{r}$ soustavy

$$(A - \lambda I)\mathbf{r} = \mathbf{0},$$

ktelé nazveme *vlastním vektorem* matice A příslušný vlastnímu číslu λ .

Problematika vlastních čísel a vlastních vektorů dané matice je v literatuře zkoumána z mnoha různých pohledů. Nás bude zajímat především rozložení vlastních čísel podle jejich velikostí.

Definice 2.2. Necht $\lambda_1, \dots, \lambda_n$ jsou vlastní čísla matice A . Vlastní číslo λ_1 nazveme *dominantním vlastním číslem*, jestliže platí

$$|\lambda_1| > |\lambda_i| \quad \text{pro všechna } i = 2, \dots, n.$$

Dominantní vlastní číslo tedy udává tzv. *spektrální poloměr* matice A , který určuje poloměr uzavřeného kruhu v komplexní rovině (se středem v počátku), ve kterém leží všechna vlastní čísla matice A .

Problém nalezení vlastních čísel a vlastních vektorů matice A patří do oblasti numerické matematiky (tyto hodnoty jsme schopni analyticky určit pouze pro matice nízkých řádů). Metody, které jsou přitom využívány, se odlišují různými hledisky. Jedním z nich je, zda hledáme všechna vlastní čísla, nebo jen některá vybraná (například dominantní). V následující kapitole připomeneme princip metody, která umožní nalézt (za jistých předpokladů) dominantní vlastní číslo dané matice, a jemu odpovídající vlastní vektor. Jak uvidíme, tento vektor může mít, při splnění určitých požadavků, významné postavení při užití některých rankingových metod.

V další části připomeneme některé speciální typy čtvercových matic.

2.2. Některé typy matic

Definice 2.3. Čtvercová matice P řádu n se nazývá permutační matice, je-li možno P získat z jednotkové matice I stejného typu postupnou výměnou řádků.

Permutační matice mají řadu užitečných vlastností, z nichž připomeneme následující:

- Výměnu řádků (resp. sloupců) dané matice A můžeme reprezentovat pomocí součinu PA (resp. AP).
- Pro každou permutační matici P platí $P^{-1} = P^T$.

Příklad 2.4. Uvažujme obecnou matici čtvrtého řádu

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix},$$

ve které bychom chtěli zaměnit druhý řádek se třetím. Provedme tuto operaci pomocí permutační matice. Zvolíme si tedy jednotkovou matici, zaměníme v ní stejné řádky, které chceme zaměnit v matici A , a vzniklou matici označíme P .

Poté touto maticí vynásobíme matici A zleva:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \implies PA = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}.$$

Došlo tedy skutečně k požadované záměně řádků. Ukažme si, že naopak při vynásobení matice A zprava touto permutační maticí P , se namísto řádků prohodi sloupce:

$$AP = \begin{pmatrix} a_{11} & a_{13} & a_{12} & a_{14} \\ a_{21} & a_{23} & a_{22} & a_{24} \\ a_{31} & a_{33} & a_{32} & a_{34} \\ a_{41} & a_{43} & a_{42} & a_{44} \end{pmatrix}.$$

Definice 2.5. Matice A je tzv. *horní trojúhelníková matice*, právě když platí

$$a_{ij} = 0 \quad \text{pro všechna } i > j.$$

Pokud je navíc nulová i hlavní diagonála, jedná se o tzv. *ryze horní trojúhelníkovou matici*.

Na závěr zavedeme speciální typ matice, která je v jistém smyslu nerozložitelná.

Definice 2.6. Matice A se nazývá *ireducibilní*, pokud je řádu jedna, nebo neexistuje žádná permutační matice P taková, že PAP^T je v tzv. blokově horním trojúhelníkovém tvaru, tj.

$$P^T AP = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

kde A_{11} (resp. A_{22}) je čtvercový blok řádu k (resp. $n - k$). Matice A_{12} je pak rozměru $k \times (n - k)$.

Poznamenejme, že ireducibilita znamená, že žádným prohozením řádků a sloupců v dané matici nedostaneme výše popsany blokově horní trojúhelníkový tvar. Dodejme ještě, že podmínku, aby matice A byla ireducibilní, můžeme formulovat ekvivalentně i v jiném tvaru (viz např. [3] a [4]).

3. PERRONOVA-FROBENIOVA VĚTA

Základní tvrzení obsažená v této kapitole byla čerpána z knih [3] a [4]. Tato tvrzení se týkají speciálních spektrálních vlastností matic. Později uvidíme, že v některých rankingových metodách hraje důležitou roli kladný vlastní vektor dané čtvercové matice A (tímto termínem rozumíme vektor pouze s kladnými souřadnicemi). Cílem této kapitoly je proto zformulovat podmínky kladené na matici A , které existenci takového vektoru zaručí.

Začneme nejprve následující početní úvahou. Vezměme pro jednoduchost matici druhého řádu a položme si otázku, za jakých podmínek, kladených na koeficienty této matice, existuje dominantní kladné vlastní číslo této matice takové, že odpovídající vlastní vektor má kladné souřadnice.

Příklad 3.1. Uvažujme obecnou matici druhého řádu

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

s charakteristickou rovnicí

$$\det(A - \lambda I) = \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0.$$

Vlastní čísla matice A tedy vypočítáme takto:

$$\lambda_{1,2} = \frac{a_{11} + a_{22} \pm \sqrt{(a_{11} + a_{22})^2 - 4a_{11}a_{22} + 4a_{12}a_{21}}}{2}$$

Požadujeme, aby λ_1 bylo dominantní, kladné, reálné vlastní číslo. Pokud má být vlastní číslo reálné, diskriminant musí být nezáporný. Aby bylo navíc dominantní, musí být jednoduché. Diskriminant musí být tudíž kladný. Máme tedy

$$D = (a_{11} + a_{22})^2 - 4a_{11}a_{22} + 4a_{12}a_{21} > 0,$$

tj.

$$D = (a_{11} - a_{22})^2 + 4a_{12}a_{21} > 0.$$

Uvedená nerovnost je zaručená mj. tehdy, je-li $a_{12}a_{21} > 0$.

Dominantní vlastní číslo má být navíc kladné, což je zaručeno tehdy, je-li $a_{11} + a_{22} > 0$.

Shrnutím dvou odvozených nerovností dostáváme, že volíme-li koeficienty dané matice jako kladná čísla, pak existuje vlastní číslo s požadovanými vlastnostmi.

Podívejme se, jak při podmínce na kladnost koeficientů bude vypadat vlastní vektor $\mathbf{r}_1$ příslušný dominantnímu vlastnímu číslu λ_1 .

Vlastní vektor $\mathbf{r}_1$ dostaneme jako řešení této soustavy rovnic:

$$(A - \lambda_1 I)\mathbf{r}_1 = \begin{pmatrix} a_{11} - \lambda_1 & a_{12} \\ a_{21} & a_{22} - \lambda_1 \end{pmatrix} \begin{pmatrix} r_{11} \\ r_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

kde r_{11} a r_{12} jsou souřadnice vektoru $\mathbf{r}_1$.

Protože jsou tyto rovnice lineárně závislé, lze je zjednodušit na tvar

$$(\lambda_1 - a_{11})r_{11} = a_{12}r_{12}.$$

Podle předpokladu jsou všechny koeficienty matice A kladné, proto, aby i souřadnice vlastního vektoru příslušného dominantního vlastního čísla byly kladné, musí platit $\lambda_1 > a_{11}$. Ekvivalentně vyjádřeno

$$\frac{a_{11} + a_{22} + \sqrt{(a_{11} - a_{22})^2 + 4a_{12}a_{21}}}{2} > a_{11},$$

tj.

$$\sqrt{(a_{11} - a_{22})^2 + 4a_{12}a_{21}} > a_{11} - a_{22}.$$

Obě strany poslední nerovnosti umocníme a dostáváme

$$(a_{11} - a_{22})^2 + 4a_{12}a_{21} > (a_{11} - a_{22})^2,$$

tj.

$$4a_{12}a_{21} > 0.$$

Tato podmínka podle předpokladu platí.

Ukázali jsme tedy, že má-li matice druhého řádu kladné prvky, pak existuje dominantní, kladné, reálné vlastní číslo, a jemu odpovídající vlastní vektor $\mathbf{r}_1$ má kladné souřadnice.

Otázkou nyní je, zda lze tuto početní úvahu rozšířit i na matice obecného řádu. Kladnou odpověď dal v roce 1907 Oscar Perron¹.

Věta 3.2 (Perronova). *Nechť A je kladná čtvercová matice. Pak existuje reálné kladné vlastní číslo λ matice A , které je jednoduchým kořenem charakteristické rovnice, a navíc představuje spektrální poloměr pro danou matici A . Tomuto vlastnímu číslu λ odpovídá vlastní vektor $\mathbf{r}$ s kladnými souřadnicemi (tj. $r_i > 0$ pro všechna $i = 1, 2, \dots, n$).*

V roce 1912 rozšířil znění Perronovy věty na obecnější případ Georg Frobenius².

Věta 3.3 (Perronova-Frobeniova). *Nechť A je nezáporná ireducibilní čtvercová matice. Pak existuje reálné kladné vlastní číslo λ , které je jednoduchým kořenem charakteristické rovnice, a navíc představuje spektrální poloměr pro danou matici A . Tomuto vlastnímu číslu λ odpovídá vlastní vektor $\mathbf{r}$ s kladnými souřadnicemi (tj. $r_i > 0$ pro všechna $i = 1, 2, \dots, n$).*

Vektor, jehož existenci zaručuje Perronova-Frobeniova věta, budeme nazývat *Perronovým vektorem*.

Otázkou, kterou Frobenius v této souvislosti položil, je, zda uvedené vlastní číslo λ , které je spektrálním poloměrem dané matice A , je navíc i dominantním vlastním číslem. Proto zavedl následující definici.

Definice 3.4. Nezáporná čtvercová matice A se nazývá *primitivní*, pokud je ireducibilní a má dominantní vlastní číslo.

Jednoduchou podmínku, aby matice A byla primitivní, udává následující tvrzení.

Věta 3.5. *Je-li čtvercová matice A nezáporná, pak je primitivní právě tehdy, když A^m je kladná matice pro vhodné přirozené m .*

Na závěr této kapitoly se zabýváme otázkou, jak lze určit Perronův vektor dané matice A numericky. Jako účinná metoda se jeví zejména mocninná metoda (viz např. [5]), která umí nalézt (v přibližném tvaru) vlastní vektor odpovídající dominantnímu vlastnímu číslu (což je za výše uvedených podmínek právě Perronův vektor).

¹Oscar Perron (1880 - 1975) byl matematik německé národnosti a profesor na univerzitě v Heidelbergu a Mnichově. Zabýval se hlavně diferenciálními, parciálními diferenciálními rovnicemi a lineární algebrou. Kromě již zmíněného teorému je například autorem tzv. Perronova paradoxu.

²Georg Frobenius (1849 - 1917) byl matematik německé národnosti, který se zabýval diferenciálními rovnicemi nebo teorií čísel. Je autorem řady známých výsledků, například jako první dokázal Cayley-Hamiltonovu větu. Byl studentem Karla Weierstrasse.

3.1. Mocninná metoda

Uvažujme čtvercovou matici A a předpokládejme, že má dominantní vlastní číslo λ_1 . Pak *mocninná metoda* vychází ze vztahu

$$\lim_{n \rightarrow \infty} A^n \mathbf{r}_0 = \mathbf{r},$$

kde $\mathbf{r}_0$ je startovací vektor, který je lineární kombinací vlastních vektorů matice A , a $\mathbf{r}$ je vlastní vektor příslušný dominantnímu vlastnímu číslu λ_1 .

Poznamenejme, že metoda má i svoje nedostatky, např. je-li $|\lambda_1| > 1$, může dojít k tzv. přetečení, a pro $|\lambda_1| < 1$ může dojít naopak k podtečení. Proto je vhodné v každém aproximačním kroku vektor $\mathbf{r}_k = A^k \mathbf{r}_0$ normovat, a použít místo toho tzv. *normalizovanou* mocninnou metodu.

3.2. Normalizovaná mocninná metoda

Normalizovaná mocninná metoda je dána vztahem

$$\lim_{n \rightarrow \infty} \frac{A^n \mathbf{r}_0}{\|A^n \mathbf{r}_0\|} = \mathbf{r},$$

kde $\mathbf{r}_0$ je startovací vektor, $\mathbf{r}$ je vlastní vektor příslušný λ_1 a $\|\cdot\|$ je euklidovská norma.

Poznamenejme ještě, že příslušné dominantní vlastní číslo λ_1 pak lze dopočítat z tzv. Rayleighova podílu

$$\lambda_1 = \frac{\mathbf{r}^T A \mathbf{r}}{\mathbf{r}^T \mathbf{r}}.$$

4. HODNOTÍCÍ A TURNAJOVÁ MATICE

V této kapitole zavedeme dva speciální typy matic, které již mají přímou vazbu na hodnocení sportovních týmů.

4.1. Hodnotící matice

Výsledky zápasů v rámci soutěže jsou vždy někde zapsány a reprezentovány různými způsoby. Na mnoha turnajích, kde se nehraje hned od začátku vyřazovací část, jsou výsledky často reprezentovány tabulkou. Pokud nepočítáme záhlaví a první sloupec s názvy týmů, tabulka pro turnaj o n týmech má obvykle n sloupců a n řádků.

Týmům se na začátku turnaje přiřadí jejich pozice v tabulce. Její buňky pak jsou body ze vzájemného střetnutí, případně skóre (například 3:5), nebo kombinace obou faktorů.

Ilustrujme tento pojem na příkladu fotbalového turnaje, jehož se zúčastní týmy T_1 až T_5 , přičemž časové okolnosti nedovolí, aby se všechny týmy navzájem utkaly. Užijeme přitom tradiční bodové hodnocení, kde za výhru se udělí tři body, za remízu jeden bod a za prohru žádný. V případě, že týmy proti sobě nehrály, políčko necháme proškrtlé. Příklad příslušné tabulky ukazuje tabulka 4.1. Ta popisuje turnaj, kdy tým T_1 porazil tým T_2 , tým T_3 remizoval s týmem T_1 , tým T_3 se neutkal s týmem T_5 , atd. Políčko tabulky v i -tém řádku a j -tém sloupci tedy

—	Tým T_1	Tým T_2	Tým T_3	Tým T_4	Tým T_5
Tým T_1	—	3	1	0	1
Tým T_2	0	—	0	3	3
Tým T_3	1	3	—	0	—
Tým T_4	—	0	3	—	0
Tým T_5	1	0	3	0	—

Tabulka 4.1. Příklad výsledkové tabulky z fotbalového turnaje.

značí výsledek zápasu týmu T_i s T_j . Diagonální pole z pochopitelných důvodů proškrtáváme.

Turnajovou tabulku lze z hlediska matematického zápisu reprezentovat tzv. *hodnotící maticí*.

Definice 4.1. Hodnotící maticí A rozumíme nezápornou čtvercovou matici řádu n s nulovými prvky na hlavní diagonále:

$$A = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{pmatrix}.$$

V této matici tedy prvek a_{ij} reprezentuje celkový bodový zisk týmu i proti týmu j . Je-li tento prvek nulový, tak to obvykle znamená, že tým i proti týmu j vůbec nehrál, nebo s ním prohrál (která z těchto variant nastává lze snadno poznat z hodnoty prvků a_{ji}). Poznamenejme, že v některých hodnotících maticích se kladný bodový zisk uděluje i za prohru.

Volbě koeficientů hodnotící matice se budeme podrobněji věnovat v následujících kapitolách. Nejprve však uvedeme dva příklady bodování z hokejového a basketbalového turnaje.

4.1.1. Hokej. Během prvního kola hokejového turnaje o pěti týmech T_1, T_2, T_3, T_4, T_5 bylo odehráno šest utkání s výsledky uvedenými v tabulce 4.2.

Domáci	Hosté	Výsledek utkání
tým T_1	tým T_2	1:2 v základní hrací době
tým T_1	tým T_3	3:1 v základní hrací době
tým T_4	tým T_5	4:1 v základní hrací době
tým T_2	tým T_5	2:1 po samostatných nájezdech
tým T_4	tým T_1	5:3 v základní hrací době
tým T_3	tým T_4	3:2 v prodloužení

Tabulka 4.2. Výsledky hokejových zápasů ilustračního příkladu.

Pokud uvažujeme typické hokejové bodování, pak prvky hodnotící matice volíme takto:

$$a_{ij} = \begin{cases} 3, & \text{tým } i \text{ zvítězil nad týmem } j \text{ v základní hrací době,} \\ 2, & \text{tým } i \text{ zvítězil nad týmem } j \text{ v prodloužení nebo SN,} \\ 1, & \text{tým } i \text{ prohrál s týmem } j \text{ v prodloužení nebo SN,} \\ 0, & \text{tým } i \text{ prohrál proti týmu } j \text{ v základní hrací době,} \\ & \text{nebo se týmy spolu nestřetly.} \end{cases}$$

Příslušná hodnotící matice tohoto turnaje má tedy tvar

$$A = \begin{pmatrix} 0 & 0 & 3 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 & 0 \\ 3 & 0 & 1 & 0 & 3 \\ 0 & 2 & 0 & 0 & 0 \end{pmatrix}.$$

4.1.2. Basketbal. Při ilustraci basketbalového turnaje předpokládáme pro lepší názornost, že zápasy pěti týmů dopadly stejně jako v případě hokejového turnaje, viz tabulku 4.3 (pouze upravíme dosažené koše oproti vstřeleným gólům do realis-

Domáci	Hosté	Výsledek utkání
tým T_1	tým T_2	88:92 v základní hrací době
tým T_1	tým T_3	94:64 v základní hrací době
tým T_4	tým T_5	106:76 v základní hrací době
tým T_2	tým T_5	87:78 v prodloužení
tým T_4	tým T_1	111:98 v základní hrací době
tým T_3	tým T_4	82:75 v prodloužení

Tabulka 4.3. Výsledky basketbalových zápasů ilustračního příkladu.

tických čísel). Při užití běžného bodování basketbalových zápasů

$$a_{ij} = \begin{cases} 2, & \text{tým } i \text{ vyhrál proti týmu } j, \\ 1, & \text{tým } i \text{ prohrál s týmem } j, \\ 0, & \text{tým } i \text{ neodehrál zápas proti týmu } j \end{cases}$$

dostaneme následující hodnotící matici

$$A = \begin{pmatrix} 0 & 1 & 2 & 1 & 0 \\ 2 & 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 2 & 0 & 2 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

4.2. Turnajová matice

V této sekci se nejdříve zaměříme na maticové popisy tzv. ideálních turnajů. Tím máme na mysli soutěže, ve kterých všechny týmy hrají proti sobě právě jednou, a každý zápas má svého vítěze (tj. remíza není možná). V anglické literatuře lze tento typ turnaje najít pod názvem „single round robin tournament“, který pro přehlednost zkrátíme pouze na „round-robin“.

Výsledky utkání z takových turnajů reprezentuje hodnotící matice, kde v případě vítězství týmu i nad týmem j klademe $a_{ij} = 1$, v případě prohry $a_{ij} = 0$. Tento speciální případ hodnotící matice (obsahující pouze jedničky nebo nuly) má mnoho zajímavých vlastností, například v souvislosti se sestavováním pořadí týmů. Proto se pro tento typ hodnotící matice zavedl speciální název, tzv. *turnajová matice*. Turnajovou maticí (popsanou v [2]) tedy rozumíme každou hodnotící matici A , jejíž prvky splňují $a_{ij} = 0$ nebo 1 , přičemž $a_{ij} + a_{ji} = 1$ pro každé $i \neq j$.

Přirozenou otázkou je, jak lze z této matice odvodit nějaké závěry o síle jednotlivých týmů. Než tuto otázku zodpovíme, shrneme některé základní vlastnosti turnajových matic.

4.2.1. Vlastnosti turnajové matice. Nejprve uvedeme čtyři základní vlastnosti turnajových matic, které budeme následně ilustrovat a komentovat na příkladu.

Vlastnost 1. Nechť I je jednotková matice a J je čtvercová matice téhož řádu, jejímiž prvky jsou samé jedničky. Pak pro každou turnajovou matici A platí

$$A + A^T = J - I.$$

Vlastnost 2. Je-li A turnajová matice, pak je A^T také turnajovou maticí.

Vlastnost 3. Je-li A turnajová matice, pak každá submatice matice A , vzniklá vynecháním řádků a sloupců stejných indexů, je turnajová matice.

Vlastnost 4. Nechť P je libovolná permutační matice. Je-li A turnajová matice, pak PAP^T je také turnajová matice.

Příklad 4.2. Uvažujme turnajovou matici

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

která reprezentuje turnaj, kde tým T_1 porazil tým T_4 , T_3 porazil T_2 , tým T_4 porazil T_3 , atd. Pro upřesnění, výsledky týmu T_i zachycuje i -tý řádek turnajové matice.

Snadno ověříme, že vlastnost 1 platí, neboť

$$A + A^T = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} = J - I.$$

Podobně, transponovaná matice

$$A^T = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

je také turnajová.

Všimněme si její souvislosti s původní turnajovou maticí A . Protože při transpozici čtvercové matice zaměníme prvky symetricky podle hlavní diagonály, matice A^T představuje turnaj, během kterého dopadly všechny výsledky zápasů přesně naopak ve srovnání s turnajem reprezentovaným původní maticí A .

Odvození vlastností 1 a 2 je triviální a je tomu i v případě vlastnosti 3. Skutečně, při vynechání libovolného počtu řádků a sloupců stejných indexů dostaneme matici s nulovou hlavní diagonálou, prvky a_{ij} rovny nule nebo jedničce, přičemž $a_{ij} + a_{ji} = 1$ pro $i \neq j$. Každá taková submatice je proto také turnajovou maticí. Pro ilustraci si uvedeme dva příklady těchto submatic:

$$A_{2,3} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

Matice vlevo značí turnajovou matici při vynechání druhého, třetího řádku a sloupce. Napravo pak při vynechání řádku a sloupce s indexem jedna.

Konečně vlastnost 4 nám říká, že že prohozením libovolného počtu řádků a sloupců stejných indexů dostaneme také turnajovou matici.

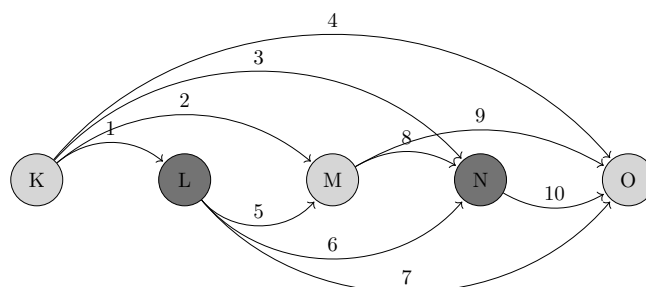
4.2.2. Počet turnajových matic a různých turnajů. Zajímavou otázkou je, jaký je celkový počet turnajových matic dimenze n , a dále kolik existuje různých turnajů. Mohlo by se zdát, že se jedná o ekvivalentní pojmy a není mezi nimi žádný rozdíl; později ukážeme, že ve skutečnosti je tomu ale naopak. Určeme celkový počet zápasů „round robin“ turnaje o n týmech. Tento počet je roven výrazu

$$(n-1) + (n-2) + (n-3) + \cdots + 3 + 2 + 1 = \sum_{k=1}^{n-1} (n-k).$$

Tato hodnota je také ilustrována na obrázku 4.1 znázorňující počet utkání turnaje pěti mužstev.

Ekvivalentně, výsledný počet všech utkání lze také vyjádřit pomocí kombinačního čísla, protože hledáme všechny různé kombinace dvou týmů z n celkových. (Jedná se o klasickou kombinaci bez opakování.) Výsledný počet je potom roven $\binom{n}{2}$.

Známe-li celkový počet utkání, můžeme určit počet všech možných turnajových matic. V uvažovaném turnaji může každý zápas dopadnout dvěma způsoby. Neznáme-li výsledek jednoho zápasu, pak máme na výběr ze dvou turnajových matic. Pokud neznáme výsledky dvou zápasů, vybíráme mezi čtyřmi různými turnajových matic, atd. Obecně, je-li m neznámých výsledků, pak dostáváme 2^m možných turnajových matic. Protože celý „round robin“ turnaj má $\binom{n}{2}$ zápasů, existuje $2^{\binom{n}{2}}$



Obrázek 4.1. Grafické znázornění počtu zápasů.

turnajových matic pro tento turnaj n týmů, kde každý hraje s každým pouze jednou.

Nyní se vraťme k otázce jednoznačného vztahu mezi výsledky daného turnaje a příslušnou turnajovou maticí. Uvažujme turnajovou matici

$$B = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Při srovnání s maticí A z Příkladu 4.2 se zdá, že matice B reprezentuje jiný turnaj. Ve skutečnosti však může reprezentovat ten stejný. Všimněme si, že matice B vznikla z matice A pouhou záměnou prvního a třetího řádku a sloupce. Pokud bychom tedy ve stejném smyslu zaměnili i řádkové a sloupcové indexy jednotlivých týmů, matice B reprezentuje stejné výsledky turnaje jako matice A .

Pro lepší přehlednost je tato operace naznačena ve dvou tabulkách. Tabulka 4.4 odpovídá turnajové matici A , tabulka 4.5 turnajové matici B (kterou jsme obdrželi výše uvedenou transformací matice A včetně záměny příslušných indexů). Jinak

—	Tým T_1	Tým T_2	Tým T_3	Tým T_4
Tým T_1	—	0	0	1
Tým T_2	1	—	0	0
Tým T_3	1	1	—	0
Tým T_4	0	1	1	—

Tabulka 4.4. Vyjádření turnajové matice A .

—	Tým T_3	Tým T_2	Tým T_1	Tým T_4
Tým T_3	—	1	1	0
Tým T_2	0	—	1	0
Tým T_1	0	0	—	1
Tým T_4	1	1	0	—

Tabulka 4.5. Vyjádření turnajové matice B .

vyjádřeno, číslo $2^{\binom{n}{2}}$ nám tedy neudává počet různých turnajů, ale pouze počet možných tvarů turnajové matice.

Z předcházející diskuze vyplývá, že dva turnaje (reprezentované turnajovými maticemi například C a D) budeme považovat za různé, jestliže žádnou záměnou řádků a sloupců matice C neobdržíme matici D , matematicky vyjádřeno: neexistuje permutační matice P taková, že

$$D = PCP^T.$$

Pod pojmem různé turnaje budeme tedy rozumět, že jakoukoliv záměnou řádků a sloupců matice C příslušící jednomu turnaji nedostaneme matici D odpovídající turnaji druhému.

Tabulka 4.6 ilustruje počet různých turnajů při n účastnících ($n = 2, 3, \dots, 8$).

Počet účastníků n	počet tříd
2	1
3	2
4	4
5	12
6	56
7	456
8	6880

Tabulka 4.6. Tabulka počtu různých „round robin“ turnajů.

5. POUŽITÍ PERRONOVY-FROBENIOVY VĚTY V RANKINGOVÝCH METODÁCH

V předcházející kapitole jsme stručně uvedli pojmy hodnotící a turnajové matice. Nyní ukážeme, jak lze těchto matic využít při sestavení pořadí na základě vzájemných utkání. Uvedeme dva různé přístupy, které vedou na zajímavou aplikaci Perronovy-Frobeniovy věty.

5.1. Stanovení pořadí pomocí turnajové matice (Kendall-Wei metoda)

Základní poznatky pro tuto sekci byly brány z článku [2]. V turnaji, kde se každý tým utká se všemi ostatními právě jednou - jedná se o typ „round robin“, skončilo poslední utkání, a my bychom chtěli za pomoci turnajové matice A sestavit celkové pořadí účastníků (v našem případě týmů), aby na každé pozici byl nejlépe pouze jeden tým. To učiníme tak, že každému týmu přisoudíme tzv. *skóre*, tedy nezáporné reálné číslo, které vhodným způsobem charakterizuje výsledky zápasů s ostatními týmy. Následně skóre porovnáme. Čím větší bude hodnota skóre týmu, tím výše bude v pořadí umístěn.

Uvažujme tedy obecnou turnajovou matici

$$A = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{pmatrix}.$$

Začneme nejjednodušší a nejpřirozenější cestou – skóre každého týmu zavedeme jako součet jeho vítězství, tedy součet všech prvků v turnajové matici A na příslušném řádku.

Pro tým i bude výsledné skóre s_i tvaru

$$s_i = a_{i1} + a_{i2} + a_{i3} + \cdots + a_{in} = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij},$$

kde člen a_{ij} značí výsledek utkání proti týmu j .

Užitím maticového zápisu lze pak skóre $\mathbf{s}$ jednotlivých týmů psát ve tvaru

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} a_{12} + a_{13} + \cdots + a_{1n} \\ a_{21} + a_{23} + \cdots + a_{2n} \\ \vdots \\ a_{n1} + a_{n2} + \cdots + a_{(n-1)n} \end{pmatrix},$$

vektorově zapsáno

$$\mathbf{s} = A\mathbf{e}, \quad (5.1)$$

kde $\mathbf{e}$ je sloupcový vektor jedniček.

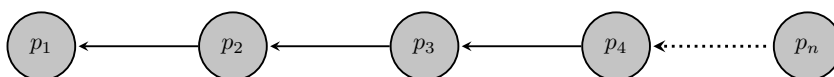
Dostali jsme tzv. *vektor skóre* $\mathbf{s}$, jehož i -tá složka odpovídá počtu vítězství i -tého týmu. Při stanovení pořadí tedy můžeme říci, že tým i je lepší než tým j , pokud platí nerovnost

$$s_i > s_j.$$

V případě, že každý tým bude mít různý počet vítězství, týmy jednoduše seřadíme na základě ostrých nerovností

$$s_{p_1} > s_{p_2} > \cdots > s_{p_n},$$

kde $((p_1 \ p_2 \ \dots \ p_n))$ je vhodná permutace prvků $1, \dots, n$. Situace, kdy každý



Obrázek 5.1. Seřazení týmů v nejjednodušším případě.

tým má odlišný počet vítězství, nastává pouze a jen pouze v případě, kdy se nám turnajovou maticí A povede, pomocí permutační matice P , převést na ryze horní trojúhelníkový tvar

$$PAP^T = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \ddots & 1 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Tento fakt si můžeme ověřit i podle vektoru skóre, kdy první tým bude mít $n - 1$ výher, druhý tým $n - 2$, atd., až dojdeme k poslednímu účastníkovi, který skončil s nula body – tedy nevyhrál ani jedno utkání. Pro vektor skóre platí

$$\mathbf{s} = \begin{pmatrix} n-1 \\ n-2 \\ \vdots \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

Často se ale ovšem stává, že více týmů má stejný počet vítězství, takže některé ostré nerovnosti v uspořádání týmů podle počtu výher jsou nahrazeny neostrými nerovnostmi. Týmům ale, jak už bylo výše zmíněno, nechceme přiřadit dělené místo, nebo jich alespoň chceme mít co nejméně. Proto se podíváme, jak při této situaci postupovat.

5.1.1. Blokový tvar matice. Turnajovou matici bychom mohli výměnou řádků a sloupců stejných indexů, tedy pomocí vhodné permutační matice P , převést na blokově horní tvar

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

kde A_{11} a A_{22} jsou čtvercové bloky řádu k a $n-k$. Matice A_{12} je dimenze $k \times (n-k)$.

Po provedení této transformace je turnajová matice tvaru

$$PAP^T = \begin{pmatrix} & & 1 & \cdots & 1 \\ & A_{11} & \vdots & & \vdots \\ & & 1 & \cdots & 1 \\ 0 & \cdots & 0 & & \\ \vdots & & \vdots & A_{22} & \\ 0 & \cdots & 0 & & \end{pmatrix}.$$

Můžeme si všimnout, že blok matice A_{12} obsahuje samé jedničky. Tato informace je důležitá především z toho hlediska, že každý z prvních k týmů porazil tým ze skupiny zbývajících $n-k$ týmů. Mohli bychom tedy konstatovat, že prvních k týmů je lepších (minimálně podle počtu vítězství), než zbylých $n-k$ týmů. Toto tvrzení si můžeme ověřit i na bodových ziscích. Prvních k týmů má alespoň $n-k$ výher, a zbytek vyhrál maximálně $n-k-1$ střetnutí.

Matematicky vyjádřeno: pro všechna $i = 1, 2, \dots, k$, a pro všechna $j = k+1, k+2, \dots, n$ platí

$$s_i > s_j.$$

Hodnotit a určovat pořadí pak stačí jen v rámci jednotlivých bloků, rozkládat je dál, a to tak dlouho, dokud je výše popsaná procedura proveditelná. Všimněme si, že horní trojúhelníkový tvar matice je vlastně pouze speciálním případem blokové matice, kde se nám matici podaří rozložit takovým způsobem, že konečné bloky jsou matice řádu jedna.

5.1.2. Ireducibilní matice. Pokud blok v nějakém kroku již dále transformovat výše popsaným způsobem nelze (speciálně, celá turnajová matice A nelze takto rozložit), budou v matici minimálně dva týmy se stejným počtem vítězství. Buď se smíříme s faktem, že některé týmy budou na stejných příčkách, nebo musíme pokračovat v hledání, jak určit jejich pořadí.

Maticí, kterou nemůžeme rozložit, je právě matice ireducibilní zavedená v druhé kapitole. V další části se proto zaměříme pouze na turnaje, kde se taková turnajová matice vyskytuje, protože v předchozích případech, kdy byla matice tzv. *reducibilní*, pořadí určit umíme.

Pojem ireducibility matic jde vysvětlit na orientovaných grafech. Potřebné pojmy zde užijeme pouze v intuitivním smyslu, protože nám především jde o názorovou ukázkou.

5.1.3. Silně souvislý graf. Graf se skládá z vrcholů a hran, které zobrazují vztahy právě mezi vrcholy. V orientovaném grafu se navíc dá určit, kde má každá hrana začátek a konec. Orientovaný graf nazveme silně souvislý, pokud se z každého vrcholu dostaneme do všech zbývajících vrcholů. Matice s nezápornými členy je pak ireducibilní pouze tehdy, pokud odpovídající orientovaný graf je silně souvislý (více podrobností viz [6]).

Nyní si ukážeme ilustrační příklad.

Příklad 5.1. Máme matice s výsledky dvou různých fotbalových turnajů pro pět týmů. Týmy označíme postupně A, B, C, D a E , a přiřadíme jim indexy v matici jedna až pět. Matice nazveme například M a O .

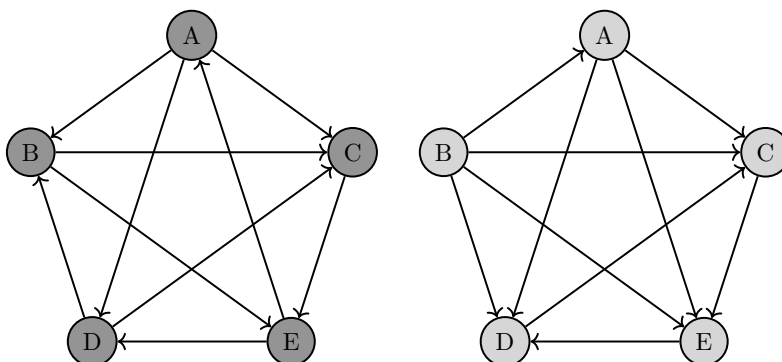
Účastníci proti sobě v každém turnaji hráli jen jednou, za výhru obdrželi jeden bod a za prohru žádný. Výsledky obou turnajů zachycují následující turnajové matice

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad O = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Orientovaný graf odpovídající příslušné turnajové matici sestrojíme tak, že si nejprve zakreslíme vrcholy odpovídající jednotlivým týmům s přidělenými indexy, viz obrázek 5.2. Kde není v matici nula uděláme hranu se začátkem v indexu řádku a koncem v indexu sloupce. Pokud tedy například $a_{51} = 1$, tzn. vyhrál tým s přiřazeným indexem pět nad celkem s indexem jedna, sestrojíme hranu s orientací z bodu E do A .

Z obrázku 5.2 lze vidět, že M je ireducibilní. Naopak turnajová matice O není, jedná se o reducibilní matici, kterou jsme schopni dokonce rozložit do tří bloků ve tvaru

$$POP^T = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$



Obrázek 5.2. Orientované grafy příslušné turnajovým maticím M a O .

Shrňme zatím nabyté poznatky. Pokud jsme turnajovou matici schopni převést na ryze horní trojúhelníkový tvar, má každý tým odlišný počet vítězství, a týmy jednoduše seřadíme právě podle počtu výher. V opačném případě jsme schopni turnajovou matici rozložit do bloků. V rámci různých bloků můžeme jednotlivé týmy porovnat, avšak uvnitř v jednotlivých blocích již ne. Budeme se tedy dále zabývat otázkou, jak určit pořadí v ireducibilních turnajových maticích.

Podívejme se na prvky turnajové matice. Pokud součin členů $a_{ij}a_{jk}$ je nenulový, vyplývá z toho, že tým i porazil j a tým j porazil k . Dá se tedy říci, že tým i je silnější než j , a ten je silnější než k . Součin můžeme zobecnit na

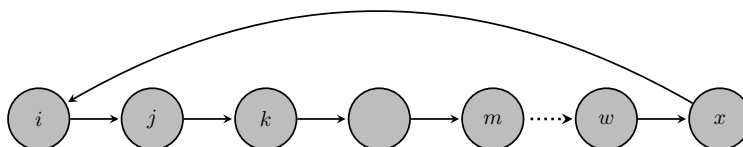
$$a_{ij}a_{jk}a_{kl}a_{lm}\dots a_{wx},$$

jehož nenulovost znamená, že tým i by měl být silnější než tým x .

Pro každou ireducibilní turnajovou matici v rámci těchto součinů narazíme na zajímavou situaci. V ireducibilní matici pro každého účastníka „round robin“ turnaje se objeví nenulový součin tvaru

$$a_{ij}a_{jk}a_{kl}a_{lm}\dots a_{wx}a_{xi},$$

z čehož vyplývá paradox, že tým i by měl být lepší než tým i . V praxi to znamená, že silný tým prohrál s outsiderem. Pro lepší představu si tento nenulový součin můžeme ukázat na grafu (šipka směřuje od vítězného týmu k poraženému), viz obrázek 5.3.



Obrázek 5.3. Grafické znázornění nenulového součinu koeficientů turnajové matice.

Toto je vlastně další zdůvodnění, proč se nám ireducibilní matice nepodaří nikdy převést na blokově horní tvar. V každém sloupci bude totiž vždy minimálně jedna jednička a dvě nuly (diagonální prvek plus minimálně jedna prohra).

5.1.4. Zohlednění výkonnosti soupeřů. Smyslem dalších úvah je navrhnout způsob hodnocení týmů v rámci ireducibilních turnajových matic.

Uvědomme si, že za silný tým se považuje takový tým, který poráží většinu svých protivníků. Naopak slabý tým moc zápasů nevyhrává, a když už se mu podaří zvítězit, tak poraženým je většinou tým podobné výkonnosti. Chtěli bychom tedy, aby výsledné skóre týmu nezáleželo pouze jen na výsledku utkání, ale i na tom, proti jakému soupeři se hrálo.

Nové skóre zadefinujeme jako součet výher všech týmů, které tým i porazil. Jelikož v našem případě tým za vítězství obdrží jeden bod, nově navržené skóre vypadá takto:

$$s_i = a_{i1}s_1 + a_{i2}s_2 + a_{i3}s_3 + \cdots + a_{in}s_n = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}s_j.$$

Výsledný počet bodů už tedy nezávisí pouze na výsledcích týmu, ale více reflektuje schopnosti protivníka – a o to nám přesně jde. Pro všechny týmy by maticový zápis takto modifikovaného skóre měl tvar

$$\mathbf{s}_2 = A^2 \mathbf{e}. \quad (5.2)$$

Abychom ještě více zohlednili kvalitu protivníků, vynásobme $\mathbf{s}_2$ opět maticí A zleva, čímž získáme nový vektor skóre

$$\mathbf{s}_3 = A^3 \mathbf{e}. \quad (5.3)$$

Jinak vyjádřeno pro i -tý tým získáme skóre

$$s_{3i} = a_{i1}s_{21} + a_{i2}s_{22} + a_{i3}s_{23} + \cdots + a_{in}s_{2n} = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}s_{2j}.$$

Slovní interpretace tohoto výrazu je, že se jedná o součet součtů všech vítězství týmů, které tým i porazil. U vektoru $A^3 \mathbf{e}$ se nám tedy ještě více projevila závislost mezi jednotlivými týmy.

Lze tedy konstatovat, že s rostoucí mocninou n dává vektor skóre

$$\mathbf{s}_n = A^n \mathbf{e}$$

stále věrohodnější výsledek. Uvážíme-li tedy vztahy (5.1), (5.2) a (5.3), pak v jistém smyslu ideální vektor skóre je dán vztahem

$$\lim_{n \rightarrow \infty} A^n \mathbf{e}.$$

Se zvyšováním mocniny matice A mohou složky vektoru skóre neúměrně narůstat. Protože nám jde pouze o porovnání jednotlivých složek, provedeme normování

$$\frac{A\mathbf{e}}{\|A\mathbf{e}\|}, \frac{A^2\mathbf{e}}{\|A^2\mathbf{e}\|}, \frac{A^3\mathbf{e}}{\|A^3\mathbf{e}\|}, \cdots, \frac{A^n\mathbf{e}}{\|A^n\mathbf{e}\|},$$

a na tuto posloupnost opět aplikujeme limitu

$$\lim_{n \rightarrow \infty} \frac{A^n \mathbf{e}}{\|A^n \mathbf{e}\|}.$$

Všimněme si, že získaný vektor skóre je právě Perronův vektor matice A vypočtený mocninnou metodou. Zdůrazněme přitom, že turnajová matice A je podle předpokladu ireducibilní a nezáporná, podle Perronovy-Frobeniovy věty musí Perronův vektor existovat.

Ilustrujme nově získaný způsob hodnocení na následujících dvou příkladech.

Příklad 5.2. V příkladu 5.1 jsme ukázali, že matice

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

která popisuje fotbalový turnaj pěti týmů A , B , C , D a E , je ireducibilní. Vektor skóre, který značí počet vítězství jednotlivých týmů je roven

$$\mathbf{s} = (3 \quad 2 \quad 1 \quad 2 \quad 2)^T.$$

Tři týmy mají stejný počet vítězství, a kdybychom je seřazovali pouze podle počtu vítězných zápasů, druhé místo by obsadily tři týmy, a tomu bychom se chtěli vyhnout. Spočítejme tedy modifikované skóre pomocí normalizované mocninné metody. Vypočtený Perronův vektor (za pomoci Matlabu při zvolené přesnosti $\epsilon = 10^{-5}$) dostáváme ve tvaru

$$\begin{pmatrix} 0,5798 \\ 0,4251 \\ 0,2765 \\ 0,3771 \\ 0,5143 \end{pmatrix}.$$

Vidíme tedy, že zohlednění kvality protivníka nám pomohlo, a týmy už můžeme seřadit na základě nerovností

$$s_A = 0,5798 > s_E = 0,5143 > s_B = 0,4251 > s_D = 0,3771 > s_C = 0,2765.$$

Na první místě se umístil tým A , druhé místo obsadil tým E , protože svá vítězství si uhrál proti silnějším soupeřům, než týmy se stejným počtem vítězství, a poslední volné místo na stupních vítězů si vybojoval tým B .

Ani výše popsaná metoda, využívající stanovení pořadí týmů podle Perronova vektoru, nemusí vést k jednoznačnému závěru. Tuto skutečnost ilustruje následující příklad.

Příklad 5.3. Druhou matici O uvažovanou v příkladu 5.1 jsme byli schopni rozložit do tří bloků, stručně

$$O = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \implies POP^T = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

kde dva čtvercové bloky jsou řádu jedna, zbývající je řádu tři. Vektor skóre turnajové matice O má tvar

$$\mathbf{s} = (3 \quad 4 \quad 1 \quad 1 \quad 1)^T.$$

První, resp. druhé místo je jasné, obsadil ho tým B , resp. tým A . Chtěli bychom určit, kdo je doplní na stupních vítězů, tedy pokusit se sestavit pořadí mezi týmy C , D a E . Můžeme si ale všimnout, že pořadí mezi nimi určit nepůjde, protože se nám zde vytvořil cyklus, v jehož rámci není možné rozhodnout, kdo je lepší, nebo horší. Vyzkoušejme, že tuto okolnost nám potvrdí i metoda založená na Perronově vektoru.

Uvažujme ireducibilní blok matice POP^T , který znázorňuje výsledky utkání mezi týmy C , D a E , jako samostatnou turnajovou matici

$$B = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

a pořadí určíme pouze mezi nimi (zdůrazněme, že výsledky jejich utkání s týmy A a B na tomto nic nemění). Přímým výpočtem jde ověřit, že Perronův vektor příslušný vlastnímu číslu $\lambda_1 = 1$ matice B (které je sice spektrálním poloměrem, nikoliv však dominantním vlastním číslem matice B) je roven

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Vidíme, že i podle Perronova vektoru složky skóre všech týmů jsou stejné, a tedy nemůžeme ani touto metodou rozhodnout o pořadí mezi nimi.

Závěr tohoto příkladu můžeme zobecnit v tom směru, že týmy nejsme schopni jednoznačně seřadit (pomocí výše popsaných metod), pokud je v každém řádku a každém sloupci nějakého ireducibilního bloku turnajové matice právě jedna jednička.

5.2. Stanovení pořadí pomocí hodnotící matice (Keenerova metoda)

Perronova-Frobeniova věta lze využít i v jiných typech soutěží než jen v „round robin“ turnajích. Jedná se především o soutěže, kde týmy proti sobě stejný počet zápasů nehrají (například NHL), nebo dokonce se některé týmy neutkají vůbec (například NFL). Velmi typické je to také pro různé univerzitní ligy napříč všemi kontinenty.

Na první pohled by se mohlo zdát, že bychom mohli bodový zisk každého týmu jednoduše vydělit počtem jeho zápasů (aby nebyly zvýhodněny týmy s větším počtem odehraných zápasů), určit průměrný zisk bodů na utkání, a podle toho sestavit pořadí. Tento způsob ale není vhodný z několika důvodů. Například týmy, které měly konkurenceschopnější soupeře, by byly znevýhodněné vůči celkům, které se převážně utkávaly se slabými týmy, a získávaly tak body snáze. Rozdíl v bodech na zápas by pak mezi těmito celky mohl být opravdu výrazný, a nezohledňoval by skutečné rozložení sil v soutěži.

Proto bychom chtěli zavést způsob hodnocení, kde by výsledný bodový zisk z utkání nezávisel pouze na výsledku, ale nějakým způsobem by reflektoval i protivníkovy schopnosti. Uvidíme, že dojdeme ke stejnému matematickému principu, jako v přechodí sekci. Výsledky, které uvedeme, jsou čerpány z článku [1].

5.2.1. Teoretické odvození. Chtěli bychom tedy, aby výsledný bodový zisk týmu z utkání nezáležel pouze na výsledku, ale i na soupeři, tedy na jakési síle protivníka. Mohli bychom ho tedy získat jako součin bodů z výsledku zápasu a právě síly soupeře. Je možné, že všechny týmy nebudou hrát proti sobě. Pokud se tak stane, počet bodů proti tomuto týmu je roven automaticky nule. Situaci, kdy má každý tým jiný počet zápasů, ošetříme vydělením celkového počtu bodů počtem odehraných zápasů n_i týmu i . Závěrečné skóre i -tého týmu bychom mohli psát tedy ve tvaru

$$s_i = \frac{1}{n_i}(b_{i1}r_1 + b_{i2}r_2 + b_{i3}r_3 + \dots + b_{in}r_n) = \frac{1}{n_i} \sum_{j=1}^n b_{ij}r_j \quad \text{pro } i = 1, 2, \dots, n, \quad (5.4)$$

kde b_{ij} značí celkový zisk bodů týmu i proti týmu j , r_j sílu týmu j , a n celkový počet týmů v soutěži.

Sestavení tvaru skóre bylo jednoduché. Nyní je potřeba vyřešit hlavní část, tedy najít vektor síly $\mathbf{r}$, který byl zatím popsán jen neurčitě. Je přirozené zavést sílu r_i i -tého týmu jako úměrnou jeho bodovému zisku s_i , tj.

$$s_i = \lambda r_i \quad \text{pro všechna } i = 1, 2, \dots, n, \quad (5.5)$$

kde $\lambda > 0$ je konstanta úměrnosti.

Do definičního vztahu (5.4) nyní za s_i dosadíme (5.5). Pro lepší přehlednost rovnici ještě zjednodušíme tak, že výraz $\frac{1}{n_i}b_{ij}$ nahradíme symbolem a_{ij} . Dostáváme tak

$$s_i = \sum_{j=1}^n a_{ij}r_j = \lambda r_i,$$

nebo-li vektorově

$$\mathbf{s} = A\mathbf{r} = \lambda\mathbf{r}, \quad (5.6)$$

kde $A = (a_{ij})$ je výše popsaná čtvercová matice řádu n .

Jinak vyjádřeno, námi hledaný vektor $\mathbf{r}$ síly týmů je vlastním vektorem matice A , podobně jako u turnajů, kde každé dva týmy mezi sebou odehrají právě jeden zápas bez možnosti remízy.

Chtěli bychom, aby se body jen přičítaly, a ne odečítaly. Proto máme požadavek, aby hledaný vlastní vektor obsahoval pouze kladné složky. Existenci takového

vlastního vektoru zajistí opět Perronova-Frobeniova věta. Všimněme si přitom, že konstanta úměrnosti λ mezi vektorem skóre a síly je právě spektrální poloměr matice A .

Zdůrazněme, že nyní už neuvažujeme turnajovou matici, ale používáme obecnější hodnotící matici, i ona je však z definice nezáporná. Pro užití Perronovy-Frobeniovy věty nám proto stačí pouze ověřit ireducibilitu A . V kladném případě výsledný vektor síly $\mathbf{r}$ vypočítáme pomocí normalizované mocninné metody, tedy

$$\mathbf{r} = \lim_{n \rightarrow \infty} \frac{A^n \mathbf{r}_0}{\|A^n \mathbf{r}_0\|},$$

kde $\mathbf{r}_0$ je vhodný startovací vektor.

5.2.2. Volba koeficientů hodnotící matice. Na začátek připomeňme, že možnou volbou koeficientů turnajové matice jsme se nezabývali, neboť její koeficienty jsou pevně dané (viz sekce 4.2). Naopak výběr koeficientů hodnotící matice je velmi subjektivní záležitostí. Omezením je pouze nezápornost každého prvku matice. Ukážeme, že i na jednoduše vypadajícím rozdělení bodů, kde po celou dobu budeme uvažovat, že součet bodů obou týmů $b_{ij} + b_{ji}$ ze vzájemného zápasu je rovný jedničce, můžeme dosáhnout velmi zajímavých interpretací.

Začneme tím, že vítěz si z utkání odnese jeden bod, poražený žádný. Při variantě, kdy zápas skončí remízou, si týmy jeden bod rozdělí rovným dílem, každý získá polovinu. Tedy

$$b_{ij} = \begin{cases} 1, & \text{tým } i \text{ vyhrál proti týmu } j, \\ 1/2, & \text{zápas skončil nerozhodně,} \\ 0, & \text{tým } i \text{ prohrál proti týmu } j. \end{cases}$$

Zajímavé údaje dostaneme, pokud se startovací vektor $\mathbf{r}_0$ (viz mocninná metoda) rovná vektoru samých jedniček, tedy $\mathbf{r}_0 = \mathbf{e}$. Při vynásobení matice A vektorem $\mathbf{r}_0$ zprava dostaneme jako v minulé sekci bodový zisk. Tentokrát se ale nebude jednat o celkový, nýbrž o průměrný zisk na zápas, protože dělíme počtem zápasů. Tento výsledek se ale dá interpretovat i trochu jinak.

V anglosaských zemích se někdy používá pojem „winning percentage“, nebo-li procento vyhraných zápasů. To dostaneme tak, že každému vítězství přidělíme váhu jedna, a každé remíze váhu poloviční. Potom tento pojem můžeme zachytit výrazem

$$w_i = \frac{W_i + \frac{1}{2}R_i}{N_i} \cdot 100\%,$$

kde W_i značí počet výher týmu i , R_i počet remíz a N_i celkový počet zápasů. Pokud se nyní podíváme detailněji, jak vypadá i -tý řádek našeho vektoru $A\mathbf{r}_0$, tak dojdeme k závěru, že díky našemu bodování (jeden bod za výhru, půl za remízu a nula za prohru) je vlastně roven výrazu w_i (vynecháme-li v něm vynásobení procenty), tedy

$$(A\mathbf{r}_0)_i = \sum_{j=1}^n a_{ij} = \frac{\sum_{j=1}^n b_{ij}}{n_i} = \frac{W_i + \frac{1}{2}R_i + 0P_i}{n_i},$$

kde písmeno P_i značí počet proher týmu i .

Podívejme se, co dále dostaneme v případě, když vektor $A\mathbf{r}_0$ vynásobíme zleva maticí A . Pro i -tou složku vektoru nyní platí

$$\begin{aligned}(A^2\mathbf{r}_0)_i &= \sum_{j=1}^n a_{ij} \frac{1}{n_j} \sum_{k=1}^n b_{jk} \\ &= a_{i1} \frac{1}{n_1} \sum_{k=1}^n b_{1k} + a_{i2} \frac{1}{n_2} \sum_{k=1}^n b_{2k} + \cdots + a_{in} \frac{1}{n_n} \sum_{k=1}^n b_{nk} \\ &= \frac{1}{n_i} \left(b_{i1} \frac{1}{n_1} \sum_{k=1}^n b_{1k} + b_{i2} \frac{1}{n_2} \sum_{k=1}^n b_{2k} + \cdots + b_{in} \frac{1}{n_n} \sum_{k=1}^n b_{nk} \right).\end{aligned}$$

Víme, že $\frac{1}{n_i} \sum_{j=1}^n b_{ij}$ je poměr počtu „vyhraných“ zápasů ku celkovému počtu zápasů (pro připomenutí, remízu uvažujeme stále jako polovinu výhry). Proto i -tá složka vektoru $A^2\mathbf{r}_0$ vyjadřuje průměr poměrů „vyhraných“ zápasů ku celkovému počtu zápasů týmů, se kterými tým i neprohrál.

Vektor $A^2\mathbf{r}_0$ je oproti $A\mathbf{r}_0$ interpretačně věrohodnější, ale pořád obsahuje nevýhodu, které bychom se rádi zbavili. Velký vliv na pořadí u $A^2\mathbf{r}_0$ má totiž náročnost programu – tedy zvýhodňuje celky, které hrají proti silnějším týmům častěji.

Tuto nevýhodu odstraníme postupným zvyšováním mocniny matice A . Proto je výsledný „ideální“ vektor síly opět roven výrazu

$$\mathbf{r} = \lim_{n \rightarrow \infty} \frac{A^n \mathbf{r}_0}{\|A^n \mathbf{r}_0\|}.$$

Bodování popsané výše můžeme bez problému použít, pokud se týmy během sezóny utkají vícekrát (při více zápasech je b_{ij} součet všech získaných bodů proti týmu j). S rostoucím počtem zápasů prvky matice b_{ij} a b_{ji} lépe vystihují vazby mezi oběma týmy.

V některých soutěžích se ale týmy potkávají pouze jednou ročně. Taková situace nastane například v některých univerzitních ligách v Severní Americe. Problémem potom je, že pokud byl zápas velmi vyrovnaný a skončil například 4:5, pak pouze vítěz bere celý bod a těsná výhra má stejnou váhu (ohodnocená jedním bodem) jako jasná převaha v utkání s výsledkem 5:0.

Mezi další nevýhody této volby koeficientů hodnotící matice bychom mohli také zařadit fakt, že tým se samými prohrami má relativní sílu nulovou a matice A v tom případě není ireducibilní. Navíc zápasy proti týmům s nulovou relativní silou jsou pro ostatní týmy nevýhodné, protože jim výhra nic nepřidá, naopak ubere, protože budou mít o zápas navíc, a tím nižší skóre.

5.2.3. Spojité rozdělování bodu. Spravedlivější by tedy bylo mezi týmy jeden bod rozdělit spojitě. Jedním z možných kritérií, jak bod rozdělit, se nabízí např. počet vstřelených gólů.

Pokud ve vzájemném zápase tým i vstřelil G_{ij} gólů, a tým j G_{ji} gólů, pak bychom mohli bod rozdělit pomocí vztahu

$$b_{ij} = \frac{G_{ij}}{G_{ij} + G_{ji}}.$$

Bodový zisk druhého účastníka utkání získáme pouhou záměnou indexů.

V případě nerozhodného stavu, $G_{ij} = G_{ji}$, si oba týmy stále odnesou po polovině, tj. $b_{ij} = b_{ji} = \frac{1}{2}$.

Nevýhodou takového rozdělení, podobně jako v diskrétním případě je, že hra může být po celou dobu utkání velmi vyrovnaná, ale skončit dejme tomu 1:0. V tom případě by vítěz bral celý bod a poražený odešel s prázdnou. Abychom takovéto situaci předešli, předchozí funkci trochu upravíme, aby si tým i přes prohru alespoň nějakou část bodu mohl odnést. Tím pádem docílíme i toho, že tým se samými prohrami bude mít nenulový bodový zisk, a tedy i nenulovou relativní sílu.

Bodové ohodnocení lze proto nahradit výrazem

$$b_{ij} = \begin{cases} \frac{G_{ij}+1}{G_{ij}+G_{ji}+2}, & \text{pokud } i \neq j, \\ \frac{1}{2}, & \text{pokud } i = j. \end{cases}$$

Pokud zápas skončí remízou, stále platí, že si týmy odnesou po polovině bodu.

Narazili jsme však na další nedokonalost, tudíž že naše nové rozdělení neošetřuje situaci, kdy si velmi silný tým bude chtít nahnat body vysokými výhrami proti slabším týmům. Hodilo by se nám tudíž rozdělit bod mezi účastníky střetnutí nelineárně. Jako možný návrh nelineárního bodového ohodnocení uveďme

$$b_{ij} = h\left(\frac{G_{ij} + 1}{G_{ij} + G_{ji} + 2}\right), \quad (5.7)$$

kde

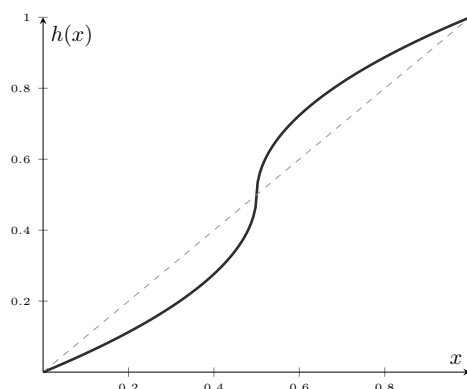
$$h(x) = \frac{1}{2} + \frac{1}{2} \operatorname{sgn}\left(x - \frac{1}{2}\right) \sqrt{|2x - 1|}. \quad (5.8)$$

Pro lepší názornost si funkci h vykreslíme do grafu, kde pro srovnání tempa růstu přidáme i graf přímé úměrnosti, viz obrázek 5.4 Jak lze vidět, remíza pořád půlí bod na dvě poloviny. Průběh této hodnotící funkce však zaručuje, že zatímco mezi výhrou a prohrou je citelný rozdíl, mezi výhrami 5:0 nebo 8:0 už tak velký rozdíl není.

Abychom ilustrovali užití této nelineární hodnotící funkce, přikládáme také příslušnou tabulku, viz tabulku 5.1. Můžeme si všimnout, že u některých výsledků utkání jsou body vítěze totožné. Je to dáno tím, že do funkce h vstupují stejné poměry. Při výhře 1:0 vítězný tým obdrží 0,7887 bodu. Podívejme se, pokud by tým už v zápase nějaké góly obdržel, kolikrát by musel celkově skórovat, aby si odnesl stejný bodový zisk jako při výhře 1:0.

Pro jednoduchost označme počet vstřelených branek vítězného týmu a , počet branek soupeře b . Při výhře 1:0 je poměr roven $2/3$. Proto můžeme psát

$$\frac{a+1}{a+b+2} = \frac{2}{3} \implies a = 2b + 1.$$



Obrázek 5.4. Graf funkce h (tmavě šedá) a funkce $y = x$.

Výsledek utkání	body vítěze	body poraženého
remíza	0,5	0,5
1:0	0,7887	0,2113
2:0	0,8536	0,1127
3:0	0,8873	0,1127
2:1	0,7236	0,2734
3:1	0,7887	0,2113
5:3	0,7236	0,2764

Tabulka 5.1. Nelineární rozdělení bodu při různých výsledcích.

Z tohoto vztahu tedy vyplývá, že pokud si chceme odnést 0,7887 bodu a už jsme branku inkasovali, musíme vstřelit tři. Pokud jsme dostali dvě, tak musíme skórovat pětkrát, atd. Pro jiné typy výsledků bychom uvažovali podobně.

5.3. Nelineární rozšíření Keenerovy metody

Metoda výpočtu relativní síly týmů, uvedená v předcházejících sekcích, je matematicky opodstatněná. V některých případech se ale může projevit její slabina ve smyslu zvýhodňování týmů s těžšími soupeři v rozpisu zápasů, a může vyvolat nespokojenost mezi fanoušky, kteří si myslí, že právě jejich tým byl znevýhodněn.

Pokud hrají všichni proti všem stejný počet zápasů, nebo se alespoň blíží podobnému počtu utkání, k této situaci nedojde. Týmy ale mohou být rozděleny například do několika divizí (skupin), a hrát převážně s týmy z té své. Pokud ovšem bude

nějaký tým ve slabé skupině, nemůže nikdy dosáhnout velkého množství bodů, i kdyby vyhrál prakticky všechny své zápasy, protože většina soupeřů má nízkou relativní sílu. Existuje několik způsobů, jak tento problém diskutovat.

Hlavní myšlenka našich úvah spočívá ve vyjádření relativní síly r_i každého týmu ve tvaru

$$r_i = \frac{1}{n_i} \sum_{j=1}^n f(p_{ij}r_j), \quad (5.9)$$

kde r_j je relativní síla soupeřů a p_{ij} značí kladné číslo charakterizující výsledek utkání mezi týmy i a j . O funkci f zde budeme předpokládat, že je spojitá a rostoucí na intervalu $\langle 0, \infty \rangle$. Navíc požadujeme, aby tato funkce splňovala

$$f(0) > 0, \quad f(\infty) = 1.$$

Z uvedených podmínek vyplývá, že silné týmy si nebudou moci jednoznačným vítězstvím zvyšovat tak rychle svůj bodový stav, a zároveň týmy, které mají naložované slabší soupeře, mohou stále dosáhnout na slušný celkový bodový zisk.

Vektor relativní síly $\mathbf{r}$ ale jednoduše dosazením do funkce nespočítáme, protože v argumentu funkce f jsou zastoupeny právě jeho složky. Naším úkolem je tedy opět nalézt kladný vektor relativní síly týmů, nyní ale odlišnou cestou než v předchozích částech.

Definujme nelineární funkci $\mathbf{F}$ argumentu $\mathbf{r}$, která má složky ve tvaru

$$F_i(\mathbf{r}) = \frac{1}{n_i} \sum_{j=1}^n f(p_{ij}r_j).$$

Všimněme si, že pravá strana předchozího vztahu je totožná s pravou stranou v rovnici (5.9), kterou lze tedy psát ve tvaru

$$r_i = F_i(\mathbf{r}),$$

nebo-li vektorově

$$\mathbf{r} = \mathbf{F}(\mathbf{r}).$$

O nelineární funkci $\mathbf{F}$ můžeme konstatovat, že se jedná o zobrazení z R_+^n do R_+^n . Zobrazení $\mathbf{F}$ je také omezené, neboť sčítáme konečný počet čísel, jejichž hodnota je mezi nulou a jedničkou, a následně součet jen vydělíme počtem zápasů. Pokud navíc budeme požadovat, aby funkce f splňovala, pro všechna nezáporná x , vztah

$$f(tx) > tf(x) \quad \text{pro všechna } t \in (0, 1),$$

pak v R_+^n existuje jediný pevný bod tohoto zobrazení $\mathbf{F}$, který získáme pomocí vztahu

$$\lim_{n \rightarrow \infty} \mathbf{F}^n(\mathbf{r}_0) = \mathbf{r}$$

s libovolným kladným startovacím vektorem $\mathbf{r}_0$. Symbolem $\mathbf{F}^n$ přitom rozumíme n -tou iteraci zobrazení $\mathbf{F}$ (podrobnější komentáře k této proceduře lze nalézt v [1]).

Jinou interpretací výše popsaného postupu je, že řešíme nelineární problém vlastních čísel, kde se opět snažíme najít vlastní vektor $\mathbf{r}$ s kladnými složkami.

Celý tento přístup může tedy být chápán jako *nelineární zobecnění Perronovy-Frobeniový věty*.

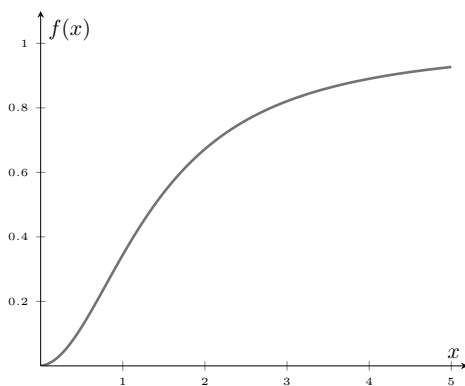
Po řadě experimentů (viz [1]) byla vybrána funkce

$$f(x) = \frac{0.05x + x^2}{2 + 0.05x + x^2}$$

a tvar zohledňující výsledek zápasu

$$p_{ij} = \frac{5 + G_{ij} + G_{ij}^{2/3}}{5 + G_{ji} + G_{ji}^{2/3}}.$$

Graf funkce f je zobrazen v obrázku 5.5. Všimněme si, že pokud počet vstřelených gólů obou týmů je stejný, tak $p_{ij} = 1$ a každý tým opět získá půl bodu. Jistým



Obrázek 5.5. Graf funkce $f(x)$.

paradoxem této volby je, že zvolená funkce f nesplňuje některé z výše uvedených podmínek (resp. tyto podmínky „téměř“ splňuje). Ukazuje se však, že metoda postupných aproximací i přesto konverguje k hledanému vektoru $\mathbf{r}$.

6. APLIKACE

V této kapitole si ukážeme použití rankingových a bodovacích systémů uvedených v sekci 5.2, a to na skutečných sportovních soutěžích týmů. Získané výsledky poté porovnáme s oficiálním pořadím. Nejprve se zaměříme na britskou fotbalovou Premier League, následně na zámořskou NHL, která je mnohými považována za nejlepší hokejovou ligu světa. Kapitulu zakončíme aplikací rankingových metod na ligu amerického fotbalu NFL.

Výběr jednotlivých soutěží, ani jejich uspořádání, není náhodný. Zvolili jsme je tak, aby nám ilustrovaly výsledky rankingových metod na různých herních systémech. Postupně si projdeme soutěže od téměř „ideálního“ rozlosování (ve smyslu, že se týmy mezi sebou utkají rovnoměrně), až po situaci, kdy se v základní části některé týmy nestřetnou vůbec.

Premier League, konkrétně sezónu 2015/2016, jsme vybrali, kromě toho, že týmy proti sobě odehrají stejný počet zápasů, také z důvodu, že zmíněný ročník vyhrál tým s jednou s nejmenších šancí na úspěch (podle sázkových kanceláří 5000-1). Proto bude zajímavé ověřit, jestli by tentýž tým vyhrál i v případě, že by k určení pořadí na konci ročníku byly použity námi zmíněné rankingové metody.

Hokejovou soutěž NHL jsme zvolili primárně ze dvou důvodů. Jedním z nich je ten, že se všechny týmy spolu utkají, ale neodehrají proti sobě stejný počet utkání. Druhým důvodem je vysoký počet celkově odehraných zápasů, který by měl zaručit, že výsledné seřazení se více blíží skutečné výkonnosti jednotlivých týmů.

V lize amerického fotbalu NFL se všechny týmy během základní části spolu neutkají, každý tým se utká přibližně pouze s polovinou. Jednotlivé celky odehrají pouze sedmnáct zápasů, bude tedy napínavé sledovat, jak významně se pořadí změní.

6.1. Premier League 2015/2016

Premier League se účastní dvacet týmů z různých koutů Anglie, případně Walesu. Sezóna začíná v srpnu a každý tým hraje vždy proti ostatním týmům jeden zápas na domácím hřišti a jeden u protivníka. V květnu na konci sezóny má tedy každý tým na svém kontě třicet osm odehraných zápasů.

Zápas může pro každý z týmů dopadnout pouze třemi způsoby – výhrou, remízou a prohrou. Za vítězství celek obdrží tři body, za prohru žádný. V případě remízy si oba týmy odnesou po jednom bodě. Průběžné i závěrečné pořadí je pak sestaveno na základě zmíněných bodových zisků. Pokud nastane shoda, je rozhodujícím faktorem např. rozdíl vstřelených a obdržených branek, nebo výsledek vzájemných zápasů.

Vítězem celé soutěže se stává tým s největším celkovým bodovým ziskem. Takto sestavené pořadí rozhoduje i o dalších věcech. Poslední tři celky, tzn. na osmnáctém až dvacátém místě, opouští ligu minimálně na jeden rok a sestupují do druhé nejvyšší ligy s názvem Championship. Naopak první čtyři týmy se kvalifikují do prestižní Ligy mistrů a určitý počet týmů od pátého místa níže pak do Evropské, nebo Konferenční ligy.

6.1.1. Absolutní senzace. Sezóna Premier League 2015/2016 nabídla neuvěřitelnou podívanou a uchvátila celý fotbalový svět. Malý klub Leicester City, který se v přechodném ročníku jen stěží v soutěži zachránil, opanoval celou ligu naprosto suverénním způsobem, a zajistil si tak místo mezi největšími senzacemi v historii fotbalu i sportu obecně.

Bylo by tedy zajímavé ověřit, jestli tým z Leicestru porážel i velmi silné protivníky, nebo jestli jen využil klopýtnutí největších favoritů s celky ze spodní části tabulky. To se nám částečně podaří díky tomu, že v systémech hodnocení zohledňujeme i sílu jednotlivých týmů.

6.1.2. Sestavení pořadí. Pro sestavení pořadí použijeme Keenerovu metodu. Nejprve s bodovými zisky, které jsou užívány v této soutěži a výše jsme je uvedli.

Příslušnou hodnotící matici označíme A . Následně použijeme spojitě, nelineární bodové ohodnocení jednotlivých zápasů, které je dáno vztahy (5.7) a (5.8), a jehož základem jsou vstřelené góly. Odpovídající hodnotící matici označíme B . Pro názornost si uvedme, jak vypadá matice A . Druhou matici uvádět nebudeme, protože obsahuje některá čísla s velmi dlouhým desetinným zápisem, a matice by byla tedy velmi nepřehledná. Hodnotící matice A je tvaru

$$A = \begin{pmatrix} 0 & 6 & 6 & 0 & 4 & 6 & 6 & 2 & 4 & 3 & 6 & 4 & 1 & 4 & 4 & 3 & 2 & 6 & 3 & 1 \\ 0 & 0 & 3 & 0 & 3 & 0 & 1 & 0 & 1 & 0 & 2 & 3 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 3 & 0 & 3 & 4 & 1 & 2 & 0 & 0 & 3 & 3 & 3 & 3 & 0 & 4 & 4 & 0 & 2 & 4 & 3 \\ 6 & 6 & 3 & 0 & 3 & 1 & 1 & 1 & 0 & 2 & 4 & 6 & 3 & 1 & 3 & 1 & 2 & 2 & 4 & 1 \\ 1 & 3 & 1 & 3 & 0 & 2 & 0 & 3 & 0 & 1 & 3 & 6 & 3 & 6 & 1 & 2 & 0 & 3 & 3 & 1 \\ 0 & 6 & 4 & 4 & 2 & 0 & 0 & 1 & 1 & 0 & 6 & 4 & 4 & 3 & 3 & 1 & 2 & 2 & 3 & 1 \\ 0 & 4 & 2 & 4 & 6 & 6 & 0 & 3 & 4 & 2 & 6 & 6 & 4 & 4 & 6 & 6 & 4 & 6 & 4 & 4 \\ 2 & 6 & 6 & 4 & 3 & 4 & 3 & 0 & 6 & 0 & 1 & 4 & 1 & 6 & 4 & 3 & 2 & 3 & 2 & 0 \\ 1 & 4 & 6 & 6 & 6 & 4 & 1 & 0 & 0 & 1 & 4 & 4 & 3 & 3 & 6 & 4 & 0 & 6 & 6 & 1 \\ 3 & 6 & 3 & 2 & 4 & 6 & 2 & 6 & 4 & 0 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 6 & 3 & 1 \\ 0 & 2 & 3 & 1 & 3 & 0 & 0 & 4 & 1 & 2 & 0 & 3 & 1 & 1 & 1 & 3 & 6 & 0 & 3 & 3 \\ 1 & 3 & 3 & 0 & 0 & 1 & 0 & 1 & 1 & 3 & 3 & 0 & 3 & 1 & 3 & 3 & 0 & 3 & 3 & 2 \\ 4 & 4 & 3 & 3 & 3 & 1 & 1 & 4 & 3 & 3 & 4 & 3 & 0 & 3 & 4 & 6 & 3 & 4 & 4 & 3 \\ 1 & 6 & 6 & 4 & 0 & 3 & 1 & 0 & 3 & 3 & 4 & 4 & 3 & 0 & 1 & 4 & 1 & 3 & 0 & 4 \\ 1 & 4 & 1 & 3 & 4 & 3 & 0 & 1 & 0 & 3 & 4 & 3 & 1 & 4 & 0 & 4 & 0 & 1 & 1 & 1 \\ 3 & 6 & 1 & 4 & 2 & 4 & 0 & 3 & 1 & 3 & 3 & 3 & 0 & 1 & 1 & 0 & 1 & 3 & 4 & 4 \\ 2 & 6 & 6 & 2 & 6 & 2 & 1 & 2 & 6 & 3 & 0 & 6 & 3 & 4 & 6 & 4 & 0 & 6 & 2 & 3 \\ 0 & 6 & 2 & 2 & 3 & 2 & 0 & 3 & 0 & 0 & 6 & 3 & 1 & 3 & 4 & 3 & 0 & 0 & 4 & 3 \\ 3 & 4 & 1 & 1 & 3 & 3 & 1 & 2 & 0 & 3 & 3 & 3 & 1 & 6 & 4 & 1 & 2 & 1 & 0 & 1 \\ 4 & 4 & 3 & 4 & 4 & 4 & 1 & 6 & 4 & 4 & 3 & 2 & 3 & 1 & 4 & 1 & 3 & 3 & 4 & 0 \end{pmatrix}.$$

Obě hodnotící matice jsou ireducibilní, takže podle Perronovy-Frobeniovy věty existuje pro matice Perronovův vektor. Výsledné vektory, vypočtené normalizovanou mocninovou metodou v Matlabu při zvolené přesnosti $\epsilon = 10^{-6}$, jsou reprezentovány v tabulce 6.1.

Z tabulky můžeme vidět, že by Leicester City vyhrál i v případě, kdybychom pro určení pořadí použili tyto rankingové systémy. Z hodnoty Perronova vektoru při použití turnajové matice A plyne, že Leicester porázel opravdu i silné týmy a titul si zaslouží. Z použití matice B plyne zase to, že nevyhrával většinou těsně o jeden gól, ale větším brankovým rozdílem.

Do Ligy mistrů by se vždy kvalifikovaly týmy Leicester, Arsenal a Tottenham. Jednou by byly doplněny Manchesterem United, podruhé West Hamem United. Ligu by pokaždé opustily dva stejné týmy - Norwich a Aston Villa. Při užití metody se standardním ohodnocením zápasů by se k nim připojil Sunderland a sestupu by se tudíž vyhnul Newcastle. Naopak pokud aplikujeme nelineární ohodnocení, Newcastle by už sestupu neunikl. Z toho můžeme dedukovat, že porázel silnější týmy než jeho konkurenti v boji o záchranu, ale malým gólovým rozdílem.

NÁZEV TÝMU	POŘ. STŽ.	KEE OB	RA	KEE NEL	RB
Leicester City	1	1	0,3366	1	0,2867
Arsenal	2	2	0,2969	3	0,2695
Tottenham Hotspur	3	3	0,2878	2	0,2701
Manchester City	4	7	0,2563	6	0,2537
Manchester United	5	4	0,2805	5	0,2539
Southampton	6	6	0,2739	7	0,2486
West Ham United	7	5	0,2754	4	0,2543
Liverpool	8	8	0,2474	8	0,2397
Stoke City	9	9	0,2126	11	0,2089
Chelsea	10	10	0,2068	9	0,2240
Everton	11	13	0,1812	10	0,2145
Swansea City	12	11	0,1947	12	0,2084
Watford	13	17	0,1686	14	0,1981
West Bromwich Albion	14	12	0,1831	13	0,2028
Crystal Palace	15	15	0,1708	15	0,1947
Bournemouth	16	14	0,1759	17	0,1900
Sunderland	17	18	0,1558	16	0,1904
Newcastle United	18	16	0,1692	18	0,1887
Norwich City	19	19	0,1438	19	0,1743
Aston Villa	20	20	0,0700	20	0,1403

Tabulka 6.1. Porovnání oficiálních výsledků s Keenerovou metodou. Významy jednotlivých zkratk: KEENER OB - pořadí určené Keenerovou metodou při ponechání oficiálního bodování; RA - složky Perronova vektoru patřící k KEENER OB; KEENER NEL - pořadí určené při nelineárním rozdělení bodů; RB - složky Perronova vektoru patřící k KEENER NEL.

Všimněme si, že mezi oficiálním pořadím a dvěma dalšími, které jsme získali Keenerovou metodou, nejsou velké rozdíly. Proto můžeme konstatovat, že na turnaje, kde týmy spolu navzájem hrají stejný počet zápasů, není nutné aplikovat metody uvedené v tomto článku.

6.2. NHL 2021/2022

Hokejové ligy se od roku 2021 účastní třicet dva týmů z USA a Kanady, které jsou rozděleny do Západní a Východní konference. Každá je ještě poté rozdělena do dvou divizí po osmi týmech, viz tabulku 6.2.

Rozdělení týmů NHL			
Západní konference		Východní konference	
Pacifická divize	Centrální divize	Atlantická divize	Metropolitní divize
Anaheim Ducks	Arizona Coyotes	Boston Bruins	Carolina Hurricanes
Calgary Flames	Chicago Blackhawks	Buffalo Sabres	Clubs Blue Jackets
Edmonton Oilers	Colorado Avalanche	Detroit Red Wings	New Jersey Devils
Los Angeles Kings	Dallas Stars	Florida Panthers	New York Islanders
San Jose Sharks	Minnesota Wild	Montreal Canadiens	New York Rangers
Seattle Kraken	Nashville Predators	Ottawa Senators	Philadelphia Flyers
Vancouver Canucks	St. Louis Blues	Tampa Bay Lightning	Pittsburgh Penguins
Vegas Gold. Knights	Winnipeg Jets	Toronto Maple Leafs	Washington Capitals

Tabulka 6.2. Rozdělení týmů NHL.

Týmy v rámci jedné divize proti sobě hrají během základní části třikrát až čtyřikrát (celkem dvacet šest zápasů). Proti soupeřům z druhé divize stejné konference se utkají třikrát (24 zápasů). S protivníky z opačné konference se potkají dvakrát. Celkově odehraje každý tým během základní části osmdesát dva zápasů. Dohromady je všech zápasů v základní části 1312.

Oficiální bodování výsledků utkání je jednoduché, a týmy si mohou odnést od nuly do tří bodů. V případě výhry v základní hrací době tým získává tři body. Pokud zápas dospěje do prodloužení, nebo dokonce do samostatných nájezdů, a tým v nich uspěje, odnese si dva body. Pokud ale prodloužení nezvládne, připíše si pouze jeden bod. Za prohru v základní hrací době se týmu nepřipíše nic.

Na konci základní části se podle počtu bodů sestavuje více pořadí – celkové, v rámci konferencí a divizí. Tým s největším počtem bodů z celé soutěže se raduje ze zisku Prezidentského poháru. Možná důležitější pořadí je to, které určuje, kdo se kvalifikuje do play-off a zahraje si o Stanley Cup. Do vyřazovací fáze postupují z každé divize tři nejlepší týmy. Následně ze zbývajících celků z každé konference postupují ještě dva týmy s největším počtem bodů na tzv. divokou kartu.

Zjistíme, jak se změní účastníci play-off při použití Keenerovy metody, pokud zachováme bodové ohodnocení výsledků jednotlivých zápasů. Například jestli Colorado Avalanche, které v tomto ročníku Stanley Cup vyhrálo, se do play-off kvalifikuje.

6.2.1. Sestavení pořadí. Hodnotící matici řádu $n = 32$ zde uvádět nebudeme, pouze si ověříme, že předpoklady Perronovy-Frobeniový věty platí. Hodnotící matice, z definice splňující nezápornost, je ireducibilní. Perronův vektor síly týmů tedy existuje. Pro zajímavost, hodnotící matice má dominantní vlastní číslo s přibližnou hodnotou $\lambda = 116,13$. Jedná se tedy dokonce o primitivní matici (viz definici 3.4). Jednotlivé složky Perronova vektoru, vypočítané znovu pomocí Matlabu při zvolené přesnosti $\epsilon = 10^{-6}$, budou uvedeny v tabulkách jednotlivých divizí.

NÁZEV	PK	RANK	PS	PB	Δ	NÁZEV	PK	RANK	PS	PB	Δ
Calgary	1	0.2154	1	111	0	Colorado	1	0.2327	1	119	0
Edmon.	2	0.2007	2	104	0	St. Louis	2	0.2152	2	109	0
Vegas	3	0.1815	4	94	-1	Minne.	3	0.2115	3	113	0
Kings	4	0.1801	3	99	1	Nashville	4	0.1864	4	97	0
Vanco.	5	0.1720	5	92	0	Dallas	5	0.1836	5	98	0
San Jose	6	0.1373	6	77	0	Winnipeg	6	0.1696	6	89	0
Anaheim	7	0.1318	7	76	0	Chicago	7	0.1114	7	68	0
Seattle	8	0.1162	8	60	0	Arizona	8	0.1041	8	57	0

Tabulka 6.3. Pořadí Pacifické divize.**Tabulka 6.4.** Pořadí Centrální divize.

NÁZEV	PK	RANK	PS	PB	Δ	NÁZEV	PK	RANK	PS	PB	Δ
Florida	1	0.2300	1	122	0	Hurric.	1	0.2315	1	116	0
Toronto	2	0.2272	2	115	0	NY Rang.	2	0.2111	2	110	0
Tampa	3	0.2065	3	110	0	Pittsb.	3	0.1940	3	103	0
Boston	4	0.1996	4	107	0	Capitals	4	0.1904	4	100	0
Ottawa	5	0.1469	7	73	-2	NY Isl.	5	0.1561	5	84	0
Buffalo	6	0.1408	5	75	1	Columbus	6	0.1473	6	81	0
Detroit	7	0.1320	6	74	1	New Jer.	7	0.1193	7	63	0
Montr.	8	0.1067	8	55	0	Flyers	8	0.1158	8	61	0

Tabulka 6.5. Pořadí Atlantické divize.**Tabulka 6.6.** Pořadí Metropolitní divize.

Významy jednotlivých zkratk a barev: NÁZEV - název týmu; PK - pořadí určené Keenerovou metodou; RANK - vypočtený ranking; tedy jednotlivé složky Perronova vektoru; PS - pořadí v soutěži; PB - počet bodů; Δ - rozdíl oproti ligu uznávanému pořadí (PK - PS), mínus značí posun směrem nahoru; světle šedá barva - postup do play-off.

Kdo čekal nějaké turbuletní změny v pořadí, musí být zklamán. Pořadí týmů v jednotlivých divizích se prakticky vůbec nezměnilo. Přesto, do boje o Stanley Cup nepůjdou stejné týmy. Z Pacifické divize se kvalifikoval tým z Las Vegas na úkor Kings z Los Angeles. Změnil se také vítěz celé základní části. Floridu Panthers vystřídal Colorado Avalanche. Otázka, jestli se Colorado vůbec dostane do play-off, je tedy zodpovězena a to jasně. Dostane.

Před vyslovením závěru, se nejprve podíváme na závěrečné pořadí celé ligy v tabulce 6.7.

Všimněme si, že pořadí se v rámci celé soutěže měnilo, a i když posuny nebyly výrazné, tak jich bylo hodně. Navíc je patrné, že tabulka se nám rozdělila do přibližně tří bloků (první, druhá a třetí desítka) a jen v rámci nich dochází k

NÁZEV TÝMU	POŘADÍ KEENER	RANKING	POŘADÍ SOUTĚŽE	POČET B.	Δ
Colorado	1	0,2327	2	119	-1
Hurricanes	2	0,2315	3	116	-1
Florida	3	0,2300	1	122	2
Toronto	4	0,2272	4	115	0
Calgary	5	0,2154	6	111	-1
St, Louis	6	0,2152	9	109	-3
Minnesota	7	0,2115	5	113	2
NY Rangers	8	0,2111	7	110	1
Tampa	9	0,2065	8	110	1
Edmonton	10	0,2007	11	104	-1
Boston	11	0,1996	10	107	1
Pittsburgh	12	0,1940	12	103	0
Capitals	13	0,1904	13	100	0
Nashville	14	0,1864	16	97	-2
Dallas	15	0,1836	15	98	0
Vegas	16	0,1815	17	94	-1
Kings	17	0,1801	14	99	3
Vancouver	18	0,1720	18	92	0
Winnipeg	19	0,1696	19	89	0
NY Islanders	20	0,1561	20	84	0
Columbus	21	0,1473	21	81	0
Ottawa	22	0,1469	26	73	-4
Buffalo	23	0,1408	24	75	-1
San Jose	24	0,1373	22	77	2
Detroit	25	0,1320	25	74	0
Anaheim	26	0,1318	23	76	3
New Jersey	27	0,1193	28	63	-1
Seattle	28	0,1162	30	60	-2
Flyers	29	0,1158	29	61	0
Chicago	30	0,1114	27	68	3
Montreal	31	0,1067	32	55	-1
Arizona	32	0,1041	31	57	1

Tabulka 6.7. Pořadí týmů NHL podle Keenerovy metody.

výměně. Tým, který si nejvíce polepšil je Ottawa, která se umístila o čtyři pozice lépe, než v původním pořadí. Znamená to, že porazila i silnější celky. Naopak týmy, které si nejvíce pohoršily, jsou Chicago, Kings a Anaheim.

Celý příklad, i když to tak na první pohled nevypadá, nám dává zajímavý výsledek. Čím více zápasů týmy odehrají, tím více se mezi nimi projeví skutečná kvalita jednotlivých mužstev. Z výsledků také plyne, že divize jsou celkem dobře rozdělené. V každé jsou velmi silné i slabé týmy. Proto „síla“ aplikovaného rankingového systému není tolik vidět.

Závěrem by se dalo konstatovat, že vliv Keenerovy rankingové metody je v případě NHL oproti Premier League větší, avšak je hodně neutralizován zmíněným vysokým počtem odehraných zápasů.

6.3. NFL 2022

Liga amerického fotbalu NFL je složena z třiceti dvou mužstev výhradně z USA. Formát ligy je velice podobný NHL, kde se po základní části hraje play-off. Týmy jsou rozděleny do dvou konferencí, které se ještě dělí na celkem osm divizí.

Výsledky zápasů jsou bodovány jednoduše. Za výhru si tým připsá jeden bod, za remízu půl bodu a za prohru žádný. Při sestavení závěrečného pořadí se body pouze sečtou a vydělí počtem zápasu. Připomeňme, že takovéto hodnocení jsme uváděli v minulé kapitole v části 5.2.2 o volbě koeficientů hodnotící matice. Pokud bychom tento tvar vynásobili sto procenty, dostali bychom již zmiňovaný „winning percentage“, nebo-li procento vyhraných zápasů.

V základní části každý tým odehraje pouze sedmnáct utkání během osmnácti týdnů. Poté jsou týmy seřazeny v příslušných divizích a z každé přímo postupuje pouze vítěz. Vítězné týmy jsou následně doplněny celky, stejně jako v případě NHL na tzv. divokou kartu, třemi nejlepšími týmy ze zbylých týmů každé konference.

6.3.1. Sestavení pořadí. V tomto případě se avšak nebudeme zabývat otázkou, kdo by se kvalifikoval do play-off přímo, nebo na divokou kartu. Zaměříme se na celkové pořadí soutěže, jeho změnu a následné srovnání při použití Keenerovy metody s nezměněným bodováním soutěže. Budeme tedy nyní taky uvažovat hodnocení, kde tým získá jeden bod, půl bodu, nebo žádný.

Nejprve musíme opět ověřit, jestli hledaný Perronův vektor, při takto zvoleném bodování zápasů, existuje. Hodnotí matice je i v tomto případě ireducibilní, a podmínky pro existenci Perronova vektoru jsou splněny. Díky tomu, že vlastní číslo (příslušné hledanému vlastnímu vektoru) je zároveň i dominantním vlastním číslem o přibližné hodnotě $\lambda = 7,3015$, můžeme použít na výpočet hledaného Perronova vektoru normalizovanou mocninovou metodu. Přesnost ponecháme nezměněnou, tedy $\epsilon = 10^{-6}$, a dostáváme tabulku 6.8.

Hned na první pohled je patrné, že došlo k velkému pohybu týmů v rámci umístění v tabulce. Je to hlavně dáno z toho důvodu, že týmy hrály jen proti některým celkům a každá výhra nebo remíza měla stejnou váhu. My jsme ale pomocí rankingového systému zohlednili i výkonnost jednotlivých protivníků. S takovýmto pořadím by určitě byly spokojeny týmy Buffalo Bills, kteří by základní část vyhráli nebo Washington Commanders, kteří se posunuli na dvanáctou příčku (v tabulce světle šedá). Výsledky však nejsou „ideální“. V sekci 5.2 jsme se zmínili o tom, že pokud bude mít tým nalosované slabší soupeře, může se díky tomu propadnout v pořadí, které bylo získáno Keenerovou metodou. Jasnou ukázkou této nevýhody je tým Kansas City Chiefs (sytle šedá), který obsadil až sedmou pozici i přesto, že v původním pořadí byl nejlepší.

Tato metoda je velmi účinný nástroj, a pokud zajistíme, že týmy budou mít podobně náročný rozpis zápasů, pak výsledky, které dostaneme, budou velmi přesné.

NÁZEV TÝMU	POŘADÍ KEENER	RANKING	POŘADÍ	WP	Δ
Buffalo Bills	1	0,3028	3	0,813	-2
Cincinnati Bengals	2	0,2881	6	0,750	-4
Philadelphia Eagles	3	0,2879	2	0,824	1
Minnesota Vikings	4	0,2643	4	0,765	0
Dallas Cowboys	5	0,2608	7	0,706	-2
Baltimore Ravens	6	0,2174	8	0,588	-2
Kansas City Chiefs	7	0,2156	1	0,824	6
San Francisco 49ers	8	0,2121	5	0,765	3
Miami Dolphins	9	0,2055	11	0,529	-2
Detroit Lions	10	0,1902	14	0,529	-4
Pittsburgh Steelers	11	0,1884	12	0,529	-1
Washington Commanders	12	0,1756	16	0,500	-4
New York Giants	13	0,1694	10	0,559	3
Green Bay Packers	14	0,1678	18	0,471	-4
New England Patriots	15	0,1654	17	0,471	-2
Jacksonville Jaguars	16	0,1651	13	0,529	3
Cleveland Browns	17	0,1639	21	0,412	-4
New York Jets	18	0,1607	20	0,412	-2
Tampa Bay Buccaneers	19	0,1438	19	0,471	0
Los Angeles Chargers	20	0,1411	9	0,588	11
Seattle Seahawks	21	0,1368	15	0,529	6
New Orleans Saints	22	0,1363	25	0,412	-3
Carolina Panthers	23	0,1306	24	0,412	-1
Atlanta Falcons	24	0,1239	23	0,412	1
Tennessee Titans	25	0,1006	22	0,412	3
Las Vegas Raiders	26	0,0916	26	0,353	0
Denver Broncos	27	0,0862	27	0,294	0
Indianapolis Colts	28	0,0801	29	0,265	-1
Los Angeles Rams	29	0,0672	28	0,294	1
Chicago Bears	30	0,0590	32	0,176	-2
Arizona Cardinals	31	0,0583	30	0,235	1
Houston Texans	32	0,0528	31	0,206	1

Tabulka 6.8. Srovnání Keenerovy metody s oficiálním pořadím soutěže NFL. Zkratka: WP - winning percentage bez vynásobení sto procenty..

V opačném případě se nám může stát, že budeme moci zpozorovat abnormality jako v tomto případě.

7. POROVNÁNÍ S VYBRANÝMI METODAMI POUŽÍVANÝMI V ŠACHU

V této kapitole porovnáme výše uvedené metody s metodami Buchholzovou a Sonnebornovou-Bergerovou, které jsou využívány k stanovení pořadí v šachových turnajích.

7.1. Buchholzova metoda

Tato metoda se používá jako pomocné hodnocení pro turnaje hrané švýcarským³ systémem. Kromě šachu se tedy také uplatňuje v petanque atd. Slouží k určení pořadí účastníků turnaje při shodě bodů a upřednostňuje hráče, kteří měli silnější protivníky.

Bodování jednotlivých zápasů je zde běžné – hráč, pokud vyhrál, získá jeden bod, pokud remizoval, přičte si půl bodu. V případě prohry si nepřipisuje nic. Celkový bodový zisk každého hráče je dán součtem získaných bodů. V případě shody se hráčům přiřadí nové skóre, které je definováno pouze jako suma celkového zisku bodů jednotlivých soupeřů. Nezáleží, jestli hráč soupeře porazil, nebo proti němu prohrál. Zde se tento způsob velmi liší od námi uvedených systémů.

Pro získání nového skóre – Buchholzova skóre – bychom museli změnit hodnotící matici, která by měla jedničku v prvcích a_{ij} a a_{ji} , pokud proti sobě účastníci i a j hráli. V opačném případě by na těchto pozicích byla nula. A to je v rozporu s našimi metodami, ve kterých používáme pouze jednu hodnotící matici.

7.2. Sonnebornova-Bergerova metoda

Více podobná námi zkoumaným metodám je Sonnebornova-Bergerova metoda. Znovu se používá pouze jako pomocná metoda při shodě bodů, nyní ale v šachových turnajích, kde každý hraje s každým právě jednou.

Všechna střetnutí bodujeme takto: výhru jedním bodem, remízu polovinou a prohru žádným. Výsledný bodový zisk každého hráče je pouhý součet bodů získaných z jednotlivých utkání a podle těchto zisků je sestaveno závěrečné pořadí. Bodový zisk b_i i -tého účastníka zapíšeme takovýmto způsobem:

$$b_i = W_i + \frac{1}{2}R_i + 0P_i,$$

kde písmeno W_i značí počet výher, R_i počet remíz a P_i počet proher i -tého účastníka.

Díky využití hodnotící matice A , kde, jak víme, prvek a_{ij} značí bodový zisk šachisty i proti j , můžeme tuto operaci pro všechny hráče zapsat maticově

$$\mathbf{b} = A\mathbf{e},$$

kde $\mathbf{e}$ je vektor jedniček.

³Hlavní myšlenka švýcarského systému je v tom, aby proti sobě v každém kole hráli hráči s co nejpodobnější úrovní a zároveň aby proti sobě nehráli vícekrát.

Pokud nastane situace, ve které má více hráčů stejný celkový bodový zisk, pokračujeme následovně. Vytvoříme nový druhý bodový zisk b_i^β , kde i -tému účastníkovi za každého protivníka, kterého porazil, přičteme soupeřův celkový počet bodů, a za každého soupeře, se kterým remizoval, polovinu jeho celkových bodů z turnaje. Za soupeře, s nimž prohrál, se mu nepřičte nic.

Díky vhodně zvolenému rozdělení bodu za zápas (jeden bod, polovina, nebo žádný) můžeme nový bodový zisk napsat maticově:

$$b^\beta = Ab = AAe = A^2e.$$

Všimněme si, že se nejedná o nic jiného než pouze o druhou iteraci Keenerovy metody, pokud za startovací vektor zvolíme e . (Také nedělíme počtem utkání – z důvodu stejného počtu zápasů to není pro druhou iteraci nutné.)

Porovnání těchto dvou metod si ukážeme na reálném příkladu a výsledky poté okomentujeme.

JMÉNO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sloth	0	0,5	0,5	1	0,5	0,5	1	1	0,5	1	0,5	1	1	1	1
Zagarovsky	0,5	0	0	0,5	1	0,5	1	1	1	0,5	1	1	1	1	1
Kosenkov	0,5	1	0	0,5	0,5	0,5	0,5	0,5	1	1	0,5	1	1	1	1
Khasin	0	0,5	0,5	0	0,5	1	0,5	0	1	1	0,5	1	0,5	1	0,5
Kletsel	0,5	0	0,5	0,5	0	0,5	0,5	0,5	0,5	0	1	1	0,5	1	1
De Carbonnel	0,5	0,5	0,5	0	0,5	0	0,5	0,5	0	1	0,5	0,5	0	1	1
Arnlin	0	0	0,5	0,5	0,5	0,5	0	0,5	1	0	0,5	0,5	1	1	0,5
Dunhaupt	0	0	0,5	1	0,5	0,5	0,5	0	0	0,5	1	0	1	0,5	1
Maedler	0,5	0	0	0	0,5	1	0	1	0	1	0,5	0,5	0,5	0,5	1
Estrin	0	0,5	0	0	1	0	1	0,5	0	0	1	1	1	0	1
Walther	0,5	0	0,5	0,5	0	0,5	0,5	0	0,5	0	0	0	1	0,5	1
Boey	0	0	0	0	0	0,5	0,5	1	0,5	0	1	0	0,5	0,5	1
Abramov	0	0	0	0,5	0,5	1	0	0	0,5	0	0	0,5	0	0,5	1
Siklos	0	0	0	0	0	0	0	0,5	0,5	1	0,5	0,5	0,5	0	1
Nun	0	0	0	0,5	0	0	0,5	0	0	0	0	0	0	0	0

Tabulka 7.1. Bodové zisky hráčů z jednotlivých utkání z následujícího šachového turnaje.

Příklad 7.1. Sestavme pořadí účastníků turnaje 8. finále mistrovství světa v korespondenčním šachu. V případě shody bodů použijme metody B-S a Keenerovu, a jejich výsledky porovnejme. Bodové zisky hráčů z jednotlivých utkání jsou zapsány v tabulce 7.1, které přepíšeme do hodnotící matice A a pořadí řádků ponecháme. Vektor s celkovým počtem bodů jednotlivých hráčů poté vyjde následovně:

$$b = Ae = (11 \quad 11 \quad 10,5 \quad 8,5 \quad 8 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 5,5 \quad 5,5 \quad 4,5 \quad 4,5 \quad 1)^T.$$

Je patrné, že několik šachistů má stejný bodový zisk, na určení jednoznačného pořadí proto použijeme pomocné metody. Nejprve si však musíme ověřit existenci Perronova vektoru. Hodnotící matice A z tohoto příkladu je ireducibilní a už z definice nezáporná, Perronův vektor síly týmů tedy existuje. Jednotlivé metody nyní už jen vypočítáme a jednotlivé složky získaných vektorů zapíšeme do tabulky 7.2.

Z výsledků je vidět, že při použití Keenerovy metody došlo oproti Sonnebornově-Bergerově metodě k prohození Arnlin a Dunhaupta. Je to z toho důvodu, že

Dunhaupt získal více bodů proti lépe postaveným hráčům v závěrečném pořadí a až při vyšší mocnině u hodnotící matice se tato skutečnost projevila.

JMÉNO	CELKEM	PT	SB RANK	PSB	KNR RANK	PK	Δ
Sloth	11	1	69.5000	1	0,3929	1	0
Zagarovsky	11	1	66.7500	2	0,3765	2	0
Kosenkov	10,5	3	67.5000	3	0,3795	3	0
Khasin	8.5	4	54.7500	4	0,3068	4	0
Kletsel	8	5	47.7500	5	0,2719	5	0
De Carbonnel	7	6	45.2500	6	0,2557	6	0
Arnlin	7	6	42.5000	7	0,2398	8	1
Dunhaupt	7	6	41.5000	8	0,2406	7	-1
Maedler	7	6	41.5000	9	0,2366	9	0
Estrin	7	6	40,5000	10	0,2282	10	0
Walther	5.5	11	33.2500	11	0,1934	11	0
Boey	5.5	11	28.5000	12	0,1638	12	0
Abramov	4.5	13	24.7500	13	0,1428	13	0
Siklos	4.5	13	22.7500	14	0,1274	14	0
Nun	1	15	7.7500	15	0,0457	15	0

Tabulka 7.2. Výsledné pořadí šachového turnaje určené dvěma metodami.

V předchozích kapitolách jsme uváděli, že čím vyšší je mocnina u hodnotící matice, tím by výsledky měli být přesnější – s rostoucí mocninou se více projeví jednotlivé závislosti mezi účastníky – a příklad z šachového turnaje nám tento fakt jen znovu potvrdil. Bylo by tedy vhodné pro dosažení přesnějších výsledků místo Sonnebornovy-Bergerovy metody zvolit Keenerovu.

8. ZÁVĚR

Cílem článku a bakalářské práce, ze které tento článek vychází, bylo vysvětlit význam Perronovy-Frobeniovovy věty a dalších důležitých teoretických výsledků v některých rankingových systémech.

Po uvedení základních pojmů z teorie matic jsme se detailněji zaměřili na zmiňovanou Perronovu-Frobeniovou větu, která udává podmínky, za kterých existuje kladný vlastní vektor příslušný kladnému vlastnímu číslu o velikosti spektrálního poloměru. Tuto problematiku jsme ilustrovali na motivačním příkladu matice druhého řádu. Nastínili jsme také numerický výpočet aproximace vlastního vektoru, příslušného dominantnímu vlastnímu číslu, a to pomocí normalizované mocninné metody.

V čtvrté kapitole jsme ukázali možný maticový zápis turnajové tabulky pomocí hodnotící a turnajové matice. Turnajovou matici jsme dále podrobně rozebrali a uvedli její zajímavé vlastnosti.

První sekce páté kapitoly byla věnována sestavení pořadí pomocí turnajové matice. Zjistili jsme, že pokud je matice ireducibilní, týmy nejsme schopni podle počtu vítězství seřadit, a tato situace vede na použití Perronovy-Frobeniovy věty. Aplikaci jsme ilustrovali na dvou jednoduchých příkladech. Z nich vyplývá, že i když nám většinou použití této věty pro určení pořadí v rámci ireducibilní matice nebo ireducibilním bloku pomůže, existují i takové situace, ve kterých bez zohlednění dalších faktorů pořadí neurčíme.

Druhá sekce kapitoly páté se zabývala hledáním vhodného rankingu, který by zohledňoval nejen výsledek utkání a schopnosti protivníka, ale také nerovnoměrný počet zápasů odehraný mezi jednotlivými soupeři. Odvodili jsme tzv. Keenerovu metodu, která vede znovu na použití Perronovy-Frobeniovy věty. Následně jsme také uvedli možné alternativní rozdělení bodů, tedy možnou volbu koeficientů hodnotící matice, a popsali jejich přednosti a nevýhody.

V poslední sekci páté kapitoly jsme naznačili myšlenku nelineárního rozšíření Keenerovy metody, a některé z toho plynoucí benefity při určování pořadí. Metoda může být chápána jako nelineární zobecnění Perronovy-Frobeniovy věty.

V šesté kapitole jsme si ukázali použití rankingových systémů, uvedených v druhé sekci kapitoly páté, na příkladech sportovních soutěží z reálného světa. Ukázalo se, že pokud týmy proti sobě hrají stejný počet zápasů, Keenerovou metodou dostaneme téměř stejné pořadí, jaké bylo to oficiální. Pokud je celkový počet utkání vysoký, vliv rankingu není tolik viditelný. U NFL se ukázala jediná velká slabina Keenerovy metody, a to v tom smyslu, že v případě, pokud má tým pouze slabé protivníky, nemůže získat vysoké ohodnocení, i kdyby téměř všechny své zápasy vyhrál.

A v závěrečné kapitole s číslem sedm jsme porovnali dvě metody možného hodnocení z šachového prostředí s námi navrženými metodami a zjistili jsme, že by místo Sonnebornovy-Bergerovy metody bylo vhodnější použít Keenerovu metodu.

Problematika byla velice zajímavá a poučná. Do budoucna by bylo zajímavé se pokusit sestavit komplexnější rozdělení bodů, které by se zohledňovalo více faktorů zároveň, např. bodový zisk z výsledků zápasu, počet vstřelených branek či střel na branku, atd. Nebo se pokusit najít jinou hodnotící funkci, která by splňovala podmínky sekce 5.3, a současně by dávala věrohodné ohodnocení týmů v rámci užitých rankingových metod.

REFERENCE

- [1] J. P. Keener: *The Perron-Frobenius Theorem and the ranking of football teams*, SIAM Rev. **35** (1993), No. 1, 80–93.
- [2] C. Eschenbach et al.: *Properties of Tournaments among Well-Matched Players*, Amer. Math. Monthly **107** (2000), No. 10, 881–892.
- [3] F. R. Gantmacher: *Applications of the theory of matrices*, Dover ed. Přeložil J. L. Brenner, Mineola: Dover Publications, 2005.
- [4] R. A. Horn, C. R. Johnson: *Matrix analysis*, New York: Cambridge University Press, 1985.

- [5] L. Čermák: *Numerické metody*, Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2020.
- [6] J. L. Gross, J. Yellen: *Graph theory and its applications*, New York: Chapman and Hall, 2006.

Lubomír Pažourek, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 616 69 Brno, Česká republika,
e-mail: `lubomir.pazourek@vutbr.cz`