UNIVERZITET U BANJOJ LUCI
PRIRODNO-MATEMATIČKI FAKULTET

BOJAN NIKOLIĆ

MJERE SLIČNOSTI NAD FAMILIJAMA STRINGOVA

DOKTORSKA DISERTACIJA

BANJA LUKA, 2024.

UNIVERZITET U BANJOJ LUCI
PRIRODNO-MATEMATIČKI FAKULTET

BOJAN NIKOLIĆ

MJERE SLIČNOSTI NAD FAMILIJAMA STRINGOVA

DOKTORSKA DISERTACIJA

BANJA LUKA, 2024.

UNIVERSITY OF BANJA LUKA
FACULTY OF NATURAL SCIENCES AND
MATHEMATICS

BOJAN NIKOLIĆ

SIMILARITY MEASURES OF FAMILIES OF STRINGS

DOCTORAL DISSERTATION

BANJA LUKA, 2024.

Mentor: dr Boris Šobot, vanredni profesor na Prirodno-matematičkom fakultetu Univerziteta u Novom Sadu

Naslov: Mjere sličnosti nad familijama stringova

Rezime

Objekti diskretnog sekvencijalnog tipa se matematički mogu modelirati kao stringovi (riječi). U nekim slučajevima, neophodna je višedimenzionalna analiza takvih podataka koja bi omogućila poređenje višečlanih familija stringova posmatranih kao cjeline. Stoga je od koristi da se definišu mjere sličnosti koje bi u što većoj mjeri odražavale zajedničke relevantne osobine posmatranih familija stringova. Činjenica da familije stringova mogu sadržavati veliki broj stringova koji dodatno mogu biti velike dužine i definisani nad alfabetom velike kardinalnosti, sugeriše da je problem znatno složeniji u odnosu na slučaj poređenja pojedinačnih stringova. Zato je prirodno uključiti u istraživanje razne matematičke oblasti koje imaju razvijenu aparaturu za izučavanje složenih struktura podataka. U ovoj disertaciji razmatrane su metode algebarske topologije, vjerovatnoće i optimizacije. Od metoda algebarske topologije razmatrana je istrajna homologija; od vjerovatnosnih metoda korišćeni su vjerovatnosno sufiksno drvo i hijerarhijski Pitman-Jorov proces, dok je od optimizacionih metoda razmatrana pretraga bima, u cilju približnog rješavanja problema nalaženja najdužeg zajedničkog podniza (LCS problema).

Ključni pojmovi: Mjere sličnosti familija stringova, istrajna homologija, tehnika razdvajanja radijusa simpleksa, vjerovatnosno sufiksno drvo, Pitman -Jorov hijerarhijski proces, LCS problem, pretraga bima

Naučna oblast: Prirodne nauke

Naučno polje: Matematika

Klasifikacioni kod prema CERIF-u šifrarniku: P 150, P 160

Supervisor: Boris Šobot, Ph.D., associate professor at Faculty of Sciences, University of Novi Sad

Title: Similarity measures of families of strings

Abstract

Discrete and sequential ordered data types can be efficiently mathematically modeled as strings (words). In some cases, a multidimensional analysis of such data is necessary to capture relevant structural based features of a multi-member string family. Therefore, it is useful to define similarity measures that would enable the comparison of these structural invariants, in order to investigate their significance and overall impact. The fact that string families can contain a large number of strings of an arbitrary (finite) lengths, together with the possibility that elements of observed strings are coming from an alphabet of high cardinality, suggests that the problem is far more complex compared to the case of individual strings. This problem can be tackled by using various mathematical fields with highly developed methods for the study of complex data structures. In this dissertation, methods of algebraic topology, probability, and optimization are considered. More precisely, string similarity measures are built using the means of the persistent homology (from the algebraic topology), and probabilistic methods of the probabilistic suffix tree and the hierarchical Pitman-Yorov process. Also, the beam search, as the prominent optimization method which efficiently solves the problem of finding the longest common subsequence (LCS problem), is elaborated.

Keywords: Similarity measures of families of strings, persistent homology, separation of simplex radii technique, probability suffix tree, Pitman-Yor hierarchical process, LCS problem, beam search

Scientific area: Natural Sciences

Scientific field: Mathematics

CERIF classification code: P 150, P 160

Zahvalnica

Iskoristio bih ovu priliku da se zahvalim pojedincima koji su svojim zalaganjem značajno uticali na moj lični i profesionalni razvoj i bez čije pomoći vjerovatno ne bih uspio završiti ovu doktorsku disertaciju.

Najprije zahvaljujem svojim najbližima: majci Vesni, ocu Milisavu i supruzi Živki. Njihova bezrezervna ljubav i podrška omogućila mi je da prebrodim sve poteškoće koje sam imao tokom svojih doktorskih studija.

Dalje, zahvaljujem odgovornim osobama iz uprave Prirodno-matematičkog fakulteta i rektorata Univerziteta u Banjoj Luci na pokazanom razumijevanju. Tu prije svega mislim na dekana Prirodno-matematičkog fakulteta prof.dr Gorana Trbića, koji mi je obezbijedio vrijeme potrebno za završetak ovog rada i doktorskih studija matematike.

Naposljetku, zahvaljujem kolegama sa Studijskog programa za matematiku i informatiku sa kojima sam sarađivao tokom svojih doktorskih studija. Prvenstveno, zahvaljujem prof.dr Draganu Matiću, koji predstavlja uzor kako profesor Univerziteta treba da uvodi mlađe kolege u kvalitetan naučno-istraživački rad. Takođe, zahvaljujem dr Marku Đukanoviću, čovjeku nepresušne stvaralačke energije koji je uvijek spreman za saradnju, naročito u pogledu razmjene ideja i unaprijeđenja istih. Na kraju, posebno zahvaljujem svom mentoru prof.dr Borisu Šobotu, na strpljivom i studioznom radu koji je rezultovao rukopisom kojeg predstavljam i čije su korisne sugestije nesumnjivo podigle njegov kvalitet.

Bojan Nikolić

Sadržaj

Glava 1

Uvod

Većina prirodnih, a i određen broj društvenih nauka ima oblasti u okviru kojih postoji potreba za proučavanjem objekata sastavljenih od gradivnih elemenata poređanih u tačno određenom redoslijedu. Analiza podataka sekvencijalnog tipa dovodi do boljeg razumijevanja i shvatanja svojstava posmatranih objekata i postuliranja zakonitosti koju oni ispoljavaju. Podaci sekvencijalnog tipa se matematički modeliraju posmatranjem stringova - konačnih nizova elemenata izabranih iz istog alfabeta. Analiza podataka ovakvog tipa bazira se na proučavanju ispoljenog poretka i može da se sprovede na više različitih nivoa: Jednodimenzionalna analiza podrazumijeva proučavanje pojedinačnih stringova, dok višedimenzionalna analiza podrazumijeva proučavanje višečlanih familija stringova, gdje je akcenat na utvrđivanju međusobne povezanosti stringova ove familije kao jedne cjeline. Višedimenzionalna analiza daje više informacija o prirodi posmatrane pojave, ali njeno sprovođenje postaje komplikovanije sa povećanjem kardinalnosti posmatrane familije stringova. Stoga je korisno uočiti referentne vrijednosti kojima se opisuje strukturalni integritet familije stringova i, uz pomoć njih, definisati mjere sličnosti kojima se izražava stepen sličnosti dvije ili više familija stringova. Izbor ovog repera zavisi od toga koja osobina se smatra esencijalnim pokazateljem povezanosti unutar jedne familije stringova.

Cilj istraživanja ove doktorske teze je ispitivanje mjera sličnosti familija stringova. Okvir ovog ispitivanja određen je matematičkim modelima koji imaju osnovu u različitim matematičkim oblastima - topologiji i vjerovatnoći. U tom smislu, u tezi je napravljen pregled poznatih rezultata iz literature u vezi poređenja familija stringova. Pored toga, razvijene su nove tehnike i dokazani određeni rezultati koji omogućavaju ne samo uvođenje novih mjera sličnosti familija stringova, već i dodatno osiguravaju da ove mjere posjeduju svojstvo stabilnosti. Na taj način, ova teza predstavlja originalan doprinos posmatranoj temi.

Nastavak ove doktorske teze je organizovan u 4 glave: Topološki metodi, Vjerovatnosni metodi, Primjena pri rješavanju LCS problema i Zaključak.

U glavi "Topološki metodi" razmotrene su mjere sličnosti familija stringova zasnovane na algebarskoj topologiji. Konkretnije, u pomenutom razmatranju koriste se sredstva i tehnike iz perzistentne (istrajne) homologije, koja predstavlja "dinamičku" varijantu simplicijalne homologije. Ključna ideja je da se posmatranoj familiji stringova pridruži filtracija, kao konačan niz simplicijalnih simpleksa koji će, u odnosu na zadanu udaljenost između stringova, sadržavati kompletnu informaciju o tome koliko su stringovi iz date familije "bliski" jedni drugima i na koji način skaliranje rastojanja dovodi do promjena u ovom obliku njihove povezanosti. Odsustvo postojanja povezanosti između nekog podskupa date familije stringova na nekom nivou filtracije slikovito se može predočiti postojanjem "rupe" odgovarajuće dimenzije. Istrajna homologija je matematički model koji prati evoluciju ovih rupa, za svaku moguću dimenziju. To se postiže tako što se, za svaku dimenziju, datoj filtraciji pridružuje istrajni modul, kao lanac homoloških grupa povezanih homomorfizmima indukovanim inkluzijama. Zatim se prati istrajnost svake postojeće rupe, kao period njenog "života" u datom istrajnom modulu, tj. raspon od nivoa filtracije u kojem nastaje homološka klasa koja predstavlja ovu rupu do nivoa filtracije u kojem ova homološka klasa nestaje. Na ovaj način se, za svaku dimenziju, posmatranoj familiji stringova može pridružiti multiskup čiji su elementi intervali istrajnosti svake rupe te dimenzije. Ovaj multiskup predstavlja bar-kod odgovarajuće dimenzije koji je pridružen posmatranoj familiji stringova. Bar-kod i istrajni modul iz koga je on proistekao su u obostrano jednoznačnoj korespondenciji. Stoga, bar-kod familije stringova može da se iskoristi kao karakteristika uz pomoć koje se mogu izgraditi mjere sličnosti familije stringova. Najvažnija mjera sličnosti ovakvog tipa razmatrana u literaturi je udaljenost uskog grla. Ova mjera se zasniva na uparivanju linija dva bar-koda iste dimenzije na način da se minimizuje potrebno "pomjeranje" linija kojim se postiže da se sve uparene linije "značajne dužine" što bolje preklope. Udaljenost uskog grla ima svojstvo stabilnosti, u smislu da male promjene strukture istrajnog modula dovode do promjena iste veličine u strukturi linija pripadnog bar-koda. Uparivanje pomoću kojeg je definisana ova udaljenost ima tendenciju uparivanja linija dva bar-koda koje imaju relativno sličan položaj u okviru odgovarajućih istrajnih modula. Nažalost, ovako uparene linije ne moraju da predstavljaju slične homološke atribute, jer se može desiti da su istrajni moduli iz kojih oni proističu strukturalno različiti. Jedan od načina da se ukloni ovaj nedostatak je posmatranje istrajnog modula u koji je moguće "potopiti" oba polazna istrajna modula i na taj način omogućiti da se njihove strukture ugrade u strukturu ovog "krovnog" istrajnog modula. Posljednje sekcije ove glave posvećene su opisu novog "hibridnog" uparivanja linija bar-koda kod kojeg se prioritet daje uparivanju onih linija dva bar-koda koje predstavljaju istu homološku karakteristiku krovnog istrajnog modula. Takođe, u ovom dijelu opisana je nova tehnika razdvajanja radijusa simpleksa, kao ključno sredstvo koje omogućava da pomenuto hibridno uparivanje ima smisla.

U glavi "Vjerovatnosni metodi" razmotrene su mjere sličnosti familija stringova zasnovane na nedeterminističkim modelima. Preciznije, string se modeluje kao diskretan stohastički proces u vidu lanca Markova (konačan niz slučajnih promjenljivih), pri čemu simboli alfabeta od kojih se string izgrađuje predstavljaju stanja ovog procesa, a svaki konkretan string predstavlja jednu njegovu trajektoriju. Ključna ideja je da se svakoj familiji stringova pridruži odgovarajuća vjerovatnosna mjera, a da se zatim mjera sličnosti dvije familije stringova izrazi poređenjem sličnosti njihovih vjerovatnosnih mjera. U tezi je izabrano poređenje čiji je temelj Kulbak-Lajblerova mjera divergencije ili relativna entropija, pri čemu se konkretno koristi očekivana vrijednost logaritamske razlike dvije posmatrane vjerovatnosne mjere. Vjerovatnosna mjera pridružena familiji stringova proističe iz klase dopustivih raspodjela trajektorija lanca Markova i ona generalno nije poznata. Stoga je potrebno izvršiti statističko modelovanje ove mjere, na osnovu stringova iz date familije kao trening podacima. Jednostavniji slučaj ovog modelovanja nastupa ako se pretpostavi da je string sastavljen od nezavisnih slučajnih veličina, jer je tada potrebno samo ocijeniti raspodjelu vjerovatnoća simbola alfabeta. Zahtjevniji (i mnogo češći) slučaj podrazumijeva zavisnost u nizanju simbola stringa. Tada je potrebno ocijeniti sve uslovne vjerovatnoće zaključno do reda lanca Markova, što znači da se, u odnosu na posmatranu poziciju u stringu, u obzir uzimaju inicijalizacije na svim prethodnim pozicijama stringa za koje još uvijek ne vrijedi "odsustvo memorije" datog lanca Markova. Za statističko ocjenjivanje ovih uslovnih vjerovatnoća mogu da se koriste dva pristupa: frekvencionistički pristup i Bejzovski pristup. Frekvencionistički pristup podrazumijeva da se posmatrane uslovne vjerovatnoće ocijene na osnovu frekvencija učestalosti pojavljivanja odgovarajućih podstringova u okviru datih trening podataka. Bejzovski pristup zasniva se na pretpostavci da posmatrane uslovne vjerovatnoće nisu parametri, već slučajne promjenljive čije raspodjele treba ocijeniti. U tezi je predložen po jedan model za svaki od dva pomenuta pristupa, pri čemu je razmotren model vjerovatnosnog sufiksnog drveta kao primjer frekvencionističkog pristupa, dok je kao primjer Bejzovskog pristupa razmotren model hijerarhijskog Pitman-Jorovog procesa. U oba tipa modela, pri ocjenjivanju uslovnih vjerovatnoća mogu se koristiti uzorci dobijeni Gibsovim metodom uzorkovanja (tzv. Gibsov sempler), koji predstavlja najjednostavniji algoritam u okviru Markov Chain Monte Carlo (MCMC) metoda. Nakon ocjenjivanja uslovnih vjerovatnoća uz pomoć nekog od ova dva modela, koristi se formula množenja vjerovatnoća da se generiše "predviđajuća" vjerovatnoća ostvarivanja proizvoljnog stringa. Ova vjerovatnoća je finalni produkt opisanog statističkog modelovanja, jer predstavlja ocjenu vjerovatnosne mjere pridružene posmatranoj familiji stringova.

U glavi "Primjena pri rješavanju LCS problema" su implementirani određeni vjerovatnosni modeli u postupku rješavanja LCS problema. LCS problem odnosi se na nalaženje najdužeg zajedničkog podniza (engleski Longest Common

Subsequence) date familije stringova definisanih nad istim alfabetom. Dužina najdužeg zajedničkog podniza predstavlja jedan oblik mjere koja je našla primjenu u raznim disciplinama npr. u bioinformatici, molekularnoj biologiji, kriptografiji, itd. Većina približnih računarskih metoda koji se koriste za rješavanje LCS problema pretpostavlja da se stringovi iz posmatrane familije biraju nezavisno i na slučajan način, kao i da se elementi svakog pojedinačnog stringa biraju nezavisno u skladu sa zadatom distribucijom vjerovatnoća simbola alfabeta. U toj postavci, raspodjela vjerovatnoća simbola alfabeta je polinomijalna raspodjela, čiji je najjednostavniji oblik uniformna raspodjela najčešće razmatrana u literaturi. U tezi je predstavljeno rješenje LCS problema uz pomoć pretrage bima (Beam Search ili BS) uz korišćenje specijalno dizajnirane novouvedene heuristike koja usmjerava pretragu ka perspektivnijim rješenjima. U kontekstu efikasnosti, ovakav pristup je u rangu postojećih tehnika iz literature, kako u slučaju uniformno raspoređenih instanci, tako i u slučaju instanci sa neuniformnim polinomijalnim distribucijama. Pomenuta heuristika, pod nazivom GMPSUM dobija se kao konveksna kombinacija dvije statistike: GM vrijednosti, koja se zasniva na geometrijskoj sredini i geometrijskoj standardnoj devijaciji frekvencija pojavljivanja simbola alfabeta u okviru ulaznih stringova odgovarajućih potproblema i PSUM vrijednosti, koja se bazira na prethodno uvedenoj matrici vjerovatnoća da, za dva slučajno izabrana i nezavisna stringa odgovarajućih dužina, prvi string bude podniz drugog stringa. Parametar koji se pojavljuje u konveksnoj kombinaciji ove dvije statistike reguliše njihovu zastupljenost u okviru GMPSUM heuristike i bira se u zavisnosti od usaglašenosti raspodjela stringova u ulaznim instancama. U tezi su detaljno opisani eksperimenti u kojima se za skupove testnih instanci vrši poređenje efikasnosti varijante BS vođene GMPSUM heuristikom sa ostalim najprestižnijim algoritmima iz literature. Rezultati ovih eksperimenata daju za pravo da se predložena varijanta BS – GMPSUM smatra novim najprestižnijim algoritmom ovog tipa za rješavanje LCS problema.

U posljednjoj glavi dat je zaključak i nagovješteni su mogući pravci daljnjeg istraživanja mjera sličnosti stringova.

Na početku svake glave naveden je spisak referenci iz literature koje se u toj glavi koriste. Za svako značajnije tvrđenje citiran je izvor iz kojeg ono potiče. To ne znači da su iz tih izvora automatski preuzeti i dokazi, već su oni uglavnom originalno rekonstruisani od strane autora ove teze. Pored toga, naglašeni su i dijelovi svake glave koji predstavljaju originalni doprinos posmatranoj temi.

Glava 2

Topološki metodi

U ovoj glavi razmatrane su mjere sličnosti familija stringova sa topološkog aspekta. U početnim sekcijama ove glave navedeni su osnovni pojmovi i rezultati koji se mogu naći npr. u [5], [6], [15], [17], [18], [20], [21], [25], [29], [30], [36], [37], [38], [45], [46], [65], [68], [77], [78] i [99]. Posljednja sekcija ove glave sadrže originalne rezultate koji su objavljeni u radu [70].

2.1 Udaljenost stringova

String je konačan niz elemenata izabranih iz istog skupa. Ovaj skup se naziva *alfabetom* i može biti proizvoljan neprazan skup sa bar dva elementa. U ovoj tezi su posmatrani alfabeti koji su konačni skupovi sa $n \geqslant 2$ elemenata (*simbola*) i, bez gubljenja na opštosti, uzima se da su oni oblika $\mathbb{N}_n := \{1, 2, \ldots, n\}$. *Dužina stringa s* predstavlja broj simbola u datom stringu i za ovu vrijednost biće korišćena oznaka $len(s)$. String dužine 0 naziva se *praznim stringom* i označava sa ϵ. Za element a_i nepraznog stringa $s = (a_i, i \in \{1, \ldots, len(s)\})$ se još kaže da se nalazi na $i-toj\ poziciji$ posmatranog stringa, pa, ukoliko se želi istaći pozicija svakog elementa ovog stringa, za zapis stringa se koristi kraća notacija $s = a_1 a_2 \ldots a_{len(s)}$. Skup svih stringova dužine l čiji simboli pripadaju alfabetu $\mathbb{N}_n$ biće označen sa $S(n, l)$.

Prirodno je posmatrati mjere koje opisuju stepen sličnosti dva stringa, pri čemu veće vrijednosti uvedene mjere sličnosti ukazuju na veći stepen sličnosti. Najjednostavniji način zadavanja mjere sličnosti dva stringa zasniva se na korišćenju neke ranije uvedene udaljenosti između njih. Pritom, ne treba smetnuti s uma da metrike generalno izražavaju stepen različitosti objekata čiju udaljenost mjere, pa se u tom smislu mogu shvatiti kao inverzne mjere sličnosti. U narednom razmatranju uvedene su neke metrike koje mjere udaljenost između dva stringa.

Hamingova udaljenost između stringova iste dužine definiše se kao broj

pozicija za koje su odgovarajući simboli ovih stringova različiti. Preciznije, za $s = a_1a_2 \ldots a_l$ i $t = b_1b_2 \ldots b_l$,

$$d_H(s,t) := |\{i \in \{1,2,\ldots,l\} : a_i \neq b_i\}| .$$

Ovaj tip udaljenosti je uveo R. W. Hamming u radu [45] i ima primjenu u nekoliko disciplina, uključujući teoriju informacija, teoriju kodiranja, kriptografiju, bioinformatiku, itd.

Lema 2.1.1. *Hamingova udaljenost je metrika na skupu $S(n,l)$.*

Dokaz. Izuzevši nejednakost trougla, sva ostala svojstva metrike se jednostavno provjeravaju. Neka su $s = a_1a_2 \ldots a_l$, $t = b_1b_2 \ldots b_l$ i $u = c_1c_2 \ldots c_l$ proizvoljni stringovi iz $S(n,l)$. Potrebno je dokazati nejednakost $d_H(s,u) \leqslant d_H(s,t) + d_H(t,u)$. Za stringove $x, y \in S(n,l)$, neka $A(x,y) \subseteq \mathbb{N}_l$ označava skup svih pozicija za koje su odgovarajući simboli stringova x i y različiti. Tada vrijedi $A(s,u) \subseteq A(s,t) \cup A(t,u)$. Zaista, za $i \in A(s,u)$, ukoliko je $b_i = a_i$, tada je $b_i \neq c_i$, odakle slijedi $i \in A(t,u)$, dok za $b_i \neq a_i$ vrijedi $i \in A(s,t)$. Na osnovu toga, dobija se

$$\begin{aligned} d_H(s,u) &= |A(s,u)| \leqslant |A(s,t) \cup A(t,u)| \leqslant |A(s,t)| + |A(t,u)| \\ &= d_H(s,t) + d_H(t,u). \end{aligned}$$

□

Druga udaljenost između stringova koja će biti korišćena u tezi zasniva se na rješavanju LCS problema. String s je *podniz stringa t*, ako postoje pozicije $i_1 < \ldots < i_{len(s)}$ koje pripadaju skupu $\mathbb{N}_{len(t)}$ tako da vrijedi $s = t_{i_1} \ldots t_{i_{len(s)}}$. Ukoliko su dodatno i_j i i_{j+1} susjedne pozicije za svako $j \in \{1,2,\ldots,len(s)-1\}$, tada je string s *podstring stringa t*. *LCS (Longest Common Subsequence) problem* odnosi se na nalaženje zajedničkog podniza maksimalne dužine za dati skup stringova, pri čemu stringovi iz ovog skupa ne moraju obavezno biti jednakih dužina. Zajednički podniz maksimalne dužine za dati skup stringova uvijek postoji, ali ne mora biti jedinstven. Ono što je svakako jedinstveno jeste dužina najdužeg zajedničkog podniza. Ako $LCS(s,t)$ predstavlja dužinu najdužeg zajedničkog podniza stringova s i t, tada se *LCS udaljenost* uvodi na sljedeći način:

$$d_{LCS}(s,t) := \begin{cases} 0, & \text{ako je } len(s) = len(t) = 0; \\ 1 - \dfrac{LCS(s,t)}{\max\{len(s), len(t)\}}, & \text{inače.} \end{cases}$$

Ovaj tip udaljenosti uveden je u radu [4]. Ideja dokaza nejednakosti trougla LCS udaljenosti preuzeta je iz tog rada, sa napomenom da je ovdje izložen nešto jednostavniji dokaz.

Lema 2.1.2. *LCS udaljenost je metrika na skupu* $S(n, l)$.

Dokaz. Tvrđenje vrijedi za $l = 0$, jer je $S(n, 0) = \{\epsilon\}$ i $d_{LCS}(\epsilon, \epsilon) = 0$. Stoga, neka je $l \geqslant 1$. Za proizvoljne $s, t \in S(n, l)$ vrijedi $LCS(s, t) \leqslant l$, što znači da je $d_{LCS}(s, t) \geqslant 0$. Jasno da vrijedi $d_{LCS}(s, s) = 0$, dok iz $d_{LCS}(s, t) = 0$ slijedi $LCS(s, t) = l$, a to je jedino moguće ako je $s = t$. Simetričnost funkcije d_{LCS} slijedi iz činjenice da je $LCS(s, t) = LCS(t, s)$. Potrebno je još dokazati nejednakost trougla. U tu svrhu, najprije će biti dokazano da za proizvoljne stringove $s, t, u \in S(n, l)$ vrijedi

$$LCS(s, t) + LCS(t, u) - LCS(s, u) \leqslant l. \tag{2.1}$$

Neka je x zajednički podniz stringova s i t, a y zajednički podniz stringova t i u, tako da su oba navedena podniza maksimalne dužine, tj. neka vrijedi $len(x) = LCS(s, t)$ i $len(y) = LCS(t, u)$. Ako je $len(x) + len(y) \leqslant l$, tada nejednakost iz formule (2.1) očigledno vrijedi, pa se u daljnjem može pretpostaviti da je $len(x) + len(y) > l$.

Stringovi x i y su podnizovi stringa t, pa postoje pozicije $i_1 < \ldots < i_{len(x)}$ i $j_1 < \ldots < j_{len(y)}$ koje pripadaju skupu $\mathbb{N}_l$ tako da važi $x = t_{i_1} \ldots t_{i_{len(x)}}$ i $y = t_{j_1} \ldots t_{j_{len(y)}}$. Iz uslova $len(x) + len(y) > l$ proističe da skupovi $I_x := \{i_1, \ldots, i_{len(x)}\}$ i $J_y := \{j_1, \ldots, j_{len(y)}\}$ nisu disjunktni. Zbog toga, postoji neprazan string v koji je podniz stringa t, a sastavljen je od svih simbola stringa t koji su na pozicijama koje pripadaju skupu $I_x \cap J_y$. String v je zajednički podniz stringova x i y, pa je samim tim on zajednički podniz stringova s i u. To znači da je $len(v) \leqslant LCS(s, u)$, odakle slijedi

$$len(x)+len(y)-LCS(s,u) \leqslant len(x)+len(y)-len(v) = |I_x \cup J_y| \leqslant len(t) = l.$$

Time je kompletiran dokaz nejednakosti iz formule (2.1), a iz ove nejednakosti direktno slijedi nejednakost trougla. Zaista, za proizvoljne stringove $s, t, u \in S(n, l)$ vrijedi

$$\begin{aligned}
&LCS(s, t) + LCS(t, u) - LCS(s, u) \leqslant l \\
&\Leftrightarrow \frac{LCS(s, t) + LCS(t, u)}{l} \leqslant 1 + \frac{LCS(s, u)}{l} \\
&\Leftrightarrow 1 - \frac{LCS(s, u)}{l} \leqslant 1 - \frac{LCS(s, t)}{l} + 1 - \frac{LCS(t, u)}{l} \\
&\Leftrightarrow d_{LCS}(s, u) \leqslant d_{LCS}(s, t) + d_{LCS}(t, u).
\end{aligned}$$

□

Primjedba 2.1.3. Hamingova udaljenost i *LCS* udaljenost se mogu svrstati u kategoriju tzv. udaljenosti editovanja. Generalno, udaljenosti editovanja izražavaju minimalan broj transformacija potrebnih da se od jednog stringa dobije

drugi string. Različite vrste udaljenosti editovanja se dobijaju preciziranjem dozvoljenog tipa transformacije stringa. Tako npr., Hamingova udaljenost dozvoljava samo zamjenu simbola, dok LCS udaljenost dozvoljava unošenje i brisanje simbola. Udaljenost editovanja koja dopušta zamjenu, unošenje i brisanje naziva se *Livenštajnova udaljenost* i označava sa d_L. Ova udaljenost takođe zadovoljava sve uslove matrike i rekurzivno se može okarakterisati na sljedeći način:

$$d_L(s,\epsilon) = d_L(\epsilon,s) = len(s),$$
$$d_L(s,t) = \min \begin{cases} d_L\left(s_{[1,len(s)-1]}, t\right) + 1, \\ d_L\left(s, t_{[1,len(t)-1]}\right) + 1, \\ d_L\left(s_{[1,len(s)-1]}, t_{[1,len(t)-1]}\right) + 1_{s_{len(s)} \neq t_{len(t)}}, \end{cases}$$

gdje je $x_{[1,i]}$ oznaka za podstring stringa x koji se sastoji od prvih i simbola stringa x (podstring ovakvog tipa još se naziva *prefiksom stringa* x), dok je $1_{s_{len(s)} \neq t_{len(t)}}$ indikator, definisan sa

$$1_{s_{len(s)} \neq t_{len(t)}} = \begin{cases} 1, & \text{ako je } s_{len(s)} \neq t_{len(t)}; \\ 0, & \text{inače.} \end{cases}$$

Pored navedenih udaljenosti, u literaturi postoje druge udaljenosti i mjere sličnosti koji porede dva stringa. Među mnoštvom se mogu izdvojiti Damerau-Livenštajnova udaljenost, Džarova mjera sličnosti, Ratklif-Obešelpova mjera sličnosti, Žakarova mjera sličnosti, Sorensen-Dajsova mjera sličnosti, indeks Tverskog, pseudometrika učestalosti $k-$grama, kosinusna mjera sličnosti, itd. Sve navedene udaljenosti i mjere sličnosti zasnovane su na konceptima koji koriste neku kombinaciju sekvencijalne i skupovne strukture stringa. Nažalost, udaljenosti između dva stringa generisane ovim mjerama sličnosti ne ispunjavaju sve uslove iz definicije metrike (najčešće ne vrijedi nejednakost trougla), pa se, u strogo matematičkom smislu, one i ne mogu smatrati udaljenostima. Stoga će, od udaljenosti između dva stringa, u nastavku teze biti posmatrane samo "prave" metrike d_H, d_{LCS} i d_L.

Poređenje višečlanih familija stringova je moguće izvesti uz pomoć Hauzdorfove udaljenosti. Najprije će biti data definicija ove udaljenosti u opštem slučaju. Neka je (X,d) proizvoljan metrički prostor i $d(x,A) = \inf\{d(x,a) : a \in A\}$ rastojanje tačke $x \in X$ do skupa $A \subseteq X$. Ako je C familija svih kompaktnih podskupova od X, tada se funkcija $D : C \times C \to [0, +\infty)$ definisana sa

$$D(A,B) := \max\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\},$$

naziva *Hauzdorfovom udaljenošću* između skupova A i B. Ograničenost skupova iz familije C garantuje postojanje supremuma iz prethodne definicije. Dakle, postoje vrijednosti $\rho(A,B) := \sup_{a\in A} d(a,B)$ i $\rho(B,A) :=$

$\sup_{b\in B} d(b, A)$, pa je i $D(A, B) = \max\{\rho(A, B), \rho(B, A)\}$ vrijednost koja uvijek postoji.

Lema 2.1.4. *Hauzdorfova udaljenost je metrika na skupu C.*

Dokaz. Očigledno da za svako $A \in C$ vrijedi $D(A, A) = \sup_{a\in A} d(a, A) = 0$. S druge strane, ako za skupove $A, B \in C$ vrijedi $D(A, B) = 0$, tada za proizvoljno $a \in A$ vrijedi $d(a, B) = 0$. To znači da za svako $\varepsilon > 0$ i otvorenu kuglu $B_d(a, \varepsilon) := \{x \in X : d(x, a) < \varepsilon\}$ vrijedi $B_d(a, \varepsilon) \cap B \neq \emptyset$, što implicira $a \in \overline{B}$, odakle, zbog zatvorenosti skupa B kao kompaktnog skupa, slijedi $a \in B$. Time je dokazano $A \subseteq B$, a na sličan način se dokazuje da je $B \subseteq A$, pa je $A = B$. Simetričnost funkcije D slijedi iz njene definicije, pa još ostaje da se dokaže nejednakost trougla. Najprije će biti dokazano da funkcija ρ zadovoljava nejednakost trougla, tj. da za proizvoljne $A, B, C \in C$ vrijedi

$$\rho(A, C) \leqslant \rho(A, B) + \rho(B, C) \tag{2.2}$$

Neka je $a \in A$ proizvoljno. Zbog kompaktnosti skupa B postoji element $b_a \in B$ sa svojstvom $d(a, b_a) = d(a, B)$. Tada je

$$\begin{aligned} d(a, C) &= \inf_{c\in C} d(a, c) \leqslant \inf_{c\in C} (d(a, b_a) + d(b_a, c)) = d(a, b_a) + \inf_{c\in C} d(b_a, c) \\ &= d(a, B) + d(b_a, C) \leqslant \rho(A, B) + \rho(B, C), \end{aligned}$$

pa je $\rho(A, B) + \rho(B, C)$ gornje ograničenje skupa $\{d(a, C) : a \in A\}$, odakle slijedi $\rho(A, C) \leqslant \rho(A, B) + \rho(B, C)$. Time je dokazana nejednakost u formuli (2.2), a iz ove nejednakosti direktno slijedi nejednakost trougla za funkciju D. Zaista, za proizvoljne skupove $A, B, C \in C$ vrijedi

$$\begin{aligned} \rho(A, C) &\leqslant \rho(A, B) + \rho(B, C) \leqslant D(A, B) + D(B, C), \\ \rho(C, A) &\leqslant \rho(C, B) + \rho(B, A) \leqslant D(C, B) + D(B, A) \end{aligned}$$

odakle, zbog simetričnosti funkcije D, slijedi $D(A, C) = \max\{\rho(A, C), \rho(C, A)\}$ $\leqslant D(A, B) + D(B, C)$. □

Neka su A i B podskupovi od $S(n, l)$ tako da je $|A| = |B| = m \geqslant 1$. Ako je $m = 1$, tada bilo koje od rastojanja d_H, d_{LCS}, d_L između stringa iz skupa A i stringa iz skupa B može generisati mjeru sličnosti ovih skupova. U slučaju $m > 1$, jedna mjera sličnosti skupova A i B može biti uvedena uz pomoć Hauzdorfove udaljenosti između ovih skupova definisane u odnosu na neku od pomenutih metrika. Naravno, Hauzdorfova udaljenost između ovih skupova uvijek postoji, jer su posmatrani skupovi konačni i samim tim kompaktni.

Primjer 2.1.5. *Neka su $A, B \subseteq S(4, 8)$ skupovi stringova, pri čemu je*

$$A = \{\underbrace{43144412}_{s_1}, \underbrace{11242443}_{s_2}, \underbrace{33314132}_{s_3}, \underbrace{43343342}_{s_4}, \underbrace{32144142}_{s_5}\}$$

i

$$B = \{\underbrace{34132432}_{t_1}, \underbrace{23432431}_{t_2}, \underbrace{33341422}_{t_3}, \underbrace{44441433}_{t_4}, \underbrace{12341233}_{t_5}\}.$$

Rastojanja u odnosu na Hamingovu metriku d_H su data u sljedećoj matrici:

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{bmatrix} 5 & 6 & 4 & 5 & 7 \\ 6 & 6 & 6 & 5 & 5 \\ 5 & 6 & 4 & 7 & 6 \\ 7 & 7 & 4 & 6 & 6 \\ 5 & 8 & 5 & 7 & 6 \end{bmatrix} \end{array}$$

Iz matrice se uočava da je $d_H(s_1, B) = 4$, $d_H(s_2, B) = 5$, $d_H(s_3, B) = 4$, $d_H(s_4, B) = 4$ i $d_H(s_5, B) = 5$, što znači da je $\rho_H(A, B) = 5$, a takođe je $d_H(t_1, A) = 5$, $d_H(t_2, A) = 6$, $d_H(t_3, A) = 4$, $d_H(t_4, A) = 5$ i $d_H(t_5, A) = 5$, što znači da je $\rho_H(B, A) = 6$. Dakle, vrijedi $D_H(A, B) = \max\{\rho_H(A, B), \rho_H(B, A)\} = 6$.

Rastojanja u odnosu na LCS metriku d_{LCS} su data u sljedećoj matrici:

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{3}{8} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{8} & \frac{5}{8} & \frac{1}{2} & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{2} & \frac{1}{4} & \frac{5}{8} & \frac{1}{2} \\ \frac{3}{8} & \frac{1}{2} & \frac{3}{8} & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{8} & \frac{1}{2} & \frac{3}{8} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array}$$

Iz matrice se uočava da je $d_{LCS}(s_1, B) = \frac{3}{8} = d_{LCS}(s_2, B) = d_{LCS}(s_4, B) = d_{LCS}(s_5, B)$ i $d_{LCS}(s_3, B) = \frac{1}{4}$, što znači da je $\rho_{LCS}(A, B) = \frac{3}{8}$, a takođe je $d_{LCS}(t_1, A) = \frac{3}{8} = d_{LCS}(t_2, A) = d_{LCS}(t_4, A) = d_{LCS}(t_5, A)$ i $d_{LCS}(t_3, A) = \frac{1}{4}$, što znači da je $\rho_{LCS}(B, A) = \frac{3}{8}$. Dakle, važi $D_{LCS}(A, B) = \frac{3}{8}$.

Hauzdorfovo rastojanje daje određenu informaciju o bliskosti dva neprazna kompaktna skupa. Međutim, ono je "jednodimenzionalno", u smislu da se ni za jedan od posmatranih skupova pojedinačno ne uzima u obzir međusobna povezanost i struktura njegovih elemenata. U ovoj glavi biće razmatrane mjere sličnosti familija stringova kod kojih je "pokriven" i ovaj aspekt sličnosti. Osnova na kojoj se ove mjere zasnivaju temelji se na pojmovima i tehnikama simplicijalne homologije i perzistentne (istrajne) homologije.

2.2 Simplicijalna homologija

U ovoj sekciji opisani su osnovni koncepti simplicijalne homologije. Prije svega, slijedi definicija simplicijalnog kompleksa kao primarne strukture na kojoj se simplicijalna homologija zasniva. U ovoj tezi se koriste isključivo apstraktni simplicijalni kompleksi, no, radi boljeg razumijevanja izloženih ideja, početak izlaganja se odnosi na intuitivno familijarnije geometrijske simplicijalne komplekse.

Neka su $u_0, u_1, \ldots, u_k$ tačke skupa $\mathbb{R}^d$. Tačka $x = \sum_{i=0}^{k} \lambda_i u_i$ je *afina kombinacija* tačaka u_i, ako je suma svih λ_i jednaka 1. *Afini omotač* je skup svih afinih kombinacija i on je $k-$*ravan*, ako je posmatranih $k+1$ tačaka *afino nezavisno*, što znači da su proizvoljne dvije afine kombinacije $x = \sum_{i=0}^{k} \lambda_i u_i$ i $y = \sum_{i=0}^{k} \mu_i u_i$ jednake ako i samo ako je $\lambda_i = \mu_i$, za svako i. Jednostavno se dokazuje da je $k+1$ tačaka $u_0, u_1, \ldots, u_k$ afino nezavisno ako i samo ako je k vektora $u_i - u_0$, $i \in \{1, 2, \ldots, k\}$, linearno nezavisno. U $\mathbb{R}^d$ postoji najviše d linearno nezavisnih vektora i samim tim najviše $d+1$ afino nezavisnih tačaka. Afina kombinacija $x = \sum_{i=0}^{k} \lambda_i u_i$ je *konveksna kombinacija*, ako su sve vrijednosti λ_i nenegativni realni brojevi. *Konveksni omotač* je skup konveksnih kombinacija. $k-$*simpleks* σ je konveksni omotač $k+1$ afino nezavisnih tačaka, $\sigma := conv\{u_0, u_1, \ldots, u_k\}$. *Dimenzija* $k-$simpleksa σ je $dim(\sigma) := k$. Za simplekse manjih dimenzija se u praksi koriste sljedeći nazivi: *vrh* za $0-$simpleks, *ivica* za $1-$simpleks, *trougao* za $2-$simpleks i *tetraedar* za $3-$simpleks. Proizvoljan podskup afino nezavisnog skupa tačaka je takođe afino nezavisan, pa i sam određuje jedan simpleks. *Strana simpleksa* σ je konveksni omotač nepraznog podskupa od σ. Ako je simpleks τ strana simpleksa σ, tada se σ još naziva *ko-stranom simpleksa* τ. Očigledno da svaki $k-$simpleks ima ukupno $2^{k+1} - 1$ različitih strana.

Geometrijski simplicijalni kompleks K je konačna, neprazna kolekcija simpleksa sa svojstvima

- Strana proizvoljnog simpleksa iz kolekcije K pripada ovoj kolekciji.
- Presjek proizvoljna dva simpleksa iz K je prazan skup ili njihova zajednička strana.

Dimenzija kompleksa K je maksimalna dimenzija svih simpleksa koji mu pripadaju. *Noseći prostor kompleksa* K, u oznaci $|K|$, je unija svih simpleksa kompleksa K, zajedno sa topologijom potprostora euklidskog prostora kojem svi ovi simpleksi pripadaju. *Triangulacija topološkog prostora* X je simplicijalni kompleks K, zajedno sa homeomorfizmom između X i $|K|$. *Potkompleks*

od K je simplicijalni kompleks $L \subseteq K$. On je *pun*, ukoliko sadrži sve simplekse u K koji su konveksni omotač vrhova iz L. Specijalna vrsta potkompleksa su $j-$*skeletoni* koji se sastoje od svih simpleksa dimenzije j ili manje, $K^{(j)} := \{\sigma \in K : dim(\sigma) \leqslant j\}$. $0-$skeleton se još naziva *skupom vrhova kompleksa* K, $Vert(K) := K^{(0)}$.

Ponekad je jednostavnije definisati simplicijalni kompleks u neodređenom okruženju, a tek naknadno brinuti kako je on smješten u nekom euklidskom prostoru (ako za tim uopšte i postoji potreba). Na taj način se dolazi do pojma apstraktnog simplicijalnog kompleksa.

Apstraktni simplicijalni kompleks $\mathcal{K}$ je uređeni par (V, Σ), gdje je V neprazan skup, a Σ konačna kolekcija nepraznih podskupova od V čiji se elementi nazivaju *simpleksima*, pri čemu je ispunjeno svojstvo da $\tau \subseteq \sigma \in \Sigma$ implicira $\tau \in \Sigma$. Za notiranje simpleksa $\sigma = \{v_0, v_1, \ldots, v_k\}$ još se koristi standardna oznaka $\sigma = [v_0, v_1, \ldots, v_k]$. *Dimenzija simpleksa* σ je $dim\,\sigma := |\sigma| - 1$, dok je *dimenzija kompleksa* maksimalna dimenzija svih njegovih simpleksa, tj. veličina $dim\mathcal{K} := \max_{\sigma\in\Sigma} dim\,\sigma$. *Pun kompleks* ili *standardni kombinatorni kompleks* je simplicijalni kompleks $(V, P(V) \setminus \{\emptyset\})$, gdje je V neprazan konačan skup. Ako je $\tau \subseteq \sigma$, tada se simpleks τ naziva *stranom simpleksa* σ, odnosno simpleks σ se naziva *ko-stranom simpleksa* τ. *Skup vrhova kompleksa* $\mathcal{K}$ je kolekcija $Vert(\mathcal{K})$ svih elemenata $v \in V$ za koje vrijedi $[v] \in \sigma$. *Potkompleks* $\mathcal{L}$ kompleksa $\mathcal{K} = (V, \Sigma)$ je apstraktni simplicijalni kompleks čiji simpleksi formiraju podfamiliju familije Σ. Činjenica da je $\mathcal{L}$ potkompleks kompleksa $\mathcal{K}$ će (neformalno) biti notirana sa $\mathcal{L} \subseteq \mathcal{K}$. Za apstraktne simplicijalne komplekse $\mathcal{K} = (V_\mathcal{K}, \Sigma_\mathcal{K})$ i $\mathcal{L} = (V_\mathcal{L}, \Sigma_\mathcal{L})$, preslikavanje $f : Vert(\mathcal{K}) \to Vert(\mathcal{L})$ takvo da $\sigma \in \Sigma_\mathcal{K}$ ako i samo ako je $f[\sigma] \in \Sigma_\mathcal{L}$ naziva se *simplicijalnim preslikavanjem*. Dva apstraktna simplicijalna kompleksa su *izomorfna*, ako postoji bijektivno simplicijalno preslikavanje između ovih kompleksa. U tom slučaju, ovo preslikavanje se naziva *simplicijalnim izomorfizmom*.

Datom geometrijskom simplicijalnom kompleksu K se može pridružiti apstraktni simplicijalni kompleks $\mathcal{K}$ tako što se odbace simpleksi iz K, a sačuvaju njihovi skupovi vrhova. Tada se $\mathcal{K}$ naziva *šemom vrhova* kompleksa K, dok se K karakteriše kao *geometrijska realizacija* kompleksa $\mathcal{K}$. Konstrukcija geometrijske realizacije je iznenađujuće jednostavna ako se dozvoli korišćenje euklidskog prostora dovoljno velike dimenzije.

Lema 2.2.1. *[36] Svaki apstraktni simplicijalni kompleks dimenzije d ima geometrijsku realizaciju u $\mathbb{R}^{2d+1}$.*

Dokaz. Neka je $\mathcal{K}$ apstraktni simplicijalni kompleks dimenzije d i $Vert(\mathcal{K}) = \{v_0, v_1, \ldots, v_k\}$ skup njegovih vrhova. Neka je $f : Vert(\mathcal{K}) \to \mathbb{R}^{2d+1}$ preslikavanje dato sa $f(v_i) = x_i$, pri čemu su $x_i \in \mathbb{R}^{2d+1}$ različite tačke izabrane tako da proizvoljan podskup skupa $\{x_0, x_1, \ldots, x_k\}$ koji sadrži $2d + 2$ ili ma-

nje elemenata jeste sastavljen od afino nezavisnih tačaka (npr. može se staviti $x_i := (i, i^2, \ldots, i^{2d+1})$). Iz definicije preslikavanja f slijedi njegova injektivnost. Neka su σ_1, σ_2 proizvoljni simpleksi kompleksa $\mathcal{K}$. Za kardinalnost unije ova dva simpleksa važi $|\sigma_1 \cup \sigma_2| = |\sigma_1| + |\sigma_2| - |\sigma_1 \cap \sigma_2| \leqslant |\sigma_1| + |\sigma_2| \leqslant 2d+2$, odakle slijedi $|f[\sigma_1 \cup \sigma_2]| \leqslant 2d+2$. To znači da su tačke iz skupa $f[\sigma_1 \cup \sigma_2]$ afino nezavisne, pa je svaka konveksna kombinacija sastavljena od tačaka iz ovog skupa jedinstvena. Zbog toga, ako je $\tau_1 := conv(f[\sigma_1])$ i $\tau_2 := conv(f[\sigma_2])$, tada $x \in \tau_1 \cap \tau_2$ ako i samo ako je x konveksna kombinacija tačaka iz $f[\sigma_1] \cap f[\sigma_2] = f[\sigma_1 \cap \sigma_2]$. Posljedično, $\tau_1 \cap \tau_2$ je ili prazan skup ili simpleks $conv(f[\sigma_1 \cap \sigma_2])$. Dakle, $K := \{conv(f[\sigma]) : \sigma$ je simpleks kompleksa $\mathcal{K}\}$ je geometrijska realizacija kompleksa $\mathcal{K}$. □

Uvođenje metrike na nekom skupu omogućava povezivanje parova najbližih tačaka. Na taj način se dobija tzv. graf susjeda kojim se iskazuje povezanost parova tačaka datog skupa. Ovaj koncept se može uopštiti u smislu da se, za dati prirodan broj $m > 1$, posmatra povezanost $k+1-$torki tačaka, za $k \in \{1, 2, \ldots, m\}$. Tako dobijena višedimenzionalna generalizacija grafa susjeda predstavlja jedan način generisanja apstraktnih simplicijalnih kompleksa. Većina (apstraktnih) simplicijalnih kompleksa u daljnjem izlaganju ove teze dobijena je na ovaj način.

Neka je $\mathcal{F}$ konačna familija nepraznih skupova. *Nerv* familije $\mathcal{F}$ čine sve neprazne podfamilije od $\mathcal{F}$ sa svojstvom da je presjek svih skupova koji pripadaju podfamiliji neprazan skup. Preciznije, $Nerv(\mathcal{F}) := \{\mathcal{A} \subseteq \mathcal{F} : \bigcap \mathcal{A} \neq \emptyset\}$. Iako familija $\mathcal{F}$ može biti proizvoljna, od interesa će biti slučaj kada je to familija kugli nekog metričkog prostora.

Neka je (X, d) metrički prostor. Za $r > 0$, neka je $B_d[x, r] := \{y \in X : d(x, y) \leqslant r\}$ *zatvorena kugla* poluprečnika r oko tačke x. Ako je $K \subseteq X$ konačan skup i ukoliko je $r > 0$ proizvoljno, jednostavno se provjerava da uređeni par $\left(K, \left\{A \in P(K) \setminus \{\emptyset\} : \bigcap_{x \in A} B_d[x, r] \neq \emptyset\right\}\right)$ predstavlja apstraktni simplicijalni kompleks. Zaista, ako $A \subseteq K$ ima svojstvo da je $\bigcap_{x \in A} B[x, r] \neq \emptyset$, tada za svaki neprazan skup $B \subseteq A$ vrijedi $\emptyset \neq \bigcap_{x \in A} B[x, r] \subseteq \bigcap_{x \in B} B[x, r]$. Ovaj kompleks se naziva *Čehovim kompleksom* i označava sa $C_K^{(r)}$. Jednostavno se uočava da "naduvavanjem" zatvorenih kugli istog poluprečnika ostaju očuvane presječne tačke ovih kugli, ukoliko one postoje. To znači da, za $0 < r_1 < r_2$, svi simpleksi koji pripadaju Čehovom kompleksu $C_K^{(r_1)}$ takođe pripadaju Čehovom kompleksu $C_K^{(r_2)}$, pa je $C_K^{(r_1)}$ potkompleks kompleksa $C_K^{(r_2)}$.

Pored Čehovog kompleksa, moguće je uvesti apstraktni simplicijalni kompleks čiji skup simpleksa nije dobijen kao nerv neke familije skupova. Npr., ako

je K konačan neprazan skup u metričkom prostoru (X, d), $r > 0$ proizvoljna vrijednost i Σ familija koja sadrži sve neprazne podskupove $A \subseteq K$ sa svojstvom da je $diam(A) := \sup\{d(a_1, a_2) : a_1, a_2 \in A\} \leqslant r$, tada je (K, Σ) apstraktni simplicijalni kompleks koji se naziva *Vietoris-Ripsovim kompleksom* i označava sa $\mathcal{VR}_K^{(r)}$. Za razliku od Čehovog kompleksa, Vietoris-Ripsov kompleks je u potpunosti određen svojim $1-$simpleksima. To efektivno znači da je za određivanje simpleksa Vietoris-Ripsovog kompleksa dovoljno znati matricu incidencije, tj. matricu koja sadrži sva rastojanja između elemenata skupa K.

Lema 2.2.2. *Neka je K konačan, neprazan skup u metričkom prostoru $(\mathbb{R}^d, e)$, gdje je e Euklidska metrika. Tada za proizvoljno $r > 0$ vrijedi $\mathcal{VR}_K^{(r)} \subseteq \mathcal{C}_K^{(r)} \subseteq \mathcal{VR}_K^{(2r)}$.*

Dokaz. Potrebno je dokazati da je, za proizvoljno $r > 0$, Vietoris-Ripsov kompleks $\mathcal{VR}_K^{(r)}$ potkompleks Čehovog kompleksa $\mathcal{C}_K^{(r)}$, kao i da je Čehov kompleks $\mathcal{C}_K^{(r)}$ potkompleks Vietoris-Ripsovog kompleksa $\mathcal{VR}_K^{(2r)}$.

Neka je $k \leqslant dim\left(\mathcal{VR}_K^{(r)}\right)$ proizvoljno i $\sigma = [v_0, v_1, \ldots, v_k]$ proizvoljan simpleks kompleksa $\mathcal{VR}_K^{(r)}$. Tada za sve $i, j \in \{0, 1, \ldots, k\}$ vrijedi $e(v_i, v_j) \leqslant r$. Ako se stavi $v := \dfrac{1}{k+1}\sum_{i=0}^{k} v_i$, tada za proizvoljno $j \in \{0, 1, \ldots, k\}$ vrijedi

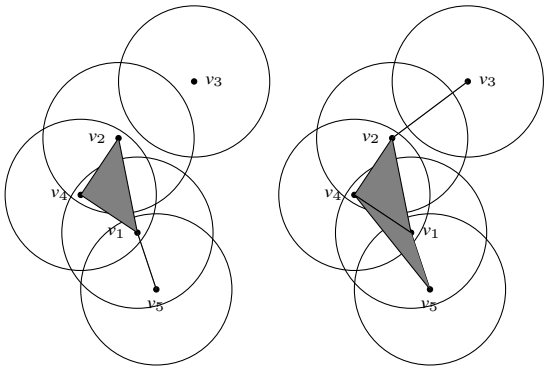
$$e(v, v_j) = e\left(\frac{1}{k+1}\sum_{i=0}^{k} v_i, \frac{1}{k+1}\sum_{i=0}^{k} v_j\right) = \frac{1}{k+1}\cdot e\left(\sum_{i=0}^{k} v_i, \sum_{i=0}^{k} v_j\right)$$
$$\leqslant \frac{1}{k+1}\cdot\sum_{i=0}^{k} e(v_i, v_j) \leqslant r,$$

što implicira $v \in \bigcap_{i=0}^{k} B_e[v_i, r]$. Dakle, $\bigcap_{i=0}^{k} B_e[v_i, r] \neq \emptyset$, što znači da je σ simpleks kompleksa $\mathcal{C}_K^{(r)}$. Time je dokazano $\mathcal{VR}_K^{(r)} \subseteq \mathcal{C}_K^{(r)}$.

Neka je $k \leqslant dim\left(\mathcal{C}_K^{(r)}\right)$ proizvoljno i $\sigma = [v_0, v_1, \ldots, v_k]$ proizvoljan simpleks kompleksa $\mathcal{C}_K^{(r)}$. Tada postoji tačka $v \in \bigcap_{i=0}^{k} B_e[v_i, r]$. Za proizvoljne $i, j \in \{0, 1, \ldots, k\}$ vrijedi $d_e(v_i, v_j) \leqslant d_e(v_i, v) + d_e(v, v_j) \leqslant 2r$, što znači da je σ simpleks kompleksa $\mathcal{VR}_K^{(2r)}$. Time je dokazano $\mathcal{C}_K^{(r)} \subseteq \mathcal{VR}_K^{(2r)}$. □

Primjer 2.2.3. *Neka je $K = \{v_1, v_2, v_3, v_4, v_5\}$, gdje su $v_1 = (0, 0)$, $v_2 = \left(-\frac{1}{4}, \frac{5}{4}\right)$, $v_3 = \left(\frac{3}{4}, 2\right)$, $v_4 = \left(-\frac{3}{4}, \frac{1}{2}\right)$ i $v_5 = \left(\frac{1}{4}, -\frac{3}{4}\right)$ tačke Euklidskog prostora $\mathbb{R}^2$. Na sljedećoj slici predstavljeni su Vietoris-Ripsov kompleks $\mathcal{VR}_K^{(1)}$*

(lijevo) i Čehov kompleks $\mathcal{C}_K^{(1)}$ (desno). Uočava se da, u odnosu na Vietoris-



Slika 2.1

Ripsov kompleks $\mathcal{VR}_K^{(1)}$, Čehov kompleks $\mathcal{C}_K^{(1)}$ dodatno sadrži $1-$simplekse $[v_2, v_3], [v_4, v_5]$ i $2-$simpleks $[v_1, v_4, v_5]$. Ovi simpleksi ne pripadaju kompleksu $\mathcal{VR}_K^{(1)}$, jer $v_3 \notin B_e[v_2, 1]$ i $v_5 \notin B_e[v_4, 1]$.

Među brojnim topološkim svojstvima simplicijalnog kompleksa izdvaja se ono koje registruje "rupe" određene dimenzije datog simplicijalnog kompleksa, a koje nastaju kao rezultat nedostatka međupovezanosti njegovih vrhova. U okvirima algebarske topologije pomenute "rupe" se opisuju korišćenjem simplicijalne homologije. Simplicijalna homologija predstavlja jedan vid "algebraizacije" datog simplicijalnog kompleksa, tako što se datom kompleksu pridružuje niz homoloških grupa.

Neka je $\mathcal{K}$ apstraktni simplicijalni kompleks i k data dimenzija. *$k-$lanac* je formalna suma $c = \sum a_i\sigma_i$, gdje su σ_i $k-$simpleksi kompleksa $\mathcal{K}$, a a_i su koeficijenti iz nekog polja ili prstena. Mada koeficijenti a_i mogu biti brojevi iz nekog od polja $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ ili iz prstena $\mathbb{Z}$, za potrebe ove teze biće podrazumijevano da ovi koeficijenti pripadaju polju $\mathbb{Z}_2$, modula pri cjelobrojnom djeljenju sa 2. Biranje ovako jednostavnih koeficijenata omogućava da se $k-$lanac $c = \sum a_i\sigma_i$ shvati kao skup simpleksa $\{\sigma_i : a_i = 1\}$. Sabiranje dva k-lanca može da se uvede slično kao kod polinoma; za $c_1 = \sum a_i\sigma_i$ i $c_2 = \sum b_i\sigma_i$, $c_1 + c_2 := \sum(a_i + b_i)\sigma_i$, pri čemu se suma $a_i + b_i$ uzima po modulu 2. U skupovnoj interpretaciji, suma dva $k-$lanca jednaka je njihovoj simetričnoj razlici.

Lema 2.2.4. *Ako je $C_k(\mathcal{K})$ skup svih $k-$lanaca simplicijalnog kompleksa $\mathcal{K}$, tada je $(C_k(\mathcal{K}), +)$ slobodna Abelova grupa, tj. Abelova grupa za koju postoji konačan minimalan skup generatora za $C_k(\mathcal{K})$.*

Dokaz. Zatvorenost, asocijativnost i komutativnost sabiranja $k-$lanaca slijedi iz istovjetnih svojstava koja vrijede za sabiranje po modulu 2. Neutralni element

je $\sum 0\sigma_i = 0$. Za proizvoljan $k-$lanac c, njegov inverzni element je $-c = c$, jer za sabiranje po modulu 2 vrijedi $c + c = 0$. Zbog toga je $(C_k(\mathcal{K}), +)$ Abelova grupa. Skup svih simpleksa kompleksa $\mathcal{K}$ je očigledno konačan skup generatora za $C_k(\mathcal{K})$, što implicira da je $(C_k(\mathcal{K}), +)$ slobodna Abelova grupa. □

Za svaku dimenziju k, grupa $(C_k(\mathcal{K}), +)$ se naziva *grupom $k-$lanaca* i u daljnjem će kraće biti notirana sa $C_k(\mathcal{K})$. U cilju povezivanja grupa lanaca, definiše se pojam *granice $k-$simpleksa*, koja predstavlja sumu svih $(k-1)-$dimenzionalnih strana ovog simpleksa. Preciznije, granica $k-$simpleksa $\sigma = [v_0, v_1, \ldots, v_k]$ je $(k-1)$-lanac $\partial_k\sigma := \sum_{j=0}^{k} [v_0, v_1, \ldots, \hat{v_j}, \ldots, v_k]$, pri čemu je $\hat{}$ oznaka koja simbolizuje odsustvo naznačenog vrha u odgovarajućem simpleksu. Granica $k-$lanca $c = \sum a_i\sigma_i$ definiše se kao suma granica njegovih simpleksa, $\partial_k\sigma := \sum a_i\partial_k(\sigma_i)$. Jednostavno se uočava da na ovaj način uvedeno preslikavanje $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$ zadovoljava uslov $\partial_k(c_1 + c_2) = \partial_k(c_1) + \partial_k(c_2)$, što znači da je, za svaku dimenziju k, ∂_k homomorfizam između grupa $C_k(\mathcal{K})$ i $C_{k-1}(\mathcal{K})$, sa napomenom da se dodefiniše $C_{-1}(\mathcal{K}) := \{0\}$. *Kompleks lanaca* pridružen kompleksu $\mathcal{K}$ je niz grupa lanaca povezanih homomorfizmima granica:

$$\ldots \xrightarrow{\partial_{k+2}} C_{k+1}(\mathcal{K}) \xrightarrow{\partial_{k+1}} C_k(\mathcal{K}) \xrightarrow{\partial_k} C_{k-1}(\mathcal{K}) \xrightarrow{\partial_{k-1}} \ldots$$

Homomorfizam ∂_k biće kraće označen sa ∂, ukoliko je iz konteksta jasno na koju grupu lanaca se on odnosi.

U grupi $k-$lanaca izdvajaju se dva tipa lanaca koji omogućavaju definisanje pojma homološke grupe. *$k-$ciklus* je $k-$lanac c za koji vrijedi $\partial c = 0$. Kako granica ∂ komutira sa sabiranjem ciklusa, skup svih $k-$ciklusa formira *grupu $k-$ciklusa*, koja se označava sa $Z_k(\mathcal{K})$ i ova grupa je podgrupa grupe $k-$lanaca, pa je i sama slobodna Abelova grupa. Uočava se da grupa $k-$ciklusa ustvari predstavlja jezgro homomorfizma granice ∂_k, tj. da je $Z_k(\mathcal{K}) = ker\partial_k$. Za $k = 0$, zbog $C_{-1}(\mathcal{K}) := \{0\}$, vrijedi $Z_0(\mathcal{K}) = ker\partial_0 = C_0(\mathcal{K})$, ali za $k > 0$ podgrupa $Z_k(\mathcal{K})$ ne mora da bude jednaka $C_k(\mathcal{K})$. *$k-$granica* je $k-$lanac c koji je slika nekog $(k+1)-$lanca funkcijom granice, tj. $c = \partial_{k+1}d$, za neki $(k+1)-$lanac d. Kako granica ∂ komutira sa sabiranjem ciklusa, skup svih $k-$granica formira *grupu $k-$granica*, koja se označava sa $B_k(\mathcal{K})$ i ova grupa je podgrupa grupe $k-$lanaca, te je i ona slobodna Abelova grupa. Uočava se da grupa $k-$granica ustvari predstavlja sliku homomorfizma granice ∂_{k+1}, tj. da je $B_k(\mathcal{K}) = im\partial_{k+1}$. Sljedeće tvrđenje iskazuje osnovno svojstvo homomorfizama granice i poznato je pod nazivom *Fundamentalna lema homologije*.

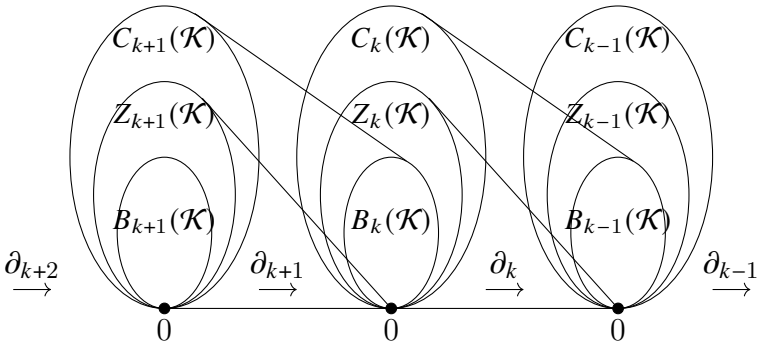
Lema 2.2.5. *[36] Za svaku dimenziju k i proizvoljan $(k+1)-$lanac d vrijedi $\partial_k\partial_{k+1}d = 0$.*

Dokaz. Dovoljno je pokazati da je $\partial_k \partial_{k+1}\tau = 0$, za svaki $(k+1)-$simpleks τ. Neka je $\tau = [v_0, v_1, \ldots, v_k, v_{k+1}]$ proizvoljno. Tada je

$$\partial_{k+1}\tau = \sum_{j=0}^{k+1} [v_0, v_1, \ldots, \hat{v_j}, \ldots, v_k, v_{k+1}],$$

pa, koristeći činjenicu da svaka $(k-1)-$strana od τ pripada tačno dvjema $k-$stranama ovog simpleksa, može se zaključiti da djelovanje homomorfizma ∂_k na prethodnu sumu dovodi do toga da se svaka $(k-1)-$strana simpleksa τ u ovoj sumi pojavljuje tačno dva puta. Uzimajući u obzir da za svaki lanac c vrijedi $c + c = 0$, slijedi $\partial_k\,(\partial_{k+1}\tau) = 0$. □

Direktna posljedica prethodne leme je da svaka $k-$granica predstavlja ujedno i $k-$ciklus, što znači da je grupa $B_k(\mathcal{K})$ podgrupa grupe $Z_k(\mathcal{K})$. Na sljedećoj slici ilustrovan je odnos između ove tri grupe, kao i njihova povezanost između susjednih dimenzija ostvarena uz pomoć homomorfizama granice.



Slika 2.2

Činjenica da je grupa $k-$granica podgrupa grupe $k-$ciklusa omogučava definisanje faktor grupe. *$k-$ta homološka grupa* kompleksa $\mathcal{K}$ označava se sa $H_k(\mathcal{K})$ i definiše sa $H_k(\mathcal{K}) := Z_k(\mathcal{K})/B_k(\mathcal{K})$. Ova grupa ima konačan skup generatora, pa je njen rank prirodan broj $\beta_k(\mathcal{K}) := rank(H_k(\mathcal{K}))$, koji se naziva *Betijevim brojem* kompleksa $\mathcal{K}$. Svaki element grupe $H_k(\mathcal{K})$ dobija se dodavanjem svih $k-$granica datom $k-$ciklusu c, tj. elementi ove grupe su oblika $c+B_k(\mathcal{K})$, za $c \in Z_k(\mathcal{K})$. Ako su c_1, c_2 dva $k-$ciklusa za koje vrijedi $c_1 = c_2+b$, za neku $k-$granicu b, tada je $c_1+B_k(\mathcal{K}) = c_2+B_k(\mathcal{K})$, jer je $b+B_k(\mathcal{K}) = B_k(\mathcal{K})$. Zbog toga, klasa $c + B_k(\mathcal{K})$, $c \in Z_k(\mathcal{K})$, predstavlja koset od $H_k(\mathcal{K})$ i klasa ovakvog oblika se još naziva *homološkom klasom*. Ciklusi koji pripadaju istoj homološkoj klasi se nazivaju *homolognim ciklusima* i svaki od njih se može izabrati kao predstavnik te klase. Zbir dvije homološke klase definiše se sa $(c_1 + B_k(\mathcal{K})) + (c_2 + B_k(\mathcal{K})) := (c_1 + c_2) + B_k(\mathcal{K})$ i ovako definisano sabiranje ne zavisi od izbora predstavnika. Jednostavno se može potvrditi da $H_k(\mathcal{K})$ u odnosu na ovu operaciju sabiranja zaista čini grupu, kao i da je ova grupa

Abelova, jer je takva grupa $Z_k(\mathcal{K})$. Homološke klase iz grupe $H_k(\mathcal{K})$ mogu da se shvate kao $k-$ciklusi koji nisu $k-$granice i slikovito se mogu predstaviti kao $k-$dimenzionalne "rupe" u kompleksu $\mathcal{K}$. U toj interpretaciji, Betijev broj $\beta_k(\mathcal{K})$ predstavlja minimalan broj $k-$dimenzionalnih rupa uz pomoć kojih je moguće "generisati" sve $k-$dimenzionalne rupe kompleksa $\mathcal{K}$.

Neka je $\mathcal{K}$ kompleks koji sadrži $m \geqslant 1$ simpleksa dimenzije $k \geqslant 0$. Tada je $rank(C_k(\mathcal{K})) = m$ i grupa $k-$lanaca $C_k(\mathcal{K})$ je izomorfna direktnoj sumi $\mathbb{Z}_2^m := \underbrace{\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \ldots \oplus \mathbb{Z}_2}_{m}$. U ovoj postavci, za grupe $k-$ciklusa i $k-$granica vrijedi $rank(B_k(\mathcal{K})) \leqslant rank(Z_k(\mathcal{K})) \leqslant m$ i ove grupe su izomorfne direktnoj sumi odgovarajućeg broja kopija grupe $\mathbb{Z}_2$. S obzirom da je za svaku homološku klasu iz $H_k(\mathcal{K})$ broj njenih ciklusa jednak broju $k-$granica, slijedi da je $|H_k(\mathcal{K})| = |Z_k(\mathcal{K})|/|B_k(\mathcal{K})|$, što se u terminima rankova može izraziti u vidu jednakosti $rank(H_k(\mathcal{K})) = rank(Z_k(\mathcal{K})) - rank(B_k(\mathcal{K}))$.

Primjer 2.2.6. *Neka je $\mathcal{K}$ Vietoris-Ripsov kompleks $\mathcal{VR}_K^{(1)}$ iz prethodnog primjera, tj. $\mathcal{K} = (K, \Sigma)$, gdje je $K = \{v_1, v_2, v_3, v_4, v_5\}$ i*

$$\Sigma = \{[v_1], [v_2], [v_3], [v_4], [v_5], [v_1, v_2], [v_1, v_4], [v_1, v_5], [v_2, v_4], [v_1, v_2, v_4]\}.$$

$0-$simpleksi $[v_1], [v_2], [v_3], [v_4], [v_5]$ generišu grupu $0-$lanaca, pa je grupa $C_0(\mathcal{K})$ izomorfna grupi $\mathbb{Z}_2^5$. $1-$simpleksi $[v_1, v_2], [v_1, v_4], [v_1, v_5], [v_2, v_4]$ generišu grupu $1-$lanaca, pa je grupa $C_1(\mathcal{K})$ izomorfna grupi $\mathbb{Z}_2^4$. Jedini $2-$simpleks $[v_1, v_2, v_4]$ generiše grupu $2-$lanaca, pa je grupa $C_2(\mathcal{K})$ izomorfna grupi $\mathbb{Z}_2$. Za $k > 2$, kompleks $\mathcal{K}$ nema $k-$simpleksa, pa je za takve dimenzije grupa $k-$lanaca trivijalna, tj. vrijedi $C_k(\mathcal{K}) \cong \{0\}$. Zbog toga, za $k > 2$ vrijedi $H_k(\mathcal{K}) \cong \{0\}$. Iz $\partial_0 = 0$ slijedi $Z_0(\mathcal{K}) = C_0(\mathcal{K})$, dok se iz

$$\begin{aligned}
&\partial_1 \left(a_1[v_1, v_2] + a_2[v_1, v_4] + a_3[v_1, v_5] + a_4[v_2, v_4]\right) \\
&= a_1\partial_1([v_1, v_2]) + a_2\partial_1([v_1, v_4]) + a_3\partial_1([v_1, v_5]) + a_4\partial_1([v_2, v_4]) \\
&= a_1([v_1] + [v_2]) + a_2([v_1] + [v_4]) + a_3([v_1] + [v_5]) + a_4([v_2] + [v_4]) \\
&= (a_1 + a_4)([v_1] + [v_2]) + (a_2 + a_4)([v_1] + [v_4]) + a_3([v_1] + [v_5]) \qquad (2.3)
\end{aligned}$$

dobija da lanci $[v_1] + [v_2], [v_1] + [v_4], [v_1] + [v_5]$ čine minimalan generator grupe $B_0(\mathcal{K})$. Zbog toga, $H_0(\mathcal{K}) \cong \mathbb{Z}_2^5/\mathbb{Z}_2^3 \cong \mathbb{Z}_2^2 = \mathbb{Z}_2 \oplus \mathbb{Z}_2$, pri čemu $[v_2]$ i

$[v_3]$ *čine minimalan generator ove grupe, jer vrijedi*

$$\underbrace{a_1[v_1] + a_2[v_2] + a_3[v_3] + a_4[v_4] + a_5[v_5]}_{\in Z_0(\mathcal{K})} = (a_4 + a_5)[v_1] + (a_1 + a_2)[v_2]$$
$$+ a_3[v_3] + \Big(a_1([v_1] + [v_2]) + a_4([v_1] + [v_4]) + a_5([v_1] + [v_5])\Big)$$
$$= (a_1 + a_2 + a_4 + a_5)[v_2] + a_3[v_3]$$
$$+ \underbrace{\Big((a_1 + a_4 + a_5)([v_1] + [v_2]) + a_4([v_1] + [v_4]) + a_5([v_1] + [v_5])\Big)}_{\in B_0(\mathcal{K})}.$$

Iz prikaza (2.3) se dobija

$$\partial_1 \left(a_1[v_1, v_2] + a_2[v_1, v_4] + a_3[v_1, v_5] + a_4[v_2, v_4]\right) = 0$$
$$\Leftrightarrow a_1 + a_4 = a_2 + a_4 = a_3 = 0 \Leftrightarrow a_1 = a_2 = a_4 \; i \; a_3 = 0,$$

odakle slijedi da je grupa $Z_1(\mathcal{K}) = ker(\partial_1)$ *generisana minimalnim generatorom koga čini lanac* $[v_1, v_2] + [v_1, v_4] + [v_2, v_4]$*, pa je* $Z_1(\mathcal{K}) \cong \mathbb{Z}_2$*. Isti ovaj lanac čini generator grupe* $B_1(\mathcal{K}) = ker(\partial_1)$*, jer je*

$$\partial_2(a[v_1, v_2, v_4]) = a([v_1, v_2] + [v_1, v_4] + [v_2, v_4]).$$

Dakle, faktor grupa $H_1(\mathcal{K})$ *je trivijalna, tj. vrijedi* $H_1(\mathcal{K}) \cong \{0\}$*. Isti zaključak vrijedi za homološku grupu* $H_2(\mathcal{K})$*, jer je* $ker(\partial_2) = \{0\} = im(\partial_3)$*. Slikovito rečeno, kompleks* $\mathcal{K}$ *ima dvije komponente povezanosti (jer je* $rank(H_0(\mathcal{K})) = 2$*) i, za proizvoljno* $k \geqslant 1$*, nema* $k-$*dimenzionalnih rupa (jer je* $rank(H_k(\mathcal{K})) = 0$*).*

Prethodni primjer upućuje da je za računanje homoloških grupa potrebno kombinovati informacije iz dva izvora, jednog koji predstavlja cikluse i drugog koji predstavlja granice. U narednom je izložen efektivan metod računanja homoloških grupa koji koristi elemente linearne algebre.

Neka je $\mathcal{K}$ simplicijalni kompleks. Za dimenziju $k \geqslant 0$, neka je $m_k := rank(C_k(\mathcal{K}))$, $z_k := rank(Z_k(\mathcal{K}))$ i $b_k := rank(B_k(\mathcal{K}))$. Kako su svi operatori granice homomorfizmi, slijedi da je $m_k = z_k + b_{k-1}$. Homomorfizmu granice ∂_k se može dodijeliti *matrica granice*, čiji su redovi $(k-1)-$simpleksi, a kolone $k-$simpleksi. Ukoliko se za svaku dimenziju zada totalni poredak na skupu odgovarajućih simpleksa, tada je matrica granice oblika $[a_{ij}]$, pri čemu $i \in \{1, 2, \ldots, m_{k-1}\}$, $j \in \{1, 2, \ldots, m_k\}$, a element a_{ij} ove matrice je definisan sa

$$a_{ij} := \begin{cases} 1, & \text{ako je } i-\text{ti } (k-1)-\text{simpleks strana } j-\text{tog } k-\text{simpleksa;} \\ 0, & \text{inače.} \end{cases}$$

Za dati $k-$lanac $c = \sum a_i \sigma_i$, njegova granica može da se izračuna uz pomoć matrice granice na sljedeći način:

$$\partial_k c = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m_k} \\ a_{21} & a_{22} & \dots & a_{2m_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m_{k-1}1} & a_{m_{k-1}2} & \dots & a_{m_{k-1}m_k} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{m_k} \end{bmatrix}$$

Drugačije rečeno, kolekcija kolona predstavlja $k-$lanac, a suma ovih kolona daje njegovu granicu.

Redovi matrice granice homomorfizma ∂_k formiraju bazu grupe $C_{k-1}(\mathcal{K})$, dok kolone ove matrice formiraju bazu grupe $C_k(\mathcal{K})$. Dvije transformacije koje se mogu izvesti sa kolonama, a da se pritom ne promijeni rank ove matrice, su zamjena dvije kolone i dodavanje jedne kolone drugoj. Obje transformacije mogu se izraziti množenjem sa matricom $V := [v_{ij}]$ sa desne strane. Kod zamjene kolona j_1 i j_2 uzima se $v_{j_1 j_2} = v_{j_2 j_1} = 1$ i $v_{jj} = 1$, za sve $j \notin \{j_1, j_2\}$, dok se na svim ostalim pozicijama matrice V stavljaju nule. Kod dodavanja kolone j_1 koloni j_2 uzima se $v_{j_1 j_2} = 1$ i $v_{jj} = 1$, za sve j, dok se na svim ostalim pozicijama matrice V stavljaju nule. Analogno transformacijama kolona, mogu se posmatrati transformacije redova matrice granice koje ne mijenjaju njen rank: zamjena dva reda i dodavanje jednog reda drugom. Obje transformacije mogu se izraziti množenjem sa matricom $U := [u_{ij}]$ sa lijeve strane. Kod zamjene redova i_1 i i_2 uzima se $u_{i_1 i_2} = u_{i_2 i_1} = 1$ i $u_{ii} = 1$, za sve $i \notin \{i_1, i_2\}$, dok se na svim ostalim pozicijama matrice U stavljaju nule. Kod dodavanja reda i_1 redu i_2 uzima se $u_{i_2 i_1} = 1$ i $v_{ii} = 1$, za sve i, dok se dok se na svim ostalim pozicijama matrice U stavljaju nule.

Koristeći opisane transformacije kolona i redova omogućavaju da se matrica granice homomorfizma ∂_k redukuje do *Smitove normalne forme*. Za aritmetiku po modulu 2, ovo znači da je rezultat redukcije matrica kod koje je početni komad dijagonale sastavljen od jedinica, a da je njen ostatak sačinjen od nula. Preciznije, iz $m_k = b_{k-1} + z_k$ slijedi da prvih b_{k-1} kolona određuje kvadratnu podmatricu sa jedinicama na dijagonali, dok je preostalih z_k kolona sastavljeno od nula. Prvopomenute kolone predstavljaju $k-$lance čije nenula granice generišu grupu $B_{k-1}(\mathcal{K})$, dok ostale kolone predstavljaju $k-$cikluse koji generišu grupu $Z_k(\mathcal{K})$. Nakon redukcije svih matrica granica, iz navedenih normalnih formi je moguće "pročitati" Betijeve brojeve kao razliku rankova, $\beta_k = rank(Z_k(\mathcal{K})) - rank(B_k(\mathcal{K})), k \geqslant 0$. Važno je istaći da se tokom postupka redukcije oznake vrsta i kolona mijenjaju u skladu sa izvršenim transformacijama. Pritom, dodavanjem $i-$te kolone $j-$toj koloni oznaka "nove" $j-$te kolone postaje zbir lanaca kojima su označeni "stara" $i-$ta i $j-$ kolona. Kod redova je obrnuto: dodavanjem $i-$tog reda $j-$tom redu oznaka "novog" $i-$tog reda postaje zbir lanaca kojima su označeni "stari" $i-$ti i $j-$ti red. U cilju

dobijanja baza za grupu granica $B_{k-1}(\mathcal{K})$ i grupu ciklusa $Z_k(\mathcal{K})$, potrebno je pratiti proizvode matrica koje predstavljaju transformacije redova i kolona.Ako se proizvod matrica sa lijeve strane označi sa U_{k-1}, a proizvod matrica sa desne strane sa V_k, normalna forma se izražava u obliku $N_k := U_{k-1}A_kV_k$, gdje je A_k matrica granice homomorfizma ∂_k. Nova baza za $Z_k(\mathcal{K})$ data je u posljednjih z_k kolona matrice V_k. Slično, nova baza za $B_{k-1}(\mathcal{K})$ je enkodirana u matrici U_{k-1}, pri čemu se bazni vektori dobijaju iz prvih b_{k-1} kolona njene inverzne matrice.

Sam postupak svođenja na normalnu formu izvodi se na sličan način kao kod Gausovog metoda eliminacije. U najviše dvije zamjene postiže se da na početnom mjestu matrice bude 1, a sa najviše $m_{k-1}-1$ dodavanja redova i najviše $m_k - 1$ dodavanja kolona se postiže da na svim ostalim mjestima u prvom redu i prvoj koloni bude 0. Zatim se postupak normalizacije rekurzivno primjenjuje na podmatricu koja se dobija uklanjanjem prvog reda i prve kolone. Ilustracija prethodnog postupka data je u sljedećem primjeru.

Primjer 2.2.7. *Neka je $\mathcal{K} = (V, \Sigma)$, gdje je $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ i*

$$\begin{aligned}\Sigma = \{ &\underbrace{[v_1]}_{\sigma_1}, \underbrace{[v_2]}_{\sigma_2}, \underbrace{[v_3]}_{\sigma_3}, \underbrace{[v_4]}_{\sigma_4}, \underbrace{[v_5]}_{\sigma_5}, \underbrace{[v_6]}_{\sigma_6}, \underbrace{[v_7]}_{\sigma_7}, \underbrace{[v_1, v_2]}_{\sigma_8}, \underbrace{[v_1, v_6]}_{\sigma_9}, \underbrace{[v_1, v_7]}_{\sigma_{10}},\\ &\underbrace{[v_2, v_6]}_{\sigma_{11}}, \underbrace{[v_2, v_7]}_{\sigma_{12}}, \underbrace{[v_6, v_7]}_{\sigma_{13}}, \underbrace{[v_1, v_2, v_6]}_{\sigma_{14}}, \underbrace{[v_1, v_2, v_7]}_{\sigma_{15}}, \underbrace{[v_1, v_6, v_7]}_{\sigma_{16}},\\ &\underbrace{[v_2, v_6, v_7]}_{\sigma_{17}}\}.\end{aligned}$$

Svi simpleksi kompleksa $\mathcal{K}$ su dimenzije najviše 2. To znači da je grupa $C_k(\mathcal{K})$, pa i $H_k(\mathcal{K})$ trivijalna za $k \geqslant 3$. Za $k = 0$, $\{\sigma_i : 1 \leqslant i \leqslant 7\}$ je baza grupe $C_0(\mathcal{K})$, pa, kako je $\partial_0 = 0$, vrijedi $Z_0(\mathcal{K}) = C_0(\mathcal{K})$, odakle slijedi $m_0 = 7 = z_0$ i $b_{-1} = 0$. Za $k = 1$, matrica homomorfizma granice ∂_1 je data sa

$$A_1 = \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_9 & \sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Redukcijom na Smitovu normalnu formu dobija se:

$$A_1 \xrightarrow{R_1+R_2\to R_2} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_9 & \sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow[K_1+K_3\to K_3]{K_1+K_2\to K_2} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_8+\sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

$$\xrightarrow{R_2+R_6\to R_6} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2+\sigma_6 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_8+\sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow[K_2+K_4\to K_4]{K_2+K_3\to K_3} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2+\sigma_6 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_9+\sigma_{10} & \sigma_8+\sigma_9+\sigma_{11} & \sigma_8+\sigma_9+\sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

$$\xrightarrow{R_3\leftrightarrow R_6} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2+\sigma_6 \\ \sigma_6 \\ \sigma_4 \\ \sigma_5 \\ \sigma_3 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_9+\sigma_{10} & \sigma_8+\sigma_9+\sigma_{11} & \sigma_8+\sigma_9+\sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow{R_3+R_7\to R_7}$$

$$\begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2+\sigma_6 \\ \sigma_6+\sigma_7 \\ \sigma_4 \\ \sigma_5 \\ \sigma_3 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_9+\sigma_{10} & \sigma_8+\sigma_9+\sigma_{11} & \sigma_8+\sigma_9+\sigma_{12} & \sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \xrightarrow[K_3+K_6\to K_6]{K_3+K_5\to K_5} \begin{array}{c} \\ \sigma_1+\sigma_2 \\ \sigma_2+\sigma_6 \\ \sigma_6+\sigma_7 \\ \sigma_4 \\ \sigma_5 \\ \sigma_3 \\ \sigma_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \sigma_8 & \sigma_8+\sigma_9 & \sigma_9+\sigma_{10} & \sigma_8+\sigma_9+\sigma_{11} & \sigma_8+\sigma_{10}+\sigma_{12} & \sigma_9+\sigma_{10}+\sigma_{13} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Na osnovu toga, vrijedi $m_1 = 6$ *i* $z_1 = b_0 = 3$*. Pritom, grupa* $B_0(\mathcal{K})$ *ima bazu* $\{\sigma_1+\sigma_2, \sigma_2+\sigma_6, \sigma_6+\sigma_7\}$*, pa je* $H_0(\mathcal{K}) \cong \mathbb{Z}_2^4$ *i* $\beta_0 = 4$ *(postoje* 4 *komponente povezanosti). Takođe, uočava se da grupa* $Z_1(\mathcal{K})$ *ima bazu* $\{\sigma_8+\sigma_9+\sigma_{11}, \sigma_8+\sigma_{10}+\sigma_{12}, \sigma_9+\sigma_{10}+\sigma_{13}\}$*. Za* $k = 2$*, matrica homomorfizma granice* ∂_2 *je data sa*

$$A_2 = \begin{array}{c} \\ \sigma_8 \\ \sigma_9 \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{15} & \sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Redukcijom na Smitovu normalnu formu dobija se:

$$A_1 \xrightarrow[R_1+R_4\to R_4]{R_1+R_2\to R_2} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9 \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{15} & \sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow{K_1+K_2\to K_2} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9 \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

$$\xrightarrow[R_2+R_5\to R_5]{R_2+R_3\to R_3 \\ R_2+R_4\to R_4} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9+\sigma_{10}+\sigma_{11}+\sigma_{12} \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow{K_2+K_3\to K_3} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9+\sigma_{10}+\sigma_{11}+\sigma_{12} \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{14}+\sigma_{15}+\sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

$$\xrightarrow{R_4\leftrightarrow R_3} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9+\sigma_{10}+\sigma_{11}+\sigma_{12} \\ \sigma_{11} \\ \sigma_{10} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{14}+\sigma_{15}+\sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \xrightarrow[R_3+R_6\to R_6]{R_3+R_5\to R_5}$$

$$\begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9+\sigma_{10}+\sigma_{11}+\sigma_{12} \\ \sigma_{11}+\sigma_{12}+\sigma_{13} \\ \sigma_{10} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{14}+\sigma_{15}+\sigma_{16} & \sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \xrightarrow{K_3+K_4\to K_4} \begin{array}{c} \\ \sigma_8+\sigma_9+\sigma_{11} \\ \sigma_9+\sigma_{10}+\sigma_{11}+\sigma_{12} \\ \sigma_{11}+\sigma_{12}+\sigma_{13} \\ \sigma_{10} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_{14} & \sigma_{14}+\sigma_{15} & \sigma_{14}+\sigma_{15}+\sigma_{16} & \sigma_{14}+\sigma_{15}+\sigma_{16}+\sigma_{17} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Na osnovu toga, vrijedi $m_2 = 4$ i $b_1 = 3, z_2 = 1$. Pritom, grupa $B_1(\mathcal{K})$ ima bazu $\{\sigma_8 + \sigma_9 + \sigma_{11}, \sigma_9 + \sigma_{10} + \sigma_{11} + \sigma_{12}, \sigma_{11} + \sigma_{12} + \sigma_{13}\}$, pa je $H_1(\mathcal{K}) \cong \{0\}$ i $\beta_1 = 0$ (nema jednodimenzionalnih rupa). Takođe, uočava se da grupa $Z_2(\mathcal{K})$ ima bazu $\{\sigma_{14} + \sigma_{15} + \sigma_{16} + \sigma_{17}\}$, a to je ujedno i baza homološke grupe $H_2(\mathcal{K})$, jer je $b_2 = 0$. To znači da je $H_2(\mathcal{K}) \cong \mathbb{Z}_2$ i $\beta_2 = 1$ (postoji jedna dvodimenzionalna rupa).

Neka su $\mathcal{K} = (V_\mathcal{K}, \Sigma_\mathcal{K})$ i $\mathcal{L} = (V_\mathcal{L}, \Sigma_\mathcal{L})$ apstraktni simplicijalni kompleksi i $f : Vert(\mathcal{K}) \to Vert(\mathcal{L})$ simplicijalno preslikavanje. Za svaku dimenziju k, ovo preslikavanje generiše preslikavanje $f_k^\# : C_k(\mathcal{K}) \to C_k(\mathcal{L})$, koje je, za $k-$lanac $c = \sum a_i\sigma_i$, definisano sa $f_k^\#(c) := \sum a_i\tau_i$, gdje je

$$\tau_i := \begin{cases} f[\sigma_i], & \text{ako je } dim(f[\sigma_i]) = k; \\ 0, & \text{ako je } dim(f[\sigma_i]) < k. \end{cases}$$

Za lance $c_1 = \sum a_i\sigma_i$ i $c_1 = \sum b_i\sigma_i$ iz $C_k(\mathcal{K})$ vrijedi $f_k^\#(c_1 + c_2) = \sum(a_i + b_i)\tau_i = \sum a_i\tau_i + \sum b_i\tau_i = f_k^\#(c_1) + f_k^\#(c_2)$, što znači da je preslikavanje $f_k^\#$ homomorfizam, za svaku dimenziju k. Homomorfizmi $f_k^\#$ omogućavaju da se kompleksi lanaca pridruženi kompleksima $\mathcal{K}$ i $\mathcal{L}$ "spoje" na način kojim se čuva saglasnost pri djelovanju odgovarajućih homomorfizama granice. Preciznije, vrijedi sljedeća lema.

Lema 2.2.8. *[36] Neka su $\partial^{\mathcal{K}}$ i $\partial^{\mathcal{L}}$ homomorfizmi granice definisani redom na lancima kompleksa $\mathcal{K}$ i $\mathcal{L}$. Tada, za svaku dimenziju $k \geqslant 0$, vrijedi $f_k^{\#} \circ \partial_{k+1}^{\mathcal{K}} = \partial_{k+1}^{\mathcal{L}} \circ f_{k+1}^{\#}$, tj. sljedeći dijagram je komutativan*

$$\begin{array}{ccc} C_{k+1}(\mathcal{K}) & \xrightarrow{\partial_{k+1}^{\mathcal{K}}} & C_k(\mathcal{K}) \\ \downarrow{\scriptstyle f_{k+1}^{\#}} & & \downarrow{\scriptstyle f_k^{\#}} \\ C_{k+1}(\mathcal{L}) & \xrightarrow{\partial_{k+1}^{\mathcal{L}}} & C_k(\mathcal{L}) \end{array}$$

Dokaz. Tvrđenje je dovoljno dokazati na nivou simpleksa kompleksa $\mathcal{K}$, tj. dokazati da za svaku dimenziju k i proizvoljni $(k+1)-$simpleks σ vrijedi $(f_k^{\#} \circ \partial_{k+1}^{\mathcal{K}})(\sigma) = (\partial_{k+1}^{\mathcal{L}} \circ f_{k+1}^{\#})(\sigma)$. Stoga, neka je σ proizvoljan $(k+1)-$simpleks kompleksa $\mathcal{K}$, tj. neka je $\sigma = [v_0, v_1, \ldots, v_{k+1}]$, za neke vrhove $v_i \in Vert(\mathcal{K})$. Moguće je:

(i) $f[\sigma]$ je $(k+1)-$simpleks kompleksa $\mathcal{L}$;

Tada je $f_{k+1}^{\#}(\sigma) = f[\sigma]$, što implicira da za sve $i \neq j$ vrijedi $f(v_i) \neq f(v_j)$. Na osnovu toga, zaključuje se da se svaka $k-$strana simpleksa σ putem preslikavanja $f_k^{\#}$ nužno preslikava u jedinstvenu $k-$stranu simpleksa $f[\sigma]$. To znači da $f_k^{\#}\left(\sum_{j=0}^{k+1}[v_0, v_1, \ldots, \hat{v_j}, \ldots, v_{k+1}]\right)$ predstavlja sumu svih $k-$strana simpleksa $f[\sigma]$, odakle slijedi tražena jednakost.

(ii) $f[\sigma]$ nije $(k+1)-$simpleks kompleksa $\mathcal{L}$;

Tada postoje različiti vrhovi v_{i_0}, v_{j_0} simpleksa σ za koje vrijedi $f(v_{i_0}) = f(v_{j_0})$, pa je $f_{k+1}^{\#}(\sigma) = 0$, što implicira $(\partial_{k+1}^{\mathcal{L}} \circ f_{k+1}^{\#})(\sigma) = 0$. S druge strane, za sve simplekse τ koji su k-strane simpleksa σ i sadrže oba vrha v_{i_0}, v_{j_0} vrijedi $f_k^{\#}(\tau) = 0$, dok za preostale dvije $k-$strane τ_{i_0} i τ_{j_0} simpleksa σ (koje sadrže tačno jedan od vrhova v_{i_0}, v_{j_0}), čak i u slučaju da je $f_k^{\#}(\tau_{i_0}) \neq 0 \neq f_k^{\#}(\tau_{j_0})$, vrijedi $f[\tau_{i_0}] + f[\tau_{j_0}] = 0$. Zbog toga, $f_k^{\#}(\partial_{k+1}^{\mathcal{K}}(\sigma)) = 0$, pa jednakost vrijedi i u ovom slučaju. □

Rezultat iz prethodne leme omogućava da se simplicijalno preslikavanje produži na nivo homoloških grupa. Preciznije, vrijedi sljedeća lema.

Lema 2.2.9. *Neka su $\mathcal{K}$ i $\mathcal{L}$ simplicijalni kompleksi, $f : Vert(\mathcal{K}) \to Vert(\mathcal{L})$ simplicijalno preslikavanje i $f_k^{\#}$ produženje ovog preslikavanja na skup $C_k(\mathcal{K})$. Tada, za svako $k \geqslant 0$ vrijedi*

$(i) f_k^{\#}[Z_k(\mathcal{K})] \subseteq Z_k(\mathcal{L})$,

$(ii) f_k^{\#}[B_k(\mathcal{K})] \subseteq B_k(\mathcal{L})$,

(iii) Preslikavanje $f_k^* : H_k(\mathcal{K}) \to H_k(\mathcal{L})$ koje je za $c \in Z_k(\mathcal{K})$ definisano sa $f_k^*(c + B_k(\mathcal{K})) = f_k^{\#}(c) + B_k(\mathcal{L})$ je homomorfizam.

Dokaz. (i) Za proizvoljno $c_{\mathcal{L}} \in f_k^{\#}[Z_k(\mathcal{K})]$ postoji $c_{\mathcal{K}} \in Z_k(\mathcal{K})$ sa svojstvom $c_{\mathcal{L}} = f_k^{\#}(c_{\mathcal{K}})$, pa na osnovu prethodne leme vrijedi $\partial_k^{\mathcal{L}}(c_{\mathcal{L}}) = \partial_k^{\mathcal{L}}(f_k^{\#}(c_{\mathcal{K}})) = f_{k-1}^{\#}(\underbrace{\partial_k^{\mathcal{K}}(c_{\mathcal{K}})}_{=0}) = 0$, odakle slijedi $c_{\mathcal{L}} \in Z_k(\mathcal{L})$.

(ii) Za proizvoljno $c_{\mathcal{L}} \in f_k^{\#}[B_k(\mathcal{K})]$ postoji $c_{\mathcal{K}} \in B_k(\mathcal{K})$ sa svojstvom $c_{\mathcal{L}} = f_k^{\#}(c_{\mathcal{K}})$. Ako je $d_{\mathcal{K}} \in C_{k+1}(\mathcal{K})$ lanac takav da je $c_{\mathcal{K}} = \partial_{k+1}^{\mathcal{K}}(d_{\mathcal{K}})$, tada se na osnovu prethodne leme dobija $c_{\mathcal{L}} = f_k^{\#}(c_{\mathcal{K}}) = f_k^{\#}(\partial_{k+1}^{\mathcal{K}}(d_{\mathcal{K}})) = \partial_{k+1}^{\mathcal{L}}(\underbrace{f_{k+1}^{\#}(d_{\mathcal{K}})}_{\in C_{k+1}(\mathcal{L})})$, odakle slijedi $c_{\mathcal{L}} \in B_k(\mathcal{L})$.

(iii) Prije svega, na osnovu prethodnog, za svako $c \in Z_k(\mathcal{K})$ vrijedi $f_k^{\#}(c) \in Z_k(\mathcal{L})$, što implicira $f_k^*(c + B_k(\mathcal{K})) \in H_k(\mathcal{L})$. Takođe, definicija preslikavanja f_k^* ne zavisi od izbora predstavnika homološke klase $H_k(\mathcal{K})$, jer za $c_1 + B_k(\mathcal{K}) = c_2 + B_k(\mathcal{K})$ vrijedi $c_1 - c_2 \in B_k(\mathcal{K})$, odakle slijedi $f_k^{\#}(c_1) - f_k^{\#}(c_2) = f_k^{\#}(c_1 - c_2) \in f_k^{\#}(B_k(\mathcal{K})) \subseteq B_k(\mathcal{L})$, što implicira $f_k^{\#}(c_2) + B_k(\mathcal{L}) = f_k^{\#}(c_2) + \Big(f_k^{\#}(c_1) - f_k^{\#}(c_2) + B_k(\mathcal{L})\Big) = f_k^{\#}(c_1) + B_k(\mathcal{L})$, te posljedično vrijedi $f_k^*(c_1 + B_k(\mathcal{K})) = f_k^*(c_2 + \mathcal{B}_k(\mathcal{K}))$. Iz date definicije i činjenice da je $f_k^{\#}$ homomorfizam direktno slijedi da je preslikavanje f_k^* takođe homomorfizam. □

Preslikavanje f^* iz prethodne leme naziva se *homomorfizmom indukovanim simplicijalnim preslikavanjem* f. Uočava se da je rank slike ovog preslikavanja ograničen odozgo sa oba Betijeva broja, tj. da vrijedi $rank(f_k^*(H_k(\mathcal{K}))) \leqslant \min\{\beta_k(\mathcal{K}), \beta_k(\mathcal{L})\}$. Najjednostavniji primjer indukovanog homomorfizma f^* dobija se kada je simplicijalno preslikavanje f inkluzija, tj. u slučaju kada je kompleks $\mathcal{K}$ potkompleks kompleksa $\mathcal{L}$.

2.3 Istrajna homologija

Konačnom podskupu K nekog metričkog prostora moguće je pridružiti Vietoris-Rips-ov kompleks $\mathcal{V}R_K^{(r)}$ ili Čehov kompleks $\mathcal{C}_K^{(r)}$, za neku vrijednost $r \geqslant 0$. Na taj način, povezanost koja eventualno postoji između elemenata tog skupa tumači se preko svojstava dobijenih kompleksa, tj. svojstava njegovih homoloških grupa. Jasno da izbor parametra r u velikoj mjeri određuje prirodu pomenutih svojstava. Tako se za dovoljno malu vrijednost r dobija kompleks sastavljen od 0-simpleksa, dok se za dovoljno veliku vrijednost r dobija standardni kombinatorni kompleks. Da li postoji optimalan izbor parametra r koji najbolje oslikava topološke osobine datog skupa podataka? Na ovo pitanje se može odgovoriti ukoliko se, umjesto jednog, posmatra familija kompleksa i prati kako se mijenjaju svojstva njima pripadnih homoloških grupa. Ovo je predmet

proučavanja *istrajne (perzistentne) homologije*. Istrajna homologija je moćan alat za ispitivanje topoloških karakteristika skupa podataka pri različitim prostornim rezolucijama. Ona se zasniva na višeslojnoj analizi topoloških osobina podataka u cilju otklanjanja mogućih grešaka u njihovoj interpretaciji do kojih dolazi zbog uzorkovanja, grešaka u prikupljanju ulaznih podataka ili pristrasnog izbora određenih parametara.

Pod *filtracijom simplicijalnog kompleksa* $\mathcal{K}$ podrazumijeva se kolekcija $\{\mathcal{K}^{(i)} : i \in \{0, 1, \ldots, t\}\}$ "rastućih" potkompleksa ovog kompleksa:

$$\emptyset = \mathcal{K}^{(0)} \subseteq \mathcal{K}^{(1)} \subseteq \cdots \subseteq \mathcal{K}^{(t)} = \mathcal{K}.$$

Ukoliko kompleks $\mathcal{K}$ ima ukupno u simpleksa, tada se njegova filtracija može shvatiti kao postepena "konstrukcija" izvedena u $t \leqslant u$ etapa, tako da se u svakoj od njih pridoda određen skup "novih" simpleksa iz $\mathcal{K}$, tj. onih simpleksa iz $\mathcal{K}$ koji nisu u sastavu nijednog od kompleksa ustanovljenih do te etape. Indeksi koji se koriste za enumeraciju kompleksa iz filtracije nazivaju se *nivoima filtracije*. Nivoi filtracije ne moraju obavezno biti elementi skupa $\{0, 1, \ldots, t\}$. Umjesto ovog skupa može se koristiti bilo koji skup $\{r_0, r_1, \ldots, r_t\} \subseteq \mathbb{R}$, za čije elemente vrijedi $r_0 < r_1 < \ldots < r_t$.

Za dati Čehov kompleks $C_K^{(r)}$, svaka kolekcija $\{r_i : i \in \{0, 1, \ldots, t\}\} \subseteq \mathbb{R}$ takva da je $r_0 < 0 \leqslant r_1 < r_2 < \cdots < r_t = r$ određuje filtraciju

$$\emptyset = C_K^{(r_0)} \subseteq C_K^{(r_1)} \subseteq \cdots \subseteq C_K^{(r_t)} = C_K^{(r)}.$$

Čehov kompleks $C_K^{(r_i)}$ predstavlja "stanje" kompleksa $C_K^{(r)}$ na nivou filtracije $r_i \leqslant r$. Važno je istaći da se nivoi filtracije mogu izabrati tako da svaka etapa konstrukcije ima tačno jednog predstavnika. Preciznije, za svako $r' > 0$ postoji jedinstvena vrijednost $r_i \leq r'$ sa svojstvom $C_K^{(r_i)} = C_K^{(r')}$. U tom slučaju, ovako dobijena filtracija naziva se *Čehovom filtracijom*. Na sličan način se može uvesti pojam *Vietoris-Ripsove filtracije*.

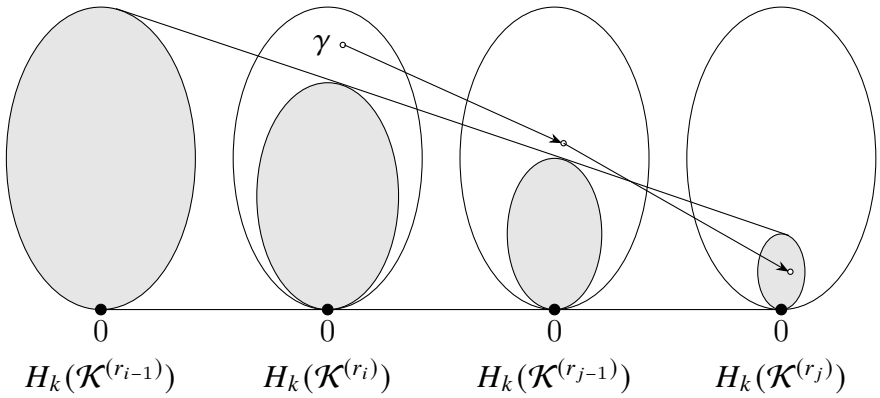
Važnije od registrovanja samih kompleksa u filtraciji jeste praćenje razvojnog puta homoloških klasa prilikom prolaska kroz nivoe filtracije. Pod ovim se podrazumijeva konstatovanje kompletne evolucije svake homološke klase, od nivoa filtracije kada se ova klasa prvi put pojavljuje do (eventualnog) nivoa filtracije u kojem ova klasa postaje trivijalna. Slikovito rečeno, za svaku dimenziju $k \geqslant 0$, pratiće se period istrajnosti svake $k-$dimenzionalne rupe koja obitava u posmatranoj filtraciji. Slijedi precizniji opis ovog postupka.

Neka je $\emptyset = \mathcal{K}^{(r_0)} \subseteq \mathcal{K}^{(r_1)} \subseteq \cdots \subseteq \mathcal{K}^{(r_t)} = \mathcal{K}$ filtracija simplicijalnog kompleksa $\mathcal{K}$. Ako su i, j nivoi date filtracije, pri čemu je $i \leqslant j$, tada, za svaku dimenziju k, inkluzija $\mathcal{K}^{(r_i)} \hookrightarrow \mathcal{K}^{(r_j)}$ generiše indukovani homomorfizam $f_k^{i,j} : H_k(\mathcal{K}^{(r_i)}) \to H_k(\mathcal{K}^{(r_j)})$. Na taj način, za svaku dimenziju k dobija se lanac

homoloških grupa povezanih odgovarajućim indukovanim homomorfizmima:

$$0 = H_k(\mathcal{K}^{(r_0)}) \overset{f_k^{r_0,r_1}}{\to} H_k(\mathcal{K}^{(r_1)}) \overset{f_k^{r_1,r_2}}{\to} \ldots \overset{f_k^{r_{t-1},r_t}}{\to} H_k(\mathcal{K}^{(r_t)}) = H_k(\mathcal{K}). \quad (2.4)$$

Za $0 \leqslant r_i \leqslant r_j \leqslant r_t$, *$k-$ta istrajna homološka grupa* $H_k^{r_i,r_j}$ definiše se kao slika homološke grupe $H_k(\mathcal{K}^{(r_i)})$ u odnosu na indukovani homomorfizam $f_k^{r_i,r_j}$, tj. ova grupa je data sa $H_k^{r_i,r_j} := f_k^{r_i,r_j}(H_k(\mathcal{K}^{(r_i)}))$. Rank ove grupe naziva se *$k-$ti istrajni Betijev broj* i označava se sa $\beta_k^{r_i,r_j}$. Primjećuje se da je $H_k^{r_i,r_i} = H_k(\mathcal{K}^{(r_i)})$ i $\beta_k^{r_i,r_i} = \beta_k(\mathcal{K}^{(r_i)})$. Elementi istrajne homološke grupe $H_k^{r_i,r_j}$ su homološke klase iz $H_k(\mathcal{K}^{(r_i)})$ koje su "istrajale" u procesu prolaska od nivoa r_i do nivoa r_j filtracije u smislu da su ove klase i dalje prisutne kao elementi grupe $H_k(\mathcal{K}^{(r_j)})$. Formalnije, ovo znači da $\gamma \in H_k^{r_i,r_j}$ ako i samo ako $\gamma \in Z_k(\mathcal{K}^{(r_i)}) / \Big(B_k(\mathcal{K}^{(r_j)}) \cap Z_k(\mathcal{K}^{(r_i)})\Big)$. Za homološku klasu γ iz $H_k(\mathcal{K}^{(r_i)})$ se kaže da je *rođena* (*nastala*) u kompleksu $\mathcal{K}^{(r_i)}$, ako $\gamma \notin H_k^{r_{i-1},r_i}$. Ako je γ homološka klasa rođena u kompleksu $\mathcal{K}^{(r_i)}$, tada ova klasa *umire* (*nestaje*) ulaskom u kompleks $\mathcal{K}^{(r_j)}$, ako se prilikom prelaska sa nivoa r_{j-1} na nivo r_j posmatrane filtracije ona spaja sa "starijom" homološkom klasom, tj. ako vrijedi $f_k^{r_i,r_{j-1}}(\gamma) \notin H_k^{r_{i-1},r_{j-1}}$ i $f_k^{r_i,r_j}(\gamma) \in H_k^{r_{i-1},r_j}$ (Slika 2.3).



Slika 2.3: Homološka klasa γ je rođena u kompleksu $\mathcal{K}^{(r_i)}$, jer ne pripada slici homološke grupe $H_k(\mathcal{K}^{(r_{i-1})})$ (osjenčeni dio). Dalje, ova klasa umire ulaskom u kompleks $\mathcal{K}^{(r_j)}$, jer je r_j najmanji nivo filtracije za koji ona pripada slici homološke grupe $H_k(\mathcal{K}^{(r_{i-1})})$.

Ako homološka klasa nastaje u kompleksu $\mathcal{K}^{(r_i)}$, a nestaje u kompleksu $\mathcal{K}^{(r_j)}$, tada se interval $[r_i, r_j)$ naziva *intervalom istrajnosti* te homološke klase. Ukoliko homološka klasa nastaje u kompleksu $\mathcal{K}^{(r_i)}$ i ostaje istrajna zaključno sa posljednjim kompleksom u filtraciji, tada se uzima da je njen interval istrajnosti $[r_i, +\infty)$. Simpleks koji ima svojstvo da njegovo dodavanje u nekom nivou filtracije dovodi do rađanja nove homološke klase se naziva *pozitivnim simpleksom*,

dok simpleks koji ima svojstvo da njegovo dodavanje u nekom nivou filtracije dovodi do smrti postojeće homološke klase se naziva *negativnim simpleksom*. Filtracija koja ima svojstvo da u svakom njenom nivou je dozvoljeno stvaranje najviše jedne nove homološke klase ili uništavanje najviše jedne postojeće homološke klase naziva se *Morzeovom filtracijom*. Ovo efektivno znači da, u Morzeovoj filtraciji, svi ograničeni intervali istrajnosti imaju različite krajnje tačke.

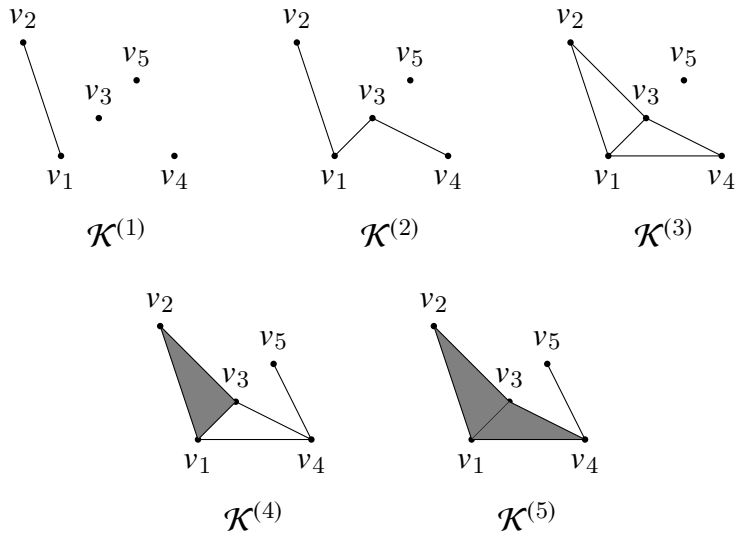
Primjer 2.3.1. *Za kompleks* $\mathcal{K} = (V, \Sigma)$*, gdje je* $V = \{v_1, v_2, v_3, v_4, v_5\}$ *i*

$$\Sigma = \{[v_1], [v_2], [v_3], [v_4], [v_5], [v_1, v_2], [v_1, v_3], [v_3, v_4], [v_1, v_4], [v_2, v_3], [v_4, v_5], [v_1, v_2, v_3], [v_1, v_3, v_4]\},$$

jedna filtracija je kolekcija $\left\{\mathcal{K}^{(j)} = (V, \Sigma^{(j)}) : j \in \{1, 2, 3, 4, 5\}\right\}$*, pri čemu je*

$$\begin{aligned}
\Sigma^{(1)} &= \{[v_1], [v_2], [v_3], [v_4], [v_5], [v_1, v_2]\}, \\
\Sigma^{(2)} &= \Sigma^{(1)} \cup \{[v_1, v_3], [v_3, v_4]\}, \\
\Sigma^{(3)} &= \Sigma^{(2)} \cup \{[v_2, v_3], [v_1, v_4]\}, \\
\Sigma^{(4)} &= \Sigma^{(3)} \cup \{[v_4, v_5], [v_1, v_2, v_3]\}, \\
\Sigma^{(5)} &= \Sigma^{(4)} \cup \{[v_1, v_3, v_4]\} = \Sigma.
\end{aligned}$$

Vizuelno, prethodna filtracija može da se prikaže kao na sljedećoj slici:



Slika 2.4

U kompleksu $\mathcal{K}^{(3)}$ *nastaju dvije jednodimenzionalne homološke klase:* $\gamma_1 = [v_1, v_2] + [v_2, v_3] + [v_1, v_3]$ *i* $\gamma_2 = [v_1, v_3] + [v_3, v_4] + [v_1, v_4]$*. U kompleksu*

$\mathcal{K}^{(4)}$ nestaje homološka klasa γ_1, a u kompleksu $\mathcal{K}^{(5)}$ nestaje homološka klasa γ_2, pa je $[3,4)$ interval istrajnosti za γ_1, a $[3,5)$ interval istrajnosti za γ_2. Simpleksi $[v_2,v_3]$ i $[v_1,v_4]$ su pozitivni simpleksi, jer se njihovim dodavanjem u nivou filtracije 3 rađaju redom homološke klase γ_1 i γ_2. Simpleksi $[v_1,v_2,v_3]$ i $[v_1,v_3,v_4]$ su negativni simpleksi, jer njihovo dodavanje u kompleksima $\mathcal{K}^{(4)}$ i $\mathcal{K}^{(5)}$ dovodi do umiranja homoloških klasa γ_1 i γ_2 respektivno. Data filtracija nije Morzeova, jer se npr. u kompleksu $\mathcal{K}^{(3)}$ pojavljuju dva pozitivna simpleksa.

Za izračunavanje istrajnih homoloških grupa i istrajnih Betijevih brojeva može se koristiti modifikacija postupka redukcije matrica homomorfizama granice, opisanog u slučaju simplicijalne homologije. Prije svega, informacije o svim homomorfizmima granice biće sadržani u jednoj matrici, umjesto da se za svaku dimenziju pojedinačno posmatra po jedna matrica. Ovo zahtjeva enumeraciju skupa simpleksa kompleksa čija se filtracija posmatra. Najprirodnije je posmatrati *kompatibilni poredak*, koji je saglasan sa datom filtracijom. Preciznije, ako je $\{\mathcal{K}^{(j)} = (V, \Sigma^{(j)}) : j \in \{1,2,\ldots,t\}\}$ filtracija kompleksa $\mathcal{K} = (V,\Sigma)$ koji ukupno ima m simpleksa, tada kompatibilni poredak podrazumijeva enumeraciju $\sigma_1, \sigma_2, \ldots, \sigma_m$ njegovih simpleksa, pri čemu, za svako $j \in \{1,2,\ldots,t\}$ vrijedi $\Sigma^{(j)} = \left\{\sigma_i : i \in \{1,\ldots,|\Sigma^{(j)}|\}\right\}$. Dakle, simpleksi sa manjim indeksima pripadaju kompleksima koji se pojavljuju na manjim nivoima filtracije, pri čemu se vodi računa da se simpleksu može dodijeliti naredni slobodan indeks i tek onda kada su svim simpleksima koji čine njegove strane već pridruženi indeksi manji od i.

Primjer 2.3.2. *Za filtraciju kompleksa iz prethodnog primjera, kompatibilni poredak je dat sa*

$$\Sigma = \{\underbrace{[v_1]}_{\sigma_1}, \underbrace{[v_2]}_{\sigma_2}, \underbrace{[v_3]}_{\sigma_3}, \underbrace{[v_4]}_{\sigma_4}, \underbrace{[v_5]}_{\sigma_5}, \underbrace{[v_1,v_2]}_{\sigma_6}, \underbrace{[v_1,v_3]}_{\sigma_7}, \underbrace{[v_3,v_4]}_{\sigma_8}, \underbrace{[v_1,v_4]}_{\sigma_9},$$
$$\underbrace{[v_2,v_3]}_{\sigma_{10}}, \underbrace{[v_4,v_5]}_{\sigma_{11}}, \underbrace{[v_1,v_2,v_3]}_{\sigma_{12}}, \underbrace{[v_1,v_3,v_4]}_{\sigma_{13}}\}.$$

Neka su $\sigma_1, \sigma_2, \ldots, \sigma_m$ simpleksi kompleksa $\mathcal{K}$ indeksirani u skladu sa kompatibilnim poretkom u odnosu na datu filtraciju ovog kompleksa. Matrica svih homomorfizama granice je kvadratna matrica $A = [a_{ij}]$ formata $m \times m$, pri čemu su njeni elementi dati sa:

$$a_{ij} := \begin{cases} 1, & \text{ako je simpleks } \sigma_i \text{ strana simpleksa } \sigma_j \text{ i } dim(\sigma_j) = dim(\sigma_i) + 1; \\ 0, & \text{inače.} \end{cases}$$

Znači, redovi i kolone matrice A su uređeni u skladu sa kompatibilnim poretkom i granica nekog simpleksa je određena kolonom u kojoj se on nalazi. Ako u $j-$toj koloni postoji jedinica, neka $low(j)$ označava indeks reda $j-$te kolone

u kojem se nalazi najniža jedinica, tj. indeks reda $j-$te kolone takav da svi redovi te kolone niži od njega sadrže isključivo nule. U suprotnom, definiše se $low(j) := 0$. Redukcija matrice A biće izvedena sa ciljem da se dobija matrica u kojoj za proizvoljne dvije različite nenula kolone j_1, j_2 vrijedi $low(j_1) \neq low(j_2)$. Ovo se može ostvariti tako što se za svaku kolonu j_1, ovoj koloni sukcesivno dodaju sve kolone $j < j_1$ koje imaju svojstvo $low(j) = low(j_1)$, ako takve postoje. Algoritam kojim se izvodi redukcija ima sljedeći oblik:

$R = A$
for $j = 1$ to m
  while $\big($postoji $j_0 < j$ sa svojstvom $low(j_0) = low(j) > 0\big)$
    dodati kolonu j_0 koloni j

Dodavanje kolone j_0 smanjuje vrijednost $low(j)$, što implicira da se algoritam izvršava nakon najviše m^2 dodavanja kolona.

Ako je R matrica koja se dobije ovom redukcijom, tada se iz ove matrice mogu "pročitati" homološke grupe kompleksa $\mathcal{K}$, kao i odgovarajući Betijevi brojevi. Naime, za dimenziju k, broj kolona matrice R koje odgovaraju $k-$simpleksima, a koje sadrže isključivo nule, jednak je $rank(Z_k(\mathcal{K}))$, dok je broj redova koji odgovaraju $k-$simpleksima, a sadrže najniže jedinice, jednak $rank(B_k(\mathcal{K}))$, te je razlika ova dva broja Betijev broj $\beta_k(\mathcal{K})$.

Ispostavlja se da redukovana matrica R sadrži informaciju i o intervalima istrajnosti homoloških klasa u odnosu na datu filtraciju kompleksa $\mathcal{K}$. Ovo postaje jasno kada se ustanovi veza između "najnižih jedinica" ove matrice i istrajnih homoloških grupa. Za početak, biće pokazano da pozicije najnižih jedinica u redukovanoj matrici R ne zavise od ove matrice, tj. od opisanog algoritma redukovanja.

Neka je $B = [b_{ij}]$ kvadratna matrica formata $m \times m$. Za $i, j \in \{1, 2, \ldots, m\}$, neka je B_i^j donja podmatrica matrice B kod koje je u gornjem desnom uglu element b_{ij}, tj. neka je B_i^j matrica koja se dobija iz matrice B kada se izostave njenih prvih $i-1$ redova i posljednjih $m-j$ kolona. Za $i, j \in \{1, 2, \ldots, m\}$, neka je $r_B(i, j) := rank(B_i^j) - rank(B_{i+1}^j) + rank(B_{i+1}^{j-1}) - rank(B_i^{j-1})$. Sljedeća lema poznata je kao *Lema uparivanja simpleksa*.

Lema 2.3.3. *[36] Neka su $\sigma_1, \sigma_2, \ldots, \sigma_m$ simpleksi kompleksa $\mathcal{K}$ indeksirani u skladu sa kompatibilnim poretkom u odnosu na datu filtraciju ovog kompleksa i A odgovarajuća kvadratna matrica svih homomorfizama granice. Ako je R matrica dobijena redukcijom matrice A, tada za sve $i \in \{1, 2, \ldots, m\}$ i sve nenula kolone $j \in \{1, 2, \ldots, m\}$ matrice R vrijedi $i = low(j)$ ako i samo ako je $r_R(i, j) = 1$. Specijalno, uparivanje redova i kolona određeno pozicijama najnižih jedinica ne zavisi od izbora matrice R.*

Dokaz.

Neka je j proizvoljna nenula kolona matrice R, tj. neka vrijedi $low(j) > 0$.

Ukoliko je $i = low(j)$, slijedi da posljednja kolona matrice R_i^j nije nula kolona. Zbog toga matrica R_i^j ima jednu nenula kolonu više od matrice R_{i+1}^j, jer je posljednja kolona matrice R_{i+1}^j sastavljena od nula. Također, to implicira da matrice R_{i+1}^j i R_{i+1}^{j-1} imaju jednak broj nenula kolona, kao i da matrica R_i^{j-1} ima jednu nenula kolonu manje od matrice R_i^j. Uzimajući sve ovo u obzir, zaključuje se da je $r_R(i,j) = rank(R_i^j) - rank(R_{i+1}^j) + rank(R_{i+1}^{j-1}) - rank(R_i^{j-1}) = 1$.

Neka je sada $i \neq low(j)$, tj. neka r_{ij} nije najniža jedinica $j-$te kolone matrice R. Moguće je da se najniža jedinica pojavljuje u i-tom redu za neku od prvih $j-1$ kolona matrice R ili da se ne pojavljuje u ovom redu.

(i) Prvih $j-1$ kolona matrice R ne sadrže najnižu jedinicu u redu i.

U ovom slučaju, matrice R_i^j i R_{i+1}^j imaju jednak broj nenula kolona, a isti zaključak vrijedi i za matrice R_{i+1}^{j-1} i R_i^{j-1}.

(ii) Postoji (jedinstvena) kolona $j_0 \in \{1, 2, \ldots, j-1\}$ matrice R takva da je $i = low(j_0)$.

U ovom slučaju, j_0-ta kolona matrice R_i^j je nenula kolona, dok je za matricu R_{i+1}^j ovo nula kolona, što implicira da matrica R_i^j ima jednu nenula kolonu više nego matrica R_{i+1}^j. Slično se zaključuje da matrica R_i^{j-1} ima jednu nenula kolonu više od matrice R_{i+1}^{j-1}.

U oba razmotrena slučaja, odgovarajući sabirci koji čine izraz $r_R(i,j)$ se anuliraju, pa je $r_R(i,j) = 0$.

S obzirom da za proizvoljne $i, j \in \{1, 2, \ldots, m\}$ dodavanje kolona s lijeva na desno ne mijenja rank matrice A_i^j, zaključuje se da za sve $i, j \in \{1, 2, \ldots, m\}$ vrijedi $rank(R_i^j) = rank(A_i^j)$. Zbog toga, za sve $i, j \in \{1, 2, \ldots, m\}$ vrijedi $r_A(i,j) = r_R(i,j)$, pa dobijena karakterizacija najnižih jedinica zavisi isključivo od elemenata matrice A. □

Posmatrajući opisani algoritam redukcije može se primijetiti da $j-$ta kolona redukovane matrice R poprima svoj finalni oblik nakon $j-$te iteracije for petlje. To znači da se u ovoj iteraciji dobija redukovana matrica za kompleks koga sačinjavaju prvih j simpleksa u odnosu na kompatibilni poredak. Ako je $low(j) = i > 0$, tada je σ_j negativan simpleks uparen sa pozitivnim simpleksom σ_i. S druge strane, ako je $low(j) = 0$, tada je σ_j pozitivan simpleks i u tom slučaju potrebno je provjeriti da li u $j-$tom redu postoji najniža jedinica. Ukoliko je $low(k) = j$, tada je σ_k negativan simpleks uparen sa σ_j, dok slučaj da u $j-$tom redu ne postoji najniža jedinica znači da simpleks σ_j nije uparen sa negativnim simpleksom, tj. homološka klasa koju on stvara postoji i u završnom kompleksu filtracije.

Primjer 2.3.4. *Za filtraciju datu u prethodna dva primjera matrica A svih*

homomorfizama granice data je sa

$$A= \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \\ \sigma_8 \\ \sigma_9 \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{ccccccccccccc} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 & \sigma_6 & \sigma_7 & \sigma_8 & \sigma_9 & \sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \left[\begin{array}{ccccccccccccc} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

Primjenom opisanog algoritma redukcije dobija se matrica

$$R= \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \\ \sigma_7 \\ \sigma_8 \\ \sigma_9 \\ \sigma_{10} \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{13} \end{array} \begin{array}{c} \begin{array}{ccccccccccccc} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 & \sigma_6 & \sigma_7 & \sigma_8 & \sigma_9 & \sigma_{10} & \sigma_{11} & \sigma_{12} & \sigma_{13} \end{array} \\ \left[\begin{array}{ccccccccccccc} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

pri čemu uokvireni elementi predstavljaju najniže jedinice. Kolone koje odgovaraju simpleksima $\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$ su nula kolone, pa su to pozitivni simpleksi. Simpleks σ_1 nije uparen sa negativnim simpleksom, te $0-$homološka klasa η_1 koju on stvara postoji i u kompleksu $\mathcal{K}^{(5)}$, te je interval istrajnosti ove klase $[1, +\infty)$. Simpleks σ_2 je uparen sa negativnim simpleksom σ_6 i oba ova simpleksa se dodaju u kompleksu $\mathcal{K}^{(1)}$, što znači da se stvorena homološka klasa odmah uništava. Simpleks σ_3 je uparen sa negativnim simpleksom σ_7 i ovaj par opisuje $0-$homološku klasu η_2, čiji je interval istrajnosti $[1, 2)$. Simpleks σ_4 je uparen sa negativnim simpleksom σ_8 i ovaj par opisuje $0-$homološku klasu η_3, čiji je interval istrajnosti $[1, 2)$. Simpleks σ_5 je uparen sa negativnim simpleksom σ_{11} i ovaj par opisuje $0-$homološku klasu η_4, čiji je interval istrajnosti $[1, 4)$. $1-$simpleksi σ_9 i σ_{10} su pozitivni simpleksi, koji su upareni redom sa

negativnim simpleksima σ_{13} *i* σ_{12}. *Odgovarajuće* $1-$*homološke klase* γ_1 *i* γ_2 *imaju redom intervale istrajnosti* $[3,5)$ *i* $[3,4)$.

Opisani postupak redukcije omogućava da se posmatranom lancu homoloških grupa pridruži multiskup intervala istrajnosti, za svaku posmatranu dimenziju. Ispostavlja se da ovako dobijena kolekcija multiskupova u potpunosti određuje dati lanac homoloških grupa. U narednoj sekciji su uvedeni potrebni pojmovi i tvrđenja uz pomoć kojih će ovo biti dokazano.

2.4 Istrajni moduli i udaljenost preplitanja

U ovoj sekciji se razmatra matematička struktura istrajnih modula. Iako su neka od tvrđenja dokazana u opštem slučaju, akcenat je na svojstvima koja vrijede za specijalan slučaj ove strukture, a koji obuhvata istrajni modul homologije određen formulom (2.4). Kroz tu prizmu treba posmatrati sav materijal koji je obuhvaćen u ovoj, a i u narednim sekcijama.

Istrajni modul M čini familija konačno-dimenzionalnih vektorskih prostora $\{M_t : t \in \mathbb{R}\}$ definisanih nad istim poljem $\mathbb{F}$ i familija linearnih preslikavanja $\{\phi_M(s,t) : M_s \to M_t | \, s \leqslant t\}$ takva da za sve realne brojeve r, s, t za koje je $r \leqslant s \leqslant t$ vrijede svojstva

$$\begin{aligned} \phi_M(t,t) &= id_{M_t}, \\ \phi_M(r,t) &= \phi_M(s,t) \circ \phi_M(r,s). \end{aligned}$$

Preslikavanja $\phi_M(s,t)$ se nazivaju *tranzicionim preslikavanjima* ili *internim morfizmima* i ona povezuju vektorske prostore modula M. Na taj način, svaki modul M se može interpretirati kao lanac vektorskih prostora prikazan na sljedećem dijagramu.

$$\ldots \longrightarrow M_r \xrightarrow{\phi_M(r,s)} M_s \xrightarrow{\phi_M(s,t)} M_t \longrightarrow \ldots$$

Primjer 2.4.1. *Najjednostavniji oblik istrajnog modula je intervalni istrajni modul. Za interval* $I \subseteq \mathbb{R}$ *(koji može biti i prazan skup) intervalni istrajni modul* $M(I)$ *sastavljen je od vektorskih prostora* $M(I)_t$ *i tranzicionih preslikavanja* $\phi_{M(I)}(s,t)$*, pri čemu, za sve* $t \in \mathbb{R}$ *vrijedi*

$$M(I)_t := \begin{cases} \mathbb{F}, & \text{ako } t \in I; \\ \{0_{\mathbb{F}}\}, & \text{inače,} \end{cases}$$

a za sve $s, t \in \mathbb{R}$*, pri čemu je* $s \leqslant t$*, vrijedi*

$$\phi_{M(I)}(s,t) := \begin{cases} id_{\mathbb{F}}, & \text{ako } s,t \in I; \\ 0, & \text{inače.} \end{cases}$$

Specijalno za $I = \emptyset$*, intervalni istrajni modul* $M(\emptyset)$ *se naziva nula (istrajnim) modulom.*

Primjedba 2.4.2. U narednom će uglavnom biti razmatrani intervalni istrajni moduli indukovani intervalima oblika $[b, d)$, gdje je $b \in \mathbb{R}$, a $d \in \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. Razlog tome je što se intervali ovakvog oblika pojavljuju kao linije bar-koda istrajne homologije. Precizniji opis bar-koda biće izložen nešto kasnije.

Za realne brojeve $r_1 < r_2 < \ldots < r_n$, neka je $\mathcal{K}^{(r_1)} \subseteq \mathcal{K}^{(r_2)} \subseteq \ldots \subseteq \mathcal{K}^{(r_n)} = \mathcal{K}$ filtracija kompleksa $\mathcal{K}$ i $H_k\left(\mathcal{K}^{(r_m)}\right)$ homološka grupa dimenzije $k \geqslant 0$ nad poljem $\mathbb{Z}_2$ koja odgovara kompleksu $\mathcal{K}^{(r_m)}$. Ako je $f_k^{r_{m_i}, r_{m_j}} : H_k\left(\mathcal{K}^{(r_{m_i})}\right) \to H_k\left(\mathcal{K}^{(r_{m_j})}\right)$, $m_i \leqslant m_j$, homomorfizam indukovan inkluzijom $\mathcal{K}^{(r_{m_i})} \hookrightarrow \mathcal{K}^{(r_{m_j})}$, tada je *istrajni modul homologije* $M\,(Hom_k)$ pridružen datoj filtraciji određen vektorskim prostorima

$$M\,(Hom_k)_t = \begin{cases} \{0\}, & \text{ako je } t < r_1; \\ H_k\left(\mathcal{K}^{(r_m)}\right), & \text{ako je } r_m \leqslant t < r_{m+1}, \text{ za } m \in \{1, 2, \ldots, n-1\}; \\ H_k\left(\mathcal{K}^{(r_n)}\right), & \text{ako je } t \geqslant r_n, \end{cases}$$

i tranzicionim preslikavanjima koja su, za $s \leqslant t$, data sa

$$\phi_{M(Hom_k)}(s,t) = \begin{cases} 0, & \text{ako je } s < r_1; \\ f_k^{r_{m_i}, r_{m_j}}, & \text{ako je } r_{m_i} \leqslant s < r_{m_i+1}, r_{m_j} \leqslant t < r_{m_j+1}; \\ f_k^{r_n, r_n}, & \text{ako je } r_n \leqslant s. \end{cases}$$

Istrajni modul homologije $M\,(Hom_k)$ "kodira" informacije o evoluciji $k-$dimenzionalnih homoloških klasa prilikom prolaska kroz filtraciju. Primjećuje se da ovaj istrajni modul zadovoljava i neka dodatna svojstva u odnosu na ona data u definiciji istrajnog modula. Konkretno, za svako $t < r_1$ vrijedi $M\,(Hom_k)_t = \{0\}$ i za svako $t \geqslant r_n$ vrijedi $M\,(Hom_k)_t = H_k\left(\mathcal{K}^{(r_n)}\right)$. Pored toga, za svaku tačku iz skupa $\mathbb{R} \setminus \{r_1, r_2, \ldots, r_n\}$ postoji otvoren skup $U \subseteq \mathbb{R}$ takva da za sve tačke $s, r \in U$ za koje je $s \leqslant r$ vrijedi da je tranziciono preslikavanje $\phi_{M(Hom_k)}(s,r)$ izomorfizam.

Primjedba 2.4.3. Uočava se da istrajni modul homologije $M\,(Hom_k)$ zadovoljava dodatna svojstva "konačnog tipa". Preciznije, za istrajni modul M se kaže da ima *svojstvo konačnog tipa*, ako postoji konačan skup $A \subseteq \mathbb{R}$ tako da za svako $t < \min A$ vrijedi $M_t = \{0\}$ i dodatno je ispunjeno

- Za svako $a \in A$ postoji $\varepsilon > 0$ tako da je tranziciono preslikavanje $\phi_M(a,t)$ izomorfizam ako $t \in [a, a+\varepsilon)$, dok $\phi_M(s,a)$ nije izomorfizam ako $s \in (a-\varepsilon, a)$.

- Za svako $x \in \mathbb{R} \setminus A$ postoji $\varepsilon > 0$ tako da je tranziciono preslikavanje $\phi_M(s,t)$ izomorfizam za sve $s \leqslant t$ koji pripadaju intervalu $(x-\varepsilon, x+\varepsilon)$.

Tačke skupa A iz prethodnog nazivaju se *spektralnim tačkama modula* M, a sam skup A *spektrom istrajnog modula*. Spektar istrajnog modula M se još označava sa $Spec(M)$. Ukratko rečeno, istrajni modul M ispunjava svojstvo konačnog tipa ako je $Spec(M)$ konačan skup. Iz konačnosti skupa $Spec(M)$ slijedi postojanje tačke "stabilizacije", tj. tačke $s_+ \in \mathbb{R}$ sa svojstvom da je za sve realne brojeve s, t za koje je $s_+ \leqslant s \leqslant t$ tranziciono preslikavanje $\phi_M(s,t)$ izomorfizam.

U daljnjem će biti podrazumijevano da su, u slučaju razmatranja više od jednog istrajnog modula, svi posmatrani istrajni moduli definisani nad istim poljem $\mathbb{F}$.

Neka su M i N istrajni moduli. *Morfizam* ili *prirodna transformacija* $\psi : M \Rightarrow N$ sastoji se od kolekcije linearnih preslikavanja $\{\psi_t : M_t \to N_t : t \in \mathbb{R}\}$ tako da za sve $s, t \in \mathbb{R}$ za koje je $s \leqslant t$ vrijedi $\phi_N(s,t) \circ \psi_s = \psi_t \circ \phi_M(s,t)$, tj. tako da je za sve $s \leqslant t$ sljedeći dijagram komutativan

$$\begin{array}{ccc} M_s & \xrightarrow{\phi_M(s,t)} & M_t \\ \downarrow{\scriptstyle \psi_s} & & \downarrow{\scriptstyle \psi_t} \\ N_s & \xrightarrow{\phi_N(s,t)} & N_t \end{array}$$

Nula morfizam je morfizam kod kojega su sva pripadna preslikavanja jednaka nula preslikavanju, a *identični morfizam* je morfizam istrajnog modula u samoga sebe kod kojega su sva pripadna preslikavanja identiteti na odgovarajućim vektorskim prostorima.

Lema 2.4.4. *Neka su* $I_1 = [b_1, d_1)$ *i* $I_2 = [b_2, d_2)$ *intervali, pri čemu* $b_1, b_2 \in \mathbb{R}, d_1, d_2 \in \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.

(i) *Ako je* $\psi : M(I_1) \Rightarrow M(I_2)$ *nenula morfizam, tada vrijedi:*

(i_1) $I_1 \cap I_2 \neq \emptyset$ *i za sve* $s, t \in I_1 \cap I_2$ *vrijedi* $\psi_s = \psi_t \neq 0$.

(i_2) $b_2 \leqslant b_1$ *i* $d_2 \leqslant d_1$.

(ii) *Ako je* $b_2 \leqslant b_1$ *i* $b_1 < d_2 \leqslant d_1$, *tada postoji nenula morfizam između istrajnih modula* $M(I_1)$ *i* $M(I_2)$.

Dokaz.

(i_1) Kako je ψ nenula morfizam, postoji $r \in \mathbb{R}$ tako da vrijedi $\psi_r \neq 0$. Tada $r \in I_1 \cap I_2$, jer bi u suprotnom ψ_r, kao preslikavanje čiji je domen ili kodomen jednak $\{0_{\mathbb{F}}\}$, moralo biti nula preslikavanje. Zbog toga je $I_1 \cap I_2 \neq \emptyset$ i ovaj presjek je ponovo interval. Neka su $s, t \in I_1 \cap I_2$ proizvoljni elementi za koje je $s < t$. Iz komutativnosti dijagrama

$$\begin{array}{ccc} \mathbb{F} & \xrightarrow{id_{\mathbb{F}}} & \mathbb{F} \\ \downarrow{\scriptstyle \psi_s} & & \downarrow{\scriptstyle \psi_t} \\ \mathbb{F} & \xrightarrow{id_{\mathbb{F}}} & \mathbb{F} \end{array}$$

slijedi $\psi_s = id_{\mathbb{F}} \circ \psi_s = \psi_t \circ id_{\mathbb{F}} = \psi_t$. Specijalno $\psi_s = \psi_r \neq 0$, pa je ispunjen i drugi dio tvrđenja.

(i_2) Iz pretpostavke da je $b_2 > b_1$ slijedi postojanje elementa $c \in (b_1, b_2) \cap I_1$, a na osnovu dijela (i_1) postoji $c' \in I_1 \cap I_2$. Kako $c \notin I_2$, vrijedi $M(I_2)_c = \{0_{\mathbb{F}}\}$, što implicira $\psi_c = 0$. Takođe, vrijedi $c < c'$, pa se dobija sljedeći komutativni dijagram

$$\begin{array}{ccc} \mathbb{F} & \xrightarrow{id_{\mathbb{F}}} & \mathbb{F} \\ \downarrow{\scriptstyle \psi_c=0} & & \downarrow{\scriptstyle \psi_{c'}} \\ \{0_{\mathbb{F}}\} & \xrightarrow{0} & \mathbb{F} \end{array}$$

Međutim, to bi značilo da je $\psi_{c'} = \psi_{c'} \circ id_{\mathbb{F}} = 0 \circ \psi_c = 0$, što, na osnovu (i_1), nije moguće. Dakle, vrijedi $b_2 \leqslant b_1$, a na sličan način se dokazuje da je $d_2 \leqslant d_1$.

(ii) Iz datog uslova slijedi da je $I_1 \cap I_2 = [b_1, d_2)$. Ako se za $c \in \mathbb{R}$ definiše linearno preslikavanje

$$\chi_c := \begin{cases} id_{\mathbb{F}}, & \text{za } c \in [b_1, d_2); \\ 0, & \text{inače,} \end{cases}$$

tada nije teško provjeriti da je kolekcijom preslikavanja $\{\chi_c : c \in \mathbb{R}\}$ određen jedan nenula morfizam $\chi : M(I_1) \Rightarrow M(I_2)$. □

Primjedba 2.4.5. Na osnovu prethodne leme, nenula morfizam intervalnih istrajnih modula $M(I_1)$ i $M(I_2)$ postoji ako i samo ako se ovi intervali sijeku i interval I_2 se ne može nalaziti "desno" od intervala I_1. Osim toga, na osnovu ove leme se zaključuje da je svaki ovakav nenula morfizam jedinstveno određen sa nenula endomorfizmom polja $\mathbb{F}$, a iz Šurove leme ([82], Propozicija 4, strana 13) slijedi da je svaki ovakav endomorfizam oblika $c \cdot id_{\mathbb{F}}$, za neko $c \in \mathbb{F}$.

Za istrajne module M, N, Q, i morfizme $\psi : M \Rightarrow N$ i $\chi : N \Rightarrow Q$, kolekcija linearnih preslikavanja $\{(\chi \circ \psi)_t : M_t \to Q_t | t \in \mathbb{R}\}$, pri čemu je $(\chi \circ \psi)_t := \chi_t \circ \psi_t$, ima svojstvo da je sljedeći dijagram komutativan za sve $s \leqslant t$

$$\begin{array}{ccc} M_s & \xrightarrow{\phi_M(s,t)} & M_t \\ \downarrow{\scriptstyle \psi_s} & & \downarrow{\scriptstyle \psi_t} \\ N_s & \xrightarrow{\phi_N(s,t)} & N_t \\ \downarrow{\scriptstyle \chi_s} & & \downarrow{\scriptstyle \chi_t} \\ Q_s & \xrightarrow{\phi_Q(s,t)} & Q_t \end{array}$$

Zbog toga, preslikavanjima $\{(\chi \circ \psi)_t : M_t \to Q_t | t \in \mathbb{R}\}$ je u potpunosti određen morfizam istrajnog modula M u istrajni modul Q, koji se označava sa $\chi \circ \psi$ i naziva *kompozicijom morfizama* ψ i χ. Istrajni moduli M i N su *izomorfni* (u oznaci $M \cong N$), ako postoje morfizmi $\psi : M \Rightarrow N$ i $\chi : N \Rightarrow M$ takvi da su $\chi \circ \psi$ i $\psi \circ \chi$ identični morfizmi na odgovarajućim istrajnim modulima. U tom slučaju morfizmi ψ i χ se nazivaju *prirodnim izomorfizmima*. Na osnovu prethodne leme jednostavno se dokazuje sljedeća lema.

Lema 2.4.6. *Intervalni istrajni moduli $M([b_1, d_1))$ i $M([b_2, d_2))$ su izomorfni ako i samo ako je $[b_1, d_1) = [b_2, d_2)$.*

Neka je M istrajni modul. *Istrajni podmodul* W modula M je istrajni modul sastavljen od vektorskih potprostora $W_t \subseteq M_t$, pri čemu su njegova tranziciona preslikavanja $\phi_W(s,t) = \phi_M(s,t)_{|W_s} : W_s \to W_t$ dobro definisana za svako $s \leqslant t$. Dva istaknuta tipa istrajnih podmodula su konstruisana u sljedećoj lemi.

Lema 2.4.7. *Neka su M i N istrajni moduli i $\psi : M \Rightarrow N$ morfizam. Tada*

(i) Kolekcija vektorskih prostora $\{ker(\psi_t) : t \in \mathbb{R}\}$ snabdjevena familijom linearnih preslikavanja $\phi_M(s,t)_{|ker(\psi_s)}$, za sve $s \leqslant t$, formira istrajni podmodul od M koji se naziva jezgrom morfizma ψ i označava sa $ker(\psi)$.

(ii) Kolekcija vektorskih prostora $\{im(\psi_t) : t \in \mathbb{R}\}$ snabdjevena familijom linearnih preslikavanja $\phi_N(s,t)_{|im(\psi_s)}$, za sve $s \leqslant t$, formira istrajni podmodul od N koji se naziva slikom morfizma ψ i označava sa $im(\psi)$.

Dokaz. (i) Očigledno da je $ker(\psi_t)$ potprostor od M_t, za svako $t \in \mathbb{R}$, pa ostaje jedino da se dokaže da za sve $s \leqslant t$ preslikavanje $\phi_M(s,t)_{|ker(\psi_s)}$ preslikava $ker(\psi_s)$ u $ker(\psi_t)$. Zaista, za proizvoljno $x \in ker(\psi_s)$, iz komutativnosti odgovarajućih dijagrama se dobija

$$\begin{aligned}\psi_t\left(\phi_M(s,t)(x)\right) &= \left(\psi_t \circ \phi_M(s,t)\right)(x) = \left(\phi_N(s,t) \circ \psi_s\right)(x) \\ &= \phi_N(s,t)\left(\psi_s(x)\right) = \phi_N(s,t)(0_{N_s}) = 0_{N_t},\end{aligned}$$

što implicira $\phi_M(s,t)(x) \in ker(\psi_t)$.

(ii) Jasno da je $im(\psi_t)$ potprostor od N_t, za svako $t \in \mathbb{R}$, pa ostaje jedino da se dokaže da za sve $s \leqslant t$ preslikavanje $\phi_N(s,t)_{|im(\psi_s)}$ preslikava $im(\psi_s)$ u $im(\psi_t)$. Zaista, za proizvoljno $y \in im(\psi_s)$ postoji $x \in M_s$ takvo da je $y = \psi_s(x)$. Iz komutativnosti odgovarajućih dijagrama se dobija

$$\begin{aligned}\phi_N(s,t)(y) &= \left(\phi_N(s,t)\right)\left(\psi_s(x)\right) = \left(\phi_N(s,t) \circ \psi_s\right)(x) = \left(\psi_t \circ \phi_M(s,t)\right)(x) \\ &= \psi_t\left(\phi_M(s,t)(x)\right),\end{aligned}$$

što implicira $\phi_N(s,t)(x) \in im(\psi_t)$. □

Primjer 2.4.8. *Za intervale $I_1 = [b, d_1)$ i $I_2 = [b, d_2)$, gdje je $b, d_1 \in \mathbb{R}, d_2 \in \overline{\mathbb{R}}$ i $d_1 < d_2$, intervalni istrajni modul $M(I_1)$ nije istrajni podmodul istrajnog modula $M(I_2)$. Zaista, za element $d \in (d_1, d_2)$ vrijedi $\phi_{M(I_2)}(b, d) = id_{\mathbb{F}}$, dok je $\phi_{M(I_1)}(b, d) = 0$, pa kako je $M(I_1)_b = M(I_2)_b = \mathbb{F}$ ne može vrijediti $\phi_{M(I_1)}(b, d) = \phi_{M(I_2)}(b, d)_{|M(I_1)_b}$.*

Još jedan način dobijanja novih istrajnih modula je putem uzimanja direktnih suma. Preciznije, *direktna suma istrajnih modula* M i N je istrajni modul $M \bigoplus N$ sastavljen od kolekcije vektorskih prostora $\{M_t \bigoplus N_t : t \in \mathbb{R}\}$ i tranzicionih preslikavanja $\phi_M(s, t) \bigoplus \phi_N(s, t)$, za $s \leqslant t$. Istrajnom modulu $M \bigoplus N$ se pridružuju kolekcije linearnih preslikavanja $\{e_t^M : t \in \mathbb{R}\}$ i $\{p_t^M : t \in \mathbb{R}\}$, pri čemu je $e_t^M : M_t \to M_t \bigoplus N_t$ prirodno potapanje prostora M_t u prostor $M_t \bigoplus N_t$, a $p_t^M : M_t \bigoplus N_t \to M_t$ projekcija prostora $M_t \bigoplus N_t$ na prostor M_t. Jednostavno se dokazuje da su ovim određeni morfizmi $e^M : M \Rightarrow M \bigoplus N$ i $p^M : M \bigoplus N \Rightarrow M$, koji se redom nazivaju *morfizmom potapanja* i *morfizmom projekcije*. Ovi morfizmi se slično definišu i u opštijem slučaju direktne sume konačno mnogo istrajnih modula. Istrajni modul koji je izomorfan direktnoj sumi nekih nenula istrajnih modula naziva se *dekompozibilnim istrajnim modulom*.

Lema 2.4.9. *Proizvoljan intervalni istrajni modul nije dekompozibilan.*

Dokaz. Neka je $I \subseteq \mathbb{R}$ proizvoljan interval. Ako je $I = \emptyset$, tada je tvrđenje trivijalno ispunjeno. Stoga, neka je $I \neq \emptyset$ i $M(I) \cong N \bigoplus Q$, za neke istrajne module N i Q. Ako je $I \subsetneq \mathbb{R}$, tada za svako $r \in \mathbb{R} \setminus I$ vrijedi $N_r \bigoplus Q_r \cong M(I)_r = \{0_{\mathbb{F}}\}$, te za takve vrijednosti r posljedično vrijedi $N_r \cong \{0_{\mathbb{F}}\} \cong Q_r$. Neka je $t \in I$ proizvoljno. Tada je $N_t \bigoplus Q_t \cong M(I)_t = \mathbb{F}$, pa, kako je $\mathbb{F}$ jednodimenzionalan vektorski prostor, slijedi da je jedan od vektorskih prostora N_t, Q_t trivijalan, a drugi izomorfan sa $\mathbb{F}$. Bez gubljenja na opštosti se može pretpostaviti da je $N_t \cong \mathbb{F}$ i $Q_t \cong \{0_{\mathbb{F}}\}$. Za svako $s \in I$ za koje je $s \geqslant t$ vrijedi $\phi_Q(t, s) = 0$, pa iz $\phi_N(t, s) \bigoplus \phi_Q(t, s) = \left(\phi_{N \bigoplus Q}\right)(t, s) \cong \phi_{M(I)}(t, s) = id_{\mathbb{F}}$ slijedi $\phi_N(t, s) \cong id_{\mathbb{F}}$. Zbog toga se iz $N_s \bigoplus Q_s \cong M(I)_s = \mathbb{F}$ dobija $N_s \cong \mathbb{F}$ i $Q_s \cong \{0_{\mathbb{F}}\}$. Na sličan način se dokazuje da za svako $s \in I$ za koje je $s \leqslant t$ vrijedi $\phi_N(s, t) \cong id_{\mathbb{F}}$, $\phi_Q(s, t) = 0$, $N_s \cong \mathbb{F}$ i $Q_s \cong \{0_{\mathbb{F}}\}$. Na taj način je dokazano da je $N \cong M(I)$ i da je Q izomorfan nula istrajnom modulu. Dakle, istrajni modul $M(I)$ nije dekompozibilan. □

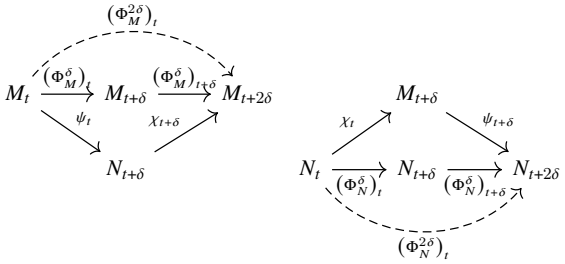
Primjer 2.4.10. *Neka su b_1, d_1, b_2, d_2 realni brojevi za koje vrijedi $b_1 < d_1 = b_2 < d_2$. Ako je $I_1 = [b_1, d_1)$ i $I_2 = [b_2, d_2)$, tada istrajni modul $M(I_1 \cup I_2)$ nije izomorfan istrajnom modulu $M(I_1) \bigoplus M(I_2)$.*

Postojanje izomorfizma između dva istrajna modula upućuje na to da su ova dva modula strukturalno slična, tj. da je lance njihovih tranzicionih preslikavanja moguće "preklopiti". Ako dva istrajna modula nisu izomorfna, postavlja se

pitanje da li je pomenute lance moguće translirati tako da se novodobijeni lanci mogu preklopiti? Prateći ovu ideju, dolazi se do pojma preplitanja i udaljenosti preplitanja.

Za istrajni modul M i $\delta \in \mathbb{R}$, sa $M(\delta)$ biće označen istrajni modul tako da je za sve $s \leqslant t$ ispunjeno $M(\delta)_t := M_{t+\delta}$ i $\phi_{M(\delta)}(s,t) := \phi_M(s+\delta, t+\delta)$. Ovaj istrajni modul se naziva *$\delta-$pomjeranjem istrajnog modula M*. Za $\delta \geqslant 0$, između istrajnih modula M i $M(\delta)$ se može uspostaviti prirodna transformacija $\Phi_M^\delta : M \Rightarrow M(\delta)$ koju čine linearna preslikavanja $\left(\Phi_M^\delta\right)_t := \phi_M(t, t+\delta)$, a koja se naziva *morfizmom $\delta-$pomjeranja*. Takođe, za morfizam $\psi : M \Rightarrow N$ istrajnih modula M i N, sa $\psi(\delta) : M(\delta) \Rightarrow N(\delta)$ biće označen morfizam između njihovih $\delta-$pomjeranja. Ovaj morfizam je sastavljen od linearnih preslikavanja $\psi(\delta)_t := \psi_{t+\delta}$.

Neka su M i N istrajni moduli. Za dato $\delta \geqslant 0$, moduli M i N su *$\delta-$prepleteni*, ako postoje morfizmi $\psi : M \Rightarrow N(\delta)$ i $\chi : N \Rightarrow M(\delta)$ za koje vrijedi $\chi(\delta) \circ \psi = \Phi_M^{2\delta}$ i $\psi(\delta) \circ \chi = \Phi_N^{2\delta}$. U tom slučaju se dati morfizmi nazivaju *$\delta-$prepletenim morfizmima*, što znači da su, za ove morfizme i svako $t \in \mathbb{R}$, sljedeći dijagrami komutativni



Primjer 2.4.11. *Intervalni istrajni moduli $M(0,1)$ i $M[0,1]$ su δ-prepleteni za svako $\delta \in \left(0, \frac{1}{2}\right)$, jer morfizmi $\psi : M(0,1) \Rightarrow M[0,1](\delta)$ i $\chi : M[0,1] \Rightarrow M(0,1)(\delta)$, pri čemu je $\psi_t = \chi_t = id_{\mathbb{F}}$, zadovoljavaju uslove iz definicije $\delta-$prepletenosti.*

Za $\delta \geqslant 0$, istrajni modul M se naziva *$2\delta-$trivijalnim*, ako je on $\delta-$prepleten sa nula istrajnim modulom. Istrajni modul koji nije $2\delta-$trivijalan se naziva *$2\delta-$značajnim*.

Lema 2.4.12. *(i) Istrajni modul M je $2\delta-$značajan ako i samo ako je $\Phi_M^{2\delta} \neq 0$.*

(ii) Nenula intervalni istrajni modul $M(I)$ je 2δ-značajan ako i samo ako postoji $t \in \mathbb{R}$ tako da je $[t, t+2\delta] \subseteq I$.

Dokaz. (i) Za svako $\delta' \geqslant 0$ je $\delta'-$pomjeranje nula istrajnog modula izomorfno nula istrajnom modulu. Takođe, svaki morfizam iz ili u nula istrajni modul je nula morfizam. Zbog toga, istrajni modul je $2\delta-$trivijalan ako i samo ako je $\Phi_M^{2\delta} = 0$.

(*ii*) Za svako $\delta' \geqslant 0$ vrijedi $\phi_{M(I)}(t, t+\delta') = \begin{cases} id_{\mathbb{F}}, & \text{ako } t, t+\delta' \in I; \\ 0, & \text{inače} \end{cases}$.

Zbog toga, $\left(\Phi^{2\delta}_{M(I)}\right)_t = \phi_{M(I)}(t, t+2\delta) \neq 0$ ako i samo ako je $t, t+2\delta \in I$. Iz ovoga i činjenice da je I konveksan skup, slijedi da je $\Phi^{2\delta}_{M(I)} \neq 0$ ako i samo ako postoji $t \in \mathbb{R}$ sa svojstvom $[t, t+2\delta] \subseteq I$. Dokaz se kompletira primjenom tvrđenja iz dijela (i). □

Funkcija d_{INT} koja istrajnim modulima M i N pridružuje vrijednost

$$d_{INT}(M, N) := \inf\{\delta \geqslant 0 : M \text{ i } N \text{ su } \delta - \text{prepleteni}\}$$

naziva se *udaljenošću preplitanja*. Kao i u slučaju nekih drugih "udaljenosti" ova funkcija ne ispunjava sva svojstva iz definicije metrike. Preciznije, u narednom će biti dokazano da je ova funkcija produžena pseudometrika. To znači da ova funkcija uzima vrijednosti na skupu $[0, +\infty]$ i može uzimati vrijednost 0 za različite istrajne module. Štaviše, za intervalne istrajne module iz prethodnog primjera vrijedi $d_{INT}\big(M[0,1], M(0,1)\big) = 0$, ali ovi moduli nisu čak ni izomorfni. U cilju dokazivanja nejednakosti trougla za funkciju d_{INT} najprije su dokazana dva pomoćna tvrđenja.

Lema 2.4.13. *Ako su M i N $\delta-$prepleteni istrajni moduli, tada su, za svako $\delta' \geqslant 0$, ovi moduli ujedno i $\delta+\delta'-$prepleteni.*

Dokaz. Neka su M i N $\delta-$prepleteni istrajni moduli i $\psi : M \Rightarrow N(\delta)$, $\chi : N \Rightarrow M(\delta)$ prirodne transformacije za koje vrijedi $\chi(\delta) \circ \psi = \Phi^{2\delta}_M$ i $\psi(\delta) \circ \chi = \Phi^{2\delta}_N$. Neka je $\delta' \geqslant 0$ proizvoljno i $\psi' : M \Rightarrow N(\delta+\delta')$, $\chi' : N \Rightarrow M(\delta+\delta')$ morfizmi definisani sa $\psi' := \Phi^{\delta'}_{N(\delta)} \circ \psi$ i $\chi' := \Phi^{\delta'}_{M(\delta)} \circ \chi$. Dovoljno je dokazati da su ψ' i χ' $\delta+\delta'-$prepleteni morfizmi, tj. da je ispunjeno $\chi'(\delta+\delta') \circ \psi' = \Phi^{2(\delta+\delta')}_M$ i $\psi'(\delta+\delta') \circ \chi' = \Phi^{2(\delta+\delta')}_N$. Za proizvoljno $t \in \mathbb{R}$ vrijedi

$$\begin{aligned}(\chi'(\delta+\delta') \circ \psi')_t &= \chi'(\delta+\delta')_t \circ \psi'_t = \chi'_{t+\delta+\delta'} \circ \psi'_t \\ &= \left(\left(\Phi^{\delta'}_{M(\delta)}\right)_{t+\delta+\delta'} \circ \chi_{t+\delta+\delta'}\right) \circ \left(\left(\Phi^{\delta'}_{N(\delta)}\right)_t \circ \psi_t\right) \\ &= \left(\phi_{M(\delta)}(t+\delta+\delta', t+\delta+2\delta') \circ \chi_{t+\delta+\delta'}\right) \\ &\circ \left(\phi_{N(\delta)}(t, t+\delta') \circ \psi_t\right) = \phi_M(t+2\delta+\delta', t+2\delta+2\delta') \\ &\circ \chi_{t+\delta+\delta'} \circ \phi_N(t+\delta, t+\delta+\delta') \circ \psi_t.\end{aligned}$$

Iz pretpostavke da je ψ morfizam slijedi da je $\phi_N(t+\delta, t+\delta+\delta') \circ \psi_t = \psi_{t+\delta'} \circ \phi_M(t, t+\delta')$, pa se na osnovu prethodnog dobija

$$(\chi'(\delta+\delta') \circ \psi')_t = \phi_M(t+2\delta+\delta', t+2\delta+2\delta') \circ \chi_{t+\delta+\delta'} \circ \psi_{t+\delta'} \circ \phi_M(t, t+\delta').$$

Dalje, koristeći $\delta-$prepletenost morfizama ψ i χ dobija se

$$\begin{aligned}\chi_{t+\delta+\delta'} \circ \psi_{t+\delta'} &= \chi(\delta)_{t+\delta'} \circ \psi_{t+\delta'} = (\chi(\delta) \circ \psi)_{t+\delta'} = \left(\Phi^{2\delta}_M\right)_{t+\delta'} \\ &= \phi_M(t+\delta', t+\delta'+2\delta),\end{aligned}$$

pa se konačno dobija

$$(\chi'(\delta+\delta')\circ\psi')_t = \phi_M(t+2\delta+\delta', t+2\delta+2\delta')\circ\phi_M(t+\delta', t+\delta'+2\delta)$$
$$\circ\,\phi_M(t, t+\delta') = \phi_M(t, t+2\delta+2\delta') = \left(\Phi_M^{2(\delta+\delta')}\right)_t.$$

Time je dokazano da vrijedi $\chi'(\delta+\delta')\circ\psi' = \Phi_M^{2(\delta+\delta')}$, a dokaz druge jednakosti se izvodi na sličan način. □

Lema 2.4.14. *Neka su $\delta', \delta'' \geqslant 0$. Ako su M i N δ'−prepleteni istrajni moduli, a N i Q δ''−prepleteni istrajni moduli, tada su M i Q $\delta'+\delta''$−prepleteni istrajni moduli.*

Dokaz. Neka su $\psi' : M \Rightarrow N(\delta')$, $\chi' : N \Rightarrow M(\delta')$, $\psi'' : N \Rightarrow Q(\delta'')$, $\chi'' : Q \Rightarrow N(\delta'')$ morfizmi za koje vrijedi $\chi'(\delta')\circ\psi' = \Phi_M^{2\delta'}$, $\psi'(\delta')\circ\chi' = \Phi_N^{2\delta'}$, $\chi''(\delta'')\circ\psi'' = \Phi_N^{2\delta''}$, $\psi''(\delta'')\circ\chi'' = \Phi_Q^{2\delta''}$. Za kompletiranje dokaza, dovoljno je verifikovati da morfizmi $\psi : M \Rightarrow Q(\delta'+\delta'')$ i $\chi : Q \Rightarrow M(\delta'+\delta'')$ dati sa $\psi = \psi''(\delta')\circ\psi'$, $\chi = \chi'(\delta'')\circ\chi''$ predstavljaju jedno $\delta'+\delta''$−preplitanje istrajnih modula M i Q. Za svako $t \in \mathbb{R}$ vrijedi

$$\begin{aligned}
(\chi(\delta'+\delta'')\circ\psi)_t &= \chi(\delta'+\delta'')_t\circ\psi_t = \chi_{t+\delta'+\delta''}\circ\psi_t \\
&= \left(\chi'(\delta'')_{t+\delta'+\delta''}\circ\chi''_{t+\delta'+\delta''}\right)\circ\left(\psi''(\delta')_t\circ\psi'_t\right) \\
&= \chi'_{t+\delta'+2\delta''}\circ\underbrace{\chi''_{t+\delta'+\delta''}\circ\psi''_{t+\delta'}}\circ\psi'_t \\
&= \chi'_{t+\delta'+2\delta''}\circ\underbrace{\phi_N(t+\delta', t+\delta'+2\delta'')\circ\psi'_t} \\
&= \underbrace{\chi'_{t+\delta'+2\delta''}\circ\psi'_{t+2\delta''}}\circ\phi_M(t, t+2\delta'') \\
&= \phi_M(t+2\delta'', t+2(\delta'+\delta''))\circ\phi_M(t, t+2\delta'') \\
&= \phi_M(t, t+2(\delta'+\delta'')) = \left(\Phi_M^{2(\delta'+\delta'')}\right)_t.
\end{aligned}$$

Time je dokazano da vrijedi $\chi(\delta'+\delta'')\circ\psi = \Phi_M^{2(\delta'+\delta'')}$, a dokaz jednakosti $\psi(\delta'+\delta'')\circ\chi = \Phi_Q^{2(\delta'+\delta'')}$ se izvodi na sličan način. □

Lema 2.4.15. *Funkcija d_{INT} je produžena pseudometrika.*

Dokaz. Očigledno da za sve istrajne module M i N vrijedi $d_{INT}(M,M) = 0$ i $d_{INT}(M,N) = d_{INT}(N,M)$. Treba dokazati nejednakost trougla, tj. da za sve istrajne module M, N i Q vrijedi $d_{INT}(M,Q) \leqslant d_{INT}(M,N)+d_{INT}(N,Q)$. Ako je bar jedan od izraza $d_{INT}(M,N)$, $d_{INT}(N,Q)$ beskonačan, tada nejednakost trivijalno vrijedi. Stoga, neka $d_{INT}(M,N)$, $d_{INT}(N,Q) \in [0,+\infty)$ i neka je $\varepsilon > 0$ proizvoljno. Kako je

$$d_{INT}(M,N) = \inf\{\delta \geqslant 0 : M \text{ i } N \text{ su } \delta-\text{prepleteni}\},$$

na osnovu prvog pomoćnog rezultata slijedi da su moduli M i N $d_{INT}(M,N)+\varepsilon$−prepleteni. Na sličan način se zaključuje da su moduli N i Q $d_{INT}(N,Q)+\varepsilon$−prepleteni. Koristeći drugi pomoćni rezultat, dalje se zaključuje da su moduli M i Q $d_{INT}(M,N)+d_{INT}(N,Q)+2\varepsilon$−prepleteni. S obzirom da je ovo ispunjeno za svako $\varepsilon > 0$, slijedi $d_{INT}(M,Q) \leqslant d_{INT}(M,N) + d_{INT}(N,Q)$. □

U narednoj lemi dato je gornje ograničenje udaljenosti preplitanja dva intervalna istrajna modula oblika $M[b_1,d_1)$ i $M[b_2,d_2)$.

Lema 2.4.16. *Za* $b_1,d_1,b_2,d_2 \in \mathbb{R}$ *i* $i_1 = \dfrac{d_1-b_1}{2}$, $i_2 = \dfrac{d_2-b_2}{2}$ *vrijedi*

$$(i)\, d_{INT}(M[b_1,d_1),M[b_2,d_2)) \leqslant \min\{\max\{i_1,i_2\},\max\{|b_1-b_2|,|d_1-d_2|\}\}, \tag{2.5}$$

(ii) $d_{INT}\left(M[b_1,d_1),M[b_2,+\infty)\right) = +\infty$,

(iii) $d_{INT}\left(M[b_1,+\infty),M[b_2,+\infty)\right) = |b_1-b_2|$.

Dokaz. (i) Dovoljno je dokazati da je udaljenost preplitanja manja ili jednaka od obje vrijednosti sa desne strane nejednakosti (2.5).

Neka je najprije $\delta := \max\{|b_1-b_2|,|d_1-d_2|\}$. Tada je $b_1-2\delta \leqslant b_2-\delta \leqslant b_1$ i $d_1-2\delta \leqslant d_2-\delta \leqslant d_1$, pa je, kao u dokazu Leme 2.4.4, moguće konstruisati morfizam $\psi : M[b_1,d_1) \Rightarrow M[b_2-\delta,d_2-\delta)$ sastavljen od linearnih preslikavanja

$$\psi_c := \begin{cases} id_{\mathbb{F}}, & \text{za } c \in [b_1,d_2-\delta); \\ 0, & \text{inače} \end{cases}$$

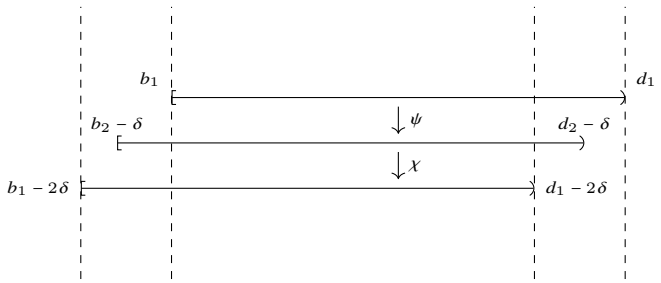
i morfizam $\chi : M[b_2,d_2) \Rightarrow [b_1-\delta,d_1-\delta)$ sastavljen od linearnih preslikavanja

$$\chi_c := \begin{cases} id_{\mathbb{F}}, & \text{za } c \in [b_2,d_1-\delta); \\ 0, & \text{inače.} \end{cases}$$

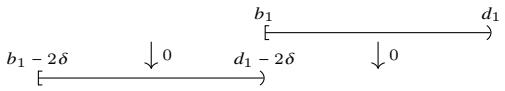
Ovi morfizmi čine δ−preplitanje istrajnih modula $M[b_1,d_1)$ i $M[b_2,d_2)$, sa napomenom da su ovo nula morfizmi u slučaju kada se odgovarajući intervali ne sijeku (npr. $\psi = 0$ ukoliko je $[b_1,d_2-\delta) = [b_1,d_1) \cap [b_2-\delta,d_2-\delta) = \emptyset$, Slika 2.5). Zbog toga je $d_{INT}\left(M[b_1,d_1),M[b_2,d_2)\right) \leqslant \delta$.

Neka je sada $\delta := \max\left\{\dfrac{d_1-b_1}{2},\dfrac{d_2-b_2}{2}\right\}$. Tada je $[b_1,d_1)\cap[b_1-2\delta,d_1-2\delta) = \emptyset$, što implicira da je $\Phi^{2\delta}_{M[b_1,d_1)} = 0$ (Slika 2.6), a slično se dobija da je $[b_2,d_2) \cap [b_2-2\delta,d_2-2\delta) = \emptyset$, tj. $\Phi^{2\delta}_{M[b_2,d_2)} = 0$. To znači da par nula morfizama predstavlja δ−preplitanje istrajnih modula $M[b_1,d_1)$ i $M[b_2,d_2)$, te je posljedično ispunjeno $d_{INT}\left(M[b_1,d_1),M[b_2,d_2)\right) \leqslant \delta$.

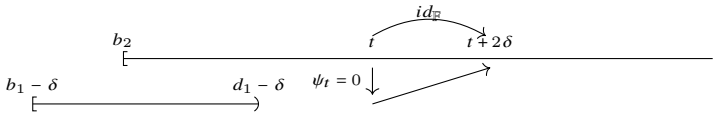
(ii) Ne može postojati konačno δ−preplitanje posmatranih istrajnih modula, jer za proizvoljno $t \in [d_1,+\infty)$ vrijedi $\left(\Phi^{2\delta}_{M[b_2,+\infty)}\right)_t = id_{\mathbb{F}}$, dok je za proizvoljan morfizam $\psi : M[b_2,+\infty) \Rightarrow M[b_1-\delta,d_1-\delta)$ ispunjeno $\psi_t = 0$ (Slika 2.7).



Slika 2.5

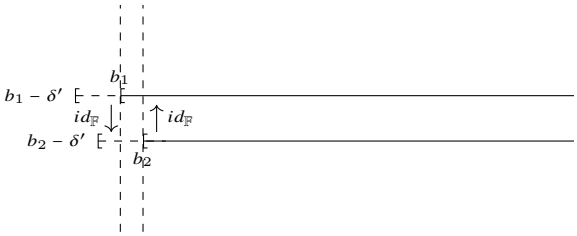


Slika 2.6



Slika 2.7

(*iii*) Neka je $\delta := |b_1 - b_2|$ i $\delta' \geqslant \delta$ proizvoljno. Tada je $[b_1, +\infty) \subseteq [b_2 - \delta', +\infty)$ i $[b_2, +\infty) \subseteq [b_1 - \delta', +\infty)$, pa morfizmi sastavljeni od identiteta predstavljaju δ'−preplitanje istrajnih modula $M[b_1, +\infty)$ i $M[b_2, +\infty)$. Zbog toga, vrijedi $d_{INT}\left(M[b_1, +\infty), M[b_2, +\infty)\right) \leqslant \delta$ (Slika 2.8).



Slika 2.8

Sa prethodne slike se uočava da, ako se bar jedan od intervala $[b_1, +\infty)$, $[b_2, +\infty)$ "pomjeri" ulijevo za manje od δ, da bar jedan od pripadnih morfizama mora biti jednak nuli. To znači da za proizvoljno $\delta'' < \delta$ ne postoji δ''−preplitanje istrajnih modula $M[b_1, +\infty)$ i $M[b_2, +\infty)$, te posljedično, udaljenost preplitanja između ovih istrajnih modula ne može biti strogo manja od

δ. □

Primjedba 2.4.17. Iz prethodne leme se primjećuje da su dva $2\delta-$trivijalna intervalna istrajna modula $M(I_1)$ i $M(I_2)$ uvijek $\delta-$prepletena bez obzira da li su intervali I_1 i I_2 "dovoljno blizu" jedan drugome.

2.5 Teorema o intervalnoj dekompoziciji istrajnog modula

U ovoj sekciji biće dokazano da je svaki istrajni modul koji ima svojstva konačnog tipa dekompozibilan, te da je izomorfan direktnoj sumi određene familije intervalnih istrajnih modula. Posljedica ovoga je da se svakom takvom modulu može pridružiti jedinstven (do na permutaciju) multiskup intervala. Na taj način, struktura istrajnog modula može da se analizira sa aspekta dobijenog multiskupa intervala.

Formalno, *multiskup* je uređena dvojka (S, m), gdje je S skup, a m funkcija koja elementima skupa S pridružuje višestrukost njihovog pojavljivanja u okviru tog multiskupa. Za potrebe ove teze smatraće se da je kodomen ove funkcije skup $\mathbb{N} \cup \{+\infty\}$. Svakom multiskupu (S, m) pridružuje se skup $Rep(S, m) := \{(s, k) : k \leqslant m(s)\}$, koji se naziva *reprezentacija multiskupa* (S, m).

Teorema 2.5.1. *(**Intervalna dekompozicija istrajnog modula**) [5] Neka je M nenula istrajni modul koji ima svojstvo konačnog tipa. Tada postoji multiskup $(\mathcal{I}, m)$ intervala, gdje je $\mathcal{I} = \{I_1, \ldots, I_K\}$ sa svojstvom*

$$M \cong \bigoplus_{j=1}^{K} M(I_j)^{m(I_j)} \tag{2.6}$$

pri čemu je svaki I_j oblika $[b_j, d_j)$ ili $[b_j, +\infty)$, za neke $b_j, d_j \in \mathbb{R}$. Štaviše, prethodna reprezentacija je jedinstvena do na permutaciju skupa $\mathcal{I}$.

Primjedba 2.5.2. Prethodna teorema vrijedi i ako se isključi uslov da istrajni modul M ima svojstva konačnog tipa. Dokaz ove opštije varijante je, naravno, dosta komplikovaniji ([27], Teorema 1.1, strana 2).

Multiskup intervala $(\mathcal{I}, m)$ iz prethodne teoreme se naziva *bar-kodom istrajnog modula M* i označava se još sa $BC(M)$. Specijalno, za bar-kod istrajne homologije M (Hom_k) biće korišćena kraća oznaka $BC(Hom_k)$.

Dokaz prethodne teoreme biće izveden u nekoliko etapa. Najprije će biti dokazano sljedeće pomoćno tvrđenje u vezi spektra istrajnog modula.

Lema 2.5.3. *Neka su M i N istrajni moduli.*
(i) Ako su M i N izomorfni, tada je $Spec(M) = Spec(N)$.
(ii) $Spec(M \bigoplus N) = Spec(M) \cup Spec(N)$.

Dokaz. (i) Neka je $\psi : M \Rightarrow N$ prirodni izomorfizam. Za sve $s < r$, preslikavanja $\psi_s : M_s \to N_s$ i $\psi_r : M_r \to N_r$ su izomorfizmi odgovarajućih vektorskih prostora i vrijedi $\psi_r \circ \phi_M(s,r) = \phi_N(s,r) \circ \psi_s$. Iz ove jednakosti se dobija $\phi_M(s,r) = \psi_r^{-1} \circ \phi_N(s,r) \circ \psi_s$, što implicira da je $\phi_M(s,r)$ izomorfizam ako i samo ako je $\phi_N(s,r)$ izomorfizam. Na osnovu toga slijedi $Spec(M) = Spec(N)$.

(ii) Kako je $\phi_{M \bigoplus N}(r,s) = \phi_M(r,s) \bigoplus \phi_N(r,s)$, zaključuje se da je tranziciono preslikavanje $\phi_{M \bigoplus N}(r,s)$ izomorfizam ako i samo ako su oba tranziciona preslikavanja $\phi_M(r,s)$, $\phi_N(r,s)$ izomorfizmi. Na osnovu toga se dobija navedena jednakost. □

Primjer 2.5.4. *Iz definicije intervalnog istrajnog modula slijedi da je, za $b_1, d_1 \in \mathbb{R}$, $Spec\,(M[b_1,d_1)) = \{b_1, d_1\}$ i $Spec\,(M[b_1,+\infty)) = \{b_1\}$. Na osnovu prethodne leme, vrijedi $Spec\left(M[b_1,d_1) \bigoplus M[b_2,d_2)\right) = \{b_1,d_1,b_2,d_2\}$, za $b_1,d_1,b_2,d_2 \in \mathbb{R}$. Na sličan način je moguće odrediti spektar bilo koje konačne direktne sume intervalnih istrajnih modula.*

Sljedeći korak dokaza Teoreme 2.5.1 podrazumijeva svojevrsnu diskretizaciju kolekcije vektorskih prostora $\{M_t : t \in \mathbb{R}\}$ istrajnog modula M na način da novodobijena diskretna familija vektorskih prostora "čuva" informaciju koju nose tranziciona preslikavanja $\phi_M(s,t)$.

Neka je M istrajni modul čiji je spektar $Spec(M) = \{a_1, a_2, \ldots, a_N\}$, pri čemu je $a_1 < a_2 < \ldots < a_N$. Tada je moguće definisati intervale

$$Q_j := \begin{cases} (-\infty, a_1), & \text{za } j = 1; \\ [a_{j-1}, a_j), & \text{za } j \in \{2, \ldots, N\}; \\ [a_N, +\infty), & \text{za } j = N+1. \end{cases} \tag{2.7}$$

Za svako $j \in \{1,2,\ldots,N+1\}$, na disjunktnoj uniji $\bigsqcup_{s \in Q_j} M_s$ posmatra se relacija ekvivalencije $\sim$ koja, za $s < t$, identifikuje elemente $v_s \in M_s$, $v_t \in M_t$ ukoliko je $v_t = \phi_M(s,t)(v_s)$. Faktor prostor u odnosu na ovu relaciju biće označen sa M^j. Kako su tranziciona preslikavanja $\phi_M(s,t)$ izomorfizmi za sve $s,t \in Q_j$, slijedi da je $M^1 \cong \{0_{\mathbb{F}}\}$ i $M^j \cong M_{a_{j-1}}$, za $j \in \{2,\ldots,N+1\}$. Kolekcija $\{M^j : j \in \{1,2,\ldots,N+1\}\}$ se može snabdjeti linearnim preslikavanjima $\pi_{j,k} : M^j \to M^k$, $j \leqslant k$, koja su indukovana preslikavanjima $\phi_M(s,t)$ na sljedeći način: Za klasu $y^j \in M^j$ i njenog predstavnika $\left(y^j\right)_{a_{j-1}} \in M_{a_{j-1}}$, $\pi_{j,k}\left(y^j\right)$ je klasa u M^k čiji je predstavnik $\phi_M(a_{j-1}, a_{k-1})\left(y^j\right)_{a_{j-1}}$. Uz to, neka je $Totaldim(M) := \sum_{j=2}^{N+1} dim\left(M^j\right)$.

Istrajni podmodul W istrajnog modula M je *semi-surjektivan*, ako postoji $r \in \mathbb{R}$ tako da vrijede svojstva

(i) Za svako $t < r$ vrijedi $W_t = M_t$,

(ii) Tranziciono preslikavanje $\phi_W(s,t) : W_s \to W_t$ je surjektivno za svako $r \leqslant s < t$.

Primjer 2.5.5. *$M[0,+\infty)$ je semi-surjektivan istrajni podmodul istrajnog modula $M[0,+\infty) \bigoplus M[1,2)$.*

Lema 2.5.6. *Neka je W semi-surjektivan istrajni podmodul modula M koji ima konačan spektar. Tada vrijedi*

(i) $Spec(W) \subseteq Spec(M)$ *i* $Totaldim(W) \leqslant Totaldim(M)$,

(ii) $\sup\{t \in \mathbb{R} : W_s = M_s,$ *za svako* $s \leqslant t\} \in Spec(M) \cup \{+\infty\}$.

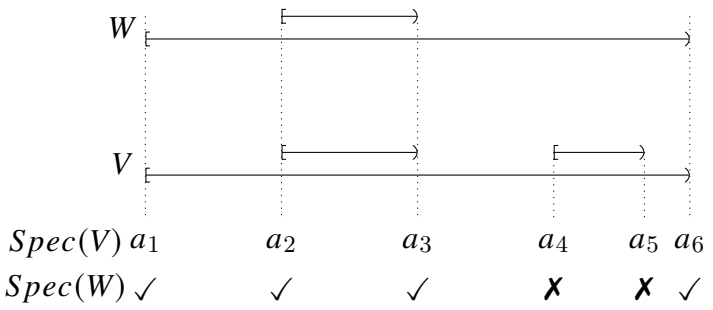
Dokaz. Neka je $r \in \mathbb{R}$ element koji zadovoljava svojstva iz definicije semi-surjektivnog istrajnog podmodula W.

(i) Iz pretpostavke $t \notin Spec(M)$ slijedi postojanje $\varepsilon > 0$, tako da je $\phi_M(s_1, s_2)$ izomorfizam za sve tačke $s_1 \leqslant s_2$ koje pripadaju intervalu $(t-\varepsilon, t+\varepsilon)$. Za takav izbor tačaka preslikavanje $\phi_W(s_1, s_2)$ je također izomorfizam, jer, u slučaju da je $s_1 < r$ vrijedi $\phi_W(s_1, s_2) = \phi_M(s_1, s_2)$, dok je za $r \leqslant s_1$ ispunjeno $\phi_W(s_1, s_2) = \phi_M(s_1, s_2)_{|W_{s_1}}$, pa je $\phi_W(s_1, s_2)$ surjektivno preslikavanje koje je i injektivno, jer je ono restrikcija injektivnog preslikavanja $\phi_M(s_1, s_2)$. Dakle, vrijedi $t \notin Spec(W)$, čime je dokazano $Spec(W) \subseteq Spec(M)$. Na osnovu ovoga slijedi drugi dio tvrđenja, ako se dodatno primijeti da za svako $j \in \{1, 2, \ldots, |Spec(W)| + 1\}$ i odgovarajuće $a_{j-1} \in Spec(W)$ vrijedi $dim(W^j) = dim(W_{a_{j-1}}) \leqslant dim(M_{a_{j-1}}) = dim(V^j)$.

(ii) Skup $U := \{t \in \mathbb{R} : W_s = M_s,$ za svako $s \leqslant t\}$ je neprazan, jer $r \in U$. Ako ovaj skup nije odozgo ograničen, tada je $\sup U = +\infty$. U suprotnom, postoji $u := \sup U$. Ako se pretpostavi $u \notin Spec(M)$, tada postoji $\varepsilon > 0$ tako da je $\phi_M(s_1, s_2)$ izomorfizam za sve tačke $s_1 \leqslant s_2$ koje pripadaju intervalu $(u - \varepsilon, u + \varepsilon)$. Neka su $s_1 \in (u - \varepsilon, u)$ i $s_2 \in (u, u + \varepsilon)$ proizvoljni elementi. Tada postoji $t \in U$ sa svojstvom $s_1 < t$, što implicira $W_{s_1} = M_{s_1}$, a iz činjenice da $s_2 \notin U$ slijedi postojanje $s_3 \in (u, s_2]$ sa svojstvom $W_{s_3} \subsetneq M_{s_3}$. Međutim, iz činjenice da je $\phi_M(s_1, s_3)$ izomorfizam, slijedi $M_{s_3} = \phi_M(s_1, s_3)(M_{s_1}) = \phi_M(s_1, s_3)(W_{s_1}) = \phi_W(s_1, s_3)(W_{s_1}) \subseteq W_{s_3}$, što je kontradikcija. □

Semi-surjektivan istrajni podmodul W istrajnog modula M biće "kodiran" na osnovu prostora $W^j \subseteq M^j$, dodjeljenih intervalima Q_j koji su uvedeni u formuli (2.7). Moguće je postojanje tačke $a_j \in Spec(M) \setminus Spec(W)$, pa u tom slučaju preslikavanje $\pi_{j,j+1} : W^j \to W^{j+1}$ može biti izomorfizam (Slika 2.9).

Lema 2.5.7. *Neka je $W \subsetneq M$ semi-surjektivan istrajni podmodul. Tada postoji semi-surjektivan istrajni podmodul $W_\sharp \subset M$ sa svojstvom $W_\sharp \cong W \bigoplus M[b,d)$, za neke $b \in Spec(M)$ i $d \in Spec(M) \cup \{+\infty\}$.*



Slika 2.9

Dokaz. Kako je W semi-surjektivan podmodul, postoji $r \in \mathbb{R}$ takvo da je $W_t = M_t$, za svako $t < r$, što znači da je i $W^j = M^j$ do određenog indeksa j. Neka je j_0 minimalan indeks za koji je $W^{j_0} \subsetneq M^{j_0}$ i $z^{j_0} \in M^{j_0} \setminus W^{j_0}$ proizvoljan element (uzimajući u obzir uvedeni način reprezentacije, ovo znači da je najmanja vrijednost t za koju vrijedi $W_t \subsetneq M_t$ jednaka a_{j_0-1}).

Za $k > j_0$, neka je $z^k := \pi_{j_0,k}(z^{j_0}) \in M^k$. U opticaju su dvije mogućnosti:

(i) Postoji $k > j_0$ tako da $z^k \in W^k$

U ovom slučaju se u direktnoj sumi pojavljuje intervalni istrajni modul određen ograničenim intervalom oblika $[b, d)$.

(ii) Za svako $k > j_0$, $z^k \notin W^k$

U ovom slučaju se u direktnoj sumi pojavljuje intervalni istrajni modul određen intervalom oblika $[b, +\infty)$.

(i) Neka je $k_0 := \min\{k > j_0 : z^k \in W^k\}$. Kako je $\pi_{j_0,k_0} : W^{j_0} \to W^{k_0}$ surjekcija, postoji $x^{j_0} \in W^{j_0}$ sa svojstvom $\pi_{j_0,k_0}(z^{j_0}) = z^{k_0} = \pi_{j_0,k_0}(x^{j_0})$. Neka je $y^{j_0} := z^{j_0} - x^{j_0}$. Zbog minimalnosti k_0, za svako $j_0 \leqslant k < k_0$ vrijedi $\pi_{j_0,k}(y^{j_0}) \notin W^k$. Također, $\pi_{j_0,k_0}(y^{j_0}) = 0$, pa za svako $k \geqslant k_0$ vrijedi $\pi_{j_0,k}(y^{j_0}) = \left(\pi_{k_0,k} \circ \pi_{j_0,k_0}\right)(y^{j_0}) = 0$. Drugačije rečeno, y^{k_0} je klasa za koju se $\pi_{j_0,k_0}(y^{j_0})$ anulira prvi put i nakon toga ostane jednako nuli (Slika 2.10).

$$
\begin{array}{ccccccccccc}
M^1 & \longrightarrow & \dots & \longrightarrow & M^{j_0} & \longrightarrow & \dots & \longrightarrow & M^{k_0} & \longrightarrow & M^{k_0+1} \longrightarrow \dots \\
 & & & & \cup & & & & \cup & & \\
 & & & & z^{j_0} & \longmapsto & \dots & \longmapsto & z^{k_0} & & \\
 & & & & \cap & & & & \cap & & \\
W^1 & \longrightarrow & \dots & \longrightarrow & W^{j_0} & \longrightarrow & \dots & \longrightarrow & W^{k_0} & \longrightarrow & W^{k_0+1} \longrightarrow \dots \\
 & & & & & & & & & & \\
 & & & & 0 \neq y^{j_0} & \longrightarrow & \dots & \longrightarrow & y^{k_0} = 0 & \longrightarrow & 0 \longrightarrow \dots
\end{array}
$$

Slika 2.10

Neka je $y^k := \pi_{j_0,k}(y^{j_0}) \in M^k$. Na osnovu ovih klasa ekvivalencije, moguće je izgraditi istrajni podmodul P modula M na sljedeći način:

Bira se predstavnik $\left(y^k\right)_s \in M_s$, za $s \in [a_{k-1}, a_k)$, i konstruišu se vektorski prostori

$$P_s := \begin{cases} \{0\}, & \text{ako } s \notin [a_{j_0-1}, a_{k_0-1}); \\ span_{\mathbb{F}}\left(\left(y^k\right)_s\right), & \text{ako } s \in [a_{k-1}, a_k) \subseteq [a_{j_0-1}, a_{k_0-1}), j_0 \leqslant k < k_0 \end{cases}$$

i tranziciona preslikavanja data sa

$$\phi_P(s,t) := \begin{cases} \phi_M(s,t)_{|P_s}, & \text{ako } s,t \in [a_{j_0-1}, a_{k_0-1}); \\ 0, & \text{inače.} \end{cases}$$

Ako se za $s \in [a_{j_0-1}, a_{k_0-1})$ izabere izomorfizam $\psi_s : \mathbb{F} \to P_s$ (koji postoji, jer su oba vektorska prostora jednodimenzionalna), a za $s \notin [a_{j_0-1}, a_{k_0-1})$ stavi $\psi_s = 0$, tada nije teško provjeriti da kolekcija preslikavanja $\{\psi_s : s \in \mathbb{R}\}$ predstavlja morfizam $\psi : M[a_{j_0-1}, a_{k_0-1}) \Rightarrow P$ koji je prirodni izomorfizam. Zbog toga su istrajni moduli $M[a_{j_0-1}, a_{k_0-1})$ i P izomorfni. Za kompletiranje dokaza dovoljno je dokazati da istrajni modul $W_\sharp := W + P$ (obična suma dva modula) ispunjava tražena svojstva, tj. verifikovati da vrijedi

(i_1) $W_\sharp = W \bigoplus P$,

(i_2) $W_\sharp$ je semi-surjektivan istrajni podmodul istrajnog modula M.

(i_1) Treba provjeriti da je za svako $s \in \mathbb{R}$ ispunjeno $W_s \cap P_s = \{0\}$. Ovo očigledno vrijedi za $s \notin [a_{j_0-1}, a_{k_0-1})$, jer je tada $P_s = \{0\}$. Stoga, neka $s \in [a_{k-1}, a_k) \subseteq [a_{j_0-1}, a_{k_0-1})$. Dovoljno je dokazati da $(y^k)_s \in M_s \setminus W_s$. Neka vrijedi suprotno, tj. neka $(y^k)_s \in W_s$. Uzimajući u obzir da je $a_{j_0-1} = \sup\{t \in \mathbb{R} : W_s = M_s, \text{ za svako } r \leqslant t\}$, kao i činjenicu da su nakon vrijednosti a_{j_0-1} tranziciona preslikavanja istrajnog podmodula W surjektivna, zaključuje se da za proizvoljno t tako da je $a_{j_0-1} \leqslant a_{k-1} \leqslant t < s$ postoji element $w_t \in W_t$ za koji je $(y^k)_s = \phi_W(t,s)(w_t)$. Klasa $\tilde{w} \in W^k$ čiji su predstavnici u svakom $W_t, t \in [a_{k-1}, a_k)$, dati sa

$$(\tilde{w})_t := \begin{cases} w_t, & \text{ako je } a_{k-1} \leqslant t < s; \\ \phi_W(s,t)\left((y^k)_s\right), & \text{ako je } s \leqslant t < a_k \end{cases}$$

je dobro definisana, u smislu da je njena definicija u saglasnosti sa djelovanjem tranzicionih preslikavanja istrajnog podmodula W. Štaviše, vrijedi $\tilde{w} = y^k$, što implicira $y^k \in W^k$. Međutim, ovo je kontradikcija sa pretpostavljenom minimalnošću vrijednosti k_0. Dakle, $(y^k)_s \in M_s \setminus W_s$, za svako $s \in [a_{j_0-1}, a_{k_0-1})$.

(i_2) Prije svega, jasno da je $W_\sharp$ istrajni podmodul istrajnog modula M, kao direktna suma dva istrajna podmodula od M. Kako je $a_{j_0-1} = \sup\{t \in \mathbb{R} : W_s = M_s, \text{ za svako } s \leqslant t\}$ iz definicije istrajnog podmodula P slijedi da za svako $t < a_{j_0-1}$ vrijedi $\left(W_\sharp\right)_t = M_t$. Takođe, za sve $a_{j_0-1} \leqslant s < t$ tranziciono

preslikavanje $\phi_{W_\sharp}(s,t) = \phi_W(s,t) \bigoplus \phi_P(s,t)$ je surjektivno, jer su $\phi_W(s,t)$ i $\phi_P(s,t)$ surjektivna preslikavanja.

(ii) U ovom slučaju, za svako $k > j_0$ vrijedi $z^k \notin W^k$. Konstrukcija istrajnog podmodula P istrajnog modula M za koji vrijedi $W_\sharp \cong W \bigoplus P$ izvodi se na sličan način kao u slučaju (i). Razlika je u tome što se istrajni modul P izomorfan intervalnom istrajnom modulu $M[a_{j_0-1}, +\infty)$ "izgrađuje" uz pomoć klase z^{j_0}: P je sastavljen od vektorskih prostora

$$P_s := \begin{cases} \{0\}, & \text{ako } s < a_{j_0-1}; \\ span_{\mathbb{F}}\left(\left(z^k\right)_s\right), & \text{ako } s \in [a_{k-1}, a_k) \subseteq [a_{j_0-1}, +\infty), k \geqslant j_0 \end{cases}$$

i tranzicionih preslikavanja

$$\phi_P(s,t) := \begin{cases} \phi_M(s,t)_{|P_s}, & \text{ako } a_{j_0-1} \leqslant s \leqslant t; \\ 0, & \text{inače.} \end{cases}$$

□

Dokaz. (Teoreme o intervalnoj dekompoziciji istrajnog modula) Prvo će biti dokazano postojanje, a zatim jedinstvenost intervalne dekompozicije.

Egzistencija slijedi na osnovu prethodne leme. Preciznije, neka je $\{W(i)\}$ familija semi-surjektivnih istrajnih podmodula istrajnog modula M induktivno izgrađena sa

$$\begin{aligned} W(0) &:= \{0\}, \\ W(i+1) &:= W(i)_\sharp, \end{aligned}$$

gdje su $W(i)_\sharp$ semi-surjektivni istrajni podmoduli iz prethodne leme. U svakom koraku dimenzija od $W(i)$ se uvećava bar za jedan, pa se ova izgradnja završava u momentu kada se dostigne $Totaldim(M)$. Na taj način se dobija konačan skup intervala $\mathcal{I}$ za koji vrijedi (2.6).

Za dokaz jedinstvenosti date reprezentacije dovoljno je pokazati da za familije nepraznih intervala $\mathcal{I} = \{I_1, \ldots, I_K\}$ i $\mathcal{J} = \{J_1, \ldots, J_L\}$ za koje vrijedi $\bigoplus_{k=1}^{K} M(I_k) \cong \bigoplus_{l=1}^{L} M(J_l)$ nužno slijedi $K = L$ i $\mathcal{I} = \mathcal{J}$. Ovo će biti dokazano indukcijom po K.

Za $K = 1$, dobija se $M(I_1) \cong \bigoplus_{l=1}^{L} M(J_l)$. Na osnovu Leme 2.4.9, intervalni istrajni modul nije dekompozibilan, što implicira $L = 1$ i $M(I_1) \cong M(J_1)$, odakle slijedi $I_1 = J_1$. Time je potvrđena baza indukcije, tj. da tvrđenje vrijedi za $K = 1$. Neka za prirodan broj $K \geqslant 1$ tvrđenje vrijedi za proizvoljnu familiju od K nepraznih intervala i neka su $\mathcal{I} = \{I_1, \ldots, I_{K+1}\}$, $\mathcal{J} = \{J_1, \ldots, J_L\}$ familije intervala za koje vrijedi $\bigoplus_{k=1}^{K+1} M(I_k) \cong \bigoplus_{l=1}^{L} M(J_l)$. Očigledno je $L > 1$. Ideja je da se dokaže da je interval I_1 jednak intervalu J_{l_0}, za neki

indeks $l_0 \in \{1, \ldots, L\}$, a zatim iskoristi indukciona pretpostavka. Neka su $\psi : \bigoplus_{k=1}^{K+1} M(I_k) \Rightarrow \bigoplus_{l=1}^{L} M(J_l)$ i $\chi : \bigoplus_{l=1}^{L} M(J_l) \Rightarrow \bigoplus_{k=1}^{K+1} M(I_k)$ prirodni izomorfizmi za koje je $\chi \circ \psi$ identični morfizam na $\bigoplus_{k=1}^{K+1} M(I_k)$. Za svako $l \in \{1, \ldots, L\}$, moguće je definisati morfizme $\psi^l : M(I_1) \Rightarrow M(J_l)$ i $\chi^l : M(J_l) \Rightarrow M(I_1)$ sa

$$\psi^l := p^{M(J_l)} \circ \psi \circ e^{M(I_1)},$$
$$\chi^l := p^{M(I_1)} \circ \chi \circ e^{M(J_l)},$$

gdje su e i p odgovarajući morfizmi potapanja i projekcije. Za svako $t \in \mathbb{R}$ i svako $w \in \left(\bigoplus_{l=1}^{L} M(J_l)\right)_t$ vrijedi $\sum_{l=1}^{L} \left(e_t^{M(J_l)} \circ p_t^{M(J_l)}\right)(w) = w$, što implicira da za svako $t \in \mathbb{R}$ vrijedi $\sum_{l=1}^{L} e_t^{M(J_l)} \circ p_t^{M(J_l)} = id_{\bigoplus_{l=1}^{L} M(J_l)_t}$. Na osnovu toga, dalje se dobija da za proizvoljne $t \in \mathbb{R}$ i $v \in M(I_1)_t$ vrijedi

$$\begin{aligned}
\left(\sum_{l=1}^{L} (\chi^l \circ \psi^l)_t\right)(v) &= \sum_{l=1}^{L} \left(\chi_t^l \circ \psi_t^l\right)(v) = \sum_{l=1}^{L} \chi_t^l \circ p_t^{M(J_l)} \circ \underbrace{\psi_t \circ e_t^{M(I_1)}(v)}_{w \in \bigoplus_{l=1}^{L} M(J_l)_t} \\
&= \sum_{l=1}^{L} p_t^{M(I_1)} \circ \chi_t \circ e_t^{M(J_l)} \circ p_t^{M(J_l)}(w) \\
&= p_t^{M(I_1)} \circ \chi_t \circ \left(\sum_{l=1}^{L} e_t^{M(J_l)} \circ p_t^{M(J_l)}\right)(w) \\
&= \left(p_t^{M(I_1)} \circ \chi_t\right)(w) = \left(p_t^{M(I_1)} \circ \underbrace{\chi_t \circ \psi_t}_{id_{\bigoplus_{k=1}^{K+1} M(I_k)_t}} \circ e_t^{M(I_1)}\right)(v) \\
&= \left(p_t^{M(I_1)} \circ e_t^{M(I_1)}\right)(v) = v,
\end{aligned}$$

odakle slijedi da je $\sum_{l=1}^{L} \left(\chi^l \circ \psi^l\right)$ identični morfizam na intervalnom istrajnom modulu $M(I_1)$. To znači da postoji $l_0 \in \{1, \ldots, L\}$ za koje je $\chi^{l_0} \circ \psi^{l_0} \neq 0$. Na osnovu Leme 2.4.4 i Primjedbe 2.4.5 slijedi da je $\left(\chi^{l_0} \circ \psi^{l_0}\right)_t = c \cdot id_{\mathbb{F}}$, za neko $c \in \mathbb{F}$, što implicira da su χ^{l_0} i ψ^{l_0} izomorfizmi. Zbog toga,

vrijedi $M(I_1) \cong M(J_{l_0})$, odakle slijedi $I_1 = J_{l_0}$. Dalje se dobija $\bigoplus_{k=2}^{K+1} M(I_k) \cong \bigoplus_{l\in\{1,\dots,L\}\setminus\{l_0\}} M(J_l)$, pa se, na osnovu indukcione pretpostavke, zaključuje $L = K+1$ i $\{I_2,\dots,I_{K+1}\} = \{J_l : l \in \{1,\dots,L\}\setminus\{l_0\}\}$, što implicira $\mathcal{I} = \mathcal{J}$. □

Primjer 2.5.8. *Za kompleks $\mathcal{K}$ i njegovu filtraciju koji su razmatrani u Primjeru 2.3.1, istrajnom modulu homologije $M(Hom_0(\mathcal{K}))$ se pridružuje intervalna dekompozicija $M[1,+\infty) \bigoplus M[1,2) \bigoplus M[1,2) \bigoplus M[1,4)$, što znači da bar-kod $BC(Hom_0(\mathcal{K}))$ pridružen ovom istrajnom modulu sadrži četiri linije određene intervalima $[1,2)$ (sa višestrukošću 2), $[1,4)$ i $[1,+\infty)$. Takođe, istrajnom modulu homologije $M(Hom_1(\mathcal{K}))$ se pridružuje intervalna dekompozicija $M[3,4) \bigoplus M[3,5)$, što znači da bar-kod $BC(Hom_1(\mathcal{K}))$ pridružen ovom istrajnom modulu sadrži dvije linije određene intervalima $[3,4)$ i $[3,5)$.*

Važno je istaći povezanost ranga preslikavanja $\pi_{i,j}$ sa višestrukošću intervala koji učestvuju u intervalnoj dekompoziciji. Naime, vrijedi $rang(\pi_{i,j}) = \sum m(I)$, pri čemu se posljednja suma uzima po svim intervalima iz reprezentacije koji počinju u ili prije vrijednosti a_{i-1}, a završavaju poslije vrijednosti a_{j-1}. U slučaju istrajnog modula homologije $M(Hom_k)$ prethodna jednakost se svodi na jednakost između istrajnog Betijevog broja $\beta_k^{i,j}$ i veličine $\mu_k^{i,j}$ koja predstavlja broj $k-$dimenzionalnih homoloških klasa koje postoje u rasponu između $i-$tog i $j-$tog kompleksa filtracije. Stoga se, primjenom Teoreme o intervalnoj dekompoziciji u slučaju istrajnog modula homologije dobija sljedeća lema, koja je u literaturi poznata kao *Fundamentalna lema istrajne homologije*.

Lema 2.5.9. *[36] Za svaki par indeksa $i \leqslant j$ i svaku dimenziju k vrijedi*

$$\beta_k^{i,j} = \sum_{i'\leqslant i} \sum_{j'>j} \mu_k^{i',j'}.$$

Primjedba 2.5.10. Linije bar-koda $BC(M)$ dobijene intervalnom dekompozicijom istrajnog modula M mogu se razvrstati u dvije skupine u zavisnosti od njihove dužine. Za proizvoljno $\delta > 0$, neka je $BC^\delta(M)$ multiskup linija iz $BC(M)$ čija je dužina veća od δ. Ako se za istrajne module M i N, svaka linija iz $BC^{2\delta}(M)$ može "upariti" sa linijom iz $BC^{2\delta}(N)$, na način da se njihove krajnje tačke ne razlikuju za više od δ, tada se ispostavlja da ovakvo uparivanje generiše jedno $\delta-$preplitanje istrajnih modula M i N. U narednoj sekciji ove glave data je detaljnija razrada ove ideje, uključujući dokaz pomenute karakterizacije.

2.6 Udaljenost uskog grla i Teorema stabilnosti

Istodimenzionalne bar-kodove dva istrajna modula homologije je moguće porediti. U ovoj sekciji se razmatra najpoznatija mjera ovakvog tipa.

Za dva konačna multiskupa X i Y, njihovo *uparivanje* je relacija $\sigma \subseteq X \times Y$ za koju postoje $X' \subseteq X$ i $Y' \subseteq Y$ takvi da je $\sigma : X' \to Y'$ bijekcija. U tom slučaju, X' se još označava sa $coim(\sigma)$, a Y' sa $im(\sigma)$ i za elemente od X' i Y' se kaže da su *upareni*. Ako se neki element u multiskupu pojavljuje više puta, tada se svaka njegova kopija tretira pojedinačno, tj. može se desiti da se zadavanjem uparivanja samo određen broj njegovih kopija bude uparen.

Za bar-kod $BC(M)$ i $\delta \geqslant 0$, neka je $BC^{\delta}(M) := \{[b, d) \in BC(M) : b + \delta < d\}$. U suštini, razmatranjem skupa $BC^{\delta}(M)$ zanemaruju se linije bar-koda $BC(M)$ čija je dužina manja ili jednaka od δ.

Neka su M, N istrajni moduli konačnog tipa i $BC(M)$, $BC(N)$ odgovarajući multiskupovi bar-kod linija. Za $\delta \geqslant 0$, uparivanje σ za $BC(M)$ i $BC(N)$ se naziva $\delta-$*uparivanjem* ako su ispunjena svojstva

$$
\begin{aligned}
&(i)\ BC^{2\delta}(M) \subseteq coim(\sigma),\\
&(ii)\ BC^{2\delta}(N) \subseteq im(\sigma),\\
&(iii)\ \text{Ako je } [b_2, d_2) = \sigma([b_1, d_1)), \text{ tada je}\\
&\qquad [b_1, d_1) \subseteq [b_2 - \delta, d_2 + \delta),\\
&\qquad [b_2, d_2) \subseteq [b_1 - \delta, d_1 + \delta).
\end{aligned}
$$

Primjećuje se da uslov (iii) prethodne definicije znači da $\delta-$uparivanje dva bar koda uparuje intervale čije se granice razlikuju za najviše δ.

Lema 2.6.1. *Neka su $BC(M)$, $BC(N)$ i $BC(Q)$ bar-kodovi. Ako postoji $\delta-$uparivanje za $BC(M)$ i $BC(N)$ i postoji $\gamma-$uparivanje za $BC(N)$ i $BC(Q)$, tada postoji $\delta+\gamma-$uparivanje za $BC(M)$ i $BC(Q)$.*

Dokaz. Neka je $\sigma \subseteq BC(M) \times BC(N)$ $\delta-$uparivanje, a $\tau \subseteq BC(N) \times BC(Q)$ $\gamma-$uparivanje. Logičan kandidat za $\delta+\gamma-$uparivanje za $BC(M)$ i $BC(Q)$ je relacija $\tau \circ \sigma \subseteq BC(M) \times BC(Q)$, gdje je

$$\tau \circ \sigma = \{(I_M, I_Q) : \text{ postoji } I_N \in BC(N) \text{ tako da } (I_M, I_N) \in \sigma, (I_N, I_Q) \in \tau\}.$$

Neka je $[b_1, d_1) \in BC^{2(\delta+\gamma)}(M)$. Iz $BC^{2(\delta+\gamma)}(M) \subseteq BC^{2\delta}(M) \subseteq coim(\sigma)$ slijedi $[b_1, d_1) \in coim(\sigma)$, pa postoji $[b_2, d_2) \in BC(N)$ takvo da je $[b_2, d_2) = \sigma([b_1, d_1))$. Tada, iz $[b_1, d_1) \subseteq [b_2-\delta, d_2+\delta)$ slijedi $b_2+2\gamma < d_2$, što implicira $[b_2, d_2) \in BC^{2\gamma}(N) \subseteq coim(\tau)$. To znači da postoji $[b_3, d_3) \in BC(Q)$ sa svojstvom $[b_3, d_3) = \tau([b_2, d_2))$, pa $[b_3, d_3) = (\tau \circ \sigma)([b_1, d_1))$, odakle slijedi $[b_1, d_1) \in coim(\tau \circ \sigma)$. Time je dokazano $BC^{2(\delta+\gamma)}(M) \subseteq coim(\tau \circ \sigma)$, a inkluzija $BC^{2(\delta+\gamma)}(Q) \subseteq im(\tau \circ \sigma)$ se dokazuje na sličan način. Potrebno je još provjeriti svojstvo (iii). Neka je $[b_2, d_2) = (\tau \circ \sigma)([b_1, d_1))$. Tada za $[b_3, d_3) := \sigma([b_1, d_1))$ vrijedi $[b_2, d_2) = \tau([b_3, d_3))$, pa se dobija

$$
\begin{aligned}
&[b_1, d_1) \subseteq [b_3 - \delta, d_3 + \delta) \subseteq [b_2 - (\delta + \gamma), d_2 + \delta + \gamma),\\
&[b_2, d_2) \subseteq [b_3 - \gamma, d_3 + \gamma) \subseteq [b_1 - (\delta + \gamma), d_1 + \delta + \gamma).
\end{aligned}
$$

□

Ako su $BC(M)$ i $BC(N)$ bar-kodovi i $U_{BC(M),BC(N)}$ skup svih nenegativnih realnih brojeva δ za koje postoji $\delta-$uparivanje za $BC(M)$ i $BC(N)$, tada se ovim bar-kodovima može pridružiti *udaljenost uskog grla*, koja se definiše na sljedeći način:

$$d_{BOT}\left(BC(M), BC(N)\right) := \begin{cases} \inf U_{BC(M),BC(N)}, & \text{ako je } U_{BC(M),BC(N)} \neq \emptyset; \\ +\infty, & \text{inače.} \end{cases}$$

Lema 2.6.2. *Udaljenost uskog grla je produžena metrika na familiji bar-kodova.*

Dokaz. Neka su $BC(M)$, $BC(N)$ i $BC(Q)$ proizvoljni bar-kodovi. Očigledno da je svaka linija bar-koda $BC(M)$ $0-$uparena sa samom sobom, što znači da je $d_{BOT}(BC(M), BC(M)) = 0$. Ako je $d_{BOT}(BC(M), BC(N)) = 0$ i ukoliko se pretpostavi da postoji interval $[b, d) \in BC(M)$ koji nije zastupljen u $BC(N)$, tada, za svako $\delta \in \left(0, \frac{d-b}{2}\right)$, postoji interval $[b_\delta, d_\delta) \in BC(N)$ različit od $[b, d)$ koji je $\delta-$uparen sa $[b, d)$, tj. za koji vrijedi $|b - b_\delta| \leqslant \delta$ i $|d - d_\delta| \leqslant \delta$. Međutim, $BC(N)$ je multiskup sa konačno mnogo intervala (uključujući i višestrukost), pa prethodne nejednakosti nije moguće postići za svako $\delta \in \left(0, \frac{d-b}{2}\right)$. Zbog toga, iz $d_{BOT}(BC(M), BC(N)) = 0$ nužno slijedi $BC(M) = BC(N)$. Ako je σ jedno $\delta-$uparivanje za $BC(M)$ i $BC(N)$, tada je relacija σ^{-1} jedno $\delta-$uparivanje za $BC(N)$ i $BC(M)$, što implicira $d_{BOT}(BC(M), BC(N)) = d_{BOT}(BC(N), BC(M))$. Za kompletiranje dokaza, potrebno je još dokazati nejednakost trougla, tj. da vrijedi

$$d_{BOT}(BC(M), BC(Q)) \leqslant d_{BOT}(BC(M), BC(N)) + d_{BOT}(BC(N), BC(Q)).$$

U slučaju da je $d_{BOT}(BC(M), BC(N)) = +\infty$ ili $d_{BOT}(BC(N), BC(Q)) = +\infty$, prethodna nejednakost trivijalno vrijedi. Stoga se može pretpostaviti da su obje vrijednosti konačne i neka je $a := d_{BOT}(BC(M), BC(N)) \in [0, +\infty)$ i $b := d_{BOT}(BC(N), BC(Q)) \in [0, +\infty)$. Za proizvoljno $\varepsilon > 0$, mogu se izabrati $\delta \in [a, a+\varepsilon)$ i $\gamma \in [b, b+\varepsilon)$ tako da postoji $\delta-$uparivanje za $BC(M)$ i $BC(N)$, kao i $\gamma-$uparivanje za $BC(N)$ i $BC(Q)$. Na osnovu prethodne leme, postoji $\delta+\gamma-$uparivanje za $BC(M)$ i $BC(Q)$, što implicira $d_{BOT}(BC(M), BC(Q)) \leqslant \delta + \gamma < a + b + 2\varepsilon$. Ova nejednakost vrijedi za svako $\varepsilon > 0$, pa iz nje slijedi nejednakost trougla. □

Primjer 2.6.3. *Za $b_1, d_1, b_2, d_2 \in \mathbb{R}$ i intervalne istrajne module $M[b_1, d_1)$, $M[b_2, d_2)$, vrijedi $BC(M[b_1, d_1)) = \{[b_1, d_1)\}$ i $BC(M[b_2, d_2)) = \{[b_2, d_2)\}$. Ako je $\delta_1 := \max\left\{\frac{d_1 - b_1}{2}, \frac{d_2 - b_2}{2}\right\}$, tada je $\sigma_1 = \emptyset$ jedno δ_1-uparivanje, jer dužine posmatranih intervala nisu veće od $2\delta_1$. Takođe, za $\delta_2 := \max\{|b_1 -$*

$b_2|, |d_1 - d_2|\}$, uparivanje σ_2 koje intervalu $[b_1, d_1)$ pridružuje interval $[b_2, d_2)$ je δ_2-uparivanje, pa je $d_{BOT}\big(BC(M[b_1, d_1)), BC(M[b_2, d_2))\big) \leqslant \min\{\delta_1, \delta_2\}$, a, s obzirom da osim datih nema drugih uparivanja, vrijedi

$$d_{BOT}\big(BC(M[b_1, d_1)), BC(M[b_2, d_2))\big) = \min\{\delta_1, \delta_2\}.$$

Lema 2.6.4. *Neka je $\delta \in [0, +\infty)$ i $[b_1, d_1)$, $[b_2, d_2)$, $b_1, b_2 \in \mathbb{R}$, $d_1, d_2 \in \overline{\mathbb{R}}$, intervali koji su $\delta-$upareni. Tada su intervalni istrajni moduli $M[b_1, d_1)$ i $M[b_2, d_2)$ $\delta-$prepleteni.*

Dokaz. Iz relacija

$$[b_1, d_1) \subseteq [b_2 - \delta, d_2 + \delta),$$
$$[b_2, d_2) \subseteq [b_1 - \delta, d_1 + \delta),$$

slijedi da su intervali $[b_1, d_1)$ i $[b_2, d_2)$ istog tipa tj. oba su ograničena ili su oba neograničena odozgo. Iz Leme 2.4.16 slijedi postojanje δ_1-preplitanja intervalnih istrajnih modula $M[b_1, +\infty)$ i $M[b_2, +\infty)$, gdje je $\delta_1 = |b_1 - b_2|$, kao i δ_2-preplitanja intervalnih istrajnih modula $M[b_1, d_1)$ i $M[b_2, d_2)$, u slučaju $b_1, b_2, d_1, d_2 \in \mathbb{R}$, gdje je $\delta_2 = \max\{|b_1 - b_2|, |d_1 - d_2|\}$. S obzirom da je $\delta_1, \delta_2 \leqslant \delta$, u oba slučaja su posmatrani intervalni istrajni moduli $\delta-$prepleteni. □

Uopštavanjem prethodne leme dobija se sljedeća lema.

Lema 2.6.5. *[6] Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. Ako postoji konačno $\delta-$uparivanje bar-kodova $BC(M)$ i $BC(N)$, tada su M i N $\delta-$prepleteni istrajni moduli. Specijalno, vrijedi $d_{INT}(M, N) \leqslant d_{BOT}\big(BC(M), BC(N)\big)$.*

Dokaz. Na osnovu Teoreme o intervalnoj dekompoziciji, vrijedi

$$M = \bigoplus_{I \in BC(M)} M(I), \quad N = \bigoplus_{J \in BC(N)} M(J).$$

Neka je σ $\delta-$uparivanje bar-kodova $BC(M)$ i $BC(N)$. Ovo uparivanje određuje istrajne module

$$M^{(1)} := \bigoplus_{I \in coim(\sigma)} M(I), \quad N^{(1)} := \bigoplus_{J \in im(\sigma)} M(J),$$
$$M^{(2)} := \bigoplus_{I \in BC(M) \setminus coim(\sigma)} M(I), \quad N^{(2)} := \bigoplus_{J \in BC(N) \setminus im(\sigma)} M(J)$$

i očigledno vrijedi $M = M^{(1)} \bigoplus M^{(2)}$, $N = N^{(1)} \bigoplus N^{(2)}$. Na osnovu prethodne leme, za svaki par (I, J) $\delta-$uparenih intervala postoji par morfizama

$\psi^{I,J} : M(I) \Rightarrow M(J)(\delta)$, $\chi^{I,J} : M(J) \Rightarrow M(I)(\delta)$ koji čine $\delta-$preplitanje intervalnih istrajnih modula $M(I)$ i $M(J)$. Uz pomoć ovih parova indukuju se par $\delta-$prepletenih morfizama $\psi^{(1)} : M^{(1)} \Rightarrow N^{(1)}(\delta)$ i $\chi^{(1)} : N^{(1)} \Rightarrow M^{(1)}(\delta)$. Za svako $I \in BC(M) \setminus coim(\sigma)$ vrijedi $I \notin BC^{2\delta}(M)$, pa, za svaki ovakav interval I, intervalni istrajni modul $M(I)$ je $\delta-$prepleten sa nula istrajnim modulom. Stoga je istrajni modul $M^{(2)}$ $\delta-$prepleten sa nula istrajnim modulom, a isti zaključak vrijedi i za istrajni modul $N^{(2)}$. Sve ovo omogućava definisanje morfizama $\psi : M \Rightarrow N(\delta)$ i $\chi : N \Rightarrow M(\delta)$ određenih restrikcijama: $\psi_{|M^{(1)}} = \psi^{(1)}$, $\psi_{|M^{(2)}} = 0$, $\chi_{|N^{(1)}} = \chi^{(1)}$, $\chi_{|N^{(2)}} = 0$. Za ovaj par morfizama se jednostavno dokazuje da čini $\delta-$preplitanje istrajnih modula M i N.

Iz pretpostavke $d_{INT}(M,N) > d_{BOT}\big(BC(M), BC(N)\big)$ bi slijedilo postojanje $\delta \in \big(d_{BOT}\big(BC(M), BC(N)\big), d_{INT}(M,N)\big)$ za koji su bar-kodovi $BC(M)$ i $BC(N)$ $\delta-$upareni, a istrajni moduli M i N nisu $\delta-$prepleteni, što, prema prethodnom, nije moguće. □

Ispostavlja se da nejednakost između udaljenosti preplitanja i udaljenosti uskog grla uspostavljena u prethodnoj lemi ustvari predstavlja jednakost. Na taj način uspostavljena "izometrija" garantuje numeričku stabilnost udaljenosti uskog grla, u smislu da su promjene istrajnog modula (izražene u terminima udaljenosti preplitanja) iste veličine kao promjene njegovog bar-koda (izražene u terminima udaljenosti uskog grla). U narednom je izložen dokaz suprotne nejednakosti koji implicira pomenutu izometriju. Ovaj dokaz je originalno izveden u [6], a njegova osnovna ideja je korišćenje tzv. *indukovanog uparivanja*.

Neka je M istrajni modul koji zadovoljava svojstvo konačnog tipa i $BC(M)$ njegov bar-kod. Za interval $[b,d)$, $b \in \mathbb{R}, d \in \overline{\mathbb{R}}$, neka je $BC^{-[b,d)}(M) \subseteq BC(M)$ multiskup koji se sastoji od intervala oblika $[b',d)$, gdje je $b' \leqslant b$. Slikovito rečeno, multiskup $BC^{-[b,d)}(M)$ sastoji se od linija čiji je početak najkasnije u b, a kraj tačno u d.

Morfizam $\psi : M \Rightarrow N$ je *monomorfizam*(*epimorfizam*), ako je za svako $t \in \mathbb{R}$ preslikavanje $\psi_t : M_t \to N_t$ injektivno (surjektivno).

Lema 2.6.6. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. Ako postoji monomorfizam $\psi : M \Rightarrow N$, tada za svaki interval $[b,d)$ vrijedi $\big|BC^{-[b,d)}(M)\big| \leqslant \big|BC^{-[b,d)}(N)\big|$.*

Dokaz. Neka su U i V istrajni moduli u čijim intervalnim dekompozicijama učestvuju isključivo linije iz multiskupova $BC^{-[b,d)}(M)$ i $BC^{-[b,d)}(N)$, tj. neka je

$$U = \bigoplus_{I \in BC^{-[b,d)}(M)} M(I),$$
$$V = \bigoplus_{I \in BC^{-[b,d)}(N)} M(I).$$

Kako istrajni moduli M i N zadovoljavaju svojstvo konačnog tipa, postoji vrijednost $t \in (b,d)$ sa svojstvom da, za svaki interval $[b',d') \in BC(M) \cup BC(N)$, gdje je $b' \leqslant b$ i $d' < d$, vrijedi $t > d'$. Očigledno vrijedi $dim(U_t) = \left|BC^{-[b,d)}(M)\right|$ i $dim(V_t) = \left|BC^{-[b,d)}(N)\right|$, pa, kako je ψ monomorfizam, dovoljno je dokazati inkluziju $\psi_t\,(U_t) \subseteq V_t$.

Uočava se da vektorski prostor U_t čine oni elementi koji dolaze iz svih prostora M_s, za $s \in [b,t)$, a koji nestaju u svim prostorima M_r, za $r \geqslant d$. Zbog toga, vrijedi $U_t = \bigcap\limits_{b \leqslant s < t} im\,(\phi_M(s,t)) \cap \bigcap\limits_{r \geqslant d} ker\,(\phi_M(d,r))$, a sličnim rezonovanjem se dobija $V_t = \bigcap\limits_{b \leqslant s < t} im\,(\phi_N(s,t)) \cap \bigcap\limits_{r \geqslant d} ker\,(\phi_N(d,r))$, gdje su ϕ_M i ϕ_N redom tranziciona preslikavanja istrajnih modula M i N. Dijagram

$$\begin{array}{ccc} M_s & \xrightarrow{\phi_M(s,t)} & M_t \\ \downarrow{\scriptstyle \psi_s} & & \downarrow{\scriptstyle \psi_t} \\ N_s & \xrightarrow{\phi_N(s,t)} & N_t \end{array}$$

je komutativan za svako $s \in [b,t]$, što implicira da za svako takvo s vrijedi $\psi_t\big(im(\phi_M(s,t))\big) \subseteq im\big(\phi_N(s,t)\big)$ i $\psi_s\big(ker(\phi_M(s,t))\big) \subseteq ker\big(\phi_N(s,t)\big)$. Analogno se zaključuje da za svako $r \geqslant d$ vrijedi $\psi_t\big(im(\phi_M(t,r))\big) \subseteq im\big(\phi_N(t,r)\big)$ i $\psi_s\big(ker(\phi_M(t,r))\big) \subseteq ker\big(\phi_N(t,r)\big)$. Zbog toga je ispunjeno $\psi_t\,(U_t) \subseteq V_t$. □

Dualno definiciji $BC^{-[b,d)}(M)$ uvodi se multiskup $BC^{[b,d)+}(M) \subseteq BC(M)$ koji se sastoji od intervala oblika $[b,d')$, gdje je $d \leqslant d'$. Postupkom sličnim kao u dokazu prethodne leme dokazuje se sljedeća lema.

Lema 2.6.7. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. Ako postoji epimorfizam $\psi : M \Longrightarrow N$, tada za svaki interval $[b,d)$ vrijedi $\left|BC^{[b,d)+}(M)\right| \geqslant \left|BC^{[b,d)+}(N)\right|$.*

Za bar-kod $BC(M)$ i $d \in \overline{\mathbb{R}}$, neka je $BC^{[\cdot,d)}(M) \subseteq BC(M)$ multiskup linija koji se sastoji od linija iz $BC(M)$ oblika $[b,d)$, za neko $b \in \mathbb{R}$. Na ovom multiskupu se može posmatrati totalno uređenje u kome je interval $[b_2,d)$ "veći" od intervala $[b_1,d)$ ukoliko je $[b_2,d) \subseteq [b_1,d)$, dok se kopije istog intervala porede u skladu sa leksikografskim poretkom na reprezentaciji multiskupa $BC^{[\cdot,d)}(M)$.

Lema 2.6.8. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. Ako postoji monomorfizam modula M u modul N, tada za svako $d \in \overline{\mathbb{R}}$ vrijedi $\left|BC^{[\cdot,d)}(M)\right| \leqslant \left|BC^{[\cdot,d)}(N)\right|$.*

Dokaz. Dovoljno je dokazati da za svaki prirodan broj $1 \leqslant i \leqslant \left|BC^{[\cdot,d)}(M)\right|$ vrijedi $i \leqslant \left|BC^{[\cdot,d)}(N)\right|$. Neka je $i \leqslant \left|BC^{[\cdot,d)}(M)\right|$ proizvoljno i $[b,d)$ $i-$ti po veličini interval iz $BC^{[\cdot,d)}(M)$ u odnosu na prethodno uvedeno totalno uređenje.

Kako za svako $1 \leqslant j \leqslant i$, multiskup $BC^{-[b,d)}(M)$ sadrži $j-$ti po veličini interval iz $BC^{[\cdot,d)}(M)$, zaključuje se da vrijedi

$$i \leqslant \left|BC^{-[b,d)}(M)\right| \leqslant \left|BC^{-[b,d)}(N)\right| \leqslant \left|BC^{[\cdot,d)}(N)\right|,$$

pri čemu druga nejednakost slijedi na osnovu Leme 2.6.6. □

Dualno definiciji $BC^{[\cdot,d)}(M)$ uvodi se multiskup $BC^{[b,\cdot)}(M) \subseteq BC(M)$ koji se sastoji od linija iz $BC(M)$ oblika $[b, d)$, za neko $d \in \overline{\mathbb{R}}$. Postupkom sličnim kao u dokazu prethodne leme dokazuje se sljedeća lema.

Lema 2.6.9. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. Ako postoji epimorfizam modula M u modul N, tada za svako $b \in \mathbb{R}$ vrijedi $\left|BC^{[b,\cdot)}(M)\right| \geqslant \left|BC^{[b,\cdot)}(N)\right|$.*

Prethodna tvrđenja omogućavaju da se uvede uparivanje bar-kodova istrajnih modula između kojih se može uspostaviti monomorfizam ili epimorfizam.

Neka su najprije M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa, pri čemu postoji monomorfizam iz M u N. Za proizvoljno $d \in \overline{\mathbb{R}}$, linije multiskupova $BC^{[\cdot,d)}(M)$ i $BC^{[\cdot,d)}(N)$ mogu se predstaviti lancima inkluzija

$$[b'_1, d) \supseteq [b'_2, d) \supseteq \ldots \supseteq [b'_k, d), \;\; b'_1 \leq b'_2 \leqslant \ldots \leqslant b'_k,$$
$$[b''_1, d) \supseteq [b''_2, d) \supseteq \ldots \supseteq [b''_l, d), \;\; b''_1 \leq b''_2 \leqslant \ldots \leqslant b''_l,$$

za neke prirodne brojeve k, l. Na osnovu Leme 2.6.8, vrijedi $k \leqslant l$, te se, za svako $i \in \{1, 2, \ldots, k\}$, intervalu $[b'_i, d)$ može pridružiti interval $[b''_i, d)$. Dakle, $i-$ta po veličini linija iz $BC^{[\cdot,d)}(M)$ se uparuje sa $i-$tom po veličini linijom iz $BC^{[\cdot,d)}(N)$. Ukoliko se ovo pridruživanje izvrši za svako $d \in \overline{\mathbb{R}}$ dobija se *injektivno indukovano uparivanje* koje će biti označeno sa σ_{inj}.

Lema 2.6.10. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa, pri čemu postoji monomorfizam iz M u N. Indukovano uparivanje $\sigma_{inj} \subseteq BC(M) \times BC(N)$ ima sljedeća svojstva*

(*i*) $coim\left(\sigma_{inj}\right) = BC(M)$,

(*ii*) *Za svako* $[b', d) \in BC(M), \sigma_{inj}\left([b', d)\right) = [b'', d)$, *gdje je* $b'' \leqslant b'$.

Dokaz. Iz definicije indukovanog uparivanja slijedi da je svaka linija iz $BC(M)$ uparena sa nekom linijom iz $BC(N)$, pa indukovano uparivanje zadovoljava svojstvo (i). Takođe, jasno je da se linija $[b', d) \in BC(M)$ uparuje sa linijom oblika $[b'', d) \in BC(N)$. Ova linija je $i-$ta po veličini iz $BC^{[\cdot,d)}(N)$, pa, iz $\left|BC^{-[b',d)}(M)\right| \leqslant \left|BC^{-[b',d)}(N)\right|$ slijedi $[b'', d) \in BC^{-[b',d)}(N)$, što implicira $b'' \leqslant b'$, te indukovano uparivanje zadovoljava i svojstvo (ii). □

Lema 2.6.11. *Neka su M, N, Q istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $\psi : M \Rightarrow N$, $\chi : N \Rightarrow Q$, $\varphi : M \Rightarrow Q$ monomorfizmi. Ako su $\sigma_{inj}(\psi)$, $\sigma_{inj}(\chi)$, $\sigma_{inj}(\varphi)$ odgovarajuća injektivno indukovana uparivanja, tada je sljedeći dijagram komutativan*

$$\begin{array}{ccccc} BC(M) & \xrightarrow{\sigma_{inj}(\psi)} & BC(N) & \xrightarrow{\sigma_{inj}(\chi)} & BC(Q) \\ & \searrow & & \nearrow & \\ & & \sigma_{inj}(\varphi) & & \end{array}$$

Dokaz. Neka je $d \in \overline{\mathbb{R}}$ proizvoljno. Na osnovu Leme 2.6.8 linije bar-kodova $BC(M)$, $BC(N)$ i $BC(Q)$ koje završavaju sa d mogu se sortirati na sljedeći način:

$$\begin{aligned} &BC(M) : [b'_1, d) \supseteq [b'_2, d) \supseteq \ldots \supseteq [b'_k, d), \\ &BC(N) : [b''_1, d) \supseteq [b''_2, d) \supseteq \ldots \supseteq [b''_k, d) \supseteq \ldots \supseteq [b''_l, d), \\ &BC(Q) : [b'''_1, d) \supseteq [b'''_2, d) \supseteq \ldots \supseteq [b'''_k, d) \supseteq \ldots \supseteq [b'''_l, d) \supseteq \ldots \supseteq [b'''_n, d), \end{aligned}$$

za neke $k \leqslant l \leqslant n$. Pritom, za svako $i \in \{1, 2, \ldots, k\}$, vrijedi $\sigma_{inj}(\psi)\big([b'_i, d)\big) = [b''_i, d)$, $\sigma_{inj}(\chi)\big([b''_i, d)\big) = [b'''_i, d)$ i $\sigma_{inj}(\varphi)\big([b'_i, d)\big) = [b'''_i, d)$. Ovo važi za svako $d \in \overline{\mathbb{R}}$, što upućuje na zaključak da je predstavljeni dijagram komutativan. □

Neka su sada M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa, pri čemu postoji epimorfizam iz M u N. Za proizvoljno $b \in \mathbb{R}$, linije multiskupova $BC^{[b,\cdot)}(M)$ i $BC^{[b,\cdot)}(N)$ mogu se predstaviti lancima inkluzija

$$\begin{aligned} &[b, d'_1) \supseteq [b, d'_2) \supseteq \ldots \supseteq [b, d'_k), \;\; d'_1 \geqslant d'_2 \geqslant \ldots \geqslant d'_k, \\ &[b, d''_1) \supseteq [b, d''_2) \supseteq \ldots \supseteq [b, d''_l), \;\; d''_1 \geqslant d''_2 \geqslant \ldots \geqslant d''_l, \end{aligned}$$

za neke prirodne brojeve k, l. Na osnovu Leme 2.6.9, vrijedi $k \geqslant l$, te se, za svako $i \in \{1, 2, \ldots, l\}$, interval $[b, d''_i)$ može pridružiti intervalu $[b, d'_i)$. Dakle, $i-$ta po veličini linija iz $BC^{[b,\cdot)}(N)$ je uparena sa $i-$tom po veličini linijom iz $BC^{[b,\cdot)}(M)$. Ukoliko se ovo pridruživanje izvrši za svako $b \in \mathbb{R}$ dobija se *surjektivno indukovano uparivanje* koje će biti označeno sa σ_{sur}. Analogno slučaju injektivnog indukovanog uparivanja dokazuju se sljedeće dvije leme.

Lema 2.6.12. *Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa, pri čemu postoji epimorfizam iz M u N. Indukovano uparivanje $\sigma_{sur} \subseteq BC(M) \times BC(N)$ ima sljedeća svojstva*

(*i*) $im\,(\sigma_{sur}) = BC(N)$,
(*ii*) *Za svako* $[b, d'') \in BC(N)$, $\sigma_{sur}\,([b, d')) = [b, d'')$, *gdje je* $d' \geqslant d''$.

Lema 2.6.13. *Neka su M, N, Q istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $\psi : M \Rightarrow N$, $\chi : N \Rightarrow Q$, $\varphi : M \Rightarrow Q$ epimorfizmi. Ako su $\sigma_{sur}(\psi)$, $\sigma_{sur}(\chi)$, $\sigma_{sur}(\varphi)$ odgovarajuća surjektivno indukovana uparivanja, tada je sljedeći dijagram komutativan*

$$BC(M) \xrightarrow{\sigma_{sur}(\psi)} BC(N) \xrightarrow{\sigma_{sur}(\chi)} BC(Q), \qquad BC(M) \xrightarrow{\sigma_{sur}(\varphi)} BC(Q)$$

Primjedba 2.6.14. Treba istaći da definicija i svojstva injektivnog i surjektivnog indukovanog uparivanja ne zavise od izbora konkretnog monomorfizma i epimorfizma, već je jedino bitno postojanje ovakvih morfizama.

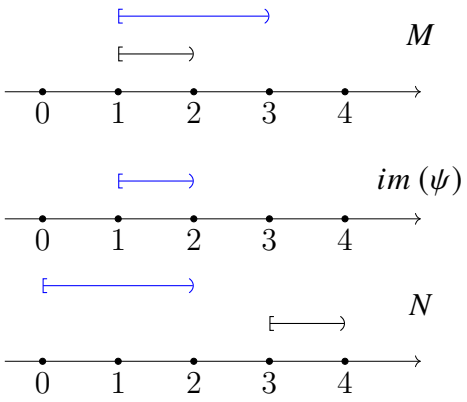
Sada će biti konstruisano indukovano uparivanje u slučaju proizvoljnog morfizma dva istrajna modula.

Neka su M i N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i neka je $\psi : M \Rightarrow N$ morfizam (koji nije obavezno monomorfizam ili epimorfizam). Ovim morfizmom su na prirodan način indukovani epimorfizam iz M u $im(\psi)$ i potapanje $im(\psi)$ u N koje je monomorfizam. Ako je $BC(im(\psi))$ bar-kod istrajnog modula $im(\psi)$, tada se mogu posmatrati surjektivno indukovano uparivanje $\sigma_{sur} \subseteq BC(M) \times BC(im(\psi))$ i injektivno indukovano uparivanje $\sigma_{inj} \subseteq BC(im(\psi)) \times BC(N)$. Uparivanje koje se definiše kao kompozicija $\sigma_{inj} \circ \sigma_{sur}$ ova dva uparivanja naziva se *uparivanje indukovano morfizmom* ψ i biće notirano sa σ_ψ.

Primjer 2.6.15. *Neka je $M = M[1,3) \bigoplus M[1,2)$, $N = M[3,4) \bigoplus M[0,2)$ i $\psi : M \Rightarrow N$ morfizam određen restrikcijama na sumandima modula M koje su date sa $\psi_{|M[1,3)} = 0$ i $\left(\psi_{|M[1,2)}\right)_t = \begin{cases} id_{\mathbb{F}}, & za\ t \in [1,2); \\ 0, & inače \end{cases}$. Tada je $im(\psi) \cong M[1,2)$, pa su odgovarajući bar-kodovi $BC(M) = \{[1,3), [1,2)\}$, $BC(N) = \{[3,4), [0,2)\}$ i $BC(im(\psi)) = \{[1,2)\}$. Dalje je $\sigma_{sur} = \{([1,3), [1,2))\}$ (jer je $[1,3)$ najduži interval iz $BC(M)$ koji počinje u 1) i $\sigma_{inj} = \{([1,2), [0,2))\}$, što znači da vrijedi $\sigma_\psi = \{([1,3), [0,2))\}$, bez obzira što je $\psi_{|M[1,3)} = 0$ (Slika 2.11).*

Neka su M i N $\delta-$prepleteni istrajni moduli koji zadovoljavaju svojstvo konačnog tipa. U narednom je opisana konstrukcija $\delta-$uparivanja između bar-kodova $BC(M)$ i $BC(N)$. Najprije će biti dokazano nekoliko pomoćnih rezultata.

Lema 2.6.16. *[6] Neka su M i N $\delta-$prepleteni istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $\psi : M \Rightarrow N(\delta)$, $\chi : N \Rightarrow M(\delta)$ par morfizama koji čine $\delta-$preplitanje, tj. za koje vrijedi $\chi(\delta) \circ \psi = \Phi_M^{2\delta}$ i $\psi(\delta) \circ \chi = \Phi_N^{2\delta}$. Ako*



Slika 2.11

je f surjektivna restrikcija morfizma ψ na $im(\psi)$, tada surjektivno indukovano uparivanje $\sigma_{sur}(f)$ ima sljedeća svojstva:

(i) $BC^{2\delta}(M) \subseteq coim(\sigma_{sur}(f))$,
(ii) $im(\sigma_{sur}(f)) = BC(im(\psi))$,
(iii) *Za* $[b,d) \in coim(\sigma_{sur}(f))$, $[b,d) \overset{\sigma_{sur}(f)}{\to} [b,d')$, *gdje je* $d - 2\delta \leqslant d' \leqslant d$.

Dokaz. (i) Neka je $g(\delta) = \chi(\delta)_{|im(\psi)} : im(\psi) \Rightarrow M(2\delta)$. Iz datih pretpostavki slijedi komutativnost dijagrama

$$M \xrightarrow{f} im(\phi) \xrightarrow{g(\delta)} im\left(\Phi_M^{2\delta}\right), \qquad M \xrightarrow{\Phi_M^{2\delta}} im\left(\Phi_M^{2\delta}\right)$$

pa, kako su f i $\Phi_M^{2\delta}$ epimorfizmi, zaključuje se da je i $g(\delta)$ epimorfizam. To, na osnovu Leme 2.6.13, implicira da je sljedeći dijagram takođe komutativan.

$$BC(M) \xrightarrow{\sigma_{sur}(f)} BC(im(\psi)) \xrightarrow{\sigma_{sur}(g(\delta))} BC(im(\Phi_M^{2\delta})), \qquad BC(M) \xrightarrow{\sigma_{sur}(\Phi_M^{2\delta})} BC(im(\Phi_M^{2\delta}))$$

Iz ove komutativnosti slijedi $coim\left(\sigma_{sur}(\Phi_M^{2\delta})\right) \subseteq coim\left(\sigma_{sur}(f)\right)$, pa je dovoljno dokazati $coim\left(\sigma_{sur}(\Phi_M^{2\delta})\right) = BC^{2\delta}(M)$. Ova jednakost slijedi iz definicije uparivanja σ_{sur}, kao i činjenice da se $im(\Phi_M^{2\delta})$ dobija "odsjecanjem" M sa

desna za 2δ, jer se linija $[b,d) \in BC(M)$ uparuje sa linijom $[b,d-2\delta) \in BC(im(\Phi_M^{2\delta}))$ ako i samo ako je $d-b>2\delta$.

(ii) Data jednakost slijedi iz dijela (i) Leme 2.6.12.

(iii) Neka je $[b,d) \in coim(\sigma_{sur}(f))$ proizvoljno. Treba ispitati da li linija $\sigma_{sur}(f)\big([b,d)\big)$ ima navedena svojstva. Moguće je

(iii_1) $d-b>2\delta$

Na osnovu dijela (ii) Leme 2.6.12, vrijedi $\sigma_{sur}(f)\big([b,d)\big) = [b,d')$, za neko $d' \leqslant d$, odnosno $\sigma_{sur}(g(\delta))\big([b,d')\big) = [b,d'')$, za neko $d'' \leqslant d'$. S druge strane, vrijedi $\sigma_{sur}(\Phi_M^{2\delta})\big([b,d)\big) = [b,d-2\delta)$, pa, iz komutativnosti prethodnog dijagrama, slijedi $[b,d'') = [b,d-2\delta)$, tj. $d'' = d-2\delta$, pri čemu je $d'' = d-2\delta \leqslant d' \leqslant d$.

(iii_2) $d-b \leqslant 2\delta$

Ponovo se, na osnovu dijela (ii) Leme 2.6.12, zaključuje $\sigma_{sur}(f)\big([b,d)\big) = [b,d')$, za neko $d' \leqslant d$, pa se direktno dobija $d \geqslant d' > b \geqslant d-2\delta$. □

Lema 2.6.17. *[6] Neka su M i N $\delta-$prepleteni istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $\psi : M \Rightarrow N(\delta)$, $\chi : N \Rightarrow M(\delta)$ par morfizama koji čine $\delta-$preplitanje, tj. za koje vrijedi $\chi(\delta)\circ\psi = \Phi_M^{2\delta}$ i $\psi(\delta)\circ\chi = \Phi_N^{2\delta}$. Ako je f prirodna injekcija $im(\psi) \hookrightarrow N(\delta)$, tada injektivno indukovano uparivanje $\sigma_{inj}(f)$ ima sljedeća svojstva:*

(i) $coim(\sigma_{inj}(f)) = BC(im(\psi))$,

(ii) $BC^{2\delta}(N(\delta)) \subseteq im(\sigma_{inj}(f))$,

(iii) *Za* $[b,d) \in BC(im(\psi))$, $[b,d) \overset{\sigma_{inj}(f)}{\rightarrow} [b',d)$, *gdje je* $b-2\delta \leqslant b' \leqslant b$.

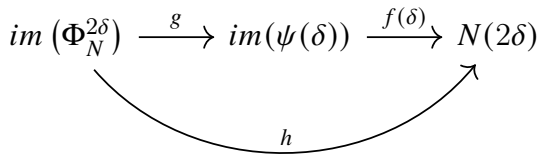
Dokaz. (i) Data jednakost slijedi iz dijela (i) Leme 2.6.10.

U svrhu nešto jednostavnije notacije, dokaz preostala dva svojstva je izložen u varijanti u kojoj je sve "pomjereno" za δ.

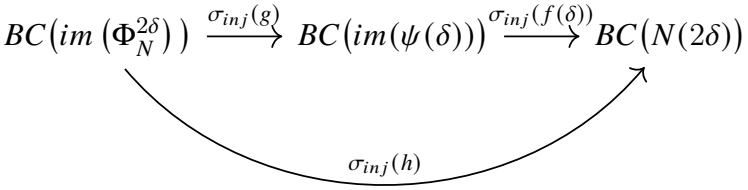
(ii) Biće verifikovana inkluzija $BC^{2\delta}(N(2\delta)) \subseteq im(\sigma_{inj}(f(\delta)))$, uz napomenu da se bar-kod $BC^{2\delta}(N(\delta))$ dobija iz bar-koda $BC^{2\delta}(N(2\delta))$ tako što se sve njegove linije transliraju ulijevo za δ. Iz pretpostavke $\psi(\delta)\circ\chi = \Phi_N^{2\delta}$ slijedi komutativnost dijagrama

$$\begin{array}{ccccc} N & \xrightarrow{\chi} & im(\chi) & \xrightarrow{\psi(\delta)} & N(2\delta) \\ & \searrow & & \nearrow & \\ & & \Phi_N^{2\delta} & & \end{array}$$

odakle slijedi $im\left(\Phi_N^{2\delta}\right) \subseteq im\left(\psi(\delta)\right) \subseteq N(2\delta)$. Neka je g prirodna injekcija $im\left(\Phi_N^{2\delta}\right) \hookrightarrow im\left(\psi(\delta)\right)$, a h prirodna injekcija $im\left(\Phi_N^{2\delta}\right) \hookrightarrow N(2\delta)$. S obzirom da je $f(\delta)$ prirodna injekcija $im\left(\psi(\delta)\right) \hookrightarrow N(2\delta)$, sljedeći dijagram je komutativan



pa, kako su svi morfizmi iz ovog dijagrama monomorfizmi, Lema 2.6.11 implicira da je sljedeći dijagram na nivou bar-kodova također komutativan.



Iz ove komutativnosti slijedi $im\big(\sigma_{inj}(h)\big) \subseteq im\big(\sigma_{inj}(f(\delta))\big)$, pa je dovoljno dokazati $im\big(\sigma_{inj}(h)\big) = BC^{2\delta}(N(2\delta))$. Ova jednakost slijedi iz definicije uparivanja $\sigma_{inj}(h)$ i činjenice da važi

$$BC\big(im\left(\Phi_N^{2\delta}\right)\big) = \{[b, d-2\delta) : [b, d) \in BC(N), d-b > 2\delta\},$$
$$BC(N(2\delta)) = \{[b-2\delta, d-2\delta) : [b, d) \in BC(N)\},$$

jer je $\sigma_{inj}(h)\big([b, d-2\delta)\big) = [b-2\delta, d-2\delta)$ ako i samo ako je $d-b > 2\delta$.

(iii) Neka je $[b, d) \in BC(im(\psi(\delta)))$ proizvoljno. Treba ispitati da li linija $\sigma_{inj}(f(\delta))\big([b, d)\big)$ ima navedena svojstva. Na osnovu dijela (ii) Leme 2.6.10, vrijedi $\sigma_{inf}(f(\delta))\big([b, d)\big) = [b', d)$, za neko $b' \leqslant b$. Ako je $d - b' \leqslant 2\delta$, tada trivijalno vrijedi $b - 2\delta < d - 2\delta \leqslant b' \leqslant b$. Ukoliko je $d - b' > 2\delta$, tada za $[b'+2\delta, d) \in BC\big(im\left(\Phi_N^{2\delta}\right)\big)$ vrijedi $\sigma_{inj}(h)\big([b'+2\delta, d)\big) = [b', d) = \sigma_{inj}(f(\delta))\big([b, d)\big)$, što implicira $b \leqslant b'+2\delta$, odakle ponovo slijedi $b-2\delta \leqslant b' \leqslant b$. □

Konačno, izvršene su sve pripreme da bi se dokazala sljedeća teorema.

Teorema 2.6.18. *[6] (**Teorema stabilnosti**) Neka su M, N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $BC(M), BC(N)$ njihovi bar-kodovi. Tada vrijedi*

$$d_{INT}(M, N) \geqslant d_{BOT}(BC(M), BC(N)).$$

Dokaz. Ideja je da se pokaže da svako $\delta-$preplitanje indukuje $\delta-$uparivanje.

Neka su M i N $\delta-$prepleteni, sa parom prepletenih morfizama $\psi : M \Rightarrow N(\delta)$ i $\chi : N \Rightarrow M(\delta)$. Kandidat za $\delta-$uparivanje bar-kodova $BC(M)$ i $BC(N)$ je $\sigma := S_\delta \circ \sigma_\psi$, gdje je $\sigma_\psi = \sigma_{inj} \circ \sigma_{sur}$ uparivanje indukovano morfizmom ψ, a $S_\delta : BC(N(\delta)) \to BC(N)$ preslikavanje definisano sa $S_\delta\big([b, d)\big) = [b+\delta, d+\delta)$ (ovo preslikavanje translira linije bar-koda $BC(N(\delta))$ udesno za δ). Potrebno je verifikovati da σ zadovoljava sva tri svojstva iz definicije $\delta-$uparivanja.

Neka je $[b,d) \in BC^{2\delta}(M)$ proizvoljna linija. Na osnovu Leme 2.6.16, ova linija se uparuje sa linijom $\sigma_{sur}\big([b,d)\big) = [b,d')$, a dalje se, na osnovu Leme 2.6.17, ova linija uparuje sa linijom $\sigma_{inj}\big([b,d')\big) = [b',d')$. Na kraju se dobija linija $S_\delta\big([b',d')\big) = [b'+\delta, d'+\delta)$, pa je $\sigma\big([b,d)\big) = [b'+\delta, d'+\delta)$, što implicira $[b,d) \in coim(\sigma)$. Time je dokazano $BC^{2\delta}(M) \subseteq coim(\sigma)$.

Neka je $[b,d) \in BC^{2\delta}(N)$ proizvoljna linija. Tada $[b-\delta, d-\delta) \in BC^{2\delta}(N(\delta))$ i vrijedi $[b,d) = S_\delta\big([b-\delta, d-\delta)\big)$. Kako je, na osnovu Leme 2.6.17, $BC^{2\delta}(N(\delta)) \subseteq im(\sigma_{inj}(f))$, gdje je f prirodna injekcija $im(\psi) \hookrightarrow N(\delta)$, zaključuje se da postoji $[b', d-\delta) \in coim(\sigma_{inj}(f))$ sa svojstvom $\sigma_{inj}(f)\big([b', d-\delta)\big) = [b-\delta, d-\delta)$. Dalje, na osnovu Leme 2.6.16, $[b', d-\delta) \in coim(\sigma_{inj}(f)) = BC(im(\psi)) = im(\sigma_{sur}(g))$, gdje je g surjektivna restrikcija morfizma ψ, što znači da postoji $[b',d') \in coim(\sigma_{sur}(g))$ sa svojstvom $\sigma_{sur}(g)\big([b',d')\big) = [b', d-\delta)$. Konačno, vrijedi $[b,d) = \big(S_\delta \circ \sigma_{inj}(f) \circ \sigma_{sur}(g)\big)\big([b',d')\big)$, što implicira $[b,d) = \sigma\big([b',d')\big)$. Time je dokazana inkluzija $BC^{2\delta}(N) \subseteq im(\sigma)$.

Neka su $[b_1,d_1) \in BC(M)$, $[b_2,d_2) \in BC(N)$, proizvoljne linije uparene sa σ, tj. linije za koje vrijedi $[b_2,d_2) = \sigma\big([b_1,d_1)\big) = \big(S_\delta \circ \sigma_{inj} \circ \sigma_{sur}\big)\big([b_1,d_1)\big)$. Na osnovu Leme 2.6.16 i Leme 2.6.17, vrijedi

$$[b_1,d_1) \overset{\sigma_{sur}}{\mapsto} [b_1,d_1') \overset{\sigma_{inj}}{\mapsto} [b_1',d_1') \overset{S_\delta}{\mapsto} [b_1'+\delta, d_1'+\delta),$$

pri čemu su ispunjeni uslovi $d_1 - 2\delta \leqslant d_1' \leqslant d_1$ i $b_1 - 2\delta \leqslant b_1' \leqslant b_1$. Zbog toga, vrijedi

$$d_1 - \delta \leqslant \underbrace{d_1' + \delta}_{d_2} \leqslant d_1 + \delta,$$

$$b_1 - \delta \leqslant \underbrace{b_1' + \delta}_{b_2} \leqslant b_1 + \delta,$$

odakle slijedi $|d_2 - d_1| \leqslant \delta$ i $|b_2 - b_1| \leqslant \delta$, te uparivanje σ zadovoljava i (iii) svojstvo iz definicije $\delta-$uparivanja.

Dobijeno $\delta-$uparivanje garantuje da između udaljenosti preplitanja i udaljenosti uskog grla vrijedi data nejednakost. Zaista, ako se pretpostavi suprotno, tj. da vrijedi nejednakost $d_{INT}(M,N) < d_{BOT}(BC(M), BC(N))$, tada, za proizvoljno $\delta \in \big(d_{INT}(M,N), d_{BOT}(BC(M), BC(N))\big)$, za istrajne module M i N postoji $\delta-$preplitanje, a za njihove bar-kodove $BC(M)$ i $BC(N)$ ne postoji $\delta-$uparivanje, što je kontradikcija, jer je u prethodnom razmatranju upravo konstruisano takvo uparivanje. □

Posljedica 2.6.19. *(**Teorema izometrije**) Neka su M, N istrajni moduli koji zadovoljavaju svojstvo konačnog tipa i $BC(M)$, $BC(N)$ njihovi bar-kodovi. Tada vrijedi*

$$d_{INT}(M,N) = d_{BOT}(BC(M), BC(N)).$$

Dokaz. Tvrđenje slijedi iz Teoreme 2.6.5 i Teoreme stabilnosti. □

2.7 Topološke mjere sličnosti familija stringova

Neka je $A \subseteq S(n,l)$ skup stringova i d neka od udaljenosti između dva stringa obuhvaćena u prvoj sekciji ove glave. Ideja realizovana u nastavku podrazumijeva da se strukturalna povezanost elemenata skupa A izrazi sredstvima istrajne homologije. Preciznije, svakom potprostoru metričkog prostora $(S(n,l),d)$ biće pridružena Čehova filtracija i odgovarajući istrajni modul homologije. To će stvoriti korespondenciju između skupa stringova i njegovog bar-koda i omogućiti da bar-kod, kao "nosilac informacije" o evoluciji međupovezanosti stringova iz datog skupa, postane reper značajnih homoloških atributa ovog skupa stringova. Stoga je razumno posmatrati mjere sličnosti dva stringova koje su zasnovane na poređenju njihovih bar-kodova. Cilj ove sekcije je uvođenje novih mjera sličnosti ovog tipa. Ove mjere biće zasnovane na modifikacijama udaljenosti uskog grla koje se baziraju na uparivanju linija bar-koda na kvalitativno višem nivou, a da se pritom ne naruši uslov stabilnosti ustanovljen u Teoremi stabilnosti.

Najprije će biti uvedena Čehova filtracija pridružena nepraznom skupu stringova $A \subseteq S(n,l)$.

Za proizvoljno $r \geqslant 0$, neka je $C_A^{(r)}$ Čehov kompleks čiji je skup vrhova sastavljen od stringova iz skupa A. Podsjećanja radi, simpleksi kompleksa $C_A^{(r)}$ su svi podskupovi $\sigma \subseteq A$ sa svojstvom $\bigcap_{s\in\sigma} B_d[s,r] \neq \emptyset$. Kako je A konačan skup, postoji najmanja vrijednost $r_t \geqslant 0$ takva da je $C_A^{(r)}$ pun kompleks za svako $r \geqslant r_t$. Filtracija od interesa je Čehova filtracija punog kompleksa $C_A := C_A^{(r_t)}$, koja formalizuje proces "izgradnje" ovog kompleksa iz inicijalnog kompleksa $C_A^{(0)}$, čiji je skup simpleksa dat sa $\Sigma_A^{(0)} := \{[s] : s \in A\}$. U prvom koraku, pronalazi se najmanja vrijednost $r_1 > 0$ sa svojstvom $C_A^{(0)} \subsetneq C_A^{(r_1)}$. Zatim, nalazi se najmanja vrijednost $r_2 > r_1$ sa svojstvom $C_A^{(r_1)} \subsetneq C_A^{(r_2)}$. Nastavljajući sa ovim postupkom, dolazi se do posljednjeg koraka $C_A^{(r_{t-1})} \subsetneq C_A^{(r_t)} = C_A$, u kojem se dodaju svi simpleksi koji su "nedostajali" u kompleksu $C_A^{(r_{t-1})}$, kompletirajući na taj način konstrukciju kompleksa C_A.

Čehova filtracija $C_A^{(0)} \subsetneq C_A^{(r_1)} \subsetneq \cdots \subsetneq C_A^{(r_t)}$ opisana u prethodnom predstavlja *filtraciju pridruženu skupu stringova* $A \subseteq S(n,l)$, a pozitivni brojevi $r_1,\ldots,r_t$ su nivoi ove filtracije. Za simpleks $\sigma \subseteq A$, najmanja vrijednost $r_\sigma \geqslant 0$ sa svojstvom $\bigcap_{s\in\sigma} B_d[s,r_\sigma] \neq \emptyset$ naziva se *radijusom simpleksa* σ. U tom slučaju, proizvoljan string $c \in \bigcap_{s\in\sigma} B[s,r_\sigma]$ se naziva *centrom simpleksa* σ.

Primjećuje se da, za razliku od radijusa simpleksa, centar simpleksa ne mora da bude jedinstven.

Lema 2.7.1. *Broj r je nivo filtracije pridružene skupu stringova A ako i samo ako je r radijus nekog simpleksa iz C_A.*

Dokaz. Neka je r nivo filtracije pridružene skupu stringova A. Ako je $r = 0$, tada je r radijus bilo kojeg $0-$simpleksa kompleksa C_A. Inače, postoji $i \in \{1, 2, \ldots, t\}$ takav da je $r = r_i$. Po načinu konstrukcije date filtracije, postoji simpleks σ koji pripada kompleksu $C_A^{(r_i)}$, a ne pripada kompleksu $C_A^{(r_{i-1})}$, što implicira da za ovaj simpleks vrijedi $r = r_\sigma$. Obrnuto, ako je r radijus simpleksa σ, tada se ovaj simpleks prvi put pojavljuje u Čehovom kompleksu $C_A^{(r)}$, te, u skladu sa definicijom filtracije pridružene skupu A, r mora biti jednak nekom nivou filtracije. □

Primjer 2.7.2. *Neka je $A \subset S(4, 8)$, pri čemu je*

$$A = \{\underbrace{12244131}_{s_1}, \underbrace{22223443}_{s_2}, \underbrace{32143431}_{s_3}, \underbrace{14443214}_{s_4}, \underbrace{22134222}_{s_5}\}$$

familija stringova iz $S(4, 8)$. Na osnovu prethodne leme, za nalaženje filtracije pridružene skupu stringova A dovoljno je naći sve parove oblika (σ, r_σ), gdje $\sigma \in C_A$, a r_σ je radijus simpleksa σ. Ako se familija A posmatra kao potprostor metričkog prostora $(S(4, 8), d_H)$, dobija se:

$$(\underbrace{[s_1]}_{\sigma_1}, 0), (\underbrace{[s_2]}_{\sigma_2}, 0), (\underbrace{[s_3]}_{\sigma_3}, 0), (\underbrace{[s_4]}_{\sigma_4}, 0), (\underbrace{[s_5]}_{\sigma_5}, 0), (\underbrace{[s_1, s_3]}_{\sigma_6}, 2),$$
$$(\underbrace{[s_1, s_2]}_{\sigma_7}, 3), (\underbrace{[s_2, s_3]}_{\sigma_8}, 3), (\underbrace{[s_1, s_2, s_3]}_{\sigma_9}, 3), (\underbrace{[s_1, s_4]}_{\sigma_{10}}, 3), (\underbrace{[s_3, s_4]}_{\sigma_{11}}, 3),$$
$$(\underbrace{[s_1, s_3, s_4]}_{\sigma_{12}}, 3), (\underbrace{[s_1, s_5]}_{\sigma_{13}}, 3), (\underbrace{[s_2, s_5]}_{\sigma_{14}}, 3), (\underbrace{[s_3, s_5]}_{\sigma_{15}}, 3), (\underbrace{[s_1, s_3, s_5]}_{\sigma_{16}}, 3);$$
$$(\underbrace{[s_2, s_4]}_{\sigma_{17}}, 4), (\underbrace{[s_1, s_2, s_4]}_{\sigma_{18}}, 4), (\underbrace{[s_2, s_3, s_4]}_{\sigma_{19}}, 4), (\underbrace{[s_1, s_2, s_3, s_4]}_{\sigma_{20}}, 4), (\underbrace{[s_1, s_2, s_5]}_{\sigma_{21}}, 4),$$

$$(\underbrace{[s_2, s_3, s_5]}_{\sigma_{22}}, 4), (\underbrace{[s_1, s_2, s_3, s_5]}_{\sigma_{23}}, 4), (\underbrace{[s_4, s_5]}_{\sigma_{24}}, 4), (\underbrace{[s_2, s_4, s_5]}_{\sigma_{25}}, 4), (\underbrace{[s_3, s_4, s_5]}_{\sigma_{26}}, 4),$$
$$(\underbrace{[s_2, s_3, s_4, s_5]}_{\sigma_{27}}, 4), (\underbrace{[s_1, s_4, s_5]}_{\sigma_{28}}, 4);$$
$$(\underbrace{[s_1, s_2, s_4, s_5]}_{\sigma_{29}}, 5), (\underbrace{[s_1, s_3, s_4, s_5]}_{\sigma_{30}}, 5), (\underbrace{[s_1, s_2, s_3, s_4, s_5]}_{\sigma_{31}}, 5).$$

Radijus simpleksa $\sigma_6 = [12244131, 32143431]$ *je* 2*, ali njegov centar nije jedinstven, npr. stringovi* 32244431 *i* 12143131 *predstavljaju centre ovog simpleksa.*

Iz prethodne karakterizacije se izvodi tražena filtracija: $C_A^{(0)} \subsetneq C_A^{(2)} \subsetneq C_A^{(3)} \subsetneq C_A^{(4)} \subsetneq C_A^{(5)} = C_A$*, pri čemu su odgovarajući skupovi simpleksa dati sa*

$$\begin{aligned}
\Sigma_A^{(0)} &= \{\sigma_i : 1 \leqslant i \leqslant 5\},\\
\Sigma_A^{(2)} &= \Sigma_A^{(0)} \cup \{\sigma_6\},\\
\Sigma_A^{(3)} &= \Sigma_A^{(2)} \cup \{\sigma_i : 7 \leqslant i \leqslant 16\},\\
\Sigma_A^{(4)} &= \Sigma_A^{(3)} \cup \{\sigma_i : 17 \leqslant i \leqslant 28\},\\
\Sigma_A^{(5)} &= \Sigma_A^{(4)} \cup \{\sigma_i : 29 \leqslant i \leqslant 31\} = P(A) \setminus \{\emptyset\}.
\end{aligned}$$

Ako se familija A posmatra kao potprostor metričkog prostora $(S(4,8), d_{LCS})$*, na sličan način se izvodi filtracija pridružena skupu stringova A:* $C_A^{(0)} \subsetneq C_A^{(\frac{1}{4})} \subsetneq C_A^{(\frac{3}{8})} = C_A$*, pri čemu su odgovarajući skupovi simpleksa dati sa*

$$\begin{aligned}
\Sigma_A^{(0)} &= \{\sigma_i : 1 \leqslant i \leqslant 5\},\\
\Sigma_A^{(\frac{1}{4})} &= \Sigma_A^{(0)} \cup \{\sigma_i : 6 \leqslant i \leqslant 15\}\\
\Sigma_A^{(\frac{3}{8})} &= \Sigma_A^{(\frac{1}{4})} \cup \{\sigma_i : 16 \leqslant i \leqslant 31\} = P(A) \setminus \{\emptyset\}.
\end{aligned}$$

U ovom slučaju, radijus simpleksa $\sigma_6 = [12244131, 32143431]$ *je* $\frac{1}{4}$ *i ispostavlja se da stringovi* 32244431 *i* 12143131 *ponovo predstavljaju centre ovog simpleksa.*

Primjedba 2.7.3. *Zbog diskretne prirode Hamingove i LCS udaljenosti na skupu* $S(n,l)$*, nivoi filtracije pridružene skupu stringova su "razmaknuti" za najmanje* 1 *u slučaju Hamingove udaljenosti, odnosno za najmanje* $\frac{1}{l}$ *u slučaju LCS udaljenosti.*

Neka su $A, B \subseteq S(n,l)$ neprazni skupovi stringova. Automorfizam metričkog prostora $(S(n,l), d)$ koji preslikava skup A u skup B se naziva $d(A \to B)$-*izomorfizmom*. U tom slučaju, skupovi A i B se nazivaju $d-$*izomorfnim*.

U narednom su opisani automorfizmi metričkog prostora $(S(n,l), d_H)$. Za datu permutaciju p skupa $\mathbb{N}_n$ i $i \in \mathbb{N}_l$, neka je $\theta_p^i : S(n,l) \to S(n,l)$ preslikavanje definisano sa $\theta_p^i(a_1a_2\ldots a_l) = a_1a_2\ldots a_{i-1}p(a_i)a_{i+1}\ldots a_l$. Takođe, za datu permutaciju q skupa $\mathbb{N}_l$, neka je preslikavanje $\varphi_q : S(n,l) \to S(n,l)$ definisano sa $\varphi_q(a_1a_2\ldots a_l) = a_{q(1)}a_{q(2)}\ldots a_{q(l)}$. Za početak, biće dokazano sljedeće pomoćno tvrđenje.

Lema 2.7.4. *Neka su $s_1, s_2, s_3, s_4 \in S(n,l)$ različiti stringovi koji se mogu predstaviti u obliku*

$$\begin{aligned} s_1 &= x_1 \dots x_{i-1} c x_{i+1} \dots x_{j-1} d x_{j+1} \dots x_l, \\ s_2 &= x_1 \dots x_{i-1} c' x_{i+1} \dots x_{j-1} d x_{j+1} \dots x_l, \\ s_3 &= x_1 \dots x_{i-1} c' x_{i+1} \dots x_{j-1} d' x_{j+1} \dots x_l, \\ s_4 &= x_1 \dots x_{i-1} c x_{i+1} \dots x_{j-1} d' x_{j+1} \dots x_l, \end{aligned}$$

za neke $i, j \in \mathbb{N}_l$ i neke $c, c', d, d', x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_l \in \mathbb{N}_n$. Ako su f i g automorfizmi metričkog prostora $(S(n,l), d_H)$ za koje vrijedi $f(s_1) = g(s_1)$, $f(s_2) = g(s_2)$ i $f(s_3) = g(s_3)$, tada je ujedno i $f(s_4) = g(s_4)$.

Dokaz. Iz definicije stringova s_1, s_2, s_3, s_4 slijedi $d_H(s_1, s_2) = d_H(s_2, s_3) = d_H(s_3, s_4) = d_H(s_4, s_1) = 1$ i $d_H(s_1, s_3) = 2$. Zbog toga, za stringove $t_1 := f(s_1)$, $t_2 := f(s_2)$ i $t_3 := f(s_3)$ vrijedi $d_H(t_1, t_2) = d_H(t_2, t_3) = 1$ i $d_H(t_1, t_3) = 2$. Ovo znači da, pored stringa t_2, postoji samo još jedan string koji je na Hamingovoj udaljenosti 1 od stringova t_1 i t_3. Neka je taj string označen sa t_4. Iz datih pretpostavki slijedi $d_H(t_1, f(s_4)) = d_H(t_3, f(s_4)) = 1$ i $d_H(t_1, g(s_4)) = d_H(t_3, g(s_4)) = 1$, pa, kako je $f(s_4) \neq t_2 \neq g(s_4)$, mora vrijediti $f(s_4) = t_4 = g(s_4)$. □

Lema 2.7.5. *Automorfizmi metričkog prostora $(S(n,l), d_H)$ su funkcije oblika $\theta^1_{p_1} \circ \theta^2_{p_2} \circ \dots \circ \theta^l_{p_l} \circ \varphi_q$, za neke permutacije $p_1, p_2, \dots, p_l$ od $\mathbb{N}_n$ i q od $\mathbb{N}_l$.*

Dokaz. Jednostavno se provjerava da su funkcije θ^i_p i φ_q automorfizmi, pa su takve i njihove kompozicije. Stoga, dovoljno je dokazati da su automorfizmi metričkog prostora $(S(n,l), d_H)$ navedenog oblika. Neka je f proizvoljan automorfizam metričkog prostora $(S(n,l), d_H)$ i neka je $a_1 a_2 \dots a_l \in S(n,l)$ string za koji vrijedi $f(\underbrace{11\dots1}_{l}) = a_1 a_2 \dots a_l$. Za proizvoljno $i \in \mathbb{N}_l$, Hamingova udaljenost između stringova oblika $f(\underbrace{11\dots1}_{i-1} c_i \underbrace{11\dots1}_{l-i})$, $c_i \in \mathbb{N}_n$, jednaka je 1. To znači da za neko $j \in \mathbb{N}_l$ i neko $b_j \in \mathbb{N}_n$ vrijedi

$$f(\underbrace{11\dots1}_{i-1} c_i \underbrace{11\dots1}_{l-i}) = a_1 a_2 \dots a_{j-1} b_j a_{j+1} \dots a_l.$$

Neka je q permutacija od $\mathbb{N}_l$ za koju vrijedi $q(i) = j$ i p_j permutacija od $\mathbb{N}_n$ za koju vrijedi $p_j(c_i) = b_j$. Konačno, neka je $g := \theta^1_{p_1} \circ \theta^2_{p_2} \circ \dots \circ \theta^l_{p_l} \circ \varphi_q$; dovoljno je dokazati da je $f = g$, tj. da za svako $s \in S(n,l)$ vrijedi $f(s) = g(s)$. Dokaz date jednakosti će biti izveden indukcijom po $r := d_H(\underbrace{11\dots1}_{l}, s)$. Za $r \leqslant 1$, jednakost slijedi iz definicije funkcije g. Indukciona pretpostavka je da za neko $r \geqslant 2$ jednakost $f(t) = g(t)$ vrijedi za svaki string t

koji ispunjava uslov $d_H(\underbrace{11\ldots1}_{l}, t) < r$. Neka je $s = c_1c_2\ldots c_l$ string za koji vrijedi $d_H(\underbrace{11\ldots1}_{l}, s) = r$. Moguće je izabrati dvije pozicije na kojima se string s razlikuje od stringa $\underbrace{11\ldots1}_{l}$. Neka su te dvije pozicije i i j, pri čemu je $i < j$. Dalje se definišu stringovi $s_1 := c_1\ldots c_{i-1}1c_{i+1}\ldots c_l$, $s_2 := c_1\ldots c_{i-1}1c_{i+1}\ldots c_{j-1}1c_{j+1}\ldots c_l$ i $s_3 := c_1\ldots c_{j-1}1c_{j+1}\ldots c_l$. Indukciona hipoteza garantuje da je $f(s_3) = g(s_3)$, $f(s_2) = g(s_2)$ i $f(s_1) = g(s_1)$. Na osnovu prethodne leme, slijedi $f(s) = g(s)$. Ovo kompletira dokaz indukcijom. □

Ukoliko filtracije pridružene skupovima A i B imaju identične skupove nivoa $\{r_1, r_2, \ldots, r_t\}$, tada se ove filtracije nazivaju *izomorfnim*, ako postoji bijekcija $f : A \to B$ takva da za svaki simpleks σ kompleksa C_A i proizvoljno $i \in \{1, 2, \ldots, t\}$ vrijedi da je σ simpleks kompleksa $C_A^{(r_i)}$ ako i samo ako je $f[\sigma]$ simpleks kompleksa $C_B^{(r_i)}$. U tom slučaju, preslikavanje f se naziva *filtracijskim izomorfizmom*.

Očigledno da d_H-izomorfni skupovi stringova imaju izomorfne pridružene filtracije. Zaista, proizvoljni $d_H(A \to B)$ izomorfizam "čuva" radijuse svih simpleksa iz C_A, što ga ujedno čini filtracijskim izomorfizmom. Obrat ovog tvrđenja generalno nije tačan, što je ilustrovano u sljedećem primjeru.

Primjer 2.7.6. *Neka je $l = 5, n = 3$ i $s_1 = 11113$, $s_2 = 22223$, $s_3 = 33333$, $s_4 = 33122$. Skupovi stringova $A = \{s_1, s_2, s_3\}$ i $B = \{s_1, s_2, s_4\}$ imaju izomorfne filtracije. Zaista, ako se izuzmu izolovani vrhovi, kompleks $C_A^{(2)}$ sadrži $1-$simplekse $[s_1, s_2]$, $[s_1, s_3]$ i $[s_2, s_3]$, dok kompleks $C_B^{(2)}$ sadrži $1-$simplekse $[s_1, s_2]$, $[s_1, s_4]$ and $[s_2, s_4]$. Ova dva kompleksa su jedini netrivijalni potkompleksi od C_A i C_B, jer su oba kompleksa $C_A^{(3)}$, $C_A^{(3)}$ puni kompleksi. Ovo znači da je preslikavanje $g : A \to B$ dato sa $g(s_i) = s_i$, $i \in \{1, 2\}$, $g(s_3) = s_4$, filtracijski izomorfizam.*

S druge strane, ne postoji $d_H(A \to B)$-izomorfizam. Zaista, ako se pretpostavi suprotno, tj. da postoji $d_H(A \to B)$-izomorfizam f, na osnovu Leme 2.7.5 vrijedi $f = \theta^1_{p_1} \circ \theta^2_{p_2} \circ \theta^3_{p_3} \circ \theta^4_{p_4} \circ \theta^5_{p_5} \circ \varphi_q$, za neke permutacije p_1, p_2, p_3, p_4, p_5 skupa $\mathbb{N}_3$ i neku permutaciju q skupa $\mathbb{N}_5$. Svako preslikavanje $\theta^i_{p_i}$ "čuva" broj različitih simbola na fiksnoj poziciji stringa, kao i vektor broja pojavljivanja svakog simbola. Kako stringovi s_1, s_2, s_3 svi završavaju simbolom 3, ovo bi značilo da postoji pozicija takva da svaki od stringova s_1, s_2, s_4 na toj poziciji ima isti simbol. Međutim, lako se provjerava da to nije zadovoljeno.

Pitanje 2.7.7. Da li postoje dodatni uslovi pod kojima pretpostavka o postojanju filtracijskog izomorfizma garantuje da su odgovarajući skupovi stringova d_H-izomorfni? Postoje naznake da bi se u metričkom prostoru $(S(2, l), d_H)$ ovi

uslovi mogli odnositi na postojanje baricentra - stringa podjednako udaljenog od svih stringova iz posmatranog skupa. Na posmatranim kontraprimjerima uočeno je da jedan od skupova stringova ima baricentar, dok ga drugi nema. Međutim, ovo i dalje ostaje na nivou hipoteze.

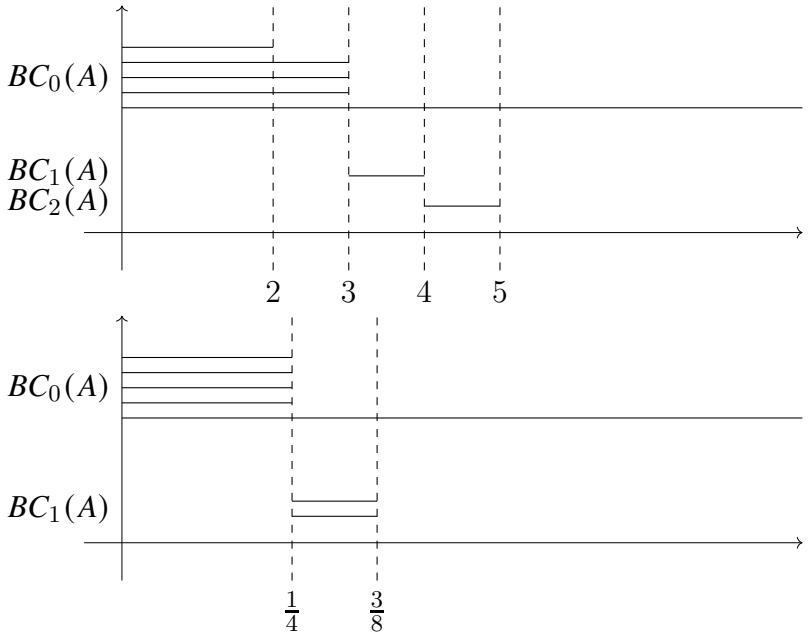
Neka je $A \subseteq S(n,l)$ neprazan skup stringova. Za datu dimenziju $k \geqslant 0$, filtraciji pridruženoj skupu A moguće je pridružiti istrajni modul homologije $M(Hom_k(C_A))$. Bar-kod ovog istrajnog modula predstavlja *bar-kod pridružen skupu stringova A*. Ovaj bar-kod biće kraće označavan sa $BC_k(A)$. Kako je A konačan skup, bar-kod $BC_k(A)$ ne sadrži linije ukoliko je $k > |A| - 1$. Bar-kod $BC_0(A)$ sadrži tačno $|A|$ linija i ove linije predstavljaju evoluciju komponenti povezanosti prilikom prolaska kroz filtraciju. S obzirom da puni kompleks C_A ima samo jednu komponentu povezanosti, zaključuje se da bar-kod $BC_0(A)$ ima tačno jednu beskonačnu liniju. Za $k \geqslant 1$, svaka $k-$dimenzionalna rupa se eventualno mora zatvoriti u nekom nivou filtracije, što znači da su sve linije bar-koda $BC_k(A)$ konačnih dužina.

Primjer 2.7.8. *Za skup stringova iz Primjera 2.7.2, neka je $M_H(Hom_k(C_A))$ istrajni modul homologije u odnosu na Čehovu filtraciju kompleksa C_A izvedenu u odnosu na Hamingovu udaljenost, a $M_{LCS}(Hom_k(C_A))$ istrajni modul homologije u odnosu na Čehovu filtraciju kompleksa C_A izvedenu u odnosu na LCS udaljenost.*

Istrajne module $M_H(Hom_k(C_A))$ "naseljavaju" 5 nuladimenzionalnih homoloških klasa, jedna jednodimenzionalna i jedna dvodimenzionalna homološka klasa. Kako $[s_1, s_3] \in C_A^{(2)}$, dvije komponente povezanosti se spajaju u jednu u ovom nivou filtracije, dok se ostale nuladimenzionalne klase spajaju u jednu u kompleksu $C_A^{(3)}$. Jedini netrivijalni jednodimenzionalni ciklus je $[s_1, s_2] + [s_2, s_5] + [s_1, s_5]$. Ovaj ciklus nastaje u kompleksu $C_A^{(3)}$ i postaje trivijalan u kompleksu $C_A^{(4)}$. Slično, jedini netrivijalni dvodimenzionalni ciklus je $[s_1, s_2, s_4] + [s_1, s_2, s_5] + [s_1, s_4, s_5] + [s_2, s_4, s_5]$. Ovaj ciklus nastaje u kompleksu $C_A^{(4)}$, a postaje trivijalan u kompleksu $C_A^{(5)} = C_A$. Dakle, istrajni moduli homologije $M_H(Hom_0(C_A))$, $M_H(Hom_1(C_A))$ i $M_H(Hom_2(C_A))$ imaju intervalne dekompozicije redom date sa $M[0,2) \bigoplus M[0,3) \bigoplus M[0,3) \bigoplus M[0,3) \bigoplus M[0,+\infty)$, $M[3,4)$ i $M[4,5)$.

Istrajne module $M_{LCS}(Hom_k(C_A))$ "naseljavaju" 5 nuladimenzionalnih homoloških klasa, dvije jednodimenzionalne klase, dok višedimenzionalnih homoloških klasa nema. Homološke klase $[s_1], [s_2], [s_3], [s_4], [s_5]$ stapaju se u jednu klasu u kompleksu $C_A^{(\frac{1}{4})}$. Dva netrivijalna jednodimenzionalna ciklusa $[s_1, s_2]+[s_2, s_5]+[s_5, s_1]$ i $[s_1, s_3]+[s_3, s_5]+[s_5, s_1]$ nastaju u kompleksu $C_A^{(\frac{1}{4})}$ i oba postaju trivijalna u punom kompleksu $C_A^{(\frac{3}{8})} = C_A$. Istrajni moduli homologije $M_{LCS}(Hom_0(C_A))$ i $M_{LCS}(Hom_1(C_A))$ imaju intervalne dekompozicije

date sa $M\left[0,\frac{3}{8}\right)\bigoplus M\left[0,\frac{3}{8}\right)\bigoplus M\left[0,\frac{3}{8}\right)\bigoplus M\left[0,\frac{1}{2}\right)\bigoplus M\,[0,+\infty)$*, odnosno* $M\left[\frac{1}{4},\frac{3}{8}\right)\bigoplus M\left[\frac{1}{4},\frac{3}{8}\right)$*, respektivno. Odgovarajući bar-kodovi su predstavljeni na sljedećoj slici.*



Slika 2.12

Svaka linija bar-koda $BC_k(A)$, $k \geqslant 1$, oslikava istrajnost $k-$dimenzionalne rupe koja "opstaje" usljed nedostatka povezanosti neke podfamilije stringova iz skupa A ispoljene u određenom dijelu filtracije pridružene skupu A. U tom smislu, bar-kod može da se upotrijebi kao sredstvo za definisanje mjere sličnosti dva skupa stringova. Ova mjera bi se zasnivala na poređenju linija bar-koda iste dimenzije, sa idejom da se uparuju linije koje su "kvalitativno slične", u smislu da predstavljaju slične homološke karakteristike. U narednom je detaljnije razrađena ova ideja.

Neka su $A, B \subseteq S(n,l)$ skupovi stringova, takvi da je $|A| = |B| = m \geqslant 2$, i neka su date filtracije pridružene ovim skupovima stringova.

Oba bar-koda $BC_0(A)$ i $BC_0(B)$ sadrže po m linija, pri čemu po $m-1$ njih su linije konačne dužine (i svaka od njih ima početak u tački 0), a po jedna linija je oblika $[0,+\infty)$. Dvije linije beskonačne dužine su savršeno uparene, pa se mogu ignorisati. Ostale linije bar-kodova $BC_0(A)$ i $BC_0(B)$ se mogu enumerisati u obliku $\left\{[0,l_i^A) : i \in \{1,2,\ldots,m-1\}\right\}$, $0 < l_1^A \leqslant l_2^A \leqslant \ldots \leqslant$

l^A_{m-1}, i $\{[0, l^B_i) : i \in \{1, 2, \ldots, m-1\}\}$, $0 < l^B_1 \leqslant l^B_2 \leqslant \ldots \leqslant l^B_{m-1}$, respektivno. Logičan izbor je uparivanje linija $[0, l^A_i)$ i $[0, l^B_i)$ i traženje maksimalne razlike između l^A_i i l^B_i.

U slučaju dimenzije $k \geqslant 1$, kao mjeru različitosti skupova A i B moguće je izabrati udaljenost uskog grla $d_{BOT}\,(BC_k(A), BC_k(B))$. Međutim, udaljenost uskog grla uparuje linije koje se "najbolje" preklapaju, bez ulaženja u njihov kontekst. Stoga će biti ispitana mogućnost uparivanja linija na kvalitativno višem nivou. Ovo ispitivanje biće sprovedeno primjenom *šeme registracije ciklusa*, tehnike opisane u [78]. Iskazano jezikom problema koji se ovdje rješava, šema registracije ciklusa može da se opiše na sljedeći način:

Zajedno sa filtracijama pridruženim skupovima A i B, posmatraće se filtracija pridružena skupu $A \cup B$. Istrajni modul $M(Hom_k(C_{A\cup B}))$ se može shvatiti kao "krovni" modul u koji se na prirodan način potapaju istrajni moduli $M(Hom_k(C_A))$ i $M(Hom_k(C_B))$. Preciznije, ova potapanja su morfizmi

$$h^A : M(Hom_k(C_A)) \Rightarrow M(Hom_k(C_{A\cup B})),$$
$$h^B : M(Hom_k(C_B)) \Rightarrow M(Hom_k(C_{A\cup B})),$$

tako da su, za svaki nivo filtracije r_i, preslikavanja $h^A_{r_i}$ i $h^B_{r_i}$ indukovana inkluzijom. Ako je γ_A $k-$ciklus u istrajnom modulu $M(Hom_k(C_A))$ i γ_B $k-$ciklus u istrajnom modulu $M(Hom_k(C_B))$, tada se ovi ciklusi nazivaju $C_{A\cup B}$*−ekvivalentnim ciklusima* (u oznaci $\gamma_A \overset{C_{A\cup B}}{\sim} \gamma_B$), ako postoji $k-$ciklus $\tilde{\gamma}_A$ u istrajnom modulu $im(h^A)$ i $k-$ciklus $\tilde{\gamma}_B$ u istrajnom modulu $im(h^B)$ za koje vrijede uslovi

- Ciklusi γ_A i $\tilde{\gamma}_A$ nastaju na istom nivou filtracije,
- Ciklusi γ_B i $\tilde{\gamma}_B$ nastaju na istom nivou filtracije,
- Ciklusi $\tilde{\gamma}_A$ i $\tilde{\gamma}_B$ nestaju na istom nivou filtracije.

Pojam $C_{A\cup B}-$ekvivalentnih ciklusa je naročito značajan u slučaju kada su filtracije kompleksa C_A, C_B i $C_{A\cup B}$ Morzeove filtracije. U ovom slučaju, prva dva uslova garantuju da su ciklusi $\tilde{\gamma}_A, \tilde{\gamma}_B$ strukturalno povezani sa ciklusima γ_A, γ_B, zbog toga što se pojavljuju na istom nivou filtracije. Treći uslov implicira da se ciklusi $\tilde{\gamma}_A$ i $\tilde{\gamma}_B$ "ubijaju" na istom nivou filtracije, što dovodi do zaključka da ovi ciklusi reprezentuju slična homološka svojstva. Posljedično, isti zaključak se prenosi i na njihove "parnjake", cikluse γ_A i γ_B. Bitno svojstvo relacije $\overset{C_{A\cup B}}{\sim}$ je njena ograničena funkcionalnost i iskazano je u sljedećoj lemi.

Lema 2.7.9. *Neka je γ_A $k-$ciklus u istrajnom modulu $M(Hom_k(C_A))$ i neka su β_B, γ_B $k-$ciklusi u istrajnom modulu $M(Hom_k(C_B))$ tako da vrijedi $\gamma_A \overset{C_{A\cup B}}{\sim} \beta_B$ i $\gamma_A \overset{C_{A\cup B}}{\sim} \gamma_B$. Ako su filtracije kompleksa C_A, C_B and $C_{A\cup B}$ Morzeove filtracije, tada je $\beta_B = \gamma_B$.*

Dokaz. Iz $\gamma_A \overset{C_{A\cup B}}{\sim} \beta_B$ slijedi postojanje $k-$ciklusa $\tilde{\gamma}_A$ u istrajnom modulu $im(h^A)$ i $k-$ciklusa $\tilde{\beta}_B$ u istrajnom modulu $im(h^B)$ tako da γ_A i $\tilde{\gamma}_A$ nastaju na istom nivou filtracije, β_B i $\tilde{\beta}_B$ nastaju na istom nivou filtracije, a da $\tilde{\gamma}_A$ i $\tilde{\beta}_B$ nestaju na istom nivou filtracije. Slično, iz $\gamma_A \overset{C_{A\cup B}}{\sim} \gamma_B$ slijedi postojanje $k-$ciklusa $\tilde{\tilde{\gamma}}_A$ u istrajnom modulu $im(h^A)$ i $k-$ciklusa $\tilde{\gamma}_B$ u istrajnom modulu $im(h^B)$ tako da γ_A i $\tilde{\tilde{\gamma}}_A$ nastaju na istom nivou filtracije, γ_B i $\tilde{\gamma}_B$ nastaju na istom nivou filtracije, a da $\tilde{\tilde{\gamma}}_A$ i $\tilde{\gamma}_B$ nestaju na istom nivou filtracije. To znači da ciklusi $\tilde{\gamma}_A$ i $\tilde{\tilde{\gamma}}_A$ nastaju na istom nivou filtracije, pa, iz pretpostavke da je filtracija kompleksa $C_{A\cup B}$ Morzeova filtracija, slijedi $\tilde{\gamma}_A = \tilde{\tilde{\gamma}}_A$. Na osnovu toga, zaključuje se da ciklusi $\tilde{\beta}_B$ i $\tilde{\gamma}_B$ nestaju na istom nivou filtracije, odakle slijedi $\tilde{\beta}_B = \tilde{\gamma}_B$. Konačno, to znači da ciklusi β_B i γ_B nastaju na istom nivou filtracije, pa, iz pretpostavke da je filtracija kompleksa C_B Morzeova filtracija, slijedi $\beta_B = \gamma_B$. □

Dakle, ako su filtracije kompleksa C_A, C_B i $C_{A\cup B}$ Morzeove filtracije, tada se poređenje linija bar-kodova $BC_k(A)$ i $BC_k(B)$ može izvesti na način koji bi favorizovao uparivanje linija koje odgovaraju $C_{A\cup B}-$ekvivalentnim ciklusima. Linije bar-kodova $BC_k(A)$ i $BC_k(B)$ koje se ne mogu upariti na ovaj način biće uparene "uobičajenim" uparivanjem iz definicije udaljenosti uskog grla. Više detalja o ovom "hibridnom" uparivanju biće pruženo na kraju ove sekcije.

Nažalost, opisana strategija je problematična u slučaju kada bar jedna od posmatranih filtracija nije Morzeova filtracija. Uzimajući u obzir diskretnu prirodu Hamingove i LCS udaljenosti, mogućnost nastajanja ili nestajanja dva ili više ciklusa na istom nivou filtracije postaje sve izvjesnija sa povećanjem broja stringova u posmatranom skupu stringova. U narednom dijelu izložena je nova tehnika koja omogućava da se u slučaju Hamingove udaljenosti ovaj problem razriješi na zadovoljavajući način.

Generalizovani string dužine $l \geqslant 1$ nad alfabetom $\mathbb{N}_n$ je funkcija $s : \mathbb{N}_l \to F_n$, gdje je F_n skup funkcija $f : \mathbb{N}_n \to [0, 1]$ koje ispunjavaju uslov $\sum_{j=1}^{n} f(j) = 1$. Skup takvih generalizovanih stringova biće notiran sa $S'(n, l)$, dok će $s[i]$ biti oznaka slike od $i \in \mathbb{N}_l$ u odnosu na generalizovani string $s \in S'(n, l)$. Svaki generalizovani string $s \in S'(n, l)$ može da se predstavi matricom $[s_{ij}]$ formata $l \times n$, gdje je $s_{ij} := s[i](j)$, pri čemu je suma elemenata svakog reda pojedinačno jednaka 1, tj. za svako $i \in \mathbb{N}_l$ vrijedi $\sum_{j=1}^{n} s_{ij} = 1$.

Generalizovana Hamingova udaljenost d_{GH} između generalizovanih strin-

gova $s, t \in S'(n, l)$ se definiše sa

$$d_{GH}(s,t) = \sum_{i=1}^{l} \left(1 - \sum_{j=1}^{n} \min\{s[i](j), t[i](j)\} \right).$$

Lema 2.7.10. *Funkcija d_{GH} je metrika na skupu $S'(n, l)$.*

Dokaz. U postupku provjeravanja svojstava metrike biće korišćen identitet

$$\min\{a, b\} = \frac{a + b - |a - b|}{2}.$$

Za proizvoljan string $s \in S'(n, l)$ i svako $i \in \mathbb{N}_l$ vrijedi $1 - \sum_{j=1}^{n} s[i](j) = 0$, pa je $d_{GH}(s, s) = 0$. Takođe, iz uslova $d_{GH}(s, t) = 0$ slijedi da za svako $i \in \mathbb{N}_l$ vrijedi $\sum_{j=1}^{n} \min\{s[i](j), t[i](j)\} = 1$, odakle se dobija $\sum_{j=1}^{n} \frac{|s[i](j) - t[i](j)|}{2} = 0$, tj. $s[i](j) = t[i](j)$, za svako $j \in \mathbb{N}_n$, što znači da je $s = t$. Uslov simetričnosti slijedi iz definicije funkcije d_{GH}, pa ostaje da se dokaže nejednakost trougla. Neka su $s, t, u \in S'(n, l)$ proizvoljni. Tada vrijedi

$$\begin{aligned}
&d_{GH}(s,t) + d_{GH}(t,u) \\
&= \sum_{i=1}^{l} \left(2 - \sum_{j=1}^{n} \min\{s[i](j), t[i](j)\} + \min\{t[i](j), u[i](j)\} \right) \\
&= \sum_{i=1}^{l} \left(1 - \sum_{j=1}^{n} \frac{s[i](j) + u[i](j) - |s[i](j) - t[i](j)| - |t[i](j) - u[i](j)|}{2} \right) \\
&\geqslant \sum_{i=1}^{l} \left(1 - \sum_{j=1}^{n} \frac{s[i](j) + u[i](j) - |s[i](j) - u[i](j)|}{2} \right) = d_{GH}(s,u),
\end{aligned}$$

pa je ispunjena i nejednakost trougla. □

Udaljenost $d_{GH}(s, t)$ mjeri preklapanje funkcija $s[i]$ i $t[i]$, $i \in \mathbb{N}_l$. String $s = a_1 a_2 \dots a_l \in S(n, l)$ se može identifikovati sa generalizovanim stringom s, pri čemu je $s[i]$ funkcija definisana sa

$$s[i](j) := \begin{cases} 1, & \text{ako je } j = a_i; \\ 0, & \text{inače.} \end{cases}$$

Koristeći ovu konvenciju, jednostavno se provjerava da za proizvoljne stringove $s, t \in S(n, l)$ vrijedi $d_{GH}(s, t) = d_H(s, t)$, što znači da je restrikcija funkcije d_{GH} na skup $S(n, l)$ "uobičajena" Hamingova udaljenost d_H. Najznačajnija razlika

između ove dvije metrike predstavlja činjenicu da udaljenost d_{GH} ne indukuje diskretnu topologiju na $S'(n,l)$ kao što to metrika d_H čini na skupu $S(n,l)$. Činjenica da se u metričkom prostoru $(S'(n,l), d_{GH})$ svakom generalizovanom stringu može "prići" dovoljno blizu je posljedica sljedeće leme.

Lema 2.7.11. *U metričkom prostoru $(S'(n,l), d_{GH})$, za svako $x \in S'(n,l)$ i proizvoljno $y \in S'(n,l)$ vrijedi $y \in \overline{B_{d_{GH}}(x, d_{GH}(x,y))}$.*

Dokaz. Neka su $x, y \in S'(n,l)$ proizvoljni elementi i neka je $r := d_{GH}(x,y) \geqslant 0$. Dovoljno je dokazati da za svako $\varepsilon > 0$ vrijedi $B_{d_{GH}}(y,\varepsilon) \cap B_{d_{GH}}(x,r) \neq \emptyset$. Neka je $\varepsilon > 0$ proizvoljno. Ukoliko je $r = 0$, tada tvrđenje trivijalno vrijedi. Stoga, neka su x i y različiti generalizovani stringovi. Tada postoje $i_0 \in \mathbb{N}_l$ i $j_0 \in \mathbb{N}_n$ tako da vrijedi $x[i_0](j_0) \neq y[i_0](j_0)$. Bez gubljenja na opštosti, može se pretpostaviti da je $x[i_0](j_0) < y[i_0](j_0)$. Kako je $\sum_{j=1}^{n} x[i_0](j) = 1 = \sum_{j=1}^{n} y[i_0](j)$, zaključuje se da postoji indeks $j_1 \in \{1,2,\ldots,n\} \setminus \{j_0\}$ takav da je $x[i_0](j_1) > y[i_0](j_1)$. Neka je $q \geqslant 1$ proizvoljan prirodan broj za koji vrijedi

$$\frac{\varepsilon}{2^q} < \min\{y[i_0](j_0), 1 - y[i_0](j_1), y[i_0](j_0) - x[i_0](j_0)\}$$

i $z \in S'(n,l)$ generalizovani string definisan na sljedeći način:

$$z[i](j) := \begin{cases} y[i_0](j_0) - \frac{\varepsilon}{2^q}, & \text{za } i = i_0, j = j_0; \\ y[i_0](j_1) + \frac{\varepsilon}{2^q}, & \text{za } i = i_0, j = j_1; \\ y[i](j), & \text{inače.} \end{cases}$$

Na osnovu ove definicije se dobija

$$\begin{aligned} d_{GH}(y,z) &= \sum_{i=1}^{l} \left(1 - \sum_{j=1}^{n} \min\{y[i](j), z[i](j)\}\right) \\ &= 1 - \left(y[i_0](1) + \ldots + \left(y[i_0](j_0) - \frac{\varepsilon}{2^q}\right) + \ldots + y[i_0](j_1) + \ldots \right. \\ &\quad \left. + y[i_0](n)\right) = 1 - \left(1 - \frac{\varepsilon}{2^q}\right) = \frac{\varepsilon}{2^q} < \varepsilon. \end{aligned}$$

Za kompletiranje dokaza, dovoljno je još dokazati da je $d_{GH}(x,z) < d_{GH}(x,y)$. Neka je $a := y[i_0](j_0) - x[i_0](j_0) > 0$ i $b := x[i_0](j_1) - y[i_0](j_1) > 0$. Tada vrijedi $\left|b - \frac{\varepsilon}{2^q}\right| < b + \frac{\varepsilon}{2^q}$, odakle slijedi $\left|a - \frac{\varepsilon}{2^q}\right| + \left|b - \frac{\varepsilon}{2^q}\right| < a + b$. Na osnovu ovoga, dalje se dobija

$$\begin{aligned} &\frac{x[i_0](j_0) + y[i_0](j_0) - a}{2} + \frac{x[i_0](j_1) + y[i_0](j_1) - b}{2} \\ &< \frac{x[i_0](j_0) + y[i_0](j_0) - \left|a - \frac{\varepsilon}{2^q}\right|}{2} + \frac{x[i_0](j_1) + y[i_0](j_1) - \left|b - \frac{\varepsilon}{2^q}\right|}{2}, \end{aligned}$$

odakle slijedi

$$x[i_0](j_0) + y[i_0](j_1) < \min\left\{x[i_0](j_0), y[i_0](j_0) - \frac{\varepsilon}{2^q}\right\} + \min\left\{x[i_0](j_1), y[i_0](j_1) + \frac{\varepsilon}{2^q}\right\}.$$

Konačno, vrijedi $\sum_{j=1}^{n} \min\{x[i_0](j), y[i_0](j)\} < \sum_{j=1}^{n} \min\{x[i_0](j), z[i_0](j)\}$, što implicira

$$\begin{aligned} d_{GH}(x,z) &= \sum_{i\in\{1,\dots,l\}\setminus\{i_0\}} \left(1 - \sum_{j=1}^{n} \min\{x[i](j), z[i](j)\}\right) \\ &\quad + \left(1 - \sum_{j=1}^{n} \min\{x[i_0](j), z[i_0](j)\}\right) \\ &< \sum_{i\in\{1,\dots,l\}\setminus\{i_0\}} \left(1 - \sum_{j=1}^{n} \min\{x[i](j), y[i](j)\}\right) \\ &\quad + \left(1 - \sum_{j=1}^{n} \min\{x[i_0](j), y[i_0](j)\}\right) = d_{GH}(x,y). \end{aligned}$$

□

Svi koncepti koji su uvedeni u slučaju skupa stringova $A \subseteq S(n,l)$ (pun kompleks C_A, filtracija pridružena skupu A, bar-kod $BC_k(A)$, itd.) mogu se uvesti i u slučaju konačnog skupa $A' \subseteq S'(n,l)$. Naravno, razlika je što se sada koristi metrika d_{GH} umjesto metrike d_H. Treba istaći da pri ovoj generalizaciji treba biti oprezan u pogledu postojanja nekih objekata čije postojanje nije bilo upitno u slučaju skupa $A \subseteq S(n,l)$.

Neka je C_A pun kompleks konačnog skupa $A \subseteq S'(n,l)$. Za $\sigma \in C_A$, vrijednost

$$r(\sigma) = \min\{r : (\exists x \in S'(n,l))(\forall y \in \sigma) d_{GH}(x,y) \leq r\}$$

se naziva *radijusom simpleksa* σ. Generalizovani string c takav da $d_{GH}(c,y) \leq r(\sigma)$ vrijedi za sve $y \in \sigma$, se naziva *centrom simpleksa* σ.

Lema 2.7.12. *Za svako $\sigma \in C_A$, minimum u definiciji $r(\sigma)$ postoji.*

Dokaz. Kako je $[0,1]$ kompaktan skup u uobičajnoj topologiji na $\mathbb{R}$, topološki proizvod $[0,1]^{\mathbb{N}_n}$ je također kompaktan skup. Najprije će biti pokazano da je $S'(n,l) = \{f \in [0,1]^{\mathbb{N}_n} : \sum_{j=1}^{n} f(j) = 1\}$ zatvoren potprostor od $[0,1]^{\mathbb{N}_n}$. Neka je $f \in \overline{S'(n,l)}$ proizvoljno. Tada postoji niz (f_q) u skupu $S'(n,l)$ koji konvergira

ka f. S obzirom da se konvergencija u topološkom proizvodu karakteriše preko koordinatnih nizova, za svako $j \in \mathbb{N}_n$ vrijedi da niz $(f_q(j))$ konvergira ka $f(j)$. Na osnovu toga, slijedi da niz $\left(\sum_{j=1}^{n} f_q(j)\right)$ konvergira ka $\sum_{j=1}^{n} f(j)$, a s druge strane, iz $f_q \in S'(n,l)$ slijedi da je $\left(\sum_{j=1}^{n} f_q(j)\right)$ niz jedinica, što implicira da je $\sum_{j=1}^{n} f(j) = 1$, odakle slijedi $f \in S'(n,l)$. Time je dokazano da je $S'(n,l)$ zatvoren potprostor kompaktnog skupa $[0,1]^{\mathbb{N}_n}$, te je ovaj skup i sam kompaktan. Metrika $d_{GH} : S'(n,l) \times S'(n,l) \to [0,+\infty)$ je neprekidna funkcija po svakoj od promjenljivih, što znači da je funkcija $\psi_\sigma : S'(n,l) \to \mathbb{R}$ definisana sa $\psi_\sigma(x) = \max\{d_{GH}(x,y) : y \in \sigma\}$ neprekidna kao maksimum konačno mnogo neprekidnih funkcija. S obzirom da je ova funkcija definisana na kompaktnom skupu $S'(n,l)$, ona dostiže minimalnu vrijednost i to je upravo $r(\sigma)$. □

Primjer 2.7.13. *Neka je* $s_1 = 111112$, $s_2 = 111113$, $s_3 = 222221$ *i* $s_4 = 333331$. *Za skup* $\sigma = \{s_1, s_2, s_3, s_4\} \subseteq S(3,6)$ *jedan centar je generalizovani string* $b \in S'(3,6)$ *dat matricom*

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc} \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ 1 & 0 & 0 \end{array}\right] \end{array}$$

Zaista, za svako $1 \le k \le 4$ *vrijedi* $d_{GH}(b, s_k) = 5 \cdot \frac{8}{15} + 1 = 5 \cdot \frac{11}{15} = \frac{11}{3}$. *Dalje, za proizvoljno* $c \in S'(3,6)$ *i svako* $1 \le i \le 5$ *vrijedi*

$$\sum_{k=2}^{4} \left(1 - \sum_{j=1}^{3} \min\{c[i](j), s_k[i](j)\}\right) = 2,$$

dok za $i = 6$ *važi*

$$\sum_{k=2}^{4} \left(1 - \sum_{j=1}^{3} \min\{c[6](j), s_k[6](j)\}\right) \ge 1,$$

pri čemu se minimum dostiže samo za $c[6] = b[6]$. *Zbog toga,* $\sum_{k=2}^{4} d_{GH}(c, s_k) \ge 11$, *što znači da je* $r(\sigma) \ge \frac{11}{3}$. *Još centara ovog simpleksa može se dobiti premještanjem težina između prvih pet pozicija, npr. jedan od centara je generalizovani string* b' *dat matricom*

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{5}{15} & \frac{5}{15} & \frac{5}{15} \\ \frac{9}{15} & \frac{3}{15} & \frac{3}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{7}{15} & \frac{4}{15} & \frac{4}{15} \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

Pri analizi filtracije pridružene skupu $A \subseteq S(n,l)$, simpleksi punog kompleksa C_A se dijele na pozitivne (one koji označavaju rađanje nove homološke klase) i negativne (one koji označavaju umiranje postojeće homološke klase). Kako $2^{|A|}-1$ simpleksa treba biti distribuirano u najviše $l+1$ nivoa filtracije, scenario u kojem dva ili više pozitivnih (ili negativnih) simpleksa imaju isti radijus je vrlo vjerovatan. Zbog toga, nema garancije da će filtracija pridružena skupu A biti Morzeova filtracija. Međutim, kako će ubrzo biti pokazano, moguće je konstruisati skup A' generalizovanih stringova tako da filtracija pridružena ovom skupu bude Morzeova filtracija. Što je još važnije, ova konstrukcija prouzrukuje striktno kontrolisana "pomjeranja" linija bar-koda $BC_k(A)$. Prije same konstrukcije, potrebni su dodatni pojmovi.

Zatvorena kugla radijusa $r \geqslant 0$ oko generalizovanog stringa $x \in S'(n,l)$ je skup $B[x,r] := \{y \in S'(n,l) : d_{GH}(x,y) \leq r\}$. $MB(\sigma) = \{B[c, r(\sigma)] : c$ je centar simpleksa $\sigma\}$ je *skup minikugli* "opisanih" oko simpleksa σ. U slučaju da centar simpleksa nije jedinstven, skup minikugli sadrži više od jedne minikugle. Pojam minikugle je detaljno ispitan u [100] u kontekstu Euklidskog prostora $\mathbb{R}^d$. Za unutrašnjost i granicu zatvorene kugle $B = B[x,r]$ u metričkom prostoru $(S'(n,l), d_{GH})$ koristiće se standardne oznake $int(B)$ i $bd(B)$.

Za konačan skup $A \subseteq S'(n,l)$, $G \subseteq A$ se naziva *skupom generatora* za A, ako postoji $B \in MB(A)$ tako da vrijedi $G \subseteq bdB$ and $A \setminus G \subseteq intB$.

Lema 2.7.14. *Svaki konačan skup $A \subseteq S'(n,l)$ ima minimalan skup generatora.*

Dokaz. Neka je $A \subseteq S'(n,l)$ proizvoljan konačan skup i G_1, G_2 skupovi generatora za A. Tada postoje odgovarajuće minikugle radijusa $r := r(A)$, sa centrima u c_1 i c_2. Ako je $G := G_1 \cap G_2$ i ukoliko se stavi $c[i](j) :=$

$\frac{c_1[i](j)+c_2[i](j)}{2}$, tada je $A \subseteq B[c,r]$. Zaista, za svako $x \in A$ vrijedi

$$\begin{aligned} d_{GH}(c,x) &= \sum_{i=1}^{l}\left(1-\sum_{j=1}^{n}\min\left\{\frac{c_1[i](j)+c_2[i](j)}{2}, x[i](j)\right\}\right) \\ &\leq \sum_{i=1}^{l}\left(1-\sum_{j=1}^{n}\frac{1}{2}\big(\min\{c_1[i](j),x[i](j)\}+\min\{c_2[i](j),x[i](j)\}\big)\right) \\ &= \frac{1}{2}\sum_{i=1}^{l}\big(1-\sum_{j=1}^{n}\min\{c_1[i](j),x[i](j)\}\big) \\ &+ \frac{1}{2}\sum_{i=1}^{l}(1-\sum_{j=1}^{n}\min\{c_2[i](j),x[i](j)\}) \\ &= \frac{1}{2}\big(d_{GH}(c_1,x)+d_{GH}(c_2,x)\big) \leq r. \end{aligned} \tag{2.8}$$

Primjećuje se da nejednakost data u (2.8) može postati jednakost samo ukoliko $x \in G$. Zbog toga, pretpostavka da su G_1 i G_2 disjunktni skupovi bi dovela do zaključka da postoji $r' < r$ takvo da $d_{GH}(c,x) \leqslant r'$ vrijedi za svako $x \in A$, što je nemoguće, jer je r radijus skupa A. To znači da je G neprazan skup koji "generiše" novu minikuglu radijusa r opisanu oko A. Dakle, presjek skupova generatora sadrži još jedan skup generatora, pa je presjek svih njih minimalan skup generatora. □

Primjer 2.7.15. *Neka su* $s_1 = 1111$, $s_2 = 2222$, $t = 1222$, $u = 1212$, $c_1 = 1122$, $c_2 = 2211$ *stringovi iz skupa* $S(2,4)$ *i neka je* $c_3 \in S'(2,4)$ *zadan matricom*

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Svaki od stringova c_1, c_2, c_3 *predstavlja centar simpleksa* $\sigma = \{s_1, s_2\}$ *i radijus odgovarajućih minikugli je* 2. *Međutim, kako* $t \in B[c_1,2] \setminus B[c_2,2]$, *prva od ovih minikugli je takođe opisana oko simpleksa* $\tau = \sigma \cup \{t\}$, *dok za drugu minikuglu to ne vrijedi. Zbog toga,* σ *je istovremeno minimalan skup generatora za* σ *i* τ. *Za simpleks* $\theta := \sigma \cup \{u\}$, *svi vrhovi od* θ *pripadaju granici kugle* $B[c_1,2]$, *što znači da je ovaj simpleks skup generatora za samoga sebe, ali je* σ *manji skup generatora:* $s_1, s_2 \in bdB[u,2]$, *dok je* $u \in intB[u,2]$.

Na konačnim podskupovima skupa $S'(n,l)$ moguće je uvesti binarnu relaciju $\approx$ na sljedeći način: $A \approx B$, ako A i B imaju isti minimalan skup generatora.

Jednostavno se provjerava da je $\approx$ relacija ekvivalencije. Ispostavlja se da, za datu filtraciju, simpleksi koji se ne mogu razdvojiti (bar u smislu metoda opisanog u nastavku) su tačno oni simpleksi koji pripadaju istoj klasi ekvivalencije relacije $\approx$.

Za $s \in S'(n,l)$ i $k < l$, neka je $s \restriction \mathbb{N}_k \in S'(n,k)$ oznaka za generalizovani string koji se sastoji od prvih k elemenata od s. Takođe, za dati simpleks $\sigma \in C_A$, neka je $C(\sigma)$ skup centara minikugli opisanih oko minimalnog skupa generatora za σ i neka je $D(\sigma,u) := \min\{d_{GH}(c,u) : c \in C(\sigma)\}$. Prije formulisanja tehnike kojom će se razdvajati radijusi simpleksa potreban je sljedeći rezultat koji predstavlja "motor" te tehnike.

Lema 2.7.16. *Neka je $A \subseteq S'(n,l)$ konačan skup generalizovanih stringova i neka su $\sigma_1, \sigma_2 \in C_A$ simpleksi takvi da je $r(\sigma_1) = r(\sigma_2) = r_0$ i $\sigma_1 \not\approx \sigma_2$. Takođe, neka je $j \in \mathbb{N}$ proizvoljno. Tada, postoji skup $B \subseteq S'(n,l+1)$, generalizovani string $z \in A$ i bijekcija $f : A \to B$ tako da vrijedi $r(f[\sigma_1]) = r(\sigma_1)$, $r(f[\sigma_2]) > r(\sigma_2)$ i*

$$r(\tau) \leq r(f[\tau]) \leq r(\tau) + \frac{1}{j} \tag{2.9}$$

za svako $\tau \in C_A$. Štaviše, vrijedi:

(i) ako $\frac{1}{j} < \min\left(\{|r(\tau) - r(\sigma)| : \sigma, \tau \in C_A\} \setminus \{0\}\right)$, tada $r(\sigma) < r(\tau)$ implicira $r(f[\sigma]) < r(f[\tau])$, za sve $\sigma, \tau \in C_A$;

(ii) ako je $\sigma \in C_A$, G minimalan skup generatora za σ, $z \notin G$ i $\frac{1}{j} < r(\sigma) - D(\sigma, z)$, tada je $f[G]$ minimalan skup generatora za $f[\sigma]$.

Dokaz. Neka su G_1 i G_2 minimalni skupovi generatora redom za σ_1 i σ_2. Uslov $\sigma_1 \not\approx \sigma_2$ povlači da su ovi skupovi različiti; bez gubljenja na opštosti može se pretpostaviti $G_2 \not\subseteq G_1$. To znači da postoji generalizovani string $z \in G_2 \setminus G_1$. Neka je funkcija f definisana na sljedeći način: za $s = a_1 a_2 \ldots a_l \in A$, $f(s) := a_1 a_2 \ldots a_l a_{l+1}$, pri čemu

- za $s \neq z$ vrijedi $a_{l+1}(1) := 1$ i $a_{l+1}(i) := 0$, za $i > 1$, i
- za $s = z$ vrijedi $a_{l+1}(1) := 1 - \frac{1}{j}$, $a_{l+1}(2) := \frac{1}{j}$ i $a_{l+1}(i) := 0$, za $i > 2$.

Sada, ako je $c = b_1 b_2 \ldots b_l$ centar minikugle $B[c, r_0]$ opisane oko σ_1, tada je $c' := b_1 b_2 \ldots b_l b_{l+1}$, gdje je $b_{l+1}(1) := 1$ i $b_{l+1}(i) = 0$ za $i > 1$, centar zatvorene kugle radijusa r_0 koja sadrži $f[\sigma_1]$, odakle slijedi $r(f[\sigma_1]) = r_0$.

Na sličan način se pokazuje da (2.9) vrijedi za proizvoljno $\tau \in C_A$.

Za proizvoljna dva generalizovana stringa $c \in S'(n, l+1)$ i $y \in \sigma_2$ vrijedi

$$\begin{aligned} d_{GH}(c, f(y)) &= \sum_{i=1}^{l+1} \left(1 - \sum_{j=1}^{n} \min\{c[i](j), f(y)[i](j)\}\right) \\ &= \sum_{i=1}^{l} \left(1 - \sum_{j=1}^{n} \min\{c[i](j), f(y)[i](j)\}\right) \\ &+ \left(1 - \sum_{j=1}^{n} \min\{c[l+1](j), f(y)[l+1](j)\}\right) \\ &= d_{GH}(c \restriction \mathbb{N}_l, y) + \left(1 - \sum_{j=1}^{n} \min\{c[l+1](j), f(y)[l+1](j)\}\right). \end{aligned}$$

Ako se pretpostavi da, za neko c_0, $d_{GH}(c_0, f(y)) \leq r_0$ vrijedi za svako $y \in \sigma_2$, tada slijedi da $d_{GH}(c_0 \restriction \mathbb{N}_l, y) \leq r_0$ vrijedi za svako $y \in G_2$, što bi značilo da $c_0 \restriction \mathbb{N}_l$ mora biti centar minikugle od σ_2. Međutim, za svako takvo c_0 vrijedi $c_0[l+1] \neq z[l+1]$, odakle slijedi $1 - \sum_{j=1}^{n} \min\{c_0[l+1](j), f(z)[l+1](j)\} > 0$ i posljedično $d_{GH}(c_0, f(z)) > r_0$. Zbog toga, vrijednost $r(f[\sigma_2])$ mora biti veća od r_0.

Tvrđenje (i) slijedi direktno iz (2.9). Konačno, za (ii), uslov $\frac{1}{j} < r(\sigma) - D(\sigma, z)$ garantuje da, s obzirom da z pripada unutrašnjosti neke minikugle $B[c, r]$, $f(z)$ pripada unutrašnjosti bar jedne minikugle (i to upravo $B[f(c), r]$) opisane oko $f[G]$. □

Važno je primijetiti da bijekcija opisana u prethodnoj lemi pomjera udesno nivoe istrajnog modula $M(Hom_k(C_A))$ za najviše $\frac{1}{j}$. Ova činjenica, zajedno sa Teoremom Stabilnosti, implicira da je udaljenost uskog grla između bar-kodova $BC_k(A)$ i $BC_k(B)$ manja ili jednaka $\frac{1}{j}$.

Nakon jedne primjene prethodne leme i dalje je moguće da postoje $\approx$-neekvivalentni simpleksi sa jednakim radijusom u punom kompleksu C_B. U cilju razdvajanja radijusa ovakvih simpleksa, nastavlja se sa sukcesivnim primjenjivanjem ove leme, sa prikladnim izborom brojeva $\frac{1}{j}$, koji će osigurati da u svakom koraku primjene, radijusi simpleksa koji su razdvojeni ranije ne postanu ponovo jednaki.

Teorema 2.7.17. ***(Tehnika razdvajanja radijusa simpleksa)*** *Neka je $A \subseteq S'(n, l)$ skup generalizovanih stringova takav da je $|A| = m$ i neka je $\varepsilon > 0$ proizvoljno. Tada postoje $Sep(A) \subseteq S'(n, l+m')$, za neko $m' \leq m$, i bijekcija $g : A \to Sep(A)$ tako da vrijedi*

(i) $r(g[\sigma]) \neq r(g[\tau])$ *za sve simplekse* $\sigma, \tau \in C_A$ *koji ispunjavaju uslov* $\sigma \not\approx \tau$,

(ii) $0 \leq r(g[\sigma]) - r(\sigma) < \varepsilon$ *za sve* $\sigma \in C_A$.

Dokaz. Prethodna lema se primjenjuje nekoliko puta, tako da se pri svakoj primjeni razdvajaju radijusi dva simpleksa, a radijusi ostalih mijenjaju za dovoljno malu veličinu. Na početku, neka je $A_0 := A$ i neka su $\sigma_1, \sigma_2 \in C_{A_0}$ simpleksi takvi da je $\sigma_1 \not\approx \sigma_2$, $r(\sigma_1) = r(\sigma_2)$. Tada je moguće izabrati z iz minimalnog skupa generatora npr. za σ_1 i neka je $j_1 \in \mathbb{N}$ tako da vrijedi

$$\frac{1}{j_1} < \min\left\{\frac{\varepsilon}{2}, h(z), \min(\{|r(\tau) - r(\sigma)| : \sigma, \tau \in C_{A_0}\} \setminus \{0\})\right\}.$$

gdje je $h(z) := \min(\{r(\sigma) - D(\sigma, z) : \sigma \in C_A\} \cap \mathbb{R}^+)$. Dalje se dobija skup $A_1 \subseteq S'(n, l+1)$ i bijekcija $f_1 : A_0 \to A_1$, tako da je ispunjeno $r(f_1[\sigma_1]) < r(f_1[\sigma_2])$ i $r(f_1[\sigma]) < r(f_1[\tau])$, kad god je $r(\sigma) < r(\tau)$, za $\sigma, \tau \in C_{A_0}$. Ovaj postupak se ponavlja i, birajući prirodne brojeve j_i za koje vrijedi $\frac{1}{j_i} < \frac{\varepsilon}{2^i}$, dobijaju se skupovi $A_2, A_3, \ldots, A_{m'}$, tako da u $Sep(A) := A_{m'}$ svi simpleksi koji nisu $\approx$-ekvivalentni imaju različite radijuse. Ovo dokazuje (i). Uslov (ii) iz prethodne leme implicira da za $\sigma \not\approx \tau$ vrijedi $f[\sigma] \not\approx f[\tau]$.

Na kraju, neka je $g := f_{m'} \circ \cdots \circ f_2 \circ f_1$. Jasno da za svako $\sigma \in C_A$ vrijedi $0 \leq r(g[\sigma]) - r(\sigma) \leq \frac{1}{j_1} + \frac{1}{j_2} + \cdots + \frac{1}{j_{m'}} < \varepsilon$, što dokazuje (ii). Takođe, m' ne može biti veće od m, jer svaki vrh z se "pomjera" najviše jednom (nakon njegovog pomjeranja on ne može biti element još jedne razlike $G_2 \setminus G_1$ skupova generatora simpleksa istog radijusa). □

Specijalno, uslovom (ii) prethodne teoreme se postiže da su "nove" linije (linije koje se pojavljuju u bar-kodu $BC_k(Sep(A))$, ali i ne u bar-kodu $BC_k(A)$) dužine manje od ε, kao i da se dužina svake "stare" linije bar-koda $BC_k(A)$ mijenja za manje od ε. Takođe, uočava se da samo $\approx$ −ekvivalentni simpleksi mogu eventualno imati jednake radijuse u punom kompleksu $C_{Sep(A)}$. U narednoj teoremi je pokazano da takve klase ekvivalencije ne utiču na bar-kod $BC_k(Sep(A))$.

Teorema 2.7.18. *Neka je* E *klasa ekvivalencije relacije* $\approx$ *sa bar dva elementa i neka je* $r_0 > 0$ *radijus svakog simpleksa iz* E. *Tada pojavljivanje simpleksa iz klase* E *ne utiče na bar-kod; preciznije, ako neki simpleks iz* E *dovodi do nastajanja* $k-$*ciklusa, za neku dimenziju* $k \geqslant 1$, *tada postoji drugi simpleks iz* E *koji će taj ciklus učiniti trivijalnim na istom* (r_0) *nivou filtracije.*

Dokaz. Kako se $C_{Sep(A)}$ dobija kao posljedica primjene Teoreme 2.7.17, jedini simpleksi sa radijusom r_0 su oni koji pripadaju klasi E. Neka je G zajednički minimalni skup generatora za simplekse iz klase E. Ovo znači da se E sastoji od

svih simpleksa σ takvih da je $G \subseteq \sigma$ i $\sigma \setminus G \subseteq intB[c, r_0]$, gdje je $B[c, r_0]$ kugla opisana oko G. Neka je $x \in intB[c, r_0] \setminus G$ proizvoljan generalizovani string koji pripada nekom od ovih simpleksa. Svi simpleksi iz E se mogu podijeliti u parove $(\sigma, \sigma \cup \{x\})$, gdje $x \notin \sigma$. Neka je $\langle(\sigma_i, \tau_i) : i < d\rangle$ jedna enumeracija svih takvih parova, koja je uređena sa $|\sigma_i| \leq |\sigma_j|$, za $i < j$. Efekat klase E biće najprije ispitan hipotetički za "scenario" kada se simpleksi iz E pojavljuju jedan po jedan u redoslijedu indeksa i, tj. kada su radijusi ovih simpleksa oblika $r(\sigma_i) = r_0 + 2i\delta$ i $r(\tau_i) = r_0 + (2i + 1)\delta$, za dovoljno malu vrijednost $\delta > 0$. Za ovu novu filtraciju (koja će biti označena sa $\mathcal{K}$) vrijedi $\mathcal{K}^{(r_0)} = C^{(r_0)}_{Sep(A)} \setminus E$ i $\mathcal{K}^{(r_0+(2d+1)\delta)} = C^{(r_0)}_{Sep(A)}$.

Dalje, neka je i fiksirano i $m := |\sigma_i|$. Svi $m-$elementni podskupovi od $\tau_i = \sigma_i \cup \{x\}$ izuzev σ_i imaju radijus manji od $r_0 + 2i\delta$. Zaista, svaki takav podskup ili ne sadrži G (što bi impliciralo da ima radijus manji od r_0: ako je G' minimalan skup generatora takvog simpleksa σ, tada, na osnovu dokaza Leme 2.7.14, $G \cap G'$ takođe sadrži skup generatora, pa je $G' \subset G$), ili je oblika τ_j, za neko $j < i$. Zbog toga, σ_i je pozitivan simpleks koji obilježava rađanje $m-$dimenzionalne homološke klase, a τ_i je negativan simpleks koji uništava istu tu klasu. Vraćajući se na situaciju u kojoj se svi simpleksi iz E pojavljuju na istom nivou filtracije, zaključuje se da ovi simpleksi nemaju nikakav uticaj na linije bar-koda. □

Napravljene su sve potrebne pripreme za uvođenje nove mjere sličnosti dva skupa stringova.

Neka su $A, B \subseteq S(n, l)$ dva skupa stringova, pri čemu je $|A| = |B| = m \geqslant 2$. Za svaku dimenziju $k \geqslant 0$, biće predloženo hibridno uparivanje k-dimenzionalnih linija i definisana udaljenost d_k između odgovarajućih bar-kodova. U ovom hibridnom uparivanju, prioritet će biti dat uparivanju linija koje odgovaraju ekvivalentnim ciklusima.

Za $k = 0$, već je ustanovljeno da oba bar-koda $BC_0(A)$ i $BC_0(B)$ sadrže m linija, pri čemu su po $m - 1$ njih konačne dužine (i sve su sa nulom kao donjom granicom), a tu je i par linija $[0, +\infty)$. Ako su $0 < l^A_1 \leqslant l^A_2 \leqslant \cdots \leqslant l^A_{m-1}$ i $0 < l^B_1 \leqslant l^B_2 \leqslant \cdots \leqslant l^B_{m-1}$ dužine linija, tada se uparuju linije $[0, l^A_i)$ i $[0, l^B_i)$ i definiše udaljenost $d_0(A, B) := \max_{i \in \{1,2,\ldots,m-1\}} |l^A_i - l^B_i|$.

Za dimenziju $k \geqslant 1$, koristi se tehnika razdvajanja radijusa simpleksa da bi se dobili $m-$elementni skupovi generalizovanih stringova $Sep(A)$ i $Sep(B)$. Ako su oba bar-koda $BC_k(Sep(A))$ i $BC_k(Sep(B))$ prazni, tj. nemaju linija, stavlja se $d_k(A, B) := 0$. U suprotnom, tehnika razdvajanja radijusa simpleksa se primjenjuje još jednom u cilju dobijanja skupa generalizovanih stringova $Sep(A \cup B)$. Pritom treba napomenuti da se separacija radijusa simpleksa iz $C_{A\cup B}$ može izvesti uključivanjem koraka koji su korišćeni pri separaciji radijusa simpleksa iz C_A i C_B, što za posljedicu ima $Sep(A) \subseteq Sep(A \cup B)$ i

$Sep(B) \subseteq Sep(A \cup B)$. Na ovaj način, obezbjeđuje se uslov da su filtracije pridružene skupovima generalizovanih stringova $Sep(A)$, $Sep(B)$ i $Sep(A \cup B)$ Morzeove filtracije, što znači da su bar-kodovi pridruženi odgovarajućim istrajnim modulima skupovi, a ne multiskupovi intervala.

Dalje, traže se potencijalni $C_{Sep(A\cup B)}-$ekvivalentni ciklusi i uparuju se linije koje im odgovaraju. Za linije koje se ne mogu upariti na ovaj način, koristi se uparivanje iz definicije udaljenosti uskog grla. Preciznije, ako su $BC'_k(Sep(A)) \subseteq BC_k(Sep(A))$ i $BC'_k(Sep(B)) \subseteq BC_k(Sep(B))$ kolekcije svih linija bar-koda za koje ne postoji $C_{Sep(A\cup B)}-$ekvivalentan "parnjak", stavlja se

$$d_k(A,B) := \sum_{\gamma_1 \sim \gamma_2} d_{BOT}\big(\{l(\gamma_1)\},\{l(\gamma_2)\}\big) + d_{BOT}\big(BC'_k(Sep(A)), BC'_k(Sep(B))\big), \tag{2.10}$$

gdje se prva suma uzima po svim parovima $C_{Sep(A\cup B)}-$ekvivalentnih $k-$ciklusa γ_1, γ_2, a $l(\gamma_1) \in BC_k(Sep(A)), l(\gamma_2) \in BC_k(Sep(B))$ su linije koje odgovaraju ovim ciklusima. Naravno, u slučaju kada ne postoje $C_{Sep(A\cup B)}-$ekvivalentni $k-$ciklusi, vrijedi $d_k(A,B) = d_{BOT}\big(BC_k(Sep(A)), BC_k(Sep(B))\big)$. Izbor da se porede linije bar-kodova $BC_k(Sep(A))$ i $BC_k(Sep(B))$ opravdan je činjenicom da, za svako $\varepsilon > 0$, skupovi $Sep(A)$ i $Sep(B)$ mogu biti izabrani tako da vrijedi

$$\begin{aligned} d_{BOT}\big(BC_k(A), BC_k(B)\big) &\leqslant \underbrace{d_{BOT}\big(BC_k(A), BC_k(Sep(A))\big)}_{\leqslant \frac{\varepsilon}{2}} \\ &+ d_{BOT}\big(BC_k(Sep(A)), BC_k(Sep(B))\big) + \underbrace{d_{BOT}\big(BC_k(Sep(B), BC_k(B)\big)}_{\leqslant \frac{\varepsilon}{2}} \\ &\leqslant d_{BOT}\big(BC_k(Sep(A)), BC_k(Sep(B))\big) + \varepsilon. \end{aligned}$$

Neka je $k_0 \geqslant 0$ minimalna dimenzija sa svojstvom $BC_k(Sep(A)) = \emptyset = BC_k(Sep(B))$, za svako $k > k_0$. Nova mjera sličnosti između skupova $A, B \subseteq S(n,l)$ iste kardinalnosti može se uvesti na sljedeći način:

$$d_{new}(A,B) := \sum_{k=0}^{k_0} \frac{2^k}{2^{k_0+1}-1} \cdot d_k(A,B).$$

Težine $\dfrac{2^k}{2^{k_0+1}-1}$, $0 \leqslant k \leqslant k_0$, se dodjeljuju u cilju davanja većeg značaja razlikama u homološkim atributima skupova A i B koje se ispoljavaju u višim dimenzijama, Udaljenost d_{new} ima svojstvo stabilnosti, jer je svaka udaljenost d_k definisana uz pomoć udaljenosti uskog grla između odgovarajućih skupova bar-kodova.

Glava 3

Vjerovatnosni metodi

U ovoj glavi razmotrene su mjere sličnosti stringova bazirane na nedeterminističkom pristupu, tj. mjere sličnosti zasnovane na vjerovatnosnim metodama. Osnovni pojmovi i rezultati navedeni u ovom dijelu mogu se naći npr. u [1], [2], [8], [9], [22], [26], [40], [41], [42], [43], [49], [52], [58], [61], [64], [74], [75], [81], [83], [89], [90], [96]. Takođe, uvedena je nova mjera sličnosti dva skupa stringova.

3.1 String kao stohastički proces

Neka je $T \subseteq \mathbb{N}$ najviše prebrojiv skup. *Diskretni stohastički proces* $\{X_n : n \in T\}$ na najviše prebrojivom skupu S je kolekcija slučajnih promjenljivih (veličina) sa vrijednostima u skupu S definisanih na prostoru vjerovatnoća $(\Omega, \mathcal{F}, P)$, gdje je P vjerovatnosna mjera uvedena na $\sigma-$algebri događaja $\mathcal{F}$ koja se posmatra u odnosu na prostor ishoda Ω. Skup T se najčešće interpretira kao vrijeme, a njegovi elementi kao *momenti* u kojima se stohastički proces posmatra. Obično se uzima da je $T = \{0, 1, \ldots, n\}$, za neko $n \in \mathbb{N}$, ili $T = \mathbb{N}$. Skup S predstavlja *skup stanja* datog stohastičkog procesa, dok je vrijednost $X_n \in S$ *stanje procesa u momentu* n. *Konačno-dimenzionalne raspodjele* stohastičkog procesa $\{X_n : n \in T\}$ određene su vjerovatnoćama

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}, i_0, i_1, \ldots, i_n \in S, n \geqslant 0.$$

Ove vjerovatnoće na jedinstven način određuju vjerovatnoće svih ishoda datog procesa. Posljedično, dva stohastička procesa (definisana na istom ili različitim prostorima vjerovatnoća) su podjednako distribuirana ukoliko imaju jednake konačno-dimenzionalne raspodjele. Jedan vid klasifikacije stohastičkih procesa podrazumijeva preciziranje zavisnosti koja postoji između slučajnih veličina koje određuju konačno-dimenzionalne raspodjele. Pored toga, stohastičke procese je moguće karakterisati preciziranjem načina na koji oni evoluiraju u vremenu.

Stohastički proces $\mathcal{X} = \{X_n : n \in T\}$ sa skupom stanja S, koji je najviše prebrojiv skup, je *lanac Markova*, ako za sve $i, j \in S$ postoje $p_{ij} \in [0, 1]$ tako da za svako $n + 1 \in T$ vrijedi

$$P\{X_{n+1} = j | X_0, X_1, \ldots, X_n\} = P\{X_{n+1} = j | X_n\}, \tag{3.1}$$

$$P\{X_{n+1} = j | X_n = i\} = p_{ij}. \tag{3.2}$$

Vrijednosti p_{ij} se nazivaju *tranzicionim vjerovatnoćama*. U principu, tranziciona vjerovatnoća p_{ij} predstavlja uslovnu vjerovatnoću prelaska procesa iz stanja i u stanje j. Za svako $i \in S$ ispunjen je uslov $\sum_{j \in S} p_{ij} = 1$. Matrica $\mathrm{P} = [p_{ij}]$ naziva se *tranzicionom matricom* ili *matricom prelaska u jednom koraku* datog lanca Markova. U ovoj tezi će isključivo biti razmatrani lanci Markova sa konačnim skupom stanja S.

Uslov (3.1) se naziva *svojstvom Markova*. Ovo svojstvo znači da, u bilo kojem momentu n, za dato trenutno stanje X_n, sljedeće stanje X_{n+1} ima uslovnu nezavisnost od stanja $X_0, X_1, \ldots, X_{n-1}$. Slikovito rečeno, svojstvo Markova je "odsustvo memorije": Sljedeće stanje (budućnost) zavisi isključivo od trenutnog stanja (sadašnjosti), a ne od stanja koja su prethodila trenutnom stanju (prošlosti). Svojstvo Markova je u prirodi većine pojava slučajnog karaktera i lanci Markova predstavljaju logičan izbor modela za ovakav tip pojava.

Uslov (3.2) garantuje da tranzicione vjerovatnoće ne zavise od izbora momenta n. U tom smislu lanac Markova je "vremenski homogen". Iako je moguće izostaviti ovaj uslov i raditi sa nehomogenim lancima Markova, u ovom radu taj slučaj neće biti razmatran.

Primjer 3.1.1. *Neka je* $(Y_n, n \geqslant 1)$ *niz nezavisnih slučajnih veličina sa cjelobrojnim vrijednostima i jednakom raspodjelom. Ako se stavi*

$$X_0 := 0, \quad X_n := \sum_{m=1}^{n} Y_m, n \geqslant 1,$$

tada se dobija stohastički proces $\{X_n : n \geqslant 0\}$ *koji se naziva slučajnim lutanjem. Slučajna veličina* Y_n *predstavlja veličinu pomaka u momentu n. Kako je* $X_{n+1} = X_n + Y_{n+1}$ *i slučajna veličina* Y_{n+1} *je nezavisna od slučajnih veličina* $X_0, X_1, \ldots, X_n$*, slijedi da za sve cijele brojeve* i, j *i svako* $n \geqslant 0$ *vrijedi*

$$\begin{aligned} P(X_{n+1} = j | X_0, X_1, \ldots, X_n = i\} &= P\{X_n + Y_{n+1} = j | X_n = i\} \\ &= P\{Y_n = j - i\} = P\{Y_1 = j - i\}. \end{aligned}$$

Zbog toga je slučajno lutanje lanac Markova sa tranzicionim vjerovatnoćama $p_{ij} = P\{Y_1 = j - i\}$*. Ako svaka slučajna veličina* Y_n *uzima isključivo vrijednosti* -1 *ili* 1*, tada je riječ o prostom slučajnom lutanju. Vjerovatnoće* $p := P\{Y_1 = 1\}$

i $q := P\{Y_1 = -1\} = 1 - p$ *u potpunosti određuju tranzicione vjerovatnoće, pri čemu je*

$$p_{i,i+1} = p, \;\; p_{i,i-1} = q, \;\; p_{i,j} = 0, j \notin \{i-1, i+1\}.$$

Primjedba 3.1.2. Svojstvo odsustva memorije je moguće generalizovati u vidu svojstva "odloženog" odsustva memorije, u smislu da na buduće stanje pored sadašnjosti utiče i određen broj stanja iz bliske prošlosti. Preciznije, stohastički proces $\{X_n : n \geqslant 0\}$ za koji postoji prirodan broj $r \geqslant 1$ tako da vrijedi

$$\begin{aligned} &P\{X_{n+1} = j | X_0, X_1, \ldots, X_{n-r+1} = i_{n-r+1}, \ldots, X_n = i_n\} \\ &= P\{X_{n+1} = j | X_{n-r+1} = i_{n-r+1}, \ldots, X_n = i_n\}, \end{aligned}$$

naziva se *lancem Markova reda* r. Specijalno, za $r = 1$ dobija se prethodno uveden lanac Markova.

Neka je $\{X_n : n \in T\}$ stohastički proces. Ako je $\{0, 1, \ldots, n\} \subseteq T$ i za ishod $\omega \in \Omega$ vrijedi $X_0(\omega) = i_0, X_1(\omega) = i_1, \ldots, X_n(\omega) = i_n$, tada se vektor $(i_0, i_1, \ldots, i_n)$ naziva *trajektorijom procesa* $\{X_n : n \in T\}$. Osnovni problem pri radu sa stohastičkim procesom je određivanje njegovih konačno-dimenzionalnih raspodjela, tj. nalaženje vjerovatnoće da slučajni vektor $(X_0, X_1, \ldots, X_n)$ poprimi konkretno datu trajektoriju. Ispostavlja se da su konačno-dimenzionalne raspodjele lanca Markova u potpunosti određene tranzicionim vjerovatnoćama i raspodjelom vjerovatnoća *početnog stanja* X_0.

Lema 3.1.3. *[81] Neka je* $\{X_n : n \geqslant 0\}$ *lanac Markova sa skupom stanja S, tranzicionim vjerovatnoćama* p_{ij} *i neka je* $\alpha_i := P\{X_0 = i\}$. *Tada, za svako* $n \geqslant 0$ *i sve* $i_0, i_1, \ldots, i_n$ *vrijedi*

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\} = \alpha_{i_0} \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-1} i_n}.$$

Dokaz. Dokaz će biti sproveden indukcijom po $n \geqslant 0$. Za $n = 0$, tvrđenje slijedi na osnovu definicije vrijednosti α_{i_0}. Neka je tvrđenje tačno za neko $n \geqslant 0$ i neka je $A_n = \{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}$. Tada je

$$\begin{aligned} &P\{\underbrace{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n}_{A_n}, X_{n+1} = i_{n+1}\} = P(A_n) \cdot P(X_{n+1} = i_{n+1} | A_n) \\ &= \alpha_{i_0} \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-1} i_n} \cdot \underbrace{P(X_{n+1} = i_{n+1} | X_n = i_n)}_{p_{i_n i_{n+1}}}, \end{aligned}$$

pri čemu posljednja jednakost slijedi iz svojstva Markova. Dakle, tvrđenje je ispunjeno i za $n + 1$, što znači da ono vrijedi za svako $n \geqslant 0$. □

Neka su $i_0, i_n \in S$ stanja lanca Markova $\{X_n : n \geqslant 0\}$. Uslovna vjerovatnoća $P\{X_n = i_n | X_0 = i_0\}$ se izražava kao vjerovatnoća realizacije trajektorije oblika

$(i_0, i_1, \ldots, i_{n-1}, i_n)$, gdje je $(i_1, i_2, \ldots, i_{n-1}) \in S^{n-1}$, što se, na osnovu prethodne leme, izražava kao suma $\sum_{(i_1,i_2,\ldots,i_{n-1})\in S^{n-1}} p_{i_0 i_1} \cdot p_{i_1 i_2} \cdot \ldots \cdot p_{i_{n-1} i_n}$. Jednostavno se uočava da posljednja suma predstavlja element matrice P^n, koja predstavlja $n-$ti stepen matrice tranzicionih vjerovatnoća P. To znači da za stanja $i, j \in S$, vjerovatnoća da se nakon realizovanja stanja i u momentu t ostvarilo stanje j u momentu $t + n$ predstavlja element matrice P^n u $i-$tom redu i $j-$toj koloni; nadalje će ova vjerovatnoća biti notirana sa P^n_{ij}. Takođe, ako je $\alpha_i := P\{X_0 = i\}$ i $\alpha := (\alpha_i)$ red-vektor, tada vrijedi $P\{X_n = j\} = (\alpha P^n)_j$, pri čemu desna strana predstavlja $j-$ti element red-vektora αP^n. Zbog toga, vjerovatnoća da lanac Markova u određenom momentu poprimi konkretno stanje zavisi isključivo od distribucije vjerovatnoća u početnom momentu i matrice tranzicionih vjerovatnoća.

Neka je $\mathcal{M} = \{X_n : n \geqslant 0\}$ lanac Markova čiji je skup stanja S i P matrica njegovih tranzicionih vjerovatnoća. Lanac $\mathcal{M}$ se naziva *ireducibilnim*, ako za svaka dva stanja $i, j \in S$ postoji prirodan broj $n = n(i, j)$ takav da je $\mathrm{P}^n_{ij} > 0$. Ovo znači da je u ireducibilnim lancima od bilo kojeg stanja moguće doći do bilo kojeg stanja koristeći isključivo prelaze sa pozitivnim tranzitnim vjerovatnoćama. Za stanje $i \in S$, neka je $\mathcal{T}(i) := \{t \geqslant 1 : \mathrm{P}^t_{ii} > 0\}$ skup vremenskih pomaka za koje je moguće da se iz stanja i lanac ponovo vrati u to stanje. Vrijednost $\mathrm{Per}(i) := \mathrm{NZD}(\mathcal{T}(i))$, gdje je NZD oznaka za najveći zajednički djelilac, naziva se *periodom stanja* i.

Lema 3.1.4. *[59] Ako je $\mathcal{M}$ ireducibilan lanac, tada za proizvoljna stanja $i, j \in S$ vrijedi $\mathrm{Per}(i) = \mathrm{Per}(j)$.*

Dokaz. Neka su $i, j \in S$ proizvoljna stanja. Zbog ireducibilnosti lanca, postoje prirodni brojevi k, l takvi da je $\mathrm{P}^k_{ij} > 0$ i $\mathrm{P}^l_{ji} > 0$. Za $k + l$ vrijedi $\mathrm{P}^{k+l}_{ii} \geqslant \mathrm{P}^k_{ij} \cdot \mathrm{P}^l_{ji} > 0$, što implicira $k + l \in \mathcal{T}(i)$, te $\mathrm{Per}(i)|(k + l)$. Neka je $m \in \mathcal{T}(j)$ proizvoljno. Tada je $\mathrm{P}^{k+m+l}_{ii} \geqslant \mathrm{P}^k_{ij} \cdot \mathrm{P}^m_{jj} \cdot \mathrm{P}^l_{ji} > 0$, pa je $k + l + m \in \mathcal{T}(i)$ i $\mathrm{Per}(i)|(k + l + m)$. Posljedično, dobija se $\mathrm{Per}(i)|m$. Dakle, $\mathrm{Per}(i)$ dijeli svaki element skupa $\mathcal{T}(j)$, odakle slijedi $\mathrm{Per}(i) \leqslant \mathrm{Per}(j)$. Zamjenom uloga i i j dobija se suprotna nejednakost. Dakle, vrijedi $\mathrm{Per}(i) = \mathrm{Per}(j)$. □

Za ireducibilan lanac, period tog lanca definiše se kao zajednički period svih njegovih stanja. Generalno, lanac se naziva *aperiodičnim*, ako sva njegova stanja imaju period 1. Ako lanac nije aperiodičan, tada se on naziva *periodičnim*. Ispostavlja se da za ireducibilan lanac koji je i aperiodičan postoji prirodan broj $k_0 \geqslant 1$ takav da su, za svako $k \geqslant k_0$, svi elementi $k-$tog stepena matrice prelaska ovog lanca pozitivne vrijednosti. Da bi se dokazala ova činjenica, potreban je pomoćni rezultat iz teorije brojeva, iskazan u sljedećoj lemi.

Lema 3.1.5. *[59] Neka je $L \subseteq \mathbb{N}$ neprazan skup, pri čemu je $\mathrm{NZD}(L) = g_L$. Tada postoji prirodan broj m_L takav da se, za svaki prirodan broj $m \geqslant m_L$,*

prirodan broj $m \cdot g_L$ može predstaviti kao linearna kombinacija elemenata iz skupa L, sa koeficijentima iz skupa $\mathbb{N}$.

Dokaz. Najprije treba primijetiti da postoji neprazan, konačan skup $K \subseteq L$ takav da je $\mathrm{NZD}(L) = \mathrm{NZD}(K)$. Zaista, ako je $n_0 := \min L$, tada nerastući niz $(\mathrm{NZD}(L \cap \mathbb{N}_n), n \geqslant n_0)$ može da ima konačno mnogo članova koji su manji od svojih prethodnika; skup K sastavljen od tih članova ima traženo svojstvo. Zbog toga, tvrđenje je dovoljno dokazati za slučaj konačnih podskupova $K \subseteq L$. Taj dokaz će biti izveden indukcijom po broju elemenata skupa K. U nastavku dokaza, podrazumijeva se da je $g_M := \mathrm{NZD}(M)$, gdje je $M \subseteq \mathbb{N}$ konačan skup.

Za $|K| = 1$ je $K = \{g_K\}$, pa u ovom slučaju tvrđenje trivijalno vrijedi. Za $K = \{a, b\}$ i dati prirodan broj $m > 0$ moguće je izabrati cijele brojeve c_m, d_m tako da vrijedi $m \cdot g_K = c_m a + d_m b$. Pritom, prethodni prikaz nije jedinstven, jer vrijedi $m \cdot g_K = (c_m + kb)a + (d_m - ka)b$, za proizvoljno $k \in \mathbb{Z}$. To se može iskoristiti da se u prikazu $m \cdot g_K = c_m a + d_m b$ broj c_m uzme tako da ima svojstvo $0 \leqslant c_m < b$. Stavljanjem $m_K := \dfrac{a(b-1)-b}{g_K} + 1$, dobija se da za svako $m \geqslant m_K$ vrijedi $c_m a + d_m b = m \cdot g_K \geqslant m_K \cdot g_K > a(b-1) - b \geqslant c_m a - b$, odakle slijedi $d_m \geqslant 0$. Time je tvrđenje dokazano za dvočlani skup K.

Neka tvrđenje vrijedi za neprazan, konačan skup F i neka je $K := F \cup \{a\}$, gdje je $a \geqslant 1$ proizvoljan prirodan broj koji ne pripada skupu F. Jasno da vrijedi $g_K = \mathrm{NZD}(a, g_F)$. Kako tvrđenje vrijedi za skupove F i $\{a, g_F\}$ (za F vrijedi zbog indukcione pretpostavke, a za skup $\{a, g_F\}$ na osnovu baze indukcije), moguće je izabrati prirodne brojeve m_F i $m_{\{a,g_F\}}$ sa odgovarajućim svojstvom. Neka je $m_K := m_{\{a,g_F\}} + \dfrac{m_F \cdot g_F}{g_K}$; dovoljno je dokazati da, u smislu posmatranog svojstva, ovaj prirodan broj odgovara skupu K.

Neka je $m \geqslant m_K$ proizvoljan prirodan broj. Tada je $m \cdot g_K \geqslant m_K \cdot g_K = m_{\{a,g_F\}} \cdot g_K + m_F \cdot g_F$, odakle slijedi $\dfrac{m \cdot g_K - m_F \cdot g_F}{g_K} \geqslant m_{\{a,g_F\}}$. To znači da je za prirodan broj $l := \dfrac{m \cdot g_K - m_F \cdot g_F}{g_K}$ moguć prikaz $l \cdot \underbrace{g_{\{a,g_F\}}}_{=g_K} = c_l \cdot a + d_l \cdot g_F$, za neke prirodne brojeve c_l i d_l. Na osnovu toga, dalje se dobija $m \cdot g_K - m_F \cdot g_F = c_l \cdot a + d_l \cdot g_F$, tj. $m \cdot g_K = c_l \cdot a + (d_l + m_F) d_F$. Kako je $q := d_l + m_F \geqslant m_F$, moguć je prikaz $q \cdot g_F = \sum_{f \in F} c_q(f) \cdot f$, za neke prirodne brojeve $c_q(f), f \in F$. Konačno, traženi prikaz je oblika $m \cdot g_K = c_l \cdot a + \sum_{f \in F} c_q(f) \cdot f$, pa tvrđenje vrijedi i za skup K. □

Lema 3.1.6. *[59] Ako je $\mathcal{M}$ ireducibilan i aperiodičan lanac, tada postoji prirodan broj $k_0 \geqslant 1$ takav da je $\mathrm{P}_{ij}^k > 0$ za sve $i, j \in S$ i svako $k \geqslant k_0$.*

Dokaz. Direktna posljedica prethodne leme je da svaki podskup skupa prirodnih brojeva koji je zatvoren u odnosu na sabiranje i ima najveći zajednički djelilac jednak 1 ima svojstvo da sadrži sve, sem konačno mnogo prvih prirodnih brojeva. Neka je $i \in S$ proizvoljno. Kako je lanac aperiodičan, vrijedi $\mathrm{Per}(i) = \mathrm{NZD}(\mathcal{T}(i)) = 1$. Skup $\mathcal{T}(i)$ je zatvoren u odnosu na sabiranje, jer za proizvoljne $k, l \in \mathcal{T}(i)$ vrijedi $\mathrm{P}_{ii}^{k+l} \geqslant \mathrm{P}_{ii}^{k} \cdot \mathrm{P}_{ii}^{l} > 0$. Zbog toga, postoji prirodan broj $k(i)$ takav da za svako $k \geqslant k(i)$ vrijedi $k \in \mathcal{T}(i)$. Koristeći ireducibilnost lanca $\mathcal{M}$, može se zaključiti da za proizvoljno $j \in S$ postoji prirodan broj $k(i,j) \geqslant 1$ takav da je $\mathrm{P}_{ij}^{k(i,j)} > 0$. To znači da za svaki prirodan broj $k \geqslant k(i) + k(i,j)$ vrijedi

$$\mathrm{P}_{ij}^{k} \geqslant \mathrm{P}_{ii}^{k-k(i,j)} \cdot \mathrm{P}_{ij}^{k(i,j)} > 0.$$

Na osnovu toga, ako je $k_0(i) := k(i) + \max\limits_{j \in S} k(i,j)$, tada za svako $k \geqslant k_0(i)$ i svako $j \in S$ vrijedi $\mathrm{P}_{ij}^{k} > 0$. Konačno, za $k_0 := \max\limits_{i \in S} k_0(i)$ i svako $k \geqslant k_0$ vrijedi $\mathrm{P}_{ij}^{k} > 0$, za proizvoljne $i, j \in S$. □

Neka je $\mathcal{M} = \{X_n : n \geqslant 0\}$ lanac Markova čiji je skup stanja S i P matrica njegovih tranzicionih vjerovatnoća. Raspodjela vjerovatnoća π na skupu stanja S naziva se *stacionarnom raspodjelom* lanca $\mathcal{M}$, ako vrijedi $\pi = \pi \cdot \mathrm{P}$, tj. ako za svako $j \in S$ vrijedi $\pi_j = \sum\limits_{i \in S} \pi_i \cdot \mathrm{P}_{ij}$. Vjerovatnoće stacionarne raspodjele nazivaju se *finalnim vjerovatnoćama*. Naziv proističe iz toga što, posmatrano na duže staze, a bez obzira na zadatu početnu raspodjelu vjerovatnoća, proporcija vremena koju lanac Markova provede u stanju $j \in S$ približno je jednaka vrijednosti $\pi(j)$, za svako stanje $j \in S$. To znači da niz raspodjela vjerovatnoća koji se dobije kada se polazna raspodjela (sastavljena od pozitivnih vjerovatnoća) "ubaci" u lanac Markova ima graničnu raspodjelu koja je jednaka stacionarnoj raspodjeli. Naravno, sve ovo vrijedi u slučaju da za posmatrani lanac Markova postoji jedinstvena stacionarna raspodjela. Ispostavlja se da ireducibilnost i aperiodičnost lanca Makova garantuju postojanje njegove stacionarne raspodjele. U dokazu te činjenice biće korišćen sljedeći pomoćni rezultat.

Lema 3.1.7. *[71] Neka je $\mathcal{M}$ lanac Markova sa konačnim, nepraznim skupom stanja S i matricom tranzicionih vjerovatnoća P čije su sve vrijednosti pozitivne. Ako su za kolonu $j \in S$, sa $m_n(j)$ i $M_n(j)$ redom označeni minimalni i maksimalni element u $j-$toj koloni matrice P^n, tada, za svako $j \in S$, vrijedi*

(i) Niz $(m_n(j), n \geqslant 1)$ je neopadajući i niz $(M_n(j), n \geqslant 1)$ je nerastući niz,
(ii) $\lim\limits_{n \to +\infty} m_n(j) = \lim\limits_{n \to +\infty} M_n(j)$.

Dokaz. Neka je $j \in S$ proizvoljno.

(i) Za svaki prirodan broj n vrijedi

$$\begin{aligned} m_{n+1}(j) &= \min_i \mathrm{P}_{ij}^{n+1} = \min_i \sum_k \mathrm{P}_{ik} \cdot \mathrm{P}_{kj}^n \geqslant \min_i \sum_k \mathrm{P}_{ik} \cdot m_n(j) \\ &= m_n(j) \cdot \min_i \underbrace{\sum_k \mathrm{P}_{ik}}_{=1} = m_n(j), \end{aligned}$$

što znači da je $(m_n(j), n \geqslant 1)$ neopadajući niz. Na sličan način se dokazuje da je $(M_n(j), n \geqslant 1)$ nerastući niz.

(ii) Na osnovu dijela (i), nizovi $(m_n(j), n \geqslant 1)$ i $(M_n(j), n \geqslant 1)$ su monotoni nizovi, a, kako su ujedno i ograničeni nizovi (jer su njihovi članovi vjerovatnoće), ovi nizovi su konvergentni. Ostaje još da se dokaže da imaju istu graničnu vrijednost. Neka je n proizvoljan prirodan broj i l_0 red matrice P^n tako da je $M_n(j) = \mathrm{P}_{l_0 j}^n$. Tada, za proizvoljno $k \in S$, vrijedi

$$\begin{aligned} \mathrm{P}_{kj}^{n+1} &= \sum_l \mathrm{P}_{kl} \cdot \mathrm{P}_{lj}^n = \mathrm{P}_{kl_0} \cdot M_n(j) + \sum_{l \neq l_0} \mathrm{P}_{kl} \cdot \mathrm{P}_{lj}^n \\ &\geqslant \mathrm{P}_{kl_0} \cdot M_n(j) + \left(1 - \mathrm{P}_{kl_0}\right) \cdot m_n(j) = m_n(j) + \mathrm{P}_{kl_0} \left(M_n(j) - m_n(j)\right) \\ &\geqslant m_n(j) + \mathrm{P}_{\min} \left(M_n(j) - m_n(j)\right), \end{aligned}$$

gdje je $\mathrm{P}_{\min} > 0$ minimum svih elemenata matrice tranzicionih vjerovatnoća P. Prethodna nejednakost vrijedi za svako k, odakle slijedi

$$m_{n+1}(j) \geqslant m_n(j) + \mathrm{P}_{\min} \left(M_n(j) - m_n(j)\right),$$

tj.

$$0 \leqslant M_n(j) - m_n(j) \leqslant \frac{m_{n+1}(j) - m_n(j)}{\mathrm{P}_{\min}}.$$

Ako se u posljednjoj nejednakosti pusti da n teži u beskonačno, njena desna strana teži ka nuli, što znači da je $\lim\limits_{n \to +\infty} \left(M_n(j) - m_n(j)\right) = 0$, tj. $\lim\limits_{n \to +\infty} m_n(j) = \lim\limits_{n \to +\infty} M_n(j)$. Važno je napomenuti da uslov pozitivnosti elemenata matrice P garantuje da je zajednička granična vrijednost pozitivan broj. □

Konačno, prethodni pomoćni rezultati omogućavaju da se dokaže sljedeća teorema.

Teorema 3.1.8. *[59] (**Fundamentalna teorema za lance Markova**) Svaki ireducibilan i aperiodičan lanac Markova čiji je skup stanja konačan ima jedinstvenu stacionarnu raspodjelu.*

Dokaz. Neka je $\mathcal{M}$ ireducibilan i aperiodičan lanac Markova sa konačnim, nepraznim skupom stanja $S = \{1, 2, \ldots, |S|\}$ i matricom tranzicionih vjerovatnoća P. Lema 3.1.6 garantuje postojanje prirodnog broja n_0 za svojstvom da je,

za svako $n \geqslant n_0$, matrica P^n sastavljena od pozitivnih vrijednosti. Zbog toga se Lema 3.1.7 može primijeniti na nizove $(m_n(j), n \geqslant n_0)$ i $(M_n(j), n \geqslant n_0)$. Za $j \in S$, neka je $\pi_j := \lim\limits_{n\to+\infty} m_n(j) = \lim\limits_{n\to+\infty} M_n(j) > 0$. Cilj je dokazati da je $\pi := \left(\pi_1, \pi_2, \ldots, \pi_{|S|}\right)$ stacionarna raspodjela za lanac Markova $\mathcal{M}$. Za početak, treba pokazati da je u pitanju distribucija vjerovatnoća na skupu stanja S. Zaista, najprije treba primijetiti da za sve $i, j \in S$ i svako $n \geqslant n_0$ vrijedi $m_n(j) \leqslant \mathrm{P}_{ij}^n \leqslant M_n(j)$, što implicira da za sve $i, j \in S$ vrijedi $\pi_j = \lim\limits_{n\to+\infty} \mathrm{P}_{ij}^n$. Na osnovu toga, fiksirajući proizvoljno $i \in S$, dobija se $\sum\limits_{j\in S} \pi_j = \sum\limits_{j\in S} \lim\limits_{n\to+\infty} \mathrm{P}_{ij}^n = \lim\limits_{n\to+\infty} \sum\limits_{j\in S} \mathrm{P}_{ij}^n = 1.$

Dalje, treba provjeriti da je distribucija vjerovatnoća π stacionarna raspodjela lanca $\mathcal{M}$. Fiksirajući proizvoljno $l \in S$, za svako $j \in S$ vrijedi

$$\pi_j = \lim_{n\to+\infty} \mathrm{P}_{lj}^{n+1} = \lim_{n\to+\infty} \sum_i \mathrm{P}_{li}^n \cdot \mathrm{P}_{ij} = \sum_i \left(\lim_{n\to+\infty} \mathrm{P}_{li}^n\right) \cdot \mathrm{P}_{ij} = \sum_i \pi_i \cdot \mathrm{P}_{ij},$$

što znači da je π stacionarna raspodjela za lanac $\mathcal{M}$.

Za dokaz jedinstvenosti date stacionarne raspodjele najprije će biti dokazano da "ubacivanje" proizvoljne početne distribucije u lanac Markova generiše niz distribucija vjerovatnoća koji konvergira ka stacionarnoj raspodjeli. Neka je $\tau := (\tau_1, \tau_2, \ldots, \tau_{|S|})$ proizvoljna distribucija vjerovatnoća početnog stanja lanca Markova $\mathcal{M}$, pri čemu je $\tau_j > 0$, za svako $j \in S$. Ubacivanje ove raspodjele u lanac Markova $\mathcal{M}$ se formalizuje uvođenjem niza raspodjela vjerovatnoća $\left(\tau^{(n)}\right)$, koji je rekurzivno definisan sa $\tau^{(0)} := \tau$, $\tau^{(n+1)} := \tau^{(n)} \cdot \mathrm{P}$, sa napomenom da je posljednji produkt ustvari rezultat matričnog množenja red-vektora $\left[\tau_1^{(n)}, \tau_2^{(n)}, \ldots, \tau_{|S|}^{(n)}\right]$ sa matricom P tranzicionih vjerovatnoća. Za svako $j \in S$ vrijedi

$$\begin{aligned}
\lim_{n\to+\infty} \tau_j^{(n+1)} &= \lim_{n\to+\infty} \sum_{i\in S} \tau_i^{(n)} \cdot \mathrm{P_{ij}} = \lim_{n\to+\infty} \sum_{i\in S} \left(\sum_{k\in S} \tau_k^{(n-1)} \cdot \mathrm{P}_{ki}\right) \mathrm{P}_{ij} \\
&= \lim_{n\to+\infty} \sum_{k\in S} \tau_k^{(n-1)} \cdot \underbrace{\sum_{i\in S} \mathrm{P}_{ki} \cdot \mathrm{P}_{ij}}_{=\mathrm{P}_{kj}^2} = \ldots = \lim_{n\to+\infty} \sum_{l\in S} \tau_l^{(0)} \cdot \mathrm{P_{lj}^n} \\
&= \sum_{l\in S} \tau_l \cdot \underbrace{\lim_{n\to+\infty} \mathrm{P}_{lj}^n}_{=\pi_j} = \pi_j \cdot \underbrace{\sum_{l\in S} \tau_l}_{=1} = \pi_j,
\end{aligned}$$

što znači da je $\lim\limits_{n\to+\infty} \tau^{(n)} = \pi$.

Na kraju, slijedi dokaz jedinstvenosti stacionarne raspodjele lanca Markova. Neka je, pored razmatrane stacionarne raspodjele π, data još jedna stacionarna

raspodjela π' lanca Markova $\mathcal{M}$, skoncentrisana na skupu stanja S. Ako se uzme da je π' distribucija vjerovatnoća početnog stanja lanca Markova $\mathcal{M}$ i $\left(\pi'^{(n)}\right)$ niz raspodjela vjerovatnoća definisan kao u prethodnom, tada se dobija

$$\pi = \lim_{n\to+\infty} \pi'^{(n)} = \lim_{n\to+\infty} \pi'^{(n-1)} \cdot \mathrm{P} = \ldots = \lim_{n\to+\infty} \pi' \cdot \mathrm{P}^n.$$

S obzirom da je i π' stacionarna raspodjela, za svaki prirodan broj $n \geqslant 1$ vrijedi

$$\pi' \cdot \mathrm{P}^n = \underbrace{\pi' \cdot \mathrm{P}}_{=\pi'} \cdot \mathrm{P}^{n-1} = \ldots = \pi' \cdot \mathrm{P} = \pi',$$

što znači da je $\pi = \pi'$. □

Neka je $\mathcal{M} = \{X_n : n \geqslant 0\}$ lanac Markova čiji je skup stanja S i $\mathrm{P}^{\mathcal{M}}$ matrica tranzicionih vjerovatnoća koja određuje sve njegove konačno-dimenzionalne raspodjele. Na taj način, određena je vjerovatnosna mjera $P^{\mathcal{M}}$ na prostoru vjerovatnoća $\prod_{m\geqslant 0} S_m$, gdje je $S_m := S$, za svako m. Ako je $t = (i_n, n \geqslant 0)$ trajektorija datog lanca Markova, tada je za prirodan broj n moguće definisati vrijednost

$$M_n(P^{\mathcal{M}}, t) := \frac{\log_2\left(P^{\mathcal{M}}\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}\right)}{n+1}.$$

Sljedeća teorema je u literaturi poznata kao **Petrijeva granična teorema**. Njen dokaz je ovdje izostavljen, jer izlazi van opsega ove teze, a može se naći npr. u [73].

Teorema 3.1.9. *$M(P^{\mathcal{M}}) := \lim_{n\to+\infty} M_n(P^{\mathcal{M}}, t)$ postoji skoro sigurno u smislu vjerovatnosne mjere $P^{\mathcal{M}}$.*

Formiranje stringa X dužine $l \geqslant 1$ nad alfabetom $\mathbb{N}_n = \{1, 2, \ldots, n\}$ se može shvatiti kao diskretni slučajni proces $(X_i, i \in \{1, \ldots, l\})$, gdje je X_i slučajna veličina koja predstavlja slučajan izbor simbola iz $\mathbb{N}_n$ koji će biti na $i-$toj poziciji posmatranog stringa. U tom kontekstu, elementi alfabeta $\mathbb{N}_n$ su stanja ovog procesa, a svaki konkretan string predstavlja jednu njegovu trajektoriju. Pritom se, zbog sekvencijalne prirode stringa, može pretpostaviti da, za svako $i \geqslant 2$, slučajna veličina X_i u nekom smislu zavisi od slučajnih veličina X_j, $j \in \{1, \ldots, i-1\}$.

Ako je $x = x_1 x_2 \ldots x_l$ konkretan string, vjerovatnoća da string X poprimi vrijednost x se označava sa $P(X = x)$ ili kraće sa $P(x)$. Na sličan način, uslovna vjerovatnoća da X_i poprimi vrijednost x_i, ukoliko, za svako $j < i$, X_j poprimi vrijednost x_j, se označava sa $P\left(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}\right)$ ili

kraće sa $P(x_i|x_1x_2\ldots x_{i-1})$. Vjerovatnoća $P(x_1x_2\ldots x_l)$ se preko pravila množenja vjerovatnoća može izraziti na sljedeći način:

$$P(x_1x_2\ldots x_l) = P(x_1)\cdot P\left(x_2|x_1\right)\cdot P\left(x_3|x_1x_2\right)\cdot\ldots\cdot P\left(x_l|x_1x_2\ldots x_{l-1}\right). \quad (3.3)$$

Kao što se može primijetiti iz prethodne formule, ključno je na neki način izraziti uslovne vjerovatnoće sa desne strane. Za diskretni slučajni proces koji opisuje formiranje stringa je prirodno pretpostaviti da predstavlja lanac Markova određenog reda, što donekle pojednostavljuje dobijanje uslovnih vjerovatnoća u formuli (3.3). Npr., ako se pretpostavi da je string X lanac Markova reda $r = 1$, tada se prethodna formula svodi na formulu $P(x_1x_2\ldots x_l) = P(x_1)\cdot P\left(x_2|x_1\right)\cdot P\left(x_3|x_2\right)\cdot\ldots\cdot P\left(x_l|x_{l-1}\right)$, pa se uslovne vjerovatnoće svode na tranzicione vjerovatnoće, tj. vjerovatnoće iz matrice prelaska u jednom koraku.

Način modelovanja pomenutih uslovnih vjerovatnoća biće izložen u kasnijim sekcijama ove glave. U narednoj sekciji je opisano kako se može doći do vjerovatnosne mjere sličnosti dva skupa stringova. Ovaj pristup podrazumijeva da je na neki način određena vjerovatnoća iz formule (3.3).

3.2 Relativna entropija i vjerovatnosne mjere sličnosti familija stringova

Neka je X diskretna slučajna veličina čiji je skup vrijednosti najviše prebrojiv skup S_X i neka je $p_X(x) := P\{X = x\}, x \in S_X$, raspodjela vjerovatnoća ove slučajne veličine. *Entropija* slučajne veličine X predstavlja mjeru neizvjesnosti ove slučajne veličine i uvodi se na sljedeći način:

$$H(X) := -\sum_{x\in S_X} p_X(x)\cdot\log_2 p_X(x).$$

Prethodna definicija je korektna i za vrijednosti X čije su vjerovatnoće jednake nuli. Zaista, i ove vrijednosti mogu biti sabirci prethodne sume ako se dodefiniše $0\cdot\log_2 0 := 0$, što je opravdano činjenicom $\lim\limits_{x\to 0+} x\cdot\log_2 x = 0$. S obzirom na prisustvo logaritma sa bazom 2, entropija se izražava u bitima. Takođe, uočava se da za svaku slučajnu veličinu X vrijedi $H(X) \geqslant 0$.

Ako je X diskretna slučajna veličina i g funkcija, tada se *matematičko očekivanje* slučajne veličine $g(X)$ dobija po formuli $E\left(g(X)\right) := \sum\limits_{x\in S_X} g(x)\cdot p_X(x)$. U tom smislu, entropija može da se shvati kao matematičko očekivanje slučajne veličine $\log_2\dfrac{1}{p(X)}$, gdje se X ravna po datoj raspodjeli vjerovatnoća $p_X(x), x \in S_X$. Na osnovu toga, slijedi da entropija ne zavisi od vrijednosti slučajne veličine X, već isključivo od raspodjele njihovih vjerovatnoća. Zbog toga, vrijednost $H(X)$ može da se notira i sa $H(p_X)$.

Primjer 3.2.1. *Ako je* $X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, *tada je*

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p).$$

Specijalno, ako je X indikator pojavljivanja pisma kod homogenog novčića, tada je $p = \frac{1}{2}$, *pa je* $H(X) = 1$.

S druge strane, ako je $X : \begin{pmatrix} a & b & c & d \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$, *tada je*

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = \frac{7}{4}.$$

Prethodna vrijednost može da se interpretira kao očekivan broj binarnih (dane) pitanja koje je potrebno postaviti kako bi se ustanovila vrijednost slučajne veličine X.

Zajednička entropija $H(X,Y)$ para diskretnih slučajnih veličina X i Y sa datom zajedničkom raspodjelom $p_{(X,Y)}$ definiše se sa

$$H(X,Y) := -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 p_{(X,Y)}(x,y).$$

Uslovna entropija $H(Y|X)$ para diskretnih slučajnih veličina X i Y sa datom zajedničkom raspodjelom $p_{(X,Y)}$ definiše se sa

$$H(Y|X) := -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 p_{Y|X}(x,y),$$

gdje je $p_{Y|X}$ uslovna raspodjela data sa $p_{Y|X}(x,y) := \dfrac{p_{(X,Y)}(x,y)}{p_X(x)}$. Veza između ove dvije entropije opisana je u narednoj lemi.

Lema 3.2.2. *[26] Za diskretne slučajne veličine X i Y vrijedi*

$$H(X,Y) = H(X) + H(Y|X).$$

Dokaz. Koristeći pravilo množenja vjerovatnoća dobija se

$$\begin{aligned} H(X,Y) &= -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 p_{(X,Y)}(x,y) \\ &= -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 \left(p_X(x) \cdot p_{Y|X}(x,y) \right) \\ &= -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 p_X(x) \\ &\quad -\sum_{x \in S_X} \sum_{y \in S_Y} p_{(X,Y)}(x,y) \cdot \log_2 p_{Y|X}(x,y) = H(X) + H(Y|X), \end{aligned}$$

što je i trebalo dokazati. □

Primjedba 3.2.3. *Prethodna lema se može induktivno uopštiti za diskretne slučajne veličine $X_1, \ldots, X_n$, $n \geqslant 2$. U tom slučaju, odgovarajuća jednakost je oblika* $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$.

Primjer 3.2.4. *Neka slučajne veličine X i Y imaju zajedničku raspodjelu vjerovatnoća*

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Marginalna raspodjela za X je $X : \begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$, *dok je marginalna raspodjela za Y uniformna raspodjela na skupu* $\{1, 2, 3, 4\}$. *Na osnovu toga, dobija se* $H(X) = \frac{7}{4}$ *i* $H(Y) = 2$. *Takođe, vrijedi* $H(X|Y) = \frac{11}{8}$, $H(Y|X) = \frac{13}{8}$ *i* $H(X, Y) = \frac{27}{8}$. *Primjećuje se da su uslovne entropije* $H(X|Y)$ *i* $H(Y|X)$ *različite.*

Relativna entropija je mjera različitosti dvije raspodjele vjerovatnoća. Preciznije, ako su p i q dvije raspodjele vjerovatnoća sa istim nosačem, tada se *Kulbak-Lajblerova mjera divergencije* ili *relativna entropija* definiše sa

$$D_{KL}(p||q) := \sum_x p(x) \cdot \log_2 \frac{p(x)}{q(x)}.$$

U prethodnoj definiciji, koristi se konvencija $0 \cdot \log_2 \frac{0}{0} := 0$, a takođe se, na osnovu neprekidnosti, koristi $0 \cdot \log_2 \frac{0}{q} := 0$ i $p \cdot \log_2 \frac{p}{0} := +\infty$. Stoga, ako postoji vrijednost x za koju je $p(x) > 0$ i $q(x) = 0$, tada je $D_{KL}(p||q) = +\infty$. Situacija da relativna entropija poprima vrijednost $+\infty$ može se izbjeći nametanjem uslova *apsolutne neprekidnosti* vjerovatnosne mjere p u odnosu na vjerovatnosnu mjeru q, što efektivno znači da za svako x za koje vrijedi $q(x) = 0$ slijedi da je takođe $p(x) = 0$.

Primjer 3.2.5. *Neka su p i q dvije raspodjele vjerovatnoća na* $\{0, 1\}$, *pri čemu je* $p(0) = 1 - r$, $p(1) = r$, $q(0) = 1 - s$, $q(1) = s$. *Tada je*

$$D_{KL}(p||q) = (1 - r) \log_2 \frac{1 - r}{1 - s} + r \log_2 \frac{r}{s},$$
$$D_{KL}(q||p) = (1 - s) \log_2 \frac{1 - s}{1 - r} + s \log_2 \frac{s}{r}.$$

Specijalno, za $r = s$ dobija se $D_{KL}(p||q) = 0 = D_{KL}(q||p)$, dok se za $r = \frac{1}{2}, s = \frac{1}{4}$ dobija $D_{KL}(p||q) = 0,2075$ i $D_{KL}(q||p) = 0,1887$. Dakle, relativna entropija nije simetrična funkcija.

Intuitivno, relativna entropija $D_{KL}(p||q)$ predstavlja očekivanu vrijednost logaritamske razlike vjerovatnosnih mjera p i q, pri čemu se očekivanje uzima u odnosu na vjerovatnoću p. U narednom je pokazano da relativna entropija posjeduje neka svojstva udaljenosti, mada nije "prava" metrika, jer ne zadovoljava uslov simetričnosti, a ne vrijedi i nejednakost trougla.

Realna funkcija f je *konveksna* na intervalu (a, b), ako za sve $x, y \in (a, b)$ i proizvoljno $\lambda \in [0, 1]$ vrijedi nejednakost

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y).$$

Ukoliko je još dodatno ispunjeno da prethodna nejednakost postaje jednakost isključivo za $\lambda = 0$ ili $\lambda = 1$, tada se funkcija f naziva *striktno konveksnom*. Realna funkcija f je *konkavna*, ako je $-f$ konveksna funkcija.

Lema 3.2.6. *[26] (**Jensenova nejednakost za matematičko očekivanje**) Ako je f konveksna funkcija i X diskretna slučajna veličina, tada vrijedi*

$$E(f(X)) \geqslant f(E(X)).$$

Štaviše, ako je f striktno konveksna funkcija, tada $E(f(X)) = f(E(X))$ implicira da je $X = E(X)$ sa vjerovatnoćom 1, tj. da je X skoro sigurno konstantna slučajna veličina.

Dokaz. Dokaz će biti izveden indukcijom po broju vrijednosti od X u kojima je raspodjela vjerovatnoća skoncentrisana, tj. broju vrijednosti od X sa pozitivnom vjerovatnoćom. Ako postoji samo jedna ovakva vrijednost, tada je X skoro sigurno konstantna slučajna veličina i trivijalno vrijedi $E(f(X)) = f(E(X))$. Takođe, nejednakost vrijedi i u slučaju dvije ovakve vrijednosti. Zaista, za $X : \begin{pmatrix} x_1 & x_2 \\ 1-p & p \end{pmatrix}$, gdje je $p \in (0,1)$, dobija se $E(f(X)) = (1-p)f(x_1) + pf(x_2)$ i $f(E(X)) = f((1-p)x_1 + px_2)$, pa tražena nejednakost slijedi iz pretpostavljene konveksnosti funkcije f. Neka posmatrana nejednakost vrijedi za svaku slučajnu veličinu koja ima $k-1$ vrijednosti sa pozitivnim vjerovatnoćama i neka je X slučajna veličina koja ima k vrijednosti $x_1, x_2, \ldots, x_k$ sa pozitivnim vjerovatnoćama $p_1, p_2, \ldots, p_k$. Tada je $E(f(X)) = \sum_{i=1}^{k} p_i f(x_i) = p_1 f(x_1) + \sum_{i=2}^{k} p_i f(x_i)$. Za slučajnu veličinu Y koja uzima vrijednosti $x_2, \ldots, x_k$ sa vjerovatnoćama $p'_2, \ldots, p'_k$, pri čemu je $p'_i := \frac{p_i}{1-p_1}$, vrijedi indukciona pretpostavka, odakle

slijedi

$$\begin{aligned} E(f(X)) &= \sum_{i=1}^{k} p_i f(x_i) = p_1 f(x_1) + \sum_{i=2}^{k} p_i f(x_i) \\ &= p_1 f(x_1) + (1-p_1) \sum_{i=2}^{k} p_i' f(x_i) \geqslant p_1 f(x_1) + (1-p_1) f\left(\sum_{i=2}^{k} p_i' x_i \right) \\ &\geqslant f\left(p_1 x_1 + (1-p_1) \sum_{i=2}^{k} p_i' x_i \right) = f\left(\sum_{i=1}^{k} p_i x_i \right) = f(E(X)). \end{aligned}$$

Ukoliko je f striktno konveksna funkcija, tada se iz prethodnog razmatranja uočava da uslov $E(f(X)) = f(E(X))$ nužno implicira da X ne može uzimati više od jedne vrijednosti sa pozitivnom vjerovatnoćom, što znači da je X skoro sigurno konstantna slučajna veličina. □

Lema 3.2.7. *[26] Neka su p i q dvije vjerovatnosne mjere definisane na istom prostoru vjerovatnoća. Tada vrijedi*

*(a) $D_{KL}(p||q) \geqslant 0$ (**Gibsova nejednakost**),*

(b) $D_{KL}(p||q) = 0$ ako i samo ako su vjerovatnosne mjere p i q jednake, sem eventualno na skupu čija je p vjerovatnoća jednaka nuli.

Dokaz.

Neka je $A := \{x : p(x) > 0\}$ nosač vjerovatnosne mjere p.

(a) Vrijedi

$$\begin{aligned} -D_{KL}(p||q) &= -\sum_{x \in A} p(x) \cdot \log_2 \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \cdot \log_2 \frac{q(x)}{p(x)} \\ &\leqslant \log_2 \left(\sum_{x \in A} p(x) \cdot \frac{q(x)}{p(x)} \right) = \log_2 \sum_{x \in A} q(x) \leqslant \log_2 \sum_{x} q(x) \\ &= \log_2 1 = 0, \end{aligned}$$

pri čemu prva nejednakost slijedi iz činjenice da je $\log_2$ konkavna funkcija i Jensenove nejednakosti koja je primjenjena na slučajnu veličinu $\dfrac{q(x)}{p(x)}$.

(b) Dovoljno je dokazati da je $D_{KL}(p||q) = 0$ ako i samo ako za svako $x \in A$ vrijedi $p(x) = q(x)$. Iz $p_{|A} = q_{|A}$ i definicije relativne entropije trivijalno slijedi $D_{KL}(p||q) = 0$. Obrnuto, neka je $D_{KL}(p||q) = 0$. To znači da sve nejednakosti iz dokaza dijela (a) postaju jednakosti. Specijalno, u Jensenovoj nejednakosti vrijedi jednakost, što je, na osnovu prethodne leme, ispunjeno ukoliko postoji konstanta c takva da za svako $x \in A$ vrijedi $q(x) = c \cdot p(x)$. Dalje je $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$, pa iz posljednje nejednakosti u dokazu dijela

(a) slijedi $c = \sum_{x \in A} q(x) = \sum_{x} q(x) = 1$. Time je dokazano da za svako $x \in A$ vrijedi $q(x) = p(x)$. □

Posljedica 3.2.8. *[26] Za svaku diskretnu slučajnu veličinu X vrijedi $H(X) \leqslant \log_2 |S_x|$, pri čemu se jednakost dostiže ako i samo ako X ima uniformnu raspodjelu na skupu S_X.*

Dokaz. Neka je p_X raspodjela vjerovatnoća slučajne veličine X i u_X ravnomjerna raspodjela vjerovatnoća na skupu S_X. Tada za svako x vrijedi $u_X(x) = \frac{1}{|S_X|}$, pa, na osnovu nenegativnosti relativne entropije, dobija se

$$0 \leqslant D_{KL}(p_X||u_X) = \sum_{x} p_X(x) \cdot \log_2 \frac{p_X(x)}{u_X(x)} = \log_2 |S_X| - H(X),$$

odakle slijedi $H(X) \leqslant \log_2 |S_x|$. Očigledno da ova nejednakost postaje jednakost ukoliko X ima uniformnu raspodjelu. Obrnuto, ako je $H(X) = \log_2 |S_x|$, tada je $D_{KL}(p_X||u_X) = 0$, što je, na osnovu prethodne leme, moguće jedino ukoliko je $p_X(x) = u_X(x)$, za svako $x \in S_X$, tj. ukoliko X ima uniformnu raspodjelu na skupu S_X. □

Lema 3.2.9. *[26] Relativna entropija je konveksna funkcija. Preciznije, ako su (p_1, q_1) i (p_2, q_2) parovi vjerovatnosnih mjera definisanih na istom mjerljivom prostoru, tada vrijedi*

$$D_{KL}\left(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2\right) \leqslant \lambda D_{KL}(p_1||q_1) + (1-\lambda) D_{KL}(p_2||q_2),$$

za svako $\lambda \in [0, 1]$.

Dokaz. Najprije će biti potvrđena sljedeća nejednakost, u literaturi poznata i kao *log-sum nejednakost*: Za sve pozitivne realne brojeve $a_1, \ldots, a_n, b_1, \ldots, b_n$ vrijedi

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geqslant \left(\sum_{i=1}^{n} a_i\right) \cdot \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}.$$

Zaista, funkcija g data sa $g(t) = t \cdot \log_2 t$ je striktno konveksna, jer za svako $t > 0$ vrijedi $g''(t) = \frac{\log_2 e}{t} > 0$. Na osnovu Jensenove nejednakosti, $\sum \alpha_i g(t_i) \geqslant g\left(\sum \alpha_i t_i\right)$, za sve $\alpha_i \geqslant 0$ za koje vrijedi $\sum \alpha_i = 1$. Stavljajući $\alpha_i := \frac{b_i}{\sum_{j=1}^{n} b_j}$ i $t_i := \frac{a_i}{b_i}$, dobija se

$$\sum_{i=1}^{n} \frac{a_i}{\sum b_j} \log_2 \frac{a_i}{b_i} \geqslant \sum_{i=1}^{n} \frac{a_i}{\sum b_j} \log_2 \sum_{i=1}^{n} \frac{a_i}{\sum b_j},$$

a ovo je upravo log-sum nejednakost.

Koristeći log-sum nejednakost, zaključuje se da za svako $\lambda \in [0,1]$ i svako x vrijedi

$$\begin{aligned}&(\lambda p_1(x) + (1-\lambda)p_2(x)) \cdot \log_2 \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\ &\leqslant \lambda p_1(x) \cdot \log_2 \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \cdot \log_2 \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}.\end{aligned}$$

Sumirajući prethodne nejednakosti po svim x dobija se tražena nejednakost. □

Posljedica 3.2.10. *[26] Entropija je konkavna funkcija.*

Dokaz. Ranije je dokazano da za svaku vjerovatnosnu mjeru p vrijedi $H(p) = \log_2 |S_X| - D_{KL}(p||u)$, gdje je u uniformna raspodjela na skupu S_X. Iz ovog prikaza i prethodno ustanovljene konveksnosti relativne entropije D_{KL} slijedi konkavnost entropije H. □

Neka su A i B neprazni, konačni skupovi sastavljeni od stringova formiranih nad istim alfabetom $\mathbb{N}_n$. Za ove skupove se smatra da su izabrani kao "trening" podaci (trajektorije) iz raspodjela vjerovatnoća $p^{(A)}$ i $p^{(B)}$ respektivno. Mjere sličnosti familija stringova A i B mogu da se uvedu na temelju poređenja vjerovatnosnih mjera $p^{(A)}$ i $p^{(B)}$. U ovoj tezi su od mjera sličnosti ovakvog tipa razmatrane one koje koriste Kulbak-Lajblerovu mjeru divergencije. Konkretno, za familije stringova A i B, od interesa će biti "prosječna" relativna entropija $\frac{D_{KL}(p^{(A)}||p^{(B)}) + D_{KL}(p^{(B)}||p^{(A)})}{2}$. Razlog posmatranja ove vrijednosti, a ne $D_{KL}(p^{(A)}||p^{(B)})$ ili $D_{KL}(p^{(B)}||p^{(A)})$ pojedinačno leži u prethodno uočenom odsustvu simetričnosti relativne entropije. Podsjećanja radi, relativna entropija $D_{KL}(p^{(A)}||p^{(B)})$ se dobija tako što se sumiraju izrazi oblika $\sum_x p^{(A)}(x) \cdot \log_2 \frac{p^{(A)}(x)}{p^{(B)}(x)}$, pri čemu se ova suma uzima po svim mogućim stringovima x izabranih iz raspodjele $p^{(A)}$. Nažalost, može se desiti da postoji veliki broj stringova koje je potrebno uzeti u razmatranje, što sa praktične strane relativnu entropiju čini neupotrebljivom vrijednošću. Stoga je neophodno naći zamjensku vrijednost koja će imati svojstvo da je "dovoljno blizu" relativnoj entropiji. U ovom postupku od koristi će biti Petrijeva granična teorema.

Neka su $\mathcal{M}_A$ i $\mathcal{M}_B$ lanci Markova iz kojih su izvučeni familije stringova A i B kao trening podaci. Takođe, neka su poznate vjerovatnosne mjere $P^{\mathcal{M}_A}$ i $P^{\mathcal{M}_B}$ koje opisuju sve konačno dimenzionalne raspodjele datih lanaca, pri čemu se pretpostavlja da je mjera $P^{\mathcal{M}_A}$ apsolutna neprekidna u odnosu na mjeru $P^{\mathcal{M}_B}$. Za neprazan string $x = x_1x_2 \ldots x_{len(x)}$, koji je trajektorija lanca $\mathcal{M}_A$, moguće je uvesti veličinu

$$d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}, x) := \frac{1}{len(x)} \cdot \log_2 \frac{P^{\mathcal{M}_A}(x)}{P^{\mathcal{M}_B}(x)}.$$

Koristeći ranije uvedenu veličinu $M_n(P^{\mathcal{M}}, t)$, dolazi se do prikaza

$$d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}, x) = M_{len(x)-1}(P^{\mathcal{M}_A}, t) - M_{len(x)-1}(P^{\mathcal{M}_B}, t),$$

pa, s obzirom na Petrijevu graničnu teoremu, postoji vrijednost

$$d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}) := \lim_{len(x)\to+\infty} d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}, x). \tag{3.4}$$

Ispostavlja se da je ova vrijednost jednaka $D_{KL}(P^{\mathcal{M}_A}||P^{\mathcal{M}_B})$. Za potvrdu ove jednakosti eksploatiše se ergodičnost stohastičkih procesa, što je materija van okvira ove teze te je dokaz izostavljen, sa napomenom da se ideja i detalji ovog dokaza mogu naći npr. u [11].

Analogno, za neprazan string $y = y_1 y_2 \dots y_{len(y)}$, koji je trajektorija lanca $\mathcal{M}_B$, moguće je uvesti veličinu

$$d(P^{\mathcal{M}_B}, P^{\mathcal{M}_A}, y) := \frac{1}{len(y)} \cdot \log_2 \frac{P^{\mathcal{M}_B}(y)}{P^{\mathcal{M}_A}(y)}$$

i, na osnovu nje, veličinu $d(P^{\mathcal{M}_B}, P^{\mathcal{M}_A}) := \lim_{len(y)\to+\infty} d(P^{\mathcal{M}_B}, P^{\mathcal{M}_A}, y)$, koja je jednaka relativnoj entropiji $D_{KL}(P^{\mathcal{M}_B}||P^{\mathcal{M}_A})$.

Veličine $d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B})$ i $d(P^{\mathcal{M}_B}, P^{\mathcal{M}_A})$ je moguće dobiti samo na osnovu beskonačnih trajektorija. Stoga je ove veličine potrebno procijeniti. Slijedi opis postupka procjene za $d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B})$, dok se procjena za $d(P^{\mathcal{M}_B}, P^{\mathcal{M}_A})$ dobija na sličan način.

Neka je $\left(x^{(1)}, \dots, x^{(m)}\right)$ slučajan uzorak konačnih trajektorija izabranih iz lanca Markova $\mathcal{M}_A$ u saglasnosti sa raspodjelom vjerovatnoća $P^{\mathcal{M}_A}$. Za svako $i \in \{1, 2, \dots, m\}$, moguće je izračunati vrijednost $d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}, x^{(i)})$ i na osnovu toga dobiti statistiku

$$\hat{d}(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}) := \frac{1}{m} \sum_{i=1}^{m} d\left(P^{\mathcal{M}_A}, P^{\mathcal{M}_B}, x^{(i)}\right). \tag{3.5}$$

Zbog jednakosti date u formuli (3.4), vrijednost $\hat{d}(P^{\mathcal{M}_A}, P^{\mathcal{M}_B})$ predstavlja procjenu veličine $d(P^{\mathcal{M}_A}, P^{\mathcal{M}_B})$ i preciznost ove procjene se poboljšava povećavanjem obima uzorka, tj. uvećavanjem broja m.

Konačno, na osnovu svega prethodnog, moguće je uvesti vjerovatnosnu mjeru sličnosti nepraznih, konačnih familija stringova A i B na sljedeći način:

$$d^{new}(A, B) := \frac{\hat{d}(P^{\mathcal{M}_A}, P^{\mathcal{M}_B})) + \hat{d}(P^{\mathcal{M}_B}, P^{\mathcal{M}_A})}{2}. \tag{3.6}$$

Prateći kompletnu strategiju dobijanja mjere $d^{new}(A, B)$ uočava se da je za njeno određivanje ključno poznavanje vjerovatnoća $P^{\mathcal{M}_A}$ i $P^{\mathcal{M}_B}$. Nažalost,

vjerovatnoće $P^{\mathcal{M}_A}$ i $P^{\mathcal{M}_B}$ u praksi nisu poznate, te je potrebno izvršiti njihovo statističko modelovanje na osnovu trening podataka datih u skupovima A i B. Rezultat ovog modelovanja biće dobijanje "zamjenskih" vjerovatnosnih mjera $\hat{P}^{\mathcal{M}_A}$ i $\hat{P}^{\mathcal{M}_B}$. U sljedećoj sekciji ove glave predložen je jedan frekvencionistički model, a u posljednjoj sekciji ove glave razmatran je jedan model Bejzovskog zaključivanja.

3.3 Frekvencionistički model - vjerovatnosno sufiksno drvo

Za nalaženje vjerovatnoće da konkretan string $x = x_1x_2 \ldots x_{len(x)}$ predstavlja konačnu trajektoriju lanca Markova $\{X_l : l \geqslant 1\}$, čija su stanja određena alfabetom $\mathbb{N}_n, n \geqslant 2$, koristi se formula (3.3). Podsjećanja radi, u pitanju je formula množenja vjerovatnoća data sa:

$$P(x) = P(x_1)\cdot P(x_2|x_1)\cdot \ldots \cdot P(x_{len(x)}|x_1x_2 \ldots x_{len(x)-1}) = \prod_{i=1}^{len(x)} P(x_i|x_1 \ldots x_{i-1}),$$

sa konvencijom $x_0 := \varepsilon$, gdje je ε prazan string, i $P(x_1|\varepsilon) := P(x_1)$. Iz prethodne formule se uočava da je za ocjenu vjerovatnoće inicijalizacije stringa x dovoljno pružiti ocjene svih uslovnih vjerovatnoća sa desne strane jednakosti, tj. dati sve ocjene oblika $\hat{P}(x_i|x_1 \ldots x_{i-1})$. U tom slučaju može se staviti $\hat{P}(x) := \prod_{i=1}^{len(x)} \hat{P}(x_i|x_1 \ldots x_{i-1})$.

Neka je A konačna familija stringova koji su konkretne trajektorije datog lanca Markova. Ideja je da se na osnovu ovih stringova, kao trening podataka, izvede ocjena odgovarajućih uslovnih vjerovatnoća, a da se potom ove uslovne vjerovatnoće iskoriste za ocjenjivanje vjerovatnoća realizacije svih stringova, uključujući i one stringove koji nisu obuhvaćeni trening podacima, tj. nisu elementi skupa A.

Svaki sufiks stringa $x_1 \ldots x_{i-1}$ koji predstavlja realizaciju određenog broja neposrednih stanja prije posmatranog stanja x_i se naziva *kontekstom simbola* x_i. U ovoj sekciji, ocjene uslovnih vjerovatnoća biće izvedene uz pomoć frekvencionističke statistike, u kojem se prati relativna frekvencija učestalosti svih konteksta u okviru trening podataka. Slijedi precizniji opis jednog ovakvog modela.

Neka je A skup trening podataka čiji su elementi stringovi sastavljeni od simbola alfabeta $\mathbb{N}_n$ i A^* familija svih podstringova stringova iz skupa A, tj. familija svih konteksta simbola alfabeta koji se pojavljuju u stringovima iz familije A. Za efikasno uređivanje svih sufiksa datog konteksta iz A^* se koristi

tzv. *sufiksno drvo*, dok se za skladištenje njihovih uslovnih vjerovatnoća koristi varijanta sufiksnog drveta koja se naziva *vjerovatnosno sufiksno drvo* i označava sa PST(A) (PST je skraćenica od Probabilistic Suffix Tree). Vjerovatnosno sufiksno drvo PST(A) pridruženo skupu trening podataka A se izgrađuje sukcesivnim dodavanjem sufiksa sve veće dužine, a koji su elementi skupa A^*. Čvorovi ovog drveta označeni su stringovima iz skupa A^*. Svaki čvor generiše najviše n nasljednih čvorova, pri čemu svaka ovakva veza (grana) sadrži informaciju o odgovarajućoj empirijskoj uslovnoj vjerovatnoći. Npr., ako je čvor drveta označen kontekstom $c_1c_2\ldots c_k \in A^*$, tada je čvor označen kontekstom $c_1c_2\ldots c_kc_{k+1} \in A^*$ njegov *nasljednik* (*potomak*) i grani između ova dva čvora se pridružuje empirijska uslovna vjerovatnoća $\hat{P}\left(c_{k+1}|c_1c_2\ldots c_k\right)$. Ova vjerovatnoća predstavlja ocjenu maksimalne vjerodostojnosti nepoznate uslovne vjerovatnoće $P\left(c_{k+1}|c_1c_2\ldots c_k\right)$ i ona se dobija po formuli:

$$\hat{P}\left(c_{k+1}|c_1c_2\ldots c_k\right) = \frac{f\left(c_1c_2\ldots c_kc_{k+1}\right)}{\sum f\left(c_1c_2\ldots c_k\sigma\right)} \tag{3.7}$$

gdje je $f(c)$ frekvencija pojavljivanja stringa $c \in A^*$, a suma u nazivniku se uzima po svim mogućim simbolima $\sigma \in \mathbb{N}_n$ za koje vrijedi $c_1c_2\ldots c_k\sigma \in A^*$. Jednostavno se provjerava da empirijske uslovne vjerovatnoće date formulom (3.7), a koje su uskladištene u granama koje polaze iz istog čvora (tzv. *čvora roditelja*) čine raspodjelu vjerovatnoća.

Dakle, vjerovatnosno sufiksno drvo PST(A) se formira kroz sljedeći niz koraka:

1. U korijen drveta se postavlja prazan string ε.

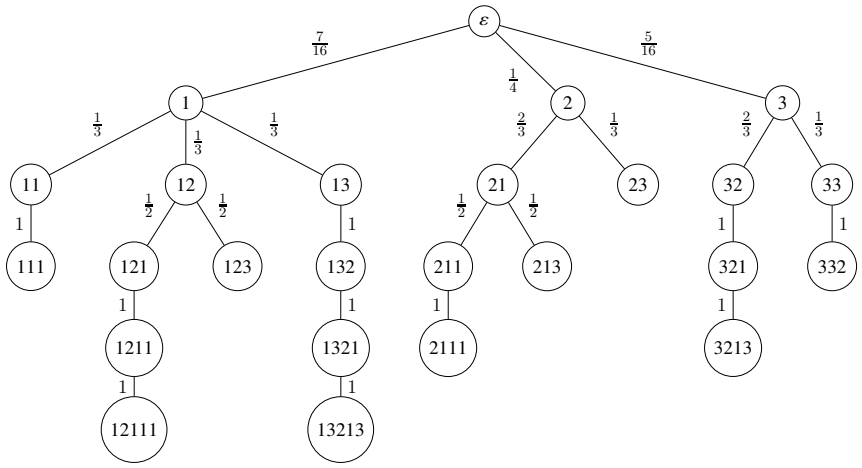
2. Potomci korijena su čvorovi označeni jednočlanim stringovima iz A^*; relativne frekvencije pojavljivanja ovih stringova u okviru skupa A predstavljaju ocjene vjerovatnoća realizacije ovih stringova.

3. Iz svakog čvora iz prethodnog koraka (čvora roditelja) polaze čvorovi potomci označeni stringovima iz A^*, tako da je za sve njih čvor roditelj njihov prefiks; svakoj vezi između roditelja i potomka dodjeljuje se odgovarajuća empirijska uslovna vjerovatnoća koja se računa na osnovu formule (3.7).

4. Za svaku granu drveta, prethodni korak se ponavlja sve dok u A^* postoji kontekst kojim se "nastavlja" kontekst iz čvora roditelja. Ukoliko se kontekst iz nekog čvora ne može nastaviti, tada je taj čvor list posmatranog drveta.

Primjer 3.3.1. *Neka je* $A = \{12111, 13213, 332, 123\}$*, pri čemu se simboli stringova iz ovog skupa uzimaju iz alfabeta* $\mathbb{N}_3$*. Uzimajući u obzir dužine svih*

konteksta iz skupa A^*, skup A^* se može particionisati na skupove $A_1^*, A_2^*, A_3^*, A_4^*$ i A_5^*, gdje je $A_m^* \subseteq A^*$ skup svih konteksta dužine m (tzv. m-grama). Na taj način, dobija se: $A_1^* = \{1, 2, 3\}$, $A_2^* = \{11, 12, 13, 21, 23, 32, 33\}$, $A_3^* = \{111, 121, 123, 132, 211, 213, 321, 332\}$, $A_4^* = \{1211, 1321, 2111, 3213\}$ i $A_5^* = \{12111, 13213\}$. U korijenu vjerovatnosnog sufiksnog drveta, kao nultom nivou, nalazi se prazan string ε, dok se čvorovi u okviru skupa A_i^* nalaze na njegovom $i-$tom nivou. U slučaju da na dva susjedna nivoa postoje čvorovi tako da je čvor na višem nivou označen prefiksom stringa kojim je označen čvor na nižem nivou, tada se ovi čvorovi povezuju granom koja sadrži odgovarajću empirijsku uslovnu vjerovatnoću. Npr. u skupu A se pojavljuju tri podstringa čiji je kontekst string 3, pri čemu se 32 pojavljuje dva puta, a 33 jednom. Zbog toga, vrijedi $\hat{P}_{PST}(2|3) = \frac{2}{3}$ i $\hat{P}_{PST}(3|3) = \frac{1}{3}$. Evidentirajući sve čvorove i grane dobija se vjerovatnosno sufiksno drvo $PST(A)$ predstavljeno na Slici 3.1.



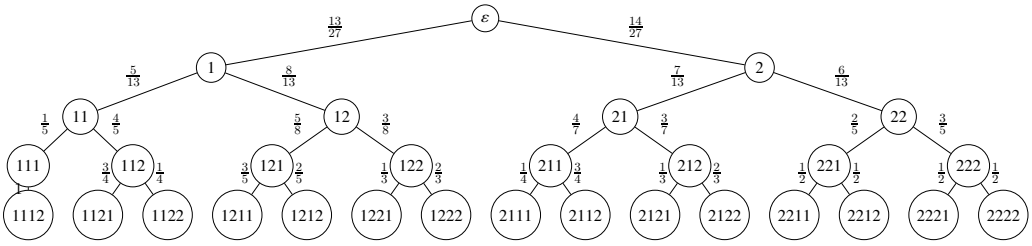
Slika 3.1

Kao što se može primijetiti iz prethodnog primjera, čak i za relativno mali obim trening podataka, vjerovatnosno sufiksno drvo može sadržavati veliki broj čvorova raspoređenih na dosta različitih nivoa. Broj čvorova vjerovatnosnog sufiksnog drveta se eksponencijalno povećava sa povećavanjem broja nivoa, a isto vrijedi i za broj empirijskih uslovnih vjerovatnoća koje se pridružuju svakoj vezi između čvorova. Broj nivoa se može smanjiti korišćenjem pretpostavke da su svi stringovi iz skupa trening podataka trajektorije lanca Markova reda r. U tom slučaju, zbog svojstva Markova, vjerovatnosno sufiksno drvo može imati najviše r različitih nivoa, jer za string $x_1x_2 \ldots x_{len(x)}$ dužine $len(x) > r$ i svako $i \in (r, len(x)]$ vrijedi $\hat{P}(x_i|x_1x_2 \ldots x_{i-1}) = \hat{P}(x_i|x_{i-r} \ldots x_{i-1})$. Efektivno, svaki kontekst dužine $i-1$ u odnosu na koji se računa empirijska uslovna vjerovatnoća, a koji je dužine veće od r, "gubi" prefiks sastavljen od prvih $i - r - 1$ simbola.

Primjer 3.3.2. *Na binarnom alfabetu* $\mathbb{N}_2$ *dat je skup*

$$A = \{12112112112222212122111121222\}$$

sastavljen od samo jednog stringa. Koristeći informaciju da je dati string izabran kao trajektorija lanca Markova reda 4*, dobija se vjerovatnosno sufiksno drvo PST(A) predstavljeno na Slici 3.2.*



Slika 3.2

Nažalost, kao što prethodni primjer ilustruje, ograničavanje reda lanca Markova samo donekle pojednostavljuje strukturu vjerovatnosnog sufiksnog drveta. Stoga je potrebno nametnuti dodatna ograničenja. S obzirom da će uz pomoć empirijskih uslovnih vjerovatnoća biti definisana vjerovatnosna mjera, u interesu je među njima prepoznati one uslovne vjerovatnoće koje su esencijalne pri ovom definisanju, a koje se mogu odbaciti. "Kresanjem" grana koje sadrže "nepotrebne" čvorove i grane vjerovatnosno sufiksno drvo se dodatno pojednostavljuje.

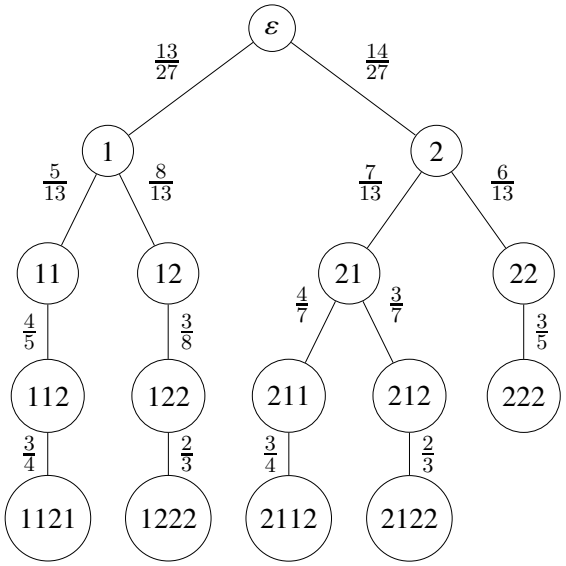
Prije svega, ako biranje skupa stringova A smatramo kao uzorkovanje, može se postaviti ograničenje da se u $PST(A)$ posmatraju samo oni čvorovi koji su označeni kontekstima iz A^* čija je frekvencija pojavljivanja veća ili jednaka od unaprijed date vrijednosti $f_{\min}$. Na taj način se eliminišu oni stringovi iz A^* čije se pojavljivanje može pripisati fluktuaciji trening podataka iz skupa A.

Drugi oblik redukovanja vjerovatnosnog sufiksnog drveta podrazumijeva uklanjanje čvorova i grana čija "moć predviđanja" nije značajno bolja od one koju imaju čvorovi roditelji. Preciznije, iz drveta $PST(A)$ se ne izbacuje čvor označen stringom $c = c_1c_2 \ldots c_k \in A^*$, ako postoji simbol $\sigma \in \Sigma$ za koji vrijedi $\dfrac{\hat{P}(\sigma|c_1c_2 \ldots c_k)}{\hat{P}(\sigma|c_2 \ldots c_k)} \geqslant C$ ili $\dfrac{\hat{P}(\sigma|c_1c_2 \ldots c_k)}{\hat{P}(\sigma|c_2 \ldots c_k)} \leqslant \dfrac{1}{C}$, gdje je $C \geqslant 1$ unaprijed zadana vrijednost, pri čemu se najčešće uzima da je $C = 1,05$ ili $C = 1,2$. Dakle, ako u vjerovatnosnom sufiksnom drvetu $PST(A)$ postoji čvor označen stringom $c = c_1c_2 \ldots c_k \in A^*$, tada se ovaj čvor ne izbacuje iz drveta $PST(A)$ ukoliko je ispunjen uslov

$$\sum_{\sigma} I\left(\frac{\hat{P}(\sigma|c_1c_2 \ldots c_k)}{\hat{P}(\sigma|c_2 \ldots c_k)} \geqslant C \text{ ili } \frac{\hat{P}(\sigma|c_1c_2 \ldots c_k)}{\hat{P}(\sigma|c_2 \ldots c_k)} \leqslant \frac{1}{C}\right) \geqslant 1, \qquad (3.8)$$

gdje se prethodna suma uzima po svim simbolima $\sigma \in \mathbb{N}_n$, za koje vrijedi $c_1c_2 \ldots c_k\sigma \in A^*$, dok je $I(\cdot)$ indikator događaja koji je naveden u zagradi.

Primjer 3.3.3. *Primjenom prethodnih restrikcija, uzimajući vrijednosti* $f_{\min} = 2$ *i* $C = 1, 2$*, vjerovatnosno sufiksno drvo* $PST(A)$ *iz prethodnog primjera se redukuje na drvo predstavljeno na Slici 3.3.*



Slika 3.3

Npr. izbačen je čvor označen sa 111*, jer se ovaj kontekst pojavljuje samo jednom u stringu iz skupa A. Takođe, izbačen je čvor označen sa* 121*, jer vrijedi* $\frac{\hat{P}(1|121)}{\hat{P}(1|21)} = \frac{21}{20} = 1,05 \in \left(\frac{5}{6}, \frac{6}{5}\right)$ *i* $\frac{\hat{P}(2|121)}{\hat{P}(2|21)} = \frac{14}{15} = 0,93 \in \left(\frac{5}{6}, \frac{6}{5}\right)$*, pa ovaj čvor ne ispunjava uslov (3.8). To znači da se empirijske uslovne vjerovatnoće* $\hat{P}(1|121)$ *i* $\hat{P}(2|121)$ *mogu aproksimirati redom empirijskim uslovnim vjerovatnoćama* $\hat{P}(1|21)$ *i* $\hat{P}(2|21)$*, tj. ulogu konteksta* 121 *uzima njegov najduži pravi sufiks* 21*. Ovakvom "zamjenom", čvorovi bliži korijenu drveta dobijaju veći značaj.*

Neka je A, konačan, neprazan skup stringova. Stringovi iz skupa A se mogu shvatiti kao konačne trajektorije lanca Markova $\mathcal{M}_A$ čije su raspodjele vjerovatnoća određene vjerovatnosnom mjerom $P^{\mathcal{M}_A}$. Vjerovatnosnom mjerom $P^{\mathcal{M}_A}$ se, uz pomoć formule (3.3), izražava mogućnost realizacije bilo koje trajektorije ovog lanca Markova. Jedna ocjena ove vjerovatnosne mjere može da se dobije uz pomoć vjerovatnosnog sufiksnog drveta $\mathrm{PST}(A)$ i ona se zasniva na empirijskim uslovnim vjerovatnoćama oblika $\hat{P}(\sigma|x_1x_2 \ldots x_i)$, koje su dodjeljene granama ovog drveta. Preciznije, za proizvoljan string $x = x_1x_2 \ldots x_{len(x)}$, definiše se

$$\widehat{P}^{\mathcal{M}_A}_{PST}(x) := \hat{P}(x_1) \cdot \hat{P}(x_2|x_1) \cdot \ldots \cdot \hat{P}\left(x_{len(x)}|x_1x_2 \ldots x_{len(x)-1}\right) \tag{3.9}$$

Ukoliko se za uslovnu vjerovatnoću $\hat{P}(x_i|x_1x_2\ldots x_{i-1})$ iz prethodne formule kontekst $x_1x_2\ldots x_{i-1}$ ne pojavljuje kao čvor vjerovatnosnog sufiksnog drveta $PST(A)$ sa čvorom potomkom $x_1x_2\ldots x_{i-1}x_i$, tada se umjesto konteksta $x_1\ldots x_{i-1}$ bira njegov najduži sufiks $x_j\ldots x_{i-1}$ takav da iz čvora drveta označenog kontekstom $x_j\ldots x_{i-1}$ polazi grana ka čvoru drveta označenim kontekstom $x_j\ldots x_{i-1}x_i$.

Primjer 3.3.4. *Neka je $\Sigma = \mathbb{N}_2$, A skup trening podataka i $PST(A)$ (redukovano) vjerovatnosno sufiksno drvo iz prethodnog primjera. Za string $x = 21122$, primjenom formule (3.9) se dobija*

$$\widehat{P}^{\mathcal{M}_A}_{PST}(x) = \hat{P}(2)\cdot\hat{P}(1|2)\cdot\hat{P}(1|21)\cdot\hat{P}(2|211)\cdot\hat{P}(2|2112)\,.$$

Čitanjem empirijskih uslovnih vjerovatnoća iz vjerovatnosnog sufiksnog drveta $PST(A)$ dobija se $\hat{P}(2) = \frac{14}{27}$, $\hat{P}(1|2) = \frac{7}{13}$, $\hat{P}(1|21) = \frac{4}{7}$ i $\hat{P}(2|211) = \frac{3}{4}$. Kontekst 2112 se pojavljuje u drvetu, ali ne kao čvor roditelj. Njegov najduži sufiks je 112, ali u drvetu ne postoji čvor označen sa 1122. Najduži sufiks od 112 je 12 i vrijedi $\hat{P}(2|12) = \frac{3}{8}$. Zbog toga je $\hat{P}(2|2112) \approx \hat{P}(2|12) = \frac{3}{8}$, te je $\widehat{P}^{\mathcal{M}_A}_{PST}(x) = \frac{7}{156} \approx 0,04487$.

Sa povećavanjem broja n, broj mogućih stringova maksimalne dužine l koji se mogu formirati od simbola alfabeta $\mathbb{N}_n$ ubrzano raste. Uslijed toga može nastati scenario u kojem slučajno izabrani trening podaci ne sadrže nikakvu informaciju o datom stringu x. To znači da podstringovi stringa x nisu konteksti iz skupa A^*, što implicira da je bar jedna od empirijskih uslovnih vjerovatnoća jednaka nuli, te posljedično vrijedi $\hat{P}^{\mathcal{M}_A}_{PST}(x) = 0$. Na taj način se za string x koji se rijetko pojavljuje, ali je njegova realizacija itekako moguća, može dobiti ocjena $P^{\mathcal{M}_A}(x) = 0$. Očigledno da ovakve ocjene ne oslikavaju realnu situaciju, te ih je potrebno poboljšati.

Prethodno opisani problem predviđanja stringova koji nisu zastupljeni u okviru trening podataka nastupa jer se težine empirijskih uslovnih vjerovatnoća raspoređuju isključivo u skladu sa "vidljivim" podacima iz skupa A. Opširna empirijska studija različitih strategija za rješavanje ovog problema može se naći u [23]. U ovoj tezi biće korišćena tehnika *izglađivanja* (na engleskom smoothing), kojom se dio vjerovatnosnih težina zastupljenih podataka preraspoređuje u korist dobijanja pozitivnih vjerovatnosnih težina nezastupljenih (ali ostvarljivih) podataka. Ova preraspodjela se izvodi u skladu sa datom pomoćnom distribucijom vjerovatnoća. Preciznije, za simbol $\sigma \in \mathbb{N}_n$ i kontekst c, "popravljene" empirijske uslovne vjerovatnoće se definišu na sljedeći način:

$$\hat{\hat{P}}(\sigma|c) := \begin{cases} \dfrac{f(c\sigma) - d_C}{\sum\limits_{\tau\in\mathbb{N}_n} f(c\tau)}, & \text{ako je } f(c\sigma) > 0; \\ \alpha(c)\cdot\beta(c,\sigma), & \text{inače,} \end{cases} \tag{3.10}$$

pri čemu je $f(c\sigma)$ uočena frekvencija pojavljivanja simbola σ nakon konteksta c; d_C je *parametar sniženja* i predstavlja vrijednost vjerovatnosne težine koja se preraspoređuje u zavisnosti od $f(c\sigma)$; $\alpha(c)$ je *normalizacioni faktor*; dok je $\beta(c,\sigma)$ *povratna distribucija*, do koje se generalno dolazi na osnovu sufiksa konteksta c. Povratna distribucija se na sličan način dalje može izgladiti. Tako se dobija rekurzivan proces koji se, u najgorem slučaju, svodi na poznavanje (bezuslovne) vjerovatnoće $\hat{P}(\sigma)$. Normalizacioni faktor $\alpha(c)$ obezbjeđuje da $\hat{\hat{P}}$ predstavlja raspodjelu vjerovatnoća, tj. da je ispunjen uslov $\sum_{\sigma\in\mathbb{N}_n} \hat{\hat{P}}(\sigma|c) = 1$.

U popravljenoj empirijskoj uslovnoj vjerovatnoći zadatoj u formuli (3.10) povratna distribucija se koristi samo u slučaju kada je odgovarajuća nepopravljena empirijska uslovna vjerovatnoća jednaka nuli. U radu [56], Knezer i Naj su pokazali da se povratna distribucija može koristiti i bez ovog uslova i da to generalno dovodi do efikasnijeg modela. Konkretno, u pomenutom radu predložen je model KN, u kojem su popravljene empirijske uslovne vjerovatnoće date sa

$$\hat{\hat{P}}_{KN}(\sigma|c) := \begin{cases} \dfrac{f(c\sigma) - d_C}{\sum_{\tau\in\mathbb{N}_n} f(c\tau)} + \alpha(c)\cdot\beta(c,\sigma), & \text{ako je } f(c\sigma) > 0; \\ \alpha(c)\cdot\beta(c,\sigma), & \text{inače,} \end{cases} \tag{3.11}$$

gdje je $d_C \leqslant f(c\sigma)$, ukoliko je $f(c\sigma) > 0$; normalizacioni faktor je dat sa

$$\alpha(c)_{KN} := \frac{d_C \cdot |\{\tau \in \mathbb{N}_n : f(c\tau) > 0\}|}{\sum_{\tau\in\mathbb{N}_n} f(c\tau)};$$

dok je povratna distribucija ocjenjena empirijskim vjerovatnoćama

$$\beta(c,\sigma)_{KN} := \frac{|F(c,\sigma)|}{\sum_{\tau\in\mathbb{N}_n} |F(c,\tau)|},$$

gdje je $F(c,\sigma) := \{\sigma' \in \mathbb{N}_n : f(\sigma'\mathrm{suff}(c)\sigma) > 0\}$ i $\mathrm{suff}(c)$ je najduži pravi sufiks konteksta c. Posljednja ocjena nije bazirana na učestalosti konteksta $c\sigma$, već na broju različitih pojavljivanja konteksta u kojima se simbol σ ostvaruje nakon realizacije konteksta $\mathrm{suff}(c)$. Važno je istaći da empirijska uslovna vjerovatnoća $\beta(c,\sigma)_{KN}$ može biti jednaka nuli, pa se za nju može iskoristiti izglađivanje dato u formuli (3.11), pri čemu se umjesto konteksta c uzima kontekst $\mathrm{suff}(c)$. Ovaj postupak se može nastaviti sve dok se u nekom koraku ne dobije pozitivna empirijska uslovna vjerovatnoća ili dok se kontekst c ne svede na simbol koji se nalazi na njegovoj posljednjoj poziciji.

Primjedba 3.3.5. Opisani KN model predstavlja jedan oblik *nelinearne interpolacije*. Naime, stavljajući $f(c\cdot) := \sum_{\tau\in\mathbb{N}_n} f(c\tau)$ i $g(c\cdot) := |\{\tau \in \mathbb{N}_n : f(c\tau) > 0\}|$, za vjerovatnoće date u formuli (3.11) vrijedi

$$\hat{\hat{P}}_{KN}(\sigma|c) = \frac{\max\{f(c\sigma) - d_C, 0\}}{f(c\cdot)} + \frac{d_C \cdot g(c\cdot)}{f(c\cdot)}\beta(c,\sigma)_{KN}.$$

Koristeći formulu (3.9) na način da se empirijske uslovne vjerovatnoće dobijene iz PST modela zamijene "popravljenim" empirijskim vjerovatnoćama iz KN modela, dolazi se do vjerovatnosne mjere $\widehat{P}_{KN}^{\mathcal{M}_A}$, koja je ocjena vjerovatnosne mjere $P^{\mathcal{M}_A}$. Uz pomoć ove ocjene, vjerovatnosna mjera sličnosti nepraznih, konačnih familija stringova A i B uvedena u formuli (3.6) ima oblik

$$d_{KN}^{new}(A,B) = \frac{\hat{d}(\hat{P}_{KN}^{\mathcal{M}_A}, \hat{P}_{KN}^{\mathcal{M}_B})) + \hat{d}(\hat{P}_{KN}^{\mathcal{M}_B}, \hat{P}_{KN}^{\mathcal{M}_A})}{2}.$$

Kada je u pitanju efikasnost pomenutih modela, ona se mjeri uz pomoć *unakrsne entropije* i *perpleksije*. Za ocjenu $\hat{P}$ vjerovatnosne mjere P, unakrsna entropija od $\hat{P}$ u odnosu na P se definiše sa $H(P,\hat{P}) := H(P) + D_{KL}(P||\hat{P})$, gdje je H entropija, a D_{KL} relativna entropija, dok se perpleksija ocjene $\hat{P}$ uvodi sa $Perp(\hat{P}) := 2^{H(P,\hat{P})}$. U radu [23] izvršena je uporedna statistička analiza efikasnosti različitih modela za ocjenjivanje vjerovatnoće $P^{\mathcal{M}_A}$ u smislu poređenja pripadnih perpleksija. Ova analiza je pokazala da, u odnosu na drugačije modele iz literature, KN model ima signifikantno najmanju perpleksiju, te je u tom smislu od njih superiorniji. Takođe, u pomenutom radu su predložena dodatna poboljšanja KN modela, koja pospješuju njegovu efikasnost. Jedno od tih poboljšanja podrazumijeva da se, umjesto fiksiranog parametra sniženja d_C, za svaku različitu dužinu konteksta razmatra zaseban parametar sniženja. Na taj način se bolje balansira preraspodjela vjerovatnosnih težina na svakom nivou na kojem se primjenjuje formula (3.11).

3.4 Bejzovski model - Hijerarhijski Pitman-Jorov proces

Bejzovski modeli koriste drugačiji pristup u odnosu na frekvencionističke modele. Fundamentalna razlika je što se u okviru frekvencionističkih modela nepoznati parametri tretiraju kao pojedinačne vrijednosti koje treba ocijeniti, dok se kod Bejzovskih modela nepoznati parametri shvataju kao slučajne promjenljive čije raspodjele vjerovatnoća treba precizirati. U narednom su opisane ideje i tehnike Bejzovskog zaključivanja na kome se zasnivaju Bejzovski modeli.

Neka je $(\Omega, \mathcal{F}, P)$ prostor vjerovatnoća. Ako je $\{B_1, B_2, \ldots, B_m\} \subseteq \mathcal{F}$ particija skupa Ω (tzv. *potpun sistem događaja*) i $A \in \mathcal{F}$ proizvoljan događaj,

tada, za svako $i \in \{1, 2, \ldots, m\}$, vrijedi *Bejzova formula*:

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{P(A)} = \frac{P(B_i) \cdot P(A|B_i)}{\sum_{j=1}^{m} P(B_j) \cdot B(A|B_j)}.$$

Naravno, pored navedene diskretne verzije, postoji i neprekidna verzija Bejzove formule kod koje se suma u nazivniku mijenja sa odgovarajućim integralom. Bejzova formula pruža jednostavno pravilo "ažuriranja" vjerovatnoća, kada su dostupne dodatne informacije date u obliku uočenih podataka. Bejzovsko zaključivanje podrazumijeva da se predubjeđenja koja su postojala u vezi raspodjele nepoznatih parametara modifikuju u skladu sa uočenim podacima.

Neka je $\boldsymbol{\theta}$ slučajna promjenljiva čija je raspodjela nepoznata. Na osnovu intuicije ili na neki drugi način, predlaže se raspodjela vjerovatnoća $\pi(\boldsymbol{\theta})$ slučajne promjenljive $\boldsymbol{\theta}$. Ova raspodjela se naziva *priornom raspodjelom*, jer se njome izražava mišljenje o distribuciji slučajne promjenljive $\boldsymbol{\theta}$ prije nego što se dođe u dodir sa empirijskim podacima. Parametri koji determinišu priornu raspodjelu nazivaju se *hiperparametrima* i za njih se smatra da su poznati i konstantni. Dalje, bira se slučajan uzorak $\mathbf{X}$, kao slučajan vektor izabran iz raspodjele vjerovatnoća u kojoj $\boldsymbol{\theta}$ učestvuje kao parametar. Ako je θ data vrijednost parametra, vjerovatnoća da je izvučen konkretan uzorak $\mathbf{x}$ je uslovna vjerovatnoća $P(\mathbf{x}|\theta)$; vjerovatnoće ovakvog oblika nalaze se uz pomoć funkcije vjerodostojnosti i one određuju uslovnu raspodjelu $P(\mathbf{x}|\boldsymbol{\theta})$, koja se naziva *vjerodostojnošću* uočenog uzorka. Na osnovu pravila množenja vjerovatnoća, zajednička raspodjela za $\boldsymbol{\theta}$ i $\mathbf{X}$ određena je sa $P(\boldsymbol{\theta}, \mathbf{x}) = \pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta})$. Iz ove raspodjele dobija se *marginalna raspodjela* za $\mathbf{X}$ data sa $P(\mathbf{x}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. *Posteriorna raspodjela* je određena uslovnom raspodjelom $P(\boldsymbol{\theta}|\mathbf{x})$; primjenom Bejzove formule, ona se dobija na sljedeći način:

$$P(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta})}{P(\mathbf{x})} = \frac{\pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{3.12}$$

Posteriorna raspodjela oslikava promjene priorne raspodjele koje su izazvali empirijski podaci izvučeni u obliku uzorka. Ona predstavlja jedan vid ujedinjenja teoretskih pretpostavki o nepoznatoj raspodjeli (iskazanih priornom raspodjelom) i praktičnog aspekta (iskazanog biranjem uzorka i registrovanja njegove vjerodostojnosti). *Predviđajuća (posteriorna) raspodjela* je distribucija novih, do tada neregistrovanih podataka $\mathbf{X}_{new}$ i definiše se sa $P(\mathbf{x}_{new}) := \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{x}) \cdot P(\mathbf{x}_{new}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. Ona se dobija uz pomoć posteriorne raspodjele $P(\boldsymbol{\theta}|\mathbf{x})$ i vjerodostojnosti $P(\mathbf{x}_{new}|\boldsymbol{\theta})$.

Primjer 3.4.1. *Neka je $q \in (0,1)$ nepoznata vjerovatnoća dobijanja pisma pri slučajnom bacanju dvostranog nočića. Frekvencionistički pristup bi podrazumijevao da je q neka vrijednost iz intervala $(0,1)$ koju treba ocijeniti na osnovu izvučenog uzorka. S druge strane, Bejzovski pristup q tretira kao slučajnu promjenljivu. Prvi korak ovog pristupa jeste biranje priorne raspodjele za q. Za priornu raspodjelu slučajne veličine q može se npr. uzeti beta raspodjela sa parametrima $\alpha > 0$ i $\beta > 0$, tj. neka je $\pi(q) = B(\alpha, \beta)$. Beta $B(\alpha, \beta)$ distribucija je raspodjela neprekidnog tipa, čija je gustina raspodjele f data sa*

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1}, & za\ 0 \leqslant x \leqslant 1; \\ 0, & inače, \end{cases}$$

gdje je B beta funkcija definisana sa $B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$. Za ovakav izbor priorne raspodjele, potrebno je izabrati brojeve $\alpha > 0$ i $\beta > 0$, koji predstavljaju hiperparametre. Dalje, bira se slučajan uzorak $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ iz raspodjele koju q određuje. Za vjerodostojnost $P(\boldsymbol{x}|q)$ se može pretpostaviti da ima binomnu $\mathcal{B}(n,q)$ raspodjelu, jer statistika $S_n := \sum_{i=1}^{n} X_i$ upravo ima tu raspodjelu. Uz ovu postavku, određuje se posteriorna raspodjela $P(q|\boldsymbol{x})$: Za proizvoljno $x \in \{0,1,\ldots,n\}$, vrijedi

$$P(q|x) = \frac{\pi(q)\cdot P(x|q)}{\int_0^1 \pi(q)\cdot P(x|q)dq} = \frac{\frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)} \cdot \binom{n}{x} q^x(1-q)^{n-x}}{\int_0^1 \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)} \cdot \binom{n}{x} q^x(1-q)^{n-x} dq}$$
$$= \frac{q^{x+\alpha-1}(1-q)^{n-x+\beta-1}}{\int_0^1 q^{x+\alpha-1}(1-q)^{n-x+\beta-1}dq} = \frac{q^{x+\alpha-1}(1-q)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)}.$$

Kao što se može primijetiti, posteriorna raspodjela se ravna ponovo po beta raspodjeli, ali sada sa parametrima $x+\alpha$ i $n-x+\beta$.

Nazivnik razlomka u formuli (3.12) je normalizaciona konstanta čija je jedina uloga da posteriorna raspodjela bude "prava" raspodjela vjerovatnoća. Stoga su vjerovatnosne težine posteriorne raspodjele proporcionalne brojiocu razlomka u formuli (3.12), što se zapisuje sa $P(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) \cdot P(\mathbf{x}|\boldsymbol{\theta})$.

Primjer 3.4.2. *Neka nepoznata slučajna veličina $\boldsymbol{\mu}$ ima priornu raspodjelu $\pi(\boldsymbol{\mu}) = \mathcal{N}(\mu_0, \sigma_0^2)$ (normalna raspodjela sa hiperparametrima μ_0 i $\sigma_0^2 > 0$) i neka za vjerodostojnost vrijedi $X|\mu \sim \mathcal{N}(\mu, \sigma^2)$, pri čemu je $\sigma^2 > 0$ poznato. To znači da je gustina f priorne raspodjele data sa $f(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot$*

$e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}$, *dok za vjerodostojnost vrijedi* $P(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x_i-\mu)^2}{\sigma^2}}$, *gdje je* $\boldsymbol{X} = (X_1, X_2 \ldots, X_n)$ *uzorak izabran iz* $\mathcal{N}(\mu, \sigma^2)$ *raspodjele. Na osnovu toga, koristeći* $P(\boldsymbol{\mu}|\boldsymbol{x}) \propto \pi(\boldsymbol{\mu}) \cdot P(\boldsymbol{x}|\boldsymbol{\mu})$, *dobija se*

$$P(\mu|\boldsymbol{x}) \propto e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \cdot \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \propto e^{-\left(\frac{1}{2\sigma_0^2}(\mu^2-2\mu_0\mu)+\frac{1}{2\sigma^2}\left(-2\mu\sum_{i=1}^{n} x_i+n\mu^2\right)\right)}$$

$$= e^{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\mu^2+\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)\mu}.$$

Stavljajući $\sigma_1^2 := \dfrac{\sigma^2 \cdot \sigma_0^2}{\sigma^2 + n\sigma_0^2}$ *i* $\mu_1 := \left(\dfrac{\mu_0}{\sigma_0^2} + \dfrac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)\sigma_1^2$, *dalje se dobija*

$$P(\mu|\boldsymbol{x}) \propto e^{-\frac{1}{2}\cdot\frac{\mu^2}{\sigma_1^2}+\frac{\mu_1}{\sigma_1^2}\mu} = e^{-\frac{1}{2\sigma_1^2}(\mu^2-2\mu_1\mu)} \propto e^{-\frac{1}{2\sigma_1^2}(\mu-\mu_1)^2},$$

a nije teško primijetiti da je gustina normalne $\mathcal{N}(\mu_1, \sigma_1^2)$ *raspodjele proporcionalna posljednjem izrazu. Dakle, posteriorna raspodjela ima takođe normalnu raspodjelu, ali sa parametrima* μ_1 *i* σ_1^2.

Za priornu raspodjelu se kaže da je *konjugovana raspodjela za vjerodostojnost*, ako je posteriorna raspodjela istog tipa kao priorna raspodjela. U prethodnim primjerima pokazano je da beta raspodjela predstavlja konjugovanu raspodjelu za binomnu raspodjelu, kao i da je normalna raspodjela konjugovana za normalnu raspodjelu kod koje je poznata disperzija. S obzirom da poznavanje posteriorne raspodjele predstavlja značajan aspekt Bejzovske analize, biranje konjugovane priorne raspodjele olakšava postupak nalaženja posteriorne raspodjele.

Bejzovsko zaključivanje se može uopštiti i na slučaj više od jedne nepoznate slučajne promjenljive. U ovom *višedimenzionalnom Bejzovskom zaključivanju*, $\boldsymbol{\theta}$ je vektor $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k})$, sastavljen od slučajnih promjenljivih $\boldsymbol{\theta_i}$ čije su raspodjele vjerovatnoća nepoznate. Naravno, ovakvo uopštenje podrazumijeva neophodne modifikacije u postupku zadavanja priorne raspodjele, vjerodostojnosti i posteriorne raspodjele. Prvi korak je zadavanje *zajedničke priorne raspodjele* $\pi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k})$; ukoliko su slučajne promjenljive $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}$ nezavisne, tada je dovoljno zadati sve pojedinačne priorne raspodjele $\pi(\boldsymbol{\theta_i})$, jer u tom slučaju vrijedi $\pi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}) = \prod_{i=1}^{k} \pi(\boldsymbol{\theta_i})$. Dalje, bira se slučajan uzorak $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ koji sadrži podatke iz svih raspodjela određenih parametrima $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}$ i nalazi se vjerodostojnost $P(\mathbf{x}|\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k})$. Posljednji (i možda najbitniji) korak je određivanje *zajedničke posteriorne raspodjele*, za koju vrijedi

$$P(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}|\mathbf{x}) \propto \pi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}) \cdot P(\mathbf{x}|\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}).$$

Ukoliko se želi ispitati posteriorna raspodjela slučajne veličine $\boldsymbol{\theta_i}$ relativno u odnosu na uočene podatke, tada je potrebno naći *marginalnu posteriornu raspodjelu* $P(\boldsymbol{\theta_i}|\mathbf{x})$, koja se dobija na sljedeći način:

$$P(\boldsymbol{\theta_i}|\mathbf{x}) = \int \cdots \int P(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_k}|\mathbf{x}) d\boldsymbol{\theta_1} d\boldsymbol{\theta_2} \ldots d\boldsymbol{\theta_{i-1}} d\boldsymbol{\theta_{i+1}} \ldots d\boldsymbol{\theta_k}.$$

Alternativno, u prethodnom integralu se zajednička posteriorna raspodjela može zamijeniti sa

$$\begin{aligned} &P(\boldsymbol{\theta_k}|\mathbf{x}) \cdot P(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_{k-1}}|\boldsymbol{\theta_k}, \mathbf{x}) \\ &= P(\boldsymbol{\theta_k}|\mathbf{x}) \cdot P(\boldsymbol{\theta_{k-1}}|\boldsymbol{\theta_k}, \mathbf{x}) \cdot P(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_{k-2}}|\boldsymbol{\theta_{k-1}}, \boldsymbol{\theta_k}, \mathbf{x}) = \ldots \\ &= P(\boldsymbol{\theta_k}|\mathbf{x}) \cdot P(\boldsymbol{\theta_{k-1}}|\boldsymbol{\theta_k}, \mathbf{x}) \cdot \ldots \cdot P(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_{k-1}}, \boldsymbol{\theta_k}, \mathbf{x}). \end{aligned}$$

Primjer 3.4.3. *Neka je slučajan uzorak* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ *izabran iz* $\mathcal{N}(\mu, \sigma^2)$ *raspodjele, pri čemu su, za razliku od prethodnog primjera, oba parametra* μ *i* σ^2 *slučajne veličine sa nepoznatom raspodjelom. To znači da je vjerodostojnost određena sa*

$$P(\boldsymbol{x}|\mu, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \propto \frac{1}{\sigma^n} \cdot e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\mu-\overline{x_n})^2\right)},$$

gdje je $\overline{x_n} := \dfrac{\sum_{i=1}^{n} x_i}{n}$ *uzoračka sredina, a* $s^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$ *popravljena uzoračka disperzija.*

Zajednička priorna raspodjela za μ *i* σ^2 *biće predstavljena hijerarhijski i to kao proizvod uslovne priorne raspodjele* $\pi(\mu|\sigma^2)$ *i priorne raspodjele* $\pi(\sigma^2)$. *Za uslovnu priornu raspodjelu* $\pi(\mu|\sigma^2)$ *biće izabrana* $\mathcal{N}\left(\mu_0, \dfrac{\sigma^2}{\nu}\right)$ *raspodjela sa hiperparametrima* μ_0 *i* ν, *dok će za priornu raspodjelu* $\pi(\sigma^2)$ *biti pretpostavljeno da je u pitanju inverzna gama distribucija sa hiperparametrima* $\dfrac{k}{2}$ *i* $\dfrac{k\sigma_0^2}{2}$, *tj. distribucija čija je gustina raspodjele* f *data sa*

$$f(x) = \begin{cases} \dfrac{(k\sigma_0^2)^{\frac{k}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \cdot \left(\dfrac{1}{x}\right)^{\frac{k}{2}+1} \cdot e^{-\frac{k\sigma_0^2}{2x}}, & za\ x > 0; \\ 0, & inače, \end{cases}$$

gdje je Γ *gama funkcija definisana sa* $\Gamma(x) := \displaystyle\int_0^{+\infty} t^{x-1}e^{-t}dt$, $x > 0$ *(naziv inverzna gama distribucija potiče iz činjenice da se recipročna vrijednost ravna*

po odgovarajućoj gama raspodjeli). Generalno, dvodimenzionalna distribucija čija je gustina raspodjele g data sa

$$g(x,y) = \begin{cases} \dfrac{\sqrt{\lambda}}{\sqrt{2\pi y}} \cdot \dfrac{\beta^\alpha}{\Gamma(\alpha)} \left(\dfrac{1}{y}\right)^{\alpha+1} \cdot e^{-\frac{2\beta+\lambda(x-\mu)^2}{2y}}, & za\ x \in \mathbb{R}, y > 0; \\ 0, & inače. \end{cases}$$

naziva se normalno-inverzna gama distribucija sa parametrima $\mu, \lambda, \alpha, \beta$ i obilježava se sa $\mathcal{N} - \Gamma^{-1}(\mu, \lambda, \alpha, \beta)$. Uočava se da zajednička priorna raspodjela $\pi(\mu, \sigma^2)$ jeste upravo $\mathcal{N} - \Gamma^{-1}\left(\mu_0, v, \dfrac{k}{2}, \dfrac{k\sigma_0^2}{2}\right)$ distribucija.

Na osnovu navedenog, moguće je pronaći zajedničku posteriornu raspodjelu:

$$\begin{aligned} P(\mu,\sigma^2|\boldsymbol{x}) &\propto \pi(\mu,\sigma^2) \cdot P(\boldsymbol{x}|\mu,\sigma^2) \\ &= \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\mu-\overline{x_n})^2\right)} \cdot \frac{\sqrt{v}}{\sqrt{2\pi\sigma^2}} \frac{(k\sigma_0^2)^{\frac{k}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{k}{2}+1} \cdot e^{-\frac{1}{2\sigma^2}\left(k\sigma_0^2+v(\mu-\mu_0)^2\right)} \\ &\propto \frac{1}{\sqrt{\sigma^2}} \cdot \left(\frac{1}{\sigma^2}\right)^{\frac{n+k}{2}+1} \cdot e^{-\frac{1}{2\sigma^2}\left(k\sigma_0^2+v(\mu-\mu_0)^2+(n-1)s^2+n(\mu-\overline{x_n})^2\right)} \\ &= \frac{1}{\sqrt{\sigma^2}} \cdot \left(\frac{1}{\sigma^2}\right)^{\alpha_1+1} \cdot e^{-\frac{1}{2\sigma^2}\left(2\beta_1+\lambda_1(\mu-\mu_1)^2\right)}, \end{aligned}$$

gdje je

$$\begin{aligned} \mu_1 &:= \frac{\mu_0 v + n\overline{x_n}}{v+n}, \\ \lambda_1 &:= v+n, \\ \alpha_1 &:= \frac{n+k}{2}, \\ 2\beta_1 &:= k\sigma_0^2 + (n-1)s^2 + \frac{nv}{n+v} \cdot \left(\overline{x_n} - \mu_0\right)^2. \end{aligned}$$

Može se prepoznati da je posljednji izraz proporcionalan gustini normalno-gama inverzne distribucije, pa vrijedi $\mu, \sigma^2|\boldsymbol{x} \sim \mathcal{N}-\Gamma^{-1}(\mu_1, \lambda_1, \alpha_1, \beta_1)$. Dakle, normalno-gama inverzna distribucija je konjugovana raspodjela za normalnu raspodjelu čija su oba parametra nepoznata.

Prethodni primjer ilustruje da proces određivanja posteriorne raspodjele kod višedimenzionalne Bejzovskog zaključivanja može biti računski zahtjevan, čak i u slučaju uzimanja konjugovane priorne raspodjele. Stoga je korisno poznavati tehnike aproksimacija koje pojednostavljuju ovaj posao. U ovoj tezi biće razmatrana *MCMC tehnika* (MCMC je skraćenica od Monte Carlo Markov Chain), koja se u literaturi najviše koristi.

MCMC tehnika predstavlja klasu algoritama koji se koriste za generisanje uzoraka iz željene (posteriorne) raspodjele. Dio "Monte Carlo" u nazivu ovih metoda ukazuje da se ovi metodi zasnivaju na slučajnim simulacijama, dok dio naziva "Markov Chain" upućuje da se ovim simulacijama generišu slučajni uzorci koji nisu nezavisni jedan od drugog, već formiraju lanac Markova kod koga je svaki naredni element generisan na osnovu svog prethodnika. Ideja koju MCMC algoritmi eksploatišu je da, uz određene uslove, ovi Markovljevi lanci konvergiraju svojoj stacionarnoj distribuciji. Dovoljni uslovi pod kojima je ovo ispunjeno su navedeni u Fundamentalnoj teoremi za lance Markova. Ako je stacionarna distribucija upravo ciljna (posteriorna) raspodjela, tada se može naći iteracija počev od koje se za sve generisane elemente može smatrati da predstavljaju uzorak iz ciljne (posteriorne) raspodjele.

Najjednostavniji algoritam koji se u okviru MCMC metoda koristi za konstrukciju Markovljevih lanaca je *Gibsov metod uzorkovanja* ili *Gibsov sempler*. Ovaj algoritam se koristi u slučaju višedimenzionalnog Bejzovskog zaključivanja i naročito ga je pogodno koristiti u slučaju kada je posteriorna raspodjela data u zatvorenoj formi, kao i u situaciji kada je uzorkovanje iz zajedničke posteriorne raspodjele komplikovanije u odnosu na uzorkovanje iz marginalnih posteriornih raspodjela. Osnovna ideja Gibsovog semplera je generisanje novih instanci pojedinačnih promjenljivih pod uslovom poznavanja trenutnih vrijednosti svih ostalih promjenljivih. U narednom je opisana konstrukcija Gibsovog semplera u opštem slučaju Bejzovskog zaključivanja.

Neka su $\mathbf{y}$ podaci koji potiču iz uzoračke distribucije $P(\mathbf{y}|\Psi)$, sa priornom raspodjelom $\pi(\Psi)$, za dati skup Ψ sastavljen od bar dvije slučajne veličine. Cilj Bejzovskog zaključivanja je aproksimacija posteriorne raspodjele $P(\Psi|\mathbf{y}) \propto P(\mathbf{y}|\Psi) \cdot \pi(\Psi)$.

Za skup Ψ sastavljen od slučajnih veličina $\psi_1, \psi_2, \ldots, \psi_k$, (prosti) Gibsov metod uzorkovanja podrazumijeva sljedeći niz koraka:

- Inicijalizuje se proizvoljna početna vrijednost $\psi^{(0)} = \left(\psi_1^{(0)}, \psi_2^{(0)}, \ldots, \psi_k^{(0)}\right)$,

- Ako je generisana iteracija $\psi^{(j)} = \left(\psi_1^{(j)}, \psi_2^{(j)}, \ldots, \ldots, \psi_k^{(j)}\right)$, tada se iteracija $\psi^{(j+1)} = \left(\psi_1^{(j+1)}, \psi_2^{(j+1)}, \ldots, \psi_k^{(j+1)}\right)$ generiše na sljedeći način:

  $\psi_1^{(j+1)}$ se uzorkuje iz raspodjele $P\left(\psi_1|\psi_2^{(j)}, \psi_3^{(j)}, \ldots, \psi_k^{(j)}, \mathbf{y}\right)$,
  $\psi_2^{(j+1)}$ se uzorkuje iz raspodjele $P\left(\psi_2|\psi_1^{(j+1)}, \psi_3^{(j)}, \ldots, \psi_k^{(j)}, \mathbf{y}\right)$,

$$\vdots$$

$\psi_i^{(j+1)}$ se uzorkuje iz raspodjele $P\left(\psi_i|\psi_1^{(j+1)},\ldots,\psi_{i-1}^{(j+1)},\psi_{i+1}^{(j)},\ldots,\psi_k^{(j)},\mathbf{y}\right)$,

$$\vdots$$

$\psi_k^{(j+1)}$ se uzorkuje iz raspodjele $P\left(\psi_k|\psi_1^{(j+1)},\psi_2^{(j+1)},\ldots,\psi_{k-1}^{(j+1)},\mathbf{y}\right)$.

Prethodno opisani postupak ima razna uopštenja i modifikacije. Npr., moguće je, umjesto jedne, posmatrati par promjenljivih i uzorkovati iz njihove zajedničke uslovne raspodjele u odnosu na prethodna stanja svih ostalih promjenljivih (to je tzv. *blokovski Gibsov sempler*). Zajedničko za sve varijante Gibsovog semplera je svojstvo da, u većini slučajeva, dobijeni lanac Markova ima jedinstvenu stacionarnu raspodjelu. To efektivno znači da niz generisanih iteracija (shvaćen kao niz stanja promjenljivih), konvergira u raspodjeli ka uzorku iz posteriorne raspodjele. U radu [79], dokazano je da lanac Markova koji se dobije primjenom Gibsovog semplera iz prethodnog razmatranja ispunjava uslove Fundamentalne teoreme za lance Markova, te za njega postoji iteracija počev od koje generisane vrijednosti mogu tretirati kao elementi uzorka iz posteriorne raspodjele. Nedostatak Gibsovog metoda uzorkovanja je što može da se desi da je potrebno izvesti veliki broj iteracija dok se do pomenutog "praga" ne dođe (u literaturi je ovo poznato kao "burn in" period).

Osnovna zamjerka protivnika Bejzovskog zaključivanja je subjektivnost izbora priorne raspodjele i njenih hiperparametara. Argument protiv ovih primjedbi je da priorna raspodjela ne treba da bude "savršena" distribucija koja će u velikoj mjeri biti saglasna sa uočenim podacima, nego početna stanica Bejzovske inferencije. Što se tiče izbora hiperparametara, do njih je moguće doći sredstvima frekvencionističkog pristupa, ali bi to donekle poremetilo koncept Bejzovske paradigme. Stoga, ideja Bejzovskog pristupa se može ponoviti i smatrati da hiperparametar nije vrijednost, već slučajna promjenljiva na čiju raspodjelu vjerovatnoća utiče bar jedan nepoznat parametar. Na taj način se dobija *hijerarhijsko Bejzovsko zaključivanje*, pri čemu je *nivo hijerarhije* određen brojem "ugnježdenih" slučajnih promjenljivih čije raspodjele treba spoznati.

Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ uzorak izabran iz raspodjele koja je određena parametrom (slučajnom veličinom) $\boldsymbol{\theta}$. Takođe, neka je $\boldsymbol{\phi}$ hiperparametar koji određuje priornu raspodjelu $\pi(\boldsymbol{\theta})$. Ukoliko je $\boldsymbol{\phi}$ poznata vrijednost, tada je u pitanju "obični" Bejzovski model, koji je do sada razmatran. Međutim, može se desiti da je $\boldsymbol{\phi}$ slučajna veličina sa nepoznatom raspodjelom vjerovatnoća i svojom priornom raspodjelom $\pi(\boldsymbol{\phi})$. Ako su hiperparametri priorne raspodjele $\pi(\boldsymbol{\phi})$ poznati, dobija se hijerarhijski Bejzovski model sa dva nivoa. Ovaj model određen je zajedničkom priornom raspodjelom

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi}) \cdot \pi(\boldsymbol{\theta}|\boldsymbol{\phi}),$$

dok je zajednička posteriorna raspodjela određena sa

$$P(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \cdot P(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \cdot P(\mathbf{x}|\boldsymbol{\theta}),$$

pri čemu se posljednja jednakost opravdava time što $\boldsymbol{\phi}$ djeluje na vjerodostojnost isključivo preko $\boldsymbol{\theta}$. Ukoliko neki od hiperparametara priorne raspodjele $\pi(\boldsymbol{\phi})$ ima svoju priornu raspodjelu, tada nastupa "spuštanje" na novi nivo hijerarhije, a sam postupak karakterisanja zajedničkih priornih i posteriornih raspodjela je sličan opisanom.

Uzorkovanje iz posteriorne raspodjele hijerarhijskog Bejzovskog modela može da se obavi Gibsovim semplerom, jer je ovaj algoritam i predviđen za korišćenje u slučaju višedimenzionalnog Bejzovskog zaključivanja.

Pitman-Jorov proces je stohastički proces koji ima osnovu u *Pitman-Jorovoj distribuciji*. Najintuitivniji način definisanja ove distribucije je uz pomoć konstrukcije "lomljenja štapa" (na engleskom "stick breaking"). Generalno, ova konstrukcija podrazumijeva sljedeći niz koraka:

1. Štap dužine 1 se na slučajan način lomi na dva dijela, čije su dužine V_1 i $1 - V_1$.

2. Dio štapa koji je dužine $1 - V_1$ se na slučajan način lomi na dva dijela čije su dužine proporcionalne veličinama V_2 i $1 - V_2$. Na taj način se od tog dijela štapa dobijaju dva manja dijela čije su dužine $(1 - V_1) \cdot V_2$ i $(1 - V_1) \cdot (1 - V_2)$.

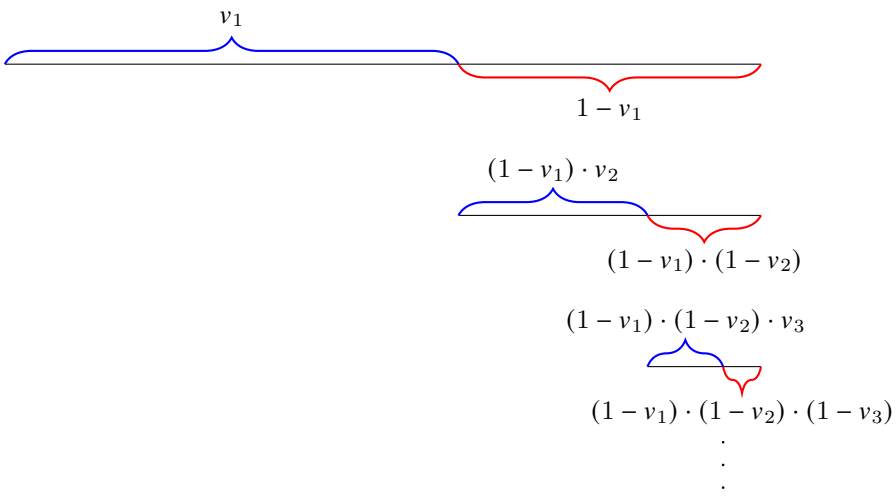
3. Dio štapa koji je dužine $(1 - V_1) \cdot (1 - V_2)$ se na slučajan način lomi na dva dijela čije su dužine proporcionalne veličinama V_3 i $1 - V_3$. Na taj način se od tog dijela štapa dobijaju dva manja dijela čije su dužine $(1 - V_1) \cdot (1 - V_2) \cdot V_3$ i $(1 - V_1) \cdot (1 - V_2) \cdot (1 - V_3)$.

   $\vdots$

Može se primijetiti da je prethodni postupak potpuno određen preciziranjem izbora niza nezavisnih slučajnih promjenljivih $(V_k, k \geqslant 1)$, u smislu zadavanja tačne raspodjele vjerovatnoća po kojoj se ravna svaka slučajna promjenljiva u ovom nizu. Ilustracija konstrukcije lomljenja štapa data je na Slici 3.4.

Neka je $\alpha \in [0, 1)$ i $\beta > 0$. Za prirodan broj $k \geqslant 1$, neka slučajna promjenljiva V_k ima Beta $(1 - \alpha, \beta + k\alpha)$ raspodjelu, tj. neka slučajna promjenljiva V_k ima funkciju gustine raspodjele f_{V_k}, koja je definisana sa

$$f_{V_k}(x) = \begin{cases} \dfrac{\Gamma(1 + \beta + (k-1)\alpha)}{\Gamma(1-\alpha) \cdot \Gamma(\beta + k\alpha)} \cdot \dfrac{(1-x)^{\beta + k\alpha - 1}}{x^{\alpha}}, & \text{za } x \in (0, 1); \\ 0, & \text{inače,} \end{cases}$$



Slika 3.4

gdje je Γ oznaka za gama funkciju. Pored datih uslova, dodatno se pretpostavlja da je $(V_k, k \geqslant 1)$ niz nezavisnih slučajnih promjenljivih. Ako se uvedu oznake

$$p_1 := V_1, p_k := (1 - V_1) \cdot \ldots \cdot (1 - V_{k-1}) \cdot V_k, \quad k \geqslant 2, \tag{3.13}$$

tada se raspodjela slučajnog vektora $(p_1, p_2, \ldots)$ naziva *GEM distribucijom* sa parametrima α i β i označava sa $\mathrm{GEM}(\alpha, \beta)$) (Naziv GEM je akronim dobijen iz imena matematičara Grifitsa, Endžena i MekKlaskija, koji su među prvima posmatrali ovu distribuciju). Ukoliko se za niz slučajnih veličina $(p_n, n \geqslant 1)$ posmatra odgovarajući niz statistika poretka (u kojem su statistike poretka zapisane u opadajućem poretku) i na taj način dobijeni niz označi sa $(\widetilde{p}_n, n \geqslant 1)$, tada se za $\vec{p} := (\widetilde{p}_n, n \geqslant 1)$ kaže da ima *Pitman-Jorovu raspodjelu* sa parametrima α i β (u oznaci $\vec{p} \sim \mathrm{PYD}(\alpha, \beta)$). Parametar α se naziva *parametrom sniženja*, dok se parametar β naziva *parametrom koncentracije*.

Neka je H raspodjela vjerovatnoća nad mjerljivim prostorom $(\mathcal{X}, \mathcal{B})$ i $\vec{p} = (p_1, p_2, \ldots) \sim \mathrm{PYD}(\alpha, \beta)$. Ako je $(X_k, k \geqslant 1)$ niz nezavisnih slučajnih promjenljivih uzorkovanih iz raspodjele H, tada je Pitman-Jorov proces (u oznaci $\mathrm{PYP}(\alpha, \beta, H)$) diskretna distribucija vjerovatnoća na $\mathcal{X}$ definisana sa

$$p\left(x|\alpha, \beta, H\right) = \sum_{k=1}^{\infty} p_k \cdot I\left(X_k = x\right), \tag{3.14}$$

gdje je $I(\cdot)$ indikator događaja navedenog u zagradi, tj. slučajna veličina koja je definisana sa

$$I\left(X_k = x\right) = \begin{cases} 1, & \text{ako je } X_k = x; \\ 0, & \text{inače.} \end{cases}$$

Raspodjela H iz prethodne definicije još se naziva *baznom* ili *početnom raspodjelom*. U principu, Pitman-Jorov proces predstavlja operator koji baznu

raspodjelu nad mjerljivim prostorom $(\mathcal{X}, \mathcal{B})$ transformiše u diskretnu raspodjelu čiji je skup mogućih vrijednosti konačan ili prebrojiv podskup od $\mathcal{X}$. Iz prikaza datog u jednačini (3.14) uočava se da se Pitman-Jorov proces može shvatiti kao izvlačenje niza uzoraka iz bazne raspodjele, što ovaj proces čini diskretnim stohastičkim procesom. Mada bazna raspodjela može biti proizvoljna, u daljnjem izlaganju se posmatraju raspodjele H diskretnog tipa, tj. koristiće se pretpostavka da je H diskretna vjerovatnosna mjera.

Najpoznatija analogija kojom se interpretira Pitman-Jorov proces $\mathrm{PYP}(\alpha, \beta, H)$ je tzv. *proces kineskog restorana* (kraće CRP, što je skraćenica od Chinese Restaurant Process):

Kineski restoran ima beskonačno mnogo stolova numerisanih skupom prirodnih brojeva, pri čemu za svakim stolom može da sjedi beskonačno mnogo gostiju. Niz gostiju pojedinačno ulazi u restoran i svaki gost može da zauzme mjesto za tačno jednim stolom. Pritom se za k-tim stolom u momentu njegovog zauzimanja na slučajan način bira jelo X_k uzorkovano iz bazne raspodjele H. Dakle, gosti za istim stolom konzumiraju isto jelo, ali se, zbog diskretnosti raspodjele H, može desiti da različiti stolovi serviraju isto jelo. Raspored sjedenja gostiju određen je sljedećim postupkom:

- Prvi gost sjeda za prvi sto i naručuje jelo X_1 koje se uzorkuje iz raspodjele H.

- Ako se n gostiju smjestilo za prvih $K_n \leqslant n$ stolova, pri čemu je n_i broj gostiju koji sjede za $i-$tim stolom i $n = \sum_{i=1}^{K_n} n_i$, tada gost $n+1$ sjeda za zauzet sto $i \in \{1, \ldots, K_n\}$ sa vjerovatnoćom $\dfrac{n_i - \alpha}{n + \beta}$ i konzumira već naručeno jelo X_i ili sjeda za prvi slobodan sto $K_n + 1$ sa vjerovatnoćom $\dfrac{\beta + K_n \alpha}{n + \beta}$ i naručuje jelo X_{K_n+1} koje se uzorkuje iz raspodjele H.

Postupak izbora stola iz prethodnog gosti mogu realizovati sekvencijalnim slučajnim izvlačenjem (sa vraćanjem) kuglice iz urne, pri čemu je, za svaku kuglicu u urni vjerovatnoća njenog izvlačenja proporcionalna njenoj težini. Inicijalno, restoran je prazan i urna sadrži samo jednu crnu kuglicu težine β. Ukoliko gost izvuče crnu kuglicu, taj gost sjeda za prvi naredni slobodan sto; crna kuglica se vraća u urnu i njena težina se povećava za α, a u urnu se dodaje bijela kuglica težine $1-\alpha$ koja je numerisana brojem stola koji je taj gost zauzeo. U suprotnom, ako gost izvuče bijelu kuglicu, on sjeda za sto označen brojem na toj kuglici; ta kuglica se vraća u kutiju i težina joj se povećava za 1.

Ako se uvede niz slučajnih veličina $(Y_n, n \geqslant 1)$ tako da je $Y_n := X_k$ ako $n-$ti gost sjeda za $k-$ti sto, tada se iz prethodnog uočava da je, za dati raspored

sjedenja prvih n gostiju, uslovna raspodjela slučajne veličine Y_{n+1} u odnosu na slučajne veličine $Y_1, \ldots, Y_n$ data sa

$$Y_{n+1}|Y_1, \ldots, Y_n, \text{ raspored sjedenja } \sim \sum_{i=1}^{K_n} \frac{n_i - \alpha}{n+\beta}\delta_{X_i} + \frac{\beta + K_n\alpha}{n+\beta} \cdot H. \quad (3.15)$$

Raspored sjedenja prvih n gostiju kineskog restorana određuje jednu slučajnu particiju Π_n skupa $\mathbb{N}_n = \{1, \ldots, n\}$. Pritom je K_n broj skupova (blokova) te particije, a n_i veličina $i-$tog bloka. Relativna veličina $i-$tog bloka u particiji Π_n jednaka je $\frac{n_i}{n}$. Raspored sjedenja svih gostiju je slučajna particija $\Pi_{\mathbb{N}}$ skupa prirodnih brojeva $\mathbb{N}$, gdje je $\Pi_{\mathbb{N}} = \{B_j : j \in \mathbb{N}\}$, pri čemu je $B_j = \bigcup \{B \in \Pi_n : j \in B, n \in \mathbb{N}\}$. Na osnovu opisanog postupka sjedenja u kineskom restoranu slijedi da niz $\left(\frac{n_i}{n}, i \geqslant 1\right)$ konvergira u raspodjeli ka vektoru $(p_1, p_2, \ldots)$, gdje je p_i definisano formulom (3.13). Stoga $(p_1, p_2, \ldots)$ ima $\mathrm{GEM}(\alpha, \beta)$ raspodjelu. Rangiranje ovih vjerovatoća u opadajućem redoslijedu dovodi do vektora $\vec{p} = (\widetilde{p_n}, n \geqslant 1)$ izabranog iz $\mathrm{PYD}(\alpha, \beta)$ raspodjele, pa se za niz slučajnih veličina $(Y_n, n \geqslant 1)$ koji je definisan u prethodnom razmatranju može smatrati da je uzorkovan iz $\mathrm{PYP}(\alpha, \beta, H)$ distribucije. U tome leži značaj analogije procesa kineskog restorana jer se, uz pravilno indeksiranje, CRP može koristiti kao alternativni metod uzorkovanja iz PYP u slučaju kada je nepoznat vjerovatnosni vektor $\vec{p}$.

Za prirodan broj n i realne brojeve x, y, *rastući faktorijel* broja x i *rastući faktorijel* broja x *sa inkrementom* y kraće se zapisuju pomoću *Pohamerovih simbola* na sljedeći način:

$$(x)^{(n)} := \begin{cases} 1, & \text{za } n = 0; \\ x(x+1) \cdot \ldots \cdot (x+n-1), & \text{za } n \geqslant 1. \end{cases}$$

$$(x|y)^{(n)} := \begin{cases} 1, & \text{za } n = 0; \\ x(x+y) \cdot \ldots \cdot (x+(n-1)y), & \text{za } n \geqslant 1. \end{cases}$$

Generalizovani Stirlingovi brojevi biće definisani u sklopu sljedeće "igre razvrstavanja":

Konačno mnogo različitih kuglica se jedna za drugom razvrstavaju po unaprijed zadatom redoslijedu. Pritom, svaka kuglica se stavlja ili u "rupu" ili u jednu od trenutno raspoloživih urni. Za prvu kuglicu, inicijalno je na raspolaganju $r \geqslant 0$ urni. Za date cijele brojeve t i q, nakon smještanja prve kuglice u neku od urni, broj urni se povećava za $-t$ (tj. smanjuje se za t), dok se u slučaju smještanja prve kuglice u rupu, broj urni se povećava za $q - t$. Sličan način razvrstavanja se dalje izvodi za drugu kuglicu, sa promjenom što se umjesto r koristi "ažurirani" broj urni. Ovaj postupak se ponavlja sve dok se ne razvrstaju sve kuglice. Opisana igra razvrstavanja u daljnjem će biti referisana kao $(t, q, r)-$*igra razvrstavanja*.

Primjer 3.4.4. *U $(-2, 18, 10)$-igri razvrstavanja inicijalno je na raspolaganju* 10 *urni. Nakon smještanja kuglice u urnu, broj urni se povećava za* $-(-2) = 2$*, dok se nakon smještanja kuglice u rupu, broj urni povećava za* $18 - (-2) = 20$.

Neka su $N, M, -t, q$ i r prirodni brojevi. *Generalizovani Stirlingov broj* $S_M^N(t, q, r)$ predstavlja broj načina da se N kuglica rasporedi u $(t, q, r)-$igri razvrstavanja tako da tačno M njih bude smješteno u rupu. Po definiciji se stavlja $S_0^0(t, q, r) := 1$.

Primjer 3.4.5. *Za nalaženje vrijednosti* $S_1^3(-2, 18, 10)$*, tj. broja načina da se* 3 *kuglice rasporede u* $(-2, 18, 10)$*-igri razvrstavanja tako da tačno jedna od njih bude smještena u rupu, može se primijetiti sljedeće: Ako je prva kuglica smještena u rupu, tada druga kuglica ide u jednu od* 30 *urni, a treća kuglica ide u jednu od* 32 *urne; ako je druga kuglica smještena u rupu, tada je broj raspoloživih urni za smještanje prve i treće kuglice jednak redom* 10 *i* 32*; a ako je treća kuglica smještena u rupu, tada je broj raspoloživih urni za smještanje prve i druge kuglice jednak redom* 10 *i* 12*. Zbog toga je* $S_1^3(-2, 18, 10) = 30 \cdot 32 + 10 \cdot 32 + 10 \cdot 12 = 1400$*. Generalno, sličnim rezonovanjem se dobija* $S_1^3(t, q, r) = (r + q - t)(r + q - 2t) + r(r + q - 2t) + r(r - t) = 3r(r - t) + (3r - 2t + q)(q - t)$.

Primjedba 3.4.6. Generalizovani Stirlingovi brojevi prebrojavaju odgovarajuće razmještaje kuglica u $(t, q, r)-$igri razvrstavanja. Ovaj broj nije jednak broju finalnih konfiguracija kuglica. Stoga je nebitno da li se nove urne nadodaju na postojeće ili se u svakom novom razmještaju sve prethodne urne zamijene sa odgovarajućim brojem novih urni. Slično, u slučajevima kada je $t > 0$ ili $q - t < 0$, broj urni se može smanjiti i nebitno je koje urne su izabrane za uklanjanje.

Rekurentna veza koja određuje generalizovane Stirlingove brojeve je iskazana u sljedećoj lemi.

Lema 3.4.7. *[64] Neka su $N, M, -t, q$ i r prirodni brojevi. Tada vrijedi*

$$
\begin{aligned}
&S_M^N(t, q, r) = 0, \text{ za } N < M,\\
&S_0^N(t, q, r) = (r|-t)^{(N)},\\
&S_{M+1}^{N+1}(t, q, r) = S_M^N(t, q, r) + (r - Nt + q(M + 1)) \cdot S_{M+1}^N(t, q, r).
\end{aligned}
$$

Dokaz. U rupu se ne može smjestiti više kuglica nego što ih se razmješta, što znači da za $M < N$ vrijedi $S_M^N(t, q, r) = 0$. Ako u rupi nema kuglica, tada se u svakom novom razmještaju broj urni povećava za $-t$, pa je

$$S_0^N(t, q, r) = r(r - t)(r - 2t) \cdot \ldots \cdot (r - (N - 1)t), \ N \geqslant 1.$$

Time je dokazano $S_0^N(t, q, r) = (r|-t)^{(N)}$, pa ostaje još da se potvrdi treća jednakost.

Neka je $N+1$ kuglica raspoređeno pomoću $(t,q,r)-$igre razvrstavanja, tako da je tačno $M+1$ njih smješteno u rupu. Posljednja, $N+1-$va kuglica može biti smještena ili u rupu ili u neku od raspoloživih urni. Ako je smještena u rupu, tada se ostalih N kuglica raspoređuje tako da se M njih smjesti u rupu, što je moguće uraditi na $S_M^N(t,q,r)$ načina. Ukoliko je posljednja kuglica smještena u urnu, tada se ostalih N kuglica raspoređuje tako da se $M+1$ njih smjesti u rupu, što je moguće uraditi na $S_{M+1}^N(t,q,r)$ načina i, u tom slučaju, za smještanje posljednje kuglice je u opticaju $r-(N-M-1)t+(q-t)(M+1)=r-Nt+q(M+1)$ raspoloživih urni. □

Generalizovani Stirlingovi brojevi mogu da se uopšte tako da jednakosti date u prethodnoj lemi vrijede za proizvoljne realne brojeve t, q i r. Dokaz postojanja i jedinstvenosti ovakvog produženja može se naći npr. u [64]. U nastavku teze, od interesa će biti isključivo generalizovani Stirlingovi brojevi za $t=-1$, $q=-\alpha$ i $r=0$, gdje je α parametar sniženja kod Pitman-Jorovog procesa. U skladu sa tim, biće korišćena kraća oznaka $S_{M,\alpha}^N := S_M^N(-1,-\alpha,0)$. Na osnovu prethodne teoreme, vrijedi

$$S_{M,\alpha}^N = 0, \text{ za } M > N, \tag{3.16}$$

$$S_{0,\alpha}^N = \begin{cases} 1, & \text{za } N=0; \\ 0, & \text{inače.} \end{cases} \tag{3.17}$$

$$S_{M,\alpha}^N = S_{M-1,\alpha}^{N-1} + (N-1-M\alpha)\cdot S_{M,\alpha}^{N-1}, \text{ za } 0 < M \leqslant N. \tag{3.18}$$

Ako je $(Y_1,\ldots,Y_n)$ konačan uzorak izabran iz $PYP(\alpha,\beta,H)$ raspodjele, tada se, zbog diskretnosti bazne raspodjele H, može desiti da uočeni uzorak sadrži jednake elemente. U terminima CRP analogije, neka je M_n broj različitih jela na uočenom uzorku od n jela i $\{Y_1^*,\ldots,Y_{M_n}^*\}$ meni na osnovu kojeg se poslužuje K_n stolova kineskog restorana. Ako l_i predstavlja višestrukost jela Y_i^*, tj. broj stolova za kojim se ovo jelo servira, pri čemu je $K_n=\sum_{i=1}^{M_n} l_i$, i ako je m_i ukupan broj gostiju koji sjedi za stolovima za kojim se servira jelo Y_i^*, tada je moguće odrediti zajedničku raspodjelu slučajnih vektora $(Y_1,\ldots,Y_n)$ i $(l_1,\ldots,l_{M_n})$. Ova raspodjela se kraće zapisuje uz pomoć rastućih faktorijela i generalizovanih Stirlingovih brojeva i određena je u sljedećoj lemi.

Lema 3.4.8. *[22] Zajednička raspodjela za $(Y_1,\ldots,Y_n)$ i $(l_1,\ldots,l_{M_n})$ je data sa*

$$P\{Y_1,\ldots,Y_n;l_1,\ldots,l_{M_n}\} = \frac{(\beta|\alpha)^{(K_n)}}{(\beta)^{(n)}}\cdot\prod_{i=1}^{M_n}\left(H(Y_i^*)^{l_i}\cdot S_{l_i,\alpha}^{m_i}\right). \tag{3.19}$$

Dokaz. Najprije treba primijetiti da na vrijednost vjerovatnoće iz formule (3.19) ne utiče raspored ulaženja gostiju u kineski restoran (ovo svojstvo uzorka

$(Y_1, \ldots, Y_n)$ izabranog iz PYP raspodjele se naziva *razmjenjivost*). Zbog toga, u daljnjem se podrazumijeva raspored u kojem u restoran najprije uđu svi gosti koji konzumiraju jelo Y_1^*, zatim svi gosti koji konzumiraju jelo Y_2^*, itd. sve dok naposlijetku u restoran ne uđu svi gosti koji konzumiraju jelo $Y_{M_n}^*$.

Sljedeći korak podrazumijeva određivanje težine vjerovatnoće iz formule (3.19) koji "otpada" na fiksirano jelo iz ponuđenog menija. Ako se za jelo Y_i^* ovaj udio označi sa $P(Y_i^*, l_i, m_i)$, tada očigledno vrijedi

$$P\{Y_1, \ldots, Y_n; l_1, \ldots, l_{M_n}\} = \prod_{j=1}^{M_n} P(Y_j^*, l_j, m_j).$$

Jelo Y_1^* se servira na l_1 stolova i konzumira ga ukupno m_1 gostiju. Neka je $A(m_1, l_1)$ skup svih mogućih rasporeda sjedenja m_1 gostiju za l_1 stolova za kojim se služi jelo Y_1^*, ali tako da za svakim od ovih stolova sjedi bar jedan gost. Tada vrijedi

$$P(Y_1^*, l_1, m_1) = \sum_{a \in A(m_1, l_1)} P\left(Y_1^*, l_1, m_1, a\right).$$

Za raspored sjedenja $a \in A(m_1, l_1)$ i $k \in \{1, \ldots, l_1\}$, neka je $m_1(a, k)$ broj gostiju koji sjedi za k-tim stolom za kojim se služi jelo Y_1^*. Za dati raspored sjedenja a, vjerovatnoću $P\left(Y_1^*, l_1, m_1, a\right)$ je moguće odrediti rekurzivnom primjenom formule (3.15). Na taj način se dobija

$$P(Y_1^*, l_1, m_1, a) = \frac{(\beta|\alpha)^{(l_1)}}{(\beta)^{(m_1)}} \cdot H(Y_1^*)^{l_1} \cdot \prod_{k=1}^{l_1} (1-\alpha)^{(m_1(a,k)-1)}.$$

Ako se sa $A(m_2, l_2)$, označi skup svih rasporeda sjedenja m_2 gostiju za l_2 stolova za kojim se služi jelo Y_2^* tako da za svakim ovakvim stolom sjedi bar jedan gost i ukoliko se za $a \in A(m_2, l_2)$ stavi da je $m_2(a, k)$ broj gostiju u grupi koji sjedaju za sto k, tada za svako $a \in A(m_2, l_2)$ vrijedi

$$P(Y_2^*, l_2, m_2, a) = \frac{(\beta + l_1\alpha|\alpha)^{(l_2)}}{(\beta + m_1)^{(m_2)}} \cdot H(Y_2^*)^{l_2} \cdot \prod_{k=1}^{l_2} (1-\alpha)^{(m_2(a,k)-1)}.$$

Slično se računaju vjerovatnoće za jela $Y_3^*, Y_4^*, \ldots$, zaključno sa vjerovatnoćom

$$P\left(Y_{M_n}^*, l_{M_n}, m_{M_n}, a\right) = \frac{\left(\beta + \left(\sum_{i=1}^{M_n-1} l_i\right)\alpha|\alpha\right)^{(l_{M_n})}}{\left(\beta + \sum_{i=1}^{M_n-1} m_i\right)^{(m_{M_n})}} \cdot H(Y_{M_n}^*)^{l_{M_n}} \cdot \prod_{k=1}^{l_{M_n}} (1-\alpha)^{(m_{M_n}(a,k)-1)}.$$

Za kompletiranje dokaza, dovoljno je dokazati da za svako $i \in \{1, \ldots, M_n\}$ vrijedi

$$\sum_{a \in A(m_i, l_i)} \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)} = S^{m_i}_{l_i,\alpha}. \tag{3.20}$$

Ovo se može dokazati indukcijom po prirodnim brojevima m_i i l_i. Preciznije, ako se za prirodne brojeve m_i, l_i definiše

$$D(m_i, l_i, \alpha) := \begin{cases} 1, & \text{za } m_i = l_i = 0; \\ 0, & \text{za } l_1 = 0 < m_i; \\ \displaystyle\sum_{a \in A(m_i, l_i)} \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)}, & \text{inače.} \end{cases}$$

biće dokazano da za sve prirodne brojeve m_i, l_i vrijedi $D(m_i, l_i, \alpha) = S^{m_i}_{l_i,\alpha}$. Iz prethodne definicije i svojstva (3.17) slijedi

$$D(m_i, 0, \alpha) = S^{m_i}_{0,\alpha} = \begin{cases} 1, & \text{za } m_i = 0; \\ 0, & \text{inače.} \end{cases}$$

Takođe, za $m_i < l_i$, iz svojstva (3.16) i činjenice da ne postoji raspored sjedenja iz skupa $A(m_i, l_i)$ u okviru kojeg bi za svakim od l_i posmatranih stolova sjedio po bar jedan gost, slijedi da u tom slučaju vrijedi $D(m_i, l_i, \alpha) = 0 = S^{m_i}_{l_i,\alpha}$. Stoga, neka su m_i, l_i prirodni brojevi za koje vrijedi $0 < l_i \leqslant m_i$, uz pretpostavku da za sve brojeve $0 < m'_i \leqslant m_i$ i $0 < l'_i \leqslant l_i$, pri čemu je bar jedna od ove dvije nejednakosti stroga (tj. vrijedi $m'_i < m_i$ ili $l'_i < l_i$), vrijedi da je $D(m'_i, l'_i, \alpha) = S^{m'_i}_{l'_i,\alpha}$. Ideja dokaza je da se skup $A(m_i, l_i)$ podijeli na $l_i + 1$ podskupova u zavisnosti od toga gdje sjeda posljednji gost od posmatranih m_i gostiju:

- Neka je $A_0(m_i, l_i) \subseteq A(m_i, l_i)$ skup svih rasporeda sjedenja za koje posljednji gost sjedi sam za nekim od m_i stolova. Uklanjanje ovog gosta dovodi do rasporeda sjedenja ostalih $m_i - 1$ gostiju za ostalih $l_i - 1$ stolova i nije teško primijetiti da je ovim uspostavljena injektivna korespondencija između skupova $A_0(m_i, l_i)$ i $A(m_i - 1, l_i - 1)$. Takođe, uočava se da dodavanje posljednjeg gosta ne doprinosi povećanju vrijednosti broja $D(m_i - 1, l_i - 1, \alpha)$.

- Za $k' \in \{1, \ldots, l_i\}$, neka je $A_{k'}(m_i, l_i) \subseteq A(m_i, l_i)$ skup svih rasporeda sjedenja za koje posljednji gost sjedi za stolom k', pri čemu za ovim stolom sjede bar dva od posmatranih m_i gostiju. Uklanjanje ovog gosta ne dovodi do toga da sto k' ostane prazan, što dovodi da raspored sjedenja ostalih $m_i - 1$ gostiju za l_i stolova. Ovim je uspostavljena bijekcija između skupova $A_{k'}(m_i, l_i)$ i $A(m_i - 1, l_i)$. Takođe, uočava se da dodavanje posljednjeg gosta dodaje faktor $m_i(a, k') - \alpha$ svakom $a \in A(m_i - 1, l_i)$.

Uzimajući u obzir prethodno razmatranje, dobija se

$$D\left(m_i, l_i, \alpha\right) = \sum_{a \in A_0(m_i, l_i)} \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)}$$

$$+ \sum_{k'=1}^{l_i} \sum_{a \in A_{k'}(m_i, l_i)} \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)} = \sum_{a \in A(m_i-1, l_i-1)} \prod_{k=1}^{l_i-1} (1-\alpha)^{(m_i(a,k)-1)}$$

$$+ \sum_{k'=1}^{l_i} \sum_{a \in A(m_i-1, l_i)} \left(m_i(a,k') - \alpha\right) \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)} = D\left(m_i - 1, l_i - 1, \alpha\right)$$

$$+ \sum_{a \in A(m_i-1, l_i)} \sum_{k'=1}^{l_i} \left(m_i(a,k') - \alpha\right) \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)} = D\left(m_i - 1, l_i - 1, \alpha\right)$$

$$+ \left(m_i - 1 - l_i\alpha\right) \cdot \sum_{a \in A(m_i-1, l_i)} \prod_{k=1}^{l_i} (1-\alpha)^{(m_i(a,k)-1)} = D\left(m_i - 1, l_i - 1, \alpha\right)$$

$$+ \left(m_i - 1 - l_i\alpha\right) \cdot D\left(m_i - 1, l_i, \alpha\right) = S^{m_i-1}_{l_i-1,\alpha} + \left(m_i - 1 - l_i\alpha\right) \cdot S^{m_i-1}_{l_i,\alpha} \overset{(3.18)}{=} S^{m_i}_{l_i,\alpha}.$$

Time je dokaz kompletiran. □

Značajna karakteristika Pitman-Jorovog procesa je transformacija bazne (ulazne) raspodjele u novu (izlaznu) diskretnu raspodjelu. Ovo svojstvo omogućava formiranje (konačnog) niza Pitman-Jorovih procesa koji su rekurzivno izgrađeni na način da bazna raspodjela jednog člana ovog niza predstavlja izlaznu raspodjelu njegovog prethodnog člana. Na taj način dobijena struktura naziva se *hijerarhijskim Pitman-Jorovim procesom*. Ovakva struktura je često prisutna kao priorna raspodjela u Bejzovskim hijerarhijskim modelima inferencije. Konkretan oblik ove hijerarhije korišćen u ovom radu opisan je u narednom izlaganju.

Neka je string X dužine T dat u obliku $X = X_1 X_2 \ldots X_T$, pri čemu su $X_1, \ldots, X_T$ diskretne slučajne veličine koje uzimaju vrijednosti iz konačnog alfabeta $\mathbb{N}_n = \{1, 2, \ldots, n\}$. Ranije je konstatovano da se vjerovatnoća da string X poprimi konkretnu vrijednost $x_{[1,T]} = x_1 x_2 \ldots x_T$ dobija računanjem uslovnih vjerovatnoća po formuli

$$P(x_{[1,T]}) = \prod_{i=1}^{T} P\left(x_i | x_{[1,i-1]}\right), \tag{3.21}$$

gdje se po konvenciji uzima da je $x_{[1,0]} = \varepsilon$ i $x_1|\varepsilon = x_1$. Takođe, opisan je postupak ocjenjivanja ovih uslovnih vjerovatnoća upotrebom sufiksnog vjerovatnosnog drveta u slučaju da je $X_1 X_2 \ldots X_T$ lanac Markova reda r. Ovaj postupak je efektivno posmatrao samo kontekste dužine r, tj. za svako i korišćena je jednakost $P\left(x_i | x_{[1,i-1]}\right) = P\left(x_i | x_{[i-r,i-1]}\right)$. Sada će biti izgrađen model

ocjenjivanja vjerovatnoće P iz formule (3.21) koji uzima u obzir kontekste proizvoljnih dužina.

Pri izgradnji takvog modela, potrebno je posmatrati vektor predviđajućih vjerovatnoća za svaki simbol relativno u odnosu na svaki postojeći kontekst koji ovom simbolu može prethoditi. Neka je $\mathbf{s}$ string proizvoljne dužine, $G_{\mathbf{s}}(v)$ vjerovatnoća pojavljivanja simbola $v \in \mathbb{N}_n$ nakon ostvarivanja konteksta $\mathbf{s}$ i $G_{\mathbf{s}}$ vektor vjerovatnoća, sa po jednom komponentom $G_{\mathbf{s}}(v)$ za svako $v \in \mathbb{N}_n$. Ideja leži u zadavanju priorne raspodjele kojom će se hijerarhijski povezivati vektor predviđajućih vjerovatnoća u odnosu na konkretan kontekst sa vektorima predviđajućih vjerovatnoća u odnosu na njegove jednostavnije i kraće podkontekste. Na taj način uspostavljena rekurentna veza omogućava praćenje zavisnosti između ocjenjenih predviđajućih vjerovatnoća u odnosu na duže kontekste i ocjenjenih predviđajućih vjerovatnoća u odnosu na odgovarajuće podkontekste.

Za skup $\mathbb{N}_n^*$ svih mogućih konteksta, posmatra se hijerahijska Bejzovska priorna raspodjela postavljena na skupu vjerovatnosnih vektora $\{G_{\mathbf{s}} : \mathbf{s} \in \mathbb{N}_n^*\}$ na sljedeći način: Za prazan kontekst ε, neka je $G_\varepsilon|\alpha_0, \beta_0, H \sim \mathrm{PYP}(\alpha_0, \beta_0, H)$, pri čemu je (globalna) bazna raspodjela H jednaka uniformnoj raspodjeli na skupu $\mathbb{N}_n$. Na prvom nivou hijerarhije, za proizvoljno $v \in \mathbb{N}_n$ se uzima da $G_v|\alpha_1, \beta_1, G_\varepsilon \sim \mathrm{PYP}(\alpha_1, \beta_1, G_\varepsilon)$. U opštem slučaju, hijerarhija je definisana rekurzivno na način da za proizvoljan neprazan kontekst $\mathbf{s}$ vrijedi

$$G_{\mathbf{s}}|\alpha_{|\mathbf{s}|}, \beta_{|\mathbf{s}|}, G_{\sigma(\mathbf{s})} \sim \mathrm{PYP}(\alpha_{|\mathbf{s}|}, \beta_{|\mathbf{s}|}, G_{\sigma(\mathbf{s})}), \tag{3.22}$$

gdje je $\sigma(\mathbf{s})$ najduži pravi sufiks konteksta $\mathbf{s}$ (tj. $\sigma(\mathbf{s})$ je string $\mathbf{s}$ bez prvog simbola), a $|\mathbf{s}|$ oznaka za dužinu konteksta $\mathbf{s}$.

Dakle, za svaki neprazan kontekst $\mathbf{s}$, priorna raspodjela vjerovatnosnog vektora $G_{|\mathbf{s}|}$ odgovara Pitman-Jorovom procesu sa parom hiperparametara $\alpha_{|\mathbf{s}|}$, $\beta_{|\mathbf{s}|}$ koji odgovaraju nivou $|\mathbf{s}|$ date hijerarhije i ulaznom (baznom) raspodjelom $G_{\sigma(\mathbf{s})}$. Ovakav izbor priora izražava predubjeđenje da simbol koji se nalazi najranije u kontekstu ima najmanji uticaj u određivanju vjerovatnoće posljednjeg simbola. Za proizvoljan prirodan broj m, neka je priorna raspodjela parametra α_m beta $B(a_m, b_m)$ raspodjela, a priorna raspodjela parametra β_m gama $\Gamma(f_m, h_m)$ raspodjela. Ovako dobijen hijerahijski Bejzovski model predstavlja jedan oblik Pitman-Jorovog hijerarhijskog procesa i u daljnjem će se za njegovo označavanje koristiti skraćenica HPYP. Ovaj proces ima strukturu drveta čiji su čvorovi vjerovatnosni vektori $G_{\mathbf{s}}$, a broj grana koje proizilaze iz svakog čvora jednak je broju simbola alfabeta $\mathbb{N}_n$.

Analogija procesa kineskog restorana može da se generalizuje kako bi se dobila interpretacija hijerarhijskog Pitman-Jorovog procesa. Ovo uopštenje se naziva procesom franšize kineskih restorana (u oznaci CRFP, što je akronim od Chinese Restaurant Franchise Process) i ono se dobija na sljedeći način:

Svakom slučajnom vektoru $G_{\mathbf{s}}$ iz prethodno uvedenog hijerahijskog Pitman-Jorovog procesa se pridružuje tačno jedan kineski restoran $\mathcal{R}_{\mathbf{s}}$ u skladu sa već opisanom CRP konstrukcijom. Pritom je svaki restoran $\mathcal{R}_{\mathbf{s}}$, izuzev restorana $\mathcal{R}_{\varepsilon}$, hijerahijski povezan sa restoranom-roditeljem $\mathcal{R}_{\sigma(\mathbf{s})}$. Niz gostiju koji ulaze u restoran $\mathcal{R}_{\mathbf{s}}$ odgovara stringovima koji počinju kontekstom $\mathbf{s}$, dok su jela koja se naručuju u ovom (a i svakom drugom restoranu) iz istog menija koji odgovara simbolima alfabeta $\mathbb{N}_n$. Kada se za novozauzetim stolom restorana $\mathcal{R}_{\mathbf{s}}$ naručuje jelo, tada se ovo jelo bira iz bazne distribucije $G_{\sigma(\mathbf{s})}$. Ovo efektivno znači da se taj sto šalje u roditeljski restoran $\mathcal{R}_{\sigma(\mathbf{s})}$, gdje se tretira kao gost tog restorana. Opisani postupak se izvodi rekurzivno sve do momenta kada u nekom od restorana-roditelja sto-gost ne sjedne za već zauzet sto i konzumira jelo koje se servira za tim stolom ili dok ne stigne do restorana praroditelja $\mathcal{R}_{\varepsilon}$ u kojem se jela (simboli) naručuju (biraju) iz menija (alfabeta Σ) po uniformnoj raspodjeli. Kada se to ostvari, za odgovarajućim stolovima restorana-potomaka ove grane, zaključno sa onim od kojeg je započeo rekurzivni poziv, se poslužuje inicijalizovano jelo. Primjećuje se da svaki restoran iz posmatrane franšize ima dvije vrste gostiju: "nezavisne" koji dolaze samostalno i mušterije koji su stolovi poslati od strane restorana-potomaka. Uslovna raspodjela sjedenja data formulom (3.15) vrijedi u svakom restoranu franšize i ona ukazuje na to da u CRFP vrijedi princip "popularnosti", u smislu da naručivanje jela v u restoranu $\mathcal{R}_{\mathbf{s}}$ povećava vjerovatnoću ponovnog naručivanja jela v u tom restoranu.

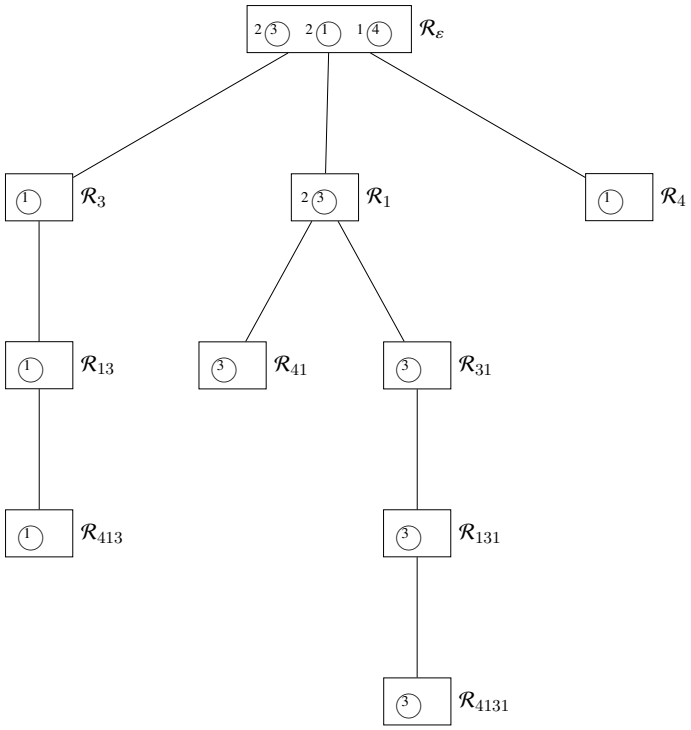
Realizacija CRFP podrazumijeva da se najprije "popune" restorani franšize koji nemaju potomaka, a da se zatim razmještanje izvede za njihove restorane-roditelje koji se nalaze na istom nivou hijerarhije (tj. za restorane koji odgovaraju kontekstima istih dužina) i da se ovaj postupak nastavi sve dok se ne stigne do restorana praroditelja. Preciznije, za familiju stringova $\mathcal{D}$, u svojstvu trening podataka, posmatra se skup konteksta $\mathcal{D}^* = \{\mathbf{s} : \mathbf{s}$ je podstring nekog stringa iz familije $\mathcal{D}\}$. Svakom kontekstu $\mathbf{s} \in \mathcal{D}^*$ odgovara kineski restoran $\mathcal{R}_{\mathbf{s}}$. Naručivanje jela v u restoranu $\mathcal{R}_{\mathbf{s}}$ odgovara realizaciji simbola v nakon konteksta $\mathbf{s}$ i ovo naručivanje je moguće samo ukoliko se kontekst $\mathbf{s}v$ pojavljuje kao podstring nekog stringa iz familije $\mathcal{D}$ (sa napomenom da se restoran u kojem nije moguće naručiti nijedno jelo može zanemariti, tj. isključiti iz franšize). Restorani franšize koji nemaju potomaka su oni koji odgovaraju kontekstima koji su neprazni prefiksi stringova iz familije $\mathcal{D}$. Nakon što se "nasele" ovakvi restorani, prelazi se na sukcesivno popunjavanje njihovih restorana-roditelja i ovaj postupak se nastavlja sve dok se ne dođe do restorana-praroditelja.

Primjer 3.4.9. *Neka je* $\mathcal{D} = \{41313\}$ *sastavljen od jednog stringa čiji simboli pripadaju alfabetu* $\mathbb{N}_4$. *Tada je*

$$\mathcal{D}^* = \{\varepsilon, 4, 1, 3, 41, 13, 31, 413, 131, 313, 4131, 1313, 41313\}.$$

Restorani bez potomaka odgovaraju kontekstima $4131, 413, 41, 4$. *U ovom slu-*

čaju se CRPF sastoji od 10 *kineskih restorana. Jedna validna realizacija CRPF za ovaj skup konteksta predstavljena je u obliku prefiksnog drveta na Slici 3.5.*



Slika 3.5: Pravougaonici predstavljaju restorane, krugovi predstavljaju stolove, brojevi unutar njih predstavljaju jela (simbole) koji se za tim stolovima serviraju, dok brojevi pored krugova predstavljaju broj gostiju za odgovarajućim stolom

Uzimajući u obzir postupak biranja stolova i biranja jela, može se zaključiti da realizacija CRFP ne zavisi u potpunosti od trening podataka. Tako u prethodnom primjeru, dva gosta restorana $\mathcal{R}_1$ konzumiraju jelo 3 za istim stolom, ali moguća je i realizacija da ova dva gosta sjede za različitim stolovima ovog restorana i konzumiraju jelo 3. Stoga, da bi realizacija CRFP bila u potpunosti određena, pored trening podataka kojim se zadaje meni u svakom restoranu franšize, potrebno je znati konfiguraciju zauzetih stolova u svakom restoranu franšize. Pod konfiguracijom zauzetih stolova nekog restorana podrazumijeva se broj zauzetih stolova, broj zauzetih stolova za kojim se servira konkretno jelo, kao i broj gostiju za svakim stolom pojedinačno. Radi lakšeg opisivanja konfiguracije zauzetih stolova u CRFP, biće korišćene dodatne statistike.

Neka je $\mathbf{s} \in \mathcal{D}^*$ proizvoljan kontekst. Za restoran $\mathcal{R}_\mathbf{s}$ biće uvedene sljedeće statistike:

- $d_\mathbf{s} := |\mathbf{s}|$ predstavlja nivo restorana $\mathcal{R}_\mathbf{s}$ u datoj hijerarhiji.

- $t_{\mathbf{s}v}$ je broj stolova u restoranu $\mathcal{R}_{\mathbf{s}}$ koji serviraju jelo $v \in \mathbb{N}_n$.
- $t_{\mathbf{s}\cdot}$ je broj zauzetih stolova u restoranu $\mathcal{R}_{\mathbf{s}}$, tj. $t_{\mathbf{s}\cdot} = \sum_{v \in \mathbb{N}_n} t_{\mathbf{s}v}$.
- $c^0_{\mathbf{s}v}$ je broj nezavisnih gostiju restorana $\mathcal{R}_{\mathbf{s}}$ koji konzumiraju jelo $v \in \mathbb{N}_n$.
- $c_{\mathbf{s}v}$ je ukupan broj gostiju restorana $\mathcal{R}_{\mathbf{s}}$ koji konzumiraju jelo $v \in \mathbb{N}_n$ (uključujući i goste-stolove poslate od strane restorana potomaka)
- $c_{\mathbf{s}vk}$ je ukupan broj gostiju restorana $\mathcal{R}_{\mathbf{s}}$ koji konzumiraju jelo $v \in \mathbb{N}_n$ i sjede za stolom k, pri čemu je $c_{\mathbf{s}vk} = 0$ ako se za stolom k ne servira jelo v.
- $c_{\mathbf{s}\cdot}$ je ukupan broj gostiju restorana $\mathcal{R}_{\mathbf{s}}$, tj. $c_{\mathbf{s}\cdot} = \sum_{v \in \mathbb{N}_n} c_{\mathbf{s}v}$.

Pored datih statistika, korisno je uvesti veličinu koja nije vezana za konkretni restoran iz franšize, a koja za svakog gosta bilježi doprinos rezervaciji novih stolova na određenom nivou hijerarhije. Preciznije, za gosta g posmatrane franšize, doprinos ovog gosta u rezervaciji novih stolova ili indikator novog stola ovog gosta (u oznaci u_g) je dat sa $u_g = d_{\mathbf{s}}$, gdje je $\mathcal{R}_{\mathbf{s}}$ restoran franšize u kojem gost g rezerviše novi sto (direktno ili rekurzivnim pozivom), a da pritom on ne rezerviše novi sto u restoranu $\mathcal{R}_{\sigma(\mathbf{s})}$. Ukoliko gost g ne rezerviše novi sto, stavlja se da je $u_g = L$, gdje je L maksimalni nivo hijerarhije koji postoji u datoj franšizi. Slikovito rečeno, u_g predstavlja najmanji nivo hijerahije u kojem je gost g "krivac" za rezervisanje novog stola. Primjera radi, ako za nekog gosta g vrijedi $u_g = 0$, tada, u odnosu na restoran u koji je ušao, ovaj gost rezerviše nove stolove u svim restoranima roditeljima zaključno sa restoranom praroditeljem.

Ako je $IG(\mathbf{s})$ skup svih nezavisnih gostiju restorana $\mathcal{R}_{\mathbf{s}}$, $C(\mathbf{s})$ grana franšize čiji je korijen restoran $\mathcal{R}_{\mathbf{s}}$ (tj. $C(\mathbf{s})$ je skup koji obuhvata sve kontekste $\mathbf{s}'$ takve da je restoran $\mathcal{R}_{\mathbf{s}'}$ potomak restorana $\mathcal{R}_{\mathbf{s}}$) i z_g jelo koje konzumira gost g, tada vrijedi sljedeća lema.

Lema 3.4.10. *[22] Za svaki restoran $\mathcal{R}_s$ i svako jelo v koje se poslužuje u tom restoranu vrijedi*

$$c_{sv} = c^0_{sv} + \sum_{s' : s = \sigma(s')} t_{s'v}. \tag{3.23}$$

$$t_{sv} = \sum_{s' \in C(s)} \sum_{g \in IG(s')} I\left(z_g = v\right) \cdot I\left(u_g \leqslant d_s\right). \tag{3.24}$$

Dokaz. Svaki od gostiju restorana $\mathcal{R}_{\mathbf{s}}$ koji konzumiraju jelo v je ili "nezavisan" gost koji konzumira ovo jelo ili je je gost-sto na kojem je inicijalizovano jelo v, a koji je poslat od strane direktnog restorana potomka ovog restorana. Zbog

toga, vrijedi jednakost (3.23). Što se tiče dokaza jednakosti (3.24), ključno je primijetiti da, na broj stolova restorana $\mathcal{R}_{\mathbf{s}}$ koji poslužuju jelo v, utiču isključivo nezavisni gosti iz bilo kojeg restorana oblika $\mathcal{R}_{\mathbf{s}'}$, za $s' \in C(\mathbf{s})$, koji konzumiraju jelo v i, u "potrazi" za ovim jelom, moraju da se u hijerarhiji franšize "popnu" bar do restorana $\mathcal{R}_{\mathbf{s}}$. □

Jednakosti (3.23) i (3.24) iz prethodne leme omogućavaju da se, uz broj nezavisnih gostiju koji se dobija iz trening podataka, konfiguracija stolova u svakom restoranu franšize izrazi u funkciji doprinosa rezervaciji novih stolova za svakog takvog gosta.

Teorema 3.4.11. *[22] Za trening podatke $\mathcal{D}$, odgovarajući skup konteksta $\mathcal{D}^*$ i $\{z_1, \ldots, z_J\}$ uzorak izabran iz CRFP, neka je u_j indikator novog stola pridružen podatku z_j. Zajednička raspodjela za $(z_1, \ldots, z_J)$ i $(u_1, \ldots, u_J)$ u HPYP čija je reprezentacija ovaj CRFP data je sa*

$$P\{z_1, \ldots, z_J; u_1, \ldots, u_J\} = \frac{1}{n^{t_{\varepsilon\cdot}}} \cdot \prod_{s \in \mathcal{D}^*} \left(\frac{\left(\beta_{|s|}|\alpha_{|s|}\right)^{(t_{s\cdot})}}{\left(\beta_{|s|}\right)^{(c_{s\cdot})}} \cdot \prod_{v \in \mathbb{N}_n} \frac{S^{c_{sv}}_{t_{sv},\alpha_{|s|}}}{\binom{c_{sv}}{t_{sv}}} \right), \quad (3.25)$$

pri čemu je $t_{s\cdot} = \sum_{v \in \mathbb{N}_n} t_{sv}$ i $c_{s\cdot} = \sum_{v \in \mathbb{N}_n} c_{sv}$ i veličine t_{sv} i c_{sv} se dobijaju iz veličina $u_1, \ldots, u_J$ primjenom formula (3.23) i (3.24).

Dokaz. Ideja na kojoj se zasniva dokaz date formule zasniva se na pojedinačnom posmatranju krajnjih restorana potomaka i svih restorana franšize u pripadnim granama, zaključno sa restoranom praroditeljem $\mathcal{R}_{\varepsilon}$. Neka je $\mathcal{R}_{\mathbf{s}}$ jedan takav restoran, $\mathbf{z} \subseteq \{z_1, \ldots, z_J\}$ skup gostiju ovog restorana i $\mathbf{u} \subseteq \{u_1, \ldots, u_J\}$ skup indikatora novog stola ovih gostiju. Iz formula (3.23) i (3.24) direktno slijedi da elementi skupa $\mathbf{u}$ na jedinstven način određuju konfiguraciju stolova u restoranu $\mathcal{R}_{\mathbf{s}}$. S druge strane, ako je data konfiguracija stolova restorana $\mathcal{R}_{\mathbf{s}}$, tj. ako su za svako $v \in \mathbb{N}$ poznate vrijednosti $t_{\mathbf{s}v}, c_{\mathbf{s}v}$, tada se može generisati $\prod_{v \in \mathbb{N}_n} \binom{c_{\mathbf{s}v}}{t_{\mathbf{s}v}}$ mogućih indikatora novih stolova. Zbog toga, veza između zajedničke raspodjele za $\mathbf{z}$ i $\mathbf{t} := \{t_{\mathbf{s}v} : v \in \mathbb{N}_n\}$ i zajedničke raspodjele za $\mathbf{z}$ i $\mathbf{u}$ data je sa

$$P(\mathbf{z}, \mathbf{t}) = \prod_{v \in \mathbb{N}_n} \binom{c_{\mathbf{s}v}}{t_{\mathbf{s}v}} \cdot P(\mathbf{z}, \mathbf{u}).$$

Zajednička raspodjela za $\mathbf{z}$ i $\mathbf{t}$ nađena je u Lemi 3.4.8 za slučaj jednog kineskog restorana. Za razliku od tog slučaja, gdje je zauzimanje novog stola podrazumijevalo direktno uzorkovanje jela iz bazne raspodjele H, u slučaju franšize restorana, zauzimanje novog stola u restoranu $\mathcal{R}_{\mathbf{s}}$ za posljedicu ima uzorkovanje

iz raspodjele $G_{\sigma(s)}$. Uzimajući u obzir ovu modifikaciju formule (3.19), kao i oznake odgovarajućih statistika, dobija se:

$$P(\mathbf{z}, \mathbf{t}) = \frac{\left(\beta_{|\mathbf{s}|}|\alpha_{|\mathbf{s}|}\right)^{(t_{\mathbf{s}\cdot})}}{\left(\beta_{|\mathbf{s}|}\right)^{(c_{\mathbf{s}\cdot})}} \cdot \prod_{v \in \mathbb{N}_n} \left(G_{\sigma(\mathbf{s})}(v)^{t_{\mathbf{s}v}} \cdot S^{c_{\mathbf{s}v}}_{t_{\mathbf{s}v}, \alpha_{|\mathbf{s}|}} \right).$$

Marginalizacijom $G_{\sigma(\mathbf{s})}(v)$, ovaj postupak može da se ponovi u restoranu $\mathcal{R}_{\sigma(\mathbf{s})}$ i da se rekurzivno izvodi sve dok se ne dođe do restorana praroditelja $\mathcal{R}_\varepsilon$. Zauzimanje novog stola u restoranu praroditelju rezultuje uzorkovanjem jela iz globalne bazne raspodjele H. Kako je H uniformna raspodjela na skupu $\mathbb{N}_n$, vjerovatnoća izbora jela za zauzetim stolovima restorana praroditelja dobija se kada se broj $\frac{1}{n}$ stepenuje sa brojem novih stolova koje je u restoranu praroditelju generisala grana franšize čiji je krajnji potomak restoran $\mathcal{R}_{\mathbf{s}}$. Ako se ovaj postupak izvede tako da se tačno jednom obuhvate svi restorani iz franšize dobija se jednakost (3.25). □

Za postavljeni hijerarhijski Pitman-Jorov proces i datu familiju trening podataka $\mathcal{D}$ moguće je odrediti posteriorne raspodjele vjerovatnosnih vektora $G_{\mathbf{s}}$, parametara sniženja $\alpha_{|\mathbf{s}|}$ i parametara koncentracije $\beta_{|\mathbf{s}|}$. Određivanje posteriornih raspodjela biće izvedeno u okviru odgovarajuće CRFP konstrukcije. U narednom je preciznije opisan ovaj postupak.

Neka je, kao do sada, $\mathcal{D}$ skup trening podataka, $\mathcal{D}^*$ skup svih podstringova stringova iz familije $\mathcal{D}$ i $\{\mathcal{R}_{\mathbf{s}} : \mathbf{s} \in \mathcal{D}^*\}$ odgovarajuća franšiza kineskih restorana. Određenosti radi, neka posmatrana franšiza ima n nivoa hijerarhije, gdje je $n-1$ dužina najdužeg konteksta iz $\mathcal{D}^*$ koji određuje neprazni restoran bez potomaka. Podsjećanja radi, restorani bez potomaka su restorani franšize koji odgovaraju kontekstima koji su neprazni prefiksi stringova iz familije $\mathcal{D}$ i ovi restorani imaju samo nezavisne goste. Uzorak iz CRFP se dobija kada se "nasele" svi restorani ove franšize počevši od restorana koji nemaju potomaka i zaključno sa restoranom praroditeljem. Ako je J ukupan broj gostiju svih restorana franšize, sa $z^{\mathcal{D}} := (z_1^{\mathcal{D}}, z_2^{\mathcal{D}}, \ldots, z_J^{\mathcal{D}})$ biće označen realizovani uzorak iz CRFP.

U odnosu na uzorak $z^{\mathcal{D}}$, posteriorna raspodjela vektora $\mathcal{G} := \{G_{\mathbf{s}} : \mathbf{s} \in \mathcal{D}^*\}$ i vektora hiperparametara $\Theta := \{\alpha_m, \beta_m : 0 \leqslant m \leqslant n-1\}$ je uslovna raspodjela data sa:

$$P\left(\mathcal{G}, \Theta | z^{\mathcal{D}}\right) = \frac{P\left(\mathcal{G}, \Theta, z^{\mathcal{D}}\right)}{P\left(z^{\mathcal{D}}\right)}. \tag{3.26}$$

Upotrebom CRFP konstrukcije se dolazi do marginalizacije vektora $G_{\mathbf{s}}$ tako što se ovaj vektor zamjenjuje indikatorima novog stola gostiju restorana $\mathcal{R}_{\mathbf{s}}$. Neka je $\mathcal{U}_{\mathbf{s}}$ niz svih tako pridruženih indikatora novog stola i $\mathcal{U}$ niz indikatora novog stola svih gostiju franšize. Tada se posteriorna distribucija data u (3.26) može

izraziti na sljedeći način:

$$P\left(\mathcal{U}, \Theta | z^{\mathcal{D}}\right) = \frac{P\left(\mathcal{U}, \Theta, z^{\mathcal{D}}\right)}{P\left(z^{\mathcal{D}}\right)} \propto P\left(\mathcal{U}, \Theta, z^{\mathcal{D}}\right). \tag{3.27}$$

Za generisanje reprezentativnih uzoraka $\{\mathcal{U}^{(i)}, \Theta^{(i)} : i \in \{1, 2, \ldots, I\}\}$ iz posteriorne raspodjele kao početni osnov može biti upotrebljena zajednička raspodjela data formulom (3.25), jer je ova raspodjela zadata u zatvorenoj formi.

Najprije će biti opisana implementacija Gibsovog semplera za određivanje uzoraka $\{\mathcal{U}^{(i)} : i \in \{1, 2, \ldots, I\}\}$ iz posteriorne raspodjele, pri čemu je, za svako $i \in \{1, 2, \ldots, I\}$, $\mathcal{U}^{(i)}$ sastavljen od indikatora zauzimanja novog stola gostiju posmatrane franšize kineskih restorana. Ova implementacija se može naći u [22] i ona koristi zajedničku raspodjelu vjerovatnoća uzoraka iz CRFP i indikatora zauzimanja novog stola, koja je data u formuli (3.25). Tokom generisanja instanci, za svakog gosta g se "stara" vrijednost indikatora zauzimanja novog stola u_g uklanja zajedno sa "starom" vrijednošću z_g. Brisanje indikatora u_g odgovara brisanju novog stola na odgovarajućem nivou hijerarhije, što može izazvati probleme ukoliko se ovaj sto šalje u restorane roditelje. U takvim situacijama prelazi se na uklanjanje drugih stolova, sve do momenta kada je moguće ukloniti sve stolove koji su povezani sa posmatranim stolom. Preciznije, procedura uzorkovanja teče na sljedeći način:

Za svakog gosta g i restoran franšize u kojem ovaj gost doprinosi rezervisanju novog stola, posmatra se skup $path(g)$ koji predstavlja putanju od restorana praroditelja do tog restorana. Neka su $t'_{\mathbf{s}\cdot}, c'_{\mathbf{s}\cdot}, t'_{\mathbf{s}v}, c'_{\mathbf{s}v}$ vrijednosti odgovarajućih statistika nakon uklanjanja gosta g, a $t''_{\mathbf{s}\cdot}, c''_{\mathbf{s}\cdot}, t''_{\mathbf{s}v}, c''_{\mathbf{s}v}$ vrijednosti tih statistika nakon dodavanja istog tog gosta u restoran $\mathcal{R}_{\mathbf{s}}$. Da bi se uzorkovao par (z_g, u_g), potrebno je razmotriti sve restorane $\mathcal{R}_{\mathbf{s}} \in path(g)$, jer bi uklanjanje gosta g moglo da promijeni konfiguraciju stolova u tim restoranima. Svaki gost g pripada tačno jednom restoranu iz skupa $path(g)$, pa se za uvedene statistike

jednostavno dokazuju sljedeća svojstva:

Za svaki restoran $\mathcal{R}_{\mathbf{s}} \in path(g)$, uklanjanjem gosta g se dobija

$$t'_{\mathbf{s}\cdot} = \begin{cases} t_{\mathbf{s}\cdot}, & \text{ako je } u_g > d_{\mathbf{s}}; \\ t_{\mathbf{s}\cdot} - 1, & \text{ako je } u_g \leqslant d_{\mathbf{s}}, \end{cases}$$

$$t'_{\mathbf{s}v} = \begin{cases} t_{\mathbf{s}v}, & \text{ako je } u_g > d_{\mathbf{s}}; \\ t_{\mathbf{s}v} - 1, & \text{ako je } u_g \leqslant d_{\mathbf{s}} \text{ i } z_g = v, \end{cases}$$

$$c'_{\mathbf{s}\cdot} = \begin{cases} c_{\mathbf{s}\cdot}, & \text{ako } g \notin IG(\mathbf{s}) \text{ i } u_g > d_{\mathbf{s}} + 1; \\ c_{\mathbf{s}\cdot} - 1, & \text{ako } (g \notin IG(\mathbf{s}) \text{ i } u_g \leqslant d_{\mathbf{s}} + 1) \text{ ili } g \in IG(\mathbf{s}), \end{cases}$$

$$c'_{\mathbf{s}v} = \begin{cases} c_{\mathbf{s}v}, & \text{ako } g \notin IG(\mathbf{s}) \text{ i } u_g > d_{\mathbf{s}} + 1; \\ c_{\mathbf{s}v} - 1, & \text{ako } z_g = v \text{ i } ((g \notin IG(\mathbf{s}) \text{ i } u_g \leqslant d_{\mathbf{s}} + 1) \text{ ili } g \in IG(\mathbf{s})) \ . \end{cases}$$

Dodavanjem gosta g i indikatora u_g se dobija

$$t''_{\mathbf{s}\cdot} = \begin{cases} t'_{\mathbf{s}\cdot}, & \text{ako je } u_g > d_{\mathbf{s}}; \\ t'_{\mathbf{s}\cdot} + 1, & \text{ako je } u_g \leqslant d_{\mathbf{s}}, \end{cases}$$

$$t''_{\mathbf{s}v} = \begin{cases} t'_{\mathbf{s}v}, & \text{ako je } u_g > d_{\mathbf{s}}; \\ t'_{\mathbf{s}v} + 1, & \text{ako je } u_g \leqslant d_{\mathbf{s}} \text{ i } z_g = v, \end{cases}$$

$$c''_{\mathbf{s}\cdot} = \begin{cases} c'_{\mathbf{s}\cdot}, & \text{ako } g \notin IG(\mathbf{s}) \text{ i } u_g > d_{\mathbf{s}} + 1; \\ c'_{\mathbf{s}\cdot} + 1, & \text{ako } (g \notin IG(\mathbf{s}) \text{ i } u_g \leqslant d_{\mathbf{s}} + 1) \text{ ili } g \in IG(\mathbf{s}), \end{cases}$$

$$c''_{\mathbf{s}v} = \begin{cases} c'_{\mathbf{s}v}, & \text{ako } g \notin IG(\mathbf{s}) \text{ i } u_g > d_{\mathbf{s}} + 1; \\ c'_{\mathbf{s}v} + 1, & \text{ako } z_g = v \text{ i } ((g \notin IG(\mathbf{s}) \text{ i } u_g \leqslant d_{\mathbf{s}} + 1) \text{ ili } g \in IG(\mathbf{s})) \ . \end{cases}$$

Koristeći prethodno, zajednička uslovna vjerovatnoća za z_g i u_g se može izraziti u obliku

$$\begin{aligned} & P\left\{z_g, u_g | \mathbf{z}_{1:J} - z_g, \mathbf{u}_{1:J} - u_g\right\} \\ & = \prod_{\mathcal{R}_{\mathbf{s}} \in path(g)} \frac{\left(\beta_{|\mathbf{s}|} + \alpha_{|\mathbf{s}|} t'_{\mathbf{s}\cdot}\right)^{I\left(t''_{\mathbf{s}\cdot} \neq t'_{\mathbf{s}\cdot}\right)}}{\left(\beta_{|\mathbf{s}|} + c'_{\mathbf{s}\cdot}\right)^{I\left(c''_{\mathbf{s}\cdot} \neq c'_{\mathbf{s}\cdot}\right)}} \cdot \left(\frac{S^{c''_{\mathbf{s}v}}_{t''_{\mathbf{s}v}, \alpha_{|\mathbf{s}|}}}{S^{c'_{\mathbf{s}v}}_{t'_{\mathbf{s}v}, \alpha_{|\mathbf{s}|}}}\right)^{I\left(n''_{\mathbf{s}v} \neq n'_{\mathbf{s}v} \vee t''_{\mathbf{s}v} \neq t'_{\mathbf{s}v}\right)} \\ & \cdot \frac{\left(t''_{\mathbf{s}v}\right)^{I\left(t''_{\mathbf{s}v} \neq t'_{\mathbf{s}v}\right)} \cdot \left(c''_{\mathbf{s}v} - t''_{\mathbf{s}v}\right)^{I\left(c''_{\mathbf{s}v} - t''_{\mathbf{s}v} = c'_{\mathbf{s}v} - t'_{\mathbf{s}v}\right)}}{\left(c''_{\mathbf{s}v}\right)^{I\left(c''_{\mathbf{s}v} \neq c'_{\mathbf{s}v}\right)}}, \end{aligned}$$

gdje je $\mathbf{z}_{1:J} - z_g$ oznaka za vektor $\mathbf{z}_{1:J}$ iz kojeg je isključena komponenta z_g, a $\mathbf{u}_{1:J} - u_g$ oznaka za vektor $\mathbf{u}_{1:J}$ iz kojeg je isključena komponenta u_g.

Kao što je već napomenuto, uklanjanjem gosta g iz odgovarajućeg restorana nije uvijek moguće promijeniti vrijednosti relevantnih statistika. Ove statistike se mogu izmjeniti ukoliko vrijedi bar jedan od sljedećih uslova:

1. Gost g ne doprinosi rezervisanju novog stola, tj. $u_g = L$.

2. Nijedan drugi gost ne sjeda za sto koji je "generisao" gost g, tj. za sve $\mathcal{R}_s \in path(g)$ vrijedi $d_{\mathbf{s}} \geqslant u_g \Rightarrow c_{\mathbf{s}v} = 1$.

3. Postoje drugi stolovi generisani od strane drugih gostiju na kojima se servira isto jelo kao na stolovima generisanim od strane gosta g, tj. za sve $\mathcal{R}_s \in path(g)$ vrijedi $d_{\mathbf{s}} \geqslant u_g \Rightarrow t_{\mathbf{s}v} > 1$.

Kada se novi gost dodaje u restoran zajedno sa odgovarajućim indikatorom novog stola, tada taj indikator ne može uvijek uzimati bilo koju vrijednost iz skupa $\{0, \ldots, L\}$. Npr., ako se gost g dodaje u restoran $\mathcal{R}_{\mathbf{s}}$ tako da konzumira jelo v, pri čemu je trenutno $t_{\mathbf{s}v} = 0$ (u tom restoranu nema stolova koji poslužuju to jelo), gost g će generisati novi sto u restoranu $\mathcal{R}_{\mathbf{s}}$, što znači da vrijedi $u_g < L$. Zbog toga je potrebno ustanoviti minimalnu moguću vrijednost indikatora u_g (u oznaci $u_g^{\min}$) i maksimalnu moguću vrijednost indikatora u_g (u oznaci $u_g^{\max}$). Relativno jednostavno se pokazuje da za gosta g koji konzumira jelo v vrijedi

$$u_g^{\min} = \begin{cases} 0, & \text{ako je } t_{\varepsilon v} = 0; \\ 1, & \text{inače.} \end{cases}$$

$$u_g^{\max} = \begin{cases} \min\{d_{\mathbf{s}} : \mathcal{R}_{\mathbf{s}} \in path(g), t_{\mathbf{s}v} = 0\}, & \text{ako postoji } \mathcal{R}_{\mathbf{s}} \text{ tako da je } t_{\mathbf{s}v} = 0; \\ L, & \text{ako za svaki } \mathcal{R}_{\mathbf{s}} \text{ vrijedi } t_{\mathbf{s}v} > 0. \end{cases}$$

Za određivanje uzoraka $\{\Theta^{(i)} : i \in \{1, 2, \ldots, I\}\}$ parametara sniženja i koncentracije biće upotrebljen metod uzorkovanja koji ove parametre optimizuje u odnosu na njihove pretpostavljene priorne raspodjele. Ideja ovog metoda je preuzeta iz [88], gdje se koriste pomoćne slučajne promjenljive čije su uslovne raspodjele određene "starim" vrijednostima parametara α_m, β_m i čijim uzorkovanjem se "popravljaju" vrijednosti hiperparametara a_m, b_m, f_m, h_m. Zatim se uz pomoć ovako dobijenih korigovanih vrijednosti hiperparametara uzorkuju "nove" vrijednosti parametara α_m, β_m i ovaj postupak se iterativno ponavlja. Podsjećanja radi, pretpostavljena priorna raspodjela parametra sniženja α_m je beta $B(a_m, b_m)$, dok je pretpostavljena priorna raspodjela parametra koncentracije β_m gama $\Gamma(f_m, h_m)$ raspodjela, pa se iz ovih priornih raspodjela uzorkuju inicijalne vrijednosti parametara α_m, β_m. Slijedi detaljniji opis ovog postupka.

Uzimajući u obzir jednakost (3.20), zajednička raspodjela vjerovatnoća data u (3.25) može da se predstavi u sljedećem obliku:

$$P(z_1, \ldots, z_J; u_1, \ldots, u_J) = \frac{1}{n^{t_{\varepsilon\cdot}}} \cdot \prod_{\mathbf{s} \in \mathcal{D}^*} \left(\frac{\left(\beta_{|\mathbf{s}|} | \alpha_{|\mathbf{s}|}\right)^{(t_{\mathbf{s}\cdot})}}{\left(\beta_{|\mathbf{s}|}\right)^{(c_{\mathbf{s}\cdot})}} \cdot \prod_{v \in \mathbb{N}_n} \prod_{k=1}^{t_{\mathbf{s}v}} \left(1 - \alpha_{|\mathbf{s}|}\right)^{(c_{\mathbf{s}vk} - 1)} \right).$$

Pomoćne slučajne veličine uvode se u cilju aproksimacije odgovarajućih faktora iz prethodne jednakosti. Neka je $\mathcal{R}_{\mathbf{s}}$ proizvoljan restoran franšize koji nije prazan. Primjenom veze između beta i gama funkcije, dobija se

$$\frac{1}{\left(\beta_{|\mathbf{s}|}\right)^{(c_{\mathbf{s}\cdot})}} = \frac{\Gamma\left(\beta_{|\mathbf{s}|}\right)}{\Gamma\left(\beta_{|\mathbf{s}|} + c_{\mathbf{s}\cdot}\right)} = \frac{1}{\Gamma\left(c_{\mathbf{s}\cdot}\right)} \int_0^1 x_{\mathbf{s}}^{\beta_{|\mathbf{s}|} - 1} (1 - x_{\mathbf{s}})^{c_{\mathbf{s}\cdot} - 1} dx_{\mathbf{s}}.$$

Na osnovu ovoga, uvodi se pomoćna slučajna veličina $x_{\mathbf{s}}$, čija je uslovna raspodjela u odnosu na stare vrijednosti parametara $\beta_{|\mathbf{s}|}$ i $c_{\mathbf{s}\cdot}$ beta $B(\beta_{|\mathbf{s}|}, c_{\mathbf{s}\cdot})$ raspodjela. Takođe, vrijedi

$$\left(\beta_{|\mathbf{s}|}|\alpha_{|\mathbf{s}|}\right)^{(t_{\mathbf{s}\cdot})} = \prod_{i=0}^{t_{\mathbf{s}\cdot}-1}\left(\beta_{|\mathbf{s}|}+\alpha_{|\mathbf{s}|}\cdot i\right) = \prod_{i=0}^{t_{\mathbf{s}\cdot}-1}\sum_{y_{\mathbf{s}i}\in\{0,1\}} \beta_{|\mathbf{s}|}^{y_{\mathbf{s}i}}\cdot\left(\alpha_{|\mathbf{s}|}i\right)^{1-y_{\mathbf{s}i}},$$

pa se mogu uvesti pomoćne slučajne veličine $y_{\mathbf{s}i}, i \in \{0,1,\ldots,t_{\mathbf{s}\cdot}-1\}$, tako da je, za stare vrijednosti $\alpha_{|\mathbf{s}|}, \beta_{|\mathbf{s}|}$, uslovna raspodjela slučajne veličine $y_{\mathbf{s}i}$ Bernulijeva raspodjela sa parametrom $q := \dfrac{\beta_{|\mathbf{s}|}}{\beta_{|\mathbf{s}|}+\alpha_{|\mathbf{s}|}\cdot i}$, tj. $y_{\mathbf{s}i}$ uzima vrijednost 0 sa vjerovatnoćom $1-q$, a vrijednost 1 sa vjerovatnoćom q.

Konačno, za $c_{\mathbf{s}vk} \geqslant 2$ vrijedi

$$\left(1-\alpha_{|\mathbf{s}|}\right)^{(c_{\mathbf{s}vk}-1)} = \prod_{j=1}^{c_{\mathbf{s}vk}-1}\left(j-\alpha_{|\mathbf{s}|}\right) = \prod_{j=1}^{c_{\mathbf{s}vk}-1}\sum_{z_{\mathbf{s}vkj}\in\{0,1\}} (j-1)^{z_{\mathbf{s}vkj}}\left(1-\alpha_{|\mathbf{s}|}\right)^{1-z_{\mathbf{s}vkj}},$$

pa se mogu uvesti pomoćne slučajne veličine $z_{\mathbf{s}vkj}, j \in \{0,1,\ldots,c_{\mathbf{s}vk}-1\}$, tako da je, za staru vrijednost $\alpha_{|\mathbf{s}|}$, uslovna raspodjela slučajne veličine $z_{\mathbf{s}vkj}$ Bernulijeva raspodjela sa parametrom $r := \dfrac{j-1}{j-\alpha_{|\mathbf{s}|}}$.

Ako se sve pomoćne slučajne promjenljive uzorkuju iz odgovarajućih raspodjela, tada su, u odnosu na dobijene vrijednosti, uslovne raspodjele novih vrijednosti parametara sniženja i koncentracije date sa

$$\alpha_m \sim B\left(a_m + \sum_{\mathbf{s}:|\mathbf{s}|=m}\sum_{i=0}^{t_{\mathbf{s}\cdot}-1}\left(1-y_{\mathbf{s}i}\right), b_m + \sum_{\mathbf{s},v,k:|\mathbf{s}|=m,c_{\mathbf{s}vk}\geqslant 2}\sum_{j=1}^{c_{\mathbf{s}vk}-1}\left(1-z_{\mathbf{s}vkj}\right)\right),$$

$$\beta_m \sim \Gamma\left(f_m + \sum_{\mathbf{s}:|\mathbf{s}|=m}\sum_{i=0}^{t_{\mathbf{s}\cdot}-1} y_{\mathbf{s}i}, h_m - \sum_{\mathbf{s}:|\mathbf{s}|=m}\log x_{\mathbf{s}}\right).$$

Uzorkovanjem vrijednosti iz prethodnih raspodjela dobijaju se nove vrijednosti parametara α_m i β_m. Postupak uzorkovanja se dalje može nastaviti uz pomoć ovih vrijednosti.

Od vrijednosti koji se mogu dobiti iz posteriorne raspodjele (3.27) svakako je najvažnija predviđajuća vjerovatnoća pojavljivanja simbola $v \in \mathbb{N}_n$ nakon ostvarivanja konteksta $\mathbf{s}$, za proizvoljan kontekst $\mathbf{s}$ koji ne mora nužno biti element skupa $\mathcal{D}^*$. Ova vjerovatnoća je data sa

$$P_{HPYP(\mathcal{D})}\left(v|\mathbf{s}, z^{\mathcal{D}}\right) = \int P\left(v|\mathbf{s},\mathcal{U},\Theta\right)\cdot P\left(\mathcal{U},\Theta|z^{\mathcal{D}}\right) d\left(\mathcal{U},\Theta\right).$$

Umjesto računanja posljednjeg integrala, iz posteriorne raspodjele se izvlače uzorci $\{\mathcal{U}^{(i)}, \Theta^{(i)} : i \in \{1, 2, \ldots, I\}\}$ i koristi se sljedeća aproksimacija

$$P_{HPYP(\mathcal{D})}\left(v|\mathbf{s}, z^{\mathcal{D}}\right) \approx \sum_{i=1}^{I} P\left(v|\mathbf{s}, \mathcal{U}^{(i)}, \Theta^{(i)}\right), \tag{3.28}$$

pri čemu se uslovna vjerovatnoća $P\left(v|\mathbf{s}, \mathcal{U}, \Theta\right)$ računa rekurzivnom primjenom formule (3.15) na sljedeći način:

$$P\left(v|\varepsilon, \mathcal{U}, \Theta\right) = \frac{1}{n}, \tag{3.29}$$

$$P\left(v|\mathbf{s}, \mathcal{U}, \Theta\right) = \frac{c_{\mathbf{s}v} - \alpha_{|\mathbf{s}|} \cdot t_{\mathbf{s}v}}{\beta_{|\mathbf{s}|} + c_{\mathbf{s}\cdot}} + \frac{\beta_{|\mathbf{s}|} + \alpha_{|\mathbf{s}|} \cdot t_{\mathbf{s}\cdot}}{\beta_{|\mathbf{s}|} + c_{\mathbf{s}\cdot}} \cdot P\left(v|\sigma(\mathbf{s}), \mathcal{U}, \Theta\right), \tag{3.30}$$

sa napomenom da se za $\mathbf{s} \notin \mathcal{D}^*$, umjesto konteksta $\mathbf{s}$ koristi njegov najduži sufiks koji pripada skupu $\mathcal{D}^*$. Način biranja uzoraka $\{\mathcal{U}^{(i)}, \Theta^{(i)} : i \in \{1, 2, \ldots, I\}\}$ iz posteriorne raspodjele je opisan u prethodnom razmatranju.

Predviđajuća vjerovatnoća $P_{HPYP(\mathcal{D})}$ je krajnji cilj kompletnog prethodnog razmatranja. Ovom vjerovatnoćom se, na osnovu trening podataka $\mathcal{D}$, može definisati ocjena $\widehat{P}_{HPYP(\mathcal{D})}$ vjerovatnoće realizacije proizvoljnog stringa $x_{[1,T]} = x_1 x_2 \ldots x_T$ iz formule (3.21) na sljedeći način:

$$\widehat{P}_{HPYP(\mathcal{D})}(x_{[1,T]}) := \prod_{i=1}^{T} P_{HPYP(\mathcal{D})}\left(x_i|x_{[1,i-1]}, z^{\mathcal{D}}\right), \tag{3.31}$$

sa dodatnom konvencijom da je $x_{[1,0]} = \varepsilon$ i $x_1|\varepsilon = x_1$.

Ovako dobijena ocjena vjerovatnoća omogućava definisanje mjere sličnosti familija stringova A i B definisanih nad istim alfabetom $\mathbb{N}_n$ na sličan način kao u slučaju modela vjerovatnosnog sufiksnog drveta. Preciznije, ako se familije A i B koriste kao skupovi trening podataka i dobiju predviđajuće vjerovatnoće $P_{HPYP(A)}$ i $P_{HPYP(B)}$, a zatim na osnovu njih, primjenom formule (3.31), izvedu ocjene $\widehat{P}_{HPYP(A)}$ i $\widehat{P}_{HPYP(B)}$, tada se na principu Kulbak-Lajblerove mjere divergencije može konstruisati nova (vjerovatnosna) mjera sličnosti familija stringova A i B:

$$d^{new}_{HPYP}(A, B) := \frac{\hat{d}(\hat{P}^{\mathcal{M}_A}_{HPYP(A)}, \hat{P}^{\mathcal{M}_B}_{HPYP(B)}) + \hat{d}(\hat{P}^{\mathcal{M}_B}_{HPYP(B)}, \hat{P}^{\mathcal{M}_A}_{HPYP(A)})}{2},$$

pri čemu se statistike $\hat{d}(\hat{P}^{\mathcal{M}_A}_{HPYP(A)})$ i $\hat{d}(\hat{P}^{\mathcal{M}_B}_{HPYP(B)})$ definišu kao u formuli (3.5), gdje ulogu vjerovatnosnih mjera $P^{\mathcal{M}_A}$ i $P^{\mathcal{M}_B}$ uzimaju redom $\hat{P}^{\mathcal{M}_A}_{HPYP(A)}$ i $\hat{P}^{\mathcal{M}_B}_{HPYP(B)}$.

HPYP model na osnovu kojeg se računaju predviđajuće vjerovatnoće zasnovan je na strukturi drveta, što može otežati njegovu implementaciju. Stoga

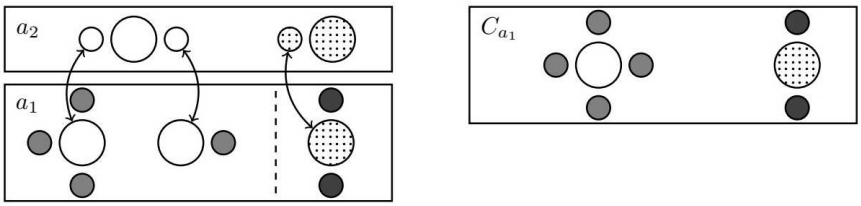
su moguće razne strategije kojima bi se pojednostavila struktura ovog drveta. Ovdje će biti razmatrana marginalizacija HPYP modela kojom će se ukloniti "unutrašnji" čvorovi drveta u kojima ne dolazi do grananja drveta, da bi se zatim neki od tih čvorova reinicijalizovali ako postoji potreba za njima u postupku nalaženja odgovarajuće predviđajuće vjerovatnoće. Na taj način se, u odnosu na polazni model, dobija model prostije strukture, a iste izražajne moći. U terminima CRFP analogije, ova marginalizacija podrazumijeva da se iz franšize najprije uklone svi restorani koji imaju tačno jednog restorana-potomka. Zatim se, na osnovu ostatka franšize formira redukovana hijerahija na osnovu koje se nalaze predviđajuće vjerovatnoće ostvarivanja datog simbola u datom kontekstu. Potreba za naknadnim dodavanjem nekog od uklonjenih restorana nastupa u slučaju kada je neki od njih označen zadanim kontekstom. Ključno u vezi prethodnog postupka je utvrđivanje HPYP strukture redukovanog modela, tj. izražavanje zavisnosti parametara sniženja i koncentracije ovog modela na svim njegovim nivoima hijerarhije u odnosu na odgovarajuće parametre polaznog modela i to u obliku koji omogućava rekonstrukciju parametara oba modela prilikom opisanog uklanjanja ili dodavanja čvorova drveta (tj. restorana iz franšize). Slijedi precizniji opis ove zavisnosti.

Ako je u kineskom restoranu koji je interpretacija Pitman-Jorovog procesa $\mathrm{PYP}(\alpha, \beta, H)$, $\alpha \in [0, 1)$, $\beta > -\alpha$, sa $A(c)$ označen skup svih rasporeda sjedenja c gostiju, tada se, za raspored sjedenja $a \in A(c)$, rekurzivnom primjenom formule (3.15) dobija

$$P(a) = \frac{(\beta + \alpha|\alpha)^{(|t(a)|-1)}}{(\beta + 1)^{(c-1)}} \cdot \prod_{b \in t(a)} (1 - \alpha)^{(c_b - 1)}, \tag{3.32}$$

gdje je $t(a)$ skup svih stolova koji se koriste u rasporedu sjedenja a, a c_b broj gostiju koji sjedi za stolom $b \in t(a)$. Ovim je određena distribucija $CRP_c(\alpha, \beta)$ raspoređivanja c gostiju ovog kineskog restorana i $a \sim CRP_c(\alpha, \beta)$, ako se vjerovatnoća realizacije rasporeda sjedenja a dobija na osnovu prethodne vjerovatnoće. Za $c \geqslant 1$, neka su data dva kineska restorana, tako da je u prvom dat raspored sjedenja $a_1 \in A(c)$, a u drugom dat raspored $a_2 \in A\left(|t(a_1)|\right)$ (Slika 3.6).

Dakle, broj gostiju u rasporedu sjedenja a_2 jednak je broju stolova u rasporedu sjedenja a_1, pa između ova dva skupa postoji bijektivna korespondencija. Koristeći ovo, moguće je na osnovu rasporeda sjedenja a_1 formirati novi raspored sjedenja $C_{a_1} \in A(c)$ u kojem se spajaju svi stolovi u rasporedu sjedenja a_1 koji odgovaraju gostima koji u rasporedu sjedenja a_2 sjede za istim stolom. Za novodobijeni raspored sjedenja C_{a_1} se još kaže da je dobijen *spajanjem* ili *koagulacijom* stolova u rasporedu a_1 saglasno sa rasporedom sjedenja a_2. Obrnuta operacija od ove je *razdvajanje* ili *fragmentacija* stolova. Ova operacija podrazumijeva particiju stolova u rasporedu sjedenja a_1 na sekcije restorana,



Slika 3.6

tako da jednu sekciju sačinjavaju svi stolovi koji odgovaraju gostima koji sjede za istim stolom u rasporedu sjedenja a_2. Jasno je da se ova particija može indeksirati skupom $t(a_2)$, tj. predstaviti u obliku $\{F_b : b \in t(a_2)\}$. Kako rasporedi sjedenja a_2 i C_{a_1} imaju isti broj stolova, posljednji skup se također može indeksirati skupom $t\left(C_{a_1}\right)$ i, za svako $b \in t\left(C_{a_1}\right)$, za sekciju F_b se može smatrati da potiče iz odgovarajuće CRP_{c_b} raspodjele, gdje je c_b broj gostiju za stolom b.

Iz prethodnih konstrukcija se jednostavno uočava da se poznavanjem rezultata operacija koagulacije i fragmentacije stolova mogu rekonstruisati rasporedi sjedenja od kojih su oni proistekli. Ovo omogućava da se uspostavi veza između odgovarajućih CRP distribucija. Uz prethodne oznake, vrijedi sljedeća teorema.

Teorema 3.4.12. *[74] Neka je*
(i) $a_1 \sim CRP_c\,(\alpha_1, \beta\alpha_1)$ *i* $a_2|a_1 \sim CRP_{|t(a_1)|}\,(\alpha_2, \beta)$,
(ii) Za svako $b \in t\left(C_{a_1}\right)$ *vrijedi*
$C_{a_1} \sim CRP_c\,(\alpha_1\alpha_2, \beta\alpha_1)$ *i* $F_b|C_{a_1} \sim CRP_{c_b}\,(\alpha_1, -\alpha_1\alpha_2)$.
Tada su distribucije date uslovima (i) *i* (ii) *ekvivalentne.*

Dokaz. Za dokaz ekvivalentnosti distribucija datih uslovima (i) i (ii) ključno je pokazati jednakost odgovarajućih zajedničkih raspodjela vjerovatnoća. Neka je ispunjen uslov (i), tj. neka su a_1 i a_2 rasporedi sjedenja za koje vrijedi $a_1 \sim CRP_c\,(\alpha_1, \beta\alpha_1)$ i $a_2|a_1 \sim CRP_{|t(a_1)|}\,(\alpha_2, \beta)$. Na osnovu formule (3.32), dobija se

$$P(a_1, a_2) = P(a_1) \cdot P\,(a_2|a_1) = \frac{(\beta\alpha_1 + \alpha_1|\alpha_1)^{(|t(a_1)|-1)}}{(\beta\alpha_1 + 1)^{(c-1)}} \cdot \prod_{d \in t(a_1)} (1-\alpha_1)^{(c_d-1)}$$
$$\cdot \frac{(\beta + \alpha_2|\alpha_2)^{(|t(a_2)|-1)}}{(\beta+1)^{(|t(a_1)|-1)}} \cdot \prod_{b \in t(a_2)} (1-\alpha_2)^{(c_b-1)} = \frac{(\beta\alpha_1 + \alpha_1\alpha_2|\alpha_1\alpha_2)^{(|t(a_2)|-1)}}{(\beta\alpha_1+1)^{(c-1)}}$$
$$\cdot \prod_{d \in t(a_1)} (1-\alpha_1)^{(c_d-1)} \cdot \prod_{b \in t(a_2)} (\alpha_1 - \alpha_1\alpha_2|\alpha_1)^{(c_b-1)},$$

pri čemu se posljednja jednakost dobija primjenom identiteta

$$(xy + y|y)^{(n-1)} = y^{n-1} \cdot (x+1)^{(n-1)}\,,\ x, y \in \mathbb{R}, n \geqslant 1,$$

zajedno sa činjenicom da vrijedi $|t(a_2)| \leqslant |t(a_1)|$ i $\sum_{b \in t(a_2)} c_b = |t(a_1)|$. Faktori dobijenog proizvoda mogu da se pregrupišu u skladu sa koagulacijom C_{a_1} i fragmentacijom $\{F_b\}$. Naime, stolu $b \in t\left(C_{a_1}\right)$ odgovara faktor $(\alpha_1 - \alpha_1\alpha_2|\alpha_1)^{(c_b-1)}$, uz koji se mogu "vezati" svi faktori iz proizvoda $\prod_{d \in t(a_1)} (1-\alpha_1)^{(c_d-1)}$ određeni stolovima $d \in t\left(F_b\right)$. Dakle, zajednička raspodjela vjerovatnoća za a_1 i a_2 se, u terminima koagulacije i fragmentacije, može izraziti na sljedeći način:

$$P(a_1, a_2) = \frac{(\beta\alpha_1 + \alpha_1\alpha_2|\alpha_1\alpha_2)^{\left(|t(C_{a_1})|-1\right)}}{(\beta\alpha_1+1)^{(c-1)}}$$

$$\cdot \prod_{b \in t\left(C_{a_1}\right)} \left((\alpha_1 - \alpha_1\alpha_2|\alpha_1)^{(|t(F_b)|-1)} \cdot \prod_{d \in t(F_b)} (1-\alpha_1)^{(c_d-1)} \right) = P\left(C_{a_1}, \{F_b\}\right).$$

Ako se u prethodnom izvrši uslovljavanje po C_{a_1}, dobija se da, za svako $b \in t(C_{a_1})$, raspored sjedenja F_b ima $CRP_{c_b}\,(\alpha_1, -\alpha_1\alpha_2)$ distribuciju. Sumiranjem po svim "blokovima" F_b, dobija se

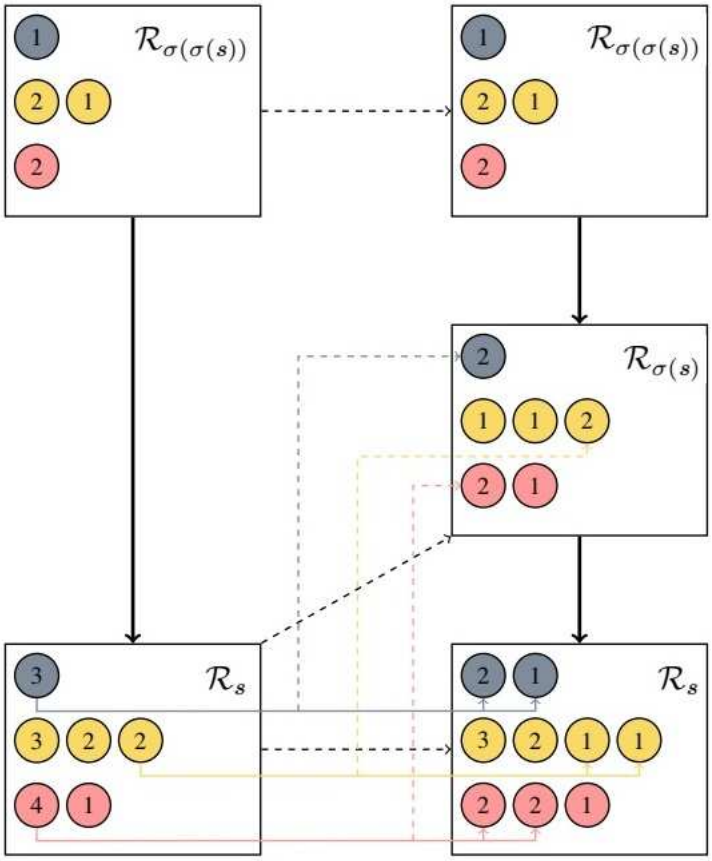
$$P(C_{a_1}) = \frac{(\beta\alpha_1 + \alpha_1\alpha_2|\alpha_1\alpha_2)^{(|t(C_{a_1})|-1)}}{(\beta\alpha_1+1)^{(c-1)}} \cdot \prod_{b \in t(C_{a_1})} (1-\alpha_1\alpha_2)^{(c_b-1)},$$

što znači da C_{a_1} ima $CRP_c\,(\alpha_1\alpha_2, \beta\alpha_1)$ raspodjelu. Time je dokazano da distribucije date uslovom (i) impliciraju distribucije date uslovom (ii). Da vrijedi i obrnuto dokazuje se na sličan način, eksploatisanjem uspostavljene jednakosti između odgovarajućih zajedničkih raspodjela. □

Prethodna teorema ima primjenu u određenim HPYP modelima. Naime, ako je za granu u kojoj se sukcesivno pojavljuju čvorovi G_0 i G_1 i G_2, pri čemu G_1 nema drugih potomaka sem G_2, ispunjeno $G_1|G_0 \sim PYP(\alpha_1, \beta, G_0)$ i $G_2|G_1 \sim PYP(\alpha_2, \beta\alpha_2, G_1)$, tada vrijedi $G_2|G_0 \sim PYP(\alpha_1\alpha_2, \beta\alpha_2, G_0)$. Ovo omogućava da nakon brisanja čvora G_1 ostanu poznati parametri sniženja i koncentracije tako redukovanog HPYP modela. Nedostatak ovog postupka je što za njegovu primjenu odgovarajući parametri sniženja i koncentracije moraju da se "poklope". Stoga je potrebno uvesti dodatna ograničenja za ove parametre kako bi se ovo postiglo. Npr. lako se vidi da se poklapanja ostvaruju ako su svi parametri koncentracije jednaki nuli. Značajno je istaći da prethodna teorema omogućava da se ovaj postupak i obrne, tj. da se izbrisani čvor reinicijalizuje, ako za njim postoji potreba. Sljedeći primjer ilustruje kako se ova reinicijalizacija obavlja u CRFP reprezentaciji.

Primjer 3.4.13. *Neka su $\mathcal{R}_{\sigma(\sigma(\boldsymbol{s}))}$, $\mathcal{R}_{\sigma(\boldsymbol{s})}$ i $\mathcal{R}_{\boldsymbol{s}}$ restorani u okviru jedne grane franšize kineskih restorana, pri čemu restoran $\mathcal{R}_{\sigma(\boldsymbol{s})}$ nema drugih potomaka sem restorana $\mathcal{R}_{\boldsymbol{s}}$. Takođe, neka je broj gostiju u restoranu $\mathcal{R}_{\sigma(\sigma(\boldsymbol{s}))}$ jednak broju stolova u restoranu $\mathcal{R}_{\boldsymbol{s}}$. To omogućava brisanje restorana $\mathcal{R}_{\sigma(\boldsymbol{s})}$ iz franšize (lijeva kolona Slike 3.7), ali i njegovu reinicijalizaciju (desna kolona Slike 3.7). Naime, ukoliko se u postupku računanja predviđajućih vjerovatnoća stvori potreba za poznavanjem konteksta $\sigma(\boldsymbol{s})$, tada je restoran $\mathcal{R}_{\sigma(\boldsymbol{s})}$ moguće reinicijalizovati u franšizu. Da bi se ovaj restoran "vratio" u franšizu, gosti za određenim stolovima restorana $\mathcal{R}_{\boldsymbol{s}}$ se razvrstavaju na više novih stolova prateći odgovarajuću CRP distribuciju. Svi novodobijeni stolovi se tretiraju kao gosti "novog" restorana $\mathcal{R}_{\sigma(\boldsymbol{s})}$. Oni se u ovom restoranu grupišu za stolove saglasno sa stolovima "originalnog" restorana $\mathcal{R}_{\boldsymbol{s}}$ na način da taj restoran i restoran $\mathcal{R}_{\sigma(\boldsymbol{s})}$ imaju jednak broj stolova, a da pritom broj gostiju u određenoj sekciji restorana $\mathcal{R}_{\sigma(\sigma(\boldsymbol{s}))}$ bude jednak broju stolova u odgovarajućoj sekciji restorana $\mathcal{R}_{\sigma(\boldsymbol{s})}$.*

Primjedba 3.4.14. U radu [88] primjećeno je da prethodno opisan HPYP model može da se poveže sa interpolisanim KN modelom koji je razmatran u prethodnoj sekciji ove glave. U suštini, oba ova modela kombinuju interpolacijski mehanizam zasnovan na parametrima sniženja i modifikovanim ocjenama iz prethodnog nivoa hijerarhije. Posljedično, oba modela daju slične distribucije predviđajućih vjerovatnoća. Odatle i proističe pomenuta efikasnost interpolacijskog KN modela, jer, u principu, svaka nova modifikacija kojom se poboljšava njegova efikasnost ide u pravcu da on sve manje liči na frekvencionistički model, a sve više podsjeća na hijerarhijski Bejzovski model.



Slika 3.7: Šematski prikaz reinicijalizacije restorana u franšizu. Pravougaonici predstavljaju restorane; krugovi predstavljaju stolove, a brojevi unutar krugova predstavljaju broj gostiju za odgovarajućim stolom. Stolovi u različitim sekcijama su označeni različitim bojama

Glava 4

Primjena pri rješavanju LCS problema

U ovoj glavi razmatran je problem nalaženja najdužeg zajedničkog podniza ili LCS problem (skraćenica LCS je akronim od Longest Common Subsequence). Ovo je istaknuti NP−težak optimizacijski problem gdje je, za datu konačnu familiju stringova, cilj naći najduži podniz (koji nije obavezno podstring!) zajednički za sve stringove iz posmatranog skupa. Pored teoretskog aspekta ovog problema, ova glava sadrži pregled postojećih pristupa i tehnika za približno rješavanje LCS problema koji se mogu naći npr. u [13], [14], [19], [32], [33], [34], [39], [50], [60], [67], [72], [87], [91], [92], [94], [95] i [98]. Svi obuhvaćeni rezultati iz literature rješavaju LCS problem koristeći instance uzorkovane iz uniformne raspodjele ili skoro uniformne raspodjele. Kako raspodjela vjerovatnoća simbola alfabeta od kojih se formiraju stringovi ne mora biti bliska uniformnoj, postavlja se pitanje efikasnosti modela iz literature za ovakav tip instanci. U ovoj tezi razmatrana je nova heuristika originalno predložena u radu [69]. Ispostavlja se da vremenski ograničena pretraga bima vođena ovom heuristikom značajno nadmašuje postojeće heurističke funkcije usmjeravanja pretrage iz literature, kada su u pitanju instance stringova sastavljenih od simbola čije relativne frekvencije učestalosti pojavljivanja nisu bliske uniformnoj raspodjeli.

4.1 Teoretski aspekti LCS problema

Najprije slijedi kratka rekapitulacija pojmova i notacije u vezi stringova iz Glave 2, uz dodavanje nekih novih definicija koji će omogućiti konciznije izražavanje rezultata navedenih u ovoj glavi.

String je konačan niz elemenata (simbola) izabranih iz konačnog, nepraznog skupa $\mathbb{N}_n := \{1, 2, \ldots, n\}, n \geqslant 2$, koji se naziva alfabetom. Dužina stringa s označava se sa $len(s)$ i predstavlja broj simbola u datom stringu. String dužine 0 naziva se praznim stringom i označava sa ε. Za element a_i nepraznog stringa

$s = (a_i, i \in \{1, \ldots, len(s)\})$ se kaže da se nalazi na $i-$toj poziciji posmatranog stringa iređanjem simbola na svim pozicijama, string s se može zapisati u obliku $s = a_1 a_2 \ldots a_{len(s)}$. Skup svih stringova dužine $l \geqslant 1$ čiji simboli pripadaju alfabetu $\mathbb{N}_n$ biće označen sa $S(n, l)$. String $t = t_1 t_2 \ldots t_{len(t)}$ je *podniz stringa* $s = s_1 s_2 \ldots s_{len(s)}$, ako postoje pozicije $1 \leqslant i_1 < i_2 < \ldots < i_{len(t)} \leqslant len(s)$ stringa s, tako da vrijedi $t_j = s_{i_j}$, za svako $j \in \{1, 2, \ldots, len(t)\}$; u tom slučaju, to se zapisuje sa $t \prec s$. *Podstring* stringa s je string oblika $s_{[i,j]}$ koji obuhvata sve simbole koji se nalaze između $i-$te i j-te pozicije stringa s (uključujući te pozicije), pri čemu se smatra da je $s_{[i,j]} := \varepsilon$, ukoliko je $i > j$. String dužine $len(s) + len(t)$ koji se dobije nadovezivanjem stringa t na string s predstavlja *konkatenaciju* ovih stringova i označava se sa st.

Za simbol $a \in \mathbb{N}_n$ i string s, neka je $s(a)$ skup svih pozicija stringa s na kojima se pojavljuje simbol a i $|s(a)|$ broj pojavljivanja simbola a u stringu s. Uopšteno, za neprazan skup $K \subseteq \mathbb{N}_n$ i string s, neka je $s(K)$ skup svih pozicija stringa s na kojima se pojavljuje neki od simbola iz skupa K. Očigledno da vrijedi $s(K) = \bigcup_{a \in K} s(a)$ i $|s(K)| = \sum_{a \in K} |s(a)|$. Za familiju stringova $S = \{s_1, s_2, \ldots, s_m\}, m \geqslant 1$, $m-$dimenzionalni vektor $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m) \in \mathbb{N}^m$ sastavljen od cjelobrojnih komponenti se naziva *vektorom pozicija*, a skup $S[\boldsymbol{\theta}] := \left\{s_{1[\theta_1, len(s_1)]}, s_{2[\theta_2, len(s_2)]}, \ldots, s_{m[\theta_m, len(s_m)]}\right\}$ se naziva *skupom sufiksnih stringova* skupa S. Ukoliko se simbol a pojavljuje u stringu $s_{i[\theta_i, len(s_i)]}$, tj. vrijedi $s_{i[\theta_i, len(s_i)]}(a) \neq \emptyset$, tada će pozicija njegovog prvog pojavljivanja u ovom stringu biti notirana sa $\theta_{i,a}$; ukratko, u tom slučaju je $\theta_{i,a} = \min s_{i[\theta_i, len(s_i)]}(a)$.

Neka je $S = \{s_1, s_2, \ldots, s_m\}, m \geqslant 1$, skup stringova definisanih nad istim alfabetom $\mathbb{N}_n$. Stringovi iz skupa S ne moraju biti jednakih dužina i podrazumijevaće se da je l dužina najdužeg stringa iz S, a $l_{\min}$ dužina najkraćeg stringa iz S. Ako je $\mathrm{PN}(S) := \{t : t \prec s, \text{ za svako } s \in S\}$ skup zajedničkih podnizova stringova iz skupa S, tada je $\mathrm{LCS}(S) = \mathrm{LCS}(s_1, s_2, \ldots, s_m) := \max\{len(t) : t \in \mathrm{PN}(S)\}$ dužina najdužeg zajedničkog podniza stringova iz skupa S.

Primjer 4.1.1. *Za* $s, t \in S(4, 17)$*, pri čemu je* $s = 14423211422233141$ *i* $t = 41113334212114121$*, stringovi* 1321211 *i* 4113341 *su zajednički podnizovi stringova s i t. Takođe, može se provjeriti da su stringovi* 411422141 *i* 142211421 *najduži podnizovi stringova s i t, što znači da je* $\mathrm{LCS}(s, t) = 9$.

U sljedećoj lemi iskazana su neka invarijantna svojstva dužine najdužeg zajedničkog podniza.

Lema 4.1.2. *[28] Za proizvoljne stringove* $s = s_1 s_2 \ldots s_{len(s)}, t = t_1 t_2 \ldots t_{len(t)}$ *nad alfabetom* $\mathbb{N}_n$ *vrijedi*

(i) $\mathrm{LCS}(s, t) = \mathrm{LCS}(t, s)$,

(ii) *Ako je* θ *permutacija skupa* $\mathbb{N}_n$ *i* $\theta(s) := \theta(s_1)\theta(s_2)\ldots\theta(s_{len(s)})$ *string dobijen iz s primjenom te permutacije, tada je* $\mathrm{LCS}(\theta(s), \theta(t)) = \mathrm{LCS}(s, t)$,

(*iii*) *Ako je* s^R *string koji se dobija iz stringa* s *"okretanjem" pozicija njegovih simbola, tj. ako je* $s^R = s_{len(s)} \dots s_2 s_1$*, tada je* $\mathrm{LCS}(s^R, t^R) = \mathrm{LCS}(s, t)$.

Dokaz. (*i*) Svojstvo da je neki string zajednički podniz stringova s i t je u svojoj prirodi simetričnog tipa, odakle slijedi data jednakost.

(*ii*) Ako je $u = u_1 u_2 \dots u_{len(u)}$ zajednički podniz stringova s i t, tada je $\theta(u) = \theta(u_1)\theta(u_2)\dots\theta(u_{len(u)})$ zajednički podniz stringova $\theta(s)$ i $\theta(t)$. Ukoliko je dodatno u najduži zajednički podniz stringova s i t, tada je posljedično $\theta(u)$ najduži zajednički podniz stringova $\theta(s)$ i $\theta(t)$, što implicira $\mathrm{LCS}(\theta(s), \theta(t)) = len(u) = \mathrm{LCS}(s, t)$.

(*iii*) Ako je $u = u_1 u_2 \dots u_{len(u)}$ zajednički podniz stringova s i t, tada je $u^R = u_{len(u)} \dots u_2 u_1$ zajednički podniz stringova s^R i t^R. Ukoliko je dodatno u najduži zajednički podniz stringova s i t, tada je posljedično u^R najduži zajednički podniz stringova s^R i t^R, što implicira $\mathrm{LCS}(s^R, t^R) = len(u) = \mathrm{LCS}(s, t)$. □

Dužina najdužeg zajedničkog podniza ima svojstvo monotonosti kada su u pitanju podnizovi i konkatenacija stringova. Preciznije, vrijedi sljedeća lema.

Lema 4.1.3. *[28] Za proizvoljne stringove* s, t, u, v *nad alfabetom* $\mathbb{N}_n$ *vrijedi*
(*i*) *Ako je* $s \prec u$ *i* $t \prec v$*, tada je* $\mathrm{LCS}(s, t) \leqslant \mathrm{LCS}(u, v)$,
(*ii*) $\mathrm{LCS}(s, t) + \mathrm{LCS}(u, v) \leqslant \mathrm{LCS}(su, tv)$.

Dokaz. (*i*) Svaki zajednički podniz stringova s i t je ujedno i zajednički podniz stringova u i v. To vrijedi i za najduži zajednički podniz stringova s i t, odakle slijedi $\mathrm{LCS}(s, t) \leqslant \mathrm{LCS}(u, v)$.

(*ii*) Ako je w proizvoljan zajednički podniz stringova s i t, a z proizvoljan zajednički podniz stringova u i v, tada je wz zajednički podniz stringova su i tv. Specijalno, konkatenacijom najdužeg zajedničkog podniza stringova s i t sa najdužim zajedničkim podnizom stringova u i v dobija se zajednički podniz stringova su i tv. Zbog toga, vrijedi $\mathrm{LCS}(s, t) + \mathrm{LCS}(u, v) \leqslant \mathrm{LCS}(su, tv)$. □

Osnovna rekurentna relacija za računanje dužine najdužeg zajedničkog podniza iskazana je u sljedećoj lemi.

Lema 4.1.4. *Za proizvoljne neprazne stringove* s *i* t *nad alfabetom* $\mathbb{N}_n$ *i proizvoljne simbole* $a, b \in \mathbb{N}_n$*, vrijedi*
(*i*) *Ako je* $a = b$*, tada je* $\mathrm{LCS}(sa, tb) = \mathrm{LCS}(s, t) + 1$,
(*ii*) *Ako je* $a \neq b$*, tada je* $\mathrm{LCS}(sa, tb) = \max\{\mathrm{LCS}(sa, t), \mathrm{LCS}(s, tb)\}$.

Dokaz. (*i*) Neka je u najduži zajednički podniz stringova s i t. Jasno da je tada ua najduži zajednički podniz stringova sa i ta, što znači da je $\mathrm{LCS}(sa, ta) = \mathrm{LCS}(s, t) + 1$.

(*ii*) Neka je $u = u_1 u_2 \dots u_{len(u)}$ najduži zajednički podniz stringova sa i tb. Ako je $u_{len(u)} = a$, tada je u podniz stringa t, što znači da je $len(u) \leqslant \mathrm{LCS}(sa, t)$. Ukoliko je $u_{len(u)} \neq a$, tada je u podniz stringa s, što znači da je $len(u) \leqslant \mathrm{LCS}(s, tb)$. Na osnovu toga, dobija se nejednakost $\mathrm{LCS}(sa, ta) =$

$len(u) \leqslant \max\{\mathrm{LCS}(sa,t), \mathrm{LCS}(s,tb)\}$. Suprotna nejednakost slijedi iz dijela (i) prehodne leme, jer je $s \prec sa$ i $t \prec tb$, pa je $\mathrm{LCS}(sa,t) \leqslant \mathrm{LCS}(sa,tb)$ i $\mathrm{LCS}(s,tb) \leqslant \mathrm{LCS}(sa,tb)$. □

Primjedba 4.1.5. Dio (ii) prethodne leme može da se uopšti na sljedeći način: ako su s i t neprazni stringovi za koje je $\mathrm{LCS}(s,t) = 0$, tada za proizvoljne neprazne stringove u i v vrijedi $\mathrm{LCS}(su,tv) = \max\{\mathrm{LCS}(su,t), \mathrm{LCS}(s,tv)\}$.

Koristeći uočenu rekurentnu relaciju, moguće je dati algoritam za rješavanje LCS problema, koji je zasnovan na dinamičkom programiranju. Za neprazne stringove s i t, neka je $d_{ij} := \mathrm{LCS}\left(s_{[1,i]}, t_{[1,j]}\right), i \leqslant len(s), j \leqslant len(t)$. Na osnovu prethodne leme, za brojeve d_{ij} vrijedi

$$\begin{aligned} &d_{i,0} = 0 = d_{0,j}; \\ &d_{i,j} = \begin{cases} d_{i-1,j-1} + 1, & \text{ako je } s_i = t_j; \\ \max\{d_{i-1,j}, d_{i,j-1}\}, & \text{inače.} \end{cases} \end{aligned} \tag{4.1}$$

Dakle, za izračunavanje $\mathrm{LCS}(s,t) = d_{len(s),len(t)}$ uz pomoć ovog pristupa, dovoljno je pronaći sve elemente matrice $D = (d_{ij})$.

Primjer 4.1.6. *Matrica D za stringove s i t iz prethodnog primjera ima sljedeći oblik:*

	4	1	1	1	3	3	3	4	2	1	2	1	1	4	1	2	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	**1**	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
4	1	1	1	1	1	1	1	**2**	2	2	2	2	2	3	3	3	3
2	1	1	1	1	1	1	1	2	**3**	3	3	3	3	3	3	4	4
3	1	1	1	1	2	2	2	2	3	3	3	3	3	3	3	4	4
2	1	1	1	1	2	2	2	2	3	3	**4**	4	4	4	4	4	4
1	1	2	2	2	2	2	2	2	3	4	4	**5**	5	5	5	5	5
1	1	2	3	3	3	3	3	3	3	4	4	5	**6**	6	6	6	6
4	1	2	3	3	3	3	3	4	4	4	4	5	6	**7**	7	7	7
2	1	2	3	3	3	3	3	4	5	5	5	5	6	7	7	8	8
2	1	2	3	3	3	3	3	4	5	5	6	6	6	7	7	8	8
2	1	2	3	3	3	3	3	4	5	5	6	6	6	7	7	**8**	8
3	1	2	3	3	4	4	4	4	5	5	6	6	6	7	7	8	8
3	1	2	3	3	4	5	5	5	5	5	6	6	6	7	7	8	8
1	1	2	3	4	4	5	5	5	5	6	6	7	7	7	8	8	9
4	1	2	3	4	4	5	5	6	6	6	6	7	7	8	8	8	9
1	1	2	3	4	4	5	5	6	6	7	7	7	8	8	9	9	**9**

Najduži zajednički podniz stringova s i t dobija se tako što se u datoj matrici pronađe trajektorija oblika 123456789*, na način da su, u svakoj etapi ove trajektorije, red i kolona odgovarajućeg elementa jednaki (upareni), a da se, u odnosu na trenutnu poziciju trajektorije, njen nastavak uvijek traži "jugoistočno",*

u odgovarajućoj podmatrici. Jedna takva trajektorija je označena uokvirenim elementima u prethodnoj matrici i ona odgovara najdužem zajedničkom podnizu 142211421.

Prethodni algoritam se može poboljšati posmatranjem samo uparenih simbola stringova s i t [47] ili posmatranjem tzv. *dominantnih uparivanja* [48], tj. uparivanja $a \leftrightarrow b$ za koje vrijedi

$$\mathrm{LCS}(sa, tb) > \max\{\mathrm{LCS}(s,t), \mathrm{LCS}(sa,t), \mathrm{LCS}(s,tb)\}.$$

Čak i sa ovim modifikacijama, algoritmi za rješavanje LCS problema na osnovu dinamičkog programiranja postaju neefikasni kada se sa slučaja dva stringa generalizuju na slučaj $m \geqslant 2$ ulaznih stringova. Osnovni problem je način skladištenja uparenih simbola, jer je u slučaju m ulaznih stringova potrebno uzeti u obzir sve $m-$torke koje se dobijaju kombinovanjem simbola na svim mogućim pozicijama odgovarajućih ulaznih stringova. Više detalja o kompleksnosti ovog, a i drugih algoritama za rješavanje LCS problema, biće dato u narednoj sekciji ove glave.

Novu dimenziju LCS problem dobija kada se dodatno pretpostavi da se ulazni stringovi za koji se on razmatra biraju na slučajan način, tj. da su simboli ovih stringova generisani u skladu sa uniformnom raspodjelom na alfabetu $\mathbb{N}_n$. Od veličina koje je korisno poznavati u ovom slučaju, svakako se ističe *očekivana dužina najdužeg zajedničkog podniza* (dva stringa) EL, koja se definiše na sljedeći način: Za prirodne brojeve $n, l \geqslant 1$,

$$\mathrm{EL}(n,l) := \frac{\sum_{s,t \in S(n,l)} \mathrm{LCS}(s,t)}{n^{2l}}.$$

Ova veličina je originalno predložena u radu [24], gdje su razmatrana neka njena svojstva. U pomenutom radu izračunate su vrijednosti $\mathrm{EL}(n,l)$, za $n \in \{1,2,\ldots,15\}$ i $l \in \{1,2,3,4,5\}$. Te vrijednosti su prikazane u Tabeli 4.1.

Tabela 4.1

	$\mathrm{EL}(n,1)$	$\mathrm{EL}(n,2)$	$\mathrm{EL}(n,3)$	$\mathrm{EL}(n,4)$	$\mathrm{EL}(n,5)$
$n=1$	1	2	3	4	5
$n=2$	0,5	1,125	1,8125	2,523438	3,246094
$n=3$	0,333333	0,888889	1,477366	2,090535	2,718742
$n=4$	0,25	0,734375	1,253906	1,801453	2,363899
$n=5$	0,2	0,624	1,096640	1,594317	2,108546
$n=6$	0,166667	0,541667	0,997109	1,435968	1,912269
$n=7$	0,142857	0,478134	0,881954	1,309838	1,754954
$n=8$	0,125	0,427734	0,803955	1,206201	1,625155
$n=9$	0,111111	0,386831	0,738692	1,119008	1,515694
$n=10$	0,1	0,353	0,683220	1,044309	1,421763
$n=11$	0,090909	0,324568	0,635470	0,979404	1,340005
$n=12$	0,083333	0,300347	0,593927	0,922366	1,267999
$n=13$	0,076923	0,279472	0,557455	0,871776	1,203953
$n=14$	0,071429	0,261297	0,525179	0,826554	1,146514
$n=15$	0,066667	0,245333	0,496417	0,785862	1,094633

Očekivana dužina najdužeg zajedničkog podniza dva stringa ima svojstvo *superaditivnosti*. Preciznije, vrijedi sljedeća lema.

Lema 4.1.7. *[24] Za sve prirodne brojeve $n, l_1, l_2 \geqslant 1$ vrijedi*

$$\mathrm{EL}(n, l_1) + \mathrm{EL}(n, l_2) \leqslant \mathrm{EL}(n, l_1 + l_2).$$

Dokaz. Za proizvoljne stringove $s, t \in S(n, l_1)$ i $u, v \in S(n, l_2)$, na osnovu dijela (ii) Leme 4.1.3, vrijedi $\mathrm{LCS}(s, t) + \mathrm{LCS}(u, v) \leqslant \mathrm{LCS}(su, tv)$, odakle se dobija

$$\begin{aligned}
\mathrm{EL}(n, l_1 + l_2) &= \frac{\sum_{x,y \in S(n, l_1+l_2)} \mathrm{LCS}(x, y)}{n^{2l_1+2l_2}} = \frac{\sum_{\substack{s,t \in S(n,l_1) \\ u,v \in S(n,l_2)}} \mathrm{LCS}(su, tv)}{n^{2l_1+2l_2}} \\
&\geqslant \frac{\sum_{\substack{s,t \in S(n,l_1) \\ u,v \in S(n,l_2)}} (\mathrm{LCS}(s, t) + \mathrm{LCS}(u, v))}{n^{2l_1+2l_2}} \\
&= \frac{\sum_{\substack{s,t \in S(n,l_1) \\ u,v \in S(n,l_2)}} \mathrm{LCS}(s, t) + \sum_{\substack{s,t \in S(n,l_1) \\ u,v \in S(n,l_2)}} \mathrm{LCS}(u, v)}{n^{2l_1+2l_2}} \\
&= \frac{n^{l_2} \cdot \sum_{s,t \in S(n,l_1)} \mathrm{LCS}(s, t) + n^{l_1} \cdot \sum_{u,v \in S(n,l_2)} \mathrm{LCS}(u, v)}{n^{2l_1+2l_2}} \\
&= \frac{\sum_{s,t \in S(n,l_1)} \mathrm{LCS}(s, t)}{n^{2l_1}} + \frac{\sum_{u,v \in S(n,l_2)} \mathrm{LCS}(u, v)}{n^{2l_2}} \\
&= \mathrm{EL}(n, l_1) + \mathrm{EL}(n, l_2).
\end{aligned}$$

□

Posljedica 4.1.8. *[28] Za sve prirodne brojeve $n, l_1, l_2 \geqslant 1$ vrijedi*

$$l_1 \cdot \mathrm{EL}(n, l_2) \leqslant \mathrm{EL}(n, l_1 \cdot l_2).$$

Dokaz. Dokaz će biti izveden indukcijom po l_1. Očigledno da je tvrđenje ispunjeno za $l_1 = 1$. Ako posmatrana nejednakost vrijedi za neko $l_1 \geqslant 1$, tada se iz prethodne leme dobija

$$\begin{aligned}
(l_1 + 1) \cdot \mathrm{EL}(n, l_2) &= l_1 \cdot \mathrm{EL}(n, l_2) + \mathrm{EL}(n, l_2) \leqslant \mathrm{EL}(n, l_1 \cdot l_2) + \mathrm{EL}(n, l_2) \\
&\leqslant \mathrm{EL}(n, l_1 \cdot l_2 + l_2) = \mathrm{EL}(n, (l_1 + 1) \cdot l_2).
\end{aligned}$$

Dakle, posmatrana nejednakost vrijedi za svaki prirodan broj $l_1 \geqslant 1$. □

Teorema 4.1.9. *[24] Za svaki prirodan broj $n \geqslant 2$ postoji vrijednost γ_n sa svojstvom*

$$\gamma_n = \lim_{l \to +\infty} \frac{\mathrm{EL}(n,l)}{l} = \sup\left\{\frac{\mathrm{EL}(n,l)}{l} : l \in \mathbb{N} \setminus \{0\}\right\}.$$

Dokaz. Prije svega, skup $\left\{\frac{\mathrm{EL}(n,l)}{l} : l \in \mathbb{N} \setminus \{0\}\right\}$ je ograničen odozgo (npr. sa 1), pa ovaj skup ima supremum. Neka je $\gamma_n := \sup\left\{\frac{\mathrm{EL}(n,l)}{l} : l \in \mathbb{N} \setminus \{0\}\right\}$ i $\varepsilon > 0$ proizvoljno. Tada postoji prirodan broj $k \geqslant 1$ takav da je $\frac{\mathrm{EL}(n,k)}{k} > \gamma_n - \varepsilon$. Za prirodan broj $l > \frac{\mathrm{EL}(n,k)}{\varepsilon}$, neka su q_l i r_l redom količnik i ostatak pri cjelobrojnom djeljenju broja l sa brojem k (pri čemu je $0 \leqslant r_l < k$). Za proizvoljno $n \geqslant 2$, na osnovu prethodne leme i njene posljedice, dobija se

$$\begin{aligned}
\gamma_n \geqslant \frac{\mathrm{EL}(n,l)}{l} &= \frac{\mathrm{EL}(n, q_l k + r_l)}{l} \geqslant \frac{\mathrm{EL}(n, q_l k)}{l} + \frac{\mathrm{EL}(n, r_l)}{l} \\
&\geqslant q_l \cdot \frac{\mathrm{EL}(n,k)}{l} + \frac{\mathrm{EL}(n,r_l)}{l} \geqslant q_l \cdot \frac{\mathrm{EL}(n,k)}{l} = \frac{l - r_l}{k} \cdot \frac{\mathrm{EL}(n,k)}{l} \\
&= \frac{\mathrm{EL}(n,k)}{k} - \frac{r_l}{kl} \cdot \mathrm{EL}(n,k) \geqslant \frac{\mathrm{EL}(n,k)}{k} - \frac{\mathrm{EL}(n,k)}{l} \\
&> \gamma_n - \varepsilon - \varepsilon = \gamma_n - 2\varepsilon.
\end{aligned}$$

Ovo znači da granična vrijednost $\lim_{l \to +\infty} \frac{\mathrm{EL}(n,l)}{l}$ postoji i jednaka je γ_n. □

Vrijednost γ_n ima praktičan značaj pri optimizovanju algoritama za rješavanje LCS problema, zato što se u okviru bilo kojeg algoritma može koristiti kao informacija o "projektovanoj" očekivanoj vrijednosti LCS. Nažalost, tačna vrijednost za γ_n nije poznata, čak ni za slučaj $n = 2$. Ocjenjivanje vrijednosti γ_n, u smislu zadavanje njene donje i gornje granice, započeto je još u radu [24], gdje je ova vrijednost i uvedena. Vremenom su različitim pristupima dobijane sve preciznije donje i gornje granice za γ_n. U radu [62], dokazano je $0,788071 \leqslant \gamma_2 \leqslant 0,826280$. U slučaju manjih vrijednosti $n \geqslant 3$, najbolje donje granice za γ_n mogu se naći u radu [54], dok se najbolje gornje granice mogu naći u radu [3]. Iz Tabele 4.1 se naslućuje da je $\lim_{n \to +\infty} \gamma_n = 0$. U radu [55] iz 2005. godine potvrđena je hipoteza koja se tiče brzine ove konvergencije; dokazano je da vrijedi $\lim_{n \to +\infty} \gamma_n \cdot \sqrt{n} = 2$.

Pojam očekivane dužine najdužeg zajedničkog podniza može na prirodan način da se uopšti na slučaj $m \geqslant 2$ stringova. U tom slučaju, veličina EL je

definisana sa

$$\mathrm{EL}(n,l,m) := \frac{\sum_{s_1,s_2,\ldots,s_m \in S(n,l)} \mathrm{LCS}(s_1,s_2,\ldots,s_m)}{n^{ml}}.$$

Specijalno, $\mathrm{EL}(n,l,2) = \mathrm{EL}(n,l)$. Svojstvo superaditivnosti se dokazuje slično i za ovu generalniju verziju veličine EL. To omogućava definisanje vrijednosti $\gamma_{n,m} := \lim_{l\to+\infty} \frac{\mathrm{EL}(n,l,m)}{l}$. Naravno, slično kao za $\gamma_n = \gamma_{n,2}$, tačne vrijednosti za $\gamma_{n,m}$ su nepoznate, te se za njihovu praktičnu upotrebu koriste razne aproksimacije. U radu [28], dokazano je da za svako $m \geqslant 2$ vrijedi $1 \leqslant \lim_{n\to+\infty} n^{1-\frac{1}{m}} \cdot \gamma_{n,m} \leqslant e$.

Većina radova iz literature posvećena je razvijanju i poboljšavanju metoda za nalaženje najdužeg zajedničkog podniza familije stringova čiji se simboli biraju na slučajan način, a u skladu sa uniformnom raspodjelom na alfabetu. U narednom će biti izložen nešto generalniji pristup, koji podrazumijeva upotrebu nekih oblika polinomijalne raspodjele na alfabetu $\mathbb{N}_n$. Ovaj pristup je predložen u radu [69] i predstavlja originalan doprinos razmatranoj temi.

Za alfabet $\mathbb{N}_n, n \geqslant 2$, kao prostor ishoda simbola stringova, *polinomijalna raspodjela* $\mathrm{MN}(p_1,\ldots,p_n)$ na skupu $\mathbb{N}_n$ određena je pozitivnim realnim brojevima p_i, koji predstavljaju vjerovatnoću ostvarivanja simbola $i \in \mathbb{N}_n$ na datoj poziciji stringa i za koje vrijedi $\sum_{i=1}^{n} p_i = 1$. Specijalno, u slučaju da su vjerovatnosne težine p_i balansirane, dobija se uniformna raspodjela, gdje je $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$.

U prethodnoj glavi ove teze, string je modelovan kao stohastički proces. U principu, to znači da se on može tretirati kao slučajni vektor sastavljen od slučajnih promjenljivih koje bilježe inicijalizaciju njegovih simbola. Međuodnos između ovih slučajnih promjenljivih može biti takav da su one međusobno nezavisne ili da između njih postoji određen oblik zavisnosti. Slučaj postojanja zavisnosti razmotren je u prethodnoj glavi ove teze, gdje je pomenuta međuzavisnost opisana uslovnim vjerovatnoćama koje su dobijene na osnovu odgovarajućeg vjerovatnosnog modela (vjerovatnosnog sufiksnog drveta ili hijerarhijskog Pitman-Jorovog procesa). Stoga je u daljnjem izlaganju pretpostavljeno da je string slučajni vektor sastavljen od nezavisnih slučajnih veličina, gdje se svaka od njih ravna po istoj polinomijalnoj raspodjeli.

Neka je t string izabran u skladu sa polinomijalnom $\mathrm{MN}(p_1,\ldots,p_n)$ raspodjelom. Cilj je opisati vjerovatnoću da string s poznate dužine, čiji simboli imaju vjerovatnosne težine iz date polinomijalne raspodjele, bude podniz stringa t. Ova vjerovatnoća će biti notirana sa $P(s \prec t)$. Jedna ideja za nalaženje ove vje-

rovatnoće podrazumijeva ispitivanje vjerovatnoća uparivanja "s lijeva na desno" pojedinačnih simbola stringova s i t, u kombinaciji sa postepenim prelaskom na ispitivanje vjerovatnoće da je odgovarajući sufiks stringa s podniz odgovarajućeg sufiksa stringa t. U sljedećoj teoremi, predložena je nova rekurentna relacija koja formalizuje ovu ideju.

Teorema 4.1.10. *Neka je $t = t_1 t_2 \ldots t_{len(t)}$ dati string izabran u skladu sa polinomijalnom raspodjelom i $s = s_1 s_2 \ldots s_{len(s)}$ string, tako da sve slučajne promjenljive s_i imaju istu polinomijalnu raspodjelu po kojoj se ravnaju simboli stringa t. Tada vrijedi*

$$P(s \prec t) = \begin{cases} 1, & za\ len(s) = 0; \\ 0, & za\ len(s) > len(t); \\ P\big(s_1 = t_1\big) \cdot P\big(s_{[2,len(s)]} \prec t_{[2,len(t)]}\big) & \\ +P\big(s_1 \neq t_1\big) \cdot P\big(s \prec t_{[2,len(t)]}\big), & inače. \end{cases} \tag{4.2}$$

Dokaz. Prazan string je sigurno podniz svakog stringa, dok je nemoguće da string bude podniz stringa manje dužine. Zbog toga, ostaje još da se jednakost potvrdi u slučaju $1 \leqslant len(s) \leqslant len(t)$. Događaji $s_1 = t_1$ i $s_1 \neq t_1$ čine potpun sistem događaja, pa se, na osnovu formule potpune vjerovatnoće, dobija

$$\begin{aligned} P(s \prec t) &= P(s_1 = t_1) \cdot P\left(s \prec t | s_1 = t_1\right) + P(s_1 \neq t_1) \cdot P\left(s \prec t | s_1 \neq t_1\right) \\ &= P\big(s_1 = t_1\big) \cdot P\big(s_{[2,len(s)]} \prec t_{[2,len(t)]}\big) + P\big(s_1 \neq t_1\big) \cdot P\big(s \prec t_{[2,len(t)]}\big). \end{aligned}$$

Posljednja jednakost vrijedi, jer se događaj $s \prec t | s_1 = t_1$ realizuje ukoliko je najduži pravi sufiks stringa s podniz najdužeg pravog sufiksa stringa t, dok su povoljni ishodi za događaj $s \prec t | s_1 \neq t_1$ oni za koji je string s podniz najdužeg pravog sufiksa stringa t. □

Vjerovatnoća $P(s \prec t)$ iz rekurentne relacije (4.2) zavisi ne samo od dužina stringova s i t, već i od vjerovatnosnih težina pridruženih simbolima alfabeta. Zbog toga je ovu vjerovatnoću teško izraziti u eksplicitnoj formi u opštem slučaju polinomijalne raspodjele $\mathrm{MN}(p_1, \ldots, p_n)$. Za neke specijalne slučajeve polinomijalne raspodjele, broj varijabli potrebnih za opis ove vjerovatnoće se drastično smanjuje, što daje veći praktični značaj date rekurentne relacije. Ovi specijalni slučajevi podrazumijevaju pružanje dodatnih informacija o vjerovatnosnim težinama $p_1, \ldots, p_n$, u pogledu njihove balansiranosti. U narednom će biti razmotrena 3 ovakva slučaja.

- Prvi slučaj: Kompletna balansiranost vjerovatnoća $p_1, \ldots, p_n$.

  Tada vrijedi $p_1 = p_2 = \ldots = p_n = \dfrac{1}{n}$ i u pitanju je uniformna raspodjela. U ovom slučaju, vjerovatnoća $P(s \prec t)$ zavisi isključivo od $k := len(s)$ i

$l := len(t)$, pa može da se notira i sa $P(k, l)$. Rekurentna relacija (4.2) se redukuje na sljedeću relaciju, koja je prvobitno predložena u radu [67]:

$$P(k,l) = \begin{cases} 1, & \text{za } k = 0; \\ 0, & \text{za } k > l; \\ \frac{1}{n} \cdot P(k-1, l-1) + \frac{n-1}{n} \cdot P(k, l-1), & \text{inače.} \end{cases} \tag{4.3}$$

U pomenutom radu prethodna rekurentna relacija je implementirana tehnikama dinamičkog programiranja.

- Drugi slučaj: Nebalansiranost jedne od vjerovatnoća $p_1, \ldots, p_n$.

  Neka simbol $i \in \mathbb{N}_n$ ima vjerovatnosnu težinu $p_i = p \neq \frac{1}{n}$, a svi ostali simboli iz $\mathbb{N}_n$ vjerovatnosne težine jednake $\frac{1-p}{n-1}$. Tada se rekurentna relacija (4.2) svodi na oblik:

$$P(s \prec t) = \begin{cases} 1, & \text{za } len(s) = 0; \\ 0, & \text{za } len(s) > len(t); \\ q \cdot P\big(s_{[2,len(s)]} \prec t_{[2,len(t)]}\big) & \\ +(1-q) \cdot P\big(s \prec t_{[2,len(t)]}\big), & \text{inače,} \end{cases} \tag{4.4}$$

  gdje je

$$q := \begin{cases} p, & \text{za } t_1 = i; \\ \frac{1-p}{n-1}, & \text{inače.} \end{cases}$$

  Primjećuje se da, pored dužina stringova s i t, vjerovatnoća $P(s \prec t)$ zavisi samo od indikatora oblika $I\{t_j = i\}$.

- Treći slučaj: Dva tipa balansiranosti vjerovatnoća $p_1, p_2, \ldots, p_n$.

  Ovaj slučaj je uopštenje prethodnog slučaja. Umjesto jednog simbola sa vjerovatnoćom $p \in (0, 1)$, neka sada svim simbolima iz višečlanog skupa $K \subsetneq \mathbb{N}_n$ odgovara ova vjerovatnoća. Pored toga, pretpostavlja se da i svim simbolima iz skupa $\mathbb{N}_n \setminus K$ odgovara ista vjerovatnoća. Efektivno, to znači da je $p_i = \frac{p}{|K|}$, za svako $i \in K$, kao i da je $p_i = \frac{1-p}{n-|K|}$, za svako $i \in \mathbb{N}_n \setminus K$. Za ovakvu polinomijalnu raspodjelu, rekurentna relacija (4.2) se svodi na oblik:

$$P(s \prec t) = \begin{cases} 1, & \text{za } len(s) = 0; \\ 0, & \text{za } len(s) > len(t); \\ q \cdot P\big(s_{[2,len(s)]} \prec t_{[2,len(t)]}\big) & \\ +(1-q) \cdot P\big(s \prec t_{[2,len(t)]}\big), & \text{inače,} \end{cases} \tag{4.5}$$

gdje je

$$q := \begin{cases} \dfrac{p}{|K|}, & \text{za } t_1 \in K; \\ \dfrac{1-p}{n-|K|}, & \text{za } t_1 \in \mathbb{N}_n \setminus K. \end{cases}$$

U odnosu na prethodni slučaj, vjerovatnoća $P(s \prec t)$ sada zavisi od indikatora oblika $I\{t_j \in K\}$.

Još jedan mogući pristup za računanje vjerovatnoće $P(s \prec t)$ podrazumijeva pretpostavku da string t nije zadat, nego da je on, kao i string s, slučajni vektor. Takođe, razumna je dodatna pretpostavka da su, za svako j_1, j_2, slučajne promjenljive s_{j_1} i t_{j_2} nezavisne. Uz ovu postavku, vrijedi sljedeća teorema.

Teorema 4.1.11. *Neka su s i t stringovi koji potiču iz iste polinomijalne raspodjele $\mathrm{MN}(p_1, p_2, \ldots, p_n)$. Tada vrijedi*

$$P(s \prec t) = \begin{cases} 1, & \textit{za } len(s) = 0; \\ 0, & \textit{za } len(s) > len(t); \\ \left(\sum_{i=1}^{n} p_i^2\right) \cdot P\left(s_{[2,len(s)]} \prec t_{[2,len(t)]}\right) & \\ + \left(1 - \sum_{i=1}^{n} p_i^2\right) \cdot P\left(s \prec t_{[2,len(t)]}\right), & \textit{inače.} \end{cases} \tag{4.6}$$

Dokaz. Prva dva slučaja su trivijalna, pa ostaje da se jednakost potvrdi u slučaju $1 \leqslant len(s) \leqslant len(t)$. Iz pretpostavljene nezavisnosti, primjenjujući formulu potpune vjerovatnoće, dobija se da za sve j_1, j_2 vrijedi

$$\begin{aligned} P\left(s_{j_1} = t_{j_2}\right) &= \sum_{i=1}^{n} P\left(t_{j_2} = i\right) \cdot P\left(s_{j_1} = t_{j_2} | t_{j_2} = i\right) \\ &= \sum_{i=1}^{n} P\left(t_{j_2} = i\right) \cdot P\left(s_{j_1} = i\right) = \sum_{i=1}^{n} p_i^2. \end{aligned}$$

Na osnovu toga, uz još jednu primjenu formule potpune vjerovatnoće, dobija se

$$\begin{aligned} P(s \prec t) &= P(s_1 = t_1) \cdot P\left(s \prec t | s_1 = t_1\right) + P(s_1 \neq t_1) \cdot P\left(s \prec t | s_1 \neq t_1\right) \\ &= P\left(s_1 = t_1\right) \cdot P\left(s_{[2,len(s)]} \prec t_{[2,len(t)]}\right) + P\left(s_1 \neq t_1\right) \cdot P\left(s \prec t_{[2,len(t)]}\right) \\ &= \left(\sum_{i=1}^{n} p_i^2\right) \cdot P\left(s_{[2,len(s)]} \prec t_{[2,len(t)]}\right) + \left(1 - \sum_{i=1}^{n} p_i^2\right) \cdot P\left(s \prec t_{[2,len(t)]}\right). \end{aligned}$$

□

Iz prethodne teoreme se uočava da, pod datim pretpostavkama, polinomijalna raspodjela $\mathrm{MN}(p_1, p_2, \ldots, p_n)$ utiče na vjerovatnoću $P(s \prec t)$ samo kroz vrijednost izraza $\sum_{i=1}^{n} p_i^2$. U tom smislu, jedini varijabilni faktori koji određuju

ovu vjerovatnoću su vrijednosti $k := len(s)$ i $l := len(t)$, te se, kao u slučaju uniformne raspodjele, za ovu vjerovatnoću može koristiti oznaka $P(k, l)$. Što je još važnije, ovo omogućava da se, tehnikama dinamičkog programiranja, za sve relevantne vrijednosti k i l vjerovatnoće $P(k, l)$ uskladište u vidu jedne matrice vjerovatnoća. Ova ideja biće eksploatisana u narednoj sekciji ove glave.

4.2 Optimizacijski aspekti LCS problema

Vrijednost $\mathrm{LCS}(S)$ za višečlanu familiju stringova $S = \{s_1, s_2, \ldots, s_m\}$ predstavlja mjeru sličnosti stringova ovog skupa. Ova mjera je u širokoj upotrebi u računarskoj biologiji [66], kompresiji podataka [86], [7], editovanju teksta [57], detektovanju presjecanja GPS putanja [97], poređenju fajlova [10], kao i u kontrolnim sistemima revizije kao što je GIT.

Za fiksiranu vrijednost m, već pominjani algoritam baziran na dinamičkom programiranju (DP) i razna njegova poboljšanja su poznata u literaturi [44]. Ovi algoritmi su polinomijalne složenosti; vrijeme izvršavanja algoritama ovakvog tipa je reda veličine $O(l^m)$, gdje je l dužina najdužeg stringa iz skupa S. Zbog toga, ovi algoritmi postaju nepraktični sa povećavanjem broja m. Ukoliko se dopusti da broj m ulaznih stringova iz skupa S bude po volji veliki broj, pokazuje se da je LCS problem NP-težak [63]. U praksi, heurističke tehnike se koriste u slučaju većih vrijednosti za m i l. Konstruktivne heuristike, kao što su algoritam ekspanzije i Best-Next heuristika ([39] i [50]), prve su se u literaturi koristile za rješavanje LCS problema. Značajno bolja rješenja se dobijaju primjenom naprednijih metaheurističkih pristupa. Većina njih se zasnivaju na *Pretrazi Bima* (BS) (na engleskom Beam-Search, vidjeti npr. [13], [32], [67], [87] i [93]), uz korišćenje različitih varijanti u zavisnosti od izbora vođenja heuristike, šeme grananja i mehanizma filtracije. U radu [33], predložen je uopšteni BS okvir u cilju ujedinjenja svih postojećih BS varijanti iz literature. Uvođenjem odgovarajuće parametrizacije izražena je svaka zasebna BS varijanta, što je omogućilo njihovo međusobno poređenje. Štaviše, u pomenutom radu predložena je heuristika evaluacije koja aproksimira očekivanu LCS vrijednost skupa stringova generisanih iz uniformne raspodjele zadate na alfabetu $\mathbb{N}_n$. Na ovaj način, dobijena je BS varijanta koja se pokazala superiornijom u odnosu na ostale varijante, u smislu uporednog testiranja na većini standardnih instanci dobijenih na osnovu uniformne ili približno uniformne raspodjele.

Kada su u pitanju egzaktni pristupi za rješavanje LCS problema, model cjelobrojnog linearnog programiranja je razmatran u [14]. Ispostavilo se da on nije dovoljno konkurentan, jer ga nije moguće primijeniti na većini standardnih instanci prisutnih u literaturi. Razlog tome je veličina ovog modela - on koristi previše mnogo binarnih promjenljivih i ograničenja, čak i u slučaju instanci malih dužina. Pristupi koji koriste dinamičko programiranje vrlo brzo iscrpe

memoriju pri radu sa instancama umjerenih dužina, a takođe vrlo često daju "slaba" rješenja. U radu [19], uveden je FAST-LCS paralelni algoritam koji ublažava nedostatke vezane za brzinu izvršavanja. U radu [93] predložen je paralelni algoritam pretrage pod nazivom QUICK-DP, koji se zasniva na pristupu dominantnih tačaka koje se izračunavaju na osnovu brzih "podijeli i vladaj" tehnika. Još jedan paralelni algoritam LEVELED-DAG, zasnovan na modelu grafa, razmatran je u radu [72]. Takođe, u radu [60] sugerisan je TOP-MLCS algoritam, baziran na modelu usmjerenog acikličnog višeslojnog grafa i navedene su topološki orijentisane strategije sortiranja za uklanjanje putanja koje odgovaraju suboptimalnim rješenjima. U skorije vrijeme, u radu [34] je predložena A^* pretraga kojom se nadmašuje TOP-MLCS algoritam i ostali navedeni egzaktni pristupi u pogledu upotrebe memorije i broja instanci za koje su nađena optimalna rješenja. Ipak, primjena ove egzaktne A^* pretrage je i dalje ograničena na slučaj instanci malih dužina. U posljednjem pomenutom radu, A^* pretraga je takođe upotrebljena kao baza za hibridni anytime algoritam koji se može zaustaviti u bilo kojem vremenu i pružiti "razumna" heuristička rješenja. Ovaj pristup podrazumijeva da se iteracije dobijene klasičnom A^* pretragom kombinuju sa iteracijama dobijenih primjenom anytime column pretrage uvedene u radu [92]. Pokazano je da dobijeni hibridni algoritam nadmašuje ostale Anytime algoritme iz literature (kao što su PRO-MLCS [98] i anytime pack search [91]) u pogledu kvaliteta rješenja.

Opisani metodi za približno rješavanje LCS problema su primarno fokusirani na nezavisne slučajne i kvazi-slučajne stringove formirane od simbola alfabeta čija je raspodjela vjerovatnoća uniformna ili skoro uniformna. Teoretski, zbog ravnomjerne raspoređenosti vjerovatnoća, uniformna raspodjela se čini kao najlogičniji izbor, ali u praksi se pojave slučajnog tipa često vjernije modeliraju uz pomoć raspodjela drugačijeg tipa (npr. raspodjela vodeće cifre slučajno izabranog skupa brojeva se ravna po Benfordovoj raspodjeli). Stoga, prirodno je postaviti pitanje koliko je efikasnost pomenutih metoda posljedica pretpostavljene uniformnosti i kakve su njihove performanse u slučaju nekih drugih distribucija? Konkretno, koliko dobro ovi metodi funkcionišu ako se za raspodjelu vjerovatnoća simbola alfabeta izabere neka od neuniformnih distribucija iz prethodne sekcije ove glave i da li je u ovoj postavci moguće pružiti efikasniji algoritam? U pokušaju da se pruži zadovoljavajući odgovor na ovo pitanje, u nastavku ove glave izloženi su originalni rezultati predstavljeni u radu [69], u kojem je predložena nova varijanta pretrage bima. Ova varijanta je pokazala bolji učinak pri radu sa neuniformnim instancama od svojih konkurenata. U njenoj osnovi je novouvedena heuristika GMPSUM, čija je implementacija jednostavnija od aproksimacija očekivane dužine LCS-a, te samim tim nema problema sa numeričkom stabilnošću u slučaju ulaznih stringova velike dužine. U narednom slijedi precizniji opis metodologije koja se u pomenutom radu koristi.

Pretraga bima (BS) je dobro poznata heuristička pretraga koja predstavlja redukovanu verziju BFS algoritma pretrage (skraćenica od Breadth-First-Search), pri čemu se, umjesto grananja po svim još uvijek nerazmatranim čvorovima drveta sa istog nivoa, bira tačno određen broj $\beta > 0$ najperspektivnijih čvorova i prate grane koje polaze iz njih. Na taj način, složenost drveta pretrage ostaje polinomijalne veličine. Izbor β čvorova koji će biti uzeti u razmatranje za daljnje grananje zavisi od konteksta problema i ovaj izbor se vrši na osnovu heurističke funkcije usmjeravanja pretrage h. Stoga, efektivnost pretrage značajno zavisi od ove funkcije. Preciznije, BS funkcioniše na sljedeći način: Prvo se postavlja inicijalni bim B sa čvorom korijenom r kao inicijalnim stanjem; u slučaju LCS problema, ovo je prazno parcijalno rješenje. U svakoj iteraciji, posmatraju se grane čvorova iz B koje se dobijaju uz pomoć svih dopustivih akcija. Dobijeni čvorovi potomci se čuvaju u skupu ekstenzija V_{ext}. Važno je istaći da, za neke probleme, postoje efikasne tehnike filtriranja koje se mogu primijeniti kako bi se iz V_{ext} izbacili čvorovi koji su "dominirani" od strane drugih čvorova, tj. izbacili oni čvorovi koji ne mogu generisati bolja rješenja. Za ovaj vid kontrole koristi se parametar k_{filter}. Ovako (moguće redukovan) skup ekstenzija se sortira u skladu sa vrijednostima čvorova u odnosu na funkciju h, i prvih β čvorova (ili manje, ako ih nema toliko u skupu V_{ext}) formiraju bim B za sljedeći nivo. Opisani postupak se ponavlja za svaki nivo, sve dok bim B ne postane prazan. U opštem slučaju, da bi se riješio problem kombinatorne optimizacije, informacija o najdužem (ili najkraćem) putu od čvora korijena do dopustivog ciljnog čvora se čuva kako bi se na kraju dobilo rješenje koje predstavlja maksimum ili minimum date funkcije cilja. Pseudo-kod koji opisuje prethodnu proceduru je dat u Algoritmu 1.

Graf stanja za LCS problem koji se koristi u svim BS varijantama je dobro poznat u literaturi (vidjeti npr. [33] i [34]). U pitanju je usmjeren aciklični graf $G = (V, A)$, gdje čvor $v = (\boldsymbol{\theta}^v, l^v) \in V$ predstavlja skup parcijalnih rješenja, koja:

1. imaju istu dužinu l^v;
2. indukuju isti podproblem označen sa $S[\boldsymbol{\theta}^v]$, u odnosu na zadati vektor pozicija $\boldsymbol{\theta}^v$.

Kaže se da parcijalno rješenje s *indukuje podproblem* $S[\boldsymbol{\theta}^v]$ ako je $s_{i\,[1,\theta_i^v-1]}$ najmanji prefiks stringa s_i među svim prefiksima koji imaju s kao podniz. Grana $a = (v_1, v_2) \in A$ postoji između dva različita čvora $v_1, v_2 \in V$ i označena je sa $l(a) \in \mathbb{N}_n$ ako vrijedi

1. $l^{v_2} = l^{v_1} + 1$;
2. parcijalno rješenje koje indukuje v_2 je dobijeno dodavanjem $l(a)$ na parcijalno rješenje koje indukuje v_1.

Algoritam 1 Pretraga Bima

Ulaz: Instanca problema, heuristika h, $\beta > 0$, k_{filter}
Izlaz: Heurističko rješenje
$B \leftarrow \{r\}$
while $B \neq \emptyset$ **do**
    $V_{\text{ext}} \leftarrow \emptyset$
    for $v \in B$ **do**
        if v je ciljni čvor **then**
            ako čvor predstavlja novo najbolje rješenje, on se skladišti
        else
            dodati još nerazmatrane čvorove potomke od v u V_{ext}
        end if
    end for
    if $k_{\text{filter}} > 0$ **then**
        $V_{\text{ext}} \leftarrow$ `Filter`$(V_{\text{ext}}, k_{\text{filter}})$ // opciono se filtriraju dominirani čvorovi
    end if
    $B \leftarrow$ `SelectBetaBest`$(V_{\text{ext}}, \beta, h)$
end while
return najbolje nađeno rješenje

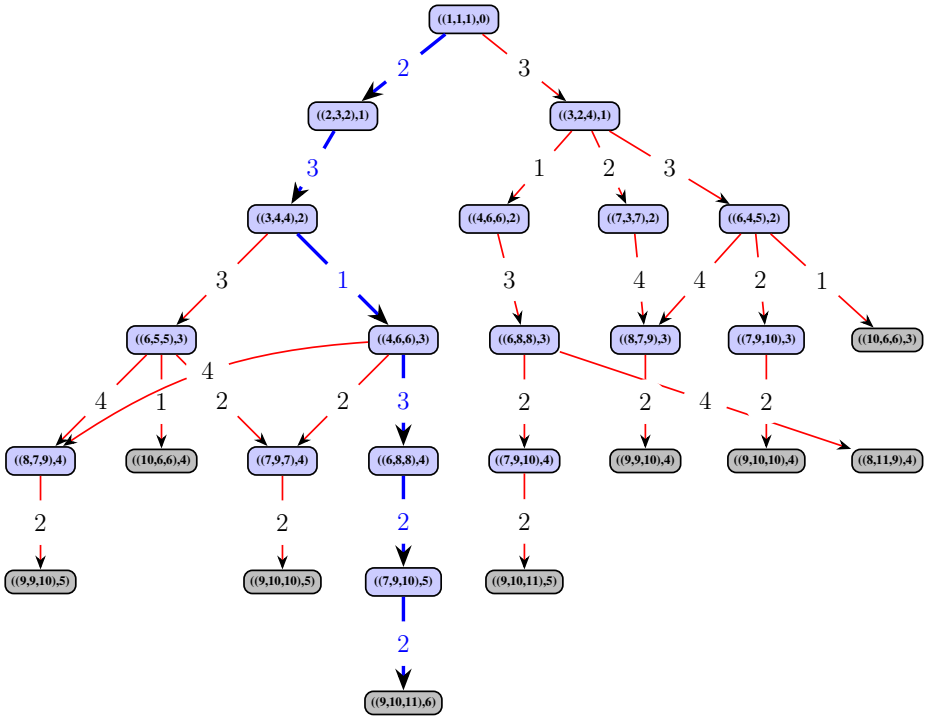
Čvor korijen $r := ((1, 1, \ldots, 1), 0)$ pripada G; on referiše na originalni LCS problem na skupu ulaznih stringova S i za njega se može smatrati da je indukovan praznim parcijalnim rješenjem ε.

Za izvođenje potomka čvora $v \in V$, prvo se pronađe skup $\mathbb{N}_n(v) \subseteq \mathbb{N}_n$ sastavljen od svih simbola alfabeta koja dopustivo proširuju parcijalno rješenje određeno sa v. Kandidati za elemente skupa $\mathbb{N}_n(v)$ su svi simboli alfabeta $\mathbb{N}_n$ koji se pojavljuju bar jednom u svakom stringu podproblema određenog skupom sufiksnih stringova $S[\boldsymbol{\theta}^v]$. Skup $\mathbb{N}_n(v)$ se nadalje redukuje određivanjem i uklanjanjem simbola koji su dominirani od strane drugih simbola iz skupa $\mathbb{N}_n(v)$. Formalno, simbol $a \in \mathbb{N}_n(v)$ *ima dominaciju* nad simbolom $b \in \mathbb{N}_n(v)$ ili simbol b je *dominiran* od strane simbola a, ukoliko za svako $i \in \{1, 2, \ldots, m\}$ vrijedi $\theta^v_{i,a} \leqslant \theta^v_{i,b}$ (podsjećanja radi, $\theta^v_{i,a}$ označava poziciju prvog pojavljivanja simbola a u stringu s_i počevši od pozicije θ^v_i).

Simboli koji su dominirani se mogu zanemariti, jer vode do suboptimalnih rješenja. Neka je $\mathbb{N}^{nd}_n(v) \subseteq \mathbb{N}_n(v)$ skup svih dopustivih simbola koji nisu dominirani od strane drugog simbola. Za svaki simbol $a \in \mathbb{N}^{nd}_n(v)$, graf G sadrži čvor potomak $v' = \left(\boldsymbol{\theta}^{v'}, l^v + 1\right)$ od v, gdje je $\theta^{v'}_i = \theta^v_{i,a} + 1$, $i \in \{1, 2, \ldots, m\}$. Čvor v koji nema čvorova potomaka, tj. za koji vrijedi $\mathbb{N}^{nd}_n(v) = \emptyset$, naziva se *neproduživim čvorom* ili *ciljnim čvorom*. Među svim ciljnim čvorovima potrebno je naći one kojima su određeni najduži stringovi koji predstavljaju

rješenja, tj. one ciljne čvorove v sa najvećom vrijednošću l^v. Važno je istaći da svaka putanja od čvora korijena r do čvora $v \in V$ predstavlja dopustivo parcijalno rješenje dobijeno prikupljanjem i konkatenacijom oznaka grana koje određuju taj put. Zbog toga, nije neophodno skladištiti tako dobijeno parcijalno rješenje s u čvorovima grafa. U grafu G, proizvoljan put koji povezuje čvor korijen sa neproduživim čvorom reprezentuje zajednički, neproduživi podniz skupa stringova iz skupa S. Svaki najduži put od čvora korijena do ciljnog čvora predstavlja *optimalno* ili *najbolje rješenje* za instancu problema S.

Primjer 4.2.1. *Neka je* $S = \{\underbrace{231132421}_{s_1}, \underbrace{3233143224}_{s_2}, \underbrace{22331234221}_{s_3}\}$ *skup stringova definisanih nad alfabetom* $\mathbb{N}_4$. *Graf stanja za LCS problem na skupu instanci S predstavljen je Slici 4.1. Sivom bojom su obojeni neproduživi ciljni čvorovi. Najduži put u ovom grafu stanja je istaknut plavom bojom; on vodi od čvora korijena do čvora* $((9, 10, 11), 6)$ *i odgovara rješenju* $s = 231322$, *što je string dužine* 6.



Slika 4.1

Potrebno je još objasniti način filtriranja dominiranih čvorova iz skupa V_{ext}, tj. pojasniti proceduru `Filter` iz Algoritma 1. Podrazumijevaće se efikasno *ograničeno filtriranje*, izloženo u radu [87], koje je parametrizovano veličinom

filtera $k_{\text{filter}} > 0$. Ideja je biranje najviše k_{filter} najboljih čvorova iz V_{ext} i provjeravanje prethodno pomenute relacije dominiranja za ovaj podskup čvorova u kombinaciji sa svim ostalim čvorovima iz V_{ext}. Ako je relacija dominiranja "pozitivno ocjenjena", dominirani čvorovi se izbacuju iz V_{ext}. Poželjno je uzeti $k_{\text{filter}} < |V_{\text{ext}}|$, jer puno filtriranje može da bude "preskupo" u smislu potrebnog vremena izvršavanja za veće širine bima.

Sada će biti prezentovana nova heuristika za evaluaciju čvorova u BS, u cilju njihovog rangiranja i biranja bima na sljedećem nivou. Ova heuristika, pod nazivom GMPSUM (što je skraćenica od Geometric Mean Probability Sum), posebno je prilagođena radu sa nebalansiranim instancama i dobija se kao konveksna kombinacija sljedeće dvije statistike:

- GM statistika (GM je skraćeno od Geometric Mean) je zasnovana na geometrijskoj sredini i geometrijskoj standardnoj devijaciji frekvencija pojavljivanja simbola svih ulaznih stringova odgovarajućih podproblema.

- PSUM statistika (PSUM je skraćeno od Probability Sum) je zasnovana na prethodno uvedenoj matrici vjerovatnoća $P(k,l)$ za proizvoljnu nebalansiranu polinomijalnu distribuciju datu rekurentnom relacijom (4.6); specijalno, mogu se zamjenski koristiti vjerovatnoće dobijene na osnovu rekurentnih relacija (4.4) ili (4.5).

Preciznije, za dati čvor v i simbol $a \in \mathbb{N}_n$, neka je

$$C_a\left(S[\boldsymbol{\theta}^v]\right) := \left(\left|s_{1[\theta_1^v, len(s_1)]}(a)\right|, \left|s_{2[\theta_2^v, len(s_2)]}(a)\right|, \ldots, \left|s_{m[\theta_m^v, len(s_m)]}(a)\right|\right)$$

vektor čije komponente predstavljaju broj pojavljivanja simbola a u svakom od stringova odgovarajućeg podproblema. Na osnovu ovog vektora moguće je definisati vrijednost $\mathrm{UB}_1(v) := \sum_{a \in \mathbb{N}_n} \min_{i \in \{1,2,\ldots,m\}} C_a\left(S[\boldsymbol{\theta}^v]\right)_i$ (UB je skraćenica od Upper Bound), koja se koristi u radu [13] i predstavlja gornju granicu dužine LCS−a za podproblem reprezentovan čvorom v.

Za vektor $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, *geometrijska sredina* i *geometrijska standardna devijacija* se redom uvode na sljedeći način:

$$\mu_g(\boldsymbol{x}) := \sqrt[m]{x_1 \cdot x_2 \cdot \ldots \cdot x_m},$$

$$\sigma_g(\boldsymbol{x}) := e^{\sqrt{\frac{\sum_{i=1}^{m} \left(\ln \frac{x_i}{\mu_g(\boldsymbol{x})}\right)^2}{m}}}.$$

Konačno, GM statistika se uvodi na sljedeći način:

$$\mathrm{GM}(v) = \mathrm{GM}\left(S[\boldsymbol{\theta}^v]\right) := \sum_{a \in \mathbb{N}_n} \frac{\mu_g\left(C_a\left(S[\boldsymbol{\theta}^v]\right)\right)}{\sigma_g\left(C_a\left(S[\boldsymbol{\theta}^v]\right)\right)} \cdot \frac{\min\limits_{i \in \{1,2,\ldots,m\}} C_a\left(S[\boldsymbol{\theta}^v]\right)_i}{\mathrm{UB}_1(v)}. \quad (4.7)$$

Suštinski, GM statistika je težinska aritmetička sredina korigovanih geometrijskih sredina $\frac{\mu_g(\cdot)}{\sigma_g(\cdot)}$ frekvencija pojavljivanja simbola, gdje su težine za svaki simbol dobijene normalizacijom minimalnog broja pojavljivanja tog simbola u svim stringovima sa sumom minimalnog broja pojavljivanja svih simbola alfabeta. Motivacija za ovu definiciju proističe iz sljedećih zapažanja:

1. Simboli sa nadprosječnom frekvencijom pojavljivanja u svim stringovima povećavaju šansu nalaženja dužeg zajedničkog podniza (sastavljenog od ovih simbola).

2. Veće odstupanje od prosjeka prirodno redukuje ovu šansu.

3. Minimalan broj pojavljivanja fiksiranog simbola u svim ulaznim stringovima je gornja granica dužine zajedničkog podniza koji se može formirati samo uz pomoć tog simbola. Zbog toga, praveći normalizaciju sa sumom svih minimalnih pojavljivanja svih simbola, svakom simbolu pojedinačno se dodjeljuje ponder koji kvantifikuje njegov značaj u ukupnoj sumi.

GM statistika ima značaj ukoliko se prateća geometrijska sredina i geometrijska standardna devijacija dobijaju na osnovu uzoraka dovoljno velikog obima. U eksperimentalnom dijelu istraživanja uzimano je bar 10 stringova u ulaznoj instanci. Rad sa instancama manjeg obima smanjuje relevantnost GM statistike.

Za dati čvor v, neka je $l_{\max}(v) := \min\limits_{i \in \{1,2,\ldots,m\}} \left(len(s_i) - \theta_i^v + 1\right)$. Tada se PSUM statistika uvodi sa

$$\mathrm{PSUM}(v) = \mathrm{PSUM}\left(S[\boldsymbol{\theta}^v]\right) := \sum_{k=1}^{l_{\max}(v)} \prod_{i=1}^{m} P\left(k, len(s_i) - \theta_i^v + 1\right). \quad (4.8)$$

Za razliku od GM statistike koja razmatra uglavnom opšti aspekat pretpostavljene raspodjele vjerovatnoća simbola alfabeta $\mathbb{N}_n$, PSUM statistika "skenira" u potrazi za specifičnijim relacijama između ulaznih stringova. Ona predstavlja sumu vjerovatnoća da je string dužine k zajednički podniz za sve odgovarajuće sufikse ulaznih stringova. Indeks k prolazi skupom $\{1, 2, \ldots, l_{\max}(v)\}$, tj. od dužine najkraćeg mogućeg nepraznog zajedničkog podniza do dužine najdužeg mogućeg, što je u ovom slučaju veličina sufiksa najkraćeg ulaznog stringa. Motivacija za upotrebu ovakvog jednostavnog (netežinskog) sumiranja proističe iz sljedećih zapažanja:

1. Tačna dužina rezultujućeg podniza nije unaprijed poznata. U slučaju HP heuristike predložene u radu [67], autori heuristički određuju odgovarajuću vrijednost k, za svaki nivo pretrage bima.

2. Sumiranje po svim k pruža uvid u ukupni "potencijal" čvora v; dobijena suma definiše mjeru kojom se poredi perspektivnost različitih "nastavljanja" na istom nivou BS drveta.

3. Dodjeljivanje različitih težina za različite dužine k bi podrazumijevalo neki oblik predviđanja očekivane dužine, što u slučaju polinomijalne raspodjele nije nimalo trivijalan zadatak.

Konačno, GMPSUM statistika se dobija kao konveksna kombinacija oblika

$$\mathrm{GMPSUM}(v, \lambda) := \lambda \cdot \mathrm{GM}(v) + (1 - \lambda) \cdot \mathrm{PSUM}(v), \tag{4.9}$$

gdje je $\lambda \in [0, 1]$ *parametar strategije*. Kako statistike GM i PSUM imaju komplementaran fokus jer bilježe različite aspekte potencijalnog produženja, smisleno ih je obje koristiti, tj. birati parametar $\lambda \in (0, 1)$. Veće vrijednosti λ se uzimaju kada je ulazni skup stringova izbalansiran u pogledu generalne raspodjele vjerovatnoća stringa, jer je tada statistika GM bolji indikator. S druge strane, statistika PSUM ima tendenciju boljeg opisa ulaznih stringova koji pokazuju značajnu mjeru odstupanja u odnosu na generalnu distribuciju vjerovatnoća stringa, što upućuje na biranje manjih vrijednosti λ.

Izračunavanje GM statistike zahtjeva vrijeme veličine $O(n \cdot m)$. Ovo se može zaključiti iz (4.7), gdje je najzahtjevniji dio iterativni prolazak kroz sve simbole alfabeta $\mathbb{N}_n$ i nalaženje minimalne frekvencije učestalosti simbola u svih m ulaznih stringova ($\mu_g(\cdot)$ i $\sigma_g(\cdot)$ imaju istu vremensku kompleksnost). Treba primijetiti da se broj pojavljivanja svakog simbola u odnosu na sve moguće sufikse svih m ulaznih stringova računa unaprijed, prije nego što se počne sa pretragom bima; ovi podaci se skladište u odgovarajući trodimenzionalni niz, vidjeti [34]. U najgorem slučaju, kompleksnost u ovom koraku je $O(n \cdot m \cdot l)$, gdje je l dužina najdužeg stringa iz skupa S. Izračunavanje PSUM statistike na osnovu (4.8) zahtjeva vrijeme veličine $O(l_{\min} \cdot m)$ (zbog definicije $l_{\max}(\cdot)$), gdje je $l_{\min}$ dužina najkraćeg stringa iz skupa S. Slično kao za GM, računanje matrice vjerovatnoća P se izvršava samo na početku i ima računsku kompleksnost $O(l \cdot l)$, vidjeti [57]. Na kraju, ukupna računska kompleksnost GMPSUM heuristike je $O((n + l_{\min}) \cdot m)$. To znači da je ukupna računska kompleksnost pretrage bima proizvod kompleksnosti GMPSUM heuristike i broja poziva ove heuristike $O(l_{\min} \cdot \beta \cdot n)$. Uočava se da je broj pozivanja GMPSUM heuristike jednak broju kreiranih čvorova prilikom trajanja pretrage bima. Umjesto nepoznate dužine LCS-a, tj. broja BS nivoa, koristi se $l_{\min}$, kao gornja granica. Dakle, BS vođen heuristikom GMPSUM se izvršava u $O(l_{\min} \cdot \beta \cdot n \cdot m \cdot (n + l_{\min}))$ vremenu, ako se ne koristi filtriranje. U slučaju

filtriranja, potrebno vrijeme je $O(\beta \cdot k_{\text{filter}} \cdot m)$ za svako nivo BS, što ukupno daje $O(l_{\min} \cdot \beta \cdot k_{\text{filter}} \cdot m)$ potrebnog vremena za izvršavanje filtriranja u okviru BS. Sveukupno, BS vođen heuristikom GMPSUM sa (ograničenim) filtriranjem zahtjeva $O(l_{\min} \cdot \beta \cdot m \cdot (k_{\text{filter}} + n^2 + n \cdot l_{\min}))$ vremena, što je polinomijalne složenosti.

Za potrebe poređenja performansi algoritama u zadanom vremenskom okviru, potrebna je modifikacija BS iz Algoritma 1. Tako se dobija varijanta BS pod nazivom TRBS (TRBS je skraćenica od Time Restricted Beam Search), koja se zasniva na dinamičkom adaptiranju širine bima u zavisnosti od pokazanog progresa ostvarenog prilikom njegovog prolaska kroz nivoe.

Slično kao i "standardna" verzija iz Algoritma 1, TRBS je parametrizovan sa instancom problema koju rješava, heurističkom funkcijom h kojom se vrši usmjeravanje i faktorom filtriranja k_{filter}. Dodatno, vrijednost parametra β koji je određivao konstantnu širinu bima sada postaje samo inicijalna vrijednost. Cilj je postići da vrijeme izvršavanja bude blisko ciljnom vremenu $t_{\max}$, koje je sada eksplicitno navedeno kao ulazni podatak. Na kraju svakog glavnog iterativnog koraka (nivoa), ako je $t_{\max} < +\infty$, tj. ako je uključeno vremensko ograničenje, širina bima za sljedeći nivo se određuje na sljedeći način:

1. Neka je t_{iter} vrijeme potrebno za izvršavanje trenutne iteracije.

2. Procjenjuje se broj preostalih nivoa uzimanjem maksimuma donjih granica za podinstance indukovane čvorovima iz V_{ext}. Preciznije,

$$LB_{\max}(V_{\text{ext}}) := \max_{(v,a) \in V_{\text{ext}} \times \mathbb{N}_n} \min_{i \in \{1,2,\ldots,m\}} C_a(S[\boldsymbol{\theta}^v])_i . \tag{4.10}$$

   Dakle, za svaki čvor $v \in V_{\text{ext}}$ i svaki simbol $a \in \mathbb{N}_n$, razmatra se minimalna frekvencija pojavljivanja ovog simbola u svim stringovima iz skupa sufiksnih stringova $S[\boldsymbol{\theta}^v]$ i među njima se bira ona koja je maksimalna. Drugačije rečeno, $LB_{\max}$ vrijednost je bazirana na obuhvatanju svih zajedničkih podnizova za koje je učestalost pojavljivanja pojedinačnog simbola što veća moguća. U literaturi, ova procedura je poznata pod nazivom *Long-run* [51].

3. Neka je t_{rem} preostalo vrijeme izvršavanja u cilju završavanja u vremenu $t_{\max}$.

4. Neka je $\overline{t_{\text{rem}}} := t_{\text{iter}} \cdot LB_{\max}(V_{\text{ext}})$ očekivano preostalo vrijeme ukoliko bi se nastavilo sa trenutnom širinom bima i ako bi, na svakom nivou, potrošeno vrijeme ostalo isto kao potrošeno vrijeme za trenutni nivo.

5. U zavisnosti od razlikovanja između aktuelnog i očekivanog preostalog

vremena, širina bima za naredni nivo se modifikuje u skladu sa formulom:

$$\beta \leftarrow \begin{cases} \lfloor \beta \cdot 1,2 \rfloor & \text{if } t_{\text{rem}}/\overline{t_{\text{rem}}} > 1,1; \\ \min(100, \lfloor \beta/1,2 \rfloor) & \text{if } t_{\text{rem}}/\overline{t_{\text{rem}}} < 0,9; \\ \beta & \text{inače.} \end{cases}$$

U opisanoj šemi, pragovi koji su indikatori postojanja značajne razlike između vrijednosti t_{rem} i $\overline{t_{\text{rem}}}$ (vrijednosti $1,1$ i $0,9$) postavljeni su empirijski. Ista napomena se odnosi na korektivni faktor $1,2$, kojim se modifikuje širina bima. Postoje efikasnije ocjene dužine LCS od $LB_{\max}$; međutim, prednost ove ocjene je njena jednostavna izračunljivost. Takođe, iako se može desiti da u ranim fazama algoritma ova ocjena poprilično odstupa od stvarne vrijednosti LCS, sa razvojem algoritma se ona približava ovoj vrijednosti. Ovo omogućava TRBS da "glatko" ažurira očekivano vrijeme izvršavanja u željeno vrijeme izvršavanja. Ovaj pristup ne pokušava da iznova određuje širinu bima za naredni nivo, već adaptira postojeću širinu bima. Na taj način se izbjegavaju nagle promjene do kojih može doći zbog fluktuacija varijanse vremena izvršavanja nivoa. Iskustva iz preliminarnih eksperimenata su pokazala da predloženi pristup funkcioniše dobro u pogledu dostizanja željenog vremenskog ograničenja, pri tome ne narušavajući kvalitet rješenja. Naravno, koliko "blizu" se želi prići ovom vremenskom limitu zavisi od stvarne dužine LCS. Za rješenja male dužine, opisani postupak uglavnom i nije modifikovao parametar β, te je precjenjivao ostatak vremena, dovodeći do situacije da se koristi manje vremena nego što je to predviđeno.

Sada će biti izloženi rezultati poređenja opisanog algoritma sa ostalim algoritmima iz literature. Svi algoritmi su implementirani u C# i izvođeni na računarima sa konfiguracijom: Intel i9-9900KF CPUs sa @ 3.6 GHz i 64 Gb RAM memorije i Microsoft Windows 10 Pro OS. Svaki eksperiment je sprovođen u single-threaded režimu (instrukcije su izvršavane sekvencijalno). Obuhvaćena su dva tipa eksperimenata:

- kratkoročni - podrazumijevaju scenario ograničenog vremena, tj. koristi se BS konfiguracija sa $\beta = 600$, izvršavana u cilju evaluacije kvaliteta vođenja svake heuristike prema perspektivnim regionima prostora pretrage.

- dugoročni - podrazumijevaju scenario fiksiranog vremena (900 sekundi) u kojem se poredi TRBS vođen heuristikom GMPSUM sa ostalim efikasnim metodima iz literature.

U izvedenim eksperimentima korišćeni su svi relevantni skupovi testnih instanci iz literature. Konkretno, podrazumijevaju se sljedeći skupovi testnih instanci:

- Skupovi testnih instanci RAT, VIRUS i RANDOM sastavljenih od po 20 pojedinačnih instanci su dobro poznati u literaturi [85]. Prva dva skupa imaju biološki kontekst i potiču iz NCBI baze podataka. Instance trećeg skupa su slučajno generisane. Dužina ulaznih stringova u ovim skupovima instanci je 600, a sami stringovi su sastavljeni od simbola iz alfabeta veličine 4 i 20 simbola.

- Skup testnih instanci ES, uveden u [35], sadrži slučajno generisane ulazne stringove čija dužina varira od 1000 do 5000, dok veličina alfabeta ima raspon od 2 do 100. Ovaj skup se sastoji od 12 grupa instanci.

- Skup testnih instanci BB, uveden u [12], drugačiji je od ostalih, jer su ulazni stringovi svake instance generisani tako da postoji visoka korelacija između njih. Naime, najprije je slučajno generisan bazni string, a zatim se na osnovu njega generišu svi ostali ulazni stringovi primjenom operacija editovanja. Ovaj skup se sastoji od 8 grupa, pri čemu se svaka sastoji od 10 pojedinačnih instanci.

- Skup testnih instanci BACTERIA, uveden u [31], sadrži podatke iz stvarne upotrebe, korišćene u kontekstu razmatranja Generalized Constrained Longest Common Subsesequence problema. Ove instance se koriste na način da se ignorišu svi šabloni koji nameću ograničenja. Ovaj skup se sastoji od 35 pojedinačnih instanci.

- Pored navedenih, biće posmatran i novi skup instanci POLY. Ovaj skup instanci sadrži instance čiji su stringovi generisani tako da je raspodjela vjerovatnoća simbola iz $\mathbb{N}_n$ saglasna sa polinomijalnom raspodjelom sa zadatim vjerovatnoćama $p_1, p_2, \ldots, p_n$; vidjeti [53] za informaciju o načinu uzorkovanja iz ove raspodjele vjerovatnoća. Za vjerovatnoće koje određuju polinomijalnu raspodjelu stavlja se $p_i := \frac{1}{2^i}, i \in \{1, 2, \ldots, n-1\}$ i $p_n := 1 - \sum_{i=1}^{n-1} \frac{1}{2^i}$. Ovim je postignuta nebalansiranost frekvencija pojavljivanja različitih simbola u generisanim ulaznim stringovima. Skup testnih stringova POLY sadrži 10 instanci za svaku kombinaciju dužine $l \in \{100, 500, 1000\}$ ulaznih stringova i broja $m \in \{10, 50\}$ ulaznih stringova, što daje ukupno 60 instanci.

Svi razmotreni algoritmi koriste efikasnu BS komponentu. U cilju testiranja kvaliteta novouvedene GMPSUM heuristike za evaluaciju parcijalnih rješenja na svakom nivou BS, izvršeno je poređenje sa ostalim heurističkim funkcijama iste namjene koje su predložene u literaturi: EX [33], POW [87] i HP [67]. Dobijene 4 BS varijante biće označene redom sa BS-GMPSUM, BS-EX, BS-POW i BS-HP. U kratkoročnom tipu eksperimenata, sve 4 varijante imaju

iste postavke parametara - $\beta = 600$ i $k_{\text{filter}} = 100$, kako bi se postiglo da im je na raspolaganju ista količina resursa. U dugoročnom tipu eksperimenata, predloženi TRBS vođen novom heuristikom GMPSUM (u nastavku notiran sa TRBS – GMPSUM) se poredi sa najprestižnijim pristupom iz literature A* + ACS [34]. Konkretno, dva algoritma se porede na sljedeći način:

- Za A* + ACS, rezultati iz skupa testnih instanci RANDOM, VIRUS, RAT, Es i BB su preuzeti iz rada [34]. Oni su dobijeni sa ograničenjem vremena izračunavanja na 900 sekundi po izvršenju. Za skupove testnih instanci POLY i BACTERIA, primjenjena je originalna implementacija od A* + ACS sa vremenskim ograničenjem od 900 sekundi na već pomenutom računaru.

- TRBS – GMPSUM je primjenjen sa vremenskim ograničenjem izračunavanja od 600 sekundi po izvršenju, za sve instance skupova testnih instanci RANDOM, VIRUS, RAT, Es i BB. Razlog redukovanja ovog vremena je korišćenje računara sa bržim procesorom od onoga koji je korišćen pri testiranju u radu [34]. Za skupove testnih instanci POLY i BACTERIA vremensko ograničenje ostaje 900 sekundi. U pogledu ograničenog filtriranja, koristi se ista postavka ($k_{\text{filter}} = 100$), kao i za kratkoročni tip eksperimenta.

Kao što je ranije istaknuto, vrijednosti parametara β (širina bima) i k_{filter} (filtriranje dominiranih čvorova) su usvojene iz [33] (za kratkoročna izvršavanja) i [34] (za dugoročna izvršavanja). Za podešavanje parametra strategije λ, koji kontroliše uticaj obje statistike iz GMPSUM heuristike, provjeravane su vrijednosti $\lambda \in \{0,\ 0,25,\ 0,5,\ 0,75,\ 1\}$. Pokazalo se da se najbolje performanse dobijaju u slučaju vrijednosti: $\lambda = 0$, za BB; $\lambda = 0,5$, za VIRUS i BACTERIA; $\lambda = 0,75$, za RANDOM, RAT i POLY; $\lambda = 1$ za ES. Iste vrijednosti parametra λ se koriste i u slučaju dugoročnog tipa eksperimenata. Pregled postavki svih parametara algoritama BS-GMPSUM i TRBS-GMPSUM dat je u Tabelama 4.2 i 4.3.

Tabela 4.2: Vrijednosti parametara β i k_{filter} za sve skupove testnih instanci

Parametar	**BS-GMPSUM**	**TRBS-GMPSUM**
β	600	20,000
k_{filter}	100	100

Tabela 4.3: Vrijednosti parametra λ za svaki skup testnih instanci

Skup testnih instanci	**BS-GMPSUM**	**TRBS-GMPSUM**
BB	0	0
VIRUS & BACTERIA	0.5	0.5
RANDOM & RAT & POLY	0.75	0.75
ES	1.0	1.0

Kompletni rezultati su dostupni na https://github.com/milanagrbic/LCSonNuD. Pregled rezultata dobijenih u eksperimentima kratkoročnog tipa predstavljen je u Tabeli 4.4.

Tabela 4.4: Pregled rezultata u eksperimentima kratkoročnog tipa

Skup testnih instanci		**BS-Ex**			**BS-Pow**			**BS-Hp**			**BS-Gmpsum**		
Ime	**#**	$\overline{len}$ **(s)**	**# b.**	$\bar{t}$	$\overline{len}$ **(s)**	**# b.**	$\bar{t}$	$\overline{len}$ **(s)**	**# b.**	$\bar{t}$	$\overline{len}$ **(s)**	**# b.**	$\bar{t}$
Random	20	108.9	16	2.7	108.1	6	1.4	108.15	6	1.1	108.95	16	6.7
RAT	20	102.8	13	2.6	101.6	4	1.2	100.95	2	0.9	102.9	14	5.5
VIRUS	20	115.85	11	2.6	114.1	6	1.5	115.35	6	1.1	116.3	17	7.4
BB	8	407.13	2	8.5	430.13	6	6.3	422.94	4	3.6	424.86	5	26.9
Es	12	242.18	8	23	241.51	0	15.6	241.14	0	13.8	242.12	4	118.8
Poly	6	232.67	0	5.6	232.27	0	3.3	231.53	0	2.7	233.02	6	6.7
Bacteria	35	809.97	12	14.7	814.86	15	8.2	830.69	22	7.9	832.09	18	29.3
Ukupno	121		62			37			40			80	

Tabela 4.4 prikazuje rezultate na način da se svaki njen red odnosi na odgovarajući skup testnih instanci. Značenje njenih kolona je sljedeće: Prva kolona sadrži ime skupa testnih instanci, dok druga kolona predstavlja broj instanci, tj. broj grupa instanci u posmatranom skupu; zatim slijede 4 bloka kolona, po jedan za svaku razmatranu BS varijantu. Prva kolona u svakom bloku prikazuje prosječni kvalitet dobijenih rješenja $\overline{len(s)}$ svih instanci iz posmatranog testnog skupa; druga kolona u bloku prikazuje broj instanci, tj. grupa instanci **# b.**, za koje je posmatrana BS varijanta došla do najboljeg rješenja; treća kolona u bloku pruža informaciju o prosječnom vremenu izvršavanja u sekundama $\bar{t}$ svih instanci iz posmatranog testnog skupa.

Mogu se izvesti sljedeći zaključci:

- Kada su u pitanju skupovi instanci RANDOM i Es, u kojima su ulazni stringovi generisani uniformno i nezavisno jedni od drugih, već od ranije je poznato da vođenje heuristikom Ex pokazuje najbolje performanse. Ipak, uočava se da BS-GMPSUM ima slične performanse kao BS-Ex, a da je očito mnogo bolji od ostale dvije BS varijante.

- U slučaju kvazi-slučajnih instanci iz skupova instanci VIRUS i RAT, BS-GMPSUM počinje da pokazuje svoj snagu dajući najkvalitetnija rješenja u 31 od 40 slučajeva. Druga najbolja varijanta je BS-Ex, koja i dalje radi dobro, te uspijeva da pruži najkvalitetnija rješenja u 24 od 40 slučajeva.

- Za skup Bb instanci, u kojem su ulazni stringovi generisani tako da postoji izražena korelacija između njih, efikasnost Gmpsum je u rangu najbolje varijante BS-Pow.

- Za skup instanci Bacteria, BS-Gmpsum je našao najbolja rješenja u 18 od 35 grupa, što je blago inferiorno u odnosu na BS-Hp, sa 22 najboljih rješenja, a superiorno u odnosu na varijante BS-Ex (12 slučajeva) i BS-Pow (15 slučajeva). U pogledu prosječnog kvaliteta dobijenih rješenja, BS-Gmpsum je najbolji od svih razmatranih pristupa.

- Za skup instanci Poly, sastavljen od ulaznih stringova čiji su simboli raspoređeni po polinomijalnoj neuniformnoj raspodjeli, BS-Gmpsum očigledno nadmašuje sve ostale konkurente. Štaviše, BS-Gmpsum uspijeva da nađe najbolja rješenja za svih 6 grupa instanci, a ima i najveću vrijednost prosječnog kvaliteta dobijenih rješenja.

- Ukupno, BS-Gmpsum nalazi najbolja rješenja u 80 (od 121) instanci ili grupa instanci. Druga najbolja varijanta je BS-Ex, koja postiže najbolje rezultate u 62 slučaja. U kontrastu sa njima, varijante BS-Hp i BS-Pow su inferiorne. Zaključak je da BS-Gmpsum posebno dobro funkcioniše u kontekstu različitih vjerovatnoća simbola alfabeta. To ovu varijantu čini najlogičnijim izborom u situaciji kada ima malo informacija o distribuciji vjerovatnoća razmatranog skupa instanci.

- Ukupno, vremena izvršavanja sve 4 BS varijante su uporediva. Najbrža varijanta je BS-Hp, dok BS-Gmpsum zahtijeva nešto više vremena od ostalih varijanti, jer koristi heuristiku koja kombinuje dvije funkcije.

Pregled rezultata dobijenih u eksperimentima dugoročnog tipa predstavljen je u Tabeli 4.5, gdje se performanse najboljeg algoritma A* + ACS porede sa učinkom TRBS-Gmpsum. Kako su skupovi testnih instanci isti kao u slučaju eksperimenata kratkoročnog tipa, prve dvije kolone ove table su iste kao prve dvije kolone Tabele 4.4. Zatim slijede dva bloka kolona, koji redom prezentuju rezultate rada A* + ACS i TRBS-Gmpsum, u pogledu prosječnog kvaliteta dobijenih rješenja $\overline{len(s)}$ svih instanci iz posmatranog testnog skupa i broja instanci (ili grupa instanci) za koje je odgovarajući algoritam postigao najbolje rješenje (**# b.**).

Tabela 4.5: Pregled rezultata u eksperimentima dugoročnog tipa

Skup testnih instanci		A* + ACS		TRBS-GMPSUM 600 s/900 s	
Ime	**#**	$\overline{len}$ **(s)**	**# b.**	$\overline{len}$ **(s)**	**# b.**
Random	20	109.9	20	109.7	16
RAT	20	104.3	17	104.4	18
VIRUS	20	117.0	14	117.3	19
BB	8	412.81	3	430.28	6
Es	12	243.82	9	243.73	4
Poly	6	234.13	4	234.23	5
Bacteria	35	829.26	10	862.63	33
Ukupno	121		77		101

Na bazi dobijenih rezultata za eksperimente dugoročnog tipa mogu se izvesti sljedeći zaključci:

- Za skupove instanci RANDOM i Es, A* + ACS je, kao što je i očekivano, bolji od TRBS-GMPSUM, u pogledu broja postignutih najboljih rješenja. Međutim, kada se porede prosječne performanse, razlika je minimalna: 109.9 protiv 109.7 za RANDOM skup testnih instanci, odnosno 243.82 protiv 243.73 za Es skup testnih instanci.

- U kontekstu skupa testnih instanci RAT i VIRUS, TRBS-GMPSUM je za nijansu bolji od A* + ACS. Ovo se ispoljava kako za brojeve dobijenih najboljih rješenja tako i za prosječne performanse algoritama.

- Za skup testnih instanci BB, TRBS-GMPSUM značajno nadmašuje A* + ACS. U 6 od 8 grupa instanci daje bolji prosječni kvalitet dobijenih rješenja, dok A* + ACS to čini samo za 3 grupe.

- Isto vrijedi za skup testnih instanci BACTERIA. Za ovaj skup instanci, TRBS-GMPSUM postiže najbolja rješenja u 33 od 35 instanci, za razliku od samo 10 najboljih rješenja u slučaju A* + ACS. Štaviše, prosječni kvalitet dobijenih rješenja je značajni veći u korist TRBS-GMPSUM, naime, 862.63 protiv 829.26.

- Konačno, performanse oba pristupa za skup testnih instanci POLY se ne razlikuju značajno.

- Ukupno, zaključak je da TRBS-GMPSUM uspijeva da ostvari najbolja rješenja u 101 od 121 slučaja, dok A* + ACS to postiže samo u 77 slučajeva. Razlog tome je što TRBS-GMPSUM pruža konzistentan kvalitet rješenja za instance karakterisane različitim distribucijama vjerovatnoća simbola alfabeta. Stoga, može se iskazati da je TRBS-GMPSUM novi najprestižniji algoritam za LCS problem.

Sveukupno, za instance (ili grupe instanci) iz skupova testnih instanci RANDOM i Es, A* + ACS pokazuje respektabilne performanse zbog pretpostavljene

slučajnosti tih instanci. I pored toga, učinak TRBS-GMPSUM za ovaj tip instanci nije zanemarljiv. Slaba strana A* + ACS postaje vidljiva kada se primjenjuje na instance koji nisu uniformno generisane. U 40 slučajeva sa kvazi-slučajnim ulaznim stringovima (skupovi testnih instanci RAT i VIRUS), TRBS-GMPSUM nalazi najbolja rješenja u 37 slučajeva, dok A* + ACS to postiže za 31 slučaj. Kada su ulazni stringovi izraženo korelisani (skup testnih instanci BB), TRBS-GMPSUM je znatno nadmoćniji od A* + ACS. Ova tendencija postaje još izraženija u slučaju skupova testnih instanci POLY i BACTERIA. Opšti utisak je da je TRBS-GMPSUM dobro "uštimovan" za rad sa širokim rasponom različitih vrsta instanci. Povrh toga, uzimajući u obzir instance poznate iz literature (80 instanci/grupa), TRBS-GMPSUM varijanta je u stanju da dobije nove najprestižnije rezultate u 13 slučajeva. Detaljnije obrazloženje je dato u narednom razmatranju.

Tabele kojima se prezentuju najprestižniji rezultati imaju sljedeću strukturu: Prva kolona sadrži ime odgovarajućeg skupa testnih instanci, dok sljedeće dvije kolone redom identifikuju instance (u slučaju skupova RAT i VIRUS) i grupe instanci (u slučaju skupova BB i Es). Nakon toga, tu su dvije kolone koje pružaju uvid u najbolja rješenja poznata iz literature. Prva od ovih kolona navodi najbolji rezultat, a druga algoritam koji je prvi postigao taj rezultat. Dalje, navedene su kolone kojima su obuhvaćeni rezultati redom za BS-EX, BS-POW, BS-HP i BS-GMPSUM u slučaju eksperimenata kratkoročnog tipa, odnosno rezultati redom za A* + ACS i TRBS-GMPSUM u slučaju eksperimenata dugoročnog tipa. Vrijeme izvršavanja dato je samo za kratkoročni tip, jer je za dugoročni tip ono unaprijed fiksirano.

Za eksperimente kratkoročnog tipa (Tabela 4.6), BS-GMPSUM je uspio da pronađe nove najbolje rezultate u 17 slučajeva. Ovo uključuje čak i 4 slučaja u okviru skupa testnih instanci Es, gdje su instance generisane slučajno u skladu sa uniformnom raspodjelom simbola alfabeta. Posebno vrijedno pažnje su 4 slučaja skupova testnih instanci RAT i VIRUS, gdje su prethodno poznata najbolja rješenja "popravljena" za dva simbola (vidjeti npr. skup testnih instanci RAT, za $n = 4, m = 40$ i $l = 600$).

Tabela 4.6: Novi najbolji rezultati za instance iz literature u eksperimentima kratkoročnog tipa

Instanca (Grupa instanci)				Najbolji $len(s)$ iz literature		BS-Ex		BS-Pow		BS-Hp		BS-Gmpsum	
Skup testnih instanci	***n***	***m***	***l***	$len(s)$	**Alg.**	$len(s)$	***t***	$len(s)$	***t***	$len(s)$	***t***	$len(s)$	***t***
Rat	4	20	600	172	BS-Ex	172	2.3	170	0.9	168	0.5	**173**	2.5
Rat	4	40	600	152	BS-Ex	152	1.8	150	1	145	0.5	**154**	3.4
Rat	4	200	600	123	BS-Ex	123	2.7	123	0.7	122	0.8	**124**	9.9
Rat	20	20	600	54	BS-Ex	54	2.5	54	1.7	54	1.2	**55**	3.5
Rat	20	40	600	49	BS-Ex	49	3	49	1.1	49	1.2	**50**	4.6
Virus	4	25	600	194	BS-Ex	194	2.2	192	1.2	194	0.7	**195**	3.1
Virus	4	40	600	170	BS-Ex	170	2.2	170	1.2	169	0.9	**172**	3.8
Virus	4	60	600	166	BS-Ex	166	2.4	165	0.8	166	0.7	**168**	5.1
Virus	4	100	600	158	BS-Ex	158	2.3	155	1.2	158	0.9	**160**	7.8
Virus	4	150	600	156	BS-Ex	156	2.4	147	1.2	156	0.7	**157**	11
Virus	4	200	600	155	BS-Hp	154	2.6	148	1.4	155	1.2	**156**	14.8
Virus	20	40	600	50	BS-Ex	50	2.9	49	1.9	50	0.9	**51**	5.5
BB	2	100	1000	560.7	BS-Pow	536.6	6.1	560.7	5.7	558.9	1.9	**560.8**	23.7
ES	2	10	1000	615.06	BS-Ex	615.06	4.4	614.2	1.4	612.5	0.9	**615.1**	5.1
ES	10	50	1000	136.32	BS-Ex	136.32	3.9	135.52	2.1	135.22	1.4	**136.34**	9.9
ES	25	10	2500	235.22	BS-Pow	231.12	19.1	235.22	10.5	233.34	8	**235.58**	29
ES	100	10	5000	144.9	BS-Pow	144.18	91.9	144.9	75.9	143.62	71.6	**145.1**	185.4

S druge strane, za eksperimente dugoročnog tipa (Tabela 4.7), prethodno poznati najbolji rezultati su poboljšani u 14 slučajeva. Izuzetno poboljšanje je postignuto u okviru skupa testnih instanci Bb, gdje se može uočiti poboljšanje prethodno najboljeg rješenja za čak 7 simbola.

Što se tiče prezentacije rezultata za skupove testnih instanci Poly i Bacteria, oni su obuhvaćeni u naredne dvije tabele. Tabele koje prezentuju rezultate za skup testnih instanci Poly imaju sličnu strukturu kao već navedene tabele rezultata za skupove testnih instanci Rat, Virus, Bb i Es. Razlika je što su grupe instanci određene sa *n* (prva kolona), *m* (druga kolona) i *l* (treća kolona). Najbolji rezultati po grupi instance, tj. u odgovarajućem redu tabele su označeni podebljanim fontom.

Tabela 4.7: Novi najbolji rezultati za instance iz literature u eksperimentima dugoročnog tipa

Instanca (Grupa instanci)				Najbolji $len(s)$ iz literature		A* + ACS	TRBS-Gmpsum
Skup testnih instanci	***n***	***m***	***l***	$len(s)$	**Alg.**	$len(s)$	$len(s)$
Rat	4	20	600	174	A* + ACS	174	**175**
Rat	4	40	600	154	A* + ACS	154	**156**
Rat	20	25	600	52	A* + ACS	52	**53**
Virus	4	10	600	228	A* + ACS	228	**229**
Virus	4	15	600	206	A* + ACS	206	**207**
Virus	4	60	600	168	A* + ACS	168	**169**
Virus	4	80	600	163	A* + ACS	163	**164**
Virus	4	100	600	160	A* + ACS	160	**162**
Virus	4	150	600	157	A* + ACS	157	**158**
BB	2	100	1000	563.6	APS	547.1	**571.1**
BB	4	100	1000	390.2	APS	344.3	**391.8**
ES	2	10	1000	618.9	A* + ACS	618.9	**619.1**
ES	10	50	1000	137.5	A* + ACS	137.5	**137.6**
ES	25	10	2500	236.6	A* + ACS-dist	235	**238**

Rezultati za eksperimente kratkoročnog tipa za skup testnih instanci Poly dati su u Tabeli 4.8. Na osnovu ovih rezultata, ubjedljivi pobjednik je BS-

Gmpsum, koji je dobio najbolji prosječni kvalitet dobijenih rješenja za svih 6 grupa instanci. Ovo pokazuje primat BS-Gmpsum u odnosu na preostale tri heuristike kada je u pitanju ovaj skup instanci. Kao što je već napomenuto, razlog ovakve situacije je što su instance iz ovog skupa uzorkovane u skladu sa polinomijalnom raspodjelom čije vjerovatnosne težine nisu ujednačene kao u slučaju uniformne raspodjele.

Tabela 4.8: Rezultati eksperimenata kratkoročnog tipa za skup testnih instanci Poly

Grupa instanci			**BS-Ex**		**BS-Pow**		**BS-Hp**		**BS-Gmpsum**	
n	***m***	***l***	$len(s)$	***t***	$len(s)$	***t***	$len(s)$	***t***	$len(s)$	***t***
4	10	100	43.2	0.5	43.2	0.3	43.1	0.3	**43.3**	0.1
4	10	500	232.5	4.1	232.7	2.6	231.3	2.1	**233**	2.5
4	10	1000	470.7	8.8	470.1	5.4	467.3	4.2	**470.9**	10.3
4	50	100	35.7	0.6	35.6	0.4	35.5	0.3	**35.8**	0.3
4	50	500	201.4	6.1	200.8	3.5	200.4	3	**202.3**	6.2
4	50	1000	412.5	13.2	411.2	7.4	411.6	6.3	**412.8**	20.9

Rezultati za eksperimente dugoročnog tipa za skup testnih instanci Poly dati su u Tabeli 4.9. Može se primijetiti da u ovom slučaju TRBS-Gmpsum i A* + ACS imaju ujednačene performanse.

Tabela 4.9: Rezultati eksperimenata dugoročnog tipa za skup testnih instanci Poly

Grupa instanci			**A* + ACS**	**TRBS-Gmpsum**	
n	***m***	***l***	$len(s)$	$len(s)$	***t***
4	10	100	**43.4**	**43.4**	580.7
4	10	500	**234.3**	**234.3**	890.5
4	10	1000	**473.9**	473.4	896.2
4	50	100	**35.9**	**35.9**	83.6
4	50	500	203	**203.5**	883.8
4	50	1000	414.3	**414.9**	892.3

Tabela 4.10: Rezultati eksperimenata kratkoročnog tipa za skup testnih instanci BACTERIA

Instanca				BS-Ex		BS-Pow		BS-Hp		BS-Gmpsum	
n	m	$l_{\min}$	l	$len(s)$	t	$len(s)$	t	$len(s)$	t	$len(s)$	t
4	383	610	1553	256	31.1	252	13.9	**279**	16.8	271	94.1
4	3	1458	1458	**1365**	2.1	**1365**	1	**1365**	1.8	**1365**	7.7
4	33	1349	1577	610	17.6	605	10.6	**755**	10.8	689	36.5
4	106	1252	1520	503	25.1	483	12	**515**	12.2	514	61.8
4	2	1502	1502	**1499**	0	**1499**	0	**1499**	0	**1499**	0.1
4	12	1274	1413	**659**	13.3	636	8.5	627	6.9	**659**	18.9
4	15	1302	1515	598	13.3	602	8.5	655	7.7	**678**	20.7
4	13	1479	1557	811	15.8	752	10.1	**1061**	10	883	21.7
4	13	1308	1507	1037	17.6	**1039**	11.1	862	8.6	882	25.9
4	44	873	1543	493	16.3	473	9.3	470	7.8	**494**	29.6
4	4	1408	1530	1204	9	**1271**	6.3	**1271**	5.8	**1271**	15.9
4	173	1234	1847	502	34.7	463	15	**541**	18.3	525	97.5
4	13	1446	1551	681	14.5	713	9.5	**794**	8.6	785	22.2
4	88	1360	1545	583	27.3	570	13.8	**667**	15.1	601	67
4	2	1540	1548	**1522**	0.2	**1522**	0.1	**1522**	0.1	**1522**	0.3
4	3	1395	1424	**1141**	11.2	**1141**	6.8	**1141**	6.1	**1141**	15
4	4	1410	1488	886	9.8	**1123**	8	**1123**	6.7	**1123**	17.4
4	51	1266	1522	**681**	25.2	552	12.3	667	12	641	48.7
4	2	1461	1539	**1354**	0.9	**1354**	0.5	**1354**	1.7	**1354**	8.6
4	13	1246	1411	687	13	662	7.4	609	6.7	**699**	19.6
4	4	1434	1478	876	9.6	**1112**	8	**1112**	6.9	**1112**	16.3
4	18	1023	1438	464	11.9	468	7.6	458	5.8	**475**	14.2
4	2	1454	1460	**1431**	0.2	**1431**	0.1	**1431**	0.1	**1431**	0.3
4	8	1401	1533	1024	15.9	**1061**	9.8	858	7.5	864	18.8
4	33	990	1483	410	12.1	**492**	8.8	467	6.9	456	16.4
4	29	1422	1549	587	16.3	581	9.8	**634**	8.9	590	26.3
4	20	571	1394	**438**	9.6	405	5.4	401	4.5	431	11.7
4	96	1270	1565	516	24	467	11	**531**	12.3	522	55.5
4	10	1322	1455	**1026**	16	**1026**	9.7	796	7.1	**1026**	19.5
4	26	1334	1596	617	16.7	584	9.4	**640**	8.6	631	26.2
4	195	1345	1547	503	38.2	448	15.3	**537**	19.2	524	100.9
4	8	1454	1532	1221	16.4	**1241**	10	**1241**	8.6	**1241**	25.5
4	8	1359	1612	555	18.9	555	11.2	600	10.3	**627**	38.4
4	89	455	1587	**251**	11.3	214	4.7	233	4.8	239	18.2
4	2	1465	1469	**1358**	0.7	**1358**	0.4	**1358**	0.5	**1358**	6.8

Tabela 4.11: Rezultati eksperimenata dugoročnog tipa za skup testnih instanci BACTERIA

Instanca				A* + ACS	TRBS-GMPSUM	
n	m	$l_{\min}$	l	$len(s)$	$len(s)$	t
4	383	610	1553	265	**273**	887.2
4	3	1458	1458	**1365**	**1365**	810.9
4	33	1349	1577	670	**723**	899
4	106	1252	1520	518	**532**	897.1
4	2	1502	1502	**1499**	**1499**	0.1
4	12	1274	1413	665	**694**	899.7
4	15	1302	1515	680	**708**	899.6
4	13	1479	1557	842	**883**	899.5
4	13	1308	1507	870	**1043**	899.7
4	44	873	1543	**514**	501	897.3
4	4	1408	1530	1204	**1271**	898.4
4	173	1234	1847	520	**528**	895.6
4	13	1446	1551	732	**816**	899.6
4	88	1360	1545	557	**634**	897.9
4	2	1540	1548	**1522**	**1522**	0.3
4	3	1395	1424	**1141**	**1141**	899.7
4	4	1410	1488	1059	**1123**	899.4
4	51	1266	1522	659	**871**	898.9
4	2	1461	1539	**1354**	**1354**	851.9
4	13	1246	1411	716	**727**	899.6
4	4	1434	1478	1030	**1112**	899.2
4	18	1023	1438	481	**488**	898.4
4	2	1454	1460	**1431**	**1431**	0.3
4	8	1401	1533	1040	**1063**	899.2
4	33	990	1483	449	**510**	899.1
4	29	1422	1549	643	**661**	899
4	20	571	1394	**439**	432	899.7
4	96	1270	1565	529	**546**	897.2
4	10	1322	1455	**1026**	**1026**	899.7
4	26	1334	1596	654	**676**	899.3
4	195	1345	1547	514	**544**	894.2
4	8	1454	1532	1204	**1241**	898.3
4	8	1359	1612	624	**644**	897.9
4	89	455	1587	250	**252**	898.2
4	2	1465	1469	**1358**	**1358**	829.6

Kao i u slučaju skupa POLY, instance iz skupa testnih instanci BACTERIA koriste se prvi put u proučavanju LCS problema. One su originalno predložene pri razmatranju LCS problema sa ograničenjima [31]. Rezultati su prezentovani na isti način kao i ranije. Ovaj skup sadrži 35 instanci. Svaka linija Tabele 4.10 (za eksperimente kratkoročnog tipa) i svaka linija Tabele 4.11 (za eksperimente dugoročnog tipa) sadrži zasebnu instancu. U obje tabele kolone su određene sa: n (uvijek jednako 4), m (varira od 2 do 383), l_{min} (dužina najkraćeg ulaznog stringa) i l (dužina najvećeg ulaznog stringa). Najbolja rješenja su istaknuta boldovanim fontom. Rezultati dobijeni u eksperimentima kratkoročnog tipa upućuju na zaključak da BS-HP prednjači za ovaj skup instanci, jer dobija najbolje rješenje u 22 od 35 slučajeva. BS-GMPSUM zaostaje sa 18 najboljih rješenja, ali u pogledu prosječnog kvaliteta dobijenih rješenja BS-GMPSUM je nešto bolji od BS-HP. Kada su u pitanju rezultati eksperimenata dugoročnog

tipa, već je ustanovljeno da TRBS-Gmpsum u potpunosti nadmašuje A* + ACS. Razlika je ponekad očigledno primjetna, npr. za instancu broj 32 (četvrti red od kraja u tabeli) TRBS-Gmpsum je pronašao rješenje dužine 1241, dok, u istom vremenu računanja, A* + ACS pronalazi rješenje dužine 1204.

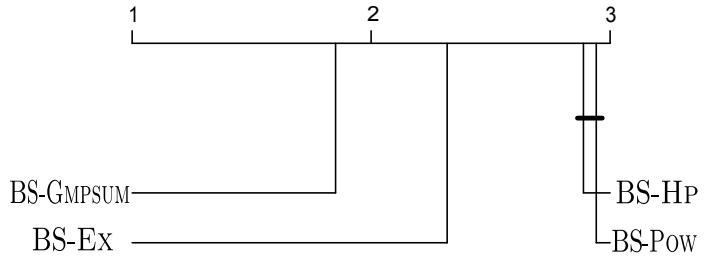
Sljedeći korak je potvrda statističke značajnosti dobijenih rezultata. U tom cilju, upotrebljen je Fridmenov test. Generalno, ovaj neparametarski statistički test predstavlja alternativu analizi varijanse i koristi se za otkrivanje eventualnih razlika u prosječnim vrijednostima $k \geqslant 3$ grupa, na osnovu N subjekata (tretmana) koji djeluju na svaku od tih grupa. Ako je x_i^j vrijednost $i-$tog tretmana na $j-$tu grupu, tada se vrijednostima iz skupa $\left\{x_i^j : j \in \{1, 2, \ldots k\}\right\}$ mogu dodijeliti *rangovi* r_i^j tako da se najvećoj među ovim vrijednostima pridružuje rang 1, sljedećoj po veličini rang 2, itd. Ako su dvije ili više vrijednosti jednake, njima se pridružuje isti rang jednak aritmetičkoj sredini pozicija koje zauzimaju u nizu vrijednosti sortiranom u opadajućem redoslijedu. Statistika Fridmenovog testa data je sa

$$\chi_F^2 := \frac{12N}{k(k+1)} \cdot \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right],$$

gdje je $R_j := \dfrac{\sum_i r_i^j}{N}$ prosječna vrijednost rangova $j-$te grupe, $j \in \{1, 2, \ldots, k\}$. Pod pretpostavkom da je tačna radna hipoteza da se grupe ne razlikuju značajno u prosječnim vrijednostima, tj. da su njihovi rangovi približno jednaki, statistika χ_F^2 ima približno χ_{k-1}^2 raspodjelu, sa $k-1$ stepeni slobode, ukoliko su N i k dovoljno velike vrijednosti (ova aproksimacija je dovoljno dobra za $N > 10$ i $k > 5$; za ostale vrijednosti N i k koriste se egzaktno izračunate kritične vrijednosti, vidjeti npr. [84]). Ukoliko se, sa datim pragom značajnosti α, radna hipoteza odbaci, tada se može primijeniti neki od post-hoc testova za utvrđivanje između kojih k grupa postoji statistički značajna razlika. U ovoj eksperimentalnoj evaluaciji koristi se *Nemenijev post-hoc test*. Ideja ovog testa je pronalaženje *kritične razlike* $\mathrm{CD} := q_\alpha \cdot \sqrt{\dfrac{k(k+1)}{6N}}$, gdje je $\alpha \in (0, 1)$ zadani prag značajnosti, a q_α kritična vrijednost koja se čita iz tablice studentizovane raspodjele raspona. Nakon toga, porede se rangovi; svake dvije grupe j_1 i j_2 za koje vrijedi $|R_{j_1} - R_{j_2}| \geqslant \mathrm{CD}$ se, u smislu datih tretmana, smatraju statistički značajno različitim.

Konkretno, posmatrane grupe su algoritmi za rješavanje LCS problema zasnovani na različitim BS varijantama, dok su subjekti svi skupovi testnih instanci kojima se tretiraju svi navedeni algoritmi (ukupno 121 grupa instanci). Sam postupak testiranja Fridmenovim testom, Nemenijevim post-hoc testom i prikaz dobijenih zaključaka izvršen je uz pomoć paketa scmamp softverskog okruženja R (pogledati [16] i [76]).

U svim slučajevima, zaključak testiranja Fridmenovim testom je da se hipoteza o ujednačenim performansama algoritama odbacuje. Stoga je smisleno izvršiti poređenje po parovima uz pomoć Nemenijevog post-hoc testa, radi utvrđivanja koji parovi algoritama ispoljavaju statistički značajnu razliku. Kritična razlika CD je računata u odnosu na prag značajnosti $\alpha = 0,05$. Rezultati izvršenog poređenja se jednostavno predstavljaju uz pomoć *dijagrama kritične razlike* ili CD dijagrama i oni su prikazani na Slici 4.2, odnosno Slici 4.3a. Za svaki algoritam, prosječna vrijednost njegovih rangova generiše poziciju ovog algoritma na segmentu. Ukoliko su pozicije dva algoritma udaljene za manje od CD, tada ne postoji statistički značajna razlika u njihovim performansama, što se na dijagramu notira stavljanjem horizontalne crtice koja spaja markere ovih algoritama.



Slika 4.2: CD dijagram za eksperimente kratkoročnog tipa

U slučaju eksperimenata kratkoročnog tipa, BS-GMPSUM je algoritam sa statistički značajnim najboljim performansama. BS-EX se nalazi na drugoj poziciji, dok razlika između performansi BS-HP i BS-POW nije statistički značajna. U slučaju eksperimenata dugoročnog tipa, najbolji prosječni rang ima TRBS-GMPSUM. Specijalno, za skup testnih instanci BACTERIA, razlika između TRBS-GMPSUM i A* + ACS je statistički značajna (Slika 4.3b). Za ostale skupove testnih instanci, razlike u performansama ova dva pristupa nisu statistički značajne.

(a) Sve instance (b) Skup testnih instanci BACTERIA.

Slika 4.3: CD dijagram za eksperimente dugoročnog tipa

Sveobuhvatno gledano, varijanta BS vođena novom heuristikom GMPSUM pokazala je konzistentnu efikasnost prilikom približnog rješavanja LCS pro-

blema. Kao kombinacija dvije komplementarne statistike GM i PSUM, čiji se uticaj kontroliše parametrom strategije, heuristika GMPSUM se može prilagoditi tako da za bilo koji skup testnih instanci produkuje kvalitetna rješenja. Ova njena prednost se i potvrdila u izvedenim eksperimentima, gdje je postignut visok prosječan kvalitet rješenja svih razmatranih instanci. Stoga, GMPSUM heuristika se s pravom može proglasiti za trenutno najrobustniju heuristiku u literaturi kada je u pitanju rješavanje LCS problema.

Glava 5

Zaključak

U većini naučnih disciplina, opisivanje strukture određenog skupa podataka predstavlja neprestani izazov postavljen pred svaku generaciju istraživača. Izbor matematičkog modela koji će interpretirati određenu strukturu je neophodan prvi korak u postupku demistifikacije pojave koja se želi objasniti i utvrđivanja obrazaca na osnovu kojih ona djeluje. Za početak, potrebno je izabrati matematičke objekte koji će reprezentovati posmatrane podatke i očuvati njihove bitne karakteristike. U slučaju podataka sekvencijalne prirode, stringovi (konačni nizovi elemenata) nameću se kao logičan izbor matematičkih predstavnika. Na taj način, ovakav skup podataka se "konvertuje" u familiju stringova i sva esencijalna pitanja u vezi strukture podataka se, u stvari, odnose na strukturu dobijene familije stringova. Nakon ove "inicijalizacije", potrebno je ustanoviti koji tip međusobne povezanosti stringova kao pojedinačnih objekata najviše oblikuje strukturu pripadne familije stringova kao grupe objekata. U ljudskoj je prirodi konstantno poređenje i evaluacija, jer to često pobuđuje intuiciju i dovodi do spoznaje o bitnim svojstvima. Zbog toga, poželjne su one vrste povezanosti familija stringova koje njihovo rangiranje i međusobno poređenje čini jednostavnim za implementaciju. Drugačije rečeno, pogodno je kvantifikovati strukturu svake familije stringova u pogledu tipa međusobne povezanosti njenih stringova. Time se generišu mjere sličnosti koje omogućavaju poređenje struktura familija stringova.

U ovoj doktorskoj tezi, razmatrane su mjere sličnosti familija stringova zasnovane na topološkim i vjerovatnosnim metodama. Dodatno, u obzir je uzeta i mjera sličnosti stringova određena dužinom najdužeg zajedničkog podniza i proučavan je optimizacioni aspekt pri rješavanju problema nalaženja ove mjere.

Od mjera sličnosti familija stringova zasnovanim na topološkim metodama, u tezi su ispitivane one mjere sličnosti bazirane na metodama i tehnikama istrajne (perzistentne) homologije, koja predstavlja dinamičku verziju simplicijalne homologije. Osnovna ideja je da se familiji stringova pridruži filtracija sastavljena od simplicijalnih kompleksa i na taj način formalizuje koncept "bliskosti", kao

oblika povezanosti koju stringovi date familije postepeno razvijaju prolaskom kroz datu filtraciju. Pomenuti tip povezanosti stringovi ispoljavaju kroz prizmu udaljenosti utvrđene odgovarajućim nivoom filtracije. Ključan momenat je praćenje onih podskupova familije stringova kod kojih povezanost postoji između svih sastavnih dijelova, ali ne i samih njih kao cjeline. Ovo odsustvo "kompletne" povezanosti stvara homološku klasu koja istrajava u određenom dijelu filtracije. Životni vijek svih homoloških klasa date familije stringova se registruje nalaženjem njenog bar-koda. Najpoznatija mjera sličnosti zasnovana na poređenju bar-kodova iste dimenzije je udaljenost uskog grla. Ova mjera, poznata u literaturi, koristi uparivanje linija bar-koda u cilju njihovog što boljeg preklapanja. Tako dobijeno uparivanje spaja one linije koje predstavljaju homološke klase relativno slične istrajnosti. Nedostatak ovog uparivanja leži u zanemarivanju kvalitativnih svojstava pripadnih homoloških klasa. U tezi je izložen originalni pristup, predložen u radu [70], koji ispravlja uočeni nedostatak. Konkretno, predložena je nova mjera sličnosti familija stringova utemeljena na udaljenosti uskog grla, sa modifikacijom kojom se prioritet daje kvalitativnom uparivanju linija bar-koda. Ono što definisanje nove mjere čini smislenim je tehnika razdvajanja radijusa simpleksa (Teorema 2.7.17), koja je takođe uvedena u radu [70]. Primjenom ove tehnike, izvodi se kontrolisano pomjeranje radijusa simpleksa, što za rezultat ima dobijanje bar-koda sličnog polaznom, ali koji ima linije sa različitim krajnjim tačkama, omogućavajući jednostavniju provjeru njihove kvalitativne saglasnosti. Na taj način, novouvedena mjera sličnosti nasljeđuje stabilnost koju ima udaljenost uskog grla, ali ima i kvalitativnu konzistentnost koju obezbjeđuje opisana modifikacija. Pored navedene mjere sličnosti, u tezi je registrovan pojam $\approx$ −ekvivalentnih simpleksa (simpleksa sa istim minimalnim skupom generatora) i u Teoremi 2.7.18 dokazana je "neutralnost" ove grupacije simpleksa u pogledu njihovog uticaja na strukturu bar-kod linija.

Od mjera sličnosti familija stringova zasnovanim na vjerovatnosnim metodama, u tezi su predstavljene one utemeljene na Kulbak-Lajblerovoj mjeri divergenciji (relativnoj entropiji). Ovaj pristup podrazumijeva da se string tretira kao stohastički proces, dok se data familija stringova shvata kao skup trening podataka, sastavljena od konačnih trajektorija ovog stohastičkog procesa. Osnovna ideja je da se datim trening podacima pridruži vjerovatnosna mjera koja će poslužiti kao pokazatelj strukture pripadne familije stringova. U tezi je predložena mjera sličnosti familija stringova koja poredi odgovarajuće vjerovatnosne mjere na principu relativne entropije. Efektivno, ovom mjerom se poredi "moć predviđanja" dvije vjerovatnosne mjere koje su definisane na istom prostoru vjerovatnoća, tako što se kvantifikuje "gubitak informacije" do kojeg dolazi kada se raspodjela određena jednom vjerovatnosnom mjerom zamjenjuje raspodjelom određenom drugom vjerovatnosnom mjerom. Nažalost, pomenute vjerovatnosne mjere u praksi nisu poznate, pa je potrebno izvesti njihovo sta-

tističko modelovanje. U tezi su detaljno obrazložena dva modela. Prvi model je vjerovatnosno sufiksno drvo utemeljen na frekvencionističkom pristupu, dok je drugi model hijerarhijski Pitman-Jorov proces utemeljen na Bejzovskom zaključivanju. Iako se ovi modeli zasnivaju na različitim konceptima, oni imaju isti cilj: ocijeniti vjerovatnoću da će se simbol alfabeta inicijalizovati na određenoj poziciji stringa, pod uslovom da su poznate sve realizacije stringa prije posmatrane pozicije. Nakon određivanja ovih vjerovatnoća, pravilom množenja vjerovatnoća se jednostavno dobijaju potrebne predviđajuće vjerovatnoće. Samo izračunavanje razmatrane mjere sličnosti uključuje uzorkovanje iz dobijene raspodjele vjerovatnoća. Time se kompletira u potpunosti vjerovatnosni orijentisan postupak zasnivanja ove mjere sličnosti familija stringova.

U tezi su razmotreni teoretski i praktični aspekti rješavanja problema nalaženja najdužeg zajedničkog podniza date familije stringova (LCS problema). Algoritmi na bazi dinamičkog programiranja se mogu iskoristiti za tačno rješavanje LCS problema, ali sa povećavanjem broja stringova njihova složenost postaje eksponencijalna, što ih čini neefikasnim. Efikasnost se može popraviti primjenom algoritama zasnovanih na optimizaciji. Od algoritama za približno rješavanje LCS problema, pretraga bima (BS) predstavlja pristup u kojem se ne izvodi kompletna pretraga za najboljim rješenjem, nego se koristi heuristika koja ovu pretragu usmjerava ka perspektivnijim regionima prostora pretrage. U tezi je izložena nova varijanta pretrage bima pod nazivom BS – GMPSUM, originalno predložena u radu [69]. Ova varijanta koristi novu GMPSUM heuristiku za rangiranje perspektivnosti čvorova u BS i utvrđivanje za koji od njih će postojeća parcijalna rješenja biti produžena na narednom nivou. Sama GMPSUM heuristika je izražena kao konveksna kombinacija dvije statistike: GM statistike i PSUM statistike. Ove statistike registruju relacije između ulaznih stringova odgovarajućih podproblema. I dok je GM statistika "odgovorna" za karakterisanje usaglašenosti ulaznih stringova podproblema sa pretpostavljenom raspodjelom vjerovatnoća simbola alfabeta, PSUM statistika bilježi vjerovatnoće da je string odgovarajuće dužine podniz svih ulaznih stringova podproblema. Komplementaran fokus ovih statistika čini GMPSUM heuristiku pogodnom za rješavanje LCS problema, kako u slučaju balansiranih instanci, tako i u slučaju nebalansiranih instanci sastavljenih od stringova čiji se simboli ravnaju po neuniformnoj polinomijalnoj raspodjeli. Radi potvrde ove argumentacije, u tezi je obavljeno upoređivanje BS – GMPSUM sa ostalim najprestižnijim BS varijantama za rješavanje LCS problema, uz pomoć metodologije opisane u radu [69]. Eksperimenti su sprovedeni na skupovima balansiranih testnih instanci standardnih za LCS problem, a takođe i na novim skupovima nebalansiranih testnih instanci. Rezultati eksperimenata za slučaj balansiranih instanci iz literature, pokazali su da je BS – GMPSUM ili u rangu ostalih varijanti ili ih nadmašuje u pogledu kvaliteta (dužina) najboljih rješenja. Ova razlika je još izraženija u rezultatima eksperimenata za slučaj (novih) nebalansiranih instanci iz literature, koji poka-

zuju superiornost $\mathrm{BS}-\mathrm{GMPSUM}$ u odnosu na ostale BS varijante. Ova superiornost je statistički značajna, što je potvrđeno kombinacijom Fridmenovog testa i Nemenijevog post-hoc testa. Razlog uočene nadmoćnosti $\mathrm{BS}-\mathrm{GMPSUM}$ je mogućnost "finog" podešavanja parametra strategije koji reguliše uticaj statistika GM i PSUM na ukupnu vrijednost GMPSUM heuristike.

Na kraju, nekoliko riječi o eventualnim budućim pravcima istraživanja razmatranih tema.

U vezi topoloških metoda, tu su sljedeći potencijalni zadaci:

- Značajno bi bilo dati odgovor na Pitanje 2.7.7, tj. naći uslove ekvivalentnosti filtracijskog izomorfizma i d_H-izomorfizma u okviru metričkog prostora $(S(2,l),d_H)$. Značaj je u tome što bi se, u slučaju afirmativnog odgovora, familija stringova, kao potprostor metričkog prostora $(S(2,l),d_H)$, uvijek mogla rekonstruisati na osnovu njoj pridružene filtracije.

- Implementacija priče o mjerama sličnosti familija stringova zahtjeva pronalaženje efikasnih algoritama za računanje radijusa i centara konačne familije generalizovanih stringova iz prostora $(S'(n,l),d_{GH})$. Takođe, opisanu tehniku razdvajanja radijusa simpleksa treba optimizovati, u smislu preciziranja mehanizma biranja simpleksa čiji se radijusi žele razdvojiti, tako da se minimizuje ukupan broj koraka potrebnih za razdvajanje radijusa svih simpleksa koji nisu $\approx-$ekvivalentni.

- Ispitivanje mogućnosti da se metodologija dobijanja nove mjere sličnosti familija stringova bazirane na Hamingovoj udaljenosti iskoristi i u slučaju LCS metrike. Konkretno, bilo bi korisno naći analogiju tehnike razdvajanja radijusa simpleksa u ovom slučaju.

- Utvrđivanje uloge $\approx-$ekvivalentnih simpleksa u opštoj postavci Čehove filtracije.

U vezi vjerovatnosnih metoda, tu su sljedeći potencijalni zadaci:

- Daljnje poboljšavanje efikasnosti vjerovatnosnog sufiksnog drveta putem podešavanja parametra sniženja na svakom nivou drveta.

- Ispitivanje efekta biranja drugačijih priornih raspodjela za hiperparametre sniženja i koncentracije kod hijerarhijskog Pitman-Jorovog procesa na stabilnost uvedene mjere sličnosti familija stringova.

U vezi optimizacionog aspekta rješavanja LCS problema, tu su sljedeći potencijalni zadaci:

- Implementacija nekog od predloženih vjerovatnosnih modela radi postizanja efekta bolje kalibrisanosti statistike PSUM, što bi rezultovalo time da heuristika GMPSUM ima veću izražajnu moć.
- Adaptacija heuristike GMPSUM radi stvaranja "hibridnog" A^* + ACS pristupa, a sve u cilju dodatnog poboljšavanja kvaliteta najboljih rješenja LCS problema.
- Primjenjivanje GMPSUM heuristike pri rješavanju problema bliskih LCS problemu. To su problemi u kojima se traži LCS, ali uz nametanje dodatnih ograničenja, npr. problem nalaženja najdužeg zajedničkog podniza koji je palindrom, problem nalaženja najdužeg zajedničkog podniza koji sadrži unaprijed zadani podniz, itd.

Bibliografija

[1] J. Albert, J. Hu, "Probability and Bayesian modeling", CRC press, 2020.

[2] S. Bacallado, S. Favaro, S. Power, L. Trippa, "Perfect sampling of the posterior in the hierarchical Pitman–Yor process", International Society for Bayesian Analysis, 2021.

[3] R. A. Baeza-Yates, R. Gavalda, G. Navarro, R. Scheihing, "Bounding the expected length of longest common subsequences and forests", Theory Comput. Syst., 32(4), pp. 435–452, 1999.

[4] D. Bakkelund, "An LCS-based string metric", University of Oslo, 2009.

[5] S. A. Barannikov, "The framed Morse complex and its invariants," Advances in Soviet Mathematics, 21, pp. 93-115, 1994.

[6] U. Bauer i M. Lesnick, "Induced matchings and the algebraic stability of persistence barcodes," Journal of Computational Geometry, 6(2), pp. 162-191, 2015.

[7] R. Beal, T. Afrin, A. Farheen, D. Adjeroh, "A new algorithm for the LCS problem with application in compressing genome resequencing data", BMC Genom., 17, 544, 2016.

[8] R. Begleiter, R. El-Yaniv, G. Yona, "On prediction using variable order Markov models", Journal of Artiflcial Intelligence Research 22, pp. 385-421, 2004.

[9] Y. Bengio, "Markovian models for sequential data", Neural computing surveys, 1999.

[10] L. Bergroth, H. Hakonen, T. Raita, "A survey of longest common subsequence algorithms", Proceedings of the SPIRE 2000—The 7th International Symposium on String Processing and Information Retrieval, Coruna, Spain, pp. 39-48, 2000.

[11] P. Billingsley, "Ergodic Theory and Information", New York: Wiley, 1965.

[12] C. Blum, M. J. Blesa, "Probabilistic beam search for the longest common subsequence problem", In Proceedings of the International Workshop on Engineering Stochastic Local Search Algorithms, Brussels, Belgium, 2007., Springer: 200, pp. 150-161, 2007.

[13] C. Blum, M. J. Blesa, M. López-Ibáñez, "Beam search for the longest common subsequence problem", Comput. Oper. Res. 2009, 36, pp. 3178–3186.

[14] C. Blum, P. Festa, "Longest common subsequence problems", Metaheuristics for String Problems in Bioinformatics, Wiley: Hoboken, NJ, USA, 2016, Chapter 3, pp. 45–60.

[15] P. Bubenik i J. A. Scott, "Categorification of persistent homology", Discrete and Computational Geometry, Vol. 51, pp. 600–627, 2014.

[16] B. Calvo, G. Santafé-Rodrigo, "Statistical comparison of multiple algorithms in multiple problems", The R Journal, Vol. 8/1, Aug. 2016.

[17] G. Carlsson, "Topology and data", Bulletin (New Series) of the American Mathematical Society, 2009.

[18] G. Carlsson, A. Zomorodian, A. Collins i L. Guibas, "Persistence barcodes for shapes", Eurographics Symposium on Geometry Processing, 2004.

[19] H. T. Chan, C. B. Yang, Y. H. Peng, "The generalized definitions of the two-dimensional largest common substructure problems", Proceedings of the 33rd Workshop on Combinatorial Mathematics and Computation Theory, Taipei, Taiwan, 13–14 May 2016, pp. 1–12.

[20] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas i S. Oudot, "Proximity of persistence modules and their diagrams," Research Report RR-6568, INRIA, 2008.

[21] F. Chazal i B. Michel, "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists", preprint 2021.

[22] C. Chen, L. Du, W. Buntime, "Sampling table configurations for the hierarchical Poisson-Dirichlet process", ECML PKDD 2011: Machine Learning and Knowledge Discovery in Databases pp. 296–311, 2011.

[23] S.F. Chen, J. Goodman, "An empirical study of smoothing tehnique for language modeling", Computer Science Group Harvard University Cambridge, Massachusetts, 1998.

[24] V. Chvatal i D. Sankoff, "Longest common subsequences of two random sequences", Journal of Applied Probability, Vol. 12, No. 2, pp. 306-315, 1975.

[25] D. Cohen-Steiner, H. Edelsbrunner i J. Harer, "Stability of persistence diagrams," Discrete and Computational Geometry, 37, pp. 103–120, 2007.

[26] T. M. Cover i J. A. Thomas, "Elements of Information Theory", Second edition, John Wiley and Sons, Inc., 2006.

[27] W. Crawley-Boevey, "Decomposition of pointwise finite-dimensional persistence modules", Journal of Algebra and Its Applications, Vol. 14, No. 5, 2015.

[28] V. Dančik, "Expected length of longest common subsequences", PhD thesis, Department of Computer Science, University of Warwick, September 1994.

[29] V. de Silva i V. Nanda, "Geometry in the space of persistence modules", Proceedings of the twenty-ninth annual symposium on Computational geometry, 2013.

[30] T. K. Dey i Y. Wang, "Computational topology for Data Analysis", Cambridge University Press, 2022.

[31] M. Djukanovic, A. Kartelj, D. Matic, M. Grbic, C. Blum, G. Raidl, "Solving the generalized constrained Longest Common Subsequence problem with many pattern strings", Technical Report AC-TR-21-008, AC, 2021.

[32] M. Djukanovic, G. R. Raidl, C. Blum, "Anytime algorithms for the longest common palindromic subsequence problem", Comput. Oper. Res. 2020, 114, 104827.

[33] M. Djukanovic, G. Raidl, C. Blum, "A Beam Search for the Longest Common Subsequence problem guided by a novel approximate expected length calculation", Proceedings of the LOD 2019—The 5th International Conference on Machine Learning, Optimization, and Data Science, Siena, Italy, 10–13 September 2019.

[34] M. Djukanovic, G. Raidl, C. Blum, "Finding longest common subsequences: new anytime A* search results", Appl. Soft. Comput. 2020, 95, 106499.

[35] T. Easton, A. Singireddy, "A large neighborhood search heuristic for the longest common subsequence problem", J. Heuristics, 14, pp. 271-283, 2008.

[36] H. Edelsbrunner i J. Harer, "Computational Topology: An Introduction", American Mathematical Society, 2010.

[37] H. Edelsbrunner i J. Harer, “Persistent homology - a survey,” Surveys on Discrete and Computational Geometry: Twenty Years Later, American Mathematical Society, 2008., pp. 257–282.

[38] R. Forman, "A user’s guide to discrete Morse theory", S´eminaire Lotharingien de Combinatoire 48, Article B48c, 2002.

[39] C.B. Fraser, "Subsequences and supersequences of strings", Ph.D. Thesis, University of Glasgow, Glasgow, UK, 1995.

[40] A. Gabadinho i G. Ritschard, "Analyzing state sequences with probabilistic suffix trees: The PST R package", Journal of Statistical Software, 2016.

[41] J.A. Gasthaus, "Hierarchical Bayesian nonparametric models for power-law sequences", PhD Thesis, 2020.

[42] J.A. Gasthaus i Y. W. Teh, "Improvements to the sequence memoizer", Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems, 2010.

[43] R. M. Gray, "Entropy and Information Theory", Second edition, Springer New York, 2011.

[44] D. Gusfield, "Algorithms on Strings, Trees, and Sequences", Computer Science and Computational Biology, Cambridge University Press: Cambridge, UK, 1997.

[45] R. W. Hamming, "Error detecting and error correcting codes," The Bell System Technical Journal, vol. 29, no. 2, pp. 147-160, April 1950.

[46] A. Hatcher, "Algebraic Topology", Cambridge University Press, 2001.

[47] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences", Communications of the Association for Computing Machinery, 18(6), pp. 341-343, 1975.

[48] D. S. Hirschberg, "Algorithms for the Longest Common Subsequence problem", Journal of the Association for Computing Machinery., 24(4), pp. 664-675, 1977.

[49] L. C. Hsu, "A unified approach to generalized Stirling numbers", Advances in Applied Mathematics 20, pp. 366-384, 1998.

[50] K. Huang, C. Yang, K. Tseng, "Fast algorithms for finding the common subsequences of multiple sequences", Proceedings of the ICS 2004—The 9th International Computer Symposium, Funchal, Portugal, 13–16 January 2004.

[51] T. Jiang, M. Li, "On the approximation of shortest common supersequences and longest common subsequences", SIAM J. Comput., 24, pp. 1122-1139, 1995.

[52] C. Kermorvant, P. Dupont, "Improved smoothing for probabilistic suffix trees seen as variable order Markov chains", Springer-Verlag Berlin Heidelberg, 2002.

[53] H. Kesten, N. Morse, "A property of the multinomial distribution", Ann. Math. Stat., 30, pp. 120-127, 1959.

[54] M. Kiwi, J. Soto, "On a speculated relation between Chvatal-Sankoff constants of several sequences", Combin. Probab. Comput., 18(4), pp. 517–532, 2009.

[55] M. Kiwi, M. Loebl, J. Matousek, "Expected length of the longest common subsequence for large alphabets", Adv. Math., 197(2), pp. 480–498, 2005.

[56] R. Kneser, H. Ney, "Improved backing-off for m-gram language modeling", International Conference on Acoustics, Speech, and Signal Processing, 1995.

[57] J. B. Kruskal, "An overview of sequence comparison: Time warps, string edits, and macromolecules", SIAM Rev., 25, pp. 201-237, 1983.

[58] S. Kullback, R. A. Leibler, "On information and sufficiency", The Annals of Mathematical Statistics, 22, pp. 79-86, 1951.

[59] D. A. Levin, Y. Peres, "Markov Chains and Mixing Times: Second Edition", American Mathematical Society, 2017.

[60] Y. Li, Y. Wang, Z. Zhang, Y. Wang, D. Ma, J. Huang, "A novel fast and memory efficient parallel MLCS algorithm for long and large-scale sequences alignments", Proceedings of the IEEE 32nd International Conference on Data Engineering, Helsinki, Finland, 16–20 May 2016, pp. 1170–1181.

[61] K. W. Lim, W. Buntime, C. Chen, L. Du, "Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes", International Journal of Approximate Reasoning, 2016.

[62] G. S. Lueker, "Improved bounds on the average length of longest common subsequences", Journal of the ACM, 56(3), pp. 1–38, May 2009.

[63] D. Maier, "The complexity of some problems on subsequences and supersequences", J. ACM, 25, pp. 322–336, 1978.

[64] M. Maltenfort, "New definitions of the generalized Stirling numbers", Aequationes mathematicae volume 94, pp. 169–200, 2020.

[65] G. Máté, A. Hofmann, N. Wenzel i D. W. Heermann, "A topological similarity measure for proteins", 2013.

[66] P. Minkiewicz, M. Darewicz, A. Iwaniak, J. Sokołowska, P. Starowicz, J. Bucholska, M. Hrynkiewicz, "Common amino acid subsequences in a universal proteome—relevance for food science", Int. J. Mol. Sci., 16, pp. 20748–20773, 2015.

[67] S. R. Mousavi, F. Tabataba, "An improved algorithm for the longest common subsequence problem". Comput. Oper. Res. 2012, 39, pp. 512–520.

[68] J. R. Munkres, "Elements of Algebraic Topology", CRC Press, 1993.

[69] B. Nikolic, A. Kartelj, M. Djukanovic, M. Grbic, C. Blum, G. Raidl, "Solving the Longest Common Subsequence problem concerning non-uniform distributions of letters in input strings", Mathematics 9, 1515, 2021.

[70] B. Nikolic, B. Sobot, "Measures of string similarities based on Hamming distance", Preprint, https://arxiv.org/abs/2211.14615, 2022.

[71] A. Panconesi, "The stationary distribution of a Markov chain", Unpublished note, Sapienza University of Rome, 2005.

[72] Z. Peng, Y. Wang, "A novel efficient graph model for the Multiple Longest Common Subsequences (MLCS) problem", Front. Genet., 8, 104, 2017.

[73] T. Petrie, "Probabilistic functions of finite state Markov chains," Ann. Math. Statist., 40, No.1, 1969, pp. 97-115.

[74] J. Pitman, "Coalescents with multiple collisions", The Annals of Probability, Vol. 27, No. 4, pp. 1870–1902, 1999.

[75] J. Pitman, M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", The Annals of Probability, vol 25., No.2, 1997.

[76] T. Pohlert, "The pairwise multiple comparison of mean ranks package (PMCMR)", R Package, 27, 9, 2014.

[77] L. Polterovich, D. Rosen, K. Samvelyan i J.Zhang, "Topological Persistence in Geometry and Analysis," American Mathematical Society, 2020.

[78] Y. Reani i O. Bobrowski, "Cycle registration in persistent homology with applications in topological bootstrap," Preprint, 2021.

[79] G. O. Roberts, "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms", Stochastic Processes and their Applications, 49, pp. 207-216, 1994.

[80] D. Sankoff, J. Kruskal, "Time warps, string edits, and macromolecules, The Theory and Practice of Sequence Comparison", CSLI Publication, 1999.

[81] R. Serfozo, "Basics of Applied Stochastic Processes, Probability and its Applications", Springer-Verlag Berlin Heidelberg, 2009.

[82] J.P. Serre, "Linear Representations of Finite Groups", Springer-Verlag New York, 1977.

[83] C. E. Shannon, "A Mathematical Theory of Communication", The Bell System Technical Journal, 1948.

[84] D. J. Sheskin, "Handbook of parametric and nonparametric statistical procedures", Chapman and Hall/CRC, 2000.

[85] S. J. Shyu, C. Y. Tsai, "Finding the longest common subsequence for multiple biological sequences by ant colony optimization", Comput. Oper. Res., 36, pp. 73-91, 2009.

[86] J. Storer, "Data Compression: Methods and Theory", Computer Science Press: MD, USA, 1988.

[87] F. S. Tabataba, S. R. Mousavi, "A hyper-heuristic for the Longest Common Subsequence problem", Comput. Biol. Chem. 2012, 36, pp. 42–54.

[88] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney", NUS School of Computing Technical Report, 2006.

[89] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 2006.

[90] Y. W. Teh i M. I. Jordan, "Bayesian nonparametrics: hierarchical Bayesian nonparametric models with applications", Cambridge Series in Statistical and Probablistic Mathematics, 2010.

[91] S. G. Vadlamudi, S. Aine, P. P Chakrabarti, "Anytime pack search", Nat. Comput. , 15, pp. 395–414. 2016.

[92] S. G. Vadlamudi, P. Gaurav, S. Aine, P. P. Chakrabarti, "Anytime column search", Proceedings of the AI'12—The 25th Australasian Joint Conference on Artificial Intelligence, Sydney, Australia, 4–7 December 2012, pp. 254–265.

[93] Q. Wang, D. Korkin, Y. Shang, "A fast multiple longest common subsequence (MLCS) algorithm", IEEE Trans. Knowl. Data Eng. 2011, 23, pp. 321–334.

[94] C. Wang, Y. Wang, Y. Cheung, "A branch and bound irredundant graph algorithm for large-scale MLCS problems", Pattern Recognit. , 119, 108059. 2021.

[95] S. Wei, Y. Wang, Y. Yang, S. Liu, "A path recorder algorithm for Multiple Longest Common Subsequences (MLCS) problems", Bioinformatics, 36, pp 3035–3042, 2020.

[96] F. Wood, C. Archambeau, J. Gasthaus, L. James, Y. W. Teh, "A stochastic memoizer for sequence data", Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.

[97] X. Xie, W. Liao, H. Aghajan, P. Veelaert, W. Philips, "Detecting road intersections from GPS traces using longest common subsequence Algorithm", ISPRS Int. J. Geo-Inf., 1, 6, 2017.

[98] J. Yang, Y. Xu, G. Sun, Y. Shang, "A new progressive algorithm for a Multiple Longest Common Subsequences Problem and its efficient parallelization", IEEE Trans. Parallel Distrib. Syst. , 24, pp. 862–870, 2013.

[99] A. Zomorodian i G. Carlsson, "Computing persistent homology," Discrete and Computational Geometry, 33(2), pp. 249–274, 2004.

[100] S. Zürcher, "Smallest enclosing ball for a point set with strictly convex level sets", MSc thesis, ETH Zurich, 2007.

Biografija

Bojan Nikolić je rođen 05.06.1980. godine u Tuzli. Osnovnu školu i Gimnaziju završio je u Zvorniku. 1999. godine upisuje opšti smjer na odsjeku za matematiku i informatiku Prirodno-matematičkog fakulteta Univerziteta u Banjoj Luci. Osnovne studije je završio jula 2004. godine stekavši zvanje diplomiranog matematičara i informatičara. Poslediplomske studije je završio na Prirodno-matematičkom fakultetu, Univerziteta u Novom Sadu, gdje je oktobra 2010. godine odbranio magistarsku tezu na temu "Prostori topologija" i time stekao zvanje magistra matematičkih nauka. Od februara 2009. godine radio je na mjestu saradnika u zvanju asistenta na Prirodno- matematičkom fakultetu u Banjoj Luci i u prethodnoj deceniji držao je vježbe iz različitih matematičkih kurseva kao što su Topologija, Osnove matematike, Elementarna matematika, Teorija vjerovatnoće, Vjerovatnoća i statistika itd.

Prvenstveno se bavi Topologijom i Teorijom skupova, a nije mu strano ni vjerovatnosno modeliranje matematičkih objekata diskretnog tipa. Autor je više knjiga i naučnih radova koji su objavljeni u domaćim i stranim časopisima. Bio je učesnik manifestacija "Dani matematike" i "Festival nauke", koji su za cilj imali promociju matematike za populaciju učenika osnovnih i srednjih škola.

Oženjen; sa ženom svakodnevno igra kviz "TV slagalica". Pasionirani je ljubitelj filmova, SF literature i društvenih igara.

УНИВЕРЗИТЕТ У БАЊОЈ ЛУЦИ
ФАКУЛТЕТ: ПРИРОДНО-МАТЕМАТИЧКИ

Образац 3

РЕПУБЛИКА СРПСКА
УНИВЕРЗИТЕТ У БАЊА ЛУЦИ
Природно-математички факултет
Број: 19/4.963-1/23
Датум: 13.11.2023 199
БАЊА ЛУКА

ИЗВЈЕШТАЈ
о оцјени урађене докторске дисертације

1. ПОДАЦИ О КОМИСИЈИ

Орган који је именовао комисију: Сенат Универзитета у Бањој Луци, на основу приједлога одлуке Наставно-Научног вијећа Природно-математичког факултета (Број 19/3.2121/23, дана 06.09.2023.)

Датум именовања комисије: 21.09.2023.

Број одлуке: 02/04-3.2036-91/23

Чланови комисије:

	Презиме и име	Звање	Научно поље и ужа научна област	Установа у којој је запослен-а	Функција у комисији
1.	Др Душко Богданић	редовни професор	Алгебра и геометрија	Природно-математички факултет Универзитета у Бањој Луци	предсједник
2.	Др Бојан Башић	редовни професор	Дискретна математика	Природно-математички факултет Универзитета у Новом Саду	члан
3.	Академик Зоран Митровић	редовни професор	Математичка анализа и примјене	Електротехнички факултет, Универзитета у Бањој Луци	члан

4. Др Марко Ђукановић	доцент	Информационе науке и биоинформатика (развој софтвера)
Презиме и име	Звање	Научно поље и ужа научна област
Природно-математички факултет Универзитета у Бањој Луци	члан	
Установа у којој је запослен-а	Функција у комисији	

2. ПОДАЦИ О СТУДЕНТУ

Име, име једног родитеља, презиме: Бојан, Милисав, Николић

Датум рођења: 05.06.1980.

Мјесто и држава рођења: Тузла, БиХ

2.1. Студије првог циклуса или основне студије или интегрисане студије

Година уписа:	1999.	Година завршетка:	2004.	Просјечна оцјена током студија:	8,19

Универзитет: Универзитет у Бањој Луци

Факултет/и: Природно-математички факултет

Студијски програм: Математика и информатика

Стечено звање: Дипломирани математичар и информатичар

2.2. Студије другог циклуса или магистарске студије

Година уписа:	2005.	Година завршетка:	2010.	Просјечна оцјена током студија:	9,67

Универзитет: Универзитет у Новом Саду

Факултет/и: Природно-математички факултет

Студијски програм: Департман за математику и информатику

Назив завршног рада другог циклуса или магистарске тезе, датум одбране: Простори топологија, октобар 2010.

Ужа научна област завршног рада другог циклуса или магистарске тезе: Математика

Стечено звање: Магистар математичких наука

2.3. Студије трећег циклуса

Година уписа:	2017.	Број ECTS остварених до сада:	120	Просјечна оцјена током студија:	10,00

Факултет/и: Природно-математички факултет

Студијски програм: Докторске студије математике

2.4. Приказ научних и стручних радова студента		
РБ		Категорија[1]
1.	Б. Николић, А. Картељ, М. Ђукановић, М. Грбић, C. Blum, G. Raidl, *Solving the Longest Common Subsequence Problem Concerning Non-Uniform Distributions of Letters in Input Strings*, Mathematics, Vol. 9, No. 13, 1515, Jun, 2021.	SCI листа IF=2,4
Кратак опис садржине: Проблем најдужег заједничког подниза (LCS проблем) представља налажење најдужег подниза, који је заједнички за све улазне низове. Овај проблем има различите примјене у биоинформатици, молекуларној биологији и провјери плагијата, између осталог. Сви претходни приступи из литературе се заснивају на претпоставци да су улазне LCS инстанце изабране из униформних или скоро униформних дистрибуција вјероватноће слова. У овом раду представљамо приступ који је у стању да се ефикасно бави општијим случајевима, гдје појављивање слова у улазним стринговима прати неке друге случајеве полиномијалне расподјеле. Предложени приступ користи временски ограничену претрагу бима, вођену новом хеуристиком по имену GMPSUM. Ова хеуристика комбинује двије комплементарне статистике у облику конвексне комбинације. У раду је дата свеобухватна емпиријска процјена у два различита подешавања: (1) краткотрајно извршавање са фиксном величином бима како би се процијениле способности вођења упоређене хеуристике претраге и (2) дуготрајна извршења са фиксним циљним временима трајања како би се добила висококвалитетна рјешења. У оба подешавања, новопредложен поступак је у случају униформне расподјеле у рангу познатих метода из литературе, док их у случају разматраних неуниформних дистрибуција надмашује.		

Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације	**ДА**	**НЕ**
РБ	Категорија	
2. B. Nikolic, B. Sobot, "Measures of string similarities based on Hamming distance", Preprint, https://arxiv.org/abs/2211.14615, 2022.		
Кратак опис садржине: У раду се уводе нове мјере сличности фамилија стрингова уз помоћ техника и метода истрајне симплицијалне хомологије. Посматра се Чехова филтрација посматране фамилије стрингова и уз помоћ удаљености уског грла се дефинише мјера сличности фамилија стрингова која, поред упаривања бар-код линија, приоритет даје упаривању оних парова линија које одговарају квалитативно сличним хомолошким класама. Да би се ово постигло, у раду је развијена нова техника раздвајање радијуса симплекса.		

Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације	**ДА**	**НЕ**
РБ	Категорија	

[1] Категорија се односи на оне часописе и научне скупове који су категорисани у складу са Правилником о публиковању научних публикација („Службени гласник РС”, бр. 77/10) и Правилником о мјерилима за остваривање и финансирање Програма одржавања научних скупова („Службени гласник РС”, бр. 102/14) односно припадност рада часописима индексираним у свјетским цитатним базама.

3.	Б. Николић, *Degenerated topological spaces*, NOVI SAD JOURNAL OF MATHEMATICS, Vol. 49, No. 2, pp. 95-107, 2019.	Scopus

Кратак опис садржине:

Коришћењем композиција оператора унутрашњости и затворења тополошког простора могуће је добити 7 различитих оператора. Тополошки простори у којима међу овим операторима има једнаких називају се дегенерисани тополошки простори. Овај рад даје класификацију и карактеризацију дегенерисаних тополошких простора.

Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације	**ДА**	**НЕ**

РБ		Категорија
4.	B. Nikolic, M. Djukanovic, D. Matic, *New mixed-integer linear programming model for solving the multidimensional multi-way number partitioning problem,* Computational and Applied Mathematics 41 (3), 119, 2022.	SCI листа IF=2,6

Кратак опис садржине:

Проблем мултидимензионалног вишеструког проблема партиционисања подразумијева прављење партиције коначног скупа вектора, тако да суме одговарајућих координата у оквиру појединачних скупова из партиције не одступају значајно. Овај проблем има значајне практичне примјене у ситуацијама када треба извршити уједначење скупа елемената у погледу два или више посматраних одлика. У раду је предложен нови MILP модел за рјешавање овог проблема који у већој мјери надмашује постојеће моделе из литературе којима се рјешава овај проблем.

Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације	**ДА**	**НЕ**

РБ		Категорија
5.	Д. Богданић, Б.Николић, D.A.Romano, *Аксиоме теорије скупова*, МАТ-КОЛ, Vol. XV(1), pp. 17-25, 2009.	

Кратак опис садржине:

Постоји више од једне могућности аксиоматизације теорије скупова. У овом раду размотрени су неки аспекти опште прихваћеног ZFC аксиоматског система теорије скупова.

Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације	**ДА**	**НЕ**

РБ		Категорија
6.	Б. Николић, Н. Елез, *О оператору границе*, МАТ-КОЛ, No. XIII(1), pp. 67-72, 2007.	

Кратак опис садржине:

У раду су испитане особине оператора границе скупа у тополошком простору, посматрају се скупови добијени примјеном тог оператора и њихов међуодноc у смислу релације инклузије. Такође, доказане су неке нове формуле у вези границе скупа у тополошком

простору.			
Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације		**ДА**	**НЕ**
РБ		Категорија	
7.	Б. Николић, Н. Елез, *Оператори пик, каро, треф, срце*, МАТ-КОЛ, No. XIII(1), pp. 73-80, 2007.		
Кратак опис садржине: У раду су, комбиновањем оператора унутрашњости и затворења, уведена 4 нова оператора, испитани су њихови међуодноси, као и топологије које ови оператори генеришу.			
Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације		**ДА**	**НЕ**
РБ		Категорија	
8.	Б. Николић, *О математичкој нотацији*, МАТ-КОЛ, Vol. XXVI, No. 1, pp. 1-15, 2020.		
Кратак опис садржине: Записивање математичких замисли подразумијева употребу неког облика математичке нотације. Избором симбола и начином њиховог записивања могу се јасније истаћи концепти који представљају суштину идеје која се жели записати. У овом раду је направљен осврт на коришћење математичке нотације, са акцентом на препознавање ситуација у којима је могуће изабрати нотацију којом се интуитивније изражава семантика текста који се жели записати.			
Припадност рада ужој научној области којој припада предмет истраживања докторске дисертације		**ДА**	**НЕ**

3. УВОДНИ ДИО ОЦЈЕНЕ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Тема докторске дисертације под насловом "Мјере сличности на фамилијама стрингова" кандидата Бојана Николића, након давања сагласности на Извјештај о оцјени подобности теме и кандидата и испуњености услова за менторство за израду докторске дисертације на Природно-математичком факултету, прихваћена је одлуком Сената Универзитета у Бањој Луци број 02/04-3.2670-34/21 од 25.11.2021. године. Истом одлуком је проф. Др Борису Шоботу, ванредном професору на Одсјеку за математику и информатику Природно-математичког факултета Универзитета у Новом Саду, дата сагласност за менторисање кандидата на задату тему.

Садржај докторске дисертације изложен је у сљедећим главама:

1) Увод (странице: 3-6)
2) Тополошки методи (странице: 7-86)

3) Вјероватносни методи (странице: 87-144)
4) Примјена при рјешавању ЛЦС проблема (странице: 145-178)
5) Закључак (странице: 179-184)
6) Литература (странице: 185-192)

Дисертација је написана на српском језику, латиничним писаним фонтом Times New Roman на 192 странице А4 формата и садржи шест глава: Увод, Тополошки методи, Вјероватносни методи, Примјена при рјешавању ЛЦС проблема, Закључак, Литература (100 референци). Такође, дисертација садржи 22 слике и 12 табела.

У уводној глави (Увод, стр. 3-6) дат је кратак преглед студираних проблема, као и основне идеје коришћене у наставку. Такође, дат је и кратак опис садржаја осталих поглавља.

У другој глави (Тополошки методи, стр. 7-86) проучавано је неколико врста метрика, као што су Хамингова, ЛЦС и Хауздорфова метрика. Дате су и основе симплицијалне хомологије, а посебна пажња је посвећена Чеховом комплексу. Затим су уведени појмови филтрације, истрајног модула, те теорема о интервалној декомпозицији која представља основ за коришћење резултата о бар кодовима. На крају поглавља су уведени појмови удаљености уског грла, те је доказана теорема стабилности. Користећи фамилије стрингова, кандидат разматра аутоморфизме одговарајућих метричких простора и описује методу шеме регистрацијских комплекса. У овој глави су дати резултати у вези са методом раздвајања радијуса симплекса.

У трећој глави (Вјероватносни методи, стр. 87-144) су проучавани стрингови из угла вјероватносних мјера. Фамилије стрингова су искоришћене да би се дефинисали одређени стохастички процеси и одговарајући простори вјероватноћа. Главни резултат овог поглавља се састоји из предложене мјере сличности ових фамилија стрингова. Ова мјера сличности је заснована на релативној ентропији. На крају поглавља су описана и два приступа за моделовање непознатих расподела. То су фреквенционистички приступ и Бејзовско закључивање.

У четвртој глави (Примјена при рјешавању ЛЦС проблема, стр. 145-178) су изведене рекурентне формуле за рачунање најдужег заједничког подниза два стринга. На почетку овог поглавља је дат преглед постојећих метода за општији случај фамилије стрингова и описан је начин на који се рјешења проблема моделирају као путеви у одговарајућем графу. У наставку поглавља се уводе нове статистике GM и PSUM, као и статистика GMPSUM која је њихова конвексна комбинација. На крају овог поглавља је дата процјена сложености добијених алгоритама, наведени су експериментални резултати који су упоређени са резултатима извршавања постојећих алгоритама.

У закључку дисертације (Закључак, стр. 179-184) је дат резиме добијених резултата и предложених поступака. Такође су наведени и евентуални будући правци разматраних тема.

4. УВОД И ПРЕГЛЕД ЛИТЕРАТУРЕ

У истраживању се посматрају стрингови који се користе за математичко моделовање

објеката дискретног секвенцијалног типа. У већини случајева је потребна вишедимезионална анализа ових објеката. Стога је корисно да при анализи ових података имамо мјере поређења фамилија стрингова који моделују те податке. Проблем поређења фамилија стрингова је много компликованији од проблема поређења појединачних стрингова, тако да је неопходно укључити технике и хеуристике неколико математичких области. Ту се прије свега издвајају методе истрајне хомологије, хијерархијски Бејзовски модели и претрага бима. Ове методе су веома актуелне, нпр. методе истрајне хомологије су се почеле развијати тек крајем прошлог и почетком овог вијека ([5, 17, 36]). Садржај ове дисертације је усмјерен не само ка добијању теоријских резултата већ има и широк спектар потенцијалних примјена. Нпр, у анализи биолошких секвенци, претраживању текста, алгоритмима за препознавање говора итд.

Циљ дисертације је да се дефинишу мјере сличности скупова стрингова које би у што већој мјери одражавале заједничке релевантне особине тих скупова. Коришћене су технике и методе алгебарске топологије ([6, 18, 25]), вјероватносно-статистичке методе ([58, 75, 90]) и рачунске методе ([24, 33, 55]).

Литература:

[1] J. Albert, J. Hu, "Probability and Bayesian modeling", CRC press, 2020.
[2] S. Bacallado, S. Favaro, S. Power, L. Trippa, "Perfect sampling of the posterior in the hierarchical Pitman–Yor process", International Society for Bayesian Analysis, 2021.
[3] R. A. Baeza-Yates, R. Gavalda, G. Navarro, R. Scheihing, "Bounding the expected length of longest common subsequences and forests", Theory Comput. Syst., 32(4), pp. 435–452, 1999.
[4] D. Bakkelund, "An LCS-based string metric", University of Oslo, 2009.
[5] S.A. Barannikov, "The framed Morse complex and its invariants,"Advances in Soviet Mathematics, 21, pp. 93-115, 1994.
[6] U. Bauer i M. Lesnick, "Induced matchings and the algebraic stability of persistence barcodes," Journal of Computational Geometry, 6(2), pp. 162-191, 2015.
[7] R. Beal, T. Afrin, A. Farheen, D. Adjeroh, "A new algorithm for the LCS problem with application in compressing genome resequencing data", BMC Genom., 17, 544, 2016.
[8] R. Begleiter, R. El-Yaniv, G. Yona, "On prediction using variable order Markov models", Journal of Artiflcial Intelligence Research 22, pp. 385-421, 2004.
[9] Y. Bengio, "Markovian models for sequential data", Neural computing surveys, 1999.
[10] L. Bergroth, H. Hakonen, T. Raita, "A survey of longest common subsequence algorithms", Proceedings of the SPIRE 2000—The 7th International Symposiumon String Processing and Information Retrieval, Coruna, Spain, pp. 39-48, 2000.
[11] P. Billingsley, "Ergodic Theory and Information", New York: Wiley, 1965.
[12] C. Blum, M. J. Blesa, "Probabilistic beam search for the longest common subsequence problem", In Proceedings of the International Workshop on Engineering Stochastic Local Search Algorithms, Brussels, Belgium, 2007., Springer: 200, pp. 150-161, 2007.
[13] C. Blum, M. J. Blesa, M. López-Ibáñez, "Beam search for the longest common subsequence problem", Comput. Oper. Res. 2009, 36, pp. 3178–3186.
[14] C. Blum, P. Festa, "Longest common subsequence problems", Metaheuristics for String Problems in Bioinformatics, Wiley:Hoboken,NJ, USA, 2016, Chapter 3, pp.45–60.
[15] P. Bubenik i J. A. Scott, "Categorification of persistent homology", Discrete and Computational Geometry, Vol. 51, pp. 600–627, 2014.
[16] B. Calvo, G. Santafé-Rodrigo, "Statistical comparison of multiple algorithms in multiple problems", The R Journal, Vol. 8/1, Aug. 2016.
[17] G. Carlsson, "Topology and data", Bulletin (New Series) of the American Mathematical Society, 2009.

[18] G. Carlsson, A. Zomorodian, A. Collins i L. Guibas, "Persistence barcodes for shapes", Eurographics Symposium on Geometry Processing, 2004.
[19] H. T. Chan, C. B. Yang, Y. H. Peng, "The generalized definitions of the twodimensional largest common substructure problems", Proceedings of the 33rd Workshop on Combinatorial Mathematics and Computation Theory, Taipei, Taiwan,
13–14 May 2016, pp. 1–12.
[20] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas i S. Oudot, "Proximity of persistence modules and their diagrams," Research Report RR-6568, INRIA, 2008.
[21] F. Chazal i B. Michel, "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists", preprint 2021.
[22] C. Chen, L. Du,W. Buntime, "Sampling table configurations for the hierarchical Poisson-Dirichlet process", ECML PKDD 2011: Machine Learning and Knowledge Discovery in Databases pp. 296–311, 2011.
[23] S.F. Chen, J. Goodman, "An empirical study of smoothing tehnique for language modeling", Computer Science Group, Harvard University, Cambridge, Massachusetts, 1998.
[24] V. Chvatal i D. Sankoff, "Longest common subsequences of two random sequences", Journal of Applied Probability, Vol. 12, No. 2, pp. 306-315, 1975.
[25] D. Cohen-Steiner, H. Edelsbrunner i J. Harer, "Stability of persistence diagrams," Discrete and Computational Geometry, 37, pp. 103–120, 2007.
[26] T. M. Cover i J. A. Thomas, "Elements of Information Theory", Second edition, John Wiley and Sons, Inc., 2006.
[27] W. Crawley-Boevey, "Decomposition of pointwise finite-dimensional persistence modules", Journal of Algebra and Its Applications, Vol. 14, No. 5, 2015.
[28] V.Dančik, "Expected length of longest common subsequences", PhD thesis, Department of Computer Science, University of Warwick, September 1994.
[29] V. de Silva i V. Nanda, "Geometry in the space of persistence modules", Proceedings of the twenty-ninth annual symposium on Computational geometry, 2013.
[30] T. K. Dey i Y. Wang, "Computational topology for Data Analysis", Cambridge University Press, 2022.
[31] M. Djukanovic, A. Kartelj, D.Matic, M. Grbic, C. Blum, G. Raidl, "Solving the generalized constrained Longest Common Subsequence problem with many pattern strings", Technical Report AC-TR-21-008, AC, 2021.
[32] M. Djukanovic, G. R. Raidl, C. Blum, "Anytime algorithms for the longest common palindromic subsequence problem", Comput. Oper. Res. 2020, 114, 104827.
[33] M. Djukanovic, G. Raidl, C. Blum, "A Beam Search for the Longest Common Subsequence problem guided by a novel approximate expected length calculation", Proceedings of the LOD 2019—The 5th International Conference on Machine Learning, Optimization, and Data Science, Siena, Italy, 10–13 September 2019.
[34] M.Djukanovic, G.Raidl, C.Blum, "Finding longest common subsequences: new anytime A* search results", Appl. Soft. Comput. 2020, 95, 106499.
[35] T. Easton, A. Singireddy, "A large neighborhood search heuristic for the longest common subsequence problem", J. Heuristics, 14, pp. 271-283, 2008.
[36] H. Edelsbrunner i J. Harer, "Computational Topology: An Introduction", American Mathematical Society, 2010.
[37] H. Edelsbrunner i J. Harer, "Persistent homology - a survey," Surveys on Discrete and Computational Geometry: Twenty Years Later, American Mathematical Society, 2008., pp. 257–282.
[38] R. Forman, "A user's guide to discrete Morse theory", Seminaire Lotharingien de Combinatoire 48, Article B48c, 2002.
[39] C.B. Fraser, "Subsequences and supersequences of strings", Ph.D. Thesis, University of

Glasgow, Glasgow, UK, 1995.
[40] A. Gabadinho i G. Ritschard, "Analyzing state sequences with probabilistic suffix trees: The PST R package", Journal of Statistical Software, 2016.
[41] J.A. Gasthaus, "Hierarchical Bayesian nonparametric models for power-law sequences", PhD Thesis, 2020.
[42] J.A. Gasthaus i Y. W. Teh, "Improvements to the sequence memoizer", Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems, 2010.
[43] R. M. Gray, "Entropy and Information Theory", Second edition, Springer New York, 2011.
[44] D. Gusfield, "Algorithms on Strings, Trees, and Sequences", Computer Science and Computational Biology, Cambridge University Press: Cambridge, UK, 1997.
[45] R. W. Hamming, "Error detecting and error correcting codes," The Bell System Technical Journal, vol. 29, no. 2, pp. 147-160, April 1950.
[46] A. Hatcher, "Algebraic Topology", Cambridge University Press, 2001.
[47] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences", Communications of the Association for Computing Machinery, 18(6), pp. 341-343, 1975.
[48] D. S. Hirschberg, "Algorithms for the Longest Common Subsequence problem", Journal of the Association for Computing Machinery., 24(4), pp. 664-675, 1977.
[49] L. C. Hsu, "A unified approach to generalized Stirling numbers", Advances in Applied Mathematics 20, pp. 366-384, 1998.
[50] K. Huang, C. Yang, K. Tseng, "Fast algorithms for finding the common subsequences of multiple sequences", Proceedings of the ICS 2004—The 9th International Computer Symposium, Funchal, Portugal, 13–16 January 2004.
[51] T. Jiang, M. Li, "On the approximation of shortest common supersequences and longest common subsequences", SIAM J. Comput., 24, pp. 1122-1139, 1995.
[52] C. Kermorvant, P. Dupont, "Improved smoothing for probabilistic suffix trees seen as variable order Markov chains", Springer-Verlag Berlin Heidelberg, 2002.
[53] H. Kesten, N. Morse, "A property of the multinomial distribution", Ann. Math. Stat., 30, pp. 120-127, 1959.
[54] M. Kiwi, J. Soto, "On a speculated relation between Chvatal-Sankoff constants of several sequences", Combin. Probab. Comput., 18(4), pp. 517–532, 2009.
[55] M. Kiwi, M. Loebl, J. Matousek, "Expected length of the longest common subsequence for large alphabets", Adv. Math., 197(2), pp. 480–498, 2005.
[56] R. Kneser, H. Ney, "Improved backing-off for m-gram language modeling", International Conference on Acoustics, Speech, and Signal Processing, 1995.
[57] J. B. Kruskal, "An overview of sequence comparison: Time warps, string edits, and macromolecules", SIAM Rev., 25, pp. 201-237, 1983.
[58] S. Kullback, R. A. Leibler, "On information and sufficiency", The Annals of Mathematical Statistics, 22, pp. 79-86, 1951.
[59] D.A. Levin, Y. Peres, "Markov Chains and Mixing Times: Second Edition", American Mathematical Society, 2017.
[60] Y. Li, Y. Wang, Z. Zhang, Y. Wang, D. Ma, J. Huang, "A novel fast and memory efficient parallel MLCS algorithm for long and large-scale sequences alignments", Proceedings of the IEEE 32nd International Conference on Data Engineering, Helsinki, Finland, 16–20 May 2016, pp. 1170–1181.
[61] K. W. Lim, W. Buntime, C. Chen, L. Du, "Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes", International Journal of Approximate Reasoning, 2016.
[62] G. S. Lueker, "Improved bounds on the average length of longest common subsequences", Journal of the ACM, 56(3), pp. 1–38, May 2009.

[63] D. Maier, "The complexity of some problems on subsequences and supersequences", J. ACM, 25, pp. 322–336, 1978.
[64] M. Maltenfort, "New definitions of the generalized Stirling numbers", Aequationes mathematicae volume 94, pp. 169–200, 2020.
[65] G. Máté, A. Hofmann, N. Wenzel i D. W. Heermann, "A topological similarity measure for proteins", 2013.
[66] P. Minkiewicz, M. Darewicz, A. Iwaniak, J. Sokołowska, P. Starowicz, J. Bucholska, M. Hrynkiewicz, "Common amino acid subsequences in a universal proteome—relevance for food science", Int. J. Mol. Sci., 16, pp. 20748–20773, 2015.
[67] S.R. Mousavi, F. Tabataba, "An improved algorithm for the longest common subsequence problem". Comput. Oper. Res. 2012, 39, pp. 512–520.
[68] J. R. Munkres, "Elements of Algebraic Topology", CRC Press, 1993.
[69] B. Nikolic, A. Kartelj, M. Djukanovic, M. Grbic, C. Blum, G. Raidl, "Solving the Longest Common Subsequence problem concerning non-uniform distributions of letters in input strings", Mathematics 9, 1515, 2021.
[70] B. Nikolic, B. Sobot, "Measures of string similarities based on Hamming distance", Preprint, https://arxiv.org/abs/2211.14615, 2022.
[71] A. Panconesi, "The stationary distribution of aMarkov chain", Unpublished note, Sapienza University of Rome, 2005.
[72] Z. Peng, Y. Wang, "A novel efficient graph model for the Multiple Longest Common Subsequences (MLCS) problem", Front. Genet., 8, 104, 2017.
[73] T. Petrie, "Probabilistic functions of finite state Markov chains," Ann.Math.Statist., 40, No.1, 1969, pp. 97-115.
[74] J. Pitman, "Coalescents with multiple collisions",The Annals of Probability, Vol. 27, No. 4, pp. 1870–1902, 1999.
[75] J. Pitman, M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", The Annals of Probability, vol 25., No.2, 1997.
[76] T. Pohlert, "The pairwise multiple comparison of mean ranks package (PMCMR)", R Package, 27, 9, 2014.
[77] L. Polterovich, D. Rosen, K. Samvelyan and J.Zhang, "Topological Persistence in Geometry and Analysis," American Mathematical Society, 2020.
[78] Y. Reani i O. Bobrowski, "Cycle registration in persistent homology with applications in topological bootstrap," Preprint, 2021.
[79] G. O. Roberts, "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms", Stochastic Processes and their Applications, 49, pp. 207-216, 1994.
[80] D. Sankoff, J. Kruskal, "Time warps, string edits, and macromolecules, The Theory and Practice of Sequence Comparison", CSLI Publication, 1999.
[81] R. Serfozo, "Basics of Applied Stochastic Processes, Probability and its Applications", Springer-Verlag Berlin Heidelberg, 2009.
[82] J.P. Serre, "Linear Representations of Finite Groups", Springer-Verlag New York, 1977.
[83] C. E. Shannon, "A Mathematical Theory of Communication", The Bell System Technical Journal, 1948.
[84] D. J. Sheskin, "Handbook of parametric and nonparametric statistical procedures", Chapman and Hall/CRC, 2000.
[85] S. J. Shyu, C.Y.Tsai, "Finding the longest common subsequence for multiple biological sequences by ant colony optimization", Comput. Oper. Res., 36, pp. 73-91, 2009.
[86] J. Storer, "Data Compression: Methods and Theory", Computer Science Press: MD, USA, 1988.
[87] F. S. Tabataba, S. R. Mousavi, "A hyper-heuristic for the Longest Common Subsequence

problem", Comput. Biol. Chem. 2012, 36, pp. 42–54.
[88] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney", NUS School of Computing Technical Report, 2006.
[89] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 2006.
[90] Y. W. Teh i M. I. Jordan, "Bayesian nonparametrics: hierarchical Bayesian nonparametric models with applications", Cambridge Series in Statistical and Probablistic Mathematics, 2010.
[91] S. G. Vadlamudi, S. Aine, P. P Chakrabarti, "Anytime pack search", Nat. Comput. , 15, pp. 395–414. 2016.
[92] S. G. Vadlamudi, P. Gaurav, S. Aine, P. P. Chakrabarti, "Anytime column search", Proceedings of the AI'12—The 25th Australasian Joint Conference on Artificial Intelligence, Sydney, Australia, 4–7 December 2012, pp. 254–265.
[93] Q. Wang, D. Korkin, Y. Shang, "A fast multiple longest common subsequence (MLCS) algorithm", IEEE Trans. Knowl. Data Eng. 2011, 23, pp.321–334.
[94] C. Wang, Y. Wang, Y. Cheung, "A branch and bound irredundant graph algorithm for large-scale MLCS problems", Pattern Recognit., 119, 108059, 2021.
[95] S. Wei, Y. Wang, Y. Yang, S. Liu, "A path recorder algorithm for Multiple Longest Common Subsequences (MLCS) problems", Bioinformatics, 36, pp 3035–3042, 2020.
[96] F. Wood, C. Archambeau, J. Gasthaus, L. James, Y. W. Teh, "A stochastic memoizer for sequence data", Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.
[97] X. Xie, W. Liao, H. Aghajan, P. Veelaert, W. Philips, "Detecting road intersections from GPS traces using longest common subsequence Algorithm", ISPRS Int. J. Geo-Inf., 1, 6, 2017.
[98] J. Yang, Y. Xu, G. Sun, Y. Shang, "A new progressive algorithm for a Multiple Longest Common Subsequences Problem and its efficient parallelization", IEEE Trans. Parallel Distrib. Syst. , 24, pp. 862–870, 2013.
[99] A. Zomorodian i G. Carlsson, "Computing persistent homology," Discrete and Computational Geometry, 33(2), pp. 249–274, 2004.
[100] S. Zürcher, "Smallest enclosing ball for a point set with strictly convex level sets", MSc thesis, ETH Zurich, 2007.

5. МАТЕРИЈАЛ И МЕТОДОЛОГИЈА РАДА

У истраживању су кориштене методе алгебарске топологије и вјероватносно-статистичке методе. Осим тога, значајну улогу у анализи проблема заузимају и напредне рачунске методе које су имале улогу у упоређивања сложености алгоритама.

Кандидат је показао да адекватно користи поменути теоријски апарат за рјешавање посматраних проблема. Није дошло до промјене плана истраживања који је дат приликом пријаве докторске тезе.

6. РЕЗУЛТАТИ И НАУЧНИ ДОПРИНОС ИСТРАЖИВАЊА

6.1. Резултати истраживања

Оригинални и најзначајнији научни резултати овог истраживања се огледају у:

1) Развијене су нове технике и доказани одређени резултати који омогућавају не само увођење нових мјера сличности фамилија стрингова, већ и додатно осигуравају да ове мјере посједују својство стабилности. Описана је нова техника раздвајања радијуса симплекса, као кључно средство које омогућава модификацију методе уског грла путем хибридног упаривања.

2) Матрица транзиционих вјероватноћа одређује вјероватносну мјеру која одговара датом скупу стрингова. За двије такве мјере можемо рачунати релативну ентропију и помоћу ње дефинисати мјеру сличности за дате фамилије стрингова. Ове расподјеле вјероватноћа у пракси нису познате, тако да се моделују разним методама. У тези су размотрене идеје које, коришћењем погодних апроксимација, знатно увећавају ефикасног овог поступка.

3) У тези је представљено рјешавање проблема проналажења најдужег заједничког подниза дате фамилије скупова над истим алфабетом (ЛЦС проблема) уз помоћ претраге бима (Beam Search или BS) који користи специјално дизајниране новоуведене хеуристике које усмјеравају претрагу ка перспективнијим рјешењима. Предности конструисаног метода у односу на постојеће су показане релативним статистичким тестовима.

4) Описани су нови правци истраживања дате тематике који су од изузетне важности у области проучавања структуре сложених скупова података.

6.2. Критичност и коректност тумачења резултата

Резултати истраживања су приказани веома компетентно, на јасан и прегледан начин.

6.3. Теоријски допринос и нови истраживачки резултати.

У дисертацији су развијене нове технике и доказани одређени резултати који омогућавају не само увођење нових мјера сличности фамилија стрингова, већ и додатно осигуравају да ове мјере посједују својство стабилности. Описана је нова техника раздвајања радијуса симплекса, као кључно средство које омогућава да поменуто хибридно упаривање има смисла. (B. Nikolic, B. Sobot, "Measures of string similarities based on Hamming distance", Preprint, https://arxiv.org/abs/2211.14615, 2022.). У тези је представљено и рјешење ЛЦС проблема уз помоћ претраге бима (Beam Search) уз коришћење специјално дизајниране новоуведене хеуристике која усмјерава претрагу ка перспективнијим рјешењима (B. Nikolic, A. Kartelj, M. Djukanovic, M. Grbic, C. Blum, G. Raidl, "Solving the Longest Common Subsequence problem concerning non-uniform distributions of letters in input strings", Mathematics 9, 1515, 2021.). На тај начин, ова теза представља оригиналан допринос посматраној теми. Презентовани резултати имају значајну примјену у већини природних наука, као и у одређеном броју друштвених наука гдје се проучавају дискретни објекти секвенцијалног типа.

7. ЗАКЉУЧАК И ПРИЈЕДЛОГ

На основу свега што је наведено у Извјештају, Комисија закључује да је докторска дисертација магистра Бојана Николића под насловом "Мјере сличности на фамилијама стрингова" израђена у складу са образложењем које је кандидат приложио приликом пријаве ове теме. Докторска дисертација је урађена према правилима и принципима научно-истраживачког рада и резултат је оригиналног научног рада кандидата. Развијене су нове технике и доказани одређени резултати који омогућавају не само увођење нових мјера сличности фамилија стрингова, већ и додатно осигуравају да ове мјере посједују својство стабилности. У тези је представљено и рјешење ЛЦС проблема уз помоћ претраге бима (Beam Search) која користи специјално дизајниране новоуведене хеуристичке функције које усмјеравају претрагу ка перспективнијим рјешењима. На тај начин, ова теза представља оригиналан допринос посматраној теми.

Будући да је кандидат показао темељно познавање предмета истраживања, те у потпуности одговорио на проблематику која се разматра у дисертацији, Комисија предлаже Научно-наставном вијећу Природно-математичког факултета Универзитета у Бањој Луци и Сенату Универзитета у Бањој Луци да прихвате овај извјештај и одобре јавну одбрану докторске дисертације.

Мјесто и датум:
Бања Лука,
13.11.2023.

Др Душко Богданић, редовни професор,
предсједник комисије

Др Бојан Башић, редовни професор, члан

Академик Зоран Митровић, редовни професор,
члан

Др Марко Ђукановић, доцент, члан

ИЗДВОЈЕНО МИШЉЕЊЕ: Члан комисије који не жели да потпише извјештај јер се не слаже са мишљењем већине чланова комисије дужан је да у извјештај унесе образложење, односно разлоге због којих не жели да потпише извјештај.

Прилог 3.

Изјава 1

ИЗЈАВА О АУТОРСТВУ

Изјављујем
да је докторска дисертација/докторски умјетнички рад

Наслов дисертације/рада МЈЕРЕ СЛИЧНОСТИ НАД ФАМИЛИЈАМА СТРИНГОВА

Наслов дисертације/рада на енглеском језику SIMILARITY MEASURES OF FAMILIES OF STRINGS

☒ резултат сопственог истраживачког/умјетничког рада,

☒ да докторска дисертација/докторски умјетнички рад, у цјелини или у дијеловима, није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,

☒ да су резултати коректно наведени и

☒ да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

У Бањој Луци: 07.12.2023.

Потпис докторанда

Бојан Николић

Изјава 2

Изјава којом се овлашћује Универзитет у Бањој Луци да докторску дисертацију/докторски умјетнички рад учини јавно доступним

Овлашћујем Универзитет у Бањој Луци да моју докторску дисертацију/докторски умјетнички рад под насловом

МЈЕРЕ СЛИЧНОСТИ НАД ФАМИЛИЈАМА СТРИНГОВА

која је моје ауторско дјело, учини јавно доступним.

Докторску дисертацију/докторски умјетнички рад са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију/докторски умјетнички рад похрањену у дигитални репозиторијум Универзитета у Бањој Луци могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (*Creative Commons*) за коју сам се одлучио/ла.

1. Ауторство
2. Ауторство – некомерцијално
3. Ауторство – некомерцијално – без прераде
4. Ауторство – некомерцијално – дијелити под истим условима
5. Ауторство – без прераде
6. Ауторство – дијелити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

У Бањој Луци: 07.12.2023.

Потпис докторанда

Бојан Ивковић

Изјава 3

Изјава о идентичности штампане и електронске верзије докторске дисертације/докторског умјетничког рада

Име и презиме аутора БОЈАН НИКОЛИЋ

Наслов дисертације/рада МЈЕРЕ СЛИЧНОСТИ НАД ФАМИЛИЈАМА СТРИНГОВА

Ментор ПРОФ. ДР БОРИС ШОБОТ

Изјављујем да је штампана верзија моје докторске дисертације/докторског умјетничког рада идентична електронској верзији коју сам предао/ла за дигитални репозиторијум Универзитета у Бањој Луци.

У Бањој Луци: 07.12.2023.

Потпис докторанда

Бојан Николић