Dinamikus rendszerek és adattömörítés

Szakdolgozat

Készítette: **Bakondi Bence Gábor**
(Matematikus MSc.)

Témavezető: **Buczolich Zoltán**
(Analízis Tanszék, Matematikai Intézet)

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Budapest, 2013.

Köszönetnyilvánítás

Hatalmas köszönettel illeti témavezetőmet, Buczolich Zoltánt az érdekes témáért és a végtelennek tűnő türelméért, mellyel munkámat segítette.

Mindenképpen hálával tartozom Buczolich Zoltán mellett Simon Péternek és Kurics Tamásnak is, hogy nagyszerű kurzusaikkal felkeltették és fenn is tartották az érdeklődésemet a dinamikus rendszerek iránt.

Tartalomjegyzék

1. Bevezetés

Szakdolgozatom célja bemutatni, hogy Ornstein és Weiss cikke [1] az entrópia és az adattömörítés kapcsolatáról milyen hatással volt a tömörítési eljárásokra, milyen általánosításai, alkalmazásai vannak és hol lehetne még használni, hogy lehetne „továbbgondolni". A Shannon által 1948-ban bevezetett entrópia [2] és az adattömörítés közötti kapcsolat nem meglepő, hiszen minél több „váratlan" része van egy adatsornak, annál nehezebb hatékonyan tömöríteni.

Bár az adattároló kapacitás nagyon gyorsan növekszik, mégis egyre szélesebb körben terjed el a tömörített állományok használata. A hangokat, képeket és videókat nagyon nagy méretük miatt már régóta szinte kizárólag tömörítve tároljuk, hiszen már egy nem túl jó minőségűnek számító, 1,3 Megapixeles kép is a leggyakrabban használt, 32 bites színfelbontással körülbelül $1280 \cdot 1024 \cdot 32 = 41\,943\,040$ bites, ami 5 242 880 bájt = 5 Mebibájt (=5,242 Megabájt), hiszen minden egyes képpontra 32 bit információt kell tárolni. A videóknál még szembetűnőbb a helyzet, a szokásos 24 vagy 25 képkocka/másodperces frekvencia azt jelenti, hogy percenként 1440 vagy 1500 képkocka jelenik meg, azaz már $512 \cdot 512$-es felbontásnál és hang nélkül is másfél Gigabájt fölött van a videó percenként, hanggal pedig még több. Ehhez képest ennél nagyobb felbontású teljes filmek mérete gyakran 1 Gigabájt körül van.

Kevésbé ismert, de a legtöbb mai irodai formátum, úgy mint a docx, xlsx, pptx vagy az iWork által használt pages kiterjesztések mind igazából zip-pelt könyvtárak.[1]

A képeknél, videóknál még könnyebb elképzelni, hogy miért lehet őket kevesebb bittel is tárolni, hiszen például a szomszédos pixelek sokszor hasonló színűek és nem is változnak minden egyes képkockával, tovbbá a legtöbb esetben „csalhatunk" is kicsit, és mondhatjuk az egymáshoz térben és időben közeli, nagyon hasonló pixelekre, hogy egyformák. Az ilyen tömörítést hívják *adatvesztéses* vagy *adatveszteséges* tömörítésnek, mert ha egyszer egy fájlon végrehajtottuk, abból az eredetit már nem nyerhetjük vissza. Képek, videók és audió fájlok esetében ez megengedhető, sőt szükséges is, mert *adatvesztés nélküli* tömörítéssel még mindig nagy fájlokat kapunk.

Jelen dolgozatban az adatvesztéses tömörítéssel nem foglalkozunk, csak a veszte-

[1] Erről bárki meggyőződhet, csak át kell nevezni a fájt úgy, hogy kiterjesztést zip-re változtatjuk, és ha van zip tömörítőnk, máris meg lehet nézni, hogy milyen fájlok találhatók a fájlnak hitt könyvtárban. Többnyire xml fájlok, de például a pages állományokban van egy pdf is, így ezek iWork használata nélkül olvashatók.

ségmentessel.

A legtöbb fájlnál nem engedhető meg adatvesztés, hiszen ha például egy szöveges fájlnál kimarad vagy megváltozik néhány karakter, egész mást jelentést vehet fel a szöveg. Egy végrehajható fájl pedig valószínűleg nem is működne. Persze a többinél gyakrabban előforduló részek szöveges fájloknál, sőt minden adatállománynál megfigyelhetők, ha a fájlt, mint bit-sorozatot nézzük: bizonyos bitsorozatok időnként visszatérnek, vannak amik gyakrabban és vannak amik ritkábban.

Minden nyelvben, még a programnyelvekben is van egy eloszlása a szavaknak, ami viszonylag pontosan megadható. Persze ez nem jelenti azt, hogy ez az eloszlás minden egyes adatállományra érvényesül, de adott fájlban a bitek sorozatáról legtöbbször feltehető, hogy „az idő előrehaladtával", ahogy a sorozaton megyünk végig, nem változik az elemek eloszlása. Az ilyen sorozatokat nevezzük *ergodikusnak*, és ez egy nagyon fontos tulajdonság, mint ahogy arra a most következők rámutatnak.

A dolgozat 2. fejezetében bemutatom, hogy működik a zip tömörítés és a legfontosabb algoritmusok, melyekből felépül. A Huffman kódolást [5] és a Lempel-Ziv algoritmus egy változatát [6] egy – a módszerek erősségeit hangsúlyozni igyekvő – saját példán részletesen kidolgozva is ismertetem. A tömörítési eljárás és az algoritmusok leírásánál felhasználtam többek között Salomon–Bryant–Motta [9] művét és a pkzip tömörítő program dokumentációját [10] is.

Ezt követően ismertetem a szakdolgozatban felhasznált, de nem feltétlenül széleskörűen ismert fogalmakat, jelöléseket. A 4. fejezetben Ornstein és Weiss cikke [1] alapján három tétel részletes bizonyítását ismertetem. Az első tétel az ergodikus sorozatok entrópiájára ad becslést, az egyre hosszabb részsorozatok visszatérési idejének segítségével. Ezzel a módszerrel egy fájl tömörítésekor elérhető legjobb tömörítési arányt is meg lehet becsülni, a teljes fájl ismerete nélkül.

A másik két tétel, melyeket blokkhosszúsági tételeknek neveztem, a tömörítés során is használt blokkfelbontásról mutatja meg, hogy nem lesz túl sok nagyon rövid, se túl sok nagyon hosszú blokk. Ennek igazi jelentősége az, hogy így a zip tömörítésben használt algoritmusok az elméletileg lehetséges legnagyobb tömörítési arányt közelítik a tömörítendő állomány méretének növekedésével.

Az entrópia becsléslésről szóló tétel nem ad becslést minden jelsorozatra, csak

azokank egy 1 valószínűségű részére. A shift-terekben viszont ki lehet számolni az irreguláris jelsorozatok Hausdorff-dimenzióját, ezt Feng és Wu cikke [4] alapján az 5. fejezetben ismertetem.

Az entrópia becslés és a blokkhosszúsági tételek nagyon erősen támaszkodnak a Shannon-McMillan-Breiman tételre, így fontosnak tartottam, hogy ennek egy bizonyítása bekerüljön a szakdolgozatba. Algoet és Cover [3] rövid, ötletes módszerét választottam, ez a 6. fejezetben található.

2. zip tömörítés

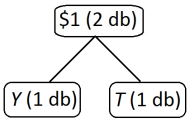
A *zip* tömörítési eljárás veszteségmentes, sok paraméterét lehet állítani, de az egyik algoritmus, amit gyakran használ a DEFLATE [8], melynek alapja a Huffman kódolás [5] és a Lempel-Ziv algortimus egy változata. Egy másik, gyakran használt tömörítési módja az LZMA algoritmus (Lempel-Ziv Markov lánc Algoritmus), ennek is a Lempel-Ziv az alapja. Bizonyos implementációkban a felhasználó beállíthatja, hogy melyik algoritmus szeretné használni.

2.1. Huffman kódolás

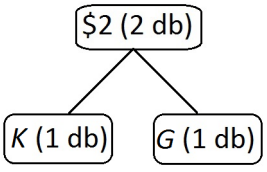
A Huffman kódolás lényege, hogy a gyakrabban előforduló jelek (amik lehetnek karakterek vagy karakter-együttesek is) rövidebb kódot kapnak, a ritkábbaknak pedig hosszabb kódja lesz. Ehhez először meg kell számolni az előforduló jelek gyakoriságát. Egy példán bemutatva:

Ha azt szeretnénk kódolni, hogy $ELEMELEK\_EGY\_ELEMET$, akkor a jelek gyakoriságai csökkenő sorrendben a következők: E 8 db, L 3 db, M 2 db, _ 2 db, K 1 db, G 1 db, Y 1 db, T 1 db (összesen 19 db). Ezeket egy listában tároljuk, és a segítségükkel rajzolunk egy gráfot, méghozzá egy bináris fát.

Kezdetben a listánk $(E, 8), (L, 3), (M, 2), (\_, 2), (K, 1), (G, 1), (Y, 1), (T, 1)$. Vesszük a két legritkább jelet (azonos gyakoriságok esetén tetszőlegesen választhatunk), ezek „testvérek” lesznek a fában, kikerülnek a listából, és a helyükre a „szülőjük” kerül, melynek gyakorisága a két „gyerek” gyakoriságának összege lesz. Ezt ismételgetjük, amíg az összes jel fel nem kerül a gráfra.

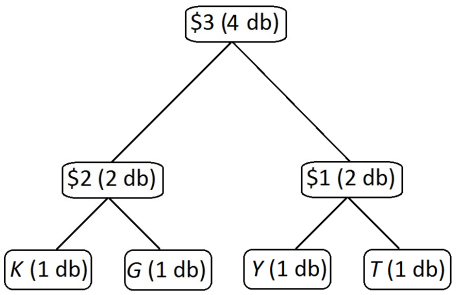


(a) 1. A legritkább jelek az Y és a T

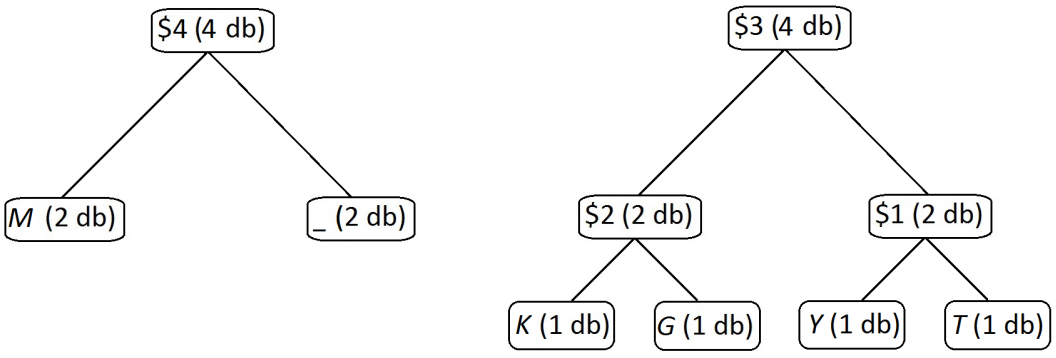


(b) K-ból és G-ből is csak 1-1 van

1. ábra. Ezután a lista így néz ki: $(E, 8), (L, 3), (M, 2), (\_, 2), (\$2, 2), (\$1, 2)$, tehát a következő lépésben $(\$1, 2)$ és $(\$2, 2)$ válnak testvérekké, $(\$3, 4)$ lesz a szülőjük.



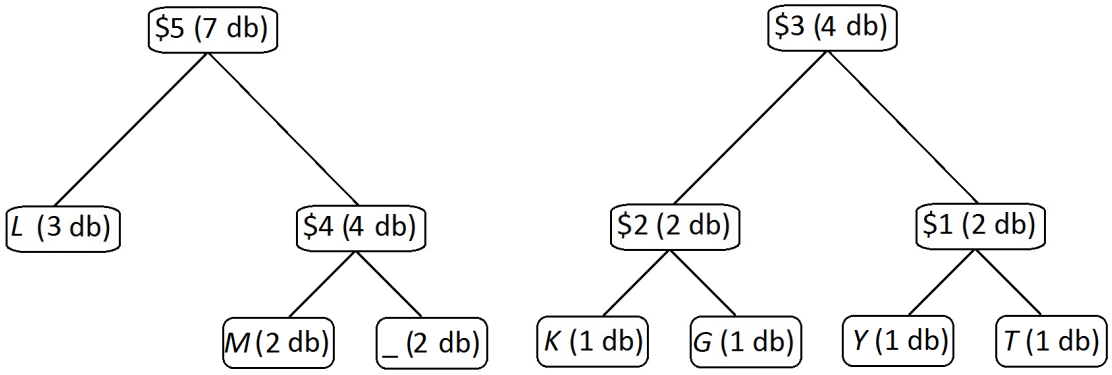
2. ábra. Az új lista $(E, 8), (\$3, 4), (L, 3), (M, 2), (\_, 2)$ lesz, így az új csúcsok $(M, 2)$ és $(\_, 2)$ lesznek, a szülőjük pedig $(\$4, 4)$ lesz.



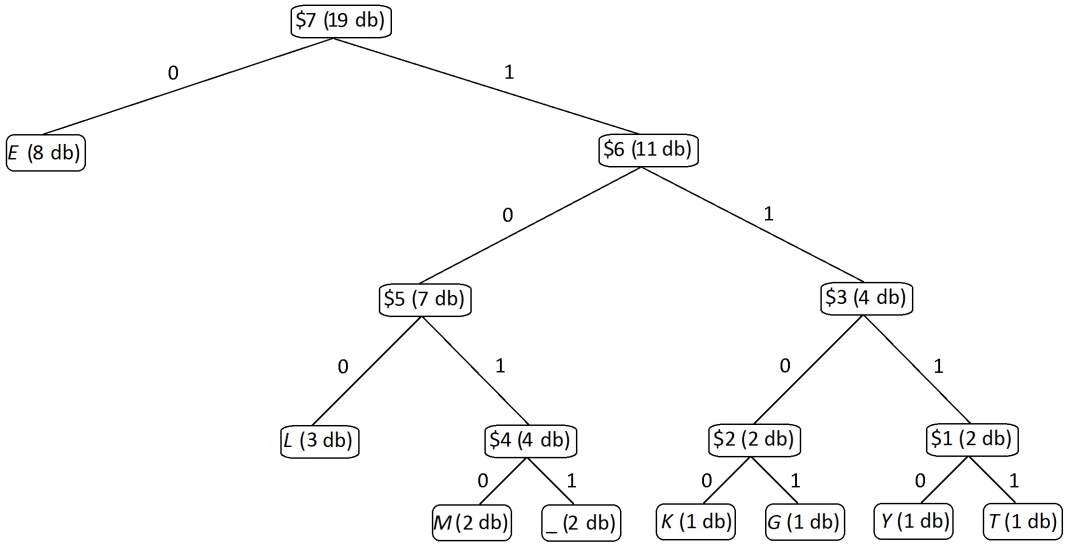
3. ábra. A lista most $(E, 8), (\$3, 4), (\$4, 4), (L, 3)$, az új testvérek tehát (\$4,4) és (L,3) lesznek.

Minden köztes csúcsban a darabszám a csúcsból származó levelekben található darabszámok összege.

A testvéreket mindig 1 bittel megkülönböztetjük, az egyik 0, a másik 1 (lásd: a 3., a 4. és az 5. ábrát). A fában a levelek felelnek meg a jeleknek, mindnek lesz egy egyértelmű bináris kódja, melyet úgy lehet kiolvasni a fáról, hogy a gyökértől az adott levélig vezető egyértelmű út csúcsainak bitjeit olvassuk be sorban. Így prefixmentes kódot kapunk, hiszen csak a levelek felelnek meg jeleknek, és egy gyökér-levél úton nem lehet több levél.



4. ábra. A lista most már csak 3 elemű: $(E, 8), (\$5, 7), (\$3, 4)$. Az utolsó két lépésben először $(\$5, 7)$ és $(\$3, 4)$ kapnak egy közös szülőt, $(\$6, 11)$-et, majd a szülőjük, $(\$6, 11)$ és $(E, 8)$ lesznek testvérek.



5. ábra. A Huffman kód gráfja.

0100010100100011001011011011110101101000101001111 tehát a Huffman kódja az $ELEMELEK\_EGY\_ELEMET$-nek (lásd az 1 táblázatot). Ez a hagyományos kódolás szerinti $8 \cdot 19 = 152$ bit helyett csak 49 bit, tehát a 19 karakteres szöveget leírtuk átlagosan nagyjából $2,5$ bit/karakteres kódolással, pedig 8 különböző karaktert használtunk.

Persze ez szándékosan egy olyan példa, ami nagyon lerövidül a Huffman kóddal, de más példa esetén is javulást láthattunk volna, de legalábbis romlást semmiképpen

Jel	E	L	M	_	K	G	Y	T
Kód	0	100	1010	1011	1100	1101	1110	1111

1. táblázat. A szótár, amit Huffman kódolással kaptunk

sem, hiszen a Huffman kód optimális prefix kód [5].

2.2. Lempel-Ziv algoritmus

A Lempel-Ziv algoritmusnak sokféle változata van, az egyik legelterjedtebb a Lempel-Ziv-Welch [6], ezt fogjuk bemutatni.

Az egyik legnagyobb előnye az, hogy a szótárt olyan módon építi fel, hogy azt a tömörített állományból is fel lehessen építeni, így nem kell a szótárat külön elküldeni, legfeljebb az elejét, bizonyos implementációk esetén. A tömörítéshez és a kicsomagoláshoz is elég egyszer végigmenni az adatállományon, így a kívánt műveletet már akkor elkezdhetjük, amikor még nem is ismerjük az egész állományt.

Tömörítés: A tömörítés során a szótár dinamikusan alakul ki, ahogy olvassuk a bitsorozatot, úgy készül. A bejegyzések olyan sorrendben kerülnek bele, ahogy a tömörítendő fájl(ok)ban előfordulnak. A szótárat nem is kell elküldeni, azt a tömörített állományból is elő lehet állítani.[2] Maga a tömörítés lényegében 4 lépés ismételgetéséből áll:

0. A tömörített állomány elejére kerül(het) egy kezdeti szótár.

1. A tömörítendő állomány beolvasása addig, amíg megtaláljuk a leghosszabb olyan sztringet, ami már szerepelt a szótárban.

2. A sztring, kiegészülve az utána következő karakterrel (tehát az első olyan sztring, amely nem szerepelt még) bekerül a szótárba, a kódja a soron következő szabad kód.

3. Az outputba a megtalált, leghosszabb, már ismert sztring kódja és az utána következő első karakter kódja (ha van) kerül.

[2]Persze a tömörített állomány fejlécében lehet információ a szótárról is, például, hogy van-e kezdeti szótár.

4. Vissza az 1. lépéshez

Azért jó, hogy a 3. lépésben nem az új szótárbejegyzést, hanem látszólag kevésbé hatékonyan a már korábban a szótárba került sztringet és az azt követő karaktert írjuk a kimenetbe, mert így a kicsomagolás során is fel lehet építeni a szótárat. Az LZW tömörítés legtöbb implementációja tudja kezelni a különböző hosszúságú kódokat, így nem tudunk „kifutni" a szabad kódokból. Ha az új kód már eggyel több bites, akkor az output is egy bittel hosszabb lesz, azaz például mikor szótárba bekerül a 32, akkor onnantól a 31 bináris kódja nem 11111, hanem 011111 stb.

Az előző példán is bemutatjuk a működését. Amit tömöríteni szeretnénk: *elemelek_egy_elemet*.

A kezdeti szótárba az egyszerűség kedvéért az angol abc betűi és a példánkban szóközként használt '_' kerülnek. A valóságban ennél jóval nagyobbak a kezdeti szótárak, legalább 255 bejegyzéssel.

Karakter	Kód	Karakter	Kód	Karakter	Kód
a	00	j	09	s	18
b	01	k	10	t	19
c	02	l	11	u	20
d	03	m	12	v	21
e	04	n	13	w	22
f	05	o	14	x	23
g	06	p	15	y	24
h	07	q	16	z	25
i	08	r	17	_	26

2. táblázat. A példánkban használt kezdeti szótár

Ez után kezdődik maga a tömörítés.

Az első karakter az *e*, ez már szerepel a szótárban, 04 a kódja, úgyhogy olvassuk tovább az inputot. A következő karakter az *l*, tehát amit eddig beolvastunk, az *el*. Ez még nem szerepel a szótárban, úgyhogy beletesszük. Ahogy a 2-es számú táblázat is mutatja, eddig 0-tól 26-ig osztottuk ki a kódokat, hogy az *el* kódja 27 lesz. Az outputba pedig beírjuk, hogy 0411, ugyanis az *e* kódja 04, az *l*-e pedig 11.

Input	Eddig beolvasva	Prefix	Prefix kódja	Új sztring?	Művelet	Új kód		Output
e	e	∅	∅	nem	tovább olvas		∅	∅
l	el	e	04	igen	Új bejegyzés	el	27	0411
e	e	∅	∅	nem	tovább olvas		∅	∅
m	em	e	04	igen	Új bejegyzés	em	28	0412
e	e	∅	∅	nem	tovább olvas		∅	∅
l	el	e	04	nem	tovább olvas		∅	∅
e	ele	el	27	igen	Új bejegyzés	ele	29	2704
k	k	∅	∅	nem	tovább olvas		∅	∅
_	k_	k	10	igen	Új bejegyzés	k_	30	1026
e	e	∅	∅	nem	tovább olvas		∅	∅
g	eg	e	04	igen	Új bejegyzés	eg	31	0406
y	y	∅	∅	nem	tovább olvas		∅	∅
_	y_	y	24	igen	Új bejegyzés	y_	32	2426
e	e	∅	∅	nem	tovább olvas		∅	∅
l	el	e	04	nem	tovább olvas		∅	∅
e	ele	el	27	nem	tovább olvas		∅	∅
m	elem	ele	29	igen	Új bejegyzés	elem	33	2912
e	e	∅	∅	nem	tovább olvas		∅	∅
t	et	e	04	igen	Új bejegyzés	et	34	0419

3. táblázat. LZW kódolás

Tovább olvasva a tömörítendő fájlt, a következő karakterek e (már benne van a szótárban) és m. Az em-nek még nincs kódja, tehát megkapja a soron következőt, a 28-at, a kimenetbe pedig beírjuk, hogy 0412 (az e és az m kódjai). A következő karakterek e (már a szótárban van) l (el is benne van már) és újra e. ele még nem szerepelt, ez lesz a következő bejegyzés, a kódja pedig 29. Az outputba persze nem a 29 kerül, akkor a kicsomagolásnál nem tudnánk, hogy az minek a kódja. Viszont így, hogy a kimenetbe 2704-et írunk (az el és az e kódja), a kicsomagolás során is tudni fogjuk, hogy itt az ele karakterlánc szerepel, ami megkapja a kövekező szabad kódot, a 29-et. Legközelebb, mikor újra szerepel az ele az állományban, már hivatkozhatunk rá, mint 29-re a tömörített állományban (ahogy azt tettük most az el-lel, ami helyett 27-et írtunk).

Így megy tovább, amíg a végére nem érünk a tömörítendő adatállománynak. A 3-adik táblázatban megtekinthető a teljes példamondat tömörítése.

A tömörített állomány, amit a végén kapunk:

041104122704102604062426291204 19

Ez kicsit hosszabb, mint az eredeti, mert (értelemszerű okokból) egy nagyon rövid sztringet választottunk példának. A tömörítő hatását csak nagyobb állományokon fejti ki, hiszen a működéséből adódóan a szótárban eleinte rövid szavak lesznek, de ahogy előrehaladunk a szövegben, egyre hosszabb sztringek kapnak (nagyjából) ugyanolyan hosszú kódokat, tehát egyre hatékonyabb lesz az egész.

Ez azért sem igazán hátrány, mert tömörítésre úgyis általában nagyobb állományok esetén van szükség. Jelen szakdolgozat forrásfájlja (ami még mindig igen kicsi állománynak számít, lényegesen kisebb, mint 1 megabájt) például körülbelül az egyharmadára csökken a tömörítés során.

A kitömörítés is egyszerű, megkapjuk a tömörített állományt, jelen esetben 041104122704102604062426291204 19-et, és végigmegyünk rajta, közben felépítve a szótárat. A kezdeti szótár ismert, így 04-ről tudjuk, hogy az e, így megvan az eredeti karakterlánc első eleme. A 11 az l-t jelöli, ez a második karakter, de most a szótár is bővül, el kapja a következő szabad kódot, 27-et. A következő két elem 04 és 12, ezek az e és az m karaktereket adják a dekódolt szövegben, az em pedig bekerül a szótárba, 28 lesz a kódja. Az input következő eleme 27, ez már nincs benne a kezdeti szótárban, de mivel a kitömörítés során is építjük a szótárat, tudjuk, hogy a 27 az el kódja, tehát a kicsomagolt szöveg is ezzel folytatódik. A kódolt karakterlánc 04-gyel

Input	Output	Szótár bővítése	
		kód	szó
04	e	∅	
11	l	27	el
04	e	∅	
12	m	28	em
27	el	∅	
04	e	29	ele
10	k	∅	
26	_	30	k_
04	e	∅	
06	g	31	eg
24	y	∅	
26	_	32	y_
29	ele	∅	
12	m	33	elem
04	e	∅	
19	t	34	et

4. táblázat. LZW dekódolás

folytatódik, így a szöveg *e*-vel, a szótár pedig a 29-cel jelölt *ele*-vel folytatódik. Így megy, amíg az egész kódot vissza nem fejtjük. A teljes kitömörítés megtekinthető a 4-es számú táblázatban.

Még említésre érdemes, hogy a zip állomány végén található egy címjegyzék a tömörített fájlokról és mappákról, így külön-külön is ki tudjuk tömöríteni a fájlokat, mappákat, ha úgy szükséges. Egy zip fájlba lehet tenni önkicsomagolót is, ekkor a fájl ki tudja csomagolni saját magát, viszont pont emiatt rosszindulatú programokat is bele lehet kódolni egy egyébként ártatlan állományba.

Még egyszer összefoglalva, lényegében 3 lépésből áll a tömörítés:

1. Blokk-felbontás: az adatsort blokkokra osztjuk aszerint, hogy az épp beolvasott sztring szerepelt-e már korábban.
2. Kódolás: minden blokk kap egy kódot
3. Kódok láncba szedése: olyan sorrendben követik egymást a kódok, ahogyan az általuk képviselt blokkok.

Ahogy Ornstein és Weiss [1] rávilágít, van két nagyon fontos tulajdonsága ennek a tömörítési eljárásnak, melyek minden változatra igazak:

1. A blokkok páronként különbözők.
2. Minden blokk, ami szerepel a felbontásban, már korábban is szerepelt a szövegben valamilyen formában[3]

A szakdolgozat szempontjából az 1. tulajdonság jelentősége az, hogy a 4.2.3 tétel alapján a felbontás során nem keletkezik sok rövid blokk. Ez azért jó hír, mert a rövid blokkok rosszul tömörítenek, sőt, a nagyon rövidek nem is tömörítenek egyáltalán.

[3]Ha például a 14-es az abcdaf betűkombináció által adott blokk kódja, akkor ha a kódban szerepel valahol a 14-es, akkor az eredeti szövegben már korábban is szerepelt az abcdaf kombináció, csak nem egy egész blokként, hanem vagy több különböző blokkban vagy egy hosszabb blokk részeként (de nem annak kezdőszeleteként)

A 2. tulajdonságból következően pedig – ahogy arra a 4.2.12 tételben rámutatunk – az is biztosított, hogy nem keletkezik sok túl hosszú blokk sem. Ez viszont azt jelenti, hogy a blokkok nagyrésze hasonló hosszúságú lesz, méghozzá egy n-hosszú, h entrópiájú sztring esetén körülbelül $\log n/h$, és ez fogja biztosítani, hogy a tömörítés a lehető leghatékonyabb.

3. Néhány alapfogalom

A 4. fejezetben szinte végig ergodikus sorozatokkal fogunk foglalkozni, ehhez viszont elengedhetetlen néhány fogalom tisztázása.

A dolgozat témájából adódóan szinte mindenhol kizárólag a diszkrét esetre lesz szükségünk.

3.0.1. Definíció (Sztochasztikus folyamat). *Véletlen változók együttese. Adott* $(\mathcal{X}, \mathcal{P}, \mu)$ *valószínűségi mező,* (S, Σ) *mérhető tér és* T *jólrendezett halmaz esetén az* $\{X_t, t \in T\} = \{X_t\}$ S*-értékű,* $\mathcal{X}$*-beli valószínűségi változók egy együttesét* S*-beli értékű sztochasztikus folyamatnak nevezzük.* S*-t nevezzük állapottérnek,* T*-t pedig időnek.*

Jelen dolgozatban csak diszkrét idejű és diszkrét állapotterű sztochasztikus folyamatokkal foglalkozunk.

3.0.2. Definíció (Stacionárius folyamat). *Az* $(\mathcal{X}, \mathcal{P}, \mu)$ *valószínűségi téren egy* $\{X_n, n \in \mathbb{Z}\}$, $\mathcal{X}$*-beli értékű folyamat pontosan akkor* stacionárius*, ha minden* $k \geq 1$*-re az* $(X_n, X_{n+1}, \ldots X_{n+k-1})$ k*-as független* $n \in \mathbb{Z}$*-től.*

Más szavakkal, egy stacionárius folyamat együttes valószínűségi eloszlása nem függ az időtől, mindig ugyanaz marad.

3.0.3. Definíció (T-invariáns halmaz). *Egy* $T : H \to H$ *mértéktartó és invertálható leképezés esetén egy* $A \subset H$ *halmazt* T*-invariánsnak nevezünk, ha* $(A \backslash T^{-1}A) \cup (T^{-1}A \backslash A)$ *nullmértékű.*

3.0.4. Definíció (Ergodikus leképezés és folyamat). *Az* $(\mathcal{X}, \mathcal{P}, \mu)$ *valószínűségi mértéktéren a* T *mértéktartó és invertálható leképezés* ergodikus*, ha minden* T*-invariáns* A *halmazra* $\mu(A) \in \{0, 1\}$.

Ergodikus folyamat *az* $(\mathcal{X}, \mathcal{P}, \mu, T)$ *együttes, ahol* T *ergodikus leképezés.*

3.0.5. Definíció (Ergodikus sorozat). *Ha* $T : \mathcal{X} \to \mathcal{X}$ *ergodikus leképezés, akkor az* $x, Tx, T^2x, \ldots$ *sorozatot* ergodikus sorozatnak *nevezzük.*

3.0.6. Megjegyzés. Egy ergodikus folyamat szükségképpen stacionárius is.

Az ergodikus folyamatok egy fontos tulajdonságát fogalmazza meg a következő tétel arról, hogy milyen értékeket vehetnek fel a folyamar elemei 1 valószínűséggel.

3.0.7. Tétel (Shannon–McMillan–Breiman).

$$-\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) = -\frac{1}{n}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\ldots,X_0) \to H \quad 1 \quad \textit{valószínűséggel}$$

Ahol $\{X_t\}$ *egy stacionárius ergodikus folyamat, melynek értékkészlete egy* $\mathcal{X}$ *megszámlálható halmaz,* p *a valószínűséget jelöli, és* $H = \lim_{k\to\infty}\frac{1}{k}H(X_1,\ldots,X_k)$ $= \lim_{k\to\infty} E(-\log p(X_k|X_{k-1},\ldots,X_0))$ *az entrópia rátája a folyamatnak.*

A 6.0.8 tétel ennek véges értékkészletű változata, bizonyítása megtalálható a dolgozat 6. fejezetében. Fő elemei a csendőrelv, Lévy martingál kovenergencia tétele és ergodelméleti összefüggések.

3.0.8. Definíció (Aszimptotikus ekvipartíciós tulajdonság). *Azon (nem feltétlenül véges értékkészletű) diszkrét idejű ergodikus folyamatokra, melyekre teljesül a 3.0.7 tétel, azaz* $-\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) \to H$*, 1 valószínűséggel, azt mondjuk, hogy teljesül rájuk az* aszimptotikus ekvipartíciós tulajdonság. [4]

A Shannon-McMillan-Breiman tétel tehát azt mondja ki, hogy minden megszámlálható értékkészletű ergodikus sorozatra teljesül az aszimptotikus ekvipartíciós tulajdonság.

A 4. fejezetben a bizonyítások során sokat fogjuk használni az alaphalmaz partícióit, illetve azok finomításait. A következő jelöléseket használom:

3.0.9. Jelölések.

1. $(\mathcal{X},\mathcal{B},\mu)$ *valószínűségi mező, ahol* $\mathcal{X}$ *az eseménytér,* $\mathcal{B}$ *a* σ*-algebra és* μ *a valószínűségi mérték.*
2. $T:\mathcal{X}\to\mathcal{X}$ *mértéktartó leképezés, ez lesz az ergodikus folyamat időbeli eltolása vagy „timeshift"-je.*
3. $\mathcal{X}$ *egy partíciója* $\mathcal{P}$*, azaz* $\mathcal{P}$ *egy* $\mathcal{X}\to\{$*a partíció elemei* $\}$ *leképezés. Például ha* $\mathcal{X} = \{x_1,x_2,x_3,x_4,x_5,x_6\}$ *és a* $\mathcal{P}$ *partíció az* $A = \{x_1,x_2\}, B = \{x_3,x_4\}, C =$

[4] Angolul: Asymptotic Equipartition Property vagy röviden AEP

$\{x_5, x_6\}$ *halmazokra osztja* $\mathcal{X}$*-et, akkor* $\mathcal{P}$*-re tekinthetünk úgy, mint egy* $\mathcal{X} \to \{A, B, C\}$ *leképezésre, ahol* $x_1 \mapsto A$*,* $x_2 \mapsto A$*,* $x_3 \mapsto B$ *stb. A partíció elemeit (jelen esetben A-t, B-t és C-t) atomoknak is nevezzük.*

4. *Ha* $\mathcal{P} = \{P_1, P_2, \ldots, P_a\}$ *és* $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_b\}$ *X partíciói, akkor ezek közös finomítása* $\mathcal{P} \vee \mathcal{Q} = \{P_i \cap Q_j : 1 \leq i \leq a, 1 \leq j \leq b\}$.

   A $\mathcal{Q}_1, \mathcal{Q}_2, \ldots \mathcal{Q}_k$ *véges sok partíció közös finomítása indukcióval történik:*

$$\vee_{i=1}^{k} \mathcal{Q}_i = \mathcal{Q}_1 \vee \mathcal{Q}_2 \vee \cdots \vee \mathcal{Q}_k = (\mathcal{Q}_1 \vee \cdots \vee \mathcal{Q}_{k-1}) \vee \mathcal{Q}_k,$$

   ahol $\mathcal{Q}_i$*-k az* $\mathcal{X}$ *egy-egy partíciói.*

5. *A* $\mathcal{P}$ *partíciónak azt az elemét jelöljük* $\mathcal{P}(T^n x)$*-szel, melyben* $T^n x$ *van.*

   $\mathcal{P}(T^n x) = \mathcal{P}(T^n y) = P_i$ pontosan akkor, ha x és y is $T^{-n}P_i$-ben van. Ez azt jelenti, hogy $\mathcal{P}(x) = \mathcal{P}(y)$ és $\mathcal{P}(T^n x) = \mathcal{P}(T^n y)$ közül egyik sem következik a másikból. Ezért érdemes külön definiálni az n-edik képek által meghatározott partíciót:

6. $T^{-n}\mathcal{P}$ *egy partíciója* $\mathcal{X}$*-nek,* x *és* y *pontosan akkor tartozik* $T^{-n}\mathcal{P}$ *ugyanazon atomjához, ha* $\mathcal{P}(T^n x) = \mathcal{P}(T^n y)$.

7. *Az* $y_n(x) = \mathcal{P}(T^n x)$ *által meghatározott stacionárius folyamat* $\{y_n\}$*. Ez tényleg stacionárius, mert minden lépésben úgy kapjuk* y_i*-t* y_{i-1}*-ből, hogy a* T *mértéktartó leképezést alkamazzuk* $\mathcal{X}$*-en, és megnézzük, hogy* x *képe a partíció melyik atomjába kerül. Ez persze független attól, hogy előtte hányszor alkalmaztuk már* T*-t.*

8. $P_n(x)$ *jelöli* $\vee_{i=1}^{n-1}(T^{-i}\mathcal{P})$ *azon atomját, melybe* x *esik.*

9. $N(n, c, \mathcal{P}) = \min\{k : \exists P_{n_1}, P_{n_2}, \ldots P_{n_k} \in \vee_{i=0}^{n-1}(T^{-i}\mathcal{P})$ *melyre* $\mu(\cup_{j=1}^{k} P_{n_j}) > c\}$,

   azaz $N(n, c, \mathcal{P})$ *a legkisebb olyan* k *szám, melyre tudunk venni* k *darab atomot* $\vee_{i=0}^{n-1}(T^{-i}\mathcal{P})$*-ből úgy, hogy ezek összmértéke* c*-nél nagyobb legyen.*[5] *A kifejezés csak* $c < 1$ *esetén értelmezhető, mivel valószínűségi mértéktéren vagyunk.*

[5] Ez persze nem jelenti azt, hogy tetszőleges k atomot kiválasztva ezek összmértéke c-nél nagyobb lesz, csak azt, hogy létezik olyan atom–k-as, hogy azok együttes mértéke nagyobb, mint c.

10. *A 3.0.7 értelmében* $h(T,\mathcal{P}) = \lim_{c\nearrow 1} \overline{\lim}_{n\to\infty} \frac{1}{n} \log N(n,c,\mathcal{P})$ *a* $(T,\mathcal{P})$ *folyamat entrópiája.*

    Mivel csak ergodikus folyamatokkal foglalkozunk, az első limeszre nincs is szükségünk.

    A továbbiakban $\mathcal{P}$ alatt olyan felosztást értünk, melyre minden atom mértéke pozitív

11. $R_n(x) = \min\{k \geq n : \mathcal{P}(T^{k+i}x) = \mathcal{P}(T^i x)$ *minden* $0 \leq i \leq n-1$*-re*$\}$ *a* $(T,\mathcal{P})$ *folyamat* visszatérési ideje.

    $R_n(x)$ *azt mutatja meg, hogy* x *„pályája" a* $\mathcal{P}$ *partíció atomjain mikor járja be leghamarabb pontosan ugyanazt az* n *atomot, ugyanabban a sorrendben, mint az első* n *lépésben.*

12. $\overline{R}(x) = \limsup_{n\to\infty} \frac{\log R_n(x)}{n}$,

13. $\underline{R}(x) = \liminf_{n\to\infty} \frac{\log R_n(x)}{n}$,

14. χ_A *az* A *halmaz* karakterisztikus függvénye, *azaz*

$$\chi_A(x) := \begin{cases} 0 & \text{ha } x \notin A \\ 1 & \text{ha } x \in A. \end{cases}$$

4. Entrópia és adattömörítés

4.1. Entrópia becslés

Ornstein és Weiss [1] adtak egy új módszert entrópia-becslésre, illetve az LZW tömörítés jóságára adtak egy új bizonyítást.

A zip tömörítésben a jelsorozat blokkokra bontása az alapján történik, hogy a legutóbbi blokk óta beolvasott szelete a jelsorozatnak szerepelt-e már korábban, és pontosan akkor „zárjuk le" az új blokkot, ha olyan jel következik, amely még nem szerepelt olyan szelet után, mint amit épp beolvasunk. Mivel a jelek (szimbólumok) halmaza véges, ezért ahogy olvassuk végig a sorozatot, egyre hosszabb blokkokat fogunk készíteni. Ha egy új, n-hosszú blokk (röviden: n-blokk) kerül a gyűjteménybe, akkor ennek a blokknak az első $n-1$ eleme ugyanebben a sorrendben már szerepelt korábban is a jelsorozatban, tehát a sorozat ezen $(n-1)$-hosszú szeletének visszatérési ideje nem lehet nagyobb, mint a két blokk első elemei közötti távolság. Másképp fogalmazva, egy $(n-1)$-blokkot a gyűjteménybe kerülése után leghamarabb csak *visszatérési ideje* elteltével „használunk" először, azaz akkor lesz először olyan n-blokkunk, melynek kezdőszelete ő.

A most következő állítás szerint ezek a visszatérési idők n-ben exponenciálisan növekednek, sőt, a logaritmusukat véve, majd n-nel osztva a határértékük pont a sorozat entrópiáját adja. Ez azért jó, mert azt egyébként csak a különböző jelek bekövetkezésének valószínűségéből tudnánk kiszámolni, amit viszont tipikusan nem ismerünk.

4.1.1. Tétel (Entrópia becslés).

$$\frac{\log R_n(x)}{n} \to H(x_1, x_2, \dots) \text{ 1 valószínűséggel,}$$

ahol $x = x_1, x_2, \dots$ *egy véges értékkészletű ergodikus sorozat, és* $R_n(x) = \min\{j \geq n : x_1 x_2 \dots x_n = x_{j+1} x_{j+2} \dots x_{j+n}\}$ *a sorozat visszatérési ideje, valamint* $H(x_1, x_2, \dots)$ *a sorozat entrópiája.*

Bizonyítás A korábbiakban definiáltak szerint a T mértéktartó leképezés által meghatározott $\{y_n\}$ ergodikus sorozatra látjuk be, hogy $\frac{R_n(y_n(x))}{n} \to H\left(y_n\left(x\right)\right)$.

$\min\left\{k \geq n-1 : \mathcal{P}\left(T^{k+1+i}x\right) = \mathcal{P}\left(T^{i+1}x\right) \text{ minden } 0 \leq i \leq n-2\text{-re}\right\} = R_{n-1}(Tx)$, ami pedig nem nagyobb, mint $R_n(x)$, hiszen ha minden $0 \leq i \leq n-1$-re $\mathcal{P}(T^{k+i}x) = \mathcal{P}(T^i x)$, akkor minden $0 \leq i \leq n-2$-re $\mathcal{P}(T^{k+i+1}x) = \mathcal{P}(T^{i+1}x)$, így a legkisebb k, melyre ez utóbbi teljesül, nem lehet nagyobb, mint a legkisebb k, melyre az előbbi igaz. Ebből következik, hogy $\overline{R}(Tx) \leq \overline{R}(x)$ és $\underline{R}(Tx) \leq \underline{R}(x)$, azaz $\underline{R}(x)$ és $\overline{R}(x)$ szubinvariáns függvények, ez pedig azt jelenti, hogy valójában invariánsak, hiszen valószínűségi mértéktérről van szó, azaz a mérték véges. Ebből pedig az következik, hogy majdnem mindentütt konstansak, mivel ergodikus leképezésekről van szó.A lim sup és a lim inf definíciója alapján azt is tudjuk, hogy $\underline{R} \leq \overline{R}$, ahol $\underline{R}$ és $\overline{R}$ rendre az $\underline{R}(x)$ és az $\overline{R}(x)$ által majdnem mindenütt felvett konstans értékek.

A tétel bizonyításához elég azt megmutatni, hogy $\underline{R} \geq h = H(x_1, x_2, \dots) \geq \overline{R}$, mert ez azt jelenti, hogy létezik a tételben szereplő határérték, és valóban az entrópiával egyenlő.

Tegyük fel indirekt, hogy hogy $\underline{R} < h$. Ekkor létezik b, b_0, melyekre $\underline{R} < b < b_0 < h$.

Definiáljuk az $A_N, B_L \subset \mathcal{X}$ részhalmazokat ($N, L \in \mathbb{Z}^+$).

$$A_N = \{x \in \mathcal{X} : \exists n \in [N_0, N] \text{ melyre } \frac{\log R_n(x)}{n} \leq b\},$$

$$B_L = \{x \in \mathcal{X} : \frac{1}{L}\sum_{k=1}^{L} \chi_{A_N}(T^k x) \geq 1 - 2\varepsilon\},$$

Azaz $A_N \subset \mathcal{X}$-ben azok az x-ek vannak, melyekre (egy rögzített $N_0 \leq N$ mellett) $(\log R_n(x))/n \leq b$ valamely $N_0 \leq n \leq N$-re, és $B_L \subset \mathcal{X}$-ben azok az x-ek vannak, melyeknek T-szerinti első L képéből legalább $(1-2\varepsilon)L$ esik A_N-be. Máshogy megfogalmazva, nevezzük *érvényesnek* azokat az n számokat, melyek N_0 és N közé esnek. A_N-ben azok az x-ek vannak, melyekhez létezik olyan érvényes n, melyre T-t n-szer iterálva x-en, $\mathcal{X}$ partíciójának különböző elemein olyan sorrendben jár, amely sorrend a további iterációi során *hamar*, azaz (A_N definíciójának megfelelően) nem több, mint $\exp(bn)$ lépés elteltével ismét előfordul.

B_L azokat a (nem feltétlenül A_N-beli) x-eket tartalmazza, melyeken T-t iterálva L-szer, az iterációk között elég sok olyan $T^i x$-et kapunk, melyből indul érvényes hosszúságú, hamar visszatérő iteráció-sorozat, string.[6]

[6]A B_L-ben azon x-ek vannak, melyek pályájának elég sok szelete hamar visszatér. Pontosabban,

Mivel $b > \underline{R} = \liminf_{n\to\infty} \frac{\log R_n(x)}{n}$, bármely rögztett $\varepsilon > 0$ esetén, elég nagy N-re $\mu(A_N) \geq 1 - \varepsilon$. Választható olyan, kicsi ε és hozzá elég nagy N_0[7] illetve L, hogy $\mu(B_L) > 0,99$ legyen.[8]

Két lépésben megszámoljuk, hogy adott $x \in B_L$ hány különböző atomhoz tartozhat $\vee_{n=0}^{L-1} T^{-n}P$ -ben. Első lépésben $\{0, 1, \ldots, L-1\}$-et rögzített blokkokra osztva nézzük meg a lehetséges esetek számát, második lépésben pedig arra adunk felső becslést, hogy blokkokra osztani hányféleképpen lehet.

Rögzített $x \in B_L$ mellett $\{0, 1, \ldots, L-1\}$-et partícionáljuk aszerint, hogy adott i-re $T^i x$ éppen A_N-beli-e: az első blokk $[0]$, ha $x \notin A_N$, illetve $[0, \ldots, n-1]$, ha $x \in A_N$ és $n > N_0$ a legkisebb olyan szám, melyre $\frac{\log R_n(x)}{n} \leq b$ vagyis n a legkisebb olyan érvényes blokkhosszúság, amire sorozat első n eleme hamar visszatér.[9]

Ha $[0, \ldots, k-1]$-et már partíciókra osztottuk, akkor a következő blokk $[k, k+1, \ldots, k+m-1]$, amennyiben $T^k x \in A_N$, és $m \geq N_0$ a legkisebb olyan szám, melyre $\frac{\log R_m(T^k x)}{m} \leq b$ és $k + m < L$ Más szavakkal, ha T-t k-szor iterálva x képe A_N-beli, és k-tól $L-1$-ig még „belefér" a legrövidebb olyan érvényes m blokkhossz, melyre $T^k x$ első m képe hamar visszatérő sorozatot ad, akkor a következő blokk m-elemű: $[k, k+1, \ldots, k+m-1]$. Ha ez már nem fér bele, vagy $T^k x$ eleve nem volt A_N-ben, akkor a következő blokk $[k]$.

Ezzel kikényszerítjük, hogy ha $T^i x \in A_N$ (aminek több, mint $1 - \varepsilon$ a valószínűsége), akkor $i < L - N$ esetén (tehát amikor $\{1, 2, \ldots L-1\}$-be még biztosan belefér bármilyen érvényes hosszúságú blokk) i egy legalább N_0-hosszú blokkba kerül.

Nézzük, hány különböző szimbólumot kaphat $\mathcal{P}(T^i x)$ egy ilyen rögzített x esetén, vagyis a $\mathcal{P}$ partíció hány különböző elemébe eshet $T^i x$. A követhetőség érdekében fontos megjegyezni, hogy egy $\mathbb{N} \supset \{1, 2, \ldots, L\} \ni i \mapsto T^i x \mapsto \{\mathcal{X}$ partíciójának elemei$\}$ leképezés-láncolat adja meg, hogy az egymást követő természetes számokból álló adott blokk milyen (a partíció különböző elemeit jelölő) jeleket kaphat a kódolás során.

---

a pálya pontjainak $(1-2\varepsilon)$-részéből indul érvényes hosszúságú, hamar visszatérő szelet.

[7] a konkrét érték $h - b$-től és a $\mathcal{P}$-beli szimbólumok számától, azaz a partíció elemszámától függ

[8] Ez nem túl meglepő, hiszen, b-t a liminf-nél nagyobbnak választottuk, így A_N nagyon nagy halmaz, elég sok x-nek lesznek a képei többnyire A_N-ben, ezért elég sok olyan x is lesz, mely képeinek nagy része A_N-ben, azaz $x \in B_L$.

[9] Tehát, ha $x \in A_N$, akkor az első blokkot a „hamar visszatérő szelet" elemeinek az indexei adják.

Most rögzítsük az $\{1, 2, \dots L-1\}$ blokkokra bontását, és nézzük meg, hányféleképpen kaphatnak ezek a blokkok jeleket.

A blokkokon visszafelé haladva keresünk felső korlátot a lehetséges jelek számára. A blokkokra osztott intervallum végén, $[L-N, L]$-ben már az egyébként gyakori $T^n x \in A_N$ esetben is előfordulhat, hogy n mégis rövid blokkba kerül, mert egy hosszú blokk vége már L után lenne. Látni fogjuk, hogy a hosszú blokkok esetén kevesebb lehetőségünk van, mint ha az adott részt csupa 1-hosszú blokkokra osztanánk, így, mivel felső becslést szeretnénk adni, feltehetjük, hogy ebben a részben csupa 1-hosszú blokkok lesznek. Ha a részre osztja $\mathcal{X}$-et $\mathcal{P}$, akkor $[L-N, L]$-hez a^{N+1}-félképpen választhatunk kódjeleket [10], hiszen $\mathcal{X}$ minden egyes elemét, így a $T^i x$-eket is a különböző részébe partícionálhatjuk. A hosszabb, m_j hosszú blokkok esetén viszont, bár $\{1, 2, \dots L\}$-ben m_j a blokk hossza, a neki adható jelek számára a^{m_j} mellett van egy másik felső becslésünk is, mégpedig $\exp(bm_j)$. i csak akkor „indíthat" $1 < m_j$-hosszú blokkot, ha $T^i x \in A_N$, azaz létezik olyan $N_0 \leq n \leq N$, melyre $\frac{\log R_n(x)}{n} \leq b$, azaz $R_n(x) \leq \exp(bn)$, és a legkisebb ilyen n lesz m_j. Ezzel $R_{m_j}(x) \leq \exp(bm_j)$, azaz x-nek a blokkbeli indexek által meghatározott képei olyan utat járnak be a partícióban, ami legkésőbb $\exp(bm_j)$ lépéssel később megismétlődik (ennyi a legnagyobb távolság, amit megtehetünk jobbra, mielőtt az egész m_j-blokkot végigvennénk). Az $[1, L-N-1]$-ben esetleg előforduló 1-hosszú blokkok természetesen most is a-féle jelet kaphatnak.

x-ről feltettük, hogy B_L-beli, de az ilyen x-ek is (legfeljebb) az iterációk 2ε-ad részénél, azaz $2\varepsilon L$ esetben A_N-en kívül lehetnek, tehát ennyi iterációhoz tartozó index $[1, 2, \dots L]$-ben 1-hosszú blokkot adhat.

Így kapjuk, rögzített partíció esetén ezt a felső becslést arra, hogy a blokkok hányféleképpen kaphatnak jeleket:

$$a^{N+1} \left(\prod_j \exp\left(bm_j\right) \right) a^{2\varepsilon L} \leq a^{\varepsilon L} a^{2\varepsilon L} \exp\left(b \sum_j m_j \right) =$$

$$a^{3\varepsilon L} \exp\left(b \sum_j m_j \right) \leq \exp(bL) a^{3\varepsilon L}$$

Az utolsó egyenlőtlenséget onnan kapjuk, hogy $\sum_j m_j \leq L$, és L-et olyan nagyra választjuk, hogy egyrészt teljesüljön az első egyenlőtlenség, azaz $N+1 \leq \varepsilon L$, másrészt az egészet felülről becsülhessük $\exp(b_0 L)$-lel.

[10] Kódjel itt a partíció egy elemét jelenti, x kódjele az annak a partíciónak a jele, amibe kerül

Most azt nézzük meg, hogy a blokkokra osztás legfeljebb hányféleképpen történhet. Rövid, azaz 1-hosszú blokkokat legfeljebb $\sum_{j\leq 3\varepsilon L}\binom{L}{j}$-féleképpen választhatunk, mert a rövid blokkok száma 0 és $3\varepsilon L$ közé esik, és ha j darab van, akkor értelemszerűen $\binom{L}{j}$-féleképpen lehet elhelyezni őket. A hosszú blokkok legalább N_0 hosszúak, így maximum L/N_0 darab hosszú blokk van, ezek kezdőeleme legfeljebb $\sum_{j\leq L/N_0}\binom{L}{j}$-féleképp választható. Ez a két becslés nagyon gyenge, hiszen egy rövid blokk, ha nem közvetlenül az előző rövid blokk után jön, akkor legalább N_0-lal később következik, egy hosszú blokk pedig mindenképpen legalább N_0-lal később kezdődik, mint az előző hosszú blokk. Ezeket a megszorításokat nem vettük figyelembe, de így is ki fog jönni az állítás.

Ezekkel a felső becslésekkel kapjuk, hogy megfelelően kicsi ε és megfelelően nagy N_0 esetén $x\in B_L$ legfeljebb $\exp(b_0L)$ atomjához tartozhat $\vee_{n=0}^{i-1}T^{-n}\mathcal{P}$-nek. Mivel $b_0<h$, ez ellentmondás.

Most már csak a másik egyenlőtlenséget kell igazolni, miszerint $h\geq\overline{R}$.

$\vee_{j=0}^{n-1}T^{-j}\mathcal{P}$ egy rögzített D atomja nagy n-re megfelel egy $u_0u_1\cdots u_{n-1}$ blokknak, és ha rögzítünk egy $c>h$-t is, akkor $D'=\{x\in D:\log R_n(x)/n>c\}$ mértéke legfeljebb $1/\exp(cn)$. Ennek az az oka, hogy $D',TD',\ldots,T^{\exp(cn)-1}D'$ mind ugyanakkora mértékű, mivel T mértéktartó, és páronként diszjunktak, ebből adódóan $1\geq\mu\left\{D'\cup TD'\cup\cdots\cup T^{\exp(cn)-1}D'\right\}=\mu(D')+\mu(TD')+\cdots+\mu(T^{\exp(cn)-1}D')=\exp(cn)\mu(D')$

A diszjunktság belátásához először vegyük észre, hogy x és y pontosan akkor vannak $\vee_{j=0}^{n-1}T^{-j}\mathcal{P}$ ugyanazon atomjában, ha a $\mathcal{P}$ partíciónak páronként ugyanabban az atomjában van $x,Tx,\ldots,T^{n-1}x$ és $y,Ty,\ldots,T^{n-1}y$ (azaz az első n képük ugyanazt az „utat” járja be $\mathcal{P}$-ben). Ez természetesen igaz D-re is, tehát ha valamely $x\in D'$-re létezne olyan $i\in[1,\exp(cn)-1]$, melyre $T^ix\in D'$ lenne, akkor $R_n(x)<\exp(cn)$ lenne, de ez ellentmond annak, hogy $x\in D'$.

A 3.0.7 tétel értelmében

$$\lim_{n\to\infty}-\frac{\log\mu(P_n(x))}{n}=h,$$

ahol $P_n(x)$ a $\vee_{j=0}^{n-1}T^{-j}P$ partíció azon atomja, melyhez x tartozik.

Vegyünk most egy olyan c_0 számot, melyre $h<c_0<c$. Ekkor, mivel az atomok

természetesen diszjunktak, és az egész tér mértéke 1,

$$\#\{D \in \vee_{j=0}^{n-1} T^{-j}P : \mu(D) > \exp(-c_0 n)\} \leq \exp(c_0 n).^{11}$$

Az ilyen, nagy D-khez tartozó D'-k összmértéke tehát nem lehet nagyobb, mint $\exp((c_0 - c)n)$.

Mivel fix $c_0 < c$ értékek mellett $\exp((c_0 - c)n)$ egy n-ben konvergens sorozat, a Borel-Cantelli lemma szerint 1 a valószínűsége, hogy x csak véges sok ilyen D' halmazhoz tartozik. Ez azt jelenti, hogy nulla mértékű az a halmaz, melyen $\overline{R}(x)$ legalább c, minden $c > h$-ra, tehát $\overline{R}(x) \leq h$.

Ezzel beláttuk a tételt.

□

4.1.2. Megjegyzés. Ha kétirányban végtelen folyamról van szó, akkor ugyanez igaz $\hat{R}_n(x) = \min\{j \geq n : x_{-n}x_{-(n-1)} \dots x_{-1} = x_{j+1}x_{j+2} \dots x_{j+n}\}$-re.

Bizonyítás A bizonyítás nagyon hasonló az előzőhöz, csak annyi a különbség, hogy $[1, L]$ balról jobbra volt diszjunkt blokkokra osztva, ezesetben viszont jobbról balra is kell, $L - N$-nel kezdve. A 3.0.7 tételt most T^{-1}-re kell alkalmazni T helyett.

□

4.1.3. Következmény.

$$\frac{\log n}{N_n(x)} \to H(x_1, x_2, \dots), \quad \textit{1 valószínűséggel,}$$

ahol $N_n(x) = \min\{j \geq 0 : x_0 x_1 \dots x_{j-1} \neq x_{-n+i}x_{-n+i+1} \dots x_{-n+i+j-1}, 0 \leq i \leq n - j\}$ [12]

Bizonyítás A 4.1.2 alkalmazható, csak azt kell észrevenni, hogy $N_n(x) = \min\{j \geq 0 : \forall k \in [j, n]$-re $\quad \forall i \in [0, j-1]$-re $\mathcal{P}(T^i x) \neq \mathcal{P}(T^{-k+i}x)\}$

□

4.1.4. Megjegyzés. Ehhez hasonló következmény egyoldali esetben nincs.

[11] A szigorú egyenlőtlenség is igaz, de nekünk elég ebben a formában.

[12] Azaz a legrövidebb „most kezdődő" olyan blokk mérete, amilyen blokk nem szerepelt a sorozat előző n tagjában részblokként

4.2. Blokkhosszúsági tételek

A 2.2. fejezetben leírt Lempel-Ziv algoritmus által készített blokkfelontás számunkra legfontosabb tulajdonságai, hogy minden blokk pontosan egyszer fordul elő, és minden blokk egy olyan sztringből áll, mely már szerepelt korábban is a jelsorozatban, csak nem úgy, hogy pontosan egy blokkot alkosson.

4.2.1. Jelölések. *I_j-vel jelöljük az {1,2,...n} részintervallumait. Ha ez az $\{i, i+1, \ldots, k\}$ intervallumot jelöli, akkor $\eta(I_j)$ alatt az $\eta_1, \eta_2, \ldots$ sorazat I_j, mint indexhalmaz által meghatározott blokkját, $\eta_i\eta_{i+1}\cdots\eta_k$-t értjük. Ebben az esetben $|I_j| = k - i + 1$.*

4.2.2. Definíció (Elhatároló felbontás). *Ha $\eta_1\eta_2\cdots\eta_n$ egy ergodikus sorozat konkrét megvalósulása, akkor ennek* elhatároló felbontása[13] *alatt az indexhalmaz, vagyis $\{1, 2, \ldots, n\}$ egy olyan $I_1, I_2, \ldots, I_k$ partícióját értjük, melyre az $\eta(I_j)$ blokkok páronként különbözők.*

Láthatjuk, hogy a zip által használt blokkfelbontás elhatároló felbontás.

4.2.3. Tétel (1. blokkhosszúsági tétel). *Legyen $(\mathcal{X}, T, \mu, \mathcal{P})$ egy ergodikus folyamat, melynek entrópiája h, továbbá $x \in X_0 \subset \mathcal{X}$, $\mu(X_0) = 1$. Ekkor adott $\varepsilon > 0$-hoz létezik olyan n_ε, hogy minden $n \geq n_\varepsilon$ -ra a $\{\mathcal{P}(T^i x) : 1 \leq i \leq n\} = \{\mathcal{P}(Tx), \mathcal{P}(T^2x), \ldots, \mathcal{P}(T^n x)\}$ tetszőleges $\{I_1, \ldots, I_m\}$ elhatároló felbontásra*

$$\sum_{\{i:|I_i| \geq \log n/(h+\varepsilon)\}} |I_i| \geq (1-\varepsilon)n,$$

4.2.4. Megjegyzés. Ez tulajdonképpen azt jelenti, hogy ha $\{\mathcal{P}(T^i x) : 1 \leq i \leq n\}$ páronként különböző blokkokra oszlik, akkor 1 valószínűséggel $\{1, 2, \ldots, n\}$ nagy részét olyan blokkokba kerül, melyek hossza legalább $\log n/h$.

4.2.5. Definíció (ε-rövid blokk). *$\{\mathcal{P}(T^i x) : 1 \leq i \leq n\}$ blokkfelbontása esetén azokat a blokkokat nevezzük* ε-rövid blokkoknak, *melynek hossza kisebb, mint $[(1-\varepsilon)\log n]/h$.*

[13] Angolul distinct parsing

4.2.6. Definíció (ε-hosszú blokk). $\{\mathcal{P}(T^i x) : 1 \leq i \leq n\}$ *blokkfelbontása esetén azokat a blokkokat nevezzük* ε-hosszú blokkoknak*, melynek hossza nagyobb, mint* $[(1+\varepsilon)\log n]/h$.

A 4.2.3 tétel szerint tehát a zip tömörírés során keletkező blokkok között csak kevés ε-rövid blokk található. A 4.2.12 tétel fogja majd megmutatni, hogy ε-hosszú blokkokból is csak kevés van.

Legyen $\mathcal{L}$ a $\vee_{i=0}^{L-1} T^{-i}\mathcal{P}$ atomjainak egy halmaza. Ennek minden l eleme egy L-blokk, melyben az i-edik helyen szereplő jel azt mutatja meg, hogy az adott $l \in \mathcal{L}$-beli pontok i-edik képe a $\mathcal{P}$ partíció melyik atomjában van.

4.2.7. Definíció (Jól fedett intervallum). *Ha* $\eta_1\eta_2\ldots\eta_n$*-nek az* $(1-\delta)$*-ad része diszjunkt,* $\mathcal{L}$*-beli* L*-blokkokkal fedett, és* $I_1, I_2, \ldots, I_k$ *egy elhatároló felbontása, akkor azokat az itervallumokat, melyeknek legalább* $(1-\sqrt{\delta})$*-ad része* $\mathcal{L}$*-beli* L*-blokkokkal fedett,* jól fedett intervallumoknak *nevezzük.*

Nekünk az $\eta_1\eta_2\cdots = \mathcal{P}(Tx)\mathcal{P}(T^2x)\ldots$ esetre lesz elsősorban szükségünk, nézzük meg, hogy itt mit jelentenek konkrétan ezek a fogalmak!

A sorozat első n elemére akkor lesz az $I_1, I_2, \ldots, I_k$ elhatároló felbontás[14], ha minden $j \in \{1, 2, \ldots, k\}$-ra az x-nek az I_j által meghatározott útja (a partíció atomjain) nem egyezik meg semelyik I_q, $q \neq j$ által meghatározott úttal sem.

Most az L-hosszú utak közül emeljünk ki néhányat úgy, hogy ezek közül bizonyos számú, egymástól diszjunkt út fedje $\eta_1\eta_2\cdots\eta_n$ $(1-\delta)$-ad részét! Ezek, és esetleg még további néhány L-blokk adják $\mathcal{L}$-t. Más szavakkal, x útjának az első n iteráció által meghatározott szakasza legalább $(1-\delta)$-részben $\mathcal{L}$-beli szakaszokból áll.

Vegyük most $\eta_1\eta_2\cdots\eta_n$ egy elhatároló felbontását, $I_1, I_2, \ldots, I_k$-t! Az $I_j = \{r, r+1, \ldots, s\}$ intervallum jól fedett, ha a hozzá tartozó $\eta_r\eta_{r+1}\cdots\eta_s$ blokk által meghatározott út „elég jól" (azaz legalább $(1-\sqrt{\delta})$-részben) lefedhető $\mathcal{L}$-beli szakaszokkal. Fedés alatt itt azt értjük, hogy az út megfelelő része pontosan egyezik az adott szakasszal. A jólfedettség tehát a blokkfelbontást jellemző tulajdonság, eleve csak akkor állhat fenn, ha a blokkfelbontás *nélkül*, az egész utat elég nagy részben fedik a kiemelt, $\mathcal{L}$-beli utak.

[14] $|I_1| + |I_2| + \cdots + |I_k| = n$

4.2.8. Tétel (4.2.3 általánosítása). *Rögzített $0 < c < 1$-re és $\varepsilon > 0$-ra, elég nagy n-re 1 a valószínűsége, hogy $\{\mathcal{P}(T^i x) : 1 \le i \le n\}$-ből legfeljebb cn-et tudunk különböző ε-rövid blokkokkal fedni.*

Bizonyítás Az entrópia definíciója miatt adott $\delta > 0$-hoz létezik elég nagy L, hogy $\vee_{i=0}^{L-1} T^{-i}\mathcal{P}$ atomjainak egy $\mathcal{L}$ halmazával le lehet fedni $\mathcal{X}$-nek $(1-\delta)$-adrészét úgy, hogy egyrészt $\sum_{A\in\mathcal{L}} \mu(A) \ge 1 - \frac{\delta}{2}$, másrészt $\#\{\mathcal{L}\} \le e^{L(h+\delta)}$.

Birkhoff ergodtételéből [7] következik, hogy legalább $(1-\frac{2}{3}\delta)n$ azon $j \le (n-L)$ indexek száma, melyekre $T^j x \in \cup_{A\in\mathcal{L}} A$[15], tehát x az útja elég nagy részét $\mathcal{L}$-beli atomokon futja (leszámítva az utolsó L lépést). Ebből mutatjuk meg, hogy van egy olyan teljes mértékű X_0 halmaz, melyre $x \in X_0$ és elég nagy n esetén az $(\eta_1\eta_2\cdots\eta_n) = (\mathcal{P}(Tx), \mathcal{P}(T^2x), \ldots, \mathcal{P}(T^n x))$ n-blokk $(1-\delta)$-részét le lehet fedni $\mathcal{L}$-beli diszjunkt L-blokkokkal.

$12\cdots n$-t összesen L-féleképpen lehet partícionálni egymást követő L-blokkokra, a blokkok első elemeit feltüntetve:

$$C_1 = \{1, L+1, 2L+1, \ldots, ([\frac{n}{L}]-1)L+1\}$$
$$C_2 = \{2, L+2, 2L+2, \ldots, ([\frac{n}{L}]-1)L+2\}$$
$$\vdots$$
$$C_L = \{L, L+L, 2L+L, \ldots, ([\frac{n}{L}]-1)L+L\}$$

,

tehát C_i-vel jelöljük azt a felbontást, ahol az első blokk előtt $i-1$ elem van. A lehetséges felbontások közül legalább egy $u \in [1, L]$-re x-nek majdnem az összes blokk (legalább $(1-\delta)\frac{n}{L}$ blokk) első eleme által meghatározott képét fedi $\mathcal{L}$ valamelyik eleme, azaz $\#\{j \in C_u : T^j x \in \cup_{A\in\mathcal{L}} A\} \ge (1-\delta)\frac{n}{L}$, különben az összes $j \in \{1, \ldots, n\}$ indexek közül is kevesebb, mint $(1-\delta)\frac{n}{L}L = (1-\delta)n$ darab esne $\cup_{A\in\mathcal{L}} A$-ba, így az előző, $(1-\frac{2}{3}\delta)n$-es becslés nem lehetne igaz. Egy ilyen u-ra C_u és az elemei, mint kezdőpontok által meghatározott L-blokkok egy megfelelő diszjunkt fedést adnak.

Most szükségünk lesz egy lemmára, ami arra ad felső becslést, hogy adott L-blokkokkal legfeljebb mennyi, rögzített hosszúságú blokkot lehet elég nagy arányban fedni.

[15] Ehhez persze elengedhetetlen, hogy $3L/2\delta \le n$ legyen.

4.2.9. Lemma. *Legyen $\mathcal{P} : X \to \{1, 2, \ldots, a\}$, és vegyük L-blokkoknak egy olyan $\mathcal{L} \subset \{1, 2, \ldots, a\}^L$ együttesét, melyre $\#\{\mathcal{L}\} \leq e^{L(h+\delta)}$. $K > L$ esetén $\kappa(\gamma)$ jelöli azokat a K-blokkokat, melyeket egymást követő L-blokkokra osztva $\mathcal{L}$ elemei legalább $(1-\gamma)$-részben fednek. Ekkor elég nagy k-ra legfeljebb $e^{k(h+\delta+\phi(\gamma))}$ ilyen K-blokk létezik, azaz $\#\{\kappa(\gamma)\} \leq e^{k(h+\delta+\phi(\gamma))}$, ahol $\phi(\gamma) \to 0$, ha $\gamma \to 0$.*[16]

Lemma bizonyítása Rögzítjük, hogy hogyan osztottuk egymást követő L-blokkokra a K-blokkot (összesen L-féleképpen lehet, ha $K \geq 2L$, egyébként $K - L + 1$ lehetőség van), és azt, hogy konkrétan melyik L-blokkok tartozzanak $\mathcal{L}$-hez. $\mathcal{L}$-ben legfeljebb $e^{L(h+\delta)}$ L-blokk van, és a K-blokk felosztása során ezek közül kell kikerülnie $(1-\gamma)\frac{K}{L}$ darabnak. A maradék $\gamma\frac{K}{L}$ darab L-blokk tetszőleges, így azok mindegyike a^L-féleképpen választható. Ez azt jelenti, hogy ekkor legfeljebb

$$(e^{L(h+\delta)})^{(1-\gamma)\frac{K}{L}}(a^L)^{\gamma\frac{K}{L}} = e^{K(h+\delta)(1-\gamma)+K\gamma\log a} = e^{K[h+\delta+\gamma(-h-\delta+\log a)]}$$

azon lehetséges különböző K-blokkok száma, melyeket, ha az adott módon osztunk L-blokkokra, akkor legalább $(1-\gamma)$-részben lesznek $\mathcal{L}$-beli blokkokkal fed ve.

Egy K-blokkot legfeljebb L-féleképpen lehet egymást követő L-blokkokra osztani, valamint $\binom{K/L}{(1-\gamma)K/L}$-féleképpen lehet kiválasztani, hogy az L-blokkok közül melyek tartozzanak mindenképpen $\mathcal{L}$-hez. A Stirling formula szerint pedig

$$L\binom{K/L}{(1-\gamma)\,K/L} \sim L\frac{\sqrt{2\pi\frac{K}{L}}\left(\frac{K/L}{e}\right)^{\frac{K}{L}}}{\sqrt{2\pi\,(1-\gamma)\,\frac{K}{L}}\left(\frac{(1-\gamma)\frac{K}{L}}{e}\right)^{(1-\gamma)\frac{K}{L}}\sqrt{2\pi\gamma\frac{K}{L}}\left(\frac{\gamma\frac{K}{L}}{e}\right)^{\gamma\frac{K}{L}}}$$

$$= e^{\ln L+1+\gamma\left(\frac{K}{L}-1\right)+\ln(1-\gamma)\cdot\left[-\frac{1}{2}+(\gamma-1)\frac{K}{L}\right]+\ln\frac{K}{L}\cdot\left[-\frac{1}{2}-1-\gamma\left(\frac{K}{L}-1\right)\right]+\ln\gamma\cdot\left[-\frac{1}{2}-\gamma\frac{K}{L}\right]-\frac{1}{2}\cdot\ln 2\pi}.$$

Ezt összeszorozva az előző becsléssel, $e^{K[h+\delta+\gamma(-h-\delta+\log a)]}$-val megkapjuk a lemma állítását.

□

Ebből következik, hogy ha m nem elég nagy, akkor a különböző m-blokkok nem tudják nagy részét fedni $\eta_1\eta_2\ldots\eta_n$-nek, így elég csak a nagyobb m-ekkel foglalkozni.

[16] $\phi(\gamma)$-ra létezik explicit becslés is.

Ha $\mathcal{I}$ jelöli a nem jól fedett intervallumok halmazát, akkor $\sum_{I_j \in \mathcal{I}} |I_j| < \sqrt{\delta} \cdot n$, mert ha $\eta_1\eta_2\cdots\eta_n$-nek legalább $\sqrt{\delta}n$-részét olyan intervallumok fednék a felbontásban, melyeknek kevesebb, mint $(1-\sqrt{\delta})$-részét fedné $\mathcal{L}$-beli intervallum, akkor a felbontás nélküli $\eta_1\eta_2\cdots\eta_n$-ből is kevesebb, mint $\sqrt{\delta}n \cdot (1-\sqrt{\delta}) = n(\sqrt{\delta}-\delta)$-t fednének $\mathcal{L}$-beliek, ez pedig kevesebb, mint $n(1-\delta)$, hiszen $\delta \in (0,1)$.

Továbbá, a lemma alapján elég nagy, rögzített m-re $\eta_1\eta_2\cdots\eta_n$-ből legfeljebb $m \cdot e^{m(h+\delta+\phi(\sqrt{\delta}))}$ fedhető m-hosszúságú, jólfedett I_j intervallumokkal. Ha adott ε mellett elég kicsit δ-t választunk, akkor ezt összegezve minden $m \in [M_0, \frac{\log n}{h+\varepsilon}]$-ra $n^{1-\beta}$-nál kisebb eredményt kapunk valamilyen β pozitív számra. Ez a nem jólfedett részekkel együtt is, εn-nél kisebbet ad.

ε-nal 0-hoz tartva, és mindig a megfelelő δ-kkal illetve L-ekkel kapjuk a tétel bizonyítását. Minden lépésben véve a Birkhoff ergodtétel [7] által adott megszámlálható sok halmazt, majd ezeket elmetszve kapjuk a teljes mértékű halmazt.

□

4.2.10. Definíció (Régi blokkokra bontás). *Az $\{1,2,\ldots,a\}$ ábécé elemeibő álló blokkok egy rögzített, véges $\mathcal{F}$ halmaza esetén $\eta_1\eta_2\cdots\eta_n$ egy felbontása* régi blokkokba *történik, ha a felbontás minden I_j intervallumára $\eta(I_j)$*

- *vagy $\mathcal{F}$-hez tartozik*
- *vagy részblokkja $\eta_1\eta_2\cdots\eta_k$-nak, ahol $I_j = \{k+1, k+2, \ldots, k+|I_j|\}$ (azaz már korábban is előfordult).*

$\mathcal{F}$ szerepe mindössze annyi, hogy valahol „elkezdődjön” a felbontás, mivel az eredmények aszimptotikusak, nem függenek $\mathcal{F}$ választásától.

A zip által használt blokk felbontás régi blokkokra bontás, ahol $\mathcal{F}$ a kezdeti szótár elemeiből áll, vagy ha az nincs, üres.

4.2.11. Definíció (Bajkeverő n-blokkok). *Az $\eta_1\eta_2\cdots\eta_n$ blokkot* (ε_0 -)bajkeverőnek *nevezzük, ha régi blokkokba bontása esetén*

$$\sum_{\mathcal{I}:=\{I_i : |I_i| \geq \frac{1+\varepsilon_0}{h}\log n\}} |I_i| \geq \varepsilon_0 n.$$

A most következő tételből kiderül, hogy miért illetjük ezzel a névvel az ilyen blokkokat.

4.2.12. Tétel (2. blokkhosszúsági tétel). *Legyen $(\mathcal{X}, \mu, \mathcal{P}, T)$ egy ergodikus folyamat h entrópiával, $X_0 \subset \mathcal{X}$ pedig egy teljes mértékű halmaz. Ekkor $x \in X_0$ esetén minden $\varepsilon > 0$-ra létezik egy $n_0(\varepsilon_x)$ küszöbszám, hogy minden $n > n_0(\varepsilon x)$-re, ha $I_1, I_2, \ldots, I_k$ egy régi blokkokra bontása $\{\mathcal{P}(T^i x) : 1 \leq i \leq n\}$-nek, akkor*

$$\sum_{\{j:|I_j| \leq \frac{(1+\varepsilon)\log n}{h}\}} |I_j| \geq (1-\varepsilon)n.$$

Ezt is megfogalmazhatjuk a 4.2.4-hoz hasonlóan, illetve ennek is egy általánosítását bizonyítjuk. Ha $\xi_1\xi_2\ldots\xi_n$ olyan blokkokra oszlik, melyek mindegyikének „tartalma" többször szerepel a felbontásban, akkor $n \to \infty$ esetén 1 valószínűséggel $(1, \ldots, n)$ majdnem mindegyike legfeljebb $\log n/h$ blokkba kerül.

Bizonyítás Az entrópia definíciójából adódóan egy h entrópiájú, $\{1, 2, \ldots, a\}$-értékű $\{y_n\}_1^\infty$ stacionárius folyamat esetén, megfelelően választott $\gamma > 0$-hoz, adott, elég nagy L-hez létezik olyan $\mathcal{L} \subset \{1, 2, \ldots a\}^L$, melyre $\#\{\mathcal{L}\} \leq e^{(h+\gamma)L}$ és $\mathcal{L}$ valószínűsége legalább $(1-\gamma)$. Azaz az $\{y_n\}_1^\infty$ sorozat majdnem minden megvalósulására a sorozat egy elég hosszú kezdőszeletének tetszőleges felbontása esetén a nem $\mathcal{L}$-beli blokkok legfeljebb δ-adrészét fedhetik összesen az adott kezdőszeletnek. A 4.2.3 tétel bizonyítása végén leírt gondolatmenet alapján, adott $\delta > 0$-hoz, elég kicsi γ és elég nagy k_0 esetén, minden $k \geq k_0$-ra, ha $\mathcal{L}_k \subset \{1, 2, \ldots a\}^k$ pontosan azokból a k-blokkokból áll, melyknek $(1-2\gamma)$-részét le lehet fedni $\mathcal{L}$-beli, diszjunkt L-blokkokkal, akkor majdnem minden $\{\eta_i\}_1^\infty$ folyamathoz létezik n_0, hogy minden $n \geq n_0$-ra $\eta_1, \eta_2, \ldots, \eta_n$ tetszőleges $I_1, I_2, \ldots I_l$ intervallumokra bontása esetén

$$\sum_{\{I_i:|I_i|=k\geq k_0, \eta(I_i)\notin\mathcal{L}_k\}} |I_i| \leq \delta n,$$

azaz az $\mathcal{L}_k$-n kívüli hosszú blokkok csak keveset fednek. Nevezzük ezt a tulajdonságot $\mathcal{L}_k$-*jó* -nak. A 4.2.9 lemmából következően elérhető, hogy $\#\{\mathcal{L}_k\} \leq e^{(h+\delta)k}$ legyen.

Legyen

$$B := \{1, 2, \ldots n\} \setminus \cup\, \mathcal{I} = \{1, 2, \ldots n\} \setminus \cup \left\{ I_i : |I_i| \geq \frac{1+\varepsilon_0}{h} \log n \right\},$$

és jelöljük ennek maximális részintervallumait $\overline{I_1}, \overline{I_2}, \dots \overline{I_l}$-lel. Az $\mathcal{I}$-t adó maximális részintervallumok száma nem több, mint $\frac{n}{\log n}\frac{h}{1+\varepsilon_0}$ (különben a hosszuk összesen nagyobb lenne, mint n, ami azért lehetetlen, mert ezek diszjunkt részintervallumai $[1,n]$-nek), és ennél legfeljebb 1-gyel lehet nagyobb l, mivel bármely $i \in [1, l-1]$-re $\overline{I_i}$ és $\overline{I_{i+1}}$ között kell lennie egy $\mathcal{I}$-beli intervallumnak. Így k_0-nál rövidebb $\overline{I_i}$ intervallumok csak kis részét fedhetik $\{1, 2, \dots, n\}$-nek. Konkrétan, elég nagy n-re[17]

$$\sum_{\{\overline{I_i}:|\overline{I_i}|\le k_0\}} \overline{I_i} \le k_0 \cdot \frac{h}{1+\varepsilon_0} \cdot \frac{n}{\log n} \le \delta n.$$

Most azt becsüljük meg, hogy hányféleképpen lehet ennek megfelelően blokkokra bontni.

1. Ha $I_i \in \mathcal{I}$, akkor felhasználva, hogy régi blokkokra bontunk, veszünk még egy, n-nél kisebb mutatót, ami megmutatja, hogy kell I_i-t kitölteni. A mutató vagy $\mathcal{F}$ valamelyik elemére mutat, vagy azt mutatja meg, hogy $\{\eta_i\}_1^n$-ben hol szerepeltek már korábban a blokk elemei.

2. Ha $\overline{I_i}$ egy k_0-nál rövidebb blokk, akkor $a^{|\overline{I_i}|}$-féleképpen lehet kitölteni.

3. Ha $\overline{I_i}$ egy k_0-nál hosszabb blokk, akkor $\eta(\overline{I_i}) \in \mathcal{L}_{|\overline{I_i}|}$ esetén természetesen legfeljebb $e^{|\overline{I_i}|(h+\delta)}$-féleképpen tölthetjük ki, egyébként pedig $A^{|\overline{I_i}|}$ különböző lehetőség van.

Az első esetben a mutató olyan, mintha $|I_i|$-blokkok egy legfeljebb $n^{\frac{2}{h}} = e^{\frac{2}{h}\log n} = e^{\frac{2}{1+\varepsilon_0}\frac{1+\varepsilon_0}{h}\log n} = e^{\frac{2}{1+\varepsilon_0}|I_i|}$ elemű halmazából választanánk $\eta(I_i)$-t, mivel $I_i \in \mathcal{I}$ esetén $|I_i| \le \frac{1+\varepsilon_0}{h}\log n$.

$\mathcal{L}_k$-jó $\eta_1\eta_2\cdots\eta_n$ esetén a lehetséges megoldások száma legfeljebb

$$e^{\frac{h}{1+\varepsilon_0}n\varepsilon_0} a^{2\delta n} e^{(h+\delta)(1-\varepsilon_0-2\delta)n} = e^{n\left(h\left(\frac{\varepsilon_0}{1+\varepsilon_0}+1-\varepsilon_0-2\delta\right)+\delta(1-\varepsilon_0-2\delta)+2\delta\log(a)\right)} = e^{(h-\gamma)n},$$

ahol $\gamma > 0$ csak ε_0-tól, δ-tól, h-tól és a-tól függ. Ez azt jelenti, hogy az ε_0-bajkeverő, $\mathcal{L}_k$-jó n-blokkok száma legfeljebb $e^{(h-\gamma)n}$. A Borel-Cantelli lemma miatt egy valószínűséggel $\mu\{x : \mathcal{P}(T^i x) = \eta_i, 1 \le i \le n\} \le e^{\left(h+\frac{\gamma}{2}\right)n}$ mértékű bajkeverő csak véges sok esetben lehet $\eta_1\eta_2\cdots\eta_n$.

[17] Olyan n-re, mely esetén $k_0 \cdot \frac{h}{(1+\varepsilon_0)\delta} \le \log n$.

A 3.0.7 tétel szerint pedig $\eta_1\eta_2\cdots\eta_n$ 1 valószínűséggel csak véges sokszor ε_0-bajkeverő. ε_0-lal 0-hoz tarva kapjuk a tétel bizonyítását.

□

Ez az eredmény akkor is igaz, ha a régi blokkokra bontás definíciójában nem követeljük meg, hogy $\eta(I_i)$ teljes egészében előforduljon már korábban is, hanem elég az is, hogy csak korábban kezdődött. Ez azért lényeges, mert a Lempel-Ziv kódolás több változatában az általunk leírtaktól egy kicsit eltérően történik történik a szótár építése. A beolvasást nem csak akkor folytatjuk, ha az eddig beolvasott sztring már szerepelt a tömörítendő állományban, mint önálló blokk, hanem akkor is, ha egyáltalán szerepelt. Az új bejegyzést a szótárban pedig csak akkor készítjük el, ha már tudjuk, hogy fel is használjuk.

4.2.13. Definíció (Régies blokkokra bontás). *Az $\{1,2,\ldots,a\}$ ábécé elemeiből álló blokkok egy rögzített, véges $\mathcal{F}$ halmaza esetén $\eta_1\eta_2\cdots\eta_n$ egy felbontása* régies blokkokba *történik, ha a felbontás minden I_j intervallumára $\eta(I_j)$*

- *vagy $\mathcal{F}$-hez tartozik*
- *vagy részblokkja $\eta_1\eta_2\cdots\eta_k$-nak, ahol $I_j = \{1+i, 2+i, \ldots, |I_j|+i\}$, $(i \geq 1)$ (azaz ugyanez a szakasz szerepelt már korábbi kezdéssel is).*

4.2.14. Tétel. *Ha $\{y_n\}_1^\infty$ egy h-entrópiájú, véges értékkészletű stacionárius ergodikus folyamat, akkor ennek egy $\{\eta\}_1^\infty$ megvalósulására [18], adott $\varepsilon > 0$-ra, minden $n \geq n_\varepsilon$-ra, $\eta_1\eta_2\ldots\eta_n$ különböző I_i blokkokba történő bármely régies felbontására 1 valószínűséggel*

$$\sum_{\left\{I_i:\frac{1}{h+\varepsilon}\log n\leq|I_i|\leq\frac{1+\varepsilon}{h}\log n\right\}} |I_i| \geq (1-\varepsilon)n.$$

Bizonyítás Ebben az esetben az új blokkoknak csak a legutolsó bitje új, de ugyanaz a bizonyítás működik itt is, mint a régi blokkoknál.

□

4.2.15. Tétel (4.2.12 általánosítása). *Rögzített $0 < c < 1$-re és $\varepsilon > 0$-ra, elég nagy n-re 1 a valószínűsége, hogy $\xi_1\xi_2\ldots\xi_n$-ből legfeljebb cn-et tudunk olyan ε-hosszú blokkokkal fedni, melyek előfordulnak $\xi_1\xi_2\ldots\xi_n$-ben több, egymástól diszjunkt helyen.*

[18] Angolul: output.

5. Entrópia és Hausdorff-dimenzió

Legyen $\Sigma := \{a_1, a_2, \ldots, a_m\}$ és $(\Sigma^{\mathbb{N}}, \sigma)$ az egyoldali eltolások tere m szimbólum felett, μ pedig egy σ-invariáns ergodikus valószínűségi mérték $\Sigma^{\mathbb{N}}$-n. d a metrika, amivel a Hausdorff-dimenziót értelmezzük, $x, y \in \Sigma^{\mathbb{N}}$ esetén $d(x,y) = m^{-\inf\{k \geq 0 : x_{k+1} \neq y_{k+1}\}}$. R_n-nel továbbra is a visszatérési időt jelöljük.

A 4.1.1 tételben megállapítottuk, hogy majdnem minden $x \in \Sigma^{\mathbb{N}}$-re a

$$\lim_{n\to\infty} \frac{\log R_n(x)}{n}$$

határérték létezik és véges. Az 5.1.1 tétel azt mondja meg, hogy mit tudunk azokról az x-ekről, melyekre nem létezik a limesz.

5.1. Visszatérő halmazok Hausdorff-dimenziója

Feng és Wu [4] $\Sigma^{\mathbb{N}}$ egy nem megszámlálható partíciójáról,

$$E_{\alpha,\beta} := \left\{ x \in \Sigma : \liminf_{n\to\infty} \frac{\log R_n(x)}{n} = \alpha, \limsup_{n\to\infty} \frac{\log R_n(x)}{n} = \beta \right\}\text{-ról}$$

mutatja meg, hogy tetszőleges $0 \leq \alpha \leq \beta \leq \infty$-ra 1 a Hausdorff-dimenziója.

5.1.1. Tétel. *Minden $\alpha, \beta \in [0, \infty]$, $\alpha \leq \beta$-ra* $\dim_H E_{\alpha,\beta} = 1$.

A bizonyításban $E_{\alpha,\beta}$ olyan részhalmazait vesszük, melyek Cantor-halmazok, és ezeknek a Hausdorff-dimenziójáról látjuk be, hogy 1-hez tartanak. Ehhez először vegyünk egy lemmát.

5.1.2. Lemma. *Legyen $\{\ell_n\}$ pozitív egészeknek egy olyan sorozata, melyre létezik n_0, hogy minden $n \geq n_0$-ra $\ell_{n+1} \geq \ell_n + 2n$ és $\lim_{n\to\infty} \frac{\ell_n}{n^2} = \infty$. Ekkor az*

$$A = A^{\{\ell_n\}} := \left\{ x \in \Sigma^{\mathbb{N}} : \exists k_0, \text{ melyre } k \geq k_0 \text{ esetén } R_k(x) = \ell_k \right\}$$

halmaznak 1 a Hausdoff-dimenziója.

Bizonyítás Elegendő megmutatni, hogy $\dim_H A \geq 1-\delta$ minden $\delta > 0$-ra, hiszen $\dim_H \Sigma^{\mathbb{N}} = 1$ és $A \subset \Sigma^{\mathbb{N}}$. Rögzített $\delta > 0$ mellett legyen $p \geq \max\{n_0, 6\}$ olyan, hogy $\frac{p-2}{p} > 1-\delta$, és legyen

$$F_p = \left\{x = \{x_i\}_1^\infty \in \Sigma^{\mathbb{N}} : x_1 = x_2 = \cdots = x_p = m \text{ és } k \in \mathbb{Z}^+ \text{ esetén } x_{pk+1} = x_{pk+p} = 1\right\},$$

azaz az első p helyen ugyanaz a szimbólum, m szerepel, majd minden p-edik, és az azt követő helyen egyes van:

$$x = \underbrace{mm\cdots m}_{p \text{ darab}} 1 \underbrace{\cdots}_{p-2 \text{ darab}} 11 \underbrace{\cdots}_{p-2 \text{ darab}} 11 \underbrace{\cdots}_{p-2 \text{ darab}}$$

Egy ilyen sorozatnak az első p elem utáni végszelete önhasonló halmaznak tekinthető m^{p-2} darab m^{-p}-arányú önhasonlósággal, így a Hausdorff-dimenziója, és ezzel $\dim_H F_p$ is $\frac{\log m^{p-2}}{\log m^p} = \frac{p-2}{p} > 1-\delta$.

Létezik olyan $g : F_p \to A$ leképezés, hogy minden $\varepsilon > 0$-ra létezik $M \in \mathbb{N}$, melyre minden $k \geq M$ esetén $d\left(g\left(x\right), g\left(y\right)\right) < m^{-k}$-ból következik, hogy $d(x,y) < m^{-(1-\varepsilon)k}$, azaz g „közel Lipschitz”[19]. Ez elég is a lemma bizonyításához, hiszen ez azt jelenti, hogy $\dim_H g(F_p) \geq \dim_H F_p$, azaz $\dim_H A \geq \dim_H g(F_p) \geq \dim_H F_p \geq 1-\delta$.

Már csak meg kell mutatni, hogy tényleg létezik ilyen g leképezés. Erre adunk egy konstrukciót.

Először definiálunk egy $x \in F_p$-től függő $\{x^{(i)}\}_{i=p}^\infty$ sorozatot Σ-n. Minden $x = \{x_i\}_{i=1}^\infty \in F_p$-re $x^{(p-1)} = x = x_1 x_2 \cdots x_n \cdots$

$j < p-1$ esetén $x^{(j)}$ definíciójára, tegyük fel, hogy már definiáltuk valahogy. Minden $j \geq 1$-re vezessük be az $x^{(j)} = x_1^{(j)} x_2^{(j)} \cdots x_n^{(j)} \cdots = x_1^{(j)} \circ x_2^{(j)} \circ \cdots \circ x_n^{(j)} \circ \cdots$ jelöléseket. Legyen

$$x^{(k)} := \begin{cases} x = \underbrace{mm\cdots m}_{\text{p darab}} 1 \cdots 11 \underbrace{\cdots}_{p-2 \text{ darab}} 11 \underbrace{\cdots}_{p-2 \text{ darab}} & \text{ha } k < n_0 \\ x_1^{(k-1)} x_2^{(k-1)} \cdots x_{\ell_k - 1}^{(k-1)} \underbrace{1 x_1^{(k-1)} \cdots x_k^{(k-1)} y_k 1}_{w_k} x_{\ell_k}^{(k-1)} x_{\ell_k+1}^{(k-1)} \cdots & \text{ha } k \geq n_0, \end{cases}$$

ahol $y_k \neq x_{k+1}^{(k-1)}$. $k \geq n_0$ miatt teljesülnek ℓ_k-ra a lemma feltételei.

Látszik, hogy $x^{(k-1)}$-ről $x^{(k)}$-ra térve nem változik az első $\ell_k - 1$ szimbólum. Tekintve, hogy $\ell_k - 1 \geq \ell_{k-1} + 2(k-1) - 1$, ez azt is jelenti egyben, hogy az összesen

[19] Nearly Lipschitz

$\ell_k - 1 + k + 3$ szimbólumból álló $x_1^{(k-1)} x_2^{(k-2)} \cdots x_{\ell_{k-1}} \circ w_k$ előtagja $x^{(k+1)}$-nek is és minden $i \geq (k+1)$-re x^i-nek is, így az $\{x^{(i)}\}_{i=p}^{\infty}$ sorozat konvergens, és a határértékének is ez a kezdőszelete. Jelöljük x^*-gal a limeszt.

Ha $k \geq p$, akkor $R_k(x^*) \leq \ell_k$, hiszen az ℓ_k-adik elemet (ami egy 1-es) követő k jel ugyanaz, mint az első k jel. Azt is tudjuk, hogy a $k+1$-edik jel viszont már különböző, direkt így választottuk y_k-t, így $d(\sigma^{\ell_k}(x^*), x^*) = m^{-k}$.

Minden $k \geq p$-re a $\theta = \overbrace{mm\cdots m}^{\text{p darab}}$ blokk csak x^k elején, illetve w_k-ban fordulhat elő, mert F_p definíciójából következően bármely p egymást követő jel közül legalább kettő 1-es.

Az $x \mapsto x^*$ leképezés jó lesz $g : F_p \to A$-nak. g injektivitása a definíciójából, g^{-1} közel Lipschitz tulajdonsága pedig az alábbiakból következik.

Létezik olyan $N > p$ egész szám, hogy minden $n \geq N$-re $\frac{n^2}{\ell_n} < \frac{\varepsilon}{2}$, mert $\frac{\ell_n}{n^2}$ végetelenhez tart. Ha $k \geq \ell_N$, és $q \in \mathbb{Z}^+$ az a szám, melyre $\ell_q \leq k \leq \ell_{q+1}$, akkor $p < N \leq q$. Ha $k \geq \ell_n$-re $d(x^*, y^*) < m^{-k}$, azaz $x_1^* x_2^* \cdots x_k^* = y_1^* y_2^* \cdots y_k^*$, akkor persze $x_1 x_2 \cdots x'_k = y_1 y_2 \cdots y'_k$, $k' = k - \sum_{j=p}^{q+1}(j+3) > k - 2q^2 \geq k - \varepsilon \ell_q \geq k(1-\varepsilon)$, ez pedig azt jelenti, hogy $d(x,y) \leq m^{-k} \leq m^{-(1-\varepsilon)k}$, tehát a lemmát bebizonyítottuk. □

A tétel bizonyítása Az előző lemmának köszönhetően már csak azt kell bizonyítani, hogy tetszőleges, adott $0 \leq \alpha \leq \beta \leq \infty$-ra létezik olyan $\{\ell_n\} \subset (\mathbb{Z}^+)^{\mathbb{N}}$ sorozat, mely megfelel a lemma feltételeinek, azaz létezik n_0, hogy minden $n \geq n_0$-ra $\ell_{n+1} \geq \ell_n + 2n$ és $\lim_{n\to\infty} \frac{\ell_n}{n^2} = \infty$, továbbá $\liminf_{n\to\infty} \frac{\log \ell_n}{n} = \alpha$ és $\limsup_{n\to\infty} \frac{\log \ell_n}{n} = \beta$. Ezt esetszétválasztással látjuk be, aszerint, hogy α és β a $[0, \infty]$ intervallumnak „végpontjai"-e, illetve egyenlők-e. 7 eset van. $[x]$ az x egészrészét jelöli.

1. $\alpha = \beta = \infty$. Ekkor $\ell_n := [e^{n^2}]$ jó.

2. $\alpha = \beta = 0$. Ekkor $\ell_n := [e^{\sqrt{n}}]$ jó.

3. $0 < \alpha = \beta < \infty$. Ekkor $\ell_n := [e^{n\alpha}]$ jó.

4. $0 < \alpha < \beta < \infty$. Ekkor $\ell_n := \sum_{i=1}^{n} u_i$, ahol

$$u_n := \begin{cases} [e^{n\alpha}] & \text{ha valamely } i \in \mathbb{N}^+\text{-ra } \sum_{j=1}^{2i-1} 2^{4^j} \leq n < \sum_{j=1}^{2i} 2^{4^j} \\ [e^{n\beta}] & \text{egyébként.} \end{cases}$$

5. $0 < \alpha < \beta = \infty$. Ekkor $\ell_n := \sum_{i=1}^{n} u_i$, ahol

$$u_n := \begin{cases} \left[e^{n\alpha}\right] & \text{ha valamely } i \in \mathbb{N}^+\text{-ra } \sum_{j=1}^{2i-1} 2^{4^j} \leq n < \sum_{j=1}^{2i} 2^{4^j} \\ \left[e^{n^2}\right] & \text{egyébként.} \end{cases}$$

6. $0 = \alpha < \beta < \infty$. Ekkor $\ell_n := \sum_{i=1}^{n} u_i$, ahol

$$u_n := \begin{cases} \left[e^{\sqrt{n}}\right] & \text{ha valamely } i \in \mathbb{N}^+\text{-ra } \sum_{j=1}^{2i-1} 2^{4^j} \leq n < \sum_{j=1}^{2i} 2^{4^j} \\ \left[e^{n\beta}\right] & \text{egyébként.} \end{cases}$$

7. $0 = \alpha < \beta = \infty$. Ekkor $\ell_n := \sum_{i=1}^{\infty}$, ahol

$$u_n := \begin{cases} \left[e^{\sqrt{n}}\right] & \text{ha } \sum_{j=1}^{2i-1} 2^{4^j} \leq n < \sum_{j=1}^{2i} 2^{4^j} \text{ valamely } i \in \mathbb{Z}^+\text{-ra} \\ \left[e^{n^2}\right] & \text{egyébként.} \end{cases}$$

□

5.1.3. Definíció. *$A = (a_{ij}) \in \{0,1\}^{m\times m}$ mátrix esetén $\Sigma_A := \left\{(x_n) \in \Sigma : a_{x_n,x_{n+1}} = 1, \quad \forall n \geq 1\right\}$, azaz egy olyan Σ-ban futó sorozat, melyben bármely két, egymást követő elemet, ha indexnek tekintünk, az általuk meghatározott eleme A-nak 1-es. Erre nyilván $\sigma\Sigma_A \subset \Sigma_A$. A (Σ_A, σ) rendszert nevezzük* véges típusú részeltolásnak.[20].

A részeltolás (topologikusan) tranzitív, *amennyiben létezik olyan $M \geq 1$, melyre A^M minden eleme szigorúan pozitív (azaz 1).*

5.1.4. Következmény. *Az 5.1.1 tétel általánosítható tranzitív véges típusú részeltolásokra:* $\dim_H E_{\alpha,\beta} = \dim_H \Sigma_A = \frac{\log p}{\log m}$, *ahol p az A spektrálsugara.*

[20]subshift of finite type

6. A Shannon-McMillan-Breiman tétel

Algoet és Cover [3] cikke alapján a Shannon-McMillan-Breiman tétel véges értékkészletű változatát bizonyítjuk ebben a fejezetben.

Egy $\mathcal{X}$ véges halmazban futó $\{X_t\}$ stacionárius ergodikus folyamat entrópia rátája a következő:

$$H = \lim_{k\to\infty} H^k = \lim_{k\to\infty} E\{-\log p(X_k|X_{k-1},\dots,X_0)\}.$$

Mivel stacionárius folyamatról beszélünk, ez egyenlő $E\{-\log p(X_0|X_{-1},\dots,X_{-k})\}$-val.

A tétel bizonyításához felhasználunk egy lemmát, de először a Markov-egyenlőtlenséget és a Borel–Cantelli lemmát mondjuk ki, mert ezekre szükség lesz a lemmában és a tételben is.

6.0.5. Tétel (Markov-egyenlőtlenség). *x valószínűségi változó és tetszőleges $\varepsilon > 0$ szám esetén $P(|x| \geq \varepsilon) \leq E(|x|)/\varepsilon$.*

6.0.6. Lemma (Borel–Cantelli). *Adott valószínűségi mezőn tetszőleges $A_1, A_2, A_3, \dots$ halmazokra:*

1. *Ha $\sum_{k=1}^{\infty} P(A_k) < \infty$, akkor 1 valószínűséggel csak véges sok A_k esemény következik be, azaz*
$$P\left(\limsup_{k\to\infty} A_k\right) = 0.$$
2. *Ha az $A_k, k = 1, 2, \dots$ események függetlenek, és $\sum_{k=1}^{\infty} P(A_k) = \infty$, akkor 1 valószínűséggel végtelen sok A_k esemény következik be, azaz*
$$P\left(\limsup_{k\to\infty} A_k\right) = 1.$$

6.0.7. Lemma. *Ha pozitív valószínűségi változók egy (g_n) sorozatára $E[(g_n)] \leq 1$ minden $n \geq 0$-ra, akkor $\limsup_{n\to\infty} n^{-1} \log g_n \leq 0$ 1 valószínűséggel.*

Bizonyítás A Markov-egyenlőtlenség miatt tetszőlegesen kicsi $\varepsilon > 0$-ra

$P(n^{-1}\log g_n \geq \varepsilon) = P(g_n \geq \exp(n\varepsilon)) \leq E[(g_n)] / \exp(n\varepsilon) \leq 1/\exp(n\varepsilon) = \exp(-n\varepsilon)$. Mivel $\sum_{n\in\mathbb{N}} \exp(-n\varepsilon)$ véges, így a 6.0.6 lemma szerint $\varepsilon \geq \limsup_{n\to\infty} \exp(-n\varepsilon) \geq \limsup_{n\to\infty} n^{-1}\log g_n$, tehát valóban 1 valószínűséggel $\limsup_{n\to\infty} n^{-1}\log g_n \leq 0$.

□

6.0.8. Tétel (Véges Shannon-McMillan-Breiman). *Ha az $\{X_t\}$ véges értékkészletű ergodikus sorozat entrópia rátája H, akkor*

$$-\frac{1}{n}\log p(X_0,\dots,X_{n-1}) = -\frac{1}{n}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\dots,X_1,X_0) \to H \text{ 1 valószínűséggel.}$$

Bizonyítás A bizonyítás lényege, hogy az $n^{-1}\log p(X_0,\dots,X_{n-1})$ sorozatnak fölső korlátja H^k és alsó korlátja H, továbbá $H^k \geq H$ és $H^k \to H$, ha k tart végtelenhez. A csendőrelv szerint így $n^{-1}\log p(X_0,\dots,X_{n-1})$ is tart H-hoz.

Elég nagy n-re, $n \geq k$ esetén

$$\begin{aligned}-\frac{1}{n}\log p^k(X_0,\dots,X_{n-1}) &:= -\frac{1}{n}\log\left(p\left(X_0,\dots,X_{k-1}\right)\left[\prod_{t=k}^{n-1}p\left(X_t|X_{t-1},\dots,X_{t-k}\right)\right]\right)\\ &= -\frac{1}{n}\log p\left(X_0,\dots,X_{k-1}\right) - \frac{1}{n}\sum_{t=k}^{n-1}\log p\left(X_t|X_{t-1},\dots,X_{t-k}\right) \to\\ &E\left[-\log p\left(X_k|X_{k-1},\dots,X_0\right)\right] = H^k \text{ 1 valószínűséggel.}\end{aligned}$$

Az első egyenlőségnél k-adrendű Markov közelítést alkalmaztunk, a másodiknál csak egy logaritmus azonosságot, a határérték pedig Birkhoff ergodtételéből következik.

Ugyanígy,

$$\begin{aligned}-\frac{1}{n}\log p(X_0,X_1,\dots,X_{n-1}|X_{-1},X_{-2},\dots) &=\\ = -\frac{1}{n}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\dots,X_1,X_0,X_{-1},X_{-2},\dots) &\to\\ E\left[-\log p\left(X_0|X_{-1},X_{-2},\dots\right)\right] = H \text{ 1 valószínűséggel.}&\end{aligned}$$

p^k szintén valószínűségi mérték, tehát az egész tér mértéke 1, ezért

$$E\left[\frac{p^k\left(X_0,\dots,X_{n-1}\right)}{p\left(X_0,\dots,X_{n-1}\right)}\right] \leq 1,$$

mivel az eredeti eloszláshoz viszonyítva egy másik mérték szerinti valószínűségeket a hányados várható értéke egyenlő az utóbbi abszolút folytonos része szerinti mértékkel.

Ugyanezért

$$E\left[\frac{p\left(X_0,\ldots,X_{n-1}\right)}{p\left(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots\right)}\right]\leq 1.$$

Így a 6.0.7 lemma alkalmazható $g_n=\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}$-ra és $h_n=\frac{p(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots)}$-ra is, tehát

$$\limsup_{n\to\infty} n^{-1}\log\left(\frac{p^k\left(X_0,\ldots,X_{n-1}\right)}{p\left(X_0,\ldots,X_{n-1}\right)}\right)\leq 0 \text{ 1 valószínűséggel,}$$

és ugyanígy

$$\limsup_{n\to\infty} n^{-1}\log\left(\frac{p\left(X_0,\ldots,X_{n-1}\right)}{p\left(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots\right)}\right)\leq 0 \text{ 1 valószínűséggel.}$$

A logaritmus azonosságokat és a fenti határértékeket felhasználva kapjuk, hogy

$$\begin{aligned}H^k&=E\left[-\log p\left(X_0|X_{-1},\ldots,X_{-k}\right)\right]\geq\limsup_{n\to\infty}-\frac{1}{n}\log p\left(X_0,\ldots,X_{n-1}\right)\\&\geq\liminf_{n\to\infty}-\frac{1}{n}\log p\left(X_0,\ldots,X_{n-1}\right)\\&\geq E\left[-\log p\left(X_0|X_{-1},X_{-2},\ldots\right)\right]=H \text{ 1 valószínűséggel.}\end{aligned}$$

minden $x_0\in\mathcal{X}$-re

$$p\left(x_0|X_{-k},\ldots,X_{-1}\right)\to p\left(x_0|X_{-1},X_{-2},\ldots\right) \text{ 1 valószínűséggel,}$$

ezért (mivel $x\log x$ abszolút korlátos és folytonos a $[0,1]$ intervallumon és $\mathcal{X}$ véges) ha k tart végtelenhez, akkor $-\sum_{x_0\in\mathcal{X}}p\left(x_0|X_{-1},\ldots,X_{-k}\right)\log p\left(x_0|X_{-1},\ldots,X_{-k}\right)$ is 1 valószínűséggel tart $-\sum_{x_0\in\mathcal{X}}p\left(x_0|X_{-1},X_{-2},\ldots\right)\log p\left(x_0|X_{-1},X_{-2},\ldots\right)$-hez, és a korlátos konvergencia tétel értelmében a várható értékeikre is igaz, hogy $E\left[-\sum_{x_0\in\mathcal{X}}p\left(x_0|X_{-1},\ldots,X\right.\right.$ $\left.\left.\log p\left(x_0|X_{-1},\ldots,X_{-k}\right)\right]\to-\sum_{x_0\in\mathcal{X}}p\left(x_0|X_{-1},X_{-2},\ldots\right)\log p\left(x_0|X_{-1},X_{-2},\ldots\right)$ 1 valószínűséggel. Ez viszont pontosan azt jelenti, hogy $H^k\to H$, így a tételt bebizonyítottuk.

□

Hivatkozások

[1] Donald S. Ornstein, Benjamin Weiss (1993): Entropy and Data Compression Schemes. IEEE vol. **39**, no. 1,78-83.

[2] Claude E. Shannon, (1948): A mathematical theory of communication. Bell System Technical Journal vol. **27**, 379-423, 623-656.

[3] Paul H. Algoet; Thomas M. Cover (1988): A Sandwich Proof of the Shannon-McMillan-Breiman Theorem. The Annals of Probability, Vol. **16**, No. 2. 899-909

[4] De-Jun Feng, Jun Wu (2001): The Hausdorff dimension of recurrent sets in symbolic spaces. Nonlinearity, vol. **14**, 81-85.

[5] David A. Huffman (1952): A Method for the Construction of Minimum-Redundancy Codes. Proc. Inst. Radio Eng. Vol. **40**, 1098-1101.

[6] Terry A. Welch: A Technic for High-Performance Data Compression Terry A. Welch (1984): A Technique for High Performance Data Compression. IEEE Computer, Vol. **17**, No. 6, 8-19.

[7] Manfred Einsiedler, Thomas Ward (2011): Ergodic theory: with a view towards Number Theory. Springer, 43-47

[8] L. Peter Deutsch (1996): DEFLATE Compressed Data Format Specification (https://tools.ietf.org/html/rfc1951)

[9] David Salomon, Mason D. Bryant, Giovanni Motta (2007): Data Compression: The Complete Reference. Springer.

[10] PKWARE Inc. File: APPNOTE.TXT - .ZIP File Format Specification. Version: 6.3.3 (http://www.pkware.com/documents/casestudies/APPNOTE.TXT)