SZÉCHENYI ISTVÁN
EGYETEM
GYŐR

KÓDOLÁSELMÉLET

Nagy Szilvia

6. Forráskódolás alapjai

2009.

Információ

Az **információ** valamely véges számú, előre ismert esemény közül annak megnevezése, hogy melyik következett be. Alternatív megfogalmazás: az információ mértéke azonos azzal a bizonytalansággal, amelyet megszűntet.

Shannon: minél váratlanabb egy esemény, bekövetkezése annál több információt jelent. Legyen $A=\{A_1, A_2, \ldots A_m\}$ eseményhalmaz, az A_1 esemény valószínűsége p_1, … az A_m-é p_m . Ekkor az A_i megnevezésekor nyert információ:

$$I(A_i) = \log_2 \frac{1}{p_i} = -\log_2 p_i .$$

Információ

1. Csak az esemény valószínűségének függvénye.
2. Nem negatív: $I \geq 0$
3. Additív: ha $m = m_1 \cdot m_2$, $I(m_1 \cdot m_2) = I(m_1) + I(m_2)$
4. Monoton: ha $p_i \geq p_j$, akkor $I(A_i) \leq I(A_j)$
5. Normálás: legyen $I(A)=1$, ha $p(A)=0{,}5$. Ekkor kettes alapú logaritmus használandó és az információegysége a **bit**.

   Megjegyzés: ha tízes alapú logaritmust (lg-t) használunk, a hartley, az egység. Ekkor a normálás: $I(p=0{,}1)=1$. Ha természetes alapú logaritmussal definiáljuk az információt ($I=-\ln p$), akkor a natban mérjük az információt, a normálás pedig $I(p=1/e)=1$.

Az entrópia

Az **entrópia** az információ várható értéke:

$$H(p_1, p_2, \ldots, p_m) = \langle I(A) \rangle = \sum_{i=1}^{m} p_i I(A_i) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Az entrópia tulajdonképpen annak a kijelentésnek az információtartalma, hogy az m db egymást kizáró esemény közül az egyik bekövetkezett.

A $p \log_2 p$ kifejezés $p \to 0$ esetén:

$$\lim_{p\to 0} p\log_2 p = \lim_{p\to 0} p\frac{\ln p}{\ln 2} = \frac{1}{\ln 2} \cdot \lim_{p\to 0} \left(\frac{\ln p}{\frac{1}{p}} \right) =$$

L'Hospital-szabály szerint

$$= \frac{1}{\ln 2} \cdot \lim_{p\to 0} \frac{\frac{1}{p}}{\frac{-1}{p^2}} = 0.$$

Az entrópia

1. Nem negatív: $H(p_1, p_2, \ldots, p_m) \geq 0$
2. Az események valószínűségeinek folytonos függvénye.
3. $H(p_1, p_2, \ldots, p_m, 0) = H(p_1, p_2, \ldots, p_m)$
4. Ha $p_i = 1$, a többi $p_k = 0$, (k=1, …, i–1, i+1 ,…, m), akkor $H(p_1, p_2, \ldots, p_m) = 0$.
5. $H(p_1, p_2, \ldots, p_m) \leq H(1/m, 1/m, \ldots 1/m)$
6. $H(p_1, \ldots, p_{k-1}, p_\ell, p_{k+1}, \ldots, p_{\ell-1}, p_k, p_{\ell+1}, \ldots, p_m) = H(p_1, p_2, \ldots, p_m)$, $\forall k, \ell$; azaz az entrópia szimmetrikus változóinak cseréjére.

A kölcsönös entrópia

Legyen $A=\{A_1, \ldots, A_{m_1}\}$ a lehetséges leadott jelek halmaza, $B=\{B_1, \ldots, B_{m_2}\}$ pedig a vehető jelek halmaza. Vizsgáljuk azt az összetett eseményt, hogy egy A-beli és egy B-beli esemény is bekövetkezik.

A_i és B_j együttes bekövetkezési valószínűsége

$$p_{i,j} = p(A_i \cdot B_j),$$

a két esemény **együttes bekövetkezésekor nyert információ**

$$I(A_i \cdot B_j) = -\log_2 p(A_i \cdot B_j) = -\log_2 p_{i,j}.$$

Mindig igaz a kölcsönös információra, hogy

$I(A_i \cdot B_j) \geq I(A_i)$, és $I(A_i \cdot B_j) \geq I(B_j)$.

A kölcsönös entrópia

Legyen $A=\{A_1, \ldots, A_{m_1}\}$ a lehetséges leadott jelek halmaza, $B=\{B_1, \ldots, B_{m_2}\}$ pedig a vehető jelek halmaza. Vizsgáljuk azt az összetett eseményt, hogy egy A-beli és egy B-beli esemény is bekövetkezik.

A_i és B_j együttes bekövetkezési valószínűsége

$$p_{i,j} = p(A_i \cdot B_j),$$

a két esemény **együttes bekövetkezésekor nyert információ**

$$I(A_i \cdot B_j) = -\log_2 p(A_i \cdot B_j) = -\log_2 p_{i,j}.$$

A és B halmazok **kölcsönös entrópiája**:

$$H(A \cdot B) = -\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} p_{i,j} \log_2 p_{i,j}.$$

A feltételes entrópia

Legyen $A=\{A_1, \ldots, A_{m_1}\}$ a lehetséges leadott jelek halmaza, $B=\{B_1, \ldots, B_{m_2}\}$ pedig a vehető jelek halmaza. Minden A-beli esemény bekövetkezése maga után von egy B-beli eseményt.

A_i-nek B_j-re vonatkoztatott feltételes valószínűsége $p(A_i \mid B_j)$.

Az A halmaz B halmazra vonatkoztatott **feltételes entrópiája**:

$$H(A|B) = -\sum_{j=1}^{m_2} p(B_j) \sum_{i=1}^{m_1} p(A_i|B_j) \log_2 p(A_i|B_j) =$$

$$= -\sum_{j=1}^{m_2} \sum_{i=1}^{m_1} p(A_i \cdot B_j) \log_2 p(A_i|B_j).$$

A feltételes entrópia

Legyen $A=\{A_1, \ldots, A_{m_1}\}$ a lehetséges leadott jelek halmaza, $B=\{B_1, \ldots, B_{m_2}\}$ pedig a vehető jelek halmaza. Minden A-beli esemény bekövetkezése maga után von egy B-beli eseményt.

A_i-nek B_j-re vonatkoztatott feltételes valószínűsége $p(A_i \mid B_j)$.

Mivel $p(A_i \cdot B_j)=p(B_j) \cdot p(A_i \mid B_j)$ minden i-re és j-re,

$$H(A \cdot B)= H(B) \cdot H(A \mid B)= H(A) \cdot H(B \mid A).$$

Így

$$H(A) \geq H(A \cdot B) \geq 0$$

A források jellemzése

Olyan forrásokkal fogunk foglalkozni, amelyek kimenetén véges sok elemből álló $A = \{A_1, \ldots, A_n\}$ halmaz elemei jelenhetnek meg. Az A halmazt ekkor **forrásábécé**nek nevezzük.

Az A elemeiből képezett véges

$A^{(1)}\ A^{(2)}\ A^{(3)}\ \ldots\ A^{(m)}$

sorozatok az **üzenet**ek. (m tetszőleges természetes szám)

A lehetséges üzenetek halmaza $\boldsymbol{A}$.

A források jellemzése

Ha egy *A* forrás által kibocsátott üzenet diszkrét jelek sorozata, akkor *A* **diszkrét információforrás**.

A forrás **emlékezet nélküli**, ha $A^{(i)}$ független $A^{(i-k)}$-től, $\forall\, i, k$.

A forrás **stacionárius**, ha $A^{(i)} \in A \quad \forall\, i$, és $p(A^{(i)} = A_j) = p_j, \quad \forall\, i, j$.

Előfordulhat, hogy a forrás által kibocsátott szimbólum függ az azt megelőző kibocsátásoktól

A források jellemzése

A rendszer lehetséges állapotainak a halmaza: $S=\{S_1, S_2, \ldots, S_n\}$. Tegyük fel, hogy a forrás egy $S_{előző}$ állapotban van, és az aktuális szimbólumkibocsátás után egy $S_{új}$ állapotba kerül.
Ha

$$p(S_{új} \mid S_{előző}, S_{előző-1}, \ldots, S_{előző-m}) = p(S_{új} \mid S_{előző}),$$

akkor a rendszer egy **Markov-folyamat**tal leírható.

A források általában jól modellezhetők Markov-folyamatokkal.

A források jellemzése

A források leírhatók Markov-folyamatokkal.

A legegyszerűbb Markov-folyamat során a forrás minden betűje azonos valószínűséggel fordul elő, és a szimbólumkibocsátások független események.

Legyen öt betűnk, A, B, C, D és E, mind 0,2 előfordulási valószínűséggel.

Egy tipikus példa így előállt szövegre (Shannon művéből):

B D C B C E C C C A D C B D D A A E C E E A A B B D A
E E C A C E E B A E E C B C E A D.

A források jellemzése

A források leírhatók Markov-folyamatokkal.

Egy összetettebb Markov-folyamat során a karakterek előfordulási valószínűsége más és más, a szimbólum-kibocsátások független események.

Legyen p_A=0,4; p_B=0,1; p_C=0,2; p_D=0,2 és p_E=0,1.

Egy tipikus példa így előállt szövegre (Shannon művéből):

A A A C D C B D C E A A D A D A C E D A E A D C A B E
D A D D C E C A A A A A D.

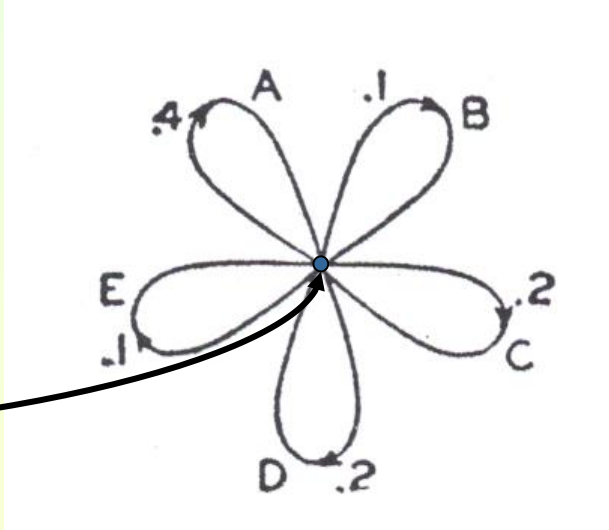
A források jellemzése

A források leírhatók Markov-folyamatokkal.

Egy összetettebb Markov-folyamat során a karakterek előfordulási valószínűsége más és más, a szimbólum-kibocsátások független események.

Legyen p_A=0,4; p_B=0,1; p_C=0,2; p_D=0,2 és p_E=0,1.

A folyamat gráfja:



Egyetlen állapot

A források jellemzése

Még összetettebb Markov-folyamatok során nem csak a karakterek előfordulási valószínűsége más és más, hanem a szimbólum-kibocsátások sem független események. A legegyszerűbb ilyen eset, ha az aktuális szimbólum csak az őt eggyel megelőzőtől függ.

Legyen $p_A=1/3$; $p_B=16/27$; $p_C=2/27$, és a feltételes valószínűségek:

$P(a_i \mid a_j)$		i = A	i = B	i = C
j	A	0	4/5	1/5
	B	1/2	1/2	0
	C	1/2	2/5	1/10

A források jellemzése

Legyen $p_A=1/3$; $p_B=16/27$; $p_C=2/27$, és a feltételes valószínűségek:

$P(a_i \mid a_j)$		i = A	i = B	i = C
j	A	0	4/5	1/5
	B	1/2	1/2	0
	C	1/2	2/5	1/10

Egy tipikus példa így előállt szövegre (Shannon művéből):

A B B A B A B A B A B A B A B B B A B B B B B A B A B A
B A B A B B B A C A C A B B A B B B B A B B A B A C B B
B A B A.

A források jellemzése

Ellenőrizhetjük a feltételes és együttes előfordulási valószínűségek közötti összefüggéseket. Az együttes valószínűségekre igaz:

$$p(a_j \cdot a_i) = p(a_j) \cdot p(a_i | a_j)$$

ahol $p_A=1/3=9/27$; $p_B=16/27$; $p_C=2/27$, és

$P(a_i \mid a_j)$		i		
		A	B	C
j	A	0	4/5	1/5
	B	1/2	1/2	0
	C	1/2	2/5	1/10

$P(a_j \cdot a_i)$		i		
		A	B	C
j	A	0	4/15	1/15
	B	8/27	8/27	0
	C	1/27	4/135	1/135

$$p(AC) = p(A) \cdot p(C|A) = \frac{1}{3} \cdot \frac{1}{5}$$

A források jellemzése

$p_A=1/3=9/27$; $p_B=16/27$; $p_C=2/27$, és

$P(a_i \mid a_j)$		i			Σ
		A	B	C	
j	A	0	4/5	1/5	1
	B	1/2	1/2	0	1
	C	1/2	2/5	1/10	1

$$\sum_i p(a_i|a_j)=1$$

$P(a_j \cdot a_i)$		i			Σ
		A	B	C	
j	A	0	4/15	1/15	1/3
	B	8/27	8/27	0	16/27
	C	1/27	4/135	1/135	2/27
Σ		9/27	16/27	2/27	1

$$\sum_i p(a_j \cdot a_i) = p(a_j)$$

$$\sum_j p(a_j \cdot a_i) = p(a_i)$$

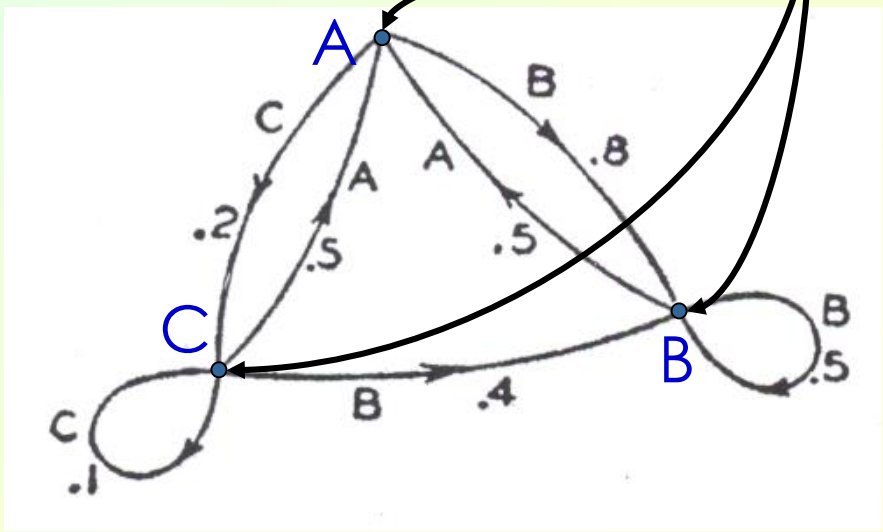
A források jellemzése

Legyen $p_A=1/3$; $p_B=16/27$; $p_C=2/27$, és a feltételes valószínűségek:

$P(a_i \mid a_j)$		i = A	i = B	i = C
j	A	0	4/5	1/5
	B	1/2	1/2	0
	C	1/2	2/5	1/10

3 állapot

A gráf:



A források jellemzése

A források leírhatók Markov-folyamatokkal.

Még magasabb összetettségi szintű Markov-folyamatok során a karakterek helyett a belőlük épített szavaknak van valamilyen előfordulási statisztikája. A szavakat is választhatjuk egymástól függetlenül.

Egy ilyen példa (Shannon művéből):
A következő 16 szó fordulhat elő az előttük álló valószínűségekkel:

.10 A	.16 BEBE	.11 CABED	.04 DEB
.04 ADEB	.04 BED	.05 CEED	.15 DEED
.05 ADEE	.02 BEED	.08 DAB	.01 EAB
.01 BADD	.05 CA	.04 DAD	.05 EE

A források jellemzése

A források leírhatók Markov-folyamatokkal.

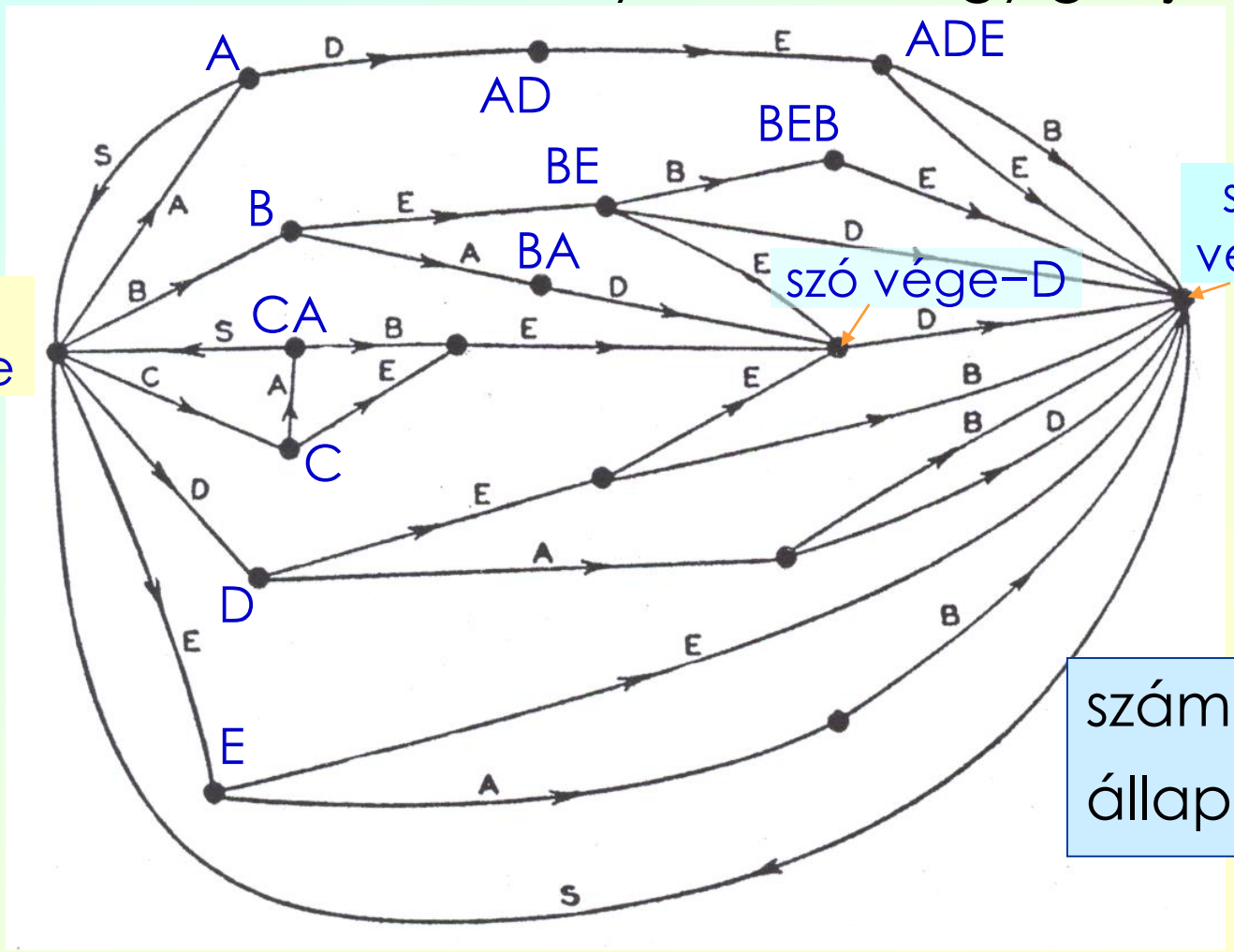
Még magasabb összetettségi szintű Markov-folyamatok során a karakterek helyett a belőlük épített szavaknak van valamilyen előfordulási statisztikája. A szavakat is választhatjuk egymástól függetlenül.

Egy tipikus példa így előállt szövegre (Shannon művéből):

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE
BEBE BEBE ADEE BED DEED DEED CEED ADEE A
DEED DEED BEBE CABED BEBE BED DAB DEED ADEB.

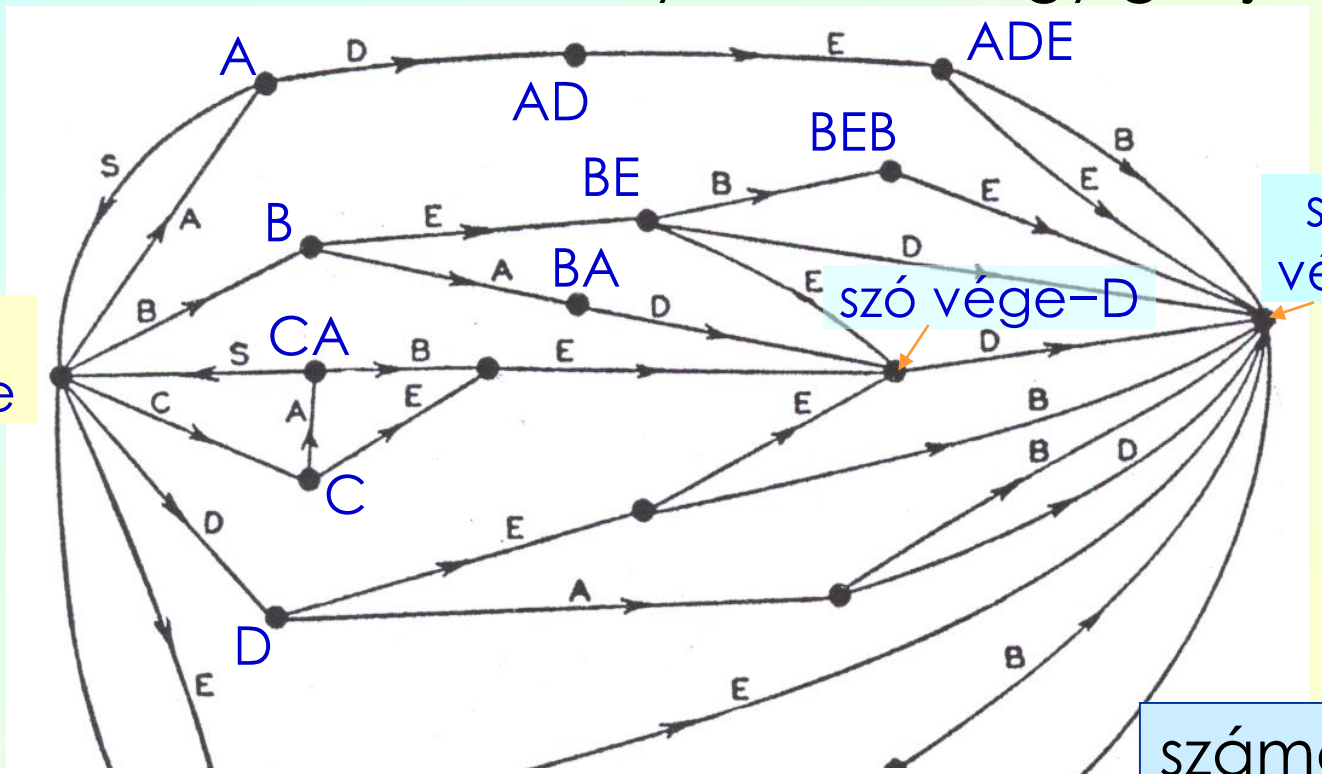
A források jellemzése

A szavakat alapul vevő, független szóválasztású Markov-folyamatunk egy gráfja:



A források jellemzése

A szavakat alapul vevő, független szóválasztású Markov-folyamatunk egy gráfja:



.10 A	.16 BEBE	.11 CABED	.04 DEB
.04 ADEB	.04 BED	.05 CEED	.15 DEED
.05 ADEE	.02 BEED	.08 DAB	.01 EAB
.01 BADD	.05 CA	.04 DAD	.05 EE

A források jellemzése – forrásentrópia

Ha a forrás az S_i állapotban van, akkor minden j-re $p(S_j \mid S_i)$ ismert, ebből

$$H(S|S_i) = \sum_j p(S_j|S_i) \cdot \log_2 p(S_j|S_i)$$

Ha ismerjük az S_i állapot P_i előfordulási valószínűségét, akkor a **forrásentrópia**:

$$H = \sum_i H(S|S_i) = \sum_{ij} P_i \cdot p(S_j|S_i) \cdot \log_2 p(S_j|S_i)$$

Ha a forrás stacionárius, azaz $p(S_j \mid S_i)=p(A_j \mid A_i)$ és $p(A_j \mid A_i)= p(A_j)$, akkor a forrásentrópia megegyezik az egyetlen szimbólum kibocsátásának entrópiájával:

$$H = H(A) = \sum_{i=1}^{n} p_i \log_2 p_i$$

A források jellemzése – forrásentrópia

Vizsgáljuk a forrás egymást követő N szimbólum-kibocsátását:
az $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ sorozatot.

Az A forrás **forrásentrópiája**:

$$H(A) = \lim_{N\to\infty} \frac{1}{N} H\left(A^{(1)}, A^{(2)}, \ldots, A^{(N)}\right)$$

Nem keverendő $H(A) = \sum_{i=1}^{n} p_i \log_2 p_i$ -vel, a forrás-ábécé entrópiájával.

Ha a forrás stacionárius, akkor H=H(A)

A forráskódok jellemzése

A kódolt üzenetek egy $B = \{B_1, \ldots, B_s\}$ szintén véges halmaz elemeiből épülnek fel, B a **kódábécé**.

A B elemeiből álló véges hosszúságú $B^{(1)}$ $B^{(2)}$ $B^{(3)}$ … $B^{(m)}$ sorozatok a **kódszavak**.

A lehetséges kódszavak halmaza $\boldsymbol{B}$.

Az $f : A \mapsto \boldsymbol{B}$, illetve $F : \boldsymbol{A} \mapsto \boldsymbol{B}$ függvényeket (forrás)**kód**oknak nevezzük. Az f leképezés a forrás egy-egy szimbólumához rendel egy-egy kódszót, az F ennél általánosabb.

Egyértelműen dekódolható kódok

Egy *f* forráskód egyértelműen dekódolható, ha minden egyes B-beli sorozatot csak egyféle A-beli sorozatból állít elő. (A neki megfeleltethető *F* invertálható. Az nem elég, hogy *f* invertálható.) Az állandó kódszóhosszú kódok egyértelműen dekódolhatók, megfejthetők, de nem elég gazdaságosak.

Egy *f* betűnkénti kód **prefix**, ha a lehetséges kódszavak közül egyik sem áll elő egy másik folytatásaként.

	prefix	posztfix	
a	0	0	0
b	10	01	01
c	110	011	011
d	1110	0111	1110

A kódszavak átlagos hossza

Az olyan $f : A \mapsto B$ kódokat, amelyek különböző A-beli szimbólumokhoz más és más hosszúságú kódszavakat rendelnek, változó szóhosszúságú kódoknak nevezzük.

Az $f(A_i) = B^{(1)} B^{(2)} \dots B^{(\ell_i)}$ B-beli sorozat, avagy kódszó **hossza** ℓ_i.

Egy f kód **átlagos szóhossza** ℓ_i várható értéke:

$$L(A) = \sum_{i=1}^{n} p(A_i)\ell_i$$

$$= \sum_{i=1}^{n} p_i \ell_i$$

A_i	kódszó	ℓ_i	p_i	$L(A)$
a	0	1	0,42	1,91
β	01	2	0,34	
γ	011	3	0,15	
δ	0111	4	0,09	

A kódszavak átlagos hossza

Az olyan $f : A \mapsto B$ kódokat, amelyek különböző A-beli szimbólumokhoz más és más hosszúságú kódszavakat rendelnek, változó szóhosszúságú kódoknak nevezzük.

Az $f(A_i) = B^{(1)} B^{(2)} \dots B^{(\ell_i)}$ B-beli sorozat, avagy kódszó **hossza** ℓ_i.

Egy f kód **átlagos szóhossza** ℓ_i várható értéke:

$$L(A) = \sum_{i=1}^{n} p(A_i)\ell_i$$

$$= \sum_{i=1}^{n} p_i \ell_i$$

A_i	kódszó	ℓ_i	p_i	$L(A)$
α	011	3	0,42	2,95
β	0111	4	0,34	
γ	0	1	0,15	
δ	01	2	0,09	

A kódszavak átlagos hossza

Az olyan $f : A \mapsto B$ kódokat, amelyek különböző A-beli szimbólumokhoz más és más hosszúságú kódszavakat rendelnek, változó szóhosszúságú kódoknak nevezzük.

Az $f(A_i) = B^{(1)} B^{(2)} \dots B^{(\ell_i)}$ B-beli sorozat, avagy kódszó **hossza** ℓ_i.

Egy f kód **átlagos szóhossza** ℓ_i várható értéke:

$$L(A) = \sum_{i=1}^{n} p(A_i)\ell_i$$

$$= \sum_{i=1}^{n} p_i \ell_i$$

A_i	kódszó	ℓ_i	p_i	$L(A)$
α	0111	3	0,42	3,09
β	011	4	0,34	
γ	01	1	0,15	
δ	0	2	0,09	

A Jensen-egyenlőtlenség

Ha f egy valós, konvex függvény az $[a,b]$ intervallumon, Z pedig valószínűségi változó, mely értékét az $[a,b]$ intervallumon veszi fel, akkor

$$f\left(\langle Z\rangle\right)\le\langle f(Z)\rangle$$

Bizonyítás: ha f konvex,

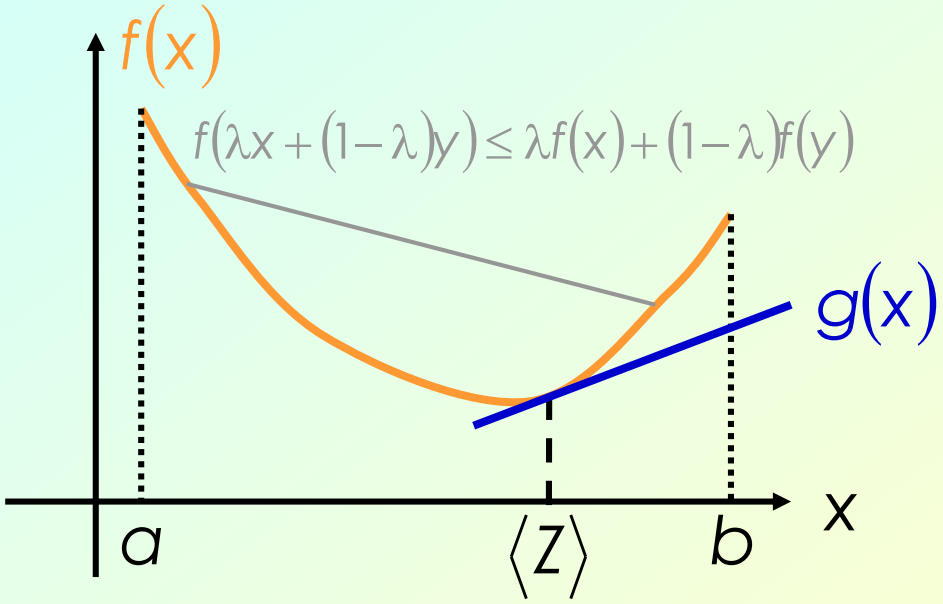
$$f\left(\lambda x+(1-\lambda)y\right)\le\lambda f(x)+(1-\lambda)f(y)\quad\forall x,y\in(a,b)$$

és $\forall x\in(a,b)$-re $\exists$ jobb és bal oldali deriváltja f-nek

$$f_+'(x)\quad\text{és}\quad f_-'(x)$$

A Jensen-egyenlőtlenség

A jobb oldali érintő: $g(x) = f(\langle Z \rangle) + f_+'(\langle Z \rangle) \cdot (x - \langle Z \rangle)$



ekkor $f(x) \geq g(x)$, így

$$f(x) \geq f(\langle Z \rangle) + f_+'(\langle Z \rangle) \cdot (x - \langle Z \rangle)$$

várható értéke 0

várható értéket véve: $\langle f(Z) \rangle \geq f(\langle Z \rangle)$

A Jensen-egyenlőtlenség

Állítás: ha $p_i \geq 0, \quad q_i > 0, \quad i=1,2,\ldots,n, \quad$ és

$$\sum_{i=1}^{n} p_i = 1 \quad \text{és} \quad \sum_{i=1}^{n} q_i = 1,$$

akkor

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i.$$

(Egyenlőség csak $p_i=q_i \ \forall i$ esetén)

Bizonyítás:

$$-\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \geq 0$$

belátása:

A Jensen-egyenlőtlenség

A belátandó: $-\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \geq 0$

Legyen egy új vsz.-i változó: $p\left(y = \frac{q_i}{p_i}\right) = p_i$,
így

$$-\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} = \left\langle -\log_2 \frac{q_i}{p_i} \right\rangle$$

Mivel a −log függvény konvex:

$$\left\langle -\log_2 \frac{q_i}{p_i} \right\rangle \geq -\log_2 \left\langle \frac{q_i}{p_i} \right\rangle = \underbrace{-\log_2 \underbrace{\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i} \right)}_{1}}_{0}$$

A McMillan-egyenlőtlenség

Minden egyértelműen dekódolható kódra igaz, hogy

$$\sum_{i=1}^{n} s^{-\ell_i} \le 1$$

ahol s a kódábécé elemszáma n pedig a forrásábécéé.

Bizonyítás: az összeg N-edik hatványa:

$$\left(\sum_{i=1}^{n} s^{-\ell_i}\right)^N = \sum_{i_1=1}^{n} \dots \sum_{i_N=1}^{n} s^{-\left(\ell_{i_1}+\dots+\ell_{i_N}\right)} = \sum_{\ell=1}^{N\cdot L_{\max}} A_\ell s^{-\ell}$$

ahol $L_{\max} = \max_{i\in[1,n]} \ell_i$, A_ℓ pedig az N kódszó egymás után íráskor keletkezhető ℓ elemű sztringek száma.

A McMillan-egyenlőtlenség

A kód egyértelműen dekódolható, így

$$s^{\ell} \geq A_{\ell}$$

ebből

$$\left(\sum_{i=1}^{n} s^{-\ell_i}\right)^{N} = \sum_{\ell=1}^{N\cdot L_{\max}} A_{\ell} s^{-\ell} \leq N \cdot L_{\max}$$

így

$$\sum_{i=1}^{n} s^{-\ell_i} \leq \sqrt[N]{N \cdot L_{\max}} = \sqrt[N]{N} \cdot \sqrt[N]{L_{\max}}$$

$N \to \infty$: $\sqrt[N]{N} \to 1$, $N \to \infty$: $\sqrt[N]{L_{\max}} \to 1$

A Kraft-egyenlőtlenség

Legyen $\ell_1, \ell_2, \ldots, \ell_n \in N$, $s > 1$ egész, és legyen rájuk érvényes, hogy

$$\sum_{i=1}^{n} s^{-\ell_i} \leq 1$$

Ekkor létezik olyan prefix kód, amelynek kódábécéje s elemű, és az n elemű forrásábécé $A_1, A_2, \ldots, A_n$ elemeihez rendelt kódszavak hossza rendre $\ell_1, \ell_2, \ldots, \ell_n$.

Bizonyítás: legyen $\ell_1 \leq \ell_2 \leq \ldots \leq \ell_n$ és

$$w_1 = 0, \quad w_j = \sum_{i=1}^{j-1} s^{\ell_j - \ell_i} \quad j = 2, \ldots, n$$

A Kraft-egyenlőtlenség

Ekkor

$$\sum_{i=1}^{n} s^{-\ell_i} \le 1 \Longrightarrow \sum_{i=1}^{n} s^{\ell_n-\ell_i} \le s^{\ell_n} \Longrightarrow w_n = \sum_{i=1}^{n-1} s^{\ell_n-\ell_i} \le s^{\ell_n} - 1$$

és $w_j \le s^{\ell_j} - 1$.

Legyen w_j szám s-alapú számrendszerbeli alakjának eléírt 0-kal ℓ_j hosszúságúra kiegészített verziója f_j. Ha ezt a kódszót rendeljük az a_j forrásábécébeli elemhez, akkor nyilván minden betűhöz más és más kódszó fog jutni, mivel

$$j < k \quad \text{esetén} \quad w_j < w_k$$

A Kraft-egyenlőtlenség

Indirekt bizonyítása annak, hogy a kapott kódrendszer prefix:

tegyük fel, hogy $\exists j,k,\ j < k$, melyre ha f_j végére megfelelő $\ell_k - \ell_j$ db számjegyet írunk f_k-t kapjuk.

Osszuk el f_k-t (w_k-t) $s^{\ell_k-\ell_j}$-nel. Ha a fenti állítás igaz, akkor

$$w_j = \left\lfloor \frac{w_k}{s^{\ell_k-\ell_j}} \right\rfloor$$

Ugyanakkor

$$\frac{w_k}{s^{\ell_k-\ell_j}} = \sum_{i=1}^{k-1} s^{\ell_j-\ell_i} = w_j + \sum_{i=j}^{k-1} s^{\ell_j-\ell_i} \geq w_j + 1$$

$$w_k = \sum_{i=1}^{k-1} s^{\ell_j-\ell_i}$$

$$\geq 1$$

A kódszavak átlagos hosszáról szóló tétel

Minden egyértelműen dekódolható $f : A \mapsto B$ kódra

$$L(A) \geq \frac{H(A)}{\log_2 s}.$$

Bizonyítás:

$$H(A) \leq L(A) \cdot \log_2 s = \sum_{i=1}^{n} p_i \ell_i \cdot \log_2 s = -\sum_{i=1}^{n} p_i \cdot \log_2 s^{-\ell_i}$$

A Jensen-egyenlőtlenség egy következménye, hogy ha $p_i \geq 0$, $q_i > 0$, és $\sum_{i=1}^{n} p_i = 1$, és $\sum_{i=1}^{n} q_i = 1$, akkor

$$\underbrace{-\sum_{i=1}^{n} p_i \log_2 p_i}_{H(A)} \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

$$q_i = \frac{s^{-\ell_i}}{\sum_{j=1}^{n} s^{-\ell_j}}$$

A kódszavak átlagos hosszáról szóló tétel

$$H(A) \le -\sum_{i=1}^{n} p_i \log_2 q_i = -\sum_{i=1}^{n} p_i \log_2 \frac{s^{-\ell_i}}{\sum_{j=1}^{n} s^{-\ell_j}} =$$

Független i-től, állandó

$$= -\sum_{i=1}^{n} p_i \log_2 s^{-\ell_i} + \sum_{i=1}^{n} p_i \underbrace{\log_2 \sum_{j=1}^{n} s^{-\ell_j}}_{\text{McMillan: } \le 1}$$

$$\underbrace{\qquad\qquad}_{< 0}$$

$$\le -\sum_{i=1}^{n} p_i \log_2 s^{-\ell_i}$$

A kódszavak átlagos hosszáról szóló második tétel

Létezik olyan $f : A \mapsto B$ prefix kód, melyre

$$L(A) < \frac{H(A)}{\log_2 s} + 1$$

Bizonyítás: Legyenek $\ell_1, \ell_2, \ldots, \ell_n$ pozitív egész számok, melyekre

$$\boxed{-\frac{\log_2 p_i}{\underbrace{\log_2 s}_{\log_s p_i}} \le \ell_i} < -\frac{\log_2 p_i}{\log_2 s} + 1, \qquad \forall i.$$

$$\log_s p_i \ge -\ell_i$$

$\cdot s^{\wedge}, \sum_i$

$$\sum_{i=1}^{n} s^{-\ell_i} \le \sum_{i=1}^{n} s^{-\log_s p_i} = \sum_{i=1}^{n} p_i = 1$$

A kódszavak átlagos hosszáról szóló második tétel

Létezik olyan $f : A \mapsto B$ prefix kód, melyre

$$L(A) < \frac{H(A)}{\log_2 s} + 1$$

Bizonyítás: Legyenek $\ell_1, \ell_2, \ldots, \ell_n$ pozitív egész számok, melyekre

$$-\frac{\log_2 p_i}{\log_2 s} \le \boxed{\ell_i < -\frac{\log_2 p_i}{\log_2 s} + 1,} \qquad \forall i.$$

$\cdot p_i$, $\sum_i$

$$\underbrace{\sum_{i=1}^{n} p_i \ell_i}_{L(A)} < \underbrace{\sum_{i=1}^{n} p_i \frac{\log_2 p_i}{\log_2 s}}_{H(A)/\log_2 s} + \underbrace{\sum_{i=1}^{n} p_i}_{1}$$

Shannon forráskódolási tétele

Minden $A = \{ A_1, A_2, \ldots, A_n \}$ véges forrásábécéjű forráshoz található olyan s elemű kódábécével rendelkező $f : A \mapsto B$ kód, amely az egyes forrásszimbólumokhoz rendre $\ell_1, \ell_2, \ldots, \ell_n$ szóhosszúságú kódszavakat rendel, és

$$\frac{H(A)}{\log_2 s} \le L(A) < \frac{H(A)}{\log_2 s} + 1$$

Az olyan kódok, amelyekre ez teljesül az **optimális kódok**.

Huffman-kód

A legrövidebb átlagos szóhosszú bináris prefix kód.

1. Valószínűségek szerint sorba rendez
2. A két legkisebb valószínűségű szimbólumot összevonja. Az összevont szimbólum valószínűsége a két eredeti szimbólum valószínűségének összege.
3. Az 1-2. lépést addig ismétli, amíg egyetlen, 1 valószínűségű összevont szimbólumot nem kap.

Huffman-kód

A legrövidebb átlagos szóhosszú bináris prefix kód.

3. Az 1-2. lépést addig ismétli, amíg egyetlen, 1 valószínűségű összevont szimbólumot nem kap.
4. A kapott gráf minden csomópontja előtti két élt megcímkézi 0-val és 1-gyel.
5. A kódfa gyökerétől elindulva megkeresi az adott szimbólumhoz tartozó útvonalat, kiolvassa az éleknek megfelelő biteket. A kapott bitsorozatot rendeli a szimbólumhoz kódszóként.

Aritmetikai kód

Legyen a forrásábécé elemszáma *n*, és *m* elemű **blokkokat** kódoljunk.

1. Felosztja a [0,1) intervallumot *n* diszjunkt részre, minden résznek megfeleltet egy-egy forrásábécébeli elemet. Célszerű a kis valószínűségű betűkhöz rövid, a gyakoriakhoz hosszú részintervallumot rendelni.
2. Kiválasztja a blokk soron következő karakterének megfelelő intervallumot.
3. Az új intervallumot felosztja ugyanolyan arányban, mint a [0,1)-et osztotta, és ugyanolyan sorrendben rendeli a részintervallumokhoz a lehetséges szimbólumokat.

Aritmetikai kód

Legyen a forrásábécé elemszáma *n*, és *m* elemű **blokkokat** kódoljunk.

3. Az új intervallumot felosztja ugyanolyan arányban, mint a [0,1)-et osztotta, és ugyanolyan sorrendben rendeli a részintervallumokhoz a lehetséges szimbólumokat.
4. A 2-3. lépéseket ismétli, amíg el nem fogy a blokk.
5. A végül maradt kis intervallumból kiválaszt egy (binárisan) jól leírható számot, az lesz a kódszó.

Lempel—Ziv-algoritmusok

Nem szükséges előre ismerni a kódolandó karakterek előfordulási valószínűségét. Az üzenet bolvasása során egy láncolt listát, ú.n. **szótár**at épít. Egy szótársornak 3 mezője van: *m* sorszám, *n* mutató és a karakter. A kódolt információ a sorszámokból álló sorozat lesz.

A kódolás során a vevő is megkapja a szükséges információt, párhuzamosan építi a szótárat, vagy pedig a tömörített fájlban szerepel maga a szótár is.

LZ78

A szótár nulladik sora adott: m=0, n=0 a karakter pedig üres. A kódolás elején a megjegyzett sorszám n_m= 0, az utolsó használt sorszám is m_U=0.

A kódoló a következő lépéseket ismétli, amíg el nem fogy az üzenet:

- Beolvassa a következő karaktert, amit nevezzünk „c”-nek
  - Ha nem szerepel a karakter a szótárban nyit neki egy új sort, a sor paraméterei: m=m_U+1, n=0, a karakter „c”.
    A megjegyzett elem n_m= 0 az utolsó sorszám m_U=m_U+1.

LZ78

- Ha már szerepel a karakter a szótárban, akkor vizsgálja azokat a sorokat, amelyeknek a megjegyzett n_m szerepel a mutató mezejükben.
  - Ha talál olyant, amelynek a karaktermezejében „c” szerepel, annak a sornak az indexe lesz az új n_m, m_U nem változik.
  - Ha nem talál olyan sort, amelyikben „c” a karakter, akkor nyit egy újat. A sorszám $m=m_U+1$, a mutató n_m, a karakter „c”.
    Az új megjegyzett sorszám 0. A használt utolsó sorszám $m_U=m_U+1$.

LZW

A szótár első *k* sora tartalmazza a használni kínánt *k* darab karaktert. A kódolás elején a megjegyzett sorszám n_m= 0, az utolsó használt sorszám m_U=k.

A kódoló a következő lépéseket ismétli, amíg el nem fogy az üzenet:

- Beolvassa a következő karaktert, amit nevezzünk „c”-nek. Vizsgálja azokat a sorokat, amelyeknek a megjegyzett n_m szerepel a mutató mezejükben.
    - Ha talál olyant, amelynek a karaktermezejében „c” szerepel, annak a sornak az indexe lesz az új n_m.

LZW

- Ha talál olyant, amelynek a karaktermezejében „c” szerepel, annak a sornak az indexe lesz az új $n_{m.}$
- Ha nem talál olyan sort, amelyikben „c” a karakter, akkor nyit egy újat. A sorszám $m=m_u+1$, a mutató n_m, a karakter „c”. Az új megjegyzett sorszám annak a sornak az m-je, ahol a „c” karakter először szerepelt. A használt utolsó sorszám $m_u=m_u+1$. Az üzenet ezen láncához rendelt kódszó n_m.