Bevezetés az információelméletbe

Csiszár Villő

2017. május 10.

1. A hírközlési rendszerek matematikai modellje

Olyan rendszerekkel foglalkozunk, amikor egy forrás által kibocsátott információt valamilyen műszaki berendezésen – csatornán – keresztül el kell juttatni egy felhasználóhoz. Ilyen kommunikációs rendszerek részei a telefon, rádió, tévé, számítógép, router, műhold, de gondolhatunk arra is, amikor adathordozón tárolunk adatokat, például könyvben, CD-n, DVD-n.

A forrás egy *véges* forrásábécébe tartozó jeleket bocsát ki egyesével (lehet a forrásábécé megszámlálhatóan végtelen, vagy folytonos is, ezekkel kevésbé fogunk foglalkozni). Olyan rendszerekre gondolunk, amikor nagyon sok ilyen jel van, és a kibocsátás véletlenszerű, azaz a forrást az $(X_i)_{i=1}^{\infty}$ valószínűségi változók együttes eloszlásával adjuk meg. A forrás emlékezet nélküli, ha az X_i jelek függetlenek, a forrás stacionárius, ha az X_i sorozat stacionárius.

A forrás által kibocsátott jeleket a csatornán való átküldés előtt kódoljuk. Ennek több oka lehet: (a) a csatorna ábécéje különbözik a forrás ábécéjétől, (b) tömörítéssel gazdaságosabb információátvitel valósítható meg, (c) zajos csatorna esetén alkalmazhatunk olyan kódolást, mely a zajhatást csökkenti.

Mint már említettük, a csatorna saját ábécével rendelkezik (sőt, lehet külön bemeneti és kimeneti ábécé). Akkor nevezzük zajmentesnek, ha a kimenő jelek egyértelműen meghatározzák a bemenő jeleket, ellenkező esetben a csatorna zajos.

A csatorna túlsó végén helyezkedik el a dekódoló, ami a kijövő jelek alapján megpróbálja megállapítani, hogy a forrás milyen jeleket adott le, majd ezt továbbítja a felhasználónak.

A hírközlési rendszernek ez a modellje blokk-diagrammal ábrázolható. Feltesszük, hogy a forrás és a csatorna sztochasztikus működését ismerjük, feladatunk pedig az, hogy ehhez keressünk bizonyos kritériumoknak megfelelő kódolást és dekódolást.

Jegyzetünk a következő témaköröket érinti:
1) Hogyan mérhető a forrás által kibocsátott információ mennyisége? Megmutatjuk, hogy minden közlemény információtartalma számszerűsíthető, és ez az információtartalom a közlemény valószínűségétől függ.
2) Hogyan tudjuk a forrás által kibocsátott közleményeket kódolni? Olyan kódokat fogunk vizsgálni, amikor a közleményt állandó hosszú blokkokra bontjuk, és a blokkokat kódoljuk. Természetesen csak olyan kódokat szeretnénk használni, melyek egyértelműen (vagy legalábbis kis hibavalószínűséggel) dekódolhatók.
3) Hogyan kódoljunk, ha a közleményt zajos csatornán kell átküldeni? Hogyan mérhetjük, hogy egy adott csatornán milyen sebességgel továbbítható megbízhatóan az információ?

2. Veszteségmentes forráskódolás

Legyen a forrásábécé az m elemű $\mathcal{X}$ halmaz. A forrásábécé betűiből alkotott véges sorozatokat közleményeknek nevezzük. Először a forráskódolás feladatát járjuk körül, ami azt jelenti, hogy a közleményeket szeretnénk kódolni egy $\mathcal{Y}$ kódábécével. Az egyszerűség kedvéért általában feltesszük, hogy $\mathcal{Y} = \{0, 1\}$, azaz bináris kódokat tekintünk. Az eredmények – a megfelelő változtatásokkal – ugyanígy bizonyíthatók az s elemű kódábécé esetére is.

Tegyük fel először, hogy a közleményeket betűnként szeretnénk kódolni a $g(x) \in \mathcal{Y}^*$ kódszavakkal, ahol $\mathcal{Y}^*$ a kódábécé betűiből alkotható véges sorozatok halmaza. A kódolást egyértelműen dekódolhatónak nevezzük, ha különböző közlemények kódja különböző. Az egyértelműen dekódolható kódok fontos speciális esete a *prefix* kódok esete.

2.1. Definíció. *A $g : \mathcal{X} \to \mathcal{Y}^*$ kód prefix tulajdonságú, ha minden $x \neq z$ esetén a $g(x)$ kódszó nem folytatása a $g(z)$ kódszónak.*

Prefix kódot kapunk például úgy, ha minden kódszó végére egy külön erre a célra fenntartott „szóköz" jelet teszünk, ez azonban nem túl gazdaságos. Az állandó hosszúságú kódok is prefix kódok, ilyenkor minden kódszó ugyanolyan hosszúságú.

2.2. Példa. Legyen a forrásábécé $\mathcal{X} = \{A, B, C\}$, nézzük meg ennek néhány bináris kódját.

1. $g(A) = 00, g(B) = 01, g(C) = 11$ egy állandó (mégpedig 2) hosszúságú kód.
2. $g(A) = 00, g(B) = 01, g(C) = 1$ nem állandó hosszúságú, de prefix kód.
3. $g(A) = 0, g(B) = 01, g(C) = 1$ nem egyértelműen dekódolható, hiszen a 01 egyaránt lehet a „B" és az „AC" közlemény kódja.
4. $g(A) = 00, g(B) = 10, g(C) = 1$ nem prefix kód, viszont egy prefix kód megfordítása, és így egyértelműen dekódolható. Ilyenkor az kódolt üzenetet hátulról kezdve lehet megfejteni, pl. az 1100000100 a „$CBAACA$" közlemény kódja.

Jelölje $L(x)$ a $g(x)$ kódszó hosszát. Azt szeretnénk elérni, hogy a kódunk minél rövidebb legyen.

2.3. Tétel. (Kraft egyenlőtlenség) *Az $\{L(x) : x \in \mathcal{X}\}$ sorozat akkor és csak akkor felel meg egy bináris prefix kód kódszóhosszainak, ha*

$$\sum_{x \in \mathcal{X}} 2^{-L(x)} \leq 1.$$

Bizonyítás. Az egyik irányban, legyen g prefix kód, feleltessük meg minden $g(x)$-nek a $t(x) = 0.g(x)_1 \dots g(x)_{L(x)}$ bináris (kettes számrendszerben felírt) számot. A prefix tulajdonság miatt minden $z \neq x$-re

$$t(z) \notin [t(x), t(x) + 2^{-L(x)}).$$

Ezért a $[t(x), t(x)+2^{-L(x)})$ intervallumok diszjunktak, és mivel mind benne vannak a $[0, 1)$ intervallumban, hosszuk összege legfeljebb 1.

Fordítva, tegyük fel, hogy teljesül a tételbeli egyenlőtlenség. Számozzuk $\mathcal{X}$ elemeit úgy, hogy

$$L(x_1) \leq L(x_2) \leq \dots \leq L(x_m).$$

Legyen $t(x_i) = \sum_{j<i} 2^{-L(x_j)}$, ez tehát egy legfeljebb $L(x_i)$ jegyű bináris törtszám a $[0, 1)$ intervallumban. Legyen a $g(x_i)$ kódszó a $t(x_i)$ szám $L(x_i)$ bináris jegyig kiírva (tehát esetleg nullákat teszünk a végére), ez prefix kód lesz. Ugyanis a kódszavak növekvő hossza miatt $g(x_j)$ csak úgy lehetne folytatása $g(x_i)$-nek, ha $j > i$, ebben az esetben azonban

$$t(x_j) - t(x_i) = \sum_{k=i}^{j-1} 2^{-L(x_j)} \geq 2^{-L(x_i)},$$

tehát $t(x_j)$ első $L(x_i)$ bináris jegye nem egyezhet meg $t(x_i)$ megfelelő bináris jegyeivel. ■

Vegyük észre, hogy a bizonyításban nem használtuk, hogy az $\mathcal{X}$ ábécé véges, tehát az állítás megszámlálhatóan végtelen ábécére is igaz. Ha az $\mathcal{Y}$ kódábécé elemszáma s, akkor a tételbeli egyenlőtlenség helyett $\sum_{x \in \mathcal{X}} s^{-L(x)} \leq 1$ lesz szükséges és elégséges feltétel. A Kraft-egyenlőtlenség egyértelműen dekódolható kódokra is igaz, ezt McMillan bizonyította.

2.4. Tétel. (McMillan) *Az $\{L(x) : x \in \mathcal{X}\}$ sorozat akkor és csak akkor felel meg egy egyértelműen dekódolható bináris kód kódszóhosszainak, ha*

$$\sum_{x\in\mathcal{X}} 2^{-L(x)} \leq 1.$$

Bizonyítás. Nyilván csak azt az irányt kell bizonyítani, hogy tetszőleges egyértelműen megfejthető kód kódszóhosszai kielégítik a tételbeli egyenlőtlenséget. Vegyünk egy egyértelműen megfejthető kódot, és legyen $L_{\max}$ az $L(x)$ kódszóhosszak maximuma, k pedig tetszőleges pozitív egész szám. Egy k hosszú $x^k = (x_1, \ldots, x_k)$ közlemény kódját úgy kapjuk, hogy az egyes betűk kódszavait egymás után írjuk, azaz a közlemény kódjának hossza

$$L(x^k) = L(x_1) + \cdots + L(x_k).$$

Ezért

$$\left(\sum_{x\in\mathcal{X}} 2^{-L(x)}\right)^k = \sum_{x^k\in\mathcal{X}^k} 2^{-L(x^k)} = \sum_{r=1}^{kL_{\max}} a(r)2^{-r},$$

ahol $a(r)$ azt jelöli, hogy hány olyan k hosszú közlemény van, melynek kódja r hosszú. Az egyértelmű dekódolhatóság miatt $a(r) \leq 2^r$. Ezért

$$\sum_{x\in\mathcal{X}} 2^{-L(x)} \leq (kL_{\max})^{1/k}$$

minden k-ra, ez pedig csak úgy lehet, ha $\sum_{x\in\mathcal{X}} 2^{-L(x)} \leq 1$. ■

McMillan tétele is igaz marad, ha megszámlálhatóan végtelen $\mathcal{X}$ forrásábécét is megengedünk (a tételbeli egyenlőtlenség a kód minden véges megszorítására igaz). Azt a némileg meglepő eredményt kaptuk, hogy általános egyértelműen dekódolható kódokkal sem tudunk „rövidebb" kódot csinálni, mint a prefix kódokkal. Tehát elegendő prefix kódokat vizsgálnunk. A Kraft-egyenlőtlenség következménye Shannon egyértelműen dekódolható kódokra vonatkozó tétele, mely már a forrásábécé betűinek valószínűségét is figyelembe véve, a várható kódszóhosszra ad alsó és felső korlátot. Ehhez be kell vezetnünk az *entrópia* fogalmát.

2.5. Definíció. *Legyen $P = \{p(x) : x \in \mathcal{X}\}$ tetszőleges eloszlás $\mathcal{X}$-en. A P eloszlás Shannon-féle entrópiája*

$$H(P) = -\sum_x p(x)\log p(x),$$

ahol a logaritmus kettes alapú, és $0\log 0 = 0$ definíció szerint.

2.6. Tétel. (Shannon) *Legyen P tetszőleges eloszlás $\mathcal{X}$-en, $E(L) = \sum_x p(x)L(x)$ pedig egy egyértelműen dekódolható kód átlagos hossza. Ekkor $E(L) \geq H(P)$. Továbbá van olyan prefix kód, melyre $L(x) = \lceil -\log p(x)\rceil$, és erre $E(L) < H(P) + 1$.*

Szükség lesz a következő log-összeg egyenlőtlenségre:

2.7. Lemma. (log-összeg egyenlőtlenség) *Legyenek $p_1, \ldots, p_n$ és $q_1, \ldots, q_n$ nemnegatív számok. Ekkor*

$$\sum_i p_i \log\frac{p_i}{q_i} \geq (\sum_i p_i)\log\frac{\sum_i p_i}{\sum_i q_i}.$$

Az egyenlőség feltétele, hogy $p_i = cq_i$ legyen. Itt $p\log(p/q)$ nulla, ha $p = 0$, és ∞, ha $p > q = 0$.

Bizonyítás. Nyilván elég azt az esetet bizonyítani, ha a bal oldalon nincs ∞ tag, és a p_i-k között sincs 0. Legyen $p = \sum_i p_i$ és $q = \sum_i q_i$. A logaritmus függvény konkáv volta miatt

$$\sum_i p_i \log\frac{q_i}{p_i} = p\sum_i \frac{p_i}{p}\log\frac{q_i}{p_i} \leq p\log\left(\sum_i \frac{p_i}{p}\cdot\frac{q_i}{p_i}\right) = p\log\frac{p}{q},$$

melyből (-1)-gyel szorozva a kívánt egyenlőtlenséget kapjuk. Az egyenlőség feltétele a logaritmus függvény szigorú konkávitásából adódik. ■

Bizonyítás. (Tételé.) Első rész: Alkalmazzuk a log-összeg egyenlőtlenséget $q(x) = 2^{-L(x)}$ szereposztással, és a Kraft-egyenlőtlenséget:

$$E(L) - H(P) = \sum_x p(x)(L(x) + \log p(x)) = \sum_x p(x) \log \frac{p(x)}{2^{-L(x)}}$$
$$\geq 1 \log \frac{1}{\sum_{x \in \mathcal{X}} 2^{-L(x)}} \geq 1 \log 1 = 0.$$

Második rész: az $L(x) = \lceil -\log p(x) \rceil$ értékek kielégítik a Kraft egyenlőtlenséget, ezért van hozzájuk prefix kód. Másrészt

$$E(L) = \sum_x p(x) \lceil -\log p(x) \rceil < \sum_x p(x)(-\log p(x) + 1) = H(P) + 1.$$

■

A bizonyításból az is látszik, hogy $E(L) = H(P)$ csak úgy lehetséges, ha minden x-re $p(x) = 2^{-n(x)}$ valamilyen $n(x)$ természetes számra, és ekkor persze $L(x) = n(x)$ adja a legjobb kódszóhosszakat.

Az entrópiát Shannon vezette be 1948-ban. Előtte 1928-ban Hartley már bevezetett egy információmennyiséget: azt mondta, hogy ha megtudjuk, hogy n lehetőség melyike következett be, azzal $\log n$ bit (binary digit) információt nyerünk. Másképpen, ahhoz, hogy n lehetőség közül egyet beazonosítsunk, $\log n$ bit információra van szükség. Egy bit információ ugyanis egy igen-nem kérdésre adott válasz, és n lehetőség közül egynek a beazonosítására tényleg $\log n$ igen-nem kérdésre (illetve ennek felső egész részére) van szükség.

Shannon vette észre, hogy a definíció hiányossága, hogy az egyes lehetőségek különböző valószínűségét nem veszi figyelembe. Ő azt javasolta, hogy egy esemény bekövetkezéséhez tartozó információmennyiség függjön az esemény valószínűségétől. Jelölje egy p valószínűségű eseményhez tartozó egyedi információt $h(p)$. A következő tétel arról szól, hogyan érdemes a $h(p)$ függvényt definiálni.

2.8. Tétel. (Egyedi információ) *Tegyük fel, hogy a $h(p)$ $(0 < p \leq 1)$ nemnegatív függvényre teljesülnek az alábbiak:*
a) $h(pq) = h(p) + h(q)$, azaz független események metszetéhez tartozó egyedi információ a két esemény egyedi információjának összege,
b) $h(1/2) = 1$ (egység megválasztása).
Ekkor $h(p) = -\log p$ (a logaritmus kettes alapú).

Bizonyítás. Egyrészt $h(1 \cdot q) = h(1) + h(q)$ miatt $h(1) = 0$. Másrészt h monoton fogyó, hiszen $p > r > 0$ esetén

$$h(r) = h(p \cdot \frac{r}{p}) = h(p) + h(r/p) \geq h(p).$$

Legyen ezután $p \in (0, 1)$. Ekkor $p = 2^{-x}$ valamilyen $x > 0$ számra. Az a) tulajdonság miatt $h(p^n) = nh(p)$, azaz

$$n = h(2^{-n}) = h\left((2^{-n/m})^m\right) = mh(2^{-n/m}),$$

azaz $h(2^{-r}) = r$ minden racionális r számra. Ha most x irracionális, akkor minden m-hez van olyan n, hogy

$$n/m < x < (n+1)/m, \text{ azaz } 2^{-(n+1)/m} < 2^{-x} < 2^{-n/m}.$$

A monotonitás miatt ebből $n/m < h(2^{-x}) < (n+1)/m$, azaz minden m-re $|h(2^{-x}) - x| < 1/m$. Ebből pedig $h(2^{-x}) = x$ adódik. ■

Így $H(P)$ éppen a $P = \{p(x) : x \in \mathcal{X}\}$ valószínűségű teljes eseményrendszerhez tartozó egyedi információk várható értéke:

$$H(P) = E(h(p(X))), \text{ ahol } P(X = x) = p(x).$$

Nézzünk meg néhány konkrét kódolási eljárást közelebbről! Legyen tehát $P = \{p_1, \ldots, p_m\}$ egy eloszlás.

Az első csokorba olyan kódok tartoznak, ahol a p_i valószínűségeket nagyság szerint csökkenő sorrendbe rendezzük, azaz tegyük fel, hogy $p_1 \geq p_2 \geq \ldots \geq p_m$. Ezeknek a kódolásoknak az a hátránya, hogy a valószínűségeket át kell rendezni, ami nagy ábécé esetén munkaigényes művelet lehet. A kódnak három változatát ismertetjük. (Shannon-kódok)

- Készítsük el az $L_i = \lceil -\log p_i \rceil$ kódszóhosszakat, ezek kielégítik a Kraft-egyenlőtlenséget, és monoton nőnek. Ezután legyen az i. kódszó a $t_i = \sum_{j<i} 2^{-L_j}$ bináris tört alakjának első L_i törtjegye. A Kraft-egyenlőtlenség bizonyításában láttuk, hogy ezzel prefix kódot kaptunk.

- A második változatban $t_i = \sum_{j<i} p_j$, az L_i mennyiségek változatlanok. Ekkor is prefix kódot kapunk, hiszen $j > i$ esetén $t_j - t_i \geq p_i \geq 2^{-L_i}$, vagyis a j. kódszó nem lehet folytatása az i.-nek .

- A harmadik változatban legyen ismét $t_i = \sum_{j<i} p_j$, de az L_i hosszakat nem adjuk meg előre. Ehelyett a $[0, 1)$ intervallumból indulva, addig felezzük az intervallumokat, amíg mindegyikben legfeljebb egy t_i pont marad (minden intervallum balról zárt, jobbról nyílt). Ezután az i. kódszó a t_i pontot tartalmazó, egyre szűkülő intervallumok kódja lesz. Így nyilván prefix kódot kapunk. Mennyi ennek a kódnak az átlagos kódszóhossza? A kód konstruálása alapján t_i az egyetlen pont az őt tartalmazó 2^{-L_i} hosszú intervallumban, viszont az eggyel előtti, 2^{-L_i+1} hosszú intervallum t_{i-1} és t_{i+1} legalább egyikét még tartalmazza. Mivel $p_i = t_{i+1} - t_i \leq t_i - t_{i-1} = p_{i-1}$, ezért

$$p_i < 2^{-L_i+1}.$$

  Innen $\log p_i < -L_i + 1$, azaz $L_i < -\log p_i + 1$, amiből $E(L) < H(P) + 1$.

A második csokorba olyan kódok tartoznak, melyekhez nem kell a valószínűségeket nagyság szerint rendezni. Ezeknek két változatát mutatjuk be. (Shannon-Fano-Elias kódok)

- Legyen $t_i = \sum_{j<i} p_j + p_i/2$, ezek $(0, 1)$-beli számok. Legyen továbbá $L_i = \lceil -\log p_i \rceil + 1$, és az i. kódszó a t_i bináris tört alakjának első L_i törtjegye. Így prefix kódot kapunk, hiszen $j > i$ esetén $t_j - t_i \geq p_i/2 \geq 2^{-L_i}$, vagyis a j. kódszó nem lehet folytatása az i.-nek .

- A második változatban legyen t_i ugyanaz, mint az előbb, de az L_i hosszakat nem adjuk meg előre. Ehelyett kezdjük el felezgetni az intervallumokat: először a $[0, 1)$ intervallumot felezzük meg, majd mindig azokat a részintervallumokat felezzük tovább, melyekben egynél több t_i van. Ha készen vagyunk, akkor t_i-hez rendeljük hozzá az őt tartalmazó egyre szűkülő intervallumok kódjait. Így nyilván prefix kódot kapunk. Mennyi ennek a kódnak az átlagos kódszóhossza? A t_i-t tartalmazó utolsó előtti, 2^{-L_i+1} hosszú intervallum t_{i-1} és t_{i+1} legalább egyikét tartalmazza. Mivel t_i a p_i hosszú intervallum felezőpontja, ebből

$$p_i/2 < 2^{-L_i+1}.$$

  Innen $\log p_i - 1 < -L_i + 1$, azaz $L_i < -\log p_i + 2$, amiből $E(L) < H(P) + 2$.

Itt jegyezzük meg, hogy minden prefix kódhoz hozzárendelhető egy kódfa. A fa minden leveléhez a levélhez vezető ághoz tartozó kódszó tartozik. Egy ilyen fa keresőfaként is értelmezhető: tegyük fel, hogy n lehetőség közül kell egyet beazonosítanunk úgy, hogy minden lépésben a még szóbajövő lehetőségeket két csoportra oszthatjuk, és megkérdezhetjük, hogy a keresett elem melyik halmazba esik. Így tehát az átlagosan legkevesebb kérdést használó keresési stratégia megegyezik a legrövidebb átlagos kódszóhosszal rendelkező prefix kódnak. A keresőfát (illetve kódot) alfabetikusnak nevezzük, ha a fa bármely csúcsán átmenő ágakhoz tartozó levelek indexei szomszédosak. Ez olyan keresési stratégiának felel meg, amikor a még szóbajöhető lehetőségeket két olyan csoportra kell osztani, melyek szomszédos sorszámú elemeket tartalmaznak. Az ilyen kódok alkalmazása az alábbi: tegyük fel, hogy a $p_1, \ldots, p_m$ eloszlásból úgy szeretnénk generálni, hogy a $[0, 1]$ intervallumon egyenletes eloszlású változóról megkeressük, hogy az $s_i = \sum_{j \leq i} p_j$ osztópontok által meghatározott intervallumok közül melyikbe esik. Ezt úgy tehetjük meg, hogy egymás után választunk osztópontokat, és azokkal összehasonlítjuk a véletlen számunkat. Könnyen látszik, hogy a fenti második csokor kódjai alfabetikusak, így közel optimális stratégiát adnak arra, hogyan válasszuk sorba ezeket az osztópontokat.

2.9. Példa. Legyen $P = (\frac{7}{32}, \frac{2}{32}, \frac{4}{32}, \frac{4}{32}, \frac{5}{32}, \frac{4}{32}, \frac{4}{32}, \frac{2}{32})$. A Shannon-Fano-Elias alfabetikus kódban a t_i osztópontok: $(\frac{7}{64}, \frac{16}{64}, \frac{22}{64}, \frac{30}{64}, \frac{39}{64}, \frac{48}{64}, \frac{56}{64}, \frac{62}{64})$. Az intervallumfelezéssel kapott kódszavak tehát:

$$00, 0100, 0101, 011, 10, 110, 1110, 1111.$$

Ennek átlagos kódszóhossza 3, ugyanakkor $H(P) = 2.898$ bit. Véletlen számot úgy generálhatunk ebből az eloszlásból, hogy a a $[0, 1]$ intervallumon egyenletes eloszlású változóról először megnézzük, hogy nagyobb-e, mint 17/32. Ha igen, akkor megnézzük, hogy nagyobb-e, mint 22/32, ellenkező esetben pedig, hogy nagyobb-e, mint 7/32. És így tovább.

A Shannon tételben az alsó korlát csak akkor érhető el, ha minden valószínűség 2^{-n} alakú valamilyen egész n-re, és ekkor a Shannon-féle kód eléri ezt a korlátot. Nézzünk most egy egyszerű példát! Legyen $|\mathcal{X}| = 5$, és a valószínűségek: 0.49, 0.24, 0.1, 0.1, 0.07. A Shannon-féle kód: $00, 010, 0110, 0111, 1000$. Ez nyilván nem optimális. Egy kód hatásfokát azzal mérhetjük, hogy mennyire közelíti meg az entrópiát, azaz az $E(L)/H(P)$ hányadossal. Száz százalék hatásfokú kód csak speciális esetben létezik (lásd fent), de minden esetben létezik maximális hatásfokú kód, ugyanis egy adott kódnál nem rosszabb hatásfokú kódból csak véges sok van, tehát van köztük legjobb.

Felmerül a kérdés, hogy általános esetben megadható-e az optimális prefix kód? Igen, egy ilyen eljárás neve Huffman-féle kódolás. Megint csak a bináris esettel foglalkozunk (tetszőleges ábécé esetén is hasonló az eljárás, bár kissé bonyolultabb). A következő észrevételeket tesszük:
1) Optimális kódra a kisebb valószínűségű jelhez legalább olyan hosszú kódszó tartozik, azaz $p_1 \geq \ldots \geq p_m$ esetén $L_1 \leq \ldots \leq L_m$.
2) Optimális kód kódfájában a gyökéren kívül minden belső pont foka három.
Ebből a két észrevételből kapjuk, hogy optimális kód esetén feltehető, hogy x_{m-1} és x_m kódszava egyforma hosszú, és a két kódszó csak az utolsó bitben különbözik. Tegyük most fel, hogy $m-1$ elemű eloszlásokra már tudunk optimális kódot készíteni. Készítsünk el egy K' optimális kódot a $p_1, \ldots, p_{m-2}, p_{m-1} + p_m$ eloszlásra, majd az utolsó kódszót nullával, illetve eggyel kiegészítve kapjunk egy K kódot az eredeti eloszlásra. Ekkor K optimális lesz. Ha ugyanis nem lenne az, akkor a nála jobb M optimális kódból a fenti észrevételek szerint készíthetnénk egy M' kódot a kisebb elemszámú eloszlásra, mely K'-nél jobb lenne:

$$E(L_{M'}) = E(L_M) - (p_{m-1} + p_m) < E(L_K) - (p_{m-1} + p_m) = E(L_{K'}).$$

A Huffman-féle optimális kódot tehát iteratívan állíthatjuk elő:
1) vonjuk össze a P eloszlás két legkisebb valószínűségét
2) az új eloszlásra konstruáljunk optimális kódfát
3) az összevont valószínűséghez tartozó levélhez toldjunk hozzá két levelet.

3. Információelméleti mennyiségek

Az előző szakaszban bevezettük egy véges halmazon megadott eloszlás Shannon-féle entrópiáját. Ebből kiindulva további hasznos információelméleti mennyiségek vezethetők le.

Előbb azonban foglaljuk össze az entrópia néhány tulajdonságát. Legyen most $P = (p_1, \ldots, p_m)$ egy valószínűség-eloszlás. Megmutatjuk, hogy a $H(P)$ entrópia a következő tulajdonságokkal rendelkezik:
1) $H(P)$ a P vektor folytonos függvénye (rögzített m-re).
2) $H(P) = H(P')$, ha P és P' csak a valószínűségek sorrendjében különbözik.
3) Minden m elemű P eloszlásra $H(P) \leq H(1/m, \ldots, 1/m)$.
4) $H(p_1, \ldots, p_m, 0) = H(p_1, \ldots, p_m)$.
5) Legyen $p_m = q_1 + \ldots + q_k$, ahol $q_i \geq 0$, és legyen $Q = (q_1/p_m, \ldots, q_k/p_m)$. Ekkor

$$H(p_1, \ldots, p_{m-1}, q_1, \ldots, q_k) = H(P) + p_m H(Q).$$

6) $H(1/2, 1/2) = 1$.

Az állítások közül csak a 3) és az 5) nem triviális. A 3) a log-összeg egyenlőtlenség következménye:

$$H(1/m, \ldots, 1/m) - H(P) = \log m + \sum_i p_i \log p_i = \sum_i p_i \log \frac{p_i}{1/m} \geq 0.$$

Az 5) állítás pedig egyszerű átalakítással kapható a definícióból. Az állítás következménye, hogy az események tovább-bontásával az eseményrendszer entrópiája növekszik.

Megmutatható, hogy ha a véges P eloszlásokon definiált H függvény rendelkezik az 1) – 6) tulajdonságokkal, akkor az csak a Shannon-féle entrópia lehet.

3.1. Példa. Hasonlítsuk össze a következő három eloszlás entrópiáját!

$$\begin{array}{lll} P=(1/256,255/256) & \Rightarrow & H(P)=0.037 \text{ bit} \\ P=(1/2,1/2) & \Rightarrow & H(P)=1 \text{ bit} \\ P=(7/16,9/16) & \Rightarrow & H(P)=0.989 \text{ bit} \end{array}$$

Ha a második esetben a második valószínűséget tovább bontjuk: $1/2 = 1/4 + 1/8 + 1/8$, akkor

$$H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) = 1.75 \text{ bit } = H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}\cdot H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) = 1 + \frac{1}{2}\cdot 1.5 = 1.75 \text{ bit}.$$

Mostantól az egyedi információ, az entrópia, és egyéb bevezetendő információelméleti mennyiségek argumentumaiba eloszlást és valószínűségi változót is írunk majd, remélve, hogy ez nem okoz nagy zavart. Legyenek tehát X, Y, Z valószínűségi változók az $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ véges halmazokon. Legyen

$$\begin{aligned} h_X(x) &= -\log P(X=x) \\ h_{X|Y}(x|y) &= -\log P(X=x|Y=y) \\ h_{X,Y}(x\wedge y) &= \log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} \\ h_{X,Y|Z}(x\wedge y|z) &= \log \frac{P(X=x,Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)} \end{aligned}$$

A bal oldalon álló mennyiségek nevei: egyedi információ, egyedi feltételes információ, egyedi kölcsönös információ és egyedi feltételes kölcsönös információ. Az első sor várható értékét véve észrevehetjük, hogy éppen az entrópiát kapjuk:

$$E(h_X(X)) = -\sum_x P(X=x)\log P(X=x) = H(X).$$

A többi sorban várható értéket véve, a kapott mennyiségek:

1) $H(X|Y)$: Az X feltételes entrópiája Y-ra nézve. Erre $H(X|Y) = \sum_y P(Y=y)H(X|Y=y)$, ahol $H(X|Y=y)$ az X változó $Y=y$ feltétel melletti feltételes eloszlásának entrópiája. Nyilván $H(X|Y)$ is nemnegatív, és a $h_{X,Y}(x,y) = h_X(x) + h_{Y|X}(y|x)$ összefüggésből $H(X,Y) = H(X) + H(Y|X)$ adódik. Hasonlóan kapható, hogy $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$. Általában pedig a következő láncszabály érvényes:

3.2. Lemma.

$$H(X_1,\ldots,X_k) = \sum_{i=1}^{k} H(X_i|X_1,\ldots,X_{i-1}),$$

és ez akkor is érvényes, ha minden entrópia feltételébe még egy Y változót is beleírunk.

2) $I(X\wedge Y)$: az X és Y kölcsönös információja. Azt fejezi ki, hogy X ismerete mennyivel csökkenti az Y változó entrópiáját, illetve fordítva. A log-összeg egyenlőtlenség szerint

$$I(X\wedge Y) = \sum_{x,y} P(X=x,Y=y)\log\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} \geq 0,$$

azaz a kölcsönös információ nemnegatív, bár adott x,y-ra $h_{X,Y}(x\wedge y)$ negatív is lehet. Az $I(X\wedge Y)=0$ feltétele az, hogy X és Y független legyen. Továbbá, mivel $h_{X,Y}(x\wedge y) = h_X(x) - h_{X|Y}(x|y)$, így

$$I(X\wedge Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y).$$

Ebből adódik, hogy $0 \leq H(X|Y) \leq H(X)$. Az első helyen csak akkor van egyenlőség, ha X az Y függvénye, a második helyen pedig csak akkor, ha X és Y függetlenek.

3) $I(X \wedge Y|Z)$: az X és Y feltételes kölcsönös információja Z-re nézve. Mivel

$$I(X \wedge Y|Z) = \sum_z P(Z = z) I(X \wedge Y|Z = z),$$

így a feltételes kölcsönös információ is nemnegatív, és csak akkor lehet nulla, ha X és Y feltételesen független Z-re. Megint csak a h mennyiségek közötti összefüggésből adódik, hogy $I(X \wedge Y|Z) = H(X|Z) - H(X|Y,Z)$. Következésképpen $H(X|Z) \geq H(X|Y,Z)$.

A kölcsönös információra is létezik láncszabály:

3.3. Lemma.

$$I(X_1, \ldots, X_k \wedge Y) = \sum_{i=1}^{k} I(X_i \wedge Y|X_1, \ldots, X_{i-1}),$$

és ez akkor is érvényes, ha minden kölcsönös információ feltételébe még egy Z változót is beleírunk.

Bizonyítás. Az $I(X_1, \ldots, X_k \wedge Y) = H(X_1, \ldots, X_k) - H(X_1, \ldots, X_k|Y)$ felírás tagjaira alkalmazzuk a 3.2. Lemmát. ■

A fenti információelméleti mennyiségek közötti számos összefüggés leolvasható egy analóg Venn-diagrammról. Feleltessünk meg minden valószínűségi változónak egy halmazt, a változó entrópiáját a halmaz területe fejezi ki. Ekkor a $H(X,Y)$ együttes entrópia az $X \cup Y$ halmaz területe, a $H(X|Y)$ feltételes entrópia az $X \setminus Y$ halmaz területe, az $I(X \wedge Y)$ kölcsönös információ pedig az $X \cap Y$ halmaz területének felel meg. Független valószínűségi változókhoz diszjunkt halmazok tartoznak, ha pedig $Y = f(X)$, akkor a halmazokra $Y \subseteq X$ teljesül.

3.4. Példa. Egy tanfolyamra nyolc gyerek jár, akiket A és B csoportokba osztottak be. Az A csoportba három fiú és egy lány, a B-be egy fiú és három lány került. Véletlenszerűen kiválasztva két gyereket, jelölje X, hogy hány fiút választottunk, Y pedig azt, hogy hány A csoportost.
Az eloszlásokat kiszámolva kapjuk, hogy $H(X) = H(Y) = 1.414$, $H(X,Y) = 2.606$, valamint $H(X|Y) = H(Y|X) = 1.192$, $I(X \wedge Y) = 0.222$. X optimális kódjára $E(L) = 1.429$, a kód hatásfoka 98.95%. Ha az (X,Y) párt ezzel a kóddal betűnként kódoljuk, akkor $E(L) = 2.858$, ennek a kódnak hatásfoka 91.18%. Ha a párt együttesen kódoljuk, akkor a Huffman kódra $E(L) = 2.643$, hatásfoka 98.60%. A kód:

(X,Y)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
esély	3/28	3/28	0	3/28	10/28	3/28	0	3/28	3/28
kódszó	010	011	–	100	00	101	–	110	111

3.5. Definíció. *Jelöljön egy forrást $\mathbb{X}$. A forrás betűnkénti entrópiája*

$$H(\mathbb{X}) = \lim_{k \to \infty} \frac{1}{k} H(X_1, \ldots, X_k),$$

ha ez a határérték létezik.

Azt mondjuk, hogy $\mathbb{X}$ stacionárius forrás, ha minden k-ra és n-re

$$\mathcal{L}(X_1, \ldots, X_k) = \mathcal{L}(X_{n+1}, \ldots, X_{n+k}),$$

ahol $\mathcal{L}$ az eloszlást jelöli.

3.6. Tétel. *Ha $\mathbb{X}$ stacionáris forrás, akkor létezik a betűnkénti entrópiája, és*

$$H(\mathbb{X}) = \lim_{k \to \infty} H(X_k|X_1, \ldots, X_{k-1}).$$

Bizonyítás. Jelölje $H_j = H(X_j|X_1,\ldots,X_{j-1})$, ekkor a láncszabály szerint $H(X_1,\ldots,X_k) = \sum_{j=1}^{k} H_j$. Elég tehát belátni, hogy H_j konvergens. A stacionaritás miatt

$$H_{j-1} = H(X_j|X_2,\ldots,X_{j-1}) \geq H(X_j|X_1,\ldots,X_{j-1}) = H_j.$$

Így a H_j sorozat monoton fogyó, és mivel nemnegatív, van határértéke. ■

A fenti bizonyításból az is látszik, hogy az $\frac{1}{k}H(X_1,\ldots,X_k)$ sorozat monoton csökken. Legyen tehát $\mathbb{X}$ stacionárius forrás, és $g_k : \mathcal{X}^k \to \{0,1\}^*$ a k elemű blokkok valamely prefix kódja. Ekkor az 1 forrásjelre eső kódbitek átlagos száma $E(L_k)/k$, melyre Shannon tétele szerint

$$\frac{E(L_k)}{k} \geq \frac{H(X_1,\ldots,X_k)}{k},$$

továbbá van olyan g_k kód, melyre

$$\frac{E(L_k)}{k} < \frac{H(X_1,\ldots,X_k)+1}{k}.$$

Tehát megkaptuk, hogy az 1 forrásjelre eső kódbitek átlagos száma legalább a forrás betűnkénti entrópiája, és ez az alsó korlát tetszőlegesen megközelíthető elég nagy k választásával.

Az X, Y, Z hármast (ebben a sorrendben) Markov láncnak hívjuk, ha X és Z feltételesen független Y-ra. Ez azzal ekvivalens, hogy $p(z|x,y) = p(z|y)$. Erre az $X \to Y \to Z$ jelölést fogjuk használni.

3.7. Tétel. *(Adatfeldolgozási egyenlőtlenség) Ha $X \to Y \to Z$, akkor $I(X \wedge Y) \geq I(X \wedge Z)$.*

Bizonyítás.

$$I(X \wedge Y, Z) = I(X \wedge Y) + I(X \wedge Z|Y) = I(X \wedge Z) + I(X \wedge Y|Z).$$

Az egyenlőtlenség abból következik, hogy $I(X \wedge Z|Y) = 0$ és $I(X \wedge Y|Z) \geq 0$. ■

Az állítást átírva kapjuk, hogy $H(X|Y) \leq H(X|Z)$. Speciálisan, ha $Z = f(Y)$, akkor $X \to Y \to Z$. Ezért kapjuk, hogy $I(X \wedge Y) \geq I(X \wedge f(Y))$, vagy másképp $H(X|Y) \leq H(X|f(Y))$. A bizonyításbeli egyenlőséget átírva kapjuk azt is, hogy Markov láncra

$$I(X \wedge Y) - I(X \wedge Y|Z) = I(X \wedge Z) \geq 0.$$

Ez az összefüggés nem mindig igaz, pedig a Venn diagramm alapján azt gondolhatnánk. Ha például X, Y független érmedobások eredményei, és $Z = X + Y$, akkor $I(X \wedge Y) = 0$, de

$$I(X \wedge Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) - 0 = 0.5,$$

hiszen

$$H(X|Z) = H(X|Z=0)\frac{1}{4} + H(X|Z=1)\frac{1}{2} + H(X|Z=2)\frac{1}{4} = 0 + 1 \cdot \frac{1}{2} + 0 = 0.5.$$

Az viszont mindig igaz, hogy

$$I(X \wedge Y) - I(X \wedge Y|Z) = I(X \wedge Z) - I(X \wedge Z|Y) = I(Z \wedge Y) - I(Z \wedge Y|X).$$

4. Állandó hosszúságú kódolás hibával

Ha egy forrás által kibocsátott üzenetet kódolás után egy zajos csatornán kell továbbítanunk, akkor a dekódolási hibákat nem tudjuk teljes mértékig kiküszöbölni. Ezért nem nagy veszteség, ha egy kis hibavalószínűséget már a kódolásnál megengedünk. Az állandó hosszúságú kódoknak az az előnyük, hogy egyszerűbb a konstruálásuk és a dekódolásuk is. Egy változó hosszúságú prefix kódnál ellenben megtörténhet, hogy egyetlen jel eltorzulása miatt egy hosszú szakasz dekódolása lehetetlenné válik, vagy elromlik. Az ilyen kódok tanulmányozásához hasznos az alábbi. Ha $p \in [0,1]$, akkor jelölje $H(p)$ a $P = (p, 1-p)$ eloszlás entrópiáját.

4.1. Lemma. *(Fano egyenlőtlenség) Legyen $X \to Y \to Z$, és $P_e = P(Z \neq X)$, jelölje $\mathcal{X}$ az X értékkészletét. Ekkor*

$$H(X|Y) \leq H(X|Z) \leq H(P_e) + P_e \log |\mathcal{X}|.$$

Bizonyítás. Az első egyenlőtlenséget már korábban bizonyítottuk. A másodikhoz legyen $E = \chi(X \neq Z)$. Ekkor

$$H(E, X|Z) = H(X|Z) + H(E|X, Z) = H(E|Z) + H(X|E, Z).$$

Egyrészt $H(E|X, Z) = 0$, másrészt $H(E|Z) \leq H(E) = H(P_e)$, valamint

$$H(X|E, Z) = (1 - P_e)H(X|Z, E = 0) + P_e H(X|Z, E = 1) \leq 0 + P_e \log |\mathcal{X}|.$$

■

4.2. Tétel. *Legyen $g : \mathcal{X}^n \to \mathcal{M}$ az $X^n = (X_1, \ldots, X_n)$ sorozat olyan kódja, melyet ε hibavalószínűséggel lehet dekódolni. Ekkor*

$$\frac{1}{n} \log |\mathcal{M}| \geq \frac{1}{n} H(X_1, \ldots, X_n) - \varepsilon \log |\mathcal{X}| - \frac{1}{n}.$$

Bizonyítás. Jelölje a dekódolót f, erre tehát

$$P(f(g(X^n)) \neq X^n) \leq \varepsilon.$$

A láncszabály szerint

$$H(X^n) = H(g(X^n), X^n) = H(g(X^n)) + H(X^n|g(X^n)).$$

Az első tagot a $\log |\mathcal{M}|$ mennyiséggel becsülhetjük felülről, míg a másodikra

$$H(X^n|g(X^n)) \leq H(X^n|f(g(X^n))) \leq \varepsilon \log |\mathcal{X}^n| + H(\varepsilon) \leq \varepsilon \log |\mathcal{X}^n| + 1$$

a Fano-egyenlőtlenség szerint. ■

Megint koncentráljunk arra az esetre, amikor a forrást bináris sorozatokkal kódoljuk, azaz a g kódra $g : \mathcal{X}^n \to \{0, 1\}^k$. Ekkor $k/n = R$ a jelsebesség vagy ráta, ennyi bit szükséges egy forrásjel kódolásához. A fenti tételből ekkor azt szűrhetjük le, hogy ha az $\mathbb{X}$ stacionárius forrásnak van egy R_n rátasorozattal rendelkező g_n kódsorozata, melyek ε hibával dekódolhatók, akkor

$$\frac{1}{n} \log |\mathcal{M}| = \frac{1}{n} \log 2^k = \frac{k}{n} = R_n$$

miatt

$$\liminf R_n \geq H(\mathbb{X}) - \varepsilon \log |\mathcal{X}|.$$

Ennél többet mondhatunk információstabilis forrásokra.

4.3. Definíció. *Legyen $\mathbb{X}$ stacionárius forrás, a véges dimenziós eloszlásokat egy közös p szimbólummal jelöljük, azaz*

$$p(x^n) = p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

A forrás információstabilis (IS), ha

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \to H(\mathbb{X}) \quad (n \to \infty)$$

sztochasztikusan.

Később látni fogjuk, hogy a gyakorlatban előforduló források rendelkeznek ezzel a tulajdonsággal. A legegyszerűbb eset az emlékezet nélküli stacionárius forrás, amikor az X_i-k függetlenek és azonos eloszlásúak. Ekkor

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} -\log p(X_i) \to E(-\log p(X_1)) = H(X_1) = H(\mathbb{X})$$

a nagy számok törvénye szerint.

4.4. Definíció. *Adott stacionárius forráshoz, ε-hoz és n-hez definiáljuk a* tipikus *jelsorozatok halmazát:*

$$A_\varepsilon^{(n)} = \{x^n \in \mathcal{X}^n : 2^{-n(H(\mathbb{X})+\varepsilon)} \le p(x^n) \le 2^{-n(H(\mathbb{X})-\varepsilon)}\}.$$

4.5. Tétel. *Legyen $\mathbb{X}$ IS forrás. Ekkor minden $\varepsilon > 0$ esetén*
(i) $P(A_\varepsilon^{(n)}) \to 1$, ha $n \to \infty$.
(ii) $|A_\varepsilon^{(n)}| \le 2^{n(H(\mathbb{X})+\varepsilon)}$.
(iii) Minden $\delta > 0$ esetén $|A_\varepsilon^{(n)}| \ge (1-\delta)2^{n(H(\mathbb{X})-\varepsilon)}$, ha n elég nagy.

Bizonyítás. Az (i) rész az információstabilitás definíciójának átfogalmazása. Az (ii) rész a tipikus sorozatok valószínűségre adott alsó korlátból következik:

$$1 \ge P(A_\varepsilon^{(n)}) = \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) \ge \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(\mathbb{X})+\varepsilon)} = 2^{-n(H(\mathbb{X})+\varepsilon)} |A_\varepsilon^{(n)}|.$$

A (iii) rész pedig a tipikus sorozatok valószínűségre adott felső korlátból következik (ii)-hez hasonlóan, felhasználva, hogy elég nagy n-re a tipikus halmaz valószínűsége már legalább $(1-\delta)$. ■

A fenti (i) angol neve „asymptotic equipartition property." Ez azt fejezi ki, hogy nagy n-re találunk egy közel 1 valószínűségű halmazt úgy, hogy a benne lévő jelsorozatok majdnem egyforma valószínűségűek.

Térjünk vissza az állandó hosszúságú kódokra. A legkisebb hibájú kódot úgy kapjuk, ha a lehetséges 2^k darab kódszót a legnagyobb valószínűségű x^n sorozatok között osztjuk ki (a többi sorozat kódja pedig tetszőleges). Ha $N(n,\varepsilon)$ jelöli azt, hogy hány x^n sorozatot kell összegyűjteni ahhoz, hogy együttes valószínűségük legalább $1-\varepsilon$ legyen, akkor az optimális ε-hibájú g_n kód rátája

$$R_n = \frac{\lceil \log N(n,\varepsilon) \rceil}{n}.$$

4.6. Tétel. *Ha $\mathbb{X}$ IS, akkor minden pozitív ε-ra $\log N(n,\varepsilon)/n \to H(\mathbb{X})$.*

Bizonyítás. A tipikus halmaz becsempészésével bizonyítunk. Legyen $0 < \delta < \min(\varepsilon, 1-\varepsilon)$ tetszőleges. Egyrészt tudjuk, hogy elég nagy n-re az $A_\delta^{(n)}$ halmaz valószínűsége már legalább $1-\varepsilon$. Ezért nagy n-re

$$N(n,\varepsilon) \le |A_\delta^{(n)}| \le 2^{n(H(\mathbb{X})+\delta)},$$

amiből

$$\limsup \frac{\log N(n,\varepsilon)}{n} \le H(\mathbb{X}).$$

A másik irányhoz legyen $B_\varepsilon^{(n)}$ a legvalószínűbb $N(n,\varepsilon)$ darab sorozat halmaza. Ekkor

$$P(A_\delta^{(n)} \cap B_\varepsilon^{(n)}) \ge 1 - \varepsilon - \delta,$$

ha n elég nagy. A tipikus sorozatok valószínűségére ismert felső korlát miatt

$$N(n,\varepsilon) \ge |A_\delta^{(n)} \cap B_\varepsilon^{(n)}| \ge (1-\varepsilon-\delta)2^{n(H(\mathbb{X})-\delta)},$$

amiből

$$\liminf \frac{\log N(n,\varepsilon)}{n} \ge H(\mathbb{X}).$$

■

Azt kaptuk tehát, hogy az optimális R_n jelsebességet aszimptotikusan a $H(\mathbb{X})$ entrópia adja meg.

4.7. Definíció. *Legyen $\mathbb{X}$ egy forrás. Egy $g : \mathcal{X}^n \to \{0,1\}^{\lceil nR \rceil}$ kódot (R,n)-kódnak nevezünk. Ha megadunk egy $f : \{0,1\}^{\lceil nR \rceil} \to \mathcal{X}^n$ dekódolót is, akkor a kódoló-dekódoló pár hibája*

$$P_e^{(n)} = P(f(g(X^n)) \ne X^n).$$

Az R rátát elérhetőnek nevezzük, ha létezik olyan (R,n)-kódok sorozata, melyre $\lim_{n\to\infty} P_e^{(n)} = 0$.

A fentiekben beláttuk, hogy ha $\mathbb{X}$ IS forrás, akkor minden $R > H(\mathbb{X})$ rátára létezik tetszőlegesen kicsi, de fix $\varepsilon > 0$ hibavalószínűségű kódsorozat. Most belátjuk, hogy nullához tartó hibavalószínűség is elérhető. Az állítást a véletlen választás módszerével bizonyítjuk, azaz nem adjuk meg a jó kódsorozatot, csak azt látjuk be, hogy léteznie kell. Ez a módszer gyakran működik az információelméletben. Válasszunk egy $g : \mathcal{X}^n \to \{0,1\}^{\lceil nR \rceil}$ kódot véletlenszerűen, azaz legyen minden $g(x^n)$ kódszó, egymástól függetlenül, egyenletes eloszlású a $\{0,1\}^{\lceil nR \rceil}$ halmazban. Dekódolásra pedig használjuk a *tipikussági dekódolót*: ha a $z^{\lceil nR \rceil} \in \{0,1\}^{\lceil nR \rceil}$ sorozathoz egyetlen olyan x^n tipikus sorozat van, hogy $g(x^n) = z^{\lceil nR \rceil}$, akkor legyen $f(z^{\lceil nR \rceil}) = x^n$, ellenkező esetben (ha nincs ilyen tipikus sorozat, vagy egynél több van), legyen mondjuk $f(z^{\lceil nR \rceil}) = 0^n$ (feltesszük, hogy $0 \in \mathcal{X}$). A tipikusságnál használjunk olyan ε-t, melyre $\varepsilon < R - H(\mathbb{X})$. Rögzített g esetén a hibávalószínűség:

$$P_e^{(n)}(g) = P_g(f(g(X^n)) \neq X^n) = \\ P(X^n \notin A_\varepsilon^{(n)}) + \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)\chi(\exists y^n \neq x^n, y^n \in A_\varepsilon^{(n)} : g(y^n) = g(x^n)).$$

Vegyünk most várható értéket a g választása szerint:

$$E(P_e^{(n)}(g)) = P\left(\overline{A_\varepsilon^{(n)}}\right) + \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) P(\exists y^n \neq x^n, y^n \in A_\varepsilon^{(n)} : g(y^n) = g(x^n)) \leq \\ P\left(\overline{A_\varepsilon^{(n)}}\right) + \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) \sum_{y^n \neq x^n, y^n \in A_\varepsilon^{(n)}} P(g(y^n) = g(x^n)).$$

A konstrukció szerint $P(g(y^n) = g(x^n)) = 2^{-\lceil nR \rceil}$, így

$$E(P_e^{(n)}(g)) \leq P\left(\overline{A_\varepsilon^{(n)}}\right) + 2^{-\lceil nR \rceil} |A_\varepsilon^{(n)}| P(A_\varepsilon^{(n)}).$$

A 4.5 Tétel (i) és (ii) pontja szerint a jobboldal nullához tart. Végül arra hivatkozunk, hogy ha a véletlenül választott kódra a várható hiba nullához tart, akkor van olyan kódsorozat, melyre a hiba nullához tart.

Fontos észrevétel, hogy a g kódról csak azt használtuk ki, hogy az $x^n \neq y^n$ sorozatokra $P(g(x^n) = g(y^n)) \leq 2^{-\lceil nR \rceil}$. Ezáltal belátható, hogy $\mathcal{X} = \{0,1\}$ esetben az R ráta *lineáris* kóddal is elérhető: legyen G egy $\lceil nR \rceil \times n$ méretű $0-1$ mátrix, ez definiálja a $g(x^n) = Gx^n$ kódot, ahol a műveleteket a kételemű véges testben végezzük. Ha most G minden elemét egymástól függetlenül, $1/2 - 1/2$ valószínűséggel választjuk 0-nak vagy 1-nek, akkor $x^n \neq y^n$ esetén

$$P(g(x^n) = g(y^n)) = P(G(x^n - y^n) = 0^{\lceil nR \rceil}) = \prod_{i=1}^{\lceil nR \rceil} P(G_i(x^n - y^n) = 0) = 2^{-\lceil nR \rceil},$$

ahol G_i a G mátrix i-edik sora. A lineáris kódok előnye, hogy nem kell egy hatalmas kódszótárat tárolni, csak a kis G mátrixot. Természetesen a lineáris kódok nagyobb véges testek felett is megadhatók.

Számoljunk most egy kicsit. Nézzük meg, hogy ha az egyre hosszabb X^n blokkokat mindig egy rögzített R jelsebességgel kódoljuk, akkor a hibavalószínűség hogyan viselkedik. Emlékezet nélküli forrásra belátjuk, hogy ha $R > H(X_1)$, akkor a hiba exponenciálisan nullához tart.

4.8. Tétel. *Legyen $\mathbb{X}$ emlékezet nélküli, stacionárius forrás, melyre az x jel valószínűsége $P(X_i = x) = p(x)$, és legyen $R > H(X_1)$ rögzített ráta. Jelölje $P_e^{(n)}$ az optimális (legkisebb hibavalószínűségű) (R,n)-kód hibavalószínűségét. Ekkor*

$$P_e^{(n)} \leq 2^{-ne(R)},$$

ahol az $e(R)$ hibaexponens az alábbi:

$$e(R) = \sup_{0<a<1} \frac{1-a}{a}\left(R - \frac{1}{1-a}\log \sum_x p^a(x)\right).$$

Bizonyítás. Jelölje B a $2^{\lceil nR\rceil}$ darab legvalószínűbb x^n sorozat halmazát Ekkor $P_e^{(n)} = 1 - P(B)$. Jelölje még b^n a B halmaz legkisebb valószínűségű elemét. A sorozatok valószínűségét most is a közös p szimbólummal jelöljük. Ekkor minden $0 < a < 1$-re

$$P_e^{(n)} = \sum_{x^n \notin B} p(x^n) \leq \sum_{x^n \notin B} p(x^n) \left(\frac{p(b^n)}{p(x^n)}\right)^{1-a} =$$

$$= p(b^n)^{1-a} \sum_{x^n \notin B} p(x^n)^a \leq p(b^n)^{1-a} \sum_{x^n} p(x^n)^a.$$

Mármost

$$p(b^n)^a \leq 2^{-\lceil nR\rceil} \sum_{x^n \in B} p(x^n)^a \leq 2^{-\lceil nR\rceil} \sum_{x^n} p(x^n)^a.$$

Ezért

$$P_e^{(n)} \leq \left(2^{-\lceil nR\rceil}\right)^{\frac{1-a}{a}} \left(\sum_{x^n} p(x^n)^a\right)^{\frac{1}{a}}$$

Mivel a függetlenség és azonos eloszlás miatt $p(x^n) = \prod_{i=1}^n p(x_i)$, ezért

$$\sum_{x^n} p(x^n)^a = \left(\sum_x p(x)^a\right)^n.$$

Visszahelyettesítve,

$$P_e^{(n)} \leq \left(2^{-\lceil nR\rceil}\right)^{\frac{1-a}{a}} \left(\sum_x p(x)^a\right)^{\frac{n}{a}} \leq 2^{-n\frac{1-a}{a}\left(R - \frac{1}{1-a}\log\sum_x p(x)^a\right)}.$$

■

A tételben szereplő $H_a(P) = \frac{1}{1-a}\log\sum_{x\in\mathcal{X}} p^a(x)$ mennyiség a *Rényi-féle entrópia*. Könnyen látszik, hogy ez a Shannon-entrópiára jellemző 6 axióma közül csak a „tovább-bontási" axiómát nem elégíti ki, a másik ötöt igen. Belátjuk, hogy rögzített eloszlás esetén a Rényi-entrópia az a monoton fogyó függvénye. Ehhez legyen $\phi(a) = \log\sum p^a(x)$, először belátjuk, hogy ez konvex függvény. Definíció szerint ez azt jelenti, hogy két függvényérték konvex kombinációja nagyobb vagy egyenlő, mint a függvény értéke a két argumentum konvex kombinációjában. Legyen $0 < c < 1$.

$$c\phi(a) + (1-c)\phi(b) = c\log\sum p^a(x) + (1-c)\log\sum p^b(x) = \log\left[\left(\sum p^a(x)\right)^c \left(\sum p^b(x)\right)^{1-c}\right].$$

A Hölder-egyenlőtlenség szerint

$$\left(\sum p^a(x)\right)^c \left(\sum p^b(x)\right)^{1-c} \geq \sum (p^a(x))^c (p^b(x))^{1-c} = \sum p(x)^{ca+(1-c)b},$$

azaz a logaritmus monoton növekedése miatt

$$c\phi(a) + (1-c)\phi(b) \geq \phi(ca + (1-c)b).$$

Nézzük a függvény egy szelőjének meredekségét:

$$\frac{\phi(a) - \phi(1)}{a-1} = \frac{\log\sum p^a(x) - 0}{a-1} = -H_a(P).$$

Mivel ϕ konvex, ezeknek a szelőknek a meredeksége monoton nő, amint a egyhez tart (balról). Az előzőből az is látszik, hogy $\phi'(1) = \lim_{a\to 1} -H_a(P)$. Rövid számolással kapjuk, hogy

$$\phi'(a) = \frac{1}{\ln 2}\frac{1}{\sum p^a(x)}\sum p^a(x)\ln p(x) = \frac{\sum p^a(x)\log p(x)}{\sum p^a(x)},$$

amiből $\phi'(1) = -H(P)$, azaz $\lim_{a\to 1} H_a(P) = H(P)$. Tehát a $H_a(P)$ Rényi entrópia a $H_0(P)$ Hartley-entrópiát köti össze monoton csökkenően a $H_1(P)$ Shannon-entrópiával.

A tételben szereplő hibaexponenst tovább vizsgálva (a Rényi entrópiák argumentumából a fix P eloszlást elhagyva), minden R-re igaz, hogy

$$e(R) \geq \lim_{a\to 1} \frac{1-a}{a}(R - H_a) = 0.$$

Az is triviális, hogy $e(R)$ monoton növő. Továbbá, ha $R > H$, akkor van olyan 1-hez közeli a is, hogy $R > H_a$, azaz $e(R) > 0$. Végül, ha $R \leq H$, akkor $R \leq H_a$ is igaz minden a-ra, így $e(R) = 0$.

5. Információstabilis források

Az emlékezet nélküli stacionárius forrás triviálisan IS a nagy számok gyenge törvénye szerint. A gyakorlatban előforduló források persze a legritkább esetben ilyenek. Viszont sok forrás jól közelíthető k-adrendű stacionárius Markov lánccal. Vezessük be egy vektor részvektorára a következő jelölést: ha $v = (v_1, \ldots, v_m)$ és $1 \leq a \leq b \leq m$, akkor $v_a^b = (v_a, \ldots, v_b)$ és $v^b = v_1^b$.

5.1. Definíció. *Az $X_1, X_2, \ldots$ folyamat k-adrendű Markov lánc, ha minden n-re és x^{n+1}-re, melyre a következő feltételes valószínűségek értelmesek, teljesül, hogy*

$$P(X_{n+1} = x_{n+1}|X^n = x^n) = P(X_{n+1} = x_{n+1}|X_{n-k+1}^n = x_{n-k+1}^n) = r(x_{n-k+1}^n, x_{n+1}),$$

ahol $r(x^k, y)$ jelöli azt az átmenetvalószínűséget, hogy az x^k sorozat után a következő jel y lesz.

Legyen $\mathbb{X}$ k-adrendű Markov lánc. Ekkor az $Y_n = X_n^{n+k-1}$ folyamat (az állapotokat k-asával összefogjuk) elsőrendű Markov lánc, melyről feltesszük, hogy irreducibilis az $\mathcal{I}$ állapottéren, mely esetleg szűkebb az $\mathcal{X}^k$ halmaznál, azaz tetszőleges $x^k \in \mathcal{I}$ kiindulási sorozat és tetszőleges $y^k \in \mathcal{I}$ sorozat esetén pozitív a valószínűsége, hogy az y^k sorozat előbb-utóbb felbukkan. Gondoljunk például a magyar nyelvre, ha ezt másodrendű Markov lánccal szeretnénk modellezni, akkor például az „xq" párt nem tesszük bele az állapottérbe, mert ez a két betű soha nem jön egymás után. $\mathcal{I}$ elemei tehát a „megengedett" sorozatok. Ismert tétel, hogy véges állapotterű, irreducibilis Markov láncnak egyértelműen létezik stacionárius eloszlása, mely szigorúan pozitív. Tehát az $\mathcal{I} \subseteq \mathcal{X}^k$ állapottéren egyértelműen megadható egy $\pi(x^k)$ eloszlás, és ha $P(X^k = x^k) = \pi(x^k)$ a forrás első k jelének az eloszlása, akkor $\mathbb{X}$ stacionárius folyamat lesz. Belátjuk, hogy ekkor $\mathbb{X}$ IS. Először is, erre a forrásra

$$H(\mathbb{X}) = \lim_{n\to\infty} H(X_{n+1}|X_1, \ldots, X_n) = H(X_{k+1}|X_1, \ldots, X_k) = E[-\log p(X_{k+1}|X_1^k)].$$

Másrészt

$$-\log p(X_1^n) = -\log p(X_1^k) + \sum_{i=k}^{n-1} -\log p(X_{i+1}|X_{i-k+1}^i) = -\log p(X_1^k) + \sum_{i=k}^{n-1} f(X_{i-k+1}^{i+1}),$$

ahol $f(x^{k+1}) = -\log p(x_{k+1}|x_1^k)$. Vegyük észre, hogy az $Y_n = X_n^{n+k}$ folyamat (az állapotokat $k+1$-esével összefogjuk) elsőrendű, stacionárius Markov lánc, és irreducibilis az

$$\mathcal{I}' = \{x^{k+1} : \ x^k \in \mathcal{I}, r(x^k, x_{k+1}) > 0\} \subseteq \mathcal{I} \times \mathcal{X}$$

állapottéren. Ezzel

$$-\frac{1}{n}\log p(X_1^n) = \frac{-\log p(X_1^k)}{n} + \frac{1}{n}\sum_{i=k}^{n-1} f(Y_{i-k+1}).$$

A jobboldalon az első tag 1 valószínűséggel nullához tart, míg a második tag a (véges állapotterű, irreducibilis) Markov láncokra vonatkozó nagy számok törvénye szerint az

$$E(f(Y_1)) = E[-\log p(X_{k+1}|X_1^k)] = H(\mathbb{X})$$

entrópiához tart, tehát a forrás információstabilis.

Egy lépéssel tovább menve, most tekintsük egy – az egyszerűség kedvéért – elsőrendű Markov forrás függvényét: legyen $\mathbb{X}$ (irreducibilis, stacionárius) Markov forrás, ϕ egy (esetleg véletlen) függvény a Markov lánc állapotterén, és $Y_n = \phi(X_n)$. Ismert, hogy ekkor $\mathbb{Y} = (Y_1, Y_2, \ldots)$ nem feltétlenül Markov folyamat, viszont az $\mathbb{Y}$ forrás nyilván stacionárius. Hogyan számolható ki ezen forrás betűnkénti entrópiája?

5.2. Tétel. *Minden n-re teljesül a*

$$H(Y_n|Y_2^{n-1}, X_1) \le H(\mathbb{Y}) \le H(Y_n|Y_1^{n-1})$$

egyenlőtlenség, valamint a két korlát különbsége nullához tart.

Bizonyítás. A felső becslés már ismert. Az alsó becslés:

$$H(Y_n|Y_2^{n-1}, X_1) = H(Y_n|Y_2^{n-1}, X_1, Y_{-k}^1) \le H(Y_n|Y_{-k}^{n-1}) \to H(\mathbb{Y}) \quad k \to \infty,$$

ahol az első egyenlőség azért teljesül, mert Y_n és Y_{-k}^1 feltételesen függetlenek X_1, Y_2^{n-1}-re nézve. A kényelem kedvéért, hogy ne kelljen a változók indexelését eltolni, feltettük, hogy $\mathbb{X}$ (és így $\mathbb{Y}$ is) kétirányban végtelen folyamatok, ez nyilván feltehető. Nézzük most a két korlát különbségét! Ismét a feltételes függetlenséget használva, $H(Y_n|Y_2^{n-1}, X_1) = H(Y_n|Y_1^{n-1}, X_1)$, és ezért a két korlát különbsége

$$H(Y_n|Y_1^{n-1}) - H(Y_n|Y_1^{n-1}, X_1) = I(X_1 \wedge Y_n|Y_1^{n-1}).$$

A láncszabályt alkalmazva, minden m-re

$$\sum_{n=1}^{m} I(X_1 \wedge Y_n|Y_1^{n-1}) = I(X_1 \wedge Y_1^m) \le H(X_1).$$

Ezért $\sum_{n=1}^{\infty} I(X_1 \wedge Y_n|Y_1^{n-1}) < \infty$, tehát a tagok nullához tartanak. ■

A következő kérdés, hogy $\mathbb{Y}$ IS-e? Egy nagyon általános tétel szerint igen. A Shannon-McMillan-Breiman tétel kimondja, hogy minden ergodikus forrás IS. Az ergodikus folyamatokra gondolhatunk úgy, mint a legáltalánosabb folyamatokra, melyekre a nagy számok erős törvénye teljesül. A pontos definíció az eseménytéren megadott stacionárius, ergodikus transzformáció segítségével történik.

5.3. Definíció. *Legyen $(\Omega, \mathcal{A}, P)$ valószínűségi mező. Az $T : \Omega \to \Omega$ (mérhető) transzformáció stacionárius ergodikus, ha minden $A \in \mathcal{A}$ eseményre $P(TA) = P(A)$, és $P(A \triangle TA) = 0$ esetén $P(A)$ nulla vagy egy.*

5.4. Definíció. *Az $X_0, X_1, \ldots$ folyamat stacionárius ergodikus, ha létezik az Ω eseménytéren X valószínűségi változó és T stacionárius ergodikus transzformáció, hogy $X_n(\omega) = X(T^n\omega)$. Ekkor persze a folyamat negatív indexekre is kiterjeszthető: $X_{-n}(\omega) = X(T^{-n}\omega)$.*

A Birkhoff-féle ergodtétel szerint ha $X_1, X_2, \ldots$ stacionárius ergodikus folyamat, akkor $\frac{1}{n}\sum_1^n X_i$ egy valószínűséggel konvergál az $E(X_i)$ várható értékhez (amennyiben ez a várható érték véges). Ha a folyamat értékkészlete véges, akkor az ergodikusság azzal ekvivalens, hogy minden k-ra a k-dimenziós tapasztalati eloszlás tart az elméletihez. Ha $\mathbb{X}$ kétirányban végtelen stacionárius, ergodikus folyamat, akkor az $Y_n = f(X_{-\infty}^n)$ $(n = 1, 2, \ldots)$ folyamat is az. Ugyanis $Y_n(\omega) = Y(T^n\omega)$, ahol $Y(\omega) = f(X(\omega), X(T^{-1}\omega), \ldots)$.

A Shannon-McMillan-Breiman tétel a következő.

5.5. Tétel. *Ha $\mathbb{X} = (X_1, X_2, \ldots)$ véges értékkészletű, stacionárius ergodikus folyamat, akkor*

$$-\frac{1}{n}\log p(X^n) \to H(\mathbb{X}) \text{ 1 valószínűséggel.}$$

A bizonyítás azon múlik, hogy a

$$-\frac{1}{n}\log p(X^n) = \frac{1}{n}\sum_{i=1}^{n} -\log p(X_i|X^{i-1}) = \frac{1}{n}\sum_{i=1}^{n} -\log Y_i$$

felírásban az Y_i folyamat nem ergodikus, de valamilyen értelemben becsülhető az $U_i = p(X_i|X_{i-k}^{i-1})$ és $V_i = p(X_i|X_{-\infty}^{i-1})$ ergodikus folyamatokkal, melyekre alkalmazható a Birkhoff ergodtétel. A részleteket nem közöljük.

A szakasz végén bevezetünk egy új információelméleti mérőszámot. Tegyük fel, hogy X és Y is az $\mathcal{X}$ halmazon veszik fel értékeiket. Vezessük be a

$$D(X\|Y) = \sum_x P(X=x)\log\frac{P(X=x)}{P(Y=x)}$$

mennyiséget, mely az X és az Y információs divergenciája, vagy Kullback-Leibler divergenciája, vagy relatív entrópiája. Ez is nemnegatív a log-összeg egyenlőtlenség szerint, és csak akkor nulla, ha a két változó eloszlása megegyezik. Vegyük észre, hogy $I(X \wedge Y) = D(X,Y\|X',Y')$, ahol X', illetve Y' ugyanolyan eloszlású, mint X, illetve Y, de függetlenek. Természetesen a divergencia csak a két valószínűségi változó eloszlásától függ, ezért az argumentumba írhatjuk az eloszlásokat is: $D(P\|Q)$.

A divergencia valamilyen értelemben a két eloszlás különbözőségét méri. Tegyük fel például, hogy egy emlékezet nélküli forrás eloszlása P, de mi azt hisszük, hogy az eloszlás $\hat{P}$ (például hosszasan megfigyeltük a forrást, és $\hat{P}$ a tapasztalati eloszlás). Ekkor a $\hat{P}$ szerinti (közel optimális) Shannon kódra $L(x) = \lceil -\log\hat{p}(x)\rceil$. Hasonlítsuk össze ennek a kódnak átlagos hosszát a forrás entrópiájával:

$$E(L)-H(P) = \sum_x p(x)\lceil -\log\hat{p}(x)\rceil + \sum_x p(x)\log p(x) \approx -\sum_x p(x)\log\hat{p}(x) + \sum_x p(x)\log p(x) = D(P\|\hat{P}).$$

Tehát betűnként átlagosan $D(P\|\hat{P})$-vel több bitet használunk, mint ha ismernénk a P eloszlást, és annak megfelelően kódolnánk.

5.6. Lemma. *A divergenciára a következő láncszabály érvényes. Legyen P_{XY}, Q_{XY} két eloszlás az $\mathcal{X}\times\mathcal{Y}$ szorzathalmazon, a marginális eloszlásokat jelölje P_X, Q_X illetve P_Y, Q_Y. Definiáljuk a feltételes eloszlások divergenciáját:*

$$D(P_{Y|X}\|Q_{Y|X}) = \sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)}.$$

Ez is nemnegatív, és

$$D(P_{XY}\|Q_{XY}) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}).$$

Bizonyítás. A nemnegativitás a log-összeg egyenlőtlenségből következik, a második állításhoz pedig csak fel kell írni a két oldalon álló mennyiségeket. A jobboldallal kezdve:

$$D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}) = \sum_x p(x)\log\frac{p(x)}{q(x)} + \sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)} =$$

$$\sum_{x,y} p(x,y)\log\frac{p(x)}{q(x)} + \sum_{x,y} p(x,y)\log\frac{p(y|x)}{q(y|x)} = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} = D(P_{XY}\|Q_{XY}).$$

■

Vegyünk most egy véges állapotterű, irreducibilis $\mathbb{X} = (X_1, X_2, \ldots)$ Markov láncot, jelölje X_n eloszlását P_n, a kezdeti eloszlás tehát P_1. Jelölje továbbá a stacionárius eloszlást π.

5.7. Tétel. *(a) Legyen P_1, Q_1 két kezdeti eloszlás. Ekkor a $d_n = D(P_n\|Q_n)$ sorozat monoton csökken. (b) Tetszőleges kezdeti eloszlásra $D(P_n\|\pi)$ monoton csökken. Továbbá, ha π az egyenletes eloszlás, akkor $H(P_n)$ monoton nő.*

Bizonyítás. (a) A Markov tulajdonság szerint $P_{n+1|n} = Q_{n+1|n}$. A 5.6 Lemma értelmében ezért

$$D(P_{n,n+1}\|Q_{n,n+1}) = D(P_n\|Q_n) + 0 = D(P_{n+1}\|Q_{n+1}) + D(P_{n|n+1}\|Q_{n|n+1}).$$

A divergencia nemnegativitásából következik az állítás.
(b) Az előző következménye, ha $Q_1 = \pi$. Végül az utolsó állítás abból következik, hogy ha π az egyenletes eloszlás, akkor

$$D(P_n\|\pi) = \log|\mathcal{X}| - H(P_n).$$

■

6. A Slepian-Wolf tétel megosztott források kódolására

Most olyan kódolásról lesz szó, amikor egyszerre több helyen keletkeznek kódolandó üzenetek, melyeket a keletkezési helyükön szeretnénk kódolni, majd a kódsorozatokat beküldeni egy központba, ahol a dekódolás történik. Gondoljunk arra, hogy több különböző állomáson végeznek méréseket, megfigyeléseket, melyeket aztán a központban értékelnek ki. Általában az egyes állomások mérései korreláltak. Meghatározandó az elérhető rátatartomány.

Az egyszerűség kedvéért csak két állomással foglalkozunk. Az eredmények általánosítását több állomásra az olvasóra bízzuk. Legyen tehát $(\mathbb{X},\mathbb{Y}) = ((X_1,Y_1),(X_2,Y_2),\ldots)$ stacionárius forrás. Feltehetjük, hogy a forrás emlékezet nélküli, vagy általánosabban, hogy IS (később pontosítjuk, hogy most mit értünk ezalatt).

6.1. Definíció. *(R_1,R_2,n)-kódnak nevezünk egy $g_1:\mathcal{X}^n\to\{0,1\}^{nR_1}$, $g_2:\mathcal{Y}^n\to\{0,1\}^{nR_2}$ párt, ahol $\mathcal{X}$ illetve $\mathcal{Y}$ az $\mathbb{X}$ illetve $\mathbb{Y}$ források véges értékkészletei. Ha megadunk egy $f:\{0,1\}^{nR_1}\times\{0,1\}^{nR_2}\to\mathcal{X}^n\times\mathcal{Y}^n$ dekódolót is, akkor a kódoló-dekódoló hibavalószínűsége*

$$P_e^{(n)} = P(f(g_1(X^n),g_2(Y^n)) \neq (X^n,Y^n)).$$

Az (R_1,R_2) rátapár elérhető, ha létezik olyan (R_1,R_2,n) kódsorozat, hogy $\lim_{n\to\infty} P_e^{(n)} = 0$. Az elérhető rátatartomány az elérhető rátapárok halmazának lezártja.

Az adott esetben pontosan meg tudjuk majd határozni az elérhető rátatartományt, melyet nem meglepő módon az entrópiák határoznak meg. Ha $R_1 > H(\mathbb{X})$ és $R_2 > H(\mathbb{Y})$, akkor az (R_1,R_2) rátapár nyilván elérhető. Ha a két forrást megengedett lenne együtt kódolni, akkor szintén az eddigiekből látszik, hogy az elérhetőséghez $R_1+R_2 \geq H(\mathbb{X},\mathbb{Y})$ szükséges. Látni fogjuk, hogy ez a feltétel nem elégséges, az R_1, R_2 rátáknak külön-külön is elég nagynak kell lenniük. Lássuk akkor a Slepian-Wolf tételt!

6.2. Tétel. *Ha $(\mathbb{X},\mathbb{Y})$ stacionárius, emlékezet nélküli megosztott források, akkor az elérhető rátatartomány:*

$$\mathcal{R} = \{(R_1,R_2):\ R_1 \geq H(X|Y), R_2 \geq H(Y|X), R_1+R_2 \geq H(X,Y)\}.$$

Bizonyítás. Először azt látjuk be, hogy a fenti $\mathcal{R}$ halmaz belső pontjai elérhetők. Ugyanúgy, mint korábban, a g_1, g_2 kódokat válasszuk véletlenül, és használjuk a tipikussági dekódolót. Lényeges, hogy az (x^n,y^n) sorozatpárt most akkor nevezzük (együttesen) tipikusnak, ha

$$\left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \varepsilon, \quad \left|-\frac{1}{n}\log p(y^n) - H(Y)\right| < \varepsilon,$$
$$\left|-\frac{1}{n}\log p(x^n,y^n) - H(X,Y)\right| < \varepsilon. \quad (1)$$

Jelölje ezt a halmazt is $A_\varepsilon^{(n)}$, vagy ha precízebbek akarunk lenni, $A_\varepsilon^{(n)}(X,Y)$. Vizsgáljuk a hibát! A hibás dekódolásnak négy oka lehet: (1) A kódolt (x^n,y^n) sorozatpár nem együttesen tipikus. A többi esetben (x^n,y^n) tipikus, de vagy (2) van olyan $u^n \neq x^n$, hogy (u^n,y^n) tipikus és $g_1(u^n) = g_1(x^n)$, vagy (3) van olyan $v^n \neq y^n$, hogy (x^n,v^n) tipikus és $g_2(v^n) = g_2(y^n)$, vagy (4) van olyan $u^n \neq x^n$ és $v^n \neq y^n$,

hogy (u^n, v^n) tipikus és $g_1(u^n) = g_1(x^n)$ és $g_2(v^n) = g_2(y^n)$. Jelölje a négy lehetőség valószínűségét (az (x^n, y^n) választása valamint a (g_1, g_2) választása szerint) rendre Q_1, Q_2, Q_3, Q_4. Kaptuk, hogy

$$E(P_e^{(n)}(g_1, g_2)) \le Q_1 + Q_2 + Q_3 + Q_4.$$

A nagy számok törvénye szerint $P(A_\varepsilon^{(n)}(X,Y)) \to 1$, így $Q_1 \to 0$. A többi esetben a korábban bemutatott triviális felső becslést alkalmazzuk. Q_4-re:

$$Q_4 \le P(A_\varepsilon^{(n)}) |A_\varepsilon^{(n)}| 2^{-n(R_1+R_2)},$$

hiszen $P(g_1(u^n) = g_1(x^n), g_2(v^n) = g_2(y^n)) = 2^{-n(R_1+R_2)}$. A 4.5 Tétel (ii) pontja szerint $Q_4 \to 0$, ha $\varepsilon < R_1 + R_2 - H(X,Y)$. Q_2 elemzéséhez vezessük be az

$$A_\varepsilon^{(n)}(X|y^n) = \{x^n : (x^n, y^n) \in A_\varepsilon^{(n)}(X,Y)\}$$

jelölést. Ekkor

$$Q_2 = \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n, y^n) P(\exists u^n \ne x^n : \ (u^n, y^n) \in A_\varepsilon^{(n)}, g_1(u^n) = g_1(x^n)) \le$$
$$\sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n, y^n) \sum_{u^n \in A_\varepsilon^{(n)}(X|y^n), u^n \ne x^n} P(g_1(u^n) = g_1(x^n)) \le P(A_\varepsilon^{(n)}) |A_\varepsilon^{(n)}(X|y^n)| 2^{-nR_1},$$

ami a 6.3 Lemma szerint nullához tart, ha $\varepsilon < (R_1 - H(X|Y))/2$. Q_3 hasonlóan nullához tart, amivel az elérhetőség bizonyítását befejeztük.

Korábban láttuk, hogy $R_1 + R_2 \ge H(X,Y)$ az elérhetőség szükséges feltétele. Másrészt megmutatjuk, hogy ha (R_1, R_2) elérhető pár, akkor $(R_1, H(Y) + \delta)$ is az (bármilyen kis pozitív δ-ra), amiből már következik $R_1 \ge H(X,Y) - H(Y) - \delta = H(X|Y) - \delta$. Létezik ugyanis egy $H(Y) + \delta$ rátájú h_2 kódsorozat $\mathbb{Y}$-ra, melynek hibája nullához tart. Ekkor a $(g_1(x^n), h_2(y^n))$ párból valóban nullához tartó hibával visszaállítható (x^n, y^n), hiszen $h_2(y^n)$-ből dekódolással kapjuk $\hat{y}^n$-et, majd azt kódolva g_2 szerint, a $(g_1(x^n), g_2(\hat{y}^n))$ párból a feltétel szerint dekódolhatjuk $(x^n, \hat{y}^n)$-et. ■

6.3. Lemma. $|A_\varepsilon^{(n)}(X|y^n)| \le 2^{n(H(X|Y)+2\varepsilon)}$.

Bizonyítás. Ha $(x^n, y^n) \in A_\varepsilon^{(n)}(X,Y)$, akkor definíció szerint egyrészt $p(x^n, y^n) \ge 2^{-n(H(X,Y)+\varepsilon)}$, másrészt $p(y^n) \le 2^{-n(H(Y)-\varepsilon)}$. Ebből osztással

$$p(x^n|y^n) \ge 2^{-n(H(X,Y)-H(Y)+2\varepsilon)} = 2^{-n(H(X|Y)+2\varepsilon)}.$$

Rögzített y^n mellett

$$1 \ge \sum_{x^n \in A_\varepsilon^{(n)}(X|y^n)} p(x^n|y^n) \ge |A_\varepsilon^{(n)}(X|y^n)| 2^{-n(H(X|Y)+2\varepsilon)},$$

amiből a lemma állítását kapjuk. ■

Megjegyezzük, hogy hasonló alsó becslés is érvényes, de arra most nem volt szükségünk. A Slepian-Wolf tétel az IS $(\mathbb{X}, \mathbb{Y})$ forrásokra is igaz (és ugyanígy bizonyítható): ezek azok a források, melyekre az (1) képlettel (csak a $H(\mathbb{X})$, $H(\mathbb{Y})$, $H(\mathbb{X}, \mathbb{Y})$ betűnkénti entrópiákkal) definiált tipikus halmaz valószínűsége 1-hez tart. Végül meggondolható, hogy a Slepian-Wolf tétel által karakterizált elérhető ráták lineáris kódokkal is elérhetők.

7. A csatornakapacitás

Térjünk most rá a csatorna vizsgálatára! Feltesszük, hogy az ismert struktúrájú stacionárius forrás által kibocsátott üzeneteket már megfelelően tömörítettük, méghozzá elhanyagolható hibavalószínűséggel.

Láttuk, hogy az összes m hosszúságú üzenetből csak nagyjából $M = 2^{mH(\mathbb{X})}$ a tipikus, és ezek összvalószínűsége majdnem 1. Kérdés, hogy ezen M különböző üzenet egyikét a csatorna hányszori használatával lehet megbízhatóan továbbítani. A továbbiakban itt csak az M szám a fontos, az már számunkra érdektelen, hogy ezek az üzenetek milyen forrásból származtak.

A csatornát az $(\mathcal{X}, \mathcal{Y}, p(y|x))$ hármas írja le, ahol $\mathcal{X}$ a csatorna (véges) bemeneti ábécéje, $\mathcal{Y}$ a kimeneti ábécé, a $p(y|x)$ sztochasztikus mátrix pedig az átmenetvalószínűségeket adja meg, azaz $p(y|x)$ annak valószínűsége, hogy az x bemeneti jelből a csatorna y kimeneti jelet ad. Egyelőre feltesszük, hogy a csatorna emlékezet nélküli, azaz ha x^n, y^n bemeneti, illetve kimeneti sorozatok, akkor $p(y^n|x^n) = \prod_i p(y_i|x_i)$.

7.1. Definíció. *A $(\mathcal{X}, \mathcal{Y}, p(y|x))$ emlékezet nélküli csatorna kapacitása $C = \max_{p(x)} I(X \wedge Y)$. A kapacitás tehát azt mondja meg, hogy ha a bemeneti eloszlást szabadon választhatjuk meg, akkor mekkora lehet a bemenet és a kimenet kölcsönös információja.*

A fenti definíció az entrópia $H(X) = -\sum p(x) \log p(x)$ alakú definíciójának analógja annyiban, hogy most még nem látszik, hogy ennek a definíciónak mi köze az információtovábbításhoz. Egyelőre higgyük el, hogy ez a szerencsés definíció, és vizsgáljuk meg néhány tulajdonságát!

7.2. Tétel. *A csatornakapacitás definíciójában a maximum létezik, és $0 \leq C \leq \min(\log|\mathcal{X}|, \log|\mathcal{Y}|)$.*

Bizonyítás. Az első állítás abból következik, hogy az $I(X \wedge Y) = H(X) + H(Y) - H(X,Y)$ kölcsönös információ nyilvánvalóan folytonos függvénye a $p(x)$ vektornak a valószínűségeloszlások konvex, zárt halmazán (mivel az entrópia az). A második állítás triviális. ■

Ennél több is igaz, méghozzá hogy az $I(X \wedge Y)$ kölcsönös információ konkáv függvénye a $p(x)$ vektornak, így minden lokális maximumhely egyben globális is, és numerikus módszerekkel viszonylag egyszerűen kereshető a maximum. Ezt a következő állítások bizonyítják.

7.3. Tétel. *A $D(P\|Q)$ divergencia konvex függvénye a (P, Q) párnak.*

Bizonyítás. Legyenek P_0, P_1, Q_0, Q_1 eloszlások, és $P_\lambda = (1-\lambda)P_0 + \lambda P_1$, $Q_\lambda = (1-\lambda)Q_0 + \lambda Q_1$. Ekkor minden x-re a log-összeg egyenlőtlenség szerint

$$p_\lambda(x) \log \frac{p_\lambda(x)}{q_\lambda(x)} \leq (1-\lambda) p_0(x) \log \frac{p_0(x)}{q_0(x)} + \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)}.$$

Ezt x-ben összegezve kapjuk, hogy $D(P_\lambda\|Q_\lambda) \leq (1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1)$. ■

7.4. Tétel. *A $H(P)$ entrópia konkáv függvénye a P-nek.*

Bizonyítás. Ha Q az egyenletes eloszlás $\mathcal{X}$-en, akkor

$$D(P\|Q) = \log|\mathcal{X}| - H(P).$$

Ezért a kívánt állítás az előző tételből következik. ■

7.5. Tétel. *Az $I(X \wedge Y)$ kölcsönös információ rögzített $p(y|x)$ mátrix mellett $p(x)$-nek konkáv függvénye, míg rögzített $p(x)$ mellett $p(y|x)$-nek konvex függvénye.*

Bizonyítás. 1) Kezdjük az első állítással! $I(X \wedge Y) = H(Y) - H(Y|X)$. Itt $H(Y|X) = \sum_x p(x) H(Y|X = x)$, ahol a $H(Y|X = x)$ mennyiségek már rögzítettek. Ezért $H(Y|X)$ a $p(x)$ eloszlás lineáris függvénye. Másrészt $H(Y)$ konkáv függvénye a $p(y)$ eloszlásnak, mely lineáris függvénye a $p(x)$ eloszlásnak a $p(y) = \sum_x p(x)p(y|x)$ összefüggés miatt.

2) A második állításra rátérve, használjuk az

$$I(X \wedge Y) = D(X, Y\|X', Y')$$

összefüggést, ahol X', Y' függetlenek. Legyen adott a $p_0(y|x)$ és a $p_1(y|x)$ eloszlás, és tekintsük ezek

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda) p_0(y|x)$$

konvex kombinációját. Legyen még $p_\lambda(x,y)$ és $p'_\lambda(x,y)$ a megfelelő együttes eloszlások, azaz

$$p_\lambda(x,y) = p(x)p_\lambda(y|x), \quad p'_\lambda(x,y) = p(x)\sum_{z\in\mathcal{X}} p(z)p_\lambda(y|z).$$

Mivel könnyen ellenőrizhető, hogy $p_\lambda(x,y) = \lambda p_1(x,y) + (1-\lambda)p_0(x,y)$ és $p'_\lambda(x,y) = \lambda p'_1(x,y) + (1-\lambda)p'_0(x,y)$, az állítás következik a divergencia konvexitásából. ■

7.6. Definíció. *A csatorna* szimmetrikus, *ha a csatornamátrix minden sora ugyanazokat a számokat tartalmazza, legfeljebb más sorrendben, és minden oszlopa is ugyanazokat a számokat tartalmazza, legfeljebb más sorrendben. A csatorna* gyengén szimmetrikus, *ha a csatornamátrix minden sora ugyanazokat a számokat tartalmazza, legfeljebb más sorrendben, és minden oszlop összege ugyanannyi.*

7.7. Tétel. *Jelölje a gyengén szimmetrikus csatorna egy tetszőleges sorának entrópiáját H_s. Ekkor a kapacitás $C = \log|\mathcal{Y}| - H_s$, mely az egyenletes bemeneti eloszláson eléretik.*

Bizonyítás.

$$I(X\wedge Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X=x) = H(Y) - H_s,$$

mivel minden x-re $H(Y|X=x) = H_s$. Ezért

$$C = (\max_{p(x)} H(Y)) - H_s \le \log|\mathcal{Y}| - H_s,$$

és a felső korlátot az egyenletes bemeneti eloszlás eléri:

$$p(y) = \sum_x p(x)p(y|x) = \frac{1}{|\mathcal{X}|}\sum_x p(y|x) = \frac{1}{|\mathcal{Y}|},$$

mivel minden oszlop összege $|\mathcal{X}|/|\mathcal{Y}|$. ■

Szimmetrikus csatornára példa a „zajos írógép". Itt $\mathcal{X} = \mathcal{Y} = \{0,1,\ldots,K-1\}$, és $p(x+1|x) = p(x|x) = 1/2$ (a műveletek mod K értendők). Tehát a kapacitás $C = \log(K/2)$. Egy másik példa a bináris szimmetrikus csatorna, melyre $\mathcal{X} = \mathcal{Y} = \{0,1\}$, a csatornamátrix pedig

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

A kapacitás ebben az esetben tehát $C = 1 - H(p)$.

Nézzünk most egy olyan példát, ahol a csatorna nem gyengén szimmetrikus! Nevezzük bináris eltörléses csatornának a következőt. Legyen megint $\mathcal{X} = \{0,1\}$, és $\mathcal{Y} = \{0,*,1\}$. A csatornamátrix:

$$\begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix}.$$

Vagyis a bemeneti jel $1-p$ valószínűséggel hibátlanul átmegy, p valószínűséggel viszont eltörlődik. A két sorban most is ugyanazok a valószínűségek szerepelnek, így $C = (\max_{p(x)} H(Y)) - H(p)$, de most az egyenletes kimeneti eloszlás nem érhető el (kivéve ha $p = 1/3$). A kimeneti eloszlás: $(p(0)(1-p), p, (1-p(0))(1-p))$. Ennek entrópiája $H(Y) = H(p) + (1-p)H(p(0))$ melynek maximális értéke $H(p) + 1 - p$, ha $p(0) = 1/2$. Tehát a csatorna kapacitása $C = 1-p$.

A csatornakapacitás meghatározása általában nehéz, van rá iteratív algoritmus (Arimoto-Blahut). Fontos speciális eset a bináris csatorna, melynek csatornamátrixa

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

Ha $p_{00} = p_{10}$, akkor a csatorna kimenete független a bemenetétől, az ilyen csatorna kapacitása nyilván nulla. Ellenkező esetben nyilván feltehetjük, hogy $p_{00} > p_{10}$ (máskülönben a kimenetnél felcseréljük a

nullát és az egyet). Legyen $p_0 = P(X = 0)$, $p_1 = P(X = 1)$ a bemeneti eloszlás, $r_0 = P(Y = 0)$, $r_1 = P(Y = 1)$ pedig a kimeneti eloszlás. Természetesen

$$\begin{array}{rcccl} r_0 & = & p_0p_{00} + p_1p_{10} & = & p_0p_{00} + (1-p_0)p_{10}, \\ r_1 & = & p_0p_{01} + p_1p_{11} & = & p_0p_{01} + (1-p_0)p_{11}. \end{array}$$

Írjuk most fel a bemenet és a kimenet kölcsönös információját, melyet maximalizálni szeretnénk!

$$I(X \wedge Y) = H(Y) - H(Y|X) = H(r_0) - (p_0H(p_{00}) + p_1H(p_{10})).$$

Ennek maximumhelyét deriválással kereshetjük meg. A kölcsönös információt p_0 szerint deriválva, a maximumhelyre kapott egyenlet:

$$H'(r_0) \cdot (p_{00} - p_{10}) - H(p_{00}) + H(p_{10}) = 0,$$

azaz

$$H'(r_0) = \frac{H(p_{00}) - H(p_{10})}{p_{00} - p_{10}}.$$

Grafikusan ez azt jelenti, hogy a legjobb bemeneti eloszlás mellett az r_0 kimeneti valószínűség által adott helyen a $H(p)$ függvény érintője párhuzamos a p_{00} és p_{10} helyek közötti szelővel. Ilyen r_0 nyilván létezik, és megadható, mint a p_{00}, p_{10} valószínűségek konvex kombinációja. A fenti egyenletből kifejezhető p_0, majd visszahelyettesítéssel megkapható a csatornakapacitás, azonban nagyon bonyolult képleteket kapunk.

A trükk, mely Murogától származik, az, hogy nem a bemeneti eloszlás szerint maximalizálunk, hanem a kimeneti szerint. (S. Muroga, On the capacity of a discrete channel, I, J. Phys. Soc. Japan 8, 4, 484-494 (1953).) Legyen q_0, q_1 a

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \end{pmatrix} = \begin{pmatrix} -H(p_{00}) \\ -H(p_{10}) \end{pmatrix}$$

egyenlet megoldása (feltevésünk szerint P invertálható mátrix). Ezzel

$$I(X \wedge Y) = H(r_0) + r_0q_0 + r_1q_1,$$

ezt szeretnénk r_0-ban maximalizálni, azaz r_0 szerint deriválni. Határozzuk meg a $H(p)$ függvény deriváltját! Térjünk át a $H_e(p) = -(p \ln p + (1-p)\ln(1-p)) = (\ln 2)H(p)$ függvényre! Erre $H_e'(p) = \ln \frac{1-p}{p}$, amiből visszatérve a kettes alapú logaritmusra, $H'(p) = \log \frac{1-p}{p}$ adódik. Tehát a maximális kölcsönös információt adó kimeneti eloszlásra kapjuk, hogy $\log(r_1/r_0) = q_1 - q_0$, átrendezve: $\log r_1 - q_1 = \log r_0 - q_0$. Ebből

$$\begin{aligned} C = H(r_0) + r_0q_0 + r_1q_1 = -r_0 \log r_0 + r_0q_0 - r_1 \log r_1 + r_1q_1 = \\ - r_0(\log r_1 - q_1) - r_1 \log r_1 + r_1q_1 = -\log r_1 + q_1 = -\log r_0 + q_0. \end{aligned}$$

Ebből kifejezhetjük a kimeneti eloszlást:

$$r_0 = 2^{q_0 - C}, \quad r_1 = 2^{q_1 - C},$$

amiből az $r_0 + r_1 = 1$ összefüggést felhasználva a kapacitás

$$C = \log(2^{q_0} + 2^{q_1}).$$

Megjegyezzük, hogy a módszer nagyobb ábécék esetén is használható, bár ott nem minden esetben működik.

Most ismertetünk egy általános algoritmust a csatornakapacitás kiszámítására. Az eljárás neve: Arimoto-Blahut algoritmus. Jelölje a csatornamátrix egy elemét $p(y|x)$, és legyen $p(x)$ egy bemeneti eloszlás. Legyen

$$W_p(x|y) = \frac{p(x)p(y|x)}{\sum_z p(z)p(y|z)},$$

a Bayes tétel szerint ez annak a valószínűsége, hogy a csatorna bemenete x volt, ha a kimenet y.

Könnyű algebrai átalakítással látszik, hogy

$$I(X \wedge Y) = \sum_{x,y} p(x)p(y|x) \log \frac{W_p(x|y)}{p(x)}.$$

Ha most $W(x|y)$ tetszőleges feltételes eloszlások gyűjteménye ($W(\cdot|y)$ minden y-ra egy eloszlás), akkor definiálható az

$$I(p, W) = \sum_{x,y} p(x)p(y|x) \log \frac{W(x|y)}{p(x)}$$

mennyiség. Felírható, hogy

$$I(p, W_p) - I(p, W) = D(P\|Q) \geq 0,$$

ahol $P(x, y) = p(x)p(y|x)$ és $Q(x, y) = W(x|y) \sum_z p(z)p(y|z)$ két eloszlás, ezért $\max_W I(p, W) = I(p, W_p)$. Következésképpen a csatornakapacitásra

$$C = \max_p I(p, W_p) = \max_{p,W} I(p, W).$$

Adódik az algoritmus ötlete: maximalizáljunk felváltva a két argumentumban! Legyen tehát p_0 egy szigorúan pozitív kezdeti eloszlás a bemeneti ábécén. Ezután $n \geq 0$-ra legyen
1) $W_n = \arg\max_W I(p_n, W) = W_{p_n}$,
2) $p_{n+1} = \arg\max_p I(p, W_n)$.

Oldjuk meg a 2)-ben szereplő maximalizálási feladatot! Rögzített W mellett legyen $\alpha(x) = \sum_y p(y|x) \log W(x|y)$. Ezzel

$$I(p, W) = \sum_x p(x)\alpha(x) + H(p),$$

ami p-nek konkáv függvénye. Tehát a maximumot deriválással találhatjuk meg. Azt találjuk, hogy a maximumhely:

$$p(x) = c2^{\alpha(x)} = \frac{2^{\alpha(x)}}{\sum_z 2^{\alpha(z)}}.$$

Legyen $\beta(x) = \beta(W, x) = 2^{\alpha(x)} = \prod_y W(x|y)^{p(y|x)}$. Tehát megoldottuk a feladatot:

$$p_{n+1} = \arg\max_p I(p, W_n) = \frac{\beta(W_n, x)}{\sum_z \beta(W_n, z)}.$$

Azt kellene még belátni, hogy az algoritmus konvergál, azaz $I(p_{n+1}, W_n) \to C$. A vizsgált kölcsönös információt megkapjuk, ha p_{n+1} helyébe behelyettesítjük a korábban kapott kifejezést. Algebrai átalakítások után:

$$I(p_{n+1}, W_n) = \sum_{x,y} \frac{\beta(W_n, x)}{\sum_z \beta(W_n, z)} p(y|x) \log \left[\frac{W_n(x|y)}{\beta(W_n, x)} \sum_z \beta(W_n, z) \right] = \log \sum_z \beta(W_n, z).$$

Belátható továbbá, hogy tetszőleges (X, Y) változókra

$$I(X \wedge Y) = D(X, Y\|\tilde{X}, \tilde{Y}) \leq D(X, Y\|\tilde{X}, Z),$$

ahol az első divergenciában $\tilde{X}$ és $\tilde{Y}$ külön-külön ugyanolyan eloszlásúak, mint X, Y, de függetlenek; a második divergenciában $\tilde{X}$ ugyanolyan eloszlású, mint X, és független Z-től, ami viszont tetszőleges eloszlású. Ezt a csatorna esetére alkalmazva kapjuk, hogy

$$I(X \wedge Y) \leq \sum_{x,y} p(x)p(y|x) \log \frac{p(y|x)}{\sum_z q(z)p(y|z)},$$

ahol $q(z)$ tetszőleges eloszlás. Felhasználva, hogy

$$\log \frac{\beta(W_n, x)}{p_n(x)} = \sum_y p(y|x) \log \frac{p(y|x)}{\sum_z p_n(z)p(y|z)},$$

a csatornakapacitást elérő p^* bemeneti eoszlásra

$$C = I(X^* \wedge Y^*) \leq \sum_x p^*(x) \log \frac{\beta(W_n, x)}{p_n(x)}.$$

Ebből

$$C - I(p_{n+1}, W_n) \le \sum_x p^*(x) \log \frac{\beta(W_n, x)}{p_n(x)} - \log \sum_z \beta(W_n, z) = \sum_x p^*(x) \log \frac{p_{n+1}(x)}{p_n(x)}.$$

Összegezve n-ben:

$$\sum_{n=0}^{m} (C - I(p_{n+1}, W_n)) \le \sum_x p^*(x) \log \frac{p_{m+1}(x)}{p_0(x)} \le \max_x \log \frac{1}{p_0(x)} = K.$$

Tehát az összeg tagjai nullához tartanak.

Lássuk most Shannon alaptételét a csatornakapacitásról! Először a definíciók:

7.8. Definíció. *(M, n)-csatornakódnak nevezünk egy $g : \{1, \dots, M\} \to \mathcal{X}^n$ leképezést. Ha megadunk egy $f : \mathcal{Y}^n \to \{1, \dots, M\}$ dekódolót is, akkor az i-edik üzenetre a dekódoló hibája $\lambda_i = P(\hat{W} \neq i | W = i)$, ahol az üzenetet a hírközlés egyes fázisaiban a*

$$W \to g(W) = X^n \to Y^n \to f(Y^n) = \hat{W}$$

betűkkel jelöljük. A maximális hiba $\lambda^{(n)} = \max_i \lambda_i$, az átlagos hiba pedig $P_e^{(n)} = \frac{1}{M} \sum_i \lambda_i$. Az R ráta elérhető, ha létezik olyan $(2^{nR}, n)$ kódsorozat, hogy $\lim_{n\to\infty} \lambda^{(n)} = 0$.

Itt tehát már mindegyik üzenet fontos, akármelyik üzenetet is adják le a csatornán, azt szeretnénk, hogy a hibás dekódolás valószínűsége elhanyagolható legyen.

7.9. Tétel. *Legyen egy emlékezet nélküli csatorna kapacitása C. Ha $R < C$, akkor az R ráta elérhető. Fordítva, ha az R ráta elérhető, akkor $R \le C$.*

Bizonyítás. Kezdjük az első állítással! Most is a véletlen kódválasztás módszerét fogjuk használni. Legyen tehát $R < C$, és $p(x)$ az a bemeneti eloszlás, melyre $C = I(X \wedge Y)$. Generáljunk eszerint egy véletlen kódot: minden $1 \le i \le 2^{nR}$ üzenetre, egymástól függetlenül, legyenek a $g(i) = X^n(i)$ kódszó betűi függetlenek, és $p(x)$ eloszlásúak. Ekkor bármelyik i üzenetre az $(X^n(i), Y^n(i))$ pár együttes eloszlása meghatározott:

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i) p(y_i | x_i),$$

ahol tehát $p(x)$ a kapacitást elérő bemeneti eloszlás, $p(y|x)$ pedig a csatorna átmenetmátrixának eleme. Vezessük be az $A_\varepsilon^{(n)}(X, Y)$ jelölést az együttesen tipikus sorozatok halmazára (lásd a Slepian-Wolf tétel bizonyítását), tehát:

$$A_\varepsilon^{(n)}(X, Y) = \{(x^n, y^n) : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \varepsilon, \quad \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \varepsilon, \quad \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \varepsilon\}.$$

Rögzített $b = (x^n(1), \dots, x^n(2^{nR}))$ kódkönyv mellett használjuk a tipikussági dekódolót! A dekódolásnál legyen $f(y^n) = j$ pontosan akkor, ha j az egyetlen olyan index, melyre $(x^n(j), y^n) \in A_\varepsilon^{(n)}(X, Y)$ (ε majd megfelelően kicsi lesz). Először az átlagos hibát vizsgáljuk. Az átlagos hiba várható értéke a B kódkönyv választása szerint:

$$E(P_e^{(n)}(B)) = \frac{1}{2^{nR}} \sum_i E(\lambda_i(B)) = E(\lambda_1(B)),$$

mivel minden üzenethez ugyanolyan mechanizmus szerint választottunk véletlen kódszót. Most már csak az első üzenetre kell koncentrálnunk: az első üzenetet elküldve, legyen $Y^n(1)$ a csatornából kijövő sorozat. Akkor lesz hibás a dekódolás, ha vagy $(X^n(1), Y^n(1))$ nem együttesen tipikus, vagy pedig van olyan $i > 1$, hogy $(X^n(i), Y^n(1))$ együttesen tipikus. Vezessük be a következő eseményeket:

$$A_i = \{(X^n(i), Y^n(1)) \in A_\varepsilon^{(n)}(X, Y)\}, \quad i = 1, \dots, 2^{nR}.$$

Ekkor

$$E(\lambda_1(B)) = P(A_1^c \cup A_2 \cup \ldots \cup A_{2^{nR}}) \leq P(A_1^c) + \sum_{i>1} P(A_i).$$

Az első tagról tudjuk, hogy nullához tart, a 7.11 Lemma szerint pedig $P(A_i) \leq 2^{-n(C-3\varepsilon)}$, ha $i > 1$. Ha tehát $R < C - 3\varepsilon$, akkor $E(P_e^{(n)}(B)) \to 0$. Ha a várható érték nullához tart, akkor kell legyen olyan b_n kódsorozat, hogy $P_e^{(n)}(b_n) \to 0$. Végül, ha ezekből a kódokból csak a $2^{nR-1} = 2^{n(R-1/n)}$ legkisebb hibájú kódszót tartjuk meg (vagyis a felét), akkor az aszimptotikus ráta nem változik, viszont az így kapott kód maximális hibája legfeljebb $2 \cdot P_e^{(n)}(b_n)$, vagyis nullához tart: legyen $M = 2^{nR}$ és tegyük fel, hogy az üzenetek úgy vannak számozva, hogy $\lambda_1 \leq ... \leq \lambda_M$. Ekkor

$$P_e^{(n)}(b_n) = \frac{1}{M} \sum_{i \leq M/2} \lambda_i + \frac{1}{M} \sum_{i > M/2} \lambda_i \geq 0 + \frac{1}{M} \cdot \frac{M}{2} \min_{i > M/2} \lambda_i \geq \frac{1}{2} \max_{i \leq M/2} \lambda_i.$$

A megfodításhoz tegyük fel, hogy R elérhető ráta, azaz létezik ilyen rátájú kódsorozat, melynek maximális hibája nullához tart. Ekkor nyilván az átlagos hiba is nullához tart. Ha deklaráljuk, hogy W egyenletes eloszlású, akkor $P_e^{(n)} = P(\hat{W} \neq W)$. A Fano-egyenlőtlenség szerint

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR.$$

Továbbá

$$nR = H(W) = H(W|\hat{W}) + I(W \wedge \hat{W}) \leq 1 + P_e^{(n)} nR + I(X^n \wedge Y^n) \leq 1 + P_e^{(n)} nR + nC,$$

ahol a 7.10 Lemmát, valamint az adatfeldolgozási egyenlőtlenséget (kétszer) használtuk. Az egyenlőtlenséget n-el osztva, majd végtelenhez tartva $R \leq C$ adódik. ■

Bizonyítható, hogy $R > C$ esetén az átlagos hiba exponenciális gyorsasággal egyhez tart.

A következő érdekes, bár nem meglepő lemma azt mondja ki, hogy a csatorna n-edik hatványának kapacitása nC.

7.10. Lemma. *Adott egy C kapacitású emlékezet nélküli csatorna. Jelölje X^n a bemenetet, Y^n a kimenetet. Ekkor*

$$\max_{p(x^n)} I(X^n \wedge Y^n) = nC.$$

Bizonyítás.

$$I(X^n \wedge Y^n) = H(Y^n) - H(Y^n|X^n) =$$

$$H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i),$$

mivel az X_i feltétel mellett Y_i minden mástól független. Az első tagot becsülve,

$$I(X^n \wedge Y^n) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i \wedge Y_i) \leq nC,$$

és ha az X_i-k függetlenek, és a C-t elérő eloszlásúak, akkor mindenhol egyenlőség áll. ■

7.11. Lemma. *Legyen $(\mathbb{X}, \mathbb{Y})$ emlékezet nélküli stacionárius forrás. Szokás szerint $p(x^n, y^n)$, $p(x^n)$, $p(y^n)$ jelöli az együttes, valamint a marginális eloszlásokat. Legyen $(\tilde{X}^n, \tilde{Y}^n)$ a fenti marginálisokkal rendelkező, de független elemű pár, azaz $P(\tilde{X}^n = x^n, \tilde{Y}^n = y^n) = p(x^n)p(y^n)$. Ekkor*

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}(X, Y)) \leq 2^{-n(I(X \wedge Y) - 3\varepsilon)}.$$

Bizonyítás. A bizonyítás igen egyszerű:

$$P((\tilde{X}^n,\tilde{Y}^n)\in A_\varepsilon^{(n)}(X,Y)) = \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}(X,Y)} p(x^n)p(y^n) \le$$
$$|A_\varepsilon^{(n)}(X,Y)|2^{-n(H(X)-\varepsilon)}2^{-n(H(Y)-\varepsilon)} \le 2^{n(H(X,Y)-H(X)-H(Y)+3\varepsilon)} = 2^{-n(I(X\wedge Y)-3\varepsilon)}.$$

■

Végül nézzünk egy érdekes esetet, a visszacsatolással rendelkező csatornát. A csatorna tulajdonképp ugyanúgy működik, mint eddig, csak a vevő a vett jeleket azonnal visszaküldi a küldőhöz (zaj nélkül), így a W üzenethez tartozó kódszó i-edik betűjének megválasztásakor az adó az Y^{i-1} sorozatot is felhasználhatja. Jelölje C_{FB} az ilyen konstrukcióval elérhető ráták szuprémumát. Shannon (1956) tétele azt mondja ki, hogy a visszacsatolás nem növeli az emlékezet nélküli csatorna kapacitását (bár a kódolást-dekódolást megkönnyítheti).

7.12. Tétel. *Adott egy emlékezet nélküli csatorna C kapacitással. Ekkor $C_{FB} = C$.*

Bizonyítás. Nyilván csak azt kell megmutatni, hogy ha R elérhető ráta, akkor $R \le C$. Ismét a Fano-egyenlőtlenséget próbáljuk alkalmazni. Legyen W egyenletes eloszlású a lehetséges 2^{nR} darab üzenet halmazán, ekkor $P(\hat{W}\neq W) = P_e^{(n)} \to 0$.

$$nR = H(W) = H(W|\hat{W}) + I(W\wedge\hat{W}) \le 1 + P_e^{(n)}nR + I(W\wedge Y^n),$$

mivel $W \to Y^n \to \hat{W}$ Markov lánc. A 7.10 Lemma bizonyítását szeretnénk lemásolni. Először is

$$I(W\wedge Y^n) = H(Y^n) - H(Y^n|W).$$

A második tag becslése:

$$H(Y^n|W) = \sum_{i=1}^n H(Y_i|Y^{i-1},W) = H(Y_i|Y^{i-1},W,X_i) = \sum_{i=1}^n H(Y_i|X_i).$$

A második egyenlőségben azt használtuk fel, hogy X_i a (W,Y^{i-1}) változók függvénye, ezért ha X_i-t is beírjuk az entrópia feltételébe, a feltételes entrópia nem csökken. A harmadik egyenlőség pedig azért teljesül, mert X_i ismerete mellett Y_i már független a (W,Y^{i-1}) változóktól, ezért ezért ezeket el lehet hagyni a feltételből. Innentől a bizonyítás ugyanúgy fejezhető be, mit a 7.10 Lemma esetében. ■

Ha visszagondolunk a korábban vizsgált zajos írógépre, és feltesszük, hogy $K = 2L$ páros, akkor a kapacitás $C = \log L$, és ez valóban elérhető, méghozzá nulla hibavalószínűséggel: ha az írógépnek csak minden második betűjét használjuk, akkor a kimenet egyértelműen meghatározza a bemenetet, és az L betű $\log L$ bitnek felel meg.

A bináris eltörléses csatorna kapacitása $1-p$ volt, ebben az esetben nem látszik ilyen egyértelműen, hogyan érhető el (vagy közelíthető meg) ez a ráta. Ha azonban van visszacsatolás, akkor megtehetjük, hogy az eltörlődött jeleket újra elküldjük. Ha az eltörlés esélye p, akkor egy jel átlagosan az $1/(1-p)$-edik kísérletre megy át. Tehát 1 bit információ a csatorna $1/(1-p)$-szeri használatával továbbítható, azaz egy használattal valóban $1-p$ bit továbbítható.

8. A bináris szimmetrikus csatorna hibaexponense

Legyen a bináris szimmetrikus csatornára a bit-torzulás valószínűsége $0 < p < 1/2$. Azt szeretnénk vizsgálni, hogy elérhető ráták esetén milyen gyorsan tart a hibavalószínűség nullához.

8.1. Definíció. *Adott emlékezet nélküli csatornára és $0 < R < C$ elérhető rátára az $E(R)$ hibaexponens a legnagyobb olyan c szám, hogy elég nagy n-re létezik R sebességű, legfeljebb $2^{-nc+o(n)}$ átlagos hibavalószínűségű n-kód.*

8.2. Tétel. *A bináris szimmetrikus csatornára ($0 < p < 1/2$) a hibaexponens függvényre teljesül, hogy $E(R) \leq D(q\|p)$, ahol $q < 1/2$ az a szám, melyre $R = 1 - H(q)$.*

Bizonyítás. Legyen $R < C = 1-H(p)$ egy elérhető ráta. Tetszőleges R rátájú n-kódra legyen $M = 2^{nR}$, és jelölje $x^n(1), \ldots, x^n(M)$ a kódszavakat. A dekódolási hibára szeretnénk alsó becslést adni, ehhez a „legjobb" (legkisebb átlagos hibavalószínűségű) dekódolót kell először meghatároznunk.

A dekódolás azt jelenti, hogy minden $y^n \in \{0,1\}^n$ sorozatra meg kell adni egy $i(y^n)$ indexet (y^n dekódoltja $i(y^n)$). Vagyis a $\{0,1\}^n$ halmazt fel kell bontani M darab diszjunkt részre:

$$\{0,1\}^n = \cup_{i=1}^M D_i : \quad y^n \in D_i \iff i(y^n) = i.$$

Vizsgáljuk meg a λ_i hibavalószínűséget:

$$1 - \lambda_i = P(\hat{W} = W|W = i) = P(Y^n \in D_i|W = i) = \sum_{y^n \in D_i} p(y^n|x^n(i)).$$

Ezeket kiátlagolva:

$$1 - P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{M}(1-\lambda_i) = \frac{1}{M}\sum_{i=1}^{M}\sum_{y^n \in D_i} p(y^n|x^n(i)) = \frac{1}{M}\sum_{y^n \in \{0,1\}^n} p(y^n|x^n(i(y^n))).$$

Ez akkor lesz maximális, ha bármely $y^n \in \{0,1\}^n$ sorozat dekódoltja az az i, melyre a $p(y^n|x^n(i))$ feltételes valószínűség maximális. Ennek elnevezése *maximum likelihood dekódoló*. A mi esetünkben ez éppen a Hamming-távolság szerinti dekódolással esik egybe: legyen $d(x^n, y^n)$ a két sorozat Hamming távolsága (azaz a különböző koordináták száma), ekkor

$$p(y^n|x^n) = p^{d(x^n,y^n)}(1-p)^{n-d(x^n,y^n)}.$$

Ez pedig akkor maximális ($p < 1/2$ miatt), ha $d(x^n, y^n)$ minimális, vagyis minden sorozat dekódoltja a hozzá legközelebb eső kódszó indexe.

Mivel összesen 2^n sorozatunk van, létezik olyan i, hogy

$$|D_i| \leq 2^n/M = 2^{n(1-R)} = 2^{nH(q)}.$$

Megmutatjuk, hogy $q < q' < 1/2$ esetén az $x^n(i)$-től $d = [nq']$ Hamming-távolságra eső sorozatok nagy része nincs benne D_i-ben. Az ilyen sorozatok száma ugyanis a Stirling formula alapján

$$|\{y^n : d(y^n, x^n(i)) = d\}| = \binom{n}{d} \sim \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{d}{e}\right)^d \sqrt{2\pi d}\left(\frac{n-d}{e}\right)^{n-d}\sqrt{2\pi(n-d)}} = 2^{o(n)}\frac{n^n}{d^d(n-d)^{n-d}}.$$

Algebrai átalakításokkal

$$\frac{n^n}{d^d(n-d)^{n-d}} = 2^{nH\left(\frac{d}{n}\right)} = 2^{nH(q')+o(n)}.$$

Mivel $H(q') > H(q)$, az $x^n(i)$-től épp d távolságra eső sorozatok jó része (mondjuk legalább fele) biztos nincs benne D_i-ben. Vagyis a hibás dekódolás valószínűségére

$$\lambda_i > p^d(1-p)^{n-d}2^{nH(q')+o(n)} = 2^{-nD(q'\|p)+o(n)}.$$

Mivel ez minden $q' > q$ esetén igaz, a maximális hibára megkaptuk az állításban szereplő alsó becslést:

$$\lambda^{(n)} \geq 2^{-nD(q\|p)+o(n)}.$$

Az átlagos hibára vonatkozó eredmény a szokásos felezéses trükkel következik: legyenek kódkönyvünkben a hibavalószínűségek $\lambda_1 \leq ... \leq \lambda_M$. Korábbi számolásunk alapján

$$P_e^{(n)} \geq \frac{1}{2}\max_{i \leq M/2} \lambda_i.$$

Tehát ha a kódkönyvünkből csak az első $M^* = M/2$ kódszót tartjuk meg, akkor az aszimptotikus ráta nem változik. Továbbá a felezett kódban a hibavalószínűségek nem nőnek: $\lambda_i^* \leq \lambda_i$, hiszen a maximum likelihood dekódolást használva, az i-edik kódszóhoz tartozó D_i tartomány csak nőhet, azaz $D_i \subseteq D_i^*$. A felezett kódra alkalmazva az eddigieket kapjuk, hogy

$$2^{-nD(q\|p)+o(n)} \leq \max_{i\leq M^*} \lambda_i^* \leq \max_{i\leq M^*} \lambda_i \leq 2P_e^{(n)},$$

vagyis az átlagos hibára is megkaptuk az állításban szereplő alsó becslést. ■

Megjegyzés: Legyen $C_1 = 1 - H\left(\frac{\sqrt{p}}{\sqrt{p}+\sqrt{1-p}}\right)$. Ha $C_1 \leq R < C$, akkor a fenti tételben valójában egyenlőség áll, ennél kisebb rátára viszont nem ismert a hibaexponens pontos értéke.

9. Hamming kód egy hiba javítására

Most már tudjuk, hogy mekkora sebességgel remélhetünk információt továbbítani egy adott csatornán, viszont nem tudjuk még, hogy ezt a sebességet milyen konkrét kódolási eljárásokkal közelíthetjük meg. Ez egyáltalán nem egyszerű kérdés, a mai napig intenzív kutatás folyik az egyre jobb és jobb, a gyakorlatban alkalmazható kódok konstruálására.

Nézzük a legegyszerűbb esetet, a bináris csatornát. Olyan blokk-kódokat keresünk, melyek paritásellenőrző bitek segítségével teszik lehetővé a hibák jelzését illetve javítását. Ez az ötlet Hammingtől származik, 1950-ből. Az alapötlet az, hogy a lehetséges kódszavak terében olyan tényleges kódszavakat válasszunk ki, melyek Hamming-távolsága elég nagy. Ha bármely két kódszó távolsága legalább $2m+1$, akkor m hiba javítható (a dekódolt üzenet ahhoz a kódszóhoz tartozik, mely a vett sorozathoz legközelebb esik). Tehát ha legfeljebb m hiba történt, akkor pontosan meg tudjuk mondani, hogy melyik kódszó volt a csatorna bemenete.

9.1. Definíció. *(n,k,d)-kódnak nevezünk egy kódot, ahol egy blokk n bitből áll (ez a kódszavak hossza), ebből k információs bit van (azaz 2^k lehetséges üzenet), és bármely két kódszó Hamming-távolsága legalább d.*

Mutatunk egy konkrét $(2^h-1, 2^h-h-1, 3)$-kódot, mely tehát egy hiba javítását garantálja. Legyen H az a $h \times 2^h-1$ méretű mátrix, melynek oszlopai a h hosszú $0-1$ vektorok, a csupa nullát kivéve. Az egyszerűség kedvéért az utolsó h oszlop által meghatározott részmátrix legyen az egységmátrix. A c kódszavak első 2^h-h-1 bitjét válasszuk tetszőlegesen, az utolsó h darab (paritásellenőrző) bitet pedig úgy, hogy $Hc = 0$ legyen. Ekkor a kódszavak lineáris alteret alkotnak. Be kell látnunk, hogy két különböző kódszó távolsága legalább három. Vegyük a $c = c_1 - c_2$ nemnulla kódszót, ebben tényleg legalább három 1-es van. Ha csak egy lenne, akkor H valamelyik oszlopa 0 lenne, ha pedig kettő lenne, akkor H két oszlopa megegyezne.

Jelölje a vett sorozatot v. Ha $Hv = 0$, akkor v kódszó, és annak is dekódoljuk. Ha $Hv = He_i$ a H i-edik oszlopa, akkor v nem kódszó, de $v + e_i$ az, és annak dekódoljuk (az i-edik biten hibát jelzünk és javítunk).

A kódunk teljes abban az értelemben, hogy az n hosszú sorozatok között nyilván legfeljebb $2^n/(n+1)$ olyat találunk, hogy a minimális távolságuk legalább három legyen. Azaz a kódszavak számára a korlát

$$\frac{2^{2^h-1}}{2^h} = 2^{2^h-h-1},$$

amit kódunk el is ér.

Egy hiba javítása nyilván nem elég a gyakorlatban. Például a p hibavalószínűségű bináris szimmetrikus csatorna esetén egy n hosszú blokkban átlagosan np hiba történik, tehát ennyit kellene kijavítani, míg a kapacitás eléréséhez $n(1-H(p))$ információs bitre lenne szükség. Az első olyan kódot, mely egy n hosszú blokkban βn információs bitet tartalmaz és αn hibát tud javítani (valamely $\alpha, \beta > 0$ értékekre), Justesen találta 1972-ben.

10. Differenciális entrópia

Ebben a szakaszban megpróbáljuk a diszkrét esetben definiált információelméleti mennyiségeket átvinni az abszolút folytonos esetre. Legyen X n-dimenziós abszolút folytonos eloszlású valószínűségi változó, sűrűségfüggvénye $f(x)$. Jelölje a sűrűségfüggvény tartóját $S_f = \{x : f(x) > 0\}$. Két természetes lehetőség kínálkozik az entrópia definiálására: a diszkrét definícióban a valószínűség helyére írjuk a sűrűségfüggvényt, vagy a diszkretizált eloszlások entrópiáinak vegyük a határértékét. Az első lehetőséget választva (ezt a definíciót is Shannon vezette be):

10.1. Definíció. *Az X valószínűségi változó* differenciális entrópiája

$$h(X) = E(-\log f(X)) = -\int_{\mathbb{R}^n} f(x)\log f(x)dx = -\int_{S_f} f(x)\log f(x)dx,$$

ha ez létezik. (Konvenciónk továbbra is az, hogy $0\log 0 = 0$.)

Vizsgáljuk azonban meg a második lehetőséget is! A teret bontsuk Δ élhosszúságú kockákra. Az X^Δ diszkretizált változó egy lehetséges értéke legyen $k\Delta = (k_1\Delta, k_2\Delta, \ldots, k_n\Delta)$, melynek valószínűsége

$$\int_{k\Delta}^{(k+1)\Delta} f(x)\mathrm{d}x = \Delta^n f^*(k\Delta).$$

Az X^Δ diszkrét valószínűségi változó entrópiája (bár nem véges az értékkészlet, a definíció kiterjeszthető erre az esetre is):

$$H(X^\Delta) = -\sum_k \Delta^n f^*(k\Delta)\log[\Delta^n f^*(k\Delta)] = -n\log\Delta\sum_k \Delta^n f^*(k\Delta) - \Delta^n\sum_k f^*(k\Delta)\log f^*(k\Delta) = -n\log\Delta - \Delta^n\sum_k f^*(k\Delta)\log f^*(k\Delta).$$

Az eredmény második tagja a $-\int f(x)\log f(x)$ integrál integrálközelítő összege, így ha a sűrűségfüggvény elég szép, akkor

$$H(X^\Delta) + n\log\Delta \to h(X), \quad \Delta \to 0.$$

Tehát a diszkretizáltak entrópiája végtelenhez tart, de hozzáadva az $n\log\Delta$ mennyiséget, már véges a határérték. Ha például X egydimenziós, és m bináris törtjegy pontosságig szeretnénk leírni ($\Delta = 2^{-m}$), ahhoz nagyjából $h(X) + m$ bitre van szükség.

10.2. Példa. Ha X egyenletes eloszlású az (a, b) intervallumon, akkor $h(X) = \log(b-a)$. Ebből láthatjuk, hogy a differenciális entrópia negatív is lehet. Ha például X egyenletes a $(0, 1/2)$ intervallumon, akkor $h(X) = -1$, és X első m bináris törtjegyének leírásához valóban $m-1$ bit kell (mivel az első bináris törtjegye biztosan 0).

Vizsgáljuk meg, hogyan változik a differenciális entrópia lineáris transzformáció során!

10.3. Tétel. *Legyen X n-dimenziós, abszolút folytonos valószínűségi változó, A $n\times n$ méretű invertálható mátrix, c pedig n dimenziós vektor. Ekkor*

$$h(AX + c) = h(X) + \log|\det A|.$$

Bizonyítás. Ismert, hogy a sűrűségfüggvények közötti összefüggés

$$f_{AX+c}(x) = \frac{1}{|\det A|} f_X(A^{-1}(x-c)).$$

Ezért

$$h(AX + c) = E\left(-\log f_{AX+c}(AX + c)\right) = E\left(-\log\frac{1}{|\det A|} f_X(X)\right) = \log|\det A| + h(X).$$

■

A differenciális entrópia definíciójából triviálisan látszik, hogy ha az $X = (X_1, \dots, X_n)$ változó koordinátái függetlenek, akkor $h(X) = \sum_{i=1}^n h(X_i)$.

10.4. Példa. Számítsuk ki $X \sim N_n(m, \Sigma)$ differenciális entrópiáját! Legyen először $Y \sim N(0,1)$ egydimenziós standard normális eloszlású. $h(Y) = E(-\log f(Y)) = \frac{1}{\ln 2} E(-\ln f(Y))$. Tovább számolva,

$$E(-\ln f(Y)) = \int_{-\infty}^{\infty} (-\ln \varphi(y))\varphi(y)\mathrm{d}y = \int_{-\infty}^{\infty} (\ln\sqrt{2\pi} + y^2/2)\varphi(y)\mathrm{d}y = \ln\sqrt{2\pi} + 1/2 = \ln\sqrt{2\pi e}.$$

Tehát $h(Y) = \log\sqrt{2\pi e} \approx 2{,}05$. Ha most $Y \sim N_n(0, I_n)$, akkor a példa előtti megjegyzés szerint $h(Y) = \frac{1}{2}\log(2\pi e)^n$. Legyen most $X = \Sigma^{1/2}Y + m$. A korábban kiszámoltak szerint

$$h(X) = h(Y) + \log|\det(\Sigma^{1/2})| = \frac{1}{2}\log\left((2\pi e)^n \det\Sigma\right).$$

Mielőtt rátérnénk a többi információelméleti mennyiség rövid tárgyalására, ejtsünk pár szót a tipikusságról! Legyen most $\mathbb{X}$ emlékezet nélküli, stacionárius, abszolút folytonos forrás, azaz X_i független, azonos eloszlású, egydimenziós változók $f(x)$ sűrűségfüggvénnyel.

10.5. Definíció. *A tipikus halmaz*

$$A_\varepsilon^{(n)} = \{x^n : |-\frac{1}{n}\log f_n(x^n) - h(X_1)| < \varepsilon\} = \{x^n : 2^{-n(h(X_1)+\varepsilon)} \le f_n(x^n) \le 2^{-n(h(X_1)-\varepsilon)}\},$$

ahol $f_n(x^n) = \prod_i f(x_i)$ az együttes sűrűségfüggvény.

A nagy számok törvénye és a definíció közvetlen következményeként kapjuk a tipikus halmaz valószínűségére és térfogatára a következő becsléseket.

10.6. Tétel. *Elég nagy n-re $P(A_\varepsilon^{(n)}) \ge 1-\varepsilon$ és $V(A_\varepsilon^{(n)}) \ge (1-\varepsilon)2^{n(h(X_1)-\varepsilon)}$, továbbá minden n-re $V(A_\varepsilon^{(n)}) \le 2^{n(h(X_1)+\varepsilon)}$.*

Tehát nagy n esetén az $X^n = (X_1, \dots, X_n)$ valószínűségi változó majdnem teljesen egy körülbelül $2^{nh(X_1)}$ térfogatú térrészre koncentrálódik, és ott közel egyenletes eloszlású. A standard normális eloszlás esetében például azt kapjuk, hogy nagy n-re az n-dimenziós standard normális eloszlású valószínűségi változó egy $4{,}13^n$ térfogatú tartományon koncentrálódik.

10.7. Definíció. *Legyen (X,Y) abszolút folytonos eloszlású. A feltételes entrópia*

$$h(X|Y) = h(X,Y) - h(Y) = E(-\log f(X|Y)).$$

10.8. Definíció. *Legyen f, g két sűrűségfüggvény ugyanazon a téren. Divergenciájuk*

$$D(f\|g) = \int_{S_f} f(x)\log\frac{f(x)}{g(x)}\mathrm{d}x = E_f\left(\log\frac{f(X)}{g(X)}\right),$$

ahol az E_f jelölés azt jelenti, hogy az X változó sűrűségfüggvénye f. Ha f nem abszolút folytonos g-re, akkor a divergenciát végtelennek definiáljuk.

Könnyű látni, hogy a divergencia nemnegatív, hiszen a Jensen egyenlőtlenségből

$$-D(f\|g) = E_f\left(\log\frac{g(X)}{f(X)}\right) \le \log E_f\left(\frac{g(X)}{f(X)}\right) = \log\int_{S_f} f(x)\frac{g(x)}{f(x)} \le \log 1 = 0.$$

Továbbá az is kiderült, hogy $D(f\|g) = 0$ csak akkor, ha $f = g$ majdnem mindenütt. A divergencia segítségével definiálhatjuk a kölcsönös információt, ugyanúgy, mint a diszkrét esetben.

10.9. Definíció. *Legyen (X,Y) abszolút folytonos, sűrűségfüggvénye $k(x,y)$. Legyenek a peremsűrűségfüggvények $f(x)$ és $g(y)$. X és Y kölcsönös információja*

$$I(X \wedge Y) = D(k(x,y)\|f(x)g(y)) = h(X) + h(Y) - h(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X).$$

Itt az első kifejezést tekinthetjük definíciónak, a másik három pedig egyszerű átalakítással adódik. Érvényben marad tehát, hogy a kölcsönös információ nemnegatív (és csak független változókra lehet nulla), valamint a feltételes entrópia legfeljebb akkora, mint a feltétel nélküli (egyenlőség megint csak függetlenség esetén van). Az entrópiára vonatkozó láncszabály is teljesül, ez szintén triviális számolással látható:

$$h(X_1,\dots,X_n)=\sum_{i=1}^{n} h(X_i|X^{i-1}).$$

Nézzük meg, hogyan változik a kölcsönös információ lineáris transzformáció esetén! Legyen (X,Y) abszolút folytonos, n, illetve m dimenziósak, A,C megfelelő méretű nemelfajuló négyzetes mátrixok, b,d vektorok. Az

$$\begin{pmatrix} AX+b \\ CY+d \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} b \\ d \end{pmatrix}$$

felírásból

$$I(AX+b \wedge CY+d) = h(AX+b)+h(CY+d)-h(AX+b,CY+d) = \\ h(X)+\log|\det(A)|+h(Y)+\log|\det(B)|-(h(X,Y)+\log|\det(A)\det(C)|)=I(X\wedge Y).$$

10.10. Példa. Legyen (X,Y) kétdimenziós normális eloszlású, a koordináták korrelációs együtthatója $R(X,Y)=\rho$. Határozzuk meg X és Y kölcsönös információját! Az előző számolás szerint standardizálhatjuk a koordinátákat, azaz feltehető, hogy X és Y is standard normális.

$$h(X)=h(Y)=\frac{1}{2}\log(2\pi e),\quad h(X,Y)=\frac{1}{2}\log((2\pi e)^2(1-\rho^2)).$$

Innen

$$I(X\wedge Y)=h(X)+h(Y)-h(X,Y)=-\frac{1}{2}\log(1-\rho^2).$$

Végül megjegyezzük, hogy a kölcsönös információ diszkretizálással is definiálható lenne. Ha ugyanis X n-dimenziós, Y m-dimenziós, akkor

$$I(X^\Delta\wedge Y^\Delta)=H(X^\Delta)+H(Y^\Delta)-H(X^\Delta,Y^\Delta)\sim h(X)-m\log\Delta+h(Y)-n\log\Delta-h(X,Y)+(n+m)\log\Delta,$$

vagyis a $\log\Delta$ tagok kiesnek. Ez általánosabban is igaz, vagyis tetszőleges valószínűségi változók kölcsönös információját definiálhatjuk úgy, hogy a véges partíciókon vett diszkretizáltjaik kölcsönös információinak szuprémumát vesszük.

11. A Gauss csatorna

Ebben a szakaszban az emlékezet nélküli, additív Gauss zajú csatornával foglalkozunk. Ha tehát X_i az i-edik bemenő jel, akkor a hozzá tartozó kimenő jel $Y_i=X_i+Z_i$, ahol a Z_i változók független, azonos eloszlásúak, mégpedig $Z_i\sim N(0,\sigma^2)$. Szeretnénk a csatorna kapacitását meghatározni. Hogyan definiáljuk a kapacitást? Próbálkozzunk először a diszkrét definíció analógjával, azaz legyen

$$C=\max_{f(x)} I(X\wedge Y).$$

Tehát olyan abszolút folytonos bemeneti eloszlást keresünk, melyre a bemenet és a kimenet kölcsönös információja a lehető legnagyobb. Számoljuk ki a kölcsönös információt!

$$I(X\wedge Y)=h(Y)-h(Y|X)=h(Y)-h(X+Z|X)=h(Y)-h(Z|X)=h(Y)-h(Z). \tag{2}$$

Látszik, hogy ez akármilyen nagy lehet, hiszen ha pl. $X\sim N(0,P)$ valamilyen P-re, akkor $Y\sim N(0,P+\sigma^2)$, és

$$I(X\wedge Y)=\frac{1}{2}\log 2\pi e(P+\sigma^2)-\frac{1}{2}\log 2\pi e\sigma^2=\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right).$$

Tehát valamilyen megkötést kell tennünk. Tegyünk felső korlátot a bemenet második momentumára, vagy (ami ezzel ekvivalens) a szórásnégyzetére!

11.1. Definíció. *Az emlékezet nélküli abszolút folytonos csatorna kapacitása P teljesítménykorlát (power) mellett* $C(P) = \max_{f(x):E(X^2)\leq P} I(X \wedge Y)$.

A (2) egyenlet szerint tehát $C(P)$ meghatározásához a $h(Y)$ entrópiát kell maximalizálni úgy, hogy

$$E(Y^2) = E((X+Z)^2) = E(X^2) + E(Z^2) + 2E(X)E(Z) = E(X^2) + E(Z^2) \leq P + \sigma^2.$$

Mivel a kölcsönös információ eltolásinvariáns, feltehető, hogy a bemeneti eloszlásra $E(X) = 0$, így a 11.2. Lemmát felhasználva $h(Y) \leq \frac{1}{2}\log 2\pi e(P+\sigma^2)$. Ráadásul a felső korlát el is érhető, ha az $X \sim N(0,P)$ bemeneti eloszlást választjuk. A vizsgált Gauss-csatorna kapacitása tehát (behelyettesítve a $h(Z) = \frac{1}{2}\log 2\pi e\sigma^2$ értéket)

$$C(P) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right).$$

11.2. Lemma. *Legyen X n-dimenziós, abszolút folytonos eloszlású, $E(X) = 0$ (ez mellékes feltétel), $\Sigma(X) = K$, ahol K rögzített szimmetrikus, pozitív definit mátrix. Ezen feltételek mellett a $h(X)$ entrópiát a normális eloszlás maximalizálja.*

Bizonyítás. Legyen $g(x)$ az X sűrűségfüggvénye, $\varphi_K(x)$ pedig a megfelelő normális sűrűségfüggvény.

$$0 \leq D(g\|\varphi_K) = \int g(x)\log\frac{g(x)}{\varphi_K(x)} = -h(X) - \int(\log\varphi_K(x))g(x).$$

A második tag $\log\varphi_K(X)$ g-szerinti várható értéke. Azonban $\log\varphi_K(x) = c_1 - c_2 x^T K^{-1} x$, ennek várható értéke csak g kovarianciamátrixától függ, ami a feltevés szerint K. Így

$$-\int(\log\varphi_K(x))g(x) = -\int(\log\varphi_K(x))\varphi_K(x) = \frac{1}{2}\log((2\pi e)^n \det K).$$

■

Most megmutatjuk, hogy a csatornakapacitás definiálható az elérhető ráták szuprémumaként is. Ismét $(2^{nR}, n)$ kódokat vizsgálunk, tehát n hosszú kódszavakból álló, 2^{nR} elemszámú kódkönyveket keresünk, melyekre még a P-feltétel is teljesül: minden x^n kódszóra $\frac{1}{n}\sum_{i=1}^n x_i^2 \leq P$. Az R ráta akkor elérhető, ha létezik $(2^{nR}, n)$ kódok olyan sorozata, hogy a maximális hibavalószínűség nullához tart.

11.3. Tétel. *A vizsgált Gauss-csatornára az elérhető ráták szuprémuma a fent kiszámolt $C(P)$.*

Bizonyítás. A bizonyítás nagyon hasonlít a diszkrét esethez.

Először megmutatjuk, hogy az $R < C(P)$ ráta elérhető. Ennek érdekében generáljunk egy véletlen $(2^{nR}, n)$ kódot, a kódszavak betűi legyenek $N(0, P-\varepsilon)$ eloszlásúak (és minden független egymástól). Itt $\varepsilon > 0$ majd megfelelően kicsi lesz. Jelölje az i-edik kódszót $X^n(i) = (X_1(i), \ldots, X_n(i))$. A dekódolást a jól ismert tipikusság szerint végezzük. Egyetlen különbség, hogy a dekódoló akkor is hibát jelez, ha pontosan egy olyan kódszó van, amely a kimenettel együttesen tipikus, viszont ez a kódszó nem teljesíti a P-feltételt. Az átlagos hiba várható értékének kiszámításához megint elég az első üzenetet figyelembe venni. Jelölje $Y^n(1)$ az első üzenet által produkált kimenetet. Ekkor

$$E(P_e^{(n)}) = P(E_0 \cup \overline{E}_1 \cup (\cup_{i>1} E_i)),$$

ahol

$$E_0 = \left\{\frac{1}{n}\sum_{j=1}^n X_j^2(1) > P\right\}, E_i = \left\{(X^n(i), Y^n(1)) \in A_\varepsilon^{(n)}(X,Y)\right\}.$$

$P(E_0)$ nullához tart a nagy számok törvénye miatt, és $P(E_1)$ egyhez tart. A maradék 2^{nR} esemény valószínűségére a diszkrét esetben látott becslés érvényes:

$$P(E_i) \leq 2^{-n(I(X\wedge Y)-3\varepsilon)}.$$

Mivel $I(X \wedge Y) = C(P-\varepsilon)$, és $R < C(P)$, így elég kis ε-ra $E(P_e^{(n)}) \to 0$. Kiválasztva egy olyan kódsorozatot, melyre $P_e^{(n)} \to 0$, a kódszavak jobbik felét megtartva ezekre a maximális hiba is nullához tart, és teljesítik a P-feltételt (különben $\lambda_i = 1$ lenne).

Most fordítva, tegyük fel, hogy R elérhető, és legyen a W üzenet eloszlása egyenletes. Az $R \le C$ egyenlőtlenséget hasonlóan bizonyítjuk, mint korábban, azaz a

$$H(W) = nR = H(W|\hat{W}) + I(W \wedge \hat{W})$$

felírás első tagjára a Fano-egyenlőtlenséget, második tagjára az adatfeldolgozási egyenlőtlenséget alkalmazzuk. Tehát az $I(X^n \wedge Y^n)$ tagot kell becsülni:

$$I(X^n \wedge Y^n) = h(Y^n) - h(Y^n|X^n) = h(Y^n) - h(Z^n) \le \sum_{i=1}^{n} h(Y_i) - h(Z^n).$$

Itt $h(Y_i) \le \frac{1}{2}\log(2\pi e(P_i + \sigma^2))$, ahol $P_i = E(X_i^2)$. Ezért

$$I(X^n \wedge Y^n) \le \sum_{i=1}^{n} \frac{1}{2}\log(1 + P_i/\sigma^2) \le \frac{n}{2}\log(1 + \frac{\sum P_i/n}{\sigma^2} \le \frac{n}{2}\log(1 + P/\sigma^2).$$

■