Teoria de la Informació

Francesc Tiñena

Departament de Matemàtiques, UPC

Assignatura TCI (FIB), curs 2025-26q1

Missatges

- Pensem en sistemes de comunicació quotidians: telèfon, radio, T.V., internet
- Aquests sistemes porten informació des de la font que la produeix fins una destinació situada a certa distància
- Tot sistema de comunicacions té la mateixa funció bàsica: transmetre informació
- El concepte d'informació és un concepte difícil de precisar. Per evitar aquesta dificultat parlem de *missatges*. Un missatge és la manifestació física de la informació produïda per la font
- Amb aquest llenguatge, l'objectiu de tot sistema de comunicacions és reproduir a destinació una rèplica acceptable del missatge generat per la font

Analògics vs Digitals

- Podem distingir dues categories de missatges:
  - analògics
  - digitals
- Un missatge analògic és una magnitud física dependent del temps (exemple: la tensió elèctrica $v(t)$ que hi ha a la sortida d'un micròfon quan algú està parlant davant d'ell)
- La informació està en la ona $v(t)$, per tant els sistemes analògics de comunicacions han de lliurar l'ona amb un grau especificat de fidelitat
- Un missatge digital és una seqüència ordenada de símbols seleccionats d'un conjunt finit, com per exemple les lletres impreses d'un diari. Com que la informació està en els símbols discrets, els sistemes digitals de comunicacions han de lliurar aquests símbols amb un grau especificat d'exactitud, en un interval de temps fixat

Limitacions dels sistemes físics

- En els sistemes elèctrics de transmissió d'informació, els missatges seran tensions o corrents que variaran amb el temps
- Tot sistema elèctric conté elements que enmagatzemen energia, energia que no es pot canviar instantàniament
- Tot sistema físic té unes limitacions en el sentit que no podrà seguir variacions massa ràpides del senyal d'entrada: tot sistema físic és *limitat en banda*
  - La comunicació en temps real necessita una certa amplada de banda, sinó hi haurà *distorsió*: la transmissió digital a r símbols/seg necessita una amplada de banda $B \geq r/2$ Hz
- *Soroll*. Els electrons dels àtoms de qualsevol conductor a temperatura $T > 0^0 K$ tenen un moviment aleatori, moviment que produeix petites fluctuacions de tensió. Aquestes fluctuacions s'afegeixen al senyal útil i si la relació senyal-soroll és massa baixa no serà possible reconstruir el senyal enviat

Què és la teoria de la Informació? (1)

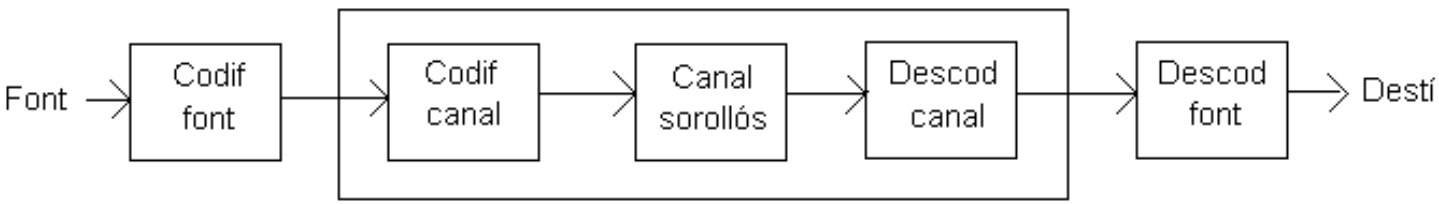
- Claude Shannon va estudiar el problema següent: donada una font productora de missatges, missatges que no escollim nosaltres, com s'han de representar aquests missatges per obtenir una transmissió fiable sobre un canal amb les seves inherents limitacions físiques (amplada de banda, soroll)?
- Shannon es va fixar en la informació del missatge més que en els senyals. Aquest punt de vista aviat va rebre el nom de *teoria de la informació*
- La teoria de la informació tracta amb tres conceptes bàsics:
  - mesura de la informació de la font
  - capacitat (d'informació) d'un canal
  - representació del missatge (codificació) per utilitzar la capacitat del canal per transmetre informació

Què és la teoria de la Informació? (2)

- El resultat bàsic de Shannon es pot enunciar de la manera següent:
  - Si la taxa d'informació de la font és menor que la capacitat del canal llavors existeixen tècniques de codificació de manera que la informació pot ser transmesa sobre el canal amb una freqüència arbitràriament baixa d'errors, amb independència de la presència de soroll.
- El més sorprenent d'aquest resultat és l'afirmació de transmissió sense errors sobre un canal sorollós, fet que s'assoleix amb l'ajut de la codificació
- La codificació serveix per *adaptar* la font al canal per obtenir una transmissió (confiable) d'informació màxima

Codificació de font

- El procés de codificació generalment involucra dues accions diferents de codificació/descodificació:



Canal equivalent sense soroll

- La codificació de canal serveix per controlar (detectar/corregir) els errors. El sistema òptim de «codificador de canal–canal sorollós–descodificador de canal» equival a un canal sense soroll amb una capacitat (per transmetre informació) ben definida
- El codificador de font serveix per adaptar la font al canal equivalent sense soroll (suposant que la taxa de producció d'informació estigui dins de la capacitat del canal)

Temari del curs

- El temari del present curs es divideix en dos blocs:
  - Teoria de la informació: mesura de la informació produida per una font i codificació (de font) dels missatges per a ser transmesos de manera eficient sobre un canal sense errors
  - Codis detectors i correctors d'errors (codificació de canal) per aumentar la fiabilitat de la transmissió sobre un canal amb errors

Fonts (d'informació) discretes i contínues

- Qualsevol font d'informació produeix una sortida que no es coneix a priori, una sortida que és aleatòria (si es coneixés de manera exacta, no hi hauria cap necessitat de transmetre-la)
- Per tant, la sortida d'una font s'ha de caracteritzar en termes estadístics
- Segons els missates que una font pot generar, podem dividir les fonts en discretes i contínues
  - Una font discreta produirà una seqüència de símbols o lletres pertanyent a un alfabet a raó de tantes lletres per segon
  - Una font contínua produirà un senyal variant en el temps $x(t)$
- Ens centrarem en la comunicació digital

La FDSM (1)

- La font discreta més senzilla és la que emet una seqüència de lletres seleccionades d'un alfabet finit i on la seqüència de sortida $a_1, a_2, a_3, \ldots$ és estadísticament independent: aquesta és la font discreta sense memòria (FDSM)
- El "sense memòria" prové de que la probabilitat d'emissió d'un símbol no depén dels símbols emesos amb anterioritat
  - Evidentment la FDSM és un model molt senzill al qual no s'hi ajusten moltes situacions habituals: una font que emet missatges en llengua catalana, per exemple, no s'adapta a aquest model (si la primera lletra d'una paraula és "q", la lletra següent és "u"; l'efecte de memòria és evident)

La FDSM (2)

- Per descriure una FDSM cal l'alfabet i les probabilitats

Example

Alfabet: $A = \{a, b, c, d\}$ $\quad p(a) = p(b) = p(c) = p(d) = 0.25$

- Des d'un punt de vista matemàtic, el primer símbol produït per la font serà una variable aleatòria x_1; el segon símbol produït per la font, x_2, serà una altra v.a. que pot prendre els mateixos valors que x_1 i amb les mateixes probabilitats. I independent amb x_1. I així successivament
- La seqüència de símbols produïts per la font, x_1, x_2, x_3, x_4, ..., és una successió de v.a. independents idènticament distribuïdes
  - En el cas de que no hi hagués independència entre lletres emeses el model matemàtic resultant seria força més complicat. Per això, tenint en compte el caràcter introductori de l'assignatura, no tractarem cap més font que la FDSM

Transmisió eficient de dades (1)

- Un aspecte pràctic molt important és intentar aconseguir una velocitat de transmissió màxima o quasi màxima
- Suposem una FDSM que genera una lletra cada τ_s segons i un canal físic concret pel qual ha de "viatjar" la seqüència generada per la font (per exemple: un parell de fils elèctrics paral.lels)
- És clar que el que "viatja" pel parell de fils elèctrics no és cap lletra sinó un senyal, per exemple, de tensió
- Cal associar a cada lletra de l'alfabet un senyal elèctric concret, senyal susceptible de "viatjar" pel canal

Transmisió eficient de dades (2)

- Fixem, per centrar idees, un exemple concret:

L'alfabet de la nostra FDSM serà $A = \{a, b, c, d\}$. Associarem, respectivament, les lletres "a, b, c, d" als senyals de tensió

$$g_a(t) = \begin{cases} 1 & \text{si} \quad 0 \leq t \leq \tau_s \\ 0 & \text{alt} \end{cases} \qquad g_b(t) = \begin{cases} -1 & \text{si} \quad 0 \leq t \leq \tau_s \\ 0 & \text{alt} \end{cases}$$

$$g_c(t) = \begin{cases} 3 & \text{si} \quad 0 \leq t \leq \tau_s \\ 0 & \text{alt} \end{cases} \qquad g_d(t) = \begin{cases} -3 & \text{si} \quad 0 \leq t \leq \tau_s \\ 0 & \text{alt} \end{cases}$$

Transmisió eficient de dades (3)

- Cal entendre que si a l'instant $t = 0$ la font genera el símbol "b", llavors el senyal que s'envia pel canal és $g_b(t)$
- Suposem que, a continuació, és a dir, a l'instant $t = \tau_s$, la font produeix la lletra "c"; en aquest cas cal enviar pel canal el senyal g_c, això sí, desplaçat τ_s unitats en el temps: $g_c(t - \tau_s)$
- Si la lletra que la font produeix a continuació (a l'instant $t = 2\tau_s$) és la lletra "d", el que caldrà enviar pel canal és g_d desplaçada en el temps $2\tau_s$ unitats: $g_d(t - 2\tau_s)$
- Representem gràficament el senyal a enviar pel canal si la seqüència generada per la font és "a, a, d, b, a":

$$\begin{cases} 1 & \text{si} \quad 0 \leq t \leq \tau_s \\ 1 & \text{si} \quad \tau_s \leq t \leq 2\tau_s \\ -3 & \text{si} \quad 2\tau_s \leq t \leq 3\tau_s \\ -1 & \text{si} \quad 3\tau_s \leq t \leq 4\tau_s \\ 1 & \text{si} \quad 4\tau_s \leq t \leq 5\tau_s \end{cases}$$

Transmisió eficient de dades (4)

- Per expressar de manera analítica l'anterior senyal ens resultarà còmode disposar del "pols bàsic"

$$g(t) = \begin{cases} 1 & \text{si} \quad 0 \leq t \leq \tau_s \\ 0 & \text{alt} \end{cases}$$

- Amb aquesta notació, i canviant, per comoditat, a, b, c, d per $1, -1, 3, -3$ el senyal que s'enviarà pel canal és:

$$s(t) = \sum_{n=0}^{4} x_n g(t - n\tau_s)$$

on els x_n són la seqüència de lletres produïda per la font

- Evidentment aquest raonament no està limitat a una producció de 5 símbols. Si la seqüència produïda per la font és $x_0, x_1, x_2, \ldots$ llavors el senyal a transmetre serà

$$s(t) = \sum_{n \geq 0} x_n g(t - n\tau_s)$$

Transmisió eficient de dades (5)

- Quina és la màxima velocitat a la que podem transmetre? (quin és el mínim temps de símbol τ_s que podem utilitzar?)
- El senyal a l'altre extrem del canal (parell de fils elèctrics) no serà idèntic al senyal injectat al canal
- A partir de certa freqüència (invers del temps de símbol) no serem capaços de reconeixer l'ona que arribarà
- Per altres tipus de canals les freqüències màximes utilitzables seràn diferents però el problema seguirà existint
- En tot canal real hi ha un mínim temps de símbol permès. Però en lloc d'enviar les lletres $\{a, b, c, d\}$ en forma de nivells de tensió $\pm 1, \pm 3$ podem enviar blocs de dues lletres

$$aa, ab, ac, ad, ba, bb, bc, bd, ca, cb, cc, cd, da, db, dc, dd$$

representant cada un d'aquests blocs per un dels nivells

$$\pm 1, \pm 3, \pm 5, \pm 7, \pm 9, \pm 11, \pm 13, \pm 15$$

Transmisió eficient de dades (6)

- A una mateixa velocitat de senyalització ($1/\tau_s$ símbols de canal per segon) es pot aconseguir transmetre molts símbols de font per segon
- Però a la pràctica això no es pot fer. Pel soroll!
- Soroll és tota forma indesitjada d'ona de natura aleatòria que s'afegeix al senyal útil en el seu pas pel canal (ex: soroll tèrmic)
- Com que la potència transmesa per tot canal real és limitada, és la relació senyal/soroll la que, juntament amb l'amplada de banda disponible, determina les prestacions del canal
- Quines són les "prestacions" màximes que es poden aconseguir d'un canal i com cal procedir per, més o menys, obtenir-les?
- Aquest problema va ser resolt per Shannon (1948), qui va introduir el concepte de *capacitat d'un canal*, concepte relacionat amb la capacitat de transmetre *informació* que té un canal

Informació (1)

- Quanta "informació" porta un missatge?
- Suposem que estem de vacances fora de Catalunya i escoltem, per televisió, el temps que ha fet durant el dia a Barcelona. El temps pot haver estat, per exemple, un dels tres que segueixen:
  - Ha fet sol
  - Ha plogut
  - Han bufat vents a 200 km/h
- Quin dels tres missatges porta més informació?
  - Sembla clar que el missatge "ha fet sol" porta molt poca informació: a Barcelona això és el que suposem que passa més o menys cada dia
  - La segona previsió porta més informació ja que, a Barcelona, hi ha molts menys dies que plou que no pas dies que fa sol
  - Finalment, la tercera previsió és la que porta més informació, és la que descriu, de les tres anteriors, el fenòmen menys probable

Informació (2)

- El concepte d'informació està relacionat amb la probabilitat d'ocurrència d'un succés: com més probable és el succés, menys informació porta el comunicat de que ha succeït
  - Per exemple, l'afirmació, relativa al sorteig de la Grossa, de que la grossa ha acabat en una xifra imparell porta menys informació que l'afirmació de que la grossa ha acabat en 37
- Seguim amb aquest exemple, però situem-nos moments abans del sorteig. I suposem que hem comprat un dècim
- Abans del sorteig tenim una incertesa sobre el nombre que sortirà, igual que, abans de tirar un dau a l'aire, tenim una incertesa sobre el resultat que es produirà. En quin dels dos casos tenim més incertesa sobre el resultat? Hi ha molta més incertesa en el nombre que sortirà premiat a la Grossa que en el nombre que sortirà del dau (en el dau sortirà 1,2,3,4,5 o 6, mentre que en la Grossa poden sortir molts nombres)

Informació (3)

- El que hem fet fins ara és parlar d'incertesa i d'informació. Abans de realitzar un experiment hi ha incertesa i després de realitzar l'experiment hi ha informació
- Des d'un punt de vista formal els conceptes d'incertesa i d'informació són equivalents, només que un és abans de realitzar l'experiment i l'altra després de realitzar l'experiment
- Volem definir el concepte d'informació (o d'incertesa). Hem començat amb una anàlisi qualitativa de la situació intentant respondre a la pregunta: què porta més informació (on hi ha més incertesa)
- A continuació caldrà passar a aspectes quantitatius: quanta informació porta un enunciat sobre el resultat d'un experiment? (quanta incertesa hi ha sobre el resultat d'un experiment?). I, en quines unitats es mesura la informació/incertesa?

Informació (4)

- La quantitat d'informació que porta el missatge "A" depen únicament de la probabilitat de que la font emeti el missatge "A"
- El concepte d'informació no té res a veure amb el concepte de significat (un missatge pot no tenir cap significat i portar molta informació i, al revés, un missatge pot estar carregat de significat i portar molt poca informació)
- Per evitar pensar en significats podem associar missatges amb successos (relacionats amb un experiment aleatori): la informació que porta un missatge m_i depén únicament de la probabilitat de que es produeixi el missatge m_i: $I(m_i) = F(p(m_i))$
- F ha de ser una funció decreixent i si m_i i m_j són missatges (successos) independents llavors la informació que porta el missatge (succés) "$m_i \wedge m_j$" ha de colincidir amb la suma d'informacions de m_i i de m_j:

$$m_i, m_j \text{ independents} \implies I(m_i \wedge m_j) = I(m_i) + I(m_j)$$

Informació (5)

- Tindrem:

$$I(m_i \wedge m_j) = F(p(m_i \wedge m_j)) = F(p(m_i) \cdot p(m_j))$$

  i, per tant, la funció F que estem buscant ha de complir:

$$F(p(m_i) \cdot p(m_j)) = F(p(m_i)) + F(p(m_j))$$

- Necessitem una funció F tal que $F(p_i \cdot p_j) = F(p_i) + F(p_j)$
- Això justifica l'elecció de la funció logaritme, canviada de signe perque sigui decreixent, per a definir la quantitat d'informació (diapositiva següent)

Informació (6)

Definition

$$I(m_i) = -\log_a(p(m_i)) = \log_a \frac{1}{p(m_i)}$$

- La base que s'agafa habitualment per definir la quantitat d'informació és $a = 2$. La quantitat d'informació es mesura en bits:

$$I(m_i) = -\log_2(p(m_i)) = \log_2 \frac{1}{p(m_i)} \text{ bits}$$

- Pels informàtics la paraula "bit" és una contracció de "BInary digiT" i té el significat de dígit binari: $0, 1$. En teoria de la informació, el bit és una mesura de quantitat d'informació i, encara que més endavant ja tindrem ocasió de relacionar ambdós conceptes, de moment quan parlarem de bits estarem parlant de bits d'informació (quantitat d'informació). Per referir-nos als dígits binaris no usarem cap tipus de contracció

Exemples (1)

Examples

Suposem una font que produeix un dígit binari cada τ_s segons. Suposem també que cada símbol és independent dels altres i que la probabilitat d'emissió del 0 és igual a la probabilitat d'emissió de l'1. Calculem la informació que porta el missatge '0':

$$I('0') = -\log_2 p('0') = -\log_2 \frac{1}{2} = 1 \text{ bit}$$

Suposem una font com l'anterior però on la probabilitat d'emissió del 1 és triple que la probabilitat d'emissió del 0. Calculem la informació que porten els missatge '0' i '1'. En aquest cas, $p(0) = \frac{1}{4}$, $p(1) = \frac{3}{4}$. Per tant,

$$I('0') = -\log_2 p('0') = -\log_2 \frac{1}{4} = 2 \text{ bits}$$

$$I('1') = -\log_2 p('1') = -\log_2 \frac{3}{4} = 0.415 \text{ bits}$$

Exemples (2)

Example

Suposem la font de l'exemple anterior però amb $p(0) = p$, $p(1) = 1 - p$. Calculem la informació que porta, per terme mig, un missatge de longitud 1, 'x_1':

La quantitat d'informació que porta un missatge concret de longitud 1 és:

$$I('x_1') = -\log_2 p('x_1') = \begin{cases} -\log_2 p & \text{si} \quad x_1 = 0 \\ -\log_2(1-p) & \text{si} \quad x_1 = 1 \end{cases}$$

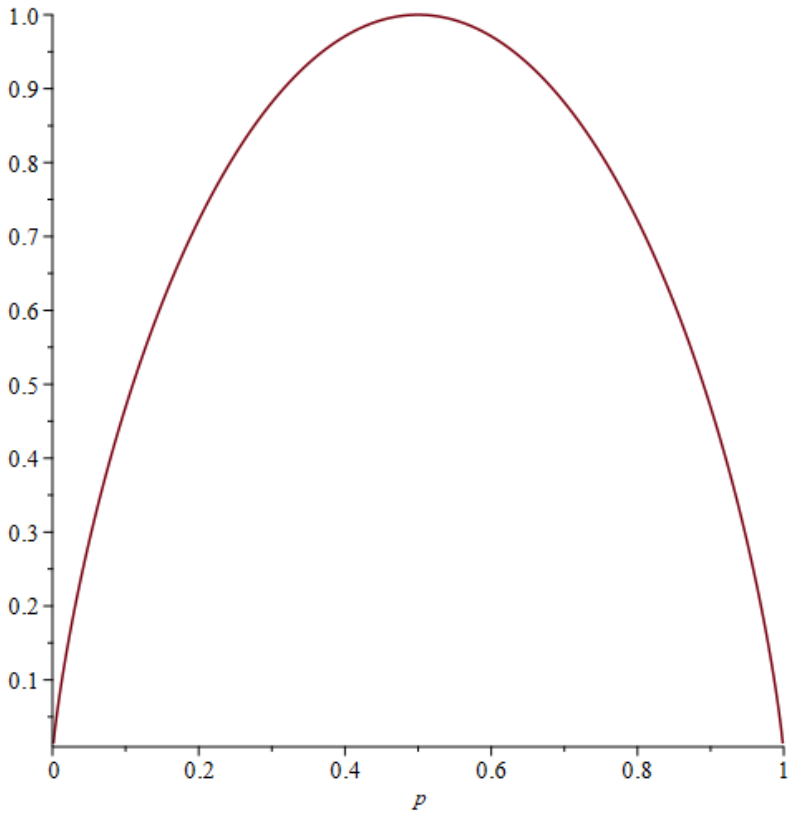
La informació mitjana la trobarem fent la mitjana ponderada de les dues informacions anteriors:

$$\text{Informació mitjana} = p('0')I('0') + p('1')I('1') = -p\log_2 p - (1-p)\log_2(1-p) \text{ bits}$$

Exemples (3)

- Representem gràficament la funció

$$\Omega(p) = -p \log_2 p - (1-p) \log_2 (1-p) :$$



Exemples (4)

- El màxim s'assoleix per $p = 0.5$ i val $\Omega(0.5) = 1$ bit
- La informació mitjana d'una FDSM binària és màxima quan els símbols són equiprobables; en aquest cas la informació generada per la font és de 1 bit/símbol
- S'ha introduït un concepte important, el concepte d'informació mitjana. La paraula tècnica que cal usar per referir-se a la informació mitjana és *entropia*. L'entropia d'una font és la informació mitjana que genera la font, mesurada en bits (d'informació) per símbol de font. L'entropia es designa amb la lletra H i és igual a

$$H = -\sum p(x_i) \log_2 p(x_i)$$

estant la suma estesa a totes les lletres de l'alfabet

Exemples (5)

Example

Suposem una FDSM amb alfabet $A = \{a, b, c, d\}$ i probabilitats $p(a) = p(b) = p(c) = p(d) = \frac{1}{4}$. Calculem l'entropia de la font:

$$H = -\sum p(x_i) \log_2 p(x_i) = -4p(x_1) \log_2 p(x_1) = -4 \times \frac{1}{4} \times \log_2 \frac{1}{4} = 2 \text{ bits/símbol}$$

- Les lletres de l'alfabet es poden codificar en binari així:

$$\begin{aligned} a &\longrightarrow 00 \\ b &\longrightarrow 01 \\ c &\longrightarrow 10 \\ d &\longrightarrow 11 \end{aligned}$$

- En aquesta codificació s'usen 2 dígits binaris per símbol de font: aquesta codificació té una longitud de 2 dígits binaris per símbol de font

Exemples (6)

Example

Suposem una FDSM amb el mateix alfabet però on les probabilitats són ara: $p(a) = \frac{1}{2}$, $p(b) = \frac{1}{4}$, $p(c) = p(d) = \frac{1}{8}$.
Calculem la entropia (informació mitjana) de la font.

$$H = -\sum p(x_i) \log_2 p(x_i) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = 1.75 \text{ bits/símbol}$$

- Té alguna cosa a veure l'entropia de la font amb la longitud d'un possible codi binari per a aquesta font? És a dir: podem codificar els símbols de font (lletres de l'alfabet) usant 1.75 dígits binaris per cada símbol?

Longitud mitjana d'un codi (1)

- Utilitzar un nombre no enter de dígits binaris per símbol pot semblar una mica estrany, però pot tenir sentit: podria ser possible una codificació (binària) on les lletres que apareixen amb més freqüència (les que tenen una probabilitat més gran) es codifiquessin amb menys dígits binaris que les lletres amb una probalitat d'ocurrència més baixa
- Estariem parlant de codis de longitud variable; en aquests casos, la longitud mitjana del codi (mitjana ponderada del nombre de dígits binaris per símbol de font) no té per què ser entera

Longitud mitjana d'un codi (2)

- Estudiem la codificació següent:

$$\begin{aligned} a &\longrightarrow 0 \\ b &\longrightarrow 10 \\ c &\longrightarrow 110 \\ d &\longrightarrow 111 \end{aligned}$$

La longitud mitjana d'aquest codi serà la mitjana ponderada de les longituds amb que es codifica cada un dels símbols:

$$\bar{l} = p(a)l(a) + p(b)l(b) + p(c)l(c) + p(d)l(d) = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 =$$

$$= 1.75 \text{ dígits binaris/símbol de font}$$

Longitud mitjana d'un codi (3)

- En aquest cas és possible aconseguir un codi amb una longitud mitjana igual a l'entropia
- Però no sempre és així, ja ho estudiarem. En tot cas, l'entropia ens donarà un límit que no és possible superar
- La teoria de la informació ens dóna un referent respecte al qual podem valorar les prestacions d'un sistema real de comunicacions
- A la diapositiva següent veurem unes propietats de l'entropia. Per a poder-les enunciar de manera precisa ens serà d'utilitat la notació següent:
- L'entropia, mesurada en bits per símbol, d'una FDSM amb alfabet $\{a_1, a_2, ..., a_L\}$ i probabilitats d'ocurrència $p(a_1) = p_1$, $p(a_2) = p_2, ..., p(a_L) = p_L$ serà designada per $H(p_1, p_2, ..., p_L)$.

Propietats de l'entropia (1)

Teorema

1. $H(p_1, p_2, ..., p_L) \geq 0$
2. $H(p_1, p_2, ..., p_L) \leq H(\frac{1}{L}, \frac{1}{L}, ..., \frac{1}{L}) = \log_2 L$
3. Si $H(p_1, p_2, ..., p_L) = H(\frac{1}{L}, \frac{1}{L}, ..., \frac{1}{L}) = \log_2 L$ llavors $p_1 = p_2 = ... = p_L = \frac{1}{L}$

Proof.

La primera afirmació és clara: cada un dels sumands que apareix a la fórmula de l'entropia és no negatiu. La demostració de la segona afirmació es basa en el resultat de la diapositiva següent □

Lema de Gibbs

Theorem

(Gibbs) Si els $(q_1, q_2, ..., q_L)$ *es corresponen amb una distribució de probabilitat, és a dir, si* $q_1, q_2, ..., q_L \geq 0$ *i* $q_1 + q_2 + ... + q_L = 1$, *llavors*

$$H = -\sum p_i \log_2 p_i \leq -\sum p_i \log_2 q_i$$

(Nota.- Aquesta proposició també és vàlida si $q_1 + q_2 + ... + q_L \leq 1$)

Proof.

És un simple càlcul, només cal utilitzar la coneguda desigualtat $\ln x \leq x - 1$:

$$H + \sum p_i \log_2 q_i = -\sum p_i \log_2 p_i + \sum p_i \log_2 q_i = \sum p_i \log_2 \frac{q_i}{p_i} = \frac{1}{\ln 2} \sum p_i \ln \frac{q_i}{p_i} \leq$$

$$\leq \frac{1}{\ln 2} \sum p_i (\frac{q_i}{p_i} - 1) = \frac{1}{\ln 2} \sum (q_i - p_i) = \frac{1}{\ln 2} \left(\sum q_i - \sum p_i \right) = 0$$

Propietats de l'entropia (2)

- Aplicant aquesta desigualtat a $q_1 = q_2 = ... = q_L = \frac{1}{L}$ s'obté la segona afirmació. Quant a la tercera,

$$H(p_1, p_2, ..., p_L) = \log_2 n \implies -\sum p_i \log_2 p_i = -\sum p_i \log_2 \frac{1}{L} \implies$$

$$\implies \sum p_i \log_2 \frac{\frac{1}{L}}{p_i} = 0 \implies \sum p_i \ln \frac{1}{Lp_i} = 0$$

Per tant,

$$0 = \sum p_i \ln \frac{1}{Lp_i} \leq \sum p_i \left(\frac{1}{Lp_i} - 1 \right) = 0$$

$$\Downarrow$$

$$\sum p_i \ln \frac{1}{Lp_i} = \sum p_i \left(\frac{1}{Lp_i} - 1 \right)$$

Propietats de l'entropia (3)

- Com que $\forall i\ \ln \frac{1}{Lp_i} \leq \frac{1}{Lp_i} - 1$, és

$$\forall i\ \ln \frac{1}{Lp_i} = \frac{1}{Lp_i} - 1$$
$$\Downarrow$$
$$\forall i\ \frac{1}{Lp_i} = 1$$
$$\Downarrow$$
$$\forall i\ p_i = \frac{1}{L}$$

Queda vista la tercera propietat de l'entropia

Informació mútua (1)

- Considerem el sistema de transmissió representat a continuació:

$$\text{font} > \boxed{\text{canal}} > \text{destinació}$$

- La font selecciona símbols d'un alfabet X per transmetre'ls
- Idealment el canal hauria de reproduïr a destinació els símbols emesos per la font
- Però el soroll i les interferències alteren els símbols de la font, resultant un alfabet diferent, Y, a destinació
- Volem mesurar la informació transmesa

Informació mútua (2)

- Per això ens caldrà disposar d'algunes probabilitats:

$p(x_i)$	probabilitat de que la font seleccioni el símbol x_i
$p(y_j)$	probabilitat de que el símbol y_j sigui rebut a destinació
$p(x_i, y_j)$	probabilitat de que x_i sigui transmès i y_j sigui rebut
$p(x_i \mid y_j)$	probabilitat condicionada de que x_i hagi estat trasmès sabent que s'ha rebut y_j
$p(y_j \mid x_i)$	probabilitat condicionada de que y_j hagi estat rebut sabent que x_i ha estat transmès

- Suposarem que el canal és invariant en el temps i sense memòria, és a dir que les probabilitats anteriors són independents del temps i dels símbols prèviament transmesos
- Les probabilitats condicionades $p(y_j|x_i)$ són les *probabilitats de transició* (cap endavant) del canal. Si el sistema està dissenyat per donar y_i quan el símbol transmès és x_i, llavors les probabilitats d'error venen donades per les $p(y_j|x_i)$, per $j \neq i$

Informació mútua (3)

Definition

La informació mútua $I(x_i, y_j)$ és

$$I(x_i, y_j) = \log_2 \frac{p(x_i|\, y_j)}{p(x_i)}$$

- $I(x_i, y_j)$ mesura la quantitat d'informació transmesa quan x_i és enviat i y_j rebut. Com que

$$p(x_i|\, y_j) = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p(x_i)p(y_j|\, x_i)}{p(y_j)},$$

$$I(x_i, y_j) = \log_2 \frac{\frac{p(x_i)p(y_j|\, x_i)}{p(y_j)}}{p(x_i)} = \log_2 \frac{p(y_j|\, x_i)}{p(y_j)} = I(y_j, x_i)$$

Informació mútua (4)

- Justifiquem la definició anterior, al menys en dos casos extrems:
  - Si el canal és ideal en el sentit de que cada y_j identifica un particular x_i, llavors $p(x_i|y_j) = 1$ i, per tant, $I(x_i, y_j) = \log_2 \frac{1}{p(x_i)} = -\log_2 p(x_i) = I(x_i)$. En aquest cas la informació transmesa quan s'envia x_i i es rep y_j coincideix amb la informació que porta x_i (no hi ha pèrdua d'informació)
  - Si el canal té un soroll tan intens que fa que el símbol rebut sigui totalment independent del símbol transmès, $p(x_i|y_j) = p(x_i)$ i, per tant, $I(x_i, y_j) = \log_2 \frac{p(x_i)}{p(x_i)} = 0$: no hi ha transmissió d'informació

Informació mútua (5)

- En el primer dels dos casos precedents la quantitat $I(x_i, y_j)$ ha sortit positiva, i en el segon, zero
- Però la quantitat $I(x_i, y_j)$ també pot ser negativa, n'hi ha prou que $0 < p(y_j|x_i) < p(y_j)$: seria el cas, per exemple, d'un canal on el zero sempre arribi bé i on l'1 arribi com a 0 o com a 1 amb la mateixa probabilitat. En aquest cas, i suposant equiprobables el 0 i l'1 a emissió,

$$p(\text{rebre } 0|\,\text{tx1}) = 0.5$$
$$p(\text{rebre } 0) = p(\text{tx}\,0) \cdot p(\text{rebre } 0|\,\text{tx0}) + p(\text{tx}\,1) \cdot p(\text{rebre } 0|\,\text{tx1}) = 0.75$$

- Per tant $I(\text{rebre } 0\ ,\ \text{enviar } 1) = -0.585$ bits.

Informació mútua (6)

- La major part dels canals de transmissió reals estan entre els dos extrems il.lustrats. Per analitzar el cas general es defineix la *informació mútua mitjana*, la qual és la mitjana ponderada de totes les informacions mútues possibles:

$$I(X,Y) = \sum_{i,j} p(x_i,y_j) I(x_i,y_j) = \sum_{i,j} p(x_i,y_j) \log_2 \frac{p(x_i|\,y_j)}{p(x_i)}$$

$$I(X,Y) = \sum_{i,j} p(x_i,y_j) I(y_j,x_i) = \sum_{i,j} p(x_i,y_j) \log_2 \frac{p(y_j|\,x_i)}{p(y_j)} = I(Y,X)$$

Informació mútua i entropia

- Podem interpretar $I(X, Y)$ com la quantitat mitjana d'informació de font guanyada per símbol rebut
- És intuïtivament clar que la quantitat d'informació de font guanyada per símbol rebut no pot superar l'entropia de la font:

$$I(X, Y) \leq H(X)$$

- Una demostració formal de l'afirmació precedent:

$$H(X) = \sum_i p(x_i) \log_2 \frac{1}{p(x_i)} = \sum_i \left(\sum_j p(x_i, y_j) \log_2 \frac{1}{p(x_i)} \right) =$$

$$= \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(x_i)} \geq \sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i | y_j)}{p(x_i)} = I(X, Y)$$

Entropia condicionada (1)

- La igualtat $I(X,Y) = H(X)$ serà vàlida quan no hi hagi pèrdua d'informació en el canal, és a dir, quan el canal sigui *sense soroll*. Per canals sorollosos serà $I(X,Y) < H(X)$, és a dir, hi haurà una pèrdua d'informació en el canal de $H(X) - I(X,Y)$ bits/símbol
- Hem interpretat $I(X,Y)$ des de l'extrem transmissor. Però també es pot interpretar des de l'extrem receptor:
- $I(X,Y)$ és la disminució de l'incertesa sobre el símbol transmès una vegada ha estat observat el símbol rebut
- La quantitat $H(X) - I(X,Y)$ pot pensar-se com la incertesa que es té sobre el símbol transmès una vegada ha estat observat el símbol rebut
- Es pot definir l'*entropia de X condicionada a Y*:

$$H(X|Y) = H(X) - I(X,Y)$$

Entropia condicionada (2)

- Si tenim en compte que $p(x_i) = \sum_j p(x_i, y_j)$,

$$H(X|Y) = -\sum_i p(x_i) \log_2 p(x_i) - \sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i|y_j)}{p(x_i)} =$$

$$= -\sum_i \left(\sum_{,j} p(x_i, y_j) \right) \log_2 p(x_i) - \sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i|y_j)}{p(x_i)} =$$

$$= -\sum_{i,j} p(x_i, y_j) \left(\log_2 p(x_i) + \log_2 \frac{p(x_i|y_j)}{p(x_i)} \right) = -\sum_{i,j} p(x_i, y_j) \log_2 p(x_i|y_j)$$

trobarem una definició per a l'entropia condicionada:

$$H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log_2 p(x_i|y_j)$$

Entropia condicionada i informació mútua

Teorema: $H(X|Y) \geq 0$

- És intuïtivament clar que la incertesa sobre el símbol transmès no pot aumentar pel fet d'haver observat el símbol rebut:

$$H(X) \geq H(X|Y)$$

i, com que $H(X|Y) = H(X) - I(X,Y)$, resulta que

$$I(X,Y) \geq 0$$

- O bé la informació mútua o bé l'entropia condicionada es pot agafar com a concepte bàsic i l'altra ser definit a partir d'aquest
- A la pràctica, si el càlcul directe de la informació mútua és massa complicat, es pot trobar abans l'entropia condicionada
- Com que la definició d'informació mútua és simètrica,

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y,X)$$

Codificació de font

- Sigui S una FDSM amb alfabet de L símbols, $X = \{x_1, ..., x_L\}$ que produeix un símbol cada τ_s segons. L'entropia de la font, en bits/símbol, es

$$H(S) = H(X) = -\sum_{i=1}^{L} p(x_i) \log_2 p(x_i) \leq \log_2 L$$

- La velocitat de producció d'informació de la font (en bits d'informació per segon) és $H(X)/\tau_s$
- Codificarem la sortida de la font, és a dir representarem la sortida de la font mitjançant una seqüència de dígits binaris
- Una mesura de l'eficiència de la codificació es pot obtenir comparant el nombre mig de dígits binaris per lletra (longitud mitjana del codi) amb l'entropia de la font

Codis de longitud fixa

- Considerem en primer lloc un esquema de codificació que assigni a cada símbol un bloc únic de R dígits binaris
- L possibles símbols $\Longrightarrow$ $R = \lceil \log_2 L \rceil$ dígits_binaris/símbol
- I com que $H(X) \leq \log_2 L$ és clar que $R \geq H(X)$
- L'*eficiència* de la codificació es defineix com

$$Eff = \frac{H(X)}{R} \quad \longleftrightarrow \quad \frac{\text{bits/símbol}}{\text{dígits\_binaris/símbol}} \quad \text{(tant per 1)}$$

Exemple_1

- Suposem una FDSM amb alfabet $\{a, b, c, d\}$ i on tots els símbols són igualment probables. Calculem l'eficiència de la codificació següent:

$$\begin{aligned} a &\rightarrow 00 \\ b &\rightarrow 01 \\ c &\rightarrow 10 \\ d &\rightarrow 11 \end{aligned}$$

- $$\left.\begin{array}{l} H(X) = 2 \ \text{ bits/símbol} \\ R = 2 \ \text{ dígits\_binaris/símbol} \end{array}\right\} \Longrightarrow \ Eff = 100\%$$

(En el cas de símbols no equiprobables l'eficiència de la codificació anterior seria inferior al 100%)

Exemple_2

- Sigui una FDSM amb alfabet $\{a, b, c\}$ i on tots els símbols són igualment probables. Calculem l'eficiència de la codificació següent:

$$\begin{aligned} a &\rightarrow 00 \\ b &\rightarrow 01 \\ c &\rightarrow 10 \end{aligned}$$

- $$\left.\begin{array}{l} H(X) = \log_2 3 = 1.58 \ \text{ bits/símbol} \\ R = 2 \ \text{ dígits\_binaris/símbol} \end{array}\right\} \Longrightarrow \ Eff = 79\%$$

(En el cas de símbols no equiprobables l'eficiència de la codificació anterior seria encara menor)

Exemple_3 (1)

- Amb últim exemple veiem que, fins i tot en el cas equiprobable, si L no és una potència de 2 llavors l'eficiència no és del 100%
- Una manera d'incrementar l'eficiència consisteix en codificar blocs de J símbols cada vegada
- Sigui la FDSM de l'exemple anterior i codifiquem els símbols de tres en tres ($J = 3$). Una possible codificació seria:

aaa	→	00000	*baa*	→	01001	*caa*	→	10010
aab	→	00001	*bab*	→	01010	*cab*	→	10011
aac	→	00010	*bac*	→	01011	*cac*	→	10100
aba	→	00011	*bba*	→	01100	*cba*	→	10101
abb	→	00100	*bbb*	→	01101	*cbb*	→	10110
abc	→	00101	*bbc*	→	01110	*cbc*	→	10111
aca	→	00110	*bca*	→	01111	*cca*	→	11000
acb	→	00111	*bcb*	→	10000	*ccb*	→	11001
acc	→	01000	*bcc*	→	10001	*ccc*	→	11010

Exemple_3 (2)

- En aquest cas, la quantitat d'informació és:

$$3 \times \log_2 3 = 4.75 \quad \text{bits/bloc\_de\_3\_símbols}$$

  i la longitud del codi és de 5 dígits_binaris/bloc. Eficiència:

$$Eff = \frac{4.75}{5} = 95\%$$

- Hem passat del 79% al 95%. Seguint aquest procediment es pot aconseguir una eficiència tan próxima al 100% com es vulgui (dispositiva següent)

Exemple_3 (3)

- Codificant en blocs de J símbols (cas equiprobable, L símbols en total), la informació és de $J \times \log_2 L$ bits/bloc, mentre que la longitud del codi (codi de longitud fixa) és $\lceil \log_2 L^J \rceil = \lceil J \times \log_2 L \rceil$ dígits_binaris/bloc. L'eficiència

$$Eff = \frac{J \times \log_2 L}{\lceil J \times \log_2 L \rceil}$$

compleix

$$\frac{J \times \log_2 L}{J \times \log_2 L + 1} \leq Eff \leq \frac{J \times \log_2 L}{J \times \log_2 L} = 1$$

Al créixer J cap a infinit, l'eficiència (cas equiprobable) tendeix cap a 1 (100%)

Exemple_4 (1)

- Sigui la FDSM de l'exemple 2, amb les probabilitats de símbol llistades a continuació:

Lletra	probabilitat
a	$1/3$
b	$1/3$
c	$1/3$

- Volem fer una codificació per blocs utilitzant un codi de longitud fixa. Si volem una eficiència superior al 99%, quina ha de ser la longitud del bloc?

Exemple_4 (2)

- Com que $Eff \geq \dfrac{J \times \log_2 3}{J \times \log_2 3 + 1}$, per garantir $Eff \geq 0.99$ és suficient que

$$\frac{J \times \log_2 3}{J \times \log_2 3 + 1} \geq 0.99$$

és a dir

$$J \times \log_2 3 \geq 0.99(J \times \log_2 3 + 1) = 0.99J \times \log_2 3 + 0.99$$
$$(1 - 0.99)J \times \log_2 3 \geq 0.99$$
$$J \geq \frac{0.99}{(1 - 0.99) \times \log_2 3} = 62.46$$

Cal agafar, per tant, una longitud de bloc de 63. Com es veu, encara que teòricament es pot assolir una eficiència del 99%, això pot no ser gaire pràctic!

Exemple_4 (3)

- La font que "emet", cada vegada, un bloc de J símbols de la FDSM S s'anomena *extensió J-èssima* de la font S i es representa per S^J
- La informació generada per S^J és $H(S^J) = J \times H(S)$ bits/bloc

Codis de longitud variable (1)

- Quan els símbols de font no són igualment probables, un mètode més eficient de codificar consisteix en utilitzar paraules-codi de longitud variable. Un exemple d'això és el codi de Morse, que data del segle XIX. En el codi de Morse, les lletres que ocurreixen més freqüentment tenen assignades paraules-codi curtes i les que ocurreixen més infreqüentment, paraules-codi llargues. Amb aquesta filosofia, podem usar les probabilitats d'ocurrència de les diferents lletres de font en l'elecció de les parauls-codi. El problema és la manera de fer-ho

Codis de longitud variable (2)

Example

Sigui una FDSM amb alfabet $\{a, b, c, d\}$ i probabilitats $p(a) = \frac{1}{2},\ p(b) = \frac{1}{4},\ p(c) = p(d) = \frac{1}{8}$. Estudiem diferents maneres de codificar:

Lletra	probabilitat	codi_1	codi_2	codi_3
a	0.5	1	0	0
b	0.25	00	10	01
c	0.125	01	110	011
d	0.125	10	111	111

Codis de longitud variable (3)

- El primer codi presenta un problema. Suposem la seqüència codificada 001001... El primer símbol de font, corresponent a 00, és '*b*'. Ara bé, els següents 4 dígits són ambigus (no unívocament descodificables): 1001 pot correspondre a '*aba*' o també a '*dc*'. És possible que l'ambigüitat pugui ser resolta esperant bits addicionals però aquest retard en la descodificació no és desitjable. Ens interessarem només per codis que siguin descodificables *instantàniament*, és a dir, sense retard en el descodificador. El codi_2 és unívocament i instantània descodificable: cap paraula-codi és prefix d'una altra. El tercer codi és un exemple de codi únicament descodificable però no instantàniament. Obviament, aquest codi no satisfà la condició del prefix.

Desigualtat de Kraft

- Construirem codis de longitud variable únivocament descodificables que siguin eficients en el sentit de que el nombre mig de dígits binaris per lletra de font sigui mínim
- Abans veiem sotes quines condicions existeix un codi prefix amb paraules de longitud donada

Theorem

(desigualtat de Kraft). Una condició necessària i suficient per a l'existència d'un codi binari amb paraules-codi de longituds $n_1 \leq n_2 \leq ... \leq n_L$ *(alfabet de L símbols) que satisfaci la condició del prefix és:*

$$2^{-n_1} + 2^{-n_2} + ... + 2^{-n_L} \leq 1$$

Demostració de la implicació directa

- Si col.loquem les paraules codi en un arbre binari complet d'alçada n_L, suprimim tots els descendents de totes les paraules-codi i compten quants vèrtexs terminals (fulles) hem eliminat d'aquesta manera, com que paraules codi diferents no tenen descendents en comú, tindrem:

$$\#\text{ vèrtexs eliminats: } \sum_{i=1}^{L} 2^{n_L - n_i} \leq 2^{n_L}$$

i, per tant,

$$\sum 2^{-n_i} \leq 1$$

Demostració de la implicació recíproca (1)

- Si $\sum 2^{-n_i} \leq 1$ construïrem un codi binari prefix amb paraules-codi de longituds $n_1, n_2, ..., n_L$. Per això pensarem cada paraula codi com un vèrtex de l'arbre binari complet d'alçada n_L (el màxim de les longituds de les paraules codi)
- Una paraula codi w_1 és prefix de w_2 si i solament si w_1 és un avantpassat de w_2 (w_2 és un descendent de w_1)
- Construirem el codi prefix escollint, per a cada n_i, un vèrtex de l'arbre de nivell n_i
  - Comencem amb n_1: elegim un vèrtex v_1 de nivell n_1 (una paraula codi) i suprimim tots els descendents de v_1
  - Dels nodes d'ordre n_2 que quedin, triem-ne un, v_2, associat a n_2. I així successivament

Demostració de la implicació recíproca (2)

- Si justifiquem que sempre ens queden vèrtexs per elegir, haurem construit un codi prefix amb paraules de longituds $n_1, n_2, ..., n_L$
- Una manera de comptar els vèrtexs que ens queden és simplement comptar els vèrtex terminals (fulles) que ens van quedant a cada pas. Després de suprimir tots els descendents de v_1 haurem eliminat $2^{n_L - n_1}$ vèrtexs terminals: encara en queden. Després d'haver eliminat els descendents de v_2 haurem eliminat $2^{n_L - n_2}$ fulles. Ens queden, a cada pas,

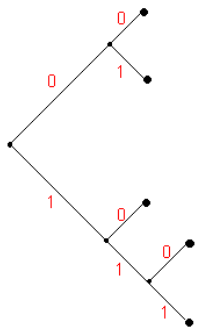
$$2^{n_L} - (2^{n_L - n_1} + 2^{n_L - n_2} + ... + 2^{n_L - n_j}) = 2^{n_L} - 2^{n_L}(2^{-n_1} + ... + 2^{-n_j}) =$$
$$= 2^{n_L}\left[1 - (2^{-n_1} + ... + 2^{-n_j})\right] > 0, \quad \forall j < L$$

Exemple (1)

- Construim, si és possible, un codi prefix per a les lletres $\{a, b, c, d, e\}$ on les longituds siguin: $2, 2, 2, 3, 3$.
- Comprovem que es satisfà la desigualtat de Kraft:

$$2^{-2} + 2^{-2} + 2^{-2} + 2^{-3} + 2^{-3} = 1 \leq 1$$

- Sobre l'arbre binari complet de profunditat 3, és clar que podem elegir tres vèrtex de nivell 2 i dos de nivell 3 que no siguin descendents dels anteriors. Una possible elecció és:



la qual dóna les paraules codi: $\{00, 01, 10, 110, 111\}$

Exemple (2)

- El codi anterior... quina longitud té?
  a) si tots els símbols són equiprobables
  b) si les probabilitats dels símbols de font són:

$$p(a) = p(b) = p(c) = 0.3, \quad p(d) = p(e) = 0.05$$

- en el cas a) $\bar{l} = \frac{1}{5}(2+2+2+3+3) = \frac{12}{5} = 2.4$ dígits_binaris/símbol
- en el cas b) i codificant

$$\begin{aligned} a &\rightarrow 00 \\ b &\rightarrow 01 \\ c &\rightarrow 10 \\ d &\rightarrow 110 \\ e &\rightarrow 111 \end{aligned}$$

$$\bar{l} = 0.3 \times (2+2+2) + 0.05 \times (3+3) = 2.1 \text{ dígits\_binaris/símbol}$$

- Comparem-ho, en aquest cas, amb l'entropia de la font:

$$H = -3 \times 0.3 \times \log_2 0.3 - 2 \times 0.05 \times \log_2 0.05 = 1.995 \text{ bits/símbol}$$

Relació entre l'entropia i la longitud mitjana d'un codi (1)

- Sigui S una FDSM amb alfabet $X = \{x_1, ..., x_L\}$ i codifiquem, en binari, les lletres del seu alfabet:

lletra	probabilitat	paraula-codi	longitud de la paraula-codi
x_1	p_1	w_1	n_1
x_2	p_2	w_2	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	p_L	w_L	n_L

(les paraules-codi són tires de 0 i 1). La longitud mitjana del codi, $\bar{l} = \sum_{i=1}^{L} p_i n_i$ satisfà

$$\bar{l} \geq H$$

Relació entre l'entropia i la longitud mitjana d'un codi (2)

- Si definim

$$q_i := \frac{2^{-n_i}}{2^{-n_1} + 2^{-n_2} + \ldots + 2^{-n_L}}$$

és clar que $q_1 + q_2 + \ldots + q_n = 1$ i, pel lema de Gibbs,

$$-\sum_{i=1}^{L} p_i \cdot \log_2 p_i \leq -\sum_{i=1}^{L} p_i \cdot \log_2 q_i = -\sum_{i=1}^{L} p_i \left(-n_i - \log_2 \left(2^{-n_1} + \ldots + 2^{-n_L}\right)\right)$$

és a dir,

$$H \leq \sum_{i=1}^{L} p_i \left(n_i + \log_2 \left(2^{-n_1} + 2^{-n_2} + \ldots + 2^{-n_L}\right)\right)$$

- Com que per Kraft és $2^{-n_1} + 2^{-n_2} + \ldots + 2^{-n_L} \leq 1$, llavors

$$H \leq \sum_{i=1}^{L} p_i n_i = \bar{\ell}$$

Relació entre l'entropia i la longitud mitjana d'un codi (3)

- Veiem ara que sempre existeix una codificació (per a una FDSM) la longitud mitjana de la qual no supera $H + 1$:
- Com que

$$H + 1 = -\sum_{i=1}^{L} p_i \log_2 p_i + \sum_{i=1}^{L} p_i = \sum_{i=1}^{L} p_i (1 - \log_2 p_i) > \sum_{i=1}^{L} p_i \lceil -\log_2 p_i \rceil$$

és suficient veure que existeix un codi

lletra	probabilitat	paraula-codi	longitud de la paraula-codi
x_1	p_1	w_1	n_1
x_2	p_2	w_2	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	p_L	w_L	n_L

on les longituds de les paraules-codi són $n_i = \lceil -\log_2 p_i \rceil$

Relació entre l'entropia i la longitud mitjana d'un codi (4)

- Però el que existeixi un codi amb longituds de les paraules-codi $n_i = \lceil -\log_2 p_i \rceil$ equival a veure que

$$\sum_{i=1}^{L} 2^{-n_i} \leq 1,$$

cosa que és òbvia ja que

$$\sum_{i=1}^{L} 2^{-n_i} = \sum_{i=1}^{L} 2^{-\lceil -\log_2 p_i \rceil} \leq \sum_{i=1}^{L} 2^{-(-\log_2 p_i)} = \sum_{i=1}^{L} 2^{\log_2 p_i} = \sum_{i=1}^{L} p_i = 1$$

Codificació de Huffman

- Acabem de provar que, donada una FDSM S, sempre existeix una codificació per a la qual la longitud mitjana del codi $\bar{l}$ (en dígits_binaris/símbol_de_font) satisfà:

$$H(S) \leq \bar{l} < H(S) + 1$$

- Veiem, a continuació i amb l'ajut d'un parell d'exemples, l'algorisme de codificació de Huffman, amb el qual algorisme s'aconsegueixen codis òptims (de longitud mínima) que satisfan, per tant, la fita anterior ($H \leq \bar{l} < H + 1$)

Exemple (1)

- Considerem una FDSM amb set símbols possibles $x_1, x_2, ..., x_7$ amb les probabilitats d'ocurrència ilustrades a la taula que segueix

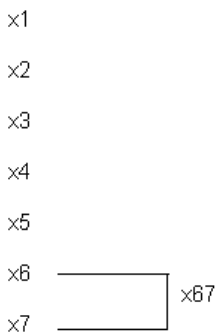
lletra	probabilitat
x_1	0.35
x_2	0.30
x_3	0.20
x_4	0.10
x_5	0.04
x_6	0.005
x_7	0.005

- Hem ordenat els símbols de font en ordre decreixent de probabilitats:

$$p(x_1) \geq p(x_2) \geq ... \geq p(x_7)$$

Exemple (2)

- Les paraules d'un codi prefix es poden situar sobre un arbre binari. Ara, però, cal construir l'arbre
- Procedirem des dels nodes terminals, els quals representaran les paraules codi i estaran, per tant, associats a les x_i (o a les p_i)
- Començarem amb els dos símbols menys probables x_6, x_7 i els ajuntarem en un node (construim un ascendent comú) de la manera indicada a la figura que segueix. Les probabilitats de les dues branques corresponents a x_6 i x_7 es sumen i s'assignen al node resultant $x_{6,7}$:

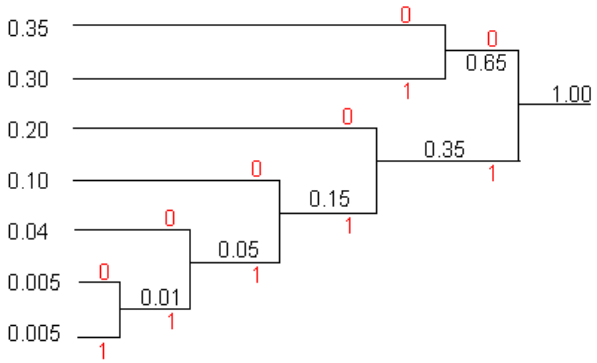


Exemple (3)

- Tenim ara els símbols de font $x_1, ..., x_5$ més el nou símbol $x_{6,7}$, obtingut per combinació de x_6 i x_7. El següent pas consisteix en tornar a fer el mateix però amb els símbols $\{x_1, x_2, x_3, x_4, x_5, x_{6,7}\}$
- Comencem amb els dos menys probables: x_5 i $x_{6,7}$, els quals es poden "combinar" en un ascendent comú, amb probabilitat associada igual a 0.05. I així, fins a arribar a l'arrel, és a dir, fins a arribar a un node amb probabilitat combinada igual a 1
- Per "llegir" el codi assignat a cada lletra procedirem de dreta a esquerra, des de l'arrel a cada una de les fulles (paraules codi). Cada vegada que elegim la branca superior posarem un 0 i cada vegada que elegim la branca inferior, un 1

Exemple (4)

- Arbre resultant i codificació corresponent:



Lletra	Probabilitat	Codi
x_1	0.35	00
x_2	0.30	01
x_3	0.20	10
x_4	0.10	110
x_5	0.04	1110
x_6	0.005	11110
x_7	0.005	11111

Exemple (5)

- La longitd mitjana de l'anterior codi és

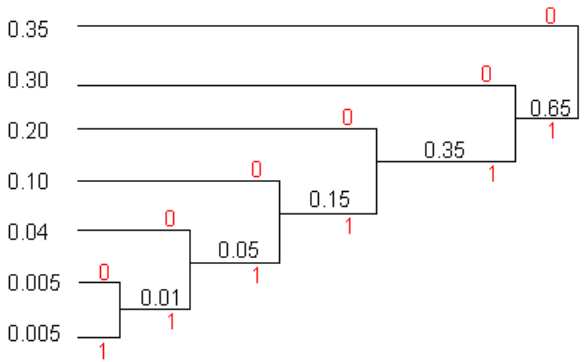
$$\bar{l} = 2\times(0.35+0.30+0.20)+3\times 0.10+4\times 0.04+5\times(0.005+0.005) = 2.21 \text{ dígits/símbol}$$

- L'entropia de la font és

$$H = -0.35 \times \log_2 0.35 - 0.30 \times \log_2 0.30 - 0.20 \times \log_2 0.20 - 0.10 \times \log_2 0.10$$
$$-0.04 \times \log_2 0.04 - 2 \times 0.005 \times \log_2 0.005 = 2.11 \quad \text{bits/símbol}$$

Exemple (6)

- El codi no és necessàriament únic: en el penúltim pas teníem dos candidats amb probabilitat 0.35 per agrupar amb el símbol de probabilitat 0.30. Si haguéssim fet l'altra agrupació:



haguéssim obtingut el codi següent:

Lletra	Probabilitat	Codi
x_1	0.35	0
x_2	0.30	10
x_3	0.20	110
x_4	0.10	1110
x_5	0.04	11110
x_6	0.005	111110
x_7	0.005	111111

Exemple (7)

- La longitud mitjana del codi és també 2.21 dígits/símbol i per, tant aquest codi és tant eficient com l'anterior
- I canviant els zeros per els uns obtindríem una altra codificació
- Es compleixen les fites:

$$H \leq \bar{l} < H + 1 \qquad \longrightarrow \qquad 2.11 \leq 2.21 < 2.11 + 1$$

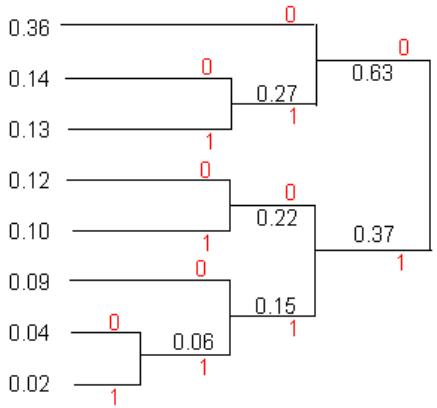
Un altre Exemple (1)

- Considerem una FDSM amb vuit símbols possibles $x_1, x_2, ..., x_8$ amb les probabilitats d'ocurrència ilustrades a la taula que segueix

lletra	probabilitat
x_1	0.36
x_2	0.14
x_3	0.13
x_4	0.12
x_5	0.10
x_6	0.09
x_7	0.04
x_8	0.02

Un altre Exemple (2)

- L'arbre amb les agrupacions corresponents és:



i el codi resultant:

Lletra	Codi
x_1	00
x_2	010
x_3	011
x_4	100
x_5	101
x_6	110
x_7	1110
x_8	1111

Un altre Exemple (3)

- L'entropia de la font és $H = 2.63$ bits/símbol i la longitud mitjana del codi de Huffman $\bar{l} = 2.70$ dígits_binaris/símbol
- Eficiència de la codificació: 97%
- L'algorisme de Huffman genera un codi prefix, la longitud mitjana del qual satisfà

$$H \leq \bar{l} < H + 1$$

- En lloc de codificar símbol a símbol es poden codificar blocs de J símbols. En aquest cas la fita anterior queda:

$$J \times H \leq \bar{l}_J < J \times H + 1,$$

on $\bar{l}_J$ representa el nombre mig de dígits binaris per bloc de J símbols i $J \times H$ és l'entropia d'un bloc de J símbols

Un altre Exemple (4)

- Dividint les anterior desigualtats entre J:

$$H \leq \frac{\bar{l}_J}{J} < H + \frac{1}{J}$$

  on $\frac{\bar{l}_J}{J}$ és el nombre mig de dígits binaris per símbol de font
- D'aquesta manera, la longitd mitjana del codi es pot fer tan propera a l'entropia com es vulgui (seleccionant J suficientment gran). Aquest és precisament el primer teorema de Shannon

Primer teorema de Shanon (teorema de Shannon de codificació de font)

Theorem

La sortida d'una FDSM es pot codificar amb una quantitat mitjana de dígits binaris per símbol de font tan propera a l'entropia de la font com es vulgui. Es pot aconseguir una longitud mitjana igual a l'entropia si i només si totes les probabilitats de símbol són potències (negatives) de 2

Exemple (1)

- Sigui una FDSM amb tres símbols possibles x_1, x_2, x_3 i probabilitats d'ocurrència donades a la taula següent

lletra	probabilitat
x_1	0.45
x_2	0.35
x_3	0.20

- L'entropia de la font és $H = 1.518$ bits/símbol. El codi de Huffman per a aquesta font és

Lletra	Codi
x_1	1
x_2	00
x_3	01

i la seva longitud mitjana és $\bar{l} = 1.55$ dígits_binaris/símbol. L'eficiència és, per tant, del 97.9%

Exemple (2)

- Si codifiquem els símbols a parells (en blocs de longitud 2, mitjançant l'algorisme de Huffman, el codi resultant serà:

Parell de lletres	Codi
x_1x_1	10
x_1x_2	001
x_1x_3	010
x_2x_1	011
x_2x_2	111
x_2x_3	0000
x_3x_1	0001
x_3x_2	1100
x_3x_3	1101

- En aquest cas el codi de Huffman té una longitud mitjana de 3.0675 dígits_binaris per parell de símbols, mentre que l'entropia és de 3.036 bits per parell de símbols. L'eficiència ha passat, doncs, al 99.0%.

L'algorisme de Lempel-Ziv (1)

- Per aplicar un Huffman cal un coneixement a priori de les estadístiques de la font, cosa que, a la pràctica, no es sol tenir
- També podem considerar algorismes que funcionin amb independència de les estadístiques de la font
- Presentem un d'aquests algorismes: l'algorisme de Lempel-Ziv
- L'algorisme LZ és un algorisme de longitud variable a fixa
- La sortida de la font es descompon en blocs de longitud variable anomenats *frases,* frases que són el més curt possible, essent cada frase diferent de les anteriors
- Les frases es van guardant en un diccionari, en el qual s'hi emmagatzema la localització de les frases existents
- Per exemple, les frases corresponents a

$$101011010010011101010000110011101011000011011$$

són: $1, 0, 10, 11, 01, 00, 100, 111, 010, 1000, 011, 001, 110, 101, 10001, 1011$

L'algorisme de Lempel-Ziv (2)

- Cada frase és la concatenació d'una frase prèvia amb una lletra
- Per codificar les frases es construeix un diccionari:

	posicions del diccionari	contingut del diccionari	paraula-codi
0	0000		
1	0001	1	00001
2	0010	0	00000
3	0011	10	00010
4	0100	11	00011
5	0101	01	00101
6	0110	00	00100
7	0111	100	00110
8	1000	111	01001
9	1001	010	01010
10	1010	1000	01110
11	1011	011	01011
12	1100	001	01101
13	1101	110	01000
14	1110	101	00111
15	1111	10001	10101

L'algorisme de Lempel-Ziv (3)

- Les posicions del diccionari es numeren de manera consecutiva
- La codificació consisteix en substituir les frases per punters al diccionari: com que cada frase és la concatenació d'una frase anterior amb un zero o un u, per codificar una frase en concret s'agafa la posició a la qual està la frase anterior i i se li afegeix un 0 o un 1 segons la frase actual acabi en 0 o en 1
- Inicialment s'usa la posició 0000 per codificar la frase buida
- El descodificador ha de construir una taula idèntica a l'extrem receptor i descodificar convenientment la seqüència rebuda

L'algorisme de Lempel-Ziv (4)

- Estem codificant 44 dígits binaris en 16 paraules-codi de 5 dígits binaris cada una, això és, en 80 dígits binaris
- L'algorisme, per tant, no proporciona compressió, cosa que és deguda a que la seqüència considerada és molt curta
- A mesura que la seqüència a codificar va creixent en longitud, el procés de codificació proporciona compressió
- Com elegir la longitud de la taula? Agafem la que agafem, acabarem per desbordar-la
- El codificador i el descodificador s'han de posar d'acord en com eliminar frases no-útils dels diccionaris respectius i substiruir-les per noves
- L'algorisme de Lempel-Ziv s'usa ampliament en la compressió de fitxers d'ordinador; en les utilitat "compress" i "uncompress" de Unix, per exemple

Capacitat d'un canal (1)

- Quina és la màxima quantitat d'informació que es pot transmetre per un canal?
- Com es defineixen matemàticament els canals?
- Per definir un canal (discret sense memòria) cal donar els alfabets d'entrada i sortida

$$X = \{x_1, x_2, ..., x_n\} \qquad Y = \{y_1, y_2, ..., y_m\}$$

i les probabilitats de transició

$$p_{ij} = p(y_j | x_i)$$

- Suposarem que el canal està pensat per donar una sortida y_i quan l'entrada és x_i
- La informació transmesa per un canal és $I(X, Y)$ (o $I(Y, X)$):

$$\sum_{i,j} p(x_i, y_j) \log_2 \frac{p(y_j | x_i)}{p(y_j)} = \sum_{i,j} p(x_i) p(y_j | x_i) \log_2 \frac{p(y_j | x_i)}{\sum_k p(x_k) p(y_j | x_k)}$$

Capacitat d'un canal (2)

- En un canal concret les probabilitats $p(y_j|\, x_i)$ són fixes
- La màxima transferència d'informació (el màxim de $I(X,Y)$) s'aconseguirà amb unes estadístiques de font $p(x_i)$ concretes. Per a unes tals estadístiques el valor màxim resultant és el que s'anomena *capacitat del canal*:

$$C_s = \max_{p(x_i)} I(X,Y) \quad \text{bits/símbol}$$

- C_s representa la quantitat màxima d'informació que es pot transmetre, en bits per símbol. Si el nostre canal és capaç d'una velocitat de s símbols/segon llavors la capacitat del canal, mesurada en bits(d'informació)/segon és:

$$C = sC_s \quad \text{bits/seg}$$

Teorema de Shanon de codificació de canal

Theorem

Si la font genera informació a una velocitat, en bits/segon, menor que la capacitat del canal llavors és possible aconseguir transmissió fiable (amb probabilitat d'error arbitràriament baixa) a través del canal amb una codificació apropiada. Per altra part, si velocitat a la que genera informació la font és superior a la capacitat del canal, llavors no és possible la transmissió fiable amb independència de la quantitat de processat realitzada al transmissor i al receptor

El canal binari simètric

- L'exemple més famós de canal és el del *canal binari simètric*:

$$\begin{aligned} X &= \{x_1, x_2\} & Y &= \{y_1, y_2\} \\ p(y_1|\,x_1) &= 1-\alpha & p(y_2|\,x_1) &= \alpha \\ p(y_1|\,x_2) &= \alpha & p(y_2|\,x_2) &= 1-\alpha \end{aligned}$$

- En paraules, el canal binari simètric (CBS) té alfabets d'entrada i sortida binaris i la probabilitat de que la transmissió d'un símbol sigui errònia és

$$\begin{aligned} p_e &= p(x_1)p(e|\,x_1) + p(x_2)p(e|\,x_2) = \\ &= p(x_1)p(y_2|\,x_1) + p(x_2)p(y_1|\,x_2) = \\ &= p(x_1)\alpha + (1-p(x_1))\alpha = \alpha \end{aligned}$$

α representa la probabilitat d'error. Suposarem $\alpha < 0.5$

Teorema de Shanon amb CBS (1)

- Aclarim el significat del teorema de Shannon amb el CBS
- Capacitat:

$$C_s = \max_{p(x_i)} I(X,Y) = \max_{p(x_i)} \left(H(Y) - H(Y|X)\right)$$

- Càlcul de $H(Y|X)$:

$$H(Y|X) = -\sum_{i,j} p(x_i,y_j)\log_2 p(y_j|x_i) = -\sum_{i,j} p(x_i)p(y_j|x_i)\log_2 p(y_j|x_i) =$$

$$= \sum_i p(x_i)\left(-\sum_j p(y_j|x_i)\log_2 p(y_j|x_i)\right) = \sum_i p(x_i)H_0 = H_0$$

  Per justificar el penúltim pas observem que tan per $i=1$ com per $i=2$ els sumands de la suma interior són els mateixos
- Per tant,

$$I(X,Y) = H(Y) - H_0$$

Teorema de Shanon amb CBS (2)

- Els valors que maximitzen $I(X,Y)$ son els que maximitzen $H(Y)$ (Y uniforme)
- La distribució de X que fa que Y sigui uniforme és la uniforme: si per comoditat notem $p(x_1) = p, p(x_2) = 1-p$ i si ha de ser $p(y_1) = \frac{1}{2}$ llavors

$$p(y_1) = p(x_1)p(y_1|\,x_1) + p(x_2)p(y_1|\,x_2)$$
$$\tfrac{1}{2} = p(1-\alpha) + (1-p)\alpha = (1-2\alpha)p + \alpha$$
$$p = \frac{\frac{1}{2}-\alpha}{1-2\alpha} = \frac{1}{2}$$

- La capacitat del CBS s'assoleix quan els símbols d'entrada són equiprobables: $p(x_1) = p(x_2) = 0.5$ i val

$$C_s = \max_{p(x_i)} I(X,Y) = H(Y)_{|p(y_1)=p(y_2)=0.5} - H_0 = \log_2 2 - H_0$$

on H_0 és

$$H_0 = -\sum_j p(y_j|\,x_i)\log_2 p(y_j|\,x_i) =$$

Teorema de Shanon amb CBS (3)

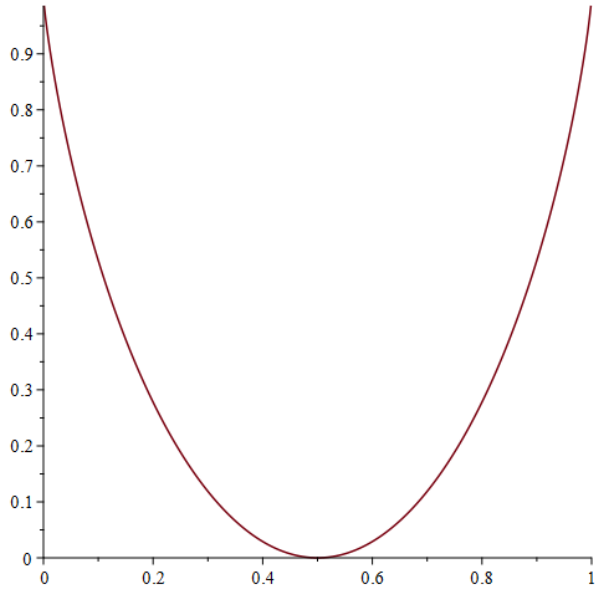
- Si, per exemple, $x_i = x_1$ (surt el mateix amb $x_i = x_2$),

  $H_0 = -p(y_1|x_1)\log_2 p(y_1|x_1) - p(y_2|x_1)\log_2 p(y_2|x_1) = -(1-\alpha)\log_2(1-\alpha) - \alpha\log_2\alpha$

  i, per tant,

  $$C_s = 1 + (1-\alpha)\log_2(1-\alpha) + \alpha\log_2\alpha \quad \text{bits/símbol}$$

- Podem fer una gràfica de la capacitat en funció de la probabilitat d'error de símbol α:



Exemple (1)

- CBS amb alfabets d'entrada i sortida

$$X = \{x_1, x_2\} \qquad Y = \{y_1, y_2\}$$

i probabilitat d'error $\alpha = 0.15$. Capacitat d'aquest canal:

$$C_s = 1+(1-0.15)\log_2(1-0.15)+0.15\log_2 0.15 = 0.39 \quad \text{bits/símbol}$$

- Si el nostre canal pot transmetre a raó de 2 símbols per segon, la seva capacitat, en bits/segon, és

$$C = 0.39 \times 2 = 0.78 \ \text{ bits/s}$$

- Suposem una font binària S amb alfabet $\{0, 1\}$, que generi un símbol cada segon. Si els símbols són independents i equiprobables llavors l'entropia de la font és

$$H(S) = \log_2 2 = 1 \ \text{ bit/símbol}$$

- La font està generant informació a una velocitat de

$$1 \text{ bit/símbol} \times 1 \text{ símbol/segon} = 1 \text{ bit/segon}$$

Exemple (2)

- Com que la capacitat del canal és de 0.78 bits/s, no és possible la transmissió fiable
- Si la probabilitat dels símbols de font és

$$p(0) = 0.93 \qquad p(1) = 0.07$$

llavors l'entropia és

$$H(S) = -0.93 \times \log_2 0.93 - 0.07 \times \log_2 0.07 = 0.37 \text{ bits/símbol}$$

que equival a una producció d'informació de

$$0.37 \text{ bits/símbol} \times 1 \text{ símbol/segon} = 0.37 \text{ bits/segon}$$

- Com que $0.37 < 0.78$, el teorema de Shannon de codificació de canal ens diu que sí que és possible la transimissió fiable
- Fiable vol dir amb una probabilitat d'error arbitràriament baixa

Sobre el teorema de Shannon de codificació de canal

- La demostració del teorema de Shannon, però, no dóna cap mètode per a aconseguir una codificació que permeti la transmissió fiable
- Caldrà considerar-lo des d'un punt de vista teòric
- A la pràctica:
  - primer es treu la «redundància descontrolada» que pugui tenir el missatge (per acostar-nos a l'entropia de la font)
  - després es divideix el missatge resultant a transmetre en blocs
  - i finalment s'afegeix a cada bloc «redundància controlada» la qual ens permetrà detectar/corregir errors (segona part del curs)
- Caldrà estudiar els codis sota la perspectiva de la seva capacitat detectora i/o correctora d'errors

Continuació de l'exemple (3)

- Continuem amb l'exemple (cas $p(0) := 0.93$, $p(1) = 0.07$, 1 símbol/segon) per veure com es pot millorar la fiabilitat de la transmissió afegint «redundància controlada»
- Pel primer teorema de Shannon podem aconseguir una codificació de longitud mitjana tan propera a l'entropia com volguem (l'entropia és 0.37)
- Suposem, per exemple, que hem aconseguit una longitud de 0.394 dígits_binaris/símbol, és a dir, 0.394 dígits_binaris/segon
- Com que el canal pot transmetre, de manera fiable, 0.78 dígits/s, tenim marge de maniobra

Continuació de l'exemple (4)

- A cada bloc de 4 dígits binaris li poden afegir una redundància de 3 dígits binaris. D'aquesta manera millorarem la probabilitat de transmetre dades errònies (CBS, prob d'error $\alpha = 0.15$):
  - Suposant independència d'errors, la probabilitat d'error en un bloc de 4 dígits és:

$$p_e = p(\text{1 o més errors}) = 1 - p(\text{0 errors}) = 1 - 0.85^4 = 0.478$$

  - Si apliquem, per exemple, una codificció de Hamming –que veurem a la segona part del curs– i a cada bloc de 4 bits li afegim 3 bits de paritat (aquests codis tenen la capacitat de corregir 1 error), llavors la probabilitat d'error en un bloc de 7 dígits serà

$$q_e = p(\text{2 o més errors}) = 1 - p(\text{0 errors}) - p(\text{1error}) =$$
$$= 1 - 0.85^7 - 7 \cdot 0.15 \cdot 0.85^6 = 0.283$$

- Amb una codificació adequada la probabilitat d'error disminueix