Εισαγωγή στη Θεωρία Πληροφορίας

- Ένα σύστημα επικοινωνιών έχει ως σκοπό τη μεταφορά μιας ακολουθίας μηνυμάτων μεταξύ πομπού και δέκτη.
- Στη πράξη, τα μηνύματα που μεταδίδει ο πομπός και αναγνωρίζει ο δέκτης **δεν είναι αυθαίρετα**.
  - Αντίθετα, δουλειά του δέκτη είναι να αναγνωρίσει σε κάθε λήψη **ποιο μήνυμα έλαβε από ένα προκαθορισμένο σύνολο** δυνατών μηνυμάτων που μπορεί να μεταδώσει ο πομπός.
    - Συνεπώς ο δέκτης ουσιαστικά δεν αναγνωρίζει **τι ήταν** ένα μήνυμα, αλλά **ποιο ήταν**.
- Η απόφαση αυτή λαμβάνεται με επιλογή από το δέκτη του μηνύματος που εμφανίζει **τη μεγαλύτερη συσχέτιση** με το μήνυμα που έλαβε,
  - δηλαδή εκείνου που **μοιάζει περισσότερο** με το ληφθέν μήνυμα.
  - Π.χ: τα δυνατά μηνύματα είναι τα 000, 111 και ελήφθη το 100.
    - *Ποιό μήνυμα εστάλη;*

Εισαγωγή στη Θεωρία Πληροφορίας

- Κάθε ένα από τα μηνύματα που μεταδίδονται μεταφέρει κάποιο **ποσό πληροφορίας**.
  - Κάθε μήνυμα δε μεταφέρει πληροφορία ίδιας σημαντικότητας ή αλλιώς την ίδια **ποσότητα πληροφορίας**.
- Το παραπάνω γίνεται εύκολα κατανοητό θεωρώντας το παράδειγμα της λήψης μετεωρολογικού δελτίου σε κάποια περιοχή στην έρημο όπου έχει να βρέξει αρκετά χρόνια.
  - Διαισθητικά καταλαβαίνουμε ότι σε περίπτωση που το δελτίο αναφέρει «βροχή-καταιγίδα» η ποσότητα πληροφορίας του μηνύματος αυτού είναι πολύ μεγαλύτερη από την αναφορά «ηλιοφάνειας»
    - καθώς η πρώτη αναφορά είναι πολύ πιο απίθανη για τη περιοχή.

Εισαγωγή στη Θεωρία Πληροφορίας

Με τι ασχολείται λοιπόν η ΘΠ;

1. Τι εννοούμε με τον όρο «πληροφορία» μηνύματος και πως μπορούμε να την μετρήσουμε.

2. Πως μπορούμε να συμπιέσουμε τα μηνύματα που παράγει μια πηγή πληροφορίας
   - Δηλ. να μεταφέρονται με τα ελάχιστα σύμβολα;
   - και ποια είναι η μέγιστη δυνατή συμπίεση;

Ποσότητα πληροφορίας-Εντροπία

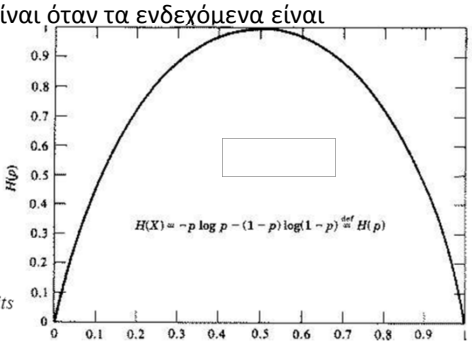
- Αν X είναι μία διακριτή τυχαία μεταβλητή με πιθανά ενδεχόμενα $\{x_1, x_2, \ldots, x_n\}$ και συνάρτηση πυκνότητας πιθανότητας $p(x_i)$,τότε η **μέση ποσότητα πληροφορίας** της X, $H(X)$, δίνεται από τη σχέση

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

- **Η Μέση Πληροφορία** ονομάζεται και **εντροπία** και εναλλακτικά συμβολίζεται με ***H(p)***
- Η εντροπία δεν εξαρτάται από τις τιμές της τ.μ X **αλλά από την κατανομή της *X!***
  - δηλ. τις τιμές που πέρνουν τα p_i
- Εάν $\log_2$: μετράται σε **bits,** εάν $\log_{10}$: μετράται σε decits
  - Στη συνέχεια θεωρούμε οτι χρησιμοποιούμε $\log_2$
- Ισχύει πάντοτε $H(X) \geq 0$

Ποσότητα πληροφορίας- Εντροπία

- **Παράδειγμα 1**
  - Έστω *X* μία τυχαία μεταβλητή **με δύο ενδεχόμενα,** πιθανότητας εμφάνισης *p* και (1-*p*) αντίστοιχα.
  - $H(p)=-p\log(p) - (1-p)\log(1-p)$
  - Η **μέγιστη τιμή της εντροπίας** είναι όταν τα ενδεχόμενα είναι **ισοπίθανα**



- **Παράδειγμα 2**

$$X=\begin{cases} a & \mu\varepsilon\ \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ 1/2 \\ b & \mu\varepsilon\ \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ 1/4 \\ c & \mu\varepsilon\ \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ 1/8 \\ d & \mu\varepsilon\ \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ 1/8 \end{cases}$$

$$\mathrm{H(X)} = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4}\,bits$$

Εντροπία Πηγής Πληροφορίας

- **Διακριτή Πηγή Πληροφορίας**
  - Παράγει ακολουθίες **συμβόλων** s_i
  - **Αλφάβητο πηγής** είναι το σύνολο των συμβόλων
    - $S=(s_1,s_2,\ldots,s_n)$, όπου *n* είναι το πλήθος των συμβόλων
  - Παράγει διαδοχικές ακολουθίες συμβόλων που ονομάζονται **μηνύματα**
    - Το πλήθος των δυνατών μηνυμάτων μήκους *l* είναι n^l
      - *Π.χ σε δυαδική πηγή (0,1) με μηνυματα μήκους 10, το πλήθος των δυνατών μηνυμάτων είναι 1024*
  - Η Παραγωγή των συμβόλων λαμβάνει χώρα με κάποια κατανομή πιθανότητας *p*
  - Η παραγωγή κάθε συμβόλου γίνεται
    - Είτε ανεξάρτητα αυτών που έχουν προηγηθεί οπότε αναφερόμαστε σε **διακριτή πηγή χωρίς μνήμη**
    - Είτε εξαρτάται στατιστικά αυτών που έχουν προηγηθεί οπότε αναφερόμαστε σε **διακριτή πηγή με μνήμη**

Εντροπία Πηγής Πληροφορίας

- Μέση ποσότητα πληροφορίας ή **εντροπία των συμβόλων** (bits/symbol)
  - $H(S) = -\sum_{i=1}^{n} p_i \log p_i$
- **Μέγιστη μέση ποσότητα πληροφορίας** (bits/symbol)
  - $\max H(S) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$
- **Πλεονασμός** διακριτής πηγής, [0,1]
  - $red = 1 - \frac{H(S)}{\max H(S)} = 1 - \frac{H(S)}{\log n}$
- **Μέσος Ρυθμός Πληροφορίας της πηγής** (bits/sec)
  - $R = symbol\ rate * H(S)$

Εντροπία Πηγής Πληροφορίας

- **Παράδειγμα:** S={0,1} με p(0)=3/4 και p(1)=1/4.
  - H(S)=0.815 bits/symbol
  - maxH(S)=log2=1 bit/symbol
  - red=1-0.815/1=0.19
  - Αν η πηγή παράγει σύμβολα με ρυθμό 1000 σύμβολα/sec R=0.815 *1000=815 bps.

Εντροπία Πηγής Πληροφορίας

- Μέσο πληροφοριακό περιεχόμενο **μηνυμάτων** της πηγής
  - Εάν γνωρίζουμε ότι η πηγή παράγει μηνύματα μήκους l, με δεδομένο ότι το πλήθος του συνόλου $M=(m_1,m_2,\ldots,m_n)$, των μηνυμάτων είναι n^l, και η πιθανότητα εμφάνισης ενός μηνύματος m_i είναι $p(m_i)$, τότε το μέσο πληροφοριακό περιεχόμενο των μηνυμάτων είναι
    - $H(M) = -\sum_{i=1}^{n^l} p(m_i)\log\, p(m_i)$
    - Αποδεικνύεται ότι ***H(M)= l *H(S)***,
      - δηλαδή το μέσο πληροφοριακό περιεχόμενο ενός μηνύματος είναι ίσο με το άθροισμα της μέσης πληροφορίας που μεταφέρουν τα σύμβολα που το αποτελούν.

- **Παράδειγμα:** Τα 4 μηνύματα μήκους 2 που δημιουργούνται από την πηγή των συμβόλων του προηγούμενου παραδείγματος είναι M={00,01,10,11} και οι πιθανότητες να συμβούν είναι p(00)=9/16, p(01)=p(10)=3/16, p(11)=1/16.
  - Τότε *H(M)*=1.63 =2*0.815 bits/μήνυμα

Εντροπία Πηγής Πληροφορίας

- ***Παράδειγμα***
  - Μια πηγή πληροφορίας παράγει σύμβολα, τα οποία ανήκουν στο αλφάβητο *S={τ, υ, φ, χ, ψ, ω}*. Οι πιθανότητες των συμβόλων αυτών είναι ¼, ¼, 1/8, 1/8, 1/8 και 1/8, αντίστοιχα.Θεωρώντας την πηγή χωρίς μνήμη, ζητείται να υπολογίσετε:
  - α) Το σύμβολο με το μεγαλύτερο και το μικρότερο πληροφορικό περιεχόμενο της πηγής.
  - β) Το μέσο πληροφορικό περιεχόμενο των **συμβόλων της πηγής**,
  - γ) Το μέσο πληροφορικό περιεχόμενο των μηνυμάτων της πηγής **αποτελούμενων από δύο σύμβολα**.
  - δ) Τον πλεονασμό της πηγής (*log6=2,585*) και
  - ε) Το μέσο ρυθμό πληροφορίας της πηγής για ρυθμό 500 συμβόλων /sec.

Εντροπία Πηγής Πληροφορίας

- ***Απάντηση***
  - α) Τα σύμβολα με το μεγαλύτερο πληροφοραικό περιεχόμενο είναι αυτά που έχουν την μικρότερη πιθανότητα
    - δηλαδή *H(i)=-log(1/8)=3 bits* όπου *i=φ,χ,ψ,ω*.
    - Αντίθετα τα σύμβολα με το μικρότερο πληροφορικό περιεχόμενο είναι αυτά που έχουν την μεγαλύτερη πιθανότητα
      - δηλαδή *H(i)=-log(1/4)=2 bits* όπου *i=τ,υ.*
  - β) Το μέσο πληροφορικό περιεχόμενο των συμβόλων της πηγής
    - $H(S) = -\sum_{i=1}^{6} p_i \log p_i = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = (20/8) = 2{,}5 \text{ bits/symbol.}$

Εντροπία Πηγής Πληροφορίας

- **Απάντηση (συνέχεια)**
  - γ) Για τον υπολογισμό του μέσου πληροφορικού περιεχομένου των μηνυμάτων της πηγής αποτελούμενων από 2 σύμβολα, υπολογίζουμε πρώτα τις (συνδυασμένες) πιθανότητες δημιουργίας των μηνυμάτων αυτών.
  - Για τον υπολογισμό της πιθανότητας κάθε μηνύματος αρκεί να πολλαπλασιάσουμε τις πιθανότητες παραγωγής των συμβόλων από τα οποία αποτελείται. Συνολικά έχουμε **6^2=36 μηνύματα**.
  - Παρατηρούμε ότι από τα 36 μηνύματα, 4 έχουν πιθανότητα παραγωγής ίση με (1/16), 16 μηνύματα έχουν πιθανότητα παραγωγής (1/64) και 16 μηνύματα έχουν πιθανότητα παραγωγής (1/32).
    - p1=p(τ,τ)=1/16, p2=p(τ,υ)=1/16, p3=p(τ,φ)=1/32, p4=p(τ,χ)=1/32, p5=p(τ,ψ)=1/32, p6=p(τ,ω)=1/32,
    - p7=p(υ,τ)=1/16, p8=p(υ,υ)=1/16, p9=p(υ,φ)=1/32, p10=p(υ,χ)=1/32, p11=p(υ,ψ)=1/32, p12=p(υ,ω)=1/32,
    - p13=p(φ,τ)=1/32, p14=p(φ,υ)=1/32, p15=p(φ,φ)=1/64, p16=p(φ,χ)=1/64, p17=p(φ,ψ)=1/64, p18=p(φ,ω)=1/64,
    - p19=p(χ,τ)=1/32, p20=p(χ,υ)=1/32, p21=p(χ,φ)=1/64, p22=p(χ,χ)=1/64, p23=p(χ,ψ)=1/64, p24=p(χ,ω)=1/64,
    - p25=p(ψ,τ)=1/32, p26=p(ψ,υ)=1/32, p27=p(ψ,φ)=1/64, p28=p(ψ,χ)=1/64, p29=p(ψ,ψ)=1/64, p30=p(ψ,ω)=1/64,
    - p31=p(ω,τ)=1/32, p32=p(ω,υ)=1/32, p33=p(ω,φ)=1/64, p34=p(ω,χ)=1/64, p35=p(ω,ψ)=1/64, p36=p(ω,ω)=1/64.

Εντροπία Πηγής Πληροφορίας

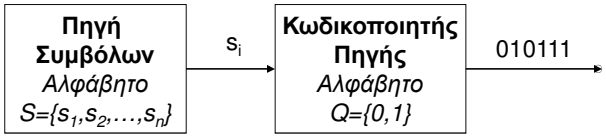
- **Απάντηση (συνέχεια)**
  - Επομένως

$$H(M) = -\sum_{i=1}^{36} p_i \log p_i = -4\frac{1}{16}\log\frac{1}{16} - 16\frac{1}{64}\log\frac{1}{64} - 16\frac{1}{32}\log\frac{1}{32} = (320/64) = 5 \text{ bits/message.}$$

  - Παρατηρούμε ότι επιβεβαιώνεται πως *H(M)=2 H(S)*
    - Αυτό συμβαίνει λόγω του ότι η πηγή είναι χωρίς μνήμη
  - δ) red=1-*H(S)*/max(*H(S)*)=1-*H(S)*/log6=1-(2,5/2,585)=1-0,967=0,0328.
  - ε) *R=rH(S)*=500*(2,5)=1250 bits/sec.

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

- Κωδικοποίηση/συμπίεση της πηγής
  - Είναι η διαδικασία αντιστοίχισης του αλφάβητου των συμβόλων σε ένα άλλο αλφάβητο.
  - Το καινούριο αυτό αλφάβητο ονομάζεται **κωδικό αλφάβητο** και τα μέλη ονομάζονται **κωδικά σύμβολα**.
  - Οι ακολουθίες των κωδικών συμβόλων που αντιστοιχούν σε σύμβολα της πηγής λέγονται **κωδικές λέξεις**



Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

- Απαιτήσεις για χρησιμότητα κωδικών
  - Κάθε ακολουθία κωδικών λέξεων πρέπει να μπορεί να αποκωδικοποιηθεί με μοναδικό τρόπο
  - Η αποκωδικοποίηση πρέπει να γίνεται εύκολα και άμεσα
  - Ο κώδικας πρέπει να πετυχαίνει τη βέλτιστη δυνατή συμπίεση

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

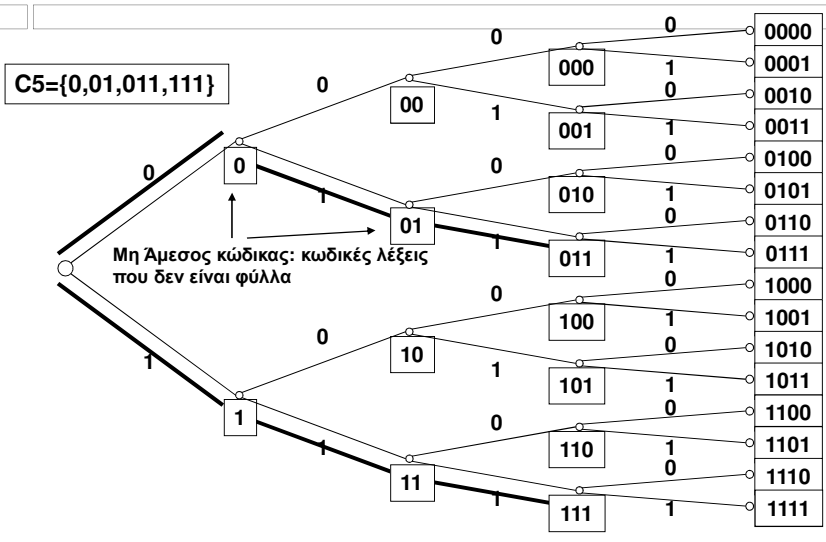
- **Μη ιδιάζων κώδικας**
  - Όταν όλες οι κωδικές λέξεις είναι διαφορετικές
- **Μοναδικά αποκωδικοποιήσιμος**
  - Όταν και οι ακολουθίες των κωδικών λέξεων είναι διαφορετικές
- **Άμεσος ή Προθεματικός κώδικας**
  - Κάθε μοναδικά αποκωδικοποίησιμος κώδικας που επιτρέπει την άμεση αποκωδικοποίηση της κωδικής λέξης χωρίς να χρειάζεται να λάβει υπόψη του τις επόμενες κωδικές λέξεις.
  - Ο άμεσος κώδικας αποτελείται από κωδικές λέξεις οι οποίες δεν αποτελούν μέρος (προθέματα άλλων)

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

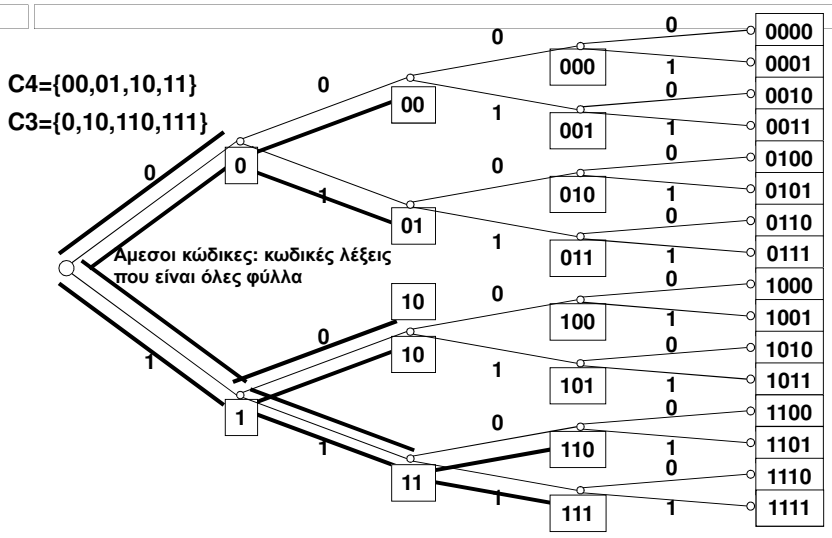
- **Παράδειγμα**
  - Μη ιδιάζων, I,II,III,IV
  - Μοναδικά αποκωδικοποιήσιμος, II,III,IV. Ο I δεν είναι αφού ΦΦ,Ψέχουν κωδική λέξη την ίδια, 00
  - Άμεσοι κώδικες, II και III
  - Ο κώδικας IV δεν είναι άμεσος αφού χρειάζεται να δούμε ψηφίο που ανήκει στην επόμενη κωδική λέξη για να καθοριστεί η τρέχουσα, π.χ. 011**0111**00

	I	II	III	IV
Φ	0	00	0	0
Χ	11	01	10	01
Ψ	00	10	110	011
Ω	01	11	1110	0111

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)



Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)



Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

Σημαντικές ερωτήσεις...

- Μπορούμε να βρούμε ένα κώδικα (άμεσο και αποκωδικοποιήσιμο) του οποίου οι κωδικές λέξεις να έχουν το βέλτιστο δυνατό μήκος; Να έχουν δηλαδή τη κατά μέσο όρο τη μικρότερη τιμή μήκους κωδικής λέξης;
  - $\min \sum_i p_i l_i$
- Υπάρχει σχέση μήκους λέξης και πιθανότητας εμφάνισης συμβόλου πηγής;
- Αν αντιστοιχίσουμε κωδικές λέξεις μικρού μήκους σε σύμβολα με μεγάλη πιθανότητα εμφάνισης, θα μειωθεί το μέσο μήκος της κωδικής λέξης;
- **Ποια είναι η βέλτιστη συμπίεση που είναι δυνατόν να επιτευχθεί;**

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

- **Θεώρημα Κωδικοποίησης Πηγής**
  - Έστω μια πηγή παράγει $S=\{s_1,s_2,\ldots,s_n\}$ σύμβολα με πιθανότητα εμφάνισης κάθε συμβόλου $\{p_1,p_2,\ldots,p_n\}$.
  - Τα σύμβολα αυτά κωδικοποιούνται από ένα κωδικό αλφάβητο q συμβόλων και αντιστοιχίζονται σε άμεσο και αποκωδικοποιήσιμο κώδικα n κωδικών λέξεων μήκους l_i η κάθε μία, $i=1,2,\ldots,n$.
  - Αν $H(C)$ είναι το μέσο πληροφοριακό περιεχόμενο των συμβόλων της πηγής τότε ισχύει

$$\frac{H(C)}{\log_2 q} \le \sum_{i=1}^{n} p_i l_i$$

- Για $q=2$, το βέλτιστο (ελάχιστο) μέσο μήκος κωδικής λέξης είναι ίσο με την εντροπία της πηγής των συμβόλων και δεν μπορεί να είναι μικρότερο από αυτή.

- **Ελάχιστο του μήκους των κωδικών λέξεων**
  - Η ισότητα ισχύει όταν για κάθε κωδική λέξη i, $l_i = \log_q\left(\frac{1}{p_i}\right)$

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

- Άρα δεν μπορούμε λοιπόν να καταλήξουμε σε **κώδικα με μέσο μήκος λέξεων μικρότερο από την εντροπία της πηγής.**

- **Όμως πόσο κοντά σε αυτή την τιμή μπορούμε να φτάσουμε;**
  - Για κάθε τ.μ Χ υπάρχει ένας άμεσος και μοναδικά αποκωδικοποιήσιμος κώδικας C του οποίου η μέση τιμή μήκους, $L(C,X)$, ικανοποιεί τη σχέση
    - $H(X) \le L(C,X) < H(X)+1$
  - Ο κώδικας αυτός έχει κωδικές λέξεις μήκους
    - $l_i^* = \lceil \log_2(1/p_i) \rceil$ όπου $\lceil \chi \rceil$ είναι ο μικρότερος ακέραιος που είναι μεγαλύτερος του χ.

Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

Επίδοση του κώδικα

Η επίδοση, a, ενός κώδικα ορίζεται ως ο λόγος του μέσου πληροφορικού περιεχομένου των συμβόλων της πηγής (ή των κωδικών λέξεων) προς το γινόμενο του μέσου μήκους των κωδικών λέξεων με το λογάριθμο του πλήθους των κωδικών συμβόλων:

$$a = \frac{H(C)}{\left(\sum_{i=1}^{n} p_i l_i\right) \log q}. \qquad (2.9)$$

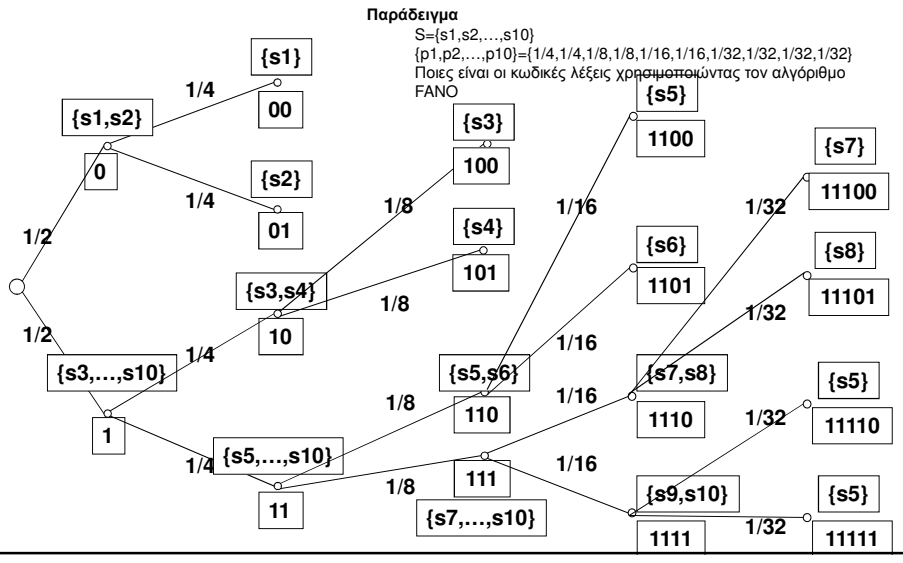
Συμπίεση Πληροφορίας (Κωδικοποίηση Πηγής)

- Αλγόριθμοι κωδικοποίησης πηγής
  - FANO
  - SHANNON
  - HUFFMAN
    - Βέλτιστος κώδικας: Παράγει το σύνολο κωδικών λέξεων με το ελάχιστο μέσο μήκος
- JPEG και MPEG χρησιμοποιούν μεταξύ άλλων και τον αλγόριθμο Huffman

Αλγόριθμος Κωδικοποίησης FANO

- **Βήμα 1ο**: Τα σύμβολα (ή τα μηνύματα) ταξινομούνται έτσι ώστε οι πιθανότητές τους είναι σε φθίνουσα ακολουθία.
- **Βήμα 2ο**: Στη συνέχεια τα σύμβολα χωρίζονται σε ομάδες ο αριθμός των οποίων είναι ίσος με τον αριθμό των κωδικών συμβόλων (στην περίπτωση δυαδικού κώδικα οι ομάδες χωρισμού συμβόλων είναι δύο).
  - Το κριτήριο σχηματισμού της κάθε ομάδας είναι τέτοιο ώστε αφενός να διατηρείται η σειρά των συμβόλων όπως αυτή έχει καθοριστεί από το βήμα 1
  - Αφετέρου δε να ελαχιστοποιείται η διαφορά στο άθροισμα πιθανοτήτων εμφάνισης συμβόλων μεταξύ των ομάδων
  - min $\left|\sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i\right|$
- **Βήμα 3ο**: Για κάθε μία ομάδα συμβόλων που δημιουργήσαμε αντιστοιχίζουμε ένα από τα κωδικά σύμβολα ως το πρώτο τον κωδικών λεξεων που θα προκύψουν
- **Βήμα 4ο**: Επαναλαμβάνουμε τα βήματα 2 & 3 για κάθε μία από τις ομάδες προσθέτοντας κάθε φορά και από ένα κωδικό σύμβολο στην κωδική λέξη μέχρι να δημιουργήσουμε ομάδες με ένα μόνο σύμβολο

Αλγόριθμος Κωδικοποίησης FANO



Αλγόριθμος κωδικοποίησης SHANNON

- **Βήμα 1ο**: Τα σύμβολα (ή τα μηνύματα) ταξινομούνται έτσι ώστε οι πιθανότητές τους είναι σε φθίνουσα ακολουθία, όπως ακριβώς και του FANO.
- **Βήμα 2ο**: Για κάθε σύμβολο s_i του οποίου η πιθανότητα εμφάνισης είναι $p(s_i)$ υπολογίζεται η αθροιστική πιθανότητα P_i ως εξής:
  - $P_i = \sum_{j=1}^{i-1} p(s_j), \quad P_1 = 0, \quad i = 2,...,n$
- **Βήμα 3ο**: Το μήκος της κωδικής λέξης που αντιστοιχεί στο σύμβολο s_i ισούται με τον ακέραιο αριθμό l_i, που πληροί τη σχέση
  - $l_i = \lceil \log_2 (1/p(s_i)) \rceil$
- **Βήμα 4ο**: Η κωδική λέξη c_i που αντιστοιχεί στο σύμβολο πηγής s_i είναι το δυαδικό ανάπτυγμα του κλάσματος P_i (μόνο τα πρώτα l_i bits αναπτύγματος)

Αλγόριθμος κωδικοποίησης SHANNON

(Για το δυαδικό ανάπτυγμα ενός κλάσματος ισχύει το εξής:

$$\frac{a_1}{2} + \frac{a_2^2}{2^2} + ... + \frac{a_k^k}{2^k} = .a_1 a_2 ... a_k \text{ όπου } a_j \text{ είναι 0 ή 1.)}$$

Π.χ Το $P_2 = 1/4$ γράφεται $(0/2^1) + (1/2^2) + (0/2^3) + (0/2^4) + (0/2^5)$ και το δυαδικό του ανάπτυγμα είναι οι αριθμητές των κλασμάτων, δηλαδή .01000 και επομένως η κωδική λέξη του συμβόλου S_2 μήκους 2 bits είναι «01».

Αλγόριθμος κωδικοποίησης SHANNON

Παράδειγμα

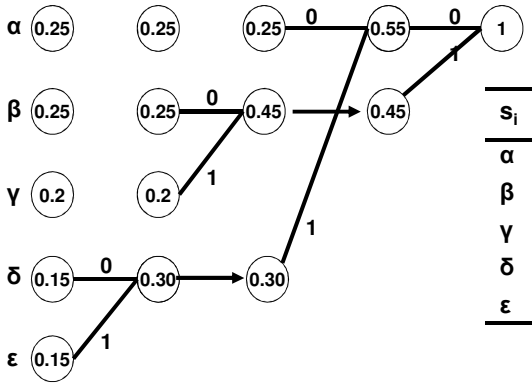
Σύμβολα Πηγής	Πιθανότητες Συμβόλων	P_i	Μήκος l_i	Ανάπτυγμα του P_i	Κωδικές Λέξεις
S_1	1/4	$P_1 = 0$	$l_1 = 2$	.00000	00
S_2	1/4	$P_2 = 1/4$	$l_2 = 2$	.01000	01
S_3	1/8	$P_3 = 1/2$	$l_3 = 3$	.10000	100
S_4	1/8	$P_4 = 5/8$	$l_4 = 3$	.10100	101
S_5	1/16	$P_5 = 3/4$	$l_5 = 4$	.11000	1100
S_6	1/16	$P_6 = 13/16$	$l_6 = 4$	.11010	1101
S_7	1/32	$P_7 = 7/8$	$l_7 = 5$	.11100	11100
S_8	1/32	$P_8 = 29/32$	$l_8 = 5$	.11101	11101
S_9	1/32	$P_9 = 15/16$	$l_9 = 5$	.11110	11110
S_{10}	1/32	$P_{10} = 31/32$	$l_{10} = 5$	.11111	11111

Αλγόριθμος κωδικοποίησης HUFFMAN

- **Αλγόριθμος κωδικοποίησης HUFFMAN**
  - Ο αλγόριθμος Huffman κατασκευάζει το δυαδικό δέντρο αρχίζοντας από τα φύλλα του και προχωράει προς τη ρίζα του.
  - **Βήμα 1ο**: Τα σύμβολα (ή τα μηνύματα) ταξινομούνται έτσι ώστε οι πιθανότητές τους είναι σε φθίνουσα ακολουθία.
  - **Βήμα 2ο**: Στη συνέχεια παίρνουμε τα δύο σύμβολα με τις μικρότερες πιθανότητες. Γι' αυτά μέσα από την διαδικασία του αλγορίθμου θα αναθέσουμε, τις μακρύτερες δυνατές κωδικές λέξεις έτσι ώστε αυτές να έχουν το ίδιο μήκος και να διαφέρουν στο τελευταίο τους ψηφίο.
    - Το βήμα αυτό θα δημιουργήσει το τελευταίο από τα ψηφία της κωδικής λέξης
  - **Βήμα 3ο**: Συνδυάζοντας τα δύο σύμβολα που επιλέξαμε στο βήμα 2 σε ένα και αναθέτοντας στο συνδυασμένο σύμβολο το άθροισμα των πιθανοτήτων των επιμέρους συμβόλων
    - επαναλαμβάνουμε τη διαδικασία από το βήμα 1 μεταξύ των συμβόλων που απομένουν και του συμβόλου που δημιουργήσαμε μέχρις ότου καταλήξουμε σε ένα σύμβολο με πιθανότητα 1.
  - **Βήμα 4ο**: Οι κωδικές λέξεις που αντιστοιχούν στο κάθε σύμβολο αποτελούνται από τις ακολουθίες 0 και 1 που δημιουργούνται αν διατρέξουμε το δένδρο που δημιουργήθηκε από τον κόμβο με το μοναδικό σύμβολο προς τα σύμβολα από τα οποία ξεκινήσαμε

Αλγόριθμος κωδικοποίησης HUFFMAN

- **Παράδειγμα**



s_i	p_i	H	l_i	C(s_i)
α	0,25	2.0	2	00
β	0.25	2.0	2	10
γ	0.2	2.3	2	11
δ	0.15	2.7	3	010
ε	0.15	2.7	3	011
		2.2855	2.30	

Αλγόριθμος κωδικοποίησης HUFFMAN

- Παρατηρήσεις σχετικά με τον αλγόριθμο Huffman
  - Αποδεικνύεται ότι κανένας άλλος αλγόριθμος δεν μπορεί να οδηγήσει στην κατασκευή κώδικα με μικρότερο μέσο μήκος κωδικών λέξεων για ένα δεδομένο αλφάβητο πηγής.
  - Η κατασκευή του δένδρου γίνεται από τα φύλλα προς τη ρίζα του δένδρου,
  - Σε αντίθεση με τον αλγόριθμο FANO όπου δημιουργεί το δένδρο από τη ρίζα του προς τα φύλλα διαιρώντας κάθε φορά το σύνολο συμβόλων σε δύο σύνολα.
    - Ένας τέτοιος αλγόριθμος είναι πάντα λιγότερο βέλτιστος.

Κώδικες Ακολουθιών Συμβόλων

- Μειονέκτημα αλγορίθμου Huffman
  - Γνωρίζουμε ότι ο αλγόριθμος Huffman παράγει βέλτιστο κώδικα και άρα βάσει του θεωρήματος ισχύει ότι
    - $H(X) \leq L(C,X) < H(X)+1$
    - Άρα κατά μέσο όρο και ανά σύμβολο έχουμε πλεονάζοντα bits μεταξύ 0 και 1.
    - Αν η εντροπία H(X) της πηγής είναι μεγάλη τότε το πλεονάζον αυτό bit, L(C,X)-H(X), θα ήταν αμελητέο στην παραγωγή μηνυμάτων. **Αν όμως η εντροπία είναι μικρότερη από 1 bit τότε το πλεονάζον bit θα ήταν καθοριστικό στην παραγωγή μηνυμάτων.**

Κώδικες Ακολουθιών Συμβόλων

- **Παράδειγμα**
  - Έστω μία πηγή με αλφάβητο $A=\{\alpha_1,\alpha_2,\alpha_3\}$ και $P=\{0.8,0.02,0.18\}$.
  - H(A)=0,816 bits/symbol
  - Εφαρμόζοντας τον αλγόριθμο του Huffman παίρνουμε τις εξής κωδικές λέξεις και μήκη
  - Παρατηρούμε ότι το μέσο μήκος κωδικής λέξης είναι 1.2 bits/symbol το οποίο απέχει κατά 47% από την εντροπία δηλαδή υπάρχει πλεονασμός κατά 0.384 bits/symbol.
  - Σε επίπεδο μηνυμάτων (ακολουθίες συμβόλων) αυτός ο πλεονασμός παίζει καθοριστικό ρόλο
    - Π.χ. Για ακολουθίες μηνυμάτων που αποτελούνται από N=1000 σύμβολα τότε σύμφωνα με την κωδικοποίηση κατά Huffman θα παράγαμε κατά μέσο όρο 384 bits περισσότερα από τα 816 που είναι τα αναγκαία
  - Τι πρέπει να γίνει;

α_i	p_i	H	l_i	$C(\alpha_i)$
α_1	0.8	0.322	1	0
α_2	0. 02	5.644	2	11
α_3	0.18	2.474	2	10
		0,816	1,2	
		H(A)	L(C,A)	

Κώδικες Ακολουθιών Συμβόλων

- **Παράδειγμα (συνέχεια)**
  - Να εφαρμόσουμε τον αλγόριθμο όχι σε επίπεδο συμβόλων **αλλά σε επίπεδο μηνυμάτων.**
  - Έτσι για μηνύματα δύο συμβόλων έχουμε
    - Το μέσο μήκος κάθε κωδικής λέξης που αντιστοιχεί σε μήνυμα 2 συμβόλων είναι 1.7228 bits το οποίο συγκρινόμενο με την εντροπία $H(A^2)$=1,632 είναι μόλις κατά 5.5% αυξημένο
  - Το πρόβλημα που παρουσιάζει αυτή η μέθοδος στην πράξη είναι ότι **χρειάζεται να υπολογίσουμε όλες τις πιθανότητες των πιθανών μηνυμάτων.**
    - Για ένα αλφάβητο με n σύμβολα και μηνύματα μήκους m τότε το σύνολο όλων των μηνυμάτων είναι n^m
      - δηλαδή για ένα αλφάβητο 2 συμβόλων και μηνύματα μήκους 20 θα χρειαστεί να υπολογίσουμε περίπου 1 εκ. Πιθανότητες για τα διαφορετικά μηνύματα

α_i	p_i	H	l_i	$C(\alpha_i)$
$\alpha_1\alpha_1$	0.64	0.644	1	0
$\alpha_1\alpha_2$	0. 016	5.966	5	10101
$\alpha_1\alpha_3$	0.144	2.796	2	11
$\alpha_2\alpha_1$	0.016	5.966	6	101000
$\alpha_2\alpha_2$	0. 0004	11.288	8	10100101
$\alpha_2\alpha_3$	0.0036	8.118	7	1010011
$\alpha_3\alpha_1$	0. 144	2.796	3	100
$\alpha_3\alpha_2$	0.0036	8.118	8	10100100
$\alpha_3\alpha_3$	0.0324	4.948	4	1011
		1,632 $H(A^2)$	1,7228 $L(C,A^2)$	