ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

# Εντροπία Διακριτής Πηγής Χωρίς Μνήμη

## Θεωρία Πληροφορίας και Κωδίκων

*Εαρινό Εξάμηνο*

Τμήμα Πληροφορικής και Τηλεπικοινωνίων

Νικόλαος Χ. Σαγιάς
*Καθηγητής*
Webpage: https://eclass.uop.gr/courses/DIT221/
e-mail: nsagias@uop.gr

# Εντροπία Πηγής

- Έστω αλφάβητο αποτελούμενο από $N$ σύμβολα, πχ $\mathscr{L} = \{\alpha, \beta, \gamma, \ldots, \varphi, \chi, \psi, \omega\}$ με $N = 24$

- Τότε, η ποσότητα πληροφορίας (*information measure*) κατά Hartley (σε nat) ορίζεται ως

$$H(\mathscr{L}) = \log_e(N) = \ln(N)$$

- Με βάση τα $N$ σύμβολα σχηματίζουμε λέξεις (ή μηνύματα) μήκους $k$ συμβόλων, πχ πληροφορια

- Τότε, το πλήθος των διαφορετικών λέξεων θα είναι $N^k$ και άρα, η ποσότητα πληροφορίας

$$H(\mathscr{L}^k) = \ln(N^k) = k \ln(N)$$

- Με τον παραπάνω ορισμό η ποσότητα πληροφορίας αυξάνει ανάλογα με το μήκος της λέξης

# Εντροπία Πηγής

- Παράδειγμα: Η λέξη `πληροφορια` (μήκος $k = 10$), με $H(\mathscr{L}^{10}) = 31.78$ nat, περιέχει 10πλάσια ποσότητα πληροφορίας σε σχέση με τη λέξη `π` (μήκος $k = 1$), με $H(\mathscr{L}) = 3.178$ nat

- Ο προηγούμενος ορισμός της ποσότητας πληροφορίας προϋποθέτει ότι όλες τα σύμβολα εμφανίζονται με την ίδια πιθανότητα

- Ανάλογα με τη βάση του λογαρίθμου, υπάρχουν διαφορετικές μονάδες ποσότητας πληροφορίας
  - Μονάδα ποσότητα πληροφορίας με βάση λογαρίθμου το 10, είναι το decit (***deci**mal unit*)
  - Μονάδα ποσότητα πληροφορίας με βάση λογαρίθμου το $e$, είναι το nat (***na**tural uni**t***)
  - Μονάδα ποσότητα πληροφορίας με βάση λογαρίθμου το 2, είναι το bit (***bi**nary uni**t***)

- Στη συνέχεια χρησιμοποιούμε το bit ως μονάδα της ποσότητας πληροφορίας

# Εντροπία Πηγής

- Ιδιότητα #1 της συνάρτησης λογάριθμος

$$\log_\beta\left(x^k\right) = k\,\log_\beta\left(x\right)$$

- Ιδιότητα #2 της συνάρτησης λογάριθμος

$$\log_\beta\left(x\,y\right) = \log_\beta\left(x\right) + \log_\beta\left(y\right)$$

$$\log_\beta\left(\frac{x}{y}\right) = \log_\beta\left(x\right) - \log_\beta\left(y\right)$$

- Ιδιότητα #3 της συνάρτησης λογάριθμος

$$\log_\beta\left(x\right) = \frac{\log_a\left(x\right)}{\log_a\left(\beta\right)}$$

# Εντροπία Πηγής

- Διαισθητικά, ένα γεγονός σπάνιο περιέχει μεγάλη ποσότητα πληροφορίας

- Για παράδειγμα, έστω τα παρακάτω γεγονότα πρόγνωσης καιρού:
  - Ενδεχόμενο #1: Ηλιόλουστη μέρα
  - Ενδεχόμενο #2: Έντονη βροχόπτωση
  - Ενδεχόμενο #3: Σφοδρή χιονόπτωση

- Το 3ο ενδεχόμενο έχει τη μικρότερη πιθανότητα εμφάνισης ανάμεσα από τα τρία γεγονότα και προκαλεί τη μεγαλύτερη έκπληξη

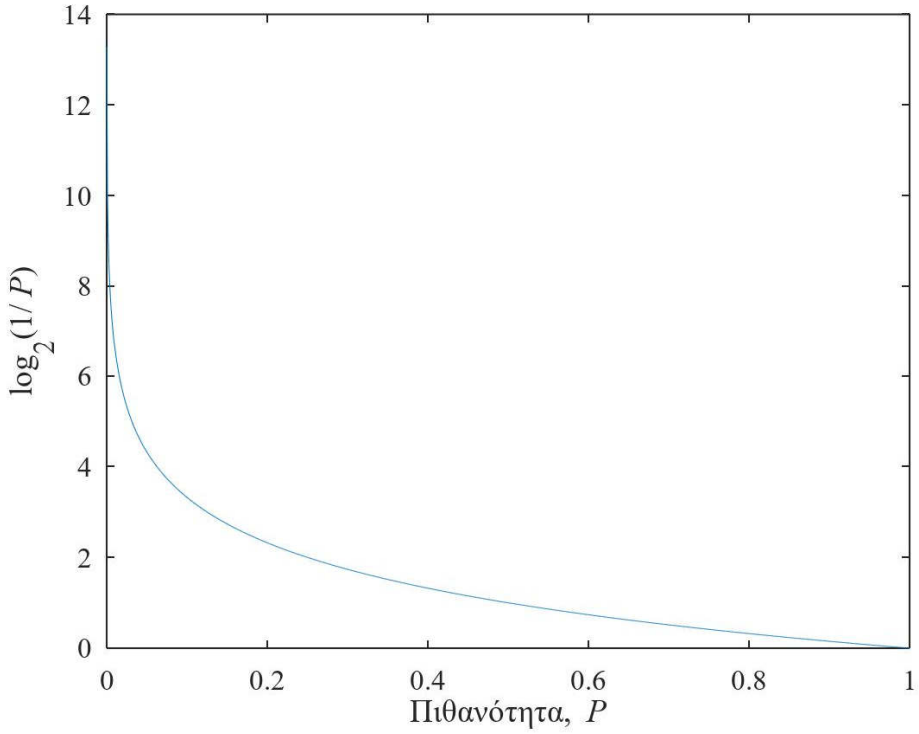
- Δηλαδή, το σπανιότερο γεγονός έχει την μεγαλύτερη αβεβαιότητα (*uncertainty*) και συνεπώς, περιέχει την μεγαλύτερη ποσότητα πληροφορίας

# Εντροπία Πηγής

- Διαισθητικά, η ποσότητα πληροφορίας δηλαδή ενέχει τα εξής χαρακτηριστικά:
  - Για πιθανότητα εμφάνισης γεγονότος $P \to 0$, η ποσότητα πληροφορίας είναι $I \to \infty$
  - Για πιθανότητα εμφάνισης γεγονότος $P = 1$, η ποσότητα πληροφορίας είναι $I = 0$

- Μία τέτοια σχέση προκύπτει από τη σχέση

$$I = \log_2\left(\frac{1}{P}\right)$$



# Εντροπία Πηγής

- Μπορούμε να οδηγηθούμε στην ίδια σχέση πριν, η οποία προέκυψε διαισθητικά, ακολουθώντας πιο αυστηρή λογική

- Έστω ένα σύνολο από ισοπίθανα σύμβολα πλήθους $n$
  - Για $n = 2$ σύμβολα $\mathscr{L} = \{m_1, m_2\}$, απαιτείται 1 bit για κωδικοποίηση κάθε συμβόλου

    $\{0, 1\}$

  - Για $n = 4$ σύμβολα $\mathscr{L} = \{m_1, m_2, m_3, m_4\}$, απαιτούνται 2 bit για κωδικοποίηση κάθε συμβόλου

    $\{00, 01, 10, 11\}$

  - Για $n = 8$ σύμβολα $\mathscr{L} = \{m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8\}$, απαιτούνται 3 bit για κάθε σύμβολο

    $\{000, 001, 010, 011, 100, 101, 110, 111\}$

# Εντροπία Πηγής

- Συνεπώς, για ένα σύνολο συμβόλων πλήθους $n$ με ίδιες πιθανότητες εμφάνισης $P = 1/n$, η απαιτούμενη ποσότητα πληροφορίας είναι και πάλι $I = \log_2(n)$ ή αλλιώς

$$I = \log_2\left(\frac{1}{P}\right)$$

- Η ποσότητα πληροφορίας, δηλαδή, ερμηνεύεται ως ο ελάχιστος αριθμός bit που χρειάζονται για την κωδικοποίηση ενός συμβόλου

# Εντροπία Πηγής

- Πώς θα είναι η σχέση για την ποσότητα πληροφορίας αν τα $n$ σύμβολα δεν είναι ισοπίθανα;

- Για να απαντηθεί το παραπάνω ερώτημα πρέπει να εισάγουμε την μέση ποσότητα πληροφορία ανά σύμβολο ή αλλιώς την εντροπία (*entropy*) της πηγής
- Έστω $n$ ανεξάρτητα σύμβολα (πηγή χωρίς μνήμη) με πιθανότητα εμφάνισης $P_i$ για το σύμβολο $m_i$, δηλαδή $\mathscr{L} = \{m_1, m_2, \ldots, m_n\}$
- Η ποσότητα πληροφορίας για το σύμβολο $m_i$ θα είναι

$$I_i = \log_2\left(\frac{1}{P_i}\right)$$

- Αφού κάθε φορά η πηγή παράγει ένα από τα $n$ σύμβολα, συνολικά ισχύει

$$\sum_{i=1}^{n} P_i = 1$$

# Εντροπία Πηγής

- Η εντροπία της πηγής θα δίδεται από τον μέσο όρο των ποσοτήτων πληροφορίας κάθε συμβόλου

$$H(\mathscr{L}) = \sum_{i=1}^{n} P_i I_i = \sum_{i=1}^{n} P_i \log_2 \left( \frac{1}{P_i} \right) \Leftrightarrow$$

$$H(\mathscr{L}) = -\sum_{i=1}^{n} P_i \log_2 (P_i)$$

- Ένα σημαντικό ζήτημα είναι η κατανομή των πιθανοτήτων που μεγιστοποιεί την εντροπία

- Δεδομένου ότι η εντροπία είναι μέτρο της αβεβαιότητας, προκύπτει ότι η κατανομή που μεγιστοποιεί την εντροπία είναι αυτή με τη μέγιστη αβεβαιότητα

- Η κατανομή με τη μέγιστη αβεβαιότητα προκύπτει όταν όλα τα σύμβολα είναι ισοπίθανα

# Εντροπία Πηγής

- Η εντροπία πηγής εκφρασμένη με βάση λογαρίθμων $b$ μπορεί να εκφραστεί με λογάριθμούς βάσης $a$ ως

$$H_a(\mathscr{L}) = \frac{H_b(\mathscr{L})}{\log_b(a)}$$

- Χρησιμοποιώντας την ιδιότητα της αλλαγής βάσης των λογαρίθμων $\log_a(x) = \log_b(x) / \log_b(a)$, εύκολα καταλήγουμε στην παραπάνω ιδιότητα

$$H_a(\mathscr{L}) = -\sum_{i=1}^{n} P_i \log_a(P_i) = -\sum_{i=1}^{n} P_i \frac{\log_b(P_i)}{\log_b(a)} = \frac{H_b(\mathscr{L})}{\log_b(a)}$$

# Εντροπία Πηγής

- Έστω ένα σύνολο $n$ συμβόλων $m_i$, με αντίστοιχες πιθανότητες εμφάνισης $P_i$ (με $i = 1, 2, \ldots, n$)

- Η μέγιστη τιμή της εντροπίας προκύπτει για ισοπίθανα σύμβολα με $\tilde{P}_i = 1/n$ να είναι

$$H_{\max} = \log_2(n)$$

- Ουσιαστικά πρόκειται για το εξής [πρόβλημα βελτιστοποίησης]

$$\max_{\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_n} \left\{ -\sum_{i=1}^{n} P_i \log_2 (P_i) \right\},$$

$$\text{υπό συνθήκη} \quad \sum_{i=1}^{n} P_i = 1 .$$

# Εντροπία Πηγής

- Για να βρούμε τη μέγιστη εντροπία αρχικά ορίζουμε την συνάρτηση

$$f\left(P_1, P_2, \ldots, P_n\right) = -\sum_{i=1}^{n} P_i \log_2\left(P_i\right) + \lambda\left(-1 + \sum_{i=1}^{n} P_i\right)$$

με παράμετρο $\lambda$ που ονομάζεται πολλαπλασιαστής Lagrange

- Τα μέγιστα της $f$ μπορεί να προκύψουν από τη λύση του συστήματος για κάθε $j = 1, 2, \ldots, n$

$$\left.\frac{\partial f\left(P_1, P_2, \ldots, P_n\right)}{\partial P_j}\right|_{\tilde{P}_j} = 0 \Leftrightarrow -\log_2\left(\tilde{P}_j\right) - \tilde{P}_j \frac{1}{\tilde{P}_j \ln(2)} + \lambda = 0 \Leftrightarrow \tilde{P}_j = 2^{\lambda - 1/\ln(2)} = \frac{1}{e} 2^{\lambda}$$

- Παρατηρούμε ότι με την παραπάνω λύση, το $\tilde{P}_j$ είναι ανεξάρτητο του $j$ και συνεπώς, η λύση είναι κοινή για όλα τα $\tilde{P}_j$

# Εντροπία Πηγής

- Δεδομένου ότι το άθροισμα των $\tilde{P}_j$ ισούται με μονάδα, εύκολα προκύπτει

$$\sum_{j=1}^{n} \tilde{P}_j = 1 \Leftrightarrow \sum_{j=1}^{n} \frac{1}{e} 2^{\tilde{\lambda}} = 1 \Leftrightarrow \tilde{\lambda} = \log_2\left(\frac{e}{n}\right)$$

- Αντικαθιστώντας το $\tilde{\lambda}$ που βρήκαμε στην προηγούμενη εξίσωση, όλα τα $\tilde{P}_j$ είναι ίδια

$$\tilde{P}_j = \frac{1}{n}$$

- Συνεπώς, η μέγιστη εντροπία της πηγής θα είναι

$$H_{\max} = -\sum_{i=1}^{n} \tilde{P}_i \log_2(P_i) = -\sum_{i=1}^{n} \frac{1}{n} \log_2\left(\frac{1}{n}\right) = -n\frac{1}{n} \log_2\left(\frac{1}{n}\right) \Leftrightarrow$$

$$H_{\max} = \log_2(n)$$

# Εντροπία Πηγής

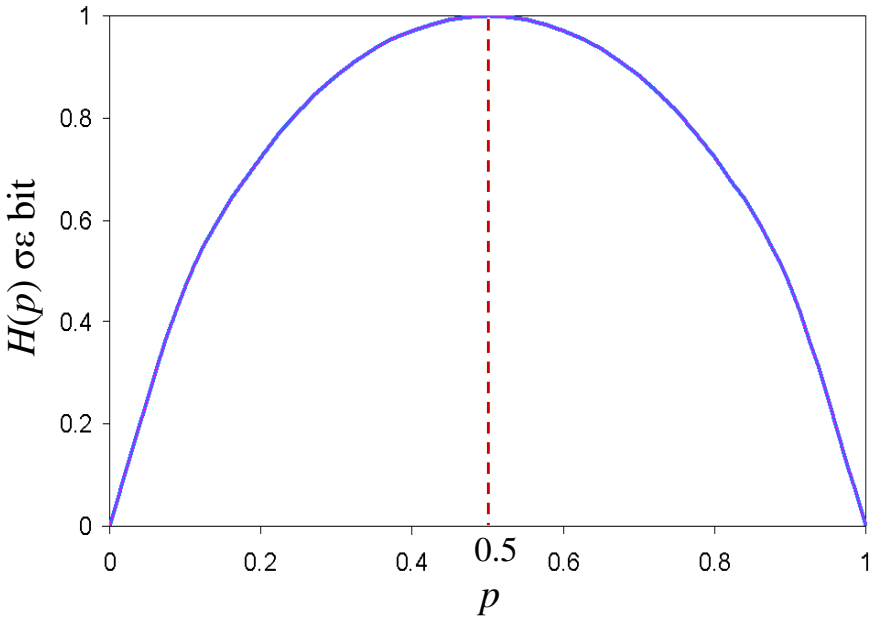
- Έστω ότι μία πηγή παράγει δύο σύμβολα $m_1$ και $m_2$ με πιθανότητες $P_1 = p$ και $P_2 = q$, αντίστοιχα

- Δεδομένου ότι $p + q = 1$, η εντροπία της πηγής είναι

$$H(p) = -\sum_{i=1}^{2} P_i \log_2(P_i) = -p\log_2(p) - q\log_2(q) = -p\log_2(p) - (1-p)\log_2(1-p)$$

- Απίθανες ($p = 0$) και βέβαιες ($p = 1$) περιπτώσεις έχουν μηδενική εντροπία

- Η μέγιστη εντροπία εμφανίζεται για $p = q = ½$, δηλαδή όταν η αβεβαιότητα είναι μέγιστη



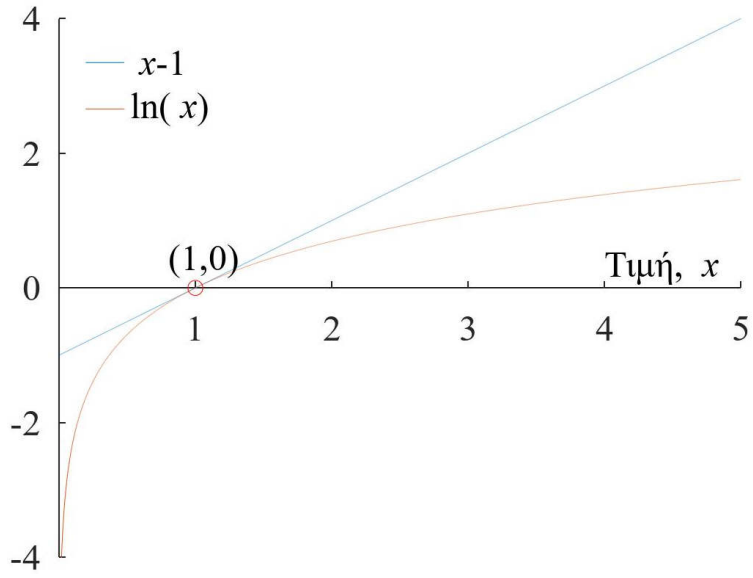
# Εντροπία Πηγής

- Εναλλακτικά, το άνω όριο της εντροπίας μπορεί να υπολογιστεί βάσει της ανισότητας

$$\ln(x) \leq x - 1, \qquad \text{με} \quad x > 0$$

- Η παραπάνω ανισότητα για τη συνάρτηση λογάριθμος με βάση 2 γράφεται

$$\log_2(x) = \frac{\ln(x)}{\ln(2)} \leq \frac{x-1}{\ln(2)} = (x-1)\,\log_2(e)$$



# Εντροπία Πηγής

- Από τον ορισμό της εντροπίας της πηγής

$$H(\mathscr{L}) - \log_2(n) = \sum_{i=1}^{n} P_i \log_2\left(\frac{1}{P_i}\right) - \log_2(n) = \sum_{i=1}^{n} P_i \log_2\left(\frac{1}{P_i}\right) + \sum_{i=1}^{n} P_i \log_2\left(\frac{1}{n}\right) =$$

$$= \sum_{i=1}^{n} P_i \log_2\left(\frac{1}{n\,P_i}\right) \leq \sum_{i=1}^{n} P_i \left(\frac{1}{n\,P_i} - 1\right) \log_2(e) =$$

$$= \log_2(e) \sum_{i=1}^{n} P_i \left(\frac{1}{n\,P_i}\right) - \log_2(e) \sum_{i=1}^{n} P_i$$

$$= \log_2(e)\,\frac{1}{n} \sum_{i=1}^{n} 1 - \log_2(e) = 0 \Leftrightarrow$$

$$H(\mathscr{L}) - \log_2(n) \leq 0 \Leftrightarrow H(\mathscr{L}) \leq \log_2(n)$$

- Συνεπώς. η μέγιστη τιμή της εντροπίας είναι $H_{\max} = \log_2(n)$

# Εντροπία Πηγής

- Ιδιότητα: Η εντροπία μιας πηγής είναι μη αρνητική ποσότητα, δηλαδή

$$H(\mathscr{L}) \geq 0$$

- Δεδομένου ότι η πιθανότητα του συμβόλου $i$ (για όλα τα $P_i$) φράσσεται ως $0 \leq P_i \leq 1$, έχουμε

$$P_i \geq 0 \Leftrightarrow \log_2(P_i) \geq -\infty \Leftrightarrow -\log_2(P_i) \leq \infty \Leftrightarrow -\frac{1}{P_i}\log_2(P_i) \leq \infty \Leftrightarrow -\sum_{i=1}^{n}\frac{1}{P_i}\log_2(P_i) \leq \infty \Leftrightarrow$$

$$H(\mathscr{L}) \leq \infty$$

και

$$P_i \leq 1 \Leftrightarrow \log_2(P_i) \leq 0 \Leftrightarrow -\log_2(P_i) \geq 0 \Leftrightarrow -\frac{1}{P_i}\log_2(P_i) \geq 0 \Leftrightarrow -\sum_{i=1}^{n}\frac{1}{P_i}\log_2(P_i) \geq 0 \Leftrightarrow$$

$$H(\mathscr{L}) \geq 0$$

# Εντροπία Πηγής

- Αν η πηγή παράγει $R_s$ σύμβολα ανά δευτερόλεπτο, ο μέσος ρυθμός πληροφορίας είναι

$$R = R_s H$$

- Ο μέσος ρυθμός πληροφορίας έχει μονάδες (symbol/sec) × (bit/symbol) = bit/sec (ή bps)

- Επίσης, ορίζουμε τον χρόνο μεταξύ δύο διαδοχικών συμβόλων (σε sec/symbol) ως

$$T_s = \frac{1}{R_s}$$

# Εντροπία Πηγής

- Παράδειγμα: Μια πηγή παράγει ένα σύμβολο κάθε 5 ms ανάμεσα από $n = 32$ ισοπίθανα σύμβολα
- Η εντροπία της πηγής είναι

$$H(\mathscr{L}) = -\sum_{i=1}^{n} P_i \log_2(P_i) = -\sum_{i=1}^{n} \frac{1}{32} \log_2\left(\frac{1}{32}\right) = 5 \text{ bit}$$

- Πράγματι, 5 bit αρκούν για να περιγράψουν 32 καταστάσεις ($2^5 = 32$)
- Ο μέσος ρυθμός μετάδοσης πληροφορίας είναι $R = R_s H = (1/0.005) \times 5$ bps = 1 kbps

| | | | |
|---|---|---|---|
| 00000 | 01000 | 10000 | 11000 |
| 00001 | 01001 | 10001 | 11001 |
| 00010 | 01010 | 10010 | 11010 |
| 00011 | 01011 | 10011 | 11011 |
| 00100 | 01100 | 10100 | 11100 |
| 00101 | 01101 | 10101 | 11101 |
| 00110 | 01110 | 10110 | 11110 |
| 00111 | 01111 | 10111 | 11111 |

# Εντροπία Πηγής

- Παράδειγμα: Μια πηγή παράγει ένα σύμβολο κάθε 5 ms ανάμεσα από ένα σύνολο $n = 8$ συμβόλων με πιθανότητες {1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64}
- Η εντροπία της πηγής είναι

$$H(\mathscr{L}) = -\sum_{i=1}^{n} P_i \log_2(P_i) =$$

$$= -\frac{1}{2}\left(\frac{1}{2}\right) - \frac{1}{4}\left(\frac{1}{4}\right) - \frac{1}{8}\left(\frac{1}{8}\right) - \frac{1}{16}\left(\frac{1}{16}\right) - \frac{1}{64}\left(\frac{1}{64}\right) - \frac{1}{64}\left(\frac{1}{64}\right) - \frac{1}{64}\left(\frac{1}{64}\right) - \frac{1}{64}\left(\frac{1}{64}\right) = 2 \text{ bit}$$

- Με 2 bit για κάθε σύμβολο δεν είναι δυνατό να κωδικοποιήσουμε 8 σύμβολα
- Τα 2 bit προκύπτουν αν αντιστοιχίσουμε τα σύμβολα με κωδικολέξεις μη σταθερού μήκους, πχ {`0, 10, 110, 1110, 111100, 111101, 111110, 111111`}, με μήκη $l_i = \{1, 2, 3, 4, 6, 6, 6, 6\}$
- Κατά μέσον όρο έχουμε 2 bit ανά κωδικολέξη αφού το μέσο μήκος είναι

$$L = \sum_{i=1}^{n} P_i l_i = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + \frac{1}{64}\times 6 + \frac{1}{64}\times 6 + \frac{1}{64}\times 6 + \frac{1}{64}\times 6 = 2 \text{ bit}$$

- Ο μέσος ρυθμός μετάδοσης πληροφορίας είναι $R = R_s H = (1/0.005) \times 2$ bps = 0.4 kbps

# Εντροπία Πηγής

- Με βάση τα $n$ σύμβολα πηγής σχηματίζονται μηνύματα μήκους $\ell$ (πλήθος μηνυμάτων $n^\ell$)

- Αν η πιθανότητα εμφάνισης του $i$-ωστού μηνύματος είναι $\mathbb{P}\{m_i\}$ (με $i = 1, 2, .., n^\ell$), ορίζουμε το μέσο πληροφοριακό περιεχόμενο των μηνυμάτων (σε bit ανά μήνυμα)

$$H(\mathscr{L}^\ell) = \sum_{i=1}^{n^\ell} \mathbb{P}\{m_i\} \log_2 \left( \frac{1}{\mathbb{P}\{m_i\}} \right)$$

- Αποδεικνύεται ότι για στατιστικά ανεξάρτητα μεταξύ τους σύμβολα, η εντροπία των μηνυμάτων είναι $\ell$ φορές μεγαλύτερη της εντροπίας των συμβόλων της πηγής

$$H(\mathscr{L}^\ell) = \ell \, H(\mathscr{L})$$

# Εντροπία Πηγής

- Παράδειγμα: Έστω πηγή παράγει σύμβολα $\mathscr{L} = \{s_1, s_2, s_3\}$ με πιθανότητες $P_1 = ¼$, $P_2 = ¼$, $P_3 = ½$
- Η εντροπία της πηγής είναι

$$H(\mathscr{L}) = -\sum_{i=1}^{3} P_i \log_2(P_i) = -\frac{1}{4}\left(\frac{1}{4}\right) - \frac{1}{2}\left(\frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2}\right) = 1.5 \text{ bit}$$

- Να βρεθεί η εντροπία των μηνυμάτων $\ell = 2$ με βάση το $\mathscr{L}$, δηλαδή $\mathscr{L}^2 = \{\sigma_1, \sigma_2, \ldots, \sigma_9\}$

$$H(\mathscr{L}^2) = \sum_{i=1}^{9} \mathbb{P}\{m_i\} \log_2\left(\frac{1}{\mathbb{P}\{m_i\}}\right) = \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{8}\log_2\left(\frac{1}{16}\right) + \frac{1}{8}\log_2\left(\frac{1}{16}\right) +$$

$$+ \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) \text{ bit} = 3 \text{ bit}$$

| Σύμβολα $\mathscr{L}^2$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ | $\sigma_8$ | $\sigma_9$ |
|---|---|---|---|---|---|---|---|---|---|
| Συνδυασμοί | $s_1s_1$ | $s_1s_2$ | $s_1s_3$ | $s_2s_1$ | $s_2s_2$ | $s_2s_3$ | $s_2s_1$ | $s_2s_2$ | $s_2s_3$ |
| Πιθανότητες | 1/16 | 1/16 | 1/8 | 1/16 | 1/16 | 1/8 | 1/8 | 1/8 | 1/4 |

- Άρα επιβεβαιώνεται ότι $H(\mathscr{L}^2) = 2\,H(\mathscr{L})$