Τμήμα Μηχανικών Πληροφορικής τ.ε.
Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας

Θεωρία Πληροφορίας

Διάλεξη 4: Διακριτή πηγή πληροφορίας χωρίς μνήμη

Δρ. Μιχάλης Παρασκευάς
Επίκουρος Καθηγητής

Ατζέντα

- Διακριτή πηγή πληροφορίας χωρίς μνήμη
- Ποσότητα πληροφορίας της πηγής
- Κωδικοποίηση πηγής
- Αλγόριθμοι κωδικοποίησης διακριτής πηγής πληροφορίας χωρίς μνήμη
  - Αλγόριθμος Fano
  - Αλγόριθμος Shannon
  - Αλγόριθμος Huffman

Διακριτή πηγή πληροφορίας χωρίς μνήμη

- **Διακριτή** ονομάζεται μια **πηγή πληροφορίας** που παράγει **ακολουθίες** συμβόλων.
- Το σύνολο των συμβόλων ονομάζεται **αλφάβητο πηγής**.
- Μια ομάδα διαδοχικών συμβόλων ονομάζεται **μήνυμα ή λέξη**.

Αν η πιθανότητα επιλογής ενός συμβόλου είναι **σταθερή** και **ανεξάρτητη** από τις επιλογές των προηγούμενων συμβόλων, τότε η πηγή ονομάζεται **διακριτή πηγή πληροφορίας χωρίς μνήμη**.

Ποσότητα πληροφορίας πηγής χωρίς μνήμη (1/3)

- Έστω $s_1, s_2, \dots, s_n$ το πλήθος των n συμβόλων του αλφάβητου S της πηγής.
- Τα μηνύματα συμβολίζονται με $m_1, m_2, \dots, m_q$, όπου q είναι το πλήθος των δυνατών μηνυμάτων. Το σύνολο όλων των μηνυμάτων συμβολίζεται με M.
- Αν κάθε μήνυμα αποτελείται από l σύμβολα, τότε το πλήθος των δυνατών μηνυμάτων είναι ίσο με n^l .

Η **μέση ποσότητα πληροφορίας** ή **εντροπία των συμβόλων** που δημιουργούνται από μια διακριτή πηγή χωρίς μνήμη με αλφάβητο $S = \{s_1, s_2, \dots s_n\}$, όπου p_i η (αμετάβλητη στο χρόνο) πιθανότητα επιλογής του συμβόλου s_i, είναι :

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i \qquad (bits/symbol)$$

Ποσότητα πληροφορίας πηγής χωρίς μνήμη (2/3)

Η **μέγιστη εντροπία** των συμβόλων της πηγής επιτυγχάνεται όταν οι πιθανότητες επιλογής τους είναι ίσες με:

$$maxH(S) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log\, n \qquad (bits/symbol\,)$$

Ο **πλεονασμός** της διακριτής πηγής ορίζεται από τη σχέση:

$$red = 1 - \frac{H(S)}{maxH(S)}$$

Ο πλεονασμός λαμβάνει τιμές στο διάστημα [0,1].

Αν η πηγή παράγει σύμβολα με ρυθμό r_s symbols/sec, τότε ο **μέσος ρυθμός πληροφορίας της πηγής** R ορίζεται από τη σχέση:

$$R = r_s\, H(S) \quad bits/\text{sec}$$

Ποσότητα πληροφορίας πηγής χωρίς μνήμη (3/3)

Το μέσο πληροφοριακό περιεχόμενο των μηνυμάτων της πηγής μπορεί να οριστεί με ανάλογο τρόπο.

- $M = \{m_1, m_2, \dots, m_q\}$ είναι το σύνολο των δυνατών μηνυμάτων
- q είναι το πλήθος των δυνατών μηνυμάτων ($q = n^l$)
- $P = \{p(m_1), p(m_2), .. p(m_q)\}$ είναι η κατανομή πιθανοτήτων των μηνυμάτων

Το **μέσο πληροφοριακό περιεχόμενο των μηνυμάτων της πηγής** είναι:

$$H(M) = -\sum_{i=1}^{q} p(m_i)\ \log p(m_i) \qquad (bits/symbol)$$

Για μηνύματα αποτελούμενα από q σύμβολα, ισχύει η σχέση $H(M) = q\ H(S)$.

Άσκηση 1

Μια δυαδική πηγή πληροφορίας χωρίς μνήμη παράγει τα σύμβολα 0 και 1 σε σταθερές ανεξάρτητες ακολουθίες με πιθανότητες 3/4 και 1/4, αντίστοιχα. Να υπολογιστούν η εντροπία, η μέγιστη μέση ποσότητα πληροφορίας και ο πλεονασμός της πηγής.

Απάντηση: Το αλφάβητο της πηγής είναι $S = \{0,1\}$ και οι αντίστοιχες πιθανότητες συμβόλων $p_1 = 3/4$ και $p_2 = 1/4$.

- Η εντροπία είναι:

$$H(S) = -\sum_{i=1}^{n} p_i \ log\, p_i = -\sum_{i=1}^{2} p_i \ log\, p_i = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) = 0{,}81\ bit/symbol$$

- Η μέγιστη μέση ποσότητα πληροφορίας είναι:

$$maxH(S) = -\sum_{i=1}^{n} \frac{1}{n} \log\frac{1}{n} = \log n = log\, 2 = 1\ bit/symbol$$

- Ο πλεονασμός της πηγής είναι:

$$red = 1 - \frac{H(S)}{maxH(S)} = 1 - \frac{0{,}81}{1} = 0{,}19$$

Άσκηση 2

Η πηγή της άσκησης 1 παράγει μηνύματα αποτελούμενα από δύο σύμβολα. Να υπολογιστεί το μέσο πληροφοριακό περιεχόμενο των μηνυμάτων της πηγής.

Απάντηση: Τα δυνατά μηνύματα είναι: $\mathrm{M} = \{00, 01, 10, 11\}$. Επειδή η πηγή είναι χωρίς μνήμη, η πιθανότητα παραγωγής καθενός από τα μηνύματα αυτά ισούται με το γινόμενο των πιθανοτήτων των συμβόλων από τα οποία αποτελείται. Επομένως ισχύει: $P = \{9/16, 3/16, 3/16, 1/16\}$.

Το μέσο πληροφοριακό περιεχόμενο των μηνυμάτων της πηγής δίνεται από τη σχέση:

$$H(M) = -\sum_{i=1}^{q} p(m_i)\ \log p(m_i) =$$

$$= -\frac{9}{16}\log\left(\frac{9}{16}\right) - \frac{3}{16}\log\left(\frac{3}{16}\right) - \frac{3}{16}\log\left(\frac{3}{16}\right) - \frac{1}{16}\log\left(\frac{1}{16}\right) =$$

$$= 1{,}63\ \ bits/message$$

Παρατηρούμε ότι ισχύει: $H(M) = q\ H(S)$

Άσκηση 3

Μια πηγή πληροφορίας παράγει σύμβολα, τα οποία ανήκουν στο αλφάβητο $S = \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta\}$. Οι πιθανότητες των συμβόλων είναι 1/32, 1/16, 1/8, 1/8, 1/8, 1/2 και 1/32, αντίστοιχα. Θεωρώντας την πηγή χωρίς μνήμη, ζητείται να προσδιορίσετε ή να υπολογίσετε:

1. Το σύμβολο της πηγής με το πιο χαμηλό πληροφορικό περιεχόμενο.
2. Τα σύμβολα της πηγής με το πιο υψηλό πληροφορικό περιεχόμενο.
3. Το μέσο πληροφορικό περιεχόμενο των συμβόλων της πηγής,
4. Το μέσο πληροφορικό περιεχόμενο των μηνυμάτων της πηγής αποτελούμενων από δύο σύμβολα.
5. Τον πλεονασμό της πηγής
6. Το μέσο ρυθμό πληροφορίας της πηγής για ρυθμό 12.500 συμβόλων/sec.

Απάντηση:

1. Το πληροφορικό περιεχόμενο ενός συμβόλου ‘x’ δίνεται από τον αρνητικό λογάριθμο της πιθανότητας παραγωγής του. Επομένως, το σύμβολο με την πιο υψηλή πιθανότητα παραγωγής έχει το πιο χαμηλό πληροφορικό περιεχόμενο. Στην προκειμένη περίπτωση, για το σύμβολο ‘ζ’ έχουμε $H(\zeta) = -\log(1/2) = 1\ bits$

Άσκηση 3 (συνέχεια)

2. Το σύμβολο με την πιο μικρή πιθανότητα παραγωγής έχει το πιο υψηλό πληροφορικό περιεχόμενο. Στην προκειμένη περίπτωση, τα σύμβολα ‘α’ και ‘η’ έχουν την πιο χαμηλή πιθανότητα παραγωγής, η οποία είναι ίση με 1/32, δηλαδή:

$$H(\alpha) = \mathrm{H}(\eta) = -\log(1/32) = 5\ bits$$

3. Το μέσο πληροφορικό περιεχόμενο των συμβόλων της πηγής υπολογίζεται από τη σχέση:

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i =$$

$$= -\frac{1}{32}\log\frac{1}{32} - \frac{1}{16}\log\frac{1}{16} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{2}\log\frac{1}{2} - \frac{1}{32}\log\frac{1}{32} =$$

$$= \frac{35}{16} = 2{,}1875\ (bits/symbol)$$

Άσκηση 3 (συνέχεια)

4. Για τον υπολογισμό του μέσου πληροφορικού περιεχομένου των μηνυμάτων της πηγής αποτελούμενων από 2 σύμβολα, αφού η πηγή είναι χωρίς μνήμη, αρκεί να πολλαπλασιάσουμε τη μέση ποσότητα πληροφορίας συμβόλων με το πλήθος των συμβόλων από τα οποία αποτελούνται τα μηνύματα. Δηλαδή:

$$H(M) = 2\ \mathrm{H}(S) = 4{,}375\ bits$$

5. Ο πλεονασμός δίνεται από τη σχέση:

$$red = 1 - \frac{H(S)}{maxH(S)} = 1 - \frac{H(S)}{\log 7} = 1 - \left(\frac{2{,}1875}{2{,}8}\right) = 1 - 0{,}781 = 0{,}2186$$

6. Η παροχή της πηγής δίνεται από τη σχέση:

$$R = r\ H(S) = 12.500\ x\ (2{,}1875) = 27.343{,}75\ bits/\mathrm{sec}$$

Κωδικοποίηση Πηγής χωρίς Μνήμη

- **Κωδικοποίηση πηγής** είναι η διαδικασία μετατροπής των ακολουθιών συμβόλων που παράγει η πηγή σε **ακολουθίες συμβόλων κώδικα** (συνήθως σε δυαδικές ακολουθίες), ώστε να αφαιρείται ο πλεονασμός και να προκύπτει συμπιεσμένη μορφή αναπαράστασης των μηνυμάτων.
- Τα διαφορετικά κωδικά σύμβολα που χρησιμοποιούνται για τη μετατροπή των δυαδικών ακολουθιών απαρτίζουν το **κωδικό αλφάβητο**.
- **Κώδικας** είναι το σύνολο των κωδικών λέξεων και η αντιστοίχιση τους με τα σύμβολα της πηγής.
- Αν όλες οι κωδικές λέξεις είναι διαφορετικές, ο κώδικας ονομάζεται ως **μη ιδιάζων**.
- Αν και οι δυνατές ακολουθίες κωδικών λέξεων είναι διαφορετικές, ο κώδικας είναι **μοναδικά αποκωδικοποιήσιμος**.
- Αν ένας μοναδικός αποκωδικοποιήσιμος κώδικας επιτρέπει την άμεση αποκωδικοποίηση κάθε συμβόλου μόλις λαμβάνεται στον προορισμό, τότε χαρακτηρίζεται ως **άμεσος κώδικας**.

Κωδικοποίηση Πηγής χωρίς Μνήμη

Για κάθε άμεσο κώδικα με πλήθος συμβόλων q και μήκος κωδικών λέξεων l_i, όπου $i = 1,2, \dots n$ και n το πλήθος των συμβόλων της πηγής, ισχύει η ακόλουθη ανισότητα (ανισότητα του Kraft):

$$\sum_{i=1}^{n} q^{-l_i} = 1$$

Εάν υπάρχει ένα σύνολο κωδικών λέξεων που ικανοποιούν την ανισότητα, τότε υπάρχει ένας άμεσος κώδικας με κωδικές λέξεις που έχουν αυτά τα μήκη.

Το **μέσο μήκος** κωδικών λέξεων δίνεται από τη σχέση $L = \sum_{i=1}^{n} p_i \, l_i$

Η **επίδοση (α) του κώδικα** ορίζεται ως ο λόγος του μέσου πληροφοριακού 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, δηλ.:

$$\alpha = \frac{\mathrm{H(C)}}{(\sum_{i=1}^{n} p_i \, l_i) \log q}$$

Άσκηση 4

Να εξετάσετε αν οι κώδικες
I, II, III και IV είναι:

(α) Μη-ιδιάζοντες

(β) Μοναδικά αποκωδικοποιήσιμοι

(γ) Άμεσοι

	I	II	III	IV
Φ	0	00	1	1
Χ	10	01	10	01
Υ	01	10	100	001
Ω	1	11	1000	0001

Απάντηση: (α) Όλοι οι κώδικες είναι μη ιδιάζοντες, αφού ο καθένας αποτελείται από διαφορετικές κωδικές λέξεις.

(β) Όλοι οι κώδικες είναι μοναδικά αποκωδικοποιήσιμοι εκτός του I. Για τον κώδικα I παρατηρούμε ότι η ακολουθία κωδικών λέξεων 1001 θα μπορούσε να προκύψει από διάφορες ακολουθίες συμβόλων όπως 'ΧΥ' ή 'ΧΦΩ' κλπ.

(γ) Μόνο οι κώδικες II και IV είναι άμεσοι. Στην περίπτωση του κώδικα I, αν ο δέκτης λάβει το '0' δεν θα ξέρει αν είναι η πρώτη κωδική λέξη ή το 1ο κωδικό σύμβολο της 3^{ης} κωδικής λέξης κοκ. Σχετικά με τον κώδικα III, όταν ο δέκτης λάβει τα '10' δεν μπορεί ξέρει αν είναι η 2η κωδική λέξη ή τα 2 πρώτα σύμβολα της 3ης κωδικής λέξης κοκ.

Άσκηση 5

Για την πηγή της άσκησης 4 να υπολογιστεί η επίδοση του κώδικα ΙΙ αν οι πιθανότητες των συμβόλων της πηγής Φ, Χ, Υ και Ω είναι 1/2, 1/4, 1/8 και 1/8, αντίστοιχα. Για την ίδια πηγή να υπολογιστεί επίσης και η επίδοση του κώδικα «1», «01», «001» και «000».

Απάντηση: Το μέσο πληροφοριακό περιεχόμενο της πηγής είναι:

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = 1{,}75\ bits$$

Η επίδοση του κώδικα είναι:

$$\alpha = \frac{\mathrm{H(C)}}{(\sum_{i=1}^{n} p_i\ l_i)\log q} = \frac{1{,}75}{2} = 0{,}875$$

Για τις κωδικές λέξεις «1», «01», «001» και «000» το μέσο πληροφοριακό περιεχόμενο της πηγής $H(S)$ είναι επίσης ίσο με 1,75 bits. Το πλήθος των κωδικών συμβόλων είναι 2 και το μέσο μήκος των κωδικών λέξεων είναι ίσο με 1,75. Επομένως η επίδοση α είναι ίση με 1.

Αλγόριθμοι κωδικοποίησης διακριτής πηγής πληροφορίας χωρίς μνήμη

- Αλγόριθμος Fano
- Αλγόριθμος Shannon
- Αλγόριθμος Huffman

Αλγόριθμος Κωδικοποίησης Fano

<u>Βήματα Υλοποίησης</u>:

1. Τα σύμβολα της πηγής διατάσσονται σε κατά **φθίνουσα τάξη** με βάση την πιθανότητα εμφάνισης.
2. Τα σύμβολα χωρίζονται σε τόσες ομάδες όσα και τα κωδικά σύμβολα, κατά τρόπο ώστε να προκύπτουν (αν αυτό είναι εφικτό) **ίσες αθροιστικές πιθανότητες** εμφάνισης των συμβόλων. Στην περίπτωση δυαδικού κώδικα, τα n σύμβολα της πηγής χωρίζονται σε 2 ομάδες, επιλέγοντας το k έτσι ώστε η διαφορά $\left|\sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i\right|$ των αθροιστικών πιθανοτήτων εμφάνισης των συμβόλων, να **ελαχιστοποιείται**.
3. Για κάθε ομάδα συμβόλων της πηγής, επιλέγεται ένα από τα κωδικά σύμβολα ως το πρώτο των κωδικών λέξεων που θα προκύψουν.
4. Για κάθε ομάδα συμβόλων της πηγής επαναλαμβάνονται τα βήματα 2 και 3 έως ότου η κάθε ομάδα αποτελείται από μόνο ένα σύμβολο. Σε κάθε επανάληψη του βήματος 3, επιλέγεται ένα ακόμα κωδικό σύμβολο για το σχηματισμό των κωδικών λέξεων.

Ο αλγόριθμος Fano οδηγεί σε **άριστους κώδικες** αν είναι δυνατή η επαναλαμβανόμενη διαίρεση των ομάδων συμβόλων σε ακριβώς ισοπίθανες ομάδες.

Άσκηση 6

Μια πηγή παράγει 8 διαφορετικά σύμβολα, τα Α, Β, Γ, Δ, Ε, Ζ, Η και Θ, με πιθανότητες 1/8, 1/4, 1/16, 1/32, 1/4, 1/32, 1/8 και 1/8, αντίστοιχα. Να σχηματιστεί κώδικας σύμφωνα με τον αλγόριθμο του Fano, με δυαδικό κωδικό αλφάβητο.

Απάντηση: Τα σύμβολα της πηγής διατάσσονται σε φθίνουσα πιθανότητα και χωρίζονται σε δύο ομάδες με το δυνατόν ίσες αθροιστικές πιθανότητες. Τα δύο πρώτα σύμβολα ανήκουν στην 1η ομάδα και τα υπόλοιπα στην 2η. Επιλέγουμε το '0' ως το πρώτο κωδικό σύμβολο των κωδικών λέξεων της 1ης ομάδας και το '1' για τις κωδικές λέξεις της 2ης ομάδας. Η πρώτη ομάδα χωρίζεται σε 2 υποομάδες με ένα σύμβολο η καθεμία. Επιλέγουμε και πάλι το '0' για την 1η υποομάδα και το '1' για τη 2. Έτσι καταλήγουμε στις κωδικές λέξεις των δύο πρώτων συμβόλων του πίνακα, τις '00' και '01'. Κατά τον ίδιο τρόπο συνεχίζουμε και σε σχέση με τη δεύτερη ομάδα, την οποία χωρίζουμε σε δύο υποομάδες, εκ των οποίων η 1η περιλαμβάνει το 3ο και το 4ο σύμβολο του πίνακα και η άλλη όλα τα υπόλοιπα σύμβολα.

Σύμβολα	Πιθανότητες	Κώδικας
$B=S_1$	1/4	00 (11)
$E=S_2$	1/4	01 (10)
$A=S_3$	1/8	100 (011)
$H=S_4$	1/8	101 (010)
$\Theta=S_5$	1/8	110 (001)
$\Gamma=S_6$	1/16	1110 (0001)
$\Delta=S_7$	1/32	11110 (00001)
$Z=S_8$	1/32	11111 (00000)

Αλγόριθμος Κωδικοποίησης Shannon

<u>Βήματα Υλοποίησης</u>:

1. Τα σύμβολα της πηγής διατάσσονται σε κατά φθίνουσα τάξη με βάση την πιθανότητα εμφάνισης.
2. Για κάθε σύμβολο S_j, του οποίου η πιθανότητα εμφάνισης είναι $p(S_j)$, υπολογίζεται η αθροιστική πιθανότητα P_i, από τη σχέση:

$$P_i = \sum_{j=1}^{i-1} p(S_j)$$

3. Το πλήθος των κωδικών συμβόλων της κωδικής λέξης που αναπαριστά το σύμβολο της πηγής S_i είναι ίσο με τον ακέραιο αριθμό l_i, που ικανοποιεί την ανισότητα:

$$\log\frac{1}{p(S_i)} \leq l_i < 1 + \log\frac{1}{p(S_i)}$$

4. Η κωδική λέξη c_i του συμβόλου S_i της πηγής είναι το δυαδικό ανάπτυγμα του κλάσματος P_i (μόνο τα πρώτα l_i bits του αναπτύγματος λαμβάνονται υπόψη), δηλαδή: $c_i = (P_i)_{binary\ l_i\ bits}$

Άσκηση 7

Για την πηγή της άσκησης 6 να σχηματιστεί κώδικας σύμφωνα με τον αλγόριθμο του Shannon, με δυαδικό κωδικό αλφάβητο.

Απάντηση:

Σύμβολα Πηγής	Πιθανότητες Συμβόλων	P_i	Μήκος l_i	Ανάπτυγμα του P_i	Κωδικές Λέξεις
$B=S_1$	1/4	$P_1 = 0$	$l_1 = 2$	.00000	00
$E=S_2$	1/4	$P_2 = 1/4$	$l_2 = 2$	.01000	01
$A=S_3$	1/8	$P_3 = 1/2$	$l_3 = 3$	.10000	100
$H=S_4$	1/8	$P_4 = 5/8$	$l_4 = 3$	.10100	101
$\Theta=S_5$	1/8	$P_5 = 6/8$	$l_5 = 3$	.11000	110
$\Gamma=S_6$	1/16	$P_6 = 7/8$	$l_6 = 4$	.11100	1110
$\Delta=S_7$	1/32	$P_7 = 15/16$	$l_7 = 5$	.11110	11110
$Z=S_8$	1/32	$P_8 = 31/32$	$l_8 = 5$	.11111	11111

Παρατηρούμε ότι οι αλγόριθμοι Shannon και Fano παρήγαγαν το ίδιο αποτέλεσμα. Αυτό όμως δεν ισχύει πάντα.

Άσκηση 8

Δίνεται μία πηγή με τα σύμβολα Φ, Χ, Υ και Ω και πιθανότητες παραγωγής τους 0,4, 0,3, 0,2 και 0,1, αντίστοιχα. Να σχηματιστούν κωδικές λέξεις με τους αλγορίθμους Shannon και Fano.

Απάντηση:

S_i	$P(S_i)$	P_i	l_i	Ανάπτυγμα του P_i	Κώδικας Shannon	Κώδικας Fano
Φ	0,4	$P_1 = 0$	$l_1 = 2$	.0000	00	0
Χ	0,3	$P_2 = 0{,}4$	$l_2 = 2$	.01100..	01	10
Υ	0,2	$P_3 = 0{,}7$	$l_3 = 3$	.10110..	101	110
Ω	0,1	$P_4 = 0{,}9$	$l_4 = 4$	.11100..	1110	111

Αλγόριθμος Κωδικοποίησης Huffman

Βήματα Υλοποίησης κωδικοποίησης HUFFMAN:

1. Τα σύμβολα της πηγής διατάσσονται κατά φθίνουσα πιθανότητα εκπομπής
2. Τα δύο τελευταία σύμβολα της πηγής με μικρότερη πιθανότητα παραγωγής ενώνονται σε ένα. Η πιθανότητα του συμβόλου είναι ίση με το άθροισμα των πιθανοτήτων των δύο συμβόλων.
3. Τα βήματα 1 και 2 επαναλαμβάνονται έως ότου το αλφάβητο της πηγής αποτελείται από δύο σύμβολα. Σε αυτά τα σύμβολα αποδίδονται το 0 και 1.
4. Τα ψηφία '0' και '1' αποδίδονται στη θέση του ενός και του άλλου συμβόλου αντίστοιχα, τα οποία στο βήμα 2 συγχωνεύτηκαν σε ένα.
5. Οι κωδικές λέξεις των συμβόλων σχηματίζονται από όλα τα ψηφία '0' και '1' που σχετίζονται με αυτά τα σύμβολα (από το τέλος προς την αρχή).

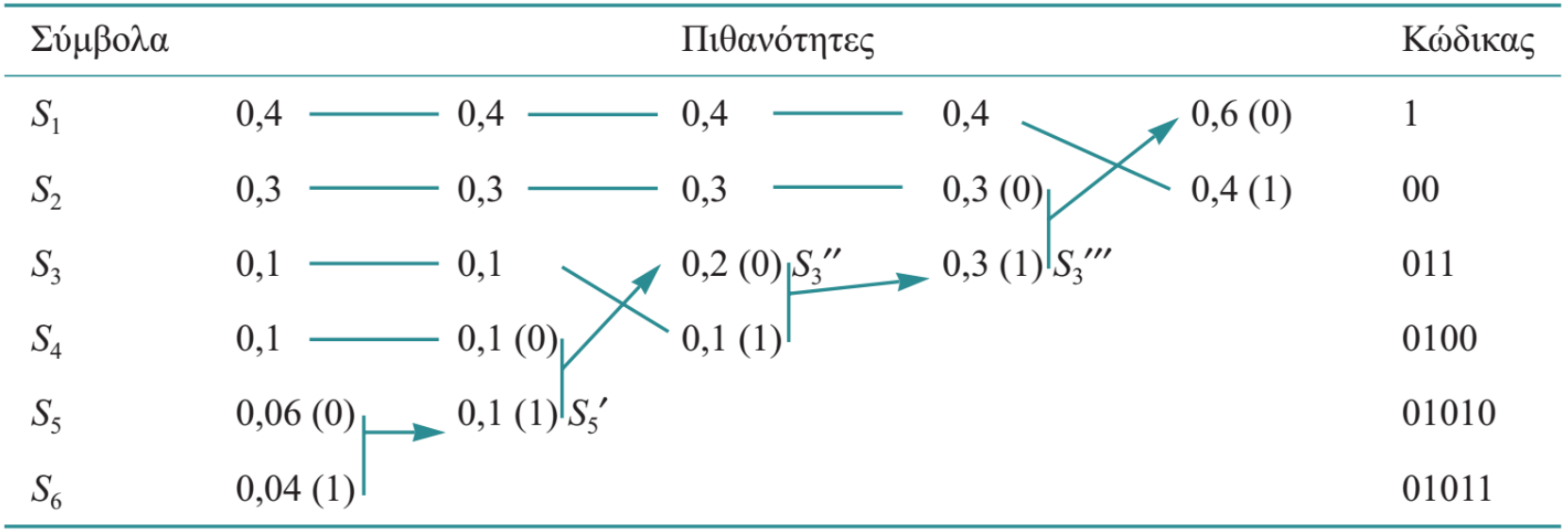
Ο κώδικας Huffman παράγει κώδικα με το μικρότερο μέσο μήκος λέξεων για δεδομένο αλφάβητο της πηγής.

Πηγή: D. Huffman, *A Method for the Construction of Minimum-Redundancy Codes*

Άσκηση 9

Μια πηγή παράγει 6 σύμβολα, τα S_1, S_2, S_3, S_4, S_5 και S_6, με πιθανότητες 0,4, 0,3, 0,1, 0,1, 0,06 και 0,04, αντίστοιχα. Να σχηματιστεί κώδικας σύμφωνα με τον αλγόριθμο του Huffman, με δυαδικό κωδικό αλφάβητο.

<u>Απάντηση</u>:



Άσκηση 10

Δίδεται διακριτή πηγή που παράγει 7 διαφορετικά σύμβολα, *Α, Β, Γ, Δ, Ε, Ζ, Η* , με πιθανότητες, αντίστοιχα: {0,2, 0,15, 0,1, 0,3, 0,06, 0,15, 0,04} και ζητείται:

1. Να σχεδιασθεί δυαδικός κώδικας σύμφωνα με τον αλγόριθμο Huffman.
2. Να σχεδιασθεί δυαδικός κώδικας σύμφωνα με τον αλγόριθμο Fano.
3. Να συγκριθούν οι κώδικες που προκύπτουν στα ερωτήματα 1 και 2 ως προς την επίδοσή τους.

<u>Απάντηση</u>:

1. **Κώδικας Huffman** με δύο κωδικά σύμβολα:

Σύμβολα							**Κώδικας**
Δ	0,3	0,3	0,3	0,3	**0,4**	**0,6** (0)	**00**
Α	0,2	0,2	0,2	**0,3**	0,3 (0)	0,4 (1)	**10**
Β	0,15	0,15	**0,2**	0,2 (0)	0,3 (1)		**010**
Ζ	0,15	0,15	0,15 (0)	0,2 (1)			**011**
Γ	0,1	0,1 (0)	0,15 (1)				**110**
Ε	0,06 (0)	**0,1** (1)					**1110**
Η	0,04 (1)						**1111**

Άσκηση 10 (συνέχεια)

2. Τα σύμβολα της πηγής κατατάσσονται σε τάξη φθίνουσας πιθανότητας (δείτε τον παρακάτω πίνακα). Χωρίζονται σε ομάδες και υποομάδες ως ακολούθως:

Τα δύο πρώτα σύμβολα περιλαμβάνονται στην 1η ομάδα και τα υπόλοιπα στη $2^{η}$ ομάδα. Επιλέγουμε το '0' ως το πρώτο κωδικό σύμβολο των κωδικών λέξεων της $1^{ης}$ ομάδας και το '1' για τις κωδικές λέξεις της $2^{ης}$ ομάδας. Η πρώτη ομάδα χωρίζεται σε 2 υποομάδες με ένα σύμβολο η πρώτη και ένα η δεύτερη. Επιλέγουμε και πάλι το '0' για την $1^{η}$ υποομάδα και το '1' για τη 2. Έτσι καταλήγουμε στην κωδική λέξη του Γ, η οποία είναι η '00' κοκ.

Κώδικας Fano

Σύμβολα	Πιθανότητες	Κώδικας
Δ	0,3	**00**
Α	0,2	**01**
Β	0,15	**100**
Ζ	0,15	**101**
Γ	0,1	**110**
Ε	0,06 (0)	**1110**
Η	0,04 (1)	**1111**

Άσκηση 10 (συνέχεια)

3. Για τον υπολογισμό της απόδοσης των κωδίκων, υπολογίζουμε την εντροπία της πηγής:

$$H(S) = \sum_{i=1}^{7} p_i \, log \, p_i =$$

$$= \frac{3}{10}\log\frac{3}{10} + 2\frac{15}{100}\log\frac{15}{100} + \frac{2}{10}\log\frac{2}{10} + \frac{1}{10}\log\frac{1}{10} + \frac{6}{100}\log\frac{6}{100} + \frac{4}{100} \; log \frac{4}{100}$$

$$= 2{,}568 \; bits/symbol$$

Υπολογίζουμε το μέσο μήκος των κωδικών λέξεων για κάθε περίπτωση.

Δυαδικός κώδικας Huffman $\sum_{i=1}^{7} p_i \; l_i = 2x0{,}5 + 3x0{,}4 + 4x0{,}1 = 2{,}6$

Η απόδοση είναι : $a = \frac{H(s)}{(\sum_{i=1}^{7} p_i l_i)\log 2} = \frac{2{,}568}{2{,}6} = 0{,}98771$

Δυαδικός κώδικας Fano $\sum_{i=1}^{7} p_i \; l_i = 2x0{,}5 + 3x0{,}4 + 4x0{,}1 = 2{,}6$

Η απόδοση είναι: $a = \frac{H(s)}{(\sum_{i=1}^{7} p_i l_i)\log 2} = \frac{2{,}568}{2{,}6} = 0{,}98771$

Παρατηρούμε ότι οι δυαδικοί κώδικες είναι σχεδόν άριστοι.

Άσκηση 11

1. Να βρείτε τον κώδικα Shannon για την κατανομή πιθανοτήτων τεσσάρων συμβόλων {1/3, 1/3, 1/4, 1/12} καθώς και το μέσο μήκος του.
2. Να βρείτε **όλα** τα δυνατά μήκη κωδικών λέξεων που μπορούν να προκύψουν με εφαρμογή του αλγόριθμου κωδικοποίησης Huffman για την ίδια κατανομή πιθανοτήτων του ερωτήματος (1). Τί παρατηρείτε σχετικά με τα μήκη των κωδικών λέξεων που αντιστοιχούν σε κάθε σύμβολο σε σχέση με τον κώδικα Shannon; Ποιός κώδικας είναι βέλτιστος;

Απάντηση: 1. Εφαρμόζοντας τον αλγόριθμο κωδικοποίησης Shannon και παρατηρώντας ότι οι πιθανότητες είναι ήδη ταξινομημένες σε φθίνουσα ακολουθία έχουμε τον παρακάτω πίνακα:

p_i	P_i	l_i	Κώδικας
1/3	0	2	00
1/3	1/3	2	01
1/4	2/3	2	10
1/12	11/12	4	1110

Το μέσο μήκος του κώδικα Shannon είναι $E[l]=2,1667$ bits

Άσκηση 11 (συνέχεια)

Παρατηρούμε ότι ο κώδικας Shannon που κατασκευάσαμε δεν είναι βέλτιστος, δεδομένου ότι μπορούμε να συντομεύσουμε την τελευταία κωδική λέξη σε 11 χωρίς να χαθεί η αμεσότητα του κώδικα.

Αποδεικνύεται ότι ο κώδικας {00, 01, 10, 11} είναι ένας από τους πιθανούς κώδικες Huffman, ο οποίος προκύπτει συνδυάζοντας πρώτα το 3ο και το 4ο ενδεχόμενο σε ένα ενδεχόμενο πιθανότητας 1/3 και στη συνέχεια συνδυάζοντας το 1ο και 2ο ενδεχόμενο σε ένα ενδεχόμενο πιθανότητας 2/3. Το μέσο μήκος του κώδικα είναι $E[l_1]=2$ bits.

Όπως γνωρίζουμε, ο βέλτιστος κώδικας που προκύπτει από Huffman δεν είναι μοναδικός. Έτσι εάν το ενδεχόμενο που προκύπτει από το συνδυασμό του 3ου και 4ου ενδεχομένου συνδυαστεί στη συνέχεια με το 1ο ή το 2ο ενδεχόμενο, προκύπτει κώδικας Huffman με μήκη (1, 2, 3, 3).

Το μέσο μήκος και σε αυτή την περίπτωση είναι $E[l_1]=1x(1/3) + 2x(1/3) + 3x(1/4+1/12) = 2$ bits, που είναι αναμενόμενο λόγω του ότι έχει προκύψει από χρήση του αλγορίθμου Huffman.

Παρατηρούμε ότι τα μήκη του κώδικα Huffman δεν είναι πάντα μοναδικά.

Άσκηση 11 (συνέχεια)

Επίσης, είναι δυνατό επιμέρους μήκη που έχουν προκύψει από τον κώδικα Shannon να είναι μικρότερα από αντίστοιχα επιμέρους μήκη που έχουν προκύψει από τον κώδικα Huffman.

Π.χ. το μήκος του κώδικα Shannon που αντιστοιχεί στο $3^{ο}$ ενδεχόμενο είναι μικρότερο από το αντίστοιχο μήκος του δεύτερου κώδικα Huffman. Ωστόσο το μέσο μήκος του κώδικα Huffman δεν υπερβαίνει σε καμία περίπτωση το μέσο μήκος του κώδικα Shannon.

Τέλος, η εντροπία της τυχαίας μεταβλητής ισούται με 1,7637 bits.

Παρατηρούμε ότι τόσο ο κώδικας Huffman όσο και ο κώδικας Shannon επιτυγχάνουν συμπίεση το πολύ 1 bit μακριά από την εντροπία.