Συμπίεση χωρίς Απώλειες

- Στόχοι της συμπίεσης δεδομένων:
  - Μείωση του απαιτούμενου χώρου αποθήκευσης των δεδομένων.
  - Περιορισμός της απαιτούμενης χωρητικότητας διαύλου επικοινωνίας για την μετάδοση.
    - μείωση του χρόνου αποστολής.
- Από τα συμπιεσμένα δεδομένα θα πρέπει να μπορούν να ανακτηθούν **πλήρως** οι αρχικές πληροφορίες
  - Ο κώδικας συμπίεσης θα πρέπει να είναι γνωστός στον πομπό και δέκτη.
  - Ο κώδικας συμπίεσης θα πρέπει να συνοδεύει τα συμπιεσμένα δεδομένα.
- Θα μας απασχολήσουν κώδικες **χωρίς απώλειες.**
- Μερικές από τις **ιδιότητες του βέλτιστου κώδικα συμπίεσης** είναι:
  - Να είναι στιγμιαία αποκωδικοποιήσιμος (άρα και μονοσήμαντος).
  - Σύμβολα του αλφαβήτου της πηγής με μεγαλύτερη πιθανότητα θα πρέπει να αντιστοιχίζονται σε κωδικές λέξεις μικρότερου μήκους.
  - Κατέχει το μικρότερο δυνατό μέσο μήκος.

Κώδικας Huffman

- Ανακαλύφθηκε από τον David Huffman το 1950.
- Αποτελεί σήμερα το βασικότερο τμήμα των περισσοτέρων αλγορίθμων συμπίεσης:
  - π.χ., **GZIP, JPEG**
- Είναι κώδικας χωρίς απώλειες.
- Η βασική ιδιότητα του είναι ότι αποτελεί τον βέλτιστο κώδικα:
  - **δεν υπάρχει κώδικας χωρίς απώλειες με μικρότερο μέσο μήκος.**

Κώδικας Huffman

• Ο αλγόριθμος Huffman (για την περίπτωση δυαδικού κώδικα) οδηγεί στην κατασκευή ενός δεδνροδιαγράμματος και περιγράφεται παρακάτω:

Βήμα 1: Θεωρούμε ότι κάθε σύμβολο πηγής είναι ένα φύλλο του δενδροδιαγράμματος. Στο κάθε φύλλο αντιστοιχίζουμε την πιθανότητα εμφάνισης του κάθε συμβόλου.

Βήμα 2: Εντοπίζουμε τους δύο κόμβους με τις μικρότερες πιθανότητες και τους συγχωνεύουμε σε έναν. Ο νέος κόμβος θα έχει το άθροισμα των πιθανοτήτων των δύο κόμβων που συγχωνεύτηκαν.

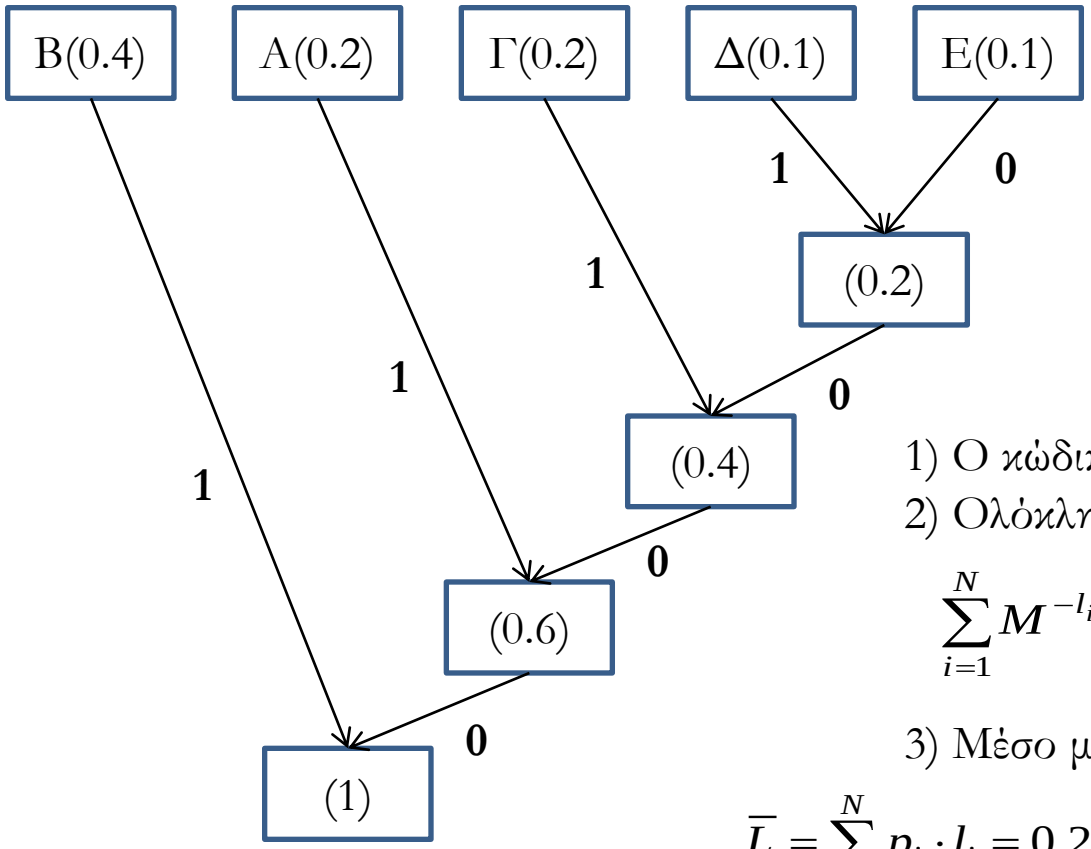
Βήμα 3: Το δεύτερο βήμα επαναλαμβάνεται μέχρι την δημιουργία του τελικού κόμβου (ρίζα).

Σε κάθε ακμή αντιστοιχίζουμε ένα κωδικό σύμβολο 0 ή 1 τυχαία.

Παράδειγμα

Έστω πηγή πληροφορίας με αλφάβητο {Α,Β,Γ,Δ,Ε} και κατανομή {0.2,0.4,0.2,0.1,0.1}. Κωδικοποίηση Huffman (1η Υλοποίηση):

Δενδροδιάγραμμα



Κώδικας

Α	**01**
Β	**1**
Γ	**001**
Δ	**0001**
Ε	**0000**

1) Ο κώδικας είναι στιγμιαία αποκωδικοποιήσιμος.

2) Ολόκληρος (complete):

$$\sum_{i=1}^{N} M^{-l_i} = \frac{1}{2^2} + \frac{1}{2^1} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^4} = 1$$

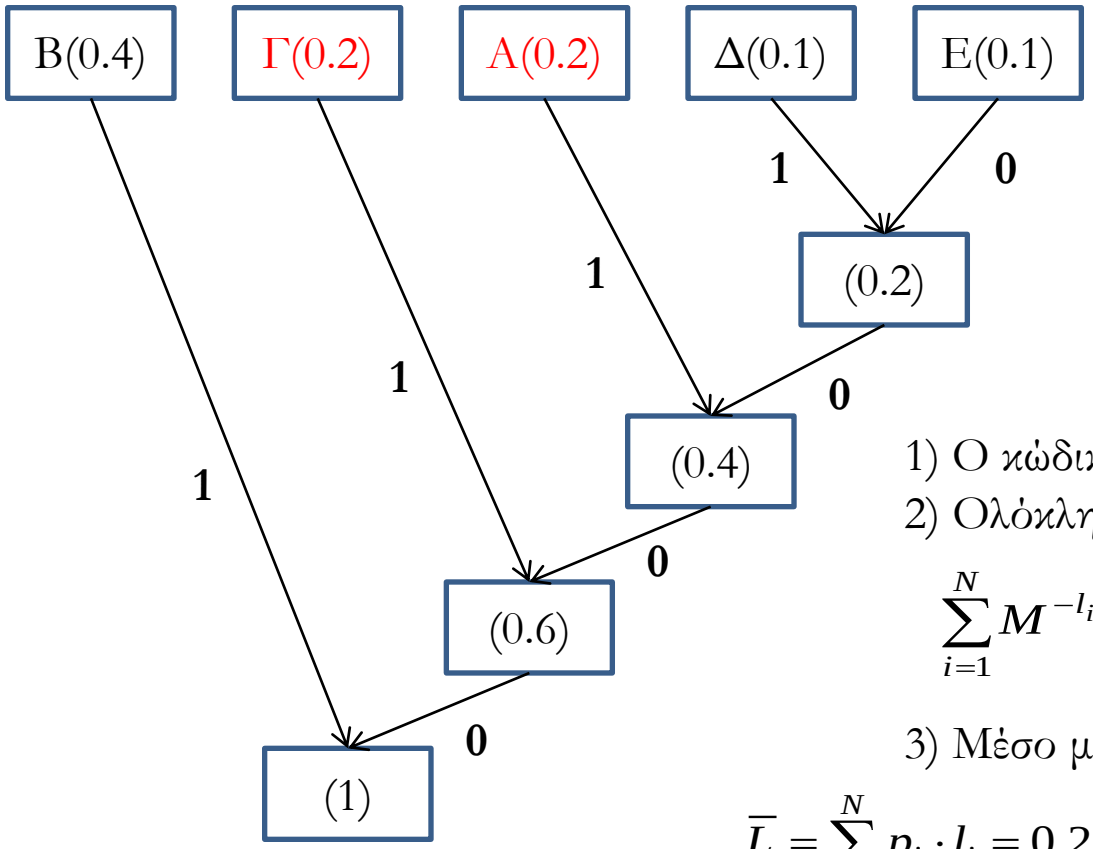
3) Μέσο μήκος:

$$\overline{L} = \sum_{i=1}^{N} p_i \cdot l_i = 0.2 \cdot 2 + 0.4 \cdot 1 + 0.2 \cdot 3 + 0.1 \cdot 4 + 0.1 \cdot 4 = 2.2$$

Παράδειγμα

Εναλλακτική κωδικοποίηση:

Δενδροδιάγραμμα



Κώδικας

A	**001**
B	**1**
Γ	**01**
Δ	**0001**
E	**0000**

1) Ο κώδικας είναι στιγμιαία αποκωδικοποιήσιμος.

2) Ολόκληρος (complete):

$$\sum_{i=1}^{N} M^{-l_i} = \frac{1}{2^3} + \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^4} = 1$$

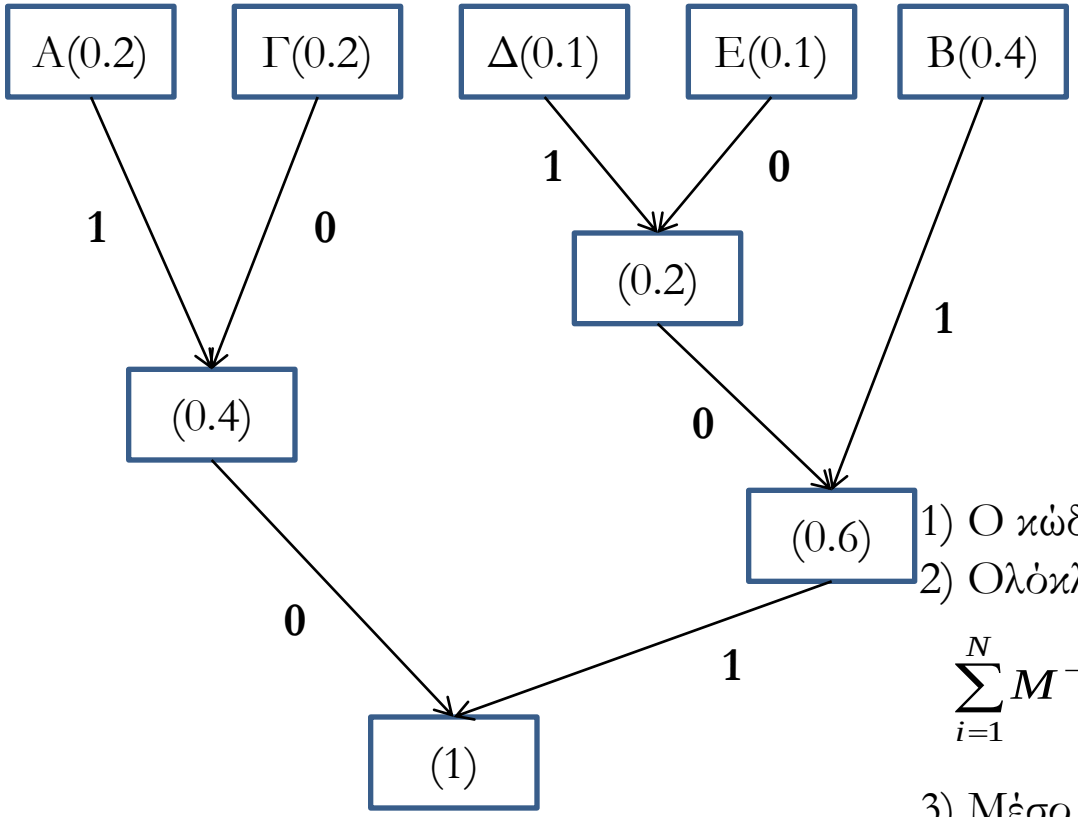
3) Μέσο μήκος:

$$\overline{L} = \sum_{i=1}^{N} p_i \cdot l_i = 0.2 \cdot 3 + 0.4 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 4 + 0.1 \cdot 4 = 2.2$$

Παράδειγμα

Τρίτη υλοποίηση:

Δενδροδιάγραμμα



Κώδικας

A	**01**
B	**11**
Γ	**00**
Δ	**101**
E	**100**

1) Ο κώδικας είναι στιγμιαία αποκωδικοποιήσιμος.

2) Ολόκληρος (complete):

$$\sum_{i=1}^{N} M^{-l_i} = \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} = 1$$

3) Μέσο μήκος:

$$\bar{L} = \sum_{i=1}^{N} p_i \cdot l_i = 0.2 \cdot 2 + 0.4 \cdot 2 + 0.2 \cdot 2 + 0.1 \cdot 3 + 0.1 \cdot 3 = 2.2$$

Παράδειγμα

Σύγκριση των τριών υλοποιήσεων:

		$1^η$		$2^η$		$3^η$	
x_i	p_i	c_i	l_i	c_i	l_i	c_i	l_i
A	0.2	10	2	111	3	01	2
B	0.4	0	1	0	1	11	2
Γ	0.2	111	3	10	2	00	2
Δ	0.1	1101	4	1101	4	101	3
E	0.1	1100	4	1100	4	100	3
$\bar{L}$		2.2		2.2		2.2	
σ^2		1.36		1.36		0.16	

$$\sigma^2 = \sum_{i=1}^{N} p_i \cdot (l_i - \bar{L})^2.$$

Ασκήσεις

Για την υλοποίηση με την μικρότερη μεταβλητότητα θα πρέπει σε κάθε βήμα του αλγορίθμου όπου προκύπτουν περισσότρες από μια επιλογές για τους κόμβους μικρότερης πιθανότητας, να επιλέγουμε για συγχώνευση τους δύο κόμβους που έχουν δημιουργηθεί παλαιότερα (ψηλότερα στο δένδρο).

- **Εφαρμογή Ι:** Έστω μια πηγή Χ που εκπέμπει 6 σύμβολα με πιθανότητες:

Α	**0.4**
Β	**0.25**
Γ	**0.1**
Δ	**0.1**
Ε	**0.1**
Ζ	**0.05**

Α) Να κατασκευάσετε μια υλοποίηση του δενδροδιαγράμματος κατά Huffman και να γράψετε τις κωδικές λέξεις για κάθε σύμβολο της πηγής.

Β) Να βρείτε το μέσο μήκος του κώδικα.

Ασκήσεις

Εφαρμογή ΙΙ: Έστω μια πηγή Χ που εκπέμπει 6 σύμβολα με πιθανότητες:

Α	**0.3**
Β	**0.15**
Γ	**0.15**
Δ	**0.15**
Ε	**0.15**
Ζ	**0.1**

Α) Να κατασκευάσετε μια υλοποίηση του δενδροδιαγράμματος κατά Huffman και να γράψετε τις κωδικές λέξεις για κάθε σύμβολο της πηγής.

Β) Να υπολογίσετε το μέσο μήκος του κώδικα.

Ασκήσεις

Εφαρμογή ΙΙΙ: (Εξεταστική 2009) Δίνεται πηγή πληροφορίας με σύμβολα πηγής και αντίστοιχες πιθανότητες X={A,B,Γ,Δ,E,Z,H}, P={0.15, 0.25, 0.05, 0.1, 0.2, 0.15, 0.1}

Α) Να κατασκευάσετε το δενδροδιάγραμμα μετά την εφαρμογή του αλγορίθμου Huffman ελάχιστης μεταβλητότητας και γράψτε τις κωδικές λέξεις για κάθε σύμβολο του X.

Β) Να υπολογίσετε το μέσο μήκος του κώδικα.

Γ) Την μεταβλητότητα του.

Στο Γ) γράψτε τον τύπο και αντικαταστήστε (μετά μην κάνετε τις πράξεις).