Δ2

Συμπίεση Δεδομένων

2014-2015

Ρυθμός κωδικοποίησης

Δ2

- Ένας κώδικας που απαιτεί L bits για την κωδικοποίηση μίας συμβολοσειράς N συμβόλων που εκπέμπει μία πηγή έχει ρυθμό κωδικοποίησης (μέσο μήκος λέξης)

$$\bar{R} = \lim_{N \to \infty} \frac{L}{N}$$

- Ποιος είναι ο ελάχιστος ρυθμός (ελάχιστο μέσο μήκος λέξης);

Θεώρημα Shannon (Κωδικοποίηση Πηγής DMS)

Δ2

- Μία πηγή εντροπίας H μπορεί να κωδικοποιηθεί με αυθαίρετα μικρό σφάλμα ($\varepsilon \rightarrow 0$) σε οποιοδήποτε ρυθμό $\overline{\mathrm{R}}$>H.
- Αντίστροφα, δεν υπάρχει κώδικαςρυθμό $\overline{\mathrm{R}}$ <H που να κωδικοποιεί μία πηγή εντροπίας H χωρίς σφάλματα.

Πλεονασμός

Δ2

- Ο πλεονασμός μίας πηγής μπορεί να οριστεί ως:

$$\pi = \bar{R} - H \geq 0$$

- Ο σχετικός πλεονασμός θα δίνεται από τη σχέση:

$$\delta = \frac{\pi}{\mathrm{H}} \cdot 100\%$$

Κώδικες Πηγής Ι

Δ2

- Κατηγοριοποιήσεις κωδίκων
  - Ανάλογα με το μήκος των κωδικών λέξεων (σταθερό, μεταβλητό)
  - Ανάλογα με τις απώλειες (Απωλεστικοί, Μη απωλεστικοί)
  - Ανάλογα με τη χρήση στατιστικής των δεδομένων (Στατικοί, Δυναμικοί)
  - …

Κώδικες Πηγής ΙΙ

- Κατηγοριοποιήσεις κωδίκων
    - Ανάλογα με το μήκος των κωδικών λέξεων (σταθερό, μεταβλητό)

Μήκος Λέξης - Συμβόλων Πηγής	Μήκος Λέξης Κώδικα
Σταθερό	Σταθερό
Σταθερό	Μεταβλητό
Μεταβλητό	Σταθερό
Μεταβλητό	Μεταβλητό

Κώδικες Πηγής ΙΙ

Δ2

- Κατηγοριοποιήσεις κωδίκων
  - Ανάλογα με το μήκος των κωδικών λέξεων (σταθερό, μεταβλητό)

	Σταθερού μήκους	**Μεταβλητού Μήκους**
Απόδοση Συμπίεσης	Εξαρτάται από το συνδυασμό πηγής-κώδικα	Εξαρτάται από το συνδυασμό πηγής-κώδικα
Πολυπλοκότητα	Μικρότερη	Μεγαλύτερη

Κώδικες Πηγής ΙΙΙ

Δ2

- Κατηγοριοποιήσεις κωδίκων
  - Ανάλογα με τις απώλειες (Απωλεστικοί, Μη απωλεστικοί)

	Απωλεστικοί	**Μη Απωλεστικοί**
Απόδοση Συμπίεσης	Μεγαλύτερη	Μικρότερη
Απώλεια Δεδομένων	Ναι	Όχι

Κώδικες Πηγής IV

Δ2

- Κατηγοριοποιήσεις κωδίκων
  - Ανάλογα με τη χρήση στατιστικής των δεδομένων (Στατικοί, Δυναμικοί)

	Στατικοί	**Δυναμικοί**
Πολυπλοκότητα	Μικρή	Μεγάλη
Προϋπολογισμένες πιθανότητες Εμφάνισης	Ναι	Όχι

Δ2

Μη απωλεστικοί κώδικες πηγής

Κώδικες Σταθερού Μήκους Λέξης

Δ2

- Σε κάθε σύμβολο της πηγής αντιστοιχίζεται σε μία κωδική λέξη σταθερού πλήθους bits.
- Για μία πηγή που διαθέτει k διακριτά σύμβολα απαιτούνται δυαδικές κωδικές λέξεις με r bits ώστε να ισχύει :

$$2^{r-1} < k \leq 2^{r}$$

- Το μέσο μήκος δυαδικής λέξης είναι r bits
- Δεν απαιτείται γνώση των στατιστικών χαρακτηριστικών της πηγής
- Κώδικας ASCII, Δυαδικός Κώδικας Gray κ.α.

Κώδικες Μεταβλητού Μήκους Λέξης Ι

Δ2

- Σε κάθε σύμβολο της πηγής αντιστοιχίζεται σε μία κωδική λέξη μεταβλητού πλήθους bits.
- Υπάρχουν πολλαπλές διαφορετικές κωδικοποιήσεις όπως και μη επιτρεπτές κωδικοποιήσεις
- Για αποτελεσματική κωδικοποίηση απαιτείται γνώση των στατιστικών χαρακτηριστικών της πηγής
- Κώδικες Shannon-Fano, Huffman κ.α.

Κώδικες Μεταβλητού Μήκους Λέξης ΙΙ

Δ2

- Έστω μία πηγή η οποία εκπέμπει σύμβολαS_i από ένα αλφάβητοC

$$\mathrm{S}_i \in C = \{S_1, S_2, \dots, S_M\}$$

- Σε κάθε σύμβολο S_i του παραπάνω αλφάβητου αναθέτουμε μία δυαδική λέξη B_i με μήκος $l(S_i) = l_i \geq 1$:

$$\mathrm{B}_i = \left\{B_0^i, \dots B_{l_i-1}^i\right\}$$

- Αν οι πιθανότητες εμφάνισης των συμβόλων είναι αντίστοιχα $\mathrm{p}_{\mathrm{S}_i} = \mathrm{p}_i$ τότε το μέσο μήκος λέξης του κώδικα θα έχει αναμενόμενη τιμή:

$$\bar{R} = \bar{l} = \sum_{i=1}^{M} p_i \cdot l_i$$

Άσκηση 2.1

Δ2

- Να υπολογίσετε το μέσο μήκος λέξης για τους παρακάτω κώδικες

S_i	p_i	Κώδικας Α	Κώδικας Β	Κώδικας Γ	Κώδικας Δ	Κώδικας Ε	Κώδικας ΣΤ
S1	0,5	00	0	0	0	00	0
S2	0,25	01	10	01	01	01	10
S3	0,125	10	11	010	011	10	110
S4	0,125	11	11	011	111	110	111

Άσκηση 2.1

Δ2

- Να υπολογίσετε το μέσο μήκος λέξης για τους παρακάτω κώδικες

S_i	p_i	Κώδικας Α	Κώδικας Β	Κώδικας Γ	Κώδικας Δ	Κώδικας Ε	Κώδικας ΣΤ
S1	0,5	00	0	0	0	00	0
S2	0,25	01	10	01	01	01	10
S3	0,125	10	11	010	011	10	110
S4	0,125	11	11	011	111	110	111
	$\bar{l}$(bits/symbol)	2	1,5	1,75	1,75	2,125	1,75

Αποκωδικοποίηση

Δ2

- Ένας κώδικας θα πρέπει να είναι μονοσήμαντα αποκωδικοποιήσιμος
  - Για κάθε σύμβολο S_iνα υπάρχει μοναδική αναπαράσταση B_i στον κώδικα
  - Αν παραχθούν πολλά σύμβολα σε μία συνεχή σειρά στην αναπαράσταση του κώδικα θα πρέπει τα σύμβολα να μπορούν να διαχωριστούν κατά την αποκωδικοποίηση

Προθεματικοί Κώδικες - Άμεσα Αποκωδικοποιήσιμοι

Δ2

- Προθεματική Συνθήκη
  - Αν μία σειρά από ψηφία του κώδικα είναι αναγνωρίσιμη κωδική λέξη, τότε δεν υπάρχει πιθανότητα αυτή η κωδική λέξη πρόθεμα μιας άλλης μεγαλύτερης κωδικής λέξης με την ίδια αρχή.
- Άμεση Αποκωδικοποίηση
  - Ένα σύμβολο μπορεί να αποκωδικοποιηθεί αμέσως μόλις ολοκληρωθεί η λήψη της κωδικοποιημένης αναπαράστασης του

Άσκηση 2.2

Δ2

- Να χαρακτηρίσετε τους παρακάτω κώδικες

S_i	p_i	Κώδικας Α	Κώδικας Β	Κώδικας Γ	Κώδικας Δ	Κώδικας Ε	Κώδικας ΣΤ
S1	0,5	00	0	0	0	00	0
S2	0,25	01	10	01	01	01	10
S3	0,125	10	11	010	011	10	110
S4	0,125	11	11	011	111	110	111
	$\bar{l}$(bits/symbol)	2	1,5	1,75	1,75	2,125	1,75

Βέλτιστοι κώδικες Μεταβλητού Μήκους

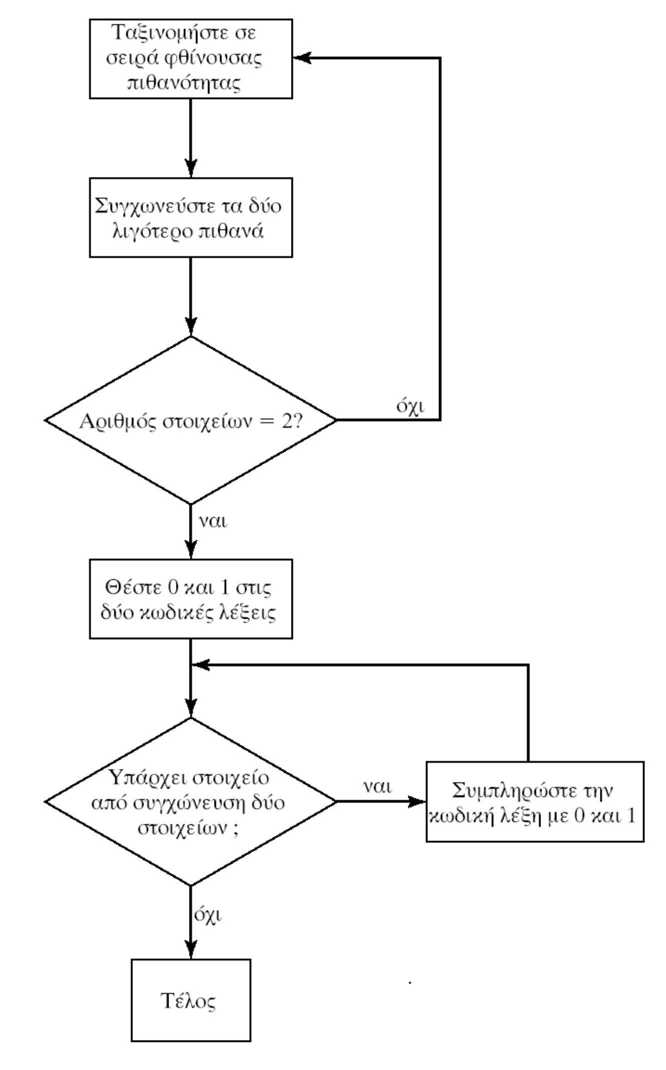
Δ2

- Ο Shannon απέδειξε ότι για τους κώδικες μεταβλητού μήκους, μπορούν να δημιουργηθούν κώδικες για τους οποίους ισχύει ότι :

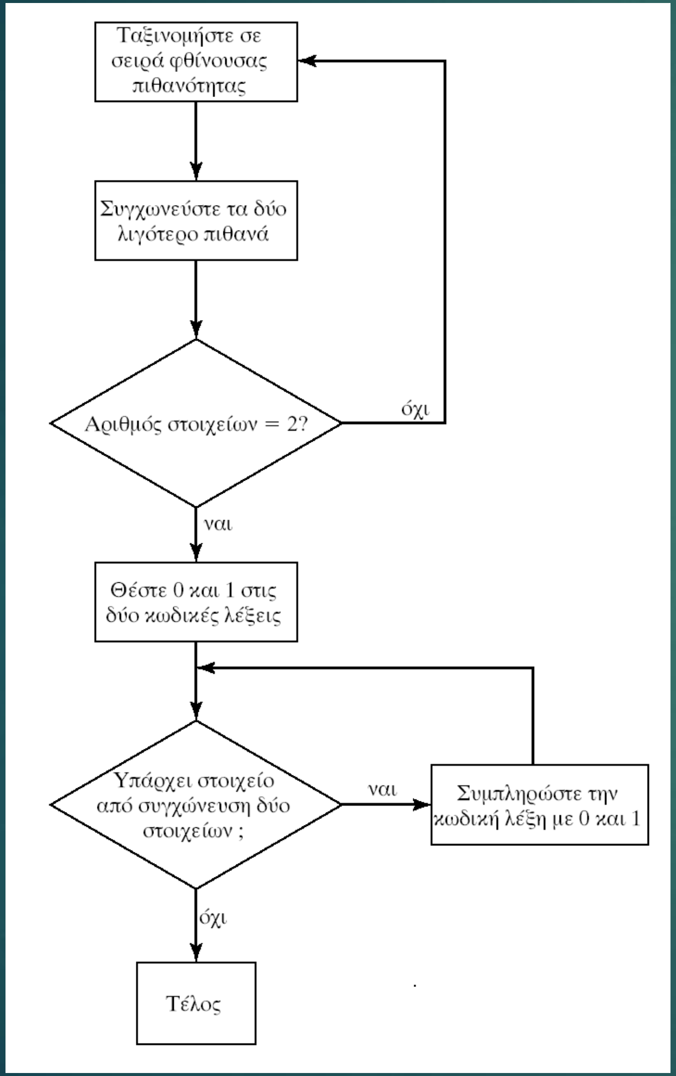
$$H \leq \overline{\mathrm{R}} < H + 1$$

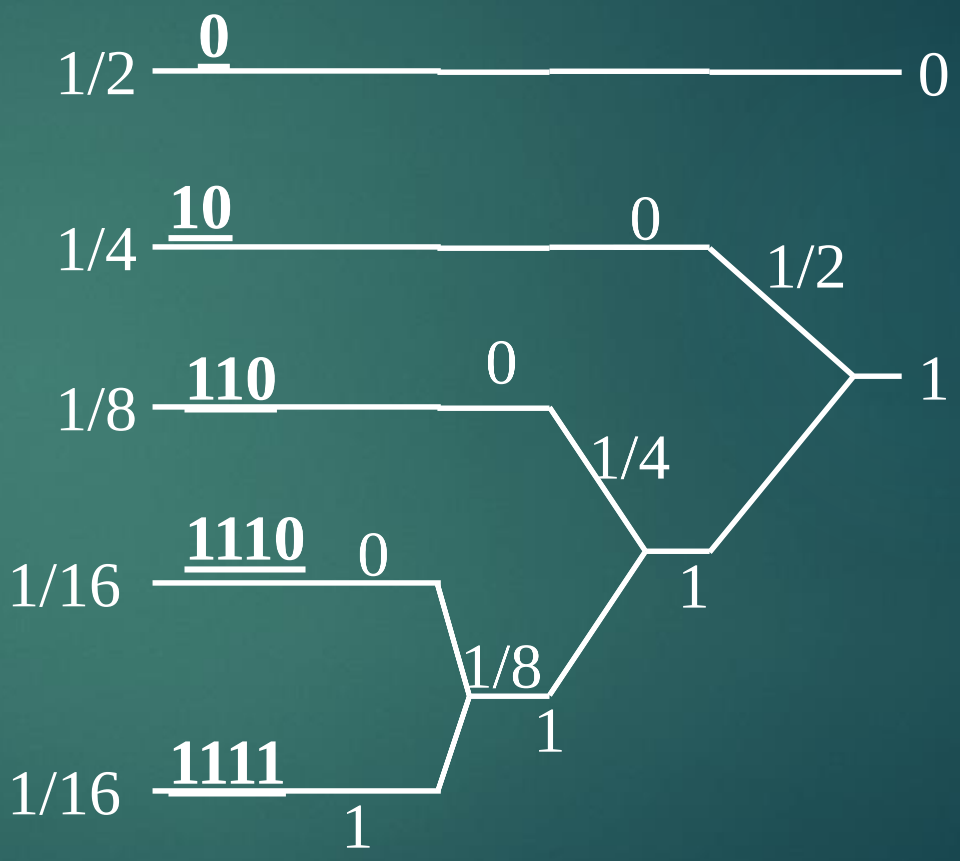
Κώδικας Huffman

Δ2



Κώδικας Huffman





Κωδικοποίηση Huffman

Δ2

- symbols = [1:6]; % Σύμβολα της πηγής
- p = [.5 .125 .125 .125 .0625 .0625]; % Πιθανότητες Συμβόλων
- [dict,avglen] = huffmandict(symbols,p); % Δη
- actualsig = randsrc(100,1,[symbols; p]); % Create data using p.
- comp = huffmanenco(actualsig,dict); % Encode the data.

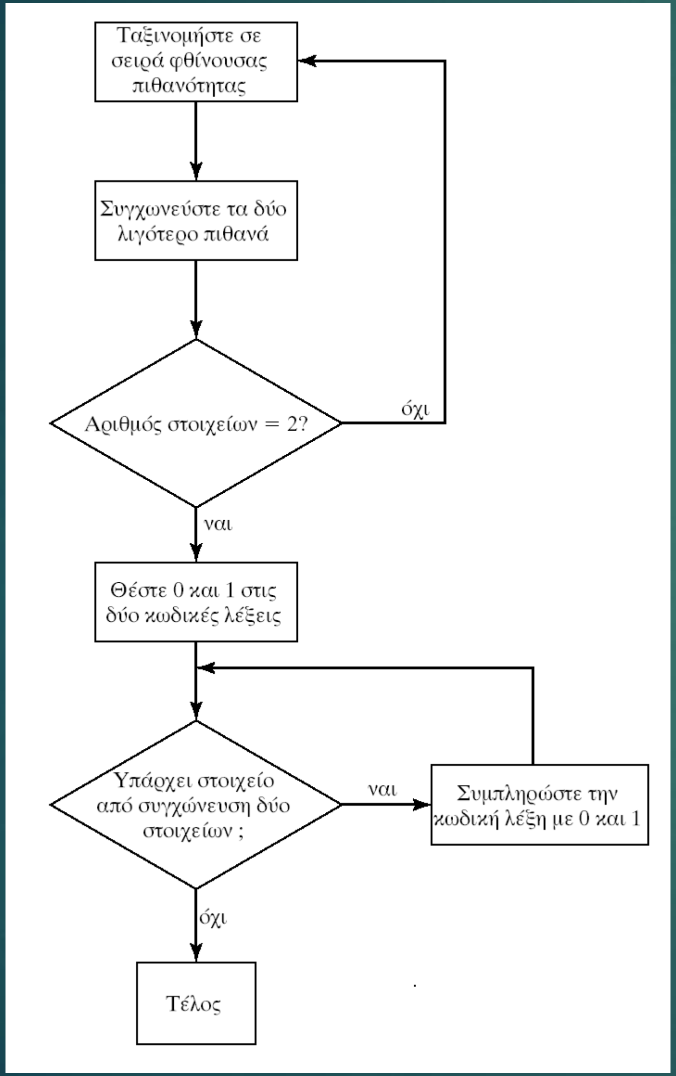
Κώδικες Lempel-Ziv

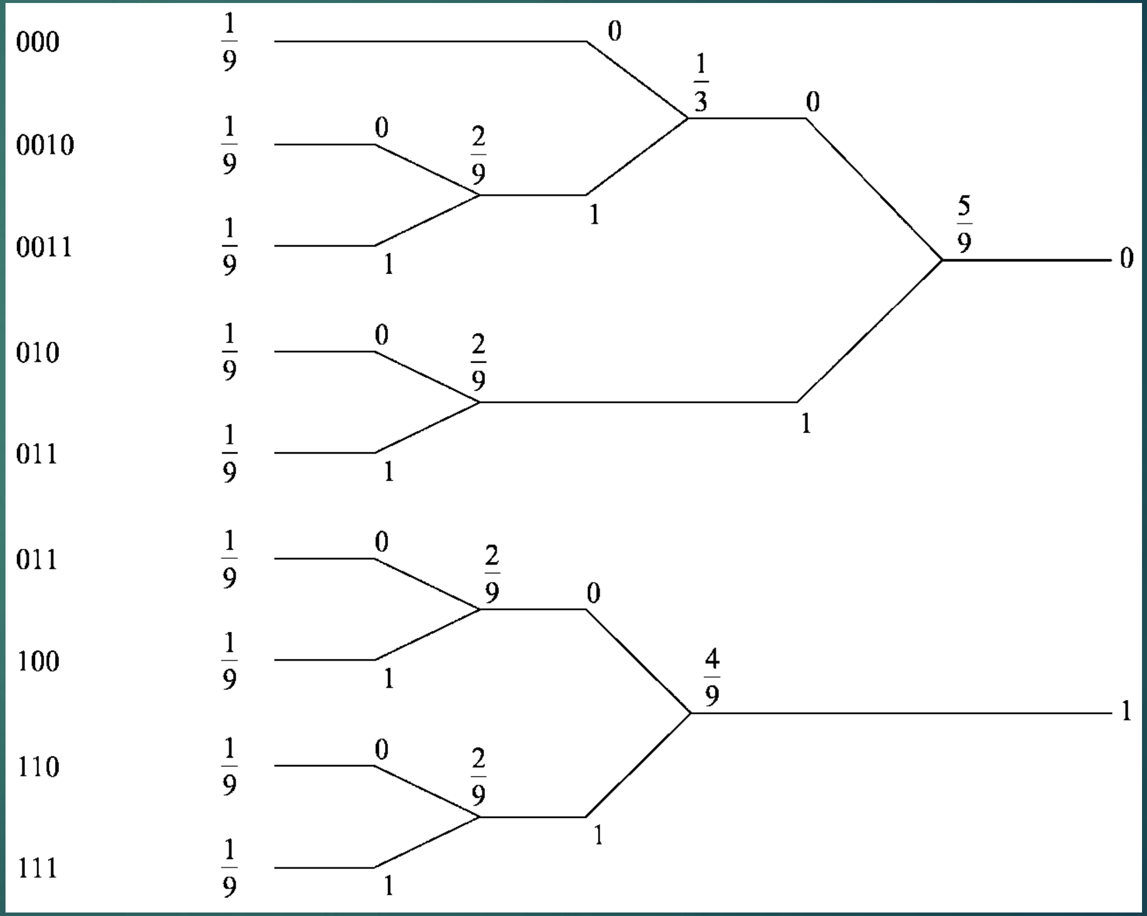
Δ2

- Έχουν ασυμπτωτική προσέγγιση στην εντροπία μιας πηγής για μεγάλες ακολουθίες συμβόλων
- Δεν απαιτείται στον αποκωδικοποιητή η γνώση του πίνακα κωδικοποίησης που έφτιαξε ο κωδικοποιητής
- Ο πίνακας κωδικοποίησης κατασκευάζεται δυναμικά
- Τα δεδομένα έχουν και ρόλο πίνακα κωδικοποίησης
- Οι κωδικές λέξεις έχουν σταθερό μήκος
- Αλγόριθμοι : LZ77(gzip,pkzip), LZ78(GIF)

Κώδικας Huffman

Δ2





Αλγόριθμος Lempel-Ziv-Welch

Αλγόριθμος 1 Αλγόριθμος LZW

```
Αρχικοποίηση λεξικού με το αλφάβητο της πηγής
p ← πρώτο σύμβολο ακολουθίας
Όσο υπάρχουν σύμβολα στην ακολουθία επανάλαβε
    c ← το επόμενο σύμβολο της ακολουθίας
    Αν το p || c υπάρχει στο λεξικό τότε
        p ← p || c
    Αλλιώς
        μετάδοση του στοιχείου p
        αποθήκευση p || c στο λεξικό
        p ← c
    Τέλος Αν
Τέλος Όσο
```

Άσκηση 2.3

Δ2

Αλγόριθμος 1 Αλγόριθμος LZW

Αρχικοποίηση λεξικού με το αλφάβητο της πηγής
$p \leftarrow$ πρώτο σύμβολο ακολουθίας
Όσο υπάρχουν σύμβολα στην ακολουθία **επανάλαβε**
    $c \leftarrow$ το επόμενο σύμβολο της ακολουθίας
    Αν το $p \parallel c$ υπάρχει στο λεξικό **τότε**
        $p \leftarrow p \parallel c$
    Αλλιώς
        μετάδοση του στοιχείου p
        αποθήκευση $p \parallel c$ στο λεξικό
        $p \leftarrow c$
    Τέλος Αν
Τέλος Όσο

Κωδικός	Σύμβολο
...	...
097	a
098	b
099	c
100	d
101	e
102	f
...	...
256	(Έναρξη)
257	(Λήξη)

Άσκηση 2.3

Δ2

- Κωδικοποίηση της ακολουθίας : abbcdeabbcdeabdeaf

Αλγόριθμος 1 Αλγόριθμος LZW

Αρχικοποίηση λεξικού με το αλφάβητο της πηγής
$p \leftarrow$ πρώτο σύμβολο ακολουθίας
Όσο υπάρχουν σύμβολα στην ακολουθία **επανάλαβε**
    $c \leftarrow$ το επόμενο σύμβολο της ακολουθίας
    Αν το $p \,||\, c$ υπάρχει στο λεξικό **τότε**
        $p \leftarrow p \,||\, c$
    Αλλιώς
        μετάδοση του στοιχείου p
        αποθήκευση $p \,||\, c$ στο λεξικό
        $p \leftarrow c$
    Τέλος Αν
Τέλος Όσο

Είσοδος		Λεξικό		Μετάδοση	
Κωδικός	Σύμβολο	Κωδικός	Σύμβολο	Κωδικός	Σύμβολο
097	a	-	-	256	(Έναρξη)
098	b	258	ab	097	a
098	b	259	bb	098	b
099	c	260	bc	098	b
100	d	261	cd	099	c
101	e	262	de	100	d
097	a	263	ea	101	e
098	b	-	-	-	-
098	b	264	abb	258	ab
099	c	-	-	-	-
100	d	265	bcd	260	bc
101	e	-	-	-	-
097	a	266	dea	262	de
098	b	-	-	-	-
100	d	267	abd	258	ab
101	e	-	-	-	-
097	a	-	-	-	-
102	f	268	deaf	266	dea
-	-	268	-	102	f
-	-	268	-	257	(Λήξη)

Άσκηση 2.3

Δ2

- Ποιο το συνολικό κόστος για την κωδικοποίηση της παραπάνω ακολουθίας με τον LZW και ποιό το κόστος μετάδοσης χωρίς κωδικοποίηση;

- Θεωρήστε ότι για κάθε σύμβολο στο 0-255 χρησιμοποιούνται 8bits ενώ το πλήθος των bits θα αυξηθεί για όλα τα μεταβιβαζόμενα κωδικοποιημένα σύμβολα με τον LZW κατά 1 bit

Επέκταση πηγής Ι

Δ2

- Με τον όρο επέκταση πηγής εννοείται η αλλαγή των συμβόλων της σε νέα σύμβολα που δημιουργούνται από τους πιθανούς συνδυασμούς των συμβόλων της

- Θεωρώντας συνδυασμούς L συμβόλων μίας πηγής που αρχικά είχε N σύμβολα τα νέα σύμβολα της πηγής θα είναι N^L

- Προκαλείται μεταβολή στις πιθανότητες εμφάνισης άρα μεταβάλλεται η εντροπία της πηγής

- Η επέκταση πηγής χρησιμοποιείται για τη βελτίωση αποδοτικότητας αλγορίθμων συμπίεσης

Επέκταση πηγής ΙΙ

Δ2

- Ο ρυθμός για την L επέκταση πηγής θα φράσσεται από την

$$H_L \leq \bar{R}_L < H_L + 1$$

- Ο μέσος ρυθμός ανά σύμβολο θα είναι $\bar{R} = \frac{1}{L}\bar{R}_L$

- Οπότε $H_L \leq \mathrm{L} \cdot \bar{R} < H_L + 1 \rightarrow \frac{H_L}{L} \leq \bar{R} < \frac{H_L}{L} + \frac{1}{L}$

- Επειδή (αποδεικνύεται) $H_L = \mathrm{L} \cdot H$ *προκύπτει ότι*

$$H \leq \bar{\mathrm{R}} < H + \frac{1}{\mathrm{L}}$$

Παράδειγμα

Δ2

- Θεωρήστε μία πηγή με 3 διαφορετικά σύμβολα $s1, s2, s3$ τα οποία έχουν πιθανότητες εμφάνισης p1=0.95, p2=0.03, p3=0.02
- Να κατασκευαστεί ο κώδικας Huffmann για την παραπάνω πηγή
- Να βρεθεί η ποσοστιαία % διαφορά του μέσου μήκους της λέξης του κώδικα από την εντροπία της πηγής

Παράδειγμα

Δ2

- Θεωρήστε μία πηγή με 3 διαφορετικά σύμβολα $s1, s2, s3$ τα οποία έχουν πιθανότητες εμφάνισης p1=0.95, p2=0.03, p3=0.02
- Να κατασκευαστεί ένας κώδικας Huffmann για την παραπάνω πηγή
- Να βρεθεί η ποσοστιαία % διαφορά του μέσου μήκους της λέξης του κώδικα από την εντροπία της πηγής

Σύμβολο	Κωδική Λέξη
s1	0
s2	11
s3	10

Παράδειγμα

Δ2

- Θεωρήστε μία πηγή με 3 διαφορετικά σύμβολα $s1, s2, s3$ τα οποία έχουν πιθανότητες εμφάνισης p1=0.95, p2=0.03, p3=0.02
- Να κατασκευαστεί ένας κώδικας Huffmann για την παραπάνω πηγή
- Να βρεθεί η ποσοστιαία % διαφορά του μέσου μήκους της λέξης του κώδικα από την εντροπία της πηγής

$$H = 0.335 \frac{bits}{symbol}, \qquad \bar{l} = \bar{R} = 1.05 \ \frac{bits}{symbol}$$

$$\delta(\%) = \frac{|\bar{l} - H|}{H} \cdot 100 = 213\%$$

Παράδειγμα

Δ2

- Θεωρώντας την L=2 επέκταση της πηγής προκύπτει ο πίνακας

Σύμβολο	Πιθανότητα εμφάνισης	Κωδικές Λέξεις Huffman
s_1s_1	0,9025	0
s_1s_2	0,0190	111
s_1s_3	0,0285	100
s_2s_1	0,0190	1101
s_2s_2	0,0004	110011
s_2s_3	0,0006	110001
s_3s_1	0,0285	101
s_3s_2	0,0006	110010
s_3s_3	0,0009	110000

Παράδειγμα

Δ2

- Υπολογίζοντας το μέσο ρυθμό (μέσο μήκος λέξης) και την εντροπία προκύπτει ότι

$$H = 0.335 \frac{bits}{symbol}, \qquad \bar{R} = \frac{\bar{\mathrm{R}}_L}{L} = 0.611 \ \frac{bits}{symbol}$$

$$\delta(\%) = \frac{|\bar{R} - H|}{H} \cdot 100 = 82\%$$

Αριθμητική Κωδικοποίηση Ι

Δ2

- Δεν ορίζονται διαφορετικές κωδικές λέξεις για τα σύμβολα της πηγής
- Μία σειρά από σύμβολα κωδικοποιείται με μία μοναδική κωδική λέξη
- Αύξηση απόδοσης από την εξάλειψη της ανάγκης ανάθεσης ακέραιου πλήθους ψηφίων στις κωδικές λέξεις κάθε συμβόλου
- Απαιτείται γνώση των πιθανοτήτων εμφάνισης των συμβόλων
- Βασίζεται στο αξίωμα πληρότητας των πραγματικών αριθμών

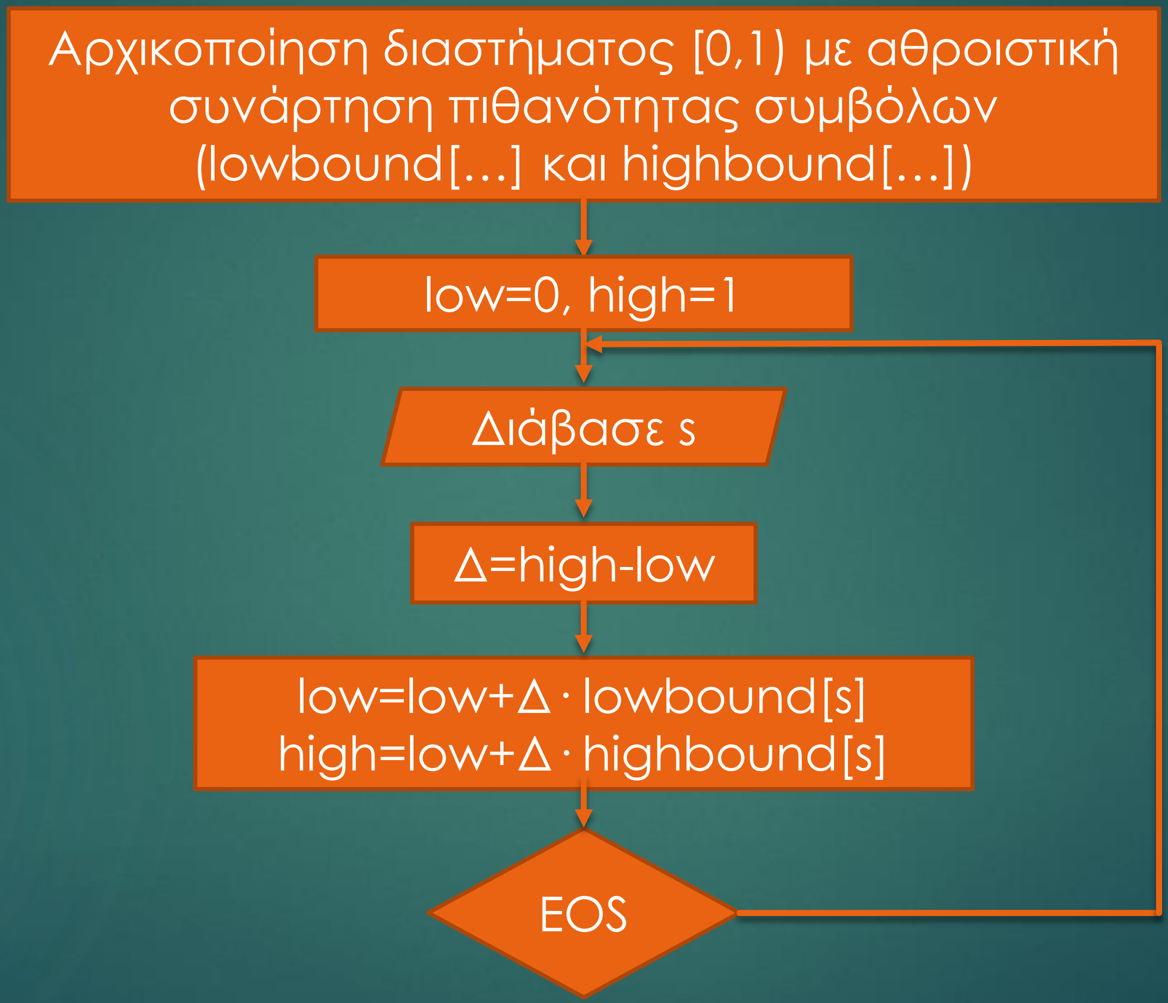
Αριθμητική Κωδικοποίηση ΙΙ

Δ2

- Διαίρεση του διαστήματος [0,1) ανάλογα με τις πιθανότητες εμφάνισης των συμβόλων (χρησιμοποιείται η συνάρτηση αθροιστικής πιθανότητας)
- Καθορίζονται διαστήματα αντί για συγκεκριμένες κωδικές λέξεις
- Το διάστημα κωδικοποίησης ενός συμβόλου καθορίζεται από το διάστημα του προηγούμενου συμβόλου που κωδικοποιήθηκε

Αριθμητική Κωδικοποίηση ΙΙ

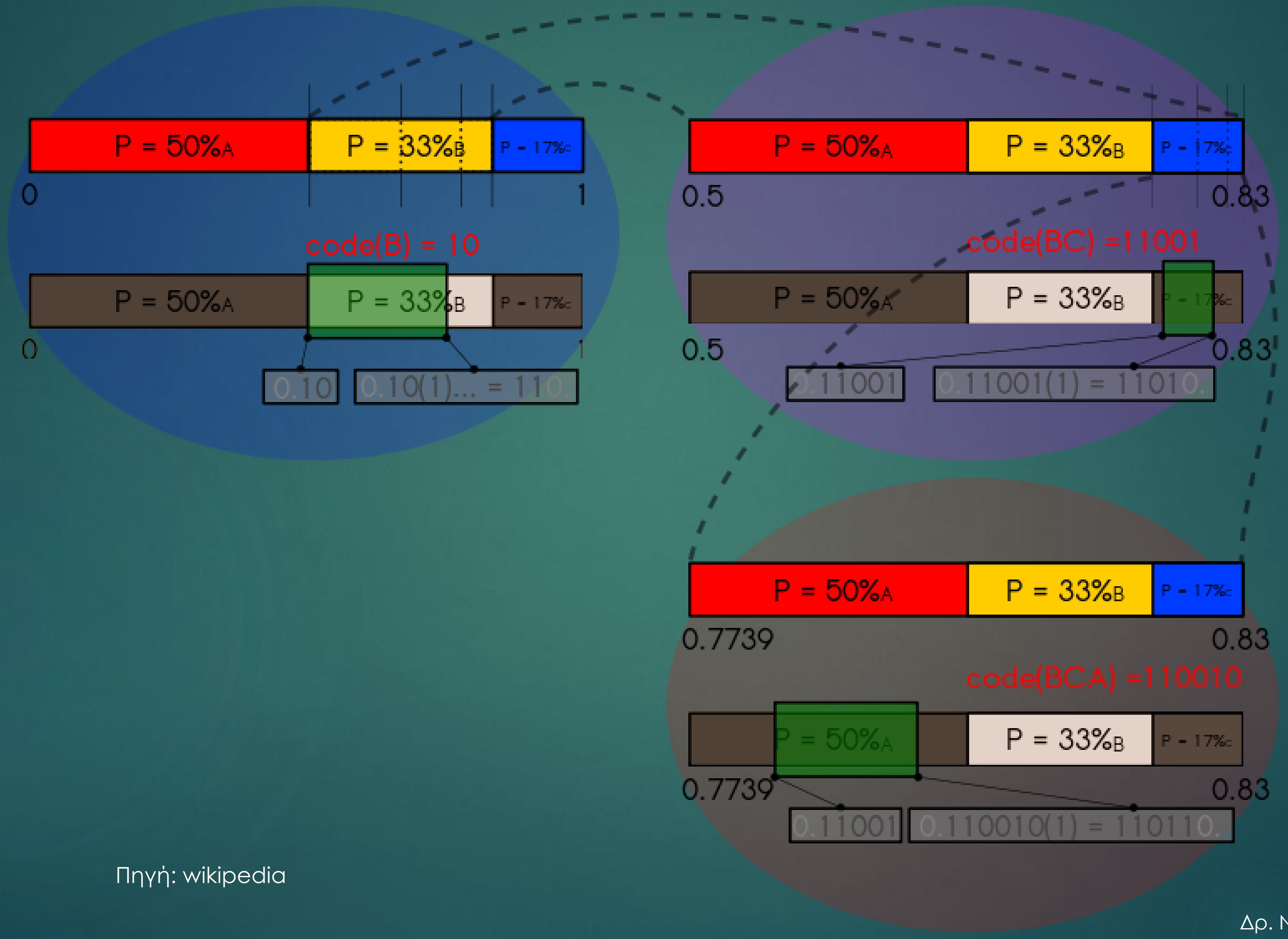
Δ2



EOS: End of Stream

Παράδειγμα

Δ2



Πηγή: wikipedia

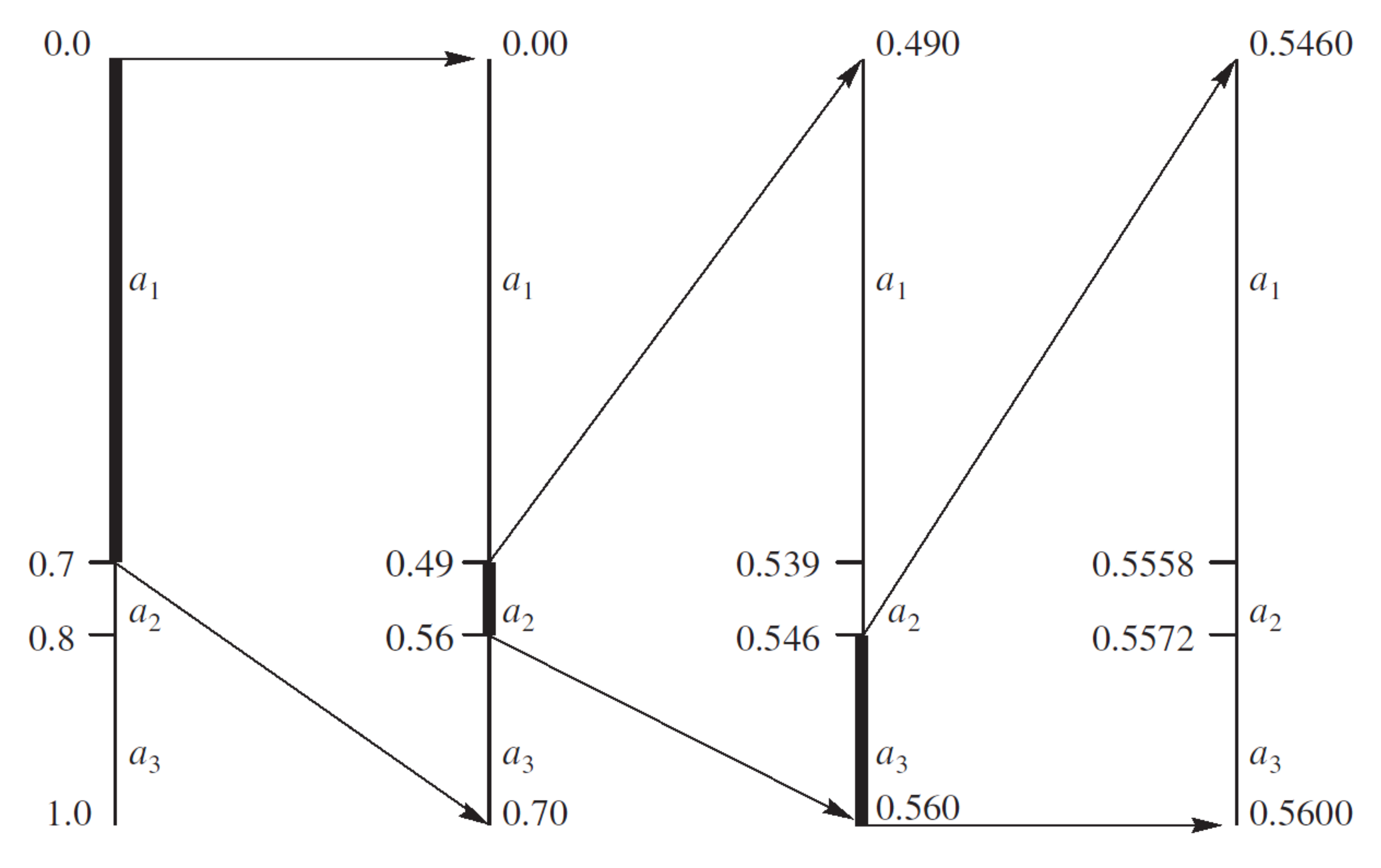
Άσκηση 2.4

Δ2

- Έστω μία πηγή με τρία σύμβολα a_1, a_2, a_3 με πιθανότητες $p_1=0.7$, $p_2=0.1$, $p_3=0.2$. Να κωδικοποιηθεί με αριθμητική κωδικοποίηση η συμβολοσειρά a_1, a_2, a_3
- Για την ίδια πηγή να αποκωδικοποιήσετε το μήνυμα 0.7538

Άσκηση 2.4

Δ2



Πηγή: Sayood

Αλγόριθμος Run Length Encoding (RLE)

Δ2

- Κωδικοποίηση επαναλαμβανόμενων χαρακτήρων στη μορφή :

  Χαρακτήρας, Πλήθος

- Λαμβάνεται ειδική μέριμνα στην περίπτωση ύπαρξης αριθμών στο κείμενο:

  Χαρακτήρας, &Πλήθος

- Μπορεί να χρησιμοποιηθεί με μεγάλη αποτελεσματικότητα σε δυαδικές ή δυαδικά κωδικοποιημένες πηγές

- Μπορεί να γίνει αποσύζευξη του ζεύγους κωδικοποίησης και επανακωδικοποίησης με άλλη μέθοδο για αύξηση του ποσοστού συμπίεσης.

Παράδειγμα Απλοποιημένης κωδικοποίησης με RLE

Δ2

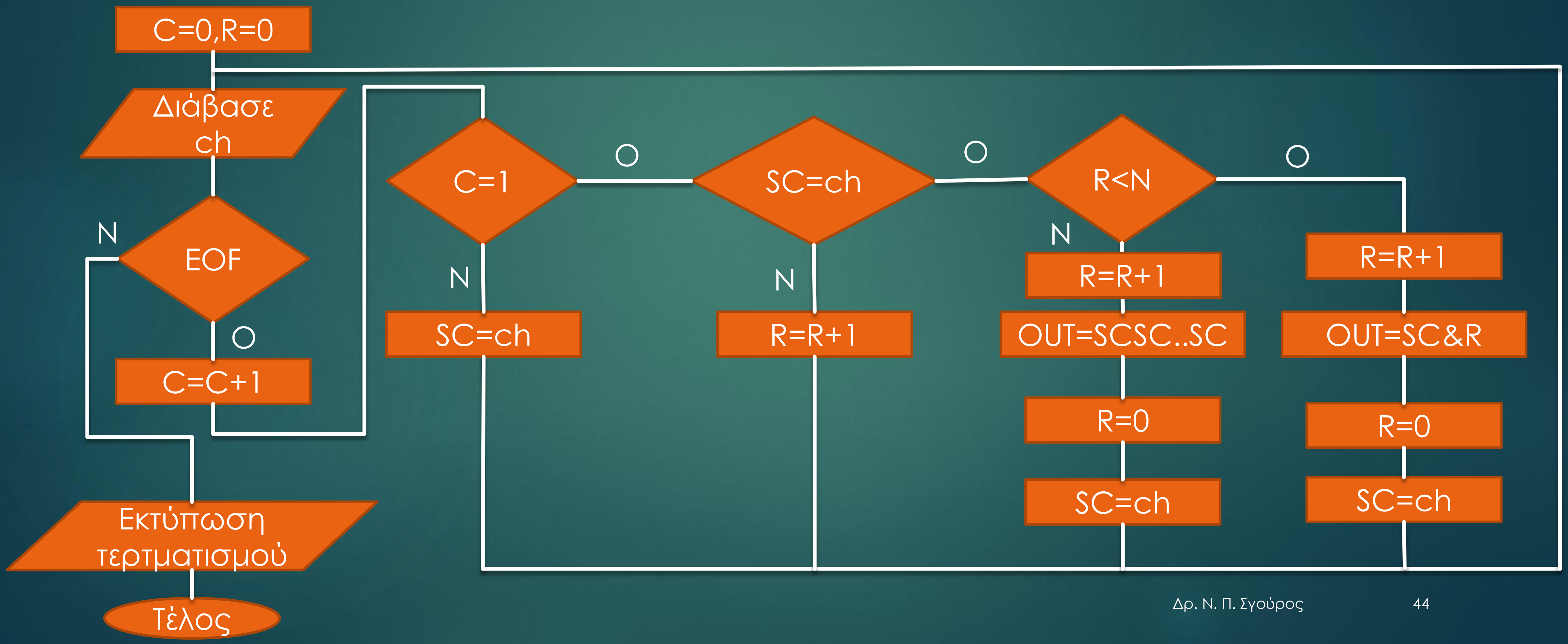
- Να κωδικοποιηθεί η συμβολοσειρά AAABBBABBBCCAACCAAB με χρήση του αλγορίθμου RLE

- AAABBBABBBCCAACCAAB →A3B3A1B3C2A2C2A2B1→(ABABCACAB)(331322221)

- Η αποσυζευγμένη ακολουθία επιδέχεται επιπλέον κωδικοποίηση

Αλγόριθμος Run Length Encoding (RLE)

Δ2



Άσκηση 2.5

Δ2

- Να υπολογιστεί ο λόγος συμπίεσης που επιτυγχάνεται για την κωδικοποίηση μιας συμβολοσειράς N χαρακτήρων στην περίπτωση που αυτή περιλαμβάνει γράμματα και αριθμούς και συνολικά περιέχει M επαναλαμβανόμενες σειρές χαρακτήρων μέσου μήκους L. Ποια η τιμή του λόγου συμπίεσης στην περίπτωση που κανένας χαρακτήρας δεν επαναλαμβάνεται;

- Σημείωση: Ο λόγος συμπίεσης είναι ίσος με $\lambda = \frac{\text{Μέγεθος συμπιεσμένης μορφής (έξοδος)}}{\text{Μέγεθος ασυμπίεστης μορφής (είσοδος)}}$

Άσκηση 2.6

Δ2

- Θεωρήστε ότι για τη μετάδοση μιας συμβολοσειράς η οποία χρησιμοποιεί 5 διαφορετικά σύμβολα χρησιμοποιείται ο κώδικας Huffman. Αν οι πιθανότητες των συμβόλων είναι a1=0.4, a2=a3=0.2 , a4=a5=0.1, να κατασκευάσετε δύο πιθανούς κώδικες Huffman για την κωδικοποίηση της παραπάνω συμβολοσειράς.

- Να αναπτύξετε κατάλληλο κριτήριο για την αξιολόγηση της επίδοσης των διαφορετικών κωδίκων Huffman για την παραπάνω συμβολοσειρά που θα εξασφαλίζει όσο το δυνατό σταθερό ρυθμό μετάδοσης σε ένα δίκτυο δεδομένων.