ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

Κωδικοποίηση Πηγής

Θεωρία Πληροφορίας και Κωδίκων

Εαρινό Εξάμηνο

Τμήμα Πληροφορικής και Τηλεπικοινωνιών

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

30/4/2020 1:44:40 μμ

Κωδικοποίηση Πηγής

- Ένας κωδικοποιητής πηγής δέχεται ως είσοδο σύμβολα από πηγή πληροφορίας και εξάγει bit

- Έστω πηγή συμβόλων $\mathscr{L} = \{s_1, s_2, \ldots, s_n\}$ τα οποία έχουν ανεξάρτητες πιθανότητες $\{P_1, P_2, \ldots, P_n\}$

- Τα bit εξόδου του κωδικοποιητή είναι οργανωμένα σε προκαθορισμένες ομάδες από bit (όχι κατ' ανάγκη ίδιου μήκους), οι οποίες ονομάζονται κωδικολέξεις (*codewords*)

- Για το σύμβολο s_i (με i=1,2,…,n), ο κωδικοποιητής παράγει την κωδικολέξη $c(s_i)$ μήκους ℓ_i σε bit

$$s_i \longrightarrow c(s_i)$$

- Προφανώς, η πιθανότητα εμφάνισης του s_i θα είναι ίδια με την πιθανότητα εμφάνισης του $c(s_i)$

Κωδικοποίηση Πηγής

- Το μέσο μήκος των κωδικολέξεων σε bit είναι

$$\overline{L} = \sum_{i=1}^{n} P_i \, \ell_i$$

- Παράδειγμα: Πηγή παράγει τα σύμβολα {Α, Β, Γ}, με πιθανότητες {0.5, 0.25, 0.25}, τα οποία κωδικοποιούνται ως εξής

$$s_1 = \mathrm{A} \rightarrow c(s_1) = 01$$
$$s_2 = \mathrm{B} \rightarrow c(s_2) = 110$$
$$s_3 = \Gamma \rightarrow c(s_3) = 00$$

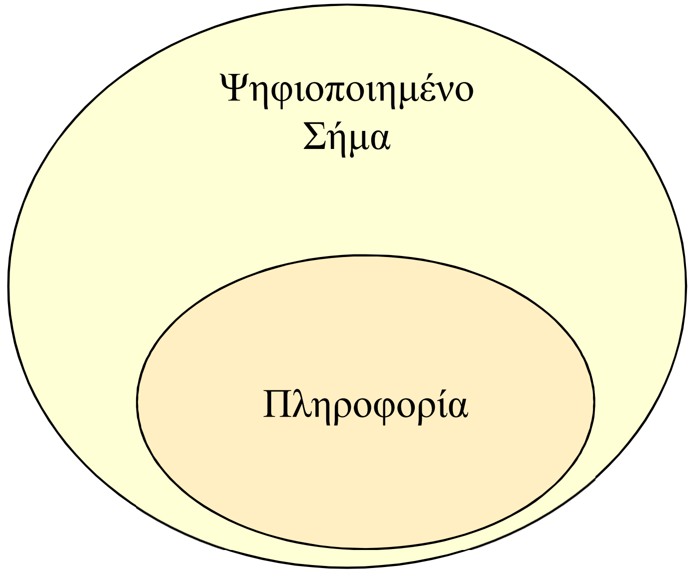
- Για την ακολουθία συμβόλων ΑΑΓΒΒΓ, ο κωδικοποιητής θα παράξει την εξής ακολουθία bit

`01 01 00 110 110 00`

- Το μέσο μήκος των κωδικολέξεων είναι $\overline{L} = 0.5\times2 + 0.25\times3 + 0.25\times2 = 2.25$ bit

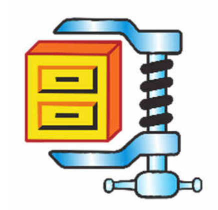
Κωδικοποίηση Πηγής

- Οι φυσικές πηγές παράγουν σήματα τα οποία έχουν μεγάλη ποσότητα πλεονασμού (*redundancy*)



- Για αποδοτική μετάδοση των σημάτων θα πρέπει πρώτα να αφαιρεθεί ο πλεονασμός

- Η (μη απωλεστική) διαδικασία αφαίρεσης του πλεονασμού ονομάζεται συμπίεση (*compression*)



- Η εντροπία της πηγής θέτει το θεμελιώδες όριο για την αφαίρεση της πλεοναστικής πληροφορίας

Κωδικοποίηση Πηγής

- Ένα κώδικας ονομάζεται αντιστρέψιμος εφόσον κάθε σύμβολο της πηγής απεικονίζεται από μία διαφορετική ακολουθία από bit

$$s_i \neq s_j \quad \Rightarrow \quad c(s_i) \neq c(s_j)$$

- Το πρόβλημα των αντιστρέψιμων κωδίκων είναι ότι δεν είναι κατ' ανάγκη μονοσήμαντα αποκωδικοποιήσιμοι

- Παράδειγμα: Η ακολουθία 0 0 1 1 0 1 αποκωδικοποιείται ως

  Γ Δ Α Β, Α Α Β Β Β Α Β, Γ Β Β Α Β ή Β Β Δ Α Β

Σύμβολα	**Κωδικολέξεις**
$s_1 = \mathrm{A}$	$c(s_1) = 0$
$s_2 = \mathrm{B}$	$c(s_2) = 1$
$s_3 = \Gamma$	$c(s_3) = 00$
$s_4 = \Delta$	$c(s_4) = 11$

Κωδικοποίηση Πηγής

- Υποσύνολο των αντιστρέψιμων κωδίκων είναι οι μονοσήμαντα αποκωδικοποιήσιμοι κώδικες

$$c(s_1\ s_2 \cdots s_n) \;\rightarrow\; c(s_1)\ c(s_2) \cdots c(s_n) \;\rightarrow\; s_1\ s_2 \cdots s_n$$

- Για την αποκωδικοποίηση των μονοσήμαντα αποκωδικοποιήσιμων κωδίκων πιθανόν να πρέπει να ληφθεί έως και όλη η ακολουθία των κωδικών λέξεων

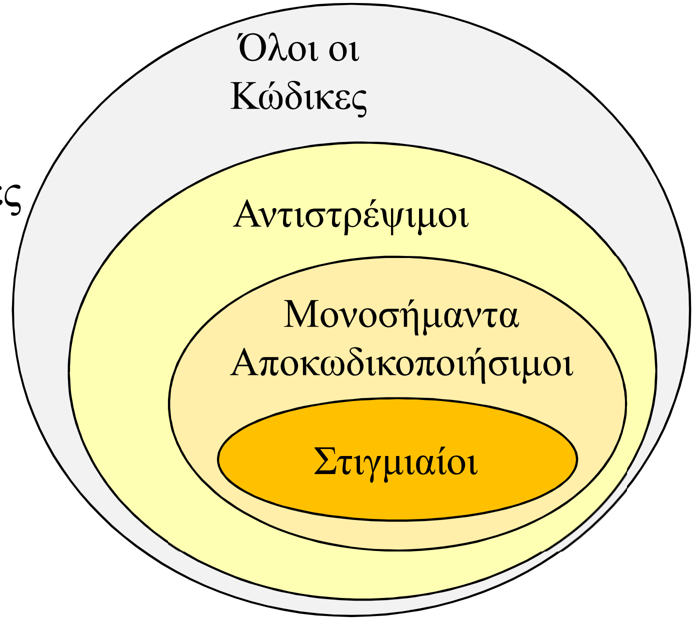
- Οι μονοσήμαντα αποκωδικοποιήσιμοι κώδικες χαρακτηρίζονται από υψηλή πολυπλοκότητα κατά την αποκωδικοποίηση

- <u>Παράδειγμα</u>: Η ακολουθία 1 1 0 0 0 1 0 0 0 μπορεί να αποκωδικοποιηθεί αρχικά ως Γ Β 0 1 0 0 0, όμως για το 0 που ακολουθεί δεν υπάρχει κατάλληλη κωδική λέξη.
  Η ορθή αποκωδικοποίηση είναι Δ Β Α Β

Σύμβολα	**Κωδικολέξεις**
$s_1 = \mathrm{A}$	$c(s_1) = 10$
$s_2 = \mathrm{B}$	$c(s_2) = 00$
$s_3 = \Gamma$	$c(s_3) = 11$
$s_4 = \Delta$	$c(s_4) = 110$

Κωδικοποίηση Πηγής

- Υποσύνολο των μονοσήμαντα αποκωδικοποιήσιμων κωδίκων είναι οι στιγμιαίοι (*instantaneous*) ή απροθεματικοί (*prefix*) κώδικες



- Οι στιγμιαίοι κώδικες χαρακτηρίζονται από το γεγονός ότι καμία κωδική λέξη δεν αποτελεί πρόθεμα κάποιας άλλης

- Οι κωδικές λέξεις ενός στιγμιαίου κώδικα μπορούν να αποκωδικοποιηθούν άμεσα μόλις ληφθούν, χωρίς πχ να απαιτείται να ληφθούν επόμενες λέξεις

- <u>Παράδειγμα</u>: Η ακολουθία 1 1 1 0 0 0 1 0 0 0 αποκωδικοποιείται μοναδικά ως Δ Α Α Α Β Α Α

Σύμβολα	Κωδικολέξεις
$s_1 = \text{A}$	$c(s_1) = 0$
$s_2 = \text{B}$	$c(s_2) = 10$
$s_3 = \Gamma$	$c(s_3) = 110$
$s_4 = \Delta$	$c(s_4) = 111$

Κωδικοποίηση Πηγής

- Η πηγή παράγει σύμβολα $\mathscr{L} = \{s_1, s_2, \ldots, s_n\}$ με ανεξάρτητες πιθανότητες $\{P_1, P_2, \ldots, P_n\}$

- Για το σύμβολο s_i ο κωδικοποιητής παράγει την κωδική λέξη $c(s_i)$ μήκους ℓ_i σε bit (με $i = 1, 2, \ldots, n$)

- Αν ο κώδικας είναι στιγμιαίος, ικανοποιεί την ανισότητα Kraft-McMillan

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

- Οι μονοσήμαντα αποκωδικοποιήσιμοι κώδικες ικανοποιούν επίσης την παραπάνω ανισότητα και άρα, αν ένας κώδικας ικανοποιεί την ανισότητα Kraft-McMillan δεν είναι κατ' ανάγκη στιγμιαίος

<table>
<tr><th colspan="4">Κωδικολέξεις</th></tr>
<tr><th>$\ell = 1$</th><th>$\ell = 2$</th><th>$\ell = 3$</th><th>$\ell = 3$</th></tr>
<tr><td rowspan="8">0</td><td rowspan="4">00</td><td rowspan="2">000</td><td>0000</td></tr>
<tr><td>0001</td></tr>
<tr><td rowspan="2">001</td><td>0010</td></tr>
<tr><td>0011</td></tr>
<tr><td rowspan="4">01</td><td rowspan="2">010</td><td>0100</td></tr>
<tr><td>0101</td></tr>
<tr><td rowspan="2">011</td><td>0110</td></tr>
<tr><td>0111</td></tr>
<tr><td rowspan="8">1</td><td rowspan="4">10</td><td rowspan="2">100</td><td>100</td></tr>
<tr><td>1001</td></tr>
<tr><td rowspan="2">101</td><td>1010</td></tr>
<tr><td>1011</td></tr>
<tr><td rowspan="4">11</td><td rowspan="2">110</td><td>1100</td></tr>
<tr><td>1101</td></tr>
<tr><td rowspan="2">111</td><td>1110</td></tr>
<tr><td>1111</td></tr>
</table>

Κωδικοποίηση Πηγής

- Ο κώδικας I δεν είναι στιγμιαίος, διότι δεν επαληθεύει την ανισότητα

$$\sum_{i=1}^{4} 2^{-\ell_i} = 2^{-1} + 2^{-1} + 2^{-2} + 2^{-2} = 1.5 > 1$$

- Οι κώδικες II και III ικανοποιούν την ανισότητα Kraft-McMillan

$$\sum_{i=1}^{4} 2^{-\ell_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$$

- Ωστόσο, μόνο ο κώδικας II είναι στιγμιαίος,
ενώ ο III είναι μονοσήμαντα αποκωδικοποιήσιμος

Σύμβολο	Πιθανότητα	Κώδικας I	Κώδικας II	Κώδικας III
s_1	0.5	0	0	0
s_2	0.25	1	10	01
s_3	0.125	00	110	011
s_4	0.125	11	111	0111

<table>
<tr><th colspan="4">Κωδικολέξεις</th></tr>
<tr><th>$\ell = 1$</th><th>$\ell = 2$</th><th>$\ell = 3$</th><th>$\ell = 3$</th></tr>
<tr><td rowspan="8">0</td><td rowspan="4">00</td><td rowspan="2">000</td><td>0000</td></tr>
<tr><td>0001</td></tr>
<tr><td rowspan="2">001</td><td>0010</td></tr>
<tr><td>0011</td></tr>
<tr><td rowspan="4">01</td><td rowspan="2">010</td><td>0100</td></tr>
<tr><td>0101</td></tr>
<tr><td rowspan="2">011</td><td>0110</td></tr>
<tr><td>0111</td></tr>
<tr><td rowspan="8">1</td><td rowspan="4">10</td><td rowspan="2">100</td><td>100</td></tr>
<tr><td>1001</td></tr>
<tr><td rowspan="2">101</td><td>1010</td></tr>
<tr><td>1011</td></tr>
<tr><td rowspan="4">11</td><td rowspan="2">110</td><td>1100</td></tr>
<tr><td>1101</td></tr>
<tr><td rowspan="2">111</td><td>1110</td></tr>
<tr><td>1111</td></tr>
</table>

Κωδικοποίηση Πηγής

- Το ελάχιστο μέσο μήκος των κωδικών λέξεων μιας πηγής δίδεται μέσω του 1ου θεωρήματος κωδικοποίησης του Shannon

- Για μία διακριτή πηγή χωρίς μνήμη εντροπίας $H(\mathscr{L})$, το μέσο μήκος $\overline{L}$ των κωδικών λέξεων οποιουδήποτε σχήματος κωδικοποίησης δεν μπορεί να είναι μικρότερο από $H(\mathscr{L})$

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

- Σε περίπτωση που χρησιμοποιηθεί σχήμα κωδικοποίησης, όπου δεν ικανοποιείται η παραπάνω συνθήκη, τα σύμβολα κατά την αποκωδικοποίηση θα είναι διαφορετικά σε σχέση με τα αρχικά

- Η απόδοση ενός κωδικοποιητή πηγής ορίζεται ως

$$\eta = \frac{H(\mathscr{L})}{\overline{L}}$$

Κωδικοποίηση Πηγής

- Για διακριτή πηγή χωρίς μνήμη εντροπίας $H(\mathscr{L})$, ποιο είναι το ελάχιστο μέσο μήκος κωδικολέξεων ενός στιγμιαίου κώδικα που μπορεί να κατασκευαστεί;

$$\overline{L} = \sum_{i=1}^{n} P_i \, \ell_i \xrightarrow[\bar{\ell}_1, \bar{\ell}_2, \ldots, \bar{\ell}_n]{} L_{\min}$$

- Ουσιαστικά πρόκειται για το εξής πρόβλημα βελτιστοποίησης:

$$\min_{\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_n} \left\{ \sum_{i=1}^{n} P_i \, \ell_i \right\}$$

$$\text{Υπό Συνθήκη:} \quad \sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

- Το παραπάνω πρόβλημα αντιμετωπίζεται χρησιμοποιώντας πολλαπλασιαστές Lagrange

Κωδικοποίηση Πηγής

- Θα ελαχιστοποιήσουμε την ποσότητα

$$J = \sum_{i=1}^{n} P_i \,\ell_i + \lambda \sum_{i=1}^{n} 2^{-\ell_i}$$

- Παραγωγίζοντας ως προς ℓ_i και εξισώνοντας με το μηδέν (με $i = 1, 2, \ldots, n$) έχουμε

$$\left.\frac{\partial J}{\partial \ell_i}\right|_{\hat{\ell}_i} = 0 \Leftrightarrow P_i - \lambda\, 2^{-\hat{\ell}_i}\ \ln(2) = 0 \Leftrightarrow 2^{-\hat{\ell}_i} = \frac{P_i}{\lambda\ \ln(2)}$$

- Με αντικατάσταση στον υπό συνθήκη περιορισμό (συγκεκριμένα στην ισότητα), προκύπτει το λ

$$\sum_{i=1}^{n} 2^{-\hat{\ell}_i} = 1 \Leftrightarrow \sum_{i=1}^{n} \frac{P_i}{\lambda\ \ln(2)} = 1 \Leftrightarrow \lambda = \frac{1}{\ln(2)}$$

- Συνεπώς,

$$P_i = 2^{-\hat{\ell}_i}$$

Κωδικοποίηση Πηγής

- Άρα, το βέλτιστο μήκος της i-ωστής κωδικολέξης $c(s_i)$ είναι ακριβώς η ποσότητα πληροφορίας του αντίστοιχου συμβόλου s_i, δηλαδή

$$\hat{\ell}_i = -\log_2(P_i)$$

- Συνεπώς, το μέσο μήκος των κωδικολέξεων (σε bit) είναι

$$\overline{L}_{\min} = \sum_{i=1}^{n} P_i\,\hat{\ell}_i = -\sum_{i=1}^{n} P_i\,\log_2(P_i) = H(\mathscr{L})$$

- Για να είναι το μήκος κάθε κωδικής λέξης ακέραιος, θα πρέπει σε όλες η κατανομή των πιθανοτήτων να είναι της μορφής $P_i \sim 2^{-k}$, με k κάποιο θετικό ακέραιο

Κωδικοποίηση Πηγής

- Στην πράξη οι πιθανότητες εμφάνισης των συμβόλων της πηγής δεν είναι κατ' ανάγκη της μορφής $P_i \sim 2^{-k}$, και συνεπώς, το μέσο μήκος των κωδικολέξεων είναι μεγαλύτερο από την εντροπία

- Για δεδομένη διακριτή πηγή χωρίς μνήμη κατανομής **P** και εντροπίας $H(\mathscr{L})$, μπορεί να κατασκευαστεί βέλτιστος κώδικας με μέσω μήκος κωδικολέξεων

$$H(\mathscr{L}) \leq \overline{L} < H(\mathscr{L}) + 1$$

- Δηλαδή, το μέσο μήκος των κωδικολέξεων μπορεί να είναι έως 1 bit άνω του ελάχιστου δυνατού

Κωδικοποίηση Πηγής

- Προκειμένου να επιτύχουμε μέσο μήκος κωδικολέξεων πολύ κοντά στην εντροπία, βάσει των *n* συμβόλων της πηγής, σχηματίζουμε υπερσύμβολα μπλοκ $\mathscr{L}^N$ μήκους *N* συμβόλων

- Με $\mathbb{E}\langle \ell(s_1, s_2 \ldots, s_N)\rangle$ συμβολίζουμε το μέσο μήκος των κωδικολέξεων ανά υπερσύμβολο

- Για το νέο κώδικα το μέσο μήκος κωδικολέξεων ανά σύμβολο πηγής θα είναι

$$\overline{L}_N = \frac{1}{N}\mathbb{E}\left\langle \ell\left(s_1, s_2, \ldots, s_N\right)\right\rangle = \frac{1}{N}\sum_{s \in \mathscr{L}^N} p\left(s_1, s_2, \ldots, s_N\right)\ell\left(s_1, s_2, \ldots, s_N\right)$$

- Βάσει προηγούμενης ανισότητας μπορεί να κατασκευαστεί κώδικας μέσου μήκους κωδικολέξεων

$$H\left(\mathscr{L}^N\right) \le \mathbb{E}\left\langle \ell\left(s_1, s_2, \ldots, s_N\right)\right\rangle < H\left(\mathscr{L}^N\right) + 1$$

Κωδικοποίηση Πηγής

- Όπως έχουμε δει, για στατιστικά ανεξάρτητα σύμβολα πηγής ισχύει $H(\mathscr{L}^N) = N\,H(\mathscr{L})$ και άρα

$$N\,H\left(\mathscr{L}\right) \leq \mathbb{E}\left\langle \ell\left(s_1, s_2, \ldots, s_n\right)\right\rangle < N\,H\left(\mathscr{L}\right) + 1 \Leftrightarrow$$

$$H\left(\mathscr{L}\right) \leq \overline{L}_N < H\left(\mathscr{L}\right) + \frac{1}{N}$$

- Επιλέγοντας μακροσκελείς ακολουθίες συμβόλων $N \rightarrow \infty$, επιτυγχάνουμε

$$\lim_{N\rightarrow\infty} \overline{L}_N = H\left(\mathscr{L}\right)$$

- Δηλαδή, όταν οι πιθανότητες εμφάνισης των συμβόλων της πηγής δεν είναι της μορφής $P_i \sim 2^{-k}$, μπορούμε να σχηματίσουμε υπερσύμβολα μπλοκ μήκους N συμβόλων και τότε, το μέσο μήκος των κωδικολέξεων ανά σύμβολο θα βρίσκεται πολύ κοντά στην εντροπία της πηγής

- Επιπλέον, το παραπάνω συμπέρασμα ισχύει και για πηγή με στατιστικά εξαρτημένα σύμβολα

Κωδικοποίηση Πηγής

Κωδικοποίηση Shannon

- Διατάσσουμε τα σύμβολα $s_1, s_2, \ldots, s_n$ κατά φθίνουσα πιθανότητα $P_1 \geq P_2 \geq \cdots \geq P_n$

- Αντιστοιχίζουμε την παράμετρο F_i στο σύμβολο s_i (με $i = 1, 2, \ldots, n$), ξεκινώντας με $F_1 = 0$ και

$$F_i = \sum_{k=1}^{i-1} P_k$$

- Το μήκος της κωδικολέξης του s_i είναι ένας θετικός ακέραιος l_i που φράσσεται από την ανισότητα

$$-\log_2(P_i) \leq l_i < 1 - \log_2(P_i)$$

- Η κωδική λέξη για το σύμβολο s_i προκύπτει από τα πρώτα l_i υψηλότερης σημαντικότητας ψηφία (MSB) του κλασματικού μέρους του F_i, αφού πρώτα έχει αυτό μετατραπεί στο 2αδικό σύστημα

Κωδικοποίηση Πηγής

- Ο κώδικας Shannon είναι μεταβλητού μήκους και υποβέλτιστος ως προς το μέσο μήκος των κωδικολέξεων

- Το μέσο μήκος των κωδικολέξεων είναι

$$H(\mathscr{L}) \leq \overline{L} < H(\mathscr{L}) + 1$$

- Ο κώδικας που παράγεται είναι στιγμιαίος

- Βάσει της $-\log_2(P_i) \leq l_i < 1 - \log_2(P_i)$, σπάνια μηνύματα κωδικοποιούνται με μεγάλο μήκος

- Κάθε κωδική λέξη διαφέρει από τις υπόλοιπες σε ένα ή περισσότερα ψηφία

Κωδικοποίηση Πηγής

- <u>Παράδειγμα</u>: Διακριτή πηγή χωρίς μνήμη παράγει ανεξάρτητα σύμβολα με πιθανότητες {0.2, 0.1, 0.2, 0.4, 0.1}

- Να βρεθεί η εντροπία της πηγής και κωδικοποιηθούν τα σύμβολα κατά Shannon, όπου να βρεθεί το μέσο μήκος των bit ανά σύμβολο

- Η εντροπία της διακριτής πηγής είναι

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

$$= 0.2 \log_2\left(\frac{1}{0.2}\right) + 0.1 \log_2\left(\frac{1}{0.1}\right) + 0.2 \log_2\left(\frac{1}{0.2}\right) + 0.4 \log_2\left(\frac{1}{0.4}\right) + 0.1 \log_2\left(\frac{1}{0.1}\right) \text{ bit} =$$

$$= 2.12193 \text{ bit}$$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2				
3	s_3	0.2				
4	s_4	0.1				
5	s_5	0.1				

Μετατροπή F_1 στο 2αδικό

$0 \times 2 = \boxed{0} + 0$
$0 \times 2 = \boxed{0} + 0$
$0 \times 2 = \boxed{0} + 0$
$0 \times 2 = \boxed{0} + 0$

- Σύμβολο s_1 με πιθανότητα εμφάνισης $P_1 = 0.4$
  - Μήκος κωδικολέξης: $-\log_2(P_1) \leq l_1 \leq 1 - \log_2(P_1) \Leftrightarrow 1.32 \leq l_1 \leq 2.32 \rightarrow l_1 = 2$
  - Αναπαράσταση του $F_1 = (0.0)_{10}$ στο 2αδικό $F_1 = (0.0000)_2$
  - $l_1 = 2$ MSB του κλασματικού μέρους: 0000
  - Κωδική λέξη: $c(s_1) = 0\ \ 0$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2	3	0.4	0.0110	0 1 1
3	s_3	0.2				
4	s_4	0.1				
5	s_5	0.1				

- Σύμβολο s_2 με πιθανότητα εμφάνισης $P_2 = 0.2$
  - Μήκος κωδικολέξης: $-\log_2(P_2) \le l_2 \le 1 - \log_2(P_2) \Leftrightarrow 2.32 \le l_2 \le 3.32 \rightarrow l_2 = 3$
  - Αναπαράσταση του $F_2 = F_1 + P_1 = (0.4)_{10}$ στο 2αδικό $F_2 = (0.0110)_2$
  - $l_2 = 3$ MSB του κλασματικού μέρους: 0110
  - Κωδική λέξη: $c(s_2) =$ 0 1 1

Μετατροπή F_2 στο 2αδικό

$$0.4 \times 2 = \boxed{0} + 0.8$$
$$0.8 \times 2 = \boxed{1} + 0.6$$
$$0.6 \times 2 = \boxed{1} + 0.2$$
$$0.2 \times 2 = \boxed{0} + 0.4$$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2	3	0.4	0.0110	0 1 1
3	s_3	0.2	3	0.6	0.1001	1 0 0
4	s_4	0.1				
5	s_5	0.1				

- Σύμβολο s_3 με πιθανότητα εμφάνισης $P_3 = 0.2$
  - Μήκος κωδικολέξης: $-\log_2(P_3) \leq l_3 \leq 1 - \log_2(P_3) \Leftrightarrow 2.32 \leq l_3 \leq 3.32 \rightarrow l_3 = 3$
  - Αναπαράσταση του $F_3 = F_2 + P_2 = (0.6)_{10}$ στο 2αδικό $F_3 = (0.1001)_2$
  - $l_3 = 3$ MSB του κλασματικού μέρους: 1001
  - Κωδική λέξη: $c(s_3) = 1\ 0\ 0$

Μετατροπή F_3 στο 2αδικό

$$0.6 \times 2 = \boxed{1} + 0.2$$
$$0.2 \times 2 = \boxed{0} + 0.4$$
$$0.4 \times 2 = \boxed{0} + 0.8$$
$$0.8 \times 2 = \boxed{1} + 0.6$$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Τα πηγαία σύμβολα αναδιατάσσονται κατά φθίνουσα πιθανότητα

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2	3	0.4	0.0110	0 1 1
3	s_3	0.2	3	0.6	0.1001	1 0 0
4	s_4	0.1	4	0.8	0.11001	1 1 0 0
5	s_5	0.1				

- Σύμβολο s_4 με πιθανότητα εμφάνισης $P_4 = 0.1$
  - Μήκος κωδικολέξης: $-\log_2(P_4) \le l_4 \le 1 - \log_2(P_4) \Leftrightarrow 3.32 \le l_4 \le 4.32 \rightarrow l_4 = 4$
  - Αναπαράσταση του $F_4 = F_3 + P_3 = (0.8)_{10}$ στο 2αδικό $F_4 = (0.11001)_2$
  - $l_4 = 4$ MSB του κλασματικού μέρους: 11001
  - Κωδική λέξη: $c(s_4) = 1\ 1\ 0\ 0$

Μετατροπή F_4 στο 2αδικό

$$\begin{aligned} 0.8 \times 2 &= \boxed{1} + 0.6 \\ 0.6 \times 2 &= \boxed{1} + 0.2 \\ 0.2 \times 2 &= \boxed{0} + 0.4 \\ 0.4 \times 2 &= \boxed{0} + 0.8 \\ 0.8 \times 2 &= \boxed{1} + 0.6 \end{aligned}$$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2	3	0.4	0.0110	0 1 1
3	s_3	0.2	3	0.6	0.1001	1 0 0
4	s_4	0.1	4	0.8	0.11001	1 1 0 0
5	s_5	0.1	4	0.9	0.11100	1 1 1 0

- Σύμβολο s_5 με πιθανότητα εμφάνισης $P_5 = 0.1$
  - Μήκος κωδικολέξης: $-\log_2(P_5) \le l_5 \le 1 - \log_2(P_5) \Leftrightarrow 3.32 \le l_4 \le 4.32 \rightarrow l_5 = 4$
  - Αναπαράσταση του $F_5 = F_4 + P_4 = (0.9)_{10}$ στο 2αδικό $F_5 = (0.11100)_2$
  - $l_5 = 4$ MSB του κλασματικού μέρους: 11100
  - Κωδική λέξη: $c(s_5) = 1\ 1\ 1\ 0$

Μετατροπή F_5 στο 2αδικό

$0.9 \times 2 = \boxed{1} + 0.8$
$0.8 \times 2 = \boxed{1} + 0.6$
$0.6 \times 2 = \boxed{1} + 0.2$
$0.2 \times 2 = \boxed{0} + 0.4$
$0.4 \times 2 = \boxed{0} + 0.8$

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

Κωδικοποίηση Shannon						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**MSB Ψηφία** ℓ_i	**10αδικό** $(F_i)_{10}$	**2αδικό** $(F_i)_2$	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	2	0.0	0.0000	0 0
2	s_2	0.2	3	0.4	0.0110	0 1 1
3	s_3	0.2	3	0.6	0.1001	1 0 0
4	s_4	0.1	4	0.8	0.1100	1 1 0 0
5	s_5	0.1	4	0.9	0.1110	1 1 1 0

- Το μέσο μήκος των κωδικολέξεων του κώδικά Shannon είναι

$$\overline{L} = \sum_{i=1}^{5} P_i\,\ell_i = 0.4 \times 2 + 0.2 \times 3 + 0.2 \times 4 + 0.1 \times 4 + 0.1 \times 4 = 2.8\,\text{bit}$$

- Η απόδοση του κώδικα είναι

$$\eta = \frac{H(\mathscr{L})}{\overline{L}} = \frac{2.12}{2.8} \approx 75.8\%$$

- Το μέσο μήκος των κωδικολέξεων είναι περίπου κατά 0.68 bit μεγαλύτερο από βέλτιστο κώδικα

Κωδικοποίηση Πηγής

Κωδικοποίηση Fano

- Διατάσσουμε τα σύμβολα $s_1, s_2, \ldots, s_n$ κατά φθίνουσα πιθανότητα $P_1 \geq P_2 \geq \cdots \geq P_n$
- Διαχωρίζουμε τη λίστα σε 2 ομάδες με κοντινές τιμές πιθανότητας αθροιστικά ανά ομάδα
- Στην άνω ομάδα τοποθετούμε το 0 και στην κάτω το 1
- Επαναλαμβάνουμε την παραπάνω διαδικασία για τις νέες υποομάδες που δημιουργούνται
- Η διαδικασία σταματά όταν δεν υπάρχουν άλλες υποομάδες προς διαίρεση
- Διατρέχουμε τα bit από την αρχή προς το τέλος για να σχηματιστούν οι κωδικολέξεις

Κωδικοποίηση Πηγής

- Ο κώδικας Fano είναι μεταβλητού μήκους και υποβέλτιστος ως προς το μέσο μήκος των κωδικολέξεων

- Το μέσο μήκος των κωδικολέξεων είναι

$$H(\mathscr{L}) \leq \overline{L} < H(\mathscr{L}) + 1$$

- Ο κώδικας που παράγεται είναι στιγμιαίος

- Σπάνια μηνύματα κωδικοποιούνται με μεγάλο μήκος

- Κάθε κωδική λέξη διαφέρει από τις υπόλοιπες σε ένα ή περισσότερα ψηφία

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Τα πηγαία σύμβολα αναδιατάσσονται κατά φθίνουσα πιθανότητα

  - Διαχωρίζουμε τη λίστα σε 2 ομάδες με κοντινές τιμές αθροιστικής πιθανότητας ανά ομάδα
  - Στην άνω ομάδα τοποθετούμε το 0 και στην κάτω το 1
  - Επαναλαμβάνουμε την παραπάνω διαδικασία για τις νέες υποομάδες που δημιουργούνται

Κωδικοποίηση Fano

α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Διαίρεση**	
1	s_1	0.4	0.4	0
2	s_2	0.2	0.2	
3	s_3	0.2	0.2	1
4	s_4	0.1	0.1	
5	s_5	0.1	0.1	

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

  - Διαχωρίζουμε τη λίστα σε 2 ομάδες με κοντινές τιμές πιθανότητας αθροιστικά ανά ομάδα
  - Στην άνω ομάδα τοποθετούμε το 0 και στην κάτω το 1
  - Επαναλαμβάνουμε την παραπάνω διαδικασία για τις νέες υποομάδες που δημιουργούνται

Κωδικοποίηση Fano

α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Διαίρεση**		**2η Διαίρεση**	
1	s_1	0.4	0.4	0	0.4	0
2	s_2	0.2	0.2		0.2	1
3	s_3	0.2	0.2	1	0.2	0
4	s_4	0.1	0.1		0.1	1
5	s_5	0.1	0.1		0.1	

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

  - Διαχωρίζουμε τη λίστα σε 2 ομάδες με κοντινές τιμές πιθανότητας αθροιστικά ανά ομάδα
  - Στην άνω ομάδα τοποθετούμε το 0 και στην κάτω το 1
  - Η διαδικασία σταματά όταν δεν υπάρχουν άλλες υποομάδες προς διαίρεση

Κωδικοποίηση Fano

α/α i	Σύμβολο s_i	Πιθανότητα P_i	1η Διαίρεση		2η Διαίρεση		3η Διαίρεση	
1	s_1	0.4	0.4	0	0.4	0		
2	s_2	0.2	0.2		0.2	1		
3	s_3	0.2	0.2	1	0.2	0		
4	s_4	0.1	0.1		0.1	1	0.1	0
5	s_5	0.1	0.1		0.1		0.1	1

Κωδικοποίηση Πηγής

- <u>Παράδειγμα (*συνέχεια*)</u>:
- Το μέσο μήκος των κωδικολέξεων του κώδικά Shannon είναι

$$\overline{L} = \sum_{i=1}^{5} P_i \, \ell_i = 0.4 \times 2 + 0.2 \times 2 + 0.2 \times 2 + 0.1 \times 3 + 0.1 \times 3 = 2.2\,\text{bit}$$

- Η απόδοση του κώδικα είναι

$$\eta = \frac{H(\mathscr{L})}{\overline{L}} = \frac{2.12}{2.2} \approx 96.4\%$$

- Το μέσο μήκος των κωδικολέξεων διαφέρει λιγότερο από 0.1 bit από την εντροπία της πηγής

Κωδικοποίηση Fano

α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Διαίρεση**	**2η Διαίρεση**	**3η Διαίρεση**	**Κωδικολέξη** $c(s_i)$	**Μήκος Κωδικολέξης** l_i
1	s_1	0.4	0.4 0	0.4 0		0 0	2
2	s_2	0.2	0.2	0.2 1		0 1	2
3	s_3	0.2	0.2	0.2 0		1 0	2
4	s_4	0.1	0.1 1	0.1 1	0.1 0	1 1 0	3
5	s_5	0.1	0.1	0.1	0.1 1	1 1 1	3

Κωδικοποίηση Πηγής

Κωδικοποίηση Huffman

- Διατάσσουμε τα σύμβολα $s_1, s_2, \ldots, s_n$ κατά φθίνουσα πιθανότητα $P_1 \geq P_2 \geq \cdots \geq P_n$
- Στα 2 χαμηλότερης πιθανότητας σύμβολα αντιστοιχούμε το 0 και 1
- Τα 2 χαμηλότερης πιθανότητας σύμβολα συνδυάζονται σε ένα νέο με πιθανότητα το άθροισμα των πιθανοτήτων και ταξινομείται στην κατάλληλη θέση της λίστας
- Η παραπάνω διαδικασία επαναλαμβάνεται έως να μείνουν μόνο 2 νούμερα στη λίστα, στα οποία αντιστοιχούμε το 0 και 1
- Τέλος, διατρέχουμε από το τέλος προς την αρχή για να βρούμε την κωδικολέξη κάθε συμβόλου

Κωδικοποίηση Πηγής

- Ο κώδικας Huffman είναι μεταβλητού μήκους και βέλτιστος ως προς το μέσο μήκος των κωδικολέξεων

- Το μέσο μήκος των κωδικολέξεων είναι

$$H(\mathscr{L}) \leq \overline{L} < H(\mathscr{L}) + 1$$

- Ο κώδικας που παράγεται είναι στιγμιαίος

Κωδικοποίηση Πηγής

- Παράδειγμα: Διακριτή πηγή χωρίς μνήμη παράγει ανεξάρτητα σύμβολα με πιθανότητες {0.2, 0.1, 0.2, 0.4, 0.1}

- Να βρεθεί η εντροπία της πηγής και κωδικοποιηθούν τα σύμβολα κατά Huffman, όπου να βρεθεί το μέσο μήκος των bit ανά σύμβολο

Κωδικοποίηση Huffman						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i				
1	s_1	0.4				
2	s_2	0.2				
3	s_3	0.2				
4	s_4	0.1				
5	s_5	0.1				

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):
  - Τα 2 χαμηλότερης πιθανότητας σύμβολα συνδυάζονται σε ένα νέο με πιθανότητα το άθροισμα των πιθανοτήτων και ταξινομείται στην κατάλληλη θέση της λίστας
  - Η παραπάνω διαδικασία επαναλαμβάνεται έως να μείνουν μόνο 2 νούμερα στη λίστα, στα οποία αντιστοιχούμε το 0 και 1

Κωδικοποίηση Huffman

α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Αναδιάταξη**			
1	s_1	0.4	0.4			
2	s_2	0.2	0.2			
3	s_3	0.2	0.2			
4	s_4	**0.1**	0.2			
5	s_5	**0.1**				

(Η τοποθέτηση των νέων πιθανοτήτων τοποθετείται στην κορυφή της λίστας μεταξύ των συμβόλων ίσης πιθανότητας)

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):
  - Τα 2 χαμηλότερης πιθανότητας σύμβολα συνδυάζονται σε ένα νέο με πιθανότητα το άθροισμα των πιθανοτήτων και ταξινομείται στην κατάλληλη θέση της λίστας
  - Η παραπάνω διαδικασία επαναλαμβάνεται έως να μείνουν μόνο 2 νούμερα στη λίστα, στα οποία αντιστοιχούμε το 0 και 1

Κωδικοποίηση Huffman

α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Αναδιάταξη**	**2η Αναδιάταξη**		
1	s_1	0.4	0.4	0.4		
2	s_2	0.2	0.2	0.4		
3	s_3	0.2	**0.2**	0.2		
4	s_4	0.1	**0.2**			
5	s_5	0.1				

(Η τοποθέτηση των νέων πιθανοτήτων τοποθετείται στην κορυφή της λίστας μεταξύ των συμβόλων ίσης πιθανότητας)

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):
  - Τα 2 χαμηλότερης πιθανότητας σύμβολα συνδυάζονται σε ένα νέο με πιθανότητα το άθροισμα των πιθανοτήτων και ταξινομείται στην κατάλληλη θέση της λίστας
  - Η παραπάνω διαδικασία επαναλαμβάνεται έως να μείνουν μόνο 2 νούμερα στη λίστα

Κωδικοποίηση Huffman

α/α i	Σύμβολο s_i	Πιθανότητα P_i	1η Αναδιάταξη	2η Αναδιάταξη	3η Αναδιάταξη	
1	s_1	0.4	0.4	0.4	0.6	
2	s_2	0.2	0.2	**0.4**	0.4	
3	s_3	0.2	0.2	**0.2**		
4	s_4	0.1	0.2			
5	s_5	0.1				

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):
  - Τέλος, διατρέχουμε από το τέλος προς την αρχή για να βρούμε την κωδικολέξη κάθε συμβόλου

Κωδικοποίηση Huffman						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Αναδιάταξη**	**2η Αναδιάταξη**	**3η Αναδιάταξη**	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	0.4	0.4	0.6 0	
2	s_2	0.2	0.2	0.4 0	0.4 1	
3	s_3	0.2	0.2 0	0.2 1		
4	s_4	0.1 0	0.2 1			
5	s_5	0.1 1				

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):
  - Τέλος, διατρέχουμε από το τέλος προς την αρχή για να βρούμε την κωδικολέξη κάθε συμβόλου

Κωδικοποίηση Huffman						
α/α i	**Σύμβολο** s_i	**Πιθανότητα** P_i	**1η Αναδιάταξη**	**2η Αναδιάταξη**	**3η Αναδιάταξη**	**Κωδική Λέξη** $c(s_i)$
1	s_1	0.4	0.4	0.4	0.6 0	0 0
2	s_2	0.2	0.2	0.4 0	0.4 1	1 0
3	s_3	0.2	0.2 0	0.2 1		1 1
4	s_4	0.1 0	0.2 1			0 1 0
5	s_5	0.1 1				0 1 1

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Το μέσο μήκος των κωδικολέξεων του κώδικά Huffman είναι

$$\overline{L} = \sum_{i=1}^{5} P_i \, \ell_i = 0.4 \times 2 + 0.2 \times 2 + 0.2 \times 2 + 0.1 \times 3 + 0.1 \times 3 = 2.2\,\text{bit}$$

- Η απόδοση του κώδικα είναι

$$\eta = \frac{H\left(\mathscr{L}\right)}{\overline{L}} = \frac{2.12}{2.2} \approx 96.4\%$$

- Το μέσο μήκος των κωδικολέξεων διαφέρει λιγότερο από 0.1 bit από την εντροπία της πηγής

- Ο κώδικας Huffman είναι βέλτιστος ως προς το μέσο μήκος των κωδικολέξεων

Κωδικοποίηση Πηγής

Κωδικοποίηση Huffman

(*Χρήση Κωδικού Δέντρου*)

- Διατάσσουμε τα σύμβολα $s_1, s_2, \ldots, s_n$ κατά φθίνουσα πιθανότητα $P_1 \geq P_2 \geq \cdots \geq P_n$
- Για κάθε σύμβολο δημιουργούμε έναν κόμβο φύλλο
- Από τους δύο χαμηλότερης πιθανότητας κόμβους φύλα δημιουργούμε ένα νέο, με πιθανότητα το άθροισμα των πιθανοτήτων των προηγούμενων δύο και αναδιατάσσουμε
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να φτάσουμε στη ρίζα του δέντρου
- Σε κάθε διακλάδωση του δέντρου τοποθετούμε αριστερά το 0 και δεξιά το 1
- Διατρέχουμε το δέντρο από την ρίζα προς τα φύλα και σχηματίζουμε τις κωδικολέξεις

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Διακριτή πηγή χωρίς μνήμη παράγει ανεξάρτητα σύμβολα με πιθανότητες {0.2, 0.1, 0.2, 0.4, 0.1}

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1

- Να βρεθεί η εντροπία της πηγής και κωδικοποιηθούν τα σύμβολα κατά Huffman, όπου να βρεθεί το μέσο μήκος των bit ανά σύμβολο

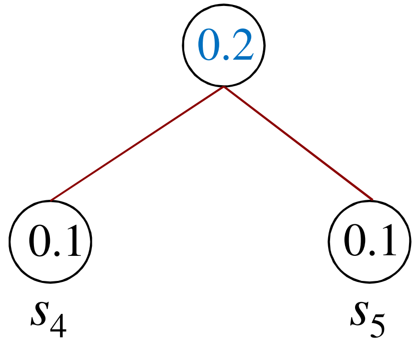
- Να γίνει επίλυση με χρήση του δέντρου του κώδικα

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1

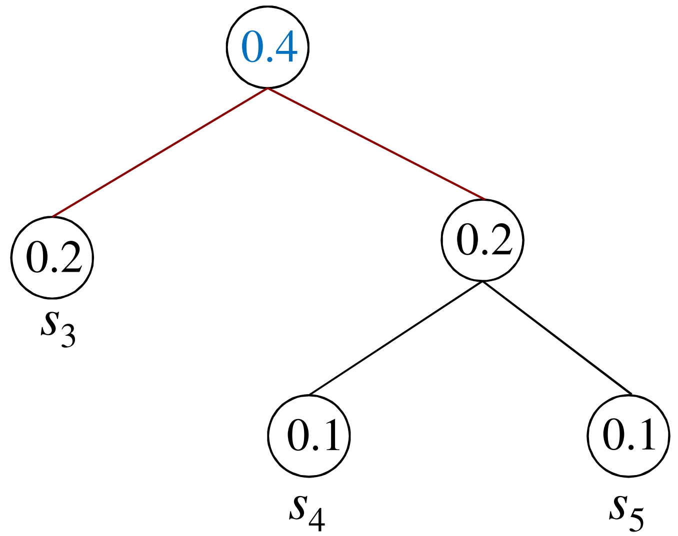
- Για κάθε σύμβολο δημιουργούμε έναν κόμβο φύλλο
- Από τους δύο χαμηλότερης πιθανότητας κόμβους φύλα δημιουργούμε ένα νέο, με πιθανότητα το άθροισμα των πιθανοτήτων των προηγούμενων δύο και αναδιατάσσουμε



Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1

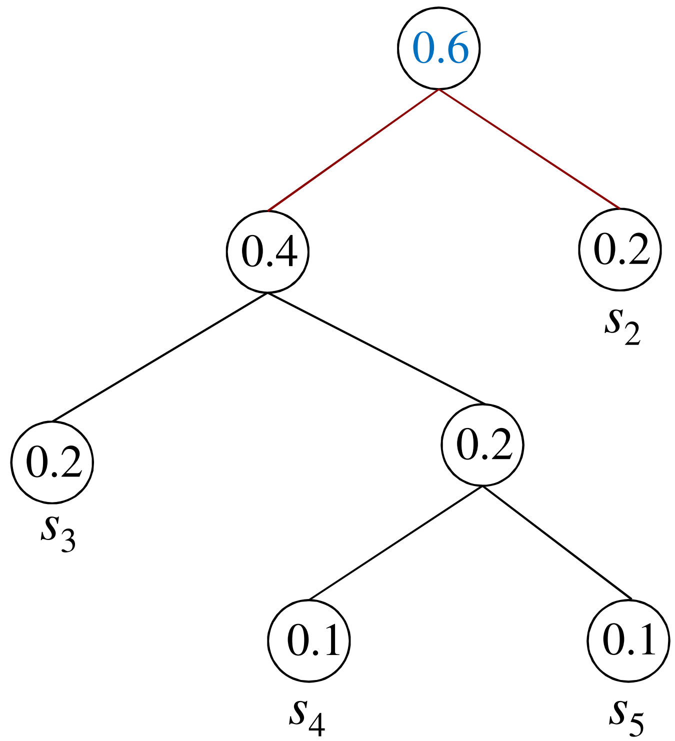


- Από τους δύο χαμηλότερης πιθανότητας κόμβους φύλα δημιουργούμε ένα νέο, με πιθανότητα το άθροισμα των πιθανοτήτων των προηγούμενων δύο και αναδιατάσσουμε
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να φτάσουμε στη ρίζα του δέντρου

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1

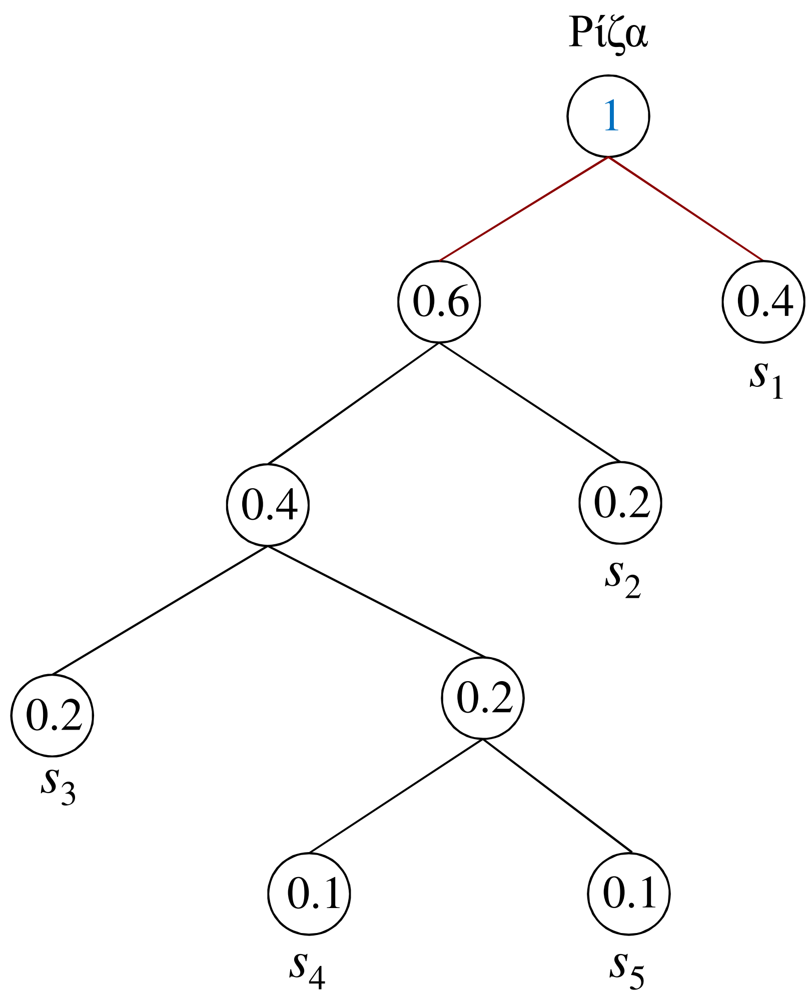


- Από τους δύο χαμηλότερης πιθανότητας κόμβους φύλα δημιουργούμε ένα νέο, με πιθανότητα το άθροισμα των πιθανοτήτων των προηγούμενων δύο και αναδιατάσσουμε
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να φτάσουμε στη ρίζα του δέντρου

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1

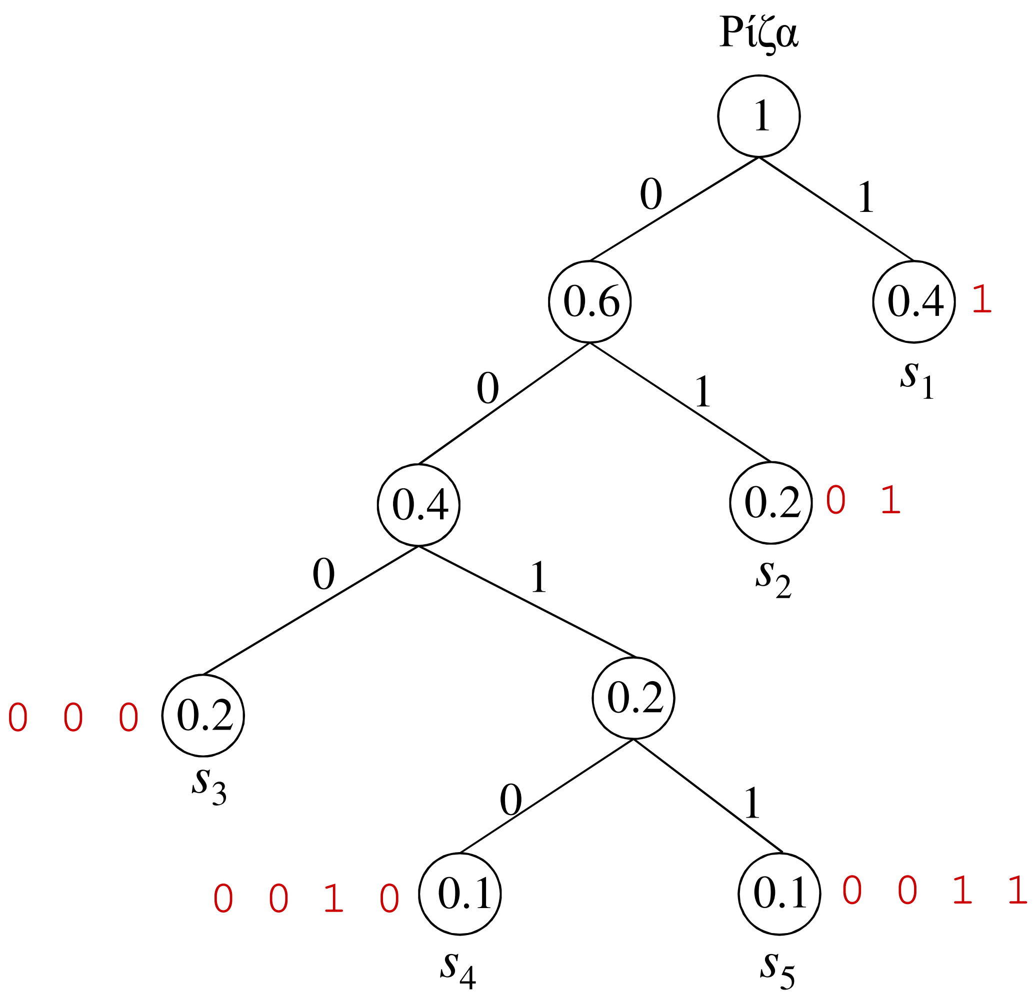


- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να φτάσουμε στη ρίζα του δέντρου

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*):

s_1	s_2	s_3	s_4	s_5
0.4	0.2	0.2	0.1	0.1



- Σε κάθε διακλάδωση του δέντρου τοποθετούμε αριστερά το 0 και δεξιά το 1
- Διατρέχουμε το δέντρο από τη ρίζα προς τα φύλα και σχηματίζουμε τις κωδικολέξεις

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Το μέσο μήκος των κωδικολέξεων του κώδικά Huffman στο παράδειγμα που προηγήθηκε είναι

$$\overline{L} = \sum_{i=1}^{5} P_i \, \ell_i = 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 0.1 \times 4 + 0.1 \times 4 = 2.2\,\text{bit}$$

- Η απόδοση του κώδικα είναι

$$\eta = \frac{H(\mathscr{L})}{\overline{L}} = \frac{2.12}{2.2} \approx 96.4\%$$

- Το μέσο μήκος των κωδικολέξεων διαφέρει λιγότερο από 0.1 bit από την εντροπία της πηγής

- Ο κώδικας που παράχθηκε είναι στιγμιαίος και διαφορετικός σε σχέση με την προηγούμενη επίλυση, αλλά η απόδοση του κώδικά είναι ακριβώς ίδια, δηλαδή οι δύο τρόποι είναι ισοδύναμοι

Κωδικοποίηση Πηγής

- Παράδειγμα: Διακριτή πηγή χωρίς μνήμη παράγει ανεξάρτητα σύμβολα $\mathscr{L} = \{s_1, s_2, s_3\}$ με πιθανότητες εμφάνισης {0.4, 0.4, 0.2}, αντίστοιχα

- Να βρεθεί η εντροπία της πηγής και κωδικοποιηθούν τα σύμβολα κατά Huffman, όπου να βρεθεί το μέσο μήκος των bit ανά σύμβολο

- Στη συνέχεια, να σχηματιστούν υπερσύμβολα αποτελούμενα από 2 σύμβολα

- Να κωδικοποιηθούν τα υπερσύμβολα κατά Huffman και να βρεθεί το μέσο μήκος των bit ανά σύμβολο

Κωδικοποίηση Πηγής

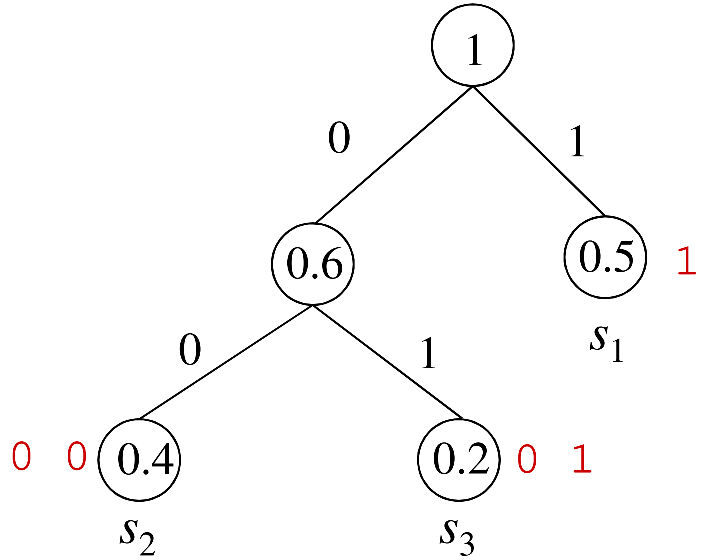
- <u>Παράδειγμα (*συνέχεια*)</u>: Η εντροπία της διακριτής πηγής χωρίς μνήμη είναι {0.4, 0.4, 0.2}

$$H(\mathscr{L}) = \sum_{i=1}^{3} P_i \log_2\left(\frac{1}{P_i}\right) = 0.4\log_2\left(\frac{1}{0.4}\right) + 0.4\log_2\left(\frac{1}{0.4}\right) + 0.2\log_2\left(\frac{1}{0.2}\right) \text{ bit} = 1.5219 \text{ bit}$$

s_1	s_2	s_3
0.4	0.4	0.2

- Το μέσο μήκος των κωδικολέξεων του κώδικά Huffman είναι

$$\overline{L} = \sum_{i=1}^{3} P_i \ell_i = (0.4\times 1 + 0.4\times 2 + 0.2\times 2) \text{ bit} = 1.6 \text{ bit}$$



- Η απόδοση του κώδικα είναι

$$\eta = \frac{H(\mathscr{L})}{\overline{L}} = \frac{1.5219}{1.6} \approx 95.121\%$$

- Το μέσο μήκος των κωδικολέξεων διαφέρει λιγότερο από 0.1 bit από την εντροπία της πηγής

Κωδικοποίηση Πηγής

- Παράδειγμα (*συνέχεια*): Κατασκευάζουμε υπερσύμβολα σε μπλοκ καθένα αποτελούμενο από δύο σύμβολα της διακριτής πηγής (πλήθος $3^2 = 9$)

- Το σύνολο των υπερσυμβόλων είναι $\mathscr{L}^2 = \{s_1 s_1, s_1 s_2, s_1 s_3, s_2 s_1, s_2 s_2, s_2 s_3, s_3 s_1, s_3 s_2, s_3 s_3\}$

- Δεδομένης της ανεξαρτησία μεταξύ των s_1, s_2, s_3, οι πιθανότητες εμφάνισης των υπερσυμβόλων θα είναι το γινόμενο των πιθανοτήτων εμφάνισης των συμβόλων που τα σχηματίζουν

- Συνεπώς, τα υπερσύμβολα χαρακτηρίζονται από τις παρακάτω πιθανότητες

Υπερσύμβολα	σ_1	σ_2	σ_5	σ_3	σ_4	σ_6	σ_7	σ_8	σ_9
	$s_1 s_1$	$s_1 s_2$	$s_1 s_3$	$s_2 s_1$	$s_2 s_2$	$s_2 s_3$	$s_3 s_1$	$s_3 s_2$	$s_3 s_3$
Πιθανότητες	$P_1 P_1$	$P_1 P_2$	$P_1 P_3$	$P_2 P_1$	$P_2 P_1$	$P_2 P_3$	$P_3 P_1$	$P_3 P_2$	$P_3 P_3$
	0.16	0.16	0.08	0.16	0.16	0.08	0.08	0.08	0.04

Κωδικοποίηση Πηγής

- Όπως αναμένεται, η εντροπία των υπερσυμβόλων θα είναι διπλάσια της εντροπίας της πηγής

- Η εντροπία μπορεί να υπολογιστεί ως

$$H\left(\mathscr{L}^2\right) = \sum_{i=1}^{9} P_i \log_2\left(\frac{1}{P_i}\right) =$$

$$= 4\times 0.16 \log_2\left(\frac{1}{0.16}\right) + 4\times 0.08 \log_2\left(\frac{1}{0.08}\right) + 0.04 \log_2\left(\frac{1}{0.04}\right) \text{ bit} = 3.0439 \text{ bit}$$

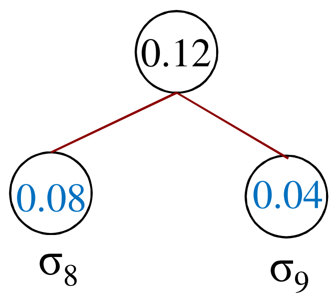
- Δηλαδή, όντως, επαληθεύεται η ιδιότητα $H(\mathscr{L}^2) = 2\, H(\mathscr{L})$

Υπερσύμβολα	σ_1	σ_2	σ_5	σ_3	σ_4	σ_6	σ_7	σ_8	σ_9
	$s_1 s_1$	$s_1 s_2$	$s_1 s_3$	$s_2 s_1$	$s_2 s_2$	$s_2 s_3$	$s_3 s_1$	$s_3 s_2$	$s_3 s_3$
Πιθανότητες	$P_1 P_1$	$P_1 P_2$	$P_1 P_3$	$P_2 P_1$	$P_2 P_1$	$P_2 P_3$	$P_3 P_1$	$P_3 P_2$	$P_3 P_3$
	0.16	0.16	0.08	0.16	0.16	0.08	0.08	0.08	0.04

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

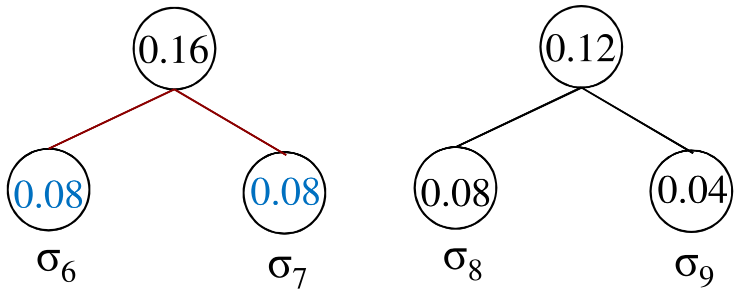


- Ο πρώτος κόμβος σχηματίζεται από τα 2 πρώτα σύμβολα με τις μικρότερες πιθανότητες

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

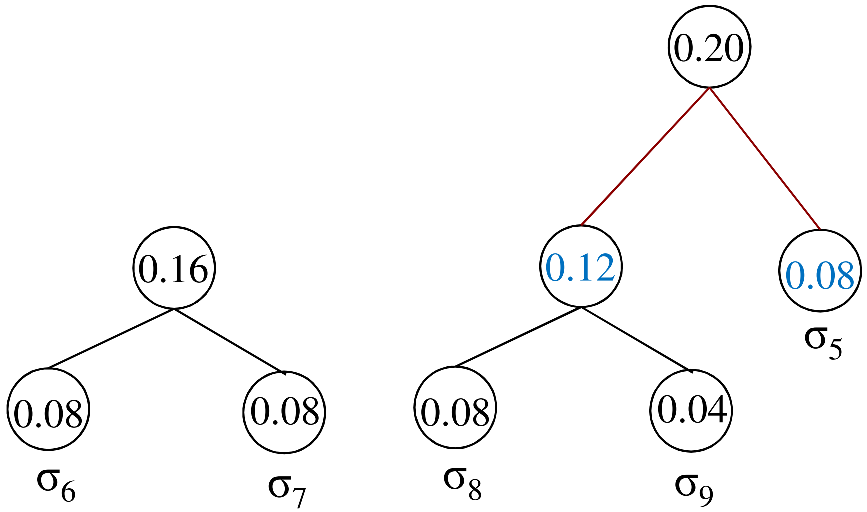


- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

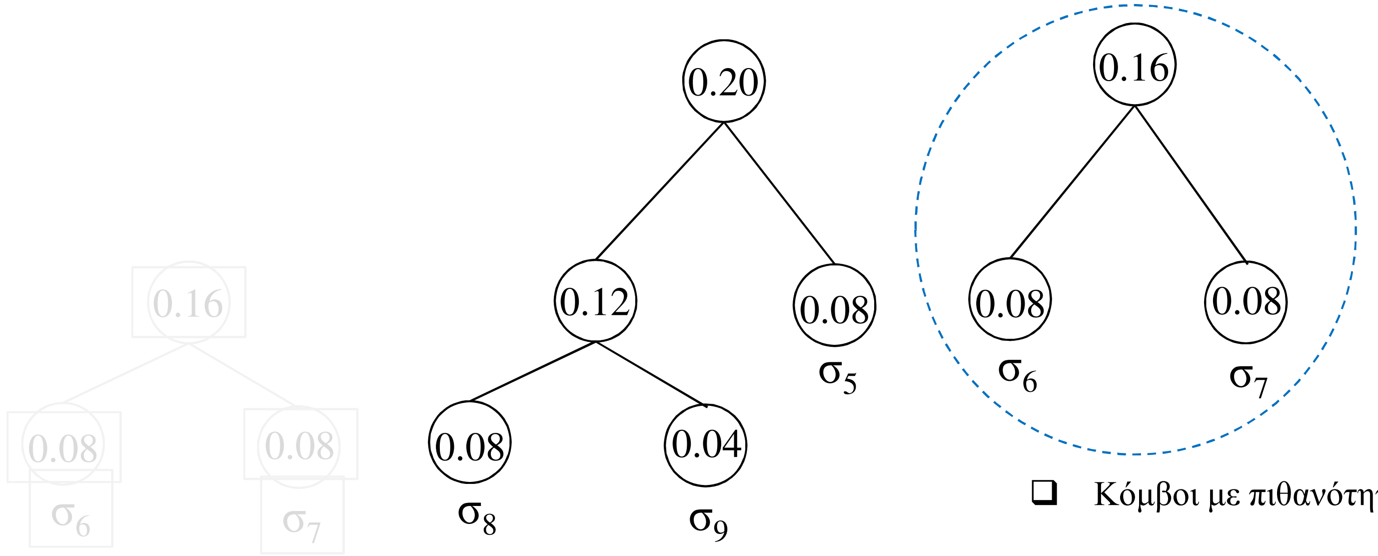


- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

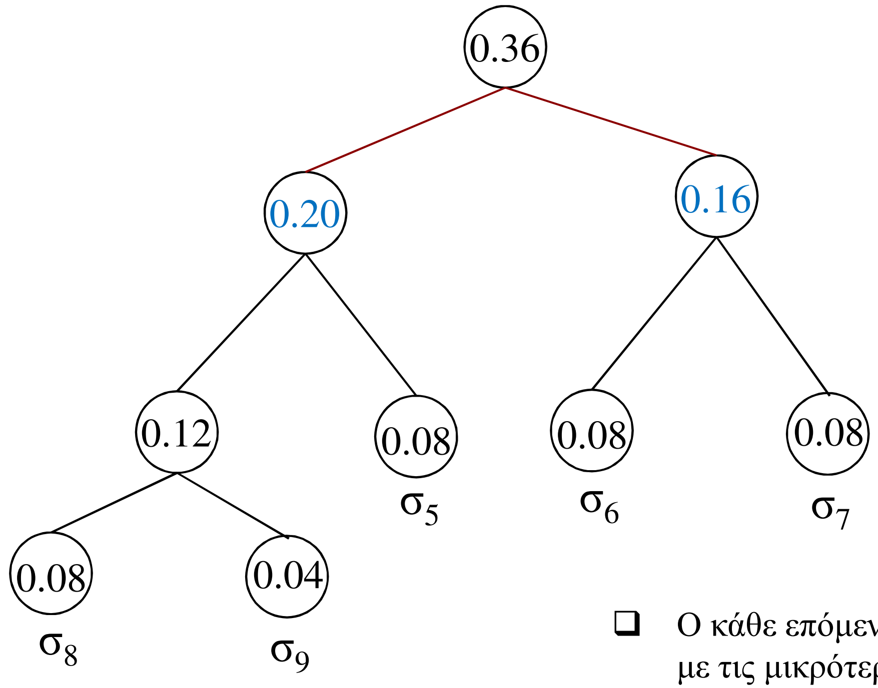


- Κόμβοι με πιθανότητες μικρότερες ή ίσες τοποθετούνται δεξιά

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

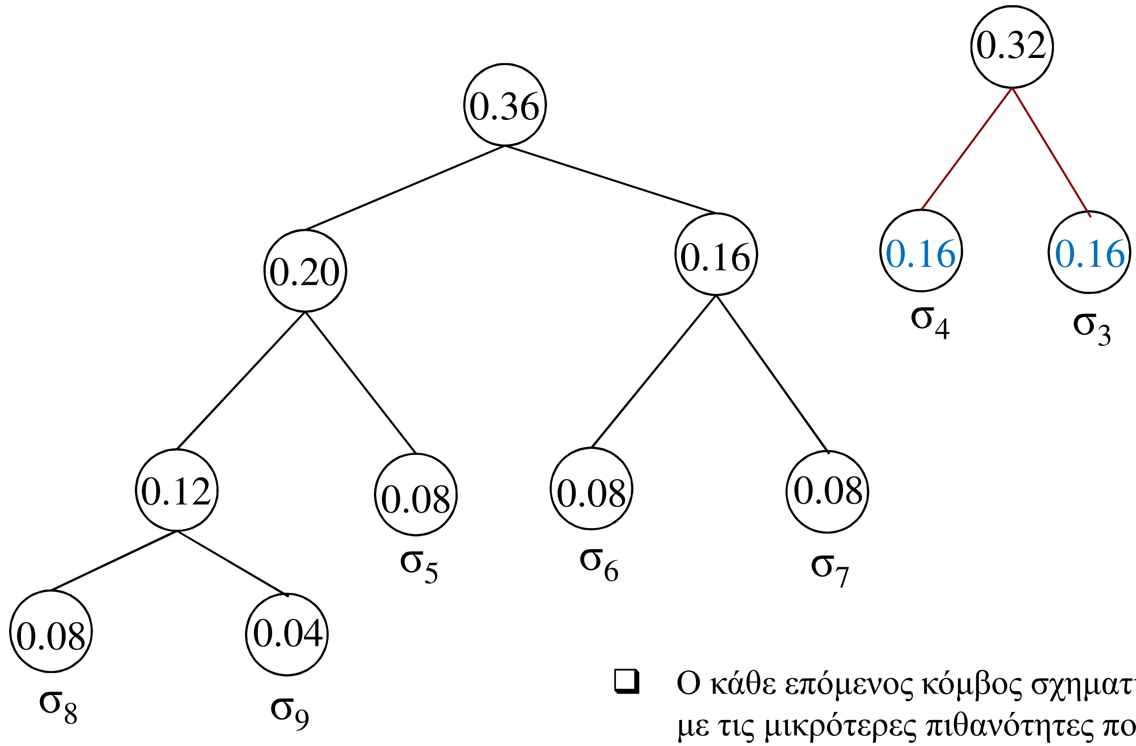


- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04

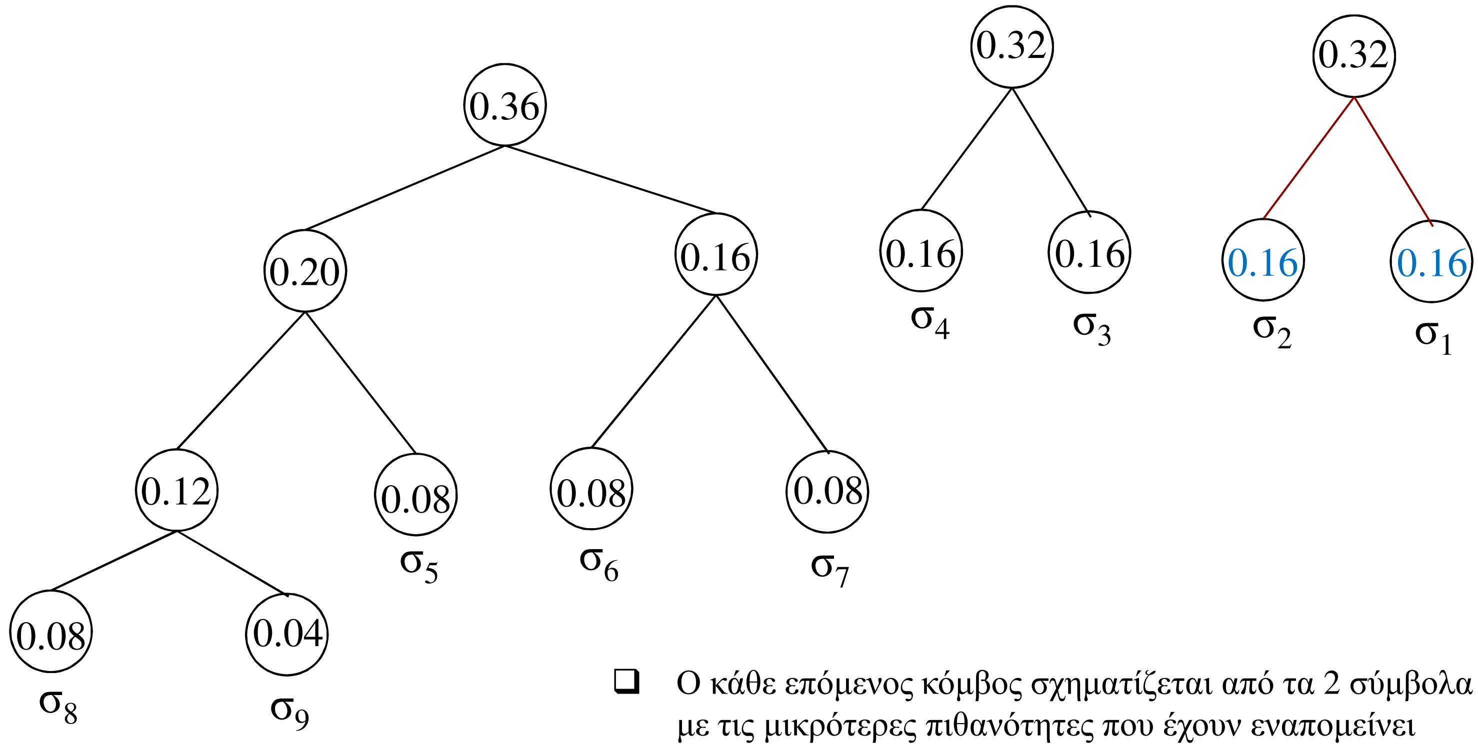


- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman

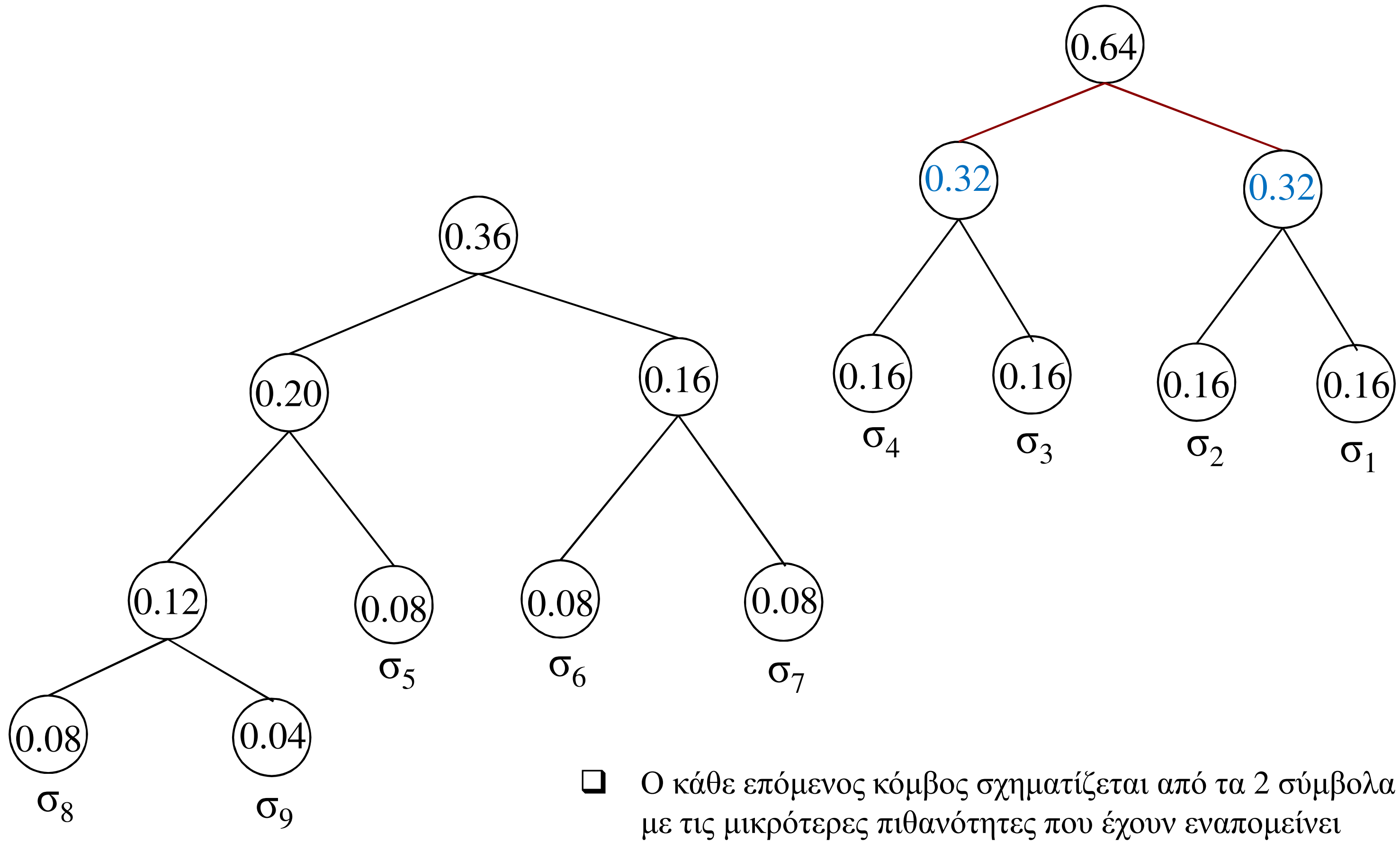
Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04



- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

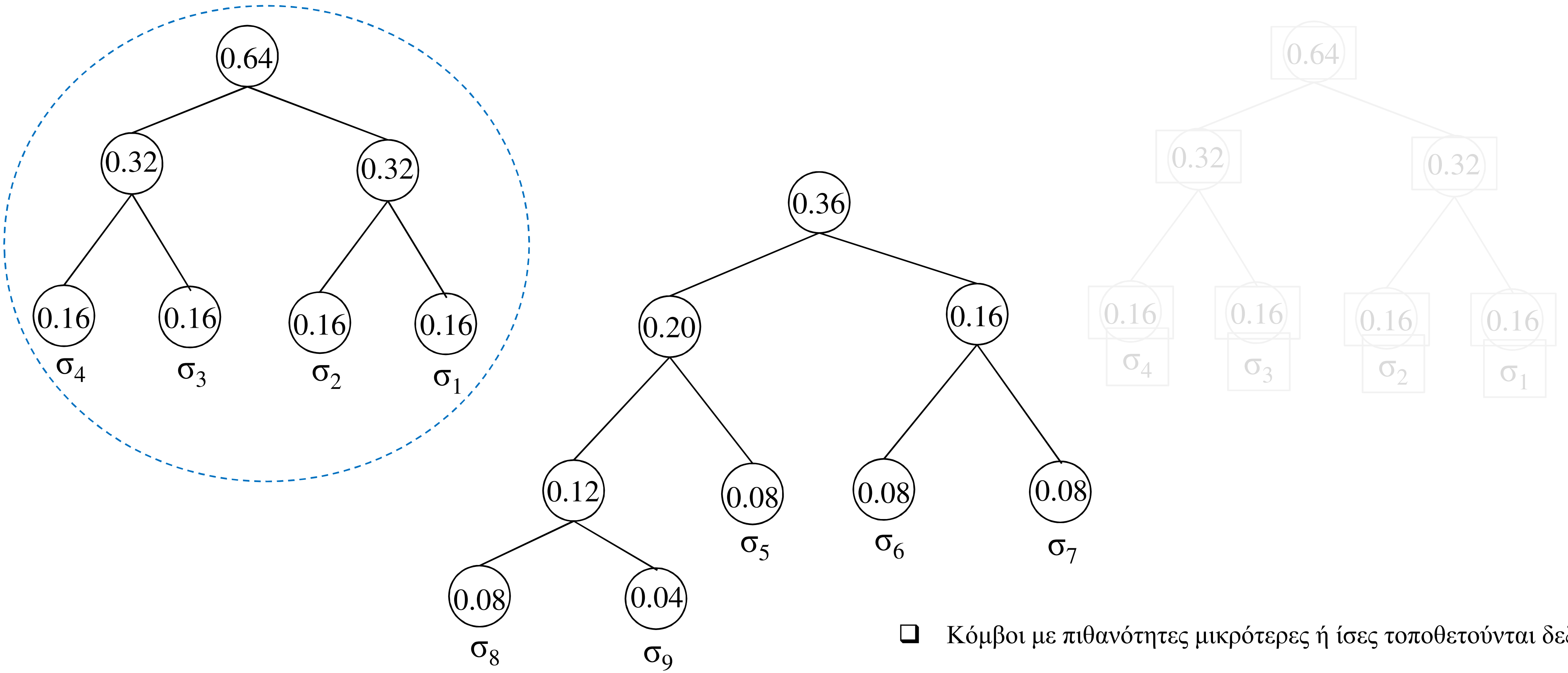
- Δέντρο Κώδικα Huffman



- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

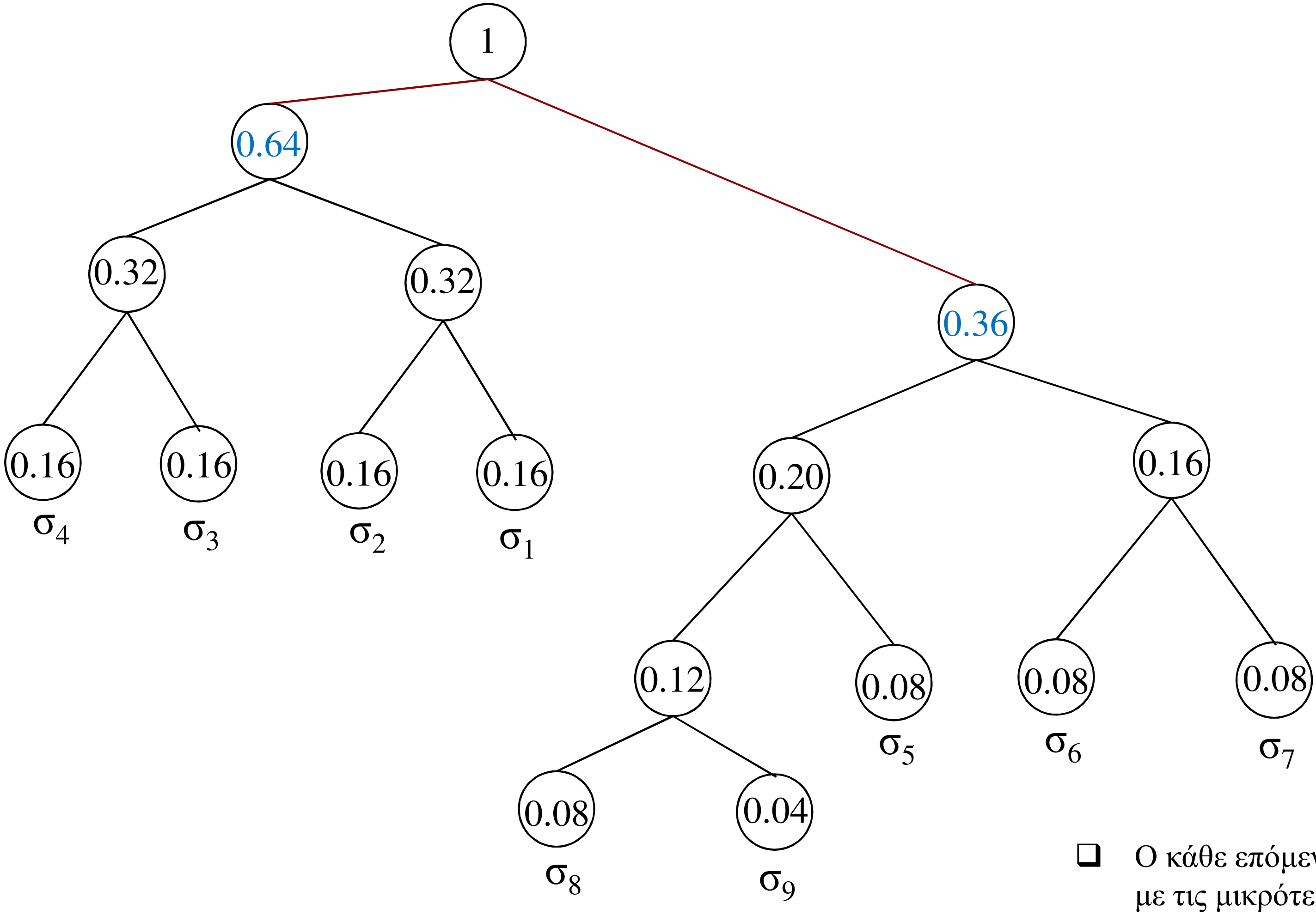
- Δέντρο Κώδικα Huffman



- Κόμβοι με πιθανότητες μικρότερες ή ίσες τοποθετούνται δεξιά

Κωδικοποίηση Πηγής

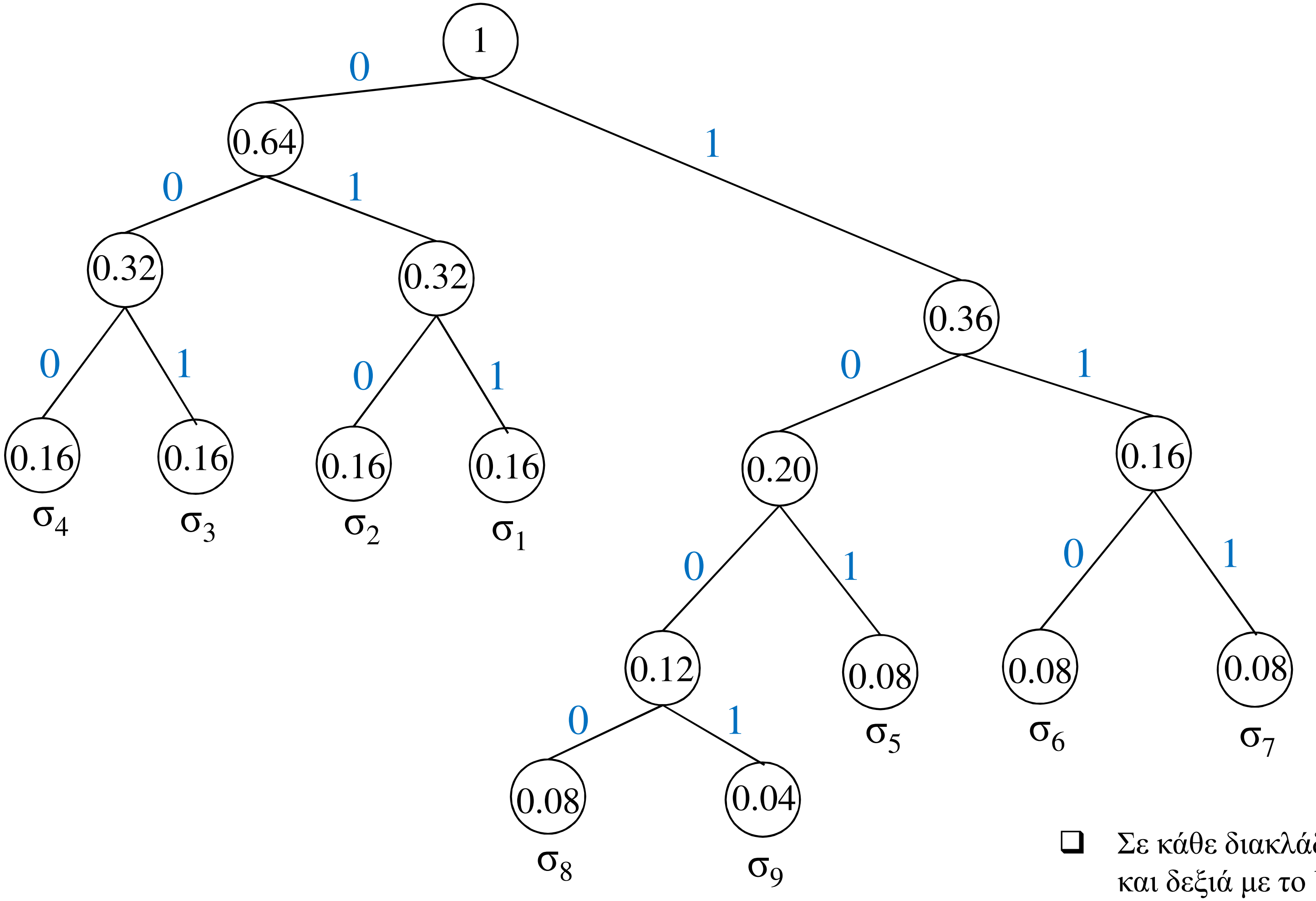
- Δέντρο Κώδικα Huffman



- Ο κάθε επόμενος κόμβος σχηματίζεται από τα 2 σύμβολα με τις μικρότερες πιθανότητες που έχουν εναπομείνει

Κωδικοποίηση Πηγής

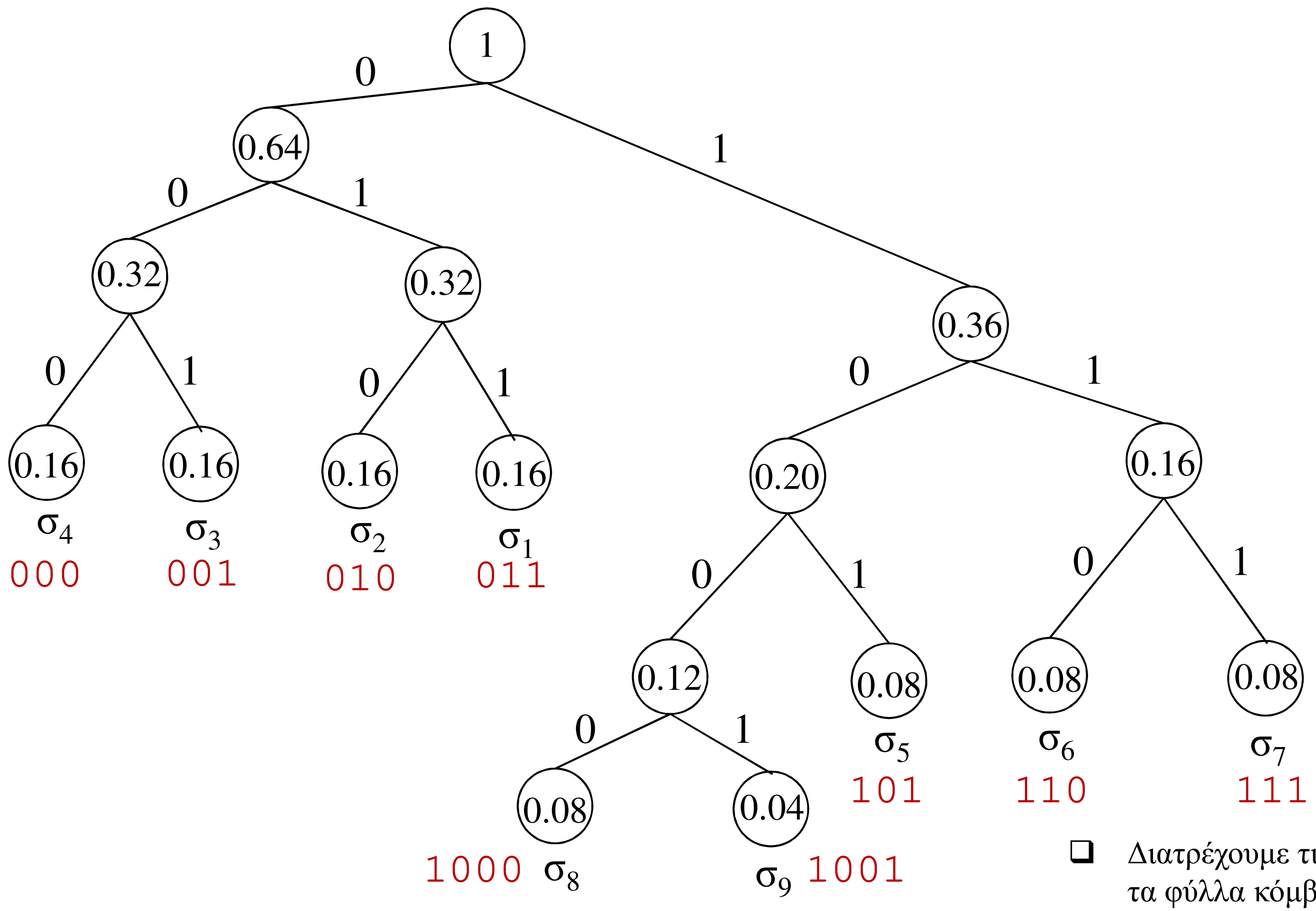
- Δέντρο Κώδικα Huffman



- Σε κάθε διακλάδωση αριστερά σημειώνουμε με το bit 0 και δεξιά με το bit 1

Κωδικοποίηση Πηγής

- Δέντρο Κώδικα Huffman



- Διατρέχουμε τις διαδρομές από τη ρίζα του δέντρου προς τα φύλλα κόμβους και βρίσκουμε τις κωδικολέξεις

Κωδικοποίηση Πηγής

- Με βάση τις πιθανότητες και τα μήκη των κωδικολέξεων, τα οποία συνοψίζονται στον παρακάτω πίνακα, μπορούμε να υπολογίσουμε το μέσο μήκος ανά υπερσύμβολο

Υπερσύμβολα	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Πιθανότητες P_i	0.16	0.16	0.16	0.16	0.08	0.08	0.08	0.08	0.04
Κωδικολέξεις $c(\sigma_i)$	011	010	001	000	101	110	111	1000	1001
Μήκος l_i	3	3	3	3	3	3	3	4	4

- Το μέσο μήκος σε bit των κωδικολέξεων ανά υπερσύμβολο είναι

$$\mathbb{E}\left\langle \ell\left(s_1, s_2\right)\right\rangle = \sum_{i=1}^{9} P_i\, \ell_i = \left(4 \times 0.16 \times 3 + 4 \times 0.08 \times 3 + 0.04 \times 1\right) \text{ bit} = 3.12 \text{ bit}$$

Κωδικοποίηση Πηγής

- Δηλαδή, δεδομένου ότι κάθε υπερσύμβολο περιέχει 2 σύμβολα, το μέσο μήκος των κωδικολέξεων ανά σύμβολο είναι

$$\overline{L}_2 = \frac{1}{2}\,\mathbb{E}\left\langle \ell\left(s_1, s_2\right)\right\rangle = 1.56 \text{ bit}$$

- Η εντροπία της πηγής έχει υπολογιστεί $H(\mathscr{L}) = 1.5219$ bit και το μέσο μήκος των κωδικολέξεων, εφαρμόζοντας κωδικοποίηση Huffman στα σύμβολα της πηγής, έχει βρεθεί 1.6 bit

- Συνεπώς, δημιουργώντας μπλοκ των 2 συμβόλων, βελτιώθηκε το μέσο μήκος των κωδικολέξεων κατά 0.04 bit