ΘΕΩΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

Κεφάλαιο 3 : Πηγές Πληροφορίας

Διάλεξη: Κώστας Μαλιάτσος

Πανεπιστήμιο Αιγαίου, Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Περιεχόμενα

- Διακριτές Πηγές Πληροφορίας χωρίς μνήμη
  - Ποσότητα πληροφορίας της πηγής
  - Κωδικοποίηση πηγής
  - Αλγόριθμοι κωδικοποίησης
- Διακριτές Πηγές Πληροφορίας με μνήμη
  - Πηγές Markoff
  - Εντροπία των πηγών Markoff
  - Ζητήματα κωδικοποίησης των πηγών Markoff

Πηγές Πληροφορίας

- Η έξοδος της πηγής είναι
  - κάτι τυχαίο και άγνωστο
  - μια τυχαία διαδικασία
- Αν είναι κάτι σταθερό ή ντετερμινιστικό, δεν υπάρχει λόγος να το μεταδώσουμε…
- Παραδείγματα πηγών πληροφορίας:
  - Ήχος, ομιλία, εικόνα, video
  - Bits, χαρακτήρες ASCII
- Διάκριση ως προς το χρόνο:
  - συνεχούς χρόνου (π.χ. αναλογικό ηχητικό σήμα)
  - διακριτού χρόνου (δειγματοληπτημένο σήμα, bits)
- Διάκριση ως προς τις δυνατές τιμές (αλφάβητο):
  - συνεχείς τιμές (π.χ. αναλογικό σήμα)
  - διακριτές τιμές (π.χ. ASCII)

Πηγές Πληροφορίας

- Μετατροπή πηγής από συνεχή σε διακριτού χρόνου
  - δειγματοληψία
  - το σήμα πρέπει να έχει πεπερασμένο εύρος ζώνης
  - αν είναι κατωπερατό με μέγιστη συχνότητα f_{max}, τότε η συνθήκη Nyquist μας λέει ότι αρκεί να το δειγματοληπτήσω* με

$$\boxed{f_s \geq 2 f_{\max}}$$

  - τότε μπορώ να ανακατασκευάσω το αναλογικό σήμα από τα δείγματα χωρίς απώλειες
- Οι πηγές που μας ενδιαφέρουν,
  - έχουν περιορισμένο εύρος ζώνης
  - ή μπορούμε να το περιορίσουμε εμείς με φιλτράρισμα
- Συμπέρασμα: αρκεί να μελετήσω τις πηγές διακριτού χρόνου

*Η συγκεκριμένη έκφανση του θ. δειγματοληψίας είναι ακριβής για πραγματικά, βαθυπερατά σήματα. Γενικευμένη μορφή του παρουσιάζεται σε μετέπειτα διαλέξεις

Διακριτή Πηγή Χωρίς Μνήμη

- **Discrete Memoryless Source (DMS):**
  - διακριτού χρόνου
  - διακριτού αλφαβήτου
  - τα σύμβολα στην έξοδό της είναι ανεξάρτητα
  - ακολουθούν την ίδια κατανομή πιθανότητας
- Περιγράφεται πλήρως από:
  - το αλφάβητο

$$\Phi = \{s_1, \cdots, s_N\}$$

  - και τις πιθανότητες εμφάνισης ανά σύμβολο του αλφαβήτου

$$\{p_1, \cdots, p_N\}$$

- Ειδικές Περιπτώσεις:
  - **Δυαδική Πηγή Χωρίς Μνήμη:** $\Phi = \{0,1\}$ $\{p, 1-p\}$
  - Για *p=0.5*, **Δυαδική Συμμετρική Πηγή Χωρίς Μνήμη**

Εντροπία

- Η εντροπία μιας DMS ορίζεται ως

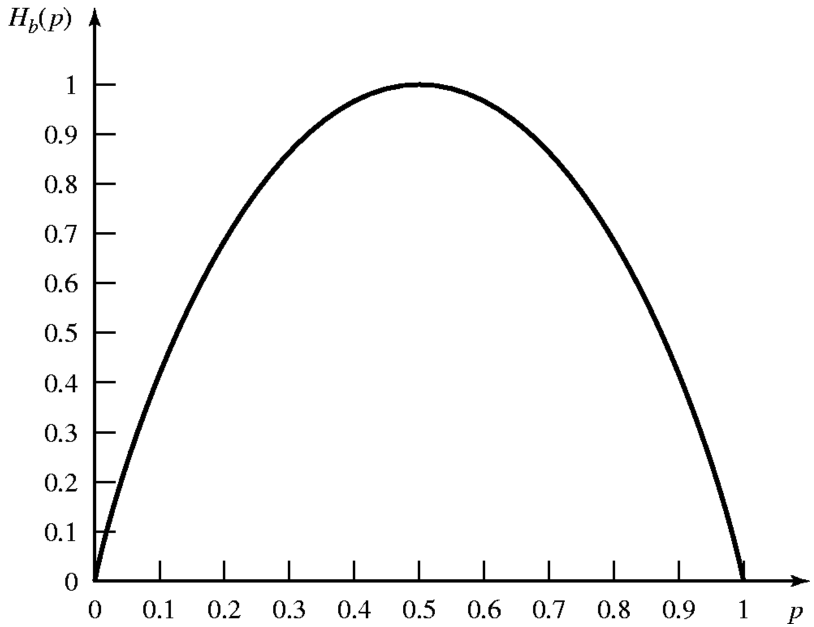
$$\boxed{H(\Phi) = \sum_{i=1}^{N} p_i I(s_i) = -\sum_{i=1}^{N} p_i \log_2 p_i}$$

- Φυσική Σημασία:
  - εκφράζει τη μέση αβεβαιότητα που έχω για την πηγή
  - είναι ο μέσος όρος της πληροφορίας των συμβόλων
- Όσο μεγαλύτερη εντροπία έχει μια πηγή,
  - τόσο περισσότερη πληροφορία φέρει, και
  - τόσο περισσότερα bits χρειάζονται για την κωδικοποίησή της (και κατά συνέπεια τόσο πιο αναποτελεσματική η συμπίεση της)

Συνάρτηση δυαδικής εντροπίας

- Αν έχω δυαδική DMS *Φ={0,1}*, με πιθανότητες εμφάνισης *{p,1-p}*, τότε ορίζεται η συνάρτηση δυαδικής εντροπίας

$$\boxed{H_b(p) = -p\log_2 p - (1-p)\log_2(1-p)}$$



Παρατηρήσεις:

1. ελαχιστοποιείται όταν *p=0 ή 1, H(0)=H(1)=0*
2. μεγιστοποιείται όταν τα σύμβολα είναι ισοπίθανα, *H(0.5)=1*

Πλεονασμός μιας πηγής πληροφορίας

$$\text{Πλεονασμός} = \frac{\max H(S) - H(S)}{\max H(S)} = 1 - \frac{H(S)}{\max H(S)} = 1 - \frac{H(S)}{\log n}$$

- Ο πλεονασμός μιας πηγής πληροφορίας οφείλεται στους παρακάτω λόγους:
  - Στο γεγονός ότι τα σύμβολα της πηγής είναι μη ισοπίθανα
  - Στη πιθανότητα η πηγή πληροφορίας να παρουσιάζει μνήμη.

Ρυθμός Παροχής Εντροπίας

- Αν μια πηγή πληροφορίας εκπέμπει σύμβολα με ρυθμό συμβόλων r_s (σε σύμβολα / sec)
- και η πηγή παρουσιάζει εντροπία H (σε bits / σύμβολο)
- τότε ο ρυθμός παροχής πληροφορίας από την πηγή (σε bits / sec) βρίσκεται άμεσα από τη σχέση:

$$R = r_s \cdot H \; bits/\mathrm{sec}$$

Κωδικοποίηση Πηγής

- **Στόχος:** η αποδοτική αναπαράσταση / κωδικοποίηση / συμπίεση της πληροφορίας/σήματος/εξόδου μιας πηγής
- Η διαδικασία μετατροπής των ακολουθιών συμβόλων που παράγει η πηγή σε ακολουθίες συμβόλων κάποιου κώδικα (συνήθως δυαδικές ακολουθίες)
- έτσι ώστε να αφαιρείται ο πλεονασμός και να προκύπτει συμπιεσμένη αναπαράσταση των μηνυμάτων ονομάζεται

κωδικοποίηση πηγής ή συμπίεση

Κωδικοποίηση Πηγής

- Βασικό χαρακτηριστικό κάθε κώδικα είναι ο αριθμός των bits που χρησιμοποιεί για να παραστήσει το κάθε σύμβολο.
- Αν ένας κώδικας χρησιμοποιεί μ αριθμό bits, ο αριθμός των δυνατών συνδυασμών των συμβόλων που μπορεί να περιγράψει με αυτά θα είναι ίσος με 2^μ.
- Αν ένας κώδικας έχει ως στόχο την κωδικοποίηση N διαφορετικών συμβόλων
- τότε ο αριθμός μ των bits που θα πρέπει να χρησιμοποιήσει δίνεται από τη σχέση:

$$2^{\mu-1} \leq N \leq 2^{\mu}$$

Κωδικοποίηση πηγής

- Στόχος: Η αποδοτική αναπαράσταση μιας πηγής Ν συμβόλων
  - **Κώδικα σταθερού μήκους:** το μήκος των κωδικών λέξεων είναι σταθερό για κάθε σύμβολο πηγής.
    - Παράδειγμα τέτοιου κώδικα είναι ο κώδικας ASCII
  - **Κώδικα μεταβλητού μήκους**: Τα σύμβολα της πηγής που έχουν μεγαλύτερη πιθανότητα εμφάνισης αντιστοιχίζονται σε μικρότερες κωδικές λέξεις και αντιστρόφως.
  - Έτσι το συνολικό μήκος του κωδικού μηνύματος μπορεί να προκύψει μικρότερο από το αρχικό μήνυμα.
    - Παράδειγμα τέτοιου κώδικα είναι ο κώδικας Morse

Κωδικοποίηση πηγής

Ανάλογα με το πόσο είναι δυνατή και εύκολη η αποκωδικοποίηση ενός κωδικού μηνύματος από το δέκτη οι κώδικες χωρίζονται σε:

- **Ευκρινείς κώδικες (non-singular):** χρησιμοποιεί διαφορετική κωδική λέξη για κάθε σύμβολο ή λέξη πληροφορίας. Η ευκρίνεια του κώδικα είναι η πρώτη προϋπόθεση για να υπάρχει δυνατότητα αποκωδικοποίησης.
- **Μονοσήμαντοι κώδικες (uniquely decodable):** χαρακτηρίζεται αν κάθε κωδική λέξη αναγνωρίζεται μέσα σε μακρά διαδοχή κωδικών συμβόλων. Δύο οποιαδήποτε μηνύματα πληροφορίας αντιστοιχίζονται με μονοσήμαντο κώδικα σε δύο διαφορετικά κωδικά μηνύματα.
- **Στιγμιαία αποκωδικοποιήσιμοι κώδικες (instantaneous code):** είναι κάθε μονοσήμαντος κώδικας (uniquely decodable), ο οποίος επιτρέπει αποκωδικοποίηση των μηνυμάτων λέξη προς λέξη χωρίς να απαιτείται εξέταση επόμενων κωδικών συμβόλων.

Ταξινόμηση κωδίκων με βάση την αποκωδικοποίηση

Κωδικοποίηση πηγής

Θεωρούμε τους ακόλουθους κώδικες I, II, III, και IV

	I	**II**	**III**	**IV**
φ	0	00	1	1
χ	10	01	10	01
ψ	01	10	100	001
ω	1	11	1000	0001

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

1. Ευκρινής
2. Μονοσήμαντοι
3. Στιγμιαία αποκωδικοποιήσημοι

Κωδικοποίηση πηγής

Λύση

- Όλοι οι κώδικες είναι ευκρινής, αφού ο καθένας αποτελείται από διαφορετικές κωδικές λέξεις
- Όλοι οι κώδικες είναι μονοσήμαντοι (μοναδικά αποκωδικοποιήσιμοι) εκτός του I.
  - Σε σχέση με τον κώδικα I, παρατηρούμε ότι η ακολουθία κωδικών λέξεων 1001 θα μπορούσε να προκύψει από διάφορες ακολουθίες συμβόλων όπως ‘χψ’ ή ‘χφω’ κ.λ.π
- Μόνο οι κώδικες II και IV είναι άμεσοι /στιγμιαία αποκωδηκοποιήσιμοι.
  - Στην περίπτωση του κώδικα I, αν ο δέκτης λάβει το ‘0’ δεν θα ξέρει αν είναι η πρώτη κωδική λέξη ή το $1^{ο}$ κωδικό σύμβολο της $3^{ης}$ κωδικής λέξης
  - Σχετικά με τον κώδικα III, όταν ο δέκτης λάβει τα ‘10’ δεν μπορεί να ξέρει αν είναι η $2^{η}$ κωδική λέξη ή τα 2 πρώτα σύμβολα της $3^{ης}$ λέξης,

Κωδικοποίηση πηγής

- Μέσο μήκος κώδικα: εκφράζει το μέσο πλήθος δυαδικών ψηφίων ανά σύμβολο πηγής τα οποία χρησιμοποιούνται στη διαδικασία της κωδικοποίησης.

$$\overline{L} = \sum_{i=1}^{N} p(s_i) l(s_i)$$

- Αποδοτικότητα κώδικα: Αν $L_{\min}$ η ελάχιστη δυνατή τιμή του μήκους, τότε μπορούμε να ορίσουμε την **αποδοτικότητα** του κώδικα ως:

$$n = \frac{L_{\min}}{\overline{L}}$$

$H(S)$

Θεώρημα Κωδικοποίησης Πηγής

- ή «Το Πρώτο Θεώρημα του Shannon» (1948)
- Χρησιμότητα: πόσο μπορούμε να συμπιέσουμε μια πηγή χωρίς να εισάγουμε σφάλματα;

Θεώρημα: Έστω πηγή με εντροπία H που κωδικοποιείται ώστε να παρέχει ρυθμό L *(bits/έξοδο πηγής)*.

- Αν $L>H$, η πηγή μπορεί να κωδικοποιηθεί με οσοδήποτε μικρή πιθανότητα σφάλματος
- Αν $L<H$, όσο πολύπλοκος κι αν είναι ο κωδικοποιητής πηγής, η πιθανότητα σφάλματος θα είναι μακριά από το 0

- Σχόλια:
  - Όπου L μπορείτε να θεωρήσετε το μέσο μήκος κώδικα $\bar{L}$
  - ο Shannon δίνει την ικανή και αναγκαία συνθήκη
  - όμως δεν προτείνει κάποιον αλγόριθμο/μεθοδολογία για να φτιάξουμε έναν κωδικοποιητή όταν $L \geq H$
  - $L<H$: *Data compression*,

Rate-Distortion Theory (lossy data compression)

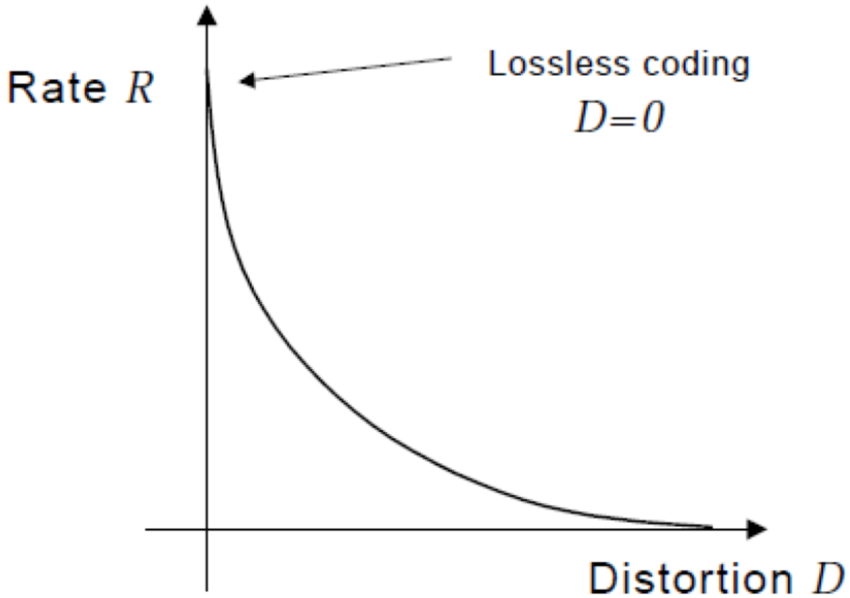
Lossy Data Compression

Low compression (High Quality)
JPEG

High compression (Low Quality)
JPEG

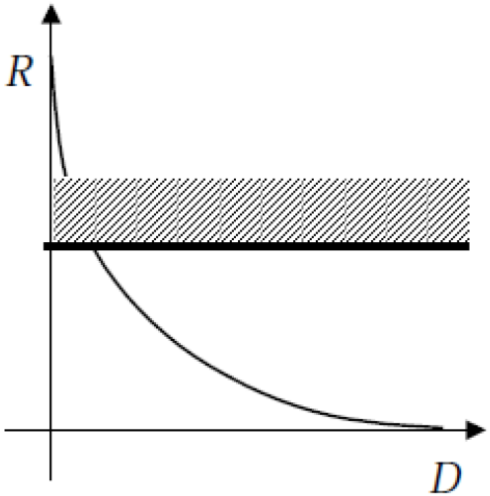
Lossy Data Compression

Lower the bit-rate R by allowing some acceptable distortion D of the signal.

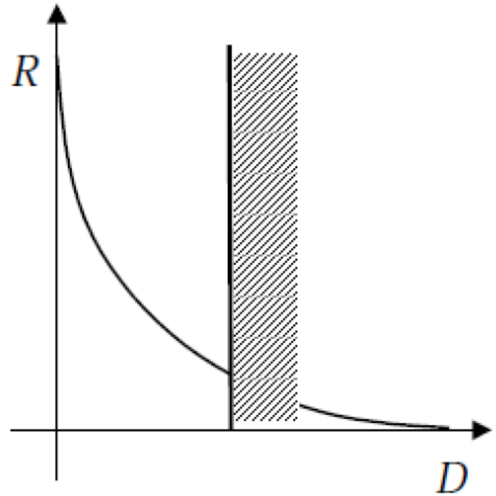


Lossy Data Compression

- Given maximum rate R, minimize distortion D



- Given distortion D, minimize rate R



Equivalent constrained optimization problems, often unwieldy due to constraint.

Προθεματικοί κώδικες

- Αλγόριθμοι κωδικοποίησης (συμπίεσης) πηγής
- Επιτυγχάνουν ρυθμούς κωδικοποίησης κοντά στην εντροπία (στο όριο συμπίεσης χωρίς απώλειες)
- Κωδικοποίηση από σταθερό σε μεταβλητό μήκος:
  - είσοδος: μπλοκ συμβόλων σταθερού μήκους

    (μήκος μπλοκ 0 1)
  - έξοδος: μπλοκ bits μεταβλητού μήκους (κωδική λέξη)
- Πρόβλημα: Συγχρονισμός
  - πώς μπορώ να βρω τα όρια των μπλοκ στην έξοδο για να γίνει η αποκωδικοποίηση
- Λύση: Προθεματικός
  - Καμία κωδική λέξη δεν αποτελεί πρόθεμα κάποιας άλλης
    - μοναδικά αποκωδικοποιήσιμος (κάθε έξοδος αντιστοιχεί σε μοναδική είσοδο)
    - άμεσος (επιτρέπει απευθείας αποκωδικοποίηση)

Παραδείγματα Κωδίκων

Source Symbol	Probability	Code 1	Code 2	Code 3
s_0	0.5	0	0	0
s_1	0.25	1	10	01
s_2	0.125	10	110	011
s_3	0.125	11	111	0111
Uniquely decodable		NO	YES	YES
Prefix		NO	YES	NO

Αποδοτικότητα Κώδικα

- Η αποδοτικότητα ενός κώδικα ορίζεται ως

$$\boxed{\eta = \frac{H(X)}{\overline{L}} \le 1}$$

και δείχνει πόσο κοντά βρίσκεται ο κωδικοποιητής στο όριο συμπίεσης της πηγής (εντροπία)

- Ένας κώδικας είναι αποδοτικός, όσο το *η* πλησιάζει στο 1
- Ίδια σχέση με πριν, όπου το ελάχιστο μέσο μήκος αντικαθίσταται με την εντροπία της πηγής.

Να βρείτε την αποδοτικότητα κάθε κώδικα

Source Symbol	**Probability**	**Code 1**	**Code 2**	**Code 3**
s_0	0.5	0	0	0
s_1	0.25	1	10	01
s_2	0.125	10	110	011
s_3	0.125	11	111	0111
Uniquely decodable				
Prefix				

Αλγόριθμος Shannon

1. Τα σύμβολα της πηγής πληροφορίας συντάσσονται στη σειρά με κριτήριο την πιθανότητα τους (από την μέγιστη στην ελάχιστη πιθανότητα).

$$
\begin{array}{cccc}
x_1 & x_2 & \cdots & x_M \\
\updownarrow & \updownarrow & & \updownarrow \\
p(x_1) \geq & p(x_2) \geq & \cdots & \geq p(x_M)
\end{array}
$$

2. Σε κάθε σύμβολο πληροφορίας αντιστοιχίζεται ένας αριθμός με την λογική που περιγράφεται στο παρακάτω σχήμα:

$$x_1 \leftrightarrow \varepsilon_1 = 0$$

$$x_2 \leftrightarrow \varepsilon_2 = p(x_1) + \varepsilon_1 = p(x_1)$$

$$x_3 \leftrightarrow \varepsilon_3 = p(x_2) + \varepsilon_2 = p(x_1) + p(x_2)$$

....

$$x_M \leftrightarrow \varepsilon_M = p(x_{M-1}) + \varepsilon_{M-1} = p(x_1) + p(x_2) + \ldots + p(x_{M-1})$$

Αλγόριθμος Shannon

3. Το μήκος της κωδικής λέξης που αντιστοιχεί στο σύμβολο πληροφορίας είναι ο ελάχιστος ακέραιος που ικανοποιεί την ταυτοανισότητα

$$l_i \geq \log\left(\frac{1}{p(x_i)}\right) \Rightarrow 2^{l_i}\, p(x_i) \geq 1$$

επιλέγεται η ελάχιστη δυνατή τιμή

4. Οι δεκαδικοί αριθμοί , όπου μετατρέπονται σε δυαδικούς αριθμούς και από τους τελευταίους διατηρούνται μόνο σημαντικά ψηφία τα οποία θα αποτελέσουν τις αντίστοιχες κωδικές λέξεις για τα σύμβολα πληροφορίας

Παράδειγμα κωδικοποίησης Shannon

Έστω πηγή πληροφορίας $X = \{x_1, x_2, \ldots\ldots, x_M\}$ με κατανομή πιθανότητας

$$p(X) = \{0.5, 0.3, 0.1, 0.1\}$$

Βήμα 1

είναι έτοιμο (σωστά διατεταγμένα σύμβολα).

Βήμα 2

$$\varepsilon_1 = 0 = 0.0 = 0.\underline{0}0000\ldots$$

$$\varepsilon_2 = 0.5 + \varepsilon_1 = 0.5 = 0.\underline{10}000\ldots$$

$$\varepsilon_3 = 0.3 + \varepsilon_2 = 0.8 = 0.\underline{1100}1\ldots$$

$$\varepsilon_4 = 0.1 + \varepsilon_3 = 0.9 = 0.\underline{1110}0\ldots$$

Παράδειγμα κωδικοποίησης Shannon

$l_i \geq \log_2\left(\frac{1}{p(x_i)}\right)$

Βήμα 3

$$2^{l_1}(0.5) \geq 1 \Rightarrow l_1 = 1$$

$l_1 \geq \log_2\left(\frac{1}{1/2}\right) \geq \log(2) = 1$

$$2^{l_2}(0.3) \geq 1 \Rightarrow l_2 = 2$$

$l_2 \geq \log_2\left(\frac{1}{0,3}\right)$

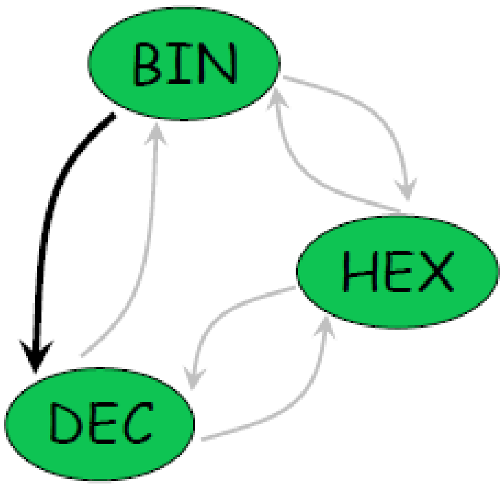
$l_2 \geq 1,737$

$$2^{l_3}(0.1) \geq 1 \Rightarrow l_3 = 4$$

$$2^{l_4}(0.1) \geq 1 \Rightarrow l_4 = 4$$

Οι λέξεις κώδικα για τα $X = \{x_1, x_2, x_3, x_4\}$ θα είναι $(0,10,1100,1110)$

Μετατροπή BIN → DEC



11101.101_2

1 1 1 0 1

2^4 2^3 2^2 2^1 2^0

1x16 1x8 1x4 0x2 1x1

16 + 8 + 4 + 0 + 1

29

1 0 1

2^{-1} 2^{-2} 2^{-3}

1x(1/2) 0x(1/4) 1x(1/8)

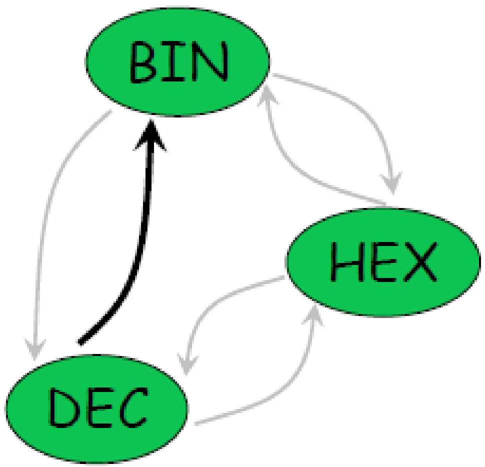
1/2 + 1/4 + 1/8

5/8

0.625

29.625_{10}

Μετατροπή DEC → BIN



29.625_{10}

29 / 2	=	14	υπόλοιπο	1
14 / 2	=	7	υπόλοιπο	0
7 / 2	=	3	υπόλοιπο	1
3 / 2	=	1	υπόλοιπο	1
1 / 2	=	0	υπόλοιπο	1

ΑΝΑΓΝΩΣΗ

.625 x 2	=	.250 +	ακέραιο	1
.250 x 2	=	.500 +	ακέραιο	0
.500 x 2	=	.000 +	ακέραιο	1

ΑΝΑΓΝΩΣΗ

11101.101_2

Κωδικοποίηση Shannon - Fano

1. Τα σύμβολα της πηγής πληροφορίας συντάσσονται στη σειρά με κριτήριο την πιθανότητα τους (από την μέγιστη στην ελάχιστη πιθανότητα). Για παράδειγμα:

$$\begin{array}{cccc} x_1 & x_2 & \dots & x_M \\ \updownarrow & \updownarrow & & \updownarrow \\ p(x_1) \geq & p(x_2) \geq & \dots & \geq p(x_M) \end{array}$$

2. Επιλέγεται συγκεκριμένη διάταξη για τα κωδικά σύμβολα η οποία δεν αλλάζει σε καμία φάση της κωδικοποίησης αλλά και κατά την αποκωδικοποίηση.

Κωδικοποίηση Shannon - Fano

3. Σχηματίζονται D ομάδες συμβόλων πηγής πληροφορίας με συγχώνευση γειτονικών συμβόλων. Οι πιθανότητες των συμβόλων που συμμετέχουν σε κάθε ομάδα αθροίζονται και το αποτέλεσμα επιδιώκεται να είναι όσο το δυνατόν πλησιέστερα στον αριθμό 1/D δηλαδή όλες οι ομάδες συμβόλων είναι κατά το δυνατόν ισοπίθανες.

4. Στα σύμβολα της πρώτης ομάδας αντιστοιχούμε σαν πρώτο κωδικό σύμβολο το γ1 , της δεύτερης το γ2 κ.ο.κ.

5. Κάθε ομάδα συμβόλων την διαιρούμε σε D υποομάδες πάλι με το ίδιο κριτήριο (κατά το δυνατόν ισοπίθανες). Στα σύμβολα της κάθε υποομάδας αντιστοιχίζεται ως δεύτερο κωδικό σύμβολο ένα κωδικό σύμβολο με την προκαθορισμένη διάταξη.

6. Η συγκεκριμένη αναδρομική διαδικασία συνεχίζεται μέχρι να προκύψουν υποομάδες με ένα μόνο σύμβολο.

Στη συνήθη περίπτωση – επιλέγεται D=2

Παράδειγμα Κωδικοποίηση Shannon - Fano

- Έστω $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ πηγή πληροφορίας με κατανομή πιθανότητας

$$p(X) = \{0.25, 0.25, 0.125, 0.125, 0.0625, 0.0625, 0.0625, 0.0625\}$$

Βήμα 1

Με βάση την κατανομή πιθανότητας βλέπουμε ότι τα σύμβολα πληροφορίας είναι διατεταγμένα σωστά (το βήμα (1) έχει πραγματοποιηθεί.

Βήμα 2

Στο βήμα (2) επιλέγεται για τα σύμβολα του δυαδικού κώδικα Shannon-Fano η διάταξη (0,1). Ο παρακάτω πίνακας περιγράφει την σταδιακή κατασκευή του κώδικα:

Παράδειγμα κωδικοποίησης Shannon - Fano

Σύμβολο	Πιθανότητα	1ο βήμα	2ο Βήμα	3ο Βήμα	4ο Βήμα
x_1	0.25	0	00	00	00
x_2	0.25	0	01	01	01
x_3	0.125	1	10	100	100
x_4	0.125	1	10	101	101
x_5	0.0625	1	11	110	1100
x_6	0.0625	1	11	110	1101
x_7	0.0625	1	11	111	1110
x_8	0.0625	1	11	111	1111

Κωδικοποίηση πηγής Huffman

- Δημιουργία Δυαδικού Δέντρου:

1. Διάταξε τις εισόδους κατά φθίνουσα σειρά πιθανοτήτων
2. Συγχώνευσε τα δύο σύμβολα με τις μικρότερες πιθανότητες και δημιούργησε νέο «σύμβολο»
3. Ανάθεσε στα δύο σύμβολα «0» και «1»
4. Ταξινόμησε εκ νέου τη λίστα των συμβόλων
5. Επανέλαβε τα παραπάνω μέχρι όλα τα σύμβολα συγχωνευτούν σε ένα τελικό σύμβολο

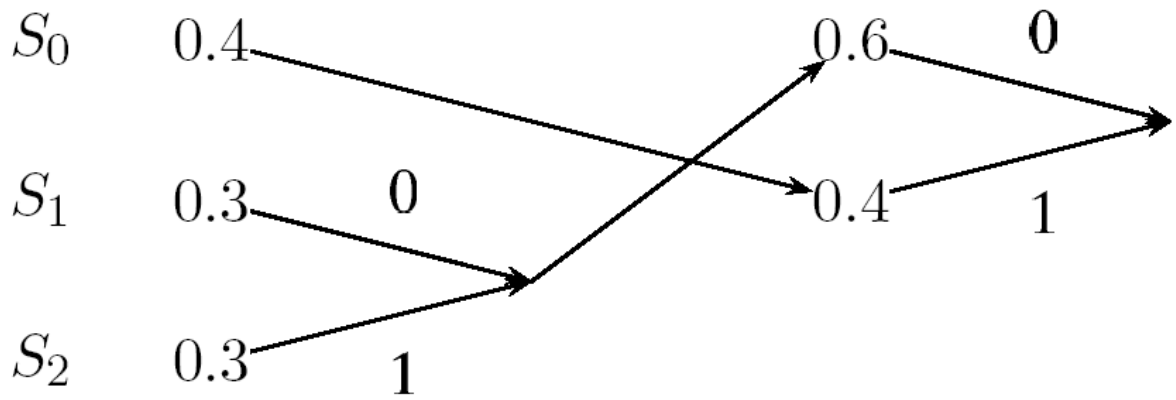
- Δημιουργήθηκε ένα δυαδικό δέντρο:
  - ρίζα: το τελικό σύνθετο σύμβολο
  - φύλλα: τα αρχικά σύμβολα
  - ενδιάμεσοι κόμβοι: σύνθετα σύμβολα

Κωδικοποίηση Πηγής Huffman

- Ανάθεση Bits σε Σύμβολα Εισόδου

1. Ξεκίνα από τη ρίζα και κινήσου προς ένα φύλλο
2. Η ακολουθία των bits που συναντώνται είναι η ακολουθία κωδικοποίησης
3. Επανέλαβε για όλα τα σύμβολα (φύλλα)

Παράδειγμα Κωδικοποίηση Πηγής Huffman



- Προθεματική αντιστοίχιση:
  - s_0: 1
  - s_1: 00
  - s_2: 01
- Μονοσήμαντη και άμεση αποκωδικοποίηση

$$\boxed{\underbrace{1}_{s_0}\,\underbrace{01}_{s_2}\,\underbrace{00}_{s_1}\,\underbrace{01}_{s_2}\,\underbrace{1}_{s_0}\,\underbrace{00}_{s_1}}$$

ΚΩΔΙΚΑΣ HUFFMAN

S		P_i				
a_1	1	0,4	0,4	0,4	0,4	0 0,6
a_2	00	0,3	0,3	0,3	00 0,3	1 0,4
a_3	011	0,1	0,1	010 0,2	01 0,3	
a_4	0100	0,1	0100 0,1	011 0,1		
a_5	01010	0,06 +	0101 0,1			
a_6	01011	0,04				
		Σ. 1.0				

ΚΩΔΙΚΑΣ HUFFMAN

S		P_i								
a_1	1	0,4		0,4		0,4		0,4	0	0,6
a_2	00	0,3		0,3		0,3	00	0,3	1	0,4
a_3	011	0,1		0,1	010	0,2	01	0,3		
a_4	0100	0,1	0100	0,1	011	0,1				
a_5	01010	0,06 +	0101	0,1						
a_6	01011	0,04								
		Σ. 1.0								

Χαρακτηριστικά Huffman

- Μειονέκτημα:
  - απαιτεί να γνωρίζει εκ των προτέρων τις πιθανότητες εμφάνισης των συμβόλων της πηγής
  - δε μπορεί να χρησιμοποιηθεί σε εφαρμογές πραγματικού χρόνου
- Βέλτιστος:
  - ανάμεσα σε όλους τους προθεματικούς κώδικες (άρα μονοσήμαντα αποκωδικοποιήσιμους και άμεσους)
  - πετυχαίνει το ελάχιστο μέσο μήκος κώδικα
- Συμβάσεις:
  - Ο τρόπος ανάθεσης 0 και 1
  - Η τοποθέτηση στην ταξινομημένη λίστα σε περίπτωση «ισοβαθμίας» σύνθετου συμβόλου με άλλο σύμβολο (σχετίζεται με τη διασπορά του κώδικα)

Άσκηση

- Για κάθε πηγή που έχουμε μελετήσει στο παρόν κεφάλαιο να εκτελέσετε τα ακόλουθα:
  - Να υπολογίσετε την Εντροπία κάθε πηγής
  - Να υπολογίσετε τον Πλεονασμό κάθε πηγής
  - Να σχεδιάσετε έναν σταθερού μήκους δυαδικό κώδικα, Shannon, Shannon-Fano, και Huffman Κώδικα.
  - Να υπολογίσετε την απόδοση κάθε κώδικα και να επιλέξετε τον περισσότερο και λιγότερο αποδοτικό κώδικα
  - Να εξετάσετε κάθε κώδικα που δημιουργήσατε εάν είναι μονοσήμαντος και στιγμιαία απόκωδικοποιήσιμος.
  - Να εξετάσετε ποιοι είναι προθεματικοί κώδικες

Έστω πηγή πληροφορίας $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ με κατανομή πιθανότητας

$$p(X) = \{0.25, 0.25, 0.125, 0.125, 0.0625, 0.0625, 0.0625, 0.0625\}$$

Εντροπία Πηγής με μνήμη

- Πηγή πληροφορίας με μνήμη **1ης τάξης** (πηγή στην οποία η εκπομπή ενός συμβόλου εξαρτάται από το προηγούμενο σύμβολο που εκπέμφθηκε):

$$H(S) = \sum_{i=1}^{m} p_i H(K_i) = -\sum_{i=1}^{m} p_i \sum_{j=1}^{m} P_{ij} \log P_{ij}$$

- Ο ρυθμός εκπομπής πληροφορίας στο κανάλι για **μία πηγή με μνήμη**:

$$R = r_s H(S)$$

Μαρκοβιανές αλυσίδες

- Μια τυχαία διαδικασία είναι μια ακολουθία τυχαίων μεταβλητών $Y1, Y2, \ldots, Yn$.
- Γενικά, μπορεί να υφίσταται οποιαδήποτε εξάρτηση μεταξύ των τυχαίων μεταβλητών της ακολουθίας.
- Μια τυχαία διαδικασία Υ1, Υ2, … , Υn χαρακτηρίζεται ως διαδικασία Markov 1ης τάξης (Μαρκοβιανή αλυσίδα), αν για n = 1,2, …ισχύει:

$$P(Y_{n+1} = y_{n+1} \mid Y_n = y_n, Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1) = P(Y_{n+1} = y_{n+1} \mid Y_n = y_n)$$

- Στην περίπτωση της διαδικασίας Markov, η συνάρτηση πιθανότητας μάζας μπορεί να γραφεί

$$p(y_1, y_2, \ldots, y_n) = p(y_1)\, p(y_2|y_1)\, p(y_3|y_2) \ldots p(y_n|y_{n-1})$$

- Η διαδικασία Markov χαρακτηρίζεται ως χρονικά αμετάβλητη αν η υπό συνθήκη πιθανότητα $\mathrm{P}(Y_{n+1} = b|Y_n = a) = \pi r^2$δεν εξαρτάται από το n. Δηλαδή για n = 1,2, …ισχύει:

$$P(Y_{n+1} = b \mid Y_n = a) = P(Y_2 = b \mid Y_1 = \alpha)$$

Εντροπία Πηγής με μνήμη

- Οι μεταπτώσεις της μπορούν να περιγράφουν με τη **μήτρα πιθανοτήτων μετάπτωσης** δηλαδή τη μήτρα των p_{ij}:
- Μία πηγή με μνήμη q συμβόλων μπορεί να περιγραφτεί με πίνακα P τα στοιχεία του οποίου είναι τιμές πιθανοτήτων μετάπτωσης:

$$P=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,q} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,q} \\ & & & \\ p_{q,1} & p_{q,2} & \cdots & p_{q,q} \end{bmatrix}$$

$$P=\begin{array}{cc} & \begin{array}{ccc} \mathrm{A} & \mathrm{B} & \Gamma \end{array} \\ \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \Gamma \end{array} & \begin{bmatrix} 3/4 & 0 & 1/4 \\ 1/4 & 2/3 & 0 \\ 0 & 1/3 & 3/4 \end{bmatrix} \end{array}$$

Μαρκοβιανές αλυσίδες

- Αν η πιθανότητα να βρίσκεται το σύστημα στην κατάσταση j κατά την αρχή του k-στου διαστήματος συμβόλου είναι $p_j(k)$, τότε οι μεταπτώσεις του συστήματος περιγράφονται από τη σχέση:

$$p_j\,(k+1) = \sum_{i=1}^{m} p_i(k) P_{ji}$$

όπου m είναι το πλήθος των δυνατών καταστάσεων της πηγής.

- Για μια στατική Μαρκοβιανή αλυσίδα, μεταξύ του διανύσματος των πιθανοτήτων των καταστάσεων $\boldsymbol{\pi}$ και του πίνακα μετάπτωσης $\boldsymbol{P}$, ισχύει η σχέση:

$$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$$

Τάξη των Μαρκοβιανών

- Η τάξη της Μαρκοβιανής αλυσίδας καθορίζεται από το πλήθος ($\boldsymbol{l}$) των προηγούμενων συμβόλων που επηρεάζουν το επόμενο σύμβολο που θα παραχθεί από την πηγή.
- Η πηγή είναι στην κατάσταση i, όταν η τελευταία εκπομπή είναι αυτή του συμβόλου s_i.
- Μετάβαση της πηγής από την κατάσταση i στην κατάσταση j, σημαίνει ότι μετά την εκπομπή του συμβόλου s_i εκπέμπεται το σύμβολο s_j.
- Σε μία Μαρκοβιανή αλυσίδα:
  - πρώτης τάξης (l = 1), το πλήθος των καταστάσεων της πηγής είναι ίσο με το πλήθος των συμβόλων του αλφαβήτου της πηγής (m =q).
  - δεύτερης τάξης (l = 2), το πλήθος των καταστάσεων της πηγής είναι ίσο με το τετράγωνο του πλήθους των συμβόλων ($m = q^2$).

Εντροπία πηγής με μνήμη

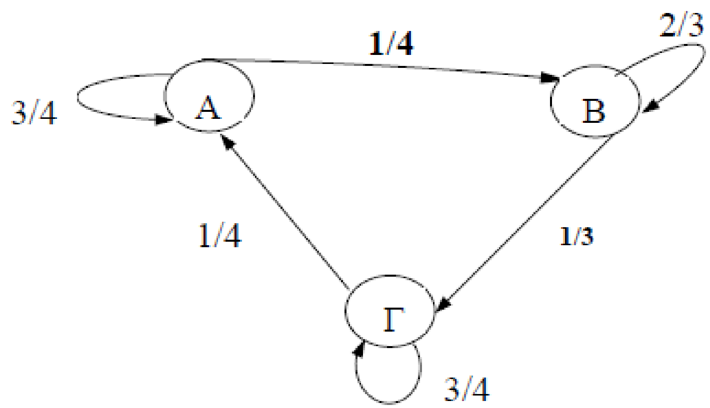
π.χ. Η πιθανότητα ¾ της πρώτης γραμμής και πρώτης στήλης είναι η πιθανότητα να σταλεί το σύμβολο Α δεδομένου ότι έχει ήδη σταλεί το σύμβολο Α, η πιθανότητα 1/3 της τρίτης γραμμής και δεύτερης στήλης ερμηνεύεται ως η πιθανότητα να σταλεί το σύμβολο Γ δεδομένου ότι έχει σταλεί το σύμβολο Β

Στην παρούσα μορφή, οι γραμμές εκφράζουν το παρόν, ενώ οι στήλες το παρελθόν (ή αντίστοιχα το μέλλον και το παρόν).

Το άθροισμα ανά στήλη είναι ίσο με 1 (από την παρελθούσα κατάσταση πρέπει να καταλήξουμε σε μια νέα κατάσταση)

Εντροπία πηγής με μνήμη

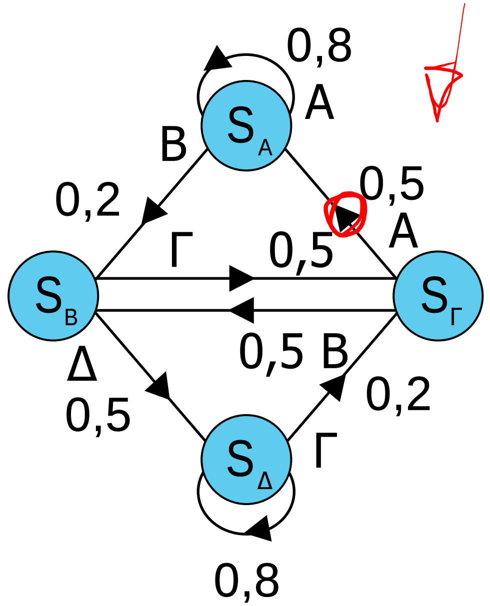
Η προηγούμενη μήτρα μεταπτώσεων ισοδυναμεί με το επόμενο **διάγραμμα καταστάσεων της πηγής**:



$$P_A = P_{A/A} \cdot P_A + P_{A/B} \cdot P_B + P_{A/\Gamma} \cdot P_\Gamma$$
$$P_B = P_{B/A} \cdot P_A + P_{B/B} \cdot P_B + P_{B/\Gamma} \cdot P_\Gamma$$
$$P_\Gamma = P_{\Gamma/A} \cdot P_A + P_{\Gamma/B} \cdot P_B + P_{\Gamma/\Gamma} \cdot P_\Gamma$$

Έχουμε ένα σύστημα 4 εξισώσεων με 3 αγνώστους (P_A, P_B, και P_Γ) το οποίο μπορεί να λυθεί. H(x) memory + H(x) memoryless

Παράδειγμα πηγής Markov 1ης τάξης



Να υπολογίσετε τη μήτρα μεταπτώσεων, και τις πιθανότητες P_A, P_B, $P_Γ$, $P_Δ$

Να υπολογίσετε την Εντροπία της εν λόγω πηγής με μνήμη και χωρίς μνήμη

	A	B	Γ	Δ
A	0,8	0	0,5	0
B	0,2	0	0,5	0
Γ	0	0,5	0	0,2
Δ	0	0,5	0	0,8

“Υπό συνθήκη” πιθανότητες

$$P_{A/A} = P_{Δ/Δ} = 0{,}8$$

$$P_{B/A} = P_{Γ/Δ} = 0{,}2$$

$$P_{Γ/B} = P_{B/Γ} = P_{A/Γ} = P_{Δ/B} = 0{,}5$$

Υπόλοιπες πιθανότητες = 0

ολοι οι κλαδοι που λείπουν

Παράδειγμα πηγής Markov 1ης τάξης

$$\left.\begin{aligned}
P_A &= 0,8\ P_A + 0,5\ P_\Gamma \\
P_B &= 0,2\ P_A + 0,5\ P_\Gamma \\
P_\Gamma &= 0,5\ P_B + 0,2\ P_\Delta \\
P_\Delta &= 0,5\ P_B + 0,8\ P_\Delta \\
P_A + P_B &+ P_\Gamma + P_\Delta = 1
\end{aligned}\right\}$$

Λύση του συστήματος

$$P_A = P_\Delta = \frac{5}{14}\ , \quad P_B = P_\Gamma = \frac{1}{7}$$

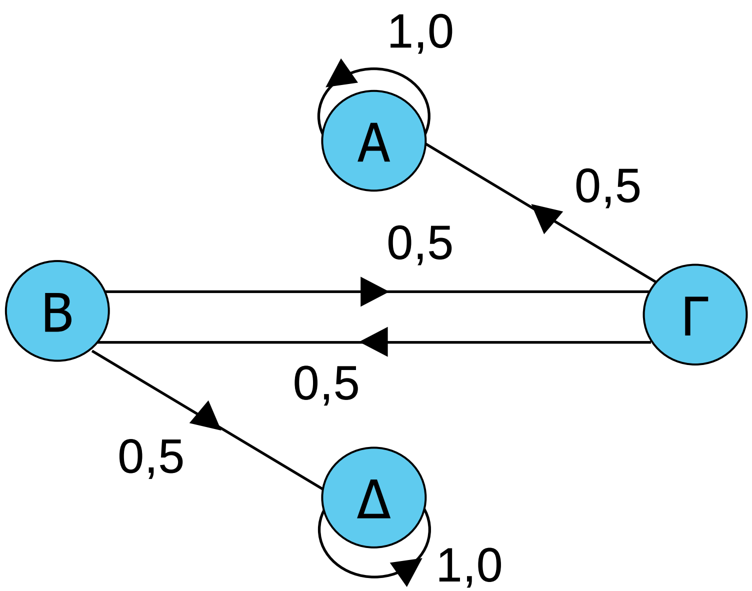
- Ποια η εντροπία της Μαρκοβιανής πηγής ;

1ης τάξης : $H = \sum_i \sum_j P_{j/i} \cdot P_i \log \frac{1}{P_{j/i}} < H_{X.M.}$

Στο παράδειγμα :

$$H_M \approx 0,8\ \text{bits}/\text{symbol} < H_{ML} \approx 1,86\ \text{bits}/\text{symbol}$$

Παράδειγμα πηγής Markov 1ης τάξης



Ποιο είναι το παράξενο στην συγκεκριμένη πηγή Markov 1ης τάξης

Παράδειγμα Μαρκοβιανής πηγής 1ης τάξης

Δημιουργήστε το διάγραμμα της Μαρκοβιανής Πηγής

$$\begin{bmatrix} 0{,}25 & 0{,}50 & 0{,}75 & 0{,}50 & 0{,}00 \\ 0{,}375 & 0{,}00 & 0{,}00 & 0{,}50 & 0{,}00 \\ 0{,}25 & 0{,}25 & 0{,}00 & 0{,}00 & 0{,}50 \\ 0{,}125 & 0{,}00 & 0{,}25 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}25 & 0{,}00 & 0{,}00 & 0{,}50 \end{bmatrix}$$

Άσκηση

- Μια διακριτή πηγή με μνήμη παράγει μια Μαρκοβιανή αλυσίδα πρώτης τάξης. Το αλφάβητο της πηγής αποτελείται από τα σύμβολα φ, χ και ψ. Ο πίνακας μετάβασης είναι:

$$\begin{bmatrix} p(\varphi|\varphi) & p(\chi|\varphi) & p(\psi|\varphi) \\ p(\chi|\varphi) & p(\chi|\chi) & p(\psi|\chi) \\ p(\psi|\varphi) & p(\chi|\psi) & p(\psi|\psi) \end{bmatrix} = \begin{bmatrix} 1/2 & 1/4 & 1/3 \\ 1/2 & 0 & 1/3 \\ 0 & 3/4 & 1/3 \end{bmatrix}$$

- Να υπολογιστούν:
  - Οι πιθανότητες εκπομπής των συμβόλων της πηγής φ, χ και ψ.
  - Οι συνδυασμένες πιθανότητες εκπομπής μηνυμάτων αποτελούμενων από δύο σύμβολα.
  - Να σχηματιστούν κωδικές λέξεις με το δυαδικό κωδικό αλφάβητο για τα δυνατά μηνύματα δύο συμβόλων σύμφωνα με τον αλγόριθμο του Fano.

- Επειδή η Μαρκοβιανή αλυσίδα είναι πρώτης τάξης, το πλήθος των καταστάσεων της πηγής είναι ίσο με το πλήθος των συμβόλων. Επομένως, η κατάσταση της πηγής περιγράφει το τελευταίο σύμβολο που παρήγαγε η πηγή και έτσι η πιθανότητα παραγωγής ενός συμβόλου ισούται με την πιθανότητα κατάστασης της πηγής.
- Από τον πίνακα μετάβασης υπολογίζουμε τις πιθανότητες εκπομπής των συμβόλων της πηγής, οι οποίες ταυτίζονται με τις πιθανότητες κατάστασης της πηγής Markov. Επιλύοντας το ακόλουθο σύστημα εξισώσεων:

$$p(\varphi) = p(\varphi)\mathrm{P}(\varphi|\varphi) + p(\chi)\mathrm{P}(\varphi|\chi) + p(\psi)\mathrm{P}(\varphi|\psi)$$
$$p(\chi) = p(\varphi)\mathrm{P}(\chi|\varphi) + p(\chi)\mathrm{P}(\chi|\chi) + p(\psi)\mathrm{P}(\chi|\psi)$$
$$p(\psi) = p(\varphi)\mathrm{P}(\psi|\varphi) + p(\chi)\mathrm{P}(\psi|\chi) + p(\psi)\mathrm{P}(\psi|\psi)$$
$$p(\varphi) + p(\chi) + p(\psi) = 1$$

Προκύπτει: $p(\varphi) = \frac{10}{27}$, $p(\chi) = \frac{8}{27}$, $p(\psi) = \frac{9}{27}$

- Το μέσο πληροφοριακό περιεχόμενο μηνυμάτων αποτελούμενων από δύο σύμβολα υπολογίζεται από τη σχέση:

$$\mathrm{H(M)} = -\sum_{\iota} p(m_i)\log(p(m_i))$$

Το πλήθος των διαφορετικών μηνυμάτων που αποτελούνται από δύο σύμβολα είναι ίσο με το τετράγωνο του πλήθους των συμβόλων της πηγής, δηλαδή ίσο με $3^2 = 9$

Οι συνδυασμένες πιθανότητες εκπομπής των μηνυμάτων αυτών μπορούν να υπολογιστούν από τον πίνακα μετάβασης και τις πιθανότητες εκπομπής των συμβόλων $p(\varphi)$, $p(\chi)$ και $p(\psi)$

Έτσι, η πιθανότητα εκπομπής του μηνύματος m_k που σχηματίζεται από τα σύμβολα s_i και s_j δίνεται από τη σχέση:

$$p(m_k = s_i s_j) = p(s_i, s_j) = p(s_i)P(s_j|s_i)$$

- Επομένως οι πιθανότητες όλων των δυνατών μηνυμάτων είναι:

$$p(\varphi\varphi) = p(\varphi)P(\varphi/\varphi) = \frac{10}{27}.\frac{1}{2} = \frac{5}{27} = 0{,}185$$

$$p(\varphi\chi) = p(\varphi)P(\chi/\varphi) = \frac{10}{27}.\frac{1}{4} = \frac{5}{54}=0{,}093$$

$$p(\varphi\psi) = p(\varphi)P(\psi/\varphi) = \frac{10}{27}.\frac{1}{3} = \frac{10}{81}=0{,}124$$

$$p(\chi\varphi) = p(\chi)P(\varphi/\chi) = \frac{8}{27}.\frac{1}{2} = \frac{4}{27}=0{,}148$$

$$p(\chi\chi) = p(\chi)P(\chi/\chi) = \frac{8}{27}.0 = 0$$

$$p(\chi\psi) = p(\chi)P(\psi/\chi) = \frac{8}{27}.\frac{1}{3} = \frac{8}{81}=0{,}099$$

$$p(\psi\varphi) = p(\psi)P(\varphi/\psi) = \frac{9}{27}.0 = 0$$

$$p(\psi\chi) = p(\psi)P(\chi/\psi) = \frac{9}{27}.\frac{3}{4} = \frac{1}{4}=0{,}250$$

$$p(\psi\psi) = p(\psi)P(\psi/\psi) = \frac{9}{27}.\frac{1}{3} = \frac{1}{9}=0{,}111$$

- Οι κωδικές λέξεις που προκύπτουν από τον αλγόριθμο Fano είναι:

Μήνυμα	Πιθανότητα	Κωδική Λέξη
ψχ	0,250	00
φφ	0,185	01
χφ	0,148	100
φψ	0,124	101
ψψ	0,111	110
χψ	0,099	1110
φχ	0,093	1111
χχ	0	-
ψφ	0	-

Παρατήρηση:

- Το μέσο μήκος των κωδικών λέξεων υπολογίζεται από το μήκος των κωδικών λέξεων λαμβανομένων υπόψη των πιθανοτήτων παραγωγής των αντίστοιχων μηνυμάτων.
- Επομένως, L = 75/27 = 2,78 bits/message ή 1,39 bits/symbol.
- Η συνδυασμένη ποσότητα πληροφορίας, H(X,Y), είναι ίση με 2,72 bits/message. Επομένως, η αποδοτικότητα, α = H(X,Y)/L, είναι περίπου ίση με 0,98.
- Αν είχαμε εφαρμόσει τον αλγόριθμο του Fano σε επίπεδο συμβόλων και όχι μηνυμάτων, θα είχε προκύψει ένα μέσο μήκος κωδικών λέξεων περίπου ίσο με 3,26 bits/message, το οποίο είναι μεγαλύτερο από το 2,78 bits/message, που λάβαμε ανωτέρω.
- Συνεπώς, η κωδικοποίηση των συμβόλων σε σύγκριση με αυτή των μηνυμάτων θα οδηγούσε σε μείωση της αποδοτικότητας (περίπου 0,76).