ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Θέματα Συστημάτων Πολυμέσων

Ενότητα # 5: Βασική Θεωρία Πληροφορίας

Διδάσκων: Γεώργιος Πολύζος

Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών “Επιστήμη των Υπολογιστών”

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες χρήσης

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σκοποί ενότητας

- Παρουσίαση του θεωρήματος Shannon
- Κατανόηση της εντροπίας μιας πηγής και τους σκοπούς που εξυπηρετεί

Περιεχόμενα ενότητας

- Τι είναι θεωρία πληροφορίας
- Στοχαστικές πηγές πληροφορίας
- Εντροπία πληροφορίας
- Εντροπία στοχαστικών πηγών
- Θεώρημα κωδικοποίησης πηγής

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Θεωρία Πληροφορίας

Μάθημα: Θέματα Συστημάτων Πολυμέσων

Ενότητα # 5: Βασική Θεωρία Πληροφορίας

Διδάσκων: Γεώργιος Κ. Πολύζος

Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών “Επιστήμη των Υπολογιστών”

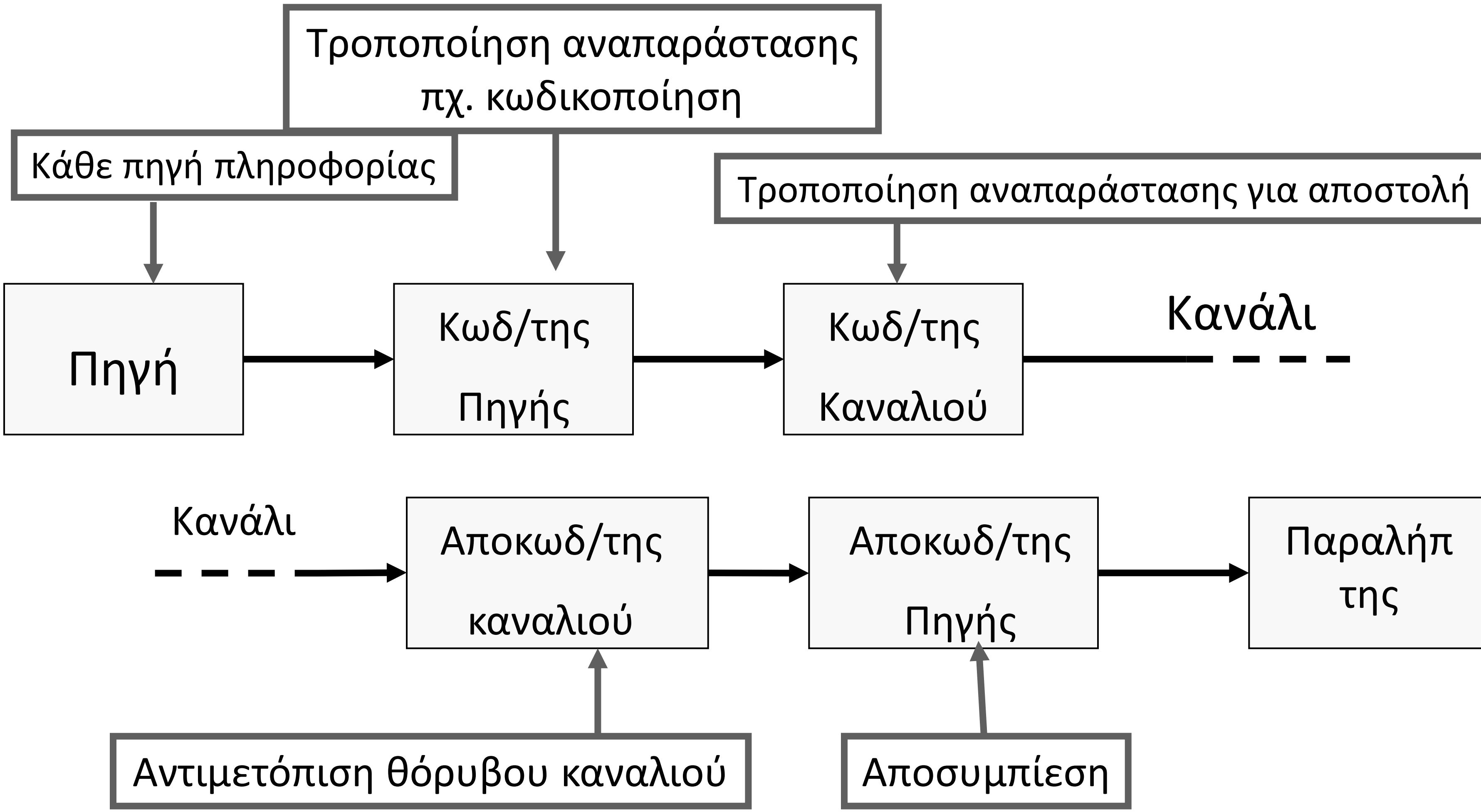
Θεωρία πληροφορίας κατά Shannon

- Claude Shannon: **Η** μαθηματική θεωρία της επικοινωνίας
  - ”A mathematical theory of communication”, Bell System Technical Journal, 1948
- Αναφέρεται και ως ”Shannon-Waver”, μιας και την ειδική δημοσίευση προλογίζει ο Weaver.

Ρητά για τον Shannon

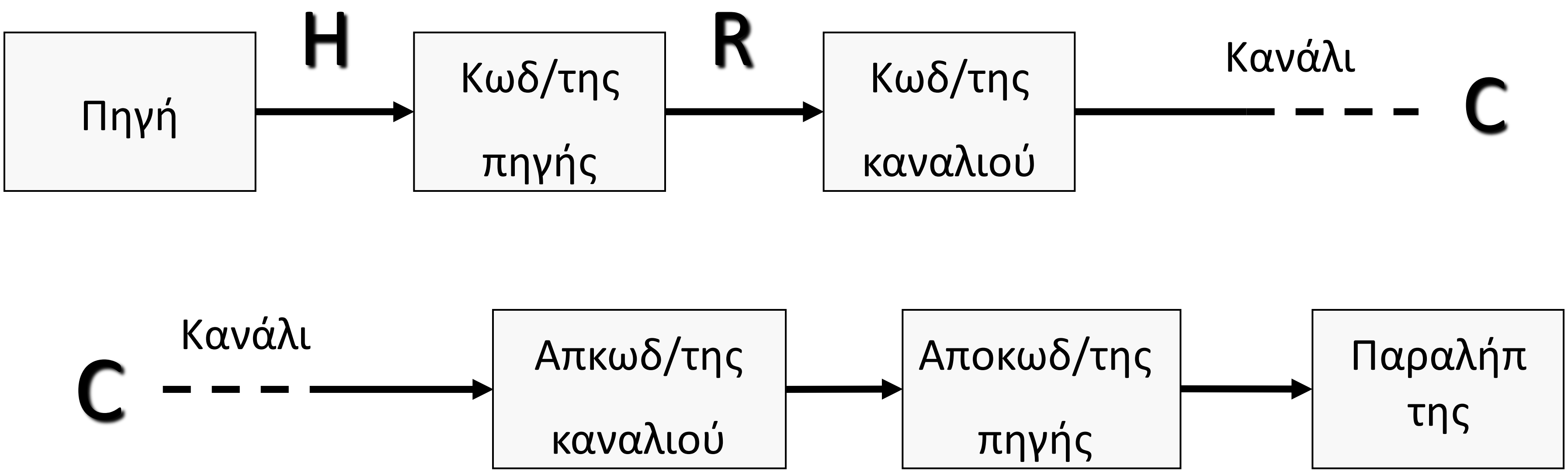
- Τι είναι πληροφορία; Αποφεύγοντας ερωτήματα σχετικά με το νόημα, ο Shannon έδειξε πως είναι μετρήσιμη ποσότητα.
- Σήμερα η διορατικότητα του Shannon βοήθησε στον εικονικό σχηματισμό όλων των συστημάτων που αποθηκεύουν, επεξεργάζονται ή ανταλλάσουν ψηφιακή πληροφορία, από τα CDs σε υπολογιστές έως αναμεταδότες σε διαστημικούς σταθμούς.
- Η θεωρία πληροφορίας έχει ξεπεράσει τα όρια των επικοινωνιών, διεισδύοντας στα πεδία της ψυχολογίας, των οικονομικών, της βιολογίας ακόμα και των τεχνών.

Στάδια μετάδοσης πληροφορίας



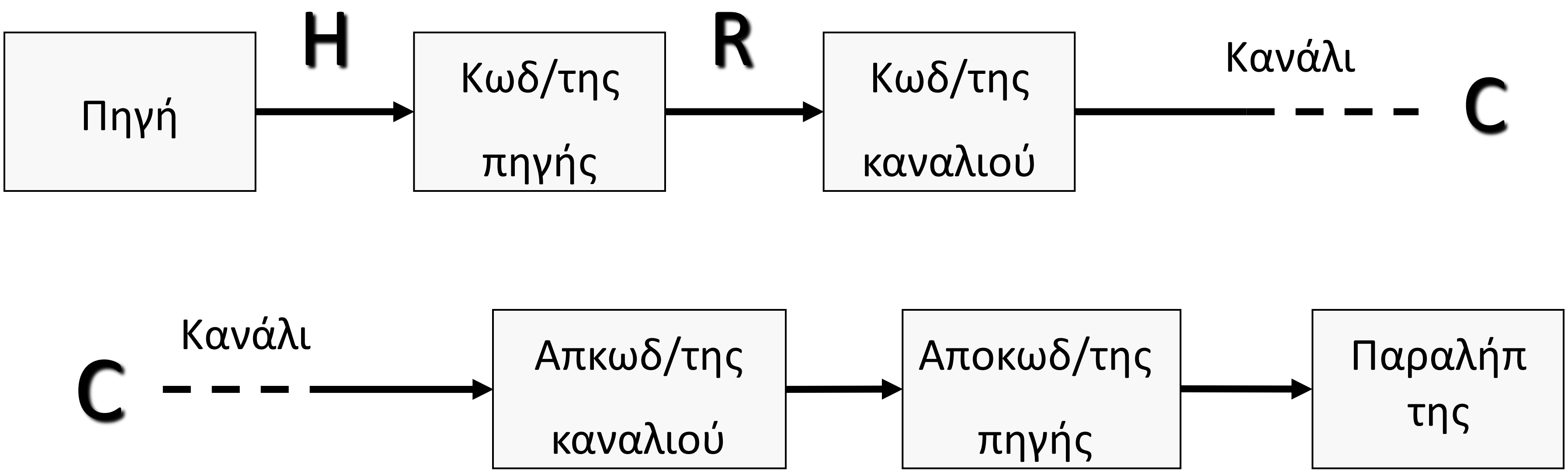
Κανάλι θεωρήται οτιδήποτε μεταφέρει ή αποθηκεύει πληροφορία,
πχ κεραία ράδιο, καλώδιο, CD, ένα κομμάτι χαρτί

Βασικές οντότητες



- **H**: Η πληροφορία της πηγής
- **R**: Ρυθμός από των κωδικοποιητή πηγής.
- **C**: Χωριτικότητα καναλιού

Βασικά Θεωρήματα



- **Shannon 1**: Αποστολή χωρίς σφάλματα μόνο αν R¸H και C¸R.
- **Shannon 2**: Η κωδικοποίηση πηγής και καναλιού βελτιστοποιήται ανεξάρτητα, και δυαδικά σύμβολα μπορούν να είναι ο ενδιάμεσος τύπος. Υπόθεση: Assumption: Αυθαίρετα μεγάλες καθυστερήσης.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Στοχαστικές Πηγές

Μάθημα: Θέματα Συστημάτων Πολυμέσων
Ενότητα # 5: Βασική Θεωρία Πληροφορίας
Διδάσκων: Γεώργιος Κ. Πολύζος
Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών "Επιστήμη των Υπολογιστών"

Στοχαστικές Πηγές (1 από 2)

- Παραγώμενα σύμβολα πηγής X_1, X_2, ...
- Κάθε σύμβολο παίρνει τιμή από το σύνολο A = (a_1, a_2, …).
  - Πχ. Το κείμενο είναι ακολουθία συμβόλων με τιμές στο αλφάβητο A = (a, …, z, A, …, Z, 1, 2, …9, !, ?, …).
  - Πχ. Μια ψηφιακή ασπρόμαυρη εικόνα είναι ακολουθία συμβόλων στο αλφάβητο A = (0,1) ή A = (0,..,255)

Στοχαστικές Πηγές (2 από 2)

- Μοντέλο: $P(X_1, X_2, \ldots, X_n)$ θεωρούνται γνωστά για όλους τους συνδυασμούς.

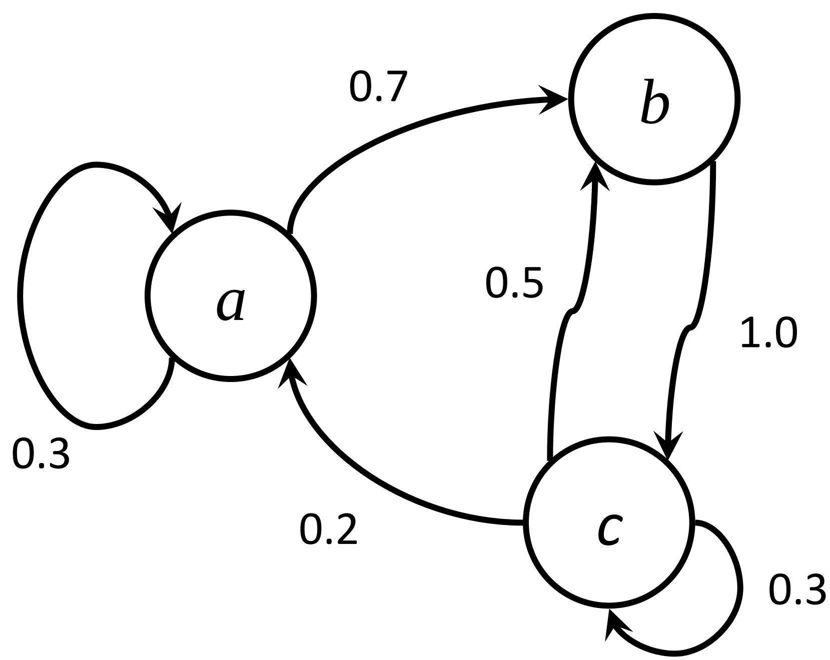
$$\boxed{\text{Πηγή}} \longrightarrow X_1, X_2, \ldots$$

Δύο ειδικές περιπτώσεις

1. Πηγή χωρίς μνήμη
   - Κάθε σύμβολο είναι ανεξάρτητο του προηγούμενου.
   - $P(X_1, X_2, \ldots, X_n) = P(X_1) \not\subset P(X_2) \not\subset \ldots \not\subset P(X_n)$
2. Μαρκοβιανή Πηγή ή Πηγή Markov
   - Κάθε σύμβολο εξαρτάται από το προηγούμενο.
   - $P(X_1, X_2, \ldots, X_n) = P(X_1) \ \ P(X_2|X_1) \not\subset P(X_3|X_2) \not\subset \ldots \not\subset P(X_n/X_{n-1})$

Πηγή Markov (1 από 4)

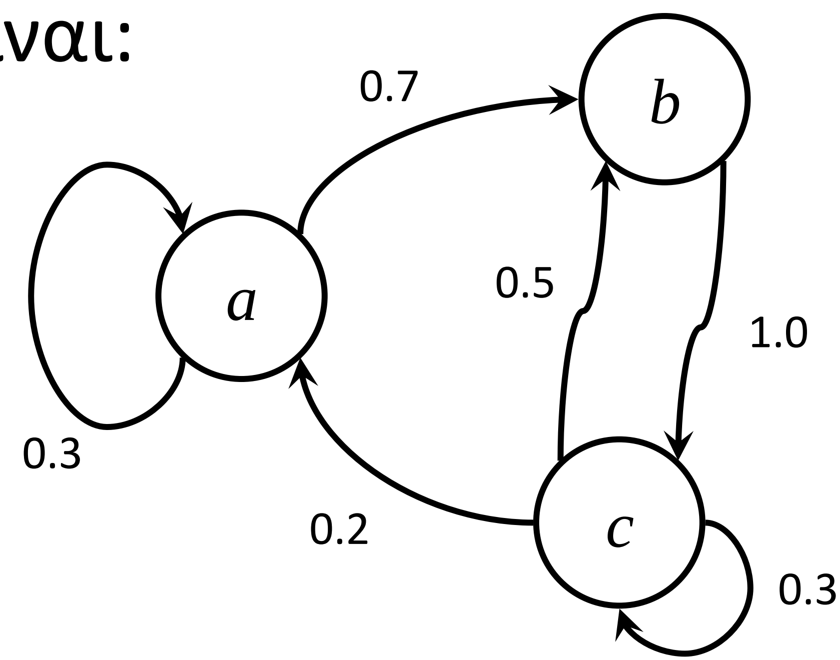
- Ένα σύμβολο εξαρτάται από το προηγούμενο σύμβολο, και η πηγή μοντελοποιείται με διάγραμμα καταστάσεων.



Τριαδική πηγή με αφάβητο A = (*a, b, c*).

Πηγή Markov (2 από 4)

- Έστω ότι βρισκόμαστε στην κατάσταση a, i.e., $X_k = a$.
- Οι πιθανότητες για το επόμενο σύμβολο είναι:



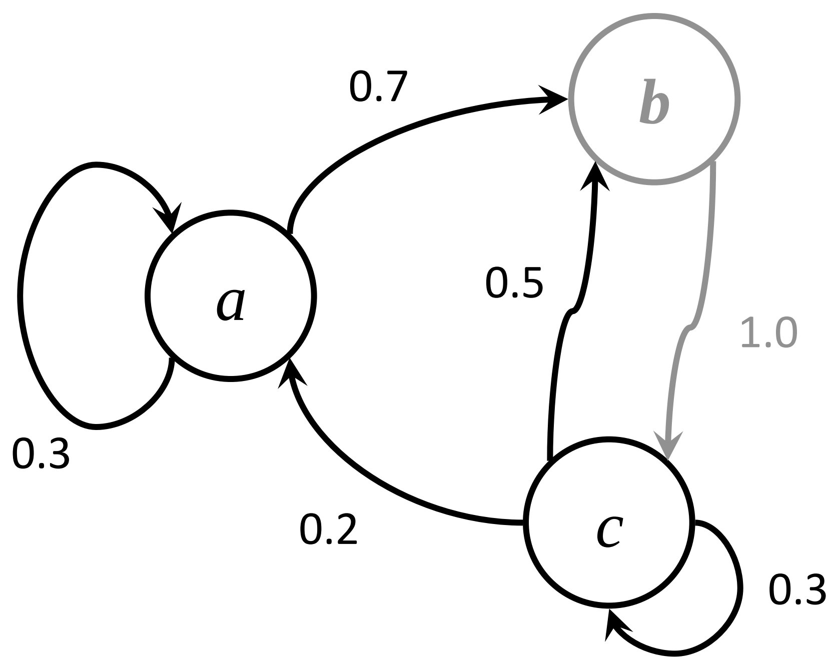
$$P(X_{k+1} = a \mid X_k = a) = 0.3$$

$$P(X_{k+1} = b \mid X_k = a) = 0.7$$

$$P(X_{k+1} = c \mid X_k = a) = 0$$

Πηγή Markov (3 από 4)

- Αν $X_{k+1} = b$, ξέρουμε ότι X_{k+2} θα ισούται με c.



$P(X_{k+2} = a \mid X_{k+1} = b) = 0$

$P(X_{k+2} = b \mid X_{k+1} = b) = 0$

$P(X_{k+2} = c \mid X_{k+1} = b) = 1$

Πηγή Markov (4 από 4)

- Αν κάθε κατάσταση είναι πιθανή, οι στατικές πιθανότητες (stationary probabilities) των καταστάσεων υπολογίζονται από τις πιθανότητες μετάβασης.
  - Στατικές πιθανότητες: $\omega_i = P(X_k = a_i)$ για κάθε k όταν $X_{k\text{-}1}, X_{k\text{-}2}, \ldots$ δεν δίνονται
- Τα μοντέλα Markov αναπαριστούν πηγές με εξαρτήσεις πάνω από ένα βήματα πίσω.
  - Χρήση διαγράμματος καταστάσεων με πολλά σύμβολα σε κάθε κατάσταση

Ανάλυση και σύνθεση

- Τα στοχαστικά μοντέλα χρησιμεύουν στην ανάλυση μιας πηγής.
  – Κατασκευή μοντέλου που αναπαριστά τον κόσμο και ανάλυση του μοντέλου αντί του πραγματικού κόσμου
- Τα στοχαστικά μοντέλα χρησιμεύουν στην σύνθεση μιας πηγής.
  – Προσομοίωση πηγής: Χρήση μήτρας τυχαίων αριθμών σε κάθε βήμα του μαρκοβιανού μοντέλου για την παραγωγή ακολουθίας συμβόλων

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Πληροφορία και εντροπία

Μάθημα: Θέματα Συστημάτων Πολυμέσων

Ενότητα # 5: Βασική Θεωρία Πληροφορίας

Διδάσκων: Γεώργιος Κ. Πολύζος

Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών “Επιστήμη των Υπολογιστών”

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Πληροφορία και εντροπία

- Έστω δυαδική πηγή χωρίς μνήμη, πχ κέρμα. Πόση πληροφορία παίρνουμε όταν μαθαίνουμε ότι ήρθε ”γράμματα”;
  - Αν είναι δίκαιο κέρμα, δηλ. *P*(*heads*) = *P* (*tails*) = 0.5, η ποσότητα της πληροφορίας είναι **1 bit**.
  - Αν γνωρίζαμε πως θα ερχόταν ”γράμματα”, δηλ. *P*(*heads*) = 1, η ποσότητα της πληροφορίας είναι **μηδέν**!
  - Αν δεν είναι δίκαιο κέρμα, δηλ. *P*(*heads*) = 0.9 , η ποσότητα της πληροφορίας είναι **πάνω από μηδέν αλλά λιγότερη από 1bit**.
  - Διαισθητικά, η ποσότητα πληροφορίας είναι **ίδια** αν *P*(*heads*) = 0.9 ή *P* (*heads*) = 0.1.

Εσωτερική πληροφορία (1 από 4)

- Μελετώντας το όπως ο Shannon.
- Θεωρούμε πηγή χωρίς μνήμη
  - Αλφάβητο A = (a_1, ..., a_n)
  - Πιθανότητες συμβόλων (p_1, ..., p_n).
- Πόση πιθανότητα παίρνουμε βρίσκοντας το επόμενο σύμβολο ίσο με a_i?
- Σύμφωνα με τον Shannon η εσωτερική πληροφορία του a_i είναι $I(a_i) = log(1/p_i)$

Εσωτερική πληροφορία (2 από 4)

- Έστω δύο ανεξάρτητα γεγονότα A και B, με πιθανότητες $P(A) = p_A$ and $P(B) = p_B$.
- Η πιθανότητα να συμβούν και τα δύο γεγονότα είναι p_A ¢ $p_{B.}$ Ωστόσο η ποσότητα της πληροφορίας πρέπει να προσταιθεί όχι να πολλαπλασιαστεί.

$$I(p_A \circ p_B) = I(p_A) + I(p_{\mathrm{B}})$$

- Οι λογάριθμοι ικανοποιούν $I(\mathrm{A}) = log(p_A)$?
- Θέλουμε η πιθανότητα να αυξάνεται για φθήνουσες πιθανότητες και χρησιμοποιούμε τον αρνητικό λογάριθμο.

$$I(\mathrm{A}) = -log(p_A) = log\ 1/p_A$$

Εσωτερική πληροφορία (3 από 4)

- Παράδειγμα 1ο:

$$p_i = 1 => I(1) = log\ 1/1 = 0$$

- Παράδειγμα 2ο:

$$p_i = 0.5 => I(0.5) = log_2\ 1/0.5 = 1\ [bit]$$

- Ποιος αλγόριθμος; Επιλέξτε έναν.
  - Στον φυσικό λογάριθμο η μονάδα μέτρησης είναι τα nats.
  - Στον δεκαδικό τα Harleys
  - Στον δυαδικό τα bits

Εσωτερική πληροφορία (4 από 4)

- Στον μέσο όρο όλων των συμβόλων, παίρνουμε

$$\text{H} = \sum_{1}^{\mathrm{N}} p_i \, I(a_i) = \sum_{1}^{\mathrm{N}} p_i \, log 1/p_i$$

- Το H(X) ονομάζεται εντροπία πηγής πρώτου βαθμού
- Εκφράζει τον βαθμό αβεβαιότητας για το επόμενο σύμβολο

Εντροπία

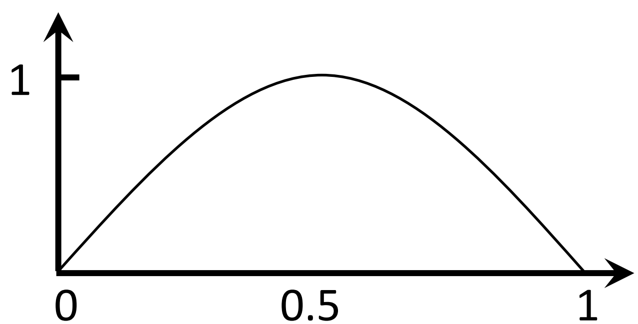
Παράδειγμα δυαδικής πηγής χωρίς μνήμη

BMS → 0 1 1 0 1 0 0 0 …

- Έστω: $\mathrm{p} = \mathrm{P}(\mathrm{X}_k = 1)$

$$\mathrm{q} = \mathrm{P}(\mathrm{X}_k = 0) = 1 - \mathrm{plog}\frac{1}{1-p}$$

- Τότε: $\mathrm{H} = \mathrm{plog}\frac{1}{p} + (1\ - p)\log\frac{1}{1-p}$



- Η αβεβαιότητα μεγιστοποιείται για p=q=1/2

3 ιδιότητες της εντροπίας

1. Αποδεικνύεται ότι 0 · *H* · *log N*.
2. Μέγιστη εντροπία (*H = log N*) έχουμε όταν τα γεγονότα είναι ισοπίθανα, δηλ $p_i = 1/N$.
3. Η διαφορά *log N – H* ονομάζεται πλεονασμός πηγής.

Εντροπία πηγής χωρίς μνήμη

Μάθημα: Θέματα Συστημάτων Πολυμέσων

Ενότητα # 5: Βασική Θεωρία Πληροφορίας

Διδάσκων: Γεώργιος Κ. Πολύζος

Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών "Επιστήμη των Υπολογιστών"

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Εντροπία πηγής χωρίς μνήμη

- Θεωρούμε ομάδα παραγόμενων συμβόλων $(X_1, \ldots, X_n)$ και ορίζουμε την ομαδική εντροπία:

$$H(X_1,..,Xn) = \sum_{1}^{N^n} P(X_1,..,Xn) log \frac{1}{P(X_1,..,Xn)}$$

- Η γενίκευση γίνεται σε κάθε πιθανό συνδυασμό των *n* συμβόλων.

- Η εντροπία πηγής με μνήμη δίνεται από τον τύπο:

$$H_\infty = \lim_{n \to \infty} \frac{1}{n} H(X_1, .., Xn)$$

- Όταν το μήκος της ομάδας απειρίζεται, διαιρούμε με *n* για να βρούμε τον αριθμό των bits ανά σύμβολο.

Εντροπία μαρκοβιανής πηγής

- Η εντροπία για την κατάσταση Sk εκφράζεται ως

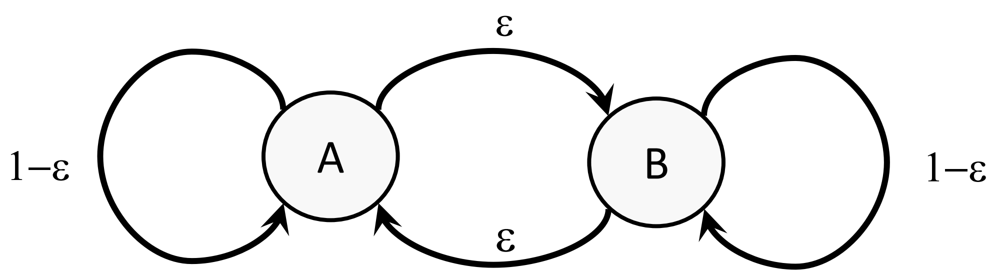
$$H(S_k) = \sum_{l=1}^{r} P_{kl} log \frac{1}{P_{kl}}$$

- Με P_{kl} την πιθανότητα μετάβασης από την κατάσταση *k* στην κατάσταση *l*.
- Η εντροπία μαρκοβιανής πηγής δίδεται από τον μέσο όρο όλων των καταστάσεων.

$$H_{\mathrm{M}} = \sum_{k=1}^{r} P(S_k) H(S_k)$$

Πηγή μήκους σειρών

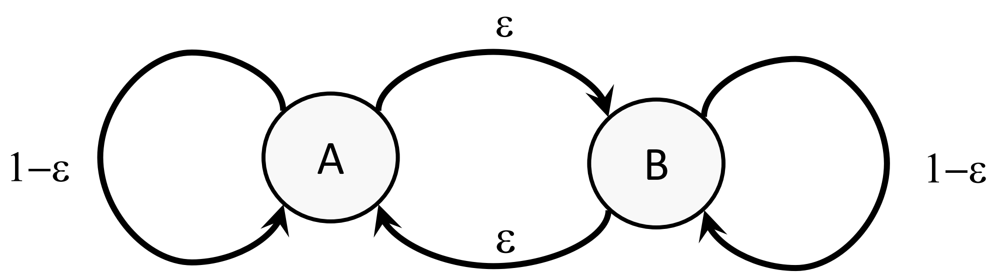
- Κάποιες πηγές παράγουν ακολουθίες ίδιων χαρακτήρων
- Πχ:



- Probability for a burst of length r: $P(r) = (1\text{-}\varepsilon)^{r\text{-}1}$¢$\varepsilon$
- Entropy: $H_R = - \sum_{\mathrm{r=1}}^{1} P(r) \log P(r)$
- If the average run length is ρ, then $H_R/\rho = H_M$.

Πηγή μήκους σειρών

- Κάποιες πηγές παράγουν σειρέςς ίδιων χαρακτήρων
- Πχ:



- Πιθανότητα παραγωγής σειράς μήκους *r*:
  $P(r) = (1-\varepsilon)^{r-1} \text{¢} \varepsilon$
- Εντροπία: $H_R = - \sum_{r=1}^{1} P(r) \log P(r)$
- Αν το μέσο μήκος σειρά είναι ρ, τότε $H_R/\rho = H_M$.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Θεώρημα κωδικοποίησης πηγής

Μάθημα: Θέματα Συστημάτων Πολυμέσων

Ενότητα # 5: Βασική Θεωρία Πληροφορίας

Διδάσκων: Γεώργιος Κ. Πολύζος

Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών "Επιστήμη των Υπολογιστών"

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Τυπικές Ακολουθίες (1 από 3)

- Έστω η μακρά ακολουθία της χωρίς μνήμης πηγής με $P(1) = p$.
- Σε n bits, θα υπάρχουν περίπου $w = n ¢ p$ *"ένα"*.
- Άρα, υπάρχουν $M = (n \text{ over } w)$ τέτοιες τυπικές ακολουθίες!
- Μας ενδιαφέρουν μόνο αυτές οι ακολουθίες. Οι υπόλοιπες εμφανίζονται με μικρότερη πιθανότητα.

Τυπικές Ακολουθίες (2 από 3)

- Πόσες τυπικές ακολουθίες υπάρχουν;

$$M = \binom{n}{w} = \frac{n!}{w!\,(n-w)!}$$

$$(Stirling\ : n! \simeq \frac{n^n}{e^n}\sqrt{2\pi n})$$

$$M = \binom{n}{w} = \frac{n!}{w!(n-w)!} \simeq \frac{n^n}{w^w (n-w)^{n-w}} * \frac{const}{\sqrt{n}}$$

- Η απαρίθμηση χρειάζεται log M bits, δηλ. $\frac{1}{n} logM$ bits ανά σύμβολο.

Τυπικές Ακολουθίες (3 από 3)

- Πόσα bits χρειαζόμαστε;

Το θεώρημα κωδικοποίησης πηγής

- Μας λέει
  - Μπορούμε να αναπαραστήσουμε τα παραγόμενα σύμβολα μιας πηγής X χρησιμοποιώντας $H(X)$ bits/σύμβολο.
  - Δε μπορεί να γίνει κάτι καλύτερο
- Δε μας λέει
  - Πως να το κάνουμε

Τέλος Ενότητας # 5

Μάθημα: Θέματα Συστημάτων Πολυμέσων
Ενότητα # 5: Βασική Θεωρία Πληροφορίας
Διδάσκων: Γεώργιος Κ. Πολύζος
Τμήμα: Μεταπτυχιακό Πρόγραμμα Σπουδών "Επιστήμη των Υπολογιστών"

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης