Δ1

Συμπίεση Δεδομένων

2013-2014

Δρ. Ν. Π. Σγούρος

Συμπίεση Δεδομένων

Δ1

- Αποτελεί ένα μέσο μείωσης του όγκου σε ένα σύνολο από δεδομένα διατηρώντας ωστόσο το νόημα των δεδομένων αυτών
- Αποτελεί μία τεχνική μείωσης του πλεονασμού σε ένα σύνολο από δεδομένα
- Αποτελεί μία τεχνική μείωσης του χρόνου και του κόστους μετάδοσης των δεδομένων
- Υπάρχουν πολλοί διαφορετικοί αλγόριθμοι συμπίεσης προσανατολισμένοι σε διαφορετικά σύνολα δεδομένων και με διαφορετικούς στόχους

Περιεχόμενα Α' Μέρους Ι

Δ1

Περίγραμμα Α' Μερους

- Εισαγωγή στη Συμπίεση Δεδομένων
- Επανάληψη Βασικών Εννοιών από Θεωρία Πληροφορίας
- Κωδικοποίηση Εξόδου Πηγής με Σύμβολα Διακριτά και Στατιστικά Ανεξάρτητα, Κώδικες Εντροπίας
- Κωδικοποίηση Εξόδου Πηγής αναλογικού Σήματος

Περιεχόμενα Α' Μέρους ΙΙ

Δ1

Βιβλιογραφία

- Διαφάνειες Μαθήματος (Μ. Σαγκριώτης , Ν. Σγούρος)
- J. G. Proakis, M. Salehi "Communication Systems Engineering" Prentice Hall, 2001 (Ελληνική Μετάφραση)
- D. MacKay "Information Theory, Inference and Learning Algorithms" Cambridge University Press, 2003
- D. Salomon, "Data Compression, The Complete Reference" Springer, 2007
- J. G. Proakis, D. K. Manolakis "Digital Signal Processing", 4th edition, Prentice Hall, 2006

Δεδομένα Ι

Δ1

- Ποσοτική Μονάδα μέτρησης μεγέθους – πλήθους - όγκου δεδομένων
  - **b**inary dig**it** - **bit** : Είναι η βασική μονάδα μέτρησης μιας ποσότητας ή όγκου δεδομένων. Δύο δυνατές τιμές (π.χ. 0/1, Yes/No, On/Off)
  - Πρώτη χρήση : Διάτρητες κάρτες (Bouchon-1732)
  - Πρώτη αναφορά ονόματος: Tuckey-1947 Shannon 1948
  - Συμβολισμός: ***bit*** ISO/IEC 80000-13 (2008) ή b IEEE 1541 (2002)

Δεδομένα ΙΙ

Δ1

- Πολλαπλάσια Μονάδων μέτρησης πλήθους-όγκου δεδομένων
- Για πολλές δεκαετίες χρησιμοποιείται διπλή αναπαράσταση για τα πολλαπλάσια π.χ. Kilo 2^{10} =1024 Kilo 10^3=1000

Πρόθεμα	Δυαδική ÷ Δεκαδική	Ποσοστιαία Διαφορά
Kilo	1.024	+2.4%
Mega	1.049	+4.9%
Giga	1.074	+7.4%
Tera	1.100	+10.0%
Peta	1.126	+12.6%
Exa	1.153	+15.3%
Zetta	1.181	+18.1%
Yotta	1.209	+20.9%

- ISO/IEC 80000-13 (2008) : Εναρμόνιση προθεμάτων όπως αναφέρονται στο SI για το σύνολο των φυσικών και μαθηματικών μεγεθών

Δεδομένα III

Δ1

- Στην περίπτωση που χρησιμοποιούνται δυνάμεις του 2 τότε αποκλειστικά δηλώνεται διαφορετικό πρόθεμα

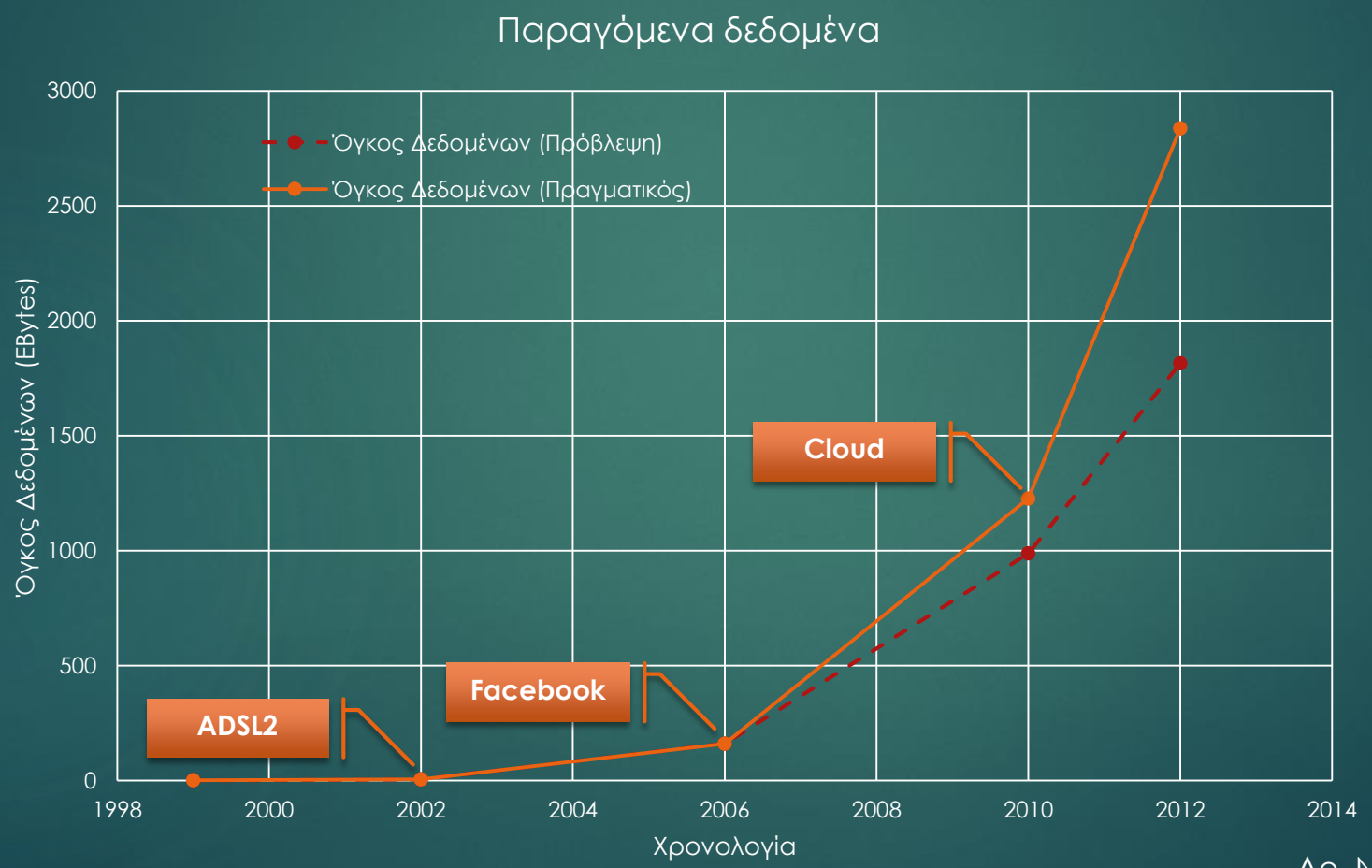
Σύμβολο	Πρόθεμα	Μονάδα	Μέγεθος
Ki	Kibi	Kibibyte (KiB)	2^{10} bytes
Mi	Mebi	Mebibyte (MiB)	2^{20} bytes
Gi	Gibi	Gibibyte (GiB)	2^{30} bytes
Ti	Tebi	Tebibyte (TiB)	2^{40} bytes
Pi	Pebi	Pebibyte (PiB)	2^{50} bytes
Ei	Exbi	Exbibyte (EiB)	2^{60} bytes
Zi	Zebi	Zebibyte (ZiB)	2^{70} bytes
Yi	Yobi	Yobibyte (EiB)	2^{80} bytes

- ISO/IEC 80000-13 (2008)

Όγκος παραγόμενων δεδομένων

Δ1

- Ετήσιος παγκόσμιος όγκος παραγόμενων δεδομένων



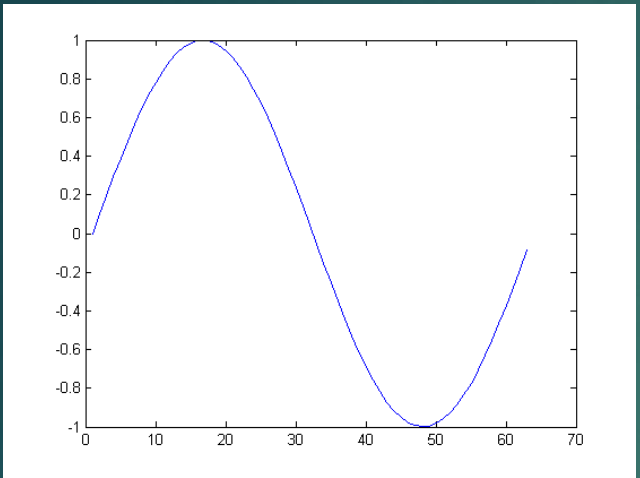
Αναλογικά – Ψηφιακά Σήματα

Δ1

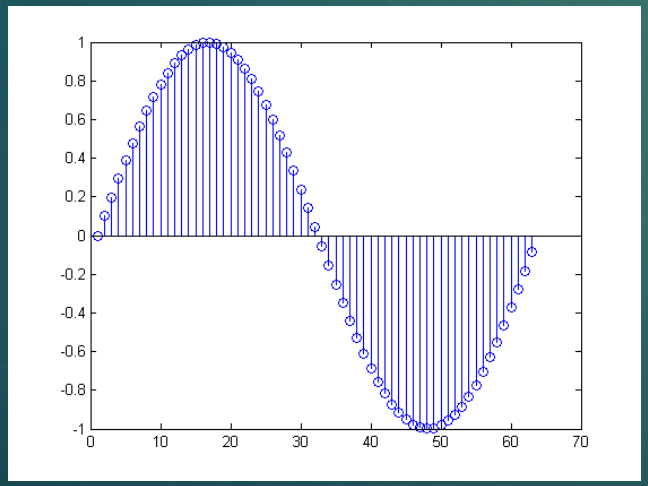
- Αναλογικό Σήμα
  - $x(t) \qquad , t \in [t_{min}, t_{max}] \, , x \in [x_{min}, x_{max}]$
- Δειγματοληψία
  - $t \to n \;\; , \;\; x(t) \to x(n) \;\; , \;\; n = 1, \dots, N$
- Κβάντιση
  - $x(n) \to \hat{x}(n)$

Παράδειγμα

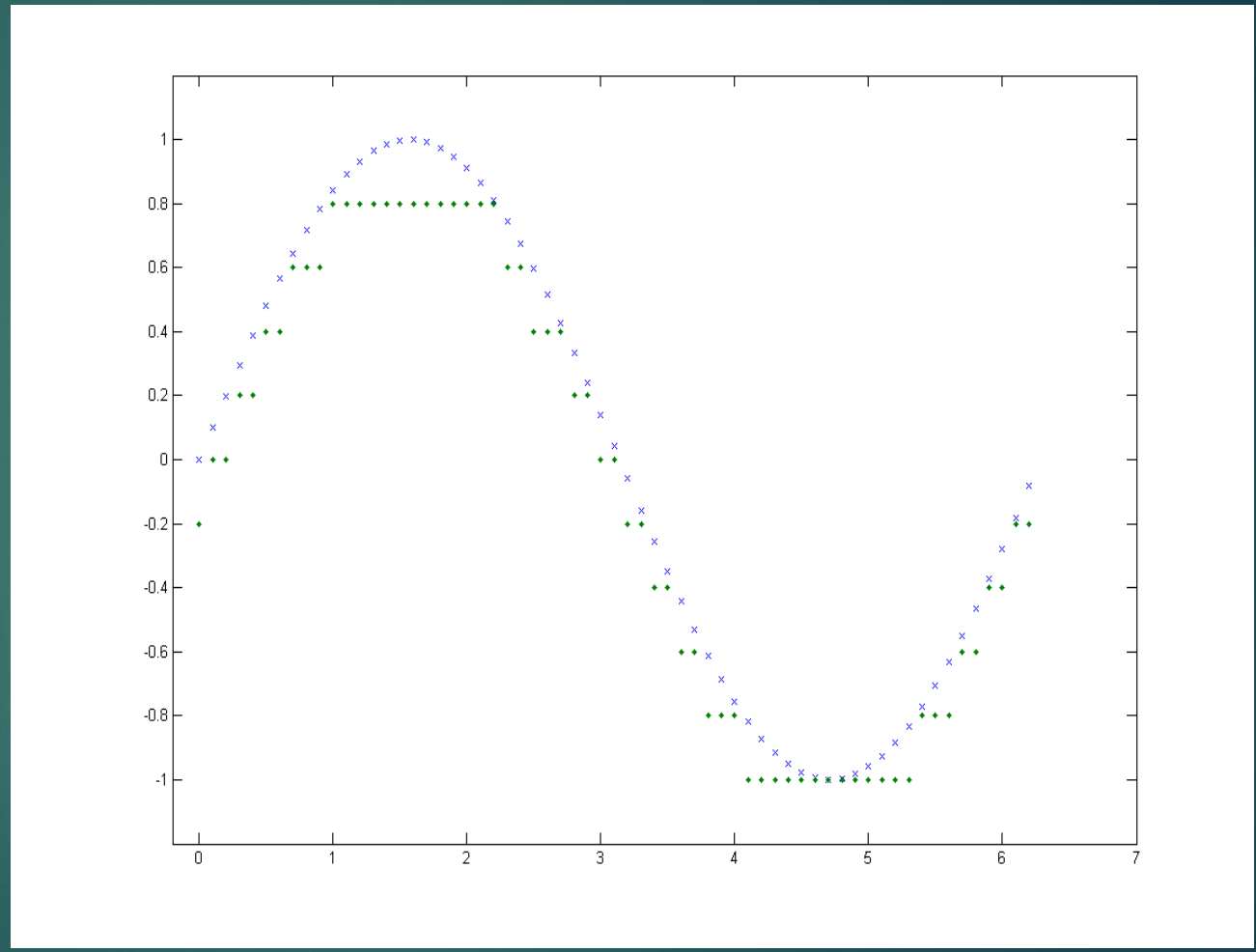
Δ1



Αναλογικό σήμα



Δειγματοληψία



Κβάντιση

Άσκηση 1.1

Δ1

- Ποια η ποσότητα δεδομένων που μεταφέρεται στον ανθρώπινο εγκέφαλο όταν κάποιος ακούει μουσική για χρονικό διάστημα μίας ώρας
  - Εύρος ακουστικών συχνοτήτων 20Hz έως 22kHz (16bits/δείγμα)
- Ποια η ποσότητα δεδομένων που μεταφέρεται στον ανθρώπινο εγκέφαλο από τα μάτια του σε διάστημα μίας ώρας
  - Πλήθος λήψεων ματιού: ~30 λήψεις /s
  - Ανάλυση αισθητήρα ματιού: ~576 Mpixels (Χρωματική απόκριση ~12bits/pixel)
- Ποια η συνολική ποσότητα δεδομένων που μεταφέρονται στον εγκέφαλο ανά δευτερόλεπτο (ρυθμός) από τις δύο παραπάνω πηγές

Άσκηση 1.2

Δ1

- Ποιος ο ρυθμός μετάδοσης για το επερχόμενο πρότυπο τηλεόρασης Ultra High Definition Television (UHDTV) – 4K;

  - **Εικόνα**
  - Πλήθος pixels: 2160 x 3840 pixels/πλαίσιο
  - Χρώμα: 10-12 bits/pixel
  - Ρυθμός Πλαισίων: 60 fps έως 120fps (120Hz)

  - **Ήχος**
  - Κανάλια : 24
  - Ρυθμός Δειγματοληψίας: 48kHz-96kHz
  - Κβάντιση: 16/20/24 bits/δείγμα

Ψηφιακά μέσα αποθήκευσης

Δ1

- Οπτικοί Δίσκοι (CD): 700 MB = 5.6 Gbit
- Δίσκοι Ψηφιακού Video (DVD): 4,7GB = 37,6 Gbit
- Δίσκοι Blu-Ray (BD-XL): 128GB = 1 Tbit
- Σκληροί Δίσκοι: 4 TB = 32 Tbit
- Χωρητικότητα 1gr DNA[1]: 700TB = 5,6 Ebit

[1] Church, G. M.; Gao, Y. & Kosuri, S. "Next-Generation Digital Information Storage in DNA," Science, 2012, 337, 1628

Τηλεπικοινωνιακά κανάλια

Δ1

- PSTN: = 56 Kbps
- ISDN: 2 κανάλια x 64Kbps+16Kbps (σημ.) = 144 Kbps
- T1: 24 κανάλια x 64Kbit/κανάλι = 1.544 Mbps
- T3: 672κανάλια x 64 KbitKανάλι = 44.7 Mbps
- ADSL2+: = 24 Mbps
- VDSL2 = 100 MBps

- GSM(2G): = 14.4Kbps
- UMTS(3G): = 384Kbps

Βασικές Αρχές Συμπίεσης Δεδομένων

Δ1

1 Ασυσχέτιστα δεδομένα

- Αντιστοίχιση μικρών κωδικών λέξεων σε δεδομένα με μεγάλη πιθανότητα εμφάνισης

2 Συσχετισμένα δεδομένα

- Αφαίρεση – Μείωση της συσχέτισης – Πλεονασμού και έπειτα χρήση της παραπάνω τεχνικής

3 Μη καταληπτά δεδομένα

- Αφαίρεση – Μείωση των μη καταληπτών δεδομένων από τον χρήστη

Πληροφορία Ι

Δ1

- Έστω πηγή που εκπέμπει διακριτά σύμβολα S_i από ένα πεπερασμένο σύνολο C: $S_i \in C = \{A, B, \dots, \Omega\}$.
- Για κάθε $S_i \in C$ ορίζουμε μία πιθανότητα εμφάνισης p_i :

$$0 \leq p_i \leq 1, \qquad \sum_i p_i = 1$$

- Λογική **θεώρηση** «μικρή πιθανότητα → μεγάλη πληροφορία»
  - «ΨΗΦΙΟ» , «Ψ_Φ_ _» , «_ _ _ΙΟ»
- $$\text{Πληροφορία} : I_{(S_i)} \equiv I_{(p_i)} \propto \frac{1}{p_i}$$
  - Θεωρώντας επιπλέον ότι τα S_i εκπέμπονται ανεξάρτητα και ακολουθούν όμοιες κατανομές (i.i.d) τότε για μια λέξη με δύο σύμβολα : $p_{ij} = p_i p_j$.
  - Λογική θεώρηση «Η πληροφορία των δύο συμβόλων είναι το άθροισμα της πληροφορίας που κουβαλά το κάθε σύμβολο»
- $$\text{Συνολική Πληροφορία} : I_{(p_{ij})} = I_{(p_i p_j)} = I_{(p_i)} + I_{(p_j)} \quad \text{(Cauchy)}$$

Πληροφορία ΙΙ

Δ1

- Επειδή η $I_{(p)}$ μονότονη (Γιατί ;)
- Με βάση τα προηγούμενα μία λογική εκλογή συνάρτησης για την ποσοτικοποίηση της πληροφορίας είναι η $I_{(p_i)} = \kappa \cdot log p_i$
- Επιλέγοντας $\kappa = -1$ (γιατί;) προκύπτει $I_{(p_i)} = log \frac{1}{p_i}$
- Ορίζουμε ως πληροφορία του συμβόλου $\boldsymbol{S_i}$ την ποσότητα

$$\boldsymbol{I_{(S_i)} \equiv I_{(p_i)} = log \frac{1}{p_i}}$$

- Ανάλογα με τη βάση του λογαρίθμου ορίζονται και οι μονάδες

Βάση	Μονάδες
10	Hartley ή dit
e	nat
2	**bit**

Εντροπία

Δ1

- Ένα καλό **στατιστικό μέτρο της πληροφορίας** μιας διακριτής πηγής χωρίς μνήμη (DMS) είναι η **εντροπία**, η οποία αντιστοιχεί στη **μέση πληροφορία ανα σύμβολο της πηγής**

$$H(s) = \sum_{i=1}^{N} p_i \cdot log_2 \frac{1}{p_i}$$

Άσκηση 1.3

Δ1

- Μία πηγή χωρίς μνήμη η οποία μπορεί να εκπέμψει Ν διαφορετικά σύμβολα $\mathrm{S}_i, \mathrm{i} = 1, \ldots, \mathrm{N}$ με πιθανότητες εκπομπής p_i , εκπέμπει μία σειρά από λ σύμβολα. Να δείξετε ότι

$$H(s) = \sum_{i=1}^{N} p_i \cdot log_2 \frac{1}{p_i}$$

Άσκηση 1.4

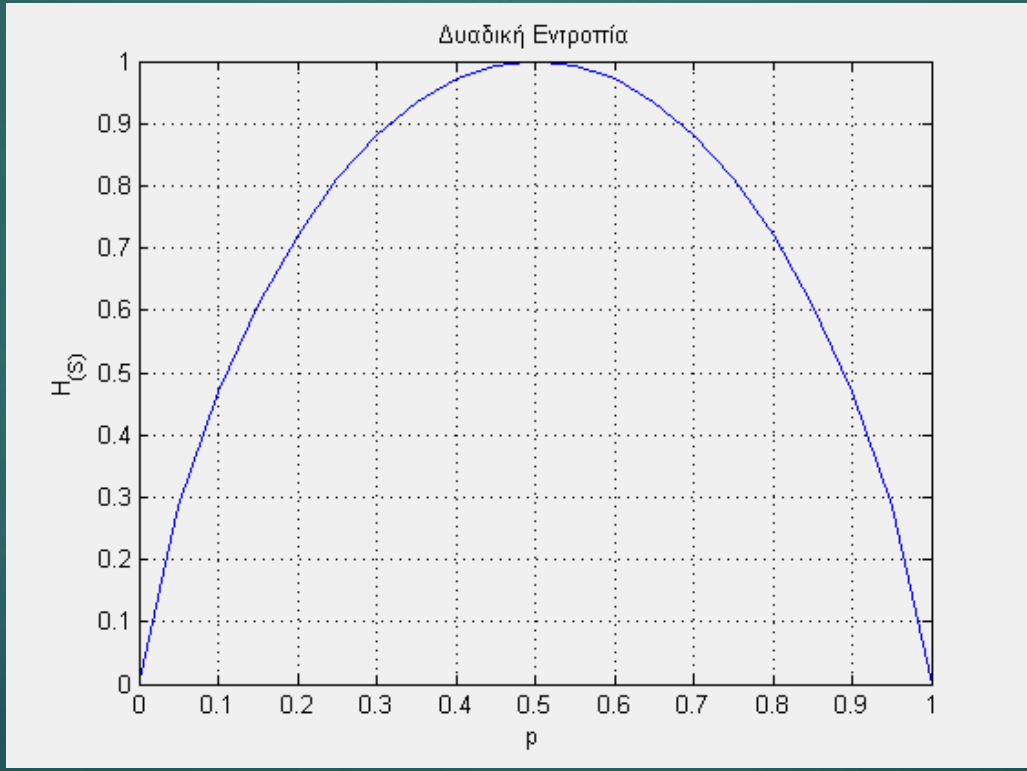
Δ1

- Να υπολογίσετε την εντροπία μιας πηγής που εκπέμπει δύο σύμβολα (Δυαδική) με πιθανότητες p και $1-p$. Ποια είναι η μέγιστη μέση ποσότητα πληροφορίας ανα σύμβολο για τη συγκεκριμένη πηγή;

Άσκηση 1.4

Δ1

- Να υπολογίσετε την εντροπία μιας πηγής που εκπέμπει δύο σύμβολα (Δυαδική) με πιθανότητες p και $1-p$. Ποια είναι η μέγιστη μέση ποσότητα πληροφορίας ανα σύμβολο για τη συγκεκριμένη πηγή;



Παράδειγμα

Δ1

- Για μία πηγή που εκπέμπει Ν σύμβολα ισχύει ότι $0 \leq \mathrm{H}_{(S)} \leq log_2 N$. Ειδικά ισχύει ότι $\mathrm{H}_{(S)} = 0$ όταν υπάρχει σύμβολο με $\mathrm{p}_i = 1$ και $\mathrm{H}_{(S)} = log_2 N$ όταν $\mathrm{p}_i = 1/\mathrm{N}$ για κάθε i.
- $\mathrm{H}_{(S)} \geq 0$
  - Επειδή για κάθε σύμβολο $\mathrm{p}_i \leq 1$ προκύπτει ότι $\mathrm{p}_i \cdot log_2 \frac{1}{p_i} \geq 0$ άρα $\mathrm{H}_{(S)} \geq 0$
  - Ειδικά $\mathrm{H}_{(S)} = 0$ όταν υπάρχει σύμβολο με $\mathrm{p}_i = 1$ (ή $\mathrm{p}_i = 0$)

Παράδειγμα

Δ1

- Για μία πηγή που εκπέμπει Ν σύμβολα ισχύει ότι $0 \leq \mathrm{H}_{(S)} \leq log_2 N$. Ειδικά ισχύει ότι $\mathrm{H}_{(S)} = 0$ όταν υπάρχει σύμβολο με $\mathrm{p}_i = 1$ και $\mathrm{H}_{(S)} = log_2 N$ όταν $\mathrm{p}_i = 1/\mathrm{N}$ για κάθε i.
- $\mathrm{H}_{(S)} \leq log_2 N$
  - Θεωρούμε την ποσότητα $\mathrm{H}_{(S)} - log_2 N$ οπότε προκύπτει ότι :
  - $\mathrm{H}_{(S)} - log_2 N = \sum_{i=1}^{N} \mathrm{p}_i \cdot log_2 \frac{1}{p_i} - log_2 N \cdot \sum_{i=1}^{N} \mathrm{p}_i \Rightarrow$

    $\mathrm{H}_{(S)} - log_2 N \sum_{i=1}^{N} \mathrm{p}_i \cdot log_2 \frac{1}{Np_i} \xrightarrow{\ln x \leq x-1 \ \forall x>0,\ log_2 x = \frac{\ln x}{log_2 e}}$
  - $\mathrm{H}_{(S)} - log_2 N \leq log_2 e \sum_{i=1}^{N} \mathrm{p}_i \cdot (\frac{1}{Np_i} - 1) \leq log_2 e \left(\sum_{i=1}^{N} \mathrm{p}_i \cdot \frac{1}{Np_i} - \sum_{i=1}^{N} \mathrm{p}_i \right) \leq 0$

Πλεονασμός

Δ1

- Ένα από τα πιθανά μέτρα πλεονασμού είναι το

$$\Pi = 1 - \frac{\mathrm{H}_{(S)}}{log_2 N}$$

Κώδικες Πηγής

Δ1

- Αποτελούν μια απεικόνιση μίας σειράς συμβόλων που εκπέμπει μία πηγή σε μία σειρά συμβόλων από ένα αλφάβητο με τρόπο ώστε να είναι δυνατή η ανάκτηση των αρχικών συμβόλων με οσοδήποτε μικρή πιθανότητα σφάλματος.

- Συνήθως χρησιμοποιούνται για το αλφάβητο δυαδικές συμβολοσειρές

Κώδικες Πηγής

Δ1

- Ακριβής ανάκτηση των αρχικών συμβόλων → Μη απωλεστική κωδικοποίηση
- Μη ακριβής ανάκτηση των αρχικών συμβόλων → Απωλεστική κωδικοποίηση
- Υπάρχουν περιορισμοί στους κώδικες που μπορούν να χρησιμοποιηθούν οι οποίοι σχετίζονται με τη δυνατότητα μοναδικής αποκωδικοποίησης

Άσκηση 1.5

Δ1

- Χρησιμοποιήστε την απεικόνιση του παρακάτω πίνακα για να κωδικοποιήσετε και να αποκωδικοποιήσετε τη συμβολοσειρά $a_1a_0a_2a_3a_0$. Τι παρατηρείτε ;

Σύμβολο	Κώδικας
a_0	0
a_1	01
a_2	10
a_3	11