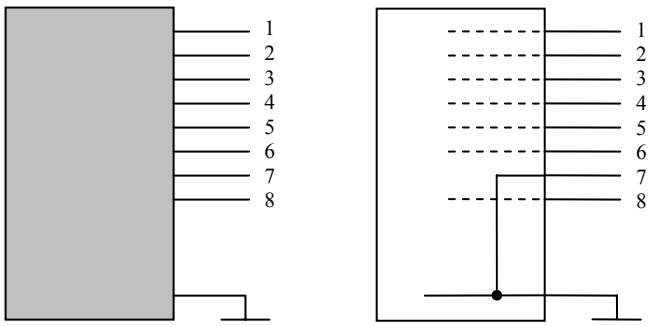
Elemente de teorie a informației

1. Câte ceva despre informație la modul subiectiv

În cele ce urmează vom face câteva considerații legate de informație şi măsurare a ei. După cum se cunoaşte informația se măsoară în biți. De asemenea şi dimensiunea unei magistrale, registru etc. se măsoară tot în biți. Deşi între cele două tipuri de biți (biți ”**de informație**” şi biți **”hardware”**) există o legătură, ele nu sunt tocmai identice. Chiar dacă cele două denumiri sunt oarecum forțate, considerăm că este foarte utilă delimitarea mai clară a lor, utilizarea noțiunii de bit în ambele contexte (hardware şi teoria informației) putând duce la confuzii importante.

Să considerăm următoarea situație. Un dispozitiv hardware are o magistrală de 8 biți. În acest caz pare corect să spunem că un fir al ei transmite un bit ”de informație” la fiecare tact. Dar dacă, privind în interiorul dispozitivului, constatăm că firul respectiv e legat la masă mai transmite el ceva informație la fiecare tact?



Să considerăm un alt exemplu: dăm cu banul. Putem spune că la fiecare aruncare cu banul obținem un nou bit ”de informație”. Dar, după ce am dat de 100 ori cu banul şi a ieşit de tot atâtea ori stema, mai obținem oare aceeaşi informație la o aruncare? Mai are rost să mai facem aruncări? Pare că informația nou obținută scade pe măsură ce stema devine în mod evident tot mai probabilă.

Să considerăm că avem o grupă de studenți şi notele acestora la două discipline la fel de importante. Care dintre note este mai relevantă (să zicem pentru un angajator): nota la disciplina la care toată lumea a obținut 10 sau la cea la care notele acoperă o plajă mai mare? Altfel zis, care notă transmite mai multă informație?

Din toate exemplele de mai sus observăm că informația este cu atât mai mare cu cât există o mai mică probabilitate a evenimentului respectiv. Un bit ”hardware” este doar **suportul fizic** pe care **s-ar putea transmite** un bit ”de informație” dacă predictibilitatea este minimă – adică dacă probabilitățile stărilor 0 şi 1 sunt egale (după cum se va demonstra la maximizarea entropiei). Dacă probabilitățile sunt mai dezechilibrate un bit ”hardware” transmite mai puțin decât un bit ”de

informație". Remarcăm faptul că noțiunea de bit "hardware" poate referi de exemplu şi un bit al unui fişier (care poate transmite şi el mai multă sau mai puțină informație).

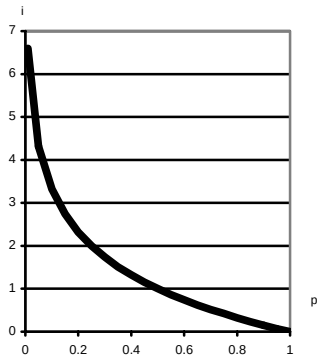
2. Definiții

Cea mai cunoscută **măsură a informației** este cea care leagă informația de **probabilitate** în mod logaritmic astfel:

$$i(x_i) = -\log p(x_i)$$

În această relație x_i indică un eveniment iar p_i probabilitatea acestuia. În aproape toate cazurile logaritmul se consideră a fi în baza 2. Această definiție a informației este datorată lui **Claude Shannon**, unanim acceptat ca "părintele teoriei informației".

În figura alăturată am reprezentat graficul acestei funcții. Se observă că această definiție corespunde cu observațiile anterioare, informația obținută pentru un eveniment sigur (de probabilitate 1) este nulă, ea crescând odată cu scăderea probabilității evenimentului.



Unitatea de măsură a informației este **bitul**. Acesta reprezintă cantitatea de informație care se obține prin producerea unui eveniment de probabilitate 0.5.

Se numesc **surse discrete** sursele care emit mesaje în formă discretă (spre deosebire de sursele care emit semnale continue, analogice).

Simbol (literă) este elementul fundamental, ireductibil, care conține o informație. Totalitatea simbolurilor care pot fi generate de o sursa constituie **alfabetul** sursei. O succesiune finită de simboluri formează un **cuvânt** iar mulțimea cuvintelor care pot fi formate cu un alfabet reprezintă o limbă.

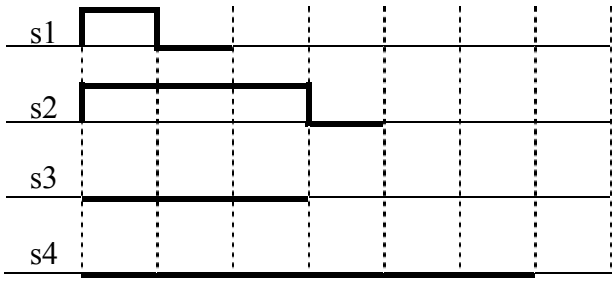
Codarea reprezintă stabilirea unei corespondențe între cuvintele formate cu un alfabet şi cuvinte formate cu alt alfabet.

În ceea ce priveşte sursele putem face următoarele caracterizări:

- o sursă discretă este **cu memorie / fără memorie** după cum probabilitatea de apariție a unui simbol depinde sau nu de simbolurile anterioare.

- o sursă este **nestaționară / staționară** după cum probabilitatea simbolurilor generate depinde sau nu de timp.
- o sursă este **cu debit controlabil / necontrolabil** după cum generarea de simboluri poate fi oprită (întârziată) din exterior sau nu. O sursă cu debit necontrolabil (de exemplu sursa obținută prin eşantionarea unui semnal analogic) ridică probleme legate de prelucrarea în timp real.

În figura următoare prezentăm cele 4 simboluri utilizate în cazul **codului Morse** (alfabetul sursei). Avem de a face cu punct (s1), linie (s2), spațiu între litere (s3) respectiv spațiu între cuvinte (s4). Acestea sunt singurele mesaje valide, care pot fi emise. Motivul pentru care am ales prezentarea acestei surse este faptul că în acest caz simbolurile au o reprezentare distinctă, foarte diferită de la un simbol la altul. Cu toate acestea considerăm în prelucrările legate de informație că sursa generează 4 simboluri fără să ne mai intereseze reprezentarea fizică a acestora. În cazul în care sursa este un fişier avem 256 simboluri diferite, dar având o reprezentare asemănătoare.



3. Entropia

Să considerăm o sursă S care emite simboluri cu probabilitățile P:

$$[S] = [s_1\ s_2\ \ldots\ s_n]$$

$$[P] = [p_1\ p_2\ \ldots\ p_n]$$

Ne interesează să evaluăm cantitatea de informație pe care o dă sursa. Pentru aceasta se defineşte **entropia sursei** astfel:

$$H(S) = \sum_{i=1}^{n} p_i\, i(s_i) = -\sum_{i=1}^{n} p_i \log p_i$$

Observăm că entropia reprezintă **informația medie pe simbol** (o medie a informației obținute pentru fiecare simbol, media fiind una ponderată cu probabilitățile de apariție ale simbolurilor). Entropia este deci egală cu incertitudinea medie apriori asupra evenimentelor [S]. Accentuăm faptul că entropia este o măsură a informației emise de sursă în ansamblul ei şi nu a informației emise de un simbol oarecare.

Din definiția entropiei observăm că ea este o funcție continuă şi simetrică în raport cu variabilele `pi`.

În continuare ne propunem să determinăm care este valoarea maximă a entropiei. Pentru aceasta căutăm maximul funcției

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i$$

între variabilele `pi` existând legătura (restricția)

$$\sum_{i=1}^{n} p_i - 1 = 0$$

Pentru a căuta un extrem cu restricții se foloseşte **metoda multiplicatorilor lui Lagrange**. Se construieşte funcția Φ

$$\Phi = -\sum_{i=1}^{n} p_i \log p_i + \lambda\left(\sum_{i=1}^{n} p_i - 1\right)$$

căreia i se caută extremul. Anulăm derivatele parțiale în raport cu variabilele `pi`

$$\frac{\partial \Phi}{\partial p_i} = 0$$

Obținem astfel:

$$\frac{\partial \Phi}{\partial p_i} = -\log p_i - \log e + \lambda = 0$$

$$\frac{\partial \Phi}{\partial p_j} = -\log p_j - \log e + \lambda = 0$$

Prin simplificare se obține:

$$\log p_i = \log p_j$$

adică:

$$p_i = p_j$$

Deoarece indicii `i` şi `j` sunt oarecare relația este valabilă pentru orice `i` şi `j` deci avem:

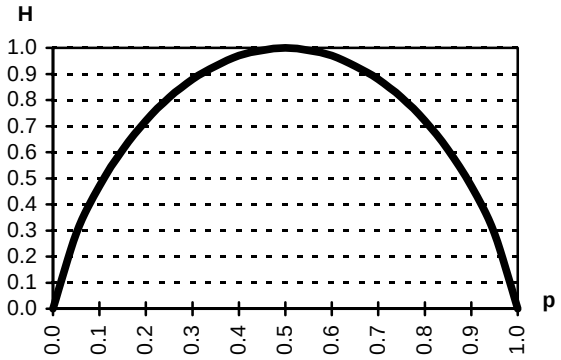
$$p_1 = p_2 = ... = p_n$$

Acest lucru dovedeşte că **informația** transmisă de o sursă este **maximă** atunci când simbolurile sunt **egal probabile**.

Să considerăm cazul în care sursa emite doar două simboluri având probabilitățile `p` şi `1-p`. Expresia entropiei devine în acest caz:

$$H(S) = -p \log p - (1-p)\log(1-p)$$

Reprezentarea grafică a acestei funcții (ca funcție de p) este dată în figura următoare. Se observă că pentru p tinzând la 0 şi la 1 funcția tinde spre 0. Deşi ea nu este definită în aceste puncte (din cauza logaritmului) putem considera prin convenție `0log0=0`. Intuitiv acest lucru este corect, în cazul în care o sursă emite un simbol cu probabilitate 0 (adică nu îl emite) acel simbol nu transmite informație.



Reținem că pentru o sursă care emite două simboluri informația este maximă (şi egală cu 1, adică un bit "de informație") când cele două simboluri sunt egal probabile. Doar în acest caz emiterea unui simbol (a unui bit "fizic") transmite şi un bit "de informație".

Putem remarca o asemănare cu noțiunea de **entropie** din **fizică**. Deşi se numesc la fel cele două entropii nu reprezintă acelaşi lucru. În fizică entropia reprezintă o măsură a agitației termice. În teoria informației entropia reprezintă o măsură a incertitudinii asupra unui eveniment. În ambele cazuri entropia constituie o **măsură a dezordinii** existente.

Se mai defineşte **redundanța sursei** ca fiind diferența între entropia maxim posibilă şi cea reală (cât emite sursa inutil):

$$R_S = H_{MAX}(S) - H(S)$$

Dacă dorim să lucrăm în termeni relativi, putem utiliza noțiunea de **redundanță relativă**:

$$r_s = 1 - \frac{H(S)}{H_{MAX}(S)}$$

În ambele relații avem:

$$H_{MAX}(S) = \log n$$

corespunzând entropiei unei surse cu n simboluri şi probabilități egale de apariție.