

Shannon Information Theory and Beyond*

W. Szpankowski

Department of Computer Science
Purdue University
USA

May 20, 2010

Profesor Wojciech Sobczak



DZIEKUJEMY

*Research supported by NSF [Science & Technology Center](#), and [Humboldt Foundation](#).

Outline of the Talk

1. W. Słobczak "Statystyczna Teoria Systemów Przekazywania Informacji"
2. Three Theorems of Shannon
3. Channel Capacity
 - Capacity of Binary Symmetric Channel
 - Temporal Capacity
 - Noisy Constrained Capacity
4. Science of Information
 - Post-Shannon Information
 - NSF Science and Technology Center

Książka Pana Profesora



Spis treści

1

Modele systemów przesyłania i wydobywania informacji/9

- 1.1 Podstawowy model otwartych systemów przesyłania informacji/9
- 1.2 Systemy ze sprzężeniem zwrotnym/12
- 1.3 Systemy o strukturze sieciowej/14
- 1.4 Systemy klasyfikacji i identyfikacji obiektów/19

2

Elementy teorii sygnałów przypadkowych/22

- 2.1 Pojęcia podstawowe/22
- 2.2 Momenty sygnałów losowych. Sygnały stacjonarne i niestacjonarne/25
- 2.3 Analiza harmoniczna sygnałów stacjonarnych/29
- 2.4 Sygnały gaussowskie/32
- 2.5 Różniczkowanie i całkowanie sygnałów przypadkowych/36
- 2.6 Przekształcenia sygnałów przypadkowych w układach liniowych/40
- 2.7 Sygnały Poissona/43
- 2.8 Sygnały dyskretne Markowa/48

3

Modele źródeł informacji, sygnałów nadawanych i kanałów/52

- 3.1 Modele źródeł informacji/52
- 3.2 Modele statystyczne ciągów informacji komputerowych/56
- 3.3 Podstawowe rodzaje sygnałów nadawanych/59
- 3.4 Modele statystyczne kanałów — zniekształcenia i szumy/67

4

Optymalizacja odbioru informacji dyskretnych/70

- 4.1 Kryteria jakości systemów przesyłania informacji/70
- 4.2 Sformułowanie i ogólne rozwiązanie problemu optymalizacji/78

6

- 4.3 Optymalizacja odbioru sygnałów binarnych w systemach synchronicznych/82
- 4.4 Jakość optymalnej reguły odbioru sygnałów binarnych w systemach synchronicznych/88
- 4.5 Optymalizacja lokalna reguł odbioru/99
- 4.6 Optymalizacja odbioru sygnałów binarnych w systemach niesynchronicznych/106
- 4.7 Ocena zniekształceń odporności na zakłócenia systemów radiolokacyjnych i radiotelegraficznych wskutek braku synchronizacji/113
- 4.8 Uogólnienia problemu optymalizacji odbioru w systemach przesyłania informacji dyskretnych/121

5

Systemy transmisji danych — kodowanie/126

- 5.1 Pojęcia wstępne i właściwości kanału cyfrowego/126
- 5.2 Problemy optymalizacji systemów kodowych/134
- 5.3 Projektowanie kodów zapewniających żądaną jakość przesyłania informacji/139
- 5.4 Przykłady kodów liniowych z wykrywaniem i korektą błędów elementarnych/149
- 5.5 Elementy ogólnej teorii liniowych kodów korekcyjnych/162
- 5.6 Kody ilorazowe/178
- 5.7 Kody cykliczne/190
- 5.8 Kody rekurencyjne/193

6

Elementy teorii informacji/201

- 6.1 Twierdzenie Shannona dla kanałów binarnych/201
- 6.2 Niektóre właściwości graniczne długich ciągów informacji i sygnałów elementarnych/211
- 6.3 Właściwości entropii informacji i sygnałów dyskretnych/218
- 6.4 Twierdzenie o kodowaniu dla kanałów bezszumowych/234
- 6.5 Pojęcie informacji wzajemnej oraz przepustowość kanałów z szumem/241
- 6.6 Twierdzenie o kodowaniu dla kanałów dyskretnych z szumem/253

7

Systemy cyfrowe ze sprzężeniem zwrotnym/259

- 7.1 Systemy z decyzyjnym sprzężeniem zwrotnym/259
- 7.2 Błędy i kryteria jakości systemów z decyzyjnym sprzężeniem zwrotnym/264
- 7.3 Problemy optymalizacji/269
- 7.4 Systemy cyfrowe z informacyjnym sprzężeniem zwrotnym/273

8

Systemy telemetryczne/279

- 8.1 Problem optymalizacji systemów telemetrycznych/279
- 8.2 Interpretacja geometryczna ogólnego problemu optymalizacji systemów telemetrycznych/287
- 8.3 Obliczanie jakości optymalnych reguł odbioru/292
- 8.4 Problem wzajemnego oddziaływania informacji przenoszonych przez ten sam sygnał na jakość decyzji/298
- 8.5 Metody wyznaczania jakości odtwarzania wielowymiarowych informacji ciągłych/303
- 8.6 Przykłady wyznaczania jakości odtwarzania nadawanych informacji/309

Three Theorems of Shannon

Theorem 1 & 3. (Shannon 1948; Lossless & Lossy Data Compression)

compression bit rate $\geq$ source entropy $H(X)$;

for distortion level D :

lossy bit rate $\geq$ rate distortion function $R(D)$.

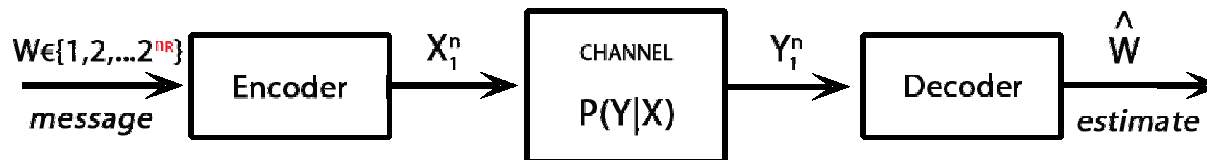
Theorem 2. (Shannon 1948; Channel Coding)

In Shannon's words:

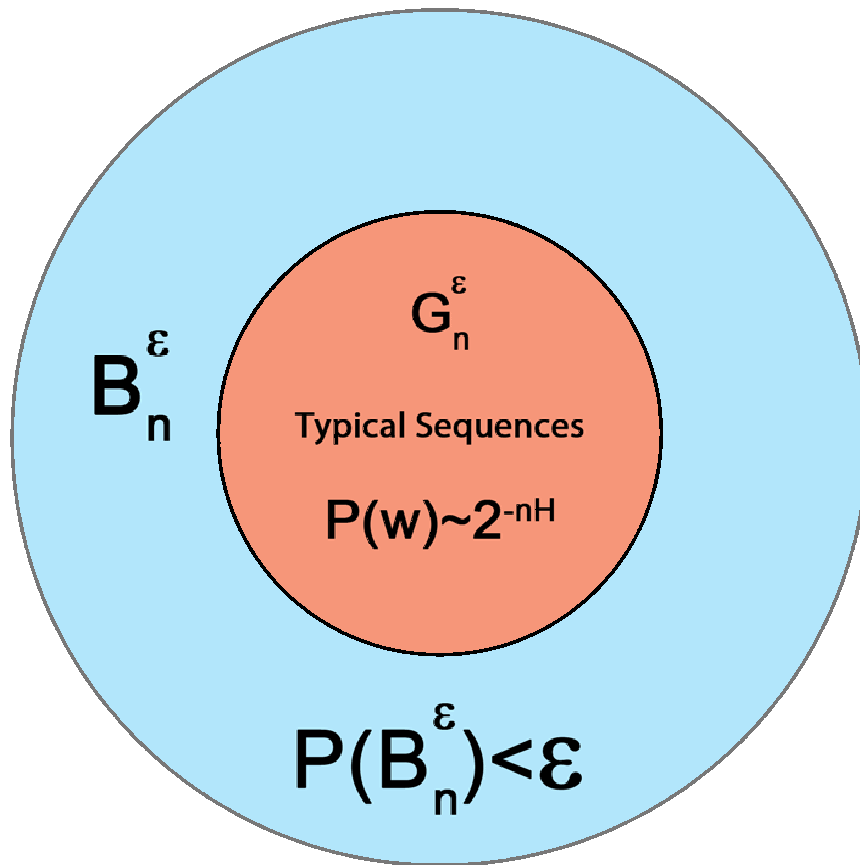


It is possible to send information at the capacity through the channel with as small a frequency of errors as desired by proper (long) encoding.

This statement is not true for any rate greater than the capacity.



Theorem 1: AEP and Typical Sequences



Shannon-McMillan-Breiman:

$$-\frac{1}{n} \log P(X_1^n) \rightarrow H(X) \quad (\text{pr.})$$

$H(X)$ is the entropy rate.

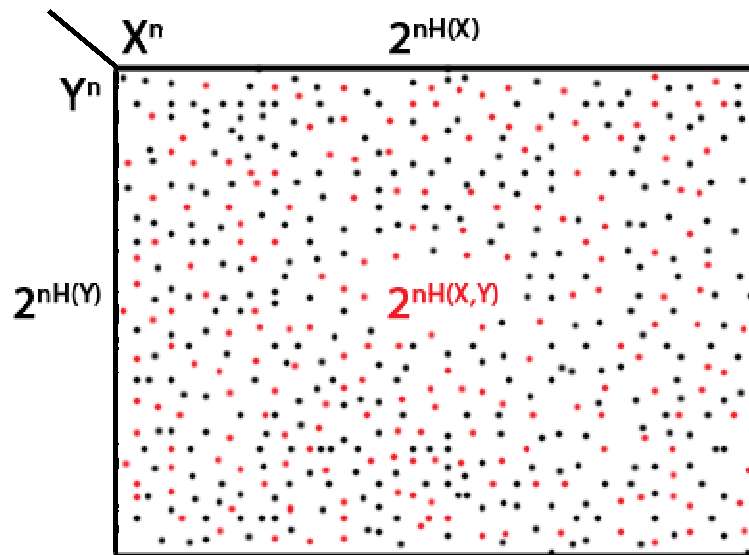
Asymptotic Equipartition Property: Sequences of length n can be partitioned into

good set G_n^ϵ $P(w) \sim 2^{-nH(X)}, \quad w \in G_n^\epsilon$

bad set B_n^ϵ $P(B_n^\epsilon) < \epsilon.$

Also, $|G_n^\epsilon| \sim 2^{nH(X)}.$

Theorem 2: Shannon Random Decoding Rule



There are $2^{nH(X)}$ X-typical sequences

There are $2^{nH(Y)}$ Y-typical sequences

There are $2^{nH(X,Y)}$ jointly X,Y-typical pair of sequences

Decoding Rule: Declare that sequence sent X is the one that is jointly typical with the received sequence Y provided there is unique X satisfying this property!

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y) .
2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$.
Indeed:
 - there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n , respectively.
 - there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y) .

The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{nH(X)+H(Y)}} = 2^{-nI(X,Y)}.$$

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y) .
2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$.

Indeed:

- there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n , respectively.
- there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y) .

The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{nH(X)+H(Y)}} = 2^{-nI(X,Y)}.$$

3. Probability of error when 2^{nR} messages are sent is approximately

$$\min P(\text{error}) \sim 2^{-n(\sup P(X) I(X,Y) - R)} = 2^{-n(C - R)}.$$

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y) .
2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$.

Indeed:

- there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n , respectively.
- there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y) .

The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{nH(X)+H(Y)}} = 2^{-nI(X,Y)}.$$

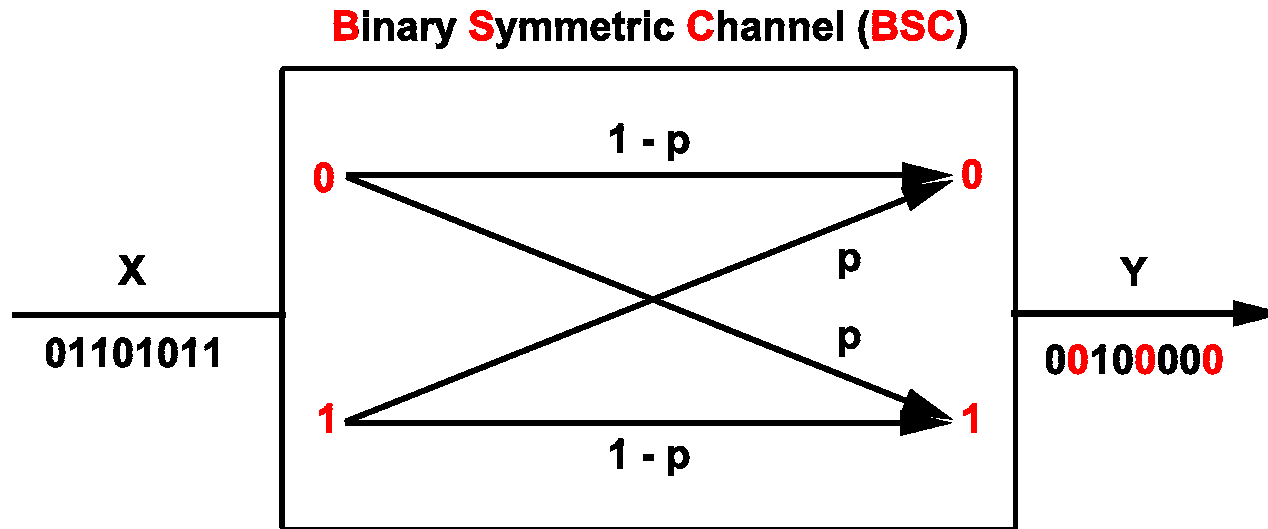
3. Probability of error when 2^{nR} messages are sent is approximately

$$\min P(\text{error}) \sim 2^{-n(\sup P(X) I(X,Y) - R)} = 2^{-n(C-R)}.$$

4. In conclusion:

$$\begin{array}{ll} R < C & P(\text{error}) \sim 2^{-n\delta} \\ R > C & P(\text{error}) \rightarrow 1. \end{array}$$

Capacity of BSC



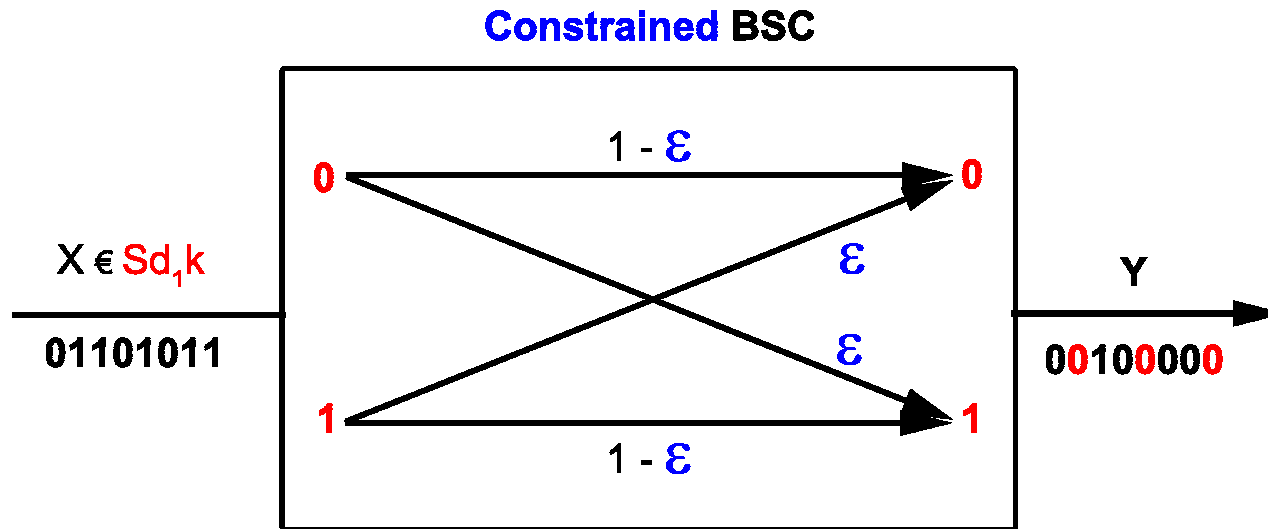
Capacity:

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p). \end{aligned}$$

The **capacity** is achieved for the **uniform** input distribution. Thus

$$C = 1 - H(p).$$

Noisy Constrained Channel

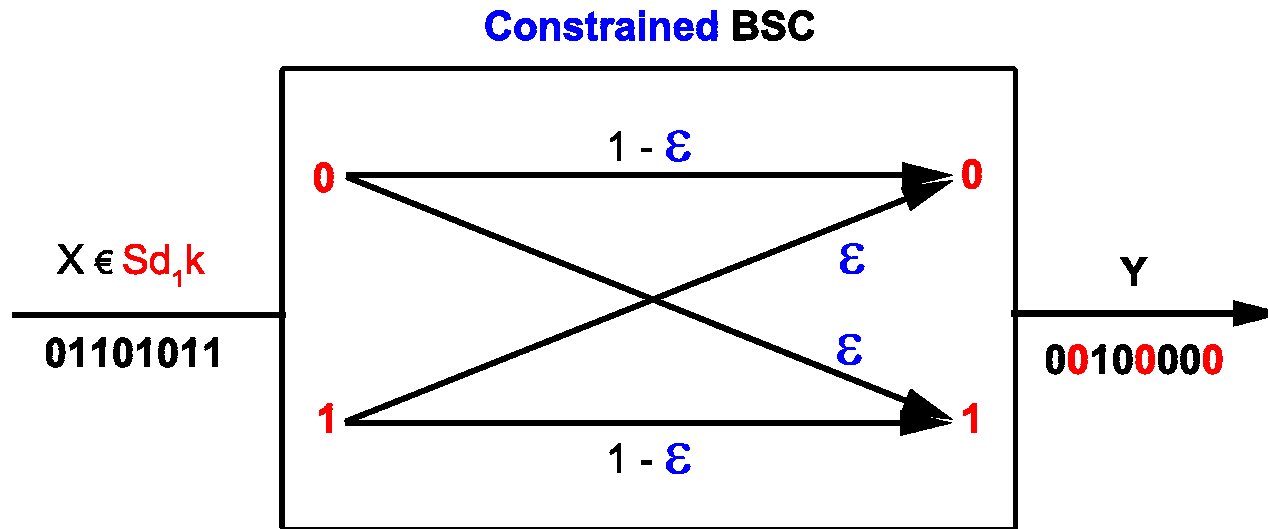


Let $\mathcal{S}$ denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(\mathbf{d}, \mathbf{k}) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d or longer than k** .

Noisy Constrained Channel



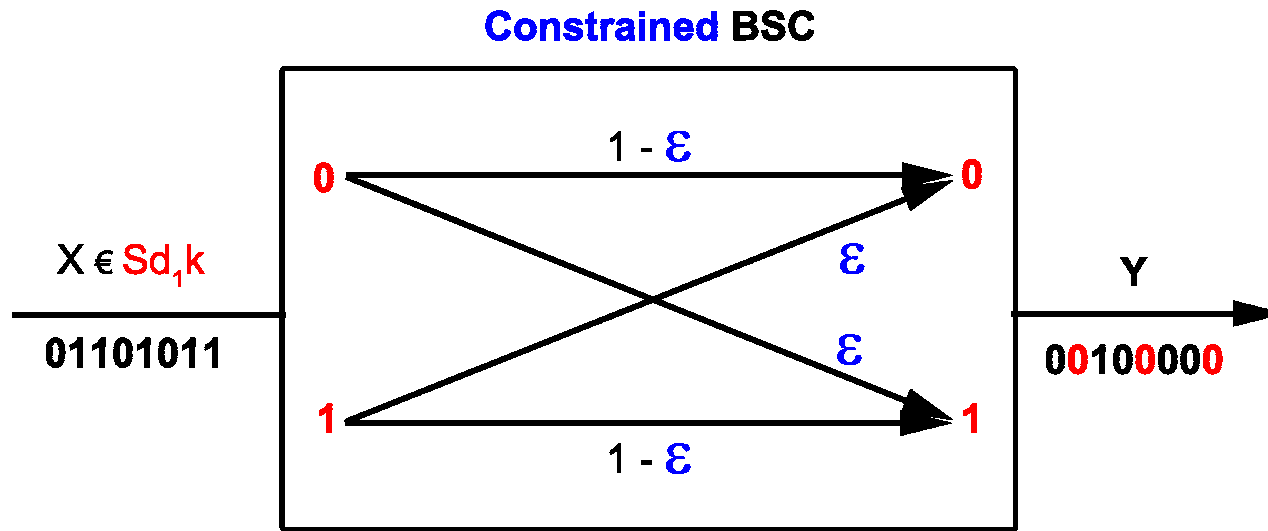
Let $\mathcal{S}$ denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(\mathbf{d}, \mathbf{k}) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d or longer than k** .

Sequence $X \in \mathcal{S}_{(d,k)}$ can be represented as a **Markov process**.

Noisy Constrained Channel



Let $\mathcal{S}$ denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(\text{d},k) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d or longer than k** .

Sequence $X \in \mathcal{S}_{(d,k)}$ can be represented as a **Markov process**.

$C(\mathcal{S}, \epsilon)$ – **noisy constrained capacity** defined as

$$C(\mathcal{S}, \epsilon) = \sup_{X \in \mathcal{S}} I(X; Y) = \lim_{n \rightarrow \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Y_1^n).$$

This is/was an open problem since Shannon.

Noisy Constrained Capacity

1. Mutual information: $I(X; Y) = H(Y) - H(Y|X)$.

$$Y_k = X_k \oplus E_k, \quad k \geq 1, \quad (\oplus \text{ is exclusive-or}),$$

$E = \{E_k\}_{k \geq 1}$ is noise, $P(E_i = 1) = \epsilon$ and $H(Y|X) = H(E) = H(\epsilon)$.

If X is Markov, then Y is a hidden Markov process (HMP).

Noisy Constrained Capacity

1. Mutual information: $I(X; Y) = H(Y) - H(Y|X)$.

$$Y_k = X_k \oplus E_k, \quad k \geq 1, \quad (\oplus \text{ is exclusive-or}),$$

$E = \{E_k\}_{k \geq 1}$ is noise, $P(E_i = 1) = \varepsilon$ and $H(Y|X) = H(E) = H(\varepsilon)$.

If X is Markov, then Y is a hidden Markov process (HMP).

2. The output process Y is hidden Markov, hence we need entropy $H(Y)$.

Theorem 1 (Jacquet, Seroussi, & Szpankowski, 2004). Consider the HMP Y as defined above. The entropy rate

$$H(Y) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{E}[-\log (\mathbf{p}_1 \mathbf{M}(Y_1, Y_2) \cdots \mathbf{M}(Y_{n-1}, Y_n) \mathbf{1}^t)] = \mu(P)$$

where $\mu(P)$ is a top Lyapunov exponent of $\mathbf{M}(Y_1, Y_2) \cdots \mathbf{M}(Y_{n-1}, Y_n)$ defined as

$$\mathbf{M}(Y_{n-1}, Y_n) = \begin{bmatrix} (1-\varepsilon)P_X(Y_n|Y_{n-1}) & \varepsilon P_X(\bar{Y}_n|Y_{n-1}) \\ (1-\varepsilon)P_X(Y_n|\bar{Y}_{n-1}) & \varepsilon P_X(\bar{Y}_n|\bar{Y}_{n-1}) \end{bmatrix}.$$

Asymptotic Expansion

3. We now assume that $P(E_i = 1) = \varepsilon \rightarrow 0$ is small.

Theorem 2 (Jacquet, Seroussi, W.S., 2004). Assume *r*th order Markov. If the conditional symbol probabilities of the underlying Markov X satisfy

$$P(a_{r+1}|a_1^r) > 0$$

for all $a_1^{r+1} \in \mathcal{A}^{r+1}$, then the *entropy rate* of Y for small ε is

$$H(Y) = \lim_{n \rightarrow \infty} \frac{1}{n} H_n(Y^n) = H(X) + f_1(P)\varepsilon + O(\varepsilon^2),$$

where

$$f_1(P) = \sum_{z_1^{2r+1}} P_X(z_1^{2r+1}) \log \frac{P_X(z_1^{2r+1})}{P_X(\bar{z}_1^{2r+1})} = \mathbb{D} \left(P_X(z_1^{2r+1}) || P_X(\bar{z}_1^{2r+1}) \right),$$

where $\bar{z}_1^{2r+1} = z_1 \dots z_r \bar{z}_{r+1} z_{r+2} \dots z_{2r+1}$. In the above, $\mathbb{D}$ denotes the *Kullback-Liebler divergence*.

Two Examples

Example 1. Markov process with symmetric transition probabilities $p_{01} = p_{10} = p$, $p_{00} = p_{11} = 1-p$; hence $P_X(0) = P_X(1) = \frac{1}{2}$. Then

$$H(\textcolor{red}{Y}) = H(\textcolor{blue}{X}) + \textcolor{blue}{f}_1(p)\varepsilon + \textcolor{blue}{f}_2(p)\varepsilon^{\textcolor{red}{2}} + O(\varepsilon^3)$$

where

$$\textcolor{blue}{f}_1(p) = 2(1 - 2p) \log \frac{1 - p}{p}, \quad \textcolor{blue}{f}_2(p) = -\textcolor{blue}{f}_1(p) - \frac{1}{2} \left(\frac{2p - 1}{p(1 - p)} \right)^2.$$

Two Examples

Example 1. Markov process with [symmetric transition](#) probabilities $p_{01} = p_{10} = p, p_{00} = p_{11} = 1-p$; hence $P_X(0) = P_X(1) = \frac{1}{2}$. Then

$$H(\textcolor{red}{Y}) = H(\textcolor{blue}{X}) + \textcolor{blue}{f}_1(p)\varepsilon + \textcolor{blue}{f}_2(p)\varepsilon^2 + O(\varepsilon^3)$$

where

$$\textcolor{blue}{f}_1(p) = 2(1 - 2p) \log \frac{1 - p}{p}, \quad \textcolor{blue}{f}_2(p) = -f_1(p) - \frac{1}{2} \left(\frac{2p - 1}{p(1 - p)} \right)^2.$$

Example 2: Consider (here $\mathbf{P} \not\propto 0$)

$$\mathbf{P} = \begin{bmatrix} 1 - p & p \\ 1 & 0 \end{bmatrix}$$

where $0 \leq p \leq 1$. [Ordentlich and Weissman](#) (2004) proved for this case

$$H(Y) = H(P) - \frac{p(2 - p)}{1 + p} \varepsilon \log \varepsilon + O(\varepsilon)$$

(e.g., [\(11 ...\)](#) will not be generated by MC, but can be outputted by HMM with probability $O(\varepsilon^\kappa)$).

Finally ...

It can be proved (Han&Marcus, 2007; Jacquet & Szpankowski,2007) that

$$H(Y) = H(P) - f_0(P)\varepsilon \log \varepsilon + f_1(P)\varepsilon + o(\varepsilon)$$

for explicitly computable $f_0(P)$ and $f_1(P)$.

Let $P^{\max}$ be the maxentropic maximizing $H(P)$. Then

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) - (1 - f_0(P^{\max}))\varepsilon \log \varepsilon + (f_1(P^{\max}) - 1)\varepsilon + o(\varepsilon)$$

where $C(\mathcal{S})$ is the capacity of noiseless system ($\varepsilon = 0$)

Theorem 4 (Jacquet & Szpankowski, 2007). For $k \leq 2d$, we shall prove that

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) + A \cdot \varepsilon + O(\varepsilon^2 \log \varepsilon).$$

For $k > 2d$, we shall prove that

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) + B \cdot \varepsilon \log \varepsilon + O(\varepsilon),$$

where A and B are explicitly computable constants.

Outline of the Talk

1. W. Sobczak "Statystyczna Teoria Systemow Przysylania Informacji"
2. Three Theorems of Shannon
3. Channel Capacity
4. Science of Information

Shannon Legacy

The Information Revolution started in 1948, with the publication of:
A Mathematical Theory of Communication.

The digital age began.



Claude Shannon:

Shannon information quantifies the extent to which a recipient of data can reduce its statistical uncertainty.

“These semantic aspects of communication are irrelevant . . .”

Applications Enabler/Driver:

CD, iPod, DVD, video games, computer communication, Internet, Facebook, Google, . . .

Design Driver:

universal data compression, data encoding, voiceband modems, CDMA, multiantenna, discrete denoising, space-time codes, cryptography, . . .

Post-Shannon Challenges

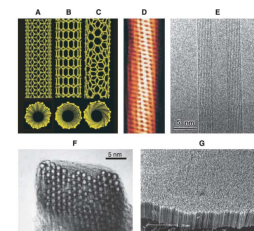
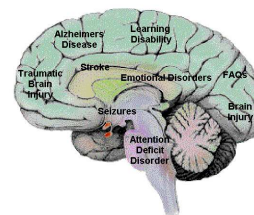
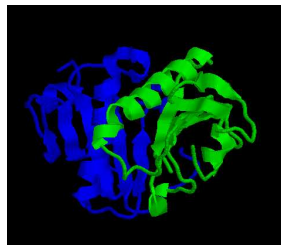
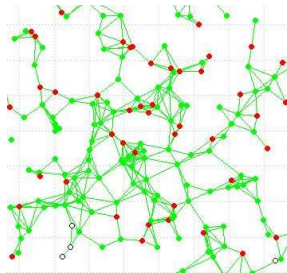
Classical Information Theory needs a **recharge** to meet new **challenges** of nowadays applications in **biology, modern communication, knowledge extraction, economics** and **physics**, ...

We need to extend traditional formalisms for **information** to include (“**meaning**”):

structure, time, space, and semantics ,

and others such as:

dynamic information, limited resources, complexity, physical information, representation-invariant information, and cooperation & dependency.



Standing on the Shoulders of Giants . . .



Manfred Eigen (Nobel Prize, 1967)

"The differentiable characteristic of the **living systems** is **Information**. **Information** assures the controlled **reproduction** of all constituents, ensuring **conservation** of viability **Information theory**, pioneered by **Claude Shannon**, **cannot** answer this question . . . in principle, the answer was formulated 130 years ago by **Charles Darwin**".



P. Nurse, (**Nature**, 2008, "Life, Logic, and Information"):

*Focusing on **information flow** will help to **understand better** cell how **cells and organisms** work.*

". . . the generation of **spatial** and **temporal** order, **memory** and **reproduction** are **not fully understood**".



A. Zeilinger (**Nature**, 2005)

". . . **reality** and **information** are two sides of the same coin, that is, they are in a deep sense **indistinguishable**.

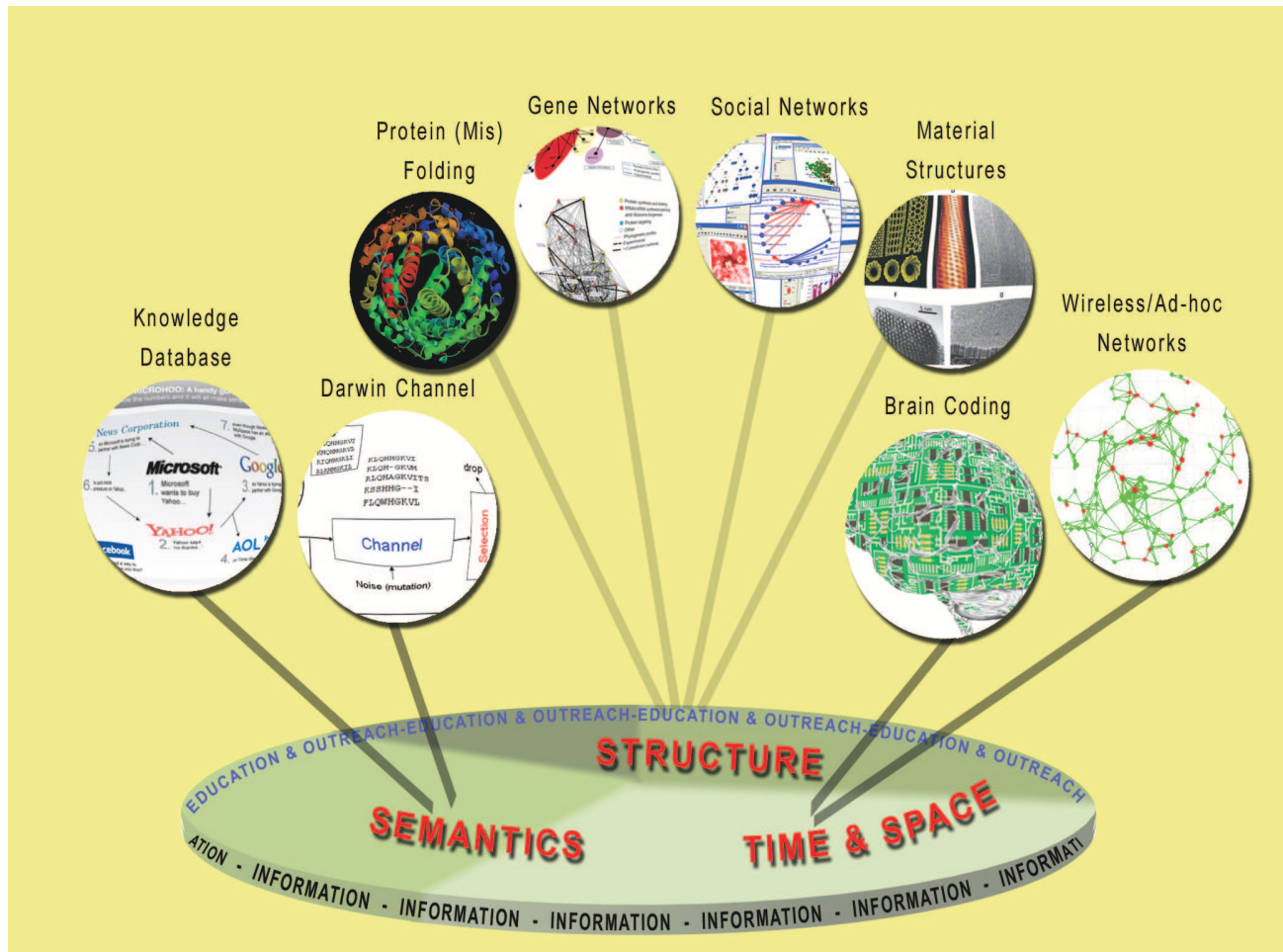


F. Brooks, jr. (JACM, 2003, "Three Great Challenges . . ."):

We have **no theory** that gives us a metric for the **Information** embodied in **structure** . . . this is the most **fundamental gap** in the theoretical underpinning of **Information** and computer science.

Science of Information

The overarching vision of **Science of Information** is to develop rigorous principles guiding the **extraction, manipulation, and exchange** of **information**, integrating elements of **space, time, structure**, and **semantics**.



Institute for Science of Information

In 2008 at Purdue we launched the

Institute for Science of Information

and in 2010 National Science Foundation established \$25M

Science and Technology Center

at Purdue to do collaborative work with Berkeley, MIT, Princeton, Stanford, UIUC and Bryn Mawr & Howard U. integrating research and teaching activities aimed at investigating the role of information from various viewpoints: from the fundamental theoretical underpinnings of green information to the science and engineering of novel information substrates, biological pathways, communication networks, economics, and complex social systems.

The specific means and goals for the Center are:

- develop post-Shannon Information Theory,
- Prestige Science Lecture Series on Information to collectively ponder short and long term goals;
- organize meetings and workshops (e.g., Information Beyond Shannon, Orlando 2005, and Venice 2008).
- initiate similar world-wide centers supporting research on information.

That's It

