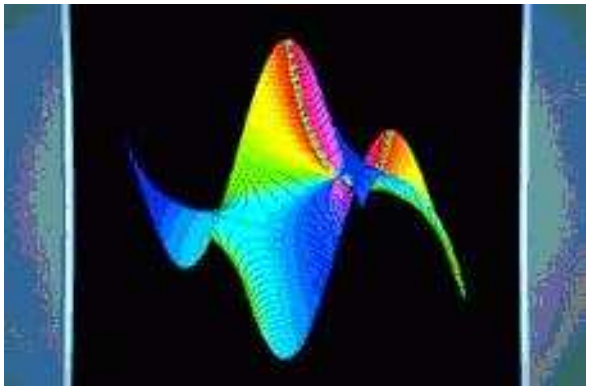
Facets of Information in Communications*

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

October 5, 2008

AofA and **IT** logos



PGTS, Berlin 2008

*Thanks to P. Jacquet, INRIA, France and J. Konorski, Gdansk, Poland.

Outline

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples
   (a) Temporal Channel
   (b) Speed of Information in Wireless Networks
   (c) Darwin Channel
5. Science of Information

Shannon Information

In 1948 C. Shannon created a powerful and beautiful **theory of information** that served as the backbone to nowadays digital communications.

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

Some Aspects of Shannon information:

objective: statistical ignorance of the recipient;
statistical uncertainty of the recipient.

cost: # binary decisions to describe E;
= $-\log P(E)$; $P(E)$ being the probability of E.

Context: "semantic aspects of communication are irrelevant"

Self-information for E_i: $\text{info}(E_i) = -\log P(E_i)$.
Average information: $H(P) = -\sum_i P(E_i) \log P(E_i)$
Entropy of $X = \{E_1, \ldots\}$: $H(X) = -\sum_i P(E_i) \log P(E_i)$
Mutual Information: $I(X;Y) = H(Y) - H(Y|X)$, (faulty channel).

Three Jewels of Shannon

Theorem 1. (**Shannon 1948; Lossless Data Compression**).

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

(There exists a codebook of size 2^{nR} of universal codes of length n with

$$R > H(X)$$

and probability of error smaller than any $\varepsilon > 0$.)

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.

(The maximum codebook size $N(n, \varepsilon)$ for codelength n and error probability ε is asymptotically equal to: $N(n, \varepsilon) \sim 2^{nC}$.)

Theorem 3. (**Shannon 1948; Lossy Data Compression**).
For distortion level D:

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

Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples

What is Information?

C. F. Von Weizsäcker:
"Information is only that which produces information" (relativity).
"Information is only that which is understood" (rationality)
"Information has no absolute meaning".

R. Feynman:
. . . Information is not simply a physical property of a message:
it is a property of the message and your knowledge about it."

Informally Speaking: A piece of data carries **information** if it can impact a **recipient's ability** to achieve the **objective** of some **activity** in a given **context** within limited resources.

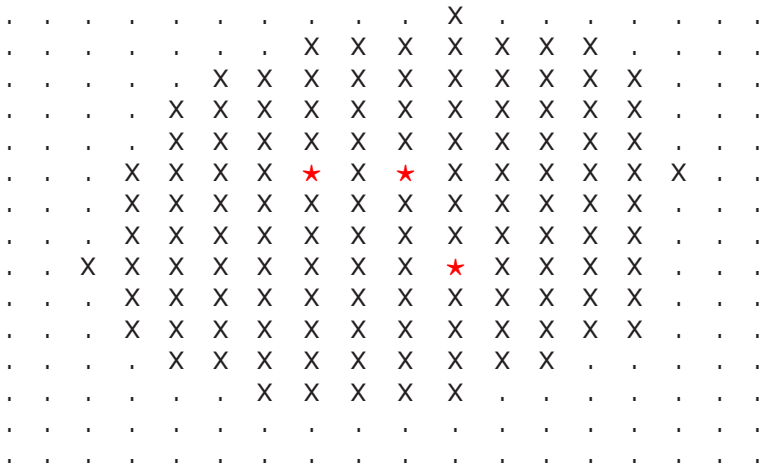
Definition 1. *The* **amount of information** *(in a faultless scenario)* $\mathrm{info}(E)$ *carried by the event* E *in the context* C *with the rules* R *is*

$$\mathrm{info}_{R,C}(E) = \mathbf{cost}[\mathrm{objective}_R(C(E)), \mathrm{objective}_R(C(E) + E)]$$

for some **cost** *(weight, distance) function.*

Example: Distributed Information

1. In an N-threshold secret sharing scheme, N subkeys of the decryption key roam among $A \times A$ stations.

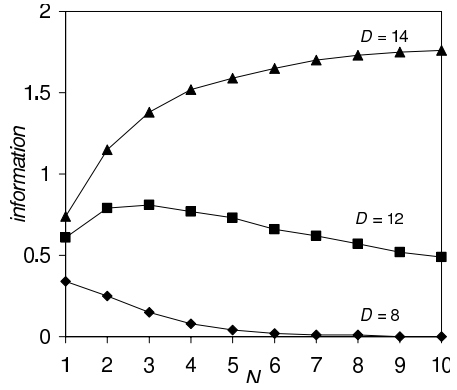


2. By protocol P a station has access:

- only it sees all N subkeys.
- it is within a distance D from all subkeys.

3. Assume that the larger N,
the more valuable the secrets.
We define the amount of information as
(cf. J. Konorski and W.S.)

info= $N \times \{\#\ \text{of stations having access}\}$.



Rissanen's MDL Principle

1. Objective(P, C) may include the cost of the very recognition and interpretation of C.

2. In 1978 Rissanen introduced the Minimum Description Length (MDL) principle (Occam's Razor) postulating that the best hypothesis is the one with the shortest description.

3. Universal data compression is used to realize MDL.

4. Normalized maximum likelihood (NML) code: Let $\mathcal{M}_k = \{Q_\theta : \theta \in \Theta\}$ and let $\hat{\theta}$ minimize $-\log Q_\theta(x)$. The minimax regret is

$$R_n^* = \min_Q \max_x \left[\log \frac{Q_{\hat{\theta}}(x)}{Q_\theta(x)}\right] = \log \sum_x Q_{\hat{\theta}}(x) = \log \sum_x \sup_\theta Q_\theta(x).$$

Rissanen (cf. W.S., 1995) proved for memoryless and Markov sources:

$$R_n^* = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_\theta \sqrt{|I(\theta)|} d\theta + o(1).$$

where $I(\theta)$ is the Fisher information.

5. Useful Information: It follows that R_n^* bits are required to describe distinguishable models with diminishing probability of making a mistake.

Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples

Beyond Shannon

Participants of the **2005 Information Beyond Shannon** workshop realize:

Delay: Delay incurred is a issue not yet adequately addressed in information theory (e.g., complete information arriving late maybe useless).

Space: In networks the spatially distributed components raise fundamental issues of limitations in information exchange since the available resources must be shared, allocated and re-used. **Information** is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective.

Structure: We still lack measures and meters to define and appraise the amount of information embodied in structure and organization.

Semantics. In many scientific contexts, one is interested in signals, without knowing precisely what these signals represent. What is semantic information and how to characterize it? How much more semantic information is there when when compared with its syntactic information?

Limited Computational Resources: In many scenarios, information is limited by available computational resources (e.g., cell phone, living cell).

Physics of Information: Information is physical (J. Wheeler).

Some Things to Think About . . .

Here is a short list of "toy problems" to think about:

- **Temporal Capacity** (e.g., assign transmission time to each symbol).
- Spatial Capacity (e.g., destination may be in different locations).
- **Darwin Channel** models flow of genetic information (e.g., a combination of a deletion/insertion channel and constrained channel).
- Distributed Information (information here/local and there/distributed is not the same?)
- **Speed of Information** (how fast information can be spread out?)
- Entropy of a structure (e.g., graph entropy)
- Representation-invariant measure of information (Shannon, 1953).

Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples
   (a) Temporal Channel
   (b) Speed of Information in Wireless Networks
   (c) Darwin Channel

Temporal Capacity

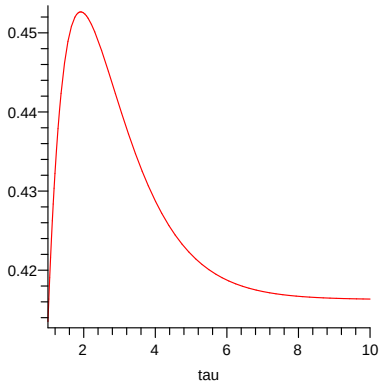
1. Binary symmetric channel (BSC): each bit incurs a delay.
2. Delay T has known probability distribution: $F(t) = P(T < t)$.

If a bit arrives after a given deadline τ, it is dropped.

3. The longer it takes to send a bit, the lower the probability of a success, which we denote by $\Phi(\varepsilon, t)$ for $t < \tau$ (e.g., $\Phi(\varepsilon, t) = (1-\varepsilon)^t$).
4. Define $P(x|x) = \int_0^\tau \Phi(\varepsilon, t) dF(t)$: prob. of a successfully transmission:

$$P(y|x) = \begin{cases} \alpha := 1 - F(\tau) & y = \text{erasure} \\ P(x|x) & \text{if } x = y \\ 1 - \alpha - P(x|x) & \text{if } x \neq y. \end{cases}$$

5. Define: $\alpha = 1 - F(\tau)$ and $\rho := \frac{P(x|x)}{(1-\alpha)}$.
Note $C(\tau) := \max_X [H(Y) - H(Y|X)]$,
$H(Y|X) = H(\alpha) + (1-\alpha)H(\rho)$
$H(Y) = H(\alpha) + (1-\alpha)H(p\rho + \bar{p}\bar{\rho})$.
Then:



$$C(\tau) = [(1 - P(T > \tau)][1 - H(\rho)].$$

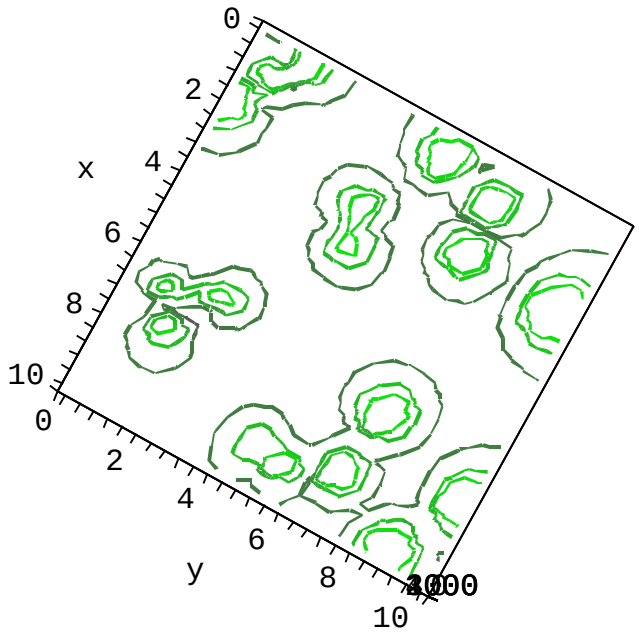
Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples
   (a) Temporal Channel
   (b) Speed of Information in Wireless Networks
   (c) Darwin Channel
5. Science of Information

Speed of Information

Based on P. Jacquet, B. Mans and G. Rodolakis, ISIT, 2008

Intermittently Connected Mobile Networks (ICN):



1. Nodes move in a space with uniform density $\nu > 0$.

2 Nodes do random walks with speed v and turn rate τ.

3. Connectivity is achieved in a unit disk.

4. Radio propagation speed is infinite.

Problem statement:
At time $t = 0$ a node at the origin broadcasts a beacon and nodes retransmit beacon immediately to neighbors in the ICN network.

Question: At what time T node at distance L from the origin will receive the beacon? Propagation speed is $\frac{L}{T}$.

Journey Analysis Through the Laplace Transform

The beacon undergoes a journey C from the origin to some point $\mathbf{z}$. Let $\mathbf{z}(C)$ be destination point reached at time $t(C)$.

Let $p(\mathbf{z}, t)$ be the space-time density of paths C that reaches location $\mathbf{z}(C)$ at time t.

The bivariate Laplace transform of $\mathbf{z}(C)$ and $t(C)$ is (P. Jacquet, B. Mans, G. Rodolakis, 2008)

$$\mathbf{E}[\exp(-\zeta \mathbf{z}(\mathbf{C}) - \theta t(C))] = \frac{1}{D(|\zeta|, \theta)}$$

with

$$D(\rho, \theta) = \sqrt{(\theta + \tau)^2 - \rho^2 v^2} - \frac{4\pi\nu v I_0(\rho)}{1 - \pi\nu\frac{2}{\rho} I_1(\rho)}$$

with I_k modified Bessel functions of order k.

In order to find $p(\mathbf{z}, t)$ one needs to inverse the Laplace transform through the saddle point method.

Information speed is σ_0 such that for all $\sigma > \sigma_0$

$$\lim p\left(\mathbf{z}, \frac{|\mathbf{z}|}{\sigma}\right) = 0.$$

Main Result on Information Speed

Let $\mathcal{K}$ be the set (ρ, θ) of all roots of $D(\rho, \theta) = 0$.

Theorem 1. *The information speed is not greater than the smallest ratio*

$$\frac{\theta}{\rho}$$

where (ρ, θ) belongs to $\mathcal{K}$.

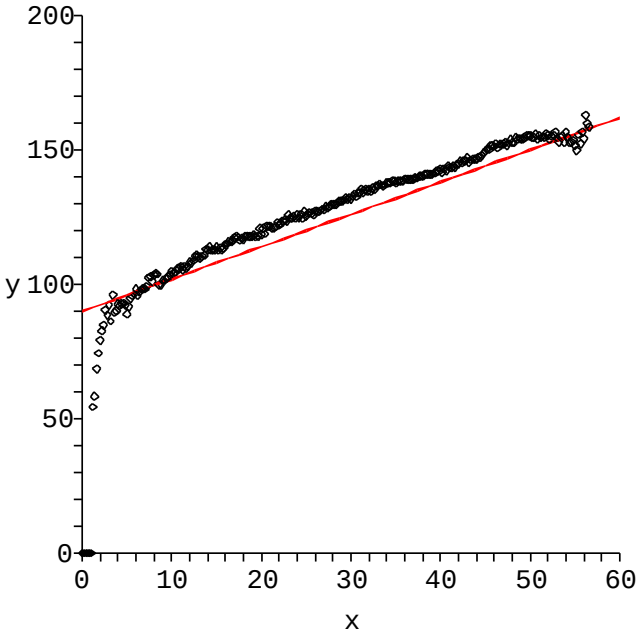


Figure 1: Time versus distance for $\nu = 0.1$, $v = 1$ and $\tau = 0.25$

.

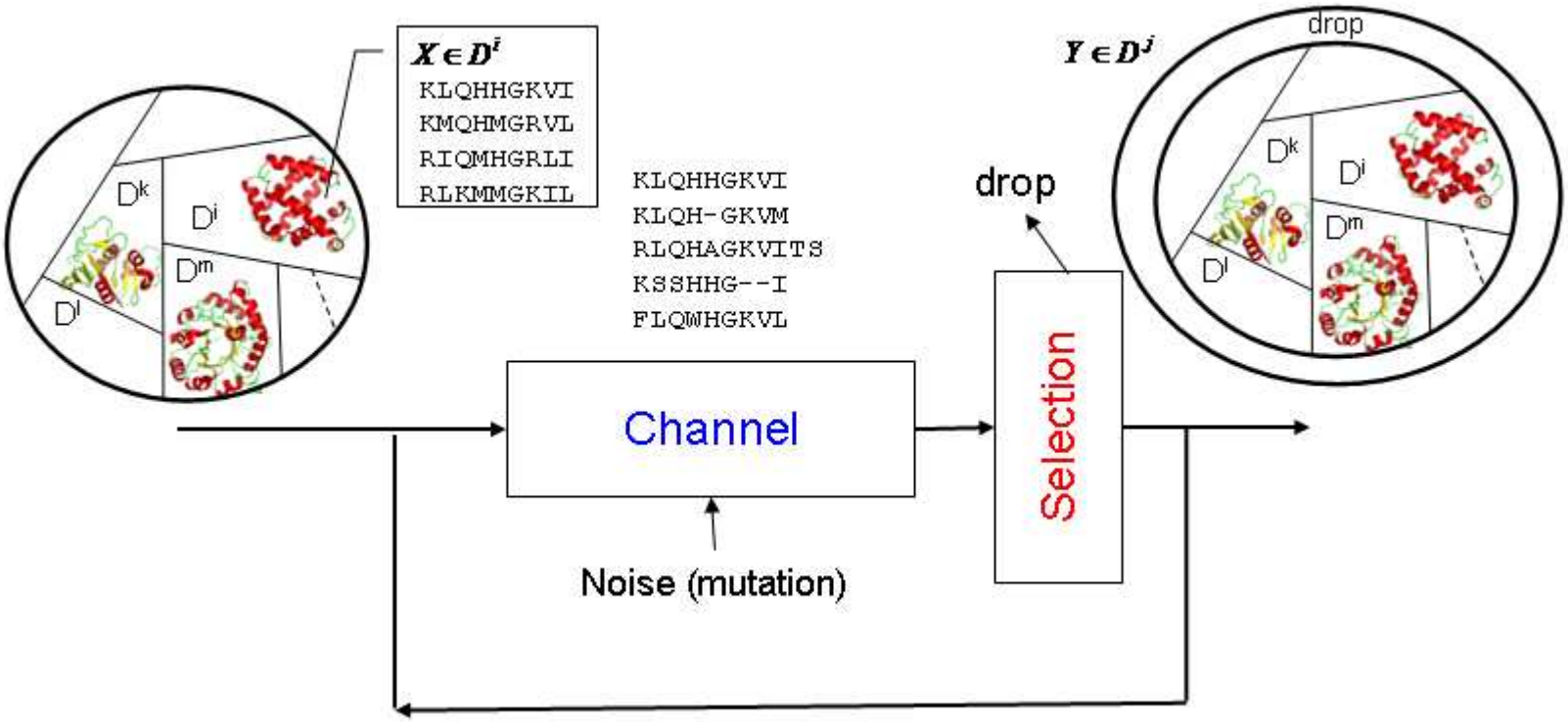
Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples
   (a) Speed of Information in Wireless Networks
   (b) Temporal Channel
   (c) Darwin Channel
5. Science of Information

Darwin Channel

Transfer of Biological Information:
Biomolecular structures (e.g., DNA, proteins) have gone through significant metamorphosis over eons through mutation and natural selection, leading to information transfer. How to assess it?

To capture mutation and natural selection we introduce **Darwin channel**.



Noisy Constrained Channel

1. Binary Symmetric Channel (BSC):
(i) crossover probability ε,
(ii) **constrained set of inputs** (Darwin preselected) that can be modeled by a **Markov Process**,
(ii) $\mathcal{S}_n$ denotes the set of binary constrained sequences of length n.

2. Channel Input and Output:
Input: Stationary process $X = \{X_k\}_{k\geq 1}$ supported on $\mathcal{S} = \bigcup_{n>0} \mathcal{S}_n$.
Channel Output: **Hidden Markov Process** (HMP)

$$Z_i = X_i \oplus E_i$$

where $\oplus$ denotes addition modulo 2, and $E = \{E_k\}_{k\geq 1}$, independent of X, with $P(E_i = 1) = \varepsilon$ is a Bernoulli process (noise).

Note: To focus, we illustrate our results on

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

i.e., no sequence in $\mathcal{S}_n$ contains a run of zeros of length ishorter than d or longer than k. Such sequences can model neural spike trains (no two spikes in a short time).

Noisy Constrained Capacity

$C(\varepsilon)$ – conventional BSC channel capacity $C(\varepsilon) = 1 - H(\varepsilon)$, where $H(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$ is the binary entropy.

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

$$C(\mathcal{S}, \varepsilon) = \sup_{X \in \mathcal{S}} I(X; Z) = \lim_{n \to \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Z_1^n),$$

where the suprema are over all stationary processes supported on $\mathcal{S}$ and $\mathcal{S}_n$, respectively. **This has been an open problem since Shannon.**

Mutual information

$$I(X; Z) = H(Z) - H(Z|X)$$

where $H(Z|X) = H(\varepsilon)$.

Thus, we must find the entropy $H(Z)$ of a **hidden Markov process** (e.g., (d, k) sequence can be generated as an output of a kth order Markov process).

Entropy Rate as a Lyapunov Exponent

Jacquet, Seroussi and W.S. (2004) proved that

$$P(Z_1^n) = \mathbf{p}_1 \mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n) \mathbf{1}^t$$

where $\mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n)$ are explicitly defined random matrices.

Theorem 2 (Furstenberg and Kesten, 1960). *Let $\mathbf{M}_1, \ldots, \mathbf{M}_n$ form a stationary ergodic sequence and $\mathbf{E}[\log^+ ||\mathbf{M}_1||] < \infty$ Then*

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{E}[\log ||\mathbf{M}_1 \cdots \mathbf{M}_n||] = \lim_{n\to\infty} \frac{1}{n} \log ||\mathbf{M}_1 \cdots \mathbf{M}_n|| = \mu \quad \text{a.s.}$$

where μ is called top Lyapunov exponent.

Corollary 1. *Consider the HMP Z as defined above. The entropy rate*

$$\begin{aligned} h(Z) &= \lim_{n\to\infty} \mathbf{E}[-\frac{1}{n} \log P(Z_1^n)] \\ &= \lim_{n\to\infty} \frac{1}{n} \mathbf{E}[-\log \left(\mathbf{p}_1 \mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n) \mathbf{1}^t \right)] \end{aligned}$$

is a top Lyapunov exponent of some random matrices $\mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n)$.

Lyapunov exponents are notoriously difficult to compute (Tsitsiklis & Blondel).

Asymptotic Expansion

We now assume that $P(E_i = 1) = \varepsilon \to 0$ is small (e.g., $\varepsilon = 10^{-12}$).

Theorem 3 (Seroussi, Jacquet and W.S., 2004). *Assume rth order Markov. If the conditional probabilities in the Markov process X satisfy*

$$P(a_{r+1}|a_1^r) > 0 \quad \textbf{IMPORTANT!}$$

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

$$h(Z) = \lim_{n\to\infty} \frac{1}{n} H_n(Z^n) = h(X) + f_1(P)\varepsilon + O(\varepsilon^2),$$

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 $\bar{z}^{2r+1} = z_1 \ldots z_r \bar{z}_{r+1} z_{r+2} \ldots z_{2r+1}$. In the above, $h(X)$ is the entropy rate of the Markov process X, $\mathbb{D}$ denotes the Kullback-Liebler divergence.

Examples

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

$$h(Z) = h(X) + f_1(p)\varepsilon + f_2(p)\varepsilon^2 + O(\varepsilon^3)$$

where

$$f_1(p) = 2(1-2p)\log\frac{1-p}{p}, \qquad f_2(p) = -f_1(p) - \frac{1}{2}\left(\frac{2p-1}{p(1-p)}\right)^2 .$$

Example 2. (Degenerate Case.) Consider the following Markov process

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

where $0 \le p \le 1$.
Ordentlich and Weissman (2004) proved for this case

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

(e.g., $(11\ldots)$ cannot be generated by MC, but can by a HMM).

Main Asymptotic Results

We observe (cf. Han and Marcus (2007))

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

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 known capacity of a noiseless channel.

Example: For (d, k) sequences, we can prove (cf. Jacquet, Seroussi, and W.S.)

(i) for $k \leq 2d$

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

(ii) For $k > 2d$

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

where $A \& B$ are computable constants (cf. also Han and Marcus (2007)).

Outline Update

1. Shannon Information
2. What is Information?
3. Beyond Shannon
4. Some Examples
5. Science of Information

Science of Information

Figure 2: Synergy of Information

Institute for Science of Information

At Purdue we initiated the

Institute for Science of Information

integrating research and teaching activities aimed at investigating the role of **information** from various viewpoints: from the fundamental theoretical underpinnings of 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:

- Prestige Science Lecture Series on Information to collectively ponder short and long term goals;
- encourage and facilitate interdisciplinary collaborations (NSF STC with Berkeley, MIT, Princeton, and Stanford).
- provide scholarships and fellowships for the best students, and support the development of new interdisciplinary courses.
- Initiate similar Institute in Europe to support TransAtlantic research on information.