# Back from Infinity: A Constrained Resources Approach to Information Theory

J. Ziv
Faculty of Electrical Engineering
Technion—Israel Institute of Technology
Haifa, 32000, Israel

*Shannon Lecture ISIT'97*

## INTRODUCTION

There are two main avenues along which classical Information Theory has progressed since 1948.

1) Bounds on Communication: Converse-to-coding theorems, the Data Processing theorem, Rate distortion theory, etc.

2) Coding theorems and algorithms: Theorems and algorithms that addresses the realization of these bounds, thus establishing their tightness and the optimality of the algorithms that are associated with the coding theorems.

Some of the classical results are asymptotic in nature and refer to cases where the "block-length" (or "constraint-length") tends to infinity. In practice, very long blocks result in causing a very large encoding and decoding delay and/or in yielding a large computational complexity.

For example, in the case of universal source coding, the classical results that establish the optimality of various universal coding theorems and algorithms are also asymptotic in nature (i.e. assuming that the amount of training data tends to infinity, or that the length of the input string to be compressed, tends to infinity).

It is therefore imperative to try to re-derive the classical results, converse theorems and coding theorems, under the assumption that parameters like delay, processor memory, and computational complexity are constrained.

Old and new results and attempts to address these problems, some more successful than others, will be critically discussed in this presentation.

## A) Universal noiseless compression with memory and latency constraints

### "In the beginning there was entropy ..."

Notation:

Let

$$X_1^\ell = X_1, X_2, \ldots, X_\ell; \qquad X_i \in \alpha$$
$$|\alpha| = A$$

Entropy (entropy-rate)

$$\boxed{H = \lim_{n\to\infty} E\frac{1}{n}\{-\log P(X_1^n)\}}$$

Conditional entropy

$$H(X_1|X_{-n}^0) = E\{-\log P(X_1|X_{-n}^0)\}$$

$$\boxed{\lim_{n\to\infty} H(X_1|X_{-n}^0) = H}$$

Consider a Fixed-to-Variable noiseless encoder for $\ell$-blocks, with limited-length history $X_{-n}^0$.

Let $L(X_1^\ell|X_{-n}^0)$ be the length function of $X_1^\ell$, given $X_{-n}^0$ (namely, the number of bits that represent $X_1^\ell$). Then

$$\begin{aligned}\text{Compression} \stackrel{\triangle}{=} \frac{1}{\ell} E\left(L(X_1^\ell|X_{-n}^o)\right) &\geq \tfrac{1}{\ell} H(X_1^\ell|X_{-n}^0) \\ &\geq H(X_1|X_{-n-\ell}^0) \\ &> H\end{aligned}$$

Thus, the total memory is

$$N = (n+1) + \ell$$

where $\ell$ is the decoding latency and where (n+1) is the size of the memory which is allocated for the past history.

Furthermore,

$$\begin{aligned} \text{Compression} &\leq \tfrac{1}{\ell} H(X_1^{\ell} | X_{-n}^{0}) + \tfrac{1}{\ell} \\ &\leq H(X_1 | X_{-n}^{0}) + \tfrac{1}{\ell} \qquad \text{(Huffman)} \end{aligned}$$

## Universal compression

Assume now that $P(X_{-n}^{\ell})$ is not known. It has been demonstrated that despite of the fact that the underlying probability law is not known, one can still achieve H asymptotically when n tends to infinity.

More precisely,

Let the length function for the vector $X_1^{N'}$ be:

$$L_u(X_1^{N'} | X_{-n}^{0}) = L(X_1^{\ell_1} | X_{-n}^{0}) + L(X_{\ell_1+1}^{\ell_2} | X_{-n+\ell_1}^{\ell_1}) + \cdots$$

where $\ell_i \leq \ell$ ($\ell$ is the maximum latency). (A "sliding -window" algorithm). Then, there exist universal coding algorithms for which

$$\lim_{n \to \infty} \frac{1}{N'} E L_u(X_1^{N'} | X_{-n+1}^{0}) = H \text{ if } \ell = O(\log n).$$

(See for example [1] [2])

**Problem:** Given a total memory constraint $N = n + \ell$, is it still possible to get a compression close to $H(X_1 | X_{-n}^{0}) > H$?

Unfortunately, the answer is negative.

**Definition:Recurrence time**

Let $N(X_{-t}^{1},\ X_{-n}^{0})$ be the *smallest* integer $i$ such that

$$X_{-t}^{1} = X_{-i-t}^{-i+1}\ ; \quad 0 < i < n - t + 1$$

if no such $i$ is found, $N(X_{-t}^{1},\ X_{-n}^{0}) \stackrel{\triangle}{=} n + 1$.

Also let $K(X_{-n}^{1})$ be the largest integer $t > 0$ such that

$$N(X_{-t}^{1},\, X_{-n}^{0}) < n + 1\ ,$$

if no such $t$ is found, $K(X_{-n}^{1}) \stackrel{\triangle}{=} 0$.

Example:

$$\underbrace{0\,1\,0\,1\,0\,0\,0}_{X_{-n}^{-3}}\ \underbrace{0\,1\,0\,|\,1}_{X_{-2}^{1}} \qquad\qquad K(X_{-n}^{1}) = 2$$

Define

$$H\left(X_1|X_{-K(X_{-n}^{1})}^{o}\right) = -E\log P\left(X_1|X_{-K(X_{-n}^{1})}^{0}\right)$$

Then,

**Claim** (converse): For *any* universal noiseless compression encoder (i.e. an encoder that does not depend on the source), with latency $\leq 0(\log n)$ and memory constraint $n$, there exist some stationary ergodic sources for which:

$$\text{Compression } \geq H(X_1|X_{-K(X_{-n}^{1})}^{0}) - \frac{0(\log\log n)}{\log n} > H$$

**Remark:** The Lempel-Ziv family of universal data compression algorithms yields a compression which approaches the entropy of the source, when n, the length of the sequence, gets large. The redundancy for the LZ algorithm was shown to be upper bounded by $0(\frac{\log\log n}{\log n})$ for "large enough" n [3].

Recently,[Szpankowski 1997, Savari 1997], the redundancy for the LZ algorithm when applied to Markov processes was shown to be bounded by $0(\frac{1}{\log n})$ as n gets large.

However, in practice we are interested in cases where n is not large enough for these asymptotic results to apply.

The lower- bound on the compression that appeared above holds for ANY vanishing-memory ergodic source [4] when applied to the family of LZ-type algorithms:

$$\text{Compression } \geq H(X_1|X_{-K(X_{-n}^{1})}^{0}) - \frac{0(\log\log n)}{\log n}$$

**Example:** Let $Z_1^n$ be a random i.i.d vector with equally probable letters (i.e. $Z_1^n$ is "purely" random). Also let

$$\mathbf{X} = \ldots,\ X_{-i}^{n-i-1},\ X_{n-i}^{2n-i-1}, \ldots$$

$$X_{-i+Kn}^{(K+1)n-i-1} = Z_1^n\ ; \quad K = -2,\ -1,\ 0,\ 1,\ 2$$

and where is a random "phase", uniformly distributed over $[0,\ n]$

$$\ldots 0100011011010001101101000110110100011011 \ldots$$

Thus, $\mathbf{X}$ is a stationary ergodic process, where

$$H(X_1|X_{-n}^o) = 0\ .$$

However, $H(X_1|X^o_{-K(X_{-n}^1)}) = 1$!

### Good news:

It is possible to achieve a universal compression such that,

$$\text{Compression}\ \ \leq H\Big(X_1|X^o_{-K(X_{-n}^1)}\Big) + 0\left(\frac{0(\log\log n)}{\log n}\right)$$

This is achieved by the HZ *conditional* string-matching context algorithm (conditional LZ) [5].

### The HZ context algorithm:

The algorithm is outlined by the following example:

$$\underbrace{0\,1\,0\,1\,0\,0\,0}_{X_{-n}^{-2}}\ \underbrace{0\,1\,|\,0\,1}_{X_{-1}^{2}} \qquad \ell = 2$$

$$X_1^\ell = 0\ 1$$

**Step 1:** Find the longest suffix $X^o_{-K}$ of $X^o_{-n}$ such that

$$X_{-K}^\ell = X_{-K-j}^{\ell-j}\ , \quad 1 \leq j \leq n-K+1$$

In our case $\boxed{K=1}$ $\qquad j = 10$

**Step 2:** Among all substrings of $X_{-n}^{\ell-1}$ of length $\ell+K+1$, list those that start with $X_{-K}^{o}$ (*e.g.*01)

**Step 3:** Generate a pointer to a substring with the one (among few, perhaps) substring that is identical to $X_{-K}^{\ell}$.

**Step 4:** The code-word is a concatenation
(in binary form) of $K$ ($K=1$) and the pointer (1; not 10!!!)

## B) Universal prediction and classification with memory constraints

Given a sequence of length $n$, $X_{-n+1}^{0}$, we would like to generate an empirical measure $Q(X_1^{\ell}|X_{-n+1}^{0})$ for the next incoming $\ell$ letters, that will be close to the "true" probability measure that generated $X_{-n+1}^{0}$.
If $X_{-n+1}^{\ell}$ is generated by $P(X_{-n+1}^{\ell})$

$$D_{X_{-n+1}^{0}}(P \parallel Q) = \frac{1}{\ell} E \log \frac{P(X_1^{\ell}|x_{-n+1}^{0})}{Q(X_1^{\ell}|X_{-n+1}^{0})} \leq \varepsilon$$

**Claim (converse)**: At least for some stationary ergodic sources

$$D_{X_{-n+1}^{0}}(P \parallel Q) \geq H(X_1|X_{-K(X_{-n}^{1})}^{0}) - H(X_1|X_{-n+1}^{0}) - 0\left(\frac{\log\log n)}{\log n}\right)$$

This follows from the fact that $-\log Q(X_1^{\ell}|X_{-n+1}^{0})$ is a proper length-function. Thus the results of the previous section may be applied.

For large $n$, $H(X_1|X_{-n+1}^{0}) \approx H \approx H(X_1|X_{-K}^{o})$, for some $K$ (the "memory" of the source). Hence, unless $n$ is large enough so as to make, with high probability, $K(X_{-n}^{1}) \geq K$, no universal prediction (or classification) is possible.

**Claim**: Let

$$Q(X_1^{\ell}|X_{-n+1}^{0}) = \frac{2^{-L_{HZ}(X_1^{\ell}|X_{-n+1}^{0})}}{\sum 2^{-L_{HZ}(X_1^{\ell}|X_{-n+1}^{0})}}$$

where $L_{HZ}(X_1^{\ell}|X_{-n+1}^{0})$ is the length function of the HZ universal encoder. Then:

$$D_{X_{-n+1}^{0}}(P \parallel Q) \leq H(X_1|X_{-K(X_{-n}^{1})}^{0}) - H(X_1|X_{-n+1}^{0}) + 0\left(\frac{\log\log n)}{\log n}\right)$$

## C) Universal "noisy" compression with memory and latency constraints ([6])

### Distortion-rate function (for mean square-error distortion)

$$\Delta_\ell(Q) = \frac{1}{\ell} E \parallel X_1^\ell - Y_1^\ell \parallel^2$$

$Y_1^\ell = Q(X_1^\ell)$ (Quantizer) ;
$Y_1^\ell \in \{Y_1, Y_2, \ldots, Y_i, \ldots, Y_{2^{\ell R}}\}$ ; $Y_i \in \mathbb{R}^\ell$ .

$$D_\ell(R) = \min_Q \Delta(Q)$$

$$\boxed{D_\ell(R) = \min_{Q: \frac{1}{\ell} I(X_1^\ell; Y_1^\ell) \leq R} \frac{1}{\ell} E \parallel X_1^\ell - Y_1^\ell \parallel^2}$$

$$\boxed{\begin{array}{c} D(R) = \lim_{\ell \to \infty} D_\ell(R) . \\ \text{is the Distortion} - \text{Rate Function} \end{array}}$$

Clearly $D_\ell(R) > D(R)$.

What happens if instead of full information about the $\ell$-th order statistics $P_\ell$, we are given only $N$ information bits of an arbitrary representation of this information?

Clearly, for each such $N$ bit vector there may correspond a particular vector-quantizer among a family of $2^N$ vector-quantizers.

It turns out that if $N < 2^{(R-\delta)l}$, it is not possible to achieve the minimal distortion that is achievable when the $\ell$-th order statistics is fully known.

**Claim:** (converse) Let $R > 0$ be given. Then, for every $\varepsilon > 0$ and $\delta > 0$, if $N < 2^{(R-\delta)l}$ and $l$ is sufficiently large, then for any deterministic $N$-bit representation $F : \mathcal{P}_l \to \{0,1\}^N$ and any set of $2^N$ rate $R$, $l$-dimensional vector quantizers $\left\{Q_b,\ b \in \{0,1\}^N\right\}$, there exists a stationary and ergodic process $\mu \in M$ whose $l$-th order marginal PDF $P$ satisfies $D_l(R) > \varepsilon$, and at the same time

$$\Delta(Q_{F(P)}) > 2D_l(R) - \varepsilon$$

The class of processes $M$ consists of processes for which, for any given $\varepsilon > 0$ and given $\ell$

$$\frac{1}{\ell} E\Big( \| X_1^\ell \|^2 \cdot 1\Big\{ \| X_1^\ell \|^2 > B(\ell, \epsilon) \cdot \ell \Big\}\Big) \leq \varepsilon$$

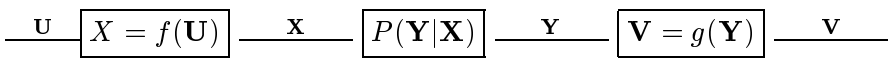
where $B(\ell, \epsilon)$ is a bounded positive number.

The good news are that for *any* $\mu \in M$, a distortion $= D_\ell(R) + \varepsilon$ is achievable with a training sequence consisting of $m$ independent drawings of $\ell$-vectors of the given process where $m = 2^{(R+\delta)\ell}$ and hence

$$N = 2^{(R+\delta)\ell} \ell \log A \, .$$

(Extension of Linder, Lugosi and Zeger, 1994). In all of the above, $N$ is finite, but is pretty large (i.e $N \approx 2^{\ell R}$).

## D) Results for "really small" values of n




The classical distortion-rate function:

$$D(R) \stackrel{\triangle}{=} \inf_{P(Y|X): I(X,Y) \leq R} E d(X, Y)$$

$$\frac{1}{n} E d(\mathbf{X}, \mathbf{Y}) \geq D(C)$$

$$C = \text{channel capacity} \stackrel{\triangle}{=} \sup_{P(X) \in \mathbb{P}(\mathbf{X})} I(X, Y)$$

### Generalized results (Z&Z '73)

Let $Q(\cdot)$ be a concave non increasing function on $(0, \infty)$ satisfying some additional constraints: (a.e. $-\log X$, $e^{-X}$, etc.). Define

$$I^Q(\mathbf{X}, \mathbf{Y}) = \iint P(\mathbf{X}, \mathbf{Y}) \, Q \frac{P(\mathbf{X}), P(\mathbf{Y})}{P(\mathbf{X}, \mathbf{Y})} \, dX \, dY$$

Remark: Note that if $(X_1, Y_1)$, $(X_2, Y_2)$ are independent pairs, then in general

$$I^Q\left[(X_1,\, Y_1),\, (X_2,\, Y_2)\right] \neq I^Q(X_1,\, Y_1) + I^Q(X_2,\, Y_2)$$

(It is for $Q(X) = \log X!$)

### A generalized data processing theorem:

$$I^Q(\mathbf{X}, \mathbf{Y}) \geq I^Q(\mathbf{U},\, \mathbf{V})$$

### A generalized distortion-rate bound:

Let

$$D_n^Q(R) = \inf_{I^Q(U_1^n, V_1^n) \leq R} \frac{1}{n}\, d(X_1^n,\, Y_1^n)$$

Let

$$C_n^Q(R) = \sup_{P(\mathbf{X}) \in \mathbb{P}} \frac{1}{n}\, I^Q(X_1,\, Y_1^n)$$

Then

$$\boxed{\tfrac{1}{n}\, d(U_1^n,\, V_1^n) \geq D_n^Q(C_n^Q)}$$

**Example:** $n = 1$ !

$$Q(X) = \begin{cases} 1 - \alpha X & ; \quad 0 \leq X \leq \frac{1}{2} \\ 0 & ; \quad X > \frac{1}{\alpha} \end{cases}$$

$U$ is evenly distributed on a circle of radius $\frac{1}{2\pi}$. The distortion between two points on the circle is the length of the shorter connecting arc, raised to the second power.

Hence, $d(U, V) \leq (\frac{1}{2})^2$

$$Q(X) = \begin{cases} 1 - \alpha X & ; \quad 0 \leq X \leq \frac{1}{2} \\ 0 & ; \quad X > \frac{1}{2} \end{cases}$$

$$
\begin{array}{c}
X_1 \longrightarrow X_1 \\
X_2 \longrightarrow X_2 \\
\vdots \\
\vdots \\
X_M \longrightarrow X_M
\end{array}
$$

channel

Let $\alpha = \frac{M}{2}$, $\varepsilon^2 = d(U,V) \geq \frac{1}{24M^2}$.

As compared with the "classical" result

$$\varepsilon^2 \geq \frac{1}{24\pi\ell M^2}$$

## References


[1] J. Ziv and A. Lempel:"A Universal Algorithm for Sequential Date-Compression" Lempel), *IEEE Trans. on Inf. Theory*, Vol. IT–23, No. 3, May 1977, pp. 337–343.

[2] A.D. Wyner and J. Ziv, "The Sliding-Window Lempel-Ziv Algorithm is Asymptotically y Optimal", (Invited paper), *Proceedings of the IEEE,* **82**, June 1994, pp. 872-877.

[3] E.Plotnik, M.Weinberger and J.Ziv "Upper Bounds on the Probability of a Sequence Emitted from a Finite-State Source and on the Redundancy of the Lempel-Ziv Data Compression Algorithm", *IEEE Trans. on Information Theory*, pp. 66-72, January 1992.

[4] A.D.Wyner and J.Ziv "Classification with Finite Memory", *IEEE Trans. on Information Theory* Vol. 42, No. 2, March 1996, pp. 337-347.

[5] Y. Hershkovits and J.Ziv "On Sliding-Window Universal Data Compression with Limited-Memory" *submitted to the IEEE Trans. on Information Theory.*

[6] N.Merhav and J.Ziv "On the Amount of Side Information Required for Lossy Data Compression" *IEEE Trans. on Information Theory, Vol 43, July 1997, pp 1112-1121.*

[7] M.Zakai and J.Ziv "On Functionals Satisfying a Data-Processing Theory" *IEEE Trans. on Inf. Theory*, Vol. IT–19, No. 3, May 1973, pp. 275–283.