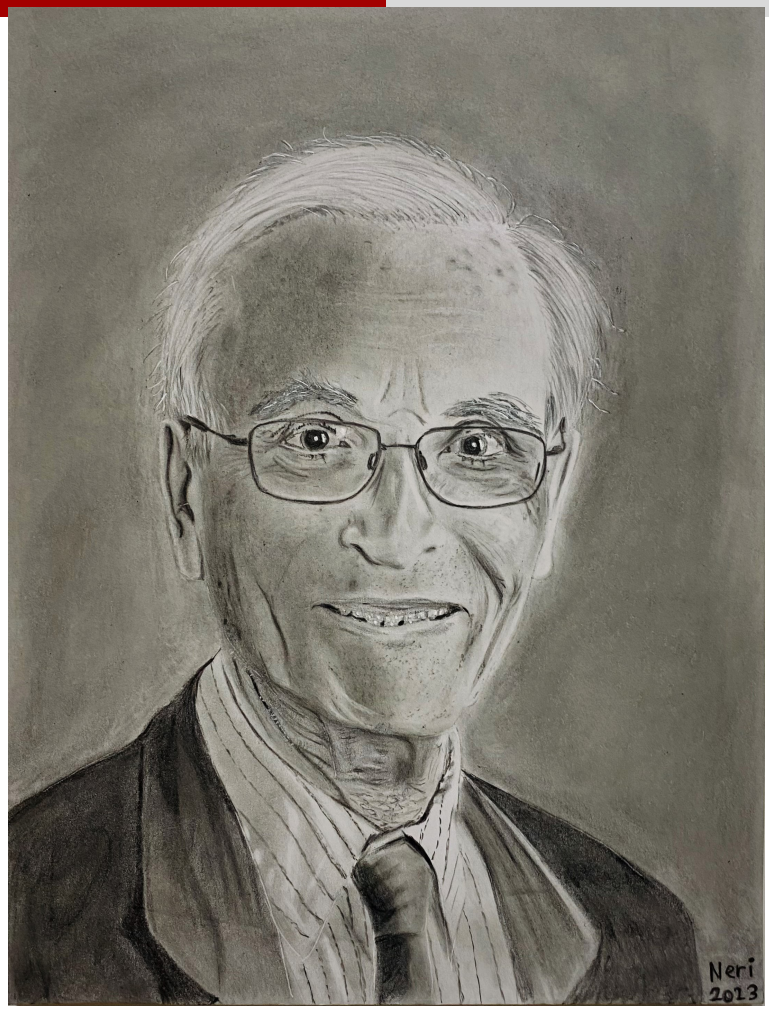
On Jacob Ziv's Individual-Sequence Approach to Information Theory

Neri Merhav

The Viterbi Faculty of Electrical & Computer Engineering
Technion—Israel Institute of Technology
Haifa, Israel

The Annual Jacob Ziv Memorial Lecture, March 21, 2024



In memory of Jacob Ziv,
a shining star in the sky of information theory
and a great inspiration to me and to many others,
for years to come.

Classical Information Theory – Shannon Theory

- Fundamental limits vs. achievable performance of information processing (compression, error correction coding, encryption, etc.).
- Based on memoryless (i.i.d.) probabilistic models of sources and channels.

Information Theory and Probabilistic Modeling

Claude Elwood Shannon

Sources: Random variables or random processes

Channels: Stochastic functions of the output, given the input

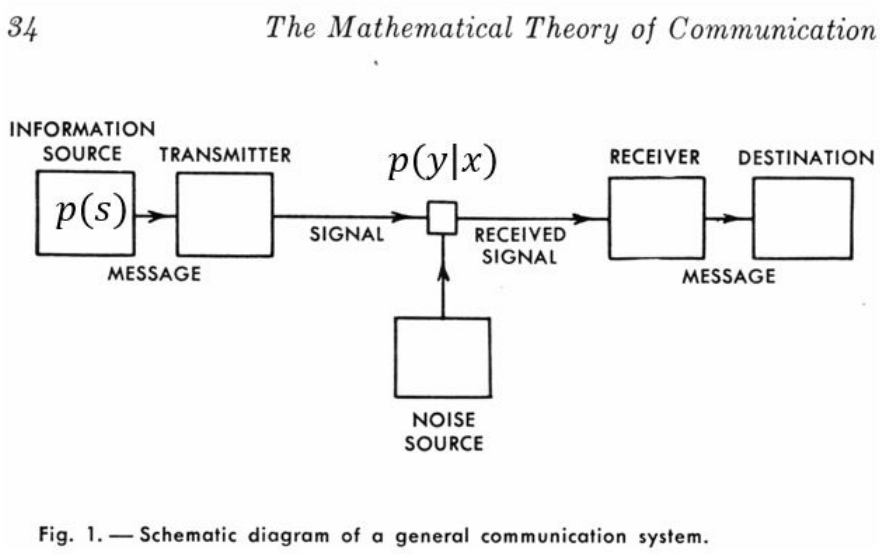
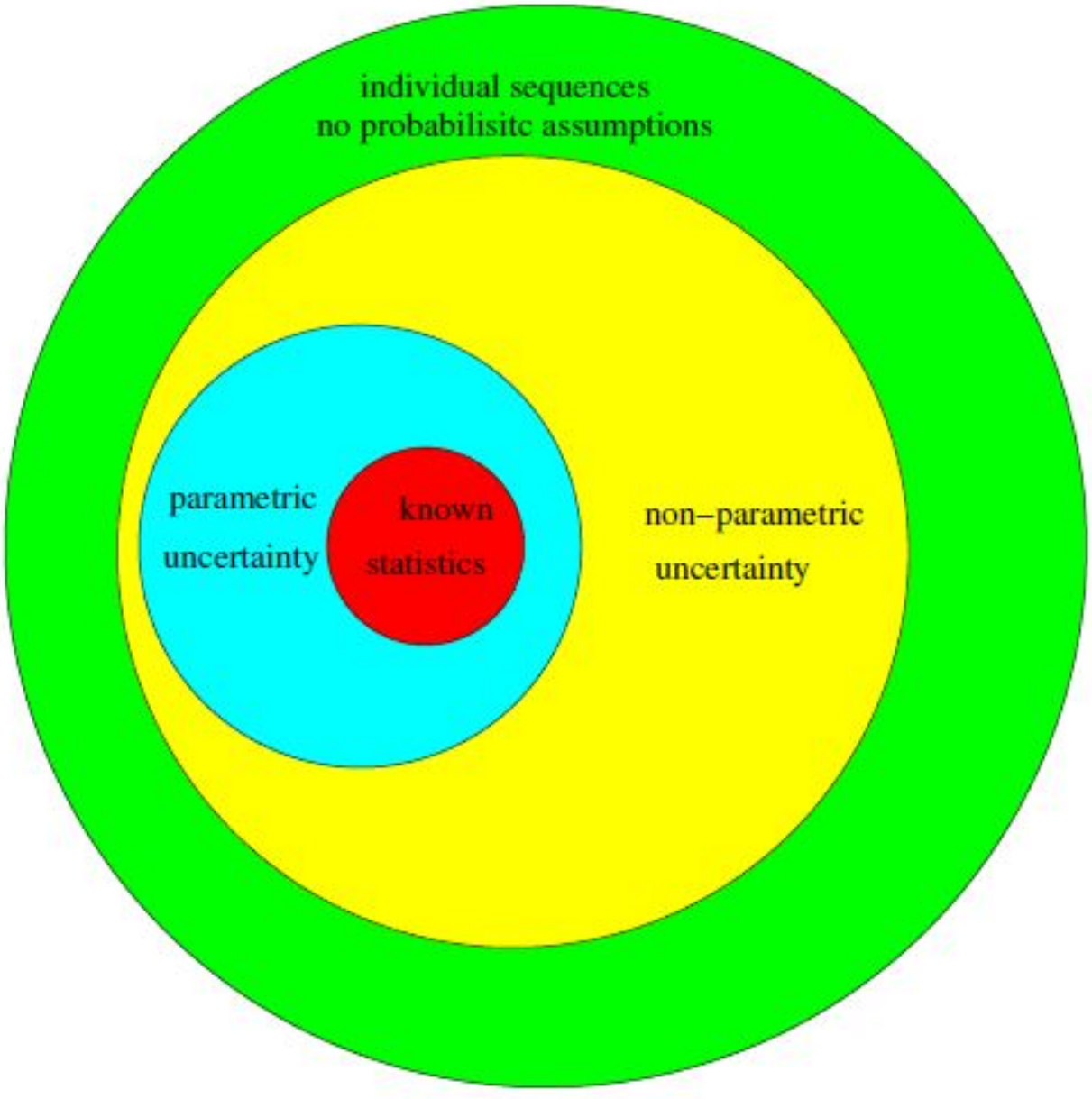


Fig. 1. — Schematic diagram of a general communication system.

Later Developments - More Realisic Assumptions

- Sources/channels with memory:
  - Markov model
  - hidden Markov model
  - finite-state source/channel model
  - general
- Relaxing the assumption of fully-known probability distributions:
  - Robustness to model uncertainty (worst-case approach):
    - robust estimation
    - robust hypothesis testing (and signal detection)
    - robust filtering
  - Universal methods:
    - data compression
    - channel coding/decoding
    - prediction
    - signal detection.

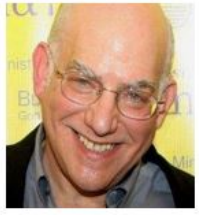

individual sequences
no probabilisitc assumptions
parametric
uncertainty
known
statistics
non−parametric
uncertainty

Claude Elwood Shannon

entropy = complexity of a random sequence

Andrei Nikolaevich Kolmogorov

Ray Solomonoff



Gregory Chaitin

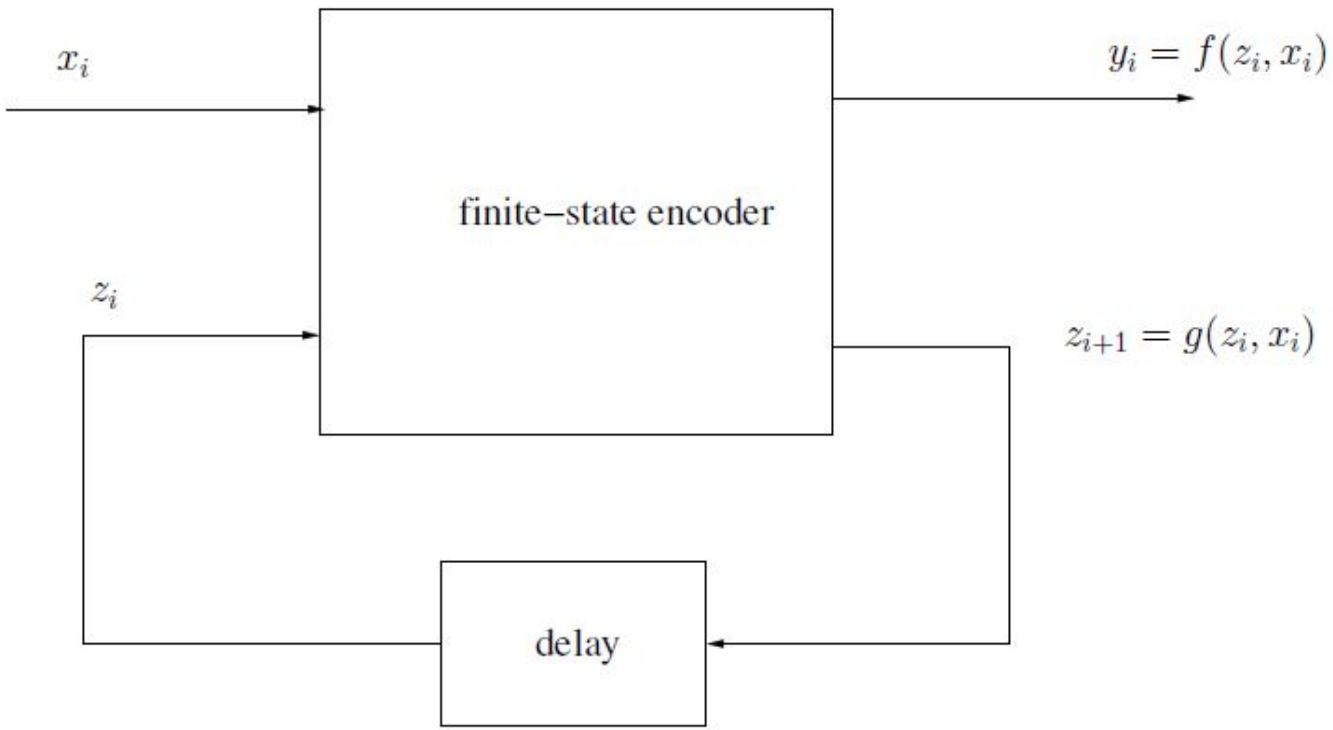
algorithmic complexity of an individual sequence

Abraham Lempel

Jacob Ziv

finite-state complexity of an individual sequence

Individual Sequences & F-S Encoders (Ziv-Lempel '78)



x_i = finite-alphabet sequence to be compressed.
z_i = encoder state. Total number of states = s
y_i = a variable-length binary string (possibly, empty for some i).

Individual Sequences & F-S Encoders (Cont'd)

For a given individual sequence, $(x_1, \ldots, x_n)$, what is the best compression ratio that can be achieved by any information-lossless F-S encoder with s states?

$$\rho_s(x_1, \ldots, x_n) = \min_{\{\text{all } s\text{-state encoders}\}} \frac{\text{length of compressed file}}{n}.$$

The choice of the best s-state encoder depends on the given $(x_1, \ldots, x_n)$.

Nevertheless, Ziv and Lempel developed a universal compression algorithm (LZ78), that always nearly attains $\rho_s(x_1, \ldots, x_n)$ as long as $s \ll n$.

Individual Sequences & F-S Encoders (Cont'd)

How can we quantify $\rho_s(x_1, \ldots, x_n)$?

Let us parse $(x_1, \ldots, x_n)$ sequentially such that each new phrase is the shortest string that has not been encountered before as a parsed phrase.

Example: $n = 99$ and $(x_1, \ldots, x_{99})$ is given by:
whatdoesitmeanwhatdoesitmeanwhatmeanswhatdoesmeansdoesmeanmeansmean whatdoesitmeanmeanswhatdoesitmean

which is parsed as:

w,h,a,t,d,o,e,s,i,tm,ea,n,wh,at,do,es,it,m,ean,wha,tme,an,sw,ha,td,oe, sm,ean,sd,oes,me,anm,eans,mea,nw,hat,doe,si,tmea,nm,eansw,hatd,oesi,tmean

Let c be the number of phrases. In our example, $c = 44$.

For repetitive/predictable sequences, the phrases grow quickly and then c is small.

For non-repetitive/unpredictable sequences, phrases grow slowly and c is large.

Individual Sequences & F-S Encoders (Cont'd)

Ziv and Lempel's 1978 article contains two main results:

1. For $s \ll n$, $\rho_s(x_1, \ldots, x_n)$ cannot be *much smaller* than $\frac{c \log c}{n}$;
2. The LZ78 algorithm achieves compression ratio *not much larger* than $\frac{c \log c}{n}$.

F-S complexity = deterministic analogue of entropy = $\frac{c \log c}{n}$.

The LZ78 algorithm compresses each phrase of length ℓ as follows:

1. First $\ell - 1$ symbols: send a pointer to its copy in the (already decoded) past.
2. Last symbol: send uncompressed.

There are several versions of the LZ algorithm, which are all based on the idea of string matching.

The Impact and the Usefulness of LZ Algorithms

The LZ compression methods are among the most popular algorithms for lossless storage.

DEFLATE is a variation on LZ optimized for decompression speed and compression ratio.

In the mid-1980s, following work by Terry Welch, the Lempel-Ziv-Welch (LZW) algorithm rapidly became the method of choice for most general-purpose compression systems.

LZW is used in GIF images, programs such as PKZIP, and hardware devices such as modems.

Also harnessed for PDF, TIFF, PNG, ZIP, video formats such as MP3, and in cellphones.

We all use the LZ algorithm on a daily basis without even being aware.

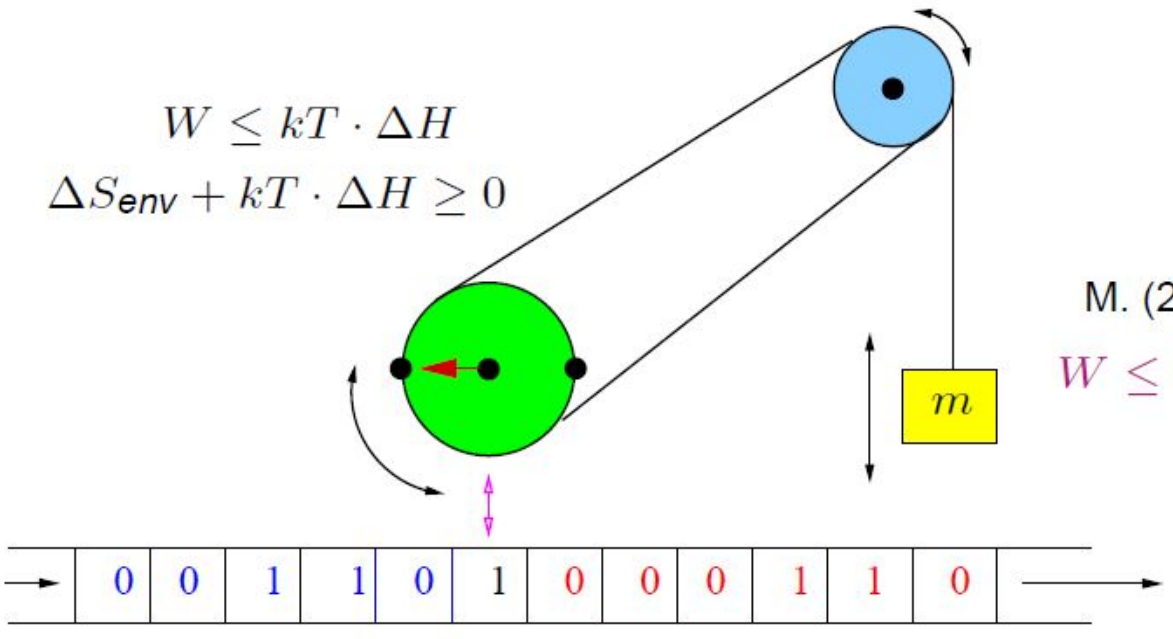
In 2004, the IEEE proclaimed the Lempel-Ziv algorithm a Milestone in Electrical Engineering and Computing.

LZ Compressibility in the Role of Thermodynamic Entropy

Mandal & Jarzynski (2012): system converting thermal fluctuations to work while writing info.

$$W \leq kT \cdot \Delta H$$

$$\Delta S_{env} + kT \cdot \Delta H \geq 0$$



M. (2015):

$$W \leq kT \cdot F[\text{LZ complexity}]$$

The LZ78 Algorithm as an Engine for Other Tasks

The LZ78 algorithm is harnessed as an engine in universally optimal methods for tasks other than data compression:

- hypothesis testing (e.g., testing for independence, testing for randomness)
- model order estimation (for Markov and hidden Markov models)
- coding and decoding for unknown channels
- encryption
- time-series prediction
- filtering
- guessing
- universal ensembles for lossy compression

In addition, the setting of finite-state encoding and decoding of individual sequences was expanded to more general scenarios, such as: lossy compression, compression with side information, joint source-channel coding, etc.

Lempel-Ziv Complexity and the Individual-Sequence Approach to Information Theory

1. M. Feder, N. Merhav, and M. Gutman, ``Universal prediction of individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1258-1270, July 1992.
2. N. Merhav and M. Feder, ``Universal schemes for sequential decision from individual data sequences,'' *IEEE Trans. Inform. Theory* vol. 39, no. 4, pp. 1280-1291, July 1993.
3. J. Ziv and N. Merhav, ``A measure of relative entropy between individual sequences with application to universal classification,'' *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1270-1279, July 1993.
4. M. J. Weinberger, N. Merhav, and M. Feder, ``Optimal sequential probability assignment for individual sequences,'' *IEEE Trans. Inform. Theory* vol. 40, no. 2, pp. 384-396, March 1994.
5. N. Merhav and M. Feder, ``On the cost of universality of block codes for individual sequences,'' *Proc. 1994 IEEE Int. Symp. on Information Theory* (ISIT `94), p. 263, Trondheim, Norway, June 1994.
6. N. Merhav and M. Feder, ``Universal prediction,'' (invited paper) *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124-2147, October 1998.
7. A. Baruch and N. Merhav, ``Universal filtering and prediction of individual sequences corrupted by noise using the Lempel-Ziv algorithm,'' *Proc. ISIT 2000*, p. 99, Sorrento, Italy, June 2000.
8. T. Weissman and N. Merhav, ``Universal prediction of individual sequences in the presence of arbitrarily varying, memoryless noise,'' *Proc. ISIT 2000*, p. 97, Sorrento, Italy, June 2000.
9. N. Merhav, ``Universal detection of messages via finite-state channels,'' *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242-2246, September 2000.
10. T. Weissman, N. Merhav, and A. Somekh-Baruch, ``Twofold universal prediction of a noisy individual sequence,'' *IEEE Trans. Inform. Theory*, vol 47, no. 5, pp. 1849-1866, July 2001.
11. T. Weissman and N. Merhav, ``Universal prediction of binary individual sequences in the presence of noise,'' *IEEE Trans. Inform. Theory*, vol. 47, no. 6, pp. 2151-2173, September 2001.
12. T. Weissman and N. Merhav, ``On limited-delay lossy coding and filtering of individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 721-733, March 2002.
13. N. Merhav, E. Ordentlich, G. Seroussi, and M. J. Weinberger, ``On sequential strategies for loss functions with memory,'' *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1947-1958, July 2002.
14. E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, ``Discrete universal filtering through incremental parsing,'' *Proc. DCC 2004*, Snowbird, Utah, March 2004.
15. N. Merhav and J. Ziv, ``On the Wyner-Ziv problem for individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 867-873, March 2006.
16. J. Ziv and N. Merhav, ``On context-tree prediction of individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1860-1866, May 2007.
17. T. Weissman, E. Ordentlich, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav," ``Universal filtering via prediction,'' *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1253-1264, April 2007.
18. N. Merhav and E. Sabbag, ``Optimal watermark embedding and detection strategies under limited detection resources,'' *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 255-274, January 2008.
19. A. Reani and N. Merhav, ``Efficient on-line schemes for encoding individual sequences with side information at the decoder,'' *Proc. ISIT 2009*, Seoul, Korea, June-July 2009.
20. A. Martin, N. Merhav, G. Seroussi, and M. J. Weinberger, ``Twice-universal simulation of Markov sources and individual sequences,'' *Proc. ISIT 2007*, pp. 2876-2880, Nice, France, June 2007.
21. N. Merhav, ``Perfectly secure encryption of individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1302-1310, March 2013.
22. N. Merhav, ``Universal decoding for arbitrary channels relative to a given class of decoding metrics,'' *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5566-5576, September 2013.
23. N. Merhav, ``On the data processing theorem in the semi-deterministic setting,'' *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6032-6040, October 2014.
24. N. Merhav, ``Sequence complexity and work extraction,'' *Journal of Statistical Mechanics: Theory and Experiment*, P06037, June 2015. doi:10.1088/1742-5468/2015/06/P06037
25. N. Merhav, ``On empirical cumulant generating functions of code lengths for individual sequences,'' *IEEE Trans. Inform. Theory*, vol. 63, no. 12, pp. 7729-7736, December 2017.
26. N. Merhav, ``Universal decoding using a noisy codebook,'' *IEEE Trans. Inform. Theory*, vol. 64, part 1, no. 4, pp. 2231-2239, April 2018.
27. N. Merhav, ``Guessing individual sequences: generating randomized guesses using finite-state machines,'' *IEEE Trans. Inform. Theory*, vol. 66, no. 5, pp. 2912-2920, May 2020.
28. N. Merhav, ``Finite-state source-channel coding for individual source sequences with source side information at the decoder,'' *IEEE Trans. Inform. Theory*, vol. 68, no. 3, pp. 1532-1544, March 2022.
29. N. Merhav, ``Encoding individual source sequences for the wiretap channel.'' *Entropy*, 23(12) 1694, December 17, 2021.
30. N. Merhav, ``A universal ensemble for sample-wise lossy compression,'' submitted for publication.
31. N. Merhav, ``Lossy compression of individual sequences revisited: fundamental limits of finite-state encoders,'' submitted for publication.

Ziv's Inequality (Plotnik, Weinberger and Ziv, '92)

Ziv's inequality is a powerful tool for proving the asymptotic optimality of the LZ mechanism at the service of some of these tasks.

It states that for every Markov process P:

$$\log[P(x_1) \cdot P(x_2|x_1) \cdot P(x_3|x_2) \cdots P(x_n|x_{n-1})] \leq -c \log c + \text{some small terms}$$

More generally, it remains true for higher order Markov processes, as well as general finite-state processes, including hidden Markov processes.

A different form of Ziv's inequality is the following:

$$\log[\#\text{ sequences with the same probability as } (x_1, \ldots, x_n)] \geq c \log c - \text{small terms.}$$

Using LZ for Hypothesis Testing

We are given a binary sequence, $(x_1, \ldots, x_n)$, and we wish to decide between two hypotheses:

$\mathcal{H}_0$: - $(x_1, \ldots, x_n)$ is a sequence of n independent fair coin tosses.
$\mathcal{H}_1$: - $(x_1, \ldots, x_n)$ is **not** a sequence of n independent fair coin tosses.

The difficulty is that under $\mathcal{H}_1$, we don't know how the sequence was generated.

It turns out that the following decision criterion gives the best trade-off between the two kinds of errors:

Compare the compression ratio, $\frac{c \log c}{n}$, to a threshold T: If $\frac{c \log c}{n} > T$, accept $\mathcal{H}_0$, otherwise, reject it.

The choice of T controls the balance between the two kinds of decision errors.

Markov Order Estimation (Merhav, Ph.D. thesis, '88)

We know that $(x_1, \ldots, x_n)$ was generated by some Markov process of order k:

$$P(x_1, \ldots, x_n) = P(x_1, \ldots, x_k) \cdot P(x_{k+1}|x_1, \ldots, x_k) \cdots P(x_n|x_{n-k}, \ldots, x_{n-1}),$$

but we don't know the order k and we wish to estimate it.

The following estimator provides the best balance between the overestimation and the underestimation errors:

- Compress $(x_1, \ldots, x_n)$ using LZ and calculate the compression ratio, $\frac{c \log c}{n}$.
- For $i = 0, 1, 2, \ldots, K$, compress $(x_1, \ldots, x_n)$ under the model of an i-th order Markov process and calculate the compression ratio, ρ_i.
- For $i = 0, 1, 2, \ldots, K$, calculate the difference, $\rho_i - \frac{c \log c}{n}$.
- The first i for which $\rho_i - \frac{c \log c}{n} \leq T$ is the estimator of k.

The choice of T controls the balance between the overestimation and underestimation errors.

‘Statistical’ Similarity of Sequences (Ziv & Merhav, ‘93)

Are the following two sequences ‘statistically’ similar?

$$0000111110000001111100000111 \qquad 11111000001111000011111000011111$$

How about the following two?

$$00000000000000000001000000010 \qquad 01000110110000010100111011101111$$

What could be a good measure for ‘statistical’ similarity/dissimiliarity between two deterministic sequences?

'Statistical' Similarity of ... (Cont'd)

For $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, let:

$$\Delta(\boldsymbol{x} \| \boldsymbol{y}) = \frac{c(\boldsymbol{x} \leftarrow \boldsymbol{y}) \log n - c(\boldsymbol{x}) \log c(\boldsymbol{x})}{n}$$

where $c(\boldsymbol{x})$ is c as before and $c(\boldsymbol{x} \leftarrow \boldsymbol{y})$ is the number of phrases of $\boldsymbol{x}$ is with respect to $\boldsymbol{y}$, created in the following manner:

- Find the longest prefix string of $\boldsymbol{x}$ that appears somewhere in $\boldsymbol{y}$, namely, the largest i such that $(x_1, x_2, \ldots, x_i) = (y_j, y_{j+1}, \ldots, y_{j+i-1})$ for some j.
- Continue from x_{i+1} in the same manner until $\boldsymbol{x}$ is exhausted.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are 'similar', the phrases of $\boldsymbol{x}$ w.r.t. $\boldsymbol{y}$ are long and then $c(\boldsymbol{x} \leftarrow \boldsymbol{y})$ is small, which implies small $\Delta(\boldsymbol{x} \| \boldsymbol{y})$.

Example: $n = 11$ and $\boldsymbol{x} = (01111000110)$ and $\boldsymbol{y} = (10010100110)$. Then parsing $\boldsymbol{x}$ with respect to $\boldsymbol{y}$ gives: $(011, 110, 00110)$, and so, $c(\boldsymbol{x} \leftarrow \boldsymbol{y}) = 3$.

‘Statistical’ Similarity of ... (Cont’d)

- $\frac{c \log c}{n} \Leftrightarrow$ entropy
  $\Delta(\boldsymbol{x} \| \boldsymbol{y}) \Leftrightarrow$ divergence between distributions.
- $\Delta(\boldsymbol{x} \| \boldsymbol{y})$ is used for universal classification using training data.
- It discriminates between statistically distinguishable sequences whenever there is some finite-state classifier that does.

Applications:

- Text classification
- ECG-based personal identification and authentication
- Anomaly detection
- Estimation of divergence - used by statistical physicists to assess entropy production and energy dissipation.

Information Theoretic Text Classification Using the Ziv-Merhav Method

David Pereira Coutinho[1] and Mário A.T. Figueiredo[2]

[1] Depart. de Engenharia de Electrónica e Telecomunicações e de Computadores
Instituto Superior de Engenharia de Lisboa
1959-007 Lisboa, Portugal
davidpc@isel.pt

[2] Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa, Portugal
mtf@lx.it.pt

Abstract. Most approaches to text classification rely on some measure of (dis)similarity between sequences of symbols. Information theoretic measures have the advantage of making very few assumptions on the models which are considered to have generated the sequences, and have been the focus of recent interest. This paper addresses the use of the *Ziv-Merhav method* (ZMM) for the estimation of relative entropy (or Kullback-Leibler divergence) from sequences of symbols as a tool for text classification. We describe an implementation of the ZMM based on a modified version of the Lempel-Ziv algorithm (LZ77). Assessing the accuracy of the ZMM on synthetic Markov sequences shows that it yields good estimates of the Kullback-Leibler divergence. Finally, we apply the method in a text classification problem (more specifically, authorship attribution) outperforming a previously proposed (also information theoretic) method.

2010 International Conference on Pattern Recognition

One-Lead ECG-Based Personal Identification Using Ziv-Merhav Cross Parsing

David Pereira Coutinho,
Instituto Superior de Engenharia de Lisboa,
Instituto de Telecomunicações,
and Instituto Superior Técnico,
Lisboa, Portugal
Email: davidpc@cc.isel.pt

Ana L. N. Fred, and Mário A. T. Figueiredo
Instituto de Telecomunicações
and Instituto Superior Técnico,
Lisboa, Portugal
Email: afred@lx.it.pt, mario.figueiredo@lx.it.pt

***Abstract*—The advance of falsification technology increases security concerns and gives biometrics an important role in security solutions. The electrocardiogram (ECG) is an emerging biometric that does not need liveliness verification. There is strong evidence that ECG signals contain sufficient discriminative information to allow the identification of individuals from a large population. Most approaches rely on ECG data and the fiducia of different parts of the heartbeat waveform. However non-fiducial approaches have proved recently to be also effective, and have the advantage of not relying critically on the accurate extraction of fiducia data. In this paper, we propose a new non-fiducial ECG biometric identification method based on data compression techniques, namely the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method relies on a string similarity measure which can be seen as a compression-based approximation of the algorithmic cross complexity.We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals. Our approach achieves 100% subject recognition rate despite the existence of differentiated stress states.**

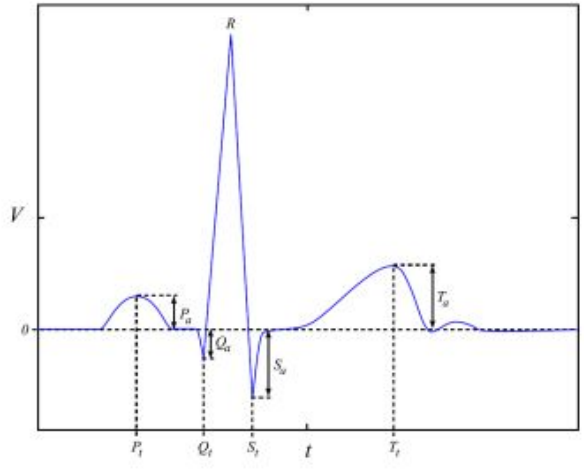


Figure 1. Example of four latency times (features) measured from the P, QRS and T complexes of an ECG heartbeat for fiducial-based feature extraction.

I. Introduction

Biometrics deals with identification of individuals based on their physiological or behavioral characteristics [1] and

A typical ECG signal of a normal heartbeat can be divided into 3 parts, as depicted in Figure 1: the P wave

Fluctuation Theorems (FT)

$$\langle \dot{W}_{diss} \rangle = \langle \dot{W} \rangle - \dot{\Delta F} = T \langle \dot{S} \rangle = \lim_{t \to \infty} \frac{kT}{t} D \left[p\left(\{x(\tau)\}_{\tau=0}^{t} \right) \middle| \middle| p\left(\{x(t-\tau)\}_{\tau=0}^{t} \right) \right]$$

Stochastic discrete processes in stationary regime: $(x_1, x_2, \ldots, x_n)$

$$D_k(p_F || p_B) = \sum_{x_1, \ldots, x_k} p(x_1, \ldots, x_k) \log \frac{p(x_1, \ldots, x_k)}{p(x_k, \ldots, x_1)}$$

$$\frac{\langle \dot{S} \rangle}{k} = d(p_F || p_B) = \lim_{n \to \infty} \frac{1}{n} D_n(p_F || p_B)$$

partial information

$$\frac{\langle \dot{S} \rangle}{k} \geq d(p_F || p_B)$$

Physics (Average)

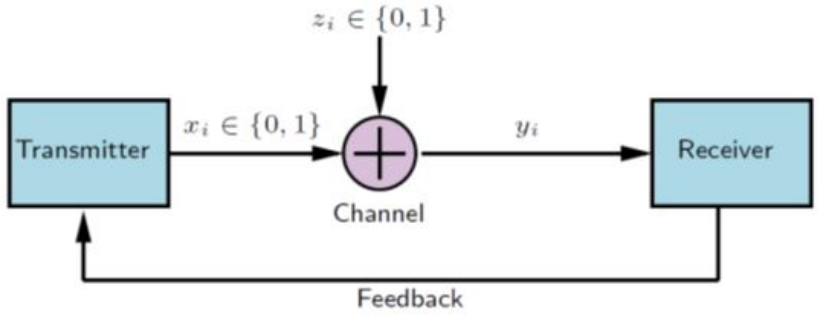
Time series (single trajectory)

Even ignoring some physical details of the system we can estimate its dissipation !

Universal Channel Coding with Feedback

By using the posterior matching scheme with randomization, for binary modulo additive channel with arbitrary noise sequence:

Shayevitz and Feder (2009) [3]

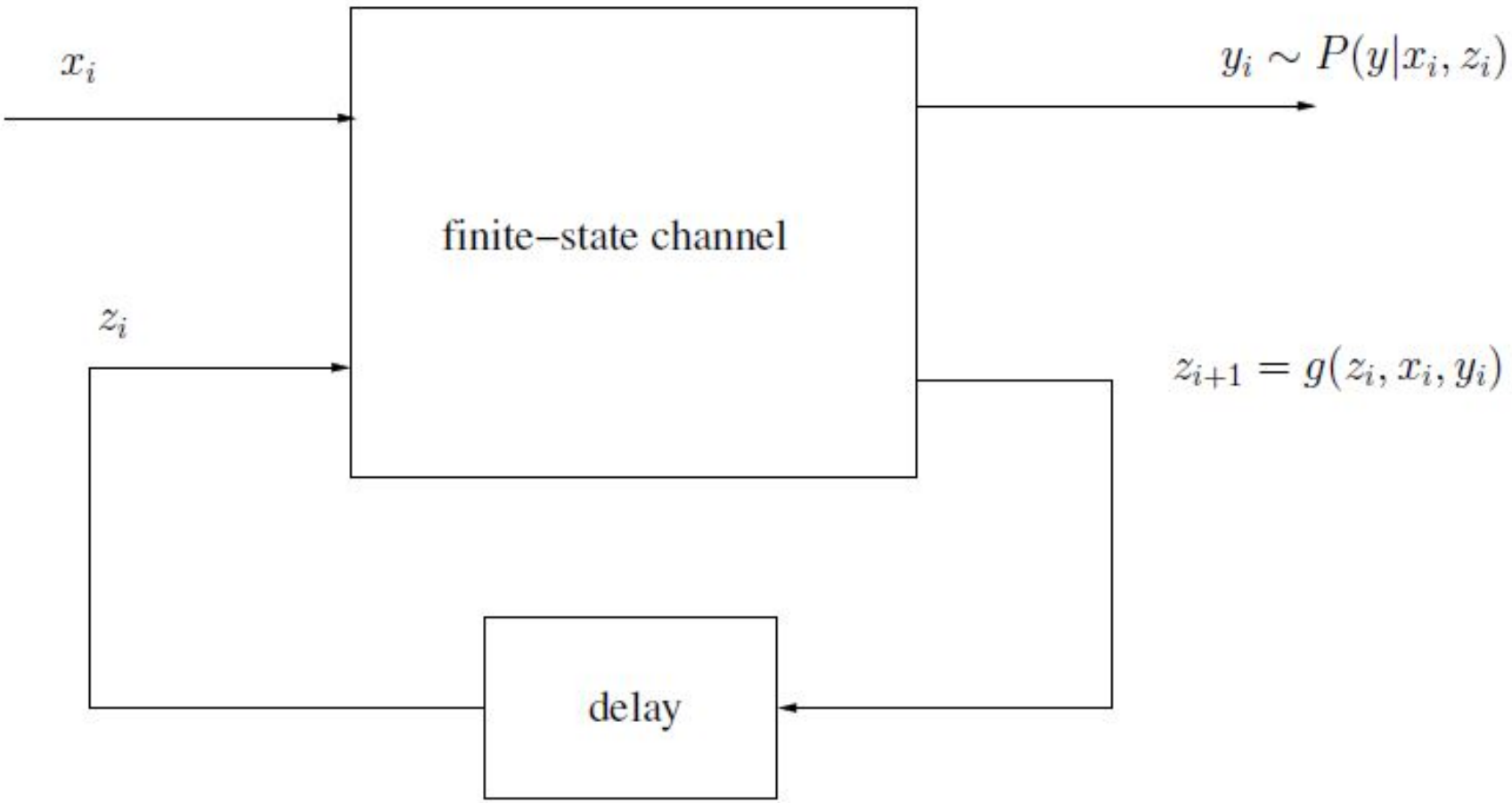


$$R = 1 - \hat{H}(\mathbf{z})$$

Following that, Lomnitz and Feder, Misra and Weissman attained: $R_{\text{emp}} \approx 1 - \rho(\mathbf{z})$

$\rho(\mathbf{z})$ compressibility of $\mathbf{z}$

Universal Decoding for Finite-State Channels (Ziv '85)



Optimal decoding requires knowledge of the channel statistics (P and g).
What if it is not known?

Universal Decoding (Cont'd)

Ziv ('85) proposed a universal decoding criterion.

For two sequences, $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, define:

$c(\boldsymbol{x}, \boldsymbol{y})$ = number of phrases in joint parsing of $((x_1, y_1), (x_2, y_2), \ldots (x_n, y_n))$.
$c(\boldsymbol{y})$ = number of distinct phrases of $\boldsymbol{y}$.
$\boldsymbol{y}(\ell)$ = the ℓ-th distinct phrase of $\boldsymbol{y}$ $(1 \leq \ell \leq c(\boldsymbol{y}))$.
$c_\ell(\boldsymbol{x}|\boldsymbol{y})$ = the number of times that $\boldsymbol{y}(\ell)$ appears in $\boldsymbol{y}$.

Example: Let $n = 6$ and

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \left(\begin{array}{c|c|cc|cc} 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right).$$

Then, $c(\boldsymbol{y}) = 3$ and

$$c_1(\boldsymbol{x}|\boldsymbol{y}) = c_2(\boldsymbol{x}|\boldsymbol{y}) = 1; \quad c_3(\boldsymbol{x}|\boldsymbol{y}) = 2.$$

Universal Decoding (Cont'd)

Now, we define

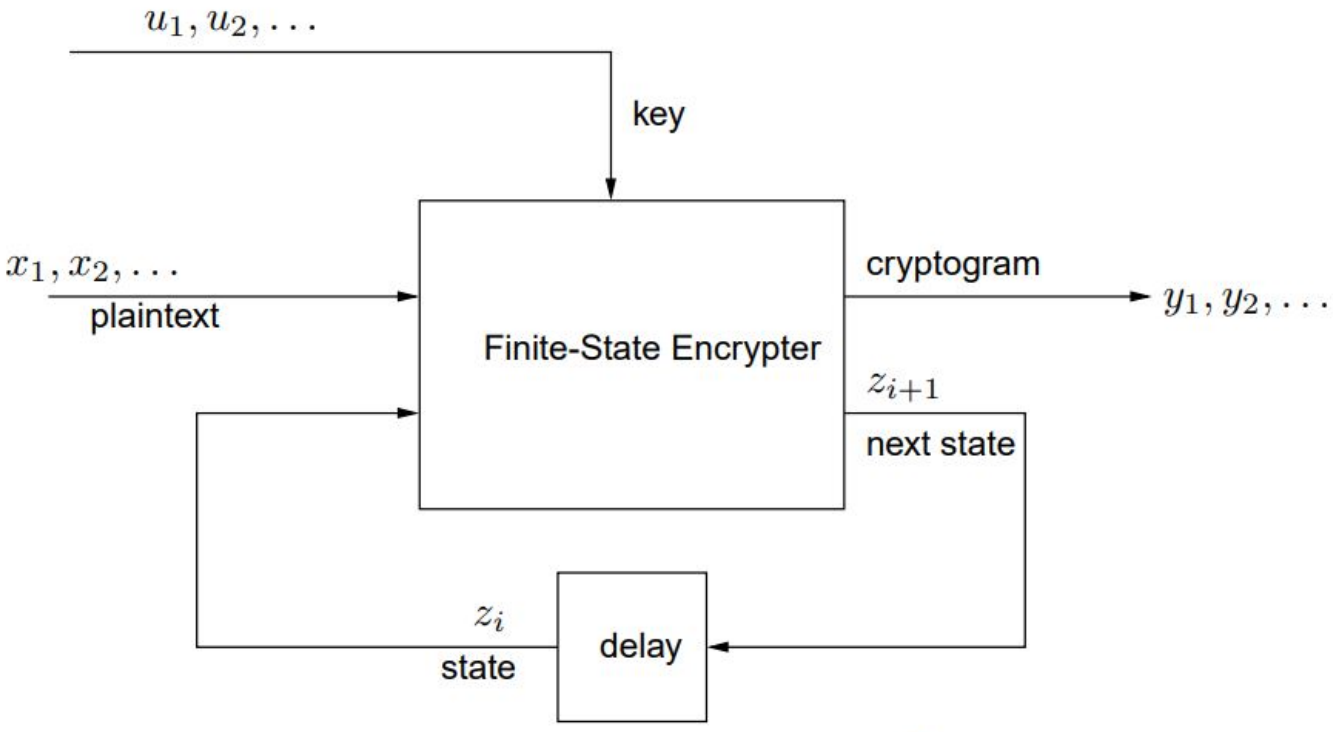
$$u(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\ell=1}^{c(\boldsymbol{y})} c_\ell(\boldsymbol{x}|\boldsymbol{y}) \log c_\ell(\boldsymbol{x}|\boldsymbol{y}).$$

The decoder receives $\boldsymbol{y}$ and calculates $u(\boldsymbol{x}_i, \boldsymbol{y})$ for all possible codewords $\{\boldsymbol{x}_i, i = 1, 2, \ldots, M\}$. The one with the smallest $u(\boldsymbol{x}_i, \boldsymbol{y})$ is the decoded message.

This works essentially as well as the optimal decoder that knows the channel for most codes.

$u(\boldsymbol{x}|\boldsymbol{y})$ has the meaning of conditional complexity of $\boldsymbol{x}$ in the presence of $\boldsymbol{y}$.

Encryption (Merhav '13)



$$
\begin{aligned}
t_i &= t_{i-1} + \Delta(z_i, x_i), \qquad t_0 \stackrel{\Delta}{=} 0 \\
k_i &= (u_{t_{i-1}+1}, u_{t_{i-1}+2}, \ldots, u_{t_i}) \\
y_i &= f(z_i, x_i, k_i) \\
z_{i+1} &= g(z_i, x_i)
\end{aligned}
$$

$$
\sigma_s(x_1, \ldots, x_n) = \min_{\{\text{all } s\text{-state encrypters}\}} \frac{1}{n} \sum_{i=1}^{n} \ell(k_i)
$$

.

Encryption (Cont'd)

It is shown that

$$\sigma_s(x_1, \ldots, x_n) \geq \frac{c \log c}{n} - \text{small terms},$$

which is achieved by an encrypter that:

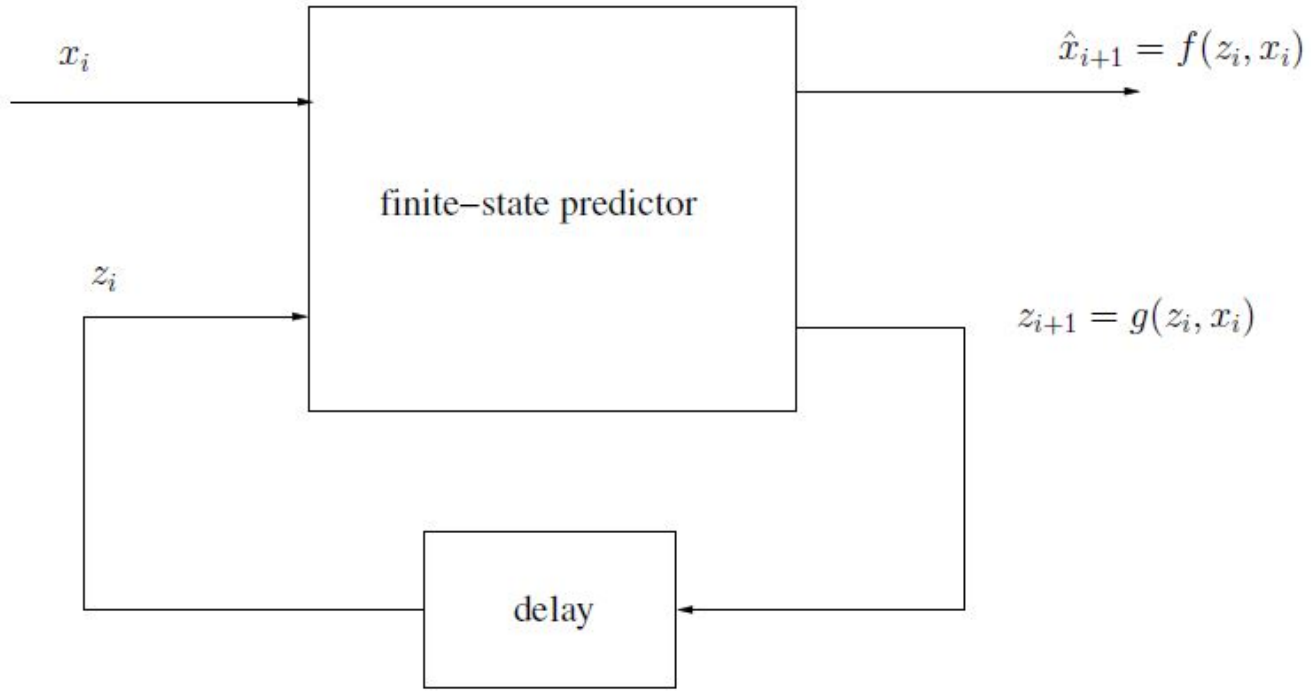
(i) applies LZ78 compression;
(ii) XORs every bit of the compressed representation with a key bit, u_i.

At the decoder:

(i) decrypt by XORing again with the corresponding key bit, u_i;
(ii) apply LZ78 decompression.

Analogous to well-known results in the probabilistic setting.

Universal Prediction (Feder, Merhav & Gutman, '92)



The s-state predictability of $(x_1, \ldots, x_n)$:

$$\pi_s(x_1, \ldots, x_n) = \min_{\{\text{all } s\text{-state predictors}\}} \frac{\text{number of prediction errors}}{n}.$$

There is no explicit expression (or tight bound) of π_s in terms of c.

Prediction Using LZ78

- LZ78 algorithm implies an efficient sequential probability assignment

- **Incremental parsing**: LZ78 parses the input sequence into *phrases*, where each new phrase is the shortest substring that has not appeared so far in the parsing
  - parsing is represented with a parsing tree
  - The tree implies a probability assignment

Universal Prediction (Cont'd)

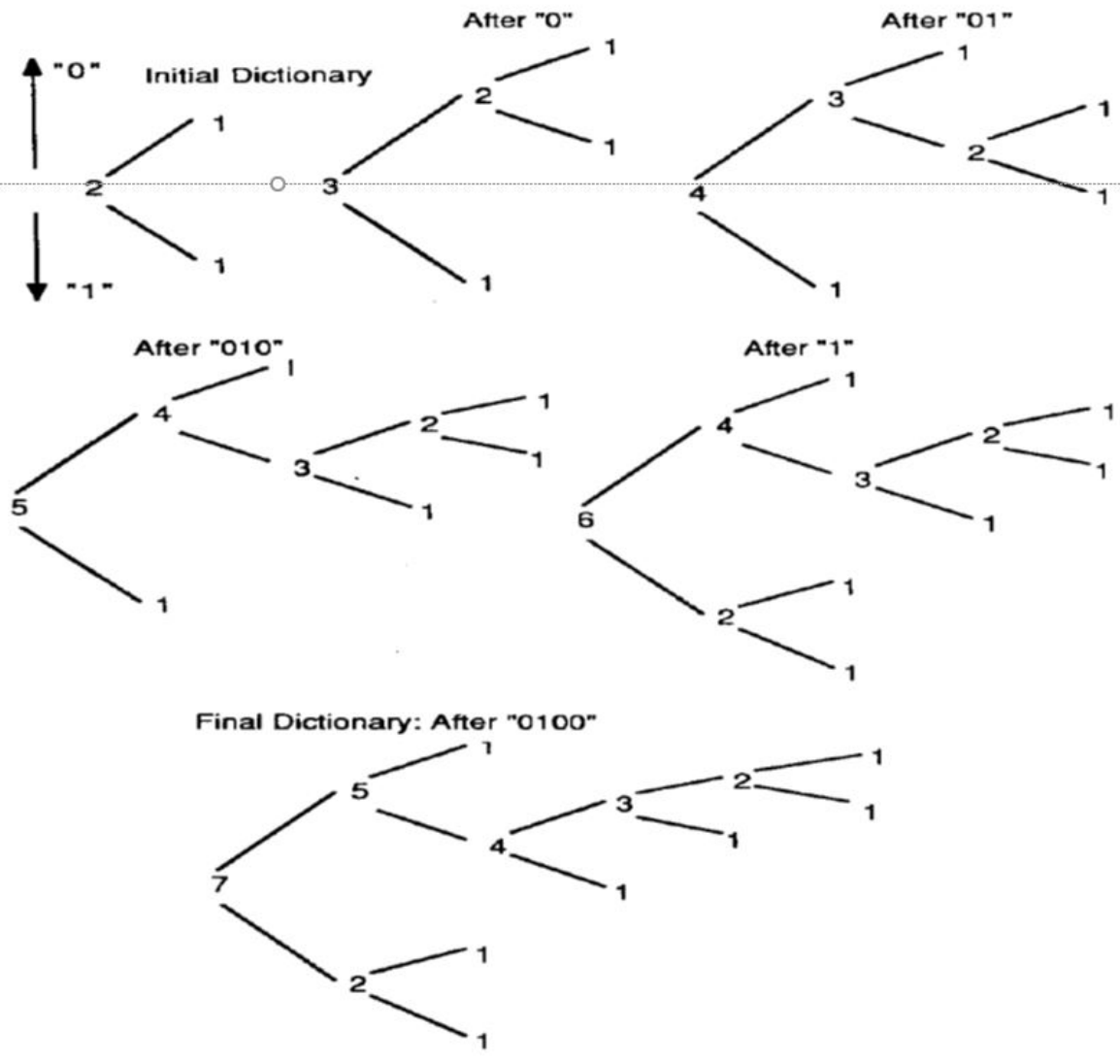
- Example: Feder-M-Gutman 92

x^{11}= 00101010100

=0,01,010,1,0100

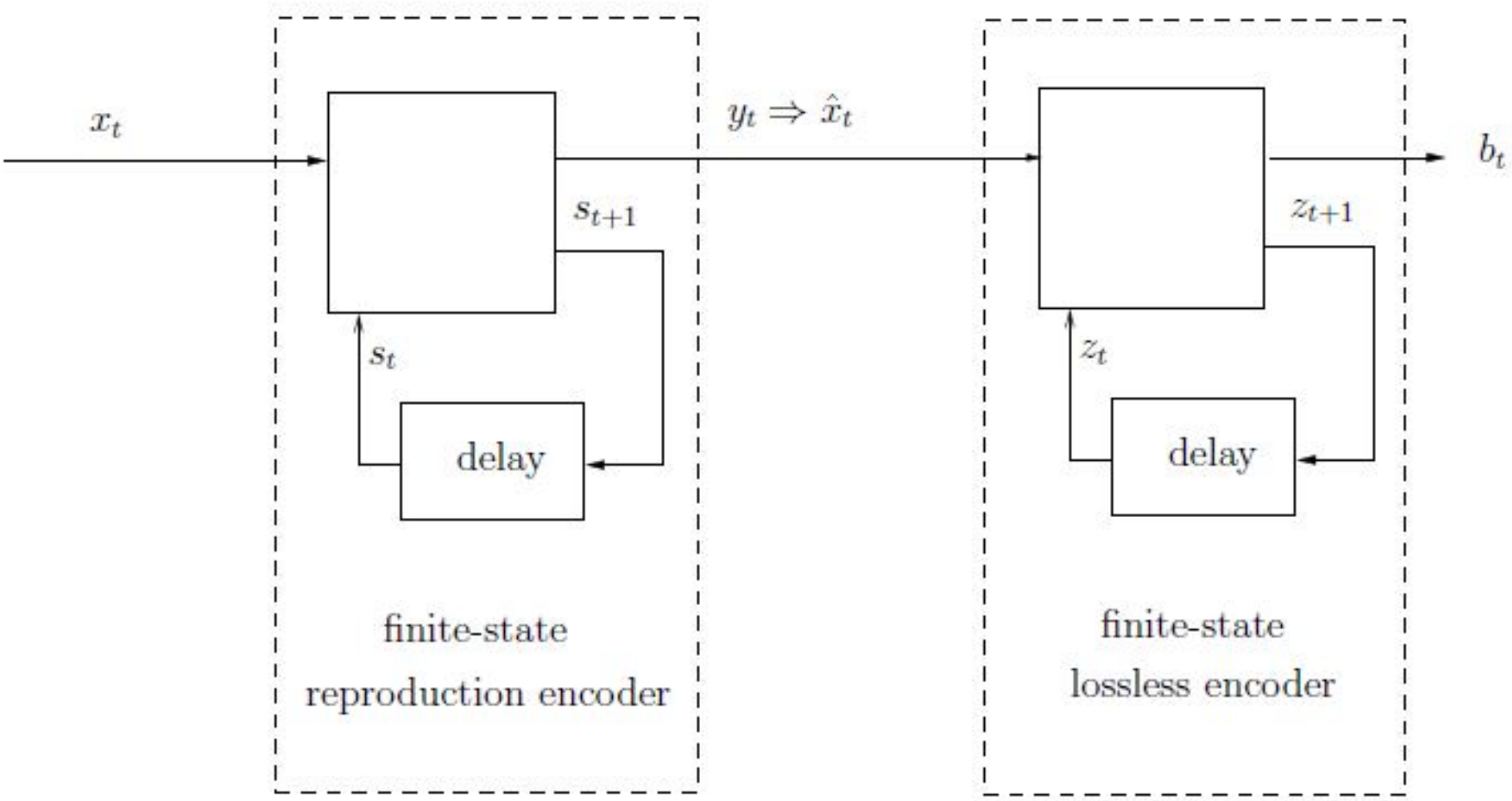
$P(x_t = 0 \mid x^{t-1}) =$

$\frac{1}{2}, \frac{2}{3}, \frac{1}{2}, \frac{3}{4}, \frac{1}{3}, \frac{1}{2}, \frac{4}{5}, \frac{2}{3}, \frac{1}{4}, \frac{2}{3}, \frac{1}{2}$



Closely related ideas appear already in Feder's 1991 paper on FS gambling.

Lossy Compression (Merhav '24)



Reproduction encoder: keeps with the distortion constraint for every k-block.
$\{y_t\}$ - variable-length strings (including $\emptyset$ of length 0) with total length $= k$.
Reproduction encoder - arbitrarily many states. Lossless encoder - s states.

Lossy Compression (Cont'd)

If $\log s \ll \log k$, the best we can do is to seek the reproduction vector with the smallest $c \log c$ within the 'sphere' in each k-block and compressing it by LZ78.

Otherwise, if s is sufficiently large in terms of k, we can do much better by generating a codebook at random using the universal probability distribution:

$$P(\hat{x}_1, \ldots, \hat{x}_n) \propto 2^{-c \log c}$$

and compressing the index of the first codeword that meets the distortion requirement.

The resulting code is universal, not only in terms of the source sequence, but also in terms of the distortion function.

The universal distribution is useful also in other tasks, such as guessing.

Summary

- We reviewed Jacob Ziv's individual-sequence approach.
- The jewel in the crown - the LZ algorithm:
  a successful marriage of a beautiful theory and great practicality.
- Ziv's inequality and its utility.
- Ziv's legacy has influenced my own research journey, as well as those of colleagues and students:
  - LZ at the service of many tasks beyond compression (we have seen just some).
  - Extensions to more general settings: lossy compression, side information, etc.
- Outlook: extensions to multiuser network configurations.