Département de Mathématiques - Laboratoire AGM
95302 Cergy-Pontoise Cedex - FRANCE
2, avenue Adolphe Chauvin

On the Ziv-Merhav estimator

Quantissima in the Serenissima V

August 13, 2024 (Venice)

Gaia Pozzoli
Project: *ENIC - Entropy, information and control of complex processes*

Joint work with:
N. Barnfield and R. Grondin (McGill University)
R. Raquépas (NY Courant Institute)

Entropies

(Shannon) entropy rate means:

$$h(\mathbb{P}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{a\in\mathcal{A}^n} \mathbb{P}[a] \log \mathbb{P}[a]$$

Cross entropy rate – when existing – means:

$$h^c(\mathbb{P}|\mathbb{Q}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{a\in\mathcal{A}^n} \mathbb{P}[a] \log \mathbb{Q}[a]$$

Relative entropy rate (or Kullback-Leibler divergence):

$$h^r(\mathbb{P}|\mathbb{Q}) = h^c(\mathbb{P}|\mathbb{Q}) - h(\mathbb{P})$$

Universal estimators: no complete knowledge of the measures!

We will always assume that $\mathbb{P}$ and $\mathbb{Q}$ are shift-invariant with respect to the left-shift operator $Tx_k := x_{k+1}$ with $x \in \mathcal{A}^{\mathbb{N}}$.

The waiting time problem

- Kac's Lemma (1947): given $R_n(x) := \inf\{\ell \geq 1 \,|\, x_{\ell+1}^{\ell+n} = x_1^n\}$, $\forall\, a \in \mathcal{A}^n$ we have

$$\log \mathbb{E}_a[R_n] := \log \int R_n(x) d\mathbb{P}(x \,|\, x_1^n = a) = -\log \mathbb{P}[a];$$

- Wyner-Ziv (1989) and Ornstein-Weiss (1993): if $\mathbb{P}$ is *ergodic*, then for $\mathbb{P}$-almost every x we have

$$\lim_{n\to\infty} \frac{\log R_n(x)}{n} = h(\mathbb{P}).$$

No strong assumptions on the measure:
access to the entropy just looking at the data!

Similarly, $W_n(x,y) := \inf_{\ell \geq 1}\{y_\ell^{\ell+n-1} = x_1^n\}$ for two sequences $x \sim \mathbb{P}$, $y \sim \mathbb{Q}$ and we expect

$$W_n(x,y) \sim e^{h^c(\mathbb{P}|\mathbb{Q})\cdot n}, \quad \text{where} \quad h^c(\mathbb{P}|\mathbb{Q}) = h(\mathbb{P}) + h^r(\mathbb{P}|\mathbb{Q})$$

for “sufficiently well-behaved” measures:

- **Kontoyiannis (1998):** $\mathbb{Q}$ is ψ-mixing or summably ϕ-mixing;
- **Cristadoro, Degli Esposti, Jakšić and Raquépas (2023):** decoupling perspective for $\mathbb{Q}$ and ergodicity for $\mathbb{P}$.

Computational drawback: given x, y *finite* sequences, even for small values of n, you may scroll all the data concerning x without having seen the prefix y_1^n ...
$\Longrightarrow$ Dual quantity: longest-match length! [Kontoyiannis (1998)]

$$\Lambda_n(x,y) := \sup_{\ell \geq 1}\{W_\ell(x,y) \leq n - \ell + 1\} \leq n$$

E.g. $x_1^{10} = \underbrace{101}\ 1100100$

$y_1^{\infty} = 01\ \underbrace{101}_{x_1^3}\ 00010\ |\ \underbrace{11120100}_{x_1^5}\ldots$

$$\lim_{n\to\infty} \frac{\log W_n(x,y)}{n} = \lim_{n\to\infty} \frac{\log \mathbf{n}}{\Lambda_n(x,y)} = h^c(\mathbb{P}|\mathbb{Q}) \quad \text{for } (\mathbb{P}\otimes\mathbb{Q})\text{-almost every } (x,y)$$

... BUT

Ziv-Merhav: a waste-free estimator!

Heuristics: Iteration!

Compute $\Lambda_n(x,y)$ at different (**dependent**) starting locations throughout x_1^n and take the average!

$$x = 0110010100010201110100100\,0210\ldots,$$

$$y = 010\underbrace{0010111}_{x_4^8}1010\;\underbrace{011}_{x_1^3}\;100100012\,2021\ldots,$$

$$x_1^{24} = 011|00101|00010|2|011101001|0 \quad \Longrightarrow \quad c_{24}^{ZM}(x|y) = 6$$

$c_n^{ZM}(x|y)$ is the number of words in the sequential parsing of x_1^n using longest-matches with y_1^n.

We expect: $\lim_{n\to\infty} Q_n^{ZM}(x,y) := \lim_{n\to\infty} \frac{c_n^{ZM}(x|y)\log n}{n} = h^c(\mathbb{P}|\mathbb{Q}) \quad (\mathbb{P}\otimes\mathbb{Q})$-a.s.

Why? Define $\ell_1 := \Lambda_n(x,y)$ and $\ell_j := \Lambda_n(T^{\ell_1+\cdots+\ell_{j-1}}x,y)$ with $\sum_{j=1}^{c_n^{ZM}} \ell_j = n$, then

$$\frac{c_n^{ZM}\log n}{n} = \frac{\log n}{\frac{1}{c_n^{ZM}}\sum_{j=0}^{c_n^{ZM}-1}\Lambda_n(T^{\ell_1+\cdots+\ell_{j-1}}x,y)} \overset{?}{\approx} \frac{\log n}{\Lambda_n(x,y)}$$

mZM estimator and decoupling perspective

Step I: do the best in the spirit of the original proof

Theorem (J. Ziv and N. Merhav, 1993)

Let $\mathbb{P}, \mathbb{Q}$ be two ergodic multi-level Markov measures. Then

$$\lim_{n\to\infty} Q_n^{ZM}(x,y) = h^c(\mathbb{P}|\mathbb{Q})$$

for $(\mathbb{P}\otimes\mathbb{Q})$-almost every $(x,y) \in \mathcal{A}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}}$.

Which properties of irreducible multi-level Markov chains do they really use?

"Strong" decoupling (for the auxiliary parsing)
[no gaps: $\tau = 0$ in **SLD**]

Step II: find new arguments which include a broader class of measures

$\implies$ mZM: we replace $\Lambda_n(x,y) \mapsto \Lambda_n(x,y) + 1$ and

$$Q_n(x,y) := \frac{c_n(x,y)\log n}{n - c_n(x,y)}$$

There exists $(k_n)_{n=1}^{\infty} \subset [0,\infty)$ and $\tau \in \mathbb{N}_0$, with $k_n = o(n)$, such that for all $n, m \in \mathbb{N}$, $a \in \mathcal{A}^n$, $b \in \mathcal{A}^m$

(UD) for all $\xi \in \mathcal{A}^\tau$

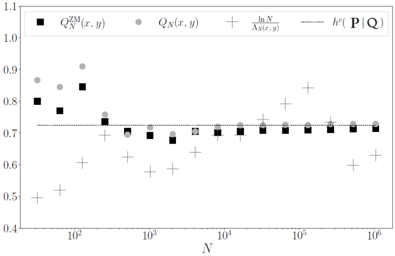
$$\mathbb{P}[a\,\xi\,b] \le e^{k_n}\mathbb{P}[a]\mathbb{P}[b];$$

(SLD) there exists $0 \le \ell(a,b) \le \tau$ and $\xi(a,b) \in \mathcal{A}^\ell$ such that

$$\mathbb{P}[a\,\xi\,b] \ge e^{-k_n}\mathbb{P}[a]\mathbb{P}[b].$$

Examples of decoupled measures:

- **Bernoulli measures;**
- **irreducible multi-level Markov chains;**
- equilibrium measures for 1D spin systems with summable interactions;
- g-measures;
- ψ-mixing measures;
- hidden-Markov measures with finite hidden alphabet;
- unravellings of irreducible repeated quantum measurements.



Theorems

Theorem (Cristadoro-Degli Esposti-Jakšić-Raquépas, 2023)

Let $\mathbb{Q}$ satisfy (UD) and (SLD), and $\mathbb{P}$ be ergodic. Then

$$\lim_{n\to\infty} \frac{\log W_n(x,y)}{n} = h^c(\mathbb{P}|\mathbb{Q}) \quad \textit{for } (\mathbb{P}\otimes\mathbb{Q})\textit{-almost every pair } (x,y) \textit{ of sequences.}$$

Theorem (Barnfield-Grondin-P.-Raquépas, 2023)

Suppose that $\mathbb{Q}$ satisfies (UD) and (SLD), and that $\mathbb{P}$ is ergodic and satisfies (UD). Then, under mild decay conditions,

$$\lim_{n\to\infty} Q_n(x,y) = h^c(\mathbb{P}|\mathbb{Q}) \quad \textit{for } (\mathbb{P}\otimes\mathbb{Q})\textit{-almost every } (x,y) \in \mathcal{A}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}}.$$

Recall that $Q_n(x,y)$ is analogous to $Q_n^{ZM}(x,y)$, but considering the *shortest* new word in x_1^n **not** found in y_1^n (as in the LZ77 algorithm!)

Some references

In the literature:

- Ziv and Lempel. *IEEE Trans. Inf. Theory* **23** (1977).
- Wyner and Ziv. *IEEE Trans. Inf. Theory* **35** (1989).
- Ziv and Merhav. *IEEE Trans. Inf. Theory* **39** (1993).
- Ornstein and Weiss. *IEEE Int. Symp. Inf. Theory* **39** (1993).
- Kontoyiannis. *J. Theor. Probab.* **11** (1998).
- Cristadoro, Degli Esposti, Jakšić and Raquépas. *Stoch. Proc. Appl.* **166** (2023).

Our results: Barnfield, Grondin, P. and Raquépas

- On the Ziv–Merhav theorem beyond Markovianity I. *CJM* **2024**, 1–25 (2024).
- On the Ziv–Merhav theorem beyond Markovianity II: leveraging the thermodynamic formalism (2023).
  arXiv:2312.02098
- Ziv–Merhav Estimation for Hidden-Markov Processes. *IEEE ISIT2024* – Athens.