Canad. J. Math. Vol. 77 (3), 2025 pp. 891–915
http://dx.doi.org/10.4153/S0008414X24000178

On the Ziv–Merhav theorem beyond Markovianity I

Nicholas Barnfield, Raphaël Grondin, Gaia Pozzoli, and Renaud Raquépas

Abstract. We generalize to a broader class of decoupled measures a result of Ziv and Merhav on universal estimation of the specific cross (or relative) entropy, originally for a pair of multilevel Markov measures. Our generalization focuses on abstract decoupling conditions and covers pairs of suitably regular g-measures and pairs of equilibrium measures arising from the "small space of interactions" in mathematical statistical mechanics.

1 Introduction

In 1993, Ziv and Merhav proposed a "new notion of empirical informational divergence," or relative-entropy estimator, based on the celebrated Lempel–Ziv compression algorithm [ZM93]. While this estimator received – to the best of our knowledge – little attention in the mathematical literature, it (and its variants) has met with success in many practical applications across fields such as linguistics, medicine, and physics (see, e.g., [BBCDE08, BCL02, CF05, CFF10, LMDEC19, RGS+22, RP12], to only cite a few). In fact, our main motivation for a more extensive rigorous treatment of the convergence of this estimator is that the very limited Markovian class of sources covered by the original result of Ziv and Merhav pales in comparison with the breadth of apparent applicability.

Ziv and Merhav's estimator is defined as follows. Given two strings x_1^N and y_1^N, let $c_N(y|x)$ be the number of words in a sequential parsing of y_1^N using the longest

Received by the editors October 2, 2023; revised January 27, 2024; accepted February 5, 2024.
Published online on Cambridge Core March 7, 2024.

The research of N.B. and R.R. was partially funded by the *Fonds de recherche du Québec – Nature et technologies* (FRQNT) and by the Natural Sciences and Engineering Research Council of Canada (NSERC). The research of R.G. was partially funded by the Rubin Gruber Science Undergraduate Research Award and Axel W Hundemer. The research of G.P. was supported by the CY Initiative of Excellence through the grant Investissements d'Avenir ANR-16-IDEX-0008, and was done under the auspices of the *Gruppo Nazionale di Fisica Matematica* (GNFM) section of the *Istituto Nazionale di Alta Matematica* (INdAM) while G.P. was a postdoctoral researcher at the University of Milano-Bicocca (Milan, Italy). Part of this work was done during a stay of the four authors in Neuville-sur-Oise, funded by CY Initiative (grant *Investissements d'avenir* ANR-16-IDEX-0008).

AMS subject classification: **94A17**, **37M25**, **37B10**.
Keywords: Cross entropy, estimator, parsing, match lengths, decoupling.

possible substrings of x_1^N; if there is no such substring of x_1^N, the parsed word is set to be one letter long. For example, if

$$x = 0100010111010011100100012202 1\ldots,$$
$$y = 0110010100010201111010010002 10\ldots,$$

and $N = 24$, then the Ziv–Merhav parsing of $y_1^{24} = 011001010001020111010010$ with respect to $x_1^{24} = 010001011101001110010001$ is

$$y_1^{24} = 011|00101|00010|2|011101001|0$$

and $c_{24}(y|x) = 6$.[1] Ziv and Merhav show that the estimator

$$\widehat{Q}_N(y,x) := \frac{c_N(y|x)\ln N}{N}$$

converges to the specific cross entropy $h_{\mathrm{c}}(\mathbb{Q}|\mathbb{P})$ – see (2.1) below – between the sources $\mathbb{P}$ and $\mathbb{Q}$ that have produced x and y, respectively, under the assumption that those measures come from irreducible multilevel Markov chains. We will refer to $(\widehat{Q}_N)_{N=1}^{\infty}$ as the ZM estimator. The relative entropy $h_{\mathrm{r}}(\mathbb{Q}|\mathbb{P})$ can then be estimated by combining the above with an estimation of the specific entropy $h(\mathbb{Q})$, say *à la* Lempel–Ziv [ZL78]. Both quantities are defined in Section 2 for the reader's convenience.

One may note that the behavior of c_N is intimately related to the so-called Wyner–Ziv problem on *waiting times*. With

$$W_\ell(y,x) := \inf\{r \in \mathbb{N} : x_r^{r+\ell-1} = y_1^\ell\},$$

the Wyner–Ziv problem concerns the convergence

$$\frac{\ln W_\ell}{\ell} \to h_{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) \tag{1.1}$$

as $\ell \to \infty$ within sufficiently nice classes of measures [Kon98, Shi93, WZ89]. To see the relation, note that the length of the first word in the ZM parsing of y_1^N with respect to x_1^N is – save some edge cases – the largest possible ℓ such that $W_\ell(y,x) \le N - \ell + 1$. This dual quantity is known as the *longest-match length*

$$\Lambda_N(y,x) := \max\{1, \sup\{\ell \in \mathbb{N} : W_\ell(y,x) \le N - \ell + 1\}\}.$$

The length of the second word in this parsing is then – again save some edge cases handled in Section 3.4 – the longest-match length $\Lambda_N(T^{\Lambda_N(y,x)}y, x)$, and so on. Any attempt at a theory of the asymptotic behavior of waiting times and its derived quantities beyond Markovianity must take two important caveats into account. First, it is known that the specific cross entropy between two ergodic sources does not always exist (see, e.g., [vEFS93, Section A.5.2]). Second, it is known that there exists a mixing measure $\mathbb{P}$ such that (1.1) fails with $\mathbb{Q} = \mathbb{P}$ (see [Shi93, Section 4]). While the precise breadth of the validity of (1.1) and its different refinements remains unknown, a focus on decoupling conditions in the spirit of [Pfi02] has recently proved effective

[1] Throughout this paper, we will refer to the partitioning symbol "|" as a separator, and we will say that a separator *falls within* a given string if the separator lies after one of the letters that make up the string.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

for making significant progress [CDEJR23a, CR23]; the present contribution follows along those lines.

Broadly speaking, the present work is part of a research program [BCJP21, BJPP18, CDEJR23a, CDEJR23b, CJPS19, CR23] whose goals include promoting the efficiency of this decoupling perspective originating in statistical mechanics in revisiting long-standing problems in dynamical systems and information theory. This efficiency concerns both the reformulation of different existing proof strategies in a common language and the generation of nontrivial extensions. In the case presently at hand, revisiting Ziv and Merhav's argument from this decoupling perspective allows us to replace the Markovianity assumption with a more permissive combination of three abstract assumptions, namely ID, KB, and FE below.

Organization of the paper.

The rest of the paper is organized as follows. In Section 2, we set the stage by properly introducing our notation, objects of interest, and assumptions. In Section 3, we state our main result, provide its proof, and make several comments. In Section 4, we discuss examples to which this result applies beyond Markovianity.

2 Setting

Let $\Omega := \{(x_k)_{k\in\mathbb{N}} : x_k \in \mathcal{A} \text{ for all } k \in \mathbb{N}\}$ be equipped with the σ-algebra generated by cylinders of the form $[a] := \{x \in \Omega : x_1^n = a\}$. The *shift* map $T : \Omega \to \Omega$ defined by $(Tx)_k := x_{k+1}$ is then a measurable surjection. Let $\mathbb{P}$ and $\mathbb{Q}$ be stationary (i.e., T-invariant) probability measures on Ω. We set

$$\operatorname{supp}\mathbb{P}_n := \{a \in \mathcal{A}^n : \mathbb{P}[a] > 0\}$$

and

$$\operatorname{supp}\mathbb{P} := \{x \in \Omega : x_1^n \in \operatorname{supp}\mathbb{P}_n \text{ for all } n \in \mathbb{N}\},$$

and similarly for $\mathbb{Q}$. We consider samples (x, y) from the product measure $\mathbb{P} \otimes \mathbb{Q}$ on $\Omega \times \Omega$, meaning that the two sources produce strings of symbols independently of each other. The (specific) *entropy* $h(\mathbb{P})$ of a measure $\mathbb{P}$ is

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

Fekete's lemma ensures that this limit always exists and lies in $[0, \ln(\#\mathcal{A})]$. The (specific) *cross entropy* of $\mathbb{Q}$ with respect to $\mathbb{P}$ is

$$h_{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) := \lim_{n\to\infty} -\frac{1}{n} \sum_{a\in\mathcal{A}^n} \mathbb{Q}[a] \ln \mathbb{P}[a], \tag{2.1}$$

when the limit exists in $[0, \infty]$. In this case, the (specific) *relative entropy* $h_{\mathrm{r}}(\mathbb{Q}|\mathbb{P})$ of $\mathbb{Q}$ with respect to $\mathbb{P}$ is then defined as

$$h_{\mathrm{r}}(\mathbb{Q}|\mathbb{P}) := h_{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) - h(\mathbb{Q}).$$

The abstract properties of stationary measures that we will work with are the following:

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

ID A measure $\mathbb{P}$ is said to be *immediately decoupled on its support* if there exists a nondecreasing,[2] nonnegative $o(n)$-sequence $(k_n)_{n=1}^{\infty}$ such that, for every $n \in \mathbb{N}$, both

$$\sup_{m\in\mathbb{N}} \max\left\{\frac{\mathbb{P}[ab]}{\mathbb{P}[a]\mathbb{P}[b]} : a \in \operatorname{supp}\mathbb{P}_n, b \in \operatorname{supp}\mathbb{P}_m\right\} \le \mathrm{e}^{k_n} \tag{2.2}$$

and

$$\inf_{m\in\mathbb{N}} \min\left\{\frac{\mathbb{P}[ab]}{\mathbb{P}[a]\mathbb{P}[b]} : a \in \operatorname{supp}\mathbb{P}_n, b \in \operatorname{supp}\mathbb{P}_m, ab \in \operatorname{supp}\mathbb{P}_{n+m}\right\} \ge \mathrm{e}^{-k_n}. \tag{2.3}$$

FE The $\mathbb{P}$-measure of cylinders is said to decay *fast enough* if there exists $\gamma_+ < 0$ such that

$$\sup_{a\in\operatorname{supp}\mathbb{P}_n} \mathbb{P}[a] \le \mathrm{e}^{\gamma_+ n} \tag{2.4}$$

for all $n \in \mathbb{N}$.

KB The measure $\mathbb{P}$ is said to satisfy *Kontoyiannis' bound on waiting times* if there exist nonnegative $o(n)$-sequences $(k_n)_{n=1}^{\infty}$ and $(\tau_n)_{n=1}^{\infty}$ such that

$$\mathbb{P}\{x : W_\ell(a,x) \ge r\} \le \exp\left(-\mathrm{e}^{-k_\ell}\mathbb{P}[a]\left\lfloor\frac{r-1}{\ell+\tau_\ell}\right\rfloor\right)$$

for every $\ell \in \mathbb{N}$, $a \in \mathcal{A}^\ell$, and $r \in \mathbb{N}$.

Let us briefly discuss these abstract assumptions. First, it is straightforward to show that if $\mathbb{P}$ is the stationary measure for an irreducible multilevel Markov chain with positive entropy, then $\mathbb{P}$ satisfies ID, FE, and KB. Already for Markov chains, we see that only requiring the lower-decoupling bound (2.3) when ab is in the support is significant: requiring the lower bound whenever a and b are in the support (separately) would be considerably more restrictive, as this would exclude all Markov measures for which some transition probability is null. Second, the bound KB was derived in [Kon98] under a ψ-mixing assumption, but the following implication seems more natural for the classes of examples we have in mind: KB will follow from the decoupling condition ID if one is willing to assume that the support of $\mathbb{P}$ satisfies – as a subshift of Ω – a suitable notion of specification.[3] Still regarding ψ-mixing, in the notation of [Bra05, Section 2], Condition ID is implied by $0 < \psi'(0) \le \psi^*(0) < \infty$, but ψ-mixing itself does not imply ID and ID does not imply ψ-mixing – or any form of mixing for that matter. Third, combining ID and FE yields Shields's finite-energy condition [Shi96, Section II.5.a], but the converse implication is not true. It is also worth mentioning the following relations to the Doeblin-type condition discussed, e.g., in [KS94]: it implies FE but not ID, and it is not implied by our assumptions; we will come back to this point in Section 4. Finally, repeated uses of (2.3) in ID implies the following property, which naturally complements FE:

[2] The requirement that $(k_n)_{n=1}^{\infty}$ be nondecreasing presents no loss of generality because one can always set $k'_n := \max\{k_1, k_2, \dots, k_n\}$ and preserve the other desired properties.

[3] For example, in the notation of [KŁO16, Section 8], Property (8) suffices. We refer the reader to [CR23] for more optimal specification properties.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

SE The $\mathbb{P}$-measure of cylinders is said to decay *slow enough* if there exists[4] $\gamma_- < 0$ such that

$$\inf_{b \in \operatorname{supp} \mathbb{P}_n} \mathbb{P}[b] \geq \mathrm{e}^{\gamma_- n}$$

for all $n \in \mathbb{N}$.

These assumptions are established and discussed in the context of important classes of examples in Section 4.

3 Main result

3.1 Statement and structure of the proof

Theorem 3.1 *Suppose that the stationary measure $\mathbb{P}$ satisfies ID, FE, and KB and that the ergodic measure $\mathbb{Q}$ satisfies ID and FE.*[5] *Then,*

$$\lim_{N \to \infty} \widehat{Q}_N(y, x) = h_c(\mathbb{Q}|\mathbb{P})$$

for $(\mathbb{P} \otimes \mathbb{Q})$-almost every (x, y).

Let us now provide the structure of the proof in the case where $\operatorname{supp} \mathbb{Q} \subseteq \operatorname{supp} \mathbb{P}$, postponing the more technical aspects to Sections 3.2 and 3.3. Throughout,

$$\ell_{-,N} := \frac{\ln N}{-2\gamma_-}$$

and

$$\ell_{+,N} := \frac{2 \ln N}{-\gamma_+}.$$

These will serve as *a priori* bounds on the lengths of the words in different auxiliary parsings.

Upper bound. Let $\varepsilon \in (0, \frac{1}{2})$ and $y \in \operatorname{supp} \mathbb{Q}$ be arbitrary. We consider an auxiliary sequential parsing $y_1^N = y^{(1,N)} y^{(2,N)} \dots y^{(\widehat{c}_N, N)}$, where each word $y^{(j,N)}$ has length $\ell_{j,N}$ and is – except possibly for $y^{(\widehat{c}_N, N)}$ – the shortest prefix of $T^{\ell_{1,N} + \dots + \ell_{j-1,N}} y_1^N$ satisfying

$$\mathbb{P}[y^{(j,N)}] \leq N^{-1+\varepsilon}, \tag{3.1}$$

where we define $\ell_{0,N} := 0$ and $T^k y_1^N := (T^k y)_1^{N-k}$ for any $k < N$. The power is chosen in the hope that the words in this auxiliary parsing will be long enough, yet likely enough for $\mathbb{P}$ that the vast majority of them find a match in x_1^N. To motivate this Ansatz, note that, by linearity of expectation, the expected number of times a

[4] In fact, by (2.3), we can set $\gamma_- := \ln(\min_{b \in \operatorname{supp} \mathbb{P}_1} \mathbb{P}[b]) - k_1$.

[5] We are using "ergodic" in the sense of dynamical systems, meaning that all shift-invariant subsets of $\mathcal{A}^{\mathbb{Z}}$ either have measure 0 or 1 according to $\mathbb{Q}$, so the following common caveat is in order: $\mathbb{Q}$ could come from a Markov chain and be ergodic in this sense even though it is periodic, which is at odds with a terminology sometimes used in the literature on Markov chains. As far as the decoupling of $\mathbb{Q}$ is concerned, we in fact only use (2.2), and not (2.3).

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

given string of $\mathbb{P}$-probability $N^{-1+\varepsilon}$ appears in a string of length N obtained from $\mathbb{P}$ grows as N^{ε}.

For N large enough, each length $\ell_{j,N}$ is between $\ell_{-,N}$ and $\ell_{+,N}$, due to Properties FE and SE, except possibly for $\ell_{\widehat{c}_N,N}$ which need only satisfy the upper bound. In particular, $\widehat{c}_N = O(\frac{N}{\ln N})$.

Note that, for each $j = 1, 2, \dots, \widehat{c}_N$, the appearance of $y^{(j,N)}$ as a substring of x_1^N – written $y^{(j,N)} \in x_1^N$ in what follows – implies the presence of at most one separator of the original ZM parsing within $y^{(j,N)}$, that is,

$$\mathbb{P}\left\{x : \#\{j \le \widehat{c}_N : y^{(j,N)} \in x_1^N\} = \widehat{c}_N\right\} \le \mathbb{P}\{x \,:\, c_N(y|x) \le \widehat{c}_N\},$$

which in turn implies

$$\mathbb{P}\{x \,:\, c_N(y|x) > \widehat{c}_N\} \le \mathbb{P}\left\{x : \#\{j \le \widehat{c}_N : y^{(j,N)} \notin x_1^N\} > 0\right\}.$$

We show in Lemma 3.3 that the probability on the right-hand side is summable in N and hence

$$\sum_{N=1}^{\infty} \mathbb{P}\left\{x : c_N(y|x) > \widehat{c}_N\right\} < \infty. \tag{3.2}$$

On the other hand, Lemma 3.10 below shows that

$$\begin{aligned}(-1+\varepsilon)(\widehat{c}_N - 1)\ln N &= \sum_{j=1}^{\widehat{c}_N - 1} \ln N^{-1+\varepsilon} \\ &\ge \sum_{j=1}^{\widehat{c}_N} \ln \mathbb{P}[y^{(j,N)}] \\ &\ge \ln \mathbb{P}[y_1^N] - o(N).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\left\{x \,:\, c_N(y|x)\ln N + \ln \mathbb{P}[y_1^N] > -\frac{\varepsilon}{1-\varepsilon}\ln \mathbb{P}[y_1^N] + \ln N + \varepsilon N\right\} \\ \le \mathbb{P}\{x \,:\, c_N(y|x)\ln N > \widehat{c}_N \ln N\}\end{aligned}$$

for all N large enough. Recall that, by Condition SE, $\ln \mathbb{P}[y_1^N] \ge \gamma_- N$ with $\gamma_- < 0$. Combining this with (3.2), we obtain

$$\sum_{N=1}^{\infty} \mathbb{P}\left\{x \,:\, \frac{c_N(y|x)\ln N}{N} + \frac{\ln \mathbb{P}[y_1^N]}{N} > -\frac{\varepsilon}{1-\varepsilon}\gamma_- + 2\varepsilon\right\} < \infty.$$

Appealing to the Borel–Cantelli lemma, using the cross entropy analogue of the Shannon–McMillan–Breiman theorem in Lemma 3.12, and then taking $\varepsilon \to 0$, we conclude that, for every $y \in \operatorname{supp} \mathbb{Q}$, we have

$$\limsup_{N\to\infty} \widehat{Q}_N(y, x) \le h^{\mathrm{c}}(\mathbb{Q}|\mathbb{P})$$

for almost every x sampled from $\mathbb{P}$.

Lower bound I. Before we obtain the almost sure lower bound required for Theorem 3.1, let us summarize Ziv and Merhav's argument for proving that the lower

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

bound holds in probability. This argument will also help the reader understand the more complicated construction used for the almost sure version. Let $\varepsilon \in (0, \frac{1}{2})$ and $y \in \operatorname{supp} \mathbb{Q}$ be arbitrary. We consider an analogous auxiliary sequential parsing $y_1^N = y^{(1,N)} y^{(2,N)} \dots y^{(\bar{c}_N,N)}$, where each word $y^{(j,N)}$ has length $\ell_{j,N}$ and is – except possibly for $y^{(\bar{c}_N,N)}$ – the shortest prefix of $T^{\ell_{1,N}+\dots+\ell_{j-1,N}} y_1^N$ that has probability

$$\mathbb{P}[y^{(j,N)}] \le N^{-1-\varepsilon}, \tag{3.3}$$

where we define $\ell_{0,N} := 0$. The power is chosen in the hope that the words in this auxiliary parsing will be numerous enough, yet unlikely enough for $\mathbb{P}$ that the vast majority of them find no match in x_1^N. To motivate this Ansatz, note that the expected number of times a given string of $\mathbb{P}$-probability $N^{-1-\varepsilon}$ appears in a string of length N obtained from $\mathbb{P}$ decays as $N^{-\varepsilon}$.

Again, for N large enough, each length $\ell_{j,N}$ in this parsing falls between $\ell_{-,N}$ and $\ell_{+,N}$, due to Properties FE and SE, except possibly for the last one, which only satisfies the upper bound. In particular, $\bar{c}_N = O(\frac{N}{\ln N})$.

The correspondence between the parsing cardinalities $c_N(y|x)$ and $\bar{c}_N$ relies on the following observation: $c_N(y|x)$ must be at least equal to the number of words in the auxiliary parsing of y_1^N that *do not* appear as strings in x_1^N. Indeed, if a word $y^{(j,N)}$ does not appear as a substring of x_1^N – written $y^{(j,N)} \notin x_1^N$ in what follows – then the ZM parsing has at least one separator within $y^{(j,N)}$. That is,

$$\begin{aligned} c_N(y|x) &\ge \#\{j : y^{(j,N)} \notin x_1^N\} \\ &\ge \#\{j \le \bar{c}_N - 1 : y^{(j,N)} \notin x_1^N\} \end{aligned}$$

and so

$$\begin{aligned} &\mathbb{P}\{x : c_N(y|x) \ge (\bar{c}_N - 1)(1-\varepsilon)\} \\ &\quad \ge \mathbb{P}\left\{x : \#\{j \le \bar{c}_N - 1 : y^{(j,N)} \in x_1^N\} \le (\bar{c}_N - 1)\varepsilon\right\} \\ &\quad \ge 1 - \mathbb{P}\left\{x : \#\{j \le \bar{c}_N - 1 : y^{(j,N)} \in x_1^N\} > (\bar{c}_N - 1)\,\varepsilon\right\}. \end{aligned}$$

One can easily show using a crude union bound and Markov's inequality that the appearance in x_1^N of more than an arbitrarily small proportion of all the words in the auxiliary parsing except for the last one has vanishing – but not necessarily summable – probability, and this enables us to conclude that

$$\lim_{N\to\infty} \mathbb{P}\{x : c_N(y|x) \ge (\bar{c}_N - 1)(1-\varepsilon)\} = 1. \tag{3.4}$$

Note that since, by construction, for any $j = 1, 2, \dots, \bar{c}_N$, the auxiliary word $y^{(j,N)}$ has *no* strict prefix with probability less than $N^{-1-\varepsilon}$, the lower bound in Condition ID implies that $\mathbb{P}[y^{(j,N)}] \ge N^{-1-2\varepsilon}$ for N large enough. Therefore, Lemma 3.10 yields

$$\begin{aligned} (-1-2\varepsilon)\bar{c}_N \ln N - o(N) &\le \sum_{j=1}^{\bar{c}_N} \ln \mathbb{P}[y^{(j,N)}] \\ &\le \ln \mathbb{P}[y_1^N] + o(N), \end{aligned} \tag{3.5}$$

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

which together with (3.4) implies

$$\lim_{N\to\infty} \mathbb{P}\left\{x : \frac{c_N(y|x)\ln N}{N} + \frac{\ln \mathbb{P}[y_1^N]}{N} \geq -3\varepsilon\frac{\bar{c}_N \ln N}{N} - \varepsilon\right\} = 1,$$

Thus, using Lemma 3.12, the fact that $\bar{c}_N = O(\frac{N}{\ln N})$ and taking $\varepsilon \to 0$, we conclude that, for all $y \in \operatorname{supp} \mathbb{Q}$, we have

$$h^{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) \leq \liminf_{N\to\infty} \widehat{Q}_N(y,x) \tag{3.6}$$

in probability with respect to x sampled from $\mathbb{P}$.

Lower bound II. Let $\varepsilon \in (0, \frac{1}{2})$ and $y \in \operatorname{supp} \mathbb{Q}$ be arbitrary, and fix $0 < \alpha < \frac{\gamma_+}{8\gamma_-} < 1$. In what follows and in the last part of Section 3.2, the number N^α is to be understood as its integer part $\lfloor N^\alpha \rfloor$. Following Ziv and Merhav's original strategy for strengthening convergence in probability to almost sure convergence, we modify the auxiliary parsing of y_1^N in "Lower Bound I" by applying the same algorithm separately to subsequent *blocks* of N^α symbols of y_1^N.

First, let $y^{(1,1,N)}$ be the shortest prefix of $y_1^{N^\alpha}$ such that $\mathbb{P}[y^{(1,1,N)}] \leq N^{-1-\varepsilon}$; it has length $\ell_{1,1,N}$, between $\ell_{-,N}$ and $\ell_{+,N}$ for N large enough due to Properties FE and SE. Now, let $y^{(2,1,N)}$ be the shortest prefix of $y_{\ell_{1,1,N}+1}^{N^\alpha}$ such that $\mathbb{P}[y^{(2,1,N)}] \leq N^{-1-\varepsilon}$, and so on until not possible. We have parsed a first *block* of size N^α:

$$y_1^{N^\alpha} = y^{(1,1,N)} y^{(2,1,N)} \cdots y^{(d_{1,N},1,N)} \xi^{(1,N)},$$

where the (possibly empty) *buffer* $\xi^{(1,N)}$ has probability at least $N^{-1-\varepsilon}$ and length at most $\ell_{+,N}$ due to Property FE.

We then repeat the procedure with $T^{N^\alpha} y_1^N$ to obtain the second block, and so on until

$$\begin{aligned} y_1^N = y^{(1,1,N)} y^{(2,1,N)} \cdots y^{(d_{1,N},1,N)} \xi^{(1,N)} y^{(1,2,N)} y^{(2,2,N)} \cdots y^{(d_{2,N},2,N)} \xi^{(2,N)} \\ \cdots y^{(1,M_N,N)} y^{(2,M_N,N)} \cdots y^{(d_{M_N,N},M_N,N)} \xi^{(M_N,N)}. \end{aligned} \tag{3.7}$$

The construction of $y^{(1,M_N,N)} y^{(2,M_N,N)} \cdots y^{(d_{M_N,N},M_N,N)} \xi^{(M_N,N)}$ may differ from that of the previous $y^{(1,s,N)} y^{(2,s,N)} \cdots y^{(d_{s,N},s,N)} \xi^{(s,N)}$ for $s < M_N$ in that it might be the parsing of a block of a length smaller than N^α if there is a remainder in the division of N by N^α. Note that, for N large enough, $N^{1-\alpha} \leq M_N \leq 2N^{1-\alpha}$ and $d_{s,N} \leq \frac{2N^\alpha}{\ell_{-,N}} =: d_{+,N}$. The number of auxiliary parsed words to be considered is

$$\tilde{c}_N := d_{1,N} + d_{2,N} + \cdots + d_{M_N,N}.$$

It follows from the above that $\tilde{c}_N = O(\frac{N}{\ln N})$, since $d_{s,N} \geq \frac{N^\alpha}{2\ell_{+,N}} =: d_{-,N}$ for any $s < M_N$. As explained in "Lower bound I," $c_N(y|x)$ must be at least equal to the number of words in the auxiliary parsing of y_1^N that *do not* appear as strings in x_1^N, that is,

$$c_N(y|x) \geq \sum_{s=1}^{M_N} \#\left\{j : y^{(j,s,N)} \notin x_1^N\right\}.$$

In order to control the latter, we prove below the two following technical estimates:

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

- Proposition 3.6: For almost every y sampled from $\mathbb{Q}$, there exists $N_\varepsilon(y)$ such that, for $N \geq N_\varepsilon(y)$, the number of indices s such that the words $y^{(1,s,N)}$, $y^{(2,s,N)}, \ldots, y^{(d_{s,N},s,N)}$ are *not* distinct is smaller than εM_N; we comment on the reason for this technical consideration in Remark 3.8.
- Proposition 3.9: Denoting by $\mathcal{S}_{\mathrm{g}}(y_1^N)$ the set of indices s whose block of y_1^N *does* consist of distinct words, we have

$$\mathbb{P}\{\#\{j : y^{(j,s,N)} \in x_1^N\} > \varepsilon d_{+,N}\} \leq \ell_{+,N}{}^2 \mathrm{e}^{\frac{\varepsilon^2 \gamma_+}{8} \frac{d_{+,N}}{\ell_{+,N}}}$$

for N large enough and all $s \in \mathcal{S}_{\mathrm{g}}(y_1^N)$. This means that, with high probability, only a small fraction of the words in these "good blocks" can appear in x_1^N (and fail to contribute to $c_N(y|x)$).

Therefore, even considering the worst-case scenario where all $y^{(i,s,N)}$ with $s \notin \mathcal{S}_{\mathrm{g}}(y_1^N)$ do appear in x_1^N, we find that, for almost every y sampled from $\mathbb{Q}$,

$$\begin{aligned}
&\sum_{N=N_\varepsilon(y)}^{\infty} \mathbb{P}\left\{x : c_N(y|x) < \tilde{c}_N - 2\varepsilon d_{+,N} M_N\right\} \\
&\quad \leq \sum_{N=N_\varepsilon(y)}^{\infty} M_N \max_{s \in \mathcal{S}_{\mathrm{g}}(y_1^N)} \mathbb{P}\{\#\{j : y^{(j,s,N)} \in x_1^N\} > \varepsilon d_{+,N}\} < \infty.
\end{aligned}$$

In fact,

$$\begin{aligned}
&\left\{x : \sum_{s=1}^{M_N} \#\left\{j : y^{(j,s,N)} \notin x_1^N\right\} < \tilde{c}_N - 2\varepsilon d_{+,N} M_N\right\} \\
&\qquad = \left\{x : \sum_{s=1}^{M_N} \#\left\{j : y^{(j,s,N)} \in x_1^N\right\} > 2\varepsilon d_{+,N} M_N\right\} \\
&\qquad \subseteq \left\{x : \sum_{s \in \mathcal{S}_{\mathrm{g}}(y_1^N)} \#\left\{j : y^{(j,s,N)} \in x_1^N\right\} > \varepsilon d_{+,N} M_N\right\},
\end{aligned}$$

and we can perform a union bound after observing that for the number of words in the auxiliary parsing of y_1^N that appear in x_1^N and belong to "good blocks" to exceed $\varepsilon d_{+,N} M_N$, at least the number of such words in one of the "good blocks" must exceed $\varepsilon d_{+,N}$.

Appealing to Lemma 3.10 and Remark 3.11, the relation (3.5) between $\tilde{c}_N$ and $\ln \mathbb{P}[y_1^N]$ remains valid and yields

$$\begin{aligned}
&\mathbb{P}\left\{x : \frac{c_N(y|x)\ln N}{N} + \frac{\ln \mathbb{P}[y_1^N]}{N} < -2\varepsilon \frac{\tilde{c}_N \ln N}{N} - \varepsilon - 8\varepsilon \frac{\ln N}{\ell_{-,N}}\right\} \\
&\qquad \leq \mathbb{P}\left\{x : \frac{c_N(y|x)\ln N}{N} < \frac{\tilde{c}_N \ln N}{N} - 2\varepsilon d_{+,N} M_N \frac{\ln N}{N}\right\}
\end{aligned}$$

for N large enough, which implies that there exists some constant $C = C(\gamma_-) > 0$ such that

$$\sum_{N=1}^{\infty} \mathbb{P}\left\{x : \frac{c_N(y|x)\ln N}{N} + \frac{\ln \mathbb{P}[y_1^N]}{N} < -C\varepsilon\right\} < \infty.$$

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

By Lemma 3.12 and the Borel–Cantelli lemma, taking $\varepsilon \to 0$, we conclude that

$$h^{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) \le \limsup_{N\to\infty} \widehat{Q}_N(y,x)$$

for almost every (x,y) sampled from $\mathbb{P}\otimes\mathbb{Q}$.

The above strategy is essentially that of Ziv and Merhav, but the lemmas and propositions on which it relies need to be adapted beyond Markovianity. Before we do so, let us state and prove a proposition that justifies our focus on situations where $\operatorname{supp}\mathbb{Q} \subseteq \operatorname{supp}\mathbb{P}$.

Proposition 3.2 *Suppose that $\mathbb{Q}$ is ergodic. If there exists $k\in\mathbb{N}$ such that $\operatorname{supp}\mathbb{Q}_k \cap \operatorname{supp}\mathbb{P}_k^{\mathrm{c}} \neq \varnothing$, then $\widehat{Q}_N \to \infty$ almost surely as $N\to\infty$, in agreement with Theorem 3.1.*

Proof Fix k as in the hypothesis and then $a \in \operatorname{supp}\mathbb{Q}_k \setminus \operatorname{supp}\mathbb{P}_k$. Because $a \notin \operatorname{supp}\mathbb{P}_k$, a crude counting argument yields that the ZM parsing satisfies

$$c_N(y|x) \ge \frac{\#\{j \le N-k+1 : T^{j-1}y \in [a]\}}{k}$$

for all $x \in \operatorname{supp}\mathbb{P}$. Because $a \in \operatorname{supp}\mathbb{Q}_k$ and k is fixed, Birkhoff's ergodic theorem applied to the function $\mathbf{1}_{[a]}$ yields

$$\liminf_{N\to\infty} \frac{c_N(y|x)}{N} > 0,$$

for almost every $y \sim \mathbb{Q}$. This allows us to conclude that, almost surely, the estimator diverges.

As for the claim that this is in agreement with Theorem 3.1, it is based on the observation that if $\operatorname{supp}\mathbb{Q}_k \cap \operatorname{supp}\mathbb{P}_k^{\mathrm{c}} \neq \varnothing$, then $\operatorname{supp}\mathbb{Q}_n \cap \operatorname{supp}\mathbb{P}_n^{\mathrm{c}} \neq \varnothing$ for all $n \ge k$. Since the existence of $a \in \operatorname{supp}\mathbb{Q}_n$ such that $\mathbb{P}_n[a] = 0$ causes at least one summand to be infinite on the right-hand side of (2.1), this allows us to conclude that $h_{\mathrm{c}}(\mathbb{Q}|\mathbb{P}) = \infty$. ∎

3.2 Properties of the auxiliary parsings

Throughout this section, $\varepsilon \in (0,1/2)$ is fixed but arbitrary. We assume that $\mathbb{P}$ and $\mathbb{Q}$ are stationary and satisfy $\operatorname{supp}\mathbb{Q} \subseteq \operatorname{supp}\mathbb{P}$. For readability, we will omit keeping track of the N-dependence in some of the notation introduced above. As foreshadowed in the introduction, our analysis of the cardinalities of the auxiliary parsings will use reformulations in terms of waiting times.

Lemma 3.3 *Suppose that $\mathbb{P}$ satisfies ID, FE, and KB. Let $y \in \operatorname{supp}\mathbb{Q}$ be arbitrary and consider the auxiliary parsing of y_1^N built around the requirement (3.1). Then,*

$$\mathbb{P}\left\{x : \#\left\{j \le \widehat{c}_N : y^{(j)} \notin x_1^N\right\} > 0\right\} \le N\mathrm{e}^{-\frac{N^{\frac{\varepsilon}{4}}}{3\ell_+}}$$

for N large enough.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

Proof Let $\underline{y}^{(j)}$ be the word that is obtained by removing the last letter from $y^{(j)}$; by construction, $\mathbb{P}[\underline{y}^{(j)}] > N^{-1+\varepsilon}$. So, in view of ID,

$$\begin{aligned}\mathbb{P}[y^{(j)}] &\geq \mathrm{e}^{-k_{\ell_{j-1}}}\mathbb{P}[\underline{y}^{(j)}]\left(\min_{a\in\operatorname{supp}\mathbb{P}_1}\mathbb{P}[a]\right)\\ &\geq N^{-1+\frac{\varepsilon}{2}}\end{aligned} \tag{3.8}$$

for N large enough. We have used the fact that $k_{\ell_{j-1}} \leq k_{\ell_+}$ with $k_\ell = o(\ell)$ and $\ell_+ = O(\ln N)$. Using KB and considering all N large enough, we have

$$\mathbb{P}\{x : W_{\ell_j}(y^{(j)}, x) > N - \ell_j + 1\} \leq \exp\left(-\frac{N^{\frac{\varepsilon}{4}}}{2\ell_+ + \tau_{\ell_+}}\right), \tag{3.9}$$

where we used the defining properties of k_{ℓ_j} and ℓ_j. Then, using that for N large enough we have $\tau_{\ell_+} \leq \ell_+$ and taking a union bound over j,

$$\mathbb{P}\left(\bigcup_{j=1}^{\widehat{c}_N}\{x : W_{\ell_j}(y^{(j)}, x) > N - \ell_j + 1\}\right) \leq N\exp\left(-\frac{N^{\frac{\varepsilon}{4}}}{3\ell_+}\right).$$

To conclude, note that $W_{\ell_j}(y^{(j)}, x) > N - \ell_j + 1$ is a necessary and sufficient condition for $y^{(j)} \notin x_1^N$. ∎

While, on the one hand, the last lemma states that the words in the auxiliary parsing built around (3.1) tend to appear in x_1^N, one can show that, on the other hand, the words in the auxiliary parsing built around (3.3) tend to *not* appear in x_1^N. However, the probabilistic estimate obtained pursuing this strategy only achieves convergence in probability of the ZM estimator; see "Lower Bound I." As Ziv and Merhav showed in their original paper in the Markovian case, this estimate can actually be refined and made summable in N using some additional combinatorial and probabilistic arguments. Such a refinement is used to go from convergence in probability to almost sure convergence in Section 3.1. We recall the following basic facts about our modified auxiliary parsing (3.7) for N large enough:

- there are $M_N \leq 2N^{1-\alpha}$ blocks, indexed by s, each of length N^α except for the last one ($s = M_N$) which possibly has length less than N^α;
- the sth block contains d_s words $y^{(i,s)}$ with

$$d_- := \frac{N^\alpha}{2\ell_+} \leq d_s \leq \frac{2N^\alpha}{\ell_-} =: d_+,$$

  except for the last one ($s = M_N$) for which the lower bound may not apply, and one (possibly empty) buffer $\xi^{(s)}$;
- each word $y^{(i,s)}$ has length $\ell_{i,s}$, with

$$\ell_- := \frac{\ln N}{-2\gamma_-} \leq \ell_{i,s} \leq \frac{2\ln N}{-\gamma_+} =: \ell_+.$$

Most of the factors of 2 in these facts are suboptimal; they are only meant to avoid having to consider integer parts or superficial dependence on ε.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

Definition 3.1 If the words $y^{(1,s)}$, $y^{(2,s)}, \dots, y^{(d_s,s)}$ in (3.7) are all distinct, we say that the sth block of y_1^N is *good* and write $s \in \mathcal{S}_{\mathrm{g}}(y_1^N)$. If that is not the case, we say that the block is *bad* and write $s \in \mathcal{S}_{\mathrm{b}}(y_1^N)$.

Lemma 3.4 *If $\mathbb{Q}$ satisfies* ID *and* FE*, then*

$$\mathbb{Q}\{y : s \in \mathcal{S}_{\mathrm{b}}(y_1^N)\} \le \mathrm{e}^{k_{\ell_-}} N^{-2\alpha},$$

for every s and every N large enough.

Proof Fix $\mathbb{Q}$ as in the statement. By shift invariance, $\mathbb{Q}\{y : s \in \mathcal{S}_{\mathrm{b}}(y_1^N)\} \le \mathbb{Q}\{y : 1 \in \mathcal{S}_{\mathrm{b}}(y_1^N)\}$.[6] For the first block to be bad, two words $y^{(i,1)}$ and $y^{(j,1)}$ need to coincide, and in particular, their ℓ_--prefixes need to coincide. Hence, considering all possible starting indices of these two words, and appealing to shift-invariance, ID and FE, we derive

$$\begin{aligned}\mathbb{Q}\{y : 1 \in \mathcal{S}_{\mathrm{b}}(y_1^N)\} &\le \sum_{r=\ell_-}^{N^\alpha} \sum_{r'=0}^{r-\ell_-} \sum_{u \in \operatorname{supp} \mathbb{Q}_{\ell_-}} \mathbb{Q}(T^{-r'}[u] \cap T^{-r}[u]) \\ &\le \binom{N^\alpha}{2} \sum_{u \in \operatorname{supp} \mathbb{Q}_{\ell_-}} \mathrm{e}^{k_{\ell_-}} \mathbb{Q}[u]^2 \\ &\le N^{2\alpha} \mathrm{e}^{k_{\ell_-}} \mathrm{e}^{\gamma_+ \ell_-}.\end{aligned}$$

To conclude, recall that we have chosen $\alpha < \frac{\gamma_+}{8\gamma_-} = -\frac{\gamma_+ \ell_-}{4 \ln N}$ and that $\gamma_\pm < 0$. ∎

Lemma 3.5 *If $\mathbb{Q}$ satisfies* ID *and* FE*, then*

$$\mathbb{Q}\{y : \#\mathcal{S}_{\mathrm{b}}(y_1^N) = m\} \le \binom{M_N}{m} \mathrm{e}^{2mk_{\ell_-}} N^{-2m\alpha},$$

for all $m \in \mathbb{N}$ and for all N large enough.

Proof Fix $\mathbb{Q}$ and m as in the statement. Let us first consider the probability that the blocks of y_1^N labeled s_m, s_{m-1} down to s_1 are bad. This event can be thought of as mth in a sequence of events defined inductively by $E'_{k+1} = T^{-N^\alpha(s_{k+1}-1)}\{1 \in \mathcal{S}_{\mathrm{b}}\} \cap E'_k$ where $E'_0 = \Omega$. It follows, by a straightforward adaptation of the strategy of Lemma 3.4, that

$$\begin{aligned}\mathbb{Q}(E'_{k+1}) &\le \binom{N^\alpha}{2} \sum_{u \in \operatorname{supp} \mathbb{Q}_{\ell_-}} \mathrm{e}^{2k_{\ell_-}} \mathbb{Q}[u]^2 \mathbb{Q}(E'_k) \\ &\le N^{2\alpha} \mathrm{e}^{2k_{\ell_-}} \max_{u \in \operatorname{supp} \mathbb{Q}_{\ell_-}} \mathbb{Q}[u] \mathbb{Q}(E'_k).\end{aligned}$$

Iterating and accounting for the different choices of $s_1, \dots, s_{m-1}, s_m$ (recall that $s \le M_N$) gives the proposed bound. ∎

Proposition 3.6 *If $\mathbb{Q}$ satisfies* ID *and* FE*, then for almost every $y \sim \mathbb{Q}$, there exists N_ε such that $\#\mathcal{S}_{\mathrm{b}}(y_1^N) < \varepsilon M_N$ for all $N \ge N_\varepsilon$.*

Proof Fixing $\mathbb{Q}$ as in the statement, using Markov's inequality, the binomial theorem, and Lemma 3.5, for every $b > 0$, we have

[6] In fact, as long as $s < M_N$, the probabilities are equal.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

$$\begin{aligned}\mathbb{Q}\left\{y : \#\mathcal{S}_{\mathrm{b}}(y_1^N) \geq \varepsilon M_N\right\} &\leq \mathbb{E}\left(\mathrm{e}^{b(\#\mathcal{S}_{\mathrm{b}}(y_1^N))}\right)\mathrm{e}^{-b\varepsilon M_N} \\ &= \mathrm{e}^{-b\varepsilon M_N}\sum_{m=1}^{M_N}\mathrm{e}^{bm}\mathbb{Q}\left\{y : \#\mathcal{S}_{\mathrm{b}}(y_1^N) = m\right\} \\ &\leq \mathrm{e}^{-b\varepsilon M_N}\left(1+\mathrm{e}^{b+2k_{\ell_-}}N^{-2\alpha}\right)^{M_N}.\end{aligned}$$

Choosing $b = 2\alpha\ln N - 2k_{\ell_-}$, recalling that $M_N/N^{1-\alpha} \in (1,2)$, and considering N large enough so that $b > 0$ gives the bound

$$\mathbb{Q}\left\{y : \#\mathcal{S}_{\mathrm{b}}(y_1^N) \geq \varepsilon M_N\right\} \leq \mathrm{e}^{-N^{1-\alpha}(2\alpha\varepsilon\ln N - 2\varepsilon k_{\ell_-} - 2\ln 2)}. \tag{3.10}$$

The proposition thus follows from the Borel–Cantelli lemma. ∎

Lemma 3.7 *Suppose that $\mathbb{P}$ satisfies* ID *and that the sth block of y_1^N is good. Given ℓ and $K \in \{1,2,\ldots,\ell\}$,*

$$\begin{aligned}&\mathbb{P}\left\{x : \#\left\{j : y^{(j,s)} = x_{K+r\ell}^{K+r\ell+(\ell-1)} \text{ for some } r \in \left\{0,1,\ldots,\left\lfloor\frac{N-K+1}{\ell}\right\rfloor - 1\right\}\right\} = m\right\} \\ &\quad\leq \binom{d_+}{m}\mathrm{e}^{mk_{\ell_+}}N^{-m\varepsilon}.\end{aligned}$$

Proof By shift invariance, we can assume that $s = 1$. Consider a set $I = \{i_k\}_{k=1}^m$ of m distinct indices such that $y^{(i_k,1)}$ has length ℓ, and let $F(I)$ denote the event that all the words $\{y^{(i_k,1)}\}_{k=1}^m$ have a match in x_1^N with a starting point equivalent to K mod ℓ. Since the words $\{y^{(i_k,1)}\}_{k=1}^m$ are distinct, the starting positions of the matches considered must be distinct. Moreover, by assumption, each such starting position is of the form $r\ell + K$ for some r at most $\lfloor\frac{N-K+1}{\ell}\rfloor - 1$. Therefore, enumerating all possibilities, we find

$$F(I) \subseteq \bigcup_{r_1,\ldots,r_m}\bigcap_{k=1}^m T^{-r_k\ell-K}[y^{(i_k,\ell)}],$$

where the union is taken over distinct nonnegative integers $r_1,\ldots,r_m$ all at most $\lfloor\frac{N-K+1}{\ell}\rfloor - 1$. Using ID, shift invariance, and subadditivity gives

$$\begin{aligned}\mathbb{P}(F(I)) &\leq m!\binom{\lfloor\frac{N-K+1}{\ell}\rfloor - 1}{m}\left(\mathrm{e}^{k_{\ell_+}}\max_{i\in I}\mathbb{P}[y^{(i,\ell)}]\right)^m \\ &\leq N^m(\mathrm{e}^{k_{\ell_+}}N^{-1-\varepsilon})^m \\ &\leq \mathrm{e}^{mk_{\ell_+}}N^{-m\varepsilon}.\end{aligned}$$

To conclude, we use a union bound, together with an upper bound on the number of sets I of this nature. ∎

Remark 3.8 The separation into fixed values of ℓ and K is a technical device to avoid overlaps that would prevent the use of ID, and will be taken care of momentarily by a union bound. For fixed ℓ, and for the purpose of relating c_N and $\tilde{c}_N$, the important quantity is the number of j such that $y^{(j,s)}$ has size ℓ and appears in x_1^N (this is the only way a separator could fail to appear within $y^{(j,s)}$), and not the number of substrings

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

of size ℓ in x_1^N that are matches for some $y^{(j,s)}$. The probability of the latter is easier to control (this is what we control in the proof), and coincides with the former when $s \in \mathcal{S}_{\mathrm{g}}(y_1^N)$.

Proposition 3.9 *Let $y \in \operatorname{supp} \mathbb{Q}$ be arbitrary and consider the modified auxiliary parsing of y_1^N in (3.7). Suppose that $\mathbb{P}$ satisfies ID and that the sth block of y_1^N is good. Then, for N large enough, the event that more than a fraction ε of the maximum number d_+ of words $y^{(i,s)}$ in the sth block appears in x_1^N satisfies*

$$\mathbb{P}\{x : \#\{j : y^{(j,s)} \in x_1^N\} > \varepsilon d_+\} \le \ell_+^2 \mathrm{e}^{\frac{\gamma_+ \varepsilon^2}{8} \frac{d_+}{\ell_+}}. \tag{3.11}$$

Proof Fix $s \in \mathcal{S}_{\mathrm{g}}(y_1^N)$. Given ℓ and $K \in \{1, \dots, \ell\}$, consider

$$\chi_{(K,\ell)} := \sum_{i : \ell_{i,s} = \ell} \mathbf{1}_{W_\ell(y^{(i,s)}, \cdot) \le N} \cdot \mathbf{1}_{W_\ell(y^{(i,s)}, \cdot) \equiv_{\mathrm{mod}\, \ell} K}. \tag{3.12}$$

Observe that, for any fixed x,

$$\#\{j : y^{(j,s)} \in x_1^N\} \le \sum_{i=1}^{d_s} \mathbf{1}_{W_{\ell_{i,s}}(y^{(i,s)}, x) \le N},$$

and so for the random variable in (3.11) to exceed εd_+, at least one of the random variables $\chi_{(K,\ell)}$ defined by (3.12) must exceed $\frac{\varepsilon d_+}{\ell_+^2}$, that is,

$$\mathbb{P}\{x : \#\{j : y^{(j,s)} \in x_1^N\} > \varepsilon d_+\} \le \mathbb{P}\left(\bigcup_{(K,\ell)} \left\{ x : \chi_{(K,\ell)}(x) > \varepsilon \frac{d_+}{\ell_+^2} \right\} \right). \tag{3.13}$$

Following the same strategy as in the proof of Proposition 3.6, we use Markov's inequality, the binomial theorem, and Lemma 3.7 to derive that, for every $b > 0$,

$$\mathbb{P}\left\{ x : \chi_{(K,\ell)}(x) > \varepsilon \frac{d_+}{\ell_+^2} \right\} \le \left(1 + \frac{\mathrm{e}^{b + k\ell_+}}{N^\varepsilon} \right)^{d_+} \mathrm{e}^{-b\varepsilon \frac{d_+}{\ell_+^2}}.$$

Choosing $b = \frac{\varepsilon}{2} \ln N$ yields

$$\begin{aligned} \mathbb{P}\left\{ x : \chi_{(K,\ell)}(x) > \varepsilon \frac{d_+}{\ell_+^2} \right\} &\le \mathrm{e}^{\frac{\gamma_+ \varepsilon^2}{4} \frac{d_+}{\ell_+} (1 - o(1))} \\ &\le \mathrm{e}^{\frac{\gamma_+ \varepsilon^2}{8} \frac{d_+}{\ell_+}} \end{aligned}$$

for N large enough, recalling that $\gamma_+ < 0$. Going back to our observation (3.13), we conclude the proof by performing a union bound over K and ℓ. ∎

3.3 Cross entropy

Lemma 3.10 *If $\mathbb{P}$ satisfies ID, $y \in \operatorname{supp} \mathbb{P}$, and y_1^N is parsed as*

$$y_1^N = y^{(1,N)} y^{(2,N)} \dots y^{(c'_N - 1, N)} y^{(c'_N, N)},$$

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

with $\ell_j := |y^{(j,N)}| \geq \lambda_N$ *for some properly diverging, nonnegative sequence* $(\lambda_N)_{N=1}^{\infty}$, *then*

$$\sum_{j=1}^{c'_N} \ln \mathbb{P}[y^{(j,N)}] = \ln \mathbb{P}[y_1^N] + o(N).$$

Proof Suppose $\mathbb{P}$ satisfies ID, $y \in \operatorname{supp} \mathbb{P}$, and y_1^N is parsed as in the statement. Both the upper and lower bounds are proved similarly, so we only provide the proof of the former. Let $\varepsilon > 0$ be arbitrary and note that ID yields

$$\begin{aligned}\ln \mathbb{P}[y_1^N] &= \ln \mathbb{P}[y^{(1,N)} y^{(2,N)} \dots y^{(c'_N-1,N)} y^{(c'_N,N)}] \\ &\leq \ln \left(\mathrm{e}^{k_{\ell_1}+\dots+k_{\ell_{c'_N-1}}} \mathbb{P}[y^{(1,N)}] \mathbb{P}[y^{(2,N)}] \dots \mathbb{P}[y^{(c'_N,N)}] \right) \\ &= \sum_{j=1}^{c'_N} \ln \mathbb{P}[y^{(j,N)}] + \sum_{j=1}^{c'_N-1} k_{\ell_j}.\end{aligned}$$

Now since $k_\ell = o(\ell)$ and $\lambda_N \to \infty$, we have $k_{\ell_j} < \varepsilon \ell_j$ for N large enough. Therefore,

$$\begin{aligned}\ln \mathbb{P}[y_1^N] &< \sum_{j=1}^{c'_N} \ln \mathbb{P}[y^{(j,N)}] + \sum_{j=1}^{c'_N-1} \varepsilon \ell_j \\ &< \sum_{j=1}^{c'_N} \ln \mathbb{P}[y^{(j,N)}] + \varepsilon N\end{aligned}$$

for N large enough. ∎

Remark 3.11 Note that the contribution coming from the buffers $\xi^{(s,N)}$, with $s \in \{1, \dots, M_N\}$, in the modified auxiliary parsing (3.7) can be embedded in the correction term $o(N)$ in the statement of Lemma 3.10. This immediately follows by observing that $M_N = o(\tilde{c}_N)$.

Lemma 3.12 *If* $\mathbb{P}$ *satisfies* ID *and* $\mathbb{Q}$ *is ergodic, and if* $\operatorname{supp} \mathbb{Q} \subseteq \operatorname{supp} \mathbb{P}$, *then*

$$-\ln \mathbb{P}[y_1^N] = N h_c(\mathbb{Q}|\mathbb{P}) + o(N)$$

for $\mathbb{Q}$*-almost every* y.

Proof Fix $\mathbb{P}$ and $\mathbb{Q}$ as in the statement. In view of the upper bound in ID, we can apply Kingman's subadditive ergodic theorem to the sequence $(f_n)_{n=1}^{\infty}$ of measurable functions on the dynamical system $(\operatorname{supp} \mathbb{P}, T, \mathbb{Q})$ defined by $f_n(x) := \ln \mathbb{P}[x_1^n]$. ∎

3.4 Comments

The following consequence of ID played an important role in the proof of the upper bound:

Ad For every $n \in \mathbb{N}$, the bound

$$\min \left\{ \frac{\mathbb{P}[ab]}{\mathbb{P}[a]} : a \in \operatorname{supp} \mathbb{P}_n, b \in \operatorname{supp} \mathbb{P}_1, ab \in \operatorname{supp} \mathbb{P}_{n+1} \right\} \geq \mathrm{e}^{-k_n} \tag{3.14}$$

holds.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

Indeed, by construction of Ziv and Merhav's auxiliary parsings, there is a lower bound on $\mathbb{P}[\underline{y}^{(j,N)}]$ and an upper bound on $\mathbb{P}[y^{(j,N)}]$, but both the bounds (3.2) and (3.5) require a lower bound on $\mathbb{P}[y^{(j,N)}]$ (see Lemma 3.3). Condition Ad serves as a way of going back and forth between the two. Unfortunately, Ad may fail upon relaxing the lower bound in ID to the more general lower-decoupling conditions that have met with success in tackling other related problems [BCJP21, CDEJR23a, CJPS19, CR23]. We will come back to this point in Section 4.4.

As for the arguments available in the literature to establish KB, we foresee no difficulty in adapting our argument to a set of hypotheses where the roles of a and b are exchanged in the decoupling inequalities. Indeed, this would not affect SE nor Ad. While the Markov property can be equivalently written in terms of conditioning on the past or conditioning on the future, the class of g-measures discussed in Section 4 and its "reverse" counterpart do not coincide (see, e.g., [BFV19, Section 4.4]).

As mentioned in the introduction, the Ziv–Merhav estimator can be written in terms of longest-match lengths:

$$\frac{c_N \ln N}{N} = \frac{\ln N}{\frac{1}{c_N}\sum_{i=1}^{c_N} \ell^{(i,N)}},$$

where[7]

$$\ell^{(i,N)} = \min\{\Lambda_N(T^{L^{(i-1,N)}} y, x), N - L^{(i-1,N)}\},$$

with

$$L^{(0,N)} = 0 \qquad \text{and} \qquad L^{(i,N)} = L^{(i-1,N)} + \ell^{(i,N)}$$

for $i = 1, 2, \ldots, c_N$. It is known that the longest-match estimator $(\ell^{(1,N)})^{-1} \ln N = \Lambda_N(y, x)^{-1} \ln N$ converges almost surely to the cross entropy, with good probability estimates, for a class of measures that is more general than that considered here (see [Kon98, Section 1.3] and [CDEJR23a, Section 3]). Hence, if each $T^{L^{(i-1,N)}} y$ were replaced by a new independent sample from $\mathbb{Q}$, or by $T^{\Delta(i-1)} y$ for some fixed deterministic $\Delta \in \mathbb{N}$, then one would expect the convergence of the Ziv–Merhav estimator to also hold considerably more generally. However, the dependence structure of the starting indices seems to be posing a serious technical difficulty for the strategy of Ziv and Merhav.

4 Examples

In this section, we discuss broad classes of measures to which our results apply. For this discussion, we need basic topological considerations that we had avoided so far. A one-sided (resp. two-sided) *subshift* is a closed and shift-invariant subset of $\mathcal{A}^{\mathbb{N}}$ (resp. $\mathcal{A}^{\mathbb{Z}}$) obtained by removing all sequences containing at least one string from some set of *forbidden strings*. Closure is understood in the product topology, and the subshift is equipped with the subspace topology inherited from that topology. A subshift is

[7] The minimum over the two terms will be given by the former as long as $i < c_N$. However, this formulation is necessary to take care of the "edge cases" alluded to in the Introduction.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

said to be *of finite type* if the list of forbidden strings that defines it can be chosen to be finite. A subshift of finite type is said to be *topologically transitive* if, for any two strings a and b with $[a]$ and $[b]$ intersecting the subshift, there exists a third string ξ such that $[a\xi b]$ also intersects the subshift. We refer the reader to [DGS76, Section 7] or [KŁO16, Section 8] for a more thorough discussion.

4.1 Markov measures

As mentioned in Section 2, if $\mathbb{P}$ is the stationary measure for an irreducible Markov chain with positive entropy, then $\mathbb{P}$ is ergodic and satisfies ID, FE, and KB. We use this setting to illustrate the role of some of our conditions.

First, it is worth noting that the stationary measure for an irreducible Markov chain need not satisfy any form of mixing, nor the Doeblin-type condition in [KS94] because it could be periodic.

In the case of a *reducible* Markov chain, a stationary measure can charge two disjoint communication classes; let us call those classes $\mathcal{A}'$ and $\mathcal{A}''$. Then, for $a \in \mathcal{A}'$, the probability $\mathbb{P}\{x : W_1(a,x) \geq r\} \geq \mathbb{P}\{x : x_1 \in \mathcal{A}''\}$ does not decay as $r \to \infty$. In terms of the language of subshifts, the failure of KB is due to the fact that $\operatorname{supp}\mathbb{P}$ does not satisfy any form of specification; it is a subshift of finite type that fails to be transitive. More concretely, if the sequence x starts in $\mathcal{A}''$, then it remains there forever and we do not expect to be able to probe any entropic quantity that also involves the behavior of $\mathbb{P}$ on $\mathcal{A}'$ using the information contained in x.

Also note that a stationary measure for an irreducible Markov chain could fail to have positive entropy if, for example, it is a convex combination of Dirac masses on periodic orbits. Such a behavior is at odds with FE and can cause the bounds on the lengths of the parsed words not to be controlled in terms of $\ell_{\pm,N}$, a fact which was used repeatedly throughout our proofs.

4.2 Regular g-measures

Let Ω' be a topologically transitive one-sided subshift of finite type. Choosing as a starting point one particular definition in the literature among others, we will say that a translation-invariant measure $\mathbb{P}$ on Ω is a *regular g-measure on* Ω' if $\operatorname{supp}\mathbb{P} = \Omega'$ and there exists a continuous function $g : \Omega' \to (0,1]$ such that

$$\sum_{\substack{y\in\Omega' \\ Ty=x}} g(y) = 1 \tag{4.1}$$

for all $x \in \Omega'$ and

$$\lim_{n\to\infty} \sup_{x\in\Omega'} \left| \frac{\mathbb{P}[x_1^n]}{\mathbb{P}[x_2^n]} - g(x) \right| = 0. \tag{4.2}$$

The convergence (4.2) can be used to show that $\mathbb{P}$ satisfies the decoupling condition ID (see [CR23, Section B.3]). Our assumption on Ω' more than suffices for ID to yield KB (see [CR23, Sections 3.1 and B.2]).

Note that the ratio being compared to g is continuous in x at finite n, and the k-level Markov condition, once written in terms of conditioning on the future, implies

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

that this ratio is eventually constant in n – starting with $n = k+1$. Hence, regular g-measures do generalize stationary k-level Markov measures.

Finally, let us discuss Condition FE in the context of regular g-measures. To do so, we will use the fact that the convergence (4.2) can also be used to establish the following weak Gibbs condition of Yuri at vanishing topological pressure: there exists an $e^{o(n)}$-sequence $(K_n)_{n=1}^{\infty}$ such that

$$K_n^{-1} e^{\sum_{j=0}^{n-1} \ln g(T^j x)} \leq \mathbb{P}[x_1^n] \leq K_n e^{\sum_{j=0}^{n-1} \ln g(T^j x)}$$

for every $x \in \Omega'$; again, see [CR23, Section B.3], but it should be noted that this can be seen as part of the "g-measure folklore" [BFV19, OST05, Wal05]. We are now ready to provide a necessary and sufficient condition on the subshift Ω' for FE to hold for all regular g-measures on Ω'. One special case will be that regular g-measures on topologically mixing subshifts of finite type with more than one letter satisfy ID and FE, allowing for an application of our main result.

Lemma 4.1 *Suppose that $\mathbb{P}$ is a regular g-measure on Ω'. Then, $\mathbb{P}$ satisfies FE if and only if there exists r with the following property: for every $y \in \Omega'$, there exists $t \leq r$ such that $T^t y$ has more than one preimage in Ω'.*

Proof Suppose that there exists r as above. Then, for every $y \in \Omega'$, there exists $t \leq r$ such that

$$g(T^{t-1}y) = 1 - \sum_{\substack{z \in \Omega' \setminus \{T^{t-1}y\} \\ Tz = T^t y}} g(z) \leq 1 - \delta,$$

where $\delta := \min g$. This number is positive by continuity and compactness. Therefore,

$$\ln g(y) + \ln g(Ty) + \cdots + \ln g(T^{t-1}y) \leq \ln(1-\delta)$$

and $\ln g(T^{t'}y) \leq \ln(1-\delta)$ for any $t' \geq t$. But then, the weak Gibbs property yields

$$\begin{aligned} \mathbb{P}[y_1^n] &\leq K_n e^{\sum_{i=0}^{\lfloor \frac{n}{r} \rfloor - 1} \sum_{t=0}^{r-1} \ln g(T^{ir+t}y)} \\ &\leq \exp\left(\ln K_n + \left\lfloor \frac{n}{r} \right\rfloor \ln(1-\delta) \right), \end{aligned}$$

with $\ln K_n = o(n)$. We conclude that Condition FE holds. Suppose now that no such r exists. Then, for every $n \in \mathbb{N}$, there exists $y \in \Omega'$ such that $T^t y$ has only one preimage in Ω' for all $t \leq n$. By the condition (4.1), this means that

$$\ln g(y) + \ln g(Ty) + \cdots + \ln g(T^{n-1}y) = 0,$$

which, together with the lower bound in the weak Gibbs property, implies

$$\mathbb{P}[y_1^n] \geq K_n^{-1} = e^{-o(n)}.$$

Since the right-hand side is eventually greater than $e^{\gamma_+ n}$ for any $\gamma_+ < 0$, FE fails as well. ∎

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

Remark 4.2 If $\mathcal{A}$ contains at least two symbols, then the hypotheses – and thus the conclusions – of Lemma 4.1 can be derived from a suitable specification property, but not any specification property from the literature.[8]

4.3 Statistical mechanics

Let $\overline{\Omega}'$ be a topologically transitive, two-sided subshift of finite type, and let Ω' be its one-sided counterpart. Consider a family $(\Phi_X)_{X \Subset \mathbb{Z}}$ of interactions with

- the continuity property that, for all $X \Subset \mathbb{Z}$, the function Φ_X – although seen as a measurable function on $\overline{\Omega}'$ – depends on the symbols with indices in the finite subset X only,
- the translation-invariance property that, for all $X \Subset \mathbb{Z}$, $\Phi_{X+1} = \Phi_X \circ T$,
- the absolute summability property that $\sum_{\substack{X \Subset \mathbb{Z} \\ X \ni 1}} \sup_{x \in \overline{\Omega}'} |\Phi_X(x)| < \infty$.

Such interactions are considered, e.g., in [Rue04, Sections 1.2 and 3.1] and are colloquially said to be in "the small space." It is well known that any equilibrium measure $\mathbb{P}$ (in the sense of the variational principle) for the energy-per-site potential

$$\phi := \sum_{\substack{X \Subset \mathbb{Z} \\ \min X = 1}} \Phi_X$$

coming from such a family of interactions is a translation-invariant Gibbs state in the sense of the Dobrushin–Lanford–Ruelle equations (see, e.g., [Rue04, Sections 3.2 and 4.2]).[9] Because we are working with a sufficiently regular subshift $\overline{\Omega}'$, the Dobrushin–Lanford–Ruelle equations and absolute summability can be used to show that $\mathbb{P}$ satisfies ID by adapting the argument of [LPS95, Section 9] for the case $\overline{\Omega}' = \mathcal{A}^{\mathbb{Z}}$. Again, the subshift is sufficiently regular for ID to yield KB (see [CR23, Sections 3.1 and B.2]).

We now turn to Condition FE, assuming a certain familiarity with the thermodynamic formalism, physical equivalence, and the Griffiths–Ruelle theorem on the reader's part (see, e.g., [Rue04, Section 4]).

Lemma 4.3 *Suppose that $\overline{\Omega}'$, Φ, and $\mathbb{P}$ are as above. If Ω' has positive topological entropy, then $\mathbb{P}$ satisfies FE.*

Proof sketch. We split the proof according to whether or not Φ is equivalent to a constant in the sense of Ruelle. Because we can always add or subtract a constant from each $\Phi_{\{i\}}$, there is no loss of generality in assuming that ϕ has topological pressure $P_{\text{top}}(\phi) = 0$.

Case 1. On the one hand, if Φ is not equivalent to a constant in the sense of Ruelle, then the Griffiths–Ruelle theorem guarantees that $\alpha \mapsto P_{\text{top}}(\phi - \alpha\phi)$ is

[8] For example, in the terminology of [KŁO16, Section 8], Property (6) suffices, but Property (8) does not, as it allows $\Omega' = \{0101010101010\ldots, 1010101010101\ldots\}$.

[9] With a slight abuse of notation, we are using $\mathbb{P}$ for both the equilibrium measure on $\overline{\Omega}' \subseteq \mathcal{A}^{\mathbb{Z}}$ and its natural marginal on $\Omega' \subseteq \mathcal{A}^{\mathbb{N}}$. Note that, by construction, the potential ϕ only depends on symbols from Ω'.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

strictly convex (see, e.g., [Rue04, Section 4.6]). On the other hand, by the weak Gibbs property established, e.g., in [PS20, Section 2], we have

$$\lim_{n\to\infty} \frac{1}{n} \ln \sum_{a\in \operatorname{supp} \mathbb{P}_n} \mathbb{P}[a]^{1-\alpha} = P_{\mathrm{top}}(\phi - \alpha\phi).$$

It is easy to see from this relation that the function $\alpha \mapsto P_{\mathrm{top}}(\phi - \alpha\phi)$ is non-decreasing. Combining these properties, we deduce that $P_{\mathrm{top}}(\phi - \alpha\phi) < 0$ for all $\alpha < 0$. Assuming for the sake of contradiction that FE fails, one easily derives a contradiction.

Case 2. If, on the contrary, Φ is equivalent to a constant the sense of Ruelle, then the weak Gibbs property reads,

$$K_n^{-1} \mathrm{e}^{vn} \le \mathbb{P}[x_1^n] \le K_n \mathrm{e}^{vn}$$

for all $x \in \Omega'$ and some constant v. Given that Ω' has positive topological entropy, summing the left-most inequality over x_1^n, the fact that $K_n = \mathrm{e}^{o(n)}$ can be used to show that $v < 0$. Then, the right-most inequality yields that FE holds, again thanks to the fact that $K_n = \mathrm{e}^{o(n)}$. ∎

Every irreducible, stationary Markov measure with stochastic matrix $[P_{a,b}]_{a,b\in\mathcal{A}}$ can be obtained in this way by considering the following nearest-neighbor interactions on its support:

$$\Phi_{\{i,i+1\}}(x) = \ln P_{x_i,x_{i+1}}$$

for $i \in \mathbb{N}$ and $\Phi_X(x) = 0$ for X not of the form $\{i, i+1\}$. To see this, one can check by direct computation that, on its support, the Markov measure satisfies the Bowen–Gibbs condition for the corresponding ϕ. For k-level Markov measures, consider instead

$$\Phi_{\{i,\dots,i+k-1,i+k\}}(x) = \ln \frac{\mathbb{P}[x_i \dots x_{i+k-1}x_{i+k}]}{\mathbb{P}[x_i \dots x_{i+k-1}]}.$$

In this sense, equilibrium measures for potentials arising from interactions that are absolutely summable do generalize stationary k-level Markov measures; we refer the reader to [BGM+21, CHM+14] for recent thorough discussions of variants and converses to this observation. This generalization is far reaching as the theory of entropy, large deviations, and phase transition is much richer in the small space of interactions than in the space of finite-range interactions.

In a similar vein, equilibrium measures (in the sense of the variational principle on Ω') for abstract potentials ϕ in the Bowen class also satisfy ID, thanks to the Bowen–Gibbs property (see [Wal01, Section 4]). We refer the reader to [Wal01, Section 1] for a definition of the Bowen class, which can be traced back to [Bow74]. This class includes potentials with summable variations, and thus Hölder-continuous potentials, and thus potentials naturally associated with stationary k-level Markov measures. A more complete discussion from the point of view of decoupling – including relaxation of the conditions on Ω' – can be found in [CR23, Section 2.3].

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

4.4 Hidden-Markov measures

While the above generalizations beyond Markovianity are often studied in the literature on mathematical physics and abstract dynamical systems, they might not be the most natural from an information-theoretic point of view; hidden-Markov models would most likely come to mind first for many practitioners. We recall that, among several equivalent representations, a stationary hidden-Markov measure $\mathbb{P}$ can be characterized by a tuple (π, P, R) where (π, P) characterizes in the usual way a stationary Markov process on a set $\mathcal{S}$, called the *hidden alphabet*, and R is a $(\#\mathcal{S})$-by-$(\#\mathcal{A})$ matrix whose rows each sum to 1:

$$\mathbb{P}[a_1^n] = \sum_{s_1^n \in \mathcal{S}^n} \pi_{s_1} R_{s_1,a_1} P_{s_1,s_2} R_{s_2,a_2} \cdots P_{s_{n-1},s_n} R_{s_n,a_n}$$

for $n \in \mathbb{N}$ and $a_1^n \in \mathcal{A}^n$. We restrict our attention to the case where $\mathcal{S}$ is a finite set and P is irreducible. We view the entry $R_{s,a}$ as the probability of observing $a \in \mathcal{A}$ at a given time step given the *hidden state* $s \in \mathcal{S}$ at that same time step – the dynamics of the latter governed by the hidden-Markov chain (π, P). There exist only very singular examples of such measures for which FE fails. As exhibited by our next lemma, this can only happen if the process is eventually almost-surely deterministic.

Lemma 4.4 *Let $\mathbb{P}$ be as above. Then, $\mathbb{P}$ satisfies FE if and only if, for each $s \in \mathcal{S}$, there exists L such that*

$$\#\{a \in \mathcal{A}^L : \mathbb{P}[a|s_1 = s] > 0\} > 1.$$

Proof Suppose that for each $s \in \mathcal{S}$ there exists L as above. By inspection of the canonical form of P provided by the Perron–Frobenius theorem, one deduces that there exists a finite set Σ' of possible row vectors σ that can arise as limit points for sequences of the form $([P^m]_{i,\cdot})_{m=1}^{\infty}$. Let $\Sigma := \Sigma' \cup \{\pi\}$ with π the unique invariant probability row vector for P. By stochasticity, each $\sigma \in \Sigma$ has nonnegative entries that sum to 1. In this context, by assumption, there exists $L \in \mathbb{N}$ such that

$$\begin{aligned}\delta &:= \max_{a \in \operatorname{supp} \mathbb{P}_L} \max_{\sigma \in \Sigma} \sum_{s \in \mathcal{S}^L} \sigma_{s_1} P_{s_1,s_2} \cdots P_{s_{L-1},s_L} R_{s_1,a_1} \cdots R_{s_L,a_L} \\ &= \max_{a \in \operatorname{supp} \mathbb{P}_L} \max_{\sigma \in \Sigma} \sum_{s \in \mathcal{S}} \sigma_s \mathbb{P}[a|s_1 = s]\end{aligned}$$

is strictly less than 1. Given $\varepsilon > 0$, by inspection of the same canonical form, there exists $m \in \mathbb{N}$ with the following property: for all i, there is $\sigma \in \Sigma$ such that

$$[P^m]_{i,\cdot} < \sigma + \varepsilon.$$

Then, taking $a \in \operatorname{supp} \mathbb{P}$ and $n \geq L$,

$$\mathbb{P}[a_1^n] \leq \mathbb{P}[a_1^{L+q(m+L)}]$$

for $q := \max\{k \in \mathbb{N}_0 : n \geq L + k(m+L)\}$. We introduce the shorthands $\mathcal{R}_0(s) = R_{s_1,a_1} \cdots R_{s_L,a_L}$,

$$\mathcal{R}_k(s) = R_{s_{(k-1)(m+L)+L+1}, a_{(k-1)(m+L)+L+1}} \cdots R_{s_{k(m+L)+L}, a_{k(m+L)+L}}$$

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

and

$$\mathcal{R}'_k(s) = R_{s_{k(m+L)+1},a_{k(m+L)+1}} \cdots R_{s_{k(m+L)+L},a_{k(m+L)+L}}$$

when $1 \le k \le q$. We also identify $s_0^1 \equiv s_L$, $s_0^2 \equiv s_{m+L}^1$, $s_0^3 \equiv s_{m+L}^2$, and so on, and so forth. One then obtains

$$\begin{aligned}
&\mathbb{P}[a_1^{L+q(m+L)}] \\
&= \sum_{\substack{s_1,\dots,s_L \\ s_1^k,\dots,s_{m+L}^k \\ \text{for } 1\le k\le q}} \pi_{s_1} P_{s_1,s_2} \cdots P_{s_{L-1},s_L} \mathcal{R}_0(s) \prod_{k=1}^{q} P_{s_0^k,s_1^k} P_{s_1^k,s_2^k} \cdots P_{s_{m+L-1}^k,s_{m+L}^k} \mathcal{R}_k(s) \\
&\le \sum_{\substack{s_1,\dots,s_L \\ s_m^k,\dots,s_{m+L}^k \\ \text{for } 1\le k\le q}} \pi_{s_1} P_{s_1,s_2} \cdots P_{s_{L-1},s_L} \mathcal{R}_0(s) \prod_{k=1}^{q} (\sigma_{s_m^k}^{(s_0^k)} + \varepsilon) P_{s_m^k,s_{m+1}^k} \cdots P_{s_{m+L-1}^k,s_{m+L}^k} \mathcal{R}'_k(s)
\end{aligned}$$

for some appropriate choices of $\sigma^{(s_0^k)} \in \Sigma$ that depend on m and the index s_0^k only. Therefore,

$$\mathbb{P}[a_1^{L+q(m+L)}] \le \delta \cdot (\delta + \varepsilon(\#\mathcal{S}))^q.$$

By taking $\varepsilon > 0$ such that $\delta + \varepsilon(\#\mathcal{S}) < 1$ and noting that q scales linearly with n, FE holds.

To see the converse implication, suppose that there exists $t \in \mathcal{S}$ such that there is no L as above. Then, there exists $a \in \Omega$ such that $\mathbb{P}[a_1^n | s_1 = t] = 1$ for all $n \in \mathbb{N}$. Since

$$\mathbb{P}[a_1^n] \ge \mathbb{P}[a_1^n | s_1 = t] \cdot \pi_t = \pi_t$$

for all $n \in \mathbb{N}$ and $e^{\gamma_+ n}$ is eventually smaller than $\pi_t > 0$ for all $\gamma_+ < 0$, FE fails. ∎

One can show that every stationary hidden-Markov measure satisfies the upper bound in ID. But in general – even if P is irreducible – only a weaker form of the lower bound, known as *selective lower decoupling*, holds (see [BCJP21, Section 2] and [CJPS19, Section 2]). The fact that selective lower decoupling implies KB but does not imply the condition called Ad in Section 3.4 seems to pose a genuine obstacle.

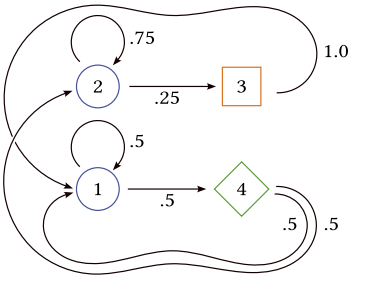

Figure 1: An example that does not satisfy Ad: a direct computation shows that $[\circ\circ\dots\circ\diamond]$ is too unlikely compared to $[\circ\circ\dots\circ]$.

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press

Determining whether the ZM estimation remains generally valid in the class of irreducible, hidden-Markov measures remains – to the best of our knowledge – an important open problem.

In the further specialized case where the elements of R are all in $\{0,1\}$ – this is sometimes called the *function-Markov* or *lumped-Markov* case – some conditions for the g-measure property (and thus ID) are discussed in [CU03, Ver11, Yoo10]. However, it is not difficult to find examples for which none of these known sufficient conditions hold.

Example 4.5 The stationary measure on $\{\circ, \diamond, \square\}^{\mathbb{N}}$ built from the four-hidden-state chain depicted in Figure 1 satisfies the upper bound in ID, as well as FE and SE, but not Ad – and hence not ID. However, note that this example satisfies the Doeblin-type condition of [KS94] with $r = 3$.

Acknowledgements The authors would like to thank G. Cristadoro, N. Cuneo, and V. Jakšić for stimulating discussions on the topic of this note, as well as the referees for comments that helped improve the manuscript.

References

[BGM+21] S. Barbieri, R. Gómez, B. Marcus, T. Meyerovitch, and S. Taati, *Gibbsian representations of continuous specifications: the theorems of Kozlov and Sullivan revisited*. Commun. Math. Phys. **382**(2021), 1111–1164.

[BBCDE08] C. Basile, D. Benedetto, E. Caglioti, and M. Degli Esposti, *An example of mathematical authorship attribution*. J. Math. Phys. **49**(2008), no. 12, 125211.

[BCL02] D. Benedetto, E. Caglioti, and V. Loreto, *Language trees and zipping*. Phys. Rev. Lett. **88**(2002), 048702.

[BCJP21] T. Benoist, N. Cuneo, V. Jakšić, and C.-A. Pillet, *On entropy production of repeated quantum measurements II examples*. J. Stat. Phys. **182**(2021), no. 3, 1–71.

[BJPP18] T. Benoist, V. Jakšić, Y. Pautrat, and C.-A. Pillet, *On entropy production of repeated quantum measurements I general theory*. Commun. Math. Phys. **357**(2018), no. 1, 77–123.

[BFV19] S. Berghout, R. Fernández, and E. Verbitskiy, *On the relation between Gibbs and g-measures*. Ergodic Theor. Dyn. Syst. **39**(2019), no. 12, 3224–3249.

[Bow74] R. Bowen, *Some systems with unique equilibrium states*. Math. Syst. Theor. **8**(1974), no. 3, 193–202.

[Bra05] R. C. Bradley, *Basic properties of strong mixing conditions. A survey and some open questions*. Probab. Surv. **2**(2005), 107–144.

[CHM+14] N. Chandgotia, G. Han, B. Marcus, T. Meyerovitch, and R. Pavlov, *One-dimensional Markov random fields, Markov chains and topological Markov fields*. Proc. Amer. Math. Soc. **142**(2014), no. 1, 227–242.

[CU03] J.-R. Chazottes and E. Ugalde, *Projection of Markov measures may be Gibbsian*. J. Stat. Phys. **111**(2003), no. 5/6, 1245–1272.

[CF05] D. P. Coutinho and M. A. Figueiredo, *Information theoretic text classification using the Ziv–Merhav method*. In: J. S. Marques, N. Pérez de la Blanca, and P. Pina (eds.), Pattern recognition and image analysis, Lecture Notes in Computer Science, 3523, Springer, Berlin, 2005, pp. 355–362.

[CFF10] D. P. Coutinho, A. L. Fred, and M. A. Figueiredo, *One-lead ECG-based personal identification using Ziv–Merhav cross parsing*. In: 20th international conference on pattern recognition, IEEE, Los Alamitos, 2010, pp. 3858–3861.

[CDEJR23a] G. Cristadoro, M. Degli Esposti, V. Jakšić, and R. Raquépas, *On a waiting-time result of Kontoyiannis: mixing or decoupling?* Stoch Proc. Appl. **166**(2023), 104222.

[CDEJR23b] G. Cristadoro, M. Degli Esposti, V. Jakšić, and R. Raquépas, *Recurrence times, waiting times and universal entropy production estimators*. Lett. Math. Phys. **113**(2023), no. 1, Article no. 19.

[CJPS19] N. Cuneo, V. Jakšić, C.-A. Pillet, and A. Shirikyan, *Large deviations and fluctuation theorem for selectively decoupled measures on shift spaces*. Rev. Math. Phys. **31**(2019), no. 10, 1950036.

[CR23] N. Cuneo and R. Raquépas, *Large deviations of return times and related entropy estimators on shift spaces*. Commun. Math. Phys. Preprint, 2023, arXiv:2306.05277 [math.PR].

[DGS76] M. Denker, C. Grillenberger, and K. Sigmund, Ergodic theory on compact spaces, Lecture Notes in Mathematics, 527, Springer, Berlin, 1976.

[Kon98] I. Kontoyiannis, *Asymptotic recurrence and waiting times for stationary processes*. J. Theor. Probab. **11**(1998), no. 3, 795–811.

[KS94] I. Kontoyiannis and Y. M. Suhov, *Prefixes and the entropy rate for long-range sources*. In: F. P. Kelly (ed.), Probability, statistics and optimization: a tribute to Peter whittle, Wiley, New York, 1994.

[KŁO16] D. Kwietniak, M. Łącka, and P. Oprocha, *A panorama of specification-like properties and their consequences*. In: S. Kolyad, M. Möller, P. Moree, and T. Ward (eds.), Dynamics and numbers, Contemporary Mathematics, 669, American Mathematical Society, Providence, RI, 2016, pp. 155–186.

[LPS95] J. T. Lewis, C.-É. Pfister, and W. G. Sullivan, *Entropy, concentration of probability and conditional limit theorems*. Markov Proc. Relat. Fields **1**(1995), no. 3, 319–386.

[LMDEC19] M. Lippi, M. A. Montemurro, M. Degli Esposti, and G. Cristadoro, *Natural language statistical features of LSTM-generated texts*. IEEE Trans. Neural Netw. Learn. Syst. **30**(2019), no. 11, 3326–3337.

[OST05] E. Olivier, N. Sidorov, and A. Thomas, *On the Gibbs properties of Bernoulli convolutions related to β-numeration in multinacci bases*. Monatshefte Math. **145**(2005), no. 2, 145–174.

[Pfi02] C.-É. Pfister, *Thermodynamical aspects of classical lattice systems*. In: V. Sidoravicius (ed.), In and out of equilibrium: probability with a physics flavor, Progress in Probability, 51, Birkhäuser, 2002, pp. 393–472.

[PS20] C.-É. Pfister and W. G. Sullivan, *Asymptotic decoupling and weak Gibbs measures for finite alphabet shift spaces*. Nonlinearity **33**(2020), no. 9, 4799–4817.

[RGS+22] S. Ro, B. Guo, A. Shih, T. V. Phan, R. H. Austin, D. Levine, P. M. Chaikin, and S. Martiniani, *Model-free measurement of local entropy production and extractable work in active matter*. Phys. Rev. Lett. **129**(2022), no. 22, 220601.

[RP12] É. Roldán and J. M. R. Parrondo, *Entropy production and Kullback–Leibler divergence between stationary trajectories of discrete systems*. Phys. Rev. E. **85**(2012), 031129.

[Rue04] D. Ruelle, Thermodynamic formalism, 2nd ed., Cambridge University Press, Cambridge, 2004.

[Shi93] P. C. Shields, *Waiting times: positive and negative results on the Wyner–Ziv problem*. J. Theor. Probab. **6**(1993), no. 3, 499–519.

[Shi96] P. C. Shields, The ergodic theory of discrete sample paths, Graduate Studies in Mathematics, 13, American Mathematical Society, Providence, RI, 1996.

[vEFS93] A. C. van Enter, R. Fernández, and A. D. Sokal, *Regularity properties and pathologies of position-space renormalization-group transformations: scope and limitations of Gibbsian theory*. J. Stat. Phys. **72**(1993), 879–1167.

[Ver11] E. Verbitskiy, *Thermodynamics of hidden Markov processes*. In: B. Marcus, K. Petersen, and T. Weissman (eds.), Entropy of hidden Markov processes and connections to dynamical systems, London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, 2011, pp. 258–272.

[Wal01] P. Walters, *Convergence of the Ruelle operator for a function satisfying Bowen's condition*. Trans. Amer. Math. Soc. **353**(2001), no. 1, 327–347.

[Wal05] P. Walters, *Regularity conditions and Bernoulli properties of equilibrium states and g-measures*. J. Lond. Math. Soc. **71**(2005), no. 2, 379–396.

[WZ89] A. D. Wyner and J. Ziv, *Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression*. IEEE Int. Symp. Inf. Theory. **35**(1989), no. 6, 1250–1258.

[Yoo10] J. Yoo, *On factor maps that send Markov measures to Gibbs measures*. J. Stat. Phys. **141**(2010), no. 6, 1055–1070.

[ZL78] J. Ziv and A. Lempel, *Compression of individual sequences via variable-rate coding*. IEEE Trans. Inf. Theory **24**(1978), no. 5, 530–536.

[ZM93] J. Ziv and N. Merhav, *A measure of relative entropy between individual sequences with application to universal classification*. IEEE Trans. Inf. Theory **39**(1993), no. 4, 1270–1279.

Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada
e-mail: nicholas.barnfield@mail.mcgill.ca

Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada
e-mail: raphael.grondin@mail.mcgill.ca

Department of Mathematics, CY Cergy Paris Université, CNRS UMR 8088, Cergy-Pontoise, France
e-mail: gaia.pozzoli@cyu.fr

Courant Institute of Mathematical Sciences, New York University, New York, NY, United States
e-mail: rr4374@nyu.edu

https://doi.org/10.4153/S0008414X24000178 Published online by Cambridge University Press