On Analytic Properties of Entropy Rate

Alexander Schönhuth, *Member, IEEE,*

***Abstract*—Entropy rate of discrete random sources are a real valued functional on the space of probability measures associated with the random sources. If one equips this space with a topology one can ask for the analytic properties of the entropy rates. A natural choice is the topology, which is induced by the norm of total variation. A central result is that entropy rate is Lipschitz continuous relative to this topology. The consequences are manifold. First, corollaries are obtained that refer to prevalent objects of probability theory. Second, the result is extended to entropy rate of dynamical systems. Third, it is shown how to exploit the proof schemes to give a direct and elementary proof for the existence of entropy rate of asymptotically mean stationary random sources.**

***Index Terms*—Entropy rate, ergodic theorems, entropy, dynamical system, asymptotically mean stationarity.**

I. Introduction

ENTROPY rate is a key quantity in information theory as it is equal to the average amount of information per symbol of a random source if it exists. Therefore, it is natural to ask how entropy rate behaves if knowledge of random sources is subject to uncertainties which, for example, may be inherent to inference processes and/or originate from noisy channels. However, closed formulas for entropy rate exist only in rare examples. For example, already hidden Markov sources (HMSs) seem to defy a convenient formula although there is one for the special case of Markov sources. Therefore, in this case, recent efforts focused on the direct investigation of analytic properties of entropy rate like smoothness or even analyticity [20], [21], [29], [30], [25], [16].

The purpose of this paper is to contribute to the issue of analytic properties of entropy rate in a more general fashion. Namely, we study the behavior of entropy rate relative to the topology induced by the norm of total variation. This topology is one of the natural choices and it is ubiquitous in both theoretical and practical work. We show that entropy rate is Lipschitzian on the whole space of discrete random sources which is, due to an elementary theorem of Rademacher, close to differentiability.

We will use this result to give a more elementary proof of the existence of entropy rate for asymptotically mean stationary sources which, for example, contain the classes of arbitrary HMSs [24] and quantum random walks (QRWs) [1], [5] as has recently been proven [6].

II. Preliminaries

As usual, $\Sigma^* = \cup_{t\geq 0}\Sigma^t$ is the set of all words (strings of finite length) over the finite alphabet Σ together with the concatenation operation

$$\bar{a} \in \Sigma^t, \bar{b} \in \Sigma^s \quad \Rightarrow \quad \bar{a}\bar{b} \in \Sigma^{t+s}.$$

For formal convenience, we have an *empty word* $\square$ such that $\Sigma^0 = \{\square\}$.

Throughout this paper we will write $\Omega = \Sigma^{\mathbb{N}} = \bigotimes_{t=0}^{\infty}\Sigma$ for the set of sequences over Σ and $\mathcal{B}$ for the σ-algebra generated by the cylinder sets. We identify cylinder sets B with sets of words $A_B \in \Sigma^t$ such that B is the set of sequences which start with the words in A_B.

As usual, we identify stochastic processes $(X_t)_{t\in\mathbb{N}}$ with values in Σ with probability measures P_X on the measurable space $(\Omega, \mathcal{B})$ and vice versa via the relationship (as discussed above, $\bar{a} = a_0...a_{t-1} \in \Sigma^t$ corresponds to the cylinder set of sequences starting with the word $\bar{a}$)

$$P_X(\bar{a}) = P(\{X_0 = a_0, X_1 = a_1, ..., X_{t-1} = a_{t-1}\}),$$

where the term on the right hand side is the probability that the random source emits the symbols $a_0, ..., a_{t-1}$ at periods $0, ..., t-1$. As a consequence of the extension theorem ([12], p. 54, th. A), a stochastic process (X_t) is uniquely determined by the values $P_X(\bar{a})$ for all $\bar{a} \in \Sigma^*$.

A. Finite signed measures

Let

$$\mathcal{P}(\Sigma) \subset \{P : \mathcal{B}(\Sigma) \to \mathbb{R}\}$$

be the set of finite, signed measures on $(\Omega, \mathcal{B}(\Sigma))$, that is, the set of σ-additive but not necesarily positive, finite set functions on $\mathcal{B}(\Sigma)$. We write $\mathcal{P}$ instead of $\mathcal{P}(\Sigma)$, if the alphabet is understood. By eventwise addition and scalar multiplication, $\mathcal{P}(\Sigma)$ becomes a vector space. The most important relevant properties of finite signed measures are summarized in the following theorem (see [12], ch. VI for proofs).

Theorem 2.1:

1) The *Jordan decomposition* theorem tells that for every $P \in \mathcal{P}$ there are finite measures P_+, P_- such that

   $$P = P_+ - P_-$$

   and for every other decomposition $P = P_1 - P_2$ with measures P_1, P_2 it holds that $P_1 = P_+ + \delta, P_2 = P_- + \delta$ for another measure δ. In this sense, P_+ and P_- are unique and called *positive* resp. *negative variation*. The measure $|P| := P_+ + P_-$ is called *total variation*.
2) In parallel to the Jordan decomposition we have the *Hahn decomposition* of Ω into two disjoint events Ω_+, Ω_-

   $$\Omega = \Omega_+ \dot\cup \Omega_-$$

A. Schönhuth is with the Center for Applied Computer Science, University Cologne, D-50931 Köln, Germany, e-mail: (see http://www.zaik.uni-koeln.de/ aschoen).

such that $P_-(\Omega_+) = 0$ and $P_+(\Omega_-) = 0$. Ω_+, Ω_- are uniquely determined up to $|P|$-null-sets.

3) The *norm of total variation* $||.||_{TV}$ on $\mathcal{P}$ is given by

$$||P||_{TV} := |P|(\Omega) = P_+(\Omega) + P_-(\Omega) \\ = P_+(\Omega_+) + P_-(\Omega_-).$$

Obviously $||\,|P|\,||_{TV} = ||P||_{TV}$.

As the norm of total variation seems to be the most natural choice for a norm, it is omnipresent in both related theoretical and practical work. Computation of the norm of total variation, however, is not always easy. The following lemma shows a practicable way.

Lemma 2.1:

$$||P||_{TV} = \sup_{t\in\mathbb{N}} \sum_{\bar{a}\in\Sigma^t} |P(\bar{a})| = \lim_{t\to\infty} \sum_{\bar{a}\in\Sigma^t} |P(\bar{a})|.$$

Again, $P(\bar{a})$ denotes the value of the finite signed measure P on the cylinder set of sequences having prefix $\bar{a}$.

Proof: As the proof is of purely measure theoretical nature, we have deferred it to appendix A. ∎

Lemma 2.2: Let $P; Q \in \mathcal{P}(\Sigma)$ be two finite, signed measures. Then it holds that

$$\sup_{B\in\mathcal{B}} |P(B) - Q(B)| \;\leq\; ||P - Q||_{TV} \\ \leq\; 2 \sup_{B\in\mathcal{B}} |P(B) - Q(B)|.$$

Hence the topology induced by the norm of total variation is that of the metric

$$d(P, Q) := \sup_{B\in\mathcal{B}} |P(B) - Q(B)|.$$

Proof: We have deferred the measure theoretical proof to appendix B.. ∎

Note further, as one can see from the proof of lemma 2.2, that for two probability measures P_1, P_2 it holds that

$$||P_1 - P_2||_{TV} = 2 \cdot \sup_{B\in\mathcal{B}} |P_1(B) - P_2(B)|.$$

B. Entropy Rates

We write $\mathcal{P}^+ \subset \mathcal{P}$ for the subset of measures, which is a convex cone in $\mathcal{P}$ and view the entropy rates as a functional on it:

$$\begin{array}{rccc} \overline{\mathbf{H}}: & \mathcal{P}^+(\Sigma) & \longrightarrow & \mathbb{R} \\ & P & \mapsto & \limsup_{t\in\mathbb{N}} \frac{1}{t} \sum_{\bar{a}\in\Sigma^t} P(\bar{a}) \log \frac{1}{P(\bar{a})} \end{array}.$$

Note that we define entropy rates, in a slightly more general fashion, not only for probability measures.

III. Analytic properties of entropy rates

A. Lipschitz continuity

Our main result is the following theorem, which tells that the entropy rates are Lipschitz continuous on the cone of measures with respect to the topology induced by the norm of total variation.

Theorem 3.1: The real valued functional $\overline{\mathbf{H}}$ is Lipschitzian with $Lip(\overline{\mathbf{H}}) = \log |\Sigma|$, i.e.

$$|\overline{\mathbf{H}}(P_1) - \overline{\mathbf{H}}(P_2)| \leq (\log |\Sigma|)\, ||P_1 - P_2||_{TV},$$

where $P_1, P_2 \in \mathcal{P}^+(\Sigma)$.

Note that this makes the entropy rates a Lipschitzian functional on the (convex) subset of probability measures of $\mathcal{P}$, which one usually is interested in.

To be prepared for the proof we present two lemmata, which incorporate the essential ideas. We write

$$||P||_{TV,t} := \sum_{\bar{a}\in\Sigma^t} |P(\bar{a})|$$

for a signed measure $P \in \mathcal{P}$ and

$$H^t(P) := -\frac{1}{t} \sum_{\bar{a}\in\Sigma^t} P(\bar{a}) \log\, P(\bar{a}),$$

for a measure $P \in \mathcal{P}^+$. Note that for a probability measure, $P(\bar{a})$ coincides with the probability that a sequence emitted by the random source relative to P starts with the word $\bar{a}$, as discussed above. Lemma 2.1 tells that

$$\lim_{t\to\infty} ||P||_{TV,t} = ||P||_{TV}.$$

Note that $||.||_{TV,t}$ is not a norm on $\mathcal{P}$.

Lemma 3.1: Let $P, Q \in \mathcal{P}^+$ be two measures such that $||P - Q||_{TV} \leq \frac{1}{e}$. Then it holds that

$$|H^t(P) - H^t(Q)| \\ \leq (\log |\Sigma| + \frac{1}{t} \log \frac{1}{||P - Q||_{TV,t}}) \cdot ||P - Q||_{TV,t},$$

where $0 \cdot \log \infty := 0$ in case of $||P - Q||_{TV,t} = 0$.

For the proof of this lemma we will need two sublemmata.

Sublemma *1:* Let

$$\begin{array}{rccc} h: & [0,1] & \longrightarrow & \mathbb{R} \\ & x & \mapsto & x \log \frac{1}{x}. \end{array}$$

Then it holds for $x, y \in [0, 1]$:

$$|x - y| \leq \frac{1}{e} \quad\Longrightarrow\quad |h(x) - h(y)| \leq h(|x - y|). \qquad (1)$$

Proof: The proof is a technical, analytic exercise. Note first that $h'(x) = \log \frac{1}{x} - 1$ and $h''(x) = -\frac{1}{x}$. Hence h is concave, has a global maximum at $\frac{1}{e}$ and $h(\frac{1}{e}) = \frac{1}{e}$. We therefore note that

$$x \leq h(x) \quad\Longleftrightarrow\quad x \leq \frac{1}{e}. \qquad (2)$$

Because of

$$\begin{aligned}|h(x)-h(y)| &= ||h(x)-h(\frac{1}{e})|-|h(\frac{1}{e})-h(y)|| \\ &\leq \max\{|h(x)-h(\frac{1}{e})|, |h(\frac{1}{e})-h(y)|\}\end{aligned} \tag{3}$$

and the fact that h is monotically increasing on $[0, \frac{1}{e}]$ we can, without loss of generality, assume that either $x, y \geq \frac{1}{e}$ or $x, y \leq \frac{1}{e}$.

As

$$\forall x \in [\frac{1}{e}, 1]: \quad |h'(x)| \leq 1, \tag{4}$$

and because of the mean value theorem, it holds that

$$\frac{1}{e} \leq x, y \leq 1 \quad \Longrightarrow \quad |h(x)-h(y)| \leq |x-y|. \tag{5}$$

Because of equation (2) we obtain the statement for the case $\frac{1}{e} \leq x, y \leq 1$.

It remains the case (w.l.o.g. $x < y$) $x < y \leq \frac{1}{e}$. Here it holds that $|h(x)-h(y)| = h(y)-h(x)$. We note that the function $\log \frac{1}{t} - 1$ is positive and monotonically decreasing on $[0, \frac{1}{e}]$ ($(*)$). We obtain the claim from the calculation

$$\begin{aligned}|h(x)-h(y)| &= \int_x^y (\log \frac{1}{t} - 1)dt \\ &\overset{(*)}{\leq} \int_x^y (\log \frac{1}{t-x} - 1)dt \\ &\overset{s=t-x}{=} \int_0^{y-x} (\log \frac{1}{s} - 1)ds \\ &= \left[s \log \frac{1}{s}\right]_0^{y-x} = h(y-x).\end{aligned} \tag{6}$$

■

Let now

$$\Delta^{n-1} := \{x = (x_1, ..., x_n) \in \mathbb{R}^n \mid x_i \geq 0, \sum_i x_i = 1\},$$

be the usual regular $n-1$-dimensional simplex in $\mathbb{R}^n$.

Sublemma *2:* Let $0 < K \in \mathbb{R}$, $2 \leq n \in \mathbb{N}$ and

$$\Delta_K^{n-1} := K \cdot \Delta^{n-1} := \{x \in \mathbb{R}^n \mid (1/K)x \in \Delta^{n-1}\},$$

The function

$$\begin{array}{rccc} h_{K,n}: & \Delta_{K,n-1} & \longrightarrow & \mathbb{R} \\ & x = (x_1, ..., x_n) & \mapsto & \sum_{i=1}^n x_i \log \frac{1}{x_i} \end{array} \tag{7}$$

attains a global maximum at $\bar{x} := (K/n, ..., K/n)$.

Proof: This is a straightforward generalization of the case $K = 1$, for which the claim is a well known result (e.g. [13]). ■

We are now able to prove lemma 3.1.

Proof: Obviously $H^t(P) = H^t(Q)$ in case of $||P-Q||_{TV,t} = 0$. In case of $||P-Q||_{TV,t} > 0$ we have

$$\begin{aligned}&|H^t(P) - H^t(Q)| \\ &= \frac{1}{t}|\sum_{\bar{a} \in \Sigma^t} P(\bar{a}) \log \frac{1}{P(\bar{a})} - Q(\bar{a}) \log \frac{1}{Q(\bar{a})}| \\ &\leq \frac{1}{t} \sum_{\bar{a} \in \Sigma^t} |P(\bar{a}) \log \frac{1}{P(\bar{a})} - Q(\bar{a}) \log \frac{1}{Q(\bar{a})}| \\ &\overset{Subl.\ 1}{\leq} \frac{1}{t} \sum_{\bar{a} \in \Sigma^t} |P(\bar{a}) - Q(\bar{a})| \log \frac{1}{|P(\bar{a}) - Q(\bar{a})|} \\ &\overset{Subl.\ 2}{\leq} \frac{1}{t} \sum_{\bar{a} \in \Sigma^t} \frac{||P-Q||_{TV,t}}{|\Sigma|^t} \log \frac{|\Sigma|^t}{||P-Q||_{TV,t}} \\ &= \frac{1}{t}||P-Q||_{TV,t}(\sum_{\bar{a} \in \Sigma^t} \frac{1}{|\Sigma|^t} t \log |\Sigma| \\ &\qquad + \sum_{\bar{a} \in \Sigma^t} \frac{1}{|\Sigma|^t} \frac{1}{||P-Q||_{TV,t}}) \\ &= \frac{1}{t}||P-Q||_{TV,t}(t \log |\Sigma| + \frac{1}{||P-Q||_{TV,t}}).\end{aligned}$$

■

To get control of the limes superior involved in the definition of entropy rates we will further need

Lemma 3.2: Let (a_t) and (b_t) two non-negative real valued sequences such that

$$|a_t - b_t| \leq c_t \quad \text{and} \quad \lim_{t \to \infty} c_t = c.$$

Then it holds that

$$|\limsup_{t \to \infty} a_t - \limsup_{t \to \infty} b_t| \leq c.$$

Proof: Without loss of generality assume $a := \limsup a_t \geq \limsup b_t =: b$. Choose a subsequence $k(t)$ such that $\lim_{t \to \infty} a_{k(t)} = a$. We obtain

$$\begin{aligned}a - b &\leq a - \limsup_{t \to \infty} b_{k(t)} = \limsup_{t \to \infty} a_{k(t)} - \limsup_{t \to \infty} b_{k(t)} \\ &\leq \limsup_{t \to \infty} |a_{k(t)} - b_{k(t)}| \leq c.\end{aligned}$$

■

We are now in position to prove theorem 3.1.

Proof: As Lipschitz continuity is a local property, we can, without loss of generality, assume that $||P-Q||_{TV} \leq \frac{1}{e}$. Putting $a_t := H^t(P)$ and $b_t := H^t(Q)$ one obtains by lemma 3.1 that

$$|a_t - b_t| \leq ||P-Q||_{TV,t}(\log |\Sigma| + \frac{1}{t}||P-Q||_{TV,t}) =: c_t.$$

One further observes that, because of the definition of $||.||_{TV,t}$ and lemma 2.1, that

$$\begin{aligned}\lim_{t \to \infty} c_t &= \lim_{t \to \infty} ||P-Q||_{TV,t}(\log |\Sigma| + \frac{1}{t}||P-Q||_{TV,t}) \\ &= ||P-Q||_{TV} \cdot \log |\Sigma|.\end{aligned}$$

Plugging $(a_t), (b_t)$ and (c_t) into lemma 3.2 then yields the desired result. ■

We conclusively remark that the idea of the proof depends to a certain extent on the choice of the norm. To demonstrate this we rephrase lemma 3.1 in a more general fashion, without the "soul" of an entropy. Therefore let $\mathbb{R}^n_+ := \{x = (x_1, ..., x_n) \in \mathbb{R}^n \mid x_i \geq 0\}$ and

$$\begin{array}{rccc} h_n: & \mathbb{R}^n_+ & \longrightarrow & \mathbb{R} \\ & x = (x_1, ..., x_n) & \mapsto & \frac{1}{\log n} \sum_{i=1}^n x_i \log \frac{1}{x_i}, \end{array}$$

where $n \geq 2$ and $0 \log \infty := 0$. A more prosaic version of lemma 3.1 then reads

$$|h_n(x) - h_n(y)| \leq ||x - y||_1 \cdot (1 + \frac{1}{\log\, n} \log \frac{1}{||x - y||_1}),$$

where $||x||_1 = \sum_i |x_i|$ as usual. An straightforward consequence of the lemma now is that

$$\forall \epsilon \in \mathbb{R}_+ \; \exists \delta \in \mathbb{R}_+ \; \forall n \geq 2 \; \forall x, y \in \Delta^{n-1}: \\ ||x - y||_1 < \delta \implies |h_n(x) - h_n(y)| < \epsilon, \quad (8)$$

which, when translated back to entropies, tells that the entropy rates are uniformly continuous on $\mathcal{P}^+$. We now note that the statement of the generalized lemma need not be true relative to norms $||.||_p$ different from $||.||_1$. More formally:

Lemma 3.3: Let $2 \leq p < \infty$ and $||x||_p = \sqrt[p]{\sum_i |x_i|^p}$ the usual p-norm on $\mathbb{R}^n$. Then it holds that

$$\exists \epsilon \in \mathbb{R}_+ \; \forall \delta \in \mathbb{R}_+ \; \exists n \geq 2 \; \exists x, y \in \Delta^{n-1}: \\ ||x - y||_p < \delta, \; |h_n(x) - h_n(y)| > \epsilon$$

which is just the negation of (8).

Proof: Indeed, let $\epsilon = 1/2$ and $\delta \in \mathbb{R}^+$ arbitrary. Choose an $m \in \mathbb{N}$, such that $m > \frac{1}{\delta}$. Then find an $N_0 > m$, such that for every $N \geq N_0$ on Δ_{N-1}

$$||(\frac{1}{N}, ..., \frac{1}{N})||_2 = \sqrt{\frac{1}{N}} < \delta.$$

It follows that

$$||(\underbrace{\frac{1}{m}, ..., \frac{1}{m}}_{m times}, 0, ..., 0) - (\frac{1}{N}, ..., \frac{1}{N})||_p \leq ||(\frac{1}{N}, ..., \frac{1}{N})||_p \\ = \sqrt[p]{\frac{1}{N^{p-1}}} = N^{-\frac{p-1}{p}} = N^{\frac{1}{p}-1} \\ \leq N^{-\frac{1}{2}} = ||(\frac{1}{N}, ..., \frac{1}{N})||_2 < \delta,$$

but

$$|h_N(\underbrace{\frac{1}{m}, ..., \frac{1}{m}}_{m\ times}, 0, ..., 0) - h_N(\frac{1}{N}, ..., \frac{1}{N})| \\ = \frac{1}{\log\, N} |\log\, m - \log\, N| \underset{N \to \infty}{\longrightarrow} 1.$$

We therefore find an $N \in \mathbb{N}$ and suitable $x, y \in \Delta_{N-1}$ which support the statement of the lemma. ∎

One could thus intuitively be led to the assumption that entropy rates need not be continuous with respect to other natural norms as, say, the norms given through the spaces $L_p(\Omega, \mathcal{B}, P)$ (definition see below). See, however, the discussion in section IV-A.

B. Differentiability

Some corollaries of the theorem above will now shed light on the issue of differentiability of entropy rates.

Corollary 3.1: Let $V \subset \mathcal{P}$ be a finite-dimensional linear subspace of $\mathcal{P}$. Let $\mathcal{B}(V)$ be the Borel-σ-algebra on V and λ the Lebesgue measure on $(V, \mathcal{B}(V))$. Then the functional

$$\overline{\mathbf{H}}: \; V \longrightarrow \mathbb{R}$$

is differentiable almost everwhere (w.r.t. λ) on V.

Proof: Identify V with an $\mathbb{R}^n$. Rademacher's theorem ([7]) tells that Lipschitzian functionals on an $\mathbb{R}^n$ are differentiable λ-a.e. This is precisely what yields the corollary. ∎

Finally, one sees that the entropy rates of a finite-dimensional set of probability measures can be approximated by polynomials.

Corollary 3.2: Let $K \subset \mathcal{P}$ a finite-dimensional compact subset (e.g. the subset of probability measures of a finite-dimensional subspace of $\mathcal{P}$). Then

$$\overline{\mathbf{H}}: \; K \longrightarrow \mathbb{R}$$

can be uniformly approximated by polynomials.

Proof: Identify K with a compact subset of $\mathbb{R}^n$. The theorem of Stone-Weierstrass ([23]) tells exactly what yields the result. ∎

REMARK. We conjecture that the differentiability of the entropy rates can be proved for the whole space $\mathcal{P}$ although the proof scheme from above cannot be exploited. However, we have not succeeded in constructing a process at which the entropy rate is not differentiable.

IV. ENTROPY RATES OF SIGNED MEASURES AND OF DYNAMICAL SYSTEMS

In this section we would like to analyze analytic properties of more general definitions of entropy rates. Before it comes to the details we would like to describe how to uncouple the entropy rates' definition from the definition of a random source and to extend it to the whole space $\mathcal{P}$ of finite, signed measures.

Definition 4.1: Let $P \in \mathcal{P}$ be a finite, signed measure. Let P_+ resp. P_- be its positive resp. negative variation. Then we define the entropy rate $\overline{\mathbf{H}}(P)$ of P to be

$$\overline{\mathbf{H}}(P) := \overline{\mathbf{H}}(P_+) - \overline{\mathbf{H}}(P_-).$$

We straightforwardly obtain

Corollary 4.1: Entropy rate viewed as a functional on the whole linear space $\mathcal{P}(\Sigma)$ is Lipschitzian with $Lip(\overline{\mathbf{H}}) = 2 \log\, |\Sigma|$.

Proof: Let $P, Q \in \mathcal{P}(\Sigma)$ be two finite signed measures. By the properties of the Jordan decomposition one obtains

$$||P_+ - Q_+||_{TV}, ||P_- - Q_-||_{TV} \quad \leq \quad ||P - Q||_{TV}.$$

Thus we have

$$\begin{aligned}|\overline{\mathbf{H}}(P) - \overline{\mathbf{H}}(Q)| &= |\overline{\mathbf{H}}(P_+) - \overline{\mathbf{H}}(P_-) - \overline{\mathbf{H}}(Q_+) + \overline{\mathbf{H}}(Q_-)| \\ &\leq |\overline{\mathbf{H}}(P_+) - \overline{\mathbf{H}}(Q_+)| + |\overline{\mathbf{H}}(Q_-) - \overline{\mathbf{H}}(P_-)| \\ &\leq \log|\Sigma|(||P_+ - Q_+||_{TV} + ||Q_- - P_-||_{TV}) \\ &\leq 2\log|\Sigma| \cdot ||P - Q||_{TV}.\end{aligned}$$

■

We will exploit this in the following subsections.

A. Entropy rates of signed measures and on $L_p(\Omega, \mathcal{B}, P)$

Let P be a measure on $(\Omega, \mathcal{B})$ and $L_1(\Omega, \mathcal{B}, P)$ be the space of P-integrable functions f (uniquely determined up to P-nullsets) equipped with the norm

$$||f||_1 := \int |f|\, dP.$$

If Q is a finite, signed measure, let $Q = Q_+ - Q_-$ its Jordan decomposition and $|Q| = Q_+ + Q_-$ its total variation (see section II). We write $P_1 << P_2$ if

$$P_2(B) = 0 \quad \Longrightarrow \quad P_1(B) = 0$$

for $B \in \mathcal{B}$ and say that P_1 is absolutely continuous w.r.t. P_2 as usual. The theorem of Radon-Nikodym tells us that the vector space of P-integrable functions is isomorphic to the subspace

$$\mathcal{P}_P := \{Q \in \mathcal{P} \mid |Q| << P\}$$

that is to the subspace of signed measures whose total variation is absolutely continuous w.r.t. P. More concretely, the isomorphism $\Phi : L_1(\Omega, \mathcal{B}, P) \to \mathcal{P}_P$ is described by the rule

$$\Phi(f)(B) = \int_B f\, dP$$

for $B \in \mathcal{B}$, and the inverse of Φ is given by

$$\begin{array}{ccc} \mathcal{P}_P & \longrightarrow & L_1 \\ Q & \mapsto & \frac{dQ_+}{dP} - \frac{dQ_-}{dP}, \end{array}$$

where $\frac{dP^*}{dP}$ is the Radon-Nikodym derivative of a measure P^*, for which $P^* << P$. Observe further that

$$||\Phi(f)||_{TV} = ||f||_1,$$

which makes Φ an isometry of normed spaces. Thus, the entropy rates $\overline{\mathbf{H}}$, viewed as a functional on L_1 by the law

$$\overline{\mathbf{H}}(f) := \overline{\mathbf{H}}(\Phi(f))$$

are, by theorem 3.1 Lipschitz continuous on $L_1(\Omega, \mathcal{B}, P)$ for all measures P.

It is now an interesting question to ask whether the entropy rates are also continuous considered as a functional on the normed spaces $L_p(\Omega, \mathcal{B}, P)$ for arbitrary $p \in]1, \infty[$, where the norms are given by

$$||f||_p := \int |f|^p\, dP.$$

This proves indeed to be true, which is quickly seen by a look at the subsequent lemma.

Lemma 4.1 ([23]): Let $1 \leq p < q < \infty$ and $P \in \mathcal{P}^+$ a measure. Then it holds that

$$||f||_p \leq P(\Omega)^{\frac{1}{p} - \frac{1}{q}} ||f||_q.$$

Note first that this makes $L_p(\Omega, \mathcal{B}, P)$ a subspace (as a vector space only) of $L_1(\Omega, \mathcal{B}, P)$ for all $p > 1$. We can therefore consider the entropy rate to be a functional on $L_p(\Omega, \mathcal{B}, P)$ as described above for L_1. Thus we obtain the following corollary.

Corollary 4.2: Let $P \in \mathcal{P}^+$ a measure. The entropy rate functional $\overline{\mathbf{H}}$ on $L_p(\Omega, \mathcal{B}, P)$ is Lipschitzian with

$$Lip(\overline{\mathbf{H}}) = 2 \cdot P(\Omega)^{1 - \frac{1}{p}} \cdot \log|\Sigma|.$$

Proof:

$$\begin{aligned}|\overline{\mathbf{H}}(f) - \overline{\mathbf{H}}(g)| &\overset{C.\ 4.1}{\leq} 2 \cdot \log|\Sigma| \cdot ||f - g||_1 \\ &\overset{L.\ 4.1}{\leq} 2 \cdot \log|\Sigma| \cdot P(\Omega)^{1 - \frac{1}{p}} ||f - g||_p.\end{aligned}$$

■

B. General dynamical systems

Let $(\Omega, \mathcal{B}, P, T)$ be a dynamical system, that is, in the most general definition a measure space $(\Omega, \mathcal{B}, P)$ with a measurable function $T : \Omega \to \Omega$. Note that, in this section, $(\Omega, \mathcal{B})$ is an **arbitrary** measurable space. The entropy rates of dynamical systems are defined in three steps [2]. We first define the entropy of a finite subfield $\mathcal{A}$ of $\mathcal{B}$ by

$$H(\mathcal{A}, P) := -\sum_{A \in \mathcal{A}} P(A) \log P(A).$$

Now let $T^{-n}\mathcal{A} := \{T^{-n}A \mid A \in \mathcal{A}\}$ and further $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{A}$ be the finite subfield of $\mathcal{B}$, which consists of all intersections of elements of the $T^{-i}\mathcal{A}$, $i = 0, ..., n-1$. The entropy of a finite field $\mathcal{A}$ relative to T is then given by

$$H(\mathcal{A}, P, T) := \limsup_{n \to \infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{A}).$$

Finally the entropy of the dynamical system is

$$\overline{\mathbf{H}}(P, T) := \sup_{\mathcal{A}} H(\mathcal{A}, T).$$

where the supremum ranges over all finite subfields $\mathcal{A}$ of $\mathcal{B}$. We moreover introduce the *k-th order entropy rate* of a dynamical system

$$\overline{\mathbf{H}}(P, T, k) := \sup_{|\mathcal{A}| = k} H(\mathcal{A}, T)$$

where here the supremum ranges only over the finite subfields $\mathcal{A}$ with k elements. As entropy increases by splitting up events we have

$$\overline{\mathbf{H}}(P, T, k) \leq \overline{\mathbf{H}}(P, T, k+1)$$

for all $k \in \mathbb{N}$. Consider now $\overline{\mathbf{H}}(P, T, k)$ as a functional of the probability measure of P. It is then a byproduct of the investigations from above that this functional is continuous with respect to the norm of total variation.

Theorem 4.1: Let $(\Omega, \mathcal{B})$ be a measurable space and $T : \Omega \to \Omega$ be a measurable function. The k-th order entropy rate

$$\begin{aligned}\overline{\mathbf{H}}(.,T,k): \quad \mathcal{P}^+(\Omega,\mathcal{B}) &\longrightarrow \quad \mathbb{R} \\ P &\mapsto \quad \overline{\mathbf{H}}(P,T,k)\end{aligned}$$

is continuous relative to the topology induced by the norm of total variation.

Proof: Let (P_n) a sequence of probability measures, which converges to a probability measure $\bar{P}$. We have from the above theorem 3.1 that $H(\mathcal{A}, P, T)$ is Lipschitzian with a Lipschitz constant $\log |\mathcal{A}|$ where $|\mathcal{A}|$ is the number of elements of $\mathcal{A}$. Let $k \in \mathbb{N}$ and $\epsilon \in \mathbb{R}_+$. There is a finite subfield $\mathcal{F}$ with $|\mathcal{F}| = k$ such that $H(\bar{P}, T, k) - H(\mathcal{F}, \bar{P}, T) \leq \epsilon$. We obtain

$$\begin{aligned}\liminf_{n\to\infty} \overline{\mathbf{H}}(P_n, T, k) &\geq \liminf_{n\to\infty} H(\mathcal{F}, P_n, T) \\ &= \lim_{n\to\infty} H(\mathcal{F}, P_n, T) \\ &\overset{(*)}{=} H(\mathcal{F}, \bar{P}, T) \geq \overline{\mathbf{H}}(\bar{P}, T) - \epsilon.\end{aligned}$$

where $(*)$ follows from the continuity of $H(\mathcal{F}, P, T)$. As ϵ was arbitrary we obtain

$$\liminf_{n\to\infty} \overline{\mathbf{H}}(P_n, T, k) \geq \overline{\mathbf{H}}(\bar{P}, T, k).$$

It remains to show

$$\limsup_{n\to\infty} \overline{\mathbf{H}}(P_n, T, k) \leq \overline{\mathbf{H}}(\bar{P}, T, k)$$

to finish the proof. We assume the contrary and let

$$\delta := \limsup_{n\to\infty} \overline{\mathbf{H}}(P_n, T, k) - \overline{\mathbf{H}}(\bar{P}, T, k) > 0.$$

We first find a subsequence $l(n)$ such that $\lim_{n\to\infty} \overline{\mathbf{H}}(P_{l(n)}, T, k)$ equals the limit superior of $\overline{\mathbf{H}}(P_n, T, k)$. With it we find a $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$

$$\overline{\mathbf{H}}(P_{l(n)}, T, k) > \overline{\mathbf{H}}(\bar{P}, T, k) + \frac{2\delta}{3}. \tag{9}$$

Second we find a $N_1 \in \mathbb{N}$, which, without loss of generality, can be assumed to be greater than N_0 such that

$$||P_{l(n)} - \bar{P}||_{TV} < \frac{\delta}{3 \log k}$$

and hence, by theorem 3.1

$$|H(\mathcal{A}, P_{l(n)}, T) - H(\mathcal{A}, \bar{P}, T)| < \frac{\delta}{3} \tag{10}$$

for all finite subfields $\mathcal{A}$ with $|\mathcal{A}| = k$. Last we find a finite subfield $\mathcal{F}$ with $|\mathcal{F}| = k$ such that

$$H(\mathcal{F}, P_{l(N_1)}, T) > \overline{\mathbf{H}}(P_{l(N_1)}, T, k) - \frac{\delta}{3}. \tag{11}$$

In sum we obtain

$$\begin{aligned}H(\mathcal{F}, \bar{P}, T) &\overset{(10)}{>} H(\mathcal{F}, P_{l(N_1)}, T) - \frac{\delta}{3} \\ &\overset{(11)}{>} \overline{\mathbf{H}}(P_{l(N_1)}, T, k) - \frac{2\delta}{3} \overset{(9)}{>} \overline{\mathbf{H}}(\bar{P}, T, k),\end{aligned}$$

which is a contradiction to the definition of $\overline{\mathbf{H}}(\bar{P}, T, k)$. ∎

As a corollary we can tell something about the behaviour of the entropy rates of dynamical systems themselves.

Corollary 4.3: Let $(\Omega, \mathcal{B})$ be a measurable space and $T : \Omega \to \Omega$ be a measurable function. The entropy rate

$$\begin{aligned}\overline{\mathbf{H}}(.,T): \quad \mathcal{P}^+(\Omega,\mathcal{B}) &\longrightarrow \quad \mathbb{R} \\ P &\mapsto \quad \overline{\mathbf{H}}(P,T)\end{aligned}$$

is lower semicontinuous relative to the topology induced by the norm of total variation.

Proof: Let P_n be a sequence of probability measures that converges in norm of total variation to a measure $\bar{P}$. We have to show that

$$\liminf_{n\to\infty} \overline{\mathbf{H}}(P_n, T) \geq \overline{\mathbf{H}}(\bar{P}, T).$$

Assume the contrary so that we find a subsequence $l(n)$ with

$$\lim_{n\to\infty} \overline{\mathbf{H}}(P_{l(n)}, T) = \overline{\mathbf{H}}(\bar{P}, T) - \delta \tag{12}$$

with $\delta > 0$. Because of $\lim_{k\to\infty} \overline{\mathbf{H}}(\bar{P}, T, k) = \overline{\mathbf{H}}(\bar{P}, T)$ we find a $K_0 \in \mathbb{N}$ such that for all $k \geq K_0$ we have

$$\overline{\mathbf{H}}(\bar{P}, T, k) > \overline{\mathbf{H}}(\bar{P}, T) - \frac{\delta}{3}. \tag{13}$$

Because of (12) we also find a $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$

$$\overline{\mathbf{H}}(P_{l(n)}, T) < \overline{\mathbf{H}}(\bar{P}, T) - \frac{2}{3}\delta. \tag{14}$$

Because of theorem 4.1 we find a $N_1 \in \mathbb{N}$, which, without loss of generality can be chosen greater than N_0 such that for all $n \geq N_1$

$$\overline{\mathbf{H}}(P_{l(n)}, T, K_0) > \overline{\mathbf{H}}(\bar{P}, T, K_0) - \frac{\delta}{3}. \tag{15}$$

Taken altogether we find that

$$\begin{aligned}\overline{\mathbf{H}}(P_{l(N_1)}, T) \geq \overline{\mathbf{H}}(P_{l(N_1)}, T, K_0) &\overset{(15)}{>} \overline{\mathbf{H}}(\bar{P}, T, K_0) - \frac{\delta}{3} \\ &\overset{(13)}{>} \overline{\mathbf{H}}(\bar{P}, T) - \frac{2}{3}\delta\end{aligned}$$

where the first inequality follows from the monotonicity of $\overline{\mathbf{H}}(P, T, k)$ in k. The resulting inequality, however, is a contradiction to (14). ∎

V. Entropy rates and asymptotic mean stationarity

Let $(\Omega, \mathcal{B}, P, T)$ be a dynamical system as explained above, that is, $(\Omega, \mathcal{B}, P)$ is a probability space and $T : \Omega \to \Omega$ is a measurable function. The dynamical system is called *stationary*, if

$$\forall B \in \mathcal{B}: \quad P(B) = P(T^{-1}B)$$

and *asymptotically mean stationary (AMS)*, if there is a measure $\bar{P}$ on $(\Omega, \mathcal{B})$ such that

$$\forall B \in \mathcal{B}: \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(T^{-i}B) = \bar{P}(B).$$

It is easily seen that the resulting dynamical system $(\Omega, \mathcal{B}, \bar{P}, T)$ is stationary. $\bar{P}$ is therefore called the *stationary*

mean of the system. Asymptotically mean stationary systems were an area of active research mostly in the late 80's [9], [18], [11]. With the help of Birkhoff's famous ergodic theorem [17], [19], [26] it can be shown that asymptotic mean stationarity is equivalent to the existence of ergodic properties with respect to bounded measurements [11].

In this section we consider the dynamical systems $(\Omega, \mathcal{B}, P, T)$, which are (canonically) associated to discrete random sources (X_t), that is, $(\Omega, \mathcal{B})$ is the sequence space from the introduction, P is the probability measure relative to (X_t) and $T : \Omega \to \Omega$ is the shift operator, that is, $(T\omega)_t = \omega_{t+1}$. Consequently, we call a random source (asymptotically mean) stationary if its dynamical system is.

In the following we will show how one can obtain a direct proof for the existence of entropy rates for AMS random sources by means of theorem 3.1. Note that it is a corollary of the famous theorem of Shannon-McMillan-Breiman in its most general form, that the entropy rates of this class of random sources exist. However, the proof given here is much more elementary. See the final remarks of this section for a more detailed comparison of the two proofs.

A. Proof for the existence of entropy rates of AMS sources

Let (X_t) be a discrete random source with values in Σ and $(\Omega, \mathcal{B}, P_X, T)$ the associated dynamical system as described above. We will again write

$$\begin{aligned} H^t(P_X) &:= H^t(X) := H(X_1 ... X_t) \\ &:= \sum_{\bar{a} \in \Sigma^t} P_X(\bar{a}) \log \frac{1}{P_X(\bar{a})}. \end{aligned}$$

Remember that $\overline{\mathbf{H}}(X)$ was defined to be $\limsup_t \frac{1}{t} H^t(X)$. The following lemma will show that the entropy rates of P_X and $P_X \circ T^{-k}$ coincide for all k, which is a well known result.

Lemma 5.1: Let (X_t) be a discrete random source and $(\Omega, \mathcal{B}, P_X, T)$ be the associated dynamical system. Then it holds that

$$\lim_{t \to \infty} \frac{1}{t} (H^t(P_X) - H^t(P_X \circ T^{-k})) = 0.$$

Proof: Using the notation ($\bar{a}\bar{b} \in \Sigma^{t+k}$ is the concatenation of the words $\bar{a} \in \Sigma^k, \bar{b} \in \Sigma^t$)

$$I_t^k(X) := I_t^k(P_X) := \sum_{\bar{a} \in \Sigma^k} \sum_{\bar{b} \in \Sigma^t} P_X(\{\bar{a}\bar{b}\}) \log \frac{P_X(T^{-k}\{\bar{b}\})}{P_X(\{\bar{a}\bar{b}\})}$$

and

$$J_t^k(X) := J_t^k(P_X) := \frac{1}{t} \sum_{\bar{a} \in \Sigma^k} \sum_{\bar{b} \in \Sigma^t} P_X(\{\bar{a}\bar{b}\}) \log \frac{P_X(\{\bar{a}\})}{P_X(\{\bar{a}\bar{b}\})}$$

one obtains

$$\begin{aligned} \frac{1}{t}(H^t(P_X) + J_t^k(X)) &\overset{(*)}{=} \frac{1}{t} H^{k+t}(X) \\ &= \frac{1}{t}(I_n^k(X) + H^t(P_X \circ T^{-k})) \end{aligned}$$

where $(*)$ follows from a well known and elementary theorem (e.g. [13], p.22, theorem 2.1) and the second equation is obvious. Because of

$$\begin{aligned} 0 \leq \frac{1}{t} J_t^k(X) &\leq \frac{1}{t} H^k(P_X \circ T^{-t}) \\ &\leq \frac{1}{t} \log \, card(\Sigma)^k \longrightarrow_{t \to \infty} 0 \end{aligned}$$

and

$$0 \leq \frac{1}{t} I_t^k(X) \leq \frac{1}{t} H^k(X) \leq \frac{1}{t} \log \, card(\Sigma)^k \longrightarrow_{t \to \infty} 0,$$

the assertion follows from an application of the sandwich theorem. ∎

In the following we will write

$$P_n := \frac{1}{n} \sum_{i=0}^{n-1} P_X \circ T^{-i}.$$

Corollary 5.1: Let (X_t) be a random source and $(\Omega, \mathcal{B}, P_X, T)$ be the associated dynamical system. Then it holds that

$$\forall n \in \mathbb{N} : \quad \lim_{t \to \infty} \frac{1}{t} (H^t(P_X) - H^t(P_n)) = 0.$$

Proof: This follows from the equation

$$\begin{aligned} \frac{1}{n} \sum_{i=0}^{n-1} H^t(P_X \circ T^{-i}) &\leq H^t(P_n) \\ &\leq \frac{1}{n} \sum_{i=0}^{n-1} H^t(P_X \circ T^{-i}) + \log n, \end{aligned}$$

which can be seen from [10], Lemma 2.3.4 and induction on n. ∎

We now define

$$\begin{array}{rccc} \underline{\mathbf{H}} : & \mathcal{P}^+(\Sigma) & \longrightarrow & \mathbb{R} \\ & P & \mapsto & \liminf_{t \in \mathbb{N}} \frac{1}{t} \sum_{\bar{a} \in \Sigma^t} P(\bar{a}) \log \frac{1}{P(\bar{a})} \end{array}.$$

One obtains the same analytic properties for $\underline{\mathbf{H}}$ by rephrasing lemma 3.2 with $\liminf$ instead of $\limsup$.

As a corollary we now obtain that $\overline{\mathbf{H}}$ as well as $\underline{\mathbf{H}}$ of P_X and the P_n coincide.

Corollary 5.2: Let (X_t) be a discrete random source and $(\Omega, \mathcal{B}, P_X, T)$ be the associated dynamical system. Then it holds that

$$\begin{aligned} \overline{\mathbf{H}}(P_X) &= \overline{\mathbf{H}}(P_n), \\ \underline{\mathbf{H}}(P_X) &= \underline{\mathbf{H}}(P_n) \end{aligned}$$

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

Proof: Use corollary 5.1 in order to apply lemma 3.2 to the sequences $(a_t := \frac{1}{t} H^t(P_X)), (b_t := \frac{1}{t} H^t(P_X \circ T^{-k}))$ for the first equation. For the second one rephrase lemma 3.2 with $\liminf$ instead of $\limsup$. ∎

The key to the proof for the existence of the entropy rates is now that the convergence involved in the definition of asymptotic mean stationarity respects the norm of total variation.

Theorem 5.1: Let $(\Omega, \mathcal{B}, P, T)$ be an AMS dynamical system and $\bar{P}$ the stationary mean. Write $P_n := \frac{1}{n}\sum_{i=0}^{n-1} P \circ T^{-i}$. Then it holds that

$$\lim_{k\to\infty} ||P_n - \bar{P}||_{TV} = 0.$$

We will not prove the theorem here, but rather tell that the theorem is a consequence of the theorems developed in [9]. Alternatively, in a more direct fashion, one can prove the theorem by means of Krengel's ergodic theorem, see [27]. We finally remark that the convergence on the sets of the underlying σ-algebra involved in the definition of asymptotic mean stationarity is usually referred to as *strong* convergence whereas convergence in norm of total variation is also referred to as *Skorokhod weak* convergence. Skorokhod weak convergence implies strong convergence as it can be translated to uniform convergence on the sets of the underlying σ-algebra, see lemma 2.2. For the above theorem however, we need the inverse direction.

Now we can prove the existence of entropy rates for AMS sources.

Theorem 5.2: Let (X_t) be a discrete AMS random source and $(\Omega, \mathcal{B}, P_X, T)$ be the associated dynamical system. Let $\bar{P}$ be the stationary mean of the system. Then it holds that

$$\overline{\mathbf{H}}(\bar{P}) = \overline{\mathbf{H}}(P_X) = \lim_{t\to\infty} \frac{1}{t} H^t(P_X),$$

that is the entropy rate of P_X exists and is equal to that of the stationary mean $\bar{P}$.

Proof: As the P_n converge in TV-norm to $\bar{P}$ (theorem 5.1) we obtain, because of the continuity of $\overline{\mathbf{H}}, \underline{\mathbf{H}}$ (theorem 3.1), that

$$\lim_{n\to\infty} \overline{\mathbf{H}}(P_n) = \overline{\mathbf{H}}(\bar{P}) \quad \text{and} \quad \lim_{n\to\infty} \underline{\mathbf{H}}(P_n) = \underline{\mathbf{H}}(\bar{P}).$$

It follows, as $\overline{\mathbf{H}}(P_n)$ and $\underline{\mathbf{H}}(P_n)$ are constant with respect to n (corollary 5.2) and $\overline{\mathbf{H}}(\bar{P}) = \underline{\mathbf{H}}(\bar{P})$ (as the entropy rates of stationary sources exist) that $\overline{\mathbf{H}}(P_X) = \underline{\mathbf{H}}(P_X) = \overline{\mathbf{H}}(\bar{P})$. ■

FINAL REMARK: As mentioned above, the result is usually obtained as a corollary of the theorem of Shannon-McMillan-Breiman [28], [22], [3], [4], which was iteratively extended [2], [14], [15] to finally hold for AMS sources in 1980 [9], [10]. The final result, however, is centered around a difficult proof for the class of stationary random sources. This proof is split up into two parts. One first shows the result for the class of ergodic, stationary sources, which, in turn, requires involved ergodic theorems. The extension to general stationary sources then needs the brilliant, but sophisticated concept of the ergodic decomposition of stationary random sources [8]. The proof given here is thus simpler from a range of aspects as it is based on the comparatively tiny proof for the existence of entropy rates of stationary sources only. Note that this way we do not even need the concept of ergodicity. It is also more elementary as the involved theorem 3.1 can be obtained by means of basic calculus alone. Only theorem 5.1 seems to require an ergodic theorem. Finally note that in [5] it was shown that the class of finite-dimensional random sources, which includes the hidden Markov sources, is AMS. We therefore would like to point out that the entropy rates for the whole class of hidden Markov sources exist, which seems to be widely unknown.

VI. CONCLUSION

The analyses presented in this article helped get a more general grip of the analytic properties of the entropy rates of discrete random sources. This may serve as an orientation when dealing with entropy rates of more special classes of random sources. We also have given rise to a range of open problems. For example, it would be interesting to know to what extent our arguments can be strengthened. To put it more concrete we raise two exemplary questions. First, do entropy rates have stronger analytic properties than Lipschitz continuity, say, differentiability? Second, do entropy rates also have nice analytic properties when considering coarser topologies than that of total variation? It is known that the subclass of stationary random sources is upper semicontinuous relative to the weak topology [10]. We believe that analogous results can be obtained on the weak topology for the subclass of "finite-dimensional" random sources [5], a class encompassing the hidden Markov models. As an additional benefit, this would, highly probably, yield new existence proofs in the style of that of section V.

APPENDIX A
PROOF OF LEMMA 2.1

For the proof, we identify, as usual, cylinder sets $B \in \mathcal{B}$ with sets of words $A_B \in \Sigma^t$ (B is the set of sequences which are the continuations of the words in A_B). In our notation, we correspondingly obtain

$$P(B) = \sum_{\bar{a}\in A_B} P(\bar{a}) \tag{16}$$

for a measure P resp.

$$P(B) = \sum_{\bar{a}\in A_B} P_+(\bar{a}) - P_-(\bar{a}) \tag{17}$$

for a signed measure P with Jordan decomposition $P = P_+ - P_-$.

The approximation theorem (see Halmos [12], p. 56, Th. D) tells that, given a measure P, an event $B \in \mathcal{B}$ and $\epsilon \in \mathbb{R}_+$, we will find a cylinder set F such that

$$P(B \triangle F) < \epsilon,$$

where $B \triangle F = (B \setminus F) \cup (F \setminus B)$ is the symmetric set difference. A straightforward consequence of this is that $|P(B) - P(F)| < \epsilon$.

Proof: The second equation follows immediately from

$$\begin{aligned}\sum_{\bar{a}\in\Sigma^t} |P(\bar{a})| &= \sum_{\bar{a}\in\Sigma^t} \underbrace{|\sum_{a\in\Sigma} P(\bar{a}a)|}_{=|P(\bar{a})|} \\ &\leq \sum_{\bar{a}\in\Sigma^t}\sum_{a\in\Sigma} |P(\bar{a}a)| = \sum_{\bar{a}\in\Sigma^{t+1}} |P(\bar{a})|\end{aligned}$$

which shows that $(\sum_{\bar{a}\in\Sigma^t} |P(\bar{a})|)_{t\in\mathbb{N}}$ is a monotonically increasing sequence. It remains to show that it converges to

$||P||_{TV}$. In the given situation, this translates to demonstrate that, given $\epsilon \in \mathbb{R}_+$, there is $T_0 \in \mathbb{N}$ with

$$\sum_{\bar{a}\in\Sigma^{T_0}} |P(\bar{a})| > ||P||_{TV} - \epsilon.$$

Therefore let P_+, P_- be the Jordan decomposition of P and, correspondingly, $\Omega = \Omega_+ \dot\cup \Omega_-$ be the Hahn decomposition. As a consequence of the approximation theorem (see above) we find $T_0 \in \mathbb{N}$ and a cylinder set corresponding to $A \subset \Sigma^{T_0}$ with

$$|P|(\Omega_+ \triangle A) < \frac{\epsilon}{4} \tag{18}$$

a straightforward ($|P| = P_+ + P_-$) consequence of which is that both

$$P_+(\Omega_+ \triangle A) < \frac{\epsilon}{4} \quad \text{and} \quad P_-(\Omega_+ \triangle A) < \frac{\epsilon}{4} \tag{19}$$

Now note that the general $\complement A \triangle \complement B = A \triangle B$ in combination with $\Omega_- = \complement\Omega_+$ and (19) yields

$$P_-(\Omega_- \triangle \complement A) = P(\Omega_+ \triangle A) < \frac{\epsilon}{4}. \tag{20}$$

(19) and (20) then yield the inequalities

$$P_+(\complement A) \overset{P_+(\Omega_-)=0}{=} P_+(\Omega_+ \setminus A) \leq P_+(\Omega_+ \triangle A) < \frac{\epsilon}{4} \tag{21}$$

and

$$P_-(A) \overset{P_-(\Omega_+)=0}{=} P_-(\Omega_- \setminus \complement A) \leq P_-(\Omega_- \triangle \complement A) < \frac{\epsilon}{4}. \tag{22}$$

Moreover, it is straightforward from (19) and (20) that

$$P_+(A) > P_+(\Omega_+) - \frac{\epsilon}{4} \tag{23}$$

as well as

$$P_-(\complement A) > P_-(\Omega_-) - \frac{\epsilon}{4}. \tag{24}$$

We finally compute

$$\begin{aligned}
\sum_{\bar{a}\in\Sigma^{T_0}} |P(\bar{a})| &= \sum_{\bar{a}\in A} |P(\bar{a})| + \sum_{\bar{a}\in\complement A} |P(\bar{a})| \\
&\geq |\sum_{\bar{a}\in A} P(\bar{a})| + |\sum_{\bar{a}\in\complement A} P(\bar{a})| \\
&= |P(A)| + |P(\complement A)| \\
&= |P_+(A) - P_-(A)| + |P_+(\complement A) - P_-(\complement A)| \\
&\geq P_+(A) - P_-(A) + P_-(\complement A) - P_+(\complement A) \\
&\overset{(21),(22)}{\underset{(23),(24)}{>}} (P_+(\Omega_+) - \frac{\epsilon}{4}) - \frac{\epsilon}{4} + (P_-(\Omega_-) - \frac{\epsilon}{4}) - \frac{\epsilon}{4} \\
&= P_+(\Omega_+) + P_-(\Omega_-) - \epsilon = ||P||_{TV} - \epsilon.
\end{aligned}$$

■

APPENDIX B
PROOF OF LEMMA 2.2

Proof: We first show the first inequality. Therefore let B be a cylinder set, that is, $B \subset \Sigma^t$ for a t. It follows that

$$\begin{aligned}
|P(B) - Q(B)| &= |\sum_{\bar{a}\in B} P(\bar{a}) - \sum_{\bar{a}\in B} Q(\bar{a})| \\
&\leq \sum_{\bar{a}\in B} |P(\bar{a}) - Q(\bar{a})| \\
&\leq \sum_{\bar{a}\in\Sigma^t} |P(\bar{a}) - Q(\bar{a})| \leq ||P - Q||_{TV}.
\end{aligned} \tag{25}$$

Let now $B \in \mathcal{B}$ be an arbitrary event. Because of the approximation theorem (see beginning of appendix A) there is a sequence of cylinder sets $(B_k)_{k\in\mathbb{N}}$ with

$$\lim_{k\to\infty} |P - Q|(B_k \triangle B) = 0.$$

A straightforward consequence is $\lim_{k\to\infty} |P - Q|(B_k) = (P - Q)(B)$ and therefore

$$\lim_{k\to\infty} |P(B_k) - Q(B_k)| = |P(B) - Q(B)|.$$

As $|P(B_k) - Q(B_k)| \leq ||P - Q||_{TV}$, also $|P(B) - Q(B)| \leq ||P - Q||_{TV}$.

Let now $\epsilon \in \mathbb{R}^+$. For the second inequality we have to prove the existence of a set $B \in \mathcal{B}$, such that

$$|P(B) - Q(B)| > \frac{1}{2}||P - Q||_{TV} - \epsilon.$$

First, because of lemma 2.1, we find $t \in \mathbb{N}$ such that

$$\sum_{\bar{a}\in\Sigma^t} |P(\bar{a}) - Q(\bar{a})| > ||P - Q||_{TV} - 2\epsilon.$$

We put

$$\begin{aligned}
A_1 &:= \{\bar{a} \in \Sigma^t \mid P(\bar{a}) > Q(\bar{a})\}, \\
A_2 &:= \{\bar{a} \in \Sigma^t \mid Q(\bar{a}) > P(\bar{a})\}
\end{aligned}$$

and obtain

$$\begin{aligned}
&(P(A_1) - Q(A_1)) + (P(A_2) - Q(A_2)) \\
&= \sum_{\bar{a}\in A_1} (P(\bar{a}) - Q(\bar{a})) + \sum_{\bar{a}\in A_2} (Q(\bar{a}) - P(\bar{a})) \\
&= \sum_{\bar{a}\in\Sigma^t} |P(\bar{a}) - Q(\bar{a})| > ||P - Q||_{TV} - 2\epsilon.
\end{aligned} \tag{26}$$

Hence, for at least one $i \in \{1, 2\}$,

$$|P(A_i) - Q(A_i)| > \frac{1}{2}||P - Q||_{TV} - \epsilon. \tag{27}$$

■

ACKNOWLEDGMENT

The author would like to thank Ulrich Faigle whose suggestions motivated much of the work presented here.

REFERENCES

[1] D. Aharonov, A. Ambainis, J. Kempe, U. Vazirani, "Quantum walks on graphs", in *Proc. of 33rd ACM STOC, New York*, 2001, pp. 50-59.

[2] P. Billingsley, *Ergodic Theory and Information*, Wiley, 1965.

[3] L. Breiman, "The individual ergodic theorem of information theory", in *Annals of Mathematical Statistics*, 1957, vol. 28, pp. 809-811.

[4] L. Breiman, A correction to 'the individual ergodic theorem of information theory'. *Annals of Mathematical Statistics*, 31:809–810, 1960.

[5] U. Faigle, A. Schönhuth, "Quantum predictor models", *Electonic Notes in Discrete Mathematics*, 2006, vol. 25, pp. 149-155.

[6] U. Faigle and A. Schoenhuth, "Asymptotic mean stationarity of sources with finite evolution dimension", *IEEE Trans. Inf. Theory*, 2007, vol. 53(7), pp. 2342-2348.

[7] H. Federer, *Geometric Measure Theory*. Springer, 1969.

[8] R. Gray and L. Davisson, "The ergodic decomposition of stationary discrete random processes", *IEEE Transactions on Information Theory*, 1974, vol. 20(5), pp. 625-636.

[9] R.M. Gray and J.C. Kieffer, "Asymptotically mean stationary measures" *Annals of Probability*, 1980, vol. 8, pp. 962–973.

[10] Robert M. Gray, *Entropy and Information Theory*. Springer, 1990.
[11] Robert M. Gray, *Probability, Random Processes and Ergodic Properties*. Springer, 2001.
[12] P.R. Halmos, *Measure Theory*. Van Nostrand, 1964.
[13] T.S. Han and K. Kobayashi, *Mathematics of Information and Coding*. American Mathematical Society, 2002.
[14] K. Jacobs, “Die bertragung diskreter Informationen durch periodische und fastperiodische Kanle”, 1959, *Mathematische Annalen*, vol. 137, pp. 125-135.
[15] K. Jacobs, “ber die Struktur der mittleren Entropie”, *Mathematisches Zentralblatt*, 1962, vol. 78, pp. 33-43.
[16] P. Jacquet, G. Seroussi, and W. Szpankowski, “On the entropy of a hidden markov process”, in *Proc. Data Compression Conf.*, Snowbird, UT, March 2004, pp. 362-371.
[17] Y. Katznelson and B. Weiss, “A simple proof of some ergodic theorems”, *Israel Journal of Mathematics*, 1982, vol. 42, pp. 291-296.
[18] J.C. Kieffer and M. Rahe, “Markov channels are asymptotically mean stationary”, *SIAM J. Math. Anal.*, 1981, vol. 12(3), pp. 293-305.
[19] U. Krengel, *Ergodic Theorems*. de Gruyter, 1985.
[20] G. Han and B. Marcus, “Analyticity of entropy rate of hidden Markov chains”, *IEEE Trans. Inf. Theory*, 2006, vol. 52(12), pp. 5251-5266.
[21] G. Han and B. Marcus, “Derivatives of entropy rate in special families of hidden Markov chains”, *IEEE Trans. Inf. Theory*, 2007, vol. 53(7), pp. 2642-2652.
[22] B. McMillan, “The basic theorems of information theory”, *Annals of Mathematical Statistics*, 1953, vol. 24, pp. 196-219.
[23] R. Meise and D. Vogt, *Einfhrung in die Funktionalanalysis*. vieweg, 1992.
[24] Y. Ephraim, N. Merhav, ”Hidden Markov processes”, *IEEE Trans. on Information Theory*, vol. 48(6), pp. 1518-1569.
[25] E. Ordentlich and T. Weissman, “On the optimality of symbol by symbol filtering and denoising”, *IEEE Trans. Inf. Theory*, 2006, 52(1), pp. 19-40.
[26] M. Pollicott and M. Yuri, *Dynamical Systems and Ergodic Theory*. Cambridge University Press, 1998.
[27] A. Schönhuth, ”The ergodic decomposition of asymptotically mean stationary dynamical systems”, submitted to *Ergodic Theory and Dynamical Systems*, 2007.
[28] C. Shannon, “A mathematical theory of communication”, *Bell System Technical Journal*, 1948.
[29] O. Zuk, E. Domany, I. Kanter, and M. Aizenman, “From finite system entropy to entropy rate for a hidden markov process”, *IEEE Signal Processing Letters*, 2006, vol. 13(9), pp. 517-520.
[30] O. Zuk, I. Kanter, and E. Domany, “The entropy of a binary hidden markov process”, *Journal of Statistical Physics*, 2005, vol. 121(3-4), pp. 343-360.

Alexander Schönhuth Biography text here.

John Doe Biography text here.