Entropy Estimators and Serial Tests for Ergodic Chains

Stefan Wegenkittl

Abstract— **In his 1992 paper entitled "A universal statistical test for random bit generators", Maurer discusses a statistic whose value is closely related to the per-bit-entropy of an ergodic stationary source. Here, we derive an entropy estimate from a class of generalized serial tests and discuss its relationship to return-time based entropy estimators and frequency-based goodness-of-fit tests. Our setup extends Kullback's I-divergence approach for independent stationary sequences to the class of ergodic Markov chains. The effects caused by the order of the source are examined theoretically and by an empirical study.**

Keywords: **entropy, ergodic chains, serial tests, universal tests**
AMS classification: **94A17 Measures of information, entropy, 62H10 Distribution of statistics, 62H15 Hypothesis testing, 62M02 Markov processes: hypothesis testing (Inference from stochastic processes), 62M05 Markov processes: estimation.**

I. Introduction

INVESTIGATIONS about randomness of given sequences of symbols, as ocurring in the quality assessment of (pseudo-) random number generators for example, mostly require a statistical tool for the comparison of distributions. Two well understood concepts are goodness-of-fit tests and entropy estimates. Whereas the former are usually built upon the occurrence rates of certain events and their comparison with the expected frequencies derived from some null-hypothesis, the latter estimate a property of the underlying common distribution of the symbols in the sequence and may be calculated from relative frequencies or return-times, alternatively.

The two approaches are linked by Kullback's I-divergence which provides for both: goodness-of-fit statistics ("distance to null-hypothesis") and an entropy-related information measure ("distance to equidistribution"). As the original I-divergence is applicable for independent, identically distributed (i.i.d.) sequences only, we develop a modified generalized I-divergence in this paper which applies to the class of finite ergodic Markov chains. Furthermore, we study the relation to return-time based entropy estimates and prove Maurer's [1] conjecture on the speed of convergence of the later.

The paper is structured as follows: After introducing some notation we review the i.i.d. case in Section II. In Section III we introduce the (modified) generalized ϕ-divergence $\tilde{I}_{\Sigma,\varphi,q}$ and give an example of its application with respect to goodness-of-fit hypothesis-testing. A linear transform of $\tilde{I}_{\Sigma,\varphi,q}$ is used to estimate the entropy of ergodic chains in Section IV. The relationship to return-time based entropy estimators and Maurer's conjecture are considered in Section V, and, finally, a comparison of the different statistics in an empirical study is done in Section VI. Section VII summarizes the results, the Appendix contains technical details.

Throughout the paper we use the following notation: For an integer $m > 1$ let S denote the state space $S = \{1, \dots, m\}$ and define $D(S) := \{\mathbf{P} = (p_1, \dots, p_m) \in \mathbb{R}^m, p_i \geq 0, i \in \{1, \dots, m\}, \sum_{i=1}^m p_i = 1\}$ the set of m-variate discrete probability distributions and $D^\circ(S) := \{\mathbf{P} \in D(S), \mathbf{P} > \emptyset\}$ the set of probability distributions with support $\{1, \dots, m\}$. We denote the Markov chain $(X_l)_{l \in \mathbb{N}}$, $X_l \in S$, with initial distribution $\mathbf{P}_0 = (p_{01}, \dots, p_{0m})$, $P[X_1 = i] = p_{0i}$, $i \in S$, and transition matrix $\mathbb{P} = (p_{ij})$, $(i,j) \in S^2$, by $(S, \mathbb{P}, \mathbf{P}_0)$. Unless stated otherwise, we assume that the chain is of order 1, i.e., $P[X_l = x_l | X_{l-1} = x_{l-1}, \dots, X_{l_1} = x_{l_1}] = P[X_l = x_l | X_{l-1} = x_{l-1}] = p_{x_l x_{l-1}}$ for $l > 2$ and $(x_1, \dots, x_l) \in S^l$. We denote the rth order transition probabilities, i.e., the elements of the rth power $\mathbb{P}^r$ of the matrix $\mathbb{P}$ by $p_{ij}^{(r)}$, $(i,j) \in S^2$. All the chains considered in this paper are finite, irreducible, and aperiodic, and thus ergodic: $\lim_{r\to\infty} p_{ij}^{(r)} = p_j$ for all $i \in S$, where we call $\mathbf{P} = (p_1, \dots, p_m) \in D^\circ(S)$ the stable distribution of the chain. We abbreviate the stationary ergodic chain $(S, \mathbb{P}, \mathbf{P}_0)$ with $\mathbf{P}_0$ equal to the stable distribution, $\mathbf{P}_0 = \mathbf{P}$, by $(S, \mathbb{P})$.

Define the counter vector $C(n) = (c_1(n), \dots, c_m(n))$, where $c_i(n) = \#\{1 \leq l \leq n : X_l = i\}$, and let $\hat{\mathbf{P}}(n)$ the normalized counter (or relative frequency) vector,

$$\hat{\mathbf{P}}(n) = (\hat{p}_1(n), \dots, \hat{p}_m(n)) = \frac{1}{n} C(n).$$

An important special case here is the independent or zero-order chain with respect to a vector $\mathbf{P} \in D^\circ(S)$ which is defined by the transition matrix $p_{ij} = p_j$, $i \in S$. This chain forgets the initial distribution $\mathbf{P}_0$ after one step and yields a sequence of i.i.d. states $(X_l)_{l>1}$, where $P[X_l = i] = p_i$ for $l > 1$. The defining vector $\mathbf{P}$ is also the stable distribution in this case. We will sometimes use the term dependent chain for ergodic chains which are not independent with respect to some $\mathbf{P}$. Chains of finite order $\kappa > 1$ on the other hand can be treated as ordinary order one chains by considering the sequence $(X_l^{(r)})_{l \in \mathbb{N}}$ of overlapping r-tuples of successive states, $X_l^{(r)} = (X_l, \dots, X_{l+r-1})$, where $r \geq \kappa$.

In order to define the relative frequency of such overlapping r-tuples for a given sample size n it is convenient to treat the first $n \in \mathbb{N}$ states of the original chain in a cyclic

Stefan Wegenkittl, Institut für Mathematik, Universität Salzburg, Hellbrunnerstr. 34, A-5020 Salzburg, Austria. e-mail: Stefan.Wegenkittl@u24.at WWW: http:random.mat.sbg.ac.at, research is supported by the Austrian Science Foundation (FWF), project no. P13480-MAT and FSP "Number-theoretic Algorithms and Their Applications".

manner: we let for $l \in \mathbb{N}$, $1 \leq k \leq n$ and $k-1 \equiv l-1 \pmod n$, $\tilde{X}_l = X_k$ and write $\tilde{X}_l^{(r)} = (\tilde{X}_l, \ldots, \tilde{X}_{l+r-1})$, $1 \leq l \leq n$. Now let $(\overline{X}_l^{(r)})_{l \in \mathbb{N}}$, $\overline{X}_l^{(r)} = (X_{(l-1)r+1}, \ldots, X_{lr})$, the sequence of nonoverlapping r-tuples of successive states treating the original sequence of the first n states in a cyclic manner. This definition depends on n.

For a chain $(S, \mathbb{P}, \mathbf{P}_0)$, an integer $r \geq 2$, and $\mathbf{i} = (i_1, \ldots, i_r) \in S^r$ we let $\pi_{\mathbf{i}}^{(r)} := p_{i_1} \cdot p_{i_1 i_2} \cdots p_{i_{r-1} i_r}$ the probability of the path $i_1, i_2, \ldots, i_r$ in the stationary chain $(S, \mathbb{P})$, and define the vector $\Pi^{(r)} = (\pi_{\mathbf{i}}^{(r)})_{\mathbf{i} \in S^r}$. We use the lexicographical order for the components of this vector. Also let $C^{(r)}(n) = (c_{\mathbf{i}}^{(r)})_{\mathbf{i} \in S^r}$, where $c_{\mathbf{i}}^{(r)} = \#\{1 \leq l \leq n : \tilde{X}_l^{(r)} = \mathbf{i}\}$. Finally, let $\hat{\mathbf{P}}^{(r)}(n) = (\hat{p}_{\mathbf{i}}^{(r)}(n))_{\mathbf{i} \in S^r} = \frac{1}{n} C^{(r)}(n)$. Note, that these counter vectors are defined using the cyclic version of the overlapping tuples of states and that, consequently, for $r > 1$, $\mathbf{i} \in S^{r-1}$, $j \in S$, and $(\mathbf{i}; j) = (i_1, \ldots, i_{r-1}, j)$,

$$\hat{p}_{\mathbf{i}}^{(r-1)}(n) = \sum_{j \in S} \hat{p}_{(\mathbf{i};j)}^{(r)}(n). \tag{1}$$

The vectors $\hat{\mathbf{P}}(n)$ and $\hat{\mathbf{P}}^{(r)}(n)$ are so-called higher order types, see Section VII.A. in [2]. The theory of types, a powerful tool in Shannon theory, can be applied to prove convergence in distribution of divergence statistics depending on these relative frequency vectors. It also provides large deviation results yielding probability one convergence for several statistics for Markov chains.

We denote convergence in distribution by $\overset{\mathrm{d}}{\to}$, convergence in probability by $\overset{\mathrm{P}}{\to}$, and almost sure convergence by $\overset{\mathrm{a.s.}}{\to}$. Further, $\log(x)$ denotes the logarithm with respect to base 2 and $\ln(x)$ denotes the natural logarithm.

II. The Independent Case

In 1963, Csiszár [3] introduced the family of φ-divergences as measures for the divergence of two probability distributions, see also [4] for the relationship to statistical information. We consider the following scaled version for discrete distributions: Let $\varphi : [0, \infty) \to (-\infty, \infty]$ be a function with continuous second derivative on some nonempty interval $I_\delta = (1-\delta, 1+\delta) \subset [0, \infty)$, such that $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(1) \neq 0$, and let φ be arbitrary outside of I_δ. Define the φ-divergence of $\hat{\mathbf{P}}(n)$ and $\mathbf{P}$ by

$$I_\varphi(\hat{\mathbf{P}}(n), \mathbf{P}) := \frac{2n}{\varphi''(1)} \sum_{i=1}^{m} p_i \, \varphi\left(\frac{\hat{p}_i(n)}{p_i}\right)$$

Under the condition that $n \cdot \hat{\mathbf{P}}(n)$ is distributed multinomial with parameters n and $\mathbf{P} \in D^\circ(S)$, $\hat{\mathbf{P}}(n) \sim \mathcal{MN}(n, \mathbf{P})$, the asymptotic distribution of I_φ as $n \to \infty$ is a chi-square with $m-1$ degrees of freedom. Two famous measures amongst this family are Pearson's [5] goodness-of-fit statistic $\mathcal{X}^2(\hat{\mathbf{P}}(n), \mathbf{P}) = n \sum_{i=1}^{m} \frac{(\hat{p}_i(n) - p_i)^2}{p_i}$, for the loss function $\varphi_2(u) := \frac{1}{2}(u-1)^2$, and the I-Divergence of Kullback-Leibler [6], $G^2 = 2n \sum_{i=1}^{m} \hat{p}_i(n) \ln(\frac{\hat{p}_i(n)}{p_i})$, for $\varphi_1(u) := 1 - u + u\ln(u)$, which is also called the log-likelihood ratio statistic. The factor 2 in the definition of G^2 is needed to get convergence in distribution in the multinomial setting. It does not appear in the standard definition of the log-likelihood ratio statistic. In the equiprobable case $\mathbf{P} = (\frac{1}{m}, \ldots, \frac{1}{m})$, G^2 is equal to $2n(\ln(m) - \ln(2) H(\hat{\mathbf{P}}(n)))$, where $H(\hat{\mathbf{P}}(n)) = -\sum_{i=1}^{m} \hat{p}_i(n) \log(\hat{p}_i(n))$ is the sample entropy.

In the setup of ergodic chains we might summarize: a stationary independent chain $(S, \mathbb{P})$ with respect to $\mathbf{P} \in D^\circ(S)$ yields a sequence $C(n)$ which is multinomial distributed with parameters n and $\mathbf{P}$, and the asymptotic distribution of $\mathcal{X}^2(\hat{\mathbf{P}}(n), \mathbf{P})$ and $G^2(\hat{\mathbf{P}}(n), \mathbf{P})$ is a chi-square with $m-1$ degrees of freedom. The strong law of large numbers guarantees that the sample entropy $H(\hat{\mathbf{P}}(n))$ converges almost sure to the entropy $H(\mathbf{P}) = -\sum_{i=1}^{m} p_i \log(p_i)$ of the distribution $\mathbf{P}$, $H(\hat{\mathbf{P}}(n)) \overset{\mathrm{a.s.}}{\to} H(\mathbf{P})$. This entropy coincides with the entropy of the chain in the independent case, see Section IV. As stated above, the sample entropy $H(\hat{\mathbf{P}}(n))$ and the goodness-of-fit statistic G^2 are equal up to a linear transformation in the equiprobable case,

$$\log(m) - \frac{G^2(\hat{\mathbf{P}}(n), (\frac{1}{m}, \ldots, \frac{1}{m}))}{2n\ln(2)} = H(\hat{\mathbf{P}}(n)) \overset{\mathrm{a.s.}}{\to} H(\mathbf{P}). \tag{2}$$

This gives a framework for relating standard serial tests - i.e. goodness-of-fit tests based on the frequency count $C(n)$ in independent chains - and the assessment of the randomness of i.i.d. sequences by their estimated entropy: rewriting (2) as

$$H\left((\frac{1}{m}, \ldots, \frac{1}{m})\right) - H(\hat{\mathbf{P}}(n)) = \frac{G^2\left(\hat{\mathbf{P}}(n), (\frac{1}{m}, \ldots, \frac{1}{m})\right)}{2n\ln(2)},$$

we observe that a properly normalized I-divergence estimates the difference between the entropy of the equidistribution and that of $\mathbf{P}$.

III. Generalized ϕ-Divergence and the Dependent Case

Various generalizations of I_φ can be considered. We refer the reader to [7] for an extensive treatment within the framework of $(\underline{h}, \underline{\phi})$-divergences. From [8] we cite without proof the following theorem on a generalized ϕ-divergence which unifies the φ-divergence approach with that of Rao [9] for quadratic forms in weak inverses:

Theorem III.1 ($\tilde{I}_{\underline{\Sigma}, \varphi}$-Divergence) Let $\varphi : [0, \infty) \to (-\infty, \infty]$ be a function with continuous second derivative on some interval $I_\delta = (1-\delta, 1+\delta) \subset [0, \infty)$, for which $\varphi(1) = \varphi'(1) = 0$ and $\varphi'' := \varphi''(1) \neq 0$, and let φ be arbitrary outside of I_δ. Let

$$\phi^\varphi(x, y) := 2\varphi\left(\frac{x+y}{2}\right) - \frac{\varphi(x) + \varphi(y)}{2},$$

and, for a matrix $\overline{\Sigma} = (\overline{\sigma}_{ij})_{(i,j)\in S^2}$, let

$$\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}(n), \mathbf{P}) = \frac{2n}{\varphi''} \sum_{i,j=1}^{m} \overline{\sigma}_{ij} p_i p_j \phi^{\varphi}\left(\frac{\hat{p}_i(n)}{p_i}, \frac{\hat{p}_j(n)}{p_j}\right).$$

On the conditions that $\sqrt{n}\left(\hat{\mathbf{P}}(n) - \mathbf{P}\right) \overset{\mathrm{d}}{\to} \mathcal{N}(\emptyset, \Sigma)$ for a covariance matrix Σ with rank $R(\Sigma)$, and that $\overline{\Sigma}$ is a weak inverse of Σ, $\Sigma\overline{\Sigma}\Sigma = \Sigma$, this statistic is asymptotically chi-square distributed with $R(\Sigma)$ degrees of freedom, $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}(n), \mathbf{P}) \overset{\mathrm{d}}{\to} \chi^2_{R(\Sigma)}$.

By the Central Limit Theorem for finite, irreducible, and aperiodic chains $(S, \mathrm{I\!P}, \mathbf{P}_0)$ (see e.g. [10, Chapter 4]) there exists a covariance matrix Σ such that the normalized counter vector $\hat{\mathbf{P}}(n)$ converges weakly to a normal distribution, $\sqrt{n}\left(\hat{\mathbf{P}}(n) - \mathbf{P}\right) \overset{\mathrm{d}}{\to} \mathcal{N}(\emptyset, \Sigma)$ as $n \to \infty$, for every initial distribution $\mathbf{P}_0 \in D(S)$. Here, $\mathbf{P}$ is the (unique) stable distribution. Letting $\overline{\Sigma}$ a weak inverse of Σ (which does always exist but need not to be unique), $\tilde{I}_{\overline{\Sigma},\varphi}$ serves as a goodness-of-fit statistic for such chains and thereby extends the notion of serial tests to observations from dependent processes with finite memory. We refer the reader to [8] for details and for the computation of the stable distribution $\mathbf{P}$ and the covariance matrix Σ in terms of determinants of minors of the transition matrix $\mathrm{I\!P}$. Among the family of $\tilde{I}_{\overline{\Sigma},\varphi}$ divergences we mention the generalized Pearson statistic $\tilde{I}_{\overline{\Sigma},\varphi_2}$, which is nothing but the quadratic form in the weak inverse $\overline{\Sigma}$, and the generalized I-Divergence $\tilde{I}_{\overline{\Sigma},\varphi_1}$. Also note, that $\tilde{I}_{\overline{\Sigma},\varphi}$ comprises I_φ in the independent case since a weak inverse of the multinomial distribution with parameters 1 and $\mathbf{P}$ is given by $diag(\frac{1}{p_1}, \ldots, \frac{1}{p_m})$ and since $\phi^\varphi(x,x) = \varphi(x)$.

Example III.2: Consider a chain $(S, \mathrm{I\!P}, \mathbf{P}_0)$ and construct the so-called overlapping chain $(S^r, \mathrm{I\!P}^{(r)}, \mathbf{P}_0^{(r)})$ with dimension r for the sequence $(X_l^{(r)})_{l\in\mathrm{I\!N}}$ of overlapping r-tuples of states. Clearly $\mathrm{I\!P}^{(r)} = (\varphi_{\mathbf{ij}}^{(r)})_{(\mathbf{i},\mathbf{j})\in S^{2r}}$ is given by $\varphi_{\mathbf{ij}}^{(r)} = p_{i_1 j_1} \cdot p_{j_1 j_2} \cdots p_{j_{r-1} j_r}$ if $i_{l+1} = j_l$ for all $1 \le l \le r-1$, and by zero, otherwise. The initial distribution $\mathbf{P}_0^{(r)} = (p_{0\mathbf{i}}^{(r)})_{\mathbf{i}\in S^r}$ is given by $p_{0\mathbf{i}}^{(r)} = p_{0 i_1} p_{i_1 i_2} \cdots p_{i_{r-1} i_r}$. We again use the lexicographical order for the components of the vectors and matrices. The chain $(S^r, \mathrm{I\!P}^{(r)}, \mathbf{P}_0^{(r)})$ is aperiodic and irreducible if and only if $(S, \mathrm{I\!P}, \mathbf{P}_0)$ is so. If $\mathbf{P}$ is the stable distribution of $(S, \mathrm{I\!P}, \mathbf{P}_0)$, $\Pi^{(r)}$ is the stable distribution of $(S^r, \mathrm{I\!P}^{(r)}, \mathbf{P}_0^{(r)})$. Note, that even if $(S, \mathrm{I\!P}, \mathbf{P}_0)$ is independent with respect to $\mathbf{P} \in D^\circ(S)$, the overlapping chain with dimension $r > 1$ is not independent any more and standard goodness-of-fit tests are not appropriate for the relative frequencies of the overlapping tuples. In this special case, there is a compact representation of a weak inverse $\overline{V}^{(r)}$ of the asymptotic covariance matrix $V^{(r)}$, say, of the normalized counter vector $\hat{\mathbf{P}}^{(r)}(n)$, however: for $\mathbf{i} \in S^r$ let $\mathbf{i}' = (i_1, \ldots, i_{r-1})$, and put $\overline{V}^{(r)} = (\overline{v}_{\mathbf{ij}})_{(\mathbf{i},\mathbf{j})\in S^{2r}}$, where

$$\overline{v}_{\mathbf{ij}} = \begin{cases} \frac{1}{\pi_{\mathbf{i}}^{(r)}} - \frac{1}{\pi_{\mathbf{i}'}^{(r-1)}} & : \quad \mathbf{i} = \mathbf{j} \\ -\frac{1}{\pi_{\mathbf{i}'}^{(r-1)}} & : \quad \mathbf{i}' = \mathbf{j}', i_r \neq j_r \\ 0 & : \quad \text{otherwise,} \end{cases} \tag{3}$$

then $V^{(r)}\overline{V}^{(r)}V^{(r)} = V^{(r)}$ and the rank of $V^{(r)}$ equals $R(V^{(r)}) = m^r - m^{r-1}$, see [11,12] for details. If $(S, \mathrm{I\!P})$ is stationary and independent with respect to $\mathbf{P}$, the probability of the sample path $\mathbf{i}$ becomes $\pi_{\mathbf{i}}^{(r)} = \prod_{l=1}^{r} p_{i_r}$. The generalized Pearson statistic $\tilde{I}_{V^{(r)},\varphi_2}$ for the *overlapping chain of such a stationary independent chain* is especially easy to evaluate (see [11-13]), since it turns out that it equals the difference of two ordinary Pearson statistics, one for dimension r and one for dimension $r-1$,

$$\begin{aligned} \tilde{I}_{\overline{V}^{(r)},\varphi_2}(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}) &= \mathcal{X}^2(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}) - \\ &- \mathcal{X}^2(\hat{\mathbf{P}}^{(r-1)}(n), \Pi^{(r-1)}). \end{aligned} \tag{4}$$

Here we rely on the cyclic definition of the counters for the overlapping tuples of states. The difference between normalized cyclic and noncyclic counters becoming arbitrarily small as n increases, the effects can be ignored for high enough sample sizes n.

The above example and the special case (4) indicate a slight generalization of Theorem III.1 such that conditions for the weak convergence to a chi-square for sums and differences of $\tilde{I}_{\overline{\Sigma},\varphi}$-divergences are given. In fact, it is easy to see, that the arguments for the proof of Theorem III.1 apply also to the following *modified* generalized ϕ-divergence

Theorem III.3 ($\tilde{I}_{(\overline{\Sigma},\varphi,q)}$-Divergence) Let $q \in \mathrm{I\!N}$ and let $\varphi^{(l)}$, $1 \le l \le q$, be a family of functions $\varphi^{(l)} : [0,\infty) \to (-\infty, \infty]$ with continuous second derivatives on some interval $I_\delta = (1-\delta, 1+\delta) \subset [0,\infty)$, such that $\varphi^{(l)}(1) = (\varphi^{(l)})'(1) = 0$ and $\varphi^{(l)''} := (\varphi^{(l)})''(1) \neq 0$. Further let $\overline{\Sigma}^{(l)} = (\overline{\sigma}_{ij}^{(l)})_{(i,j)\in S^2}$, $1 \le l \le q$, be a family of $m \times m$ real matrices. Define

$$\phi^{(l)}(x,y) := 2\varphi^{(l)}\left(\frac{x+y}{2}\right) - \frac{\varphi^{(l)}(x) + \varphi^{(l)}(y)}{2},$$

and let

$$\begin{aligned} &\tilde{I}_{(\overline{\Sigma},\varphi,q)}(\hat{\mathbf{P}}(n), \mathbf{P}) = \\ &= 2n \sum_{l=1}^{q} \frac{1}{\varphi^{(l)''}} \sum_{i,j=1}^{m} \overline{\sigma}_{ij}^{(l)} p_i p_j \phi^{(l)}\left(\frac{\hat{p}_i(n)}{p_i}, \frac{\hat{p}_j(n)}{p_j}\right). \end{aligned}$$

On the conditions that $\sqrt{n}\left(\hat{\mathbf{P}}(n) - \mathbf{P}\right) \overset{\mathrm{d}}{\to} \mathcal{N}(\emptyset, \Sigma)$ for a covariance matrix Σ, and that $\overline{\Sigma} := \sum_{l=1}^{q} \overline{\Sigma}^{(l)}$ is a weak inverse of Σ, $\Sigma\overline{\Sigma}\Sigma = \Sigma$, this statistic is asymptotically distributed chi-square with $R(\Sigma)$ degrees of freedom, $\tilde{I}_{(\overline{\Sigma},\varphi,q)}(\hat{\mathbf{P}}(n), \mathbf{P}) \overset{\mathrm{d}}{\to} \chi^2_{R(\Sigma)}$.

Proof: We consider a Taylor expansion of $\phi^{(l)}(x,y)$ at the point $(1,1)$ inside I_δ^2. The terms of order zero and one vanish according to the conditions imposed on $\varphi^{(l)}$. Setting $h = (x,y) - (1,1)$, the second order remainder term becomes $\frac{\delta^2\phi^{(l)}((1,1)+\vartheta h;h)}{2!}$ for a $\vartheta \in (0,1)$, where δ^2 denotes the second variation of $\phi^{(l)}$ in the point $(1,1)+\vartheta h$ and in direction h. Now let $R(x,y) = \frac{\delta^2\phi^{(l)}((1,1)+\vartheta h;h)}{2!} - \frac{\delta^2\phi^{(l)}((1,1);h)}{2!}$. Since the derivatives $\frac{\partial^2\phi^{(l)}(a,b)}{\partial a^2}$ and $\frac{\partial^2\phi^{(l)}(a,b)}{\partial b^2}$ vanish in $(1,1)$, we get

$$\phi^{(l)}(x,y) = \frac{\varphi^{(l)''}}{2}(x-1)(y-1) + R(x,y).$$

From $\sqrt{n}\left(\hat{\mathbf{P}}(n) - \mathbf{P}\right) \overset{\mathrm{d}}{\to} \mathcal{N}(\emptyset, \Sigma)$ it follows by standard estimation techniques (see e.g. [14]), first, that $P\left[\left(\frac{\hat{p}_i(n)}{p_i}, \frac{\hat{p}_j(n)}{p_j}\right) \notin I_\delta^2\right]$ converges to zero, and, secondly, that all remainder terms $R(x,y)$ in a complete expansion of $\tilde{I}_{(\overline{\Sigma},\varphi,q)}$ converge in probability to zero, so that the asymptotic distribution of $\tilde{I}_{(\overline{\Sigma},\varphi,q)}(\hat{\mathbf{P}}(n), \mathbf{P})$ equals that of $n\sum_{l=1}^{q} \sum\limits_{i,j=1}^{m} \overline{\sigma}_{ij}^{(l)} \left(\hat{p}_i(n) - p_i\right)\left(\hat{p}_j(n) - p_j\right)$. The Theorem follows by exchanging the order of summation, which yields a standard quadratic form in the weak inverse. ■

Example III.4:
Recall the overlapping chain $(S^r, \mathbb{P}^{(r)}, \mathbf{P}_0^{(r)})$ of dimension $r > 1$ of an independent chain $(S, \mathbb{P}, \mathbf{P}_0)$ with respect to a vector $\mathbf{P} \in D^\circ(S)$ considered in Example III.2. We decompose the weak inverse $\overline{V}^{(r)}$ given in (3) with respect to its additive structure into $\overline{V}^{(r)} = \overline{V}^{(r,1)} + \overline{V}^{(r,2)}$ with $\overline{V}^{(r,l)} = (\overline{v}_{\mathbf{ij}}^{(r,l)})_{(\mathbf{i},\mathbf{j})\in S^{2r}}$, $l \in \{1,2\}$,

$$\overline{v}_{\mathbf{ij}}^{(r,1)} = \begin{cases} \frac{1}{\pi_{\mathbf{i}}^{(r)}} & : \ \mathbf{i} = \mathbf{j} \\ 0 & : \ \text{otherwise, and} \end{cases}$$

$$\overline{v}_{\mathbf{ij}}^{(r,2)} = \begin{cases} -\frac{1}{\pi_{\mathbf{i}'}^{(r-1)}} & : \ \mathbf{i}' = \mathbf{j}' \\ 0 & : \ \text{otherwise.} \end{cases}$$

Now let $\varphi^{(1)}(u) = \varphi^{(2)}(u) = \varphi(u)$ be a function fulfilling the conditions in Theorem III.3. From the Central Limit Theorem for ergodic Markov chains it follows that

$$\tilde{I}_{(\overline{V}^{(r)},\varphi,2)}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) =$$

$$= \tilde{I}_{(\overline{V}^{(r,1)},\varphi)}\left(\cdot,\cdot\right) + \tilde{I}_{(\overline{V}^{(r,2)},\varphi)}\left(\cdot,\cdot\right) =$$

$$= I_\varphi\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) - I_\varphi\left(\hat{\mathbf{P}}^{(r-1)}(n), \Pi^{(r-1)}\right) =: \tilde{I}_\varphi^{(r)} \quad (5)$$

converges to a chi-square with $m^r - m^{r-1}$ degrees of freedom, $\tilde{I}_\varphi^{(r)} \overset{\mathrm{d}}{\to} \chi^2_{m^r - m^{r-1}}$, as n approaches infinity.
Similar statistics for overlapping tuples of states from independent chains and their application in the area of empirical assessment of random number generators have been studied in [15].

The essential point here is, that (5) is an appropriate analogue to equation (4) for the generalized I-divergence based on the function φ_1. The difference between the I-divergence for overlapping r-tuples and that for overlapping $(r-1)$-tuples gives a modified generalized I-divergence if the data stems from sampling successive overlapping tuples of states from an independent chain. In particular, we will consider the accordingly modified generalized I-divergence which we denote by $\tilde{I}_{\varphi_1}^{(r)}$ for $r > 1$ and let $\tilde{I}_{\varphi_1}^{(1)} := I_{\varphi_1}$. We will see below, that this statistic has an interesting application even if the data is sampled from a dependent chain, although the convergence to a chi-square is not guaranteed then.

IV. Entropy Estimates

So far we have a proper generalization from the independent to the dependent case concerning goodness-of-fit tests and a special representation in the case of overlapping tuples of successive states sampled from an independent chain. Recalling the relation between goodness-of-fit tests and entropy estimates in the independent case from Section II, we might suspect that we will have to compare $\hat{\mathbf{P}}(n)$ to the equidistribution $(\frac{1}{m^r}, \ldots, \frac{1}{m^r})$ in order to relate the statistic $\tilde{I}_{\varphi_1}^{(r)}$ to appropriate entropy estimators in the case of *dependent chains*. In the following we study this relation in more detail.

Denote by $H(S, \mathbb{P})$ the entropy of the process of states $(X_l)_{l\in\mathbb{N}}$ of the ergodic chain $(S, \mathbb{P}, \mathbf{P}_0)$,

$$H(S,\mathbb{P}) = -\sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{ij}\log(p_{ij}),$$

where $\mathbf{P} = (p_1, \ldots, p_m)$ is the stable distribution of the chain. The Theorem of Shannon-Breiman-McMillan states that for almost every sequence $(x_n)_{n\in\mathbb{N}}$ and for every initial distribution $\mathbf{P}_0 \in D(S)$, the average $-\frac{1}{n}\log(P[X_1 = x_1, \ldots, X_n = x_n])$ converges to $H(S,\mathbb{P})$. Note, that $H(S,\mathbb{P}) = H(\mathbf{P})$ for independent chains with respect to $\mathbf{P} \in D(S)$.

A standard estimator of the entropy $H(S,\mathbb{P})$ can be derived from the normalized counter vector $\hat{\mathbf{P}}^{(r)}(n)$ which converges almost surely to the stable distribution $\Pi^{(r)}$ of the overlapping chain of dimension r due to the strong law for ergodic chains. Consider the following inductive rule for the entropies of stable distributions of overlapping chains:

Lemma IV.1 (Chain rule) Let $(S, \mathbb{P}, \mathbf{P}_0)$ be an ergodic chain with state space $S = \{1, \ldots, m\}$ and stable distribution $\mathbf{P}$, and let $(S^r, \mathbb{P}^{(r)}, \mathbf{P}_0^{(r)})$ be the overlapping chain with dimension r as defined in Example III.2. Denote the entropy of the stable distribution of $(S^r, \mathbb{P}^{(r)}, \mathbf{P}_0^{(r)})$ by $H(\Pi^{(r)})$, $H(\Pi^{(r)}) := -\sum_{\mathbf{i}\in S^r} \pi_{\mathbf{i}}^{(r)} \log(\pi_{\mathbf{i}}^{(r)})$, $r > 1$, and let $H(\Pi^{(1)}) := H(\mathbf{P})$. Then for $r > 1$,

$$H(S,\mathbb{P}) = H(\Pi^{(r)}) - H(\Pi^{(r-1)}).$$

Proof: Let $l_{\mathbf{i}}^{(r)} = log(p_{i_1}) + \sum_{k=2}^{r} log(p_{i_{k-1}i_k})$ and $\mathbf{i}' = (i_1, \ldots, i_{r-1})$, then

$$H(\Pi^{(r)}) - H(\Pi^{(r-1)}) =$$

$$= -\sum_{\mathbf{i}\in S^r} \pi_{\mathbf{i}}^{(r)} l_{\mathbf{i}}^{(r)} - H(\Pi^{(r-1)}) =$$

$$= -\sum_{\mathbf{i}\in S^r} \pi_{\mathbf{i}}^{(r)} \left(l_{\mathbf{i}'}^{(r-1)} + \log(p_{i_{r-1}i_r})\right) - H(\Pi^{(r-1)}) =$$

$$= H(\Pi^{(r-1)}) - \sum_{\mathbf{i}\in S^r} \pi_{\mathbf{i}'}^{(r-1)} p_{i_{r-1}i_r} \log(p_{i_{r-1}i_r}) - H(\Pi^{(r-1)})$$

where the last line follows from the fact that

$$\sum_{i_r\in S} p_{i_{r-1}i_r} = 1.$$

Since $\mathbf{P}$ is the stable distribution of the chain $(S, \mathbb{P}, \mathbf{P}_0)$, $\mathbf{P}\mathbb{P}^k = \mathbf{P}$ for every $k \in \mathbb{N}$ and

$$H(\Pi^{(r)}) - H(\Pi^{(r-1)}) =$$

$$= - \sum_{\mathbf{i}\in S^{r-1}} \pi_{\mathbf{i}'}^{(r-1)} \sum_{i_r\in S} p_{i_{r-1}i_r} \log(p_{i_{r-1}i_r}) =$$

$$= - \sum_{i_{r-1}\in S} p_{i_{r-1}} \sum_{i_r\in S} p_{i_{r-1}i_r} \log(p_{i_{r-1}i_r}) = H(S, \mathbb{P})$$

■

With the additional definition $H(\Pi^{(0)}) := 0$, the chain rule also holds for $r = 1$, if $(S, \mathbb{P}, \mathbf{P}_0)$ is independent with respect to some $\mathbf{P} \in D^\circ(S)$. As mentioned in the introduction, chains with order $\kappa > 1$ can be treated as ordinary chains by analyzing overlapping κ-tuples of successive states, so that we restrict ourselves to the case of ordinary chains to avoid excessive notation. It is clear however, how to extend the definitions of $H(S, \mathbb{P})$ and $\Pi^{(r)}$ such that Lemma IV.1 holds for every chain of order $\kappa \in \mathbb{N}$ provided that $r > \kappa$.

The aforementioned almost sure convergence of the normalized counters and the continuity of the entropy (where we define $0\log(0) = 0$) imply

$$H(\hat{\mathbf{P}}^{(r)}(n)) - H(\hat{\mathbf{P}}^{(r-1)}(n)) \overset{\text{a.s.}}{\to} H(S, \mathbb{P}), \tag{6}$$

so that the difference $H(\hat{\mathbf{P}}^{(r)}(n)) - H(\hat{\mathbf{P}}^{(r-1)}(n))$ may be used to estimate the entropy of the chain $(S, \mathbb{P})$. By (1), this entropy estimator is based on counting overlapping r-tuples of successive states. Estimating entropies from relative frequency histograms is a standard technique. A nice example concerning the entropy of English is given in [16, Section 6.6]. The relation to serial testing can now be established by observing that

$$I_{\varphi_1}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) = 2n \sum_{\mathbf{i}\in S^r} \pi_{\mathbf{i}}^{(r)} \varphi_1\left(\frac{\hat{p}_{\mathbf{i}}^{(r)}(n)}{\pi_{\mathbf{i}}^{(r)}}\right) =$$

$$= 2n \sum_{\mathbf{i}\in S^r} \hat{p}_{\mathbf{i}}^{(r)}(n) \ln\left(\hat{p}_{\mathbf{i}}^{(r)}(n)\right) - 2n \sum_{\mathbf{i}\in S^r} \hat{p}_{\mathbf{i}}^{(r)}(n) \ln\left(\pi_{\mathbf{i}}^{(r)}\right),$$

and that – for the independent equiprobable case with $p_{ij} = p_i = \frac{1}{m}$, $(i, j) \in S^2$ – we get $\pi_{\mathbf{i}}^{(r)} = \frac{1}{m^r}$, $\mathbf{i} \in S^r$, so that

$$I_{\varphi_1}\left(\hat{\mathbf{P}}^{(r)}(n), (\frac{1}{m^r}, \ldots, \frac{1}{m^r})\right) =$$

$$= -2n\ln(2) H(\hat{\mathbf{P}}^{(r)}(n)) + 2n\ln(m^r). \tag{7}$$

We summarize the results in the following Corollary:

Corollary IV.2 (Entropy Estimates vs. Serial Tests) Let $r > 1$, consider the loss function $\varphi_1(u) = 1 - u + u\ln(u)$ and the statistic (5),

$$\tilde{I}_{\varphi_1}^{(r)}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) =$$

$$= I_{\varphi_1}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) - I_{\varphi_1}\left(\hat{\mathbf{P}}^{(r-1)}(n), \Pi^{(r-1)}\right).$$

i) If $(S, \mathbb{P}, \mathbf{P}_0)$ is ergodic and independent with respect to $\mathbf{P} \in D^\circ(S)$, then

$$\tilde{I}_{\varphi_1}^{(r)}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right) \overset{\text{d}}{\to} \chi^2_{m^r - m^{r-1}},$$

providing for an asymptotically chi-square distributed goodness-of-fit test for the hypothesis that the normalized overlapping counter vector has been sampled from an independent chain with respect to $\mathbf{P}$.
ii) If $(S, \mathbb{P}, \mathbf{P}_0)$ is an ergodic, not necessarily independent chain, then

$$\log(m) - \frac{\tilde{I}_{\varphi_1}^{(r)}\left(\hat{\mathbf{P}}^{(r)}(n), (\frac{1}{m^r}, \ldots, \frac{1}{m^r})\right)}{2n\ln(2)} \overset{\text{a.s.}}{\to} H(S, \mathbb{P}),$$

This statistic thus provides for an entropy estimate for ergodic chains $(S, \mathbb{P}, \mathbf{P}_0)$ of order 1.

Proof: Item i) follows directly from Theorem III.3, ii) is deduced from (7) by observing that

$$H\left((\frac{1}{m}, \ldots, \frac{1}{m})\right) - \left(H(\hat{\mathbf{P}}^{(r)}(n)) - H(\hat{\mathbf{P}}^{(r-1)}(n))\right) =$$

$$= \frac{\tilde{I}_{\varphi_1}^{(r)}\left(\hat{\mathbf{P}}^{(r)}(n), \Pi^{(r)}\right)}{2n\ln(2)}. \tag{8}$$

■

Note, that both applications of $\tilde{I}_{\varphi_1}^{(r)}$ actually compare data sampled from an (imaginary) chain process to a specified chain: In i), the chain is specified by the null-hypothesis of the test out of the family of independent chains, whereas in ii) the data is always compared to the independent equiprobable chain. In perfect accordance with the results from Section II, the properly normalized modified I-divergence for overlapping chains estimates the difference between the entropy of the equidistribution (i.e., the entropy of an i.i.d. equidistributed process) and that of the data generating chain.

Also note, that if $r = 1$ in ii), the estimator estimates the entropy of the stable distribution of the data generating chain, so that ii) also holds if $r \geq 1$ and $(S, \mathbb{P}, \mathbf{P}_0)$ is an *independent* chain. By this, the special case (2) is also included. If the order κ, say, of the data-generating chain was greater than 1, one would have to increase r such that $r > \kappa$ in order to get an estimator for its entropy, compare the above note on the definition of the entropy of such chains. As to i) again, if the chain was not independent, $\tilde{I}_{\varphi_1}^{(r)}$ would have to be replaced by the modified generalized $\tilde{I}_{(\Sigma,\varphi,q)}$-divergence with an *appropriate* weak inverse for the asymptotic covariance matrix of $\hat{\mathbf{P}}^{(r)}(n)$ in order to get convergence in distribution to a chi-square. An example for the application of the $\tilde{I}_{\overline{\Sigma},\varphi}$-divergence in the analysis of a gambling strategy is given in [8,Chp.5.2] where we consider the quality of huge-period pseudorandom number generators. Further examples for the application of the statistics (4) and $\tilde{I}^{(r)}_{\varphi_1}$ in the area of quality assessment of pseudorandom number generators are discussed in [15,17]. From Corollary IV.2 we conclude that similar results may be obtained by the corresponding entropy estimates.

V. Relationship to the First Return-Time

In his 1992 paper [1], U. M. Maurer discussed a "universal statistical test" f_{T_U} for random bit generators. The statistic is presented in a setup of stationary ergodic chains with two states and finite memory, the generalization to arbitrary finite state chains being obvious. A related statistic and its application to testing pseudorandomness is considered in [18,19].

These statistics are based on the observation, that the logarithm of the first return-time is related to the entropy of the chain: let

$$T^{(r)} = \min\left\{l \geq r \ : \ X^{(r)}_{l+1} = X^{(r)}_1\right\}$$

denote the first return-time to the vector $X^{(r)}_1$ of the initial r states. In [20] it is shown, that $\frac{\log(T^{(r)})}{r}$ converges in probability to the entropy $H(S,\mathbb{P})$ as $r \to \infty$ if the chain is ergodic and stationary, see also [21] for a pointwise theorem. Now let $\overline{X}^{(r)}_l = (X_{(l-1)r+1}, \ldots, X_{lr})$ be vectors of nonoverlapping r-tuples of successive states of the chain. The sequence $(\overline{X}^{(r)}_l)_{l\in\mathbb{N}}$ clearly is a Markov chain itself and has $\Pi^{(r)}$ as stable distribution. In the following, we consider the modified first return-time

$$\overline{T}^{(r)} = \min\left\{l \in \mathbb{N} \ : \ \overline{X}^{(r)}_{l+1} = \overline{X}^{(r)}_1\right\}$$

For stationary ergodic chains $(S,\mathbb{P})$, the chain $(\overline{X}^{(r)}_l)_{l\in\mathbb{N}}$ is ergodic itself and the conditional expectation

$$E[\overline{T}^{(r)}|\overline{X}^{(r)}_1 = \mathbf{i}]$$

consequently equals the inverse of the according probability in the stable distribution of the nonoverlapping chain,

$$E[\overline{T}^{(r)}|\overline{X}^{(r)}_1 = \mathbf{i}] = \frac{1}{\pi^{(r)}_{\mathbf{i}}}. \tag{9}$$

Maurer conjectures a relation between the expectation of the logarithm of the estimated return-time $\overline{T}^{(r)}$ and the entropy of the form

$$\lim_{r\to\infty}\left(E[\log(\overline{T}^{(r)})] - r\cdot H(S,\mathbb{P})\right) = C,$$

where we let the function

$$\nu(\epsilon) = \left(\epsilon\sum_{k\geq 1}\log(k)(1-\epsilon)^{k-1} + \log(\epsilon)\right)$$

and define

$$C = \lim_{\epsilon\to 0}\nu(\epsilon) = \lim_{\epsilon\to 0}\int_\epsilon^\infty e^{-\xi}\log(\xi)d\xi =$$
$$= -\gamma/\ln(2) = -0.832746\ldots \tag{10}$$

Here, γ denotes Euler's constant. Consider the following, slightly modified theorem:

Theorem V.1: Let $(S,\mathbb{P})$ be ergodic and stationary, and let $H(S,\mathbb{P})$, $H(\mathbf{P})$, $\overline{T}^{(r)}$, and C be defined as above. Then

$$\lim_{r\to\infty}\left(E[\log(\overline{T}^{(r)})] - (r-1)\cdot H(S,\mathbb{P}) - H(\mathbf{P})\right) = C.$$

This theorem was proved for binary sources in [22]. Here, we prove the general stationary case and thereby also extend the results obtained in [23].

Proof: First observe that by the chain rule (Lemma IV.1),

$$H(\Pi^{(r)}) = H(S,\mathbb{P}) + H(\Pi^{(r-1)}) =$$
$$= H(S,\mathbb{P}) + H(S,\mathbb{P}) + H(\Pi^{(r-2)}) = \ldots =$$
$$= (r-1)H(S,\mathbb{P}) + H(\Pi^{(1)}) = (r-1)H(S,\mathbb{P}) + H(\mathbf{P}),$$

so that $H(\Pi^{(r)}) - (r-1)\cdot H(S,\mathbb{P}) - H(\mathbf{P}) = 0$, and the theorem holds, if we can show that

$$\lim_{r\to\infty}\left(E[\log(\overline{T}^{(r)})] - H(\Pi^{(r)})\right) = C. \tag{11}$$

We write

$$E[\log(\overline{T}^{(r)})] = \sum_{\mathbf{i}\in S^r}\pi^{(r)}_{\mathbf{i}}\sum_{k\geq 1}\log(k)\cdot$$
$$\cdot P\left[\min\left\{l\in\mathbb{N} \ : \ \overline{X}^{(r)}_{l+1} = \mathbf{i}\right\} = k\middle|\overline{X}^{(r)}_1 = \mathbf{i}\right] =$$
$$=: \sum_{\mathbf{i}\in S^r}\pi^{(r)}_{\mathbf{i}}\sum_{k\geq 1}\log(k)P(\mathbf{i},k,r).$$

and let $\delta^{(r)}_{\mathbf{i}} = \frac{p_{\mathbf{i}_r\mathbf{i}_1}}{p_{\mathbf{i}_1}}$, $\epsilon^{(r)}_{\mathbf{i}} = \frac{\pi^{(r)}_{\mathbf{i}}}{1-\pi^{(r)}_{\mathbf{i}}}(1-\delta^{(r)}_{\mathbf{i}}\pi^{(r)}_{\mathbf{i}})$, and $\tau^{(r)}_{\mathbf{i}} = \frac{1-\delta^{(r)}_{\mathbf{i}}\pi^{(r)}_{\mathbf{i}}}{1-\epsilon^{(r)}_{\mathbf{i}}}$, so that $P(\mathbf{i},k,r) = \tau^{(r)}_{\mathbf{i}}\epsilon^{(r)}_{\mathbf{i}}(1-\epsilon^{(r)}_{\mathbf{i}})^{(k-1)}$ for $k > 1$, see the first lemma in the Appendix. The convergence of $\pi^{(r)}_{\mathbf{i}} \to 0$ as $r\to\infty$, which is uniform in $\mathbf{i}$, implies that of $\epsilon^{(r)}_{\mathbf{i}} \to 0$ and $\tau^{(r)}_{\mathbf{i}} \to 1$, see the second lemma in the Appendix. Letting

$$C(\mathbf{i},r) := \sum_{k\geq 1}\log(k)P(\mathbf{i},k,r) + \log\left(\pi^{(r)}_{\mathbf{i}}\right) =$$
$$= \left(\sum_{k\geq 2}\log(k)\tau^{(r)}_{\mathbf{i}}\epsilon^{(r)}_{\mathbf{i}}\left(1-\epsilon^{(r)}_{\mathbf{i}}\right)^{(k-1)} + \tau^{(r)}_{\mathbf{i}}\log\left(\epsilon^{(r)}_{\mathbf{i}}\right)\right) +$$
$$+\left(\log\left(\pi^{(r)}_{\mathbf{i}}\right) - \tau^{(r)}_{\mathbf{i}}\log\left(\epsilon^{(r)}_{\mathbf{i}}\right)\right)$$

we get, after some analysis, the limit $\lim_{r\to\infty}C(\mathbf{i},r) = C$, which is uniform in $\mathbf{i}$, too. The Theorem now follows owing to the fact that $\sum_{\mathbf{i}\in S^r}\pi^{(r)}_{\mathbf{i}} = 1$, by

$$\lim_{r\to\infty}\left(E[\log(\overline{T}^{(r)})] - H(\Pi^{(r)})\right) = \lim_{r\to\infty}\sum_{\mathbf{i}\in S^r}\pi^{(r)}_{\mathbf{i}}C(\mathbf{i},r) = C.$$

■

More on this topic can be found in [23]. Again, the arguments extend easily to chains of order $\kappa > 1$ by replacing the elements of the tuples $\overline{X}_l$ by overlapping κ-tuples of successive states and considering the according chains. Note, that in the special case that $(S, \mathbb{P})$ is independent with respect to $\mathbf{P} \in D^\circ(S)$, we have $H(S, \mathbb{P}) = H(\mathbf{P})$ so that Theorem V.1 simplifies to

$$\lim_{r\to\infty} \left(E[\log(\overline{T}^{(r)})] - r \cdot H(S, \mathbb{P}) \right) = C,$$

in accordance with the results in [1].

Remark V.2: According to (9), the expectation $E[\overline{T}^{(r)}]$ equals the number of paths $\mathbf{i} \in S^r$ with positive probability $\pi_{\mathbf{i}}^{(r)}$ for both the nonoverlapping chain $(\overline{X}_l)_{l\in\mathbb{N}}$, and a stationary independent chain with respect to $\Pi^{(r)}$. Similar arguments as employed in the proof of Theorem V.1 show that the expectation of the *logarithm* of the return-time also converges to the same limit for both chains as $r \to \infty$. This seems clear intuitively from the small deviation of the single waiting-time probabilities from the geometric distribution.

Corollary V.3: Let $(S, \mathbb{P})$ be stationary and ergodic, then

$$\lim_{r\to\infty} \frac{E[\log(\overline{T}^{(r)})]}{r} = H(S, \mathbb{P}). \qquad (12)$$

Proof: The corollary follows trivially from Theorem V.1 which also gives the speed of convergence for (12). ■

VI. Empirical Results

Viewing the unknown probabilities $(\pi_{\mathbf{i}}^{(r)})_{\mathbf{i}\in S^r}$ as expectations of the corresponding relative frequences, we might infer from Equation (9), that entropy estimates built on the estimation of the return-time of nonoverlapping r tuples are on a par with the corresponding estimator of Corollary IV.2 ii) which is built on the estimation of the stable distribution itself. There might be differences in the speed of convergence and power of the tests for finite sample sizes and dimensions, however. In this section we report some empirical results.

In all the examples, we select an integer $m \geq 2$ denoting the size of the state space $S = \{1, \ldots, m\}$, and generate an $m \times m$ transition matrix $\mathbb{P}$ depending on two parameters δ_1 and δ_2, such that for every $(i, j) \in S^2$,

$$p_{ij} = \lambda_i^{-1} f\left(\frac{i}{m}, \frac{j}{m}, \delta_1, \delta_2\right),$$

where $\lambda_i = \sum_{j\in S} f\left(\frac{i}{m}, \frac{j}{m}, \delta_1, \delta_2\right)$ and for $\delta_1 \geq 0$ we let $f(x, y, \delta_1, \delta_2) = y^{\delta_1}$, and for $\delta_1 < 0$ we let $f(x, y, \delta_1, \delta_2) = 1$ if $\delta_2 x + y \pmod 1 = 0$, and $f(x, y, \delta_1, \delta_2) = (\delta_2 x + y \pmod 1)^{-\delta_1}$, otherwise. Although this seems a somewhat ad hoc construction of test chains, the parameters give a good control over the entropies: parameters $\delta_1 \geq 0$ yield independent chains, whereas parameters $\delta_1 < 0$ yield dependent chains where we control the difference between $H(S, \mathbb{P})$ and $H(\mathbf{P})$ by δ_2.

We now choose a dimension $r \geq 1$ and a sample size $N \gg 1$ and calculate a sample path $\tau := (x_l)_{1\leq l\leq r\cdot N}$, $x_l \in S$, of the chain $(S, \mathbb{P})$ by straightforward simulation. To do so, we start with an initial distribution $\mathbf{P}_0 = (\frac{1}{m}, \ldots, \frac{1}{m})$ and simulate 2^{10} transitions before actually sampling τ. Since the influence of $\mathbf{P}_0$ wears off exponentially, it can be neglected. The simulation is based on the very stable high-performance pseudorandom number generator $TT800$, see [24,25,8]. We use the notation $x_l^{(r)}$ and $\overline{x}_l^{(r)}$ in accordance to the definitions in the introduction to denote overlapping and nonoverlapping r-tuples of successive states in the sample path.

Our goal is to analyze the behavior of the following three statistics:

1. Maurer's statistic $\hat{F}_c := \frac{1}{r\log(m)}(f_{T_U} - C)$. The dimension r is equal to the parameter L in [1]. Besides the sample size N we need an additional parameter $Q \ll N$ denoting the "warm-up". Define
$$A_l(\tau, r) = \begin{cases} l\text{: if there exists no } 1 \leq i \leq l \\ \qquad \text{such that } \overline{x}_i^{(r)} = \overline{x}_l^{(r)}, \text{ and} \\ \max\{i \geq 1 \,:\, \overline{x}_i^{(r)} = \overline{x}_l^{(r)}\} \,:\, \text{otherwise.} \end{cases}$$
and let $f_{T_U}(\tau, r, Q, N) = \frac{1}{N-Q}\sum_{l=Q+1}^{N} \log(A_l(\tau, r))$. Note that $\hat{F}_c$ is calculated from nonoverlapping vectors $(\overline{x}_l^{(r)})$ and that the total number of sample states needed for the test equals $N \cdot r$. For $N \to \infty$ and fixed Q, f_{T_U} converges to the expectation $E[\log(\overline{T}^{(r)})]$ almost surely by the linearity of the expectation and by the Birkhoff Ergodic Theorem.
2. the statistic $\hat{F}_e := \frac{1}{r\log(m)}(f_{T_U} - E_{r\log(m)} + r\log(m))$. This one is similar to $\hat{F}_c$, but instead of using the asymptotic value C we correct f_{T_U} by the numerically computed expectations $E_{r\log(m)} \approx E[f_{T_U}]$ of an independent equidistributed chain from column 2 in Table I in [1]. For small values of $r\log(m)$, this seems to provide higher accuracy, especially if the entropy is nearly maximal. In all our examples, $\log(m) \in \mathbb{N}$ is an integer.
3. the statistic $\hat{I} := \frac{1}{m}(H(\hat{\mathbf{P}}^{(r)}(N)) - H(\hat{\mathbf{P}}^{(r-1)}(N)))$, $N \in \mathbb{N}$, where $\hat{\mathbf{P}}^{(r)}(N)$ is calculated from the sequence $(x_l^{(r)})_{1\leq l\leq N}$ of overlapping vectors of successive states. The total number of sample states needed for $\hat{I}$ equals N, only, due to the cyclic definition of the overlapping $x_l^{(r)}$.

All the statistics can be used to estimate $Hc := \frac{1}{\log(m)}H(S, \mathbb{P})$ for large enough r. We use the normalization $\frac{1}{\log(m)}$ in order to get the per-bit-entropy which equals 1 for any independent equidistributed chain $(S, \mathbb{P})$ regardless of $m \geq 2$. The statistic $\hat{I}$ estimates $Hp := \frac{1}{\log(m)}H(\mathbf{P})$ if $r = 1$. Note that for a fixed sample size $N \in \mathbb{N}$ and parameter $Q \in \mathbb{N}$, each statistic scans the same number of r-dimensional vectors, but f_C and f_E need an r-times longer sample path to do so. Whereas the time complexity of $\hat{F}_c$ and $\hat{F}_e$ is $O(N)$ regardless of r and m, that of $\hat{I}$ should be written $O(N) + O(r \cdot m)$.

For any setting of the parameters m, δ_1, and δ_2 we calculate the per-bit-entropies Hc and Hp. The vector of parameters $par = (m, \delta_1, \delta_2, Hc, Hp)$ is indicated below the single plots. We use GNU C to implement the chain simulation and the calculation of the sample values of $\hat{F}_c$, $\hat{F}_e$, and $\hat{I}$, and Wolfram's *Mathematica* 3.0 system for the calculation of Hc and Hp and for the graphical output.

Our first goal is to get an impression of the speed of convergence as we increase the total sample size N of the test. We vary the parameters $(\log(N), \log(Q))$ in the set $\{(6,5),(8,7),(10,8),(12,10),(14,11),(16,12),(18,13),(20,14),(22,15),(24,16)\}$.
Here, the values of Q have been chosen such, that the statistics $\hat{F}_c$ and $\hat{F}_e$ have a reasonable "warm-up" before actually scanning return-time values. For every such pair, we simulate 32 sample paths $\tau_1, \ldots, \tau_{32}$ of length $N \cdot r$ and calculate the mean values $Fc := \frac{1}{32}\sum_{i=1}^{32} \hat{F}_c(\tau_i)$, $Fe := \frac{1}{32}\sum_{i=1}^{32} \hat{F}_e(\tau_i)$, and $I := \frac{1}{32}\sum_{i=1}^{32} \hat{I}(\tau_i)$ (where the latter uses only one rth of the data in each sample path τ_i), and the according sample standard deviations, $\sigma(\hat{F}_c) = (\frac{1}{31}\sum_{i=1}^{32}(\hat{F}_c(\tau_i) - Fc)^2)^{1/2}$, $\sigma(\hat{F}_e)$, and $\sigma(\hat{I})$.

In the plots, the values of Fc, Fe, and I are indicated by the symbols triangle, diamond, and star, respectively, and the sample standard variation is plotted as error bar. The left hand plot in Figure 1 shows that in an independent chain, the parameter $r = 1$ yields correct estimates for reasonable sample sizes with the statistic I, whereas r is to small to get the correct entropy from Fc or Fe. The sample standard deviation decreases as N increases for all three statistics. The right hand plot shows that the speed of convergence in N for higher r is a little better for the statistics Fc and Fe, but recall, that they scan an r times longer sample path. Clearly, one has to choose a much larger sample size for dimension $r = 4$ as for dimension $r = 1$ to obtain reasonable estimates. In both Figures, Hp equals Hc since the chains are independent.

In a second series of examples we study the impact of the parameter r in Figures 2 and 3. Recall, that we suspect I to yield the per-bit-entropy of the stable distribution, Hp, if $r = 1$, and that of the chain, Hs, if $r > 1$, which is well demonstrated in Figure 3, where Hc differs significantly from Hp. The statistics Fc and Fe ignore Hp and approach the entropy Hc as r increases, which is exactly what we await from Corollary V.3. The sample sizes N have been chosen with respect to r in such a way, that the estimators yielded reasonable values and did not change significantly on further increase of N.

VII. Conclusion

In our paper we study the relation between serial- or relative frequency- tests out of the family of (modified) generalized ϕ-divergences on the one hand, and the assessment of randomness based on estimators for the entropy, on the other hand. We select the loss function $\varphi_1(u) = 1 - u + u\ln(u)$ which originally led to Kullback's I-divergence and have the following three applications: For goodness-of-fit tests in the first place, one has to parameterize the generalized ϕ-divergence $\tilde{I}_{\overline{\Sigma},\varphi}$ with a weak inverse of the asymptotic covariance matrix of the vector of relative frequencies as specified by the null hypothesis. Such an inverse exists e.g. for data sampled from an irreducible aperiodic finite chain. If the null hypothesis is an independent chain with respect to a vector $\mathbf{P}$, a simple weak inverse $\overline{V}^{(r)}$ can not only be given for the process $(X_l)_{l\in\mathbb{N}}$ of states (where $r = 1$ and $\overline{V}^{(1)}$ is diagonal and leads to the original Kullback I-divergence measure), but also for the process $(\tilde{X}_l^{(r)})_{1\le l\le n}$ of cyclic overlapping r-tuples of successive states.

The modified generalized divergence $\tilde{I}_{(\overline{\Sigma},\varphi,q)}$, secondly, gives a slightly different framework and allows for the additive structure of the aforementioned inverses $\overline{V}^{(r)}$. $\tilde{I}_{(\overline{\Sigma},\varphi,q)}$ comprises the well-known overlapping serial test and extends it to more general loss functions. If we again employ $\varphi_1(u)$, we arrive at the modified generalized I-divergence, $\tilde{I}_{\varphi_1}^{(r)}$, for short. Finally, the linear transform $\log(m) - \frac{\tilde{I}_{\varphi_1}^{(r)}}{2n\ln(2)}$ of this goodness-of-fit statistic with the null-hypothesis of an independent equidistributed process gives an estimator for the entropy of data sampled from an ergodic chain, provided that the tuple size is larger than the order of the chain. Thus, the whole setup which was already well known for i.i.d. processes is extended to the class of processes with finite memory and time independent transition probabilities.

We also studied the relationship to entropy estimates built on the notion of return-time and reported some empirical results from a sample study. Since expected return-time and expected relative frequency are inversely proportional in ergodic chains, we did not expect much difference in terms of efficiency between the two types of entropy estimates which was well demonstrated in our samples. The main difference lies in the meaning of the dimension parameter, which is clearly connected to the memory or order of the chain for the frequency based estimator, whereas this connection is blurred in the case of the return-time based statistics. Further theoretical work has to be done with respect to the estimation of the variance and asymptotic efficiency of these statistics.

VIII. Appendix

For $\mathbf{i} \in S^r$, let $\delta_{\mathbf{i}}^{(r)} = \frac{p_{\mathbf{i}_r\mathbf{i}_1}}{p_{\mathbf{i}_1}}$, $\epsilon_{\mathbf{i}}^{(r)} = \frac{\pi_{\mathbf{i}}^{(r)}}{1-\pi_{\mathbf{i}}^{(r)}}(1 - \delta_{\mathbf{i}}^{(r)}\pi_{\mathbf{i}}^{(r)})$, and $\tau_{\mathbf{i}}^{(r)} = \frac{1-\delta_{\mathbf{i}}^{(r)}\pi_{\mathbf{i}}^{(r)}}{1-\epsilon_{\mathbf{i}}^{(r)}}$.

Lemma VIII.1: For $k > 1$, $P(\mathbf{i}, k, r) = \tau_{\mathbf{i}}^{(r)}\epsilon_{\mathbf{i}}^{(r)}(1 - \epsilon_{\mathbf{i}}^{(r)})^{(k-1)}$.

Proof: By the Markov property,

$$P\left[\min\left\{l \in \mathbb{N} \;:\; \overline{X}_{l+1}^{(r)} = \mathbf{i}\right\} = k \middle| \overline{X}_1^{(r)} = \mathbf{i}\right] =$$

$$= P\left[\overline{X}_2^{(r)} \neq \mathbf{i} \middle| \overline{X}_1^{(r)} = \mathbf{i}\right] \cdot \cdot P\left[\overline{X}_3^{(r)} \neq \mathbf{i} \middle| \overline{X}_2^{(r)} \neq \mathbf{i}\right] \cdots$$

$$\cdots P\left[\overline{X}_k^{(r)} \neq \mathbf{i} \middle| \overline{X}_{k-1}^{(r)} \neq \mathbf{i}\right] \cdot P\left[\overline{X}_{k+1}^{(r)} = \mathbf{i} \middle| \overline{X}_k^{(r)} \neq \mathbf{i}\right].$$

As to the middle terms, for $k > 2$,

$$P\left[\overline{X}_k^{(r)} \neq \mathbf{i} \middle| \overline{X}_{k-1}^{(r)} \neq \mathbf{i}\right] = 1 - P\left[\overline{X}_k^{(r)} = \mathbf{i} \middle| \overline{X}_{k-1}^{(r)} \neq \mathbf{i}\right] =$$

$$= 1 - \frac{1}{1-\pi_{\mathbf{i}}^{(r)}} \sum_{\mathbf{j}\in S^r\setminus\{\mathbf{i}\}} \left(\frac{\pi_{\mathbf{i}}^{(r)} p_{\mathbf{j}_r \mathbf{i}_1}}{p_{\mathbf{i}_1}} \pi_{\mathbf{j}}^{(r)} \right) = 1 - \epsilon_{\mathbf{i}}^{(r)},$$

owing to the stationarity. Thus the last term becomes $P\left[\overline{X}_{k+1}^{(r)} = \mathbf{i} \middle| \overline{X}_k^{(r)} \neq \mathbf{i}\right] = 1 - (1 - \epsilon_{\mathbf{i}}^{(r)}) = \epsilon_{\mathbf{i}}^{(r)}$. We also get $P\left[\overline{X}_2^{(r)} \neq \mathbf{i} \middle| \overline{X}_1^{(r)} = \mathbf{i}\right] = 1 - \delta_{\mathbf{i}} \cdot \pi_{\mathbf{i}}^{(r)} = (1 - \epsilon_{\mathbf{i}}^{(r)})\tau_{\mathbf{i}}^{(r)}$, for the first term, so that $P(\mathbf{i}, k, r) = \tau_{\mathbf{i}}^{(r)} \epsilon_{\mathbf{i}}^{(r)} (1 - \epsilon_{\mathbf{i}}^{(r)})^{(k-1)}$. The same formula holds in the case $k = 2$, where no middle terms are present. ■

Lemma VIII.2: As $r \to \infty$, $\pi_{\mathbf{i}}^{(r)} \to 0$, $\epsilon_{\mathbf{i}}^{(r)} \to 0$, and $\tau_{\mathbf{i}}^{(r)} \to 1$ uniformly in $\mathbf{i}$.

Proof: By the ergodicity and finiteness of the chain $(S, \mathbb{P})$ there exists an integer $k \in \mathbb{N}$ such that for every $(i, j) \in S^2$, the kth order transition probability $p_{ij}^{(k)}$ is strictly positive. Thus $\max\{p_{ij}^{(k)} : (i, j) \in S^2\} =: \Delta < 1$ and, consequently, $\pi_{\mathbf{i}}^{(k)} \leq p_{\mathbf{i}_1} p_{\mathbf{i}_1 \mathbf{i}_k}^{(k)} \leq \Delta$ for every $\mathbf{i} \in S^k$. By induction one immediately gets $\pi_{\mathbf{i}}^{(l\cdot k)} \leq \Delta^l$ for every integer $l \in \mathbb{N}$ and every $\mathbf{i} \in S^{l\cdot k}$. Finally, if $l \cdot k \leq r < (l+1) \cdot k$ and $\mathbf{i} \in S^r$, we have $\pi_{\mathbf{i}}^{(r)} \leq \pi_{\mathbf{i}}^{(l\cdot k)} \leq \Delta^l$, so that for arbitrary $r \in \mathbb{N}$ and $\mathbf{i} \in S^r$, $\pi_{\mathbf{i}}^{(r)} \leq \Delta^{\lfloor \frac{r}{k} \rfloor} \to 0$, as $r \to \infty$. Clearly, the continuity of $\epsilon_{\mathbf{i}}^{(r)}$ and $\tau_{\mathbf{i}}^{(r)}$ viewn as functions of $\pi_{\mathbf{i}}^{(r)}$ in the point $\pi_{\mathbf{i}}^{(r)} = 0$ implies the uniform convergence claimed in the Lemma. This completes the proof. ■

Acknowledgments

The author wishes to thank Peter Hellekalek (head of the author's supporting FWF project P13480-MAT), and Ferdinand Österreicher, both from the Department of Mathematics, University of Salzburg, Austria, for the many fruitful discussions and hints.

References

[1] U.M. Maurer, "A universal statistical test for random bit generators," *J. Cryptology*, vol. **5**, pp. 89–105, 1992.

[2] I. Csiszár, "The method of types," *IEEE Transactions on Information Theory*, vol. **44**, no. 6, pp. 2505–2523, 1998.

[3] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Magyar Tud. Akad. Mat. Kutató Int. Közl*, vol. **8**, pp. 85–108, 1963.

[4] F. Österreicher and I. Vajda, "Statistical information and discrimination," *IEEE Trans. on Information Theory*, vol. 39, no. 3, pp. 1036–1039, 1993.

[5] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.," *Philos. Magazine Series*, vol. **50** , no. 5, pp. 157–172, 1900.

[6] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[7] M. Salicrú, D. Morales, M. L. Menéndez, and L. Pardo, "On the applications of divergence type measures in testing statistical hypotheses," *Journal of Multivariate Analysis*, vol. **51**, pp. 372–391, 1994.

[8] S. Wegenkittl, "Generalized ϕ-Divergence and Frequency Analysis in Markov Chains," Ph.D. thesis, Universität Salzburg, Österreich, 1998, HTML version: http://random.mat.sbg.ac.at

[9] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*, Wiley Series in Probability and Mathematical Statistics. Wiley and Sons, 1971.

[10] V. Romanovsky, *Discrete Markov Chains*, Wolters-Noordhoff Publishing, Groningen, Netherlands, 1970.

[11] I. J. Good, "The serial test for sampling numbers and other tests for randomness," *Proc. Cambridge Philosophical Society*, vol. **49**, pp. 276–284, 1953.

[12] S. Wegenkittl, "Empirical testing of pseudorandom number generators," M.S. thesis, Universität Salzburg, Österreich, 1995, HTML version: `http://random.mat.sbg.ac.at/~ste/dipl`.

[13] G. Marsaglia, "A current view of random number generators," in *Computer Science and Statistics: The Interface*, L. Billard, Ed. 1985, pp. 3–10, Elsevier Science Publishers B.V.

[14] S. Wegenkittl, "A generalized ϕ-divergence for asymptotically multivariate normal models," 1999, Submitted for publication to Journal of Multivariate Analysis.

[15] P. L'Ecuyer, R. Simard, and S. Wegenkittl, "Sparse serial tests of uniformity for random number generators," Submitted for publication, 1998.

[16] T. M. Cover and J. A. Thomas, *Elements of Information theory*, Wiley and Sons, New York, 1991.

[17] H. Leeb and S. Wegenkittl, "Inversive and linear congruential pseudorandom number generators in empirical tests.," *ACM Trans. Modeling and Computer Simulation*, vol. **7**, no. 2, pp. 272–286, 1997.

[18] G.H. Choe and D.H. Kim, "The probability distribution of the first return time," Submitted for publication, 1999.

[19] G.H. Choe and D.H. Kim, "The first return time test of pseudorandom number generators," Submitted for publication, 1999.

[20] A. Wyner and J. Ziv, "Some asymptotic properties of the entropy of stationary ergodic data source with applications to data compression," *IEEE Trans. Information Theory*, vol. **35**, pp. 1250–1258, 1989.

[21] D. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. Information Theory*, vol. **39**, pp. 78–93, 1993.

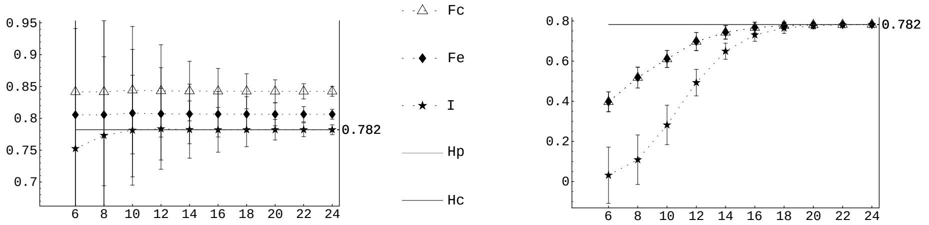
[22] J. S. Coron and D. Naccache, "An Accurate Evaluation of Maurer's Universal Test," in *Proceedings of Selected Areas in Cryptography 98*. 1998, Lecture Notes in Computer Science 1556, pp. 57–71, Springer.

[23] G.H. Choe and D.H. Kim, "Average convergence rate of the first return time," Submitted for publication, 1999.

[24] P. Hellekalek, "Good random number generators are (not so) easy to find," *Mathematics and Computers in Simulation*, vol. **46**, pp. 485–505, 1998.

[25] S. Wegenkittl, "The pLab picturebook: Load tests and ultimate load tests, part I," Report no. 1, pLab – reports, University of Salzburg, 1997.

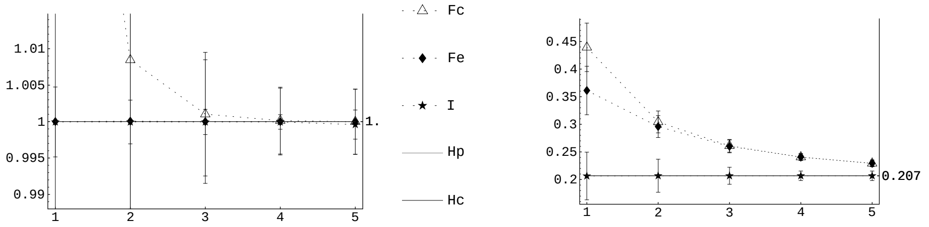
Stefan Wegenkittl received the B.S. and M.S. degrees (both in Mathematics) from the University of Salzburg and completed his Ph.D. on Mathematics at the same University. His research interests include mathematical statistics (φ-divergence) and metric number theory, special areas of interest being stochastic processes (dynamic systems and chaos), distance measures for probability measures, stochastic simulation, Markov chain Monte Carlo, and empirical tests for pseudorandom number generators, see `http://random.mat.sbg.ac.at`. He is currently working as Manager of Applied Mathematics in the area of Bioinformatics and Medical Sciences.



$par = (2^4, 3, 0, 0.7821, 0.7821),$
$r = 1$

$par = (2^4, 3, 0, 0.7821, 0.7821),$
$r = 4$

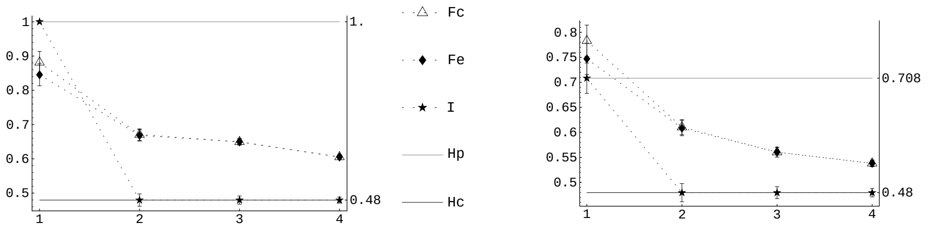
Fig. 1. Samples of Fc, Fe, and I, where $\log(N) \in \{6, 8, \ldots, 24\}$ on the horizontal axis.



$par = (2^3, 0, 0, 1.0, 1.0)$

$par = (2^3, 15, 0, 0.2068, 0.2068)$

Fig. 2. Samples of Fc, Fe, and I, where $r \in \{1, \ldots, 5\}$ on the horizontal axis.



$par = (2^4, -10, 1, 0.4796, 1.0)$

$par = (2^4, -10, 8, 0.4796, 0.7084)$

Fig. 3. Samples of Fc, Fe, and I, where $r \in \{1, \ldots, 4\}$ on the horizontal axis.