Chapter 6
Metric Permutation Entropy

The word "entropy" was coined by the German physicist R. Clausius (1822–1888), who introduced it in thermodynamics in 1865 to measure the amount of energy in a system that cannot produce work. The fact that the entropy of an isolated system never decreases constitutes the second law of thermodynamics and clearly shows the central role of entropy in many-particle physics. The direction of time is then explained as a consequence of the increase of entropy in all irreversible processes. Later on the concept of entropy was given a microscopic interpretation in the foundational works of L. Boltzmann (1844–1906) on gas kinetics and statistical mechanics [184]. The celebrated Boltzmann's equation reads in the usual physical notation

$$S = k_{\mathrm{B}} \ln \Omega, \tag{6.1}$$

where here S is the entropy of the thermodynamical system, k_{B} is a physical constant (called Boltzmann's constant, $k_{\mathrm{B}} = 1.3806504(24) \times 10^{-23}$ J/K) and Ω is the number of microscopic states consistent with the macroscopic constraints. In this realm, the entropy is a measure of the microscopic *disorder* of the system, the entropy being higher the more disordered the system.

In 1948 the word entropy came to the fore in the new context of information theory, coding theory, and cryptography through the seminal papers of C.E. Shannon[1] (1916–2001) [186]. This time, entropy measures the average *uncertainty* about the outcome of a random variable. More generally, the *entropy rate* measures the uncertainty per symbol (time unit, channel use, etc.) of a stationary stochastic process, eventually modeling an information source. Instead of associating entropy with uncertainty, one can alternatively speak of the average information gained by performing a random experiment. Entropy plays a paramount role in all information-related fields, being at the heart of the fundamental results.

[1] According to [64] "When Shannon had invented his quantity and consulted von Neumann on what to call it, von Neumann replied: 'Call it entropy. It is already in use under that name and besides, it will give you a great edge in debates because nobody knows what entropy is anyway.' "

J.M. Amigó, *Permutation Complexity in Dynamical Systems,*
Springer Series in Synergetics, DOI 10.1007/978-3-642-04084-9_6,

Shannon's ideas, properly transformed, were incorporated by A.N. Kolmogorov (1903–1987) into ergodic theory in 1958 [126] to measure the *randomness* of deterministic dynamical systems. Kolmogorov's proposal was improved a short time later by Sinai [189]. The result became the most important invariant in the theory of discrete and continuous dynamical systems.

Since then the concept of entropy has evolved along different ways: Rényi entropy, topological entropy, sequence entropy, Tsallis entropy, directional entropy, permutation entropy, epsilon–tau entropy, etc. The basics of Shannon entropy, metric (Kolmogorov–Sinai or measure-theoretical) entropy, and topological entropy are systematized in Annex B.

Permutation entropy, both in the metric version (this chapter) and in the topological version (next chapter), was introduced by Bandt, Keller, and Pompe in [29] (see [28] as well). The main ingredient of permutation entropy is the ordinal patterns we studied in Chap. 3. As we shall see below, the definition of the metric permutation entropy of an information source is formally the same as Shannon's entropy, except for the fact that now probabilities refer not to length-L blocks of symbols but to the length-L ordinal patterns realized by them (assuming, of course, that those symbols can be ordered).

On defining the metric permutation entropy of maps, we depart from [29] to follow basically Kolmogorov's strategy: coarse-grain the state space with a partition, apply the definition of (in our case, permutation) entropy to the resulting symbolic dynamics, and then refine successively the original partition into the partition into separate points. Moreover, the partitions used may be taken to be product, uniform partitions, making possible the numerical estimation of metric permutation entropy under rather general conditions. Most importantly, we shall show that metric permutation entropy converges to the conventional metric entropy for ergodic self-maps of n-dimensional intervals.

6.1 The Metric Permutation Entropy of a Finite-State Process

Let $\mathbf{X} = \{X_n\}_{n\in\mathbb{N}_0}$ be a random process with finite state space S (see Annex A.3). We take without restriction $S = \{1, 2, \ldots, |S|\}$. As noted in Example 2, the relation between length-L words and length-L ordinal patterns is in general many-to-one. This is due to the fact that ordinal patterns do not take into account the sizes of the elements being compared, but only their relative order. The same happens with the *ranks* or *rank variables*, which are the outputs of a random process $\mathbf{R} = \{R_n\}_{n\in\mathbb{N}_0}$ subsidiary of $\mathbf{X}$, defined as follows:

$$R_n = |\{X_i, 0 \le i \le n : X_i \le X_n\}| = \sum_{i=0}^{n} \delta(X_i \le X_n),$$

where as usual the δ-function of a proposition is 1 if it holds and 0 otherwise. By definition, R_n is a *discrete* random variable with range $\{1, \ldots, n+1\}$, and the sequence

$\mathbf{R} = \{R_n\}_{n\in\mathbb{N}_0}$ builds a discrete-time, non-stationary stochastic process. The point about introducing rank variables is that the relation between length-L ordinal patterns $\pi(x_n^{n+L-1})$ and length-L ranks $r_n^{n+L-1} = r_n, r_{n+1}, \ldots, r_{n+L-1}$ is one-to-one. The many-to-one relation between X_0^{L-1} and R_0^{L-1} will be written as

$$R_0^{L-1} = \text{rank}\,(X_0^{L-1}). \tag{6.2}$$

Ranks are specially useful in proofs.

Example 10 If, as in Example 2, $S = \{a, b, c\}$ with $a < b < c$ and $x_0^2 = c, a, a$, then $r_0^2 = 1, 1, 2$. All other words defining the same ordinal pattern $\pi(x_0^2) = \langle 1, 2, 0\rangle$ define also the same rank variables:

$$r_0^2 = 1, 1, 2 = \text{rank}\,(c, b, b) = \text{rank}\,(c, a, b) = \text{rank}\,(b, a, a).$$

Having defined the sibling concepts of ordinal patterns and rank variables of finite-alphabet sequences, we can proceed now very much the same way as we did when defining Shannon's entropy (rate) of stochastic processes or information sources in Sect. 1.1.1 (see also Annex B.1), this time though bookkeeping ordinal patterns instead of symbol blocks.

In this spirit, the *metric permutation entropy* of a stochastic process $\mathbf{X} = \{X_n\}_{n\in\mathbb{N}_0}$ is defined as

$$h^*(\mathbf{X}) = \lim_{L\to\infty} h^*(X_0^{L-1}), \tag{6.3}$$

provided the limit exists, where

$$h^*(X_0^{L-1}) = -\frac{1}{L}\sum_{x_0,\ldots,x_{L-1}} p(\pi(x_0^{L-1}))\log p(\pi(x_0^{L-1}))$$

is the *metric permutation entropy of order* $L \geq 2$ of $\mathbf{X}$. Here $p(\pi(x_0^{L-1}))$ is the probability for the length-L block $x_0^{L-1} = x_0, \ldots, x_{L-1}$ to be of type $\pi(x_0^{L-1}) \in \mathcal{S}_L$. Alternatively,

$$h^*(X_0^{L-1}) = -\frac{1}{L}\sum_{r_0,\ldots,r_{L-1}} p(r_0^{L-1})\log p(r_0^{L-1}) = h(R_0^{L-1}), \tag{6.4}$$

where $p(r_0^{L-1})$ is the probability for the block x_0^{L-1} to define the rank vector $r_0^{L-1} = r_0, \ldots, r_{L-1}$ (remember that the relation between $\pi(X_0^{L-1})$ and $R_0^{L-1} = \text{rank}(X_0^{L-1})$ is one-to-one). In both cases,

$$h^*(\mathbf{X}) = h(\pi(\mathbf{X})) = h(\mathbf{R}),$$

where $h(\,\cdot\,)$ denotes the Shannon entropy of the corresponding stochastic process.

In case that the random process $\mathbf{X}$ is stationary, there is still a third way to look at its metric entropy permutation. If $(S^{\mathbb{N}_0}, \mathcal{B}_\Pi(S), m, \Sigma)$ is the sequence space model of $\mathbf{X}$ (see Annex A.3), then the non-empty cylinder sets

$$C_\pi = \{(x_n) \in S^{\mathbb{N}_0} : x_0^{L-1} \text{ is of type } \pi \in \mathcal{S}_L\}$$

build a partition of $(S^{\mathbb{N}_0}, \mathcal{B}_\Pi(S), m)$ with $m(C_\pi) = \Pr\{\pi(X_0^{L-1}) = \pi\} = \Pr\{R_0^{L-1} = r_0^{L-1}\}$, where $R_0^{L-1} = \text{rank}\,(X_0^{L-1})$, and $1 \le r_k \le k+1$ for $k = 0, \ldots, L-1$. Therefore

$$h^*(X_0^{L-1}) = -\frac{1}{L} \sum_{\pi \in \mathcal{S}_L} m(C_\pi) \log m(C_\pi). \tag{6.5}$$

As a result, the permutation entropy is sensitive to the measures of non-trivial order relationships observed in a word, as the Shannon entropy is sensitive to the measures of the different word values themselves.

When stationarity is important, as in (6.5), we call $\mathbf{X}$ an information source or just a source.

In the next lemma we use the conditional entropy of a random variable Y given another random variable X, $H(Y|X)$, which is the expected value of the entropies of the conditional distributions averaged over the conditioning variable X (see Annex B, (B.5)).

Lemma 7 *Given an ergodic source* $\mathbf{X} = \{X_n\}_{n \in \mathbb{N}_0}$*, the equality*

$$\lim_{k \to \infty} H(R_k^{k+l} | X_0^{k-1}) = \lim_{k \to \infty} H(X_k^{k+l} | X_0^{k-1})$$

holds for all $l \ge 0$*.*

That is, given a sufficiently long tail of previously observed symbols, the later ranks can be predicted virtually as well as the symbols themselves. Heuristically, this is because the rank of a late variable is sensitive effectively to the cumulative distribution function of the source, approximated by the normalized sum of X_0^{k-1}. In turn, this means that the information contained in R_k is the same as the information in X_k.

Proof Consider $R_k = \sum_{i=0}^{k} \delta(X_i \le X_k)$. For $a \in S = \{1, \ldots, |S|\}$ define the *sample frequency* of the letter a in the word x_0^k, $k \ge 0$, to be

$$\vartheta_k(a) = \frac{1}{k+1} \sum_{i=0}^{k} \delta(X_i = a).$$

With the help of $\vartheta_k(a)$ we may express R_k in terms of X_i, $0 \le i \le k$, namely,

$$R_k(X_k) = (k+1) \sum_{a=1}^{X_k} \vartheta_k(a),$$

where we assume the outcomes $X_0, \ldots, X_k$ to be known. Then, the identity

$$\Pr\{R_k = y\} = \sum_{q=1}^{|S|} \Pr\{X_k = q\}\delta\left(R_k(q) = y\right) \tag{6.6}$$

gives us the probability for observing some R_k with value $y \in \{1, \ldots, k+1\}$ by means of $\Pr\{X_k = q\}$, $1 \leq q \leq |S|$. Since, given X_0^{k-1} ($k \geq 1$), R_k is a deterministic function of the random variable X_k, i.e., $\Pr\{R_k = y | X_k = q\} = \delta(R_k(q) = y)$, (6.6) can be seen as an application of the law of total probability.

Without loss of generality, we may first rearrange the sum in (6.6) to consider only those symbol values q with non-zero $\Pr\{X_k = q\}$, summing to $N \leq |S|$. Expand the sum,

$$\begin{aligned}\Pr\{R_k = y\} = {} & \Pr\{X_k = 1\}\delta\left[y = (k+1)\vartheta_k(1)\right] \\ & + \Pr\{X_k = 2\}\delta\left[y = (k+1)(\vartheta_k(1) + \vartheta_k(2))\right] \\ & + \cdots + \Pr\{X_k = N\}\delta\left[y = (k+1)(\vartheta_k(1) + \cdots + \vartheta_k(N))\right].\end{aligned}$$

Suppose all the relevant sample frequencies $\vartheta_k(1), \ldots, \vartheta_k(N)$ are greater than zero. This means that for any y, only a single one of the δ-functions can be non-zero, and hence we have a one-to-one transformation taking non-zero elements from the distribution $\Pr\{X_k\}$ without change into some bin for $\Pr\{R_k\}$. Since entropy is invariant to a renaming of the bins, and the remaining zero probability bins add nothing to the entropy, we conclude that, if $\vartheta_k(a) > 0$ for all a where the true probability $\Pr\{X_k = a\} > 0$ (i.e., $a = 1, \ldots, N$ after a hypothetical rearrangement), then $H(R_k|X_0^{k-1}) = H(X_k|X_0^{k-1})$ for $k \geq 1$. Because of the assumed ergodicity, we can make the probability that $\vartheta_k(a) = 0$ when $\Pr\{X_k = a\} > 0$ to be arbitrarily small by taking k to be sufficiently large, and the claim follows for $l = 0$.

This construction can be extended without change to words X_k^{k+l} of arbitrary length $l + 1 \geq 1$ via

$$\begin{aligned}&\Pr\{R_k^{k+l} = y_0 \ldots y_l\} \\ &= \sum_{q_0,\ldots,q_l=1}^{N} \Pr\{X_k^{k+l} = q_0 \ldots q_l\}\delta(R_k(q_0) = y_0)\ldots\delta(R_{k+l}(q_l) = y_l).\end{aligned}$$

Observe that if $\vartheta_k(a) > 0$ for $1 \leq a \leq N$, then the same happens with $\vartheta_{k+1}(a), \ldots, \vartheta_{k+l}(a)$ and $H(R_k^{k+l}|X_0^{k-1}) = H(X_k^{k+l}|X_0^{k-1})$ follows. Again, ergodicity guarantees that there exist realizations of X_0^{k+l} with sufficiently large k, whose sample frequencies fulfill the said condition. □

Example 11 As way of illustration, suppose that $X_n = 0, 1$ are independent random variables with probability $\Pr\{X_n = 0\} = \Pr\{X_n = 1\} = \frac{1}{2}$. Given $x_0^{k-1} =$

$x_0 \dots x_{k-1} \in \{0,1\}^k$, set $N_0 = \left|\{i{:}x_i = 0 \text{ in } x_0^{k-1}\}\right|$, $0 \le N_0 \le k$. Consider the case $l = 1$ in Lemma 1. There are two possibilities:

(i) $0 \le N_0 \le k-1$. Then

$$\begin{aligned} x_k^{k+1} &= 0,0 \quad \Rightarrow \quad r_k^{k+1} = N_0+1, N_0+2, \\ x_k^{k+1} &= 0,1 \quad \Rightarrow \quad r_k^{k+1} = N_0+1, k+2, \\ x_k^{k+1} &= 1,0 \quad \Rightarrow \quad r_k^{k+1} = k+1, N_0+1, \\ x_k^{k+1} &= 1,1 \quad \Rightarrow \quad r_k^{k+1} = k+1, k+2. \end{aligned}$$

Each of these events has the joint probability

$$\Pr\{N_0 = \nu, R_k^{k+1} = r_k^{k+1}\} = \frac{\binom{k}{\nu}}{2^k} \cdot \frac{1}{4} = \frac{1}{2^{k+2}} \binom{k}{\nu}$$

and conditional probability

$$\Pr\{R_k^{k+1} = r_k^{k+1} | N_0 = \nu\} = \frac{1}{4},$$

where $0 \le \nu \le k-1$ and $r_k^{k+1} = (\nu+1, \nu+2)$, $(\nu+1, k+2)$, $(k+1, \nu+1)$, or $(k+1, k+2)$.

(ii) $N_0 = k$. Then

$$\begin{aligned} x_k^{k+1} = 0,0 \,\&\, x_k^{k+1} = 0,1 \,\&\, x_k^{k+1} = 1,1 \quad &\Rightarrow \quad r_k^{k+1} = k+1, k+2, \\ x_k^{k+1} = 1,0 \quad &\Rightarrow \quad r_k^{k+1} = k+1, k+1. \end{aligned}$$

These events have the joint probabilities

$$\begin{aligned} \Pr\left\{N_0 = k, R_k^{k+1} = (k+1, k+2)\right\} &= \frac{1}{2^k} \cdot \frac{1}{4} \cdot 3 = \frac{3}{2^{k+2}}, \\ \Pr\left\{N_0 = k, R_k^{k+1} = (k+1, k+1)\right\} &= \frac{1}{2^k} \cdot \frac{1}{4} = \frac{1}{2^{k+2}} \end{aligned}$$

and conditional probabilities

$$\begin{aligned} \Pr\left\{R_k^{k+1} = (k+1, k+2) | N_0 = k\right\} &= \frac{3}{4}, \\ \Pr\left\{R_k^{k+1} = (k+1, k+1) | N_0 = k\right\} &= \frac{1}{4}. \end{aligned}$$

From Annex (B.5) and (i)–(ii), we get

$$\begin{aligned} H(R_k^{k+1}|X_0^{k-1}) &= -4 \times \sum_{\nu=0}^{k-1} \frac{1}{2^{k+2}} \binom{k}{\nu} \log \frac{1}{4} - \frac{3}{2^{k+2}} \log \frac{3}{4} - \frac{1}{2^{k+2}} \log \frac{1}{4} \\ &= 4 \times \frac{2}{2^{k+2}}(2^k - 1) + \frac{8}{2^{k+2}} - \frac{3}{2^{k+2}} \log 3 \\ &= 2\left(1 - \frac{3}{2^{k+3}} \log 3\right). \end{aligned}$$

On the other hand, since the random variables X_n are independent,

$$H(X_k^{k+1}|X_0^{k-1}) = H(X_k^{k+1}) = 2.$$

It follows that $H(R_k^{k+1}|X_0^{k-1})$ and $H(X_k^{k+1}|X_0^{k-1})$ coincide in the limit $k \to \infty$, as guaranteed by Lemma 7.

With Lemma 7 in hand, we turn to the main result.

Theorem 8 *For a finite-alphabet ergodic source* $\mathbf{X}$*, the permutation entropy exists and equals the metric entropy:* $h^*(\mathbf{X}) = h(\mathbf{X})$*.*

Proof We prove inequalities in both directions.

(a) $\limsup_{L\to\infty} h^*(X_0^{L-1}) \leq h(\mathbf{X})$. Given X_0^{L-1}, the corresponding rank variables are uniquely determined via $R_0^{L-1} = \text{rank}\,(X_0^{L-1})$. By [59, Chap. 2, Exercise 5], $H(\varphi(Z)) \leq H(Z)$ for any discrete random variable Z and function φ, so $H(R_0^{L-1}) \leq H(X_0^{L-1})$ and thus (see (6.4)),

$$\limsup_{L\to\infty} h^*(X_0^{L-1}) = \limsup_{L\to\infty} h(R_0^{L-1}) \leq \limsup_{L\to\infty} h(X_0^{L-1}) = h(\mathbf{X}).$$

(b) $\liminf_{L\to\infty} h^*(X_0^{L-1}) \geq h(\mathbf{X})$. There are several ways to prove this inequality. Consider, for instance,

$$\begin{aligned} &\liminf_{L\to\infty} h^*(X_0^{L-1}) \\ &= \liminf_{L\to\infty} \frac{1}{L} H(R_0^{L-1}) \\ &= \liminf_{L\to\infty} \frac{1}{L} \left(\left[H(R_{L-1}|R_0^{L-2}) + \cdots + H(R_{L^*+1}|R_0^{L^*}) \right] + H(R_0^{L^*}) \right) \end{aligned}$$

for any $L^* < L - 1$, where we have applied the chain rule for entropy (B.9). As $R_1^k = \text{rank}\,(X_1^k)$ we apply the data processing inequality $H(Y|\varphi(Z)) \geq H(Y|Z)$ [59] to all elements of the first term on the right-hand side:

$$\begin{aligned} &\liminf_{L\to\infty} h(X_0^{L-1}) \\ &\geq \liminf_{L\to\infty} \frac{1}{L} \left(\left[H(R_{L-1}|X_0^{L-2}) + \cdots + H(R_{L^*+1}|X_0^{L^*}) \right] + H(R_0^{L^*}) \right). \end{aligned}$$

By Lemma 7 with $l = 0$, for any $\varepsilon > 0$ there exists some L^* such that

$$\left| H(X_L|X_0^{L-1}) - H(R_L|X_0^{L-1}) \right| < \varepsilon$$

for $L > L^*$, so

$$\begin{aligned}
&\liminf_{L\to\infty} h(X_0^{L-1}) \\
&> \liminf_{L\to\infty} \left(\frac{1}{L}\left[H(X_{L-1}|X_0^{L-2}) + \cdots + H(X_1|X_0) + H(X_0) \right] \right. \\
&\left. + \frac{1}{L}\left[H(R_0^{L^*}) - H(X_0^{L^*}) \right] - \left(\frac{L - L^* - 1}{L} \right) \varepsilon \right) \\
&= h(\mathbf{X}) - \varepsilon,
\end{aligned}$$

since $H(X_0^{L^*}) = H(X_0) + H(X_1|X_0) + \cdots + H(X_{L^*}|X_0^{L^*-1})$ (B.9).
The existence of the limit and equality follows from (a) and (b). □

Observe in the proof of Theorem 8 that the ergodicity hypothesis was used only in part (b) via Lemma 7, while part (a) is completely general. We highlight this particular result in the following corollary for further reference.

Corollary 4 *For finite-alphabet sources* **X**,

$$\limsup_{L\to\infty} h^*(X_0^{L-1}) \leq h(\mathbf{X})$$

holds.

In order to deal further with the general, nonergodic case, we appeal to the theorem on ergodic decompositions [114]: if Ω is a compact metrizable space and $T{:}(\Omega, \mathcal{B}, \mu) \to (\Omega, \mathcal{B}, \mu)$ is a continuous transformation, then there is a partition of Ω into T-invariant subsets Ω_w, each equipped with a sigma-algebra $\mathcal{B}_w$ and a probability measure μ_w, such that T acts ergodically on each probability space $(\Omega_w, \mathcal{B}_w, \mu_w)$, the indexing set being another probability space $(W, \mathcal{F}, \nu)$. Furthermore,

$$\mu(E) = \int_W \int_E d\mu_w d\nu(w) = \int_W \mu_w(E) d\nu(w) \quad (E \in \mathcal{B}).$$

The family $\{\mu_w{:}w \in W\}$ is called the *ergodic decomposition* of μ.

If Σ is the shift on the (compact, metric) sequence space $(S^{\mathbb{N}_0}, \mathcal{B}_\Pi(S), m)$, the indexing set can be taken to be $S^{\mathbb{N}_0}$, i.e.,

$$m(C) = \int_{S^{\mathbb{N}_0}} \int_C dm_s dm(s) = \int_{S^{\mathbb{N}_0}} m_s(C) dm(s) \quad (C \in \mathcal{B}_\Pi(S)), \tag{6.7}$$

where $m_{\Sigma(s)} = m_s$ [89]. This result shows that any source which is not ergodic can be represented as a mixture of ergodic subsources. The next lemma states that such a decomposition holds also for the entropy.

Lemma 8 (Ergodic Decomposition of the Entropy) [89] *Let* $(S^{\mathbb{N}_0}, \mathcal{B}_\Pi(S), m, \Sigma)$ *be the sequence space model of a stationary finite-alphabet random process* $\mathbf{X} = \{X_n\}_{n\in\mathbb{N}_0}$. *Let* $\{m_s : s \in S^{\mathbb{N}_0}\}$ *be the ergodic decomposition of m. If* $h_{m_s}(\mathbf{X})$ *is m-integrable, then*

$$h(\mathbf{X}) = \int_{S^{\mathbb{N}_0}} h_{m_s}(\mathbf{X}) dm(s). \tag{6.8}$$

Theorem 9 *Under the assumptions of Lemma 8,*

$$\liminf_{L\to\infty} h^*(X_0^{L-1}) \geq h(\mathbf{X}) \tag{6.9}$$

for any finite-alphabet source $\mathbf{X}$.

Proof Fix $L \geq 2$. From (6.5) and (6.7),

$$\begin{aligned} h^*(X_0^{L-1}) &= -\frac{1}{L} \sum_{\pi\in\mathcal{S}_L} \left(\int_{S^{\mathbb{N}_0}} m_s(C_\pi) dm(s) \right) \log \left(\int_{S^{\mathbb{N}_0}} m_s(C_\pi) dm(s) \right) \\ &\geq -\frac{1}{L} \sum_{\pi\in\mathcal{S}_L} \left(\int_{S^{\mathbb{N}_0}} m_s(C_\pi) \log m_s(C_\pi) dm(s) \right) \qquad (6.10) \\ &= \int_{S^{\mathbb{N}_0}} \left(-\frac{1}{L} \sum_{\pi\in\mathcal{S}_L} m_s(C_\pi) \log m_s(C_\pi) \right) dm(s) \\ &= \int_{S^{\mathbb{N}_0}} h^*_{m_s}(X_0^{L-1}) dm(s), \end{aligned}$$

where in (6.10) we have used Jensen's inequality,

$$\Phi\left(\int_{S^{\mathbb{N}}} f d\mu \right) \leq \int_{S^{\mathbb{N}}} \Phi \circ f d\mu,$$

with $\Phi(t) = t \log t$ convex in $[0, \infty)$ and $f(s) = m_s(C_\pi) \geq 0$.

Therefore,

$$\begin{aligned}\liminf_{L\to\infty} h^*(X_0^{L-1}) &\geq \liminf_{L\to\infty} \int_{S^{\mathbb{N}_0}} h^*_{m_s}(X_0^{L-1}) dm(s) \\ &\geq \int_{S^{\mathbb{N}_0}} \left(\liminf_{L\to\infty} h^*_{m_s}(X_0^{L-1}) \right) dm(s) \qquad (6.11) \\ &= \int_{S^{\mathbb{N}_0}} h^*_{m_s}(\mathbf{X}) dm(s),\end{aligned}$$

where we have applied Fatou's lemma in (6.11) to the sequence of positive and (by hypothesis) m-measurable functions $h^*_{m_s}(X_0^{L-1})$. Observe that $h^*_{m_s}(\mathbf{X})$ exists for all $s \in S^{\mathbb{N}_0}$ (and is m-integrable as a function of s) since $h^*_{m_s}(\mathbf{X}) = h_{m_s}(\mathbf{X})$ by Theorem 8 ($\mathbf{X}$ is ergodic with respect to m_s). Therefore,

$$\liminf_{L\to\infty} h^*(X_0^{L-1}) \geq \int_{S^{\mathbb{N}_0}} h_{m_s}(\mathbf{X}) dm(s) = h(\mathbf{X})$$

by (6.8). □

Corollary 4 and Theorem 9 yield the following result.

Corollary 5 *Under the assumptions of Lemma 8, $h^*(\mathbf{X}) = h(\mathbf{X})$ holds for any finite-alphabet source* $\mathbf{X}$.

6.2 Permutation Metric Entropy of Maps

In this section we shall use the previous results on finite-alphabet stochastic processes to show that the equality between permutation and metric entropies holds also for ergodic self-maps on domains homeomorphic to q-dimensional compact intervals.

We say that a set $D \subset \mathbb{R}^q$ is a (q-dimensional) *simple domain* if it is homeomorphic to a q-dimensional compact interval (hence D is compact). In particular, one-dimensional simple domains are close intervals. As a subset of $\mathbb{R}^q$, D is also ordered. Let D be a q-dimensional simple domain and $f{:}D \to D$ a μ-preserving map, with μ being a probability measure on $(D, \mathcal{B} \cap D)$ and $\mathcal{B}$ being the Borel sigma-algebra of $\mathbb{R}^q$. In order to define the permutation entropy of f, consider a q-dimensional compact interval $I \supset D$ and product partitions

$$\iota = \prod_{k=1}^{q} \{I_{1,k}, \ldots, I_{N_k,k}\} \qquad (6.12)$$

of I into $|\iota| = N_1 \cdots N_q$ subintervals of lengths $\Delta_{j,k}$, $1 \leq j \leq N_k$, in each coordinate k. As for the norm of ι (see (1.13)), the perhaps most popular are the *Euclidean norm*,

$$\|\iota\| = \max_{j_1,\dots j_q} \left(\sum_{k=1}^{q} \Delta_{j_k,k}^2 \right)^{1/2} =: \|\iota\|_2 \tag{6.13}$$

(i.e., $\|\iota\|_2$ is the longest diagonal of the bins $I_{j_1,1} \times \cdots \times I_{j_q,q} \in \iota$) and the *supremum norm*,

$$\|\iota\| = \max_{j,k} \Delta_{j,k} =: \|\iota\|_\infty . \tag{6.14}$$

For definiteness, the intervals are lexicographically ordered in each dimension, that is, points in $I_{j,k}$ are smaller than points in $I_{j+1,k}$ and, for the multiple dimensions, $I_{j,k} < I_{j,k+1}$, so there is an order relation between all the N partition elements, and we can enumerate them with a single index $i \in \{1, \dots, |\iota|\}$:

$$\iota = \{I_i : 1 \le i \le |\iota|\}, \quad I_i < I_{i+1}$$

(i.e., points in I_i are smaller than points in I_{i+1}).

Below we shall consider refinements of product and general partitions. As usual we write $\alpha \le \beta$ to mean that the partition β is a *refinement* of the partition α (of $(D, \mathcal{B} \cap D)$ or of any other measurable space for that matter), meaning that the elements of α are unions of the elements of β. By an *increasing sequence of partitions* we mean therefore a sequence of partitions, $(\alpha_n)_{n\in\mathbb{N}}$, such that $\alpha_n \le \alpha_{n+1}$ for all n. If, as in the present case, the state space is a product space, then by a *product refinement* of partition (6.12) we mean any product partition of I obtained by subdividing some or all of the intervals $\{I_{1,k}, \dots, I_{N_k,k}\}$, $1 \le k \le q$.

Furthermore, let κ be the partition of D defined as

$$\kappa = \iota \cap D = \{I_i \cap D \ne \emptyset : 1 \le i \le |\iota|\} = \{K_j : 1 \le j \le |\kappa|\}.$$

In words, κ consists of all subintervals $I_i \in \iota$ contained in the interior of D, together with the overlaps with D of those I_i that intersect the boundary of D. Partitions κ of the form $\kappa = \iota \cap D$, where ι is a product partition and D a simple domain, will be called *quasi-product partitions*; if, moreover, ι is a box (i.e., uniform) partition, κ will be called a *quasi-box partition*. For simplicity, we set $\|\kappa\| = \|\iota\|$.

Next let $\mathbf{X}^\kappa = \{X_n^\kappa\}_{n\in\mathbb{N}_0}$ be the symbolic dynamics associated with $f : D \to D$ with respect to the partition κ:

$$X_n^\kappa(x) = j \quad \text{if} \quad f^n(x) \in K_j,\ n = 0, 1, \dots .$$

Hence $\mathbf{X}^\kappa$ is a stationary, $|\kappa|$-state random process on $(D, \mathcal{B} \cap D, \mu)$ with alphabet $S^\kappa = \{1, \dots, |\kappa|\}$.

Example 12 If $I = [0, 1]$ and $\kappa = \{K_j : 1 \le j \le 10^k\}$, with $K_j = [(j-1)10^{-k}, j10^{-k})$ for $1 \le j \le 10^k - 1$ and $K_{10^k} = [1 - 10^{-k}, 1]$, then $\mathbf{X}^\kappa$ can be written as follows: $X_n^\kappa(x) = \lfloor f^n(x) \cdot 10^k \rfloor + 1$ for $0 \le x < 1$ and $X_n^\kappa(1) = 10^k$.

According to (B.16) (with $\alpha = \kappa$), the entropy of the symbolic dynamics $\mathbf{X}^\kappa$ equals the metric entropy of f with respect to κ:

$$h_\mu(f,\kappa) = h_\mu(\mathbf{X}^\kappa). \tag{6.15}$$

If we take now an increasing sequence of product refinements $\kappa \equiv \kappa_0 \leq \kappa_1 \leq \cdots$ such that $\|\kappa_n\| \to 0$, then we deduce from Theorem 25 that $h_\mu(f) = \lim_{n\to\infty} h_\mu(\mathbf{X}^{\kappa_n})$. This suggests to define the metric permutation of f as $h^*_\mu(f) = \lim_{n\to\infty} h^*_\mu(\mathbf{X}^{\kappa_n})$. The fact that the limit $n \to \infty$ proceeds by successive refinements of κ_0 and the way product partitions are being numbered guarantees that the order relations are preserved. This means, in particular, that if $X_k^{\kappa_n}(x) = i < j = X_{k+1}^{\kappa_n}(x)$ $(1 \leq i,j \leq |\kappa_n|)$, then $X_k^{\kappa_{n+1}}(x) = i' < j' = X_{k+1}^{\kappa_{n+1}}(x)$ $(1 \leq i',j' \leq |\kappa_{n+1}|)$ for all $x \in D$ and $k \in \mathbb{N}_0$. Thus $h^*_\mu(f)$ has a good chance to exist.

Definition 3 Given a measure-preserving dynamical system $(D, \mathcal{B} \cap D, \mu, f)$, and a lexicographically ordered, quasi-product partition κ_0 of $(D, \mathcal{B} \cap D, \mu)$, the metric permutation entropy of f with respect to the measure μ is defined by

$$h^*_\mu(f) = \lim_{n\to\infty} h^*_\mu(\mathbf{X}^{\kappa_n}) \tag{6.16}$$

(provided the limit exists), where $(\kappa_n)_{n\in\mathbb{N}}$ is a sequence of successive product refinements of κ_0 such that $\|\kappa_n\| \to 0$ and $\mathbf{X}^{\kappa_n}$ is the symbolic dynamics of f with respect to κ_n.

It is plain that this definition is independent from the auxiliary interval $I \supset D$ used to construct κ_0 and also independent from the particular collection of product refinements κ_n used, as long as $\|\kappa_n\| \to 0$. This being the case, we may take quasi-box partitions in (6.16).

One practical reason for using product partitions is that they make numerical calculations much easier. But most importantly, we claim that $\lim_{\|\alpha_n\|\to 0} h^*_\mu(\mathbf{X}^{\alpha_n})$ does not depend on the particular increasing sequence $(\alpha_n)_{n\in\mathbb{N}_0}$ of successive refinements of a general finite partition α_0 of $(D, \mathcal{B} \cap D, \mu)$, as long as (i) they converge to the point partition of D, $\epsilon = \{\{x\}{:}x \in D\}$, and (ii) the numbering of the elements of $\alpha_1, \alpha_2, \ldots$ preserves the order relations through the process of refinement. Condition (i) requires that α_n consists of connected sets for all n and $\lim_{n\to\infty} \|A\| = 0$ for all $A \in \alpha_n$. Condition (ii) means that if $A_i, A_j \in \alpha_n$ and $i < j$, then $i' < j'$ whenever $A_i \supset A'_{i'} \in \alpha_{n+1}$ and $A_j \supset A'_{j'} \in \alpha_{n+1}$ (this is automatically satisfied by the lexicographically ordered, product refinements ι_n).

Lemma 9 *Let $(D, \mathcal{B}\cap D, \mu, f)$ be a measure-preserving dynamical system, α_0 a finite partition of $(D, \mathcal{B} \cap D, \mu)$, and $(\alpha_n)_{n\in\mathbb{N}}$ a sequence of successive refinements of α_0 preserving the order relations and converging to the point partition. Then*

$$h^*_\mu(f) = \lim_{n\to\infty} h^*_\mu(\mathbf{X}^{\alpha_n}),$$

where $\mathbf{X}^{\alpha_n}$ is the symbolic dynamics of f with respect to the partition α_n.

Proof Roughly speaking, the increasing sequences $\cdots \leq \kappa_n \leq \kappa_{n+1} \leq \cdots$ and $\cdots \leq \alpha_n \leq \alpha_{n+1} \leq \cdots$ are equivalent in the sense that, given κ_n there is a partition α_m with $\|\alpha_m\| \lesssim \|\kappa_n\|$ which can resolve the orbits of f with the same precision as κ_n does—and reciprocally. Of course, the ordinal patterns of length $L = 2, 3, \ldots$ of a given orbit will be, in general, different, depending on the partitions used. Nevertheless, there will be a one-to-one relation between the ordinal L-patterns realized by $\mathbf{X}^{\alpha_n}$ and $\mathbf{X}^{\kappa_n}$ in the limit $n \to \infty$, and the same holds for the corresponding probabilities. Therefore,

$$\lim_{n\to\infty} h_\mu^*(\mathbf{X}^{\alpha_n}) = \lim_{n\to\infty} h_\mu^*(\mathbf{X}^{\kappa_n}) = h_\mu^*(f).$$

□

The partitions $\mathcal{P}_L$, Eq. (3.5) build a sequence of successive refinements, but they do not preserve in general the order relations because their elements eventually decompose into different components. For the same reason, they cannot converge in general to the partition of D into separate points, ϵ, nor are their norms otherwise expected to vanish as $L \to \infty$.

Having shown that the metric permutation entropy does not depend on the partitions used in its calculation (with the provisos stated in Lemma 9), we turn to the main result of this chapter.

Theorem 10 *Let $f{:}D \to D$ be ergodic with respect to the measure μ, and suppose that $h_\mu^*(f)$ exists. Then $h_\mu^*(f) = h_\mu(f)$.*

Proof Let κ_0 be a quasi-box partition of $(D, \mathcal{B} \cap D, \mu)$ and $(\kappa_n)_{n\in\mathbb{N}}$ a sequence of successive product refinements of κ_0. Then,

$$h_\mu(f, \kappa_n) = h_\mu(\mathbf{X}^{\kappa_n})$$

by (6.15), where $\mathbf{X}^{\kappa_n} = \{X_k^{\kappa_n}\}_{k\in\mathbb{N}_0}$ is the symbolic dynamics of f with respect to the partition κ_n. Furthermore, $h_\mu(\mathbf{X}^{\kappa_n}) = h_\mu^*(\mathbf{X}^{\kappa_n})$ by Theorem 8, since $\mathbf{X}^\kappa$ is ergodic with respect to the measure μ if f is ergodic with respect to μ. Putting together, we have so far

$$h_\mu^*(f) = \lim_{n\to\infty} h_\mu^*(\mathbf{X}^{\kappa_n}) = \lim_{n\to\infty} h_\mu(\mathbf{X}^{\kappa_n}) = \lim_{n\to\infty} h_\mu(f, \kappa_n).$$

From Theorem 25 (Annex B) it follows then

$$\lim_{n\to\infty} h_\mu(f, \kappa_n) = h_\mu(f)$$

and we are done. □

If instead of Theorem 8, we use Corollary 5 in the previous proof for every process $\mathbf{X}^\kappa$, we conclude also $h_\mu^*(f) = h_\mu(f)$ for μ-preserving maps. This requires the technical assumption that $h_{m_s}(\mathbf{X}^\kappa)$ is m-integrable, where $\{m_s{:}s \in S^{\mathbb{N}_0}\}$, $S =$

$\{1, \ldots, |\kappa|\}$, is the ergodic decomposition of m, and m the shift-invariant measure of the sequence space model $(S^{\mathbb{N}_0}, \mathcal{B}_\Pi(S), m, \Sigma)$ of $\mathbf{X}^\kappa$—and this for every partition κ.

Theorem 11 *Let $f: D \to D$ be μ-preserving, and suppose that $h_\mu^*(f) = \lim_{n\to\infty} h_\mu^*(\mathbf{X}^{\kappa_n})$ exists. Under the assumptions of Lemma 8 for each $\mathbf{X}^{\kappa_n}$, the equality $h_\mu^*(f) = h_\mu(f)$ holds.*

6.3 On the Definition of Metric Permutation Entropy for Maps

The original definition of permutation entropy by Bandt, Keller, and Pompe [29] was presented in Sect. 1.2. Recall that it involves closed *one-dimensional* intervals I, maps $f{:}I \to I$, and sets of the form

$$P_\pi = \left\{x \in I{:}f^{\pi_0}(x) < f^{\pi_1}(x) < \cdots < f^{\pi_{L-1}}(x)\right\},$$

where $\pi = \langle \pi_0, \ldots, \pi_{L-1}\rangle \in \mathcal{S}_L$, $L \geq 2$. Recall once again that

$$\mathcal{P}_L = \{P_\pi \neq \emptyset{:}\pi \in \mathcal{S}_L\}.$$

In most situations of interest, $\mathcal{P}_L$ will be a partition of $(I, \mathcal{B} \cap I, \mu)$, where $\mathcal{B}$ is the Borel sigma-algebra of $\mathbb{R}$ and μ is an f-invariant measure. This is going to be our setting throughout this section.

Bandt, Keller, and Pompe define then the metric permutation entropy of order L as[2]

$$h_\mu^{*\mathrm{BKP}}(f, L) = -\frac{1}{L-1} \sum_{\pi\in\mathcal{S}_L} \mu(P_\pi) \log \mu(P_\pi) \tag{6.17}$$

and the permutation entropy of f to be

$$h_\mu^{*\mathrm{BKP}}(f) = \lim_{L\to\infty} h_\mu^{*\mathrm{BKP}}(f, L), \tag{6.18}$$

provided the limit exists.

As compared to conventional entropy, $h_\mu^{*\mathrm{BKP}}(f)$ has at least one remarkable feature: it involves only one infinite limit over the length of the word, while $h_\mu(f)$ involves additionally a second infinite process, namely, a supremum over partitions—unless a generating partition is known. This fact can be rephrased by saying that the sequence $\mathcal{P}_L$ builds a "generator" for $h_\mu^{*\mathrm{BKP}}$.

Let us highlight at this point the main result concerning $h_\mu^{*\mathrm{BKP}}(f)$:

Theorem 12 [29] *If $f{:}I \to I$ is piecewise monotone, then $h_\mu^{*\mathrm{BKP}}(f) = h_\mu(f)$.*

---

[2] Bandt, Keller, and Pompe chose the factor $1/(L-1)$ instead of $1/L$ (see (1.30)) because $\pi(x_0^0)$ contributes nothing to the entropy. Of course, either choice yields the same limit when $L \to \infty$.

Example 13 For the symmetric tent map (1.17), the elements of $\mathcal{P}_2$ are

$$P_{\langle 0,1\rangle} = (0, \tfrac{2}{3}), \qquad P_{\langle 1,0\rangle} = (\tfrac{2}{3}, 1)\,;$$

the elements of $\mathcal{P}_3$ are

$$P_{\langle 0,1,2\rangle} = (0, \tfrac{1}{3}), \qquad P_{\langle 0,2,1\rangle} = (\tfrac{1}{3}, \tfrac{2}{5}), \qquad P_{\langle 2,0,1\rangle} = (\tfrac{2}{5}, \tfrac{2}{3}),$$
$$P_{\langle 1,0,2\rangle} = (\tfrac{2}{3}, \tfrac{4}{5}), \qquad P_{\langle 1,2,0\rangle} = (\tfrac{4}{5}, 1);$$

and the elements of $\mathcal{P}_4$ are

$$P_{\langle 0,1,2,3\rangle} = (0, \tfrac{1}{6}), \quad P_{\langle 0,1,3,2\rangle} = (\tfrac{1}{6}, \tfrac{1}{5}), \quad P_{\langle 0,3,1,2\rangle} = (\tfrac{1}{5}, \tfrac{2}{9}) \cup (\tfrac{2}{7}, \tfrac{1}{3}),$$
$$P_{\langle 3,0,1,2\rangle} = (\tfrac{2}{9}, \tfrac{2}{7}), \quad P_{\langle 0,2,1,3\rangle} = (\tfrac{1}{3}, \tfrac{2}{5}), \quad P_{\langle 2,0,3,1\rangle} = (\tfrac{2}{5}, \tfrac{4}{9}) \cup (\tfrac{4}{7}, \tfrac{3}{5}),$$
$$P_{\langle 2,3,0,1\rangle} = (\tfrac{4}{9}, \tfrac{4}{7}), \quad P_{\langle 2,0,1,3\rangle} = (\tfrac{3}{5}, \tfrac{2}{3}), \quad P_{\langle 3,1,0,2\rangle} = (\tfrac{2}{3}, \tfrac{4}{5}),$$
$$P_{\langle 1,3,2,0\rangle} = (\tfrac{4}{5}, \tfrac{5}{6}), \quad P_{\langle 1,2,0,3\rangle} = (\tfrac{6}{7}, \tfrac{8}{9}), \quad P_{\langle 1,2,3,0\rangle} = (\tfrac{5}{6}, \tfrac{6}{7}) \cup (\tfrac{8}{9}, 1).$$

See Fig. 6.1 and compare with Fig. 1.7; owing to the order isomorphy of the symmetric tent map and the logistic map, there is a one-to-one relation between their admissible ordinal L-patterns. Computation of the metric permutation entropies of orders 2, 3, and 4 of the symmetric tent map Λ (the invariant measure μ is here the Lebesgue measure) yields the following results:

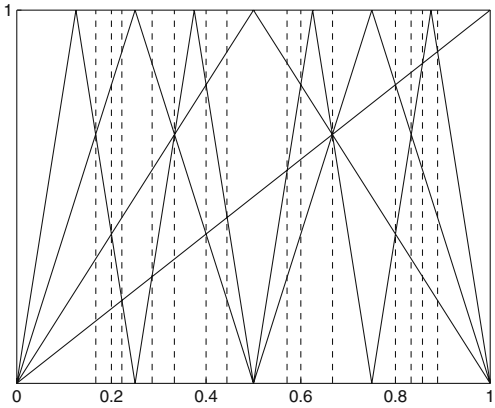

Fig. 6.1 Graphs of the identity, Λ, Λ^2, and Λ^3. The *vertical, dashed lines* separate different P_π, $\pi \in \mathcal{S}_4$

$$h_\mu^{*\mathrm{BKP}}(\Lambda,2) = \tfrac{2}{3}\log\tfrac{3}{2} + \tfrac{1}{3}\log 3 = 0.9183 \text{ bit/symbol},$$
$$h_\mu^{*\mathrm{BKP}}(\Lambda,3) = 1.0746 \text{ bit/symbol},$$
$$h_\mu^{*\mathrm{BKP}}(\Lambda,4) = 1.1807 \text{ bit/symbol}.$$

By Theorem 12,

$$h_\mu^{*\mathrm{BKP}}(\Lambda) = h_\mu(\Lambda) = \log 2 = 1 \text{ bit/symbol}.$$

But in the case of general maps, it seems that only inequality (6.19) below (formally similar to (6.9)) can be proved. Comparing such one-dimensional results with the dimensional generality of Theorem 10, we may conclude that the definition (6.16) of permutation entropy offers some advantages.

Note that the central distinction, which makes formulation (6.16) easier and more natural, is that (6.16) takes the limit of infinite long conditioning ($L \to \infty$) first and the discretization limit ($\|\kappa_n\| \to 0$) last, similar to Kolmogorov–Sinai entropy, and as opposed to (6.18), where an explicit discretization is not taken. Thus we have two limits to take (while $h_\mu^{*\mathrm{BKP}}(f)$ involves only one limit), but the second, $\|\kappa_n\| \to 0$, is harmless and, in principle, can be numerically approximated. We conjecture that for "non-pathological" dynamical systems of the sort one might observe in nature, the two formulations are equivalent, but there are likely to be some non-trivial technicalities involved in a rigorous analysis. More on this, in the next chapter.

Transformations with an infinite number of monotonicity segments are not unusual in ergodic theory.

Example 14 The *Gauss transformation*, $f{:}[0,1) \to [0,1)$ with

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \\ \frac{1}{x} \pmod 1 & \text{if } x \neq 0 \end{cases},$$

is an ergodic map [52, Chap. 5] with infinitely many monotonicity segments, see Fig. 6.2.

The next theorem shows that, in general, $h_\mu^{*\mathrm{BKP}}(f)$ can only be expected to be an upper bound of $h_\mu(f)$.

Theorem 13 [29] *If $f{:}I \to I$ is a μ-preserving map with $h_\mu(f) < \infty$, then*

$$\liminf_{L\to\infty} h_\mu^{*BKP}(f,L) \geq h_\mu(f). \tag{6.19}$$

It follows $h_\mu^{*BKP}(f) \geq h_\mu(f)$, provided $h_\mu^{*BKP}(f)$ exists.

Proof Let $\iota = \{I_j, 1 \leq j \leq |\iota|\}$ be a partition of $(I, \mathcal{B} \cap I, \mu)$, where $I_j \subset I$ are intervals. This being the case, let $c_1 < c_2 < \cdots < c_{|\iota|-1}$ be the points that subdivide the interval $I = [a,b]$ into the $|\iota|$ intervals I_j of the partition ι. We consider a fixed

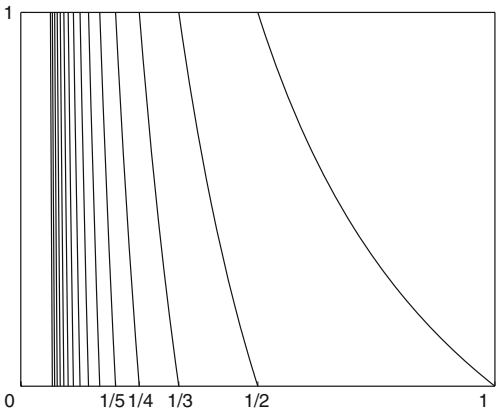

Fig. 6.2 Some monotony intervals of the Gauss transformation

$P_\pi \in \mathcal{P}_L$ and show that it can intersect at most $(L+1)^{|\iota|-1}$ sets of the partition $\iota_0^{L-1} := \vee_{i=0}^{L-1} f^{-i}(I_{j_i})$ with $I_{j_0}, \ldots, I_{j_{L-1}} \in \iota$. For $x \in P_\pi$, let $\Delta_L[x]$ denote the set in ι_0^{L-1} that contains x. Thus, $\Delta_L[x]$ can be written as $I_{j_0} \cap f^{-1}(I_{j_1}) \cap \cdots \cap f^{-(L-1)}(I_{j_{L-1}})$ with $I_{j_0}, \ldots, I_{j_{L-1}} \in \iota$, so that it can be specified by the n-tupel $j[x] = (j_0, \ldots, j_{L-1}) \in \{1, \ldots, |\iota|\}^L$.

Now, π is given by inequalities $x_{k_1} < \cdots < x_{k_L}$ with $\{k_1, \ldots, k_L\} = \{0, \ldots, L-1\}$ and $x_k = f^k(x)$. For each $x \in P_\pi$ we can extend these inequalities so that they give the common order of the c_r and the x_{k_l}, where $1 \le r \le |\iota| - 1$ and $1 \le l \le L$. It follows that there are at most $(L+1)^{|\iota|-1}$ possible extended orders since each c_r has $L+1$ possible bins to go among the x_{k_l}. Moreover, when we know the common order of the c_r and x_{k_l}, then $j[x]$ is uniquely determined (since $c_{j-1} < x_k < c_j$ implies $x_k \in I_j$ and thus $x \in f^{-k}(I_j)$, with $1 \le j \le |\iota|$, $c_0 = a$, and $c_{|\iota|} = b$).

Each $P_\pi \in \mathcal{P}_L$ is then the union of at most $(L+1)^{|\iota|-1}$ sets $V_k \in \iota_0^{L-1} \vee \mathcal{P}_L$ with total measure $\mu(P_\pi)$. Hence,

$$
\begin{aligned}
&- \sum_{k=1}^{(L+1)^{|\iota|-1}} \mu(V_k) \log \mu(V_k) \\
&\le - \sum_{k=1}^{(L+1)^{|\iota|-1}} \frac{\mu(P_\pi)}{(L+1)^{|\iota|-1}} \log \frac{\mu(P_\pi)}{(L+1)^{|\iota|-1}} \\
&= -\mu(P_\pi) \log \mu(P_\pi) + (|\iota| - 1)\mu(P_\pi) \log (L+1)
\end{aligned}
$$

and summing over all $\pi \in \mathcal{S}_L$,

$$
H_\mu(\iota_0^{L-1}) \le H_\mu(\iota_0^{L-1} \vee \mathcal{P}_L) \le H_\mu(\mathcal{P}_L) + (|\iota| - 1) \log (L+1). \tag{6.20}
$$

It follows that

$$\frac{1}{L-1}H_\mu(\mathcal{P}_L) \geq \frac{1}{L-1}\left[H_\mu(\iota_0^{L-1}) - (\,|\iota| - 1)\log(L+1)\right]$$

and

$$\liminf_{L\to\infty}\frac{1}{L-1}H_\mu(\mathcal{P}_L) \geq \liminf_{L\to\infty}\frac{1}{L-1}H_\mu(\iota_0^{L-1}), \tag{6.21}$$

since $\frac{1}{L-1}\log(L+1) \to 0$ as $L \to \infty$.

On the other hand, the sequence $\frac{1}{L-1}H_\mu(\iota_0^{L-1})$ converges to $h_\mu(f,\iota)$ when $L \to \infty$, hence

$$\liminf_{L\to\infty} h_\mu^{*\mathrm{BKP}}(f,L) = \lim_{L\to\infty}\inf\frac{1}{L-1}H_\mu(\mathcal{P}_L) \geq h_\mu(f,\iota),$$

for any partition ι. Finally,

$$\liminf_{L\to\infty} h_\mu^{*\mathrm{BKP}}(f,L) \geq \sup_\iota h_\mu(f,\iota) = h_\mu(f).$$

□

6.4 Numerical Issues

Our way to the metric permutation entropy of maps was paved by partitions of the state space and the corresponding symbolic dynamics, very much the same way as it happens with the Kolmogorov–Sinai entropy. Therefore, calculating the metric permutation entropy of maps and information sources turns out to be essentially the same task, except for the fact that in the first case this calculation has, in principle, to be repeated with ever finer partitions. In practice, one estimates the true value of the permutation entropy by taking a "sufficiently" fine partition once and for all. This corresponds, by the way, to the numerical practice, as we shall presently explain. If, furthermore, the map (and hence the ensuing source) is ergodic, then it suffices to consider one or a small sample of coarse-grained orbits.

As a by-product of the previous results on metric permutation entropy, the practitioner of time-series analysis will find an alternative way to envision or, eventually, numerically estimate the Kolmogorov–Sinai entropy of real sources. It is worth reminding (see Chap. 1) that the entropy of information sources can be measured by a variety of techniques that go beyond counting word statistics and comprise different definitions of "complexities" such as, for example, counting the patterns along a digital (or digitalized) data sequence [137, 211, 6]. Bandt and Pompe refer in [28] to the permutation entropy of time series as complexity. That the entropy can also be computed by counting ordinal patterns shows once again that it is a so general concept that can be captured with different and seemingly blunt approaches.

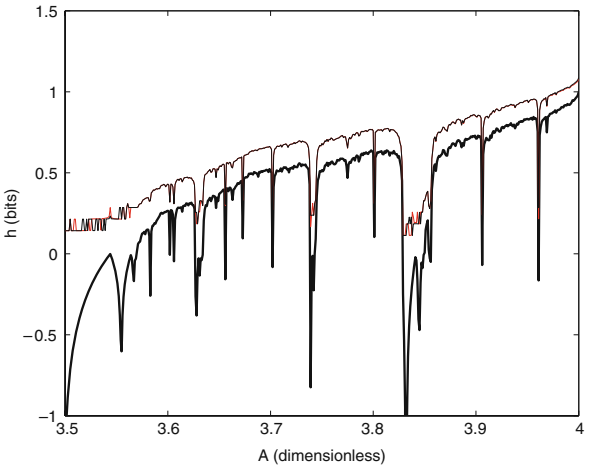

Fig. 6.3 Lyapunov exponent (*black thick line*) of the logistic map g_A, $3.5 \leq A \leq 4$, and metric permutation entropy (rate) estimates $\hat{h} = h^*(X_0^{13})$ in bits/symbol for $N = 10^6$ length time series from the map (*black thin lines*). The metric permutation entropy estimate tracks changes in the Lyapunov exponent well, with a nearly constant bias. Periodic orbits give a finite permutation entropy, but the rate estimate would tend to zero given a sufficiently long word

We demonstrate numerical results on time series $x_{n+1} = g_A(x_n)$ from the logistic map $g_A(x) = Ax(1-x)$, where $0 \leq A \leq 4$ and $0 \leq x \leq 1$. Figure 6.3 shows an estimate of the metric permutation entropy on noise-free data as a function of A, comparing the Lyapunov exponent $L_\mu(g_A)$ (computed from the orbit knowing the equation of motion) to the metric permutation entropy of g_A for $3.5 \leq A \leq 4$. To be precise, we are estimating $h^*_\mu(\mathbf{X})$ with $\mathbf{X}$ discretized from the logistic map iterated at the discretization of double-precision numerical representation, i.e., $\mathbf{X}$ is the output of a standard numerical iteration and μ is the natural invariant measure with density $d\mu/dx = \frac{1}{\pi\sqrt{x(1-x)}}$. The entropy estimator of the block ranks was the plug-in estimator (substituting observed frequencies for probabilities) plus the classical bias correction, first order in $1/N$, N being here the number of samples (which can be taken, for instance, from sliding windows of fixed length L along the orbit/orbits considered) [167]. Let us remind that

$$h_\mu(g_4) = L_\mu(g_4) = \int_0^1 \log \left|g_4'(x)\right| d\mu(x) = \log 2.$$

Thus, in practice the BKP approach (Sect. 6.3) and our approach (Sect. 6.2) boil down to the same recipe: generate orbits and count ordinal patterns in sliding windows of increasing sizes; for more details, see Chap. 9. The most intriguing characteristic of order relations is that they define, on their own, partitions $\mathcal{P}_L$ for the mapping from continuous values (as the discretization level $\|\kappa_n\|$ goes to zero) to a lower precision symbolic representation which has the natural structure for entropy. When estimating entropy from the discrete information source induced

from a *fixed* discretization, the entropy of the symbol stream will not generally equal the Kolmogorov–Sinai entropy unless a generating partition is used, and that can be difficult to find, especially for observed data alone, although some recent works show progress in this direction (e.g., [40] and references therein). The "magic" in using ordinal patterns is that the self-defined partitions $\mathcal{P}_L$ give the Kolmogorov–Sinai entropy, at least asymptotically. Permutation entropy may offer a significant opportunity to advance analytical computations of entropies for various dynamical systems, where generating partitions might be too difficult to find rigorously.

It turns out that using metric permutation entropy to accurately estimate the Kolmogorov–Sinai entropy is more difficult than using it as a very rapid and easy-to-compute *relative* quantification of entropy or complexity which can be computed without requiring a fixed partition (see, e.g., [45]). The key issue in using permutation entropy for empirical data analysis as an entropy estimator is the same as with standard Shannon entropy estimation: balancing the tension between larger word lengths L, to capture more dependencies, and the loss of sufficient sampling for good statistics in the ever larger discrete space. Extracting permutation entropies is rapid and easy—but taking the limits is not at all simple numerically. The finite L performance and convergence rate and bias of any specific computational method are major issues when it comes to accurately estimating the entropy of a source from observed data. It is now appreciated that numerically estimating the Shannon block entropy from finite data and, especially, the asymptotic entropy can be surprisingly tricky [195, 127, 6, 121, 122]. The theoretical definitions of entropy do not necessarily lead to good statistical methods, and superior alternatives have been developed over the many years since Shannon. We believe that some of these ideas may similarly be applicable to the permutation entropy situation, either in terms of using some of the superior entropy estimation methods for block entropies or developing algorithms based on more sophisticated data compression principles to extract the entropy itself.

Also important for practical time-series analysis is the usual situation where observations of a predominantly deterministic source is contaminated with a small level of observational noise. Here, we recommend that the user *fix* some discretization level $\|\kappa_n\|$ characteristic of the noise and evaluate the permutation entropies via entropies of rank words evaluated from the discretized observables.

In regard to vector-valued sources, we used (without restriction) lexicographic ordering in the theoretical part because of definiteness and simplicity. For analyzing chaotic observed data, however, it may be acceptable to still use but one scalar projection subject to the traditional caveats of time-delay embedology. We would expect that for appropriately mixing sources and generic observation functions, the Kolmogorov–Sinai entropy estimated through that scalar still equals the true value, and likewise so might permutation entropy. We have found that numerically this appears to work in practice. Moreover, the lexicographic ordering will effectively reduce to this case anyway except for the few cases where the symbols on the dominant coordinate match, which will be less frequent as L increases. More on this in Chaps. 7 and 9.