On the Rate of Information Loss in Memoryless Systems

Bernhard C. Geiger*, Gernot Kubin*
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{geiger.g.kubin}@ieee.org

Abstract—In this work we present results about the rate of (relative) information loss induced by passing a real-valued, stationary stochastic process through a memoryless system. We show that for a special class of systems the information loss rate is closely related to the difference of differential entropy rates of the input and output processes. It is further shown that the rate of (relative) information loss is bounded from above by the (relative) information loss the system induces on a random variable distributed according to the process’s marginal distribution.

As a side result, in this work we present conditions such that for a Markovian input process also the output process possesses the Markov property.

Index Terms—data processing inequality, information loss, entropy dimension, system theory, lumpability

I. INTRODUCTION

Signal processing, as defined by many textbooks, is related to the “representation, transformation, and manipulation of signals and the information the signals contain” [1] emphasis added. Yet, most of these textbooks leave the notion of information completely aside and focus, instead, on purely energetic aspects or second-order statistics: transfer functions for linear filters, their effect on the auto-correlation function of its output signal, and similar results for nonlinear, memoryless systems (e.g., [2]) are popular characterizations. However, except for the purely Gaussian case, energy (or second-order statistics) and information show an inherently different behavior. It is therefore desirable to extend current system theory by information-theoretic aspects.

While the data processing inequality (e.g., [3] p. 35]) captures the fact that deterministic functions of random variables (RVs) destroy information, relatively little has been done to quantify this information loss. Pinsker showed that the entropy rate of a function of a stationary stochastic process on a finite alphabet is bounded from above by the entropy rate of the original process [4] Ch. 6.3]. Similarly, Watanabe and Abrahám analyzed the rate of information loss for functions of stationary stochastic processes, introducing also a relative version of information loss in [5]. Results on the information loss rate in dynamical systems, together with an upper bound, were presented in [6].

While these works focus on finite or countable alphabets, [7] analyzes the absolute and, in case the latter is infinite, relative information loss induced by passing a real-valued RV through a memoryless system. In this work we extend [7] to real-valued, stationary stochastic processes. In particular we show that the information loss for RVs distributed according to the marginal distribution of the process is an upper bound on the information loss rate (Section III). A similar result is shown also for the relative information loss rate, although there the bound is tight in many more cases (Section V).

While redundancy helps to reduce the rate of information loss, it often fails to reduce the rate of relative information loss. The connection between the rate of information loss and the differential entropy rates of the input and output processes shown in Section III is remarkably similar to the corresponding result for information loss presented in [7].

In search for processes which are simple to analyze, we found a set of sufficient conditions such that for a Markovian input process also the output process has the Markov property (Section IV). This extends the notion of lumpability (cf. [8]) from discrete-time and continuous-time, homogeneous Markov chains to discrete-time, homogeneous, real-valued Markov processes. These conditions, together with our other theoretical findings, are illustrated with the help of examples in Section VI.

II. PRELIMINARIES & NOTATION

Throughout this work we consider discrete-time, stationary stochastic processes \( X \) with alphabet \( \mathcal{X} \subseteq \mathbb{R} \). Let \( X_n \) be the \( n \)-th sample of the process, and let \( X^j = \{ X_i, X_{i+1}, \ldots, X_j \} \). By stationarity, the distribution of \( X_n \), \( P_{X_n} \), equals the marginal distribution \( P_X \). We assume that \( P_X \) is absolutely continuous w.r.t. the Lebesgue measure, and that it thus possesses a probability density function (PDF) \( f_X \). Similarly, we assume that for all \( n \), the joint PDF \( f_{X^n} \) and the conditional PDF \( f_{X^n|X^n_{1:n-1}} \) exist.

Let \( H(\cdot), h(\cdot), \overline{H}(\cdot), \), and \( \overline{h}(\cdot) \) denote the entropy, the differential entropy, the entropy rate, and the differential entropy rate of the RVs and stochastic processes in the argument (see [3] or [9] for definitions). We assume that the joint differential entropy of an arbitrary collection of RVs exists and is finite, and that also the entropy rate [9] Thm. 14.7]

\[
\overline{h}(X) := \lim_{n \to \infty} h(X_n|X^n_{1:n-1}) = \lim_{n \to \infty} \frac{1}{n} h(X^n_n) \tag{1}
\]

exists and is finite. The logarithm for the entropies is taken to the base 2.
III. INFORMATION LOSS RATE PIECEWISE BIJECTIVE FUNCTIONS

In this section we devote our attention to a specific class of functions for which the preimage of every point of its range is an at most countable set:

Definition 1 (Piecewise Bijective Function). A piecewise bijective function \( g: \mathcal{X} \to \mathcal{Y} \), \( \mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N \), is a surjective, measurable function defined piecewise on an at most countable partition \( \{ \mathcal{X}_i \} \) of its domain:

\[
g(x) = \begin{cases} 
g_1(x), & \text{if } x \in \mathcal{X}_1 \\
g_2(x), & \text{if } x \in \mathcal{X}_2 \\
\vdots
\end{cases}
\]

where each \( g_i: \mathcal{X}_i \to \mathcal{Y}_i \) is bijective. Furthermore, the derivative \( g' \) exists on the closures of \( \mathcal{X}_i \), and its magnitude is non-zero \( P_X \)-a.s.

Feeding the stationary stochastic process \( X \) through a memoryless system described by such a function \( g \) gives rise to another stationary stochastic process \( Y \) defined by \( Y_n := g(X_n) \), which, intuitively, conveys less information. In order to analyze the amount of information lost per sample we introduce

Definition 2 (Information Loss Rate). The information loss rate is

\[
\overline{L}(X \to Y) := \lim_{n \to \infty} \frac{1}{n} L(X^n \to Y^n) = \lim_{n \to \infty} \frac{1}{n} H(X^n|Y^n)
\]

i.e., the average of the block information loss.

We showed in [7] that the information loss in systems described by functions satisfying Definition 1 can be computed as

\[
L(X \to Y) = H(X|Y) = h(X) - h(Y) + E \{ \log |g'(X)| \}
\]

where \( Y = g(X) \) and where the expectation is taken w.r.t. \( X \). We now present a corresponding result for stationary stochastic processes:

Proposition 1 (Information Loss Rate for PBFs). The information loss rate induced by feeding a stationary stochastic process \( X \) through a PBF \( g \) is

\[
\overline{L}(X \to Y) = \overline{h}(X) - \overline{h}(Y) + E \{ \log |g'(X)| \}.
\]

Proof: For the proof we note that the \( n \) RVs \( X^n \) can be interpreted as a single, \( n \)-dimensional RV; similarly, we can define an extended function \( g^n: \mathcal{X}^n \to \mathcal{Y}^n \), applying \( g \) coordinate-wise. The Jacobian matrix of \( g^n \) is a diagonal matrix constituted of the elements \( g'(x_i) \). With the extension of (5) to multivariate functions we thus obtain

\[
L(X^n \to Y^n) = h(X^n) - h(Y^n) + E \left\{ \log \prod_{i=1}^{n} |g'(X_i)| \right\}
\]

where the first line is because the determinant of a diagonal matrix is the product of its diagonal elements, and where we employed stationarity of \( X \) to obtain the second line. Dividing by \( n \) and taking the limit completes the proof.

In [7] we showed that the information loss of a cascade of systems equals the sum of the information losses induced in the systems constituting the cascade. Indeed, this result can be carried over to the information loss rate as well:

Proposition 2 (Cascade of Systems). Let \( X \) be fed through a PBF \( g \) to obtain \( Y \), and let \( Y \) be fed through a PBF \( h \) to obtain \( Z \). The information loss rate of the cascade is given as the sum of the individual information loss rates:

\[
\overline{L}(X \to Z) = \overline{L}(X \to Y) + \overline{L}(Y \to Z).
\]

Proof: The proof follows from the fact that the cascade is described by the function \( h \circ g \), and that

\[
E \{ \log |(h \circ g)'(X)| \} = E \{ \log |g'(X)| h'(g(X)) \}
\]

\[
= E \{ \log |g'(X)| \} + E \{ \log |h'(Y)| \}.
\]

It is often not possible to obtain closed-form expressions for the information loss rate induced by a system. Moreover, estimating the information loss rate by simulations soon suffers the curse of dimensionality, as, in principle, infinitely long random sequences have to be drawn and averaged. Much simpler is an estimation of the information loss, since a single realized, sufficiently long sequence allows for an estimation of the latter. As the next proposition shows, this relatively simple estimation delivers an upper bound on the information loss rate:

Proposition 3 (Loss > Loss Rate). Let \( X \) be a stationary stochastic process and \( X \) an RV distributed according to the process’s marginal distribution. The information loss induced by feeding \( X \) through a PBF \( g \) is an upper bound on the information loss rate induced by passing \( X \) through \( g \), i.e.,

\[
\overline{L}(X \to Y) \leq L(X \to Y).
\]

Proof: The inequality holds trivially if \( L(X \to Y) = \infty \).

The rest of the proof follows from the chain rule and the fact that conditioning reduces entropy:

\[
\overline{L}(X \to Y) = \lim_{n \to \infty} \frac{1}{n} H(X^n|Y^n)
\]

\[
= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_i^{i-1}, Y^n)
\]

\[
\leq \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|Y_i)
\]

\[
= L(X \to Y).
\]

Clearly, this bound is tight whenever the input process \( X \) is an iid process. Moreover, it is trivially tight whenever the function is bijective, i.e., when \( L(X \to Y) = 0 \). In
Section 4 we present an example which renders this bound tight in the general case.

Intuitively, this bound suggests that redundancy of a process, i.e., the statistical dependence of its samples, reduces the amount of information lost per sample when fed through a deterministic system. The same connection between information loss and information loss rate has already been observed in [8] for stationary stochastic processes with finite alphabets.

The next bound again extends a result from [7], bounding the information loss rate by the entropy rate of a stationary stochastic process on an at most countable alphabet. As such, it presents a different way to estimate the information loss rate efficiently using numerical simulations.

**Proposition 4** (Upper Bound). Let \( W \) be a stationary stochastic process defined by \( W_n := i \) if \( X_n \in X_i \). Then,
\[
\overline{L}(X \to Y) = \overline{H}(W|Y) \leq \overline{H}(W). \tag{14}
\]

**Proof:** We again treat \( X^n \) as an \( n \)-dimensional RV; \( g^n \) induces a partition of its domain \( X^n \), which is equivalent to the \( n \)-fold product of the partition \( \{X_i\} \). Letting \( \tilde{W} \) be the RV obtained by quantizing \( X^n \) according to this partition, it is easy to see that \( W^n \) is equivalent to \( \tilde{W} \). Thus, with [7],
\[
H(X^n|Y^n) = H(\tilde{W}|Y^n) = H(W^n|Y^n) \tag{15}
\]
for all \( n \). This, together with the fact that conditioning reduces entropy, completes the proof. ■

For the case that the input process is a Markov process, i.e., if \( f_{X_n|X_{n-1}} = f_{X_n|X_{n-1}} \) for all \( n \), an additional, sharper, upper bound can be presented:

**Proposition 5** (Upper Bound for Markovian \( X \)). Let \( X \) be a Markov process, and let \( W \) be as in Proposition 2. Then, for finite \( L(X \to Y) \),
\[
\overline{L}(X \to Y) \leq H(W_2|X_1). \tag{16}
\]

**Proof:** We again apply the chain rule, Markovity of \( X \), and the fact that conditioning reduces entropy to arrive at
\[
H(X^n|Y^n) \leq \sum_{i=1}^{n} H(X_i|X_{i-1}, Y_i). \tag{17}
\]
By stationarity we obtain
\[
\overline{L}(X \to Y) \leq H(X_2|X_1, Y_2) \tag{18}
\]
\[
\overset{(a)}{=} H(W_2|X_1, Y_2) \tag{19}
\]
\[
\leq H(W_2|X_1) \tag{20}
\]
where \((a)\) holds since, for all \( x \in X \), \( H(X_2|Y_2, X_1 = x) = H(W_2|Y_2, X_1 = x) \) [7]. The last inequality is due conditioning [3] Thm. 2.6.5] and completes the proof. ■

That the bound is sharper than the one of Proposition 4 follows from observing that
\[
H(W_n|X_{n-1}) = \lim_{n \to \infty} H(W_n|X_{n-1}) \leq \lim_{n \to \infty} H(W_n|W_1^{n-1}) = \overline{H}(W). \tag{21}
\]

The interpretation of this result is that a function destroys little information if the process is such that, given the current sample \( X_{n-1} \), the next sample \( X_n \) falls within some element of the partition with a high probability. The question whether, and under which conditions, this bound is tight is related to the phenomenon of lumpability and will be answered in the following section.

**IV. LUMPABILITY FOR CONTINUOUS-VALUED MARKOV PROCESSES**

It is well-known that the function of a Markov process need not possess the Markov property itself. However, as it is known for Markov chains, there exist conditions on the function and/or the chain such that the output is Markov. In [8] this has been termed lumpability and subsequently investigated by numerous researchers. While most results are given for finite Markov chains (e.g., [10], [11]) relatively little is known in the general case of an uncountable alphabet (see [12] for an exception). Our small contribution to this field of research lies in presenting sufficient conditions for lumpability of continuous-valued Markov processes.

Let \( f_{X_n|X_{n-1}} \) for all \( n \), i.e., let \( X \) be a Markov process. We maintain

**Proposition 6. If**
\[
\forall y_1^2 \in \mathcal{Y}^2 : \forall x \in g^{-1}[y_1] : \quad f_{Y_2|X_1}(y_2|x) > 0 \Rightarrow f_{Y_2|X_1}(y_2|x) = f_{Y_2|Y_1}(y_2|y_1) \tag{22}
\]
then \( X \) is lumpable w.r.t. \( g, i.e., Y \) is Markov.

**Proof:** See Appendix. ■

As a corollary, we next make the conditions on the function \( g \), the marginal distribution \( f_X \), and the conditional distribution \( f_{X_n|X_1} \) explicit. By adding a further condition, we gain tightness of Proposition 5 in addition to Markovity:

**Corollary 1. If for all \( y_1^2 \in \mathcal{Y}^2 \) and all \( x, x' \in g^{-1}[y_1] \) such that \( f_X(x) > 0 \) and \( f_X(x') > 0 \) the following holds**
\[
\sum_{x \in g^{-1}[y_1]} \frac{f_{X|X_1}(x|x)}{|g'(x)|} = \sum_{x \in g^{-1}[y_1]} \frac{f_{X|X_1}(x|x')}{|g'(x')|} \tag{23}
\]
then the condition of Proposition 4 is fulfilled and \( Y \) is Markov.

**If, additionally, for all \( y \in \mathcal{Y} \), all \( x \) within the support of \( f_X \), and all \( w, w' \) such that \( \Pr(W_2 = w|X_1 = x) > 0 \) and \( \Pr(W_2 = w'|X_1 = x) > 0 \)**
\[
\frac{f_{X|X_1}(w|y_2|x)}{|g'(g^{-1}(y_2))|} = \frac{f_{X|X_1}(w'|y_2|x)}{|g'(g^{-1}(y_2))|} \tag{24a}
\]
and
\[
\Pr(W_2 = w'|X_1 = x) = \Pr(W_2 = w|X_1 = x) \tag{24b}
\]
then the bound of Proposition 5 holds with equality.

**Proof:** See Appendix. ■
V. RELATIVE INFORMATION LOSS RATE FOR FUNCTIONS WHICH REDUCE DIMENSIONALITY

Not all systems can be described by functions satisfying Definition 1. In particular, a simple quantizer already violates this definition and suffers from infinite information loss. To analyze the information processing characteristics of a broader class of systems, in [7] the notion of relative information loss was introduced, capturing the percentage of information available at the input lost in the system. To extend this notion to stochastic processes, we introduce

Definition 3 (Relative Information Loss Rate). The relative information loss rate is

\[ l(X \rightarrow Y) := \lim_{n \to \infty} \frac{l(X^n \rightarrow Y^n)}{n} \]

whenever the limit exists.

The limit \( \hat{X} \rightarrow X \) is equivalent to \( \lim_{k \to \infty} [2^k X]/2^k \), where flooring and scalar multiplication are applied element-wise (cf. [7]).

Based on

\[ l(X^n_i \rightarrow Y^n_i) \leq \frac{1}{n} \sum_{i=1}^{n} l(X_i \rightarrow Y_i) = l(X \rightarrow Y) \]

from [7] and from stationarity of \( X \), which yields (a), one can show that \( l(X \rightarrow Y) \leq l(X \rightarrow Y) \), complementing Proposition 3. However, in many cases this inequality is an equality, as we show in

Proposition 7 (Redundancy won’t help). Let \( X \) be a stationary stochastic process and \( X \) an RV distributed according to the process’s marginal distribution. Let further \( g \) be defined on a finite partition \( \{X_i\} \) of \( X \) into non-empty sets as in [4], where \( g_i \in C^\infty \) is either injective or constant (i.e., \( g_i(x) = c_i \) for all \( x \in X_i \)). Then,

\[ l(X \rightarrow Y) = l(X \rightarrow Y) = P_X(X_i) \]

where \( X_i \) is the union of all elements \( X_i \) of the partition on which \( g \) is constant.

Proof: See Appendix.

Indeed, we conjecture that equality is indeed the “usual” case, prevailing in most practical scenarios. Thus, while redundancy can help reduce information loss, it may be useless when it comes to relative information loss. Applications of this result may be the scalar quantization of a stochastic process (leading to a relative information loss rate of 1, i.e., 100% of the information is lost [7]) and system blocks for multirate signal processing (see the example in Section VI-D).

VI. EXAMPLES

A. AR-Process and Magnitude Function

In this example we assume that a first-order, zero-mean, Gaussian auto-regressive process \( X \) is fed through a magnitude function (see Fig. 1). Let the AR process be generated by the following difference equation:

\[ X_n = aX_{n-1} + Z_n \]

where \( a \in (0,1) \) and where \( Z_n \) are samples drawn independently from a Gaussian distribution with zero mean and variance \( \sigma^2 \). It follows immediately that the process \( X \) is also zero mean and has variance \( \sigma_X^2 = \frac{1}{a^2} \) [7 Ex. 6.11]. Let \( Y \) be defined by \( Y_n = |X_n| \).

For the sake of brevity we define \( \phi(\mu, \sigma^2; x) \) as the PDF of a Gaussian RV with mean \( \mu \) and variance \( \sigma^2 \), evaluated at \( x \). Thus, we get

\[ f_X(x) = \phi(0, \sigma_X^2; x) \]

and

\[ f_{X_2|X_1}(x_2|x_1) = \phi(ax_1, \sigma_2^2; x_2). \]

It follows that (23) is satisfied with \(|g'(x)| = 1\) and since \( \phi(ax_1, \sigma_2^2; x_2) = \phi(-ax_1, \sigma_2^2; -x_2) \),

\[ \sum_{x_2 \in g^{-1}[y_2]} \frac{f_{X_2|X_1}(x_2|y_1)}{|g'(x_2)|} = \phi(ax_1, \sigma_2^2; y_2) + \phi(ax_1, \sigma_2^2; -y_2) = \phi(-ax_1, \sigma_2^2; -y_2) + \phi(-ax_1, \sigma_2^2; y_2) \]

As a consequence, the output process \( Y \) is Markov.

We performed a series of simulations, as the information loss rate for this example cannot be expressed in closed form. Rewriting, e.g., the lower bound on the information loss rate as

\[ \tilde{L}(X \rightarrow Y) \]

\[ \geq h(X_2|X_1) - h(Y_2|X_1) + E \{ \log |g'(X)| \} \]

\[ = h(X) - I(X_1; X_2) - h(Y) + I(X_1; Y_2) \]

\[ + E \{ \log |g'(X)| \} \]

\[ = L(X \rightarrow Y) - I(X_1; X_2) + I(X_1; Y_2) \]

1Note that also Watanabe and Abraham defined the fractional information loss for stochastic processes on finite alphabets [5], for these types of processes, however, the relative information loss can be smaller or larger than the information loss.
allowed us to employ the histogram-based mutual information estimation from [13] together with \( L(X \rightarrow Y) \) = 1, as shown in [7]. The upper bound \( H(W_2|X_1) \) from Proposition 5 was computed using numerical integration. In Fig. 2 one can see that the first-order upper and lower bounds on the information loss rate from Lemma 1 in the proof of Proposition 5 are indistinguishable, which suggests that the output process is indeed Markov. Moreover, it can be seen that a higher value for the magnitude of the pole \( a \) leads to a smaller information loss rate. This can be explained by the fact that the redundancy of the process \( X \) increases with increasing \( a \), which helps preventing information loss.

Generally, while redundancy reduces the information loss rate compared to an iid process (cf. Proposition 3), it is not necessarily true that more redundancy leads to a smaller information loss rate than less redundancy. Indeed, one can create examples where a process with a higher redundancy suffers from a higher information loss rate than a process with less redundancy. This suggests that the redundancy of a process has to be matched to the function \( g \) in order to efficiently prevent information from being lost; in that sense, this parallels the field of channel coding, where the code needs to be matched to the characteristics of the channel (noise, fading, burst errors) in order to successfully reduce the bit error rate.

B. Cyclic Random Walk

We next consider a scenario where our process is a cyclic random walk on a subset \([-M, M]\) of the real line. Assume that for a given state \( X_1 \) the following state is uniformly distributed on a cyclically shifted subset of \([-M, M]\) of length \( 2a \leq 2M \), i.e.,

\[
f_M(x_2|x_1) := f_{X_2|X_1}(x_2|x_1) = \begin{cases} \frac{1}{2a}, & \text{if } d(x_2, x_1) \leq a \\ 0, & \text{else} \end{cases}
\]

where \( d(x, y) = \min_k |x - y - 2kM| \). Intuitively, \( X_n \) is the sum of \( n \) independent RVs uniformly distributed on \([-a, a]\), where sums outside of \([-M, M]\) are mapped back into this interval via the modulo operation. It is easy to verify that the marginal distribution of \( X \) is the uniform distribution, i.e., \( f_X(x) = \frac{1}{M} \) for all \( x \in [-M, M] \) and zero otherwise. The function we feed the process through shall again be the magnitude function, i.e., \( Y_n = |X_n| \).

Since \( d(x, y) = d(-x, -y) \) and since \( |g'(x)| \equiv 1 \) for all \( x \), it follows that \( (23) \) is fulfilled, and that thus \( Y \) is Markov. Moreover, we have \( \bar{h}(Y) = \bar{h}(Y_2|X_1) \), and obtain for the information loss rate with Proposition 1

\[
\bar{L}(X \rightarrow Y) = \bar{h}(X) - \bar{h}(Y) + E \{ \log |g'(X)| \}
\]

\[
= h(X_2|X_1) - h(Y_2|X_1)
\]

\[
= \int_{-M}^{M} \int_{-M}^{M} f_X(x_1)f_M(x_2|x_1) \log \frac{f_{Y_2|X_1}(|x_2| |x_1|)}{f_M(x_2|x_1)} \, dx_2 \, dx_1
\]

\[
= \int_{-M}^{M} \int_{-M}^{M} f_M(x_2|x_1) \frac{1}{2M} \log \left( 1 + \frac{f_M(-x_2|x_1)}{f_M(x_2|x_1)} \right) \, dx_2 \, dx_1.
\]

The logarithm evaluates to zero if \( f_M(-x_2|x_1) = 0 \) and to one otherwise (the logarithm is taken to base 2). Therefore, we can write

\[
\bar{L}(X \rightarrow Y) = \frac{4a}{2M} \int_{0}^{M} \int_{-M}^{M} f_M(x_2|x_1)f_M(-x_2|x_1) \, dx_2 \, dx_1
\]

where we exploited the symmetry of \( f_M \). It can be shown that the integral evaluates to \( \frac{1}{2} \), so the information loss rate is \( \bar{L}(X \rightarrow Y) = \frac{1}{2} \).

This result has a nice geometric interpretation: It quantifies the expected overlap of two segments of length \( 2a \) randomly placed on a circle with circumference \( 2M \); due to the modulo operation the point \(-M\) is equivalent to the point \( M \), and the conditional PDFs \( f_M(x_2|x_1) \) and \( f_M(-x_2|x_1) \) represent the segments (see Fig. 3).

Finally, we evaluated the upper bound from Proposition 6.

Letting \( X_1 = [-M, 0) \) and \( X_2 = [0, M) \) and abbreviating \( p(1|x) := \Pr(W_2 = 1|X_1 = x) \) we obtained

\[
p(1|x) = \begin{cases} \frac{a-M-x}{2a}, & -M \leq x < -M + a \\ 0, & -M + a \leq x < -a \\ \frac{a+x}{2a}, & -a \leq x < a \\ 1, & a \leq x < M - a \\ \frac{a+M-x}{2a}, & M - a \leq x < M \end{cases}
\]

2The redundancy is defined as the difference between the entropy of the marginal distribution and the entropy rate of the process. The former increases due to increasing variance \( \sigma_X^2 \), while the latter remains constant and equal to \( \bar{h}(Z) \) (cf. [13]).
Fig. 3. Interpreting the information loss rate of a cyclic random walk in a magnitude function. The depicted scenario corresponds to $a = M/3$.

![L(X → Y)](image)

Fig. 4. Information loss rate of a cyclic random walk $X$ on $[-M, M]$ in a magnitude function as a function of the support $[-\alpha, \alpha]$ of the uniform input PDF.

If $M > 2\alpha$ and

$$p(1|x) = \begin{cases} \frac{2-M}{2\alpha}, & -M \leq x < -\alpha \\ \frac{2a-M}{2\alpha}, & -\alpha \leq x < -M + a \\ \frac{a}{2a-M}, & -M + a \leq x < M - a \\ \frac{2M-a}{2a-M}, & M - a \leq x < a \\ \frac{M-a}{M}, & a \leq x < M \end{cases}$$

if $M \leq 2\alpha$. (Naturally, $p(2|x) = 1 - p(1|x)$.) Computing the entropy $H(W_2|X_1 = x)$ based on these probabilities and taking the expectation w.r.t. $X_1$ yields

$$H(W_2|X_1) = \left\{ \begin{array}{ll} \frac{a}{m} - \frac{M-a}{M} & M > 2a \\ \frac{a}{m} + \log \frac{2a}{M} & M \leq 2a \end{array} \right.$$ (43)

The analytic result for the information loss rate and the bound, numerically validated using the same procedure as in Section VI-A are depicted in Fig. 4.

C. An Example illustrating the Tightness of the Bounds

The following example illustrates the tightness of the presented bounds and also satisfies the condition of lumpability. Assume that $X$ is a Markov process with conditional distribution

$$f_M(x_2|x_1) = \frac{1}{2} \begin{cases} \mathbb{1}_{[1,2]}(x_2) + \mathbb{1}_{[3,4]}(x_2), & x_1 \in [0, 1) \cup [2, 3) \\ \mathbb{1}_{[0,1]}(x_2) + \mathbb{1}_{[2,3]}(x_2), & x_1 \in [1, 2) \cup [3, 4) \end{cases}$$ (44)

where $\mathbb{1}_a(x) = 1$ iff $x \in A$. As it can be shown easily, the stationary distribution is the uniform distribution on $[0, 4)$, thus, $h(X) = \log 4 = 2$ and $\hat{F}(X) = 1$.

We analyze the system mapping the interval $X_2 = [2, 4)$ onto the interval $X_1 = [0, 1)$, i.e.,

$$g(x) = \begin{cases} x, & x \in [0, 2) \\ x-2, & x \in [2, 4) \end{cases}$$ (45)

which yields the conditional distribution of $Y_2$ given $X_1$

$$f_{Y_2|X_1}(y_2|x_1) = \begin{cases} \mathbb{1}_{[1,2]}(y_2), & x_1 \in [0, 1) \cup [2, 3) \\ \mathbb{1}_{[0,1]}(y_2), & x_1 \in [1, 2) \cup [3, 4) \end{cases}$$ (46)

The derivative of this function is identical to one, the stationary distribution of the output process $Y$ is the uniform distribution on $[0, 2)$; thus, $h(Y) = \log 2 = 1$ and $L(X \to Y) = 1$.

The output process $Y$ can be shown to be Markov: Assuming $X_1 = x \in [0, 1)$, it follows that $x' \in [2, 3)$; since these conditions are equivalent in the definition of $f_M$, (44) is fulfilled.

From $f_M$ one can see that $\Pr(W_2 = 1|X_1 = x) = \Pr(W_2 = 2|X_1 = x) = \frac{1}{2}$ regardless of $x$, which satisfies (24a) and renders the upper bound from Proposition 5 as

$$H(W_2|X_1) = 1.$$ (47)

The bound can be shown to be tight, since also (24a) is fulfilled: Given, e.g., $X_1 = x \in [0, 1)$ and $Y_2 = y \in [1, 2)$, it follows that $X_2 \in \{y, y+2\}$ and $f_M(y|x) = f_M(y+2|x) = \frac{1}{2}$. Thus, we are led to the following conclusion:

$$1 = L(X \to Y) \geq \hat{L}(X \to Y) = H(W_2|X_1) = 1$$ (48)

This is an example not only for tightness of Proposition 5 but also of Proposition 3. Interestingly, neither is the function information-preserving, nor is the input process $X$ iid. Consequently, one can interpret this example as a worst-case, where redundancy is not matched to the system (the “channel”), failing to alleviate the adverse effects of the system.

D. Multirate Systems

Although strictly speaking not time-invariant, also multirate systems can be analyzed with the proposed relative information loss rates. In particular, we will show that for an $M$-fold downsampler, which is described by the input-output relation $Y_n = X_{nM}$, the information loss rate equals

$$l(X \to Y) = \frac{M-1}{M}.$$ (49)

To this end, note that the stationary output process $Y$ is equivalent to the cyclo-stationary process $\tilde{Y}$, whose samples are defined as

$$\tilde{Y}_n = \begin{cases} X_n, & \text{if } n/M \in \mathbb{Z} \\ 0, & \text{else} \end{cases}.$$ (50)

In essence, the function in (50) implements a projection on a subspace of lower dimensionality. For these type we showed in [7] that the relative information loss is related to the
information dimension of the output, which in our case is given by the number of its non-zero entries, i.e., by
\[ d(Y^n) = \frac{n}{M} . \tag{51} \]
With \( d(X^n) = n \) and by the fact that \( |n/M| = n/M + \{n/M\} \), where \( \{ \cdot \} \) denotes the fractional part, we obtain
\[ \lim_{n \to \infty} l(X^n \to Y^n) = 1 - \lim_{n \to \infty} \frac{|n/M + \{n/M\}|}{n} = 1 - \frac{1}{M} = \frac{M - 1}{M} . \tag{52} \]
The second equality follows because the magnitude of the fractional part is bounded by unity and that, thus, this term vanishes in the limit.

VII. CONCLUSION

In this work we extended previous results about the information loss induced by deterministic, memoryless input-output systems from random variables to stationary stochastic processes with continuous distribution. Notably, we showed a connection between the rate of information loss and the differential entropy rates of the input and output processes for a special class of functions. While redundancy decreases the information loss rate for this class of systems, systems which destroy an infinite amount of information do not benefit from redundancy of the process in most practical cases. Future investigations shall focus on the extension to systems with memory and on the problem of reconstructing the input process.

As side results, we presented sufficient conditions for the Markovian input process and the system function such that the output process is Markov.

APPENDIX A

PROOFS

A. Proof of Proposition 6

Note that a possible definition of Markovity is given by

Definition 4 (Markov Process [15 II.6, p. 80]). A process \( X \) is a Markov process iff for all \( i \in \mathbb{N} \), \( \alpha \in \mathbb{R} \), and integers \( n_1 < n_2 < \cdots < n_i < n \), with probability one,
\[ \Pr(X_n \leq \alpha|X_{n_1}, \ldots, X_{n_i} = x_{n_i}) = \Pr(X_n \leq \alpha|X_{n_i} = x_{n_i}) . \tag{53} \]
Clearly, a process is Markov if, for all \( n \),
\[ f_{X_n|X_{n-1}}(x) \overset{a.e.}{=} f_{X_n|X_{n-1}} \quad \text{holds } P_{X_{n-1}}-a.s. \text{, because} \tag{54} \]
results from integrating the densities over \((-\infty, \alpha]\).

The proof of the proposition follows along the same lines as the proof for Markov chains given in [16], and is built on the following Lemma, which is an extension of [15 Thm. 4.5.1]:

Lemma 1 (Bounds on the differential entropy rate). Let \( X \) be a stationary Markov process with differential entropy rate \( \overline{h}(X) = h(X_2|X_1) \) and let \( Y \) be a stationary process derived from \( X \) by \( Y_n = g(X_n) \). Then,
\[ h(Y_n|Y_{n-1}^{2-1}, X_1) \leq \overline{h}(Y) \leq h(Y_n|Y_{n-1}^{1}) . \tag{55} \]

Proof: The upper bound follows from the fact that conditioning reduces entropy, so we only have to show the lower bound. For this, note that by Markovity of \( X \),
\[ h(Y_n|Y_{n-1}^{2-1}, X_1) = h(Y_n|Y_{n-1}^{2-1}, X_1^1) \tag{56} \]
for all \( k < 1 \). Let \( U_k = \{ Y_{n-1}^{k-1}, X_1^1 \} \) and \( V_k = Y_{n-1}^{1} \). Obviously, there exists a function \( f \) such that \( V_k = f(U_k) \), namely the function which is the identity function on the first \( n-2 \), and the function \( g \) on the last \( 2-k \) elements. By showing that
\[ h(Y_n|U_k) \leq h(Y_n|V_k) \tag{57} \]
the lower bound is proved by [9 Thm. 14.7]
\[ h(Y_n|Y_{n-1}^{2-1}, X_1) = \lim_{k \to -\infty} h(Y_n|U_k) \leq \lim_{k \to -\infty} h(Y_n|V_k) = \overline{h}(Y) . \tag{58} \]
Thus, we write
\[ h(Y_n|V_k) - h(Y_n|U_k) = h(Y_n, V_k) - h(Y_n, U_k) + h(U_k) \tag{59} \]
\[ \overset{(a)}{=} H(U_k|V_k) - E \{ \log |det J_f(U_k)| \} \]
\[ - H(U_k, Y_n|V_k, Y_n) + E \{ \log |det J_f(U_k)| \} \tag{60} \]
\[ = H(U_k|V_k) - H(U_k|V_k, Y_n) \tag{61} \]
\[ \geq 0 \tag{62} \]
where \((a)\) is due the multivariate extension of (3) and since the determinant of the Jacobian matrix is the same for the function \( f \), and for a function which applies \( f \) to some, and the identity function to the rest of the elements. This completes the proof.

We now turn to the

Proof of Proposition 6. Note that the assumption implies that
\[ \int_X \int_Y f_{Y_2, X_1}(y, x) \log \left( \frac{f_{Y_2|X_1}(y|x)}{f_{Y_2|Y_1}(y|g(x))} \right) dy dx \]
\[ = h(Y_2|Y_1) - h(Y_2|X_1) = 0 \tag{63} \]
which renders the upper bounds of Lemma 1 equal for \( n = 2 \). Thus, \( \overline{h}(Y) = h(Y_n|Y_{n-1}^{1}) = h(Y_2|Y_1) \) for all \( n \). By stationarity,
\[ 0 = h(Y_n|Y_{n-1}^{1}) - h(Y_n|Y_{n-1}^{1-1}) \tag{64} \]
\[ = I(Y_n; Y_{n-1}^{1-1}|Y_{n-1}^{1}) \tag{65} \]
\[ = E \left\{ \log \frac{f_{Y_n, Y_{n-1}^{1-1}, Y_{n-1}^{1}|Y_{n-1}^{1}}}{f_{Y_n, Y_{n-1}^{1-1}}(Y_n^{1}) \cdot f_{Y_{n-1}^{1-1}, Y_{n-1}^{1}}(Y_{n-1}^{1-1})} \right\} \tag{66} \]
\[ = E \left\{ \log \frac{f_{Y_n, Y_{n-1}^{1-1}}(Y_n^{1})}{f_{Y_{n-1}^{1-1}}(Y_{n-1}^{1-1})} \right\} \tag{67} \]
\[ = E \left\{ D(f_{Y_n, Y_{n-1}^{1-1}}(\cdot, Y_{n-1}^{1-1})||f_{Y_{n-1}^{1-1}}(\cdot, Y_{n-1}^{1-1})) \right\} \tag{68} \]
where $D(\cdot \| \cdot)$ is the Kullback-Leibler divergence and where in the last line the expectation is taken w.r.t. $Y_1^{n-1}$.

The expectation of a non-negative RV, such as the Kullback-Leibler divergence above, can only be zero if this RV is almost surely zero. Together with the fact that the Kullback-Leibler divergence between two PDFs vanishes iff the PDFs are equal almost everywhere, the assumption of the proposition implies that

$$P_{Y_1^{n-1}} = P_{Y_1^{n-1}} \quad \text{w.r.t.} \quad Y_1^{n-1} = a.e.$$  \hfill (69)

But this implies Markovity by Definition 4 (cf. [54]) and completes the proof. \hfill \blacksquare

B. Proof of Corollary 7

Note that (23) implies $f_{Y_2|X_1}(y_2|x) = f_{Y_2|X_1}(y_2|x')$ for all $x, x'$ within the support of $f_X$. Now

$$f_{Y_2|X_1}(y_2|y_1) = \frac{1}{f_Y(y_1)} \sum_{x_1 \in g^{-1}_1[y_1]} f_{Y_2|X_1}(y_2|x_1) f_X(x_1) \quad \text{w.r.t.} \quad g'|g.$$ \hfill (70)

Let $g^{-1}_1[y_1] := \{ x \in g^{-1}[y_1] : f_X(x) > 0 \}$ and let $\hat{x}$ be an arbitrary element of this set. We proceed

$$f_{Y_2|X_1}(y_2|y_1) = \frac{1}{f_Y(y_1)} \sum_{x_1 \in g^{-1}_1[y_1]} f_{Y_2|X_1}(y_2|x_1) f_X(x_1) \quad \text{w.r.t.} \quad g'|g.$$ \hfill (71)

Let $f_{Y_2|X_1}(y_2|x) = f_{Y_2|X_1}(y_2|x')$ we can apply Proposition 6 to complete the first part of the proof.

For the second part, note that we have with Proposition 6

$$\hat{h}(Y) = h(Y_2|X_1)$$ \hfill (72)

and thus, with Proposition 1 and 4,

$$\hat{L}(X \rightarrow Y) = h(X_2|X_1) - h(Y_2|X_1) + E \{ \log |g'(X)| \} = H(X_2|X_1, Y_2).$$ \hfill (73)

It remains to show that $\hat{L}$ is equal to $\hat{L}$ in the proof of Proposition 5. To this end, observe that

$$H(W_2|X_1) - H(W_2|X_1, Y_2) = I(W_2; Y_2|X_1)$$ \hfill (74)

vanishes if for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$ such that $f_X(x) > 0$ and for all $w$ such that $Pr(W_2 = w|X_1 = x) > 0$

$$f_{Y_2|X_1}(y|x) = f_{Y_2|W_2, X_1}(y|w, x).$$ \hfill (75)

But

$$f_{Y_2|W_2, X_1}(y|w, x) = f_{Y_2|X_1}(g^{-1}_w(y)|x) \quad \text{w.r.t.} \quad p(w|x) |g'(g^{-1}_w(y))| \quad \text{w.r.t.} \quad g(w)$$ \hfill (76)

where $p(w|x) = Pr(W_2 = w|X_1 = x)$. Let, for a given $x$, $\hat{w}$ satisfy $p(\hat{w}|x) > 0$. The proof is completed by recognizing that

$$f_{Y_2|X_1}(y|x) = \sum_w p(w|x) f_{Y_2|W_2, X_1}(y|w, x)$$ \hfill (77)

and

$$f_{X_2|X_1}(g^{-1}_w(y)|x) = \frac{f_{X_2|X_1}(g^{-1}_w(y)|x)}{|g'(g^{-1}_w(y))|} \quad \text{w.r.t.} \quad g(w)$$ \hfill (78)

where $a$ is due to (23) and $b$ is due to (24a).

C. Proof of Proposition 7

We start with showing $l(X \rightarrow Y) = P_X(\mathcal{X}_c)$. To this end, letting $d(Z)$ denoting the Rényi information dimension of $Z$ [17] and employing [18, 19], we write

$$d(X|Y = y) = \sum_{i=1}^K d(X|y \in \mathcal{X}_c) P_X|Y=y(\mathcal{X}_i)$$ \hfill (79)

where $K = \text{card}(\{ \mathcal{X}_i \})$. W.l.o.g., the partition is indexed such that the first $L$ elements correspond to subsets $\mathcal{X}_i$ on which $g$ is constant. Thus, $\mathcal{X}_i = \bigcup_{i=1}^L \mathcal{X}_i$. It follows for $i > L$, from the bijectivity of $g_i$, that $d(X|Y = y, \mathcal{X}_c = \mathcal{X}_i) = 0$. Moreover, if for $i \leq L$ we have $X \in \mathcal{X}_i$, it follows that $Y = y_i$, and that thus

$$d(X|Y = y) = \sum_{i=1}^L d(X|Y = y, \mathcal{X}_c = \mathcal{X}_i) P_X|Y=y(\mathcal{X}_i)$$ \hfill (80)

$$= \sum_{i=1}^L P_X|Y=y(\mathcal{X}_i)$$ \hfill (81)

It remains to show that (24) is equality in (20) in the proof of Proposition 5. To this end, observe that

$$H(W_2|X_1) - H(W_2|X_1, Y_2) = I(W_2; Y_2|X_1)$$ \hfill (82)

vanishes if for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$ such that $f_X(x) > 0$ and for all $w$ such that $Pr(W_2 = w|X_1 = x) > 0$

$$f_{Y_2|X_1}(y|x) = f_{Y_2|W_2, X_1}(y|w, x).$$ \hfill (83)

But

$$f_{Y_2|W_2, X_1}(y|w, x) = f_{Y_2|X_1}(g^{-1}_w(y)|x) \quad \text{w.r.t.} \quad p(w|x) |g'(g^{-1}_w(y))| \quad \text{w.r.t.} \quad g(w)$$ \hfill (84)

where $a$ is due to the fact that the RV $X$ restricted to the non-empty set $\mathcal{X}_i$ possesses a density and $b$ follows from the fact that the partition consists only disjoint sets. We now combine $d(X) = 1$ and

$$d(X|Y) = \frac{d(X|Y = y)d_P(Y)(y)}{d(X)}$$ \hfill (85)

with the fact that $l(X \rightarrow Y) = \frac{d(X|Y = y)}{d(X)}$ [17] and obtain the first part of the proof.
Now take a finite sequence $X^n$ obtained from the stochastic process $X$ and look at the relative information loss incurred in $g$. Similarly as in the proof of Proposition 3, $g^n$ induces a finite partition of $X^n$. Moreover, for every element of this partition, $g^n$ is a composition of a bijective, differentiable function and, possibly, a projection. We can thus apply the result about dimensionality reduction presented in [7] which leads to (90) where $X^n_c = X^n \setminus X_c$. Compactly written, we get

$$l(X^n_1 \rightarrow Y^n_1) = \frac{1}{n} \sum_{i=1}^{n} \Pr(\text{card}(\{X_j \in X^n_1 : X_j \in X_c\}) = i).$$

Defining

$$V_n := \begin{cases} 1, & \text{if } X_n \in X_c \\ 0, & \text{else} \end{cases}$$

and $Z_n := \sum_{j=1}^{n} V_j$, and with the linearity of expectation we get

$$l(X^n_1 \rightarrow Y^n_1) = \frac{1}{n} \sum_{i=1}^{n} \Pr(\sum_{j=1}^{n} V_j = i)$$

(94)

$$= \frac{1}{n} \sum_{i=1}^{n} \Pr(Z_n = i)$$

(95)

$$= \frac{1}{n} \mathbb{E}\{Z_n\}$$

(96)

$$= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\{V_j\}$$

(97)

$$= \frac{1}{n} \mathbb{E}\{V\}$$

(98)

$$= \mathbb{E}\{\mathbb{E}\{V\}\}$$

(99)

where (a) is due to stationarity of $X$. This completes the proof.

REFERENCES

[1] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, 3rd ed. Upper Saddle River, NJ: Pearson Higher Ed., 2010.

[2] R. Price, “A useful theorem for nonlinear devices having Gaussian inputs,” IEEE Trans. Inf. Theory, vol. 4, pp. 69–72, Jun. 1958.

[3] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ: Wiley Interscience, 2006.

[4] M. S. Pinsker, Information and Information Stability of Random Variables and Processes. San Francisco, CA: Holden Day, 1964.

[5] S. Watanabe and C. T. Abraham, “Loss and recovery of information by coarse observation of stochastic chain,” Information and Control, vol. 3, no. 3, pp. 248–278, Sep. 1960.

[6] B. C. Geiger and G. Kubin, “Some results on the information loss in dynamical systems,” in Proc. IEEE Int. Symp. Wireless Communication Systems (ISWCS), Aachen, Nov. 2011, pp. 794–798, extended version available: arXiv:1106.2404 [cs.IT].

[7] ——, “Information measures for deterministic input-output systems,” Mar. 2013, submitted; preprint available: arXiv:1303.6409 [cs.IT].

[8] J. G. Kemeny and J. L. Snell, Finite Markov Chains, 2nd ed. Springer, 1976.

[9] A. Papoulis and U. S. Pillai, Probability, Random Variables and Stochastic Processes, 4th ed. New York, NY: McGraw Hill, 2002.

[10] C. J. Burke and M. Rosenblatt, “A Markovian function of a Markov chain,” Ann. Math. Stat., vol. 29, no. 4, pp. 1112–1122, Dec. 1958.

[11] L. C. G. Rogers and J. W. Pitman, “Markov functions,” Ann. Prob., vol. 9, no. 4, pp. 573–582, 1981.

[12] M. Rosenblatt, “Functions of a Markov process that are Markovian,” in Selected Works of Murray Rosenblatt, K.-S. Lii, and D. N. Politis, Eds. Springer, 2011, pp. 134–146.

[13] R. Modde, “Information measures for deterministic input-output systems,” Mar. 2013, submitted; preprint available: arXiv:1303.6409 [cs.IT].

[14] M. Dumitrescu and G. Popovici, “Entropy invariance for autoregressive processes constructed by linear filtering,” International Journal of Computer Mathematics, vol. 88, no. 4, pp. 864–880, Mar. 2011.

[15] J. L. Doob, Stochastic Processes, ser. Wiley Classics Library. New York, NY: Wiley Interscience, 1990.

[16] B. C. Geiger and C. Temmel, “Lumping of Markov chains and entropy rate loss,” Dec. 2012, submitted to J. Appl. Prob.; preprint available: arXiv:1212.4375 [cs.IT].

[17] A. Rényi, “On the dimension and entropy of probability distributions,” Acta Mathematica Hungarica, vol. 10, no. 1-2, pp. 193–215, Mar. 1959.

[18] Y. Wu and S. Verdú, “Rényi information dimension: Fundamental limits of almost lossless analog compression,” IEEE Trans. Inf. Theory, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[19] M. Śmieja and J. Tabor, “Entropy of the mixture of sources and entropy dimension,” IEEE Trans. Inf. Theory, vol. 58, no. 5, pp. 2719–2728, May 2012.