Information Loss in Static Nonlinearities

Bernhard C. Geiger*, Christian Feldbauer*, Gernot Kubin*
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{geiger,feldbauer,gernot.kubin}@tugraz.at

***Abstract*—In this work, conditional entropy is used to quantify the information loss induced by passing a continuous random variable through a memoryless nonlinear input-output system. We derive an expression for the information loss depending on the input density and the nonlinearity and show that the result is strongly related to the non-injectivity of the considered system. Tight upper bounds are presented, which can be evaluated with less difficulty than a direct evaluation of the information loss, which involves the logarithm of a sum. Application of our results is illustrated on a set of examples.**

I. Introduction

Information processing, in the sense of changing the information of or retrieving information from a signal, can only be accomplished by nonlinear systems, while causal, stable, linear systems do not affect a signal's entropy rate [1], [2, pp. 663]. As a consequence, in the past information-theoretic measures in system analysis were almost exclusively used for highly nonlinear, chaotic systems, mainly motivated by the works of Kolmogorov [3], [4] and Sinai [5]. On the contrary, linear systems and relatively simple nonlinear systems (e.g., containing static nonlinearities) usually lack information-theoretic descriptions and are often characterized by second-order statistics or energetic measures (e.g., transfer function, power spectrum, signal-to-distortion ratio, mean square error between input and output, correlation functions, etc.).

In this work, we characterize the amount of information lost by passing a signal through a static nonlinearity. These systems, although simple, are by no means irrelevant in technical applications: One of the major components of the energy detector, a low-complexity receiver architecture for wireless communications, is a square-law device. Rectifiers, omnipresent in electronic systems are another example for static nonlinearities, which further constitute the nonlinear components in Wiener and Hammerstein systems. This work thus acts as a first step towards the goal of a comprehensive information-theoretic framework for more general nonlinear systems, providing an alternative to the prevailing energetic descriptions. While an analysis of information rates will be left for future work, this paper is concerned with zeroth-order entropies only.

Information loss can most generally be expressed as the difference of mutual informations,

$$I(\hat{X};X) - I(\hat{X};Y) \tag{1}$$

where the random variables (RV) X and Y are two descriptions for another RV $\hat{X}$. In words, the difference in (1) is the information lost by changing the description from X to Y (cf. Fig. 1). This kind of information loss is of particular interest for learning/coding/clustering (e.g., word clustering [6]) and triggered the development of optimal representation techniques [7].

In case $\hat{X}$ is identical to the RV X itself, the information loss simplifies to (cf. proof of Theorem 1)

$$H(X|Y) \tag{2}$$

i.e., to the conditional entropy of X given the description Y. This equivocation, as Shannon termed it in his seminal paper [1], was originally used to describe the information loss for stochastic relations between the RVs X and Y. In contrary to that, we are concerned with deterministic functions $Y = g(X)$.

To our knowledge, little work has been done in this regard. Some results are available for the capacity of nonlinear channels [8], [9], and recently the capacity of a noisy (possibly nonlinear and non-injective) function was analyzed [10], [11]. Considering deterministic systems, we found that Pippenger used equivocation to characterize the information loss induced by multiplying two integer numbers [12], while the coarse observation of discrete stochastic processes is analyzed in [13]. An analysis of how much information is lost by passing a continuous RV through a static nonlinearity cannot be found in the literature.

Aside from providing information-theoretic descriptions for the nonlinear systems mentioned above, our results also apply to different fields of signal processing and communication theory. To be specific, the information loss may prove useful to compute error bounds for the reconstruction of nonlinearly distorted signals and in capacity considerations for nonlinear channels.

After introducing the problem statement in Section II, an expression for the information loss is derived and related to the non-injectivity of the system in Section III, while bounds on the information loss are presented in Section IV. Section V illustrates the theoretical results with the help of examples. An extended version of this paper, containing some additional proofs and examples, is available in [14].

II. Problem Statement

We focus our attention on a class of systems whose input-output behavior can be described by a piecewise strictly monotone function. While this excludes functions which are constant on some proper interval (e.g., limiters or quantizers, for which it can be shown that the information loss becomes infinite for a continuous-valued input), many well-behaved

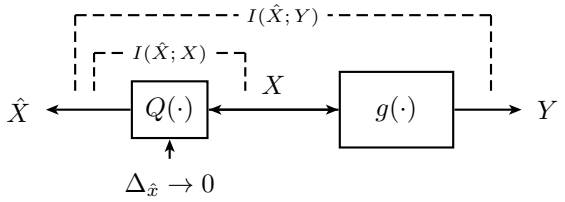

Fig. 1. Equivalent model for computing the information loss of an input-output system with static nonlinearity $g(\cdot)$. $Q(\cdot)$ is a quantizer with quantization step size $\Delta_{\hat{x}}$.

functions can be interpreted in the light of the forthcoming Definition:

Definition 1. Let g: $\mathcal{X} \to \mathcal{Y}$, $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$, be a bounded, surjective, Borel measurable function which is piecewise strictly monotone on L subdomains $\mathcal{X}_l$

$$g(x) = \begin{cases} g_1(x), & \text{if } x \in \mathcal{X}_1 \\ g_2(x), & \text{if } x \in \mathcal{X}_2 \\ \vdots & \\ g_L(x), & \text{if } x \in \mathcal{X}_L \end{cases} \tag{3}$$

where g_l: $\mathcal{X}_l \to \mathcal{Y}_l$ are bijective. We assume without loss of generality that the subdomains are an ordered set of disjoint, proper intervals with $\bigcup_{l=1}^{L} \mathcal{X}_l = \mathcal{X}$ and $x_i < x_j$ for all $x_i \in \mathcal{X}_i$, $x_j \in \mathcal{X}_j$ whenever $i < j$. We further require all $g_l(\cdot)$ to be differentiable on the interval enclosure of $\mathcal{X}_l$.

Note that $\mathcal{X}$ does not need to be an interval itself. Strict monotonicity implies that the function is invertible on each interval $\mathcal{X}_l$, i.e., there exists an inverse function g_l^{-1}: $\mathcal{Y}_l \to \mathcal{X}_l$, where $\mathcal{Y}_l$ is the image of $\mathcal{X}_l$. However, the function $g(\cdot)$ needs not be invertible on $\mathcal{X}$, i.e., it can be non-injective. Equivalently, the images of the intervals, $\mathcal{Y}_l$, unite to $\mathcal{Y}$, but need not be disjoint. Let $g(\cdot)$ describe the input-output behavior of the system under consideration (see Fig. 1).

As an input to this system consider a sequence of independent samples, identically distributed with continuous cumulative distribution function (CDF) $F_X(x)$ and probability density function (PDF) $f_X(x)$. Without loss of generality, let the support of this RV be $\mathcal{X}$, i.e., $f_X(x)$ is positive on $\mathcal{X}$ and zero elsewhere.

As an immediate consequence of this system model, the conditional PDF of the output Y given the input X can be written as [15]

$$f_{Y|X}(x,y) = \delta(y - g(x)) = \sum_{i \in \mathbb{I}(y)} \frac{\delta(x - x_i)}{|g'(x_i)|} \tag{4}$$

where $\delta(\cdot)$ is Dirac's delta distribution, $\mathbb{I}(y) = \{i : y \in \mathcal{Y}_i\}$ and $x_i = g_i^{-1}(y)$ for all $i \in \mathbb{I}(y)$, and $g'(\cdot)$ is the derivative of $g(\cdot)$. In other words, $\{x_i\}$ is the preimage of y or the set of roots satisfying $y = g(x)$. The marginal PDF of Y is thus given as [2, pp. 130], [15]

$$f_Y(y) = \sum_{i \in \mathbb{I}(y)} \frac{f_X(x_i)}{|g'(x_i)|}. \tag{5}$$

III. Information Loss of Static Nonlinearities

In what follows we quantify the information loss induced by $g(\cdot)$, and we show that this information loss is identical to the remaining uncertainty from which interval $\mathcal{X}_l$ the input x originated after observing the output y. The main contribution of this work is thus concentrated in the following two Theorems.

Theorem 1. *The information loss induced by a function $g(\cdot)$ satisfying Definition 1 is given as*

$$H(X|Y) = \int_{\mathcal{X}} f_X(x) \log \left(\frac{\sum_{i \in \mathbb{I}(g(x))} \frac{f_X(x_i)}{|g'(x_i)|}}{\frac{f_X(x)}{|g'(x)|}} \right) dx. \tag{6}$$

Proof: Using identities from [16] and the model in Fig. 1 the conditional entropy $H(X|Y)$ can be calculated as

$$\begin{aligned} H(X|Y) &= \lim_{\hat{X} \to X} \left(H(\hat{X}|Y) - H(\hat{X}|X) \right) \\ &= \lim_{\hat{X} \to X} \left(H(\hat{X}) - H(\hat{X}|X) - H(\hat{X}) + H(\hat{X}|Y) \right) \\ &= \lim_{\hat{X} \to X} \left(I(\hat{X};X) - I(\hat{X};Y) \right). \end{aligned} \tag{7}$$

where $\hat{X}$ is a discrete RV converging surely to X. One can interpret this convergence as $\hat{X}$ being the quantization of X, where the quantization intervals are made increasingly fine (see Fig. 1). Here, motivated by the data processing inequality [16, pp. 34], we have related the conditional entropy to a difference of mutual informations, which we have introduced as the most general notion of information loss in Section I.

For the mutual information between X and $\hat{X}$ we can write with [16, pp. 251]

$$I(\hat{X}, X) = \int_{\mathcal{X}} \int_{\hat{\mathcal{X}}} f_{\hat{X}X}(\hat{x}, x) \log \left(\frac{f_{X|\hat{X}}(\hat{x}, x)}{f_X(x)} \right) d\hat{x} dx. \tag{8}$$

Similarily, with [2, pp. 142, Theorem 5-1] (the logarithm and all PDFs are measurable) we get for $I(\hat{X}, Y)$

$$I(\hat{X}; Y) = \int_{\mathcal{X}} \int_{\hat{\mathcal{X}}} f_{\hat{X}X}(\hat{x}, x) \log \left(\frac{f_{Y|\hat{X}}(\hat{x}, g(x))}{f_Y(g(x))} \right) d\hat{x} dx. \tag{9}$$

After subtracting these expressions according to (7) we can exchange limit and integration (for a rigorous proof see [14]). In the limit the conditional PDFs assume $f_{X|\hat{X}}(\hat{x}, x) = \delta(x - \hat{x})$ and $f_{Y|\hat{X}}(\cdot,\cdot) = f_{Y|X}(\cdot,\cdot)$, thus (4), and using these we obtain (10) at the bottom of the next page. Since the integral over $\hat{x}$ is zero for $\hat{x} \neq x$ due to $\delta(x - \hat{x})$, only the term satisfying $x_k = x$ remains from the sum over Dirac's deltas in the denominator; this term cancels with the delta in the numerator. Integrating over $\hat{x}$ and substituting (5) for $f_Y(\cdot)$ finally yields

$$H(X|Y) = \int_{\mathcal{X}} f_X(x) \log \left(\frac{\sum_{i \in \mathbb{I}(g(x))} \frac{f_X(x_i)}{|g'(x_i)|}}{\frac{f_X(x)}{|g'(x)|}} \right) dx \tag{11}$$

and completes the proof. ■

Note that for $\hat{X} \to X$ both $I(\hat{X};X)$ and $I(\hat{X};Y)$ diverge to infinity, but their difference not necessarily does. Further, if for all $y \in \mathcal{Y}$ the preimage is a singleton ($|\mathbb{I}(g(x))| = 1$ for all $x \in \mathcal{X}$), $g(\cdot)$ is bijective and the information loss $H(X|Y) = 0$.

To underline the dependency of $H(X|Y)$ on the injectivity of $g(\cdot)$, we will now show that the information loss is identical to the uncertainty about the interval $\mathcal{X}_l$ from which the input x originated given the output value y. To this end, let us introduce the following Definition:

Definition 2. Let W be a discrete RV with $|\mathcal{W}| = L$ mass points which is defined as

$$W = w_i \quad \text{if} \quad x \in \mathcal{X}_i \tag{12}$$

for all $i = 1, \ldots, L$.

In other words, W is a discrete RV which depends on the interval $\mathcal{X}_l$ of x, and not on its actual value. We are now ready to state the following Theorem:

Theorem 2. *The uncertainty about the input value x after observing the output y is identical to the uncertainty about the interval $\mathcal{X}_l$ from which the input was taken, i.e.,*

$$H(X|Y) = H(W|Y). \tag{13}$$

Proof: See [14]. ■

The information loss induced by a function satisfying Definition 1 is thus only related to the roots of the equation $y = g(x)$. Conversely, if the interval $\mathcal{X}_l$ of x is known, the exact value of x can be reconstructed after observing y.

This Theorem, relating the information loss for a continuous RV to the conditional entropy of a discrete RV not only provides insight in the cause of information loss. It also suggests the development of Fano-type inequalities [16] to bound the error for reconstructing nonlinearly distorted signals.

Aside from the properties of conditional entropies (non-negativity [16, pp. 15], asymmetry in its arguments, etc.) the information loss has an important property concerning the cascade of deterministic, static systems, which is not shared by the conditional entropy in general.

Theorem 3. *Consider two functions $g\colon \mathcal{X} \to \mathcal{Y}$ and $h\colon \mathcal{Y} \to \mathcal{Z}$ satisfying Definition 1 and a cascade of systems implementing these functions. Let $Y = g(X)$ and $Z = h(Y)$. Then, the information loss induced by this cascade, or equivalently, by the composition $(h \circ g)(\cdot) = h(g(\cdot))$ is given by:*

$$H(X|Z) = H(X|Y) + H(Y|Z) \tag{14}$$

Proof: See [14]. ■

This result does not imply that the order in which the functions can be arranged has no influence on the information loss of the cascade, as one would expect from stable, linear systems. Illustrative examples showing that this does not hold can be found, e.g., in [17]. Moreover, calculating the individual information losses requires in each case the PDF of the input to the function under consideration. While this does not seem to yield an improvement compared to a direct evaluation of (6), Theorem 3 can be used to bound the information loss of the cascade efficiently whenever bounds on the individual information losses are available. We will introduce such bounds in the next Section.

IV. Upper Bounds on the Information Loss

In many situations it might be inconvenient, or even impossible, to evaluate the information loss (6) analytically since it involves the logarithm of a sum, for which only inequalities exist [16]. Therefore, one has to resort to numerical integration or use bounds on the information loss which are simpler to evaluate. In this Section we derive an upper bound which requires only minor knowledge about the function $g(\cdot)$ – namely, the number of intervals L – and we show that this bound is tight.

Theorem 4. *The information loss induced by a function $g(\cdot)$ satisfying Definition 1 can be upper bounded by the following ordered set of inequalities:*

$$\begin{aligned} H(X|Y) &\leq \int_{\mathcal{Y}} f_Y(y) \log\left(|\mathbb{I}(y)|\right) dy && (15)\\ &\leq \log\left(\sum_{l=1}^{L} \int_{\mathcal{Y}_l} f_Y(y) dy\right) && (16)\\ &\leq \log L && (17) \end{aligned}$$

These bounds hold with equality if and only if

$$\sum_{k \in \mathbb{I}(g(x))} \frac{f_X(x_k)}{|g'(x_k)|} \frac{|g'(x)|}{f_X(x)} = L \;\forall x \in \mathcal{X}. \tag{18}$$

Proof: We give here only a sketch of the proof, the detailed proof can be found in [14]: The first inequality is obtained by upper bounding $H(W|Y = y)$ using the maximum entropy property of the uniform distribution, the second is due to Jensen. The last inequality is tight if $\mathcal{Y}_l = \mathcal{Y}$ for all l. ■

An example of a function $g(\cdot)$ satisfying (18) assumes on each interval $\mathcal{X}_l$ the cumulative distribution function $F_X(x)$, modified with an additive constant and, possibly, with a sign. In other words, for all $l = 1, \ldots, L$

$$g_l(x) = b_l F_X(x) + c_l \tag{19}$$

where $b_l \in \{1, -1\}$ and $c_l \in \mathbb{R}$ are appropriate constants such that $g_l\colon \mathcal{X}_l \to \mathcal{Y}$, i.e., $\mathcal{Y}_l = \mathcal{Y}$ for all $l = 1, \ldots, L$. In order

$$H(X|Y) = \lim_{\hat{X} \to X} \Big(I(\hat{X};X) - I(\hat{X};Y) \Big) = \int_{\mathcal{X}} \int_{\hat{\mathcal{X}}} f_X(x) \delta(x - \hat{x}) \log\left(\frac{\delta(x-\hat{x}) f_Y(g(x))}{f_X(x) \sum_{k \in \mathbb{I}(g(x))} \frac{\delta(\hat{x} - x_k)}{|g'(x_k)|}} \right) d\hat{x} dx \tag{10}$$

that such constants exist, all intervals $\mathcal{X}_l$ have to contain the same probability mass, i.e.,

$$\int_{\mathcal{X}_l} f_X(x)dx = \frac{1}{L}. \tag{20}$$

For example, if $b_l = 1$ for all l, the constants c_l have to be set to

$$c_l = -\sum_{i=1}^{l-1} \int_{\mathcal{X}_i} f_X(x)dx = -\frac{l-1}{L}. \tag{21}$$

Other examples of functions satisfying the tightness conditions of Theorem 4 are given in Section V-A and in [14].

V. Examples

In this Section, the application of the obtained expression for the information loss and its upper bounds is illustrated. We start with a square-law function, as it is used, e.g., in energy detection receivers. Next, for an asymmetric function it is shown that different PDF shaping on different intervals can reduce information loss. In the last example we consider a third-order polynomial. This is of particular practical importance, since according to Weierstrass all continuous functions on bounded intervals can be approximated by polynomials.

Unless otherwise noted, the logarithm is taken to base 2.

A. Example 1: Even PDF, Square Function

Consider a continuous RV X with an even PDF, i.e., $f_X(-x) = f_X(x)$. Let the support $\mathcal{X} = \mathbb{R}$ and let this RV be the input to the square function, i.e.,

$$g(x) = x^2. \tag{22}$$

The square function is piecewise strictly monotone on $\mathcal{X}_1 = (-\infty, 0)$ and $\mathcal{X}_2 = [0, \infty)$, and with $L = 2$ we obtain the largest bound from Theorem 4 as

$$H(X|Y) \leq \log 2 = 1. \tag{23}$$

Both intervals are mapped to the positive (non-negative) real axis, i.e., $\mathcal{Y}_1 \cup \{0\} = \mathcal{Y}_2 = \mathcal{Y} = [0, \infty)$, which implies that the second bound in Theorem 4 also yields $H(X|Y) \leq 1$. There are two partial inverses mapping $\mathcal{Y}$ to the subdomains of $\mathcal{X}$:

$$x_1 = g_1^{-1}(y) = -\sqrt{y} = -|x|, \text{ and} \tag{24}$$

$$x_2 = g_2^{-1}(y) = \sqrt{y} = |x|. \tag{25}$$

Thus for all $x \in \mathcal{X}$ we have $|\mathbb{I}(g(x))| = 2$, which renders the smallest bound of Theorem 4 as $H(X|Y) \leq 1$. The magnitude of the first derivative $|g'(x)| = 2|x|$ is the same for all elements of the preimage and thus cancels in (6). By combining (22) with the two partial inverses and using the result in (6) we obtain the information loss

$$\begin{aligned} H(X|Y) = &\int_{\mathcal{X}_1} f_X(x) \log\left(\frac{f_X(x) + f_X(-x)}{f_X(x)}\right) dx \\ &\int_{\mathcal{X}_2} f_X(x) \log\left(\frac{f_X(-x) + f_X(x)}{f_X(x)}\right) dx \\ = &\log 2 \int_{\mathcal{X}} f_X(x)dx = 1 \end{aligned}$$

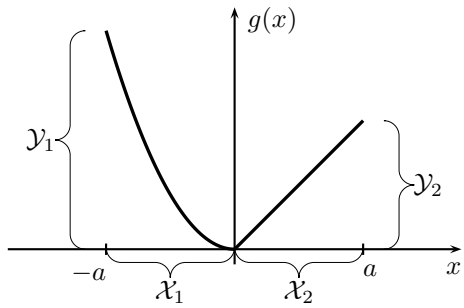

Fig. 2. Piecewise strictly monotone function of Example 2

which shows that all bounds of Theorem 4 are tight in this example.

The conditional entropy is identical to one bit. In other words, if an RV with an even PDF (thus, with equal probability masses for positive and negative values) is fed through a magnitude function, one bit of information is lost.

B. Example 2: Zero-Mean Uniform PDF, Piecewise Strictly Monotone Function

Consider an RV X uniformly distributed on $[-a, a]$, $a \geq 1$, and a function $g(\cdot)$ defined as:

$$g(x) = \begin{cases} x^2, & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases}. \tag{26}$$

This function, depicted in Fig. 2, is piecewise strictly monotone on $[-a, 0)$ and $[0, a]$. As shown in [14] evaluating (6) results in

$$\begin{aligned} H(X|Y) = &\frac{4a + 4\sqrt{a} + 1}{8a} \log(2\sqrt{a} + 1) \\ &- \frac{\log(2\sqrt{a})}{2} - \frac{1}{4\sqrt{a}\ln 2} \end{aligned}$$

where $\ln$ is the natural logarithm. For $a = 1$ this approximates to $H(X|Y) \approx 0.922$ bits. The information loss is slightly less than one bit, despite the fact that two equal probability masses collapse and the complete sign information is lost. As it can be seen from this result and Fig. 2, different PDF shaping effects for probability masses from different subdomains allow for the implicit retrieval of parts of the sign information even for $a = 1$.

By evaluating the bounds from Theorem 4 we obtain

$$H(X|Y) \leq \frac{1 + \sqrt{a}}{2\sqrt{a}} \leq \log\left(\frac{3\sqrt{a} + 1}{2\sqrt{a}}\right) \leq 1 \tag{27}$$

which for $a = 1$ all evaluate to 1 bit. The bounds are not tight as the conditions of Theorem 4 are not met in this case.

C. Example 3: Normal PDF, Third-order Polynomial

Finally, consider a Gaussian RV $X \sim \mathcal{N}(0, \sigma^2)$ and the function

$$g(x) = x^3 - 100x \tag{28}$$

depicted in Fig. 3. An analytic computation of the information loss is prevented by the logarithm of a sum in (6). Still, we

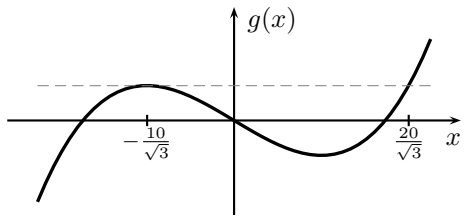

Fig. 3. Third-order polynomial of Example 3

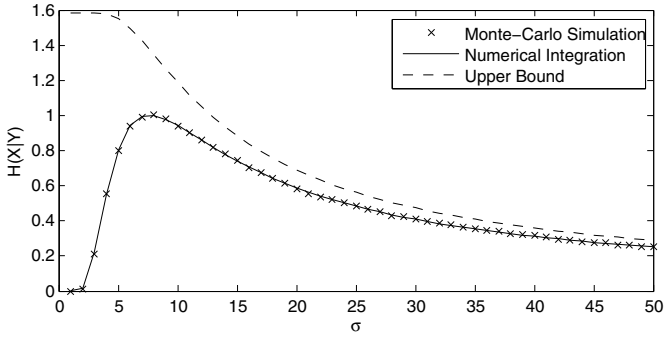

Fig. 4. Information loss of Example 3

will show that with the help of Theorem 4 at least a bound on the information loss can be computed.

Judging from the extrema of this function, three piecewise monotone intervals can be defined. Further, the domain which is mapped bijectively can be shown to be identical to $\mathcal{X}_b = \left(-\infty, -\frac{20}{\sqrt{3}}\right] \cup \left[\frac{20}{\sqrt{3}}, \infty\right)$ and contains a probability mass of

$$P_b = 2F_X\left(-\frac{20}{\sqrt{3}}\right) = 2Q\left(\frac{20}{\sqrt{3}\sigma}\right) \tag{29}$$

where $Q(\cdot)$ is the Q-function. With this result and the fact that

$$P_b = \int_{\mathcal{X}_b} f_X(x)dx = \int_{\mathcal{Y}_b} f_Y(y)dy \tag{30}$$

for a bijective mapping $g(\cdot)$ between $\mathcal{X}_b$ and $\mathcal{Y}_b$ we can upper bound the information loss by Theorem 4:

$$H(X|Y) \leq \int_{\mathcal{Y}} f_Y(y)\log(|\mathbb{I}(y)|)dy \tag{31}$$

$$= \int_{\mathcal{Y}\setminus\mathcal{Y}_b} f_Y(y)\log 3dy = (1-P_b)\log 3 \tag{32}$$

since $|\mathbb{I}(y)| = 1$ if $y \in \mathcal{Y}_b$ and $|\mathbb{I}(y)| = 3$ if $y \in \mathcal{Y}\setminus\mathcal{Y}_b$. In Fig. 4, this bound is illustrated together with the results from numerical integration of (6) and from Monte-Carlo simulations of the information loss.

VI. Conclusion

We presented an analytic expression for the information loss induced by a static nonlinearity. It was shown that this information loss is strongly related to the non-injectivity of the system, i.e., to the fact that a particular output can result from multiple possible inputs. Conversely, given a certain output, the input to the system under consideration is uncertain only with respect to the roots of the equation describing the system. The information loss can be utilized, e.g., for estimating the reconstruction error for nonlinearly distorted signals.

Since the obtained expression involves the integral over the logarithm of a sum, bounds on the information loss were derived which can be computed with much less difficulty. In particular, it was shown that the information loss for a piecewise strictly monotone function is upper bounded by the logarithm of the number of subdomains, and that this bound is tight.

Generalizations of these results to rates of information loss and nonlinear systems with memory, as well as the extension to discrete random variables are the object of future work.

Acknowledgments

The authors gratefully acknowledge discussions with Sebastian Tschiatschek concerning mathematical notation, and his comments improving the quality of this manuscript.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 623–656, Oct. 1948.

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

[3] A. N. Kolmogorov, "A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces," *Dokl. Akad. Nauk. SSSR*, vol. 119, pp. 861–864, 1958.

[4] ——, "Entropy per unit time as a metric invariant of automorphisms," *Dokl. Akad. Nauk. SSSR*, vol. 124, pp. 754–755, 1959.

[5] Y. Sinai, "On the concept of entropy for a dynamic system," *Dokl. Akad. Nauk. SSSR*, vol. 124, pp. 768–771, 1959.

[6] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, Mar. 2003.

[7] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Allerton, Sep. 1999, pp. 368–377.

[8] P. Zillmann and G. P. Fettweis, "On the capacity of multicarrier transmission over nonlinear channels," in *Proc. IEEE Vehicular Technology Conf. (VTC)*, Stockholm, Jun. 2005, pp. 1148–1152.

[9] I. Abou-Faycal and J. Fahs, "On the capacity of some deterministic nonlinear channels subject to additive white Gaussian noise," in *Proc. IEEE Int. Conf. on Telecommunications (ICT)*, Doha, Nov. 2009, pp. 63–70.

[10] F. Simon, "Capacity of a noisy function," in *Proc. IEEE Information Theory Workshop (ITW)*, Dublin, Aug.-Sep. 2010.

[11] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, Oct. 2007.

[12] N. Pippenger, "The average amount of information lost in multiplication," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 684–687, Feb. 2005.

[13] 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.

[14] B. C. Geiger, C. Feldbauer, and G. Kubin, "Information loss in static nonlinearities," in *Proc. IEEE Int. Sym. Wireless Communication Systems (ISWSC)*, Aachen, Nov. 2011, accepted; preprint available: `arXiv:1102.4794 [cs.IT]`.

[15] A. Chi and T. Judy, "Transforming variables using the Dirac generalized function," *The American Statistician*, vol. 53, no. 3, pp. 270–272, Aug. 1999.

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

[17] D. H. Johnson, "Information theory and neural information processing," *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 653–666, Feb. 2010.