Chapter 2
Piecewise Bijective Functions and Continuous Inputs

In this section, we treat the class of systems that can be described by *piecewise bijective functions*. We call a function piecewise bijective if every output value originates from at most countably many input values, i.e., if the preimage of every output value is an at most countable set. The full-wave rectifier, stripping the sign off the input, for example, is piecewise bijective. The quantizer, mapping intervals to points, is not. Piecewise bijective functions often lead to finite information loss, even if the input has a continuous distribution. In the full-wave rectifier, for example, not more than one bit can be lost.

Throughout this section, we assume that the input RV $X := (X_1, \ldots, X_N)$ is N-dimensional and continuous, i.e., its probability measure P_X is absolutely continuous w.r.t. the N-dimensional Lebesgue measure λ^N ($P_X \ll \lambda^N$). It can thus be described by a PDF f_X with support $\mathcal{X} \subseteq \mathbb{R}^N$.

Definition 2.1 (*Piecewise Bijective Function (cf. [GFK11, Definition 1])* Let $\{\mathcal{X}_i\}$ be a countable partition of $\mathcal{X}$. A piecewise bijective function (PBF) $g\colon \mathcal{X} \to \mathcal{Y}$, $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N$, is a surjective function defined in a piecewise manner:

$$g(x) = \begin{cases} g_1(x), & \text{if } x \in \mathcal{X}_1 \\ g_2(x), & \text{if } x \in \mathcal{X}_2 \\ \vdots \end{cases} \tag{2.1}$$

where each $g_i\colon \mathcal{X}_i \to \mathcal{Y}_i$ is bijective. If x_i and y_j are the i-th and j-th coordinate of x and y, respectively, then we can write $y_j = g^{(j)}(x)$, and obtain the Jacobian matrix as

B.C. Geiger and G. Kubin, *Information Loss in Deterministic Signal Processing Systems*, Understanding Complex Systems,
DOI 10.1007/978-3-319-59533-7_2

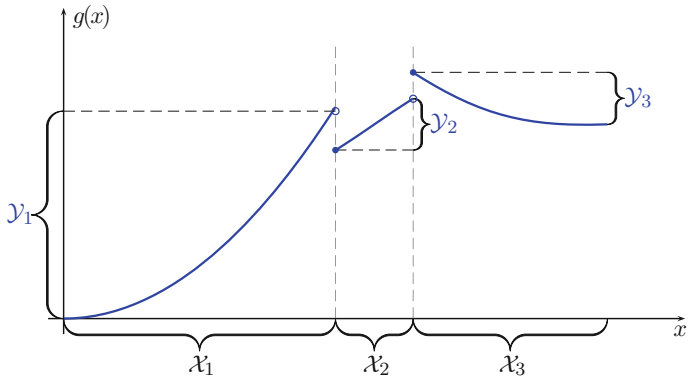


Fig. 2.1 A piecewise bijective functions with $\mathrm{card}(\{\mathcal{X}_i\}) = 3$

$$\mathcal{J}_g(x) := \begin{bmatrix} \frac{\partial g^{(1)}}{\partial x_1} & \frac{\partial g^{(1)}}{\partial x_2} & \cdots & \frac{\partial g^{(1)}}{\partial x_N} \\ \frac{\partial g^{(2)}}{\partial x_1} & \frac{\partial g^{(2)}}{\partial x_2} & \cdots & \frac{\partial g^{(2)}}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g^{(N)}}{\partial x_1} & \frac{\partial g^{(N)}}{\partial x_2} & \cdots & \frac{\partial g^{(N)}}{\partial x_N} \end{bmatrix}. \tag{2.2}$$

We assume that $\mathcal{J}_g(\cdot)$ exists P_X-almost surely (a.s.) and that its determinant, $\det\mathcal{J}_g(\cdot)$, is non-zero P_X-a.s.

An example for a PBF is shown in Fig. 2.1.

2.1 The PDF of $Y = g(X)$

Lemma 2.1 (Change of Variables [PP02, p. 244]) *Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$, let g be a PBF, and let $Y = g(X)$. The PDF of $Y = g(X)$ is given by*

$$f_Y(y) = \sum_{i=1}^{\mathrm{card}(\{\mathcal{X}_i\})} \frac{f_X(g_i^{-1}(y))}{|\det\mathcal{J}_g(g_i^{-1}(y))|} = \sum_{x\in g^{-1}[y]} \frac{f_X(x)}{|\det\mathcal{J}_g(x)|}. \tag{2.3}$$

Note that if $y \notin g_i(\mathcal{X})$, then of course $g_i^{-1}(y)$ is empty. We omit the proof of this lemma, but wish to give an intuitive explanation. The probability mass contained in a differential area is invariant under a change of variables. Suppose that for a given y, its preimage under g contains two elements x and x' corresponding to g_1 and g_2, and that both g_1 and g_2 are increasing functions. We must ensure that

$$\mathbb{P}(y \le Y \le y + \mathrm{d}y) = \mathbb{P}(x \le X \le x + \mathrm{d}x) + \mathbb{P}\left(x' \le X \le x' + \mathrm{d}x'\right). \tag{2.4}$$

Here, $\mathrm{d}y$ is an N-dimensional hypercube, while $\mathrm{d}x$ and $\mathrm{d}x'$ are N-dimensional parallelepipeds spanned by vectors $\mathrm{d}x_1, \ldots, \mathrm{d}x_N$ and $\mathrm{d}x'_1, \ldots, \mathrm{d}x'_N$. These paralellepipeds are generated by a linear transform of the hypercube $\mathrm{d}y$ via $\mathcal{J}_{g_1^{-1}}(y)$ and $\mathcal{J}_{g_2^{-1}}(y)$, the Jacobian matrices of g_1^{-1} and g_2^{-1}, respectively. Since for a linear transform $\mathbf{T}$ and a set A, the Lebesgue measure satisfies $\lambda^N(\mathbf{T}A) = |\det\mathbf{T}|\lambda^N(A)$, we obtain

$$\begin{aligned}\mathbb{P}(y \le Y \le y + \mathrm{d}y) &\approx f_Y(y)\lambda^N(\mathrm{d}y) \\ &= f_X(x)|\det\mathcal{J}_{g_1^{-1}}(y)|\lambda^N(\mathrm{d}y) + f_X(x')|\det\mathcal{J}_{g_2^{-1}}(y)|\lambda^N(\mathrm{d}y)\end{aligned} \tag{2.5}$$

and hence

$$f_Y(y) = \frac{f_X(x)}{|\det\mathcal{J}_g(x)|} + \frac{f_X(x')}{|\det\mathcal{J}_g(x')|}. \tag{2.6}$$

We present a short example to illustrate this result:

Example 2 **Full-Wave Rectifier** Let X be one-dimensional and have PDF f_X supported on $\mathbb{R}$, and let $Y = |X|$. We have $\mathcal{X}_1 = [0, \infty)$ and $\mathcal{X}_2 = (-\infty, 0)$. The Jacobian determinant degenerates to the derivative. For a given $y > 0$, the preimage consists of the elements y and $-y$. The derivative at these points exist: $g'(y) = 1$ and $g'(-y) = -1$. We hence obtain the PDF of Y as

$$f_Y(y) = f_X(y) + f_X(-y). \tag{2.7}$$

Indeed, we also have

$$\begin{aligned}f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \le y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(|X| \le y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(-y \le X \le y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}F_X(y) - \frac{\mathrm{d}}{\mathrm{d}y}F_X(-y) \\ &= f_X(y) - (-1)f_X(-y)\end{aligned}$$

which gives the same result.

2.2 The Differential Entropy of $Y = g(X)$

The differential entropy of an RV X with PDF f_X supported on $\mathcal{X}$ is given as

$$h(X) := -\int_{\mathcal{X}} f_X(x) \log f_X(x) \mathrm{d}x = -\mathbb{E}\left(\log f_X(X)\right) \tag{2.8}$$

provided that the (N-dimensional) integral exists (which we will assume throughout this chapter). We will later show that differential entropy has its justification[1] as the N-dimensional entropy of X [Ré59].

To calculate the differential entropy of $Y = g(X)$, assume first that g is bijective, i.e., $\{\mathcal{X}_i\} = \mathcal{X}$. Using Lemma 2.1 we thus obtain

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|\det \mathcal{J}_g(g^{-1}(y))|}. \tag{2.9}$$

Hence,

$$\begin{aligned} h(Y) &= -\mathbb{E}\left(\log f_Y(Y)\right) \\ &= -\mathbb{E}\left(\log f_Y(g(X))\right) \\ &= -\int_{\mathcal{X}} f_X(x) \log \frac{f_X(g^{-1}(g(x)))}{|\det \mathcal{J}_g(g^{-1}(g(x)))|} \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \log \frac{f_X(x)}{|\det \mathcal{J}_g(x)|} \mathrm{d}x \\ &= h(X) + \mathbb{E}\left(\log |\det \mathcal{J}_g(X)|\right). \end{aligned}$$

Example 3 **Linear Functions** Take $g(x) = ax$, where a is real and non-zero. It follows that $|g'(x)| = |a|$ for every $x \in \mathcal{X}$, and $\mathbb{E}\left(\log |a|\right) = \log |a|$. Hence, $h(aX) = h(X) + \log |a|$.

Now assume that g is a PBF. The PDF f_Y is a sum, and evaluating the differential entropy involves the logarithm of this sum, for which in general no closed-form solution exists. We therefore try to present a bound on $h(Y)$. It is easy to see that

[1] Edwin T. Jaynes expressed his dislike of differential entropy with the following words: "[...] the entropy of a continuous probability distribution is *not* an invariant. This is due to the historical accident that in his original papers, Shannon assumed, without calculating, that the analog of $\sum p_i \log p_i$ was $\int w \log w \mathrm{d}x$ [...] we have realized that mathematical deduction from the uniqueness theorem, instead of guesswork, yields [an] invariant information measure" [Jay63, p. 202]. At this time, Rényi's work had already been published, justifying differential entropy from a completely different point-of-view.

$x \in g^{-1}[g(x)]$, from which follows that

$$f_Y(g(x)) = \frac{f_X(x)}{|\det \mathcal{J}_g(x)|} + \sum_{x' \in g^{-1}[g(x)] \setminus \{x\}} \frac{f_X(x')}{|\det \mathcal{J}_g(x')|} \geq \frac{f_X(x)}{|\det \mathcal{J}_g(x)|}. \tag{2.10}$$

Using the monotonicity of the logarithm, we thus obtain [PP02, Eq. (14–113)]

$$\begin{aligned} h(Y) &= -\mathbb{E}\left(\log f_Y(g(X))\right) \leq -\mathbb{E}\left(\log \frac{f_X(X)}{|\det \mathcal{J}_g(X)|}\right) \\ &= h(X) + \mathbb{E}\left(\log |\det \mathcal{J}_g(X)|\right). \end{aligned} \tag{2.11}$$

2.3 Information Loss in PBFs

It is not clear if the *data processing inequality* (2.11) for continuous RVs is an appropriate measure for the non-negative number of bits that are lost in the system. The reason is, again, that $h(X)$ and $h(Y)$ are not invariant under a change of variables. The following results clarify the situation, stating that the difference between the right-hand side and the left-hand side of (2.11) is indeed a valid measure of information loss.

Definition 2.2 (*Information Loss in PBFs*) Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$, let g be a PBF, and let $Y = g(X)$. The information loss in g is

$$L(X \to Y) := H(X|Y) = \int_{\mathcal{Y}} H(X|Y = y) f_Y(y) \mathrm{d}y. \tag{2.12}$$

Note that this definition is meaningful, because $H(X|Y = y)$ is the entropy of an RV with an at most countable alphabet.

2.3.1 Elementary Properties

Intuitively, the information loss is due to the non-injectivity of g, which, employing Definition 2.1, is invertible if the set $\mathcal{X}_i$ from which the input X originated is already known. The following statements put this intuition on solid ground.

Definition 2.3 (*Partition Indicator*) The *partition indicator* W is a discrete RV that satisfies

$$W = i \text{ if } X \in \mathcal{X}_i \tag{2.13}$$

for every $i \in \{1, \ldots, \mathrm{card}(\mathcal{X})\}$. In other words, W is obtained by quantizing X according to the partition $\{\mathcal{X}_i\}$.

Proposition 2.1 ([GFK11, Theorem 2]) *Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$, let g be a PBF, and let $Y = g(X)$. The information loss is identical to the uncertainty about the set $\mathcal{X}_i$ from which the input was taken, i.e.,*

$$L(X \to Y) = H(W|Y). \tag{2.14}$$

Proof Since g is piecewise bijective, for every output $y \in \mathcal{Y}$, the conditional distribution $P_{X|Y=y}$ is discrete with $\mathrm{card}(g^{-1}[y])$ mass points. By piecewise bijectivity, every mass point lies in a different set $\mathcal{X}_i$, hence

$$P_{X|Y=y}(g^{-1}[y] \cap \mathcal{X}_i) = P_{X|Y=y}(\mathcal{X}_i) = P_{W|Y=y}(i) \tag{2.15}$$

Hence, $H(X|Y = y) = H(W|Y = y)$, for every $y \in \mathcal{Y}$. □

Thus, knowing the output value and the element of the partition from which the input originated, perfect reconstruction is possible:

Corollary 2.1 *System output Y and partition indicator W together are a sufficient statistic of the system input X, i.e.,*

$$H(X|Y, W) = 0. \tag{2.16}$$

Proof Since W is a function of X,

$$\begin{aligned} L(X \to Y) = H(X|Y) &= H(X, W|Y) \\ &= H(X|W, Y) + H(W|Y) = H(X|W, Y) + L(X \to Y) \end{aligned} \tag{2.17}$$

from which $H(X|Y, W) = 0$ follows. □

We are now ready to present:

Theorem 2.1 (Information Loss and Differential Entropy [GK12a, Corollary 1]) *Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$, let g be a PBF, and let $Y = g(X)$. The information loss in g is*

$$L(X \to Y) = h(X) - h(Y) + \mathbb{E}\left(\log|\det \mathcal{J}_g(X)|\right) \tag{2.18}$$

provided the quantities on the right-hand side exist.

Proof We start with a sketch of the proof that makes use of Dirac delta functions. A different proof, which requires some additional notation, is deferred to Sect. 3.3.

From Proposition 2.1 follows that $L(X \to Y) = H(W|Y)$. The latter can be computed as

$$H(W|Y) = \int_{\mathcal{Y}} H(W|Y=y)\mathrm{d}P_Y(y) = \int_{\mathcal{Y}} H(W|Y=y) f_Y(y)\mathrm{d}y. \tag{2.19}$$

We thus need to compute

$$p(i|y) := P_{W|Y=y}(i) = P_{X|Y=y}(\mathcal{X}_i). \tag{2.20}$$

For the sake of simplicity, we permit the Dirac delta function δ as a PDF for discrete probability measures. In particular, since Y is a function of X, we have $P_{Y|X=x}(A) = 1$ if and only if $g(x) \in A$; hence, we may write

$$f_{Y|X}(y|x) = \delta(y - g(x)) = \sum_{j=1}^{\mathrm{card}(\{\mathcal{X}_i\})} \frac{\delta(x - g_j^{-1}(y))}{|\det \mathcal{J}_g(g_j^{-1}(y))|}. \tag{2.21}$$

Using Bayes' theorem for densities, we get

$$\begin{aligned} p(i|y) &= \int_{\mathcal{X}_i} \mathrm{d}P_{X|Y=y}(x) = \int_{\mathcal{X}_i} f_{X|Y}(x|y)\mathrm{d}x && (2.22)\\ &= \int_{\mathcal{X}_i} \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}\mathrm{d}x && (2.23)\\ &= \frac{1}{f_Y(y)} \int_{\mathcal{X}_i} \sum_{j=1}^{\mathrm{card}(\{\mathcal{X}_i\})} \frac{\delta(x - g_j^{-1}(y))}{|\det \mathcal{J}_g(g_j^{-1}(y))|} f_X(x)\mathrm{d}x && (2.24)\\ &= \begin{cases} \frac{f_X(g_i^{-1}(y))}{|\det \mathcal{J}_g(g_i^{-1}(y))| f_Y(y)}, & y \in \mathcal{Y}_i = g(\mathcal{X}_i) \\ 0, & \text{else.} \end{cases} && (2.25) \end{aligned}$$

We can now write

$$\begin{aligned} H(W|Y) &= -\int_{\mathcal{Y}} \sum_{i=1}^{\mathrm{card}(\{\mathcal{X}_i\})} p(i|y)\log(p(i|y)) f_Y(y)\mathrm{d}y && (2.26)\\ &= -\sum_{i=1}^{\mathrm{card}(\{\mathcal{X}_i\})} \int_{\mathcal{Y}} p(i|y)\log(p(i|y)) f_Y(y)\mathrm{d}y && (2.27)\\ &\overset{(a)}{=} -\sum_{i=1}^{\mathrm{card}(\{\mathcal{X}_i\})} \int_{\mathcal{Y}_i} p(i|y)\log(p(i|y)) f_Y(y)\mathrm{d}y && (2.28) \end{aligned}$$

$$= -\sum_{i=1}^{\operatorname{card}(\{\mathcal{X}_i\})} \int_{\mathcal{Y}_i} \frac{f_X(g_i^{-1}(y))}{|\det \mathcal{J}_g(g_i^{-1}(y))|} \log\left(\frac{f_X(g_i^{-1}(y))}{|\det \mathcal{J}_g(g_i^{-1}(y))| f_Y(y)}\right) \mathrm{d}y \tag{2.29}$$

$$= -\sum_{i=1}^{\operatorname{card}(\{\mathcal{X}_i\})} \int_{\mathcal{X}_i} f_X(x) \log\left(\frac{f_X(x)}{|\det \mathcal{J}_g(x)| f_Y(g(x))}\right) \mathrm{d}x \tag{2.30}$$

$$= -\int_{\mathcal{X}} f_X(x) \log\left(\frac{f_X(x)}{|\det \mathcal{J}_g(x)| f_Y(g(x))}\right) \mathrm{d}x \tag{2.31}$$

$$= h(X) - h(Y) + \mathbb{E}\left(\log|\det \mathcal{J}_g(X)|\right) \tag{2.32}$$

were we exchanged sum and integral by Tonelli's theorem (the term $H(W|Y=y)$ is non-negative), and where in (a) we used the fact that $p(i|y) = 0$ if $y \notin \mathcal{Y}_i$ together with $0 \log 0 = 0$. □

Note that if g is bijective, i.e., invertible, (2.11) becomes an equality and the information loss vanishes: *bijective functions describe lossless systems*.

Example 4 **Full-Wave Rectifier** Consider the full-wave rectifier, i.e., we have $Y = |X|$. We have immediately that $|g'(x)| = 1$ for every $x \in \mathcal{X}$, hence $L(X \to Y) = h(X) - h(Y)$.

Example 5 **Square-Law Device and Gaussian Input [GK16, Sect. 5.4]** Let X be a zero-mean, unit variance Gaussian RV and let $Y = X^2$. Switching to nats, the differential entropy of X is $h(X) = \frac{1}{2}\ln(2\pi\mathrm{e})$. The output Y is χ^2-distributed with one degree of freedom and has differential entropy [VLR78]

$$h(Y) = \frac{1}{2}(1 + \ln\pi - \gamma) \tag{2.33}$$

where γ is the Euler-Mascheroni constant [AS72, pp. 3]. We moreover get

$$\mathbb{E}\left(\ln|g'(X)|\right) = \mathbb{E}\left(\ln|2X|\right) = \frac{1}{2}(\ln 2 - \gamma)\,. \tag{2.34}$$

Applying Theorem 2.1 and switching back to the binary logarithm yields $L(X \to Y) = 1$. Indeed, the information loss in a square-law device is always one bit if the PDF of the input RV has even symmetry [GFK11, Sect. V.A].

It can be shown that the information loss of a cascade is again the sum of the information lost in each constituting system.

Proposition 2.2 (Information Loss of a Cascade [GFK11, Theorem 3]) *Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$. Consider two functions $g\colon \mathcal{X} \to \mathcal{Y}$ and $f\colon \mathcal{Y} \to \mathcal{Z}$ and a cascade of systems implementing these functions. Let $Y = g(X)$ and $Z = f(Y)$. The information loss induced by this cascade, or equivalently, by the system implementing the composition $(f \circ g)(\cdot) = f(g(\cdot))$ is given by:*

$$L(X \to Z) = L(X \to Y) + L(Y \to Z) \tag{2.35}$$

Proof We apply Theorem 2.1 and the chain rule for Jacobian matrices

$$\mathcal{J}_{f\circ g}(x) = \mathcal{J}_f(g(x))\mathcal{J}_g(x)$$

to get

$$\begin{aligned}
&L(X \to Z)\\
&= h(X) - h(Z) + \mathbb{E}\left(\log|\det \mathcal{J}_{f\circ g}(X)|\right)\\
&= h(X) - h(Z) + \mathbb{E}\left(\log|\det \mathcal{J}_f(g(X))|\right) + \mathbb{E}\left(\log|\det \mathcal{J}_g(X)|\right)\\
&= L(X \to Y) + h(Y) - h(Z) + \mathbb{E}\left(\log|\det \mathcal{J}_f(g(X))|\right)\\
&= L(X \to Y) + L(Y \to Z).
\end{aligned}$$

□

While the information loss in most practically relevant PBFs will be a finite quantity, this does not always have to be the case.

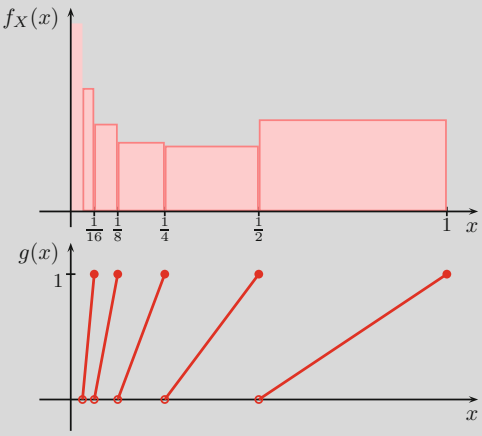
Example 6 **Infinite Loss in a PBF** Consider the scalar function $g\colon (0, 1] \to (0, 1]$ depicted below, mapping every interval $(2^{-n}, 2^{-n+1}]$ onto the interval $(0, 1]$:

$$g(x) = 2^n(x - 2^{-n}) \text{ if } x \in \mathcal{X}_n := (2^{-n}, 2^{-n+1}],\ n \in \mathbb{N} \tag{2.36}$$

The PDF of the input X is given as

$$f_X(x) = 2^n\left(\frac{1}{\log(n+1)} - \frac{1}{\log(n+2)}\right),\ \text{if } x \in (2^{-n}, 2^{-n+1}],\ n \in \mathbb{N}. \tag{2.37}$$

and also depicted below. It follows that the output RV Y is uniformly distributed on $(0, 1]$.



To apply Proposition 2.1, one needs

$$\mathbb{P}(W = n|Y = y) = \mathbb{P}(W = n) = \frac{1}{\log(n+1)} - \frac{1}{\log(n+2)}. \tag{2.38}$$

For this distribution, the entropy is known to be infinite [Bae08], and thus

$$L(X \to Y) = H(W|Y) = H(W) = \infty. \tag{2.39}$$

2.3.2 *Upper Bounds on the Information Loss*

In many cases one cannot directly evaluate the information loss according to Theorem 2.1, since the differential entropy of Y involves the logarithm of a sum. This section presents upper bounds on the information loss which are comparably easy to evaluate.

A particularly simple example for an upper bound–which is exact in Examples 5 and 6—is the following corollary to Proposition 2.1, which follows from the fact that conditioning reduces entropy.

Corollary 2.2 $L(X \to Y) \leq H(W)$.

More interesting is the following list of inequalities: All of these involve the cardinality of the preimage of the output. The further down one moves in this list, the simpler is the expression to evaluate; the last two bounds do not require any knowledge about the PDF of the input X. Nevertheless, the bounds are tight, as Examples 5 and 6 show.

Proposition 2.3 (Upper Bounds on Information Loss [GFK11, Theorem 4], [GK12a, Theorem 4]) *Let X be a continuous RV with PDF f_X supported on $\mathcal{X}$, let g be a PBF, and let $Y = g(X)$. The information loss can be bounded as follows:*

$$L(X \to Y) \leq \int_{\mathcal{Y}} f_Y(y) \log \operatorname{card}(g^{-1}[y]) \mathrm{d}y \tag{2.40}$$

$$\leq \log \mathbb{E}\left(\operatorname{card}(g^{-1}[Y])\right) = \log\left(\sum_i \int_{\mathcal{Y}_i} f_Y(y) \mathrm{d}y\right) \tag{2.41}$$

$$\leq \operatorname*{ess\,sup}_{y \in \mathcal{Y}} \log \operatorname{card}(g^{-1}[y]) \tag{2.42}$$

$$\leq \log \operatorname{card}(\{\mathcal{X}_i\}) \tag{2.43}$$

Proof We only sketch the proof here. The details can be found in the proof of [GFK11, Theorem 4]. The first inequality is due to the maximum entropy property of the uniform distribution, the second inequality is due to Jensen's inequality. The third inequality results from replacing the expected cardinality of the preimage by its essential supremum. This value can never exceed the cardinality of the partition $\{\mathcal{X}_i\}$, which gives the last bound. □

In the arXiv-version of [GFK11] we prove conditions under which these bounds hold with equality and illustrate them with intuitive examples.

Although the bounds of Proposition 2.3 are more elaborate than the one of Corollary 2.2, one can show that the latter is not necessarily useless.

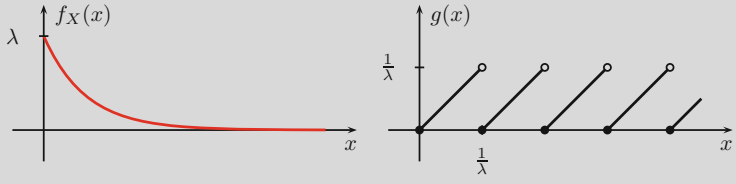
Example 7 **Exponential RV and Infinite Bounds** Consider an exponential input X with PDF

$$f_X(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases} \tag{2.44}$$

and the piecewise linear function

$$g(x) = x - \frac{\lfloor \lambda x \rfloor}{\lambda} \tag{2.45}$$

depicted below.



Obviously, $\mathcal{X} = [0, \infty)$ and $\mathcal{Y} = [0, \frac{1}{\lambda})$, while g partitions $\mathcal{X}$ into countably many intervals of length $\frac{1}{\lambda}$. In other words,

$$\mathcal{X}_i = \left[\frac{i-1}{\lambda}, \frac{i}{\lambda}\right) \tag{2.46}$$

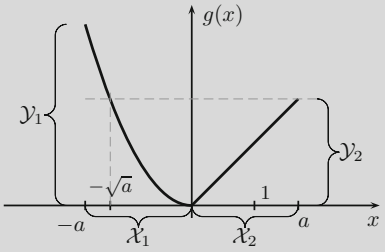
and $g(\mathcal{X}_i) = \mathcal{Y}$ for all $i = 1, 2, \ldots$. From this follows that for every $y \in \mathcal{Y}$ the preimage contains an element from each $\mathcal{X}_i$; thus, the bounds from Proposition 2.3 all evaluate to infinity. However, it can be shown that Corollary 2.2 is tight, i.e., $L(X \to Y) = H(W)$: With

$$P_X(\mathcal{X}_i) = \int_{\mathcal{X}_i} f_X(x)\mathrm{d}x = (1 - \mathrm{e}^{-1})\mathrm{e}^{-i+1} \tag{2.47}$$

one gets $H(W) = -\log(1 - \mathrm{e}^{-1}) + \frac{\mathrm{e}^{-1}}{1-\mathrm{e}^{-1}} \approx 1.24$. The same result is obtained by a direct evaluation of Theorem 2.1.

Example 8 **The Square-Linear Function [GFK11, Sect. V.B]** Consider an RV X uniformly distributed on $[-a, a]$, $a \geq 1$, and the function g depicted below. (© 2011 IEEE. Reprinted, with permission, from [GFK11].)

$$g(x) = \begin{cases} x^2, & \text{if } x < 0 \\ x, & \text{if } x \geq 0. \end{cases} \tag{2.48}$$



The information loss is

$$L(X \to Y) = \frac{4a + 4\sqrt{a} + 1}{8a} \log(2\sqrt{a} + 1) - \frac{\log(2\sqrt{a})}{2} - \frac{1}{4\sqrt{a} \ln 2} \tag{2.49}$$

where ln is the natural logarithm.

For $a = 1$, both sets $\mathcal{X}_1$ and $\mathcal{X}_2$ not only contain the same probability mass, but also map to the same image. Despite this fact, the information loss evaluates to $L(X \to Y) \approx 0.922$ bits. This suggests that by observing the output, part of the sign information can be retrieved. Looking at the picture, one can see that from $\mathcal{X}_1$ more probability mass is mapped to small output values than to large outputs. Thus, for a small output value y it is more likely that the input originated from $\mathcal{X}_1$ than from $\mathcal{X}_2$ (and vice-versa for large output values).

The bounds from Proposition 2.3 are not tight in this case, as they yield

$$L(X \to Y) \leq \frac{1 + \sqrt{a}}{2\sqrt{a}} \leq \log\left(\frac{3\sqrt{a} + 1}{2\sqrt{a}}\right) \leq 1 \leq 1 \tag{2.50}$$

which for $a = 1$ all evaluate to 1 bit.

2.3.3 Computing Information Loss Numerically

In this section, we briefly mention how to compute information loss numerically from sufficiently many realizations of X and $Y = g(X)$. We assume throughout that the function g is known, i.e., we know the partition $\{\mathcal{X}_i\}$. Note that this assumption is relatively unproblematic, since at least if X is one-dimensional, assuming that we *only* have vectors $\mathtt{x}$ and $\mathtt{y}$ corresponding to X and $g(X)$, respectively, we can "draw" the graph of g by plotting pairs $(\mathtt{x}, \mathtt{y})$ in a two-dimensional plane.

We thus assume that we can use the partition $\{\mathcal{X}_i\}$ to compute a vector $\mathtt{w}$ corresponding to W from Definition 2.3. Using the equivalence from Proposition 2.1, to compute the information loss it thus suffices to compute the conditional entropy of a *discrete* RV given a continuous observation, i.e., $H(W|Y)$. To further simplify the problem and make it accessible to numerical tools, we further quantize Y to $\hat{Y}$, converting the problem to that of computing the conditional entropy of two discrete RVs, i.e., $H(W|\hat{Y})$. The data processing inequality dictates that this quantity is an upper bound on the information loss $L(X \to Y)$, but estimation from numerical data adds further errors, cf. [SKD+02]. For PBFs with not too many branches (i.e., $\mathrm{card}(\{\mathcal{X}_i\})$ is small), this method gives reasonable approximations of information loss. Improvements are possible by *rank ordering* of $\mathtt{x}$ and $\mathtt{y}$ prior to quantization to reduce finite-sample effects [SKD+02], or to rely on more sophisticated numerical estimators for entropy and mutual information (see, e.g., [KSG04]).

The following code snippet for GNU Octave outlines computing the information loss in a full-wave rectifier, carried out for a Gaussian input RV X with unit variance and mean `m=1.5`. The missing part is a function computing the conditional entropy based on two vectors of realizations.

```
N=1e5;
m=1.5;
x=randn(1,N)+m;

w=x>0;
y=abs(x);

  %% Divide range of Y into sqrt(N) bins:
range=max(y)-min(y);
edges=min(y)-0.01*min(y):range/sqrt(length(x)):max(y)+0.01*max(y);

y_quant=zeros(size(y));
for ind=length(edges):-1:1
  y_quant(y<=edges(ind))=ind;
end

  %% Use pre-defined function for discrete RVs:
loss=condEntropy(w,y_quant);
```

2.3.4 Application: Polynomials

Many nonlinear systems consist of a memoryless nonlinearity in combination with a linear filter, e.g., Wiener and Hammerstein systems. The memoryless nonlinearity is often a polynomial (e.g., the square-law device in an energy detector) or can at least be approximated by one (cf. the Stone-Weierstrass theorem [Rud76, Theorem 7.26, p. 159]). We now show that our theory can be successfully applied to these systems. To this end, consider the third-order polynomial[2] depicted in Fig. 2.2, which is defined as

$$g(x) = x^3 - 100x. \tag{2.51}$$

Let X be Gaussian with zero mean and variance σ^2. We are not aware of closed-form expressions for either $h(Y)$ nor for $\mathbb{E}\left(\log|g'(X)|\right)$, hence we have to bound the information loss. To this end, note that some of the probability mass is mapped bijectively, i.e., there are $x \in \mathcal{X}$ such that $\operatorname{card}(g^{-1}[g(x)]) = 1$. We have

$$\mathcal{X}_b := \{x \in \mathcal{X} \colon \operatorname{card}(g^{-1}[g(x)])\} = \left(-\infty, -\frac{20}{\sqrt{3}}\right] \cup \left[\frac{20}{\sqrt{3}}, \infty\right). \tag{2.52}$$

Letting P_b denote the bijectively mapped probability mass, i.e., $P_b := P_X(\mathcal{X}_b) = P_Y(\mathcal{Y}_b)$, where $\mathcal{Y}_b = g(\mathcal{X}_b)$, we get $P_b = 2Q\left(\frac{20}{\sqrt{3}\sigma}\right)$, where Q denotes the Q-function [AS72, 26.2.3]. With this, the bounds from Proposition 2.3 evaluate to

[2] This example appeared in slightly different forms in [GFK11, Sect. V.C] and [GK16, Sect. 5.5].

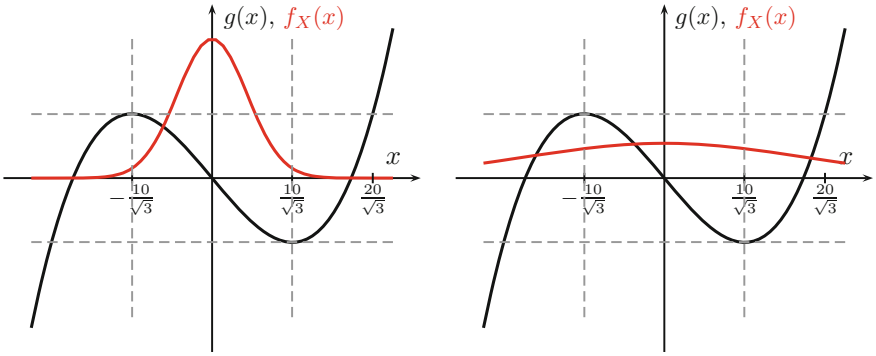


Fig. 2.2 Third-order polynomial, shown once with a Gaussian input with a small variance σ^2, and once with a large variance. *Vertical dashed lines* indicate the partition $\{\mathcal{X}_i\}$. For $g(x)$ above the *upper* and below the *lower horizontal dashed line*, we have $\text{card}(g^{-1}[g(x)]) = 1$

$$L(X \to Y) \leq (1 - P_b) \log 3 \leq \log (3 - 2P_b) \leq \log 3 \tag{2.53}$$

where we used the fact that $\text{ess sup}_{y \in \mathcal{Y}} \text{card}(g^{-1}[y]) = \text{card}(\{\mathcal{X}_i\}) = 3$.

Moreover, the three sets

$$\mathcal{X}_1 = \left(-\infty, -\frac{10}{\sqrt{3}}\right)$$
$$\mathcal{X}_2 = \left[-\frac{10}{\sqrt{3}}, \frac{10}{\sqrt{3}}\right]$$
$$\mathcal{X}_3 = \left(\frac{10}{\sqrt{3}}, \infty\right)$$

can be used to calculate to other upper bound $H(W)$ (cf. Corollary 2.2).

Figure 2.3 shows the information loss together with the derived upper bounds. It can be seen that the information loss is small if the input signal has small variance. This is quite intuitive, since most of the probability mass is concentrated in the interval $[-\frac{10}{\sqrt{3}}, \frac{10}{\sqrt{3}}]$, and the input can be reconstructed with high probability. After an increase in information loss, the loss decreases again, owing to the fact that a greater probability mass is mapped bijectively.

Problems

Problem 5 Let X_1 and X_2 be two independent, Gaussian RVs with zero mean and unit variance. The output Y is defined as $Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)^2$ and $Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2)^2$. Complete the following tasks:

- Determine the partition $\{\mathcal{X}_i\}$ for this PBF.
- Calculate the Jacobian matrix of g and its determinant.
- For a given y, compute $g^{-1}[y]$.

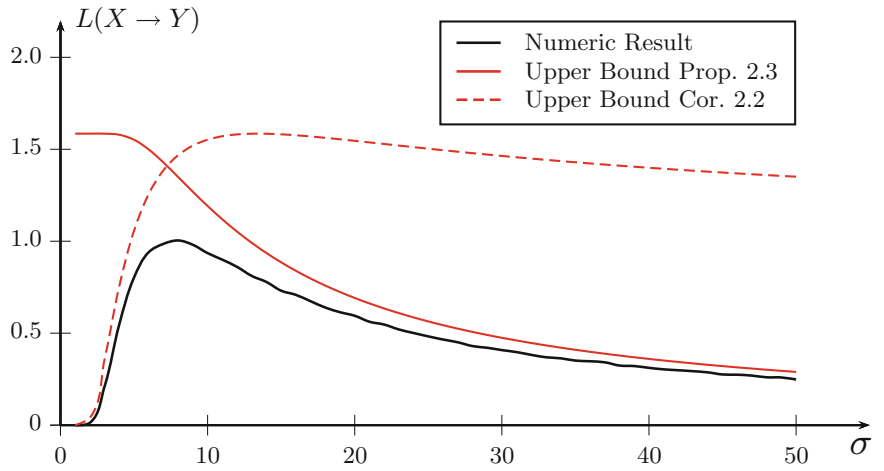


Fig. 2.3 Information loss for the third-order polynomial as a function of the input variance σ^2. The information loss, evaluated numerically using the scheme outlined in Sect. 2.3.3 (with rank ordering), is displayed together with the upper bounds

- Compute the PDF f_Y.
- Compute the preimage of the following points: $\{(y_1, y_2), (y_1 + \mathrm{d}y, y_2), (y_1, y_2 + \mathrm{d}y), (y_1 + \mathrm{d}y, y_2 + \mathrm{d}y)\}$. The area of the square defined by these four points is $\mathrm{d}y^2$. Draw the preimage of this square in a two-dimensional plane, for $(y_1, y_2) = (1, 4)$. Explain the connection between the area of the preimage and the Jacobian determinant.
- Compute the information loss $L(X \to Y)$.
- Suppose Z_1 and Z_2 are such that $Y_1 = Z_1^2$ and $Y_2 = Z_2^2$. It can be shown that X and Z are connected via a linear transform, i.e., $Z = \mathbf{T}X$. Compute $\mathbf{T}$ and $L(X \to Z)$.
- Let $Y_1 = X_1^2$ and $Y_2 = X_2^2$. Compute $L(X \to Y)$. What do you observe? Can, in this case, linear pre-processing reduce the information loss?

Problem 6 Suppose that X is a one-dimensional RV with even PDF, i.e., $f_X(x) = f_X(-x)$. Suppose further that g has even symmetry, and is bijective on $[0, \infty)$ and $(-\infty, 0]$. Show that the information loss satisfies $L(X \to Y) = 1$.

Problem 7 Show that for an RV X with even PDF, the sign and the magnitude are independent. **Hint:** Show that for all $a > 0$

$$\mathbb{P}\left(|X| < a | \mathrm{sgn}(X) = 1\right) = \mathbb{P}\left(|X| < a | \mathrm{sgn}(X) = -1\right).$$

Problem 8 Suppose X is uniformly distributed on $[-a + m, a + m]$ and let $g(x) = |x|$. Compute the information loss $L(X \to Y)$ as a function of the expected value m of X.

Problem 9 For the previous problem, implement a numerical estimator of $L(X \to Y)$ as described in Sect. 2.3.3. Plot the computed information loss as a function of m for $a = 1$ and $m \in [-1.5, 1.5]$ and compare it to the analytical results obtained in the previous problem.

Problem 10 Show that the order of two systems has an influence on the information loss of the cascade. **Hint:** Take a uniformly distributed input $X \sim \mathcal{U}([-a, a])$, and let the two systems be a rectifier $g(x) = |x|$ and an offset device $f(x) = x + m$ for some constant $m \neq 0$.

Problem 11 Suppose X is normally distributed with mean m and unit variance. Let again $g(x) = |x|$. Compute an upper bound on the information loss by

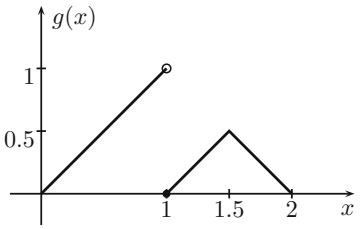
1. the entropy of the partition indicator, W
2. the expected cardinality of the preimage.

Problem 12 For the previous problem, implement a numerical estimator of $L(X \to Y)$ as described in Sect. 2.3.3. Plot the computed information loss as a function of m, for $m \in [-3, 3]$ and compare it to the bounds obtained in the previous problem.

Problem 13 For Example 8, verify the information loss for a general $a > 1$. Compute the information loss also for $0 < a < 1$.

Problem 14 For Example 8, verify the bounds from Proposition 2.3 for a general $a > 1$. Moreover, compute these bounds for $a < 1$.

Problem 15



Compute the information loss for the function depicted above and an input RV X with PDF

$$f_X(x) = \begin{cases} \frac{x}{2}, & \text{if } 0 \leq x \leq 2 \\ 0, & \text{elsewhere.} \end{cases} \tag{2.54}$$

Problem 16 For the function g of the previous problem and an input RV X uniformly distributed on $[0, 2]$, compute the information loss. Furthermore, compute the bounds of Proposition 2.3. What can you observe?

Problem 17 Consider Example 8. For every a, we have $H(W) = 1$. Note, however, that for $a > 1$, the probability mass in $\mathcal{X}_b = [-a, -\sqrt{a})$ is mapped bijectively to $\mathcal{Y}_b = (a, a^2]$. Using this knowledge, we can strengthen the bound from Corollary 2.2 to

$$\begin{aligned} H(W|Y) &\leq P_Y(\mathcal{Y}_b)H(W|Y \in \mathcal{Y}_b) + (1 - P_Y(\mathcal{Y}_b))H(W|Y \notin \mathcal{Y}_b) \\ &= P_b H(W|X \in \mathcal{X}_b) + (1 - P_b)H(W|X \notin \mathcal{X}_b) \\ &\leq H(W). \end{aligned}$$

Compute this bound for Example 8 and $a > 1$, and compare it to the bounds from Proposition 2.3.

Problem 18 Use the third-order polynomial from Sect. 2.3.4 to explain why, in general, $H(W|X \in \mathcal{X}_b) > 0$.

Problem 19 Let X have a probability measure $P_X \ll \lambda$ with PDF f_X and cumulative distribution function (CDF) F_X. Show that $Y = F_X(X)$ is uniformly distributed on $[0, 1]$. (This is also called histogram normalization or histogram equalization and is used, e.g., in image processing.)

http://www.springer.com/978-3-319-59532-0

Information Loss in Deterministic Signal Processing Systems

Geiger, B.C.; Kubin, G.

2018, XIII, 145 p. 16 illus., 9 illus. in color., Hardcover
ISBN: 978-3-319-59532-0