Graz University of Technology

PhD Thesis

Information Loss in Deterministic Systems

Connecting Information Theory, Systems Theory, and Signal Processing

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by
Dipl.-Ing. Bernhard C. Geiger

Supervisor:
Univ-.Prof. Dipl.-Ing. Dr. Gernot Kubin

Assessors/Examiners:
Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin
Prof. Dr.sc.techn. Gerhard Kramer, TU Munich

Graz, June 3, 2014

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRA-Zonline is identical to the present doctoral dissertation.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

date	(signature)

Contents

Abstract

A fundamental theorem in information theory – the data processing inequality – states that deterministic processing cannot increase the amount of information contained in a random variable or a stochastic process. The task of signal processing is to operate on the physical representation of information such that the intended user can access this information with little effort. In the light of the data processing inequality, this can be viewed as the task of *removing irrelevant information*, while preserving as much relevant information as possible.

This thesis defines information loss for memoryless systems processing random variables or stochastic processes, both with and without a notion of relevance. These definitions are the basis of an information-theoretic systems theory, which complements the currently prevailing energy-centered approaches. The results thus developed are used to analyze various systems in the signal processor's toolbox: polynomials, quantizers, rectifiers, linear filters with and without quantization effects, principal components analysis, multirate systems, etc. The analysis not only focuses on the information processing capabilities of these systems: It also highlights differences and similarities between design principles based on information-theoretic quantities and those based on energetic measures, such as the mean-squared error. It is shown that, at least in some cases, simple energetic design can be justified information-theoretically.

As a side result, this thesis presents two approaches to model complexity reduction for time-homogeneous, finite Markov chains. While one approach preserves full model information with the cost of losing the (first-order) Markov property, the other approach yields a Markov chain on a smaller state space with reduced model information. Finally, this thesis presents an information-theoretic characterization of *strong lumpability*, the case where the function of a Markov chain is Markov (of some order).

Acknowledgments

A lot of persons directly or indirectly influenced this thesis, and this is the perfect opportunity to say thank you.

First and foremost, I want to thank you, Gernot, for being the best PhD advisor I could wish for: During the many discussions we had, you always treated me like a peer, not like a student. You granted me the freedom to find (and pursue!) my own research ideas and to present my results as I liked – and you always had great research ideas when I was lacking them and you made sense of my results when I couldn't. Every time we met you astonished me with your immense knowledge in every field relevant for our joint work. I sincerely hope we can continue researching together some time in the future.

Thank you, Klaus, for teaching me how to write papers, and for supervising my Master's thesis. Without you, I would probably not have become a member of this lab. Thank you, Christian V., for teaching me how to write papers, for employing me as a study assistant in 2007, and for the beers. Thank you particularly for the many chats we had about (scientific) life, the universe, and everything. Thank you, Christian F., for your open door during the difficult first months of my thesis. I really appreciate the time you spent explaining things to me.

I want to thank all the SPSC lab members I met during these four-and-a-half years – you're a great bunch, I really had a lot of fun with you! Particularly, I want to thank the members of the 2-pm-coffee-party: Paul, my office "roomie", Shuli, Andi P., Erik, Kathi, Michi S., Shahzad, and Stefan M. Thank you for all the technical and, more importantly, non-technical chats!

Thank you, Christoph, Tanja, and Georg, for collaborating with me despite the distance that separated us: From you, I learned (again) writing papers, and how mathematicians, computer scientists, and information-theorists think. I hope we will stay in contact!

I also want to thank all the students that attended my classes: You were a great source of motivation for me. Most importantly, I want to thank my excellent Master's students Erik, Gebhard, and Christian K. It was a pleasure working with all of you!

Thank you, Mum, Berni, Peter, and Michi, for your endless support: I always knew you would never let me down if I needed your help. You always stood behind my decisions, and I am extremely grateful for that. Thank you, Berni, for tolerating our almost entirely disgusting meals at Mensa!

Finally, I want to thank you, Babsi, for being there for me the last ten years. Thank you for patiently listening to my mathematical stories of success and failure – they must have been infinitely boring for you! Thank you for sharing my joy when I succeeded, and for cheering me up when I failed. You all did this while you were struggling with your own thesis, and I really appreciate that. You are one of the strongest women I know, be proud of yourself. I love you and I look forward to our future together.

Preface

A typical PhD thesis is a highly specialized piece of work: Metaphorically speaking, the scientist has to climb a mountain (the scientific field), find a tree on this mountain (the research topic), and pick some tasty fruits (the contribution). This PhD thesis is different: Instead of a deep investigation of a narrow field, this thesis is broad and, consequently, rather shallow. Sticking to the metaphor of the tree, my journey during the last four years may be described as follows:

After an equally demanding and exhausting ascent halfway up the Mountain of Information Theory (I never hoped for getting higher), I saw a tree with small, sour fruits, that have not been touched by previous scientists on their way to the top: This tree represents information loss in deterministic functions. I climbed it, to harvest my first fruits, and from one of its branches I discovered a beautiful orchard in a valley, surrounded by the Mountains of Signal Processing, Machine Learning, Control Theory, and Mathematics. I ventured into this orchard and discovered lots of trees, hybrids of the one I just climbed and of seeds blown down from the surrounding mountains. And so I came to analyzing the effect of deterministic processing in various disciplines from an information-theoretic point-of-view. Admittedly, I grabbed mainly the low-hanging, exotic fruits, but there were many of them, and there are still many of them left for future scientists.

The topics considered in my thesis range from anti-aliasing filters, over lumpability of Markov chains, to Rényi's information dimension, the information bottleneck method, and principal components analysis. It was astonishing for me that *information* can play a fundamental role in so many disciplines, but in hindsight this is only natural: Nowadays information is one of the most important entities one has to consider in the design and the application of systems. While for communications, this importance was recognized more than 60 years ago, other areas are only recently getting aware of this concept. I can only hope that my thesis makes other scientists eager to wander through the same valley that I have wandered, and to pick the fruits for which I was simply too small.

1 Introduction

"Information is information,
not matter or energy."
– Norbert Wiener, "Cybernetics"

1.1 Motivation

Information is everywhere. All there is to know about the world is already in it, we just have to discover it. And signal processing helps us in discovering it, by processing the *signals*, the physical carriers of information, in such a way that we can access this information with little effort. The information in a string of zeros and ones is the same as in the digital image, or in the snippet of recorded speech transmitted from one cell phone to the other. But it is signal processing which makes this information usable to the person in front of the computer screen, or to the one holding the cell phone.

Signal processing is usually done by deterministic systems, taking an input signal and responding with an output signal. And the systems used today come in great variety: Taking the example of the cell phone, there is an antenna converting the electromagnetic wave to an electronic signal, a sampler and a quantizer converting this electronic signal into a string of zeros and ones, some digital filters removing noise and suppressing interference, a demodulator and a decoder, and again a system converting the resulting string of zeros and ones first again into an electronic signal, and then into acoustic waves travelling into our ear. Of course, this list is not exhaustive, but it suggests that there are lots of systems involved when it comes to preparing information for its sink, i.e., the human listener in this case. Thus, signal processing and systems theory are closely related disciplines.

But neither signal processing nor systems theory provide means to quantify the *information* contained in these signals; they provide quantitative measures for its physical representation, and for how the systems act on it, though. This physical representation is either energetic or material, but – and this is the link to the quote of Norbert Wiener at the beginning of this chapter – the representation is not identical to the represented information. Information theory *does* provide a measure of information – entropy – but is often not concerned with deterministic systems. Thus, there are two large, flourishing disciplines which have been largely developed independently from one another, but which pursue – should pursue – the same goal: Transmission and processing of information. Only recently, information theory embraced concepts from signal processing,

e.g., [43,67,155], and only recently signal processing based on information-theoretic cost functions has been gaining momentum, e.g., [8, 9, 35, 36].

That information theory is not concerned with deterministic systems at all is not entirely true: All textbooks on information theory include the well-known *data-processing inequality*, stating in one way or another that deterministic processing of random variables or stochastic processes cannot increase information, but decreases it or at best leaves it unmodified. In fact, this data-processing inequality for functions and/or deterministic systems is a direct consequence of Shannon's third axiom characterizing entropy [138]. Nevertheless, aside from this theorem, the question *how much* information is lost during deterministic processing has not been answered yet. Generally, despite recent advances mentioned above, the link between information theory, deterministic signal processing, and systems theory is weak. In fact, Johnson mentioned in [81] that "classic information theory is silent on how to use information theoretic measures (or if they can be used) to assess actual system performance". It is exactly the purpose of this work to make information theory speak up in this regard, for, as will be seen below, it has a lot to say.

Among the few published results about the information processing behavior of deterministic input-output systems are Pippenger's analysis of the information lost by multiplying two integer random variables [117] and the work of Watanabe and Abraham concerning the *rate* of information loss caused by feeding a discrete-time, finite-alphabet stationary stochastic process through a static, non-injective function [161]. All these works, however, focus only on finite-alphabet random variables and stochastic processes.

Slightly larger, but still focused on discrete random variables only, is the field concerning information-theoretic cost functions for the design of intelligent systems: The infomax principle [100], the information bottleneck method [145] using the Kullback-Leibler divergence as a distortion function, and system design by minimizing the error entropy (e.g., [120]) are just a few examples of this recent trend. Additionally, Lev's approach to aggregating accounting data [96], and, although not immediately evident, the work about macroscopic descriptions of multi-agent systems [93] belong to that category.

A systems theory for neural information processing has been proposed by Johnson in [81]. The assumptions made there (information need not be stochastic, the same information can be represented by different signals, information can be seen as a parameter of a probability distribution, etc.) suggest the Kullback-Leibler divergence as a central quantity. Although these assumptions are incompatible with the ones in this work, some similarities exist (e.g., the *information transfer ratio* of a cascade in [81] and Proposition 2.2).

Another, completely different connection between information theory and systems theory is in the field of autonomous dynamical systems or iterated maps. There, different measures for information flow within these systems have been proposed, e.g., [97, 103, 153, 164], most notably the definition of *transfer entropy* in [84, 135]. This connection clearly follows the spirit of Kolmogorov and Sinaï, who characterized dynamical systems exhibiting chaotic behavior with entropy [91, 92, 141], cf. [31].

1.2 Five Influential PhD Theses

During his work, the author this theses drew from a vast literature – but it is probably the following five PhD theses that had most influence. Therefore, this section gives a short overview of these works, and indicates which tools they provided or which research ideas they sparked. The discussion in Chapter 6 contains the counterpart: A list of contributions, and in which sense they complement, generalize, or extend these five PhD theses.

A very influential text was the PhD thesis of **Yihong Wu** [168, 169], submitted 2011 at Princeton University. He analyzed compressed sensing from the viewpoint of analog compres-

sion, i.e., of representing a set of real numbers by another, smaller set of real numbers, with restricted encoder and decoder (e.g., linear encoder, Lipschitz decoder). He showed that the measurement rate (or the ratio of the mentioned set cardinalities) is closely related to the Rényi information dimension [123] of the input random variables; thus, aside from discussing many of its properties, he gave information dimension an operational characterization. In this work, especially in Sections 2.4 and 3.4, information dimension is used to characterize the relative information loss (rate) in deterministic systems. Many results from [168, Chapter 2] served as a basis for the results developed in this work.

William S. Evans used information theory to analyze system design in his thesis [37, 38], submitted at the University of California in Berkeley in 1994. In particular, he analyzed the ratio of information transferred over a noisy circuit with binary input and m-ary output, $m \geq 2$. This "signal strength ratio" is large if the information at the input of the circuit is small; hence, the greatest information loss occurs at the beginning of a cascade of noisy circuits. These results connect to the measures of relative information loss proposed in Section 2.4, to the results about cascades of systems, and to this work's general approach of using information theory for system analysis and design. With his results, Evans gave lower bounds on the depth and size of electronic circuits to ensure reliable computation.

A connection between information theory and automatic control has been made by **Kun Deng**, who submitted his dissertation [26,27] at the University of Illinois at Urbana-Champaign in 2012. His topic was the aggregation of Markov chains, i.e., defining a Markov chain on a smaller state space which is close to the original Markov chain in terms of the Kullback-Leibler divergence rate. He showed that the problem of bi-partitioning the state space is solved by spectral theory, at least for nearly completely decomposable Markov chains. In addition to this spectral-based aggregation, he introduced a simulation-based aggregation, which requires only a single realized sequence of the Markov chain for aggregation. For hidden Markov models, he suggested to reduce the state space by keeping the Kullback-Leibler divergence rate between the observation processes small. In the present work, the Kullback-Leibler divergence rate is used for Markov chain aggregation too, although its computation is done differently; see Section 4.3. As a particular application of his theoretical results, Deng considered reducing the complexity of thermal models of buildings.

Connected also to Chapter 4 in this work, but to a different section and a completely different topic is the PhD thesis [131,132] of **Manuel Antonio Sánchez-Montañés**, submitted in 2003 at the Universidad Autónoma de Madrid. In his work, he proposed an effective information-processing measure trading model complexity for preservation of information relevant for a specific task. The motivation comes from neuro-biological systems, e.g., the auditory system, which are known not only to communicate, but to actually process information already at very early stages. This information-processing measure does not rule out non-Shannon information measures a priori; e.g., Bayes' error could very well take the place of Shannon entropy according to Sánchez-Montañés. His work, which is very similar to the information bottleneck method [145], is strongly connected to this work's Sections 4.2 and 4.4: Also here, the trade-off between preservation of relevant information and reduction of spurious information is an important concept; the author even believes that this trade-off is equivalent to the problem of signal enhancement or signal processing, not only for biological systems, but essentially for all systems which are designed to prepare an information-bearing signal for its sink. As applications, Sánchez-Montañés analyzed, e.g., linear systems with linear objectives (to which the principal components analysis is the optimal solution) and induction of decision trees.

Similar in concept is the thesis [72, 73] of **Gustav Eje Henter**, submitted to the KTH Royal Institute of Technology in 2013. He discussed the general problem of solving tasks (e.g., classification, synthesis, etc.) using data, and quantifying the performance using loss or cost functions. The solution of a task is in many cases the definition of a proper model for the data, be it the estimate of just a single parameter (e.g., the maximum likelihood estimate), or more complicated, generative or discriminative models. To get such a model, features have to

be extracted out of the data, which Henter adequately called *information removal*: Redundant information has to be removed, and the remaining information has to be transformed to allow efficient model identification – these are exactly the two steps required for signal enhancement, too, and Section 4.2 makes this explicit. Specifically, in [73] Henter used an information-theoretic formulation to reduce the complexity of stationary stochastic processes: Minimize the process' entropy rate while keeping the Kullback-Leibler divergence rate to the original process small. He presented closed-form solutions for Gaussian and Markov processes, and showed that the former are related to Wiener filtering. Henter applied his results to speech and language models.

1.3 Contributions and Outline

The contributions of this work are all motivated by Wiener's quote: Information is information, not matter or energy. Therefore, whenever we build a system to process information, the objective function shall capture the *information-processing capabilities* of the system. Since the systems we build are usually deterministic, information loss lends itself as an appropriate measure. But Wiener's quote also allows a second interpretation in the light of this work: More than once it will be shown that standard "energetic" system design, i.e., design based on energetic cost functions such as the mean-squared error, fails information-theoretically.

The larger part of **Chapter 2** provides general results on information loss, assuming that the input to the system is a multidimensional random variable. The system there is described either by a function which has a countable preimage for each output value, or by projections to lower dimensions. Both classes are important, and, although not exhaustive, allow an analysis of many practically important systems. As an example the principal components analysis (PCA) is considered. It is shown that PCA in general does not lead to a reduction of information loss, if it is, as usual, used prior to dimensionality reduction. This marks the first of many instances in this work where common design practices are overthrown because information-theoretic and energetic designs differ. Moreover, while Chapter 2 is closest to systems theory in its general treatment of systems, the section on PCA represents a first connection to machine learning.

Extending information loss from random variables to stationary stochastic processes is the focus of **Chapter 3**. Because this more general case is also more difficult to treat, the results presented there are focused on more specific scenarios than those of the previous chapter. In particular, and representing a connection to applied probability, finite-state Markov chains are characterized from an information-theoretic point-of-view. Not only equivalent conditions for information-preserving *lumpings*, i.e., state space reductions, are presented, but also an information-theoretic formulation of *strong lumpability*. One of the fruits mentioned in the preface has the flavor of natural language processing, since the presented theory is applied to state space reduction of a letter bi-gram model. After presenting some results for continuous-valued processes, multirate signal processing is considered: Section 3.5 proves that anti-aliasing filters cannot reduce the amount of information lost in the subsequent downsampling device. In analogy to PCA, this is another instance where an energetically optimal system fails to have superior performance in terms of information theory. The end of Chapter 3 contains a brief outlook to systems with memory, showing, quite counterintuively, that digital filter implementations with round-off errors need not lose information.

The large discrepancy between energy and information, particularly for the PCA and anti-aliasing filtering example, is resolved by bringing the notion of *relevance* into the game in **Chapter 4**: Considering again random variables only, not all information in a data vector is relevant. By focusing only on the relevant information in the data vector, a new information-processing measure is suggested. And indeed, this measure is shown to best correspond to the signal processor's task of *signal enhancement*, i.e., of preparing the signal such that the sink can

retrieve the relevant information with least effort. The connection to machine learning is made by showing that the information bottleneck method exactly minimizes *relevant information loss*. As a consequence, PCA is better understood by showing that, under specific signal model assumptions, it minimizes the relevant information loss in the following dimensionality reduction. Moreover, relevant information loss is also shown to be an adequate cost function for state space reduction of Markov chains, where this time the focus is not on preserving all information, but on obtaining a good first-order Markov model on a smaller state space. Possible applications for this type of state space reduction lie in the field of automatic control.

Finally, **Chapter 5** extends results of the previous chapter from random variables to stochastic processes. For this most general case, only few specific results could be obtained. And again, the connection to signal processing is strong: By introducing a specific signal model of data superimposed by a Gaussian noise process, anti-aliasing filters are justified information-theoretically, resolving the counterintuivity from Section 3.5. As a second example, filter design prior to quantization is analyzed, essentially justifying linear prediction under specific assumptions. That this chapter discusses mainly filter design is no accident: Especially filter design, an elementary tool of the signal processor, is often based on energetic considerations. It is, therefore, of prime importance to know when, and when not, this energetic design coincides with the information-theoretic optimum.

Naturally, this thesis is by no means complete. There are many open questions, the most immediate of which are mentioned at the end of each chapter. **Chapter 6** finally points at larger areas left uncovered, and suggests the most fruitful – or most important – directions for future research.

2 Information Loss

The results of this chapter have almost exclusively been obtained by the author, owing to fruitful discussions with Christian Feldbauer and Gernot Kubin. The results about piecewise bijective functions in Section 2.3 have been partly published in [49] and [51]. Relative information loss (Section 2.4) and its application to principal components analysis (Section 2.6) was treated in [52]. Furthermore, parts of this chapter constitute [46].

The result that the sub-optimal reconstructor coincides with the maximum a-posteriori reconstructor in the example in Section 2.3.4 is due to the author's student Stefan Wakolbinger. The significance of Theorem 2.2 in the light of chaotic systems was investigated by the author's student Gebhard Wallinger.

2.1 What Is Information Loss? – The Discrete Case

What is information loss? In order to analyze the central quantity of this work one needs to define it properly. And in defining a quantity one typically has two options: The first is to present a definition for the general case and then apply it to each special case. The second, more instructive option – the one taken up in this work – is to define the quantity for a special case and continue with generalizing it to larger and larger classes, making sure that the generalizations are consistent.

Intuitively, the information loss in a deterministic system is the same as the water loss in a system of (possibly corroded) pipes[1]: Water that flows into the system at the well may either leave it at the faucet (the desired output), or through holes or leaking adaptors. The water loss – for the person trying to wash its hands – is just the difference between the amount of water leaving the well and the amount of water flowing out of the faucet. In analogy, the information loss in a deterministic input-output system may be defined as the difference between the information at its input and the information at its output.

To make this precise, let $(\Omega, \mathfrak{A}, \Pr)$ be a probability space and let $X\colon \Omega \to \mathcal{X}$ be a discrete random variable, taking values from a finite set $\mathcal{X} \subset \mathbb{N}$. It induces a new probability space $(\mathcal{X}, \mathfrak{P}(\mathcal{X}), P_X)$, where $\mathfrak{P}(\mathcal{X})$ is the power set of $\mathcal{X}$ and where

$$\forall B \in \mathfrak{P}(\mathcal{X})\colon P_X(B) = \Pr(X^{-1}[B]). \tag{2.1}$$

[1] This analogy is not new; in his thesis [37], Evans writes: "[For functions,] Pippenger first showed that the total information sent is bounded by the sum of the information sent over each path [...] This supports the view of information as a kind of fluid which flows from the input X to the output [...] At each gate, several paths combine, but the fluid flowing out of the gate is no more than the sum of the fluid flowing in."

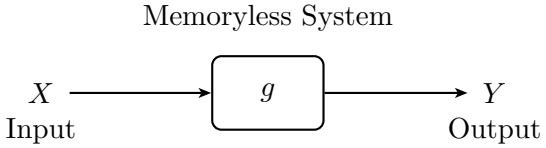


Figure 2.1: A model for information loss. The system is memoryless and time-invariant, and described by the system function g. Both input X and output Y are discrete RVs.

Here, $X^{-1}[B] = \{\omega \in \Omega\colon X(\omega) \in B\}$ denotes the preimage[2] of B under X. Abusing notation, write $P_X(x)$ for the probability measure of a single point instead of $P_X(\{x\})$. Since $\mathcal{X}$ is finite, one can define the probability mass function (PMF) p_X as

$$\forall x \in \mathcal{X}\colon \quad p_X(x) := P_X(x) = \Pr(X = x), \tag{2.2}$$

abusing notation by writing $\Pr(X = x)$ instead of $\Pr(X^{-1}[x])$.

In this work, the information contained in an RV shall be measured by its Shannon entropy [21, p. 14],

$$H(X) := -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \tag{2.3}$$

where log denotes the binary logarithm in this work.

Let now the RV X describe the input to a deterministic system (see Fig. 2.1). Let further $g\colon \mathcal{X} \to \mathcal{Y}$, $\mathcal{Y} \subset \mathbb{N}$ be a surjective, measurable (w.r.t. the power sets of $\mathcal{X}$ and $\mathcal{Y}$) function describing the system behavior. Thus, the output of the system is the RV $Y := g(X)$, which induces a probability measure P_Y with PMF p_Y. The information contained in Y is of course its entropy, $H(Y)$.

This mathematical basis allows the definition of information loss for discrete RVs as the difference between the entropies of the system input and system output, respectively:

Definition 2.1 (Information Loss for Discrete RVs)**.** Let X be an RV with finite alphabet $\mathcal{X}$, and let $Y := g(X)$. The information loss induced by g is

$$L(X \to Y) := H(X) - H(Y) = H(X|Y). \tag{2.4}$$

As this definition shows, information loss is the conditional entropy of the input given the output. This follows simply from the chain rule of entropy [21, Thm. 2.2.1] and the fact that $H(X, g(X)) = H(X)$ [21, Problem 2.4].

The following result is taken from [7]: It justifies Definition 2.1 by showing that it is the only measure satisfying a set of desirable axioms.

Theorem 2.1 (Axiomatization of Information Loss, [7, Thm. 2])**.** *Suppose $F(X, g)$ a measure of information loss satisfying the following three axioms:*

1. ***Functorality:*** *Let $g\colon \mathcal{X} \to \mathcal{Y}$ and $h\colon \mathcal{Y} \to \mathcal{Z}$. Then,*

$$F(X, (g \circ h)) = F(X, g) + F(g(X), h). \tag{2.5}$$

2. ***Convex Linearity:*** *Let a coin with probability p decide whether $g_1\colon \mathcal{X} \to \mathcal{Y}$ or $g_2\colon \mathcal{X} \to \mathcal{Y}$*

[2] Throughout this work, the preimage under a function is indicated with square brackets rather than parenthesis to avoid confusion with the inverse function $g^{-1}(\cdot)$ (if it exists). Moreover, for singletons the curly brackets will be omitted, i.e., $g^{-1}[y] := g^{-1}[\{y\}]$.

describes the system. Then,

$$F(X, pg_1 \oplus (1-p)g_2) = pF(X, g_1) + (1-p)F(X, g_2). \tag{2.6}$$

3. ***Continuity:*** *Let (X_n) be a sequence of RVs with alphabets $(\mathcal{X}_n)$ and PMFs (p_{X_n}) and let $(g_n\colon \mathcal{X}_n \to \mathcal{Y}_n)$ be a sequence of system functions. Let for sufficiently large n, $\mathcal{X}_n = \mathcal{X}$, $\mathcal{Y}_n = \mathcal{Y}$, and $g_n(x) = g(x)$ for all $x \in \mathcal{X}$. Let further $p_{X_n} \to p_X$ and $p_{Y_n} \to p_Y$ (pointwise). Then,*

$$F(X_n, g_n) \to F(X, g). \tag{2.7}$$

Then, there exists a constant $c \geq 0$ such that $F(X, g) = cL(X \to g(X))$.

The property of functorality will be dealt with later; it will be shown that it not only holds for discrete RVs with finite alphabets, but for all scenarios investigated in this work. The other two properties, while interesting, are not generalized here. Particularly, one cannot expect that continuity still holds for countable alphabets, due to the discontinuity of entropy [75].

2.2 Generalization to Continuous Random Variables

Let X still be an RV, and let $Y := g(X)$. Thus, still $H(Y|X) = 0$. Assume now, however, that X is not discrete, but an N-dimensional RV taking values from $\mathcal{X} \subseteq \mathbb{R}^N$. Its probability measure, P_X, need not be supported on a countable set, but may, generally, decompose[3] into a discrete, atomic component P_X^d, a singularly continuous component P_X^{sc}, and a component P_X^{ac} absolutely continuous w.r.t. the N-dimensional Lebesgue measure λ^N. If P_X consists only of the latter, which is denoted by $P_X \ll \lambda^N$, it possesses a probability density function (PDF) f_X, the Radon-Nikodym derivative of P_X w.r.t. λ^N.

Assume that $g\colon \mathcal{X} \to \mathcal{Y}$ is measurable w.r.t. the Borel-algebras $\mathfrak{B}_\mathcal{X}$ and $\mathfrak{B}_\mathcal{Y}$ of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, the probability distribution of Y is

$$\forall B \in \mathfrak{B}_\mathcal{Y}\colon \quad P_Y(B) = P_X(g^{-1}[B]). \tag{2.9}$$

As soon as P_X has a non-atomic component, it follows that $H(X) = \infty$, and the same holds for $H(Y)$. Computing the information loss as the difference between the entropy of the input and the entropy of the output fails. As a remedy, the following approach is proposed: Quantize X by partitioning its alphabet $\mathcal{X}$ uniformly; this defines

$$\hat{X}^{(n)} := \frac{\lfloor 2^n X \rfloor}{2^n} \tag{2.10}$$

where the floor operation is taken element-wise. The elements $\{\hat{\mathcal{X}}_k^{(n)}\}$, $k \in \mathbb{Z}$, of the induced partition $\mathcal{P}_n$ of $\mathcal{X}$ are N-dimensional hypercubes[4] of side length $\frac{1}{2^n}$. Obviously, the partitions refine with increasing n, i.e., $\mathcal{P}_n \succ \mathcal{P}_{n+1}$.

[3] To be specific, according to the Lebesgue decomposition theorem [128, pp. 121] every (probability) measure can be decomposed into

$$P_X = P_X^{ac} + P_X^{sc} + P_X^d \tag{2.8}$$

with $P_X^{ac}(\mathcal{X}) + P_X^{sc}(\mathcal{X}) + P_X^d(\mathcal{X}) = 1$. Here, $P_X^{ac} \ll \lambda^N$ while the other two measures are singular to λ^N. P_X^d is a discrete probability measure, i.e., it consists of point masses, while P_X^{sc} is called singular continuous (single points have zero probability, but the probability mass is concentrated on a Lebesgue null set). In the one-dimensional case, P_X^{sc} would account for, e.g., fractal probability measures such as, e.g., the Cantor distribution; in higher dimensions P_X^{sc} also accounts for probability masses concentrated on smooth submanifolds of $\mathbb{R}^N$ of lower dimensionality.

[4] Specifically, for $N = 1$ the element $\hat{\mathcal{X}}_k^{(n)}$ corresponds to the interval $[\frac{k}{2^n}, \frac{k+1}{2^n})$.

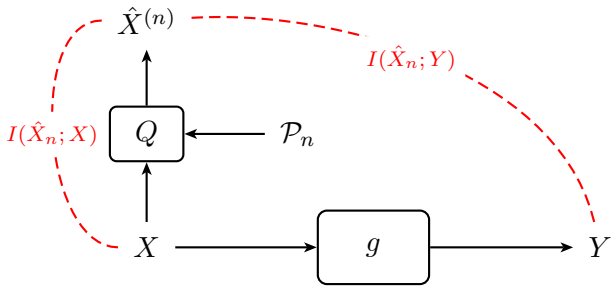


Figure 2.2: Model for computing the information loss of a memoryless input-output system g. Q is a quantizer with partition $\mathcal{P}_n$. The input X is not necessarily discrete.

One can now measure the mutual information between X and its quantization $\hat{X}^{(n)}$, as well as the mutual information between the system output Y and $\hat{X}^{(n)}$ (see Fig. 2.2). While the former is an *approximation* of the information available at the system input, the latter approximates the information at the system output. By the data processing inequality [21, p. 35], the former quantity cannot be smaller than the latter. Moreover, the finer the quantization, the better are these approximations. This leads to

Definition 2.2 (Information Loss)**.** Let X be an RV with alphabet $\mathcal{X}$, and let $Y := g(X)$. The information loss induced by g is

$$L(X \to Y) := \lim_{n\to\infty} \left(I(\hat{X}^{(n)};X) - I(\hat{X}^{(n)};Y) \right) = H(X|Y). \tag{2.11}$$

Before addressing the second equality, note that for a bijective system the two mutual informations are equal for all n, and that thus the information loss evaluates to zero: *bijective functions describe lossless systems.*

For the second equality, note that with [66, Lem. 7.20]

$$\begin{aligned} I(\hat{X}^{(n)};X) - I(\hat{X}^{(n)};Y) &= H(\hat{X}^{(n)}) - H(\hat{X}^{(n)}|X) - H(\hat{X}^{(n)}) + H(\hat{X}^{(n)}|Y) && (2.12)\\ &= H(\hat{X}^{(n)}|Y) && (2.13) \end{aligned}$$

since $\hat{X}_n$ is a function of X. Since furthermore $H(\hat{X}_n|Y=y) \nearrow H(X|Y=y)$ monotonically [66, Lem. 7.18], the second equality follows.

For a discrete input RV X one obtains $L(X \to Y) = H(X) - H(Y)$. Thus, the extension to general RVs does not conflict with Definition 2.1 and remains justified by [7]. Clearly, for a discrete input RV (with finite entropy) the information loss will always be a finite quantity. The same does not hold for a continuous input X:

Proposition 2.1 (Infinite Information Loss)**.** *Let P_X have an absolutely continuous component $P_X^{ac} \ll \lambda^N$ which is supported on $\mathcal{X}$. If there exists a set $B \subseteq \mathcal{Y}$ of positive P_Y-measure such that the preimage $g^{-1}[y]$ is uncountable for every $y \in B$, then*

$$L(X \to Y) = \infty. \tag{2.14}$$

Proof.

$$L(X \to Y) = H(X|Y) = \int_{\mathcal{Y}} H(X|Y=y) dP_Y(y) \geq \int_B H(X|Y=y) dP_Y(y) \tag{2.15}$$

since $B \subseteq \mathcal{Y}$. Since, for all $y \in B$, $g^{-1}[y]$ is uncountable and since P_X has an absolutely continuous component on $\mathcal{X}$, the conditional probability measure $P_{X|Y=y}$ cannot be supported on a countable set of points; thus one obtains $H(X|Y=y) = \infty$ [116, Ch. 2.4] for all $y \in B$. The proof follows from $P_Y(B) > 0$. □

Now let $P_X \ll \lambda^N$ and take $y^* \in \mathcal{Y}$. Since $P_Y(y^*) = P_X(g^{-1}[y^*])$, $P_Y(y^*) > 0$ is only possible if $g^{-1}[y^*]$ is uncountable (it cannot be a λ^N null set). This proves

Corollary 2.1. *Let $P_X \ll \lambda^N$. If there exists a point $y^* \in \mathcal{Y}$ such that $P_Y(y^*) > 0$, then $L(X \to Y) = \infty$.*

Example 1: Quantizer.

Look at the information loss of a scalar quantizer, i.e., of a system described by a function

$$g(x) = \lfloor x \rfloor. \tag{2.16}$$

With the notation introduced above, $Y = g(X) = \hat{X}^{(0)}$. Assuming $P_X \ll \lambda$, which is fulfilled by every univariate distribution described by a PDF, there will be at least one point y^* for which $\Pr(Y = y^*) = P_Y(y^*) = P_X([y^*, y^* + 1)) > 0$. The conditions of Corollary 2.1 are fulfilled and one obtains

$$L(X \to \hat{X}^{(0)}) = \infty. \tag{2.17}$$

Note that due to the continuity of the input distribution, $H(X) = \infty$, while in all practically relevant cases $H(Y) < \infty$. Thus, in this case the information loss truly corresponds to the difference between input and output entropies.

There obviously exists a class of systems with finite information transfer ($I(X;Y) < \infty$) and infinite information loss. Contrarily, Section 2.3 treats systems with finite information loss for which $I(X;Y) = \infty$; finally, Section 2.4 analyzes systems for which both quantities are infinite. As the next example shows, not every function with an uncountable preimage leads to infinite information loss:

Example 2: An Almost Invertible Transform.

Consider a two-dimensional RV X which places probability mass uniformly on the unit disc, i.e.,

$$f_X(x) = \begin{cases} \frac{1}{\pi}, & \text{if } ||x|| \leq 1 \\ 0, & \text{else} \end{cases} \tag{2.18}$$

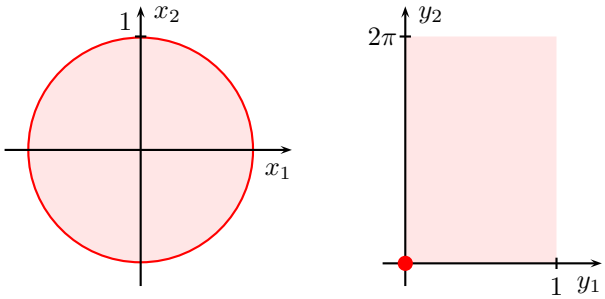
where $||\cdot||$ is the Euclidean norm. Thus, $\mathcal{X} = \{x \in \mathbb{R}^2 : ||x|| \leq 1\}$. The Cartesian coordinates x are now transformed to polar coordinates in a special way, namely:

$$y_1 = \begin{cases} ||x||, & \text{if } ||x|| < 1 \\ 0, & \text{else} \end{cases} \tag{2.19}$$

$$y_2 = \begin{cases} \arctan(\frac{x_2}{x_1}) + \frac{\pi}{2}(2 - \operatorname{sgn}(x_2) - \operatorname{sgn}(x_1 x_2)), & \text{if } 0 < ||x|| < 1 \\ 0, & \text{else} \end{cases} \tag{2.20}$$

where x_1 and x_2 are the first and second coordinate of x.

The mapping together with the domains of X and Y is illustrated below: The solid red circle in the left diagram and the red dot in the right diagram correspond to each other, illustrating the mapping of an uncountable P_X-null set to a point. The lightly shaded areas are mapped bijectively.



As a direct consequence one has $\mathcal{Y} = (0,1) \times [0,2\pi) \cup \{(0,0)\}$. Observe that not only the point $x = \{(0,0)\}$ is mapped to the point $y = \{(0,0)\}$, but that also the unit circle $\mathcal{S} = \{x : ||x|| = 1\}$ is mapped to $y = \{(0,0)\}$. As a consequence, the preimage of $\{(0,0)\}$ under g is uncountable. However, since a circle in $\mathbb{R}^2$ is a Lebesgue null-set and thus $P_X(\mathcal{S}) = 0$, also $P_Y(\{(0,0)\}) = 0$ and the conditions of Proposition 2.1 are not met. Indeed, since $H(X|Y = y) = 0$ P_Y-a.s., it can be shown that $L(X \to Y) = 0$.

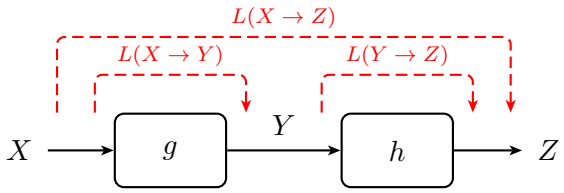


Figure 2.3: Cascade of systems: The information loss of the cascade equals the sum of the individual information losses of the constituent systems.

The following proposition about the cascade of systems (see Fig. 2.3) generalizes the property of functorality from Theorem 2.1 to continuous RVs:

Proposition 2.2 (Information Loss of a Cascade)**.** *Consider two functions g: $\mathcal{X} \to \mathcal{Y}$ and h: $\mathcal{Y} \to \mathcal{Z}$ and a cascade of systems implementing these functions. Let $Y := g(X)$ and $Z := h(Y)$. The information loss induced by this cascade, or equivalently, by the system implementing the composition $(h \circ g)(\cdot) = h(g(\cdot))$ is given by:*

$$L(X \to Z) = L(X \to Y) + L(Y \to Z) \tag{2.21}$$

Proof. Referring to Definition 2.2 and [116, Ch. 3.9],

$$L(X \to Z) = H(X|Z) = H(Y|Z) + H(X|Y,Z) = H(Y|Z) + H(X|Y) \tag{2.22}$$

since Y and (Y,Z) are mutually subordinate. □

Naturally, by induction this result can be generalized to an arbitrary number of systems, which is not only meaningful when all systems have finite information loss: As soon as in a system an infinite amount of information is lost, the full cascade also loses an infinite amount of information. Intuitively, the consequence of this proposition is that information already lost in a system cannot be recovered by subsequent systems.

Finally, Proposition 2.2 has a beautiful correspondence to the theory of linear systems: The (logarithmic) transfer function of a cascade of linear filters equals the product (sum) of the individual (logarithmic) transfer functions. The essential difference is that the order of stable linear filters has no influence on the frequency response of their cascade, whereas the same is not true for the information loss of nonlinear systems. There, even a simple scalar multiplication or addition can change the information loss occurring in the subsequent system. In that sense,

nonlinear systems do not necessarily commute w.r.t. information loss, while stable linear systems do w.r.t., e.g., the frequency response. Consequently, while post-processing cannot recover information already lost, pre-processing can prevent it from *getting lost*, cf. [81]. A significant part of this work deals with designing such pre-processing systems for this very purpose.

2.3 Information Loss in Piecewise Bijective Functions

As Example 1 showed, there certainly exist functions for which the information loss is infinite; for these, our measure of information loss will be of little help, and other, yet to be developed measures shall be applied. However, as we show in this section, quite a large class of systems loses only a finite amount of information. As a prime example, take the recitfier: Stripping the sign off a number cannot destroy more than 1 bit of information.

Throughout this section, let the probability distribution of the system input X satisfy $P_X \ll \lambda^N$ for some N, and let $\{\mathcal{X}_i\}$ be a partition of $\mathcal{X}$ into non-null sets, i.e., for all i, $P_X(\mathcal{X}_i) > 0$. The class of systems considered in this section is described by

Definition 2.3 (Piecewise Bijective Function)**.** 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:

$$g(x) = \begin{cases} g_1(x), & \text{if } x \in \mathcal{X}_1 \\ g_2(x), & \text{if } x \in \mathcal{X}_2 \\ \vdots \end{cases} \tag{2.23}$$

where $g_i\colon \mathcal{X}_i \to \mathcal{Y}_i$ bijectively. The Jacobian matrix $\mathcal{J}_g(\cdot)$ exists on the closures of $\mathcal{X}_i$, and its determinant, $\det\mathcal{J}_g(\cdot)$, is non-zero P_X-a.s.

2.3.1 Elementary Properties

A direct consequence of Definition 2.3 is that the preimage $g^{-1}[y]$ is countable for all $y \in \mathcal{Y}$ and that the probability measure of the output Y satisfies $P_Y \ll \lambda^N$ and has a PDF

$$f_Y(y) = \sum_{x_i \in g^{-1}[y]} \frac{f_X(x_i)}{|\det\mathcal{J}_g(x_i)|} \tag{2.24}$$

by the method of transformation [114, p. 244].

One of the main results of this work is the following, connecting information loss with differential entropies. The differential entropy of an RV X with PDF f_X is given as

$$h(X) := -\int_{\mathcal{X}} f_X(x) \log f_X(x) dx \tag{2.25}$$

provided that the (N-dimensional) integral exists. It equals the N-dimensional entropy of X [123].

Theorem 2.2 (Information Loss and Differential Entropy)**.** *Let $P_X \ll \lambda^N$ and let g be a PBF. The information loss induced by g is*

$$L(X \to Y) = h(X) - h(Y) + \mathbb{E}\left(\log|\det\mathcal{J}_g(X)|\right) \tag{2.26}$$

provided the quantities on the right exist.

Proof. See Appendix A.1. □

This result is interesting because it connects a quantity which is invariant under a change of variables (the information loss is defined via mutual information, which exhibits this invariance)

and a quantity which is not: differential entropy changes under a bijective variable transform and can thus not be regarded as a measure of information[5]. As such, the difference $h(X)-h(Y)$ is meaningless: Changing g to $c \cdot g$ with c being a real number changes the aforementioned difference. The third term, the *expected logarithmic differential gain* $\mathbb{E}\left(\log\left|\det\mathcal{J}_g(X)\right|\right)$, however, compensates for this variation – the effect from shaping the PDF is mitigated, what remains is a measure of information loss.

From a different point-of-view, Theorem 2.2 extends [114, pp. 660], which claims that the differential entropy of a function g of an RV X is bounded by

$$h(Y) \leq h(X) + \mathbb{E}\left(\log\left|\det\mathcal{J}_g(X)\right|\right) \tag{2.27}$$

where equality holds if and only if g is bijective: Theorem 2.2 explains the difference between the right-hand side and the left-hand side of (2.27) as the information lost due to data processing for functions which are only piecewise bijective.

It is also worth noting that Theorem 2.2 has a tight connection to the theory of iterated function systems. In particular, [164] analyzed the *information flow* in one-dimensional maps, which is the difference between information generation via stretching (corresponding to the term involving the Jacobian determinant) and information reduction via folding (corresponding to information loss). Ruelle [129] later proved that for a restricted class of systems the *folding entropy* $L(X \to Y)$ cannot fall below the information generated via stretching, and therefore speaks of *positivity of entropy production*. He also established a connection to the Kolmogorov-Sinaï entropy rate. In [82], both components constituting information flow in iterated function systems are described as ways a dynamical system can lose information. This highly interesting connection was recently investigated in a Master's thesis for one-dimensional, discrete-time chaotic systems [160].

Example 3: Square-Law Device and Gaussian Input.

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 a χ^2-distributed RV with one degree of freedom, for which the differential entropy can be computed as [156]

$$h(Y) = \frac{1}{2}\left(1 + \ln\pi - \gamma\right) \tag{2.28}$$

where γ is the Euler-Mascheroni constant [3, pp. 3]. The Jacobian determinant degenerates to the derivative, and using some calculus yields

$$\mathbb{E}\left(\ln\left|g'(X)\right|\right) = \mathbb{E}\left(\ln\left|2X\right|\right) = \frac{1}{2}\left(\ln 2 - \gamma\right). \tag{2.29}$$

With Theorem 2.2 it follows that $L(X \to Y) = \ln 2$, which after changing the base of the logarithm amounts to one bit. Indeed, the information loss induced by a square-law device is always one bit if the PDF of the input RV has even symmetry [49].

Intuitively, the information loss is due to the non-injectivity of g, which, employing Definition 2.3, is only invertible if the partition $\mathcal{X}_i$ from which the input X originated is already known. The following statements will put this intuition on solid ground.

Definition 2.4 (Partition Indicator)**.** The *partition indicator* W is a discrete RV which satisfies

$$W = i \quad \text{if} \quad X \in \mathcal{X}_i \tag{2.30}$$

[5] In fact, this invariance under a variable change tempted Edwin T. Jaynes to write "[...] that 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 $\int w \log w dx$ [...] we have realized that mathematical deduction from the uniqueness theorem, instead of guesswork, yields [an] invariant information measure" [78, p. 202].

for all i. In other words, W is obtained by quantizing X according to the partition $\{\mathcal{X}_i\}$.

Proposition 2.3. *The information loss is identical to the uncertainty about the set $\mathcal{X}_i$ from which the input was taken, i.e.,*

$$L(X \to Y) = H(W|Y). \tag{2.31}$$

Proof. See Appendix A.2. □

The mathematical justification of the intuition-based claim is actually contained in

Corollary 2.2. *System output Y and partition indicator W together are a sufficient statistic of the system input X, i.e.,*

$$H(X|Y,W) = 0. \tag{2.32}$$

Proof. Since W is a function of X,

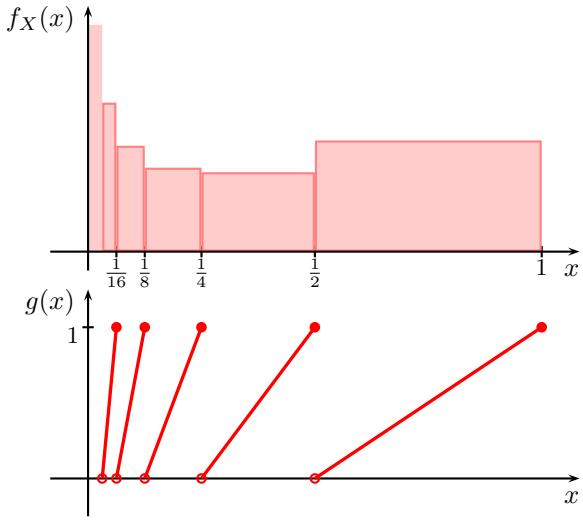
$$L(X \to Y) = H(X|Y) = H(X,W|Y) = H(X|W,Y) + H(W|Y) = H(X|W,Y) + L(X \to Y) \tag{2.33}$$

from which $H(X|Y,W) = 0$ follows. □

Knowing the output value, and the element of the partition from which the input originated, perfect reconstruction is possible. In all other cases, reconstruction will fail with some probability, cf. Section 2.3.3.

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

Example 4: Infinite Loss in a PBF.



Assume the scalar function $g\colon (0,1] \to (0,1]$ depicted above, mapping every interval $(2^{-n}, 2^{-n+1}]$ onto the interval $(0,1]$:

$$g(x) = 2^n(x - 2^{-n}) \text{ if } x \in (2^{-n}, 2^{-n+1}],\ n \in \mathbb{N} \tag{2.34}$$

The PDF of the input X is given as

$$f_X(x) = 2^n \left(\frac{1}{\log(n+1)} - \frac{1}{\log(n+2)} \right), \text{ if } x \in (2^{-n}, 2^{-n+1}],\ n \in \mathbb{N}. \tag{2.35}$$

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

To apply Proposition 2.3, one needs

$$\Pr(W = n|Y = y) = \Pr(W = n) = \frac{1}{\log(n+1)} - \frac{1}{\log(n+2)}. \tag{2.36}$$

For this distribution, the entropy is known to be infinite [6], and thus

$$L(X \to Y) = H(W|Y) = H(W) = \infty. \tag{2.37}$$

2.3.2 Bounds on the Information Loss

In many cases one cannot directly evaluate the information loss according to Theorem 2.2, since the differential entropy of Y involves the logarithm of a sum. This section therefore 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 3 and 4 – is the following corollary to Proposition 2.3, which is due to the fact that conditioning reduces entropy:

Corollary 2.3. $L(X \to Y) \leq H(W)$.

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 3 and 4 show.

Proposition 2.4 (Upper Bounds on Information Loss)**.** *The information loss induced by a PBF can be upper bounded by the following ordered set of inequalities:*

$$\begin{aligned} L(X \to Y) &\leq \int_{\mathcal{Y}} f_Y(y) \log \operatorname{card}(g^{-1}[y]) dy && (2.38)\\ &\leq \log\left(\sum_i \int_{\mathcal{Y}_i} f_Y(y) dy\right) && (2.39)\\ &\leq \operatorname*{ess\,sup}_{y \in \mathcal{Y}} \log \operatorname{card}(g^{-1}[y]) && (2.40)\\ &\leq \log \operatorname{card}(\{\mathcal{X}_i\}) && (2.41) \end{aligned}$$

where $\operatorname{card}(B)$ *is the cardinality of the set* B*. Bound* (2.38) *holds with equality if and only if*

$$\sum_{x_k \in g^{-1}[g(x)]} \frac{f_X(x_k)}{|\det \mathcal{J}_g(x_k)|} \frac{|\det \mathcal{J}_g(x)|}{f_X(x)} \overset{P_X\text{-a.s.}}{=} \operatorname{card}(g^{-1}[g(x)]). \tag{2.42}$$

If and only if this expression is constant P_X*-a.s., bounds* (2.39) *and* (2.40) *are tight. Bound* (2.41) *holds with equality if and only if additionally* $P_Y(\mathcal{Y}_i) = 1$ *for all* i*.*

Proof. See Appendix A.3. □

If the PDF of X and the absolute value of the Jacobian determinant (but not necessarily the cardinality of the preimage) are constant on $\mathcal{X}$, the first bound (2.38) holds with equality (cf. [51, conference version, Sect. VI.A]). But also two other scenarios, where these bounds hold with equality, are worth mentioning: First, for functions $g\colon \mathbb{R} \to \mathbb{R}$ equality holds if g is related to the cumulative distribution function F_X of the input RV such that, for all x, $|g'(x)| = f_X(x)$. From this immediately follows that g assumes, for all i,

$$g_i(x) = b_i F_X(x) + c_i \tag{2.43}$$

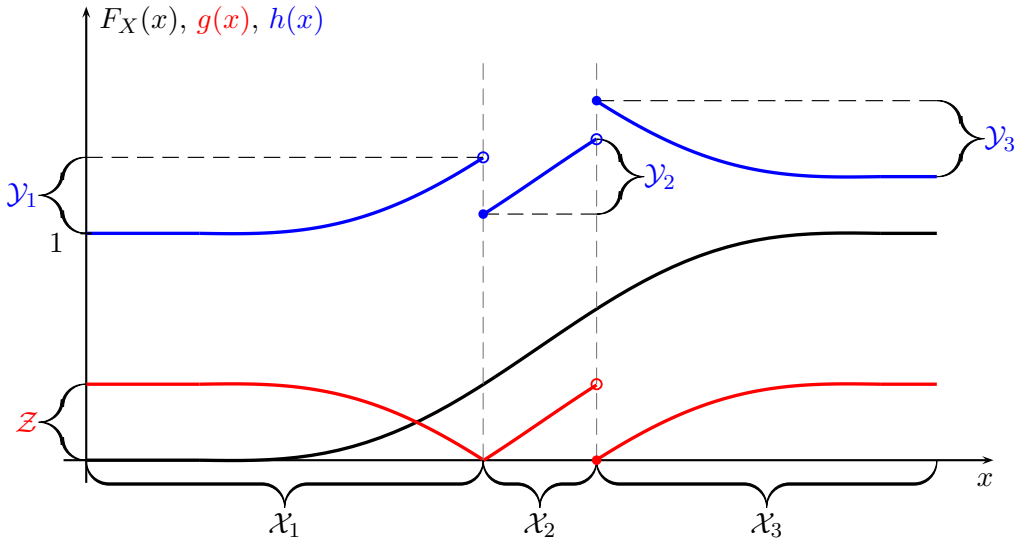


Figure 2.4: Piecewise bijective functions with $\mathrm{card}(\{\mathcal{X}_i\}) =: L = 3$ *illustrating tightness of Proposition 2.4. The black curve is the cumulative distribution function* F_X *of* X*. The function in blue,* $h\colon \mathcal{X} \to \mathcal{Y}$*, renders* (2.42) *piecewise constant but not constant due to improper setting of the constants* c_i*. Tightness is achieved in the smallest bound,* (2.38)*, only. The function in red,* $g\colon \mathcal{X} \to \mathcal{Z}$*, satisfies all conditions (i.e.,* (2.42) *is constant and* $\mathcal{Z}_l = \mathcal{Z}$ *for all* l*) and thus achieves equality in the largest bound* (2.41)*. The constants for* g *are* $(b_i) = (-1, 1, 1)$ *and* $(c_i) = (1/3, 1/3, 2/3)$*. Note that the elements* $\mathcal{X}_i$ *are chosen such that each contains the same probability mass.*

where $b_i \in \{1, -1\}$ and $c_i \in \mathbb{R}$ are multiplicative and additive constants which may differ on each element of the partition $\{\mathcal{X}_i\}$. The function $h\colon \mathcal{X} \to \mathcal{Y}$ depicted in Fig. 2.4 satisfies this condition.

The constants c_i and the probability masses in each interval are constrained if equality in (2.41) is desired. Specifically, in (2.42) $\mathrm{card}(g^{-1}[g(x)]) = \mathrm{card}(\{\mathcal{X}_i\}) =: L$ shall hold P_X-a.s., essentially stating that the image of every $\mathcal{X}_i$ has to coincide with $\mathcal{Y}$ except on a set of measure zero. Assuming that $\mathcal{X}_i$ are intervals containing the same probability mass, i.e., $P_X(\mathcal{X}_i) = 1/L$ for all i, (2.41) holds with equality for additive constants

$$c_i = -\sum_{l=1}^{i-1} P_X(\mathcal{X}_l) = -\frac{i-1}{L} \tag{2.44}$$

for $b_i = 1$ and

$$c_i = -\frac{i}{L} \tag{2.45}$$

for $b_i = -1$, respectively. A function $g\colon \mathcal{X} \to \mathcal{Z}$ satisfying these requirements is shown in Fig. 2.4.

The second case occurs when both function and PDF are "repetitive", in the sense that their behavior on $\mathcal{X}_1$ is copied to all other $\mathcal{X}_i$, and that, thus, $f_X(x_i)$ and $|\det J_g(x_i)|$ are the same for all elements of the preimage $g^{-1}[y]$. A corresponding function is depicted in Fig. 2.5. Less obviously, Example 3 also represents such a case.

Although the bounds of Proposition 2.4 are more elaborate than the one of Corollary 2.3, one can show that the latter is not necessarily useless. In particular, as the example below shows, it might yield a helpful result even when Proposition 2.4 completely fails to do so.

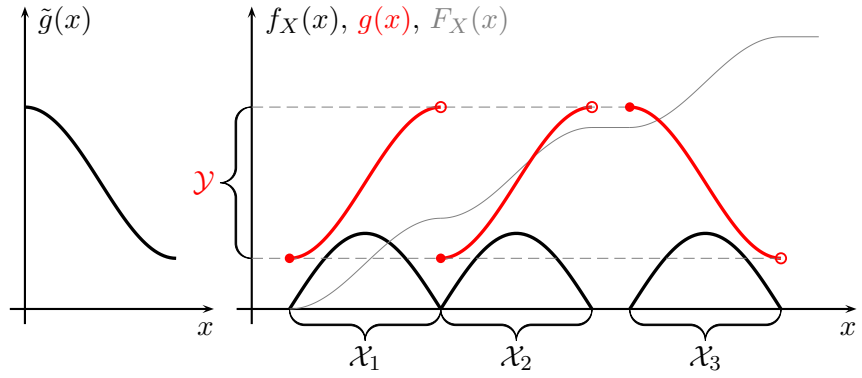
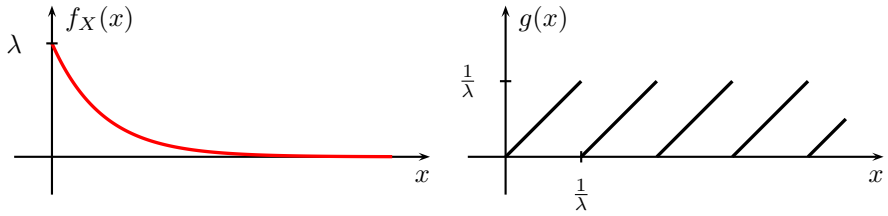


Figure 2.5: Piecewise bijective function with $L = 3$ satisfying all conditions of Proposition 2.4. The black and the grey functions are the PDF f_X and cumulative distribution function F_X of X, respectively. The function $g\colon \mathcal{X} \to \mathcal{Y}$ consists of L shifted copies of the prototype function $\tilde{g}$, possibly modified with different signs b_i on different sets $\mathcal{X}_i$. Moreover, the PDF f_X is identical on all elements $\mathcal{X}_i$ (equal probability mass in each interval). Note that in this example $\mathcal{X}$ is not an interval.

Example 5: Exponential RV and Infinite Bounds.



Consider an exponential input X with PDF

$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} \tag{2.46}$$

and the piecewise linear function

$$g(x) = x - \frac{\lfloor \lambda x \rfloor}{\lambda} \tag{2.47}$$

depicted above.

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,

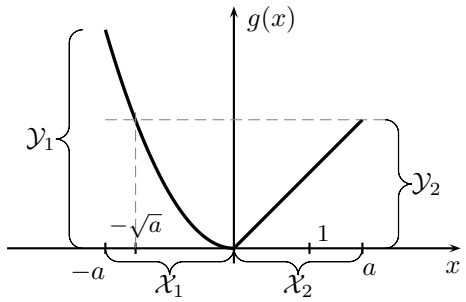
$$\mathcal{X}_k = \left[\frac{k-1}{\lambda}, \frac{k}{\lambda}\right) \tag{2.48}$$

and $g(\mathcal{X}_k) = \mathcal{Y}$ for all $k = 1, 2, \ldots$. From this follows that for every $y \in \mathcal{Y}$ the preimage contains an element from each $\mathcal{X}_k$; thus, the bounds from Proposition 2.4 all evaluate to $L(X \to Y) \leq \infty$. However, it can be shown that Corollary 2.3, $L(X \to Y) \leq H(W)$, is tight in this case: With

$$p_k = P_X(\mathcal{X}_k) = \int_{\mathcal{X}_k} f_X(x)dx = (1 - \mathrm{e}^{-1})\mathrm{e}^{-k+1} \tag{2.49}$$

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 2.2.

Example 6: The Square-Linear Function.



Consider an RV X uniformly distributed on $[-a, a]$, $a \geq 1$, and the function g depicted above,

$$g(x) = \begin{cases} x^2, & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases}. \tag{2.50}$$

The information loss computes to

$$L(X \to Y) = \frac{4a + 4\sqrt{a} + 1}{8a} \log(2\sqrt{a} + 1) - \frac{\log(2\sqrt{a})}{2} - \frac{1}{4\sqrt{a} \ln 2} \tag{2.51}$$

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 smaller output values than to higher 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 2.4 are not tight in this case, as they yield

$$L(X \to Y) \leq \frac{1 + \sqrt{a}}{2\sqrt{a}} \leq \log\left(\frac{3\sqrt{a} + 1}{2\sqrt{a}}\right) \leq 1 \leq 1 \tag{2.52}$$

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

2.3.3 Reconstruction Error and Information Loss

The fact that the preimage of every output value is an at most countable set suggests that – with a certain probability of error – the input can be perfectly reconstructed. Assuming a finite partition $\{\mathcal{X}_i\}$ (which by Proposition 2.4 corresponds to a finite information loss $L(X \to Y)$), the *maximum a-posteriori* (MAP) estimate of the input is correct with positive probability, cf. [20, 40]. In particular, and in accordance with intuition, if the information loss vanishes, perfect reconstruction is possible with probability one; see also Proposition 2.6 below. It is therefore justified to investigate the interplay between information loss in a PBF and the error probability for reconstructions in terms of Fano-type bounds, despite that it seems counter-intuitive to define a reconstruction error probability for a real-valued RV.

Pushing these concerns aside, there are practical reasons which justify this investigation too: Since the information loss of Theorem 2.2 involves the logarithm of a sum, it may not always be computable in closed form. If one can design a simple reconstruction scheme and calculate its error probability, an upper bound on the information loss can be obtained, which might outperform those of Proposition 2.4.

Conversely, given one already knows the information loss of a system, the presented inequalities bound the probability of perfect reconstruction. This is of interest, e.g., in semi-coherent broadcast scenarios such as in the IEEE 802.15.4a standard [85]. In these scenarios, the transmitter employs a mixture of coherent and non-coherent modulation (e.g., pulse-position modulation

and phase-shift keying). A non-coherent receiver, such as the energy detector, might under certain circumstances be able to decode also part of the coherently transmitted message; the presented inequalities can yield performance bounds.

To simplify analysis and to avoid the peculiarities of entropy pointed out in [75], the focus is on PBFs with finite partition. For a possible generalization to countable partitions the reader is refered to [74]. While the results in this section were derived with PBFs in mind, they are also valid for the more general case where the cardinality of the reconstruction alphabet depends on the observed output.

Definition 2.5 (Reconstructor & Reconstruction Error)**.** A *reconstructor* is a function $r\colon \mathcal{Y} \to \mathcal{X}$ mapping each system output to a value inside the domain of the system function. E denotes the event of a *reconstruction error*, i.e.,

$$E := \begin{cases} 1, & \text{if } r(Y) \neq X \\ 0, & \text{if } r(Y) = X \end{cases}. \tag{2.53}$$

The probability of a reconstruction error is thus given by

$$P_e := \Pr(E = 1) = \int_{\mathcal{Y}} P_e(y) dP_Y(y) \tag{2.54}$$

where $P_e(y) := \Pr(E = 1|Y = y)$.

As mentioned in the beginning of this section, there are two applications for the Fano-type bounds being derived below: First, to bound the error probability by the information loss from above and below; to this end, the MAP reconstructor will be analyzed. And second, the probability of a reconstruction error shall be used to bound the information loss from above; in this case, a simpler, sub-optimal reconstructor will be designed, for which the error probability is easier to obtain.

Generally, the MAP reconstructor is defined by

$$r_{\mathrm{MAP}}(y) := \operatorname*{argmax}_{x \in g^{-1}[y]} \Pr(X = x|Y = y). \tag{2.55}$$

In other words, with Definition 2.5, the MAP reconstructor minimizes $P_e(y)$. For the envisaged scenario, where $y = g(x)$, this simplifies to

Proposition 2.5 (MAP Reconstructor)**.** *The MAP estimator for a PBF is*

$$r_{\mathrm{MAP}}(y) = g_k^{-1}(y) \tag{2.56}$$

where

$$k = \operatorname*{argmax}_{i\colon\, g_i^{-1}(y) \neq \emptyset} \left\{ \frac{f_X(g_i^{-1}(y))}{|\det \mathcal{J}_g(g_i^{-1}(y))|} \right\}. \tag{2.57}$$

Proof. The proof follows from Corollary 2.3, stating that, given $Y = y$ is known, reconstructing the input is possible if the partition indicator W is known. The MAP reconstructor thus depends on the most likely partition indicator (given the output), i.e.,

$$r_{\mathrm{MAP}}(y) = \operatorname*{argmax}_i p(i|y) \tag{2.58}$$

where we used the notation from the proof of Proposition 2.3 (cf. Appendix A.2). The proof is

completed with

$$p(i|y) = \begin{cases} \frac{f_X(g_i^{-1}(y))}{|\det \mathcal{J}_g(g_i^{-1}(y))| f_Y(y)}, & \text{if } g_i^{-1}(y) \neq \emptyset \\ 0, & \text{if } g_i^{-1}(y) = \emptyset \end{cases}. \tag{2.59}$$

□

As a direct consequence, for the MAP reconstructor one has $r_{\mathrm{MAP}}(y) \in g^{-1}[y]$. Therefore, for this and any other reconstructor choosing the reconstruction from the preimage of the output, Fano's inequality holds and one gets [21, pp. 39]

$$L(X \to Y) \leq H_2(P_e) + P_e \log \left(\mathrm{card}(\{\mathcal{X}_i\}) - 1\right) \tag{2.60}$$

where $H_2(p) := -p \log p - (1-p) \log(1-p)$ is the entropy of a Bernoulli-p random variable. Trivially, one can exchange the cardinality of the partition $\{\mathcal{X}_i\}$ by the essential supremum over all preimage cardinalities (cf. Appendix A.4), i.e.,

$$L(X \to Y) \leq H_2(P_e) + P_e \log \left(\operatorname*{ess\,sup}_{y \in \mathcal{Y}} \mathrm{card}(g^{-1}[y]) - 1 \right). \tag{2.61}$$

Definition 2.6 (Bijective Part)**.** $\mathcal{X}_b$ is the maximal set which is mapped injectively by g, and $\mathcal{Y}_b$ is its image under g. Thus, g: $\mathcal{X}_b \to \mathcal{Y}_b$ bijectively, where

$$\mathcal{X}_b := \{x \in \mathcal{X} \colon \mathrm{card}(g^{-1}[g(x)]) = 1\}. \tag{2.62}$$

Then $P_b := P_X(\mathcal{X}_b) = P_Y(\mathcal{Y}_b)$ denotes the bijectively mapped probability mass.

Theorem 2.3 (Fano-Type Bound)**.** *For the MAP reconstructor – or any reconstructor r for which $r(y) \in g^{-1}[y]$ – the information loss $L(X \to Y)$ is upper bounded by*

$$L(X \to Y) \leq \min\{1 - P_b, H_2(P_e)\} - P_e \log P_e + P_e \log \left(\mathbb{E}\left(\mathrm{card}(g^{-1}[Y]) - 1 \right) \right). \tag{2.63}$$

Proof. See Appendix A.4. □

Comparing this result with Fano's original bound (2.60), one sees that the cardinality of the partition is replaced by the expected cardinality of the preimage. Due to the additional term $P_e \log P_e$ this improvement is only potential, since there exist cases where Fano's original bound is better. An example is the square-law device of Example 3, for which Fano's inequality is tight, but for which Theorem 2.3 would yield $L(X \to Y) \leq 2$ (but see also Section 2.3.4).

For the sake of completeness, note that the MAP reconstruction error admits also a lower bound on the information loss:

Proposition 2.6 (Feder & Merhav, [40])**.** *The information loss $L(X \to Y)$ is lower bounded by the error probability P_e of a MAP reconstructor by*

$$\phi(P_e) \leq L(X \to Y) \tag{2.64}$$

where $\phi(x)$ is the piecewise linear function

$$\phi(x) := \left(x - \frac{i-1}{i} \right) (i+1) i \log \left(1 + \frac{1}{i} \right) + \log i \tag{2.65}$$

for $\frac{i-1}{i} \leq x \leq \frac{i}{i+1}$.

Presenlty, it is not clear if this bound can be improved for the present context, since the cardinality of the preimage has no influence on ϕ.

In concrete examples, the MAP reconstructor is not always easy to find. For bounding the information loss of a system (rather than reconstructing the input), it is therefore desirable to introduce a simpler, sub-optimal reconstructor:

Proposition 2.7 (Suboptimal Reconstruction)**.** *Consider the following sub-optimal reconstructor*

$$r_{\text{sub}}(y) = \begin{cases} g^{-1}(y), & \text{if } y \in \mathcal{Y}_b \\ g_k^{-1}(y), & \text{if } y \in \mathcal{Y}_k \setminus \mathcal{Y}_b \\ x\text{: } x \in \mathcal{X}_k, & \text{else} \end{cases} \tag{2.66}$$

where

$$k = \operatorname*{argmax}_i P_X(\mathcal{X}_i \cup \mathcal{X}_b) \tag{2.67}$$

and where $\mathcal{Y}_k = g(\mathcal{X}_k)$.

Letting $\overline{K} = \operatorname{ess\,sup}_{y\in\mathcal{Y}} \operatorname{card}(g^{-1}[y])$ *and with the error probability*

$$\hat{P}_e = 1 - P_X(\mathcal{X}_k \cup \mathcal{X}_b) \tag{2.68}$$

of this reconstructor, the information loss is upper bounded by the following, Fano-type inequality:

$$L(X \to Y) \leq 1 - P_b + \hat{P}_e \log\left(\overline{K} - 1\right) \tag{2.69}$$

Proof. See Appendix A.5. □

This reconstructor is *simple* in the sense that the reconstruction is always chosen from the element $\mathcal{X}_k$ containing most of the probability mass, after considering the set on which the function is bijective. This allows for a simple evaluation of the reconstruction error probability $\hat{P}_e$, which is independent of the Jacobian determinant of g.

It is interesting to see that the Fano-type bound derived here permits a similar expression as was derived in Theorem 2.3, despite the fact that the sub-optimal reconstructor not necessarily satisfies $r_{\text{sub}}(y) \in g^{-1}[y]$. For this type of reconstructors, $(\operatorname{card}(\cdot) - 1)$ typically has to be replaced by $\operatorname{card}(\cdot)$. Note that, thus, also the following bounds hold:

$$\begin{aligned} L(X \to Y) &\leq H_2(\hat{P}_e) + \hat{P}_e \log\left(\operatorname*{ess\,sup}_{y\in\mathcal{Y}} \operatorname{card}(g^{-1}[y])\right) && (2.70)\\ &\leq H_2(\hat{P}_e) + \hat{P}_e \log\left(\operatorname{card}(\{\mathcal{X}_i\})\right) && (2.71)\\ L(X \to Y) &\leq \min\{1 - P_b, H_2(P_e)\} - P_e \log P_e + P_e \log\left(\mathbb{E}\left(\operatorname{card}(g^{-1}[Y])\right)\right) && (2.72) \end{aligned}$$

Before proceeding, briefly reconsider

Example 4: Infinite Loss (revisited).

For this example it was shown that the information loss was infinite. By recognizing that the probability mass contained in $\mathcal{X}_1 = (\frac{1}{2}, 1]$ exceeds the mass contained in all other subsets, one obtains an error probability for reconstruction equal to

$$\hat{P}_e = \frac{1}{\log 3} \approx 0.63. \tag{2.73}$$

In this particular case $P_e = \hat{P}_e$ holds, since the MAP reconstructor coincides with the suboptimal reconstructor. But since $\operatorname{card}(g^{-1}[y]) = \infty$ for all $y \in \mathcal{Y}$, all Fano-type bounds evaluate to infinity.

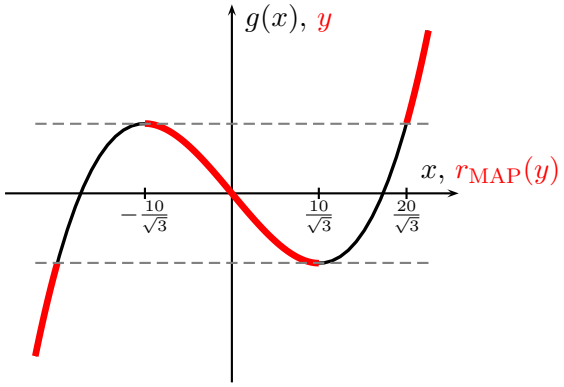


Figure 2.6: Third-order polynomial and its MAP reconstructor (indicated with a bold red line).

Example 6: Square-Linear Function (revisited).
It was shown that the information loss calculates to $L(X \to Y) \approx 0.922$ bits. The error probability of a MAP reconstructor can thus be bounded via Fano's bound (which in this case is better than Theorem 2.3) and Proposition 2.6:

$$0.337 \leq P_e \leq 0.461 \tag{2.74}$$

A simple analysis shows that the MAP reconstructor is

$$r_{\mathrm{MAP}}(y) = \begin{cases} -\sqrt{y}, & \text{if } 0 \leq y < 1/\sqrt{2} \\ y, & \text{else} \end{cases} \tag{2.75}$$

and that it has an error probability of $P_e \approx 0.433$. This value can be used to bound the information loss:

$$0.866 \leq L(X \to Y) \leq 0.987. \tag{2.76}$$

2.3.4 Example: Third-Order Polynomial

Many non-linear functions used in practice are polynomials (e.g., the square-law device in the energy detector) or can at least be approximated by polynomials (cf. the Stone-Weierstrass theorem [127, Thm. 7.26, p. 159]). Moreover, despite the fact that static, i.e., memoryless, functions as they are discussed in this work, are a minority among practically relevant systems, they still form an important constituent: Wiener and Hammerstein systems, for example, containing just a static nonlinearity and a linear filter, are very often used for system approximation, too.

Consider the third-order polynomial depicted in Fig. 2.6, which is defined as

$$g(x) = x^3 - 100x. \tag{2.77}$$

The input to this function is a zero-mean Gaussian RV X with variance σ^2. A closed-form evaluation of the information loss is not possible, since the integral involves the logarithm of a sum. However, note that

$$\mathcal{X}_b = \left(-\infty, -\frac{20}{\sqrt{3}}\right] \cup \left[\frac{20}{\sqrt{3}}, \infty\right) \tag{2.78}$$

and thus $P_b = 2Q\left(\frac{20}{\sqrt{3}\sigma}\right)$, where Q denotes the Q-function [3, 26.2.3]. With a little algebra the

bounds from Proposition 2.4 evaluate to

$$L(X \to Y) \leq (1 - P_b) \log 3 \leq \log (3 - 2P_b) \leq \log 3 \tag{2.79}$$

where $\operatorname{ess\,sup}_{y \in \mathcal{Y}} \operatorname{card}(g^{-1}[y]) = \operatorname{card}(\{\mathcal{X}_i\}) = 3$.

The sub-optimal reconstructor $r_{\text{sub}}(y)$ assigns every ambiguous output to the interval with largest probability mass (given that all bijectively mapped mass has already been considered). Since this is the mass contained in the interval $[-\frac{10}{\sqrt{3}}, \frac{10}{\sqrt{3}}]$, the error probability of the sub-optimal reconstructor is

$$\hat{P}_e = 2Q\left(\frac{10}{\sqrt{3}\sigma}\right) - 2Q\left(\frac{20}{\sqrt{3}\sigma}\right). \tag{2.80}$$

Interestingly, the sub-optimal reconstructor coincides with the MAP reconstructur $r_{\text{MAP}}(y)$ in this example: Clearly, the PDF f_X is maximized for the element of the preimage lying in the interval $[-\frac{10}{\sqrt{3}}, \frac{10}{\sqrt{3}}]$. Showing that the derivative of g is *minimized* for this particular element proves that the two reconstructors coincide (cf. Proposition 2.5).

To this end note that, by the trigonometric method to solve a cubic equation, the preimage of $y \in \mathcal{Y} \setminus \mathcal{Y}_b$ contains the elements

$$x_k = 2\sqrt{\frac{100}{3}} \cos\left(\frac{\varphi + 2k\pi}{3}\right) \qquad k = 0, 1, 2 \tag{2.81}$$

where

$$\varphi = \arccos\left(\frac{y}{2}\sqrt{\frac{27}{10^6}}\right). \tag{2.82}$$

From the fact that $x = 0$ implies $y = 0$ it can be verified that the desired element of the preimage has index $k = 1$, i.e., is x_1. By reasons of symmetry it thus suffices to check whether

$$|g'(x_1)| \leq |g'(x_0)| \tag{2.83}$$

holds for all $y \in \mathcal{Y} \setminus \mathcal{Y}_b$. But this is equivalent to

$$\frac{1}{2} \leq \cos^2\left(\frac{\varphi}{3}\right) + \cos^2\left(\frac{\varphi + 2\pi}{3}\right) \tag{2.84}$$

$$= 1 + \frac{1}{2}\cos\left(\frac{2\varphi}{3}\right) + \frac{1}{2}\cos\left(\frac{2\varphi + 4\pi}{3}\right) \tag{2.85}$$

$$= 1 + \cos\left(\frac{2\pi}{3}\right)\cos\left(\frac{2\varphi + 2\pi}{3}\right) \tag{2.86}$$

$$= 1 - \frac{1}{2}\cos\left(\frac{2\varphi + 2\pi}{3}\right) \tag{2.87}$$

which obviously holds for all $\varphi \in [-\pi, \pi)$.

Thus, instead of the bounds of Proposition 2.7, those of Theorem 2.3 and Proposition 2.6 can be applied. Using $P_e = \hat{P}_e$ the bounds are displayed, together with Fano's bound and the bounds from Proposition 2.4, in Fig. 2.7.

It can be seen that the information loss is small for small variance of the input signal. This is quite intuitive, since in this case most of the probability mass is concentrated on 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 more and more probability mass is mapped bijectively. Moreover, it can be seen that the bound of Proposition 2.4 is not

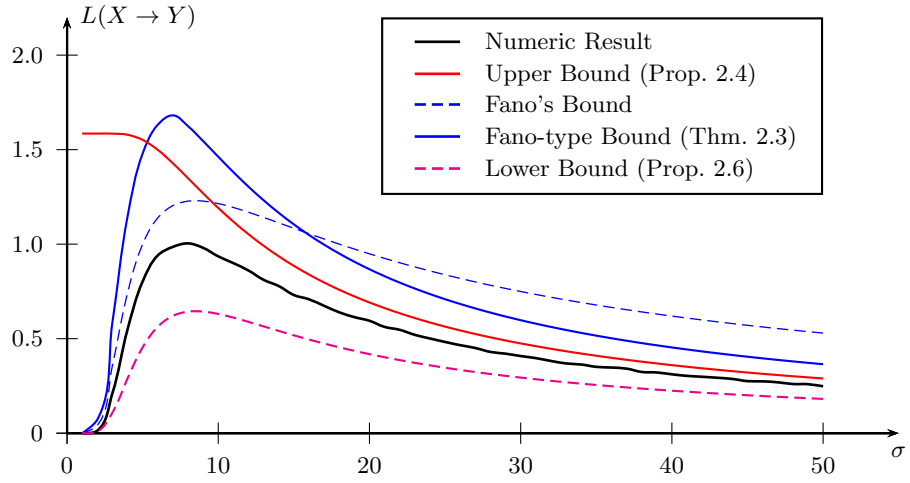


Figure 2.7: Information loss for third-order polynomial as a function of input variance σ^2.

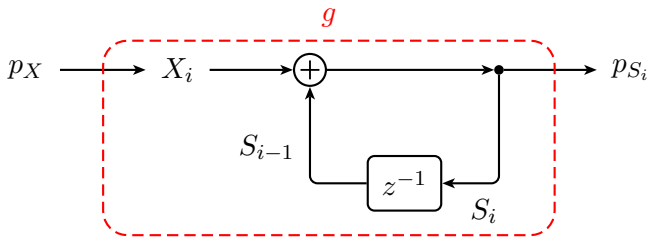


Figure 2.8: Accumulating a sequence of independent, identically distributed RVs X_i

helpful for small input signal variances. The reason is that the discussed bound does not take the reconstruction error probability into account, but only the expected cardinality of the preimage. Thus, bounds depending on the reconstruction error probability perform better for small σ^2.

What becomes apparent from this example is that the bounds from Proposition 2.4 and Theorem 2.3 cannot form an ordered set; the same holds for Fano's inequality, which can be better or worse than the Fano-type bound.

2.3.5 Example: Accumulator

As a further example, consider the system depicted in Fig. 2.8: How much information about a probability measure of an RV X is lost if one observes the probability measure of a sum of independent, identically distributed (iid) copies of this RV? For simplicity, assume that the probability measure of X is supported on a finite field $\mathcal{X} = \{0, \dots, N-1\}$, where N is even.

The input to the system is thus the PMF p_X. Since its elements must sum to unity, the vector p_X can be chosen from the $(N-1)$-simplex in $\mathbb{R}^N$; assume $P_{p_X} \ll \lambda^{N-1}$. The output of the system at a certain discrete time index i is the PMF p_{S_i} of the sum S_i of i iid copies of X:

$$S_i = \bigoplus_{k=1}^{i} X_k \tag{2.88}$$

where $\bigoplus$ denotes modulo-addition.

Given the PMF p_{S_i} of the output S_i at some time i (e.g., by computing the histogram of multiple realizations of this system), how much information is lost about the PMF of the input

X? Mathematically, this information loss is represented by the following quantity[6]:

$$L(p_X \to p_{S_i}) \tag{2.89}$$

Note that the transition from S_i to S_{i+1} can be modeled as a cyclic random walk; the transition matrix of the corresponding Markov chain is a positive circulant matrix built[7] from p_X. As a consequence of the Perron-Frobenius theorem (e.g., [114, Thm. 15-8]) there exists a unique stationary distribution which, for a doubly stochastic matrix as in this case, equals the uniform distribution on $\mathcal{X}$ (cf. [87, Thm. 4.1.7]).

To attack the problem, note that the PMF of the sum of independent RVs is given as the convolution of the PMFs of the summands. In case of the modulo sum, the circular convolution needs to be applied instead, as can be shown by computing one Markov step. Using the discrete Fourier transform (DFT), the circular convolution turns into a multiplication of the corresponding spectra; in particular (see, e.g., [111, Sec. 8.6.5])

$$(p_X * p_{S_i}) \longleftrightarrow F_{p_X} F_{p_{S_i}} \tag{2.90}$$

where $F_{p_X} = \mathsf{DFT}(p_X)$. Since p_X is a real vector, F_{p_X} will be Hermitian (circularly) symmetric (see, e.g., [111, Sec. 8.6.4]); moreover, using indices from 0 to $N-1$, $F_{p_X}^{(0)} = 1$ and $F_{p_X}^{(\frac{N}{2})} \in \mathbb{R}$. Iterating the system i times and repeating this analysis yields

$$p_{S_i} = \mathsf{DFT}^{-1}\big(F_{p_X}^i\big) \tag{2.91}$$

where the i-th power is taken element-wise. Neither DFT nor inverse DFT lose information; information is only lost in taking F_{p_X} to the i-th power. Taking the i-th power of a real number β loses at most one bit of information for even i and nothing for odd i; thus

$$L(\beta \to \beta^i) \leq \cos^2\left(\frac{i\pi}{2}\right). \tag{2.92}$$

Taking the power of a complex number α corresponds to taking the power of its magnitude (which can be inverted by taking the corresponding root) and multiplying its phase. Only the latter is a non-injective operation, since the i-th root of a complex number yields i different solutions for the phase. Thus, invoking Proposition 2.4

$$L(\alpha \to \alpha^i) \leq \log i. \tag{2.93}$$

Applying (2.92) to index $\frac{N}{2}$ and (2.93) to the indices $\{1, \ldots, \frac{N}{2} - 1\}$ reveals that

$$L(p_X \to p_{S_i}) = L(F_{p_X} \to F_{p_{S_i}}) \leq \left(\frac{N}{2} - 1\right) \log i + \cos^2\left(\frac{i\pi}{2}\right). \tag{2.94}$$

Thus, the information loss increases linearly with the number N of components, but sublinearly with the number if iterations i. Moreover, since for $i \to \infty$, p_{S_i} converges to a uniform distribution, it is plausible that $L(p_X \to p_{S_i}) \to \infty$ (at least for $N > 2$). Intuitively, the earlier one observes the output of such a system, the more information about the unknown PMF p_X can be retrieved.

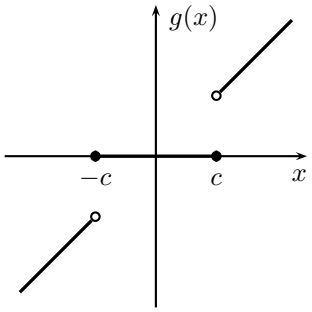
[6] Note that $L(p_X \to p_{S_i})$ is not to be confused with $H(S_1|S_i)$, a quantity measuring the information loss about the initial *state* of a Markov chain.

[7] In this particular case it can be shown that the first row of the transition matrix consists of the elements of p_X, while all other rows are obtained by circularly shifting the first one.

2.4 Relative Information Loss

In some cases, neither the information loss nor the information transfer, i.e., the mutual information between input and output of a system, are sufficient to fully characterize its information-processing behavior. As a particular example, consider the center clipper:

Example 7: Center Clipper.



The center clipper, used for, e.g., residual echo suppression [152], can be described by the following function:

$$g(x) = \begin{cases} x, & \text{if } |x| > c \\ 0, & \text{otherwise} \end{cases} \tag{2.95}$$

Assuming again that $P_X \ll \lambda$ and that $0 < P_X([-c, c]) < 1$, with Corollary 2.1 the information loss becomes infinite. However, the set $\{(x, y) \in \mathcal{X} \times \mathcal{Y} : x = y, |x| > c\}$ has positive P_{XY}-measure, but vanishing $P_X P_Y$-measure. Thus, with [116, Thm. 2.1.2]

$$I(X;Y) = \infty. \tag{2.96}$$

As this example shows, there are systems for which neither information transfer (i.e., the mutual information between input and output) nor information loss provides sufficient insight. For these systems, a different characterization is necessary, which leads to

Definition 2.7 (Relative Information Loss)**.** The *relative information loss* induced by g is

$$l(X \to Y) = \lim_{n\to\infty} \frac{H(\hat{X}^{(n)}|Y)}{H(\hat{X}^{(n)})} \tag{2.97}$$

provided the limit exists.

2.4.1 Elementary Properties

One elementary property of the relative information loss is that $l(X \to Y) \in [0, 1]$, due to the non-negativity of entropy and the fact that $H(\hat{X}^{(n)}|Y) \leq H(\hat{X}^{(n)})$. More interestingly, the relative information loss is related to the Rényi information dimension:

Definition 2.8 (Rényi Information Dimension [123])**.** The *information dimension* of an RV X is

$$d(X) = \lim_{n\to\infty} \frac{H(\hat{X}^{(n)})}{n} \tag{2.98}$$

provided the limit exists and is finite.

This definition is adopted from Wu and Verdú, who showed in [169, Prop. 2] that it is equivalent to the one given by Rényi in [123]. The case that the information dimension is infinite,

which may occur if $H(\hat{X}^{(0)}) = \infty$ [169, Prop. 1], is excluded here. Conversely, if the information dimension of an RV X exists, it is guaranteed to be finite if $H(\hat{X}^{(0)}) < \infty$ [123] or if $\mathbb{E}(|X|^{\epsilon}) < \infty$ for some $\epsilon > 0$ [169]. Aside from that, the information dimension exists for discrete RVs and RVs with probability measures absolutely continuous w.r.t. the Lebesgue measure on a sufficiently smooth manifold [123], for mixtures of RVs with existing information dimension [123, 144, 169], and self-similar distributions generated by iterated function systems [169]. Finally, the information dimension exists if the MMSE dimension exists [170, Thm. 8]. The remainder of this work is based on the assumption that the information dimension of all considered RVs exists and is finite.

Proposition 2.8 (Relative Information Loss and Information Dimension)**.** *Let X be an N-dimensional RV with positive information dimension $d(X)$ and finite $H(\hat{X}^{(0)})$. If $d(X|Y=y)$ and $H(\hat{X}^{(0)}|Y=y)$ exist and are finite P_Y-a.s., the relative information loss equals*

$$l(X \to Y) = \frac{d(X|Y)}{d(X)} \tag{2.99}$$

where $d(X|Y) = \int_{\mathcal{Y}} d(X|Y=y) dP_Y(y)$.

Proof. From Definition 2.7,

$$l(X \to Y) = \lim_{n\to\infty} \frac{H(\hat{X}^{(n)}|Y)}{H(\hat{X}^{(n)})} \tag{2.100}$$

$$= \lim_{n\to\infty} \frac{\int_{\mathcal{Y}} H(\hat{X}^{(n)}|Y=y) dP_Y(y)}{H(\hat{X}^{(n)})} \tag{2.101}$$

$$= \lim_{n\to\infty} \frac{\int_{\mathcal{Y}} \frac{H(\hat{X}^{(n)}|Y=y)}{n} dP_Y(y)}{\frac{H(\hat{X}^{(n)})}{n}}. \tag{2.102}$$

By assumption, the limit of the denominator and of the expression under the integral both exist and correspond to $d(X)$ and $d(X|Y=y)$, respectively. By assumption, $H(\hat{X}^{(0)}|Y=y)$ is finite P_Y-a.s., thus, with [123, eq. (11)], the expression under the integral is finite for all $n \geq 1$:

$$\frac{H(\hat{X}^{(n)}|Y=y)}{n} \leq \frac{H(\hat{X}^{(0)}|Y=y) + N(n+1)}{n} < H(\hat{X}^{(0)}|Y=y) + 2N < \infty \tag{2.103}$$

The latter is an integrable function, integrating to $H(\hat{X}^{(0)}|Y) + 2N \leq H(\hat{X}^{(0)}) + 2N < \infty$. One can thus apply Lebesgue's dominated convergence theorem (e.g., [128, Thm. 1.34, p. 26]) to exchange the order of the limit and the integral:

$$l(X \to Y) = \frac{\lim_{n\to\infty} \int_{\mathcal{Y}} \frac{H(\hat{X}^{(n)}|Y=y)}{n} dP_Y(y)}{d(X)} = \frac{\int_{\mathcal{Y}} d(X|Y=y) dP_Y(y)}{d(X)} = \frac{d(X|Y)}{d(X)}. \tag{2.104}$$

This completes the proof. □

The relative information loss shall be accompanied by its logical complement, the relative information transfer:

Definition 2.9 (Relative Information Transfer)**.** The *relative information transfer* through the system g is

$$t(X \to Y) = \lim_{n\to\infty} \frac{I(\hat{X}^{(n)}; Y)}{H(\hat{X}^{(n)})} = 1 - l(X \to Y) \tag{2.105}$$

provided the limit exists.

The relative information loss was introduced to characterize systems for which the absolute information loss from Definition 2.2 is infinite. The following result shows that, at least for input RVs with infinite entropy, an infinite absolute information loss is a prerequisite for positive relative information loss:

Proposition 2.9 (Positive Relative Loss leads to Infinite Absolute Loss)**.** *Let X be such that $H(X) = \infty$ and let $l(X \to Y) > 0$. Then, $L(X \to Y) = \infty$.*

Proof. The proposition is proved by contradiction: Assume that $L(X \to Y) = H(X|Y) = \kappa < \infty$. Thus,

$$\begin{aligned} l(X \to Y) &= \lim_{n\to\infty} \frac{H(\hat{X}^{(n)}|Y)}{H(\hat{X}^{(n)})} && (2.106)\\ &\overset{(a)}{\leq} \lim_{n\to\infty} \frac{H(X|Y)}{H(\hat{X}^{(n)})} && (2.107)\\ &\overset{(b)}{=} 0 && (2.108) \end{aligned}$$

where (a) is due to data processing and (b) follows from $H(X|Y) = \kappa < \infty$ and from $H(\hat{X}^{(n)}) \to H(X) = \infty$ (e.g., [66, Lem. 7.18]). □

Note that the converse is not true: There exist examples where an infinite amount of information is lost, but for which the relative information loss nevertheless vanishes, i.e., $l(X \to Y) = 0$ (see Example 4).

2.4.2 Relative Information Loss for System Reducing the Dimensionality of Continuous Random Variables

Systems for which the information loss $L(X \to Y)$ is infinite subsume, among others, those which reduce the dimensionality of the input signal, e.g., by dropping coordinates or by keeping the function constant on a subset of its domain. This section is devoted to an investigation of this particular class of systems. Interestingly, a reduction of the dimension of the support $\mathcal{X}$ does not necessarily lead to a positive relative information loss, nor does its preservation guarantee vanishing relative information loss.

Assume that $\mathcal{X} \subseteq \mathbb{R}^N$ and $P_X \ll \lambda^N$, thus $d(X) = N$. Assume further that g is piecewise defined, as in Definition 2.3. Here, however, $g_i\colon \mathcal{X}_i \to \mathcal{Y}_i$ are not necessarily bijective, but projections.

Proposition 2.10 (Relative Information Loss in Dimensionality Reduction)**.** *Let $\{\mathcal{X}_i\}$ be a partition of $\mathcal{X}$. Let g be such that $g_i = g|_{\mathcal{X}_i}$ are projections to M_i coordinates. Then, the relative information loss is*

$$l(X \to Y) = \sum_{i=1}^{K} P_X(\mathcal{X}_i) \frac{N - M_i}{N}. \tag{2.109}$$

Proof. See Appendix A.6. □

As it is argued in the proof, the result can be generalized to *submersions*, i.e., smooth functions between smooth manifolds whose pushforward is surjective everywhere (see, e.g., [95]). This generalization is, however, not carried out in this work.

Corollary 2.4. *Let g be any projection of X onto M of its coordinates. Then, the relative information loss is*

$$l(X \to Y) = \frac{N - M}{N}. \tag{2.110}$$

Corollary 2.5. *Let g be constant on a set $A \subseteq \mathcal{X}$ with positive P_X-measure. Let furthermore g be such that* $\operatorname{card}(g^{-1}[y]) < \infty$ *for all $y \notin g(A)$. Then, the relative information loss is*

$$l(X \to Y) = P_X(A). \tag{2.111}$$

The first of these two corollaries will be applied to principal components analysis in Section 2.6, while the second solves Example 7 (center clipper): While both the information loss and the information transfer are infinite, the relative information loss corresponds to the probability mass contained in the clipping region, i.e., $l(X \to Y) = P_X([-c, c])$. Yet another illustration of these corollaries is given in

Example 8: Adding Two RVs.
Consider two N-dimensional input RVs X_1 and X_2, and assume that the output of the system under consideration is

$$Y = X_1 + X_2 \tag{2.112}$$

i.e., the sum of these two RVs.

For the moment, assume that X_1 and X_2 have a joint probability measure $P_{X_1,X_2} \ll \lambda^{2N}$. As it can be shown rather easily by transforming X_1, X_2 invertibly to $X_1 + X_2, X_1$, dropping the second coordinate leads to

$$l(X_1, X_2 \to Y) = \frac{1}{2}. \tag{2.113}$$

Things look different if the joint probability measure P_{X_1,X_2} is supported on some lower-dimensional submanifold of $\mathbb{R}^{2N}$. Consider, e.g., the case where $X_2 = -X_1$, thus $Y \equiv 0$, and $l(X_1, X_2 \to Y) = 1$. In contrary to this, assume that both input variables are one-dimensional, and that $X_2 = -0.01X_1^3$. Then, as it turns out,

$$Y = X_1 - 0.01X_1^3 = -0.01(X_1^3 - 100X_1) \tag{2.114}$$

which is a piecewise bijective function. As the analysis of the third-order polynomial in Section 2.3.4 shows, $l(X_1, X_2 \to Y) = 0$ in this case.

The somewhat surprising consequence of these results is that the shape of the PDF (if one exists) has no influence on the relative information loss; whether the PDF is peaky in the clipping region or flat, or whether the omitted coordinates are highly correlated to the preserved ones does neither increase nor decrease the relative information loss. As Example 8 shows, things become more complicated when the distribution of the input does not have a PDF: The choice of the set on which the probability mass is supported (be it either of the same or of a smaller dimension as the function's domain) can have a large influence on the relative information loss.

Proposition 2.10 permits another corollary, one which is of great interest and which is also proved in Appendix A.6:

Corollary 2.6. *In the setting of Proposition 2.10,*

$$t(X \to Y) = \frac{d(Y)}{N}. \tag{2.115}$$

Indeed, the author believes that this behavior is the "general" behavior for functions reducing the dimensionality of the data. This leads to the following

Conjecture 2.1 (Relative Information Transfer and Information Dimension)**.** *Let X be an RV with positive information dimension $d(X)$ and let g be a Lipschitz function. Then, the relative*

information transfer through this function is

$$t(X \to Y) = \frac{d(Y)}{d(X)}. \tag{2.116}$$

The author is currently not able to prove this conjecture, but a possible line of reasoning is outlined in Appendix A.7. Moreover, this conjecture suggests that, at least for a restricted class of functions,

$$d(X) = d(X, Y) = d(X|Y) + d(Y) \tag{2.117}$$

holds. This complements the results of [22], where it was shown that the *point-wise* information dimension satisfies this chain rule (second equality) given that the conditional probability measure satisfies a Lipschitz property. Moreover, the first equality was shown to hold for the point-wise information dimension given that Y is a Lipschitz function of X; for non-Lipschitz functions, the point-wise information dimension of (X, Y) may exceed the one of X. If the same holds for the information dimension (which is the expectation over the point-wise information dimension) is an interesting question for future research.

A corollary to Conjecture 2.1 is the counterpart of Proposition 2.2 for relative information loss:

Conjecture 2.2 (Relative Information Loss of a Cascade)**.** *Consider two Lipschitz functions $g\colon \mathcal{X} \to \mathcal{Y}$ and $h\colon \mathcal{Y} \to \mathcal{Z}$ and a cascade of systems implementing these functions. Let $Y := g(X)$ and $Z := h(Y)$. For the cascade of these systems the relative information transfer and relative information loss are*

$$t(X \to Z) = t(X \to Y)t(Y \to Z) \tag{2.118}$$

and

$$l(X \to Z) = l(X \to Y) + l(Y \to Z) - l(X \to Y)l(Y \to Z) \tag{2.119}$$

respectively.

This conjecture about cascades can be proved for cascades of projections (with absolutely continuous inputs) and also for discrete RVs.

2.4.3 A Bound on the Relative Information Loss

Complementing the results from Section 2.3.2, this subsection presents a bound for the relative information loss. Clearly, from the trivial bounds on the information dimension ($d(X) \in [0, N]$ if $\mathcal{X}$ is a subset of the N-dimensional Euclidean space or a sufficiently smooth N-dimensional manifold) simple bounds on the relative information loss can be computed. The bound on the relative information loss presented here for an N-dimensional input RV is formulated by the corresponding coordinate-wise quantities.

Proposition 2.11 (Upper Bound on the Relative Information Loss)**.** *Let X be an N-dimensional RV with a probability measure $P_X \ll \lambda^N$ and let Y be N-dimensional. Then,*

$$l(X \to Y) \leq \frac{1}{N} \sum_{i=1}^{N} l(X_i \to Y) \leq \frac{1}{N} \sum_{i=1}^{N} l(X_i \to Y_i) \tag{2.120}$$

where X_i and Y_i are the i-th coordinates of X and Y, respectively.

Proof. The proof follows from the fact that $d(X) = N$ and, for all i, $d(X_i) = 1$. From the definition of relative information loss,

$$l(X \to Y) = \lim_{n\to\infty} \frac{H(\hat{X}_1^{(n)}, \ldots, \hat{X}_N^{(n)}|Y)}{H(\hat{X}_1^{(n)}, \ldots, \hat{X}_N^{(n)})} \tag{2.121}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=1}^N H(\hat{X}_i^{(n)}|\hat{X}_1^{(n)}, \ldots, \hat{X}_{i-1}^{(n)}, Y)}{H(\hat{X}_1^{(n)}, \ldots, \hat{X}_N^{(n)})} \tag{2.122}$$

$$\leq \lim_{n\to\infty} \frac{\sum_{i=1}^N H(\hat{X}_i^{(n)}|Y)}{H(\hat{X}_1^{(n)}, \ldots, \hat{X}_N^{(n)})} \tag{2.123}$$

Exchanging limit and summation yields

$$l(X \to Y) \leq \frac{\sum_{i=1}^N d(X_i|Y)}{d(X)} = \frac{\sum_{i=1}^N d(X_i|Y)}{N}. \tag{2.124}$$

But since $N = Nd(X_i)$ for all i, this is equivalent to

$$l(X \to Y) \leq \frac{1}{N}\sum_{i=1}^N \frac{d(X_i|Y)}{d(X_i)} = \frac{1}{N}\sum_{i=1}^N l(X_i \to Y). \tag{2.125}$$

This proves the first inequality. The second is obtained by removing conditioning again in (2.123), since $Y = \{Y_1, \ldots, Y_N\}$. □

Example 9: Projection.
Let X be an N-dimensional RV with probability measure $P_X \ll \lambda^N$ and let X_i denote the i-th coordinate of X. Let g be a projection onto the first $M < N$ coordinates. The information loss is given as

$$l(X \to Y) = \frac{N - M}{N} \tag{2.126}$$

by Corollary 2.4. Note that $l(X_i \to Y) = 0$ for $i \in \{1, \ldots, M\}$, while $l(X_i \to Y) = 1$ for $i \in \{M + 1, \ldots, N\}$. This shows tightness of Proposition 2.11.

Example 8: Adding Two RVs (revisited).
Take the adder $Y = X_1 + X_2$ with $P_{X_1,X_2} \ll \lambda^2$: Since in the general case the probability measure $P_{X_i|Y=y}$ possesses a density, one gets $l(X_i \to Y) = 1$ for $i = 1, 2$. The bound thus evaluates to 1 and the inequality is strict, since $l(X \to Y) = \frac{1}{2} < 1$.

2.4.4 Reconstruction Error and Relative Information Loss

Complementing Section 2.3.3, this subsection presents a Fano-type relation between the relative information loss and the probability of a reconstruction error. While for piecewise bijective functions this relation was justified by the fact that for every output value the preimage under the system function is a countable set, the case is completely different here: Quantizers, for example, characterized by relative information loss, are typically evaluated based on energetic measures (e.g., the mean-squared reconstruction error). As the following example shows, the relative information loss does not permit a meaningful interpretation in energetic terms, underlining the intrinsically different behavior of information and energy measures.

Example 1: Quantizer (revisited).
Consider a continuous one-dimensional RV X ($P_X \ll \lambda$) and the quantizer introduced in Section 2.2. Since the quantizer is constant P_X-a.s., one obtains with Corollary 2.5

$$l(X \to \hat{X}_n) = 1. \tag{2.127}$$

In other words, the quantizer destroys 100% of the information available at its input. This naturally holds for all n, so a finer partition $\mathcal{P}_n$ cannot decrease the relative information loss. Contrarily, the mean-squared reconstruction error decreases with increasing n.

Definition 2.10 (Minkowski Dimension)**.** The Minkowski or box-counting dimension of a compact set $\mathcal{X} \subset \mathbb{R}^N$ is

$$d_B(\mathcal{X}) = \lim_{n\to\infty} \frac{\log \operatorname{card}(\mathcal{P}_n)}{n} \tag{2.128}$$

where the partition $\mathcal{P}_n$ is induced by a uniform vector quantizer with quantization interval $\frac{1}{2^n}$.

The Minkowski dimension of a set equals the information dimension of a uniform distribution on that set (e.g., [39]), and is a special case of Rényi information dimension where the entropy is replaced with the Rényi entropy of zeroth order [64].

Proposition 2.12. *Let X be an RV with probability measure P_X with positive information dimension $d(X)$. Let the support $\mathcal{X} \subset \mathbb{R}^N$ of P_X be compact and have positive Minkowski dimension $d_B(\mathcal{X})$. Then, the error probability bounds the relative information loss from above by*

$$l(X \to Y) \leq P_e \frac{d_B(\mathcal{X})}{d(X)}. \tag{2.129}$$

Proof. Note that by the compactness of $\mathcal{X}$ the quantized input $\hat{X}^{(n)}$ has a finite alphabet, which allows employing Fano's inequality

$$H(\hat{X}^{(n)}|Y) \leq H_2(P_{e,n}) + P_{e,n} \log \operatorname{card}(\mathcal{P}_n) \tag{2.130}$$

where

$$P_{e,n} = \Pr(r(Y) \neq \hat{X}^{(n)}). \tag{2.131}$$

Since Fano's inequality holds for arbitrary reconstructors, let r be the composition of the MAP reconstructor r_{MAP} and the quantizer introduced in Section 2.2. Consequently, $P_{e,n}$ is the probability that $r_{\mathrm{MAP}}(Y)$ and X do not lie in the same quantization bin. Since the bin volume shrinks with increasing n, $P_{e,n}$ increases monotonically to P_e. Thus, with $H_2(p) \leq 1$ for $0 \leq p \leq 1$,

$$H(\hat{X}^{(n)}|Y) \leq 1 + P_e \log \operatorname{card}(\mathcal{P}_n). \tag{2.132}$$

With the introduced definitions,

$$l(X \to Y) = \lim_{n\to\infty} \frac{H(\hat{X}^{(n)}|Y)}{H(\hat{X}^{(n)})} \leq \lim_{n\to\infty} \frac{1 + P_e \log \operatorname{card}(\mathcal{P}_n)}{H(\hat{X}^{(n)})} \overset{(a)}{=} P_e \frac{d_B(\mathcal{X})}{d(X)} \tag{2.133}$$

where (a) is obtained by dividing both numerator and denominator by n and evaluating the limit. This completes the proof. □

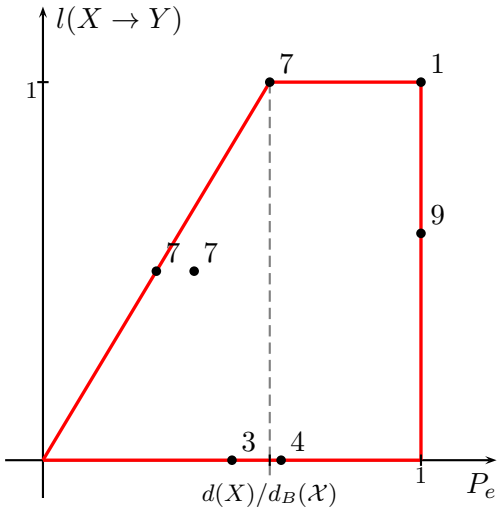


Figure 2.9: The accessible region for a $(P_e, l(X \to Y))$-pair for $d(X) = 0.6 d_B(\mathcal{X})$. Points with numbers indicate references to the numbered examples in this work. Note that the center clipper (Example 7) occurs three times in the plot, representing each instance in the text.

Note that always $d(X) \leq d_B(\mathcal{X}) \leq N$ if $\mathcal{X} \subset \mathbb{R}^N$, e.g., by [168, Thm. 1 and Lem. 4]. Therefore, in the case where $\mathcal{X} \subset \mathbb{R}^N$ and $P_X \ll \lambda^N$ it can be shown that this result simplifies to $l(X \to Y) \leq P_e$. Comparing this to the results of Example 1, it turns out that for a quantizer the reconstruction error probability is always $P_e = 1$.

The possible region for a $(P_e, l(X \to Y))$-pair is depicted in Fig. 2.9. Note that, to the best of the author's knowledge, this region cannot be restricted further. For example, with reference to Section 2.3 there exist systems with $l(X \to Y) = 0$ but with $P_e > 0$. Conversely, for a simple projection one will have $P_e = 1$ while $l(X \to Y) < 1$. Finally, that $l(X \to Y) = 1$ need not imply $P_e = 1$ can be shown by revisiting the center clipper:

Example 7: Center Clipper (revisited).
Assume that the input probability measure P_X is mixed with an absolutely continuous component supported on $[-c, c]$ ($0 < P_X([-c, c]) < 1$) and a point mass at an arbitrary point $x_0 \notin [-c, c]$. According to [123, 169], $d(X) = P_X([-c, c])$. The output probability measure P_Y has two point masses at 0 and x_0 with $P_Y(0) = P_X([-c, c])$ and $P_Y(x_0) = 1 - P_X([-c, c])$, respectively. Clearly, $d(X|Y = 0) = 1$ while $d(X|Y = x_0) = 0$. Consequently,

$$l(X \to Y) = \frac{d(X|Y)}{d(X)} = 1. \tag{2.134}$$

In comparison, $P_e \leq P_X([-c, c])$, since one can always use the reconstructor $r(y) = x_0$ for all y.

This is a further example where Conjecture 2.1 holds, despite that neither the center clipper is Lipschitz, nor that the requirement of a continuously distributed input RV in Proposition 2.10 is met.

It is worth mentioning that Proposition 2.12 together with Conjecture 2.1 allows to prove converses to lossless analog compression, as investigated in [169,171]. To this end, and borrowing the terminology and notation from [169], let us *encode* a length-n block X of independent realizations of a real-valued input RV Z with information dimension $0 < d(Z) \leq 1$ via a Lipschitz mapping to the Euclidean space of dimension $\lfloor \mathsf{R} n \rfloor \leq n$. Let $\mathsf{R}(\epsilon)$ be the infimum of R such that there exists a Lipschitz $g\colon \mathbb{R}^n \to \mathbb{R}^{\lfloor \mathsf{R} n \rfloor}$ and an arbitrary (measurable) reconstructor r such that $P_e \leq \epsilon$.

Conjecture 2.3 (Converse for Lipschitz Encoders; connection to [169], [171, eq. (26)])**.** *For a memoryless source with compactly supported marginal distribution P_Z and information dimension $0 < d(Z) \leq 1$, and a Lipschitz encoder function g,*

$$\mathsf{R}(\epsilon) \geq d(Z) - \epsilon. \tag{2.135}$$

Proof. Since X is the collection of n real-valued, independent RVs Z_i with probability measure P_Z supported on $\mathcal{Z}$, it follows that $\mathcal{Z}^n \subset \mathbb{R}^n$ and $d_B(\mathcal{Z}^n) = n$. With Proposition 2.12 it thus follows

$$\begin{aligned} nP_e &\geq d(X)l(X \to Y) && (2.136)\\ &\overset{(a)}{=} d(X) - d(Y) && (2.137)\\ &\overset{(b)}{=} nd(Z) - d(Y) && (2.138) \end{aligned}$$

where (a) is due to Conjecture 2.1 and (b) is due to the fact that the information dimension of a set of independent RVs is the sum of the individual information dimensions (see, e.g., [22] or [168, Lem. 3]). Since Y is an $\mathbb{R}^{\lfloor \mathsf{R}n \rfloor}$-valued RV, $d(Y) \leq \lfloor \mathsf{R}n \rfloor$. Thus,

$$nP_e \geq nd(Z) - \lfloor \mathsf{R}n \rfloor \geq nd(Z) - \mathsf{R}n. \tag{2.139}$$

Dividing by the block length n and rearranging the terms completes the proof. □

While this result – compared with those presented in [169, 171] – is rather weak, it suggests that the presented theory has relationships with different topics in information theory, such as compressed sensing. Note further that the reconstructor is arbitrary, since already the encoder – the function g – loses information. The restriction of Lipschitz continuity cannot be dropped, since, as stated in [169], there are non-Lipschitz bijections from $\mathbb{R}^n$ to $\mathbb{R}$.

2.4.5 Relative Information Loss for 1D-Systems of Mixed Random Variables

Briefly consider the case where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$, but where not necessarily $P_X \ll \lambda$. Instead, let P_X be a mixture of continuous and discrete probability measures, i.e., assume that P_X has no singular continuous component. Thus, [128, pp. 121]

$$P_X = P_X^{ac} + P_X^{d}. \tag{2.140}$$

According to [123, 169] the information dimension of X equals the probability mass with absolutely continuous distribution, $d(X) = P_X^{ac}(\mathcal{X})$.

Proposition 2.13 (Relative Information Loss for Mixed RVs)**.** *Let X be a mixed RV with a probability measure $P_X = P_X^{ac} + P_X^{d}$, $0 < P_X^{ac}(\mathcal{X}) \leq 1$. Let $\{\mathcal{X}_i\}$ be a finite partition of $\mathcal{X} \subseteq \mathbb{R}$ into compact sets. Let g be a bounded function such that $g|_{\mathcal{X}_i}$ is either injective or constant. The relative information loss is given as*

$$l(X \to Y) = \frac{P_X^{ac}(A)}{P_X^{ac}(\mathcal{X})} \tag{2.141}$$

where A is the union of sets $\mathcal{X}_i$ on which g is constant.

Proof. Since the partition is finite, using [169, Thm. 2] yields

$$d(X|Y=y) = \sum_i d(X|Y=y, X \in \mathcal{X}_i) P_{X|Y=y}(\mathcal{X}_i). \tag{2.142}$$

If g is injective on $\mathcal{X}_i$, the intersection $g^{-1}[y] \cap \mathcal{X}_i$ is a single point[8], thus $d(X|Y=y, X \in \mathcal{X}_i) = 0$.

[8] We do not consider the case here that the preimage of y does not intersect $\mathcal{X}_i$, since in this case $P_{X|Y=y}(\mathcal{X}_i) = 0$.

Conversely, if g is constant on $\mathcal{X}_i$, the preimage is $\mathcal{X}_i$ itself, so one obtains

$$d(X|Y) = \int_{\mathcal{Y}} \sum_i d(X|Y=y, X \in \mathcal{X}_i) P_{X|Y=y}(\mathcal{X}_i) dP_Y(y) \tag{2.143}$$

$$= \int_{\mathcal{Y}} \sum_{i:\mathcal{X}_i \subseteq A} \frac{P_X^{ac}(\mathcal{X}_i)}{P_X(\mathcal{X}_i)} P_{X|Y=y}(\mathcal{X}_i) dP_Y(y) \tag{2.144}$$

$$\overset{(a)}{=} \sum_{i:\mathcal{X}_i \subseteq A} \frac{P_X^{ac}(\mathcal{X}_i)}{P_X(\mathcal{X}_i)} \int_{\mathcal{Y}} P_{X|Y=y}(\mathcal{X}_i) dP_Y(y) \tag{2.145}$$

$$\overset{(b)}{=} P_X^{ac}(A) \tag{2.146}$$

where in (a) summation and integration were exchanged with the help of Fubini's theorem [128, Thm. 8.8, p. 164] and (b) is due to the fact that the sum runs over exactly the union of sets on which g is constant, A. Proposition 2.8 completes the proof. □

Comparing this result with Corollary 2.5, one can see that the former implies the latter for $P_X^{ac}(\mathcal{X}) = 1$. Moreover, as in Example 7, the relative information loss induced by a function g can increase if the probability measure is not absolutely continuous: In this case, from $P_X([-c,c])$ (where $P_X \ll \lambda$) to 1. As will be shown next, the relative information loss can also decrease:

Example 7: Center Clipper (revisited).
Assume that $P_X^{ac}(\mathcal{X}) = 0.6$ and $P_X^{ac}([-c,c]) = 0.3$. The remaining probability mass is a point mass at zero, i.e., $P_X(0) = P_X^d(0) = 0.4$. It follows that $d(X) = 0.6$ and, from Proposition 2.13, $l(X \to Y) = 0.5$. Fixing $r(0) = 0$ gives a reconstruction error probability $P_e = P_X^{ac}([-c,c]) = 0.3$. Using Proposition 2.12 yields

$$0.5 = l(X \to Y) \leq \frac{d_B(\mathcal{X})}{d(X)} P_e = \frac{1}{0.6} 0.3 = 0.5 \tag{2.147}$$

which shows that in this case the bound holds with equality. Moreover, here $l(X \to Y) < P_X([-c,c]) = 0.7$.

Consider now the case that the point mass at 0 is split into two point masses at $a, b \in [-c,c]$, where $P_X^d(a) = 0.3$ and $P_X^d(b) = 0.1$. Using $r(0) = a$ the reconstruction error increases to $P_e = P_X([-c,c]) - P_X^d(a) = 0.4$. The inequality in Proposition 2.12 is now strict.

2.5 Application: Multi-Channel Autocorrelation Receiver

The multi-channel autocorrelation receiver (MC-AcR) was introduced in [166] and analyzed in [105, 115] as a non-coherent receiver architecture for ultrawide band communications. In this receiver the decision metric is formed by evaluating the autocorrelation function of the input signal for multiple time lags (see Fig. 2.10).

To simplify the analysis, assume that the input signal is a discrete-time, complex-valued N-periodic signal superimposed with independent and identically distributed complex-valued noise. The complete analysis can be based on N consecutive values of the input, denoted by X_1 through X_N (X is the collection of these RVs). The real and imaginary parts of X_i are $\Re X_i$ and $\Im X_i$, respectively. Assume further that $P_X \ll \lambda^{2N}$, where P_X is compactly supported. The analysis considers only three time lags $k_1, k_2, k_3 \in \{1, \ldots, N-1\}$. Furthermore, for the sake of simplicity, the notation assumes that Conjecture 2.1 holds; in this case, this is unproblematic, since the relevant operations are linear, and can thus be represented as a cascade of a bi-Lipschitz function (which does not affect the information dimension) and a projection (for which Proposition 2.10 can be applied).

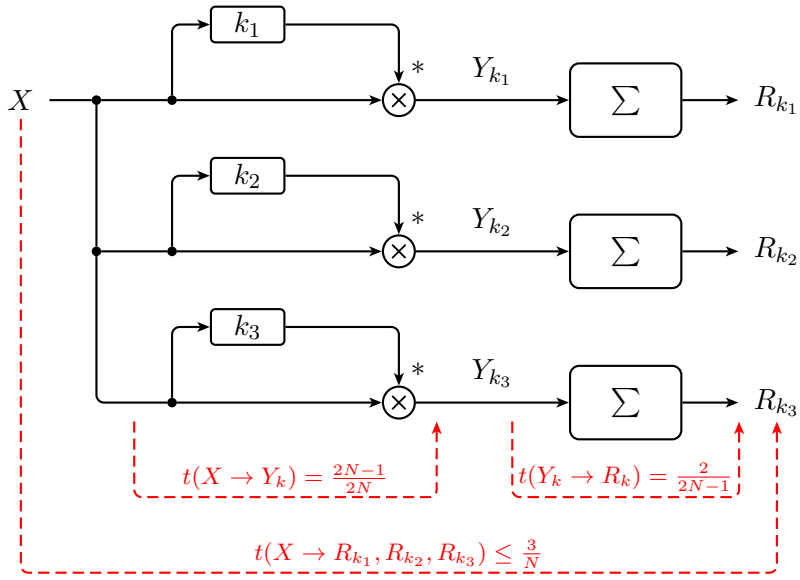


Figure 2.10: Discrete-time model of the multi-channel autocorrelation receiver: The elementary mathematical operations depicted are the complex conjugation ($*$), the summation of vector elements ($\sum$), and the circular shift (blocks with k_i). The information flow is illustrated using red arrows, labeled according to the relative information transfer.

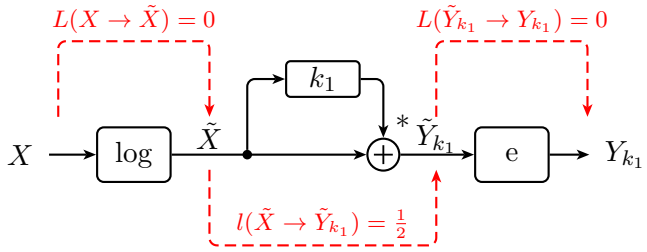


Figure 2.11: Equivalent model for multiplying the two branches in Fig. 2.10.

By the periodicity of the input signal, the linear autocorrelation is equivalent to the circular autocorrelation (e.g., [111, pp. 655])

$$R_k = \sum_{n=0}^{N-1} X_n X^*_{(n+k)} = \sum_{n=0}^{N-1} Y_{k,n} \tag{2.148}$$

where $*$ denotes complex conjugation, $Y_{k,n}$ is the n-th element of Y_k, and where k assumes any value of the set $\{k_1, k_2, k_3\}$.

For $k = 0$, $l(X \to Y_0) = \frac{1}{2}$, since

$$Y_{0,n} = X_n X^*_n = |X_n|^2 \tag{2.149}$$

is real and, thus, $P_{Y_0} \ll \lambda^N$. Since R_0 is the sum of the components of Y_0, it is real as well and one gets

$$t(X \to R_0) = \frac{1}{2N}. \tag{2.150}$$

For non-zero k note that (see Fig. 2.11)

$$X_n X_{n+k}^* = \mathrm{e}^{\log X_n X_{n+k}^*} = \mathrm{e}^{\log X_n + \log X_{n+k}^*} = \mathrm{e}^{\log X_n + (\log X_{n+k})^*}$$
$$= \mathrm{e}^{\tilde{X}_n + \tilde{X}_{n+k}^*} = \mathrm{e}^{\tilde{Y}_{k,n}} = Y_{k,n}. \quad (2.151)$$

In other words, one can write the multiplication as an addition (logarithm and exponential function are invertible and, thus, information lossless). Letting $\tilde{X}_k$ denote the vector of elements indexed by $\tilde{X}_{n+k}$, $n = 0, \ldots, N-1$,

$$\tilde{X}_k = \mathbf{C}_k \tilde{X}_0 \quad (2.152)$$

where $\mathbf{C}_k$ is a circulant permutation matrix. Thus, $\tilde{Y}_k = \tilde{X}_0 + \tilde{X}_k^*$ and

$$\Re\tilde{Y}_k = (\mathbf{I} + \mathbf{C}_k)\Re\tilde{X}_0 \quad (2.153)$$
$$\Im\tilde{Y}_k = (\mathbf{I} - \mathbf{C}_k)\Im\tilde{X}_0 \quad (2.154)$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Since $\mathbf{I} + \mathbf{C}_k$ is invertible, $P_{\Re\tilde{Y}_k} \ll \lambda^N$. In contrast, the rank of $\mathbf{I} - \mathbf{C}_k$ is $N-1$ and thus $d(\Im\tilde{Y}_k) = N-1$. It follows that

$$t(\Re\tilde{X}_0, \Im\tilde{X}_0 \to \Re\tilde{Y}_k, \Im\tilde{Y}_k) = t(\tilde{X}_0 \to \tilde{Y}_k) = t(X \to Y_k) = \frac{2N-1}{2N} \quad (2.155)$$

and $d(Y_k) = 2N-1$.

For $k \neq 0$ the autocorrelation will be a complex number a.s., thus $d(R_k) = 2$. Since the summation is a bi-Lipschitz function followed by a projection,

$$t(Y_k \to R_k) = \frac{2}{2N-1} \quad (2.156)$$

and by the result about the cascades in Conjecture 2.2,

$$t(X \to R_k) = \frac{1}{N}. \quad (2.157)$$

Finally, by bounding the information dimension of $\{R_{k_1}, R_{k_2}, R_{k_3}\}$ from above by the dimension of its support, one obtains

$$t(X \to R_{k_1}, R_{k_2}, R_{k_3}) \leq \frac{3}{N}. \quad (2.158)$$

Note that this analysis would imply that, if all values of the autocorrelation function would be evaluated, the relative information transfer would increase to

$$t(X \to R) \leq \frac{2N-1}{2N}. \quad (2.159)$$

The autocorrelation function of a complex, periodic sequence is Hermitian and periodic with the same period, from which follows that $P_R \ll \lambda^N$ and $t(X \to R) = \frac{1}{2}$. The bound is thus not tight in this case. Furthermore, applying the same bound to the relative information transfer from X to $Y_{k_1}, Y_{k_2}, Y_{k_3}$ yields a number greater than one. This is due to the three output vectors having lots of information in common, prohibiting simply adding the dimensions of their supports: The support of their joint distribution has a dimension strictly smaller than the Cartesian product of the supports of the marginal distributions.

A slightly different picture is revealed by looking at an equivalent signal model, where the circular autocorrelation is computed via the discrete Fourier transform (DFT, cf. Fig. 2.12): Let $F_X = \mathsf{DFT}(X)$. Doing a little algebra, the DFT of the autocorrelation function is obtained as

$$F_R = |F_X|^2. \quad (2.160)$$

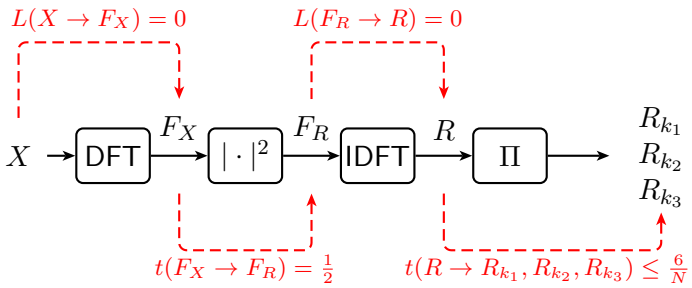


Figure 2.12: Equivalent model of the MC-AcR of Fig. 2.10, using the DFT to compute the circular autocorrelation. IDFT *denotes the inverse DFT and* Π *a projection onto a subset of coordinates. The information flow is indicated by red arrows labeled according to the relative information transfer.*

Since the DFT is an invertible transform, one has $d(F_X) = d(X) = 2N$. The squaring of the magnitude can be written as a cascade of a coordinate transform (from Cartesian to polar coordinates; invertible), a dimensionality reduction (the phase information is dropped), and a squaring function (invertible, since the magnitude is a non-negative quantity). It follows that

$$l(X \to R) = l(F_X \to F_R) = \frac{1}{2}. \tag{2.161}$$

since the inverse DFT is again invertible.

Since F_R is a real RV and thus $P_{F_R} \ll \lambda^N$, it follows that also $P_R \ll \lambda^N$, despite the fact that R is a vector of N complex numbers. The probability measure of this RV is concentrated on an N-dimensional submanifold of $\mathbb{R}^{2N}$ defined by the periodicity and Hermitian symmetry of R. Choosing three coordinates of R as the final output of the system amounts to upper bounding the information dimension of the output by

$$d(R_{k_1}, R_{k_2}, R_{k_3}) \leq 6. \tag{2.162}$$

Thus,

$$t(X \to R_{k_1}, R_{k_2}, R_{k_3}) \leq \frac{3}{N} \tag{2.163}$$

where equality is achieved if, e.g., all time lags are distinct and smaller than $\frac{N}{2}$. The information flow for this example – computed from the relative information loss and information transfer – is depicted in Figs. 2.10 and 2.12.

2.6 Application: Principal Components Analysis

As another illustration for the results about the relative information loss in systems which reduce the dimensionality of the data, this section investigates the information loss in principal components analysis (PCA). In particular, it will be shown that – without any specific signal model in mind – PCA is *not* a transform which minimizes the information loss occurring in subsequent dimensionality reduction (although this statement can be qualified by assuming a signal model, as in Section 4.4). Furthermore, in case of PCA based on data samples, information is lost even without deliberately reducing the dimensionality of the data.

2.6.1 PCA with Population Covariance Matrix

In PCA one uses the eigenvalue decomposition (EVD) of the covariance matrix of a multivariate input to obtain a different representation of the input vector. Specifically, let X be an RV with distribution $P_X \ll \lambda^N$ and information dimension $d(X) = N$. Assume further that X has zero mean and a positive definite population covariance matrix $\mathbf{C}_X = \mathbb{E}\left(XX^T\right)$ which is known a priori. The case where $\mathbf{C}_X$ is not known but has to be estimated from the data is considered in Section 2.6.2.

The EVD of the covariance matrix yields

$$\mathbf{C}_X = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^T \tag{2.164}$$

where $\mathbf{W}$ is an orthogonal matrix (i.e., $\mathbf{W}^{-1} = \mathbf{W}^T$) and $\mathbf{\Sigma}$ is a diagonal matrix consisting of the N positive eigenvalues of $\mathbf{C}_X$. The PCA is defined as the following linear transform:

$$Y = g(X) = \mathbf{W}^T X. \tag{2.165}$$

Since this linear transform is bi-Lipschitz and invertible, the absolute and relative information loss vanish.

Often, however, the PCA is used for dimensionality reduction, where after the linear transform in (2.165) the elements of the random vector Y with the smallest variances are discarded (thus, preserving the subspace with the largest variance). Essentially, the mapping from Y to, e.g., $Y_1^M := [Y_1, \ldots, Y_M]$ is a projection onto the first $M < N$ coordinates. With $\mathbf{I}_M$ being a rectangular identity matrix with M rows and N columns, one can write this dimensionality reduction as $Y_1^M = \mathbf{I}_M Y$. By applying Corollary 2.4 the relative information loss is

$$l(X \to Y_1^M) = l(Y \to Y_1^M) = \frac{N-M}{N}. \tag{2.166}$$

Let us extend this analysis to the case where from Y_1^M an N-dimensional estimate $\tilde{X}$ of the original data X is reconstructed. This estimate is obtained using the linear transform

$$\tilde{X} = \mathbf{W}\mathbf{I}_M^T Y_1^M. \tag{2.167}$$

The (full-rank) matrix $\mathbf{I}_M^T$ is a mapping to a higher-dimensional space (i.e., from $\mathbb{R}^M$ to $\mathbb{R}^N$) and is thus bi-Lipschitz; so is the rotation with the matrix $\mathbf{W}$. Furthermore, the transform from Y_1^M to $\tilde{X}$ is invertible and, as a consequence, no additional information is lost. Thus,

$$l(X \to \tilde{X}) = \frac{N-M}{N} \tag{2.168}$$

where, using above notation, $\tilde{X} = \mathbf{W}\mathbf{I}_M^T\mathbf{I}_M\mathbf{W}^T X$.

Indeed, the same result would have been obtained if the rotation would have been performed using any other orthogonal matrix and regardless which elements of the rotated vector were discarded. In particular, also if just the first M components of X would have been preserved, one would get $l(X \to X_1^M) = \frac{N-M}{N}$.

Of course, PCA is known to be optimal in the sense that, by discarding the elements of Y with the smallest variances, the mean-squared reconstruction error for $\tilde{X}$ is minimized [25]. For this interpretation, measuring the information loss *with respect to a relevant random variable* may do the trick, providing us with a statement about the optimality of PCA in an information-theoretic sense, cf. [25, 118]. This approach will be taken up again in Chapter 4. Conversely, if one cannot determine which information at the input is relevant, one has no reason to perform PCA prior to reducing the dimension of the data.

Specifically, these results are in stark contrast with intuition, originating from our energy-

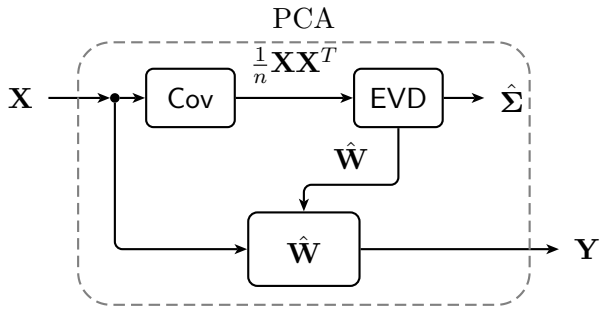


Figure 2.13: The PCA as a nonlinear input-output system. "Cov" denotes the computation of the sample covariance matrix and "EVD" stands for eigenvalue decomposition.

focused view on phenomena. Even when referring to Wikipedia, where it is stated, among other things, that "[...] PCA can supply the user with a lower-dimensional picture, a 'shadow' of this object when viewed from its (in some sense) most informative viewpoint" and that the variances of the dropped coordinates "tend to be small and may be dropped with minimal loss of information" [1]. Nothing tells us, however, that it is the strong components which convey most of the information; by coincidence, it might be the smallest eigenvalues representing the information which is of importance for us. This, seemingly counter-intuitive, example shall remind us that, without prior knowledge about the significant components of the data, PCA prior to dimensionality reduction, as it is often applied in practice, might not yield the desired result of *preserving information*[9].

2.6.2 PCA with Sample Covariance Matrix

In Section 2.6.1 the fact that PCA without dimensionality reduction is an invertible transform was employed. This, however, only holds if one has access to the orthogonal matrix $\mathbf{W}$; if not, i.e., if one feeds a system with the data matrix $\mathbf{X}$ and just receives the rotated matrix $\mathbf{Y}$ (see Fig. 2.13), it can be shown that information is lost. This is due to PCA now being a non-linear operation:

$$\mathbf{Y} = \hat{\mathbf{W}}^T\mathbf{X} = \mathbf{w}^T(\mathbf{X})\mathbf{X} \tag{2.169}$$

The author believes that the following analysis can be generalized to all data matrices with a continuous joint distribution, i.e., $P_{\mathbf{X}} \ll \lambda^{nN}$. For the sake of simplicity, however, the focus is on a particularly simple scenario: Let $\mathbf{X}$ denote a matrix where each of its n columns represents an *independent* sample of an N-dimensional Gaussian RV X. Again, let X have zero mean and positive definite population covariance matrix $\mathbf{C}_X$. As a consequence, the probability distribution of the data matrix $\mathbf{X}$ is absolutely continuous w.r.t. the nN-dimensional Lebesgue measure ($P_{\mathbf{X}} \ll \lambda^{nN}$).

The sample covariance matrix $\hat{\mathbf{C}}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ is symmetric and almost surely positive definite. In the usual case where $n \geq N$ one can show that $N(N+1)/2$ entries can be chosen and that the remaining entries depend on these in a deterministic manner. Indeed, since in this case the distribution of $\hat{\mathbf{C}}_X$ possesses a density (the Wishart distribution, cf. [107]), the distribution is absolutely continuous w.r.t. the Lebesgue measure on an $N(N+1)/2$-dimensional submanifold of the N^2-dimensional Euclidean space. With some abuse of notation one can thus write $P_{\hat{\mathbf{C}}_X} \ll \lambda^{\frac{N(N+1)}{2}}$.

[9] In some cases, the relevant part of the information is not known a priori. Henter [72, p. 54] takes unprecedented care in formulating his justification for PCA as a transform *preserving variability*: "In situations where the experimenter does not know a-priori what information to keep, feature extractors can be made to incorporate unsupervised dimensionality-reduction techniques such as [PCA] to discard information while retaining most of the empirical variability."

The orthogonal matrix $\hat{\mathbf{W}}$ for PCA (see (2.165); now applied to the matrix $\mathbf{X}$ instead of the vector X) is obtained from the EVD of the sample covariance matrix, i.e.,

$$\hat{\mathbf{C}}_X = \hat{\mathbf{W}}\hat{\mathbf{\Sigma}}\hat{\mathbf{W}}^T \tag{2.170}$$

where $\hat{\mathbf{\Sigma}}$ is the diagonal matrix containing the eigenvalues of $\hat{\mathbf{C}}_X$. The joint distribution of the N eigenvalues of $\hat{\mathbf{C}}_X$ possesses a density [107, Ch. 9.4]; thus, the distribution of $\hat{\mathbf{\Sigma}}$ is absolutely continuous w.r.t. the Lebesgue measure on an N-dimensional submanifold of the N^2-dimensional Euclidean space, or

$$P_{\hat{\mathbf{\Sigma}}} \ll \lambda^N. \tag{2.171}$$

Clearly, the entries of $\hat{\mathbf{C}}_X$ are smooth functions of the eigenvalues and the entries of $\hat{\mathbf{W}}$. Images of Lebesgue null-sets under smooth functions between Euclidean spaces of same dimension are null-sets themselves; were the probability measure $P_{\hat{\mathbf{W}}}$ supported on some set of dimensionality lower than $N(N+1)/2$, the image of the product of this set and $\mathbb{R}^N$ (for the eigenvalues) would be a Lebesgue null-set with positive probability measure. Since this contradicts the fact that $\hat{\mathbf{C}}_X$ is continuously distributed, it follows that

$$P_{\hat{\mathbf{W}}} \ll \lambda^{\frac{N(N-1)}{2}}. \tag{2.172}$$

It is shown in Appendix A.8 that the rotated data does not tell us anything about the rotation, hence

$$P_{\hat{\mathbf{W}}|\mathbf{Y}=\mathbf{y}} \ll \lambda^{\frac{N(N-1)}{2}}. \tag{2.173}$$

Knowing $\mathbf{Y} = \mathbf{y}$, $\mathbf{X}$ is a linear, bi-Lipschitz function of $\hat{\mathbf{W}}$, and thus the information dimension remains unchanged; with Proposition 2.8,

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{d(\mathbf{X}|\mathbf{Y})}{d(\mathbf{X})} = \frac{d(\hat{\mathbf{W}}|\mathbf{Y})}{d(\mathbf{X})} = \frac{N(N-1)}{2nN} = \frac{N-1}{2n}. \tag{2.174}$$

The more samples are collected in the data matrix, the smaller is the relative information loss.

For the sake of completeness and dropping a little of the mathematical rigor, the less common case where there are less data samples than there are dimensions for each sample ($n < N$) shall be discussed. In this case, the sample covariance matrix is not full rank, which means that the EVD yields $N - n$ vanishing eigenvalues. Assuming that still $P_{\hat{\mathbf{C}}_X} \ll \lambda^{n\frac{2N-n+1}{2}}$, one finds along the same lines as in the case $n \geq N$ that the loss evaluates to

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{2N-n-1}{2N}. \tag{2.175}$$

The behavior of the relative information loss as a function of n is shown in Fig. 2.14 for different choices of N.

The relative information loss induced by PCA results from the fact that one cannot know *which rotation* led to the output data matrix. As a consequence, the relative information loss decreases with a larger number of samples: the total information increases while the uncertainty about the rotation remains the same. Note further that the relative information loss cannot exceed $(N-1)/N$ (for $n = 1$), which is due to the fact that the rotation preserves the norm of the sample.

Example 10: PCA with Singular Sample Covariance Matrix.
Consider the – admittedly less common – case of a singular sample covariance matrix. Let X be a two-dimensional Gaussian RV, and let $n = 1$, i.e., $\mathbf{X} = X$. The sample covariance

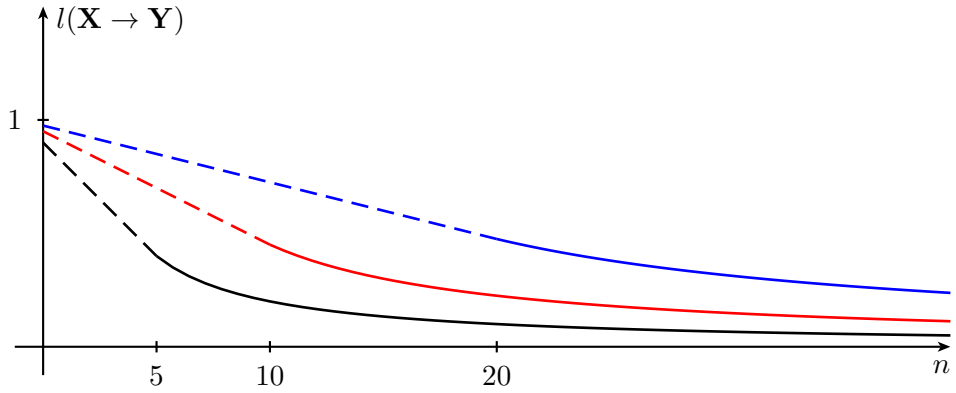


Figure 2.14: Relative information loss in the PCA with sample covariance matrix as a function of the number n of independent measurements. The cases $N = 5$ (black), $N = 10$ (red), and $N = 20$ (blue) are shown. The dashed lines indicate the conjectured loss for singular sample covariance matrices.

matrix is given by

$$\hat{\mathbf{C}}_X = \begin{bmatrix} X_1^2 & X_1 X_2 \\ X_1 X_2 & X_2^2 \end{bmatrix} \tag{2.176}$$

and has eigenvalues $|X|^2$ and 0. The corresponding (normalized) eigenvectors are then given by

$$p_1 = \left[\frac{X_1}{|X|}, \frac{X_2}{|X|}\right]^T \tag{2.177}$$

and

$$p_2 = \left[-\frac{X_2}{|X|}, \frac{X_1}{|X|}\right]^T. \tag{2.178}$$

Performing the rotation $Y = \hat{\mathbf{W}}^T X$ with $\hat{\mathbf{W}} = [p_1, p_2]$ one obtains

$$Y = [|X|, 0]^T. \tag{2.179}$$

The fact that the second component of Y is zero regardless of the entries of X makes it obvious that exactly one half of the information is lost, i.e., $l(X \to Y) = \frac{1}{2}$.

The PCA with sample covariance matrix also allows different interpretations of information loss, in addition to the information lost in the rotation: First, by the fact that the sample covariance matrix of $\mathbf{Y}$ is a diagonal matrix, the possible values of $\mathbf{Y}$ are restricted to a submanifold of dimensionality smaller than nN. Naturally, this also restricts the amount of information which can be conveyed in the output. In contrary to this, in PCA using the population covariance matrix the sample covariance matrix of $\mathbf{Y}$ (almost surely) does not contain zeros, so the entries of $\mathbf{Y}$ will not be restricted deterministically.

Finally, it is interesting to observe that in this case the *absolute* information loss in the PCA is infinite (see Proposition 2.9), even if no additional dimensionality reduction is performed. Moreover, this analysis not only holds for the PCA, but for *any* rotation which depends on the input data in a similar manner – in this sense, the PCA is not better than any other rotation. In fact, *there is no* rotation to be preferred, because doing nothing at least does not destroy valuable information. In practice, however, this advice is often ignored: In many cases, data is analyzed by performing PCA to see how the principal components behave. In the particular case where the rotation itself is the data relevant to the recipient, such an approach is fatal.

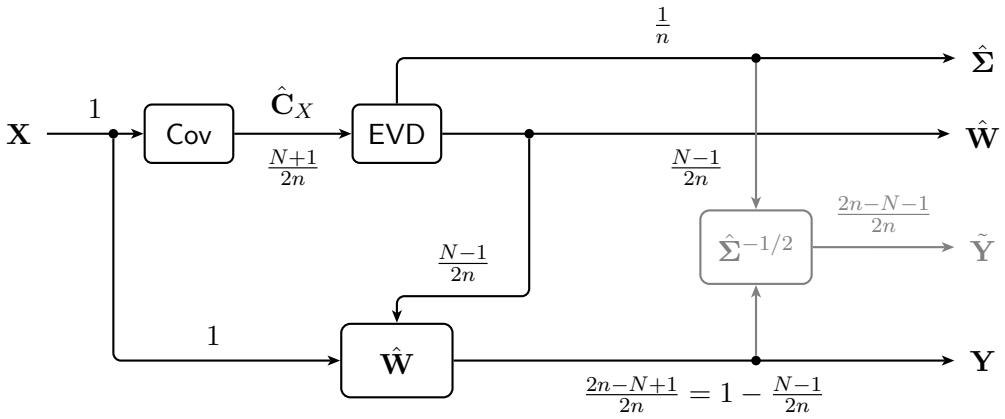


Figure 2.15: Information propagation in PCA with sample covariance matrix. The values on the arrows indicate the relative information transfer $t(\mathbf{X} \to \cdot)$. We assume $n > N$ in this case. The gray part considers the sphering transform (see text). Note that the separate information transfers to $\hat{\mathbf{\Sigma}}$, $\hat{\mathbf{W}}$, and $\mathbf{Y}$ add up to $1+\frac{1}{n} > 1$. This is because the information in $\hat{\mathbf{\Sigma}}$ is already contained in $\mathbf{Y}$. Note further that $t(\mathbf{X} \to \mathbf{Y}, \hat{\mathbf{\Sigma}}, \hat{\mathbf{W}}) = 1$, as expected.

2.6.3 Information Propagation in PCA

It is now time to investigate a different aspect of this work about information loss – be it absolute or relative – in deterministic systems. The definition of these quantities allows us to quantify and, hopefully, understand the propagation of information in a network of systems.

Take, for example, the PCA: The information at the input is split and propagates through the system, with parts of it being lost in some paths and preserved in others. By applying the relative information transfer from Definition 2.9, one can redraw the system model and obtain Fig. 2.15, where the arrows are labeled according to the relative information transfer along them. In particular, the eigenvalue decomposition splits the information contained in the covariance matrix into a part describing the eigenvectors and a part describing the eigenvalues. The former part is lost in PCA, while the second is preserved in the output. However, as mentioned before, the rotation (i.e., the multiplication with the orthogonal matrix) removes exactly as much information from the input as is contained in the orthogonal matrix. The information contained in both $\mathbf{Y}$ and $\hat{\mathbf{W}}$ suffices to reconstruct $\mathbf{X}$. After performing a sphering transform on $\mathbf{Y}$ (thus making all eigenvalues unity), one needs the triple $\tilde{\mathbf{Y}}$, $\hat{\mathbf{W}}$ and $\hat{\mathbf{\Sigma}}$ to reconstruct $\mathbf{X}$.

It is obvious that this transfer graph must not be understood in the sense of a preservation theorem, similar to Kirchhoff's current law: Information can be split and fused, and after splitting, the sum of information at the output needs not equal the sum of information at the input (as it is, by coincidence, for the EVD). If the output information is less than the input information, the intermediate system was lossy (as in the rotation). Conversely, if the sum of output information exceeds the input information, this only means that parts of the information must be the same (as, e.g., $\mathbf{Y}$ completely determines $\hat{\mathbf{\Sigma}}$).

2.7 Open Questions

Although this is probably one of the most evolved chapters in this work, the results are far from exhaustive. One will agree that information loss for discrete RVs, for absolutely continuous RVs in piecewise bijective functions, and for absolutely continuous RVs in functions reducing dimensionality have been treated to some satisfaction. But what about functions which reduce the dimensionality *and* are non-injective on their dimension-preserving domain? Consider, for

example, a symmetrized version of the center clipper (cf. Example 7): Is the statement "The system destroys 25% of the information plus one bit" satisfactory?

One of the most pressing needs is, of course, a proof of Conjecture 2.1 and its corollaries. That such a proof exists is strongly suggested by intuition.

Furthermore, the case where the input probability measure is neither discrete nor absolutely continuous has been treated only marginally. Not only input measures supported on a lower-dimensional manifold of the domain (e.g., on a line in $\mathbb{R}^2$), but also fractal input measures, like the Cantor measure, would be interesting to investigate – both for piecewise bijective functions and for systems reducing dimensionality.

While information loss was given an operational meaning by its connection to the reconstruction error, the operational meaning of relative information loss is not as clear. A connection to the reconstruction error was found in Section 2.4.4, but the problem of reconstruction and – in particular – the definition of a reconstruction error, seems ill-posed.

Whether the information measures proposed in this work suffice to develop a theory of information propagation in deterministic systems remains an open question. "Information flow graphs" or "information link budgets" may be the results; a hint in this direction was given in Section 2.6, where the PCA was analyzed.

Also interesting, but probably not at the core of information or systems theory, is the question whether relative information loss can be given an interpretation in thermodynamics, similar to absolute information loss – Landauer's principle shall be named here. Speaking about Landauer's principle, it is worth investigating where exactly, in, e.g., a rectifier, the one bit of information loss is converted to thermodynamical entropy – or if this principle applies at all. Closely related is also the connection between information loss and the Kolmogorov-Sinaï entropy rate in iterated function systems: Also this connection has not been investigated to its full extent, at least in the opinion of the author.

3 Information Loss Rate

The signal model and the mathematical preliminaries, the style of which is strongly influenced by Tatjana Petrov, were taken from [47] unless noted otherwise. The results of Section 3.2 were obtained in collaboration with Christoph Temmel, and partly published in [48, 58]; a part of its introduction is taken from a joint work with Tatjana Petrov, Heinz Köppl, and Gernot Kubin [47], cf. Section 4.3. Christoph Temmel devised the characterization for the lossless case and the current proof idea is due to him; the sufficient criteria for lumpability and lossless compression are due to the present author. In both cases, the results were obtained jointly. The results of Section 3.2.6 are mainly due to the present author, while the idea to use the adjacency matrix of a graph for the characterization was again from Christoph Temmel. Sections 3.3 and 3.4 have been derived by the present author, thanks to suggestions from Gernot Kubin; they have been published on arXiv [54]. The contents of Section 3.5 constitute part of [56] and [53]. Finally, [50] comprises Section 3.6, where again Gernot Kubin made many suggestions (to investigate finite-precision effects in digital filters).

3.1 What Is Information Loss for Stochastic Processes? – The Discrete Case

In extending information loss from random variables to stochastic processes, it is instructive to take up the water analogy from Section 2.1; indeed, the analogy is even more appropriate here. Observing the flow of the liquid (the stochastic process) through the pipes (the system), one can measure the amount of liquid lost *per time unit*. Depending on the flow rate (the process' entropy rate) and the viscosity of the liquid (the process' correlation or redundancy), the corrosion of the pipes will lead to a larger or smaller loss. As the loss rate of the liquid is just the difference between the flow rates at the input and the output of the pipe system, the *information loss rate* shall be defined likewise: as the difference between the *information rates* at the input and the output of the system under consideration.

This shall be made precise now. A *discrete-time, one-sided random process* $\mathbf{X}$ is a sequence of RVs $\{X_0, X_1, X_2, \dots\}$ defined on a common probability space. Each RV X_i shall take values from the same finite alphabet $\mathcal{X}$.

For a finite index set $\mathbb{I} \subset \mathbb{N}_0$, the joint PMF of $X_\mathbb{I} = \{X_i\}_{i\in\mathbb{I}}$ is

$$\forall z_\mathbb{I} \in \mathcal{X}^{\mathrm{card}(\mathbb{I})}\colon\ p_{X_\mathbb{I}}(z_\mathbb{I}) := \Pr(X_i = x_i, i \in \mathbb{I}). \tag{3.1}$$

In particular, for $\mathbb{I} = \{m, m+1, \dots, n\}$, $X_{\mathbb{I}}$ is abbreviated by $X_m^n := \{X_m, X_{m+1}, \dots, X_n\}$ and

$$p_{X_m^n}(z_m^n) := \Pr(X_m = z_m, \dots, X_n = z_n). \tag{3.2}$$

Along the same lines one obtains the marginal PMF p_{X_n} of X_n and the conditional PMF $p_{X_n|X_1^{n-1}}$ of X_n given its past, X_1^{n-1}, where the latter is assumed to be well-defined (i.e., the conditioning event has positive probability).

The random processes considered in this work are *stationary*, i.e., for an arbitrary, finite $\mathbb{I}$ the corresponding joint PMF is shift-invariant,

$$\forall k \in \mathbb{N}_0\colon p_{X_{\mathbb{I}}} = p_{X_{\mathbb{I}+k}}. \tag{3.3}$$

In particular, stationarity implies that the marginal distribution of X_k is equal for all k and shall be denoted as p_X. For simplicity, whenever a quantity depends only on the marginal distribution of the process, it shall be given as a function of an RV X with PMF p_X. In particular, due to stationarity,

$$\forall k \in \mathbb{N}_0\colon H(X_k) = H(X). \tag{3.4}$$

The information conveyed by a stationary, stochastic process per time unit is measured by its entropy rate [21, Thm. 4.2.1]

$$\overline{H}(\mathbf{X}) := \lim_{n\to\infty} \frac{1}{n} H(X_1^n) = \lim_{n\to\infty} H(X_n|X_1^{n-1}). \tag{3.5}$$

By the fact that conditioning reduces entropy, one always has $\overline{H}(\mathbf{X}) \leq H(X)$, with equality if $\mathbf{X}$ is a sequence of independent, identically distributed (iid) RVs. The difference between these two quantities leads to the definition of redundancy:

Definition 3.1 (Redundancy Rate)**.** The *redundancy rate* of a stationary stochastic process $\mathbf{X}$ is

$$\bar{R}(\mathbf{X}) := H(X) - \overline{H}(\mathbf{X}). \tag{3.6}$$

The redundancy rate is a measure of statistical dependence between the current sample and its past: For an iid process $\overline{H}(\mathbf{X}) = H(X)$ and $\bar{R}(\mathbf{X}) = 0$. Conversely, for a completely predictable process $\overline{H}(\mathbf{X}) = 0$ and $\bar{R}(\mathbf{X}) = H(X)$. In other words, the higher the redundancy rate, the lower the entropy rate and, thus, the less information is conveyed by the process in each time step.

Let $\mathbf{X}$ be the stochastic process at the input of a system described by a measurable function $g\colon \mathcal{X} \to \mathcal{Y}$. The output of the system is again a stochastic process $\mathbf{Y}$, whose n-th sample is $Y_n := g(X_n)$. From stationarity of $\mathbf{X}$ it follows that $\mathbf{Y}$ is stationary (and jointly stationary with $\mathbf{X}$). The information $\mathbf{Y}$ conveys in each time unit is measured by its entropy rate $\overline{H}(\mathbf{Y})$.

These preliminaries admit a definition of the information lost per time unit for discrete-time, discrete-valued stationary stochastic processes:

Definition 3.2 (Information Loss Rate, [161])**.** Let $\mathbf{X}$ be a stationary stochastic process with finite alphabet $\mathcal{X}$, and let $\mathbf{Y}$ be defined via $Y_n := g(X_n)$. The information loss induced by g in each time unit is

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) := \lim_{n\to\infty} \frac{1}{n} L(X_1^n \to Y_1^n) = \overline{H}(\mathbf{X}) - \overline{H}(\mathbf{Y}). \tag{3.7}$$

For the last equality, combine Definition 2.1 with the definition of entropy rate given in (3.5). Before proceeding, it is worth mentioning that this quantity was defined by Watanabe and Abraham in [161], albeit using non-standard notation. That above difference is non-negative

was observed by Pinsker (cf. [116, eq. (6.3.4)]), thus establishing a data processing inequality for stochastic processes. Another interesting result from [161] shall be reproduced here:

Lemma 3.1 (Upper Bound on the Information Loss Rate [161, Thm. 3])**.** *The information loss rate is bounded from above by the information loss, i.e.,*

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) \leq L(X \to Y) \tag{3.8}$$

where X is distributed according to the marginal distribution of $\mathbf{X}$.

Proof. The proof follows from the chain rule of entropy, conditioning, and stationarity:

$$\begin{aligned}
\bar{L}(\mathbf{X} \to \mathbf{Y}) &= \lim_{n\to\infty} \frac{1}{n} \left(H(X_1^n) - H(Y_1^n)\right) && (3.9)\\
&= \lim_{n\to\infty} \frac{1}{n} H(X_1^n | Y_1^n) && (3.10)\\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_1^{i-1}, Y_1^n) && (3.11)\\
&\leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | Y_i) && (3.12)\\
&= L(X \to Y) && (3.13)
\end{aligned}$$

□

It seems as if redundancy in the process can help prevent information from being lost, compared to the case where the process has no redundancy. This is intuitive, since part of the information is stored in the temporal structure of the process. However, one should not conclude that a process with a higher redundancy rate suffers from less information loss than a process with a lower redundancy rate: The redundancy has to be matched to the function in order to be effective. This parallels coding theory, where the code has to be matched to the channel in order to reduce the bit error rate. Example 18 shows that a process with higher redundancy can suffer from higher information loss; since the considered processes are Markov chains, the example is deferred to the end of Section 3.2.4.

3.2 Information Loss Rate for Markov Chains

This section starts with a particularly simple class of stochastic processes: Markov chains. Markov models are ubiquitous in scientific and engineering disciplines, for example, to understand chemical reaction systems, to perfprm speech recognition and to model data sources, or in Markov decision processes in automated control. The popularity of these models arises because the Markov property often renders model analysis tractable and their simulation efficient. However, the state space of a Markov model (i.e., its alphabet) is sometimes too large to permit simulation, even when harnessing today's computing power. Indeed, in stochastic modeling in computational biology [165], or in n-gram word models in speech recognition [102], dealing with the state space explosion is a major challenge. Also in control theory, particularly for nearly completely decomposable Markov chains, state space reduction is an important topic [5, 27].

Feeding the Markov chain through a function will be called a *lumping* here, referring to the fact that the non-injective function g *lumps* together states of the chain's transition graph. Assuming stationarity, the resulting (stationary) *lumped stochastic process* is also called a *functional hidden Markov model* [33]. One can transform an arbitrary hidden Markov model on finite state and observation spaces into this setting [33, Section IV.E]. In general, the lumped process loses the Markov property [68, 87] and has a lower entropy rate than the original Markov chain due to the aggregation of states [116, 161].

The present section investigates the structure of information-preserving lumpings of stationary Markov chains over a finite state space. The central result characterizes information preservation in two ways. First, by a structural property of the transition graph associated with the Markov chain. Second, by the growth of the cardinality of the realizable preimage of a realization of the lumped process. This analysis reveals a strong dichotomy between the preservation and loss case: In the former, the information loss of sequences of arbitrary length is finitely bounded, and the cardinality of the realizable preimage has finite growth a.s. In the latter, the information loss grows linearly in the sequence length and the cardinality of the realizable preimage grows exponentially a.s.

In particular, a positive transition matrix always implies information loss for a non-trivial lumping. For Markov chains with non-positive transition matrices, a sufficient condition for a lumping to preserve information is presented. Carlyle [16] investigated the representation of a finite-state stationary stochastic process as a lumping of a Markov chain on an at most countable state space; his representation fulfills the sufficient condition for information preservation.

Preserving the Markov property during lumping is highly desirable from a simulation point-of-view. It is shown that lumpings resulting in higher-order Markov chains are characterized by equality in natural entropic bounds on the entropy rate of the lumped process. The equality holding only for entropies depending on the lumped process is equivalent to *weak lumpability*, i.e., the lumped process is a higher-order Markov chain in the stationary setting. A second equality involving also entropies using the underlying Markov chain in the stationary case is equivalent to *strong lumpability*, i.e., the lumped process is a higher-order Markov chain *for every* initial distribution. This characterization is an information-theoretic complement to Gurvits & Ledoux's [68] linear algebraic approach to characterize lumpability.

Finally, a sufficient condition on a lumping of a Markov chain to preserve the entropy rate *and* be strongly k-lumpable is given. The condition is fulfilled on non-trivial lower-dimensional subspaces of the space of transition matrices[10]. The latter condition will be employed for lossless compression of letter n-gram models in Section 3.2.6.

A different notion of information loss is worth mentioning, namely the loss of information about the initial state X_0 (or its distribution) occuring over time. Lindqvist quantified the amount of information about X_0 contained in the observation of X_n without using information-theoretic methods in [98]. If instead of X_n only a function $Y_n := g(X_n)$ is observed, the information about X_0 is decreased further (although in some cases this additional loss decreases as a function of n) [99]. In case the *distribution* of X_0 can be inferred by the joint distribution of $Y_0, Y_1, Y_2, \ldots$, the lumping is called *g-observable* (see [68, Section 3] and the references therein). However, neither g-observability nor Lindqvist's results appear to have a direct correspondence in this work.

3.2.1 Preliminaries

Let $\mathbf{X}$ be an irreducible, aperiodic, time-homogeneous Markov chain on the finite state space $\mathcal{X}$. It has transition matrix $\mathbf{P} = \{P_{ij}\}$ with an invariant probability vector μ satisfying $\mu^T = \mu^T\mathbf{P}$. Assume that $\mathbf{X}$ is stationary, i.e., $X_0 \sim \mu$. The function g shall in this section be called the *lumping function*; it is non-trivial, i.e., $2 \leq \mathrm{card}(\mathcal{Y}) < \mathrm{card}(\mathcal{X})$. Without loss of generality, the function g is extended to $\mathcal{X}^n \to \mathcal{Y}^n$ coordinate-wise, for arbitrary $n \in \mathbb{N}$. The stationary stochastic process $\mathbf{Y}$ defined by $Y_n := g(X_n)$ shall be called the *lumped process*, and the total setup is referred to as the *lumping* $(\mathbf{P}, g)$.

The lumping of course induces an information loss rate $\bar{L}(\mathbf{X} \to \mathbf{Y})$. The main question is whether $\bar{L}(\mathbf{X} \to \mathbf{Y})$ is positive or zero (speaking of *information loss* or *information preservation*, respectively). Note that information preservation does not imply that the original process can be reconstructed from the lumped process (see Example 13).

[10] This does not conflict with Gurvits & Ledoux's [68] result that lumpings having higher-order Markov behavior are nowhere dense.

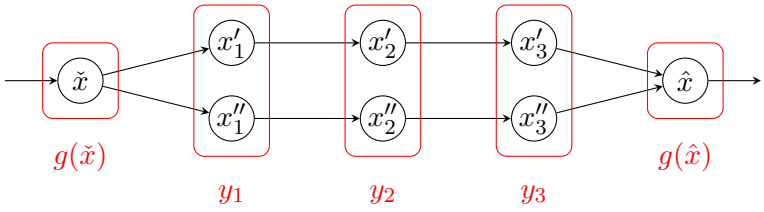


Figure 3.1: A section of trajectory space, with time running left-to-right. The two realizable length-5 trajectories $(\check{x}, x'_1, x'_2, x'_3, \hat{x})$ and $(\check{x}, x''_1, x''_2, x''_3, \hat{x})$ have the same lumped image $(g(\check{x}), y_1, y_2, y_3, g(\hat{x}))$. Thus the split-merge index $\mathcal{K} \leq 3$. Note that the lumped states $\{g(\check{x}), y_1, y_2, y_3, g(\hat{x})\}$ do not need to be distinct; e.g., it might be that $y_1 = y_2 = g(\hat{x})$.

The *transition graph* of the Markov chain $\mathbf{X}$ is the directed graph with vertex set $\mathcal{X}$ and an edge (x, x') iff $P_{xx'} > 0$. A length n trajectory $x_1^n \in \mathcal{X}^n$ is *realizable*, iff $\Pr(X_1^n = x_1^n) > 0$, equivalent to being a directed path in the transition graph. A key structural property of this graph is its *split-merge index* with respect to g:

$$\mathcal{K} := \inf \left\{ n \in \mathbb{N} \;\middle|\; \begin{array}{l} \exists \check{x}, \hat{x} \in \mathcal{X}, y_1^n \in \mathcal{Y}^n : \exists (x')_1^n, (x'')_1^n \in g^{-1}[y_1^n], (x'')_1^n \neq (x')_1^n : \\ \text{such that both } \begin{cases} \Pr(X_0 = \check{x}, X_1^n = (x')_1^n, X_{n+1} = \hat{x}) > 0 \\ \Pr(X_0 = \check{x}, X_1^n = (x'')_1^n, X_{n+1} = \hat{x}) > 0 \end{cases} \end{array} \right\} . \tag{3.14}$$

The split-merge index can be interpreted as follows: Take a pair of different, realizable trajectories with the same lumped image and common start- and endpoints (if such a pair exists). Then, $\mathcal{K}$ is the length of the shortest path on which the two trajectories differ in every coordinate. Thus $\mathcal{K} \in \mathbb{N} \cup \{\infty\}$. Figure 3.1 gives an example of $\mathcal{K} \leq 3$.

The *realizable preimage* of a lumped trajectory $y_1^n \in \mathcal{Y}^n$ are the realizable trajectories in its preimage, i.e.

$$R(y_1^n) := \{x_1^n \in g^{-1}[y_1^n] : x_1^n \text{ is realizable}\} . \tag{3.15}$$

The *preimage count of length* n of the lumping $(\mathbf{P}, g)$ is a random variable defined by the cardinality of the *realizable preimage* of a random lumped trajectory of length n:

$$T_n := \operatorname{card}(R(Y_1^n)) = \sum_{x_1^n \in g^{-1}[Y_1^n]} \lceil \Pr(X_1^n = x_1^n) \rceil , \tag{3.16}$$

where $[A] = 1$ if A is true and zero otherwise (Iverson bracket).

3.2.2 Equivalent Conditions for Information-Preservation

The first main result of this section is

Theorem 3.1.

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) > 0 \quad \Leftrightarrow \quad \mathcal{K} < \infty \quad \Leftrightarrow \quad \exists C > 1 : \Pr(\liminf_{n\to\infty} \sqrt[n]{T_n} \geq C) = 1 , \tag{3.17a}$$

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) = 0 \quad \Leftrightarrow \quad \mathcal{K} = \infty \quad \Leftrightarrow \quad \exists C < \infty : \sup_{n\to\infty} T_n \leq C . \tag{3.17b}$$

Proof. See Appendix B.1. □

The constants C in Theorem 3.1 are explicit functions of $(\mathbf{P}, g)$ and can be found in the proof. Likewise, an explicit lower bound for the information loss rate in case (3.17a) is stated in (B.15), implying that the information loss grows at least linearly in the sequence length.

Theorem 3.1 reveals a dichotomy in the information-theoretic behavior of the lumping. If $\mathcal{K}$ is infinite, then no split-merge situations as in Figure 3.1 occur. Thus, all finite trajectories of $\mathbf{X}$ can be reconstructed from their lumped image and knowledge of their endpoints. Therefore, the only information loss occurs at those endpoints and is finite. This yields uniform finite bounds on the conditional block entropies, i.e., the information loss, and the preimage count. If $\mathcal{K}$ is finite, then at least two different, realizable length-$(\mathcal{K}+2)$ trajectories of $\mathbf{X}$ with the same lumped image split and merge (see Figure 3.1). Such a split-merge leads to a finite information loss. The ergodic theorem ensures that this situation occurs linearly often in the block length, thus leading to a linear growth of the information loss. This in turn implies an information loss rate greater than zero. In particular, the information loss of a lumping can never exhibit sublinear (unbounded) growth.

If no split-merge situation occurs, then realisable trajectories with the same lumped image move in parallel and their number is constrained. This yields a uniform bound on the conditional block entropies for lengths smaller than $\mathcal{K}$:

Proposition 3.1.

$$\forall n: \quad n-2<\mathcal{K} \Rightarrow L(X_1^n \to Y_1^n) \leq 2\log(\mathrm{card}(\mathcal{X}) - \mathrm{card}(\mathcal{Y}) + 1)\,. \tag{3.18}$$

Proof. See Appendix B.1.1. □

An information-preserving lumping of course needs to have an output alphabet sufficiently large to represent the information of the process. This motivates the necessary condition presented in

Proposition 3.2. *An information-preserving lumping* $(\mathbf{P}, g)$ *satisfies*

$$\mathrm{card}(\mathcal{Y}) \geq \min_i d_i \tag{3.19}$$

where $d_i := \sum_{j=1}^{N} \lceil P_{ij} \rceil$ *is the out-degree of state* i*.*

Proof. See Appendix B.2. □

Since for a positive transition matrix $\mathbf{P}$ the out-degree of the associated transition graph is $\mathrm{card}(\mathcal{X})$ for all states, there is no (non-trivial) lumping which can preserve information:

Corollary 3.1. *If* $\mathbf{P}$ *is positive, i.e., all its entries are positive, then* $\bar{L}(\mathbf{X} \to \mathbf{Y}) > 0$*.*

Thus, information preserving lumpings must have sufficiently sparse transition matrices $\mathbf{P}$. Conversely, for a sufficiently sparse transition matrix $\mathbf{P}$ one can hope to find a more compact representation of the original Markov chain.

3.2.3 Excursion: k-lumpability of Markov Chains

Functions of Markov chains have been considered in the literature for a long time: Kemeny & Snell [87] coined the term *lumpability* for retaining the Markov property and presented necessary and sufficient conditions. Burke and Rosenblatt [15] analyzed the Markovity of a function of a Markov chain without requiring that the resulting process is a *homogeneous* Markov chain. They stated that the condition in [87] is sufficient for all stationary, regular Markov chains with arbitrary initial probability, but necessary only for reversible chains with an initial distribution equal to the invariant distribution. Also Rogers and Pitman worked on basic results on lumpability [124]. Higher-order lumpability, as it used in this section, has been analyzed by Gurvits and Ledoux [68], and they showed that the class of Markov chains being lumpable is nowhere dense.

Abdel-Moneim and Leysieffer [2] extended the work by Kememy and Snell [87] on weakly lumpable Markov chains, i.e., where the derived process is Markov only for a specific set of initial conditions. Their work was later extended and corrected by Rubino and Sericola [126].

Gilbert [59], Dharmadhikari [28], and Heller [71] took the reverse approach and analyzed under which conditions a stochastic process with a finite number of states can be represented as a function of a finite Markov chain.

Buchholz [14] investigated what can be inferred about the stationary and transient behavior of the underlying Markov chain by observing the derived chain. In particular, the conditions for lumpability are relaxed in the sense that the derived process is only nearly Markov and bounds on the error are presented.

The entropy rate of functions of a Markov chain was originally considered by Blackwell [12], who concluded that the obtained expression is an intrinsically complicated function. Approximations of the entropy rate were proposed and analyzed w.r.t. converge rates by Birch [11] and in [21].

The case of the lumped process retaining the Markov property is desirable from a computational and modelling point of view. However, in general, the lumped process $\mathbf{Y}$ does not possess the Markov property [68, 87]. Nevertheless, one may hope that the lumped process belongs to the larger and still desirable class of higher-order Markov chains.

Definition 3.3 (k-th order Markov Chain)**.** A stochastic process $\mathbf{Z}$ is a k-th order homogeneous Markov chain (short: $\mathbf{Z}$ is $\mathsf{HMC}(k)$), iff

$$\forall n, m \in \mathbb{N} \setminus \{1, \ldots, k-1\}, m \leq n, z_n \in \mathcal{Z}, z_{n-m}^{n-1} \in \mathcal{Z}^m : \\ \Pr(Z_n = z_n | Z_{n-m}^{n-1} = z_{n-m}^{n-1}) = \Pr(Z_n = z_n | Z_{n-k}^{n-1} = z_{n-k}^{n-1}) \,. \quad (3.20)$$

The entropy rate of a $\mathsf{HMC}(k)$ is as straightforward as one would expect:

Proposition 3.3. *Let $\mathbf{Z}$ be a stationary stochastic process on $\mathcal{Z}$. Then*

$$\mathbf{Z} \text{ is } \mathsf{HMC}(k) \Leftrightarrow \overline{H}(\mathbf{Z}) = H(Z_k | Z_0^{k-1}) \,. \quad (3.21)$$

Proof. See Appendix B.3. □

Definition 3.4 (k-lumpability; extension of [87, Def. 6.3.1])**.** A lumping $(\mathbf{P}, g)$ of a stationary Markov chain $\mathbf{X}$ is *weakly k-lumpable*, iff $\mathbf{Y}$ is $\mathsf{HMC}(k)$. It is *strongly k-lumpable*, iff this holds for each distribution of X_0 and iff the transition probabilities of $\mathbf{Y}$ are independent of this distribution.

A direct expression of the entropy rate of the lumped process $\mathbf{Y}$ was shown to be intrinsically complicated [12]. However, there are upper and lower bounds, which are asymptotically tight:

Lemma 3.2 ([21, Thm. 4.5.1, pp. 86])**.** *In the setup of this section, the following bounds on the entropy rate of $\mathbf{Y}$ hold:*

$$\forall n \in \mathbb{N} : \quad H(Y_n | Y_1^{n-1}, X_0) \leq \overline{H}(\mathbf{Y}) \leq H(Y_n | Y_0^{n-1}) \,. \quad (3.22)$$

In the stationary setting, equality on the r.h.s., for $n = k$, together with Proposition 3.3, implies that $\mathbf{Y}$ is $\mathsf{HMC}(k)$, i.e., $(\mathbf{P}, g)$ is weakly k-lumpable. If there is also equality on the l.h.s., for $n = k$, then knowledge of the distribution of X_0 delivers no additional information about Y_k. In other words, $\mathbf{Y}$ is $\mathsf{HMC}(k)$, independently of the starting distribution. In [48] the following result has been proved:

Theorem 3.2 (Information-Theoretic Characterization of Strong k-Lumpability)**.** *The following facts are equivalent:*

$$H(Y_k | Y_1^{k-1}, X_0) = H(Y_k | Y_0^{k-1}) \,, \quad (3.23a)$$

$$\mathbf{X} \text{ is strongly } k\text{-lumpable.} \quad (3.23b)$$

It is worth noting that (3.23a) is a condition only on the stationary setting, whereas (3.23b) is a statement about the lumping $(\mathbf{P}, g)$. Theorem 3.2 is thus an information-theoretic equivalent to Gurvits & Ledoux's characterization [68, Thms. 2 & 6]. Example 11 shows that weak k-lumpability alone is not sufficient for (3.23).

Example 11: taken from [87, pp. 139].
Consider the following transition matrix, where the lines divide lumped states:

$$\mathbf{P} := \left[\begin{array}{c|ccc} 1/4 & 1/16 & 3/16 & 1/2 \\ \hline 0 & 1/12 & 1/12 & 5/6 \\ 0 & 1/12 & 1/12 & 5/6 \\ 7/8 & 1/32 & 3/32 & 0 \end{array}\right] .$$

This lumping (and its time-reversal) is weakly 1-lumpable [87, pp. 139]. However, with an accuracy of 0.0001,

$$0.5588 = H(Y_1|X_0) < \overline{H}(\mathbf{Y}) = H(Y_1|Y_0) = 0.9061$$

and

$$0.9048 = H(Y_0|X_1) < \overline{H}(\hat{\mathbf{Y}}) = H(Y_0|Y_1) = 0.9061 ,$$

where $\hat{\mathbf{Y}}$ is the time-reversed process. Hence, weak k-lumpability alone does not imply (3.23).

3.2.4 Sufficient Conditions for Information-Preservation and k-Markovity

This section presents easy-to-check sufficient conditions on a lumping to have vanishing information loss rate and satisfy strong k-lumpability. These conditions depend only on the transition graph and the lumping function g. The conditions are only stated in a "forward form"; applying the conditions to the transition graph with the direction of the edges reversed yields a set of mirrored sufficient conditions, which are omitted.

The first sufficient condition is

Definition 3.5 (Single Entry Property (SE))**.** A lumping $(\mathbf{P}, g)$ is *single entry* (short: SE), iff

$$\forall y \in \mathcal{Y}, x \in \mathcal{X} : \exists x' := x'(x,y) \in g^{-1}[y] : \forall x'' \in g^{-1}[y] \setminus \{x'\} : \\ \Pr(X_1 = x''|X_0 = x) = 0 , \quad (3.24)$$

i.e., there is *at most* one edge from a given state x into the preimage $g^{-1}[y]$.

The SE lumpings have a vanishing information loss rate:

Proposition 3.4. *If* $(\mathbf{P}, g)$ *is* SE*, then* $\overline{L}(\mathbf{X} \to \mathbf{Y}) = 0$.

Proof.

$$H(Y_k|X_{k-1}) \le H(Y_k|Y_1^{k-1}, X_0) \le \overline{H}(\mathbf{Y}) \le \overline{H}(\mathbf{X}) = H(X_k|X_{k-1}) ,$$

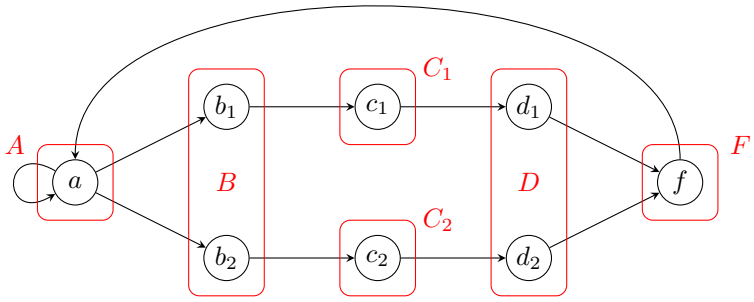
where the first and the second inequality are due to [21, Thm. 4.5.1, pp. 86] (cf. Lemma 3.1) and the third inequality is due to data processing. The SE property implies that, $p_{X_k,X_{k-1}}$-almost surely,

$$p_{Y_k|X_{k-1}}(y|x) = p_{X_k|X_{k-1}}(x'(x,y)|x) ,$$

where $x'(x, y)$ is from (3.24). Thus, the outer terms in the above chain of inequalities coincide, yielding $\overline{H}(\mathbf{Y}) = \overline{H}(\mathbf{X})$. □

The following and Example 17 show that SE is not necessary for entropy rate preservation:

Example 12: SE is not necessary for entropy rate preservation.



The transition graph of a Markov chain with the lumping represented by red boxes. The lumping is neither SE (violated by transitions from a into B) nor its mirror condition for the time-reversed process (violated by transitions from D to f). On the other hand, the existence of the uniquely represented states C_1 and C_2 allows to distinguish between the trajectories (a, b_1, c_1, d_1, f) and (a, b_2, c_2, d_2, f). Therefore the lumping preserves the entropy rate. Furthermore, this lumping is weakly 1-lumpable and strongly 2-lumpable. Hence it shows that SE is neither necessary for entropy rate preservation nor for weak k-lumpability. This also applies to SFS(k) (see Definition 3.6), which is a subclass of SE.

Corollary 3.2. *If $(\mathbf{P}, g)$ is SE and weakly k-lumpable, then it is strongly k-lumpable.*

Proof. The proof of Proposition 3.4 shows that SE implies equality on the l.h.s. of (3.22), for all n. Weak k-lumpability implies equality on the r.h.s. of (3.22) for $n = k$. Therefore, Theorem 3.2 applies. □

There is one interesting result about SE-lumpings, which is similar to the one in Proposition 3.2:

Proposition 3.5. *An SE-lumping satisfies*

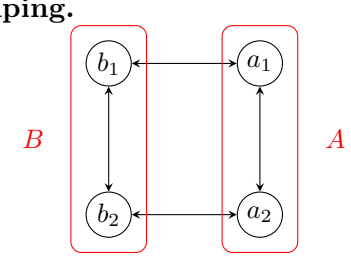
$$\operatorname{card}(\mathcal{Y}) \geq \max_i d_i \tag{3.25}$$

where d_i is the out-degree of state i.

In particular, this implies that a transition matrix with at least one positive row does not admit an SE-lumping.

Proof. Evaluate the rows of $\mathbf{P}$ separately: All states x_2 accessible from state x_1 are characterized by $P_{x_1,x_2} > 0$. Any two states accessible from x_1 cannot be merged, since this would contradict Definition 3.5. Thus, all states accessible from x_1 must have different images, implying $\operatorname{card}(\mathcal{Y}) \geq d_{x_1}$. The result follows by considering all states x_1. □

Example 13: An SE lumping.



The transition graph of a Markov chain with the lumping represented by red boxes. The lumping is SE and thus preserves the entropy rate. Furthermore, if all transitions have probability 1/2, it is strongly 1-lumpable and thus $H(Y_1|X_0) = H(Y_1|Y_0)$ (see Theorem 3.2). However, observing an arbitrarily long trajectory of the lumped process does not determine the current preimage state. Whence $(\mathbf{P}, g)$ is not $\mathsf{SFS}(k)$ (see below), for every k. Therefore, $\mathsf{SFS}(k)$ is neither necessary for entropy rate preservation nor for strong lumpability.

That a lumping can be SE without being strongly lumpable, or strongly lumpable without being SE is shown in Examples 15 and 16, respectively.

Interestingly, Proposition 3.5 has a source coding theorem [21, Thm. 5.3.1, p. 111 or Thm. 5.4.2, p. 114] as a consequence. Clearly, the entropy rate of $\mathbf{X}$ satisfies $\overline{H}(\mathbf{X}) \leq \log(\max_i d_i)$. The lumping corresponds to a fixed-length coding, so the expected codelength per symbol is $\log \operatorname{card}(\mathcal{Y})$. Proposition 3.5 therefore implies that the expected code length cannot be smaller than the entropy rate of $\mathbf{X}$.

The restriction to symbol-by-symbol encodings is, in general, not optimal for the compression of Markov sources. Conditions for optimality would be, e.g., equality in Proposition 3.5 and equally probable transitions from each state. In Example 19 these conditions are fulfilled.

The second sufficient condition is:

Definition 3.6 (Single Forward k-Sequence Property ($\mathsf{SFS}(k)$))**.** For $k \geq 2$, a lumping $(\mathbf{P}, g)$ has the *single forward k-sequence property* (short: $\mathsf{SFS}(k)$), iff

$$\begin{aligned}
&\forall y_1^{k-1} \in \mathcal{Y}^{k-1}, y \in \mathcal{Y} : \exists {x'}_1^{k-1} \in g^{-1}[y_1^{k-1}] : \\
&\forall x \in g^{-1}[y], x_1^{k-1} \in g^{-1}[y_1^{k-1}] \setminus \left\{ {x'}_1^{k-1} \right\} : \\
&\qquad\qquad \Pr(X_1^{k-1} = x_1^{k-1} | Y_1^{k-1} = y_1^{k-1}, X_0 = x) = 0\,, \quad (3.26)
\end{aligned}$$

i.e., there is *at most* one realizable sequence in the preimage $g^{-1}[y_1^{k-1}]$ starting in $g^{-1}[y]$.

The $\mathsf{SFS}(k)$ property implies entropy rate preservation and strong k-lumpability:

Proposition 3.6. *If $(\mathbf{P}, g)$ is $\mathsf{SFS}(k)$, then it is strongly k-lumpable and* SE*.*

Sketch of proof. First, show that SE contains $\mathsf{SFS}(k)$, which implies preservation of entropy. To this end

$$p_{X_1^{k-1}|Y_1^{k-1},X_0}(x_1^{k-1}|y_1^{k-1}, x_0) = \begin{cases} 1 & \text{if } x_1^{k-1} = {x'}_1^{k-1} \text{ from (3.26)}\,, \\ 0 & \text{else.} \end{cases}$$

If SE does not hold, then there exist states $y^\star \in \mathcal{Y}$ and $x^\star \in \mathcal{X}$ such that at least two states in $g^{-1}[y^\star]$ can be reached from $x^\star$. Assume that $y_{k-2}^{k-1} = (g(x^\star), y^\star)$. Thus, observing y_0^{k-1} does not determine x_1^{k-1}.

To show that $\mathsf{SFS}(k)$ implies strong k-lumpability, a technical lemma from [48] is required. The proof proceeds by showing that $\mathsf{SFS}(k)$ eliminates the dependence on the initial distribution, and that the condition of Theorem 3.2 holds. □

The $\mathsf{SFS}(k)$ is a property of the combinatorial structure of the transition matrix $\mathbf{P}$, i.e., it only depends on the location of its non-zero entries, and can be checked with a complexity of $\mathcal{O}(N^k)$. It also has practical significance: Besides preserving, if possible, the information of the original model, those lumpings which possess the Markov property of any (low) order are preferable from a computational perspective. The corresponding conditions for the more desirable first-order Markov output, not necessarily information-preserving, are too restrictive in most scenarios, cf. [87, Sec. 6.3]).

The next result investigates a cascade of lumpings. Below, identify a function with the partition it induces on its domain. Let $h\colon \mathcal{X} \to \mathcal{Z}$, $f\colon \mathcal{Z} \to \mathcal{Y}$, and $g := h \circ f$ be $\mathcal{X} \to \mathcal{Y}$. Clearly, (the partition induced by) g is coarser than (the partition induced by) h because of the intermediate application of f. In other words, h is a *refinement* of g.

Proposition 3.7 (SFS(k) & Refinements)**.** *If a lumping* $(\mathbf{P}, g)$ *is* SFS(k)*, then so is* $(\mathbf{P}, h)$*, for all refinements* h *of* g*.*

Proof. The proof follows from contradiction: Assume $(\mathbf{P}, h)$ violates SFS(k). Then there exist $z_1^{k-1} \in \mathcal{Z}^{k-1}, z \in \mathcal{Z}$ such that there exist two distinct $x'^{k-1}_1, x''^{k-1}_1 \in h^{-1}[z_1^{k-1}]$ and two, not necessarily distinct $x', x'' \in h^{-1}[z]$ such that

$$\Pr(X_2^k = x'^{k-1}_1 | Z_2^k = z_1^{k-1}, X_1 = x') > 0 \tag{3.27}$$

and

$$\Pr(X_2^k = x''^{k-1}_1 | Z_2^k = z_1^{k-1}, X_1 = x'') > 0\,. \tag{3.28}$$

In other words, there are two different sequences x'^{k-1}_1, x''^{k-1}_1 accessible from either the same ($x' = x''$) or from different ($x' \neq x''$) starting states.

Now take $y_1^{k-1} = f(z_1^{k-1})$ and $y = f(z)$. Since h is a refinement of g, we have $h^{-1}[z_1^{k-1}] \subseteq g^{-1}[y_1^{k-1}]$ and $h^{-1}[z] \subseteq g^{-1}[y]$. As a consequence, $x'^{k-1}_1, x''^{k-1}_1 \in g^{-1}[y_1^n]$ and $x', x'' \in g^{-1}[y]$, implying that $(\mathbf{P}, g)$ violates SFS(k). This proves

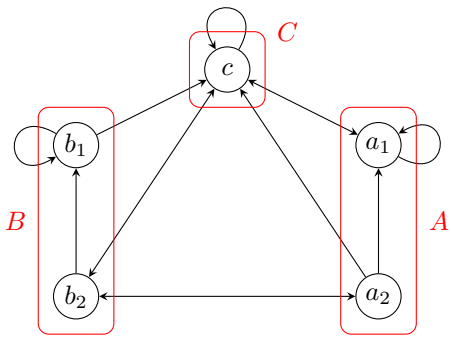
$$(\mathbf{P}, h) \text{ violates } \mathsf{SFS}(k) \Rightarrow (\mathbf{P}, g) \text{ violates } \mathsf{SFS}(k)\,. \tag{3.29}$$

The negation of these statements completes the proof. □

A refinement does not increase the loss of information, so information-preservation is preserved under refinements. In contrast, a refinement of a lumping yielding a kth-order Markov process $\mathbf{Y}$ need not possess that property; the lumping to a single state has the Markov property, while a refinement of it generally has not.

The following example satisfies the SFS(2) property:

Example 14: An Example satisfying SFS(2).

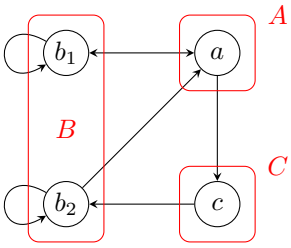


The transition graph of a Markov chain with the lumping represented by red boxes. After at most two steps one either enters a new lumped state at a unique original state or is circling in either b_1 or a_1. Hence, this lumping is SFS(2). The space of Markov chains with this transition graph contains the interior of a multi-simplex in $\mathbb{R}^{13}$, parametrized by 8 parameters (13 directed edges minus 5 nodes).

Example 12 shows that SFS(2) is neither necessary for weak 1-lumpability nor for entropy rate preservation, while Example 13 demonstrates that SFS(2) is neither necessary for SE nor

for strong 1-lumpability. The following example shows that SE does neither imply SFS(k) nor strong k-lumpability, for every k. Finally, Example 16 gives a strongly 2-lumpable lumping which is not SFS(2).

Example 15: SE but not SFS(k).



The transition graph of a Markov chain with the lumping represented by red boxes. The lumping is SE. The loops at b_1 and b_2 imply that the lumped process is not HMC(k), for every k and regardless of the distribution of X_0. This is easily seen by the inability to differentiate between n consecutive b_1's and n consecutive b_2's. When starting in B and as long as $\Pr(X_1 = a|X_0 = b_1) \neq \Pr(X_1 = a|X_0 = b_2)$ and $\Pr(X_1 = b_1|X_0 = b_1) \neq \Pr(X_1 = b_2|X_0 = b_2)$, this long sequence of Bs prevents determining the probability of entering A. Thus this is neither SFS(k) nor strongly k-lumpable, for each k.

Example 16: A strongly 2-lumpable chain.

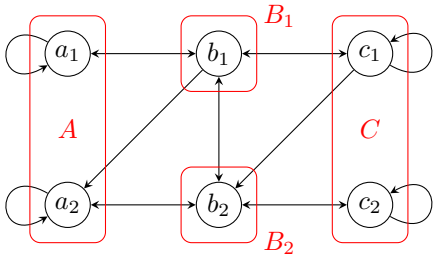
Consider the following transition matrix, where the lines divide lumped states:

$$\mathbf{P} := \left[\begin{array}{cc|cc} 0.6 & 0.4 & 0 & 0 \\ 0.3 & 0.2 & 0.1 & 0.4 \\ \hline 0.2 & 0.05 & 0.375 & 0.375 \\ 0.2 & 0.05 & 0.375 & 0.375 \end{array}\right] .$$

This lumping is strongly 2-lumpable and satisfies (3.23a) with $\overline{H}(\mathbf{Y}) = H(Y_2|Y_{[0,1]}) = H(Y_2|Y_1, X_0) = 0.733$ (with an accuracy of 0.001). However, it does not preserve entropy: $1.480 = \overline{H}(\mathbf{X}) > \overline{H}(\mathbf{Y})$, whence it is neither SE nor SFS(2).

Examples 12 and 17 have vanishing information loss rate without satisfying any of the sufficient conditions.

Example 17: An information-preserving lumping.



The transition graph of a Markov chain with the lumping represented by red boxes. The lumping preserves the entropy rate without satisfying SE or its mirror condition for the reversed transition graph. The loops at a_1 and a_2, and those at c_1 and c_2 prevent that the lumped process is HMC(k), for every k, given that the loop probabilities are different.

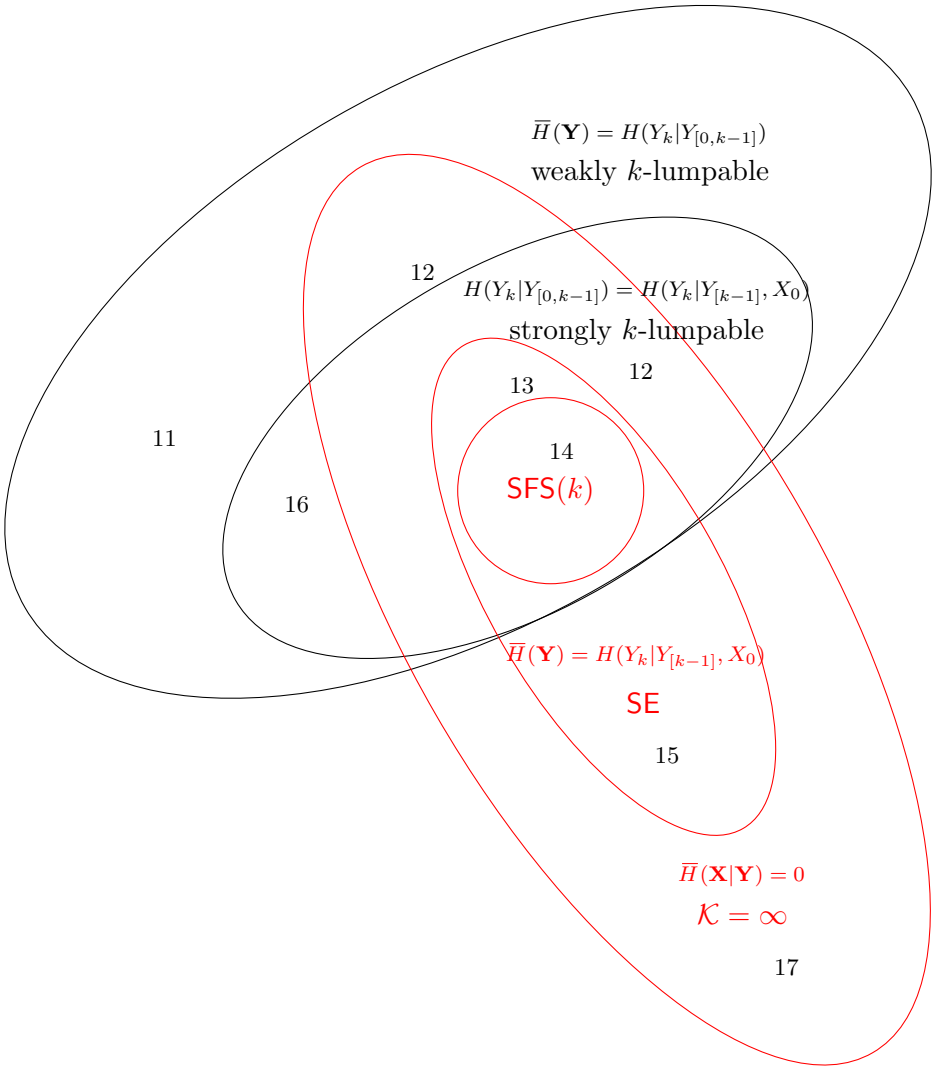


Figure 3.2: Venn diagram of the relation between different classes. Numbers indicate the examples in this work.

Finally, the following example shows that higher redundancy not necessarily reduces the information loss rate, connecting this section to the discussion after Lemma 3.1.

Example 18: Redundancy not always helps.
Consider the following two Markov chains $\mathbf{X}$ and $\mathbf{X}'$, with transition matrices

$$\mathbf{P} = \left[\begin{array}{cc|cc} 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right] \quad \text{and} \quad \mathbf{P}' = \left[\begin{array}{cc|cc} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ \hline 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{array}\right]. \tag{3.30}$$

Both of these Markov chains exhibit a uniform invariant distribution, thus $H(X) = H(X') = 2$. The entropy rates compute to $\overline{H}(\mathbf{X}) = 0.5$ and $\overline{H}(\mathbf{X}') = 1$, leaving the process $\mathbf{X}$ with the higher redundancy rate.

Collapsing the states as indicated yields first-order Markov chains $\mathbf{Y}$ and $\mathbf{Y}'$. In both cases, the information loss is equal to $H(X|Y) = H(X'|Y') = 1$. However, the process $\mathbf{Y}$ alternates between the two lumps, resulting in a vanishing output entropy rate ($\overline{H}(\mathbf{Y}) = 0$).

Surprisingly, the entropy rate of $\mathbf{Y}'$ is equal to the entropy rate of $\mathbf{X}'$, resulting in zero information loss rate. The process with the higher redundancy rate suffered from a higher information loss rate.

Looking at the redundancy rates, one observes that $\bar{R}(\mathbf{X}) = 1.5$ and $\bar{R}(\mathbf{X}') = 1$ while $\bar{R}(\mathbf{Y}) = 1$ and $\bar{R}(\mathbf{Y}') = 0$. Thus, there exist nonlinearities which destroy the full information, as for the process $\mathbf{X}$, while there also exist nonlinearities which destroy the full redundancy, as for the process $\mathbf{X}'$.

3.2.5 An Algorithm for Obtaining SFS(2)-Lumpings

Clearly, losing the Markov property during lumping states is not desirable from a computational point-of-view, as Markov chains are especially simple to simulate numerically and also permit parameter estimation from data without suffering too much from the curse of dimensionality. The sufficient conditions presented in Definition 3.6 are thus highly relevant from a practical perspective. Among these, especially the class SFS(2) is interesting: Second-order Markov chains are a good trade-off between model complexity and restrictiveness of the conditions. Moreover, whether a lumping $(\mathbf{P}, g)$ satisfies the SFS(2) property can be checked directly from the transition matrix $\mathbf{P}$; SFS(2)-lumpings have the property that, for all y_1, $y_2 \in \mathcal{Y}$, from within a set $g^{-1}[y_1]$ at most one element in the set $g^{-1}[y_2]$ is accessible:

$$\forall y_1, y_2 \in \mathcal{Y} : \exists! x_2' \in g^{-1}[y_2] : \forall x_1 \in g^{-1}[y_1], x_2 \in g^{-1}[y_2] \setminus \{x_2'\} : P_{x_1,x_2} = 0. \quad (3.31)$$

An algorithm listing all SFS(2)-lumpings, or SFS(2)-partitions, for a given transition matrix $\mathbf{P}$ has to check the SFS(2)-property for all partitions of $\mathcal{X}$ into at least $\max_i d_i$ non-empty sets. The number of these partitions can be calculated from the Stirling numbers of the second kind [13, Thm. 8.2.5] and is typically too large to allow an exhaustive search. As it is shown below, Proposition 3.7 can be used to reduce the search space.

Starting from the trivial partition with $N := \mathrm{card}(\mathcal{X})$ elements, evaluate all possible merges of two states, i.e., all possible partitions with $N - 1$ sets, of which there exist $N(N-1)/2$. Out of these, drop those from the list which do not possess the SFS(2)-property. The remaining set of *admissible pairs* is a central element of the algorithm.

Now proceed iteratively: To generate all candidate partitions with $N - i$ sets, perform all admissible pair-wise merges on all SFS(2)-partitions with $N - i + 1$ sets. An admissible pair-wise merge is a merge of two sets of a partition, where either set contains one element of an admissible pair. From the resulting partitions one drops those violating SFS(2) before performing the next iteration. These partitions can be dropped, since they are common refinements of all partitions obtained by merging some of their elements, i.e., of their descendents; hence, if these partitions violate SFS(2), by Proposition 3.7 all their descendents will as well. Since the algorithm generates some partitions multiple times (see Example 19 below), in every iteration all duplicates are removed. The algorithm is presented in Algorithm 1; note that a similar algorithm has been proposed in [163] to determine all lumping functions w.r.t. which a given Markov chain is strongly lumpable.

The application of Proposition 3.7 reduces the number of partitions to be searched (see Fig. 3.3). If the number of admissible pairs is small compared to $N(N-1)/2$, then this reduction is significant. Inefficiencies in the algorithm caused by multiple considerations of the same partitions could be alleviated by adapting the classical algorithms for the partition generation problem [34, 137].

The actual choice of one of the obtained SFS(2)-partitions for model order reduction requires additional model-specific considerations: A possible criterion could be maximum compression (i.e., smallest entropy of the marginal distribution), the smallest cardinality of the output alphabet, or semantic properties of the partition.

Algorithm 1 Algorithm for listing all SFS(2)-lumpings

```
1: function ListLumpings(P)
2:     admPairs ← GetAdmissiblePairs(P)
3:     Lumpings(1) ← merge(admPairs)                    ▷ Convert pairs to functions
4:     n ← 1
5:     while notEmpty(Lumpings(n)) do
6:         n ← n + 1
7:         Lumpings(n) ← [ ]
8:         for h ∈ Lumpings(n − 1) do
9:             for {i1, i2} ∈ admPairs do
10:                g ← h
11:                g(h^{-1}(h(i2))) ← g(i1)             ▷ i1 and i2 have same image.
12:                if g is SFS(2) then
13:                    Lumpings(n) ← [Lumpings(n); g]
14:                end if
15:            end for
16:        end for
17:        Remove duplicates from Lumpings
18:    end while
19:    return Lumpings
20: end function
21: function GetAdmissiblePairs(P)
22:    Pairs ← [ ]
23:    N ← dim(P)
24:    for i1 = 1 : N − 1 do
25:        for i2 = i1 : N do
26:            f ← merge(i1, i2)                        ▷ f merges i1 and i2
27:            if f is SFS(2) then
28:                Pairs ← [Pairs; {i1, i2}]
29:            end if
30:        end for
31:    end for
32:    return Pairs
33: end function
```

In practice, if N is large and if the number of admissible pairs is of the same order as $N(N-1)/2$, listing all SFS(2)-partitions might be computationally expensive. One can trade optimality for speed by greedy selection of the best h in line 8, given a specific criterion, or by evaluating only a (random) subset of admissible pairs in line 9 (cf. Fig. 3.3).

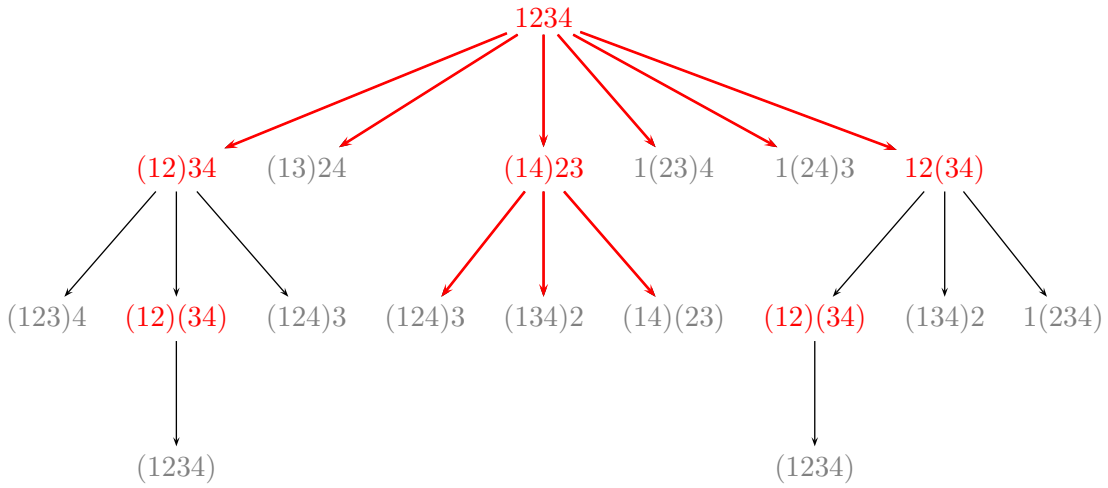
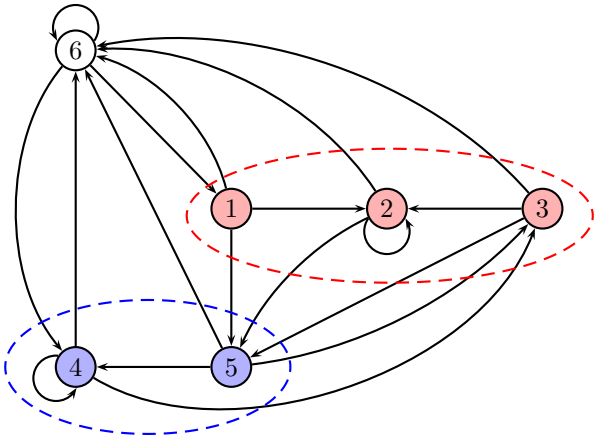


Figure 3.3: Illustration of how Algorithm 1 searches for partitions satisfying the SFS(2)*-property; here, the set* $\mathcal{X} = \{1, 2, 3, 4\}$ *is partitioned, red partitions satisfy the* SFS(2)*-property. Partitions are indicated using short-hand notation, e.g.,* $(ij)kl := \{i, j\}, \{k\}, \{l\}$*. It can be seen that the algorithm prunes the tree whenever the desired property is violated (Proposition 3.7). Moreover, if the computational complexity is still too large, instead of an exhaustive search a greedy one can be conducted, continuing the search at each level of the tree only from the most promising partition satisfying the* SFS(2)*-property (see red arrows).*

Example 19: A Toy Example.



The algorithm is illustrated at the hand of a small example. Consider the six-state Markov chain with transition graph depicted above, with adjacency matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} . \tag{3.32}$$

Since all states have out-degree $d_i = 3$, lumpings to at least $M = 3$ states are considered. The indicated lumping satisfies the SFS(2)-property.

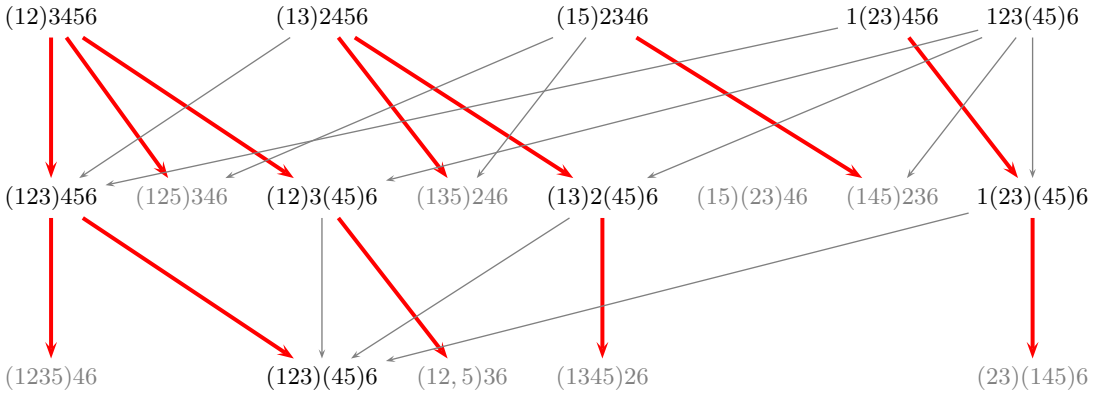
During initialization all 15 possible pair-wise merges are evaluated. Of these, all pairs where both members are accessible from the same state are excluded, i.e., $\{2, 5\}$, $\{2, 6\}$, $\{5, 6\}$, $\{3, 4\}$, $\{3, 6\}$, $\{4, 6\}$, $\{1, 4\}$, and $\{1, 6\}$. Furthermore, $\{2, 4\}$ and $\{3, 5\}$ are excluded too; the former because both states have self-loops, the latter because both states are connected in either direction. Only five pairs are admissible.

One admissible pair is $\{1,2\}$, i.e., the function h merging $\{1,2\}$ and, thus, inducing the partition $\mathcal{Z}_5 = \{\{1,2\},\{3\},\{4\},\{5\},\{6\}\}$, satisfies SFS(2). With this h enter the algorithm in the innermost loop (Algorithm 1, line 9). The algorithm performs pair-wise merges according to the five admissible pairs and obtains the following merged sets: $\{1,2\}$, $\{1,2,3\}$, $\{1,2,5\}$, $\{\{1,2\},\{4,5\}\}$; the first is a (trivial) duplicate (by performing a pair-wise merge according to $\{1,2\}$) and the second is obtained twice (by pairing $\{1,2\}$ with $\{1,3\}$ and $\{2,3\}$). Only $\{1,2,5\}$ violates SFS(2). The functions merging $\{1,2,3\}$ and $\{\{1,2\},\{4,5\}\}$ are added to the list of lumping functions to four states, and the procedure is repeated for a different admissible pair.

For the next iteration, fix h such that it induces the partition $\mathcal{Z}_4 = \{\{1,2,3\},\{4\},\{5\},\{6\}\}$. The five admissible pairs yield the merged sets $\{1,2,3\}$, a duplicate which is obtained three times, $\{1,2,3,5\}$, which violates SFS(2), and $\{\{1,2,3\},\{4,5\}\}$, which is the solution depicted above. The algorithm terminates now, since every pair-wise merge of $\mathcal{Z}_3 = \mathcal{Y} = \{\{1,2,3\},\{4,5\},\{6\}\}$ either violates SFS(2) or is a duplicate. The list of all SFS(2)-lumpings found by the algorithm is given in the table below.

M	Partition $\mathcal{Z}_M$
6	$\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}$
5	$\{1,2\},\{3\},\{4\},\{5\},\{6\}$
	$\{1,3\},\{2\},\{4\},\{5\},\{6\}$
	$\{1,5\},\{2\},\{3\},\{4\},\{6\}$
	$\{1\},\{2,3\},\{3\},\{4\},\{6\}$
	$\{1\},\{2\},\{3\},\{4,5\},\{6\}$
4	$\{1,2,3\},\{4\},\{5\},\{6\}$
	$\{1,2\},\{3\},\{4,5\},\{6\}$
	$\{1,3\},\{2\},\{4,5\},\{6\}$
	$\{1\},\{2,3\},\{4,5\},\{6\}$
3	$\{1,2,3\},\{4,5\},\{6\}$

Furthermore, the figure below shows the derivation of the lumping by Algorithm 1. The first row shows all admissible pairs, the algorithm runs through all rows (top to bottom) by merging according to the admissible pairs (left to right). Bold, red arrows indicate newly generated partitions, gray arrows indicate that this partition was already found and is thus removed as a duplicate. Gray partitions violate the SFS(2)-property. Partitions are again indicated using short-hand notation (see Fig. 3.3).



Interestingly, if for the given transition graph all transition probabilities are set to 1/3, it can be shown that the lumped process $\mathbf{Y}$ is a sequence of iid random variables. This observation does not conflict with the SFS(2)-property, since an iid process is Markov of every order. Furthermore, since the lumping is information-preserving and since the redundancy of $\mathbf{Y}$ vanishes, one has $H(Y) = \overline{H}(\mathbf{Y})$. The compression achieved by this simple symbol-by-symbol encoding is optimal for this example.

3.2.6 Application: n-gram models

Algorithm 1 is now applied to a bi-gram[11] letter model. Commonly used in natural language processing [102, Ch. 6], n-grams (of which bi-grams are a special case) are $(n-1)$th-order Markov models for the occurrence of letters or words. From a set of training data the relative frequency of the (co-)occurrence of letters or words is determined, yielding the maximum likelihood estimate of their (conditional) probabilities. In practice, for large n, even large training data cannot contain all possible sequences, so the n-gram model will contain a considerable amount of zero transition probabilities. Since this would lead to problems in, e.g., a speech recognition system, those entries are increased by a small constant to *smooth* the model, for example using Laplace's law [102, pp. 202].

Since, by Proposition 3.2, an information-preserving lumping is more efficient for a sparse transition matrix, in this example the maximum likelihood estimates of the model parameters are used instead. To this end, a bi-gram letter model of F. Scott Fitzgerald's *"The Great Gatsby"*, a text containing roughly 270000 letters, was trained. To reduce the alphabet size and, thus, the run-time of the algorithm, all numbers and all upper case letters are replaced by '`#`' and the corresponding lower case letters, respectively. Punctuations were left unchanged, yielding a total alphabet size of $\operatorname{card}(\mathcal{X}) = 41$. The adjacency matrix of the bi-gram model can be seen in Fig. 3.4; the maximum out-degree of the Markov chain is 37.

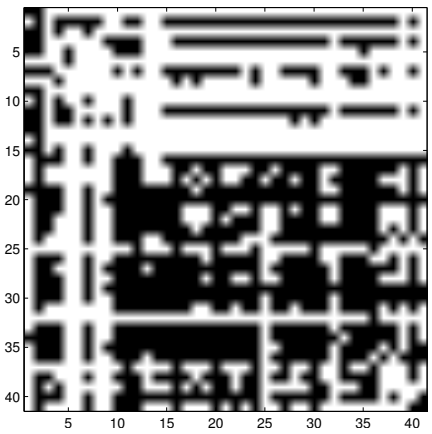


Figure 3.4: The adjacency matrix of the simplified bi-gram model of "The Great Gatsby". The first two states are line break (LB) and space (' '), followed by punctuations. The block in the lower right corner indicates interactions of letters and punctuation following letters.

Of the 820 possible merges only 21 are admissible. Furthermore, there are 129, 246, and 90 SFS(2)-lumpings to sets of cardinalities 39, 38, and 37, respectively. Only two triples can be merged, namely {LB, '`$`', '`x`'} and {LB, '`(`', '`x`'}, where LB denotes the line break. Among the more notable pair-wise merges are {'`(`','`)`'}, {'`(`','`z`'}, and the merges of '`#`' with colon, semicolon, and exclamation mark. Especially the first is intuitive, since parentheses can be replaced by, e.g., '`|`' while preserving the meaning of the symbol[12].

Finally, the lumping yielding maximum compression was determined, i.e., the one for which $H(Y)$ is minimal. This lumping, merging {LB, '`$`', '`x`'}, {'`!`', '`#`'}, and {'`(`', '`,`'}, decreases the entropy from 4.3100 to 4.3044 bits. These entropies roughly correspond to the 4.03 bits derived for Shannon's first-order model, which contains only 27 symbols [21, p. 170].

[11] Shannon used bi-grams, or *digrams* as he called them, as a second-order approximation of the English language [138]. A few decades prior to that, Markov used bigrams to analyze the sequence of vowels and consonants in the Russian text *"Eugen Onegin"* [69].

[12] Whether the symbol initiates or terminates a parenthetic expression is determined by whether the symbol is preceded or succeeded by a blank space. Unless parenthetic expressions are nested, also simple counting distinguishes between initiation and termination.

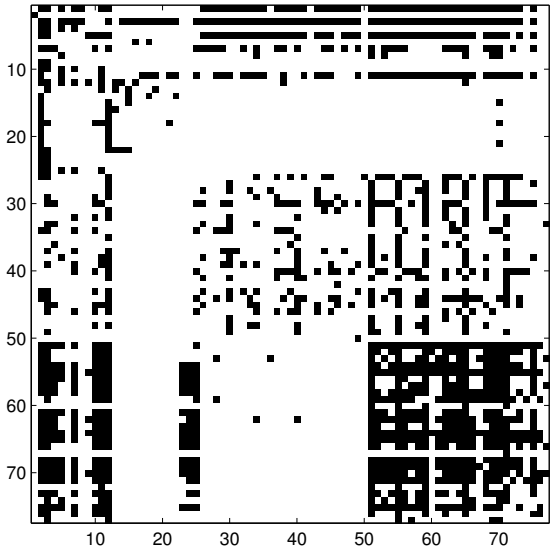


Figure 3.5: The adjacency matrix of the bi-gram model of "The Great Gatsby". The first two states are line break (LB) and space (' '), followed by punctuations, numbers, upper case and lower case letters. The four blocks in the lower right corner indicate the four types of interactions between upper case and lower case letters.

Without preprocessing, an exhaustive search is significantly more expensive: With an alphabet size of $\operatorname{card}(\mathcal{X}) = 77$ (see Fig. 3.5 for the adjacency matrix) and 357 admissible pairs, the first iteration of the algorithm already checks roughly 60000 *distinct* partitions with 75 elements, most of them satisfying the $\mathsf{SFS}(2)$-condition. Proposition 3.5 yields $\operatorname{card}(\mathcal{Y}) \geq 65$.

A modified algorithm only retains the best (in terms of the entropy of the marginal distribution) partition in each iteration. This greedy heuristic achieves a compression from 4.5706 to 4.4596 bits with $\operatorname{card}(\mathcal{Y}) = 66$. It is presently not known if in this example $\operatorname{card}(\mathcal{Y}) = 65$ can be attained, even by an exhaustive search.

Finally, a tri-gram model of the same text was trained without preprocessing the alphabet. The resulting second-order Markov chain was lifted to a first-order Markov chain on $\mathcal{X}^2$ (states are now letter tuples). Eliminating all non-occurring tuples reduces the alphabet size to $\operatorname{card}(\tilde{\mathcal{X}}) = 1347$. Proposition 3.5 yields $\operatorname{card}(\mathcal{Y}) \geq 53$. There are roughly 900000 admissible pairs. A further modified algorithm, considering only 10 random admissible pairs in each iteration, achieves a compression from 8.0285 to 7.1781 bits. The algorithm terminates at $\operatorname{card}(\mathcal{Y}) = 579$. The question, whether a reduction to $\operatorname{card}(\mathcal{Y}) < 77$ is possible (thus replacing a second-order Markov model by one on a smaller state space) remains open.

3.3 Information Loss Rate for Continuous Processes

Extending the notion of an information loss rate to general processes is not trivial; it is even more difficult than generalizing the concept of information loss from discrete to continuous RVs. However, in this and the following section this generalization will be made, making similar restrictions as in Chapter 2. In particular, in this section the focus will be on piecewise bijective functions (PBFs) and continuous-valued, one-dimensional, discrete-time stationary stochastic processes. Thus, the process $\mathbf{X}$ is such that its n-th sample, X_n, has a distribution $P_{X_n} = P_X$ independent of n which is supported on $\mathcal{X} \subset \mathbb{R}$ and satisfies $P_X \ll \lambda$. Generally, for every finite set $\mathbb{I} \subset \mathbb{N}_0$, the joint distribution $P_{X_\mathbb{I}} \ll \lambda^{\operatorname{card}(\mathbb{I})}$. This (almost-surely) guarantees the existence of all joint (conditional) PDFs. Specifically, for all n, the joint PDF $f_{X_1^n}$ and the conditional

PDF $f_{X_n|X_1^{n-1}}$ exist.

Additionally, it shall be assumed that the joint differential entropy of an arbitrary collection of RVs exists and is finite, and that, thus, also the differential entropy rate [114, Thm. 14.7]

$$\bar{h}(\mathbf{X}) := \lim_{n\to\infty} h(X_n|X_1^{n-1}) = \lim_{n\to\infty} \frac{1}{n} h(X_1^n) \tag{3.33}$$

exists and is finite.

The system under consideration shall be described by a PBF satisfying Definition 2.3. The output process $\mathbf{Y}$ is obtained by feeding $\mathbf{X}$ through this system, i.e., $Y_n := g(X_n)$, and is jointly stationary with $\mathbf{X}$. The information loss occurring in this system *per sample* shall be quantified by

Definition 3.7 (Information Loss Rate)**.** Let $\mathbf{X}$ be a continuous-valued, one-dimensional, discrete-time stationary stochastic process, and let $\mathbf{Y}$ be a jointly stationary process defined by $Y_n := g(X_n)$. The information loss rate induced by g is

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) := \lim_{n\to\infty} \frac{1}{n} L(X_1^n \to Y_1^n) \tag{3.34}$$

where $L(\cdot \to \cdot)$ follows Definition 2.2.

It can be shown that many properties of information loss for PBFs can be generalized also to the information loss rate. Most surprisingly, the information loss rate allows a simple expression in terms of the differential entropy rates of the input and output processes:

Proposition 3.8 (Information Loss Rate and Differential Entropy Rate)**.** *The information loss rate induced by feeding a stationary stochastic process through a PBF is*

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) = \bar{h}(\mathbf{X}) - \bar{h}(\mathbf{Y}) + \mathbb{E}\left(\log|g'(X)|\right). \tag{3.35}$$

Proof. For the proof note that the RVs $\{X_1, \ldots, X_n\}$ can be interpreted as a single, n-dimensional RV; similarily, one 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 by the elements $g'(x_i)$. With Theorem 2.2,

$$\begin{aligned} L(X_1^n \to Y_1^n) &= h(X_1^n) - h(Y_1^n) + \mathbb{E}\left(\log\left|\prod_{i=1}^n g'(X_i)\right|\right) \\ &= h(X_1^n) - h(Y_1^n) + n\mathbb{E}\left(\log\left|g'(X)\right|\right) \end{aligned} \tag{3.36}$$

where the first line is because the determinant of a diagonal matrix is the product of its diagonal elements, and where stationarity of $\mathbf{X}$ yields the second line. Dividing by n and taking the limit completes the proof. □

Proposition 2.2 shows 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 carries over to the information loss rate as well:

Proposition 3.9 (Cascade of Systems)**.** *Consider two functions g: $\mathcal{X} \to \mathcal{Y}$ and h: $\mathcal{Y} \to \mathcal{Z}$ and a cascade of systems implementing these functions. Let $Y_n := g(X_n)$ and $Z_n := h(Y_n)$, where X_n are samples from a continuous-valued stationary stochastic process $\mathbf{X}$. The information loss rate induced by this cascade, or equivalently, by the system implementing the composition $(h \circ g)(\cdot) = h(g(\cdot))$ is given as the sum of the individual information loss rates:*

$$\bar{L}(\mathbf{X} \to \mathbf{Z}) = \bar{L}(\mathbf{X} \to \mathbf{Y}) + \bar{L}(\mathbf{Y} \to \mathbf{Z}). \tag{3.37}$$

Proof. The proof follows from the fact that the cascade is described by the function $h \circ g$, and that

$$\mathbb{E}\left(\log|(h\circ g)'(X)|\right) = \mathbb{E}\left(\log|g'(X)h'(g(X))|\right) = \mathbb{E}\left(\log|g'(X)|\right) + \mathbb{E}\left(\log|h'(Y)|\right). \quad (3.38)$$

□

Finally, an extension of Proposition 2.3 is possible: To this end, let the stationary stochastic process $\mathbf{W}$ be given according to Definition 2.4, applied to each sample pair (X_n, W_n).

Proposition 3.10. *Let* $\mathbf{W}$ *be a stationary stochastic process defined by* $W_n := i$ *if* $X_n \in \mathcal{X}_i$. *Then,*

$$\overline{L}(\mathbf{X} \to \mathbf{Y}) = \overline{H}(\mathbf{W}|\mathbf{Y}). \quad (3.39)$$

Proof. Again, treat X_1^n as an n-dimensional RV; g^n induces a partition of its domain $\mathcal{X}^n$, which is equivalent to the n-fold product of the partition $\{\mathcal{X}_i\}$. Letting $\tilde{W}$ be the RV obtained by quantizing X_1^n according to this partition, it is easy to see that W_1^n is equivalent to $\tilde{W}$. Thus, with Proposition 2.3,

$$H(X_1^n|Y_1^n) = H(\tilde{W}|Y_1^n) = H(W_1^n|Y_1^n) \quad (3.40)$$

for all n. The proof is completed with $\overline{H}(\mathbf{W}|\mathbf{Y}) := \lim_{n\to\infty} \frac{1}{n} H(W_1^n|Y_1^n)$. □

3.3.1 Upper Bounds on the Information Loss Rate

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, and thus extends Lemma 3.1 to continuous-valued processes.

Proposition 3.11 (Loss ≥ Loss Rate)**.** *Let* $\mathbf{X}$ *be a stationary stochastic process and* X *an RV distributed according to the process' 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* $\mathbf{X}$ *through* g*, i.e.,*

$$\overline{L}(\mathbf{X} \to \mathbf{Y}) \leq L(X \to Y). \quad (3.41)$$

Proof. The inequality holds trivially if $L(X \to Y) = \infty$. The rest of the proof follows along the same lines as in Lemma 3.1. □

Clearly, this bound is tight whenever the input process $\mathbf{X}$ is an iid process. Moreover, it is trivially tight whenever the function is bijective, i.e., when $L(X \to Y) = 0$. Example 20 renders this bound tight in a more 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 [161] for stationary stochastic processes with finite alphabets.

The next bound extends Corollary 2.3, 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 3.12 (Upper Bound)**.** *Let* $\mathbf{W}$ *be as in Proposition 3.10. Then,*

$$\overline{L}(\mathbf{X} \to \mathbf{Y}) \leq \overline{H}(\mathbf{W}). \quad (3.42)$$

The proof follows from the fact that conditioning reduces entropy.

For the case that the input process is a stationary Markov process, i.e., if $f_{X_n|X_1^{n-1}} = f_{X_n|X_{n-1}} = f_{X_2|X_1}$ $P_{X_1^{n-1}}$-a.s. for all n, an additional, sharper, upper bound can be presented:

Proposition 3.13 (Upper Bound for Markovian $\mathbf{X}$)**.** *Let* $\mathbf{X}$ *be a stationary Markov process, and let* $\mathbf{W}$ *be as in Proposition 3.10. Then, for finite* $L(X \to Y)$*,*

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) \leq H(W_2|X_1). \tag{3.43}$$

Proof. Applying the chain rule, Markovity of $\mathbf{X}$, and the fact that conditioning reduces entropy yields

$$H(X_1^n|Y_1^n) \leq H(X_1|Y_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}, Y_i). \tag{3.44}$$

With stationarity and with $H(X_1|Y_1) = L(X \to Y) < \infty$ one can continue

$$\begin{aligned}
\bar{L}(\mathbf{X} \to \mathbf{Y}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1^n|Y_1^n) && (3.45)\\
&\leq H(X_2|X_1, Y_2) && (3.46)\\
&\overset{(a)}{=} H(W_2|X_1, Y_2) && (3.47)\\
&\leq H(W_2|X_1) && (3.48)
\end{aligned}$$

where (a) holds due to Proposition 2.3 since, for all $x \in \mathcal{X}$, $H(X_2|Y_2, X_1 = x) = H(W_2|Y_2, X_1 = x)$. The last inequality is due to conditioning [21, Thm. 2.6.5] and completes the proof. □

That the bound is sharper than the one of Proposition 3.12 follows from observing that

$$H(W_n|X_{n-1}) = \lim_{n\to\infty} H(W_n|X_1^{n-1}) \leq \lim_{n\to\infty} H(W_n|W_1^{n-1}) = \overline{H}(\mathbf{W}). \tag{3.49}$$

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.

3.3.2 Excursion: Lumpability of Continuous 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 (see Section 3.2), there exist conditions on the function and/or the chain such that the output is Markov. While most results are given for finite Markov chains (e.g., [15, 124]) relatively little is known in the general case of an uncountable alphabet (see [125] for an exception). This section is a small contribution to this field of research by presenting sufficient conditions for lumpability of continuous-valued Markov processes.

Let $f_{X_n|X_1^{n-1}} = f_{X_n|X_{n-1}} = f_{X_2|X_1}$ for all n, i.e., let $\mathbf{X}$ be a stationary Markov process.

Proposition 3.14. *If*

$$\forall y_1^2 \in \mathcal{Y}^2 : \forall x \in g^{-1}[y_1] : f_{Y_2,X_1}(y_2, x) > 0 \Rightarrow f_{Y_2|X_1}(y_2|x) = f_{Y_2|Y_1}(y_2|y_1) \tag{3.50}$$

then $\mathbf{X}$ *is lumpable w.r.t.* g*, i.e.,* $\mathbf{Y}$ *is Markov.*

Proof. See Appendix B.4. □

As a corollary, the conditions on the function g, the marginal distribution f_X, and the conditional distribution $f_{X_2|X_1}$ can be made explicit. By adding a further condition, Proposition 3.13 becomes tight:

Corollary 3.3. *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*

$$\sum_{x_2 \in g^{-1}[y_2]} \frac{f_{X_2|X_1}(x_2|x)}{|g'(x_2)|} = \sum_{x_2 \in g^{-1}[y_2]} \frac{f_{X_2|X_1}(x_2|x')}{|g'(x_2)|} \tag{3.51}$$

then the condition of Proposition 3.14 is fulfilled and $\mathbf{Y}$ is Markov.

Let, additionally, for all $y \in \mathcal{Y}$, all x within the support $\mathcal{X}$ of f_X, and all w, w' with $\Pr(W_2 = w|X_1 = x) > 0$ and $\Pr(W_2 = w'|X_1 = x) > 0$

$$\frac{f_{X_2|X_1}(g_w^{-1}(y_2)|x)}{|g'(g_w^{-1}(y_2))|} = \frac{f_{X_2|X_1}(g_{w'}^{-1}(y_2)|x)}{|g'(g_{w'}^{-1}(y_2))|} \tag{3.52a}$$

and

$$\Pr(W_2 = w'|X_1 = x) = \Pr(W_2 = w|X_1 = x). \tag{3.52b}$$

Then, the bound of Proposition 3.13 holds with equality.

Proof. See Appendix B.5. □

Example 20: An Example for Tightness of the Bounds.

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

$$f_{X_2|X_1}(x_2|x_1) =: f_M(x_2|x_1) = \frac{1}{2} \begin{cases} \mathbb{I}_{[1,2)}(x_2) + \mathbb{I}_{[3,4)}(x_2), & x_1 \in [0,1) \cup [2,3) \\ \mathbb{I}_{[0,1)}(x_2) + \mathbb{I}_{[2,3)}(x_2), & x_1 \in [1,2) \cup [3,4) \end{cases} \tag{3.53}$$

where $\mathbb{I}_A(x) = 1$ iff $x \in A$. The stationary distribution is the uniform distribution on $[0, 4)$, thus, $h(X) = \log 4 = 2$ and $\bar{h}(\mathbf{X}) = 1$.

The system maps the interval $\mathcal{X}_2 = [2, 4)$ onto the interval $\mathcal{X}_1 = [0, 2)$, i.e.,

$$g(x) = \begin{cases} x, & x \in [0,2) \\ x - 2, & x \in [2,4) \end{cases} \tag{3.54}$$

which yields the conditional distribution of Y_2 given X_1,

$$f_{Y_2|X_1}(y_2|x_1) = \begin{cases} \mathbb{I}_{[1,2)}(y_2), & x_1 \in [0,1) \cup [2,3) \\ \mathbb{I}_{[0,1)}(y_2), & x_1 \in [1,2) \cup [3,4) \end{cases}. \tag{3.55}$$

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

The output process $\mathbf{Y}$ 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, (3.51) holds.

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 (3.52b) and renders the upper bound from Proposition 3.13 as

$$H(W_2|X_1) = 1. \tag{3.56}$$

The bound can be shown to be tight, since also (3.52a) 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,

$$1 = L(X \to Y) \geq \bar{L}(\mathbf{X} \to \mathbf{Y}) = H(W_2|X_1) = 1. \tag{3.57}$$

This is an example for tightness not only of Proposition 3.13 but also of Proposition 3.11. Interestingly, neither is the function information-preserving, nor is the input process $\mathbf{X}$ iid. Consequently, one can interpret this example as a worst-case, where redundancy is not matched to the system (the "channel"), failing to alleviate its adverse effects.

3.3.3 Example: AR(1)-Process in a Rectifier

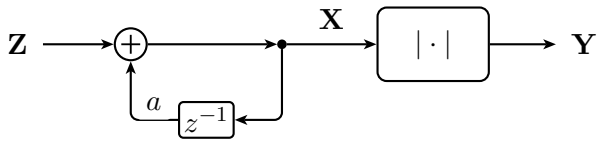


Figure 3.6: AR(1)-process with magnitude function. The input $\mathbf{Z}$ *is a sequence of iid Gaussian RVs with zero mean and variance* σ^2*; thus, the process* $\mathbf{X}$ *is Gaussian with zero mean and variance* $\sigma^2/(1-a^2)$*. The process generator filter is a first-order all-pole filter with a single pole at* a*.*

In this example a first-order, zero-mean, Gaussian auto-regressive process $\mathbf{X}$ is fed through a magnitude function (see Fig. 3.6). Let the AR process be generated by the following difference equation:

$$X_n = aX_{n-1} + Z_n \tag{3.58}$$

where $a \in (0, 1)$ and where Z_n are samples drawn independently from a Gaussian distribution with zero mean and variance σ^2. It follows immediately that the process $\mathbf{X}$ is also zero mean and has variance $\sigma_X^2 = \sigma^2/(1-a^2)$ [111, Ex. 6.11]. Let $\mathbf{Y}$ be defined by $Y_n := |X_n|$.

For the sake of brevity let $\phi(\mu, \sigma^2; x)$ denote the PDF of a Gaussian RV with mean μ and variance σ^2, evaluated at x. Thus,

$$f_X(x) = \phi(0, \sigma_X^2; x) \tag{3.59}$$

and

$$f_{X_2|X_1}(x_2|x_1) = \phi(ax_1, \sigma^2; x_2). \tag{3.60}$$

It follows that (3.51) is satisfied with $|g'(x)| \equiv 1$ and since $\phi(ax_1, \sigma^2; x_2) = \phi(-ax_1, \sigma^2; -x_2)$,

$$\sum_{x_2 \in g^{-1}[y_2]} \frac{f_{X_2|X_1}(x_2|y_1)}{|g'(x_2)|} = \phi(ay_1, \sigma^2; y_2) + \phi(ay_1, \sigma^2; -y_2) \tag{3.61}$$

$$= \phi(-ay_1, \sigma^2; -y_2) + \phi(-ay_1, \sigma^2; y_2) \tag{3.62}$$

$$= \sum_{x_2 \in g^{-1}[y_2]} \frac{f_{X_2|X_1}(x_2|-y_1)}{|g'(x_2)|}. \tag{3.63}$$

As a consequence, the output process $\mathbf{Y}$ is Markov.

As the information loss rate for this example cannot be expressed in closed form, numerical experiments were made. Rewriting, e.g., the lower bound on the information loss rate

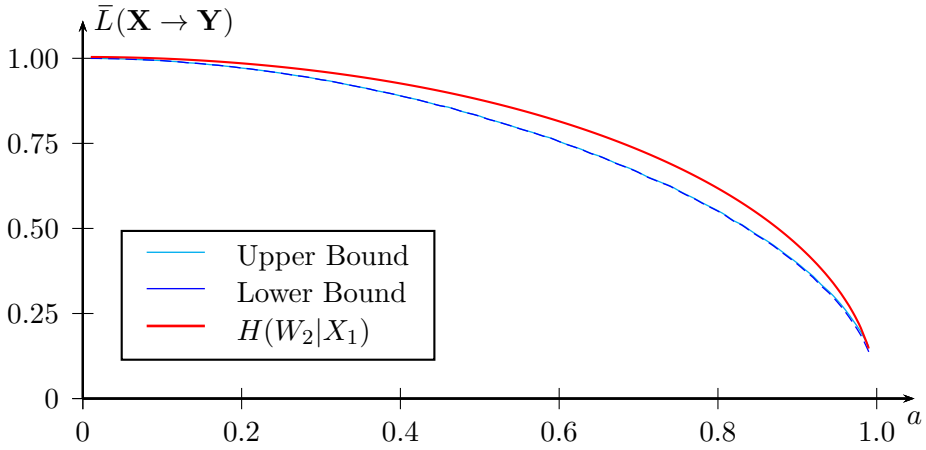


Figure 3.7: Information Loss Rate of an AR(1)-process $\mathbf{X}$ in a magnitude function as a function of the pole a of the process generator difference equation. A larger pole, leading to a higher redundancy of $\mathbf{X}$, reduces the information loss rate.

(cf. Lemma B.1) as

$$\begin{aligned}\bar{L}(\mathbf{X} \to \mathbf{Y}) &\geq h(X_2|X_1) - h(Y_2|X_1) + \mathbb{E}\left(\log|g'(X)|\right) && (3.64)\\ &= h(X) - I(X_1;X_2) - h(Y) + I(X_1;Y_2) + \mathbb{E}\left(\log|g'(X)|\right) && (3.65)\\ &= L(X \to Y) - I(X_1;X_2) + I(X_1;Y_2) && (3.66)\end{aligned}$$

admits employing the histogram-based mutual information estimation from [106] together with $L(X \to Y) = 1$, as shown in Example 3. The upper bound $H(W_2|X_1)$ from Proposition 3.13 was computed using numerical integration. In Fig. 3.7 one can see that the first-order upper and lower bounds on the information loss rate from Lemma B.1 in the proof of Proposition 3.14 are indistinguishable, which suggests that the output process is Markov. Moreover, it can be seen that a higher value for the magnitude a of the pole leads to a smaller information loss rate. This can be explained by the fact that the redundancy[13] of the process $\mathbf{X}$ increases with increasing a, which helps preventing information loss.

3.3.4 Example: Cyclic Random Walk in a Rectifier

Consider a cyclic random walk on a subset $\mathcal{X} = [-M, M]$ 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.,

$$f_M(x_2|x_1) := f_{X_2|X_1}(x_2|x_1) = \begin{cases} \frac{1}{2a}, & \text{if } d(x_2, x_1) \leq a \\ 0, & \text{else} \end{cases} \tag{3.67}$$

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 $\mathbf{X}$ is the uniform distribution[14], i.e., $f_X(x) = \frac{1}{2M}$ for all $x \in [-M, M]$ and zero otherwise. The function g 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 (3.51) is fulfilled, and that thus $\mathbf{Y}$ is Markov. Moreover, $\bar{h}(\mathbf{Y}) = h(Y_2|X_1)$, and the information loss rate reads with

[13] The 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 σ_X^2, while the latter remains constant and equal to $h(Z)$ (cf. [32]).

[14] The discrete-valued equivalent is a Markov chain with a doubly stochastic transition matrix, for which it is known that the stationary distribution is the uniform distribution [114, p. 732].

Proposition 3.8

$$\begin{aligned}\bar{L}(\mathbf{X}\to\mathbf{Y}) &= \bar{h}(\mathbf{X}) - \bar{h}(\mathbf{Y}) + \mathbb{E}\left(\log|g'(X)|\right) && (3.68)\\ &= h(X_2|X_1) - h(Y_2|X_1) && (3.69)\\ &= \int_{-M}^{M}\int_{-M}^{M} f_X(x_1) f_M(x_2|x_1) \log\frac{f_{Y_2|X_1}(|x_2||x_1)}{f_M(x_2|x_1)} dx_2 dx_1 && (3.70)\\ &= \int_{-M}^{M}\int_{-M}^{M} \frac{f_M(x_2|x_1)}{2M} \log\left(1 + \frac{f_M(-x_2|x_1)}{f_M(x_2|x_1)}\right) dx_2 dx_1. && (3.71)\end{aligned}$$

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,

$$\bar{L}(\mathbf{X}\to\mathbf{Y}) = \frac{4a}{2M}\int_0^M \int_{-M}^{M} f_M(x_2|x_1) f_M(-x_2|x_1) dx_2 dx_1 \tag{3.72}$$

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

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. 3.8).

Finally, evaluate the upper bound from Proposition 3.13: Letting $\mathcal{X}_1 = [-M, 0)$ and $\mathcal{X}_2 = [0, M]$ and abbreviating $p(1|x) := \Pr(W_2 = 1|X_1 = x)$ one gets

$$p(1|x) = \begin{cases} \frac{a-M-x}{2a}, & -M \le x < -M+a \\ 0, & -M+a \le x < -a \\ \frac{x+a}{2a}, & -a \le x < a \\ 1, & a \le x < M-a \\ \frac{a+M-x}{2a}, & M-a \le x < M \end{cases} \tag{3.73}$$

if $M > 2a$ and

$$p(1|x) = \begin{cases} \frac{a-M-x}{2a}, & -M \le x < -a \\ \frac{2a-M}{2a}, & -a \le x < -M+a \\ \frac{x+a}{2a}, & -M+a \le x < M-a \\ \frac{M}{2a}, & M-a \le x < a \\ \frac{a+M-x}{2a}, & a \le x < M \end{cases} \tag{3.74}$$

if $M \le 2a$. (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

$$H(W_2|X_1) = \begin{cases} \frac{a}{M\ln 2}, & M > 2a \\ \frac{M-a}{M\ln 2} + \log\frac{2a}{M}, & M \le 2a \end{cases}. \tag{3.75}$$

The analytic result for the information loss rate and the bound, numerically validated using the same procedure as in Section 3.3.3, are depicted in Fig. 3.9.

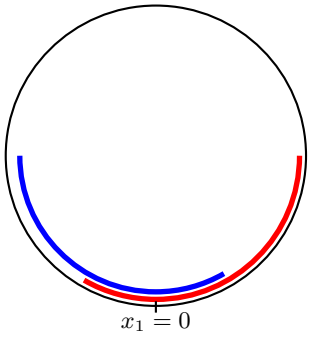


Figure 3.8: Interpreting the information loss rate of a cyclic random walk in a magnitude function. The depicted scenario corresponds to $a = M/3$.

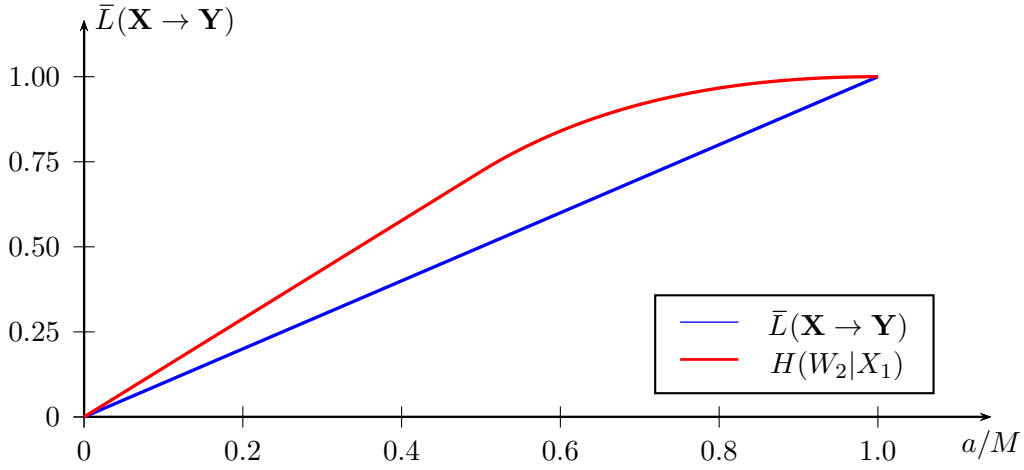


Figure 3.9: Information loss rate of a cyclic random walk $\mathbf{X}$ on $[-M, M]$ in a magnitude function as a function of the support $[-a, a]$ of the uniform input PDF.

3.4 Relative Information Loss Rate for Continuous Processes

As it was already shown in Section 2.4, not all systems can be described by piecewise bijective functions satisfying Definition 2.3. To analyze the information processing characteristics of a broader class of systems, the notion of relative information loss, capturing the *percentage* of information available at the input lost in the system, shall be extended to stochastic processes:

Definition 3.8 (Relative Information Loss Rate)**.** The relative information loss rate is

$$l(\mathbf{X} \to \mathbf{Y}) := \lim_{n\to\infty} l(X_1^n \to Y_1^n) = \lim_{n\to\infty} \lim_{k\to\infty} \frac{H((\hat{X}^{(k)})_1^n | Y_1^n)}{H((\hat{X}^{(k)})_1^n)} \tag{3.76}$$

whenever the limit exists.

The first result is an extension of Proposition 2.9, suggesting that if a system destroys a positive fraction of an infinite-information rate process, then the loss rate needs to be infinite as well:

Proposition 3.15 (Positive Relative Loss Rate leads to Infinite Absolute Loss Rate)**.** *Let $\mathbf{X}$ be such that $\overline{H}(\mathbf{X}) = \infty$ and let $l(\mathbf{X} \to \mathbf{Y}) > 0$. Then, $\bar{L}(\mathbf{X} \to \mathbf{Y}) = \infty$.*

Proof. Since, by assumption, $\overline{H}(\mathbf{X}) = \infty$, it follows that for all n, $H(X_1^n) = \infty$. Moreover, since $l(\mathbf{X} \to \mathbf{Y}) > 0$, for all $\epsilon > 0$ there exists an n_0 such that $l(X_1^n \to Y_1^n) \geq \epsilon$ for all $n \geq n_0$. With

Proposition 2.9 it follows that $L(X_1^n \to Y_1^n) = \infty$. Since this holds for all $n \geq n_0$, this completes the proof. □

For the results following below, let us again assume that the stationary process $\mathbf{X}$ is such that all joint and conditional PDFs exist, cf. Section 3.3. Thus, $d(X_1^n) = nd(X) = n$, for all n. Clearly, if $\mathbf{X}$ is iid, so is $\mathbf{Y}$, and $l(\mathbf{X} \to \mathbf{Y}) = l(X \to Y)$. In general, from stationarity, existence of the joint PDFs, and Proposition 2.11 the following holds for all n:

$$l(X_1^n \to Y_1^n) \leq \frac{1}{n} \sum_{i=1}^{n} l(X_i \to Y_i) = l(X \to Y) \tag{3.77}$$

Therefore one can show[15] that $l(\mathbf{X} \to \mathbf{Y}) \leq l(X \to Y)$, complementing Proposition 3.11. In many cases, this inequality is tight, as presented in the following

Proposition 3.16 (Redundancy won't help)**.** *Let $\mathbf{X}$ be a stationary stochastic process and X an RV distributed according to the process' marginal distribution. Let further g be defined on a finite partition $\{\mathcal{X}_i\}$ of $\mathcal{X}$ into non-empty sets as in Definition 2.3, where g_i is either bi-Lipschitz or constant (i.e., $g_i(x) = c_i$ for all $x \in \mathcal{X}_i$). Then,*

$$l(\mathbf{X} \to \mathbf{Y}) = l(X \to Y) = P_X(\mathcal{X}_c) \tag{3.78}$$

where $\mathcal{X}_c$ is the union of all elements $\mathcal{X}_i$ of the partition on which g is constant.

Proof. See Appendix B.6 □

Indeed, the author conjectures 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, cf. Example 1) and system blocks for multirate signal processing. The latter application is discussed in the next section.

3.5 Application: Multirate Signal Processing & Sampling

Consider the system depicted in Fig. 3.10, i.e., a decimation system consisting of a linear filter H and an M-fold downsampler. The statements below will make it clear that no linear filter H – regardless whether it is stable and causal or not – can reduce the amount of information lost in such a system. While this is surprising in itself (in general, pre-processing a signal should be able to reduce the amount of information lost in a consecutive system; cf. channel coding), it suggests that *all linear filters are equivalent in this sense*: Whether an ideal low-pass filter is employed or no filter at all, the relative information loss rate remains unchanged. This in some sense parallels the analysis of PCA in Section 2.6. Again, the reason for this seemingly counter-intuitive behavior is that every bit of input information is treated equally. As soon as one assumes a portion of the input process being *relevant*, the counter-intuitivity is removed and filtering *does* make sense (cf. Section 5.2).

The stochastic processes dealt with here satisfy

Assumption 1. $\mathbf{Z}$ is stationary, has finite marginal differential entropy $h(Z_n) = h(Z)$, finite Shannon entropy of the quantized RV $\lfloor Z_n \rfloor$, and finite differential entropy rate $\bar{h}(\mathbf{Z})$.

Lemma 3.3 (Finite differential entropy (rate) and information dimension)**.** *Let $\mathbf{X}$ be a stationary stochastic process satisfying Assumption 1. Then, for every finite set $\mathbb{J}$, $d(X_{\mathbb{J}}) = \operatorname{card}(\mathbb{J})$.*

[15] Note that also Watanabe and Abraham defined the *fractional information loss* for stochastic processes on finite alphabets [161]; for these types of processes, however, the relative information loss rate can be smaller *or* larger than the relative information loss.

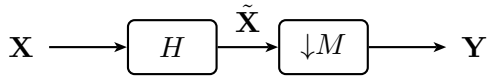


Figure 3.10: Simple decimation system. H is a linear filter.

Proof. See Appendix B.7. □

Lemma 3.4 (Filter's Do Nothing)**.** *Let* $\mathbf{X}$ *be a finite-variance stationary stochastic process satisfying Assumption 1 and let* H *be a stable, causal linear, time invariant filter with input* $\mathbf{X}$*; then, the output process* $\tilde{\mathbf{X}}$ *of the filter satisfies Assumption 1.*

Proof. See Appendix B.8. □

There is one pathological exception to this lemma, namely the case of a zero-output filter: Such a filter has a frequency response $H(\mathrm{e}^{\jmath\theta}) \equiv 0$ and hence violates the Paley-Wiener condition (see Appendix B.8), despite being stable and causal. The output process $\tilde{\mathbf{X}}$ of this system has a differential entropy rate $\bar{h}(\tilde{\mathbf{X}}) = -\infty$ and thus violates Assumption 1. Hence, in the remainder of this work, Lemma 3.4 is guaranteed to hold by requiring a stable, causal filter to be *non-trivial*.

It is not yet clear whether in Lemma 3.4 the requirement of finite variance can be dropped: As an example, the Pareto distribution [45, p. 145] has finite differential entropy and, for a specific parameter choice, finite mean but infinite variance. Assume the input process is iid, thus $\bar{h}(\mathbf{X}) = h(X)$. Moreover, since the mean is finite, $H(\hat{X}^{(0)}) < \infty$ and the information dimension exists. The filter H has the property of making the distribution of $\tilde{\mathbf{X}}$ close to a Gaussian distribution. Clearly, the variance of the output is infinite, so the Gaussian upper bound on $h(\tilde{X})$ evaluates to infinity.

With Lemma 3.3 and Lemma 3.4 it is possible to analyze multirate systems such as the one depicted in Fig. 3.10. Although strictly speaking not time-invariant, these systems can still be analyzed with relative information loss rates. To this end, assume that $\mathbf{X}$ is a stationary stochastic process satisfying Assumption 1.

For an M-fold downsampler, which is described by the input-output relation $Y_n := X_{nM}$, the information loss rate is simple to compute by introducing the M-fold blocked input process $\mathbf{X}^{(M)}$, whose samples are given by the M-dimensional RVs

$$X_n^{(M)} := X_{(n-1)M+1}^{nM}. \tag{3.79}$$

Since $Y_n := X_{nM}$ is a projection of $X_n^{(M)}$ to a single coordinate, the relative information loss for a specific sample is obtained via Corollary 2.4

$$l((X^{(M)})_1^n \to Y_1^n) = \frac{d((X^{(M)})_1^n | Y_1^n)}{d((X^{(M)})_1^n)} = \frac{n(M-1)}{nM} = \frac{M-1}{M}. \tag{3.80}$$

The relative information loss rate thus evaluates to

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) = \frac{M-1}{M}. \tag{3.81}$$

But how does a linear filter affect the relative information loss rate? In case the filter H is stable and causal, Lemma 3.4 guarantees that the information loss in the downsampler remains unchanged, i.e.,

$$l(\tilde{\mathbf{X}}^{(M)} \to \mathbf{Y}) = \frac{M-1}{M}. \tag{3.82}$$

Presently, the introduced measures of information loss are not directly applicable to linear filters (see Example 22). However, as intuition suggests and as made more explicit in Lemma 5.2, causal

and stable filters do not change the information content of the signal. Thus, one can abuse notation and exchange stochastic processes (or collections of stochastic processes) if they are equivalent from an informational perspective. In particular, this abuse will be made whenever there is a bijection[16] between the two processes. As a consequence, also for a stable and causal filter H one has

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) = l(\tilde{\mathbf{X}}^{(M)} \to \mathbf{Y}) = \frac{M-1}{M}. \tag{3.83}$$

The question remains whether an ideal anti-aliasing low-pass filter (or any filter violating the Paley-Wiener condition[17]) can eliminate, or at least reduce, the relative information loss rate. Clearly, as an ideal anti-aliasing filter limits the bandwidth of the process, the downsampler will not cause any information loss. In this case, however, the information is lost already in the filter:

Theorem 3.3 (Anti-Aliasing Filters are Useless)**.** *Let* $\mathbf{X}$ *be a finite-variance, Gaussian stationary stochastic process satisfying Assumption 1. The relative information loss rate in the multirate system depicted in Fig. 3.10 is*

$$l(\mathbf{X} \to \mathbf{Y}) \geq \frac{M-1}{M} \tag{3.84}$$

for all linear, time-invariant filters H *with finitely many stop-bands.*

Proof. See Appendix B.9. □

Note that a general filter can destroy an arbitrarily large amount of information, hence the inequality: Taking the pathological zero-output filter, for example, the relative information loss rate evaluates to 100% for every downsampling factor M.

While the proof is given only for Gaussian processes, the author believes that Theorem 3.3 is valid *for all* finite-variance processes satisfying Assumption 1. The proof in Appendix B.9 discusses a promising way to prove this result.

The intuition behind this result is that the ideal anti-aliasing low-pass eliminates a relative fraction of $(M-1)/M$ of the spectral components[18] of $\mathbf{X}$, which yields the according relative

[16] Note that a stable and causal linear filter may not have a stable and causal inverse. A particularly simple example to illustrate this claim is when H has a rational transfer function: The inverse transfer function may not be stable if required to be causal, but a non-causal, stable inverse will always exist. Thus, taking the complete process $\tilde{\mathbf{X}}$, it is possible to infer the input process $\mathbf{X}$. Only zeros in the frequency response lead to components which cannot be reconstructed. However, isolated zeros in the frequency response cancel only complex exponentials with corresponding frequencies; components which do not contribute to the *entropy rate* of the process.

[17] If the PSD $S_X(\mathrm{e}^{\jmath\theta})$ of a stationary process violates the Paley-Wiener condition, according to Martin Schetzen [134, p. 169]: "[The future of the signal violating]

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S_X(\mathrm{e}^{\jmath\theta}) d\theta > -\infty$$

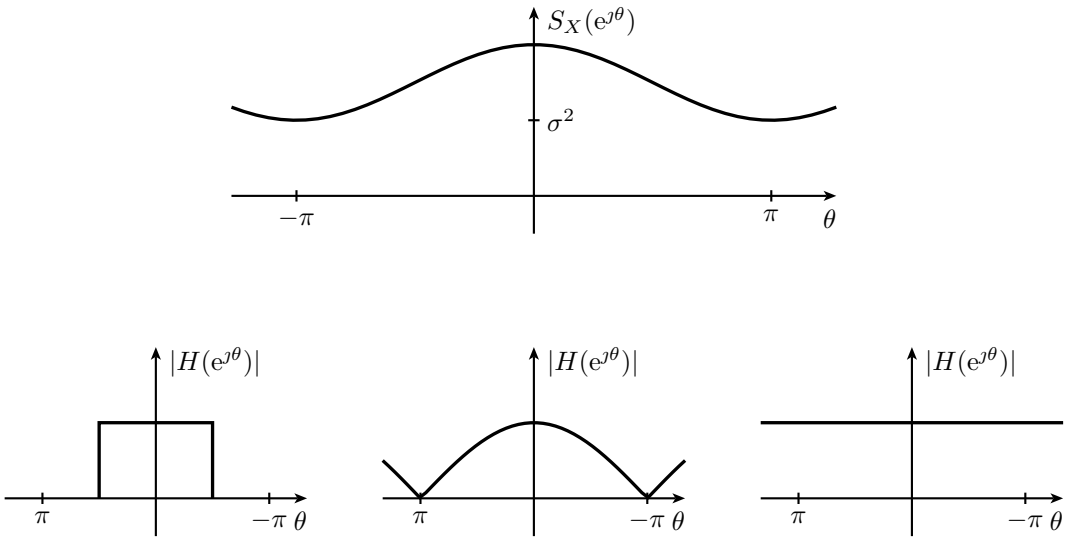
can be completely determined from its own past with arbitrarily small error [...] If the future of your speech waveform were predictable with arbitrary small error, then all that you will say in the future is predetermined and you would not be able to change it. Thus your free will would be definitely limited." Schetzen thus concludes that the PSD of our speech waveform cannot be nonzero in any band. The author of this work is not sure whether he agrees with this conclusion.

[18] An interesting aspect of these spectral considerations is the following: Eliminating spectral components leads to a band-limited process, for which the differential entropy rate evaluates to $-\infty$. In other words, the process becomes predictable [134, p. 169]. This suggests that the information dimension of n consecutive samples of the process will be strictly smaller than n, for n large enough. One might now ask if the concept of information dimension introduced by Rényi [123] can be generalized to stochastic processes, e.g., by $d(\mathbf{X}) := \lim_{n\to\infty} d(X_1^n)/n$. Furthermore, is there a relation between the information dimension of the process and its bandwidth, at least if the process is Gaussian? Simple examples (a predictable, sinusoidal process or an ideally low-pass filtered process) at least suggest so.

information loss rate. Although no information is lost in the downsampler, the total relative information loss rate stays the same: Here, information is already lost in the linear filter.

The consequence of these considerations is quite interesting, and parallels the reasoning for the PCA example in Section 2.6: When no signal model is available for the stochastic process $\mathbf{X}$, i.e., when it is not clear which part of the input information is *relevant*, anti-aliasing low-pass filtering is completely neutral in information-theoretic terms. The following example makes this explicit and summarizes the previous analysis:

Example 21: Anti-Aliasing Filters are Useless.



Let $\mathbf{X}$ be a stationary Gaussian process with PSD $S_X(\mathrm{e}^{\jmath\theta}) = \sigma^2 + 1 + \cos(\theta)$ depicted above. It shall be the input process to the multirate system in Fig. 3.10 with $M = 2$. To minimize the information loss rate across the downsampler, H can be any of the following three options (see above):

(a) An ideal anti-aliasing low-pass filter with cut-off frequency $\theta_c = \frac{\pi}{2}$,

(b) An FIR low-pass filter satisfying the Paley-Wiener conditions,

(c) A direct connection, i.e., $H(\mathrm{e}^{\jmath\theta}) = 1$ for all θ.

Energetic considerations would rank the three options in the given order; however, all are equivalent in terms of the relative information loss rate, which equals

$$l(\mathbf{X}^{(2)} \to \mathbf{Y}) = \frac{1}{2} \tag{3.85}$$

in all cases.

While the results of this section focus on discrete-time processes and decimation systems, the analysis can be extended to sampling continuous-time processes as well. This extension may be based on the sampling expansion of a bandlimited input process and successive downsampling. If the input process is not bandlimited, it is the author's conjecture that the relative information loss rate is unity, i.e., that 100% of the information is lost.

3.6 Outlook: Information Loss Rate in Systems with Memory

The measures of information loss presented in this work are not directly applicable to systems with memory, as the following example shows:

Example 22: Linear Filters and Information Loss Rate.
Consider the simple FIR filter given by the input-output difference equation $Y_n = aX_{n-1} + X_n$, where $|a| < 1$. Assume that the input process $\mathbf{X}$ is iid, and that the marginal probability measure, supported on $\mathbb{R}$, satisfies $P_X \ll \lambda$. First, note that

$$H(X_0^n|Y_1^n) = H(X_0|Y_1^n, X_1^n) + H(X_1^n|Y_1^n) \tag{3.86}$$
$$\overset{(a)}{=} H(X_0|Y_1, X_1^n) + H(X_1^n|Y_1^n) \tag{3.87}$$
$$\overset{(b)}{=} H(X_1^n|Y_1^n) \tag{3.88}$$

where (a) follows from the definition of the output and where (b) follows from the fact that X_0 can be computed from Y_1 and X_1. But the mapping from X_0^n to Y_1^n is a mapping from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$, and can be written as the cascade of a linear map and a projection. Therefore, and by the assumption that $\mathbf{X}$ is iid and has a marginal PDF,

$$l(X_0^n \to Y_1^n) = \frac{1}{n+1} > 0. \tag{3.89}$$

By Proposition 2.9 it follows that $H(X_0^n|Y_1^n) = \infty$ and, therefore,

$$H(X_1^n|Y_1^n) = L(X_1^n \to Y_1^n) = \infty \tag{3.90}$$

for all n, and hence $\bar{L}(\mathbf{X} \to \mathbf{Y}) = \infty$. But this naturally conflicts with the intuition that stable and causal filters do not change the information content of a signal (cf. Lemma 5.2). Moreover, it conflicts with the fact that the filter is minimum-phase and has, therefore, a stable and causal inverse (invertable systems cannot induce an information loss). Hence, the proposed measures of information loss rate cannot immediately be applied to system with memory, e.g., linear filters.

At least in case the involved processes have finite or countable alphabets, some preliminary results on the information loss rate can be presented. To this end, let $\mathbf{X}$ be a *two-sided* stationary stochastic process with finite alphabet $\mathcal{X}$. Consider the following class of systems:

Definition 3.9 (Finite-Dimensional Dynamical System)**.** Let $Y_n = f(X_{n-N}^n, Y_{n-M}^{n-1})$, $0 \leq M, N < \infty$, be the RV of the n-th output sample of a dynamical system with a finite-dimensional state vector subject to the input process $\mathbf{X}$. Here, $f\colon \mathcal{X}^{N+1} \times \mathcal{Y}^M \to \mathcal{Y}$ is a function such that the sequence of output samples, Y_n, constitutes a two-sided stochastic process $\mathbf{Y}$ jointly stationary with $\mathbf{X}$.

Lemma 3.5. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly stationary stochastic processes related as in Definition 3.9. Then, for* $M < \infty$*,*

$$\overline{H}(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n|Y_1^M) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n, Y_1^M). \tag{3.91}$$

Proof. Clearly,

$$H(X_1^n|Y_1^M) \leq H(X_1^n) \leq H(X_1^n, Y_1^M) \tag{3.92}$$

for all n, thus also in the limit. Now, since $H(X_1^n, Y_1^M) = H(X_1^n|Y_1^M) + H(Y_1^M)$ and since all involved entities are non-negative,

$$\overline{H}(\mathbf{X}) \leq \lim_{n\to\infty} \frac{1}{n} H(X_1^n|Y_1^M) + \underbrace{\lim_{n\to\infty} \frac{1}{n} H(Y_1^M)}_{\to 0} \tag{3.93}$$

if the first limit exists. But by assumption and (3.5), $H(X_1^n|Y_1^M) \geq H(X_1^n|X_{-\infty}^M) = (n -$

$M)\overline{H}(\mathbf{X})$, and therfore the limit exists:

$$\lim_{n\to\infty}\frac{n-M}{n}\overline{H}(\mathbf{X}) \leq \lim_{n\to\infty}\frac{1}{n}H(X_1^n|Y_1^M) \leq \overline{H}(\mathbf{X}) \tag{3.94}$$

Thus, in the limit the upper and lower bound are equal and the proof is completed. □

Intuitively, one would suspect that for dynamical systems, i.e., systems with memory, the last equality of Definition 3.2, stating that $\bar{L}(\mathbf{X}\to\mathbf{Y}) = \overline{H}(\mathbf{X}) - \overline{H}(\mathbf{Y})$ would not hold. As the next result shows, this intuition is wrong:

Proposition 3.17. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly stationary processes related as in Definition 3.9. Then, the information loss rate is given by the difference of entropy rates:*

$$\bar{L}(\mathbf{X}\to\mathbf{Y}) = \overline{H}(\mathbf{X}) - \overline{H}(\mathbf{Y}). \tag{3.95}$$

Proof. The proof follows with Definition 3.2,

$$\bar{L}(\mathbf{X}\to\mathbf{Y}) = \lim_{n\to\infty}\frac{1}{n}H(X_1^n|Y_1^n) = \lim_{n\to\infty}\frac{1}{n}H(X_1^n) - \lim_{n\to\infty}\frac{1}{n}H(Y_1^n) \tag{3.96}$$

where the second equality needs to be proved. Consider that, for $n > \max\{M, N\}$

$$\begin{aligned} H(X_1^n, Y_1^n) &= H(Y_n, X_1^n, Y_1^{n-1}) && (3.97)\\ &= H(f(X_{n-N}^n, Y_{n-M}^{n-1}), X_1^n, Y_1^{n-1}) && (3.98)\\ &= H(X_1^n, Y_1^{n-1}) && (3.99)\\ &\vdots \\ &= H(X_1^n, Y_1^{\max\{M,N\}}). && (3.100) \end{aligned}$$

Lemma 3.5 and the chain rule complete the proof. □

This result is interesting since it allows to employ Proposition 3.9 to the cascade of two or more systems also to the case where the systems are dynamic. In general, however, the computation of entropy rates is a non-trivial problem, where closed-form solutions exist only for simple processes (e.g., Markov chains). Since even functions of Markov chains only rarely allow such a simplified treatment (cf. Section 3.2), the availability of bounds is of great importance.

Proposition 3.18 (Upper Bound)**.** *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly stationary processes related as in Definition 3.9. Then, the information loss rate is bounded by*

$$\bar{L}(\mathbf{X}\to\mathbf{Y}) \leq \max_{(x,\theta)\in\mathcal{X}\times\mathcal{T}} \log \operatorname{card}(f_\theta^{-1}[f_\theta(x)]) \tag{3.101}$$

where $\mathcal{T} = \mathcal{X}^N \times \mathcal{Y}^M$, $\theta \in \mathcal{T}$ *are the possible values of the RV* $\Theta_n = \{X_{n-N}^{n-1}, Y_{n-M}^{n-1}\}$, *and* $f_\theta^{-1}[\cdot]$ *denotes the preimage under* f_θ, *an instantiation of the function* $f_{\Theta_n}(\cdot) = f(\cdot, \Theta_n)$.

Proof.

$$\begin{aligned} \bar{L}(\mathbf{X}\to\mathbf{Y}) &= \lim_{n\to\infty}\frac{1}{n}H(X_1^n|Y_1^n) && (3.102)\\ &\overset{(a)}{=} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1^{i-1}, Y_1^n) && (3.103)\\ &\overset{(b)}{\leq} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1^{i-1}, Y_1^i) && (3.104) \end{aligned}$$

where (a) is due to the chain rule of entropy and (b) is due to conditioning. The expression under the sum in (3.104) is a non-negative decreasing sequence in i and thus has a limit. Its

Cesáro mean exists [21, Thm. 4.2.3, p. 76], and one obtains

$$\begin{aligned}\bar{L}(\mathbf{X}\to\mathbf{Y}) &\le \lim_{n\to\infty} H(X_n|X_1^{n-1},Y_1^n) && (3.105)\\ &\le H(X_n|X_{n-N}^{n-1},Y_{n-M}^n) && (3.106)\\ &= H(X_n|Y_n,\Theta_n). && (3.107)\end{aligned}$$

One can now write $Y_n = f(X_n,\Theta_n) = f_{\Theta_n}(X_n)$ by treating the collection of all previous RVs influencing Y_n as a (random) parameter Θ_n of the function. This approach interprets the dynamical system as a parameterized static system $f_{\Theta_n}\colon \mathcal{X}\to\mathcal{Y}$, where Θ_n takes values θ from $\mathcal{T} = \mathcal{X}^N\times\mathcal{Y}^M$.

$$\begin{aligned}\bar{L}(\mathbf{X}\to\mathbf{Y}) &\le H(X_n|f_{\Theta_n}(X_n),\Theta_n) && (3.108)\\ &= \sum_{(x,\theta)\in\mathcal{X}\times\mathcal{T}} H(X_n|f_{\Theta_n}(X_n)=f_\theta(x),\Theta_n=\theta)\Pr(X_n=x,\Theta_n=\theta) && (3.109)\\ &\overset{(c)}{\le} \sum_{(x,\theta)\in\mathcal{X}\times\mathcal{T}} \log\operatorname{card}(f_\theta^{-1}[f_\theta(x)])\Pr(X_n=x,\Theta_n=\theta) && (3.110)\\ &\le \max_{(x,\theta)\in\mathcal{X}\times\mathcal{T}} \log\operatorname{card}(f_\theta^{-1}[f_\theta(x)]) && (3.111)\end{aligned}$$

where (c) is due the maximum entropy property of the uniform distribution over an alphabet size equal to the cardinality of the preimage under f_θ. Maximizing over all possible x and parameter values θ completes the proof. □

3.6.1 Partially Invertible Systems

To investigate yet another class of systems, consider an additional restriction on the system function in Definition 3.9. This additional restriction defines a class for which the information loss rate can be shown to vanish.

Definition 3.10 (Partially invertible system)**.** A system satisfying Definition 3.9 is *partially invertible* if there exists a function f_{inv} such that

$$X_n = f_{\text{inv}}(X_{n-N}^{n-1},Y_{n-M}^n) = f_{\text{inv}}(Y_n,\Theta_n) = f_{\Theta_n}^{-1}(Y_n). \tag{3.112}$$

In other words, a system is partially invertible if its parameterized static function f_{Θ_n} is invertible for all possible parameter values $\theta\in\mathcal{T}$.

It will next be shown that partially invertible systems are information lossless. To this end, consider

Definition 3.11 (Information-Lossless System)**.** A system satisfying Definition 3.9 is called *information-lossless*, if and only if $\bar{L}(\mathbf{X}\to\mathbf{Y}) = 0$ for all stationary stochastic processes $\mathbf{X}$ taking values from the set $\mathcal{X}$, for which the system responds with an output process $\mathbf{Y}$ jointly stationary with $\mathbf{X}$.

Proposition 3.19. *Partially invertible systems are information-lossless, i.e.,*

$$\bar{L}(\mathbf{X}\to\mathbf{Y}) = 0. \tag{3.113}$$

Proof. For all $\theta\in\mathcal{T}$, the parameterized function f_θ is invertible at all points of the image of $\mathcal{X}$ (which may be a strict subset of $\mathcal{Y}$ in this case). Thus, $\operatorname{card}(f_\theta^{-1}[f_\theta(x)]) = 1$ for all $x\in\mathcal{X}$ and $\theta\in\mathcal{T}$. Proposition 3.18 completes the proof. □

Note that the converse is not true, i.e., there exist information-lossless systems which are not partially invertible. For example, let f be such that, for all input processes $\mathbf{X}$ on $\mathcal{X}$, the

sequence of M consecutive output samples Y_{n-M}^{n-1} takes values from a strict subset of $\mathcal{Y}^M$. As a consequence, the parameterized function f_θ needs not be invertible for *all* parameters $\theta \in \mathcal{T}$, but only for values which are achievable:

Example 23: Information Lossless but not Partially Invertible.
Consider a linear filter described by the difference equation

$$Y_n = X_n + \alpha Y_{n-1} \tag{3.114}$$

with $\alpha > 0.5$ and an input process $\mathbf{X}$ taking values from the set $\mathcal{X} = [-a, a] \cap \mathbb{Q}$. Consequently, the output process takes values from the set $\mathcal{Y} = [\frac{-a}{1-\alpha}, \frac{a}{1-\alpha}]$. However, the maximum amount by which two consecutive output values can differ is given by

$$\max |Y_n - Y_{n-1}| = 2a < \frac{a}{1-\alpha}. \tag{3.115}$$

By Definition 3.10, the linear filter described by (3.114) is partially invertible, thus also information-lossless. However, let us now modify the system equation such that the output is given as

$$Y_n = f_{Y_{n-2}^{n-1}}(X_n) = \begin{cases} X_n + \alpha Y_{n-1}, & \text{if } |Y_{n-1} - Y_{n-2}| \leq 2a \\ 0, & \text{else} \end{cases}. \tag{3.116}$$

Since the condition above is always (i.e., for all possible input processes and initial values of the system states) fulfilled, the system is information-lossless; however, the system is not partially invertible, since there exist points in $\mathcal{Y}^2$ for which the parameter values θ of $\Theta_n = Y_{n-2}^{n-1}$ have a difference greater than $2a$. And since for these parameter values the function delivers zero as output regardless of the input, the preimage is not a singleton.

Next, consider the following system:

$$Y_n = \begin{cases} X_n + \alpha Y_{n-1}, & \text{if } |Y_{n-1} - Y_{n-2}| \leq 2a \\ Y_{n-2}, & \text{else} \end{cases} \tag{3.117}$$

Again, this system is not partially invertible, since for some parameter values the system responds with a single value regardless of the input. However, depending on the initial conditions the system may not even be information-lossless: Assume that initially $|Y_{n-1} - Y_{n-2}| > 2a$. Then, as it can be verified easily, the output oscillates between these two initial values and all information is lost. This suggests a connection between preservation of information and the stability properties (asymptotic stability, input-to-state stability, etc.) of the nonlinear system.

Indeed, the class of partially invertible systems not only proves to have vanishing information loss rate, but also finite absolute information loss: By observing a sequence of output samples Y_1^n, the total information loss corresponds to the uncertainty about the initial values of the corresponding input sequence. Indeed, for $n > \max\{M, N\}$, it turns out that [50, Thm. 4]

$$H(X_1^n | Y_1^n) = H(X_1^{\max\{M,N\}} | Y_1^n). \tag{3.118}$$

Note that even though f_{Θ_n} is invertible for all parameter values θ, this does not mean that $H(X_n | f_{\Theta_n}(X_n)) = 0$; however, $H(X_n | f_{\Theta_n}(X_n), \Theta_n) = 0$ holds. This is due to the fact that, for $n < \max\{M, N\}$, Θ_n is unknown.

The class of information lossless systems contains the class of systems which permit perfect reconstruction of the input as a proper subclass: While the possibility of perfect reconstruction automatically yields a vanishing information loss rate, the converse is not true, even if reconstruction errors are allowed in the first $\max\{M, N\}$ samples. The following example illustrates

this fact:

Example 24: Multiplying Consecutive Inputs.
The following system can be shown to be partially invertible in the sense of Definition 3.10:

$$Y_n = X_n X_{n-1} \tag{3.119}$$

The partial inverse is $X_n = \frac{Y_n}{X_{n-1}}$ if $X_{n-1} \neq 0$, while for $X_{n-1} = 0$ no such inverse exists. Therefore, this example represents a class of systems whose partial invertibility depends on the alphabet $\mathcal{X}$ of the stochastic process. If the process $\mathbf{X}$ is such that $\mathcal{X}$ does not contain the element 0, the partial inverse exists and one obtains for X_n, $n > 1$:

$$X_n = \begin{cases} X_1 \prod_{k=1}^{\frac{n-1}{2}} \frac{Y_{2k+1}}{Y_{2k}}, & \text{for odd } n \\ \frac{Y_n}{X_1} \prod_{k=1}^{\frac{n}{2}-1} \frac{Y_{2k}}{Y_{2k+1}}, & \text{for even } n \end{cases} \tag{3.120}$$

Indeed, since all X_n, $n > 1$, can be computed from X_1 and Y_1^n, it follows that $H(X_1^n|Y_1^n) = H(X_1|Y_1^n)$. Reconstruction of $\mathbf{X}$ is thus possible up to an unknown X_1. Note, however, that this unknown sample influences the whole reconstructed sequence as shown in (3.120). Thus, even though the information loss rate vanishes, perfect reconstruction of any subsequence of $\mathbf{X}$ is impossible by observing the corresponding output sequence only.

The notion of partial invertibility might play a role in the equalization or linearization of mildly nonlinear Volterra systems [77, 110]. In particular, [77] presented a sufficient condition for an iterative equalization method to converge. To fulfill the condition, the linear part of the Volterra system needs to be dominant and invertible. It can be shown that the sufficient condition of [77] is sufficient but not necessary for partial invertibility according to Definition 3.10.

3.6.2 Example: Fixed-Point Implementation of a Linear Filter

An important subclass of discrete-time stable causal linear filters falls in the category of partially invertible systems, as long as the input and output alphabets are countable. An example where the latter condition is satisfied is given if the input process and the coefficients take values from the field of rational numbers. This subclass, powerful enough to cover most applications [29], comprises filters with a finite-dimensional state vector described by constant-coefficient difference equations:

$$Y_n = \sum_{k=0}^{N} b_k X_{n-k} + \sum_{l=1}^{M} a_l Y_{n-l} \tag{3.121}$$

As noted in [113], stability of the filter guarantees that for a stationary input process the output process is stationary and that Definition 3.9 applies. By rearranging the terms in (3.121) it can be verified that this subclass of linear systems satisfies the definition of partially invertible systems and, thus, has a vanishing information loss rate.

It is noteworthy that this property is independent of the minimum-phase property (cf. [111, pp. 280]) of linear filters, which ensures that the filter has a stable and causal inverse. Indeed, for filters which are not minimum-phase, the partial inverse function f_{inv} in Definition 3.10 describes a causal, but unstable linear filter. As a consequence, to an arbitrary stationary stochastic input process, the inverse filter described by f_{inv} may respond with a non-stationary output process; however, the response to $\mathbf{Y}$ will be $\mathbf{X}$.

A signal space model may effectively illustrate these considerations: Let $\mathcal{X}^\infty$ and $\mathcal{Y}^\infty$ be the spaces of stationary input and output processes $\mathbf{X}$ and $\mathbf{Y}$, respectively, and let $F\{\cdot\}$ be the (linear) operator mapping each element of $\mathcal{X}^\infty$ to $\mathcal{Y}^\infty$. By restricting our attention to regular

stochastic processes, i.e., processes which cannot have periodic components, the operator $F\{\cdot\}$ is injective. As a consequence, for each element of $\mathcal{Y}^\infty$ there exists at most one element in $\mathcal{X}^\infty$ such that $\mathbf{Y} = F\{\mathbf{X}\}$. Note, however, that there are stationary stochastic processes in $\mathcal{Y}^\infty$ which are not images of elements in $\mathcal{X}^\infty$. Only if $F\{\cdot\}$ is such that it describes a stable, causal minimum-phase system, i.e., has a *stable* and causal inverse, $\mathcal{Y}^\infty$ contains only images of elements from $\mathcal{X}^\infty$.

Linear filters, if stable and causal, can thus be assumed to be lossless from an information-theoretic point-of-view. It is surprising, though, that even filters with finite precision, i.e., non-linear filters, can preserve information:

Example 25: Finite-Precision Linear Filters.

In many practical applications in digital signal processing linear filters are implemented with finite-precision number representations only. Thus assume that both input process and filter coefficients take values from a finite set. For example, $\mathcal{X}$ may be a finite subset of the rational numbers $\mathbb{Q}$, closed under modulo-addition. Multiplying two values from that set, e.g., by multiplying an input sample with a filter coefficient, typically yields a result not representable in $\mathcal{X}$. As a consequence, after every multiplication a quantizer is necessary, essentially truncating the additional bits resulting from multiplication. Let the quantizer be described by a function $Q\colon \mathbb{R} \to \mathcal{X}$ with $Q(a + X_n) = Q(a) \oplus X_n$ if $X_n \in \mathcal{X}$, where $\oplus$ denotes modulo-addition (e.g., [111, pp. 373]). With this, (3.121) changes to

$$Y_n = \bigoplus_{k=0}^{N} Q(b_k X_{n-k}) \oplus \bigoplus_{l=1}^{M} Q(a_l Y_{n-l}) \tag{3.122}$$

or

$$Y_n = Q\left(\bigoplus_{k=0}^{N} b_k X_{n-k} \oplus \bigoplus_{l=1}^{M} a_l Y_{n-l}\right) \tag{3.123}$$

depending whether quantization is performed after multiplication or after accumulation (in the latter case, the intermediate results are represented in a larger set $\mathcal{X}'$). Note that due to modulo-addition the result Y_n remains in $\mathcal{X}$.

Focus on filters with $b_0 = 1$. For filters with infinite precision this can be done without loss of generality by considering a constant gain factor b_0 and by normalizing all b_k coefficients. However, this gain normalization poses a restriction in the finite-precision case since b_k/b_0 is not necessarily an element of $\mathcal{X}$. With $b_0 = 1$, (3.122) and (3.123) change to

$$Y_n = X_n \oplus \left(\bigoplus_{k=1}^{N} Q(b_k X_{n-k}) \oplus \bigoplus_{l=1}^{M} Q(a_l Y_{n-l})\right) \tag{3.124}$$

and

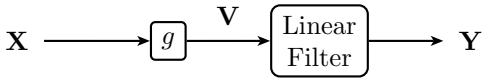
$$Y_n = X_n \oplus Q\left(\bigoplus_{k=1}^{N} b_k X_{n-k} - \bigoplus_{l=1}^{M} a_l Y_{n-l}\right) \tag{3.125}$$

by the property of the quantizer. From this it can be seen that either implementation is partially invertible (the terms in parentheses in (3.124) and (3.125) are both in $\mathcal{X}$, and modulo-addition has an inverse element). Consequently, even filters with nonlinear elements can be shown to preserve information under certain circumstances despite the fact that the quantizer function is non-injective.

How does the fact that linear filters – finite-precision or not – preserve information connect to Example 22, which claims that even an invertible filter destroys an infinite amount of information *in each time unit*? The answer is that Example 22 just illustrated that the proposed measure of information loss rate is not suitable for systems with memory in the general case. It is well-

suited, however, for input processes with a finite or countable alphabet (at least if the entropy of the marginal distribution remains finite). The problem in the general case is probably due to the particular choice of the order of limits in the definition: The information loss rate is a limit of information losses for blocks of samples divided by the block lengths; the information losses themselves being limits of quantizations, cf. Definition 2.2. For input processes with finite or countable alphabets, the order of these limits is immaterial. Whether some of these problems resolve by exchanging the order of the limits is within the scope of future work.

Example 26: Hammerstein System.



Consider the cascade of a static nonlinearity and a linear filter depicted above; such a cascade is usually referred to as Hammerstein system [139]. A practical example is the energy detector, a low-complexity receiver in wireless communications. In the discrete-time case the input-output relationship is given by

$$Y_n = \sum_{k=0}^{N} b_k g(X_{n-k}) + \sum_{l=1}^{M} a_l Y_{n-l}. \tag{3.126}$$

This system is partially invertible if and only if the function g has an inverse. If g is not invertible, one obtains in the light of Proposition 3.18:

$$Y_n = f_{\Theta_n}(X_n) = b_0 g(X_n) + C_{\Theta_n} \tag{3.127}$$

where C_{Θ_n} depends on the random parameter Θ_n. With this and $f_\theta^{-1}[f_\theta(x)] = g^{-1}[g(x)]$ for all $x \in \mathcal{X}$, $\theta \in \mathcal{T}$ one obtains an upper bound on the information loss rate:

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) \leq \max_{x \in \mathcal{X}} \log \operatorname{card}(g^{-1}[g(x)]) \tag{3.128}$$

Interestingly, the structure of this system allows a simplified analysis: Since the information loss rate of a cascade of systems is equal to the sum of individual information loss rates (cf. Proposition 3.9 in combination with Proposition 3.17) one can analyze the constituting systems separately. The linear filter was already shown to preserve full information, so any information loss will be caused by the static nonlinearity, i.e., $\bar{L}(\mathbf{X} \to \mathbf{Y}) = \bar{L}(\mathbf{X} \to \mathbf{V})$. This is in accordance with the observation that the Hammerstein system is partially invertible if the static nonlinearity is invertible.

3.7 Open Questions

The extension from information loss of RVs to information loss rates of stochastic processes is not unique. Specifically, in Definition 3.7, one has the choice of exchanging the limit over the sequence length with the limit over quantizations, yielding different properties of the defined quantity. Specifically, it may be the case that this exchange is better suited for systems with memory, which are not considered in this work (except briefly in Section 3.6). This is only one aspect worth investigating in the future.

A second one concerns the information loss rate in Markov chains (cf. Section 3.2). Christoph Temmel and the present author believe that the single entry property of Definition 3.5 can be given a graph-theoretic interpretation, and that lumpings satisfying this property can be "read off" a specific graph with node set $\mathcal{X}$. This is clearly a problem worth investigating. Moreover, the application of the presented results to problems other than n-gram clustering, and the

improvement of the algorithm should be part of future work. The present author also found different sufficient conditions for 2-lumpability, which should be investigated; in particular, the interplay between these conditions and the $\mathsf{SFS}(2)$-property should be analyzed. Furthermore, performance bounds for $\mathsf{SFS}(2)$-lumpings and a complexity analysis of the algorithm as a function of the graph degree could be developed. Finally, the optimality of $\mathsf{SFS}(k)$- lumpings in terms of compression, and the generalization to non-stationary, time-homogeneous Markov chains would be important.

A connection between the information loss rate and the probability of reconstruction error would give the former quantity an operational meaning. Reconstruction of nonlinearly distorted sequences is difficult, and performance bounds would be helpful in this regard.

In Section 3.5, relative information loss was employed for analyzing decimation systems. The generalization to sampling devices was hinted at, but not thoroughly analyzed. Future work shall fill this gap, which will probably need an information measure for continuous-time, continuous-amplitude stochastic processes. It is also of great interest if there are systems and/or processes, for which the relative information loss rate is *strictly smaller* than the relative information loss, cf. Proposition 3.16.

Another interesting question is whether, at least for discrete-time Gaussian processes, the definition of an information dimension rate makes sense, and if it is connected to the bandwidth of the process. The tools for such an analysis might be, again, a filterbank decomposition, Rényi's information dimension, and the Paley-Wiener theorem.

Paralleling the analysis of sample covariance-based PCA in Section 2.6.2, it might be interesting if similar results can be obtained for adaptive filters, i.e., filters which depend on the input signal. In other words, assuming that the adapted filter coefficients are not transmitted, what is the (relative) information loss (rate) in, e.g., and LMS-adapted linear predictor?

Of course, the extension to systems with memory is of great importance, but, as the author believes, also one of great difficulty. Probably, results can be obtained for restricted classes of systems, such as Volterra systems. The relevance of these results in practical applications (e.g., linearization or equalization of systems, as mentioned in Section 3.6) shall also be investigated.

4 Relevant Information Loss

The notion of relevant information loss was first employed in [55]; an essentially identical quantity was introduced earlier by Plumbley in [118]. The idea for this metric was written down simultaneously by the present author and Gernot Kubin, although the latter had had it in mind probably much longer. It was essentially his idea to apply this quantity for signal enhancement; the present author found its connection to the information bottleneck method, as indicated in Section 4.2.

The part on Markov chain aggregation in Section 4.3 was developed in cooperation with Tatjana Petrov, Gernot Kubin, and Heinz Köppl [47]. In particular, the theoretic results have mainly been discovered by the present author; much of the notation is due to Dr. Petrov. The example on bio-molecular systems in Section 4.3.6 has also been written by her, while the simulations have been carried out together with the present author.

The results on PCA in Section 4.4, again influenced by Gernot Kubin, are published in [55]. Inaccurate function knowledge was considered by the main author only; the results of the corresponding Section 4.5 are unpublished.

4.1 The Problem of Relevance - A General Definition

As the examples in Sections 2.6 and 3.5 showed, both the absolute and the relative definition of information loss have their shortcomings, especially when it comes to systems g used for signal enhancement: Since the expressions only consider the RV X at the input of the system, they do not take into account that not all of the information contained in X is *relevant*, often leading to counter-intuitive results. The main contribution of this chapter lies thus in analyzing the implications of

Definition 4.1 (Relevant Information Loss)**.** Let X be an RV with alphabet $\mathcal{X}$, and let $Y = g(X)$. Let S be another RV with alphabet $\mathcal{S}$ representing *relevant information*. The information loss *relevant w.r.t.* S is

$$L_S(X \to Y) = \lim_{n\to\infty} \Big(I(\hat{S}^{(n)}; X) - I(\hat{S}^{(n)}; Y) \Big) = I(X; S|Y). \tag{4.1}$$

The last equality holds because Y is a function of X and, thus,

$$\lim_{n\to\infty} \Big(I(\hat{S}^{(n)}; X) - I(\hat{S}^{(n)}; Y) \Big) = \lim_{n\to\infty} \Big(I(\hat{S}^{(n)}; X, Y) - I(\hat{S}^{(n)}; Y) \Big) \tag{4.2}$$

$$= \lim_{n\to\infty} I(\hat{S}^{(n)}; X|Y) \tag{4.3}$$

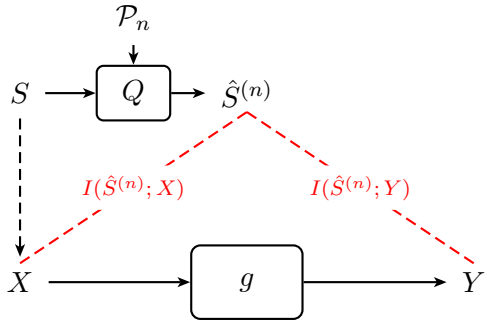


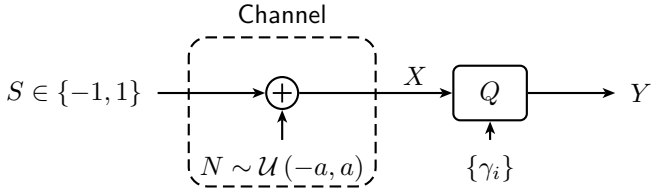
Figure 4.1: Model for computing the relevant information loss of a memoryless input-output system g. Q is a quantizer with partition $\mathcal{P}_n$.

by the chain rule of information [21, Thm. 2.5.2, p. 24]. Application of [66, Lemma. 7.22] yields the desired result.

The information loss relevant w.r.t. an RV S is thus the difference of mutual informations between S and the input and output of the system (see Fig. 4.1). Equivalently, the relevant information loss is exactly the information that X contains about S *which is not contained in* Y. Due to the data processing inequality, it is a non-negative quantity, i.e., a deterministic system cannot increase the amount of available relevant information.

This definition of relevant information loss is not altogether new: Plumbley already introduced this quantity (named $\Delta I_S(X;Y)$, and omitting the limit over quantizations assuming finite mutual information between S and X) in the context of unsupervised learning in neural networks [118]. The rationale for this new measure was to circumvent some shortcomings of Linsker's principle of information maximization [100]: While infomax works well for Gaussian RVs and linear systems, applying the same algorithms to non-Gaussian data just maximizes an upper bound on the information. Plumbley's information loss, conversely, also yields closed form solutions for Gaussian data, but in addition to that one can derive upper bounds on the relevant information loss whenever the data is non-Gaussian. Naturally, minimizing an upper bound on the information loss is more promising than maximizing an upper bound on the transferred information [118].

Example 27: A Simple Digital Communication System.



Consider the digital communication system depicted above. Let the data symbols S be uniformly distributed on $\{-1,1\}$, thus $H(S)=1$. Let further N be a noise signal which, for the sake of simplicity, is uniformly distributed on $[-a,a]$, $a>1$. Clearly, $X=S+N$ is a continuous-valued signal with infinite entropy. During quantization an infinite amount of information is lost, $L(X\to Y)=\infty$ (cf. Example 1). The main point here is that most of the information at the input of the quantizer is information about the noise signal N.

With the differential entropy of X given by

$$h(X)=\frac{1}{a}\log 4a+\frac{a-1}{a}\log 2a \tag{4.4}$$

and with $h(X|S)=\log 2a$, the amount of relevant information available at the input of the quantizer is $I(X;S)=1/a<1$.

Choosing a single quantizer threshold $\gamma_1=0$, i.e., $Y=\operatorname{sgn}(X)$, yields a binary symmetric

channel with cross-over probability $P_e = \frac{a-1}{2a}$. With $H_2(\cdot)$ being the binary entropy function the mutual information thus computes to $I(Y;S) = 1 - H_2(P_e)$ [21, p. 187]. Consequently, one gets a relevant information loss of

$$L_S(X \to Y) = H_2(\frac{a-1}{2a}) - \frac{a-1}{a}. \tag{4.5}$$

Using two quantizer thresholds $\gamma_1 = 1 - a$ and $\gamma_2 = a - 1$ yields a ternary output RV Y. Interpreting any value $\gamma_1 \leq X \leq \gamma_2$ as an erasure, one obtains a binary erasure channel with erasure probability $\frac{a-1}{a}$. The corresponding mutual information computes to $I(Y;S) = \frac{1}{a}$ [21, p. 188], and the relevant information loss vanishes.

As this example shows, while the quantizer destroys an infinite amount of information (to be precise, the relative information loss is 100%), the relevant information loss can still be zero. Signal enhancement, in the sense of removing irrelevant information, was thus successful.

4.1.1 Elementary Properties

Before proceeding, the elementary properties of relevant information loss will be analyzed:

Proposition 4.1 (Elementary Properties)**.** *The relevant information loss from Definition 4.1 satisfies the following properties:*

1. $L_S(X \to Y) \leq I(S;X) \leq H(S)$
2. $L_S(X \to Y) \geq L_Y(X \to Y) = 0$
3. $L_S(X \to Y) \leq L_X(X \to Y) = L(X \to Y)$*, with equality if* X *is a function of* S*.*
4. $L_S(X \to Y) = H(S|Y)$ *if* S *is a function of* X*.*

Proof. The first property results immediately from the definition, while the second property is due to the fact that Y is a function of X. The third property results from making X the relevant RV, thus making Definition 4.1 equal to Definition 2.2. Note that, with $L(X \to Y) = H(X|Y)$,

$$L_S(X \to Y) = I(S;X|Y) = H(X|Y) - H(X|Y,S) \leq L(X \to Y) \tag{4.6}$$

with equality if X is a function of S. For $S = f(X)$, the last property follows by expanding $I(X;S|Y)$ as $H(S|Y) - H(S|Y,X) = H(S|Y)$. □

The third property is of particular interest. Essentially, it states that the relevant information loss cannot exceed the total information loss, so upper bounds (e.g., those presented in Section 2.3.2) for the latter can be used for the former as well.

Since by Definition 4.1 relevant information loss is represented by a conditional mutual information, it inherits all of its properties. In particular, a data processing inequality holds:

Proposition 4.2 (Data Processing Inequality)**.** *Let* $V - W - X - Y$ *be a Markov tuple. Then,*

$$L_W(X \to Y) \geq L_V(X \to Y). \tag{4.7}$$

Proof. The proof follows along the same lines as the proof of the data processing inequality for mutual information in [21, Thm. 2.8.1, p. 34]. Expanding $I(X;W,V|Y)$ yields

$$\begin{aligned} I(X;W,V|Y) &= I(X;V|Y) + I(X;W|V,Y) && (4.8)\\ &= I(X;W|Y) + I(X;V|W,Y) && (4.9)\\ &= I(X;W|Y) && (4.10) \end{aligned}$$

where the last line follows from the fact that V and X are conditionally independent given W. With Definition 4.1 one then obtains

$$\begin{aligned} L_W(X \to Y) &= I(X;W|Y) && (4.11)\\ &= I(X;V|Y) + I(X;W|V,Y) && (4.12)\\ &= L_V(X \to Y) + I(X;W|V,Y) && (4.13)\\ &\geq L_V(X \to Y) && (4.14) \end{aligned}$$

which completes the proof. □

The usefulness of this data processing inequality (especially in the context of this work) relies on the fact that both $S - X - g(X)$ and $f(S) - S - X$ are Markov tuples. Comparing this to Proposition 4.2 one is tempted to believe that the direction of the inequality depends on the fact whether $f(S) - S - X - Y$ or $S - f(S) - X - Y$ is a Markov tuple. The following corollary, a generalization of the third property of Proposition 4.1, resolves this complication.

Corollary 4.1. *Let f be a measurable function defined on the sample space of S. Then,*

$$L_S(X \to Y) \geq L_{f(S)}(X \to Y) \tag{4.15}$$

with equality if $S - f(S) - X - Y$ is a Markov tuple.

Proof. The corollary follows immediately from [116, Thm. 3.7.1]. Specifically, if $f(S) - S - X - Y$ is a Markov tuple, the case is trivial. For $S - f(S) - X - Y$ being a Markov tuple, one gets with the proof of Proposition 4.2

$$L_{f(S)}(X \to Y) = L_S(X \to Y) + I(X; f(S)|S,Y) = L_S(X \to Y). \tag{4.16}$$

□

Inherited from the properties of mutual information, also the relevant information loss obeys a chain rule:

Lemma 4.1 (Chain Rule of Information Loss)**.** *The information loss $L_{S_1^n}(X \to Y)$ induced by a function g, relevant w.r.t. a collection $S_1^n = \{S_1, \ldots, S_n\}$ of RVs, satisfies*

$$L_{S_1^n}(X \to Y) = \sum_{i=1}^{n} L_{S_i|S_1^{i-1}}(X \to Y). \tag{4.17}$$

The proof follows immediately from the chain rule of (conditional) information and is thus omitted. More interestingly, this chain rule justifies our intuitive understanding of the nature of information loss:

Corollary 4.2. *The information loss $L(X \to Y)$ induced by a function g can be split into* relevant *(w.r.t. S) and* irrelevant *information loss:*

$$L(X \to Y) = L_S(X \to Y) + L_{X|S}(X \to Y) \tag{4.18}$$

Proof. The proof follows from Lemma 4.1 and from the fact that $L_{XS}(X \to Y) = L(X \to Y)$, since S and Y are conditionally independent given X. □

Note that even in a simple scenario with additive noise, i.e., $X = S + N$, it is not straightforward to just identify the noise N with the irrelevant information:

Example 27: A Simple Digital Communication System (revisited).
In this example it is tempting to identify the noise variable N with the irrelevant information $X|S$. However, this not necessarily leads to a correct result:

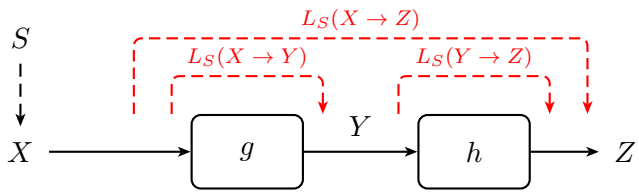


Figure 4.2: Cascade of two systems: The relevant information loss of the cascade is the sum of relevant information losses of the constituting systems.

To this end, substitute the quantizer in the system by a magnitude device, i.e., $Y' = |X|$. By the fact that the probability density function of X has even symmetry, one has $L(X \to Y') = H(X|Y') = 1$. Further, since the marginal distribution of Y' coincides with the conditional distributions $Y|S = -1$ and $Y|S = 1$, the mutual information $I(Y'; S) = 0$. Thus, $L_S(X \to Y') = \frac{1}{a}$ and $L_{X|S}(X \to Y') = \frac{a-1}{a}$.

Let us now determine the information loss relevant w.r.t. N. Showing that $L_N(X \to Y') \neq L_{X|S}(X \to Y')$ proves that noise and irrelevant information are not identical. Observe that

$$\begin{aligned} L_N(X \to Y') &= I(X; N|Y') && (4.19)\\ &= H(X|Y') - H(X|N, Y') && (4.20)\\ &= 1 - H(S + N|N, Y') && (4.21)\\ &= 1 - H(S|N, |S + N|). && (4.22) \end{aligned}$$

Given N, S is uncertain only if $|S + N|$ yields the same value for both $S = 1$ and $S = -1$. In other words, only for $|N - 1| = |N + 1|$. Squaring both sides translates to requiring $N = -N$, which is fulfilled only for $N = 0$. Since $\Pr(N = 0) = 0$ (N is a continuous RV), it follows that

$$L_N(X \to Y') = 1 \neq \frac{a-1}{a} = L_{X|S}(X \to Y'). \tag{4.23}$$

The reason why one cannot identify the noise with the irrelevant information can be related to N and S not being conditionally independent given X. The findings are summarized in the table below.

Loss	$Y = \operatorname{sgn}(X)$	$Y = \lvert X\rvert$
$L(X \to Y)$	∞	1
$L_S(X \to Y)$	$H_2(\frac{a-1}{2a}) - \frac{a-1}{a}$	$\frac{1}{a}$
$L_{X\lvert S}(X \to Y)$	∞	$\frac{a-1}{a}$
$L_N(X \to Y)$	∞	1
$L_{X\lvert N}(X \to Y)$	$\frac{a-1}{2a}$	0

A final property worth presenting concerns the cascade of systems (see Fig. 4.2). Assume that the output Y of the first system g is used as the input to a second system h, which responds with the RV Z. In Proposition 2.2 it was shown that the information loss of this cascade is additive, i.e.,

$$L(X \to Z) = L(X \to Y) + L(Y \to Z). \tag{4.24}$$

The same holds for the relevant information loss, as Plumbley showed in [118]:

Proposition 4.3 (Information Loss in a Cascade is Additive, [118])**.** *Consider two functions $g\colon \mathcal{X} \to \mathcal{Y}$ and $h\colon \mathcal{Y} \to \mathcal{Z}$ and a cascade of systems implementing these functions. Let $Y := g(X)$ and $Z := h(Y)$. The relevant information loss w.r.t. S induced by this cascade, or equivalently,*

by the system implementing the composition $(h \circ g)(\cdot) = h(g(\cdot))$ *is given by*

$$L_S(X \to Z) = L_S(X \to Y) + L_S(Y \to Z). \tag{4.25}$$

Proof. Note that with Definition 4.1 one gets

$$\begin{aligned} L_S(X \to Z) &\overset{(a)}{=} I(S; X, Y|Z) && (4.26)\\ &\overset{(b)}{=} I(S; X|Y, Z) + I(S; Y|Z) && (4.27)\\ &\overset{(a)}{=} I(S; X|Y) + I(S; Y|Z) && (4.28) \end{aligned}$$

where (a) is due to the fact that $Y = g(X)$ and $Z = h(Y)$, respectively, and (b) is the chain rule of information. □

4.1.2 A Simple Upper Bound

It was already mentioned that, by the fact $L_S(X \to Y) \leq L(X \to Y)$, the upper bounds of Section 2.3 can also act as upper bounds on the relevant information loss. In addition to these, the relevant information loss can – under specific assumptions – be upper bounded by looking at the joint second-order statistics only.

Definition 4.2 (Negentropy)**.** Let X be an RVs with distribution $P_X \ll \lambda$, and let X_G be a Gaussian RVs with the same first and second moments. The *negentropy* of X is

$$J(X) := D(f_X || f_{X_G}) = h(X_G) - h(X) \tag{4.29}$$

whenever the quantity exists.

Clearly, negentropy is non-negative; it is commonly used as a measure of Gaussianity, as it is zero if and only if X is Gaussian.

Proposition 4.4 (Gaussian Upper Bound)**.** *Let* S *and* X *be jointly Gaussian RVs and let* $Y := g(X)$*, where* g *is piecewise bijective. Let further* Y_G *(*Y_G, S*) be a Gaussian RV with the same (joint) first and second moments as* Y *(*Y, S*). Then,*

$$L_S(X \to Y) \leq L_S(X \to Y_G). \tag{4.30}$$

Proof. By the chain rule of relative entropy [21, Thm. 2.5.3, p. 24] and Gaussianity of S,

$$\begin{aligned} J(S, Y) &= D(f_{S,Y} || f_{S,Y_G}) && (4.31)\\ &= \mathbb{E}\left(D(f_{S|Y}(\cdot|Y) || f_{S|Y_G}(\cdot|Y))\right) + J(Y) && (4.32)\\ &= \mathbb{E}\left(D(f_{Y|S}(\cdot|S) || f_{Y_G|S}(\cdot|S))\right) + J(S) && (4.33)\\ &= \mathbb{E}\left(D(f_{Y|S}(\cdot|S) || f_{Y_G|S}(\cdot|S))\right) && (4.34)\\ &= h(Y_G|S) - h(Y|S) \qquad =: J(Y|S) && (4.35) \end{aligned}$$

from which follows that $J(Y) \leq J(Y|S)$. The proof is completed with Definitions 4.2 and 4.1,

$$\begin{aligned} L_S(X \to Y) &= I(S; X) - I(S; Y) && (4.36)\\ &= I(S; X) - h(Y) + h(Y|S) && (4.37)\\ &= I(S; X) + J(Y) - h(Y_G) - J(Y|S) + h(Y_G|S) && (4.38)\\ &\leq I(S; X) - I(S; Y_G) = L_S(X \to Y_G). && (4.39) \end{aligned}$$

□

Note that the requirement of piecewise bijectivity of g in general cannot be dropped, since it guarantees that Y has a PDF. One could extend the previous result to let X and S have

arbitrary distribution and qualify the bound as being an upper or a lower bound by refering to the negentropy of the involved RVs. This does not yield general results, however, since a given (non-linear) function g can both increase *or* decrease negentropy.

4.2 Signal Enhancement and the Information Bottleneck Method

Consider the discrete case: X and S are RVs with finitely many states, and one is interested in a lossy compression Y of X which contains as much information about S as possible. Compression is related to removing irrelevant information, while preserving relevant information. In fact, this goal is not unique to compression, but is also found in signal enhancement: Often, the information one wants to use is not directly available, but only through some corrupted observation. Information-carrying signals are affected by noise, distorted through nonlinear systems, or time-varying, dynamic effects. It is essentially the goal of the signal processing engineer to mitigate all these adverse effects; to improve the quality of the observation such that as much information as possible can be retrieved from it with little effort. Quite similar is the problem of feature extraction from data for, e.g., model identification, cf. [72, p. 53] or pattern recognition.

As the data processing inequality dictates, signal enhancement does *not* mean that one increases the amount of information in the observation. At best, one can build a system which *preserves* as much information as possible. Conversely, noise, distortion, and other *irrelevant* components of the observation should be removed such that the relevant information can be retrieved easily[19].

This is where relevant information loss comes in: When S is the information carrying signal and X its corrupted observation, the goal is to find a system g such that the relevant information loss is minimized while simultaneously maximizing the irrelevant information loss. One may cast the resulting optimization problem as follows:

$$\begin{aligned} \max_{g} \quad & L_{X|S}(X \to Y) \\ \text{s.t.} \quad & L_S(X \to Y) \leq C \end{aligned} \tag{4.40}$$

where C is some constant. Alternatively, one may minimize the relevant information loss subject to a lower bound on the reduction of irrelevant information, or use a variational formulation.

Let us now investigate whether a particular, variational formulation for maximizing relevant information can help to solve the abovementioned problem. Consider the following formulation of the information bottleneck method (IB) [145]

$$\min_{p(y|x)} \; I(Y;X) - \beta I(S;Y) \tag{4.41}$$

where the minimization is performed over all relations between the (discrete) RVs Y and X and where β is a design parameter[20] trading compression for preservation of relevant information, $I(S;Y)$. While in principle the relation $p(y|x)$ can be stochastic, in many cases an algorithm is used for (hard) clustering using a deterministic function. By the definition of relevant information loss,

$$I(S;Y) = I(S;X) - L_S(X \to Y) \tag{4.42}$$

[19] Note that compression often removes redundancy, especially when the compression is lossless. Redundancy can, however, be quite important when it comes to retrieving relevant information: While a bit stream might contain almost the same relevant information as a speech waveform, the human listener can extract this information more easily from the latter, redundant, representation. Hence, *relevance* and *redundancy* are not interchangable in this context.

[20] By the fact that $S - X - Y$ is a Markov chain and by the data processing inequality, $I(X;Y) \geq I(S;Y)$. As a consequence, equation (4.41) is lower bounded by $(1-\beta)I(X;Y)$, which is a non-negative quantity for $\beta < 1$. For this parameter choice the minimum is thus achieved for Y being independent of X, i.e., $I(X;Y) = 0$. This can be achieved, for example, by Y being constant. Hence, in this work it is assumed that $\beta > 1$.

where the first term is independent of $p(y|x)$.

Expressing (4.41) in terms of relevant and irrelevant information loss and with the restriction to the clustering problem, i.e., to deterministic functions $Y = g(X)$, one gets

$$g^\circ = \operatorname*{argmin}_g \; I(Y;X) - \beta I(S;Y) \tag{4.43}$$

$$= \operatorname*{argmin}_g \; -H(X) + I(Y;X) - \beta I(S;X) + \beta L_S(X \to Y) \tag{4.44}$$

$$= \operatorname*{argmin}_g \; -L(X \to Y) + \beta L_S(X \to Y) \tag{4.45}$$

$$= \operatorname*{argmin}_g \; -L_S(X \to Y) - L_{X|S}(X \to Y) + \beta L_S(X \to Y) \tag{4.46}$$

$$= \operatorname*{argmin}_g \; (\beta - 1) L_S(X \to Y) - L_{X|S}(X \to Y) \tag{4.47}$$

and thus, the optimization problem can be cast as

$$\min_g \; (\beta - 1) L_S(X \to Y) - L_{X|S}(X \to Y). \tag{4.48}$$

The information bottleneck method thus solves the signal enhancement problem by minimizing relevant and maximizing irrelevant information loss. Note that for large β stronger emphasis is placed on minimizing relevant information loss, and a trivial solution to this latter problem is obtained by any bijective g.

IB solves the signal enhancement problem not by coincidence: Minimizing relevant information loss *lies at its core* and is inherent in its formulation, as will be shown immediately. Tishby et al. [145] placed the IB method in the broader context of rate-distortion theory, with the goal to minimize

$$I(X;Y) + \beta \mathbb{E}\left(d(X,Y)\right) \tag{4.49}$$

where d is a distortion function. In other words, one wants to minimize the transmitted information while also minimizing the loss in fidelity of the representation X. As measures of distortion, the Hamming distance (i.e., the bit error rate) or the mean-squared reconstruction error are commonly chosen [21, Ch. 10]. The authors of [145] showed that the Kullback-Leibler divergence emerged as a distortion measure from comparing (4.41) and (4.49); in particular,

$$d(x,y) = D(p_{S|X}(\cdot|x)||p_{S|Y}(\cdot|y)). \tag{4.50}$$

Employing Markovity of $S - X - Y$ in (a) below and taking the expection w.r.t. the joint distribution of these three RVs, one obtains

$$\mathbb{E}\left(d(X,Y)\right) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) D(p_{S|X}(\cdot|x)||p_{S|Y}(\cdot|y)) \tag{4.51}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) \sum_{s\in\mathcal{S}} p_{S|X}(s|x) \log \frac{p_{S|X}(s|x)}{p_{S|Y}(s|y)} \tag{4.52}$$

$$\overset{(a)}{=} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} p_{X,Y}(x,y) p_{S|X,Y}(s|x,y) \log \frac{p_{S|X}(s|x)}{p_{S|Y}(s|y)} \tag{4.53}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} p_{S,X,Y}(s,x,y) \log \frac{p_{S|X}(s|x)}{p_{S|Y}(s|y)} \tag{4.54}$$

$$= H(S|Y) - H(S|X) \tag{4.55}$$

$$\overset{(a)}{=} H(S|Y) - H(S|X,Y) \tag{4.56}$$

$$= I(S;X|Y) \tag{4.57}$$
$$= L_S(X \to Y). \tag{4.58}$$

In other words, the relevant information loss *is* the distortion the IB method tries to minimize. This naturally applies to all other variants of this method, some of which will be discussed now.

The first variant considered is the agglomerative IB (AIB) method proposed in [143]. It starts with $Y = X$ and iteratively merges elements of the state space $\mathcal{Y}$ of Y such that the relevant information loss is minimized in each step, thus enforcing a functional dependence between X and Y. By stopping this algorithm as soon as during some merging step $L_S(X \to Y) > C$, at least a local optimum of the original signal enhancement problem (4.40) can be found.

Consider (4.41) with hard clustering and $\beta \to \infty$, define $[i] := \{1, \dots, i\}$, and let$[i] \to [i-1]$ denote the set of functions from $[i]$ to $[i-1]$. Let $\mathcal{X} = [N]$ and $\mathcal{Y} = [M]$. While the IB method solves

$$\min_{g \in [N] \to [M]} L_S(X \to Y) \tag{4.59}$$

the agglomerative version solves

$$\min_{g_i \in [N-i] \to [N-i-1]} L_S(Y_i \to Y_{i+1}) \tag{4.60}$$

in each iteration. Thus, employing the result about cascades, Proposition 4.3, one gets with $Y_0 = X$ and $Y_{N-M} = Y$,

$$L_S(X \to Y) = \sum_{i=0}^{N-M-1} \min_{g_i \in [N-i] \to [N-i-1]} L_S(Y_i \to Y_{i+1}). \tag{4.61}$$

Comparing (4.59) and (4.61) shows that AIB is inferior to IB.

A further variant of IB has been proposed by the authors of [17], extendeding it to incorporate knowledge about irrelevant signal components (e.g., signal components which are already known to the user and, hence, irrelevant for the current processing step). They accompany the relevant information S by an irrelevance variable $\bar{S}$, assuming conditional independence between S and $\bar{S}$ given X, and minimize the following functional:

$$I(Y;X) - \beta \left[I(Y;S) - \gamma I(Y;\bar{S}) \right] \tag{4.62}$$

At the same time, the entropy of the system output Y and the irrelevant information $I(Y;\bar{S})$ are minimized, and the relevant information $I(Y;S)$ is maximized. As in IB, β and γ are the weights for these three conflicting goals. Rewriting this in terms of relevant information loss yields

$$(\beta - 1) L_S(X \to Y) - L_{X|S}(X \to Y) - \beta\gamma L_{\bar{S}}(X \to Y). \tag{4.63}$$

Taking $\bar{S}$ as relevant information one already has (which becomes irrelevant information for the current goal of signal enhancement), this cost function reflects the desire to remove this available knowledge from the compressed representation Y. This is particularly interesting in the field of co-clustering. Moreover, unlike for IB, by employing side information it does make sense to let $\beta \to \infty$: Compression emerges naturally from maximizing the information loss w.r.t. the side information (due to $\gamma > 0$). A few other variants of the IB method and other cost functions for information-theoretic clustering are given in Table 4.1.

A similar approach has been taken up by the authors of [131,132], who introduced the following

Table 4.1: Different information-theoretic clustering algorithms and their formulation in terms of relevant information loss. S represents relevant information, while $\bar{S}$ represents irrelevant information (e.g., side information obtained from a previous clustering). A $(*)$ after the cost function indicates that it should be maximized.

Method	**Cost function**
Information Bottleneck [145]	$I(Y;X) - \beta I(S;Y)$ $(\beta-1)L_S(X \to Y) - L_{X\|S}(X \to Y)$
IB w. Side Information [17]	$I(Y;X) - \beta\left[I(Y;S) - \gamma I(Y;\bar{S})\right]$ $(\beta-1)L_S(X \to Y) - L_{X\|S}(X \to Y) - \beta\gamma L_{\bar{S}}(X \to Y)$
Conditional IB [61]	$I(Y;X) - \beta I(S;Y\|\bar{S})$ $(\beta-1)L_{S\|\bar{S}}(X \to Y) - L_{X\|S,\bar{S}}(X \to Y) - L_{\bar{S}}(X \to Y)$
Coordinated Cond. IB [62]	$I(Y;X) - \beta I(S;Y\|\bar{S}) - \gamma I(S;Y)$ $(\beta-1)L_{S\|\bar{S}}(X \to Y) - L_{X\|S,\bar{S}}(X \to Y) - L_{\bar{S}}(X \to Y)$ $+\gamma L_S(X \to Y)$
Alternative Clusterings [24]	$I(Y;X) - \eta I(Y\|\bar{S})$ $(*)$ $(\beta-1)L_{X\|\bar{S}}(X \to Y) - L_{\bar{S}}(X \to Y)$, where $\beta = 1/(\eta-1)$

information processing measure for discrete RVs X, Y, and S:

$$\Delta P(X \to Y|S) = H(X|S) - H(Y|S) - \alpha\left(H(S|Y) - H(S|X)\right) \tag{4.64}$$

The authors argue that the first difference in (4.64) corresponds to complexity reduction (i.e., reduction of irrelevant information), while the second term accounts for loss of relevant information[21]. Indeed, with

$$L_{X|S}(X \to Y) = I(X;X|Y,S) = H(X|Y,S) \tag{4.65}$$

$$= H(X|S) - H(Y|S) \tag{4.66}$$

one can rewrite $\Delta P(X \to Y|S)$ and state a variational utility function (to be maximized) for the signal enhancement problem (4.40):

$$\Delta P(X \to Y|S) = L_{X|S}(X \to Y) - \alpha L_S(X \to Y) \tag{4.67}$$

Here, $\alpha > 0$ is a design parameter trading between the loss of relevant and irrelevant information. In essence, this cost function is equivalent to the information bottleneck formulation.

Aside from the IB method and its variants widely used in machine learning, many other functions and algorithms inherently solve the problem of signal enhancement: Quantizers used in digital communication systems and regenerative repeaters have the goal to preserve the information-carrying (discrete) RV as much as possible while removing continuous-valued noise (see Example 27). Bandpass filters (although not yet directly tractable using relevant information loss) remove out-of-band noise while leaving the information signal unaltered (cf. Section 5.2). And finally, methods for dimensionality reduction (such as PCA) remove redundancy and irrelevance while trying to preserve the interesting part of the observation (cf. Section 4.4). Taking a look at these signal enhancement methods from an information-theoretic perspective, as it is done

[21] A few things have to be considered there: First of all, they argue that the complexity reduction (irrelevant information loss) can be negative for stochastic systems, which should be re-evaluated in the light of the presented results. Secondly, they argue that not only the conditional entropy can act as an uncertainty measure, but, e.g., also Bayes error. Thirdly, by concentrating on entropy, they do not take into account differential entropy and thus quantize continuous RVs to make them accessible in their framework. Finally, the authors argue that their cost function leads to PCA, Fisher discrimination, and C4.5. In their work, relevance is defined w.r.t. some goal a (neural) processing system shall achieve.

in this work, is essential; not only for a better understanding of these methods, but also for an improved design of *information* processing systems.

4.3 Application: Markov Chain Aggregation

This section is again devoted to reduction of Markov models, but the approach here is inherently different from the one in Section 3.2: There, a function of a Markov chain – a lumping – was analyzed, and conditions were presented such that the resulting process is Markov (of higher order) and/or information-preserving. The approach here is different: Here, the reduced-complexity model is required to be a first-order Markov chain, which makes it necessary to accept some information loss. Clearly, only in very limited cases a bijection between the original Markov chain and the Markov chain on a smaller state space will be possible. However, as will be shown later, one can define a reduced-complexity Markov model which minimizes information loss *relevant in a specific aspect*.

One way of reducing large Markov chains is state space aggregation. More precisely, the transition graph of the original Markov chain is aggregated to a smaller graph with a given number of nodes, through a lumping function. The aggregated process, or *aggregation*, can be any Markov process over this smaller transition graph, depending on how the transition probabilities are computed. The stochastic process obtained by projecting realizations of the original chain through the lumping function shall be called the lumped process. Ideally, the aggregated and lumped process should coincide. However, as the lumped process is generally not Markov, the aggregation "closest" to it is sought instead, where closeness has to be defined appropriately.

For the problem treated in this section, closeness is defined via the Kullback-Leibler divergence rate (KLDR) between the lumped and the aggregated process, which acts a cost function. Although, for a given lumping function, the optimal aggregation is easy to obtain (cf. [27] or Lemma 4.4 in this work), finding the optimal lumping function remains computationally expensive because it requires both an exhaustive search among all lumping functions to a given alphabet size, and an exact evaluation of the aggregation cost for each candidate function.

Here, the problem is relaxed to minimizing an upper bound on the aggregation cost instead of the exact cost. More precisely, the aggregated Markov chain is *lifted* to the original alphabet, and then compared to the original Markov chain. The KLDR between these Markov chains can be evaluated analytically [27, 121]. Further relaxing the problem allows its expression in terms of the information bottleneck method [145]. The latter method can be employed for finding a (sub-)optimal lumping function.

4.3.1 Contributions and Related Work

In control theory, state space aggregation of Markov models is an important topic: For example, White et al. analyzed aggregation of Markov and hidden Markov models in [163]. In particular, they presented a linear algebraic condition for lumpable chains (see Definition 3.4) and determined, for a given lumping function, the best aggregation in terms of the Frobenius norm. For a given transition matrix of the original Markov chain, they obtained a bi-partition of the state space via alternating projection. Aldhaheri and Khalil considered optimal control of nearly completely decomposable Markov chains and adapted Howard's algorithm to work on an aggregated model [5]. The work of Jia considers state aggregation of Markov decision processes optimal w.r.t. the value function and provides algorithms which perform this aggregation [80]. Aggregation of Markov chains with information-theoretic cost functions was considered by Deng et al. [27] and Vidyasagar [158], the first reference being the main inspiration of this section.

The idea of lifting the aggregated chain to the original state space, was used in, e.g., Deng et al. [27] and Katsoulakis et al. [86]. In [86], the authors realized that the Kullback-Leibler divergence between the resulting Markov chains provides an upper bound on the reduction

cost; however, their work is focused on continuous-time Markov chains, which makes a detailed comparison with our work difficult. Compared to [27], the approach in this work is different in the definition of the lifting and its consequences. More precisely, the lifting used here incorporates the one-step transition probabilities of the original chain, while the authors of [27] define lifting based only on the stationary distribution of the original chain. Consequently, while Deng et al. maximize the redundancy of the aggregated Markov chain, the lifting proposed here leads to a minimization of *information loss*. Moreover, the upper bound in this section is better than the upper bound obtained in [27], and it is tight in the special case where the original chain is lumpable.

The connection to spectral graph theory observed in [27] does not apply for the proposed method, to the best of the author's knowledge. More precisely, for Markov chains with strongly interacting groups of states (sometimes called nearly completely decomposable), the optimal partition of the alphabet is known to be determined by the sign structure of the Fiedler vector. Despite spectral graph theory being employed for model reduction and Markov chain aggregation for some time (e.g., [104, 130]), the authors of [27] first showed a connection between this eigenvalue-based aggregation method and an information-theoretic cost function.

In summary, by introducing a different lifting, the connection to eigenvalue-based aggregation is lost, but instead the following is obtained:

1. The lifting that minimizes an upper bound on the KLDR between the lumped process and the aggregation, subject to the requirement that the lifted chain is *lumpable*. The aggregation turns out to be the best Markov approximation for the lumped process.

2. The upper bound is tight in the special case where the original chain is strongly lumpable.

3. Minimizing the proposed upper bound minimizes information loss in a well-defined sense; Moreover, this minimization, loosely speaking, yields the partition w.r.t. which the original chain is "most lumpable".

4. A slight relaxation of the cost function allows the application of the information bottleneck method for Markov chain aggregation.

The connection to the information bottleneck method is most interesting: Recently, Vidyasagar investigated a metric between distributions on sets of different cardinalities, a problem very similar to the one considered in this work [159]. He proposed an information-theoretic metric called the variation of information, and showed that the optimal reduced-order distribution on a set of given cardinality is obtained by *aggregating* the original distribution. Specifically, the reduced-order distribution should have maximal entropy, which is equivalent to requiring that the lumping function induces the minimum information loss; a sub-optimal solution to this problem is given by the information bottleneck method (cf. Section 4.2).

In works related to (graph) clustering, information-theoretic cost functions are often used for error quantification. In particular, in [122], the authors use the information bottleneck method for partitioning a graph via assuming continuous-time graph diffusion. Moreover, in [146] and [44] pairwise distance measures between data points were used to define a stationary Markov chain, whose statistics are then used for clustering the data points. While [146] applies the information bottleneck method and obtains a result very similar to the one presented here, the authors do not describe its importance for Markov chain aggregation. In [44], the authors employ the same cost function as [27] and present an iterative algorithm similar to the agglomerative information bottleneck method. While their work focuses on pairwise clustering, they conclude by stating that their results can be employed for Markov chain aggregation as well.

4.3.2 Preliminaries

Let $\mathcal{X} = \{1, \dots, N\}$ be the alphabet of a stationary, irreducible and aperiodic Markov chain $\mathbf{X}$ with transition matrix $\mathbf{P}$ and invariant distribution μ, i.e., $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \mu)$. Let $g\colon \mathcal{X} \to \mathcal{Y}$,

where $\mathcal{Y} = \{1, \dots, M\}$, be the lumping function. Projecting $\mathbf{X}$ through the function, i.e., $Y_n := g(X_n)$, defines another stochastic process $\mathbf{Y}$, the *lumped process* $\mathbf{X}$ (see Fig. 4.3).

It is well known that $\mathbf{Y}$ is not necessarily Markov, except when $\mathbf{X}$ is lumpable (see Definition 3.4 with $k = 1$). In the latter case, $\mathbf{Y}$ is a Markov chain on $\mathcal{Y}$ with transition matrix $\mathbf{Q}$ and invariant distribution ν, i.e., $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \nu)$.

Let $\mathbf{V}$ be an $N \times M$ matrix with $V_{ij} := 1$ if $i \in g^{-1}[j]$ and zero otherwise (thus, every row contains exactly one 1). Furthermore, $\mathbf{U}^\pi$ is an $M \times N$ matrix with zeros in the same positions as $\mathbf{V}^T$, but with otherwise positive row entries which sum up to one. In other words, with π being a positive probability vector,

$$U^\pi_{ij} := \frac{\pi_j}{\sum_{k \in g^{-1}[i]} \pi_k} \tag{4.68}$$

if $j \in g^{-1}[i]$ and zero otherwise. If the superscript is omitted, let π be arbitrary.

Lemma 4.2 (Conditions for Lumpability)**.** *A stationary Markov chain* $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \mu)$ *is lumpable w.r.t.* g *if either*

$$\mathbf{VUPV} = \mathbf{PV} \tag{4.69}$$

or

$$\mathbf{U}^\mu\mathbf{PVU}^\mu = \mathbf{U}^\mu\mathbf{P}. \tag{4.70}$$

In both cases, $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \nu)$ *with* $\nu^T = \mu^T\mathbf{V}$ *and*

$$\mathbf{Q} = \mathbf{U}^\mu\mathbf{PV}. \tag{4.71}$$

Since, by assumption, $\mathbf{X}$ is stationary (i.e., its initial distribution equals the invariant distribution) one needs not distinguish between strong and weak lumpability: A chain which is weakly lumpable is also lumpable if its initial distribution is the invariant distribution (cf. [87, Thm. 6.4.3]). The corresponding result for continuous-time Markov chains on a countable alphabet has been proven in [41, Thm. 2].

Proof. For the first condition – the condition for strong lumpability in the sense of [87] – see [87, Thm. 6.3.5 & Example 6.3.3].

For the second condition, assume that

$$\mathbf{U}^\mu\mathbf{PVU}^\mu = \mathbf{U}^\mu\mathbf{P}. \tag{4.72}$$

is fulfilled. By [87, Thm. 6.4.4], it follows that the Markov chain $\mathbf{X}$ with transition matrix $\mathbf{P}$ is *weakly* lumpable w.r.t. g and that thus, if the initial distribution among the states equals the invariant distribution, the projected process $\mathbf{Y}$ is a Markov chain with transition matrix [87, Thm. 6.4.3]

$$\mathbf{Q} = \mathbf{U}^\mu\mathbf{PV}. \tag{4.73}$$

Doing a little algebra

$$\nu^T = \nu^T\mathbf{Q} = \nu^T\mathbf{U}^\mu\mathbf{PV} = \mu^T\mathbf{PV} = \mu^T\mathbf{V} \tag{4.74}$$

establishes the result about the invariant distribution. □

For the development of the results another definition is needed:

Definition 4.3 (Kullback-Leibler Divergence Rate)**.** The Kullback-Leibler divergence rate (KLDR) between two stationary stochastic processes $\mathbf{Z}$ and $\mathbf{Z}'$ on the same finite alphabet $\mathcal{Z}$ is [66, Ch. 10]

$$\bar{D}(\mathbf{Z}||\mathbf{Z}') := \lim_{n\to\infty} \frac{1}{n} D(p_{Z_1^n}||p_{Z'^n_1}) = \lim_{n\to\infty} \frac{1}{n} \sum_{z_1^n \in \mathcal{Z}^n} p_{Z_1^n}(z_1^n) \log \frac{p_{Z_1^n}(z_1^n)}{p_{Z'^n_1}(z_1^n)} \tag{4.75}$$

whenever the limit exists and if $p_{Z_1^n} \ll p_{Z'^n_1}$ for all n.

The limit exists, e.g., between a stationary stochastic process and a time-homogeneous Markov chain [66] as well as between Markov chains (not necessarily stationary or irreducible) [121]:

$$\bar{D}(\mathbf{X}||\mathbf{X}') = \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{P'_{ij}} \tag{4.76}$$

if $P'_{ij} = 0$ implies $P_{ij} = 0$ ($\mathbf{P} \ll \mathbf{P}'$).

Roughly speaking, the KLDR between a process $\mathbf{Z}$ and its model $\mathbf{Z}'$ quantifies the number of bits necessary per time step to correct the model distribution to arrive at the true process distribution. The KLDR also satisfies a data processing inequality:

Lemma 4.3. *Let $\mathbf{X}$ and $\mathbf{X}'$ be stationary, time-homogeneous, regular Markov chains with transistion matrices $\mathbf{P}$ and $\mathbf{P}'$ on the same alphabet $\mathcal{X}$. Let $\mathbf{P} \ll \mathbf{P}'$. Define two processes $\mathbf{Y}$ and $\mathbf{Y}'$ by $Y_n := g(X_n)$ and $Y'_n := g(X'_n)$, $g\colon \mathcal{X} \to \mathcal{Y}$. Let additionally $\mathbf{X}'$ be* lumpable *w.r.t. g. Then,*

$$\bar{D}(\mathbf{X}||\mathbf{X}') \geq \bar{D}(\mathbf{Y}||\mathbf{Y}'). \tag{4.77}$$

Proof. All involved processes are stationary. Since $\mathbf{P} \ll \mathbf{P}'$, $\bar{D}(\mathbf{X}||\mathbf{X}')$ exists and equals (4.76). Since $\mathbf{X}'$ is lumpable, $\mathbf{Y}'$ is a regular, time-homogeneous Markov chain. Moreover, from $\mathbf{P} \ll \mathbf{P}'$ and, thus, $P_{\mathbf{X}} \ll P_{\mathbf{X}'}$, it follows that $P_{\mathbf{Y}} \ll P_{\mathbf{Y}'}$. This ensures the existence of $\bar{D}(\mathbf{Y}||\mathbf{Y}')$ [66, Lem. 10.1].

The proof is completed by the fact that the Kullback-Leibler divergence reduces under measurements (e.g., [66, Cor. 3.3] or [116, Ch. 2.4]), i.e., that for all n,

$$D(p_{X_1^n}||p_{X'^n_1}) \geq D(p_{Y_1^n}||p_{Y'^n_1}). \tag{4.78}$$

□

4.3.3 An Information-Theoretic Aggregation Method

In order to make the notation more specific, let $\mathbf{Y}_g$ be the lumped process obtained by projecting the Markov chain $\mathbf{X}$ through the lumping function g, i.e., $Y_{g,n} := g(X_n)$.

Definition 4.4 (M-partition problem)**.** Given $\mathbf{X}$, the *M-partition problem* searches for the lumping function g such that the KLDR between the g-lumping of $\mathbf{X}$ and its best Markov approximation is minimal, i.e., it solves

$$\operatorname*{argmin}_{g\in[\mathcal{X}\to\mathcal{Y}]} \min_{\mathbf{Y}'} \{\bar{D}(\mathbf{Y}_g||\mathbf{Y}') \mid \mathbf{Y}' \text{ is Markov}\}. \tag{4.79}$$

For a fixed partition function g, this optimal $\mathbf{Y}'$, i.e., the "best" Markov approximation (in the sense of the KLDR) of the lumped process $\mathbf{Y}_g$ can be found analytically:

Lemma 4.4. *Given $\mathbf{X}$, let $\mathbf{Y}'_g$ denote the best Markov approximation of $\mathbf{Y}_g$ in the sense of the KLDR, i.e.,*

$$\mathbf{Y}'_g := \operatorname*{argmin}_{\mathbf{Y}'} \{\bar{D}(\mathbf{Y}_g||\mathbf{Y}') \mid \mathbf{Y}' \textit{ is Markov}\}. \tag{4.80}$$

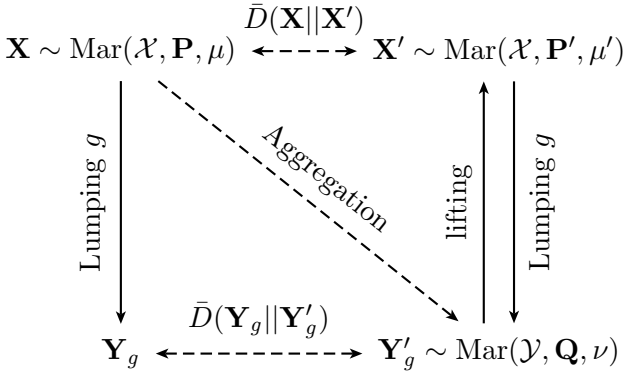


Figure 4.3: Illustration of the problem: Assume a Markov chain $\mathbf{X}$ is given. One is interested in finding an aggregation of $\mathbf{X}$, i.e., a Markov chain $\mathbf{Y}'_g$ on a partition of the alphabet of $\mathbf{X}$. This partition defines a function g (and vice-versa), which allows to define the lumped process $\mathbf{Y}_g$ via $Y_{g,n} := g(X_n)$. Note that $\mathbf{Y}_g$ need not be Markov. Lifting $\mathbf{Y}'_g$ yields a Markov chain on the original alphabet, which can be lumped to $\mathbf{Y}'_g$ using the function g.

Then, $\mathbf{Y}'_g \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \nu)$ *with* $\nu^T = \mu^T \mathbf{V}$ *and*

$$\mathbf{Q} = \mathbf{U}^{\mu} \mathbf{P} \mathbf{V}, \tag{4.81}$$

which is a matrix notation for

$$Q_{kl} = \frac{\sum_{i \in g^{-1}[k]} \sum_{j \in g^{-1}[l]} \mu_i P_{ij}}{\sum_{i \in g^{-1}[k]} \mu_i}, \quad k, l \in \mathcal{Y}. \tag{4.82}$$

From now on, $\mathbf{Y}'_g$ denotes the optimal aggregation (see Fig. 4.3) of $\mathbf{X}$, given a lumping function g. The same aggregation was declared being optimal in [162], although by using a different cost function.

Proof. See [66, Cor. 10.4] or [27, Thm. 1] and the references therein. □

One thus obtains the transition matrix $\mathbf{Q}$ of the optimal Markov model $\mathbf{Y}'_g$ from the joint distribution of two consecutive samples of $\mathbf{Y}_g$. If this joint distribution completely specifies the process $\mathbf{Y}_g$, then $\mathbf{Y}_g = \mathbf{Y}'_g$, i.e., $\mathbf{X}$ is lumpable (cf. Lemma 4.2). Note further that since $\mathbf{P}$ is the transition matrix of a regular Markov chain, so is $\mathbf{Q}$ [87, p. 140].

This suggests the definition of the aggregation error of $\mathbf{X}$ w.r.t. g:

Definition 4.5 (Aggregation error)**.** The *aggregation error* of $\mathbf{X}$ w.r.t. g is $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$.

It immediately follows that if $\mathbf{X}$ is lumpable, the aggregation error will be zero.

Following [27], one can split the M-partition problem into two sub-problems: finding the best Markov approximation of a lumped process $\mathbf{Y}_g$ (to which Lemma 4.4 provides the solution), and minimizing the aggregation error over all lumping functions g with a range of cardinality M. Thus, the optimization problem stated in (4.79) translates to finding

$$\operatorname*{argmin}_{g \in [N] \to [M]} \bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g). \tag{4.83}$$

Often, a direct evaluation of the aggregation error in Definition 4.5 is mathematically cumbersome, since $\mathbf{Y}_g$ is not necessarily Markov[22]. The authors of [27] therefore suggested to *lift*

[22] The expression for the aggregation error involves the entropy rate of a function of a Markov chain, which is a simple expression only if the lumped process is a Markov chain (of some order).

the aggregation $\mathbf{Y}'_g$ to a Markov chain $\mathbf{X}'$ over the alphabet $\mathcal{X}$, which subsequently allows a computation of the KLDR (an equivalent lifting method is suggested in [162] and [157, 158]):

Definition 4.6 (π-lifting [27, Def. 2])**.** Let $\mathbf{Y}'_g \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \nu)$ be as in Lemma 4.4 and let π be a positive probability distribution over the alphabet $\mathcal{X}$. The *π-lifting* of $\mathbf{Y}'_g$ w.r.t. g, denoted by $\mathbf{X}'^{\pi}_g$, is a Markov chain over the alphabet $\mathcal{X}$ with transition matrix

$$\mathbf{P}' := \mathbf{V}\mathbf{Q}\mathbf{U}^{\pi}, \tag{4.84}$$

which is a matrix notation for

$$P'_{ij} = \frac{\pi_j}{\sum_{k \in g^{-1}[g(j)]} \pi_k} Q_{g(i)g(j)}, \quad i, j \in \mathcal{X}. \tag{4.85}$$

Proposition 4.5 (Properties of π-lifting)**.** *The π-lifting $\mathbf{X}'^{\pi}_g$ satisfies*

1. *$\mathbf{X}'^{\pi}_g$ is lumpable w.r.t. g (and $\mathbf{Y}'_g$ is the resulting g-projection);*
2. *The invariant distribution of $\mathbf{X}'^{\mu}_g$ is μ;*
3. *$\mu = \mathrm{argmin}_{\pi} \bar{D}(\mathbf{X}||\mathbf{X}'^{\pi}_g)$;*
4. *$\mathbf{P}' \gg \mathbf{P}$;*
5. *$\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g)$.*

Properties 1), 2), and 3) appear also in [27].

Proof. See Appendix C.1. □

The last property of the previous proposition shows that the KLDR between $\mathbf{X}$ and the μ-lifting[23] $\mathbf{X}'^{\mu}_g$ provides an upper bound on the aggregation error for a given lumping function g. Unfortunately, the bound is loose in the sense that for $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) = 0$, one may have $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g) > 0$; see also [158]. One of the reasons for this disadvantage of π-lifting is that, by construction, the lifted process $\mathbf{X}'^{\pi}_g$ does not contain information about the transition probabilities between states of $\mathbf{X}$. To remove this problem, consider

Definition 4.7 ($\mathbf{P}$-lifting)**.** The *$\mathbf{P}$-lifting* of $\mathbf{Y}'_g$ w.r.t. g, denoted by $\mathbf{X}'^{\mathbf{P}}_g$, is a Markov chain over the alphabet $\mathcal{X}$ with a transition matrix $\hat{\mathbf{P}}$ given by

$$\hat{P}_{ij} := \begin{cases} \frac{P_{ij}}{\sum_{k \in g^{-1}(g(j))} P_{ik}} Q_{g(i)g(j)}, & \text{if } \sum_{k \in g^{-1}(g(j))} P_{ik} > 0 \\ \frac{1}{\mathrm{card}(g^{-1}(g(j)))} Q_{g(i)g(j)}, & \text{if } \sum_{k \in g^{-1}(g(j))} P_{ik} = 0 \end{cases}. \tag{4.86}$$

Theorem 4.1 (Properties of $\mathbf{P}$-lifting)**.** *The $\mathbf{P}$-lifting $\mathbf{X}'^{\mathbf{P}}_g$ satisfies*

1. *$\mathbf{X}'^{\mathbf{P}}_g$ is lumpable w.r.t. g (and $\mathbf{Y}'_g$ is the resulting lumped process);*
2. *$\hat{\mathbf{P}} \gg \mathbf{P}$;*
3. *$\mathbf{X}'^{\mathbf{P}}_g = \arg \min_{\hat{\mathbf{X}}: \mathbf{Y}'_g \text{ is lumping of } \hat{\mathbf{X}}} \bar{D}(\mathbf{X}||\hat{\mathbf{X}})$*
4. *$\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g)$*

[23] i.e., the π-lifting obtained by chosing π to be the invariant distribution μ of $\mathbf{X}$

5. If $\mathbf{X}$ *is strongly lumpable w.r.t.* g*, then*

$$\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) = 0 \Leftrightarrow \bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) = 0. \tag{4.87}$$

Proof. See Appendix C.2. □

The tightness result follows from the fact that for a strongly lumpable $\mathbf{X}$, the $\mathbf{P}$-lifting yields $\hat{\mathbf{P}} = \mathbf{P}$; as a direct consequence, in this case also the invariant distribution of $\hat{\mathbf{P}}$ trivially coincides with μ, the invariant distribution of $\mathbf{P}$. In general, however, the invariant distribution of $\hat{\mathbf{P}}$ differs from μ, contrasting the corresponding result for π-lifing (cf. Proposition 4.5, property 2).

Interestingly, the restriction to strongly lumpable chains for the tightness result cannot be dropped: There are Markov chains $\mathbf{X}$ which are weakly lumpable (i.e., not for all initial distributions but, e.g., for the invariant distribution) for which consequently the aggregation error vanishes, but for which the $\mathbf{P}$-lifting does not yield $\hat{\mathbf{P}} = \mathbf{P}$. A simple example for a transition matrix of such a chain is

Example 28: A weakly lumpable Markov chain.

Taking the example from [87, pp. 139] shows that the upper bound on the aggregation error is not tight in general, but only for *strongly* lumpable $\mathbf{X}$. To this end, let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{6} & \frac{5}{6} \\ \frac{7}{8} & \frac{1}{8} & 0 \end{bmatrix}. \tag{4.88}$$

This chain is lumpable w.r.t. the partition $\{\{1\},\{2,3\}\}$, but not strongly lumpable; i.e., the matrix fulfils (4.70) but not (4.69). To show that the bound is not tight, observe that $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) = 0$ but that, with

$$\hat{\mathbf{P}} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{7}{12} & \frac{5}{72} & \frac{25}{72} \\ \frac{7}{12} & \frac{5}{12} & 0 \end{bmatrix} \neq \mathbf{P} \tag{4.89}$$

one gets $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) = 0.347 > 0$.

As this theorem shows, $\mathbf{P}$-lifting yields the best upper bound on the aggregation error achievable for Markov chains over the alphabet $\mathcal{X}$. This can also be explained intuitively, by expanding the KLDR as

$$\begin{aligned} \bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) &= \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{\hat{P}_{ij}} && (4.90)\\ &= \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{\sum_{k\in\mathcal{S}_j} P_{ik}}{Q_{g(i)g(j)}} && (4.91)\\ &\overset{(a)}{=} \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{\left(\sum_{k\in\mathcal{S}_i} \mu_k\right)\left(\sum_{l\in\mathcal{S}_j} P_{il}\right)}{\sum_{k\in\mathcal{S}_i} \mu_k \sum_{l\in\mathcal{S}_j} P_{kl}} && (4.92)\\ &= H(Y_{g,n}|Y_{g,n-1}) - H(Y_{g,n}|X_{n-1}) && (4.93) \end{aligned}$$

where (a) is due to Lemma 4.4. Note that the last line corresponds to the difference between the upper and lower bounds on the entropy rate of a function of a Markov chain [21, Thm. 4.5.1]; according to Theorem 3.2, equality of these bounds implies Markovity of $\mathbf{Y}_g$, i.e., strong lumpability of $\mathbf{X}$ w.r.t. g. In other words, minimizing this cost function yields the function g for which the projected process $\mathbf{Y}_g$ is "as Markov as possible".

Example 29: Comparison of the two lifting methods.

Consider the transition matrix given in [27, Section V.A]

$$\mathbf{P} = \begin{bmatrix} 0.97 & 0.01 & 0.02 \\ 0.02 & 0.48 & 0.50 \\ 0.01 & 0.75 & 0.24 \end{bmatrix} \tag{4.94}$$

and use three different functions g inducing the following partitions of $\mathcal{X}$: $\{\{1,2\},\{3\}\}$, $\{\{1,3\},\{2\}\}$, and $\{\{1\},\{2,3\}\}$.

For all the resulting aggregations one can compute upper bounds on the aggregation error using both the μ-lifting and the $\mathbf{P}$-lifting. In addition to that, the invariant distributions $\hat{\mu}$ of the $\mathbf{P}$-lifted Markov chains $\mathbf{X}_g'^{\mathbf{P}}$ are computed and compared to μ, the invariant distribution of the original chain $\mathbf{X}$. The results are shown in the table below.

Partition	$\bar{D}(\mathbf{X}\|\|\mathbf{X}_g'^{\mu})$ bit/sample	$\bar{D}(\mathbf{X}\|\|\mathbf{X}_g'^{\mathbf{P}})$ bit/sample	$\hat{\mu}$ $[0.347, 0.388, 0.265]^T$
$\{\{1,2\},\{3\}\}$	0.823	0.185	$[0.077, 0.658, 0.265]^T$
$\{\{1,3\},\{2\}\}$	0.808	0.317	$[0.065, 0.388, 0.546]^T$
$\{\{1\},\{2,3\}\}$	0.037	0.001	$[0.347, 0.388, 0.265]^T$

As it can be seen, the partition $\{\{1\},\{2,3\}\}$ yields the best results in terms of KLDR. Moreover, it can be seen that the KLDR using $\mathbf{P}$-lifting is smaller than the KLDR using μ-lifting in all three cases, as suggested by Theorem 4.1. However, while the μ-lifting yields the same invariant distribution as the original chain has, $\mathbf{P}$-lifting obtains quite different values. An exception is the optimal partition, where $\mathbf{Y}_g$ and $\mathbf{Y}_g'$ are very close in terms of the KLDR, i.e., where $\mathbf{X}$ is "nearly" lumpable w.r.t. g.

Example 30: A "nearly" periodic Markov chain.

In another bi-partition problem the Markov chain $\mathbf{X}$ has a highly periodic structure:

$$\mathbf{P} = \begin{bmatrix} 0.03 & 0.92 & 0.00 & 0.05 \\ 0.92 & 0.03 & 0.05 & 0.00 \\ 0.00 & 0.05 & 0.03 & 0.92 \\ 0.05 & 0.00 & 0.92 & 0.03 \end{bmatrix} \tag{4.95}$$

Consider two different functions g and h, inducing the partitions $\{\{1,2\},\{3,4\}\}$ and $\{\{1,4\},\{2,3\}\}$. Both partitions lead to Markov chains, i.e., the original Markov chain $\mathbf{X}$ is strongly lumpable w.r.t. both g and h. However, the former partition leads to a nearly decomposable chain $\mathbf{Y}_g'$, while the latter leads to a chain $\mathbf{Y}_h'$ with highly periodic structure (cf. [27, Section. V.C]). The resulting transition matrices are given by

$$\mathbf{Q}_g = \begin{bmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{bmatrix}, \quad \mathbf{Q}_h = \begin{bmatrix} 0.03 & 0.97 \\ 0.97 & 0.03 \end{bmatrix}. \tag{4.96}$$

Since the aggregation error is zero for both functions and since the upper bound obtained via $\mathbf{P}$-lifting is tight in this particular case, one obtains

$$\bar{D}(\mathbf{X}||\mathbf{X}_g'^{\mathbf{P}}) = \bar{D}(\mathbf{X}||\mathbf{X}_h'^{\mathbf{P}}) = 0 \tag{4.97}$$

i.e., both aggregations are equivalent. The μ-lifting, however, is in favor of the function h, for which one obtains $\bar{D}(\mathbf{X}||\mathbf{X}_h'^{\mu}) = 0.716$ compared to $\bar{D}(\mathbf{X}||\mathbf{X}_g'^{\mu}) = 0.808$.

The reason why one function is preferred w.r.t. the other is not clear at present, except that the partition w.r.t. h leads to a Markov chain which has a higher redundancy. In addition to that, the bi-partition does not admit an interpretation via the spectral theory of Markov chains: the aggregation via h does not correspond to the Fiedler vector, but to the

- eigenvector associated with the eigenvalue with the second largest *magnitude* (cf. [27]).

4.3.4 Interpreting the KLDR as Information Loss

There is a connection between the KLDR of the original process and a lifted process (be it either by π- or $\mathbf{P}$-lifting) and the information loss induced by the lumping function g. In particular, the cost function from π-lifting is – except for its connection to spectral theory [27] – counter-intuitive, since it *maximizes* information loss. Contrarily, the cost function induced by $\mathbf{P}$-lifting *minimizes* information loss relevant in a specific aspect.

Since for the π-lifting exclusively the minimizer $\pi = \mu$ is used (cf. Proposition 4.5), this section will also refer to it as μ-lifting. Recall that the goal is to find a partition of the original alphabet $\mathcal{X}$ (induced by a functiom g) such that the KLDR is minimized:

$$g^\circ := \operatorname*{argmin}_g \bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g) \overset{(a)}{=} \operatorname*{argmax}_g \bar{R}(\mathbf{Y}'_g) \tag{4.98}$$

where (a) is due to [27, Lem. 3]: A minimization of $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g)$ can be stated as a maximization of $\bar{R}(\mathbf{Y}'_g)$, since the redundancy rate of $\mathbf{X}$ is independent of g.

Note that $H(Y'_g) = H(Y_g)$ since $\mathbf{Y}_g$ and $\mathbf{Y}'_g$ have the same marginal distribution. However,

$$\bar{H}(\mathbf{Y}'_g) = H(Y_{g,1}|Y_{g,0}) \geq \bar{H}(\mathbf{Y}_g) \tag{4.99}$$

by the fact that conditioning reduces entropy [21, Thm. 2.6.5]. With Definition 3.1, a maximization of $\bar{R}(\mathbf{Y}'_g)$ minimizes an upper bound on the entropy rate $\bar{H}(\mathbf{Y}_g)$ of the lumped process $\mathbf{Y}_g$ while simultaneously maximizing its marginal entropy. Since the entropy rate of a process is a measure of the average amount of information the process conveys in each time step, the solution to above minimization problem yields a function g° with as little information at its output as possible. In other words, one tries to *maximize information loss*. This is in line with the reasoning in [27], where the optimal solution was characterized as the model that is most "predictable". Note further that essentially the same cost function was suggested by Friedman and Goldberger [44], although the focus of their work was on pairwise clustering.

Despite this counter-intuitivity of the cost function, the model reduction method proposed by [27] works and, given some assumptions on the eigenstructure of the transition matrix[24], has a justification in spectral theory. In particular, the bi-partition problem is sub-optimally solved employing the sign structure of the eigenvector associated to the second-largest eigenvalue, the Fiedler vector. The authors of [27] then solve the M-partition problem by recursively applying the bi-partition problem, i.e., by refining the partitions iteratively.

To analyze the $\mathbf{P}$-lifting, one needs the definition of relevant information loss. In addition, let

$$g^\bullet := \operatorname*{argmin}_g \bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g). \tag{4.100}$$

Recall from (4.93) that

$$\bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) = H(Y_{g,n}|Y_{g,n-1}) - H(Y_{g,n}|X_{n-1}) \tag{4.101}$$

which, by adding and subtracting $H(Y_{g,n})$ can be rewritten as

$$\bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) = I(Y_{g,n}; X_{n-1}) - I(Y_{g,n}; Y_{g,n-1}) = L_{Y_{g,n}}(X_{n-1} \to Y_{g,n-1}) \tag{4.102}$$

where $L_{Y_{g,n}}(X_{n-1} \to Y_{g,n-1})$ is the *information loss relevant w.r.t.* $Y_{g,n}$ induced by projecting

[24] to be precise, on its *additive reversibilization*

the previous sample X_{n-1} through the function g. Finding the optimal function $g^\bullet$ thus amounts to *minimizing* information loss.

To the present date, the author could not verify if this cost function has an interpretation in spectral theory. However, as mentioned above, it minimizes the difference between first-order upper and lower bounds on the entropy rate of the lumped process $\mathbf{Y}_g$. Upper and lower bounds are equal if and only if the original process is strongly lumpable w.r.t. g (cf. Theorem 3.2). Minimizing this cost function thus amounts to making the lumped process as Markov as possible.

4.3.5 Employing the Information-Bottleneck Method for Aggregation

This section shows that the model reduction problem can be solved by the information bottleneck method [145]. Specifically, one needs to connect the results from Section 4.2, where it was shown that the information bottleneck method can be used to minimize relevant information loss, with those of Section 4.3.4. Unfortunately, in the cost function (4.102), the relevant information $Y_{g,n}$ depends on g, i.e., on the object to be optimized. Since in such a case the IB method is not applicable directly, let us relax the problem by applying Proposition 4.2:

$$L_{Y_{g,n}}(X_{n-1} \to Y_{g,n-1}) \leq L_{X_n}(X_{n-1} \to Y_{g,n-1}) \tag{4.103}$$

Instead of minimizing $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g)$, one minimizes an upper bound given by (4.103), and thus gets

$$g^{IB} := \operatorname*{argmin}_g L_{X_n}(X_{n-1} \to Y_{g,n-1}). \tag{4.104}$$

The possibility to apply the IB method and its algorithms (e.g., AIB) for model order reduction comes at the cost of optimality. This cost is not as high as one would expect, since the obtained upper bound is still better than $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g)$:

$$\begin{aligned}
L_{X_n}(X_{n-1} \to Y_{g,n-1}) &= H(X_n|Y_{g,n-1}) - H(X_n|X_{n-1}) && (4.105)\\
&= H(X_n, Y_{g,n}|Y_{g,n-1}) - \overline{H}(\mathbf{X}) && (4.106)\\
&= H(X_n|Y_{g,n}, Y_{g,n-1}) + \underbrace{H(Y_{g,n}|Y_{g,n-1})}_{=\bar{H}(\mathbf{Y}'_g)} - \overline{H}(\mathbf{X}) && (4.107)\\
&\leq H(X_n|Y_{g,n}) + \overline{H}(\mathbf{Y}'_g) - \overline{H}(\mathbf{X}) && (4.108)\\
&= H(X) - H(Y'_g) + \overline{H}(\mathbf{Y}'_g) - \overline{H}(\mathbf{X}) && (4.109)\\
&= \bar{R}(\mathbf{X}) - \bar{R}(\mathbf{Y}'_g) && (4.110)\\
&= \bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g) && (4.111)
\end{aligned}$$

where the last line is due to [27, Lem. 3].

The solution of the relaxed problem might not coincide with the solution of the original problem. To be specific: Even if a Markov chain $\mathbf{X}$ is lumpable, neither the AIB nor the IB method implementing the relaxed optimization problem necessarily find the optimal M-partition:

Example 31: Sub-Optimality of the Relaxation.
Let $\mathbf{X}$ be a Markov chain with state space $\mathcal{X}$, $N = 3$, and transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.0475 & 0.9025 & 0.05 \\ 0.9025 & 0.0475 & 0.05 \\ 0.95 & 0.05 & 0 \end{bmatrix}. \tag{4.112}$$

Consider the bi-partition problem (i.e., $M = 2$). Since this chain is lumpable for the partition

$\{\{12\},\{3\}\}$ (induced by the optimal function g°), one obtains

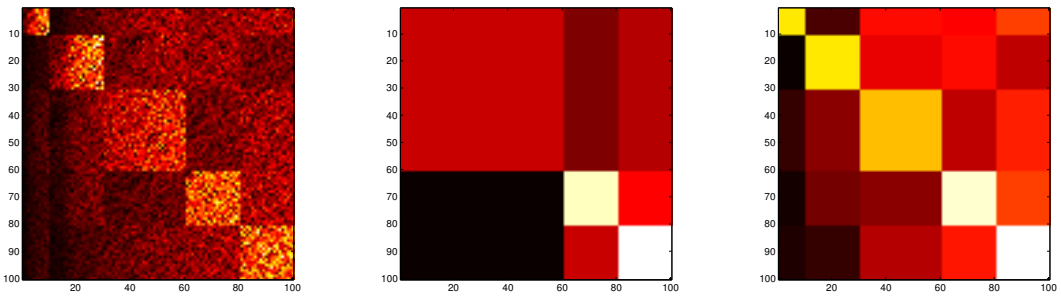
$$L_{Y_{g^\circ,n}}(X_{n-1} \to Y_{g^\circ,n-1}) = 0. \tag{4.113}$$

Computer simulations show, however, that this partition leads to a larger value of $H(X_n|Y_{g^\circ,n-1})$ than the other two options (namely, 1.19 bit compared to 0.55 and 0.69 bit, respectively). Since here the AIB and IB methods coincide (the bi-partition is obtained by merging two states), this example shows that the relaxation of the optimization problem not necessarily leads to the optimal partition.

It is interesting to observe, however, that the information bottleneck method provides the same partition function as the method introduced in [27], namely $\{\{1\},\{2,3\}\}$. The eigenvalues of the additive reversibilization of $\mathbf{P}$ are $\lambda_1 = 1$, $\lambda_2 = -0.038$, and $\lambda_3 = -0.867$, the latter two inducing the partitions $\{\{1,2\},\{3\}\}$ and $\{\{1\},\{2,3\}\}$, respectively. Hence, IB and the method in [27] respond with the solution related to the eigenvalue with the second-largest modulus, while the optimal solution remains to be related to the second-largest eigenvalue. This suggests a closer investigation of the interplay between the proposed cost function, its relaxation, and spectral theory, especially when the relevant eigenvalues are negative, cf. [27].

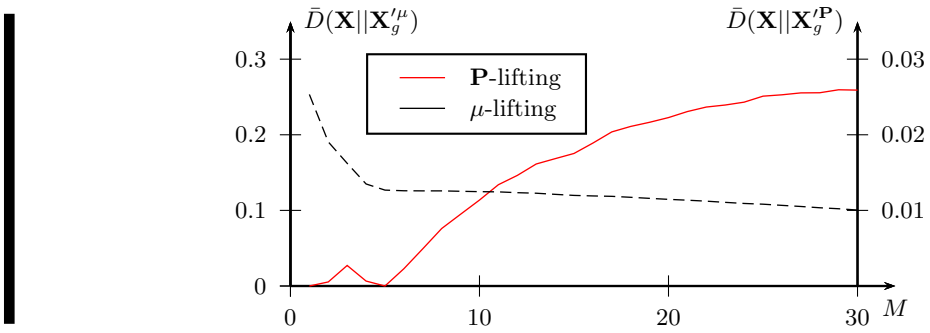
Example 32: Applying AIB to Markov aggregation.

This example concerns the transition matrix $\mathbf{P}$ from [27, Fig. 7] and uses the agglomerative information bottleneck method [143] to aggregate the chain. Simulations where done using the VLFeat Matlab implementation [154] of the agglomerative IB method. As it can be seen below, the partitions of the alphabet appear to be reasonable and, for $M = 5$, coincide with the solution obtained in [27]. In essence, the aggregation reduces the alphabet to groups of strongly interacting states. The figure below shows the original transition matrix and the transition matrices obtained by aggregation using the agglomerative information bottleneck method for $M = 3$ and $M = 5$, respectively. Blocks of the same color indicate that the corresponding states are mapped to the same output.



An interesting fact can be observed by looking at the comparison of the KLDR curves for both lifting methods (the aggregation was obtained using AIB in both cases). While for μ-lifting the KLDR seems to be a function decreasing with increasing M, the same does not hold for the $\mathbf{P}$-lifting: If a certain partition is "nearly" strongly lumpable, the KLDR curve exhibits a local minimum (cf. Theorem 4.1). Trivially, global minima with value zero are obtained for $M = 1$ and $M = N$; thus, the curve depicted below will decrease eventually if M is further increased.

These results are relevant for properly choosing the order of the reduced model: For μ-lifting, it was suggested that a change in slope of the KLDR indicates that a meaningful partition was obtained [27, Section V.D]. Utilizing the tighter bound from $\mathbf{P}$-lifting allows to choose the model order by detecting local minima.



4.3.6 Example: Models of Bio-Molecular Systems

Recent advances in measurement techniques brought the need for quantitative modeling in biology [165]. Markov models are a major tool used for modeling the stochastic nature of bio-molecular interactions in cells. However, even the simplest networks with only a few interacting species can result in very large Markov chains, in which case their analysis becomes computationally inefficient or prohibitive. In these cases, reducing the state space of the model, with minimal information loss, is an important challenge.

For a well-mixed reaction system with molecular species $S_1, \ldots, S_n$, the state of a system is typically modeled by a multiset of species' abundances $x := (x_1, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{N}_0^n$. The dynamics of such a system are determined by a set of reactions. The k-th reaction, e.g., reads

$$\nu_{1k}S_1, \ldots, \nu_{nk}S_{nk} \xrightarrow{c_k} \nu'_{1k}S_1, \ldots, \nu'_{nk}S_{nk}, \tag{4.114}$$

where $\nu_{ik} \in \mathbb{N}_0$ and $\nu'_{ik} \in \mathbb{N}_0$ denote the substrate and product stoichiometric coefficients of species i, respectively, and where c_k is the rate with which the reaction occurs. If the k-th reaction occurs, after being in the state x, the next state will be $x + (\nu'_k - \nu_k) = x + \mu_k$, where μ_k is referred to as the stoichiometric change vector.

The species multiplicities follow a continuous-time Markov chain (CTMC), where the state of the system is described by the t-indexed random vector $X(t) := (X_1(t), ..., X_n(t))$. Hence, the probability of moving to the state $x + \mu_k$ from x after time Δ is

$$\Pr(X(t + \Delta) = x + \mu_k | X(t) = x) = \lambda_k(x)\Delta + o(\Delta) \tag{4.115}$$

with λ_k the propensity of reaction k, the functional form of which is assumed to follow the principle of mass-action $\lambda_k(x) = c_k \prod_{i=1}^{n} \binom{x_i}{\nu_{ik}}$ [60]. The generator matrix $\mathbf{R}$: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the CTMC is determined by $R_{x,x+\mu_k} = \lambda_k(x)$, $R_{x,x} = -\sum_{k=1}^{r} R_{x,x+\mu_k}$, and zero otherwise.

To illustrate, assume that a gene G spontaneously turns on and off at rates c_1 and c_2 respectively, and that it regulates the expression of protein P. More precisely, whenever a gene is turned on, the protein is synthesized at a rate c_3, with $c_1, c_2 \ll c_3$. Such a system requires a stochastic model and it can be specified with the following set of reactions:

$$G_0 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} G_1, \quad G_1 + P_0 \xrightarrow{c_3} G_1 + P, \quad P \xrightarrow{c_4} P_0,$$

Here, P_0 is introduced to simplify the system so that the total number of proteins is n_P, and is arguably more realistic than the unlimited birth-death process, as P_0 represents the (limited) pool of amino-acid building blocks for the proteins. The protein can spontaneously degrade at rate c_4.

To apply the presented results to the biochemical reaction network evolving in continuous-time, one needs to work on the subordinated DTMC [109]:

Definition 4.8 (Subordinated DTMC)**.** Let $\mathbf{X}$ be a CTMC over the state space $\mathcal{X}$ with generator matrix $\mathbf{R}$ and marginal distribution $\pi_i(t) = \Pr(X(t) = i)$. The subordinated DTMC with uniformization constant $\lambda \geq \sup_{i\in\{1,\dots,N\}} |\mathbf{R}_{ii}|$ is denoted by $\mathbf{X}_\lambda$ and has the transition matrix

$$\mathbf{P} := \mathbf{R}/\lambda + \mathbf{I}_N \tag{4.116}$$

where $\mathbf{I}_N$ is an $N \times N$ identity matrix.

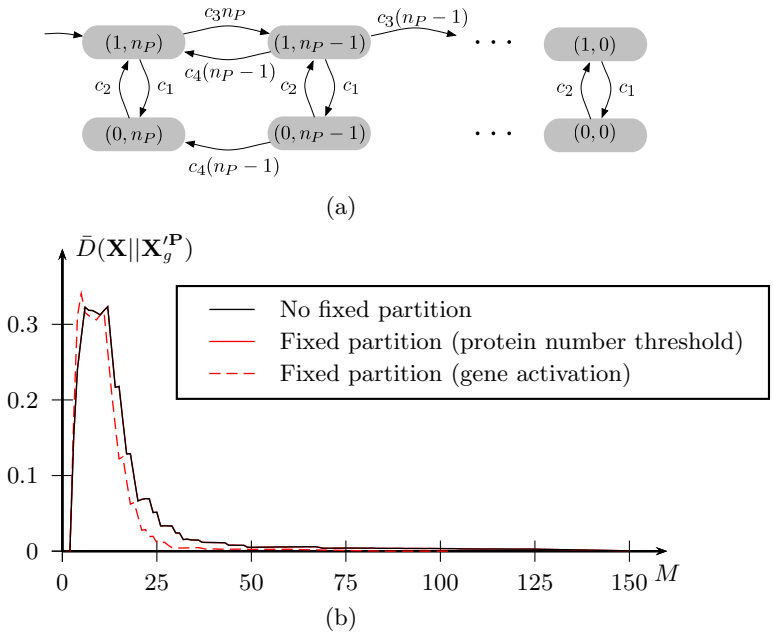


Figure 4.4: Application to modeling bio-molecular interactions: (a) The continuous-time Markov chain assigned to the gene expression example; (b) The upper bound on the KLDR obtained by the agglomerative information bottleneck method, for partition sizes up to $M = 150$: (black) no partition class is fixed, (red) a partition class, where the number of proteins is bigger than a threshold $T = 0.9n = 180$ is fixed, (red, dashed) all states where the gene is turned on (a cluster with 101 states) are fixed. The red line is not visible because it falls together with the black line.

For the initial vector $X(0) = (1, 0, n_P, 0)$, the CTMC has $N = 2(n_P + 1)$ states (Fig. 4.4(a)). After the chain exhibits stationary behavior, the algorithm is applied for $n_P = 100$ and $M = 1, 2, \dots, 202$. Moreover, the algorithm was adapted so to search for the optimal partition after one partition class is fixed. This is desirable in scenarios where the modeler a priori wants to track the joint probability of all the states that satisfy a certain property.

For example, one may be interested in clustering the states where the number of proteins is bigger than a given threshold $T = 0.9n_P$, or all the states where the gene is turned on (depicted in the top row in Fig. 4.4(a)). In Fig. 4.4(b), the upper bounds on the KLDR for the optimal partition and for the optimal partition after fixing one of the mentioned partition classes are compared. The results confirm that the proposed method provides only sub-optimal solutions, because, for example, lumping a priori all states with an activated gene indicates a better bound than when no partition class is fixed. In particular, for $M = 2$, lumping all states where the gene is turned on satisfies the criterion of strong lumpability, rendering the upper bound on the error to be within numerical precision.

4.4 Application: PCA with Specific Signal Models

This section is devoted to the analysis of PCA; it closely follows Plumbley's approach which showed that the PCA in some cases minimizes the relevant information loss [118]. At the end of

this section, Plumbley's result will be slightly extended; eventually, it will be shown that even lossless dimensionality reduction is possible in some cases.

To this end, define the PCA as

$$Y_1^M = g(X) = \mathbf{I}_M Y = \mathbf{I}_M \mathbf{W}^T X \tag{4.117}$$

where X is an N-dimensional continuous-valued input RV, Y_1^M an M-dimensional output RV (composed of the first M elements of Y), and $\mathbf{W}$ the matrix of eigenvectors of the covariance matrix of $\mathbf{X}$, $\mathbf{C}_X = \mathbb{E}\left(XX^T\right)$. Thus, $\det\mathbf{W} = 1$ and $\mathbf{W}^{-1} = \mathbf{W}^T$. Furthermore, let $\mathbf{I}_M$ be an $(M \times N)$-matrix with ones in the main diagonal. The PCA is used here for dimensionality reduction and, assuming that one has perfect knowledge of the rotation matrix $\mathbf{W}$, the *relative* information loss equals $(N - M)/N$, while the absolute information loss is infinite (cf. Section 2.6).

It is known that among all rotations prior to dimensionality reduction, the PCA minimizes the mean squared error for a reconstruction $\tilde{X} = \mathbf{W}\mathbf{I}_M^T Y_1^M$ of X [25]. In addition to that, as Linsker pointed out in [100], given that X is an observation of a Gaussian RV S corrupted by iid Gaussian noise, the PCA maximizes the mutual information $I(S;Y)$. For non-Gaussian S (and iid Gaussian noise), the PCA not only maximizes an upper bound on the mutual information, but also minimizes an upper bound on the relevant information loss [25, 118].

This section generalizes these results such that non-spherical and non-Gaussian noise is taken into account as well. Let

$$X = S + N \tag{4.118}$$

where S and N are the relevant information and the noise, respectively, with covariance matrices $\mathbf{C}_S$ and $\mathbf{C}_N$. Assume further that S and N are independent and, as a consequence, $\mathbf{C}_X = \mathbf{C}_S + \mathbf{C}_N$. Let $\{\lambda_i\}$ $\{\nu_i\}$, and $\{\mu_i\}$ be the sets of eigenvalues of $\mathbf{C}_X$, $\mathbf{C}_S$, and $\mathbf{C}_N$, respectively. Let the eigenvalues be ordered descendingly, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N. \tag{4.119}$$

One can now write

$$Y_1^M = \mathbf{I}_M \mathbf{W}^T X = \mathbf{I}_M \mathbf{W}^T S + \mathbf{I}_M \mathbf{W}^T N = \tilde{S}_1^M + \tilde{N}_1^M. \tag{4.120}$$

As mentioned before, Y_1^M is composed of the first M elements of the vector Y, which is a sum of the rotated signal vector $\tilde{S}$ and the rotated noise vector $\tilde{N}$. Conversely, let Y_{M+1}^N, $\tilde{S}_{M+1}^N$, and $\tilde{N}_{M+1}^N$ denote the last $N - M$ elements of the corresponding vectors. The covariance matrix of Y_1^M is a diagonal matrix with the M largest eigenvalues of $\mathbf{C}_X$.

Lemma 4.5. *For above signal model, the relevant information loss in the PCA is given by*

$$L_S(X \to Y_1^M) = h(Y_{M+1}^N | Y_1^M) - h(\tilde{N}_{M+1}^N | \tilde{N}_1^M). \tag{4.121}$$

Proof. Note that $L_S(X \to Y_1^M) = L_S(Y \to Y_1^M)$ since Y and X are related by an invertible transform. Thus,

$$\begin{aligned}
L_S(Y \to Y_1^M) &= I(Y;S) - I(Y_1^M;S) && (4.122)\\
&= h(Y) - h(Y|S) - h(Y_1^M) + h(Y_1^M|S) && (4.123)\\
&= h(Y_{M+1}^N|Y_1^M) - h(\mathbf{W}^T S + \mathbf{W}^T N|S) + h(\mathbf{I}_M\mathbf{W}^T S + \mathbf{I}_M\mathbf{W}^T N|S) && (4.124)\\
&\overset{(a)}{=} h(Y_{M+1}^N|Y_1^M) - h(\mathbf{W}^T N) + h(\mathbf{I}_M\mathbf{W}^T N) && (4.125)\\
&= h(Y_{M+1}^N|Y_1^M) - h(\tilde{N}_{M+1}^N|\tilde{N}_1^M) && (4.126)
\end{aligned}$$

where (a) is due to independence of S and N. This completes the proof. □

Definition 4.9 (Conditional Divergence)**.** Let X and Y be continuous RVs with arbitrary continuous (joint) distribution, and let X_G and Y_G be Gaussian RVs with same (joint) first and second moments. The conditional divergence is

$$\tilde{J}(X|Y) := \mathbb{E}\left(D(f_{X|Y}(\cdot|Y)||f_{X_G|Y_G}(\cdot|Y))\right) = h(X_G|Y_G) - h(X|Y). \tag{4.127}$$

Clearly, this quantity inherits all its properties from the Kullback-Leibler divergence, e.g., non-negativity, and can be considered as a measure of Gaussianity. Despite the similarity, this quantity is not to be confused with *negentropy*. For negentropy, instead of using a jointly Gaussian distribution for (X_G, Y_G) for computing $h(X_G|Y_G)$, one uses a Gaussian distribution for $X_G|Y = y$ to compute $h(X_G|Y = y)$ before taking the expectation w.r.t. the true marginal distribution of Y.

While most results about the optimality of the PCA are restricted to the Gaussian case, conditional divergence can be employed to generalize some of these results. In particular, one obtains

Theorem 4.2. *Let $X = S + N$, with N and S independent, and let Y_1^M be obtained by performing dimensionality-reducing PCA. If N is iid (i.e., $\mathbf{C}_N$ is a scaled identity matrix) and more Gaussian than Y in the sense*

$$\tilde{J}(\tilde{N}_{M+1}^N|\tilde{N}_1^M) \leq \tilde{J}(Y_{M+1}^N|Y_1^M) \tag{4.128}$$

then the PCA minimizes the Gaussian upper bound on the relevant information loss $L_S(X \to Y_1^M)$.

Proof. See Appendix C.3. □

This theorem generalizes the result by Plumbley [25, 118], who claimed that PCA minimizes the Gaussian upper bound on the information loss for iid Gaussian noise (i.e., for Gaussian N with $\mathbf{C}_N$ being a scaled identity matrix). In addition to that, it justifies the use of information loss instead of information transfer; an upper bound on the latter would not be useful in the signal enhancement problem.

As a next step, assume that the relevant information is concentrated on an $L \leq M$-dimensional subspace, and drop the requirement that N is iid. It is still assumed, however, that $\mathbf{C}_N$ (and, thus, $\mathbf{C}_X$) is full rank. Note that due to these assumptions $\lambda_i > 0$ and $\mu_i > 0$ for all i, while $\nu_i = 0$ for $i > L$.

Theorem 4.3 (Bounds for the PCA)**.** *Assume that S has covariance matrix $\mathbf{C}_S$ with at most rank M, and assume that N is independent of S and has (full-rank) covariance matrix $\mathbf{C}_N$. Let further N be more Gaussian than Y in the sense*

$$\tilde{J}(\tilde{N}_{M+1}^N|\tilde{N}_M) \leq \tilde{J}(Y_{M+1}^N|Y_1^M) \tag{4.129}$$

where Y_1^M is obtained by employing the PCA for dimensionality reduction. Then, the relevant information loss in nats is bounded from above by

$$L_S(X \to Y_1^M) \leq \frac{1}{2} \ln \left(\prod_{i=M+1}^{N} \frac{\mu_1}{\mu_i} \right) \tag{4.130}$$

where $\{\mu_i\}$ is the set of decreasing eigenvalues of $\mathbf{C}_N$ and where ln *is the natural logarithm.*

Proof. See Appendix C.4. □

Note that this upper bound is non-negative, since – by assumption on the ordering of the eigenvalues – any term in the product cannot be smaller than one. Along the same lines a lower

bound can be derived; by similar reasons, this bound is non-positive and thus inactive (the relevant information loss is a non-negative quantity).

The previous theorem shows that there are cases where, despite the fact that the dimensionality of the data is reduced, all of the relevant information can be preserved:

Corollary 4.3. *Assume that S has covariance matrix $\mathbf{C}_S$ with at most rank M, and assume that N is zero-mean Gaussian noise independent of S with covariance matrix $\sigma_N^2\mathbf{I}$. Let Y_1^M be obtained by employing PCA for dimensionality reduction. Then, the relevant information loss $L_S(X \to Y_1^M)$ vanishes.*

Proof. The proof follows from the fact that as $\mathbf{C}_N = \sigma_N^2\mathbf{I}$, all eigenvalues $\mu_1 = \cdots = \mu_N = \sigma_N^2$. □

As mentioned before, due to dimensionality reduction the absolute information loss $L(X \to Y_1^M)$ is infinite; a direct consequence is that the irrelevant information loss, $L_{X|S}(X \to Y_1^M)$, is infinite as well. Given the assumptions of Corollary 4.3 hold, PCA is a good solution to the signal enhancement problem:

Example 33: PCA with Non-Gaussian Data.

Assume that two independent data sources, S_1 and S_2, are observed with three sensors which are corrupted by independent, unit-variance Gaussian noise N_i, $i = 1, 2, 3$. The sensor signals shall be defined as

$$X_1 = S_1 + N_1, \tag{4.131}$$
$$X_2 = S_1 + S_2 + N_2, \text{ and} \tag{4.132}$$
$$X_3 = S_2 + N_3. \tag{4.133}$$

Assume further that the data sources have variances σ_1^2 and σ_2^2 and are non-Gaussian, but that they still can be described by a (joint) probability density function. The covariance matrix of $X = [X_1, X_2, X_3]^T$ is

$$\mathbf{C}_X = \begin{bmatrix} \sigma_1 + 1 & \sigma_1 & 0 \\ \sigma_1 & \sigma_1 + \sigma_2 + 1 & \sigma_2 \\ 0 & \sigma_2 & \sigma_2 + 1 \end{bmatrix}. \tag{4.134}$$

Performing the eigenvalue decomposition yields three eigenvalues,

$$\{\lambda_1, \lambda_2, \lambda_3\} = \{\sigma_1 + \sigma_2 + 1 + C, \sigma_1 + \sigma_2 + 1 - C, 1\} \tag{4.135}$$

where $C = \sqrt{\sigma_1^2 + \sigma_2^2 - \sigma_1\sigma_2}$.

Let us reduce the dimension of the output vector Y from $N = 3$ to $M = 2$ by dropping the component corresponding to the smallest eigenvalue. By Corollary 4.3 no relevant information is lost since the relevant information is concentrated on a two-dimensional subspace.

Note that N is iid Gaussian, and thus $h(\tilde{N}_3|\tilde{N}_1^2) = h(\tilde{N}_3)$. By assumption, $\mathbf{C}_N = \mathbf{I}$, and by the orthogonality of the transform,

$$h(\tilde{N}_3) = \frac{1}{2}\ln(2\pi\mathrm{e}). \tag{4.136}$$

Conversely, by the maximum entropy property of the Gaussian distribution,

$$h(Y_3|Y_1^2) \leq h(Y_{3,G}|(Y_1^2)_G) = h(Y_{3,G}) = \frac{1}{2}\ln(2\pi\mathrm{e}\lambda_3). \tag{4.137}$$

With $\lambda_3 = 1$ and Lemma 4.5 one thus gets

$$L_S(X \to Y_1^2) \leq 0. \tag{4.138}$$

The relevant information loss vanishes. Indeed, the eigenvector corresponding to the

smallest eigenvalue is given as $p_3 = \frac{1}{\sqrt{3}}[1, -1, 1]^T$; thus, for Y_3 one would obtain

$$Y_3 = \frac{X_1 + X_3 - X_2}{\sqrt{3}} = \frac{N_1 + N_3 - N_2}{\sqrt{3}}. \tag{4.139}$$

Since this component does not contain any relevant signal component, dropping it does not lead to a loss of relevant information.

Note that in case the noise sources N_i do not have the same variances, the application of PCA may lead to a loss of information, even though the relevant information is still concentrated on a subspace of lower dimensionality. This is made precise in the next example.

Example 34: PCA with Large Noise Variances.

Again assume that three sensors observe two independent, non-Gaussian data sources which are corrupted by independent Gaussian noise:

$$X_1 = S_1 + N_1, \tag{4.140}$$
$$X_2 = S_2 + N_2, \text{ and} \tag{4.141}$$
$$X_3 = N_3 \tag{4.142}$$

This time, however, assume that the data sources have unit variance, and that the variance of noise source N_i is i, thus $\{\mu_1, \mu_2, \mu_3\} = \{3, 2, 1\}$. With this, the covariance matrix of X is given as

$$\mathbf{C}_X = \begin{bmatrix} 1+1 & 0 & 0 \\ 0 & 1+2 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \tag{4.143}$$

and has eigenvalues $\{\lambda_1, \lambda_2, \lambda_3\} = \{3, 3, 2\}$.

Since $\mathbf{C}_X$ is already diagonal, PCA only leads to an ordering w.r.t. the eigenvalues; dropping the component of Y corresponding to the smallest eigenvalue, i.e., dimensionality reduction from $N = 3$ to $M = 2$, yields $Y_1 = N_3$ and $Y_2 = S_2 + N_2$. Since the output does not depend on S_1 anymore, information is lost – in fact, PCA suggested to drop the component of X with the highest SNR.

With Lemma 4.5 one can compute the Gaussian upper bound on the relevant information loss: Following the proof of Theorem 4.2,

$$L_S(X \to Y_1^2) \leq h(Y_{3,G}) - h(N_1) = \frac{1}{2} \ln 2. \tag{4.144}$$

The bound obtained from Theorem 4.3 is loose here, evaluating to $L_S(X \to Y_1^2) < \frac{1}{2} \ln 3$.

Since the noise is not iid, the condition of Theorem 4.2 is not fulfilled, thus the Gaussian upper bound is not minimized by the PCA. By preserving X_1 and X_2 (and dropping X_3) the relevant information loss would vanish.

4.5 Relevant Information Loss in Inaccurate Functions

In this section, the results of Section 2.3 will be extended to a practically relevant scenario: To the case where the piecewise bijective function g is not exactly known. Clearly, in such a case, which corresponds in some sense to adding noise to the signals at the system's input or output, the information loss for a continuous input RV will be infinite. If the input needs to be known only with a specific accuracy, one can at least provide bounds on the information loss relevant w.r.t. the accuracy requirements.

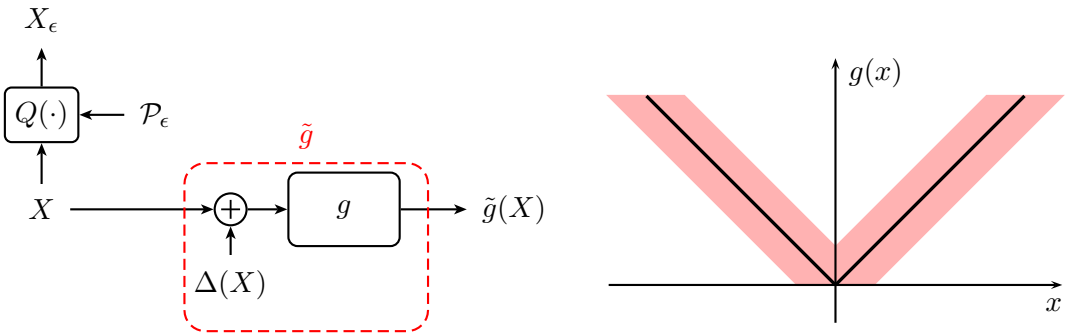


Figure 4.5: Model for the unknown function of Proposition 4.6. The shaded area in the graph on the right illustrates the uncertainty region for g being the magnitude function.

The focus on this section will be on piecewise bijective functions (according to Definition 2.3) and one-dimensional input RVs with distributions absolutely continuous w.r.t. the Lebesgue measure.

Definition 4.10 (ϵ-quantization)**.** Let $\mathcal{P}_\epsilon$ be a partition of $\mathcal{X}$ into intervals of length ϵ. An *ϵ-quantization* of X, X_ϵ, is obtained by quantizing X with the quantizer partition $\mathcal{P}_\epsilon$ where the values of X_ϵ are the interval midpoints of $\mathcal{P}_\epsilon$.

Proposition 4.6. *Let g be a PBF and let $\tilde{g}$ be an (unknown) function satisfying*

$$\forall x \in \mathcal{X} : \exists x' \in [x - \delta/2, x + \delta/2] : \tilde{g}(x) = g(x'). \tag{4.145}$$

Then, the information loss w.r.t. the ϵ-quantization X_ϵ is upper bounded by

$$L_{X_\epsilon}(X \to \tilde{g}(X)) \leq L_{\max} + \log\left(\left\lceil \frac{\delta}{\epsilon} \right\rceil + 1\right) \tag{4.146}$$

where

$$L_{\max} = \sup_{y \in \mathcal{Y}} \log \operatorname{card}(g^{-1}[y]). \tag{4.147}$$

The unknown function $\tilde{g}$ can be interpreted as one adding a random variable $\Delta = \Delta(X)$ distributed on $[-\delta/2, \delta/2]$ to the input of the known function g (see Fig. 4.5). Note that Δ need not be independent from X; in fact, Δ might even be a function of X. This way, both "noisy" functions with bounded noise and incomplete function knowledge can be modeled.

Proof. For the proof note that $L_{X_\epsilon}(X \to \tilde{g}(X)) = H(X_\epsilon|\tilde{g}(X))$ by Proposition 4.1. Applying the model from Fig. 4.5, bounding this quantity from above, and applying the chain rule one obtains

$$\begin{aligned} H(X_\epsilon|g(X+\Delta)) &\leq H(X_\epsilon, X+\Delta|g(X+\Delta)) \\ &= H(X_\epsilon|X+\Delta) + H(X+\Delta|g(X+\Delta)). \end{aligned} \tag{4.148}$$

The first quantity can be bounded from above by $\log(\lceil \delta/\epsilon \rceil + 1)$: This is the maximum number an interval of length δ can intersect a partition of ϵ-intervals. The second quantity can be bounded from above by the maximum cardinality of the preimage. This completes the proof. □

Note that $H(X+\Delta|g(X+\Delta)) \neq H(X|g(X))$, since adding this, possibly correlated, noise term Δ changes the distribution of the input. Due to the same reason, only a very coarse bound of Proposition 2.4 can be used, namely, one which does not depend on the distribution of X. In

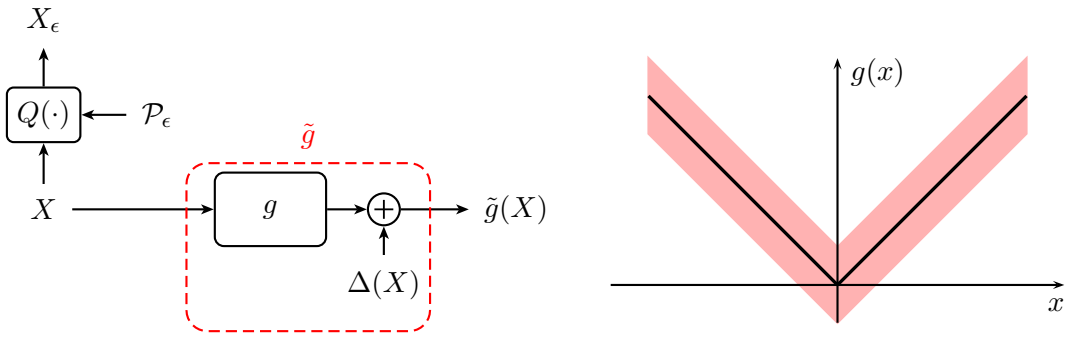


Figure 4.6: Model for the unknown function of Proposition 4.7. The shaded area in the graph on the right illustrates the uncertainty region for g being the magnitude function.

Figure 4.7: Illustration of the partitioning of the real axis into segments of length ϵ (indicated by ticks). The red bold lines indicate the preimage of the output interval, $g^{-1}[I_\delta(y)]$. The total length of the preimage is 2.8ϵ, $L'(y) = 4$. In this example, (4.156) evaluates to 10, which coincides exactly with the number of segments which intersect with the preimage.

case $\tilde{g}(X) = g(X'_\epsilon)$ for an epsilon quantization X'_ϵ possibly different from X_ϵ (i.e., the partitions are shifted w.r.t. each other) one can show

Corollary 4.4. *If $\tilde{g}(X) = g(X'_\epsilon)$, then*

$$L_{X_\epsilon}(X \to \tilde{g}(X)) \leq L_{\max} + 1. \tag{4.149}$$

In particular, if $\tilde{g}(X) = X'_\epsilon$, then

$$L_{X_\epsilon}(X \to \tilde{g}(X)) \leq 1. \tag{4.150}$$

There is a second class of inaccurately known or noisy functions one can consider: namely, one whose output differs from the output of the known function g by not more than a specific deviation. This class of functions can be interpreted as adding a random variable $\Delta = \Delta(X)$ to the output Y of the original function g, where again Δ may depend deterministically on X. For the model and an easy example see Fig. 4.6.

Proposition 4.7. *Let g be a PBF (bijective on every element of the partition $\{\mathcal{X}_j\}$ of $\mathcal{X}$) with minimum slope $\kappa := \inf_{x\in\mathcal{X}} |\tilde{g}'(x)|$ and let $\tilde{g}$ be an (unknown) function satisfying*

$$\forall x \in \mathcal{X} : |\tilde{g}(x) - g(x)| < \frac{\delta}{2}. \tag{4.151}$$

Then, the information loss w.r.t. the ϵ-quantization X_ϵ is upper bounded by

$$L_{X_\epsilon}(X \to \tilde{g}(X)) \leq \log\left(2L'_{\max} + \left\lceil \frac{L'_{\max}\delta}{\epsilon\kappa} \right\rceil - 1\right) \tag{4.152}$$

where

$$L'_{\max} = \sup_{y\in\tilde{g}(\mathcal{X})} \operatorname{card}(\{j : I_\delta(y) \cup g(\mathcal{X}_j) \neq \emptyset\}). \tag{4.153}$$

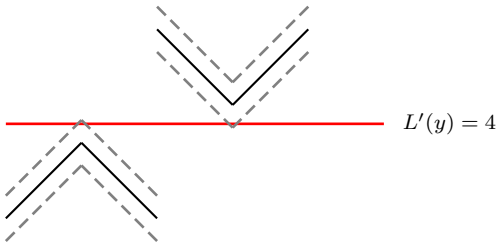


Figure 4.8: Illustration of the impact of unknown functions. If y assumes the value indicated by the red line, the preimage may intersect up to four elements $\mathcal{X}_i$ of the partition upon which g is piecewise bijective, thus $L'(y) = L'_{\max} = 4$. Contrarily, $L_{\max} = 2$.

Proof. For the proof let λ denote the Lebesgue measure. By knowing the output $y = \tilde{g}(x)$, one knows the output $g(x)$ with an accuracy of $\pm\delta/2$. One can now look at the preimage of the interval $I_\delta(y) := [y - \delta/2, y + \delta/2]$, $g^{-1}[I_\delta(y)]$. If the preimage is an interval itself, i.e., if g is invertible on $I_\delta(y)$, then the preimage intersects with at most

$$\left\lceil \frac{\lambda(g^{-1}[I_\delta(y)])}{\epsilon} \right\rceil + 1 \tag{4.154}$$

ϵ-quantization intervals. If the preimage consists of multiple intervals, the total Lebesgue measure of the preimage remains $\lambda(g^{-1}[I_\delta(y)])$, but this might now be spread over $L'(y)$ intervals, where

$$L'(y) := \mathrm{card}(\{j : I_\delta(y) \cup g(\mathcal{X}_j) \neq \emptyset\}). \tag{4.155}$$

The maximum number of intersections with the ϵ-quantization intervals is then given by (see Figs. 4.7 and 4.8)

$$\left\lceil \frac{\lambda(g^{-1}[I_\delta(y)])}{\epsilon} \right\rceil + 1 + 2(L'(y) - 1). \tag{4.156}$$

One further has, with g_j being the restriction of g to $\mathcal{X}_j$,

$$\lambda(g^{-1}[I_\delta(y)]) = \sum_j \lambda(g_j^{-1}[I_\delta(y)]) \leq L'(y)\frac{\delta}{\kappa} \tag{4.157}$$

by the fact that the slope is bounded from below. With $L'_{\max} = \sup_{y\in\tilde{g}(\mathcal{X})} L'(y)$ one thus gets

$$H(X_\epsilon|\tilde{g}(X)) \leq \log\left(2L'_{\max} + \left\lceil \frac{L'_{\max}\delta}{\epsilon\kappa} \right\rceil - 1\right). \tag{4.158}$$

□

In particular, above result holds if $\tilde{g}(X)$ is a δ-quantization of $g(X)$. If $\tilde{g}(X) = X'_\epsilon$, the result of Corollary 4.4 is recovered with $\delta = \epsilon$, $L'_{\max} = 1$, and $\kappa = 1$ ($g(x) = x$). Note, finally, that $L'_{\max} \geq L_{\max}$, since the former degenerates to the latter for $\delta = 0$ only.

4.6 Open Questions

Relevant information loss was introduced as a measure for signal enhancement, connecting the problem to clustering methods developed by the machine learning community. However, very little signal enhancement was done in this chapter, except of course in the application of PCA in Section 4.4. The reason is that most other signal enhancement methods work on processes, not

random variables, so part of this topic was deferred to (and almost exclusively constitutes) the following chapter. However, the author fears that the importance of relevant information loss for signal enhancement is presently underestimated in the field of signal processing, and that significant examples of its usability have to be given in the near future.

Markov chain aggregation by employing the information bottleneck method is one of the main contributions of the present author. However, the method can still be improved by applying the information bottleneck method in some iterative scheme, where the relevant signal is exchanged after every iteration; this way, part of the sub-optimality pointed out in Section 4.3.5 can be overcome. Moreover, it would be interesting if the new cost function – the relevant information loss – has an interpretation in terms of spectral theory of Markov chains.

Lindqvist [98] analyzed the information lost about the initial state of a Markov chain if the chain is observed at some later time, although he did not quantify this loss in information-theoretic terms. For a specific type of Markov chain this was also considered in Section 2.3.5, where the information loss in an accumulator was analyzed. Lindqvist also analyzed the case in which only a function of a Markov chain is observed [99]. To complement his statistical analysis with an information-theoretic one, the notion of relevant information loss could be used. Specifically, the quantities of interest could be $L_{p_{X_0}}(p_{X_n} \to p_{Y_n})$ or $L_{X_0}(X_n \to Y_n)$, if one is interested in the *additional* loss incurred by the function. If, and how, these quantities relate to the statistical measures used by Lindqvist would be of great interest.

Yet another open question is the applicability of PCA to non-additive signal models; while PCA was shown to be a smart choice for data superimposed with iid Gaussian noise, the same may not be the case if the data is affected by multiplicative noise or nonlinear distortion. Since PCA is a widely used method, finding signal models in which this method is justified from an information-theoretic point-of-view is of great importance. Equally interesting is an information-theoretic analysis of nonlinear methods for dimensionality reduction, such as kernel PCA, independent components analysis [25], or Kohonen's self-organizing feature maps.

5 Relevant Information Loss Rate

The contents of this section have been exclusively developed by the present author – the inspiration for it mainly coming from Gernot Kubin, who from the very beginning desired an information-theoretic analysis of the Wiener filter. In Section 5.3, which constitutes a large portion of [57], the Wiener filter is briefly analyzed in Example 38. The results of Section 5.2 are also available in [56] and [53] and have been presented at the IEEE Forum on Signal Processing for Radio Frequency Systems 2013 in Linz.

5.1 A Definition, its Properties, and a Simple Upper Bound

This section contains the last generalization of information loss in this work: The generalization of relevant information loss from random variables to stationary stochastic processes. This most general scenario is clearly the one where one needs to take greatest care to guarantee that the introduced quantities exist. Moreover, while it is also the scenario with the greatest practical importance, it is hard to obtain results for large *classes* of systems, as in Chapters 2 and 3. Therefore, in what follows, only two specific applications are considered: That of anti-aliasing filtering in Section 5.2 and that of prefiltering for quantizers in Section 5.3. This focus on design of linear filters gives an information-theoretic justification for the standard, i.e., linear filtering approach to the problem of signal enhancement. At least, this justification will be given in specific signal model scenarios.

Before presenting a definition of *relevant information loss rate*, the notion of information rate shall be introduced, cf. [66, 116]:

Definition 5.1 (Information Rate)**.** The information rate between two jointly stationary stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ is

$$\bar{I}(\mathbf{X};\mathbf{Y}) := \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n) \tag{5.1}$$

whenever the limit exists.

The limit exists, e.g., when at least one of the process alphabets is finite, or when at least one of the processes has a finite redundancy rate $\lim_{n\to\infty} I(X_n; X_1^{n-1})$, cf. [66, Thm. 8.3, p. 229].

Definition 5.2 (Relevant Information Loss Rate)**.** Let $\mathbf{S}$ and $\mathbf{X}$ be jointly stationary stochastic processes with at least one of them having finite redundancy rate, and let $\mathbf{Y}$ be a process defined

by $Y_n := g(X_n)$. Then, if $\bar{I}(\mathbf{S};\mathbf{Y}) < \infty$, the information loss rate relevant w.r.t. $\mathbf{S}$ is

$$\bar{L}_{\mathbf{S}}(\mathbf{X} \to \mathbf{Y}) := \lim_{n\to\infty} \frac{1}{n} L_{S_1^n}(X_1^n \to Y_1^n) = \bar{I}(\mathbf{S};\mathbf{X}) - \bar{I}(\mathbf{S};\mathbf{Y}). \tag{5.2}$$

If $\mathbf{S}$ has finite redundancy rate, then both involved information rates exist. If $\mathbf{X}$ has finite redundancy rate, then so has $\mathbf{Y}$ (see [161] for the discrete-alphabet case; in general, the result can be obtained via the data processing inequality) – again, both information rates exist. Because, by assumption, the limit $\frac{1}{n} I(S_1^n; Y_1^n)$ exists and is finite, the limit of the difference equals the difference of limits; hence the second equality.

The relevant information loss rate inherits many properties from relevant information loss discussed in Section 4.1.1. In particular, $\bar{L}_{\mathbf{S}}(\mathbf{X} \to \mathbf{Y}) \leq \bar{L}(\mathbf{X} \to \mathbf{Y})$, and the property for cascades from Proposition 4.3.

It is currently not clear whether also the information loss rate can be split into a relevant and an irrelevant portion, similar to Corollary 4.2. The reason is that the *irrelevant process*, $\mathbf{X}|\mathbf{S}$, need not be stationary for all conditioning events. Moreover, it is not yet clear whether the following result holds for all jointly stationary processes $\mathbf{S}$ and $\mathbf{X}$:

Lemma 5.1. *Let $\mathbf{S}$ and $\mathbf{X}$ be two jointly stationary processes satisfying Assumption 1 on page 88, and let them be connected by a memoryless channel, i.e., $S_1^n - S_i - X_i - X_1^n$ is a Markov tuple for all n and $1 \leq i \leq n$. Then, for $Y_n := g(X_n)$,*

$$\bar{L}_{\mathbf{S}}(\mathbf{X} \to \mathbf{Y}) \leq L_S(X \to Y) \tag{5.3}$$

where S, X, and Y are RVs with the same joint distribution as S_n, X_n, and Y_n for arbitrary n.

Proof. The assumption that $S_1^{i-1} - S_i - X_i$ is a Markov tuple and the fact that Y_1^{i-1} is a function of X_1^{i-1} gives equality in (a) below

$$L_{S_1^n}(X_1^n \to Y_1^n) = I(X_1^n; S_1^n | Y_1^n) = \sum_{i=1}^{n} I(X_i; S_1^n | X_1^{i-1}, Y_1^n) \overset{(a)}{=} \sum_{i=1}^{n} I(X_i; S_i | X_1^{i-1}, Y_i^n). \tag{5.4}$$

Since $S_i - X_i - X_1^n - (X_1^{i-1}, Y_i^n) - Y_i$ is a Markov tuple, for all $0 \leq i \leq n$,

$$\begin{aligned} I(X_i; S_i | X_1^{i-1}, Y_i^n) &= I(S_i; X_1^i, Y_i^n) - I(S_i; X_1^{i-1}, Y_i^n) && (5.5)\\ &= I(S_i; X_i) - I(S_i; X_1^{i-1}, Y_i^n) && (5.6)\\ &\leq I(S_i; X_i) - I(S_i; Y_i) && (5.7)\\ &= L_S(X \to Y). && (5.8) \end{aligned}$$

Thus $L_{S_1^n}(X_1^n \to Y_1^n) \leq n L_S(X \to Y)$, which completes the proof. □

The larger portion of this chapter is devoted to the case where the involved processes – $\mathbf{S}$ and $\mathbf{X}$ – have an absolutely continuous distribution. Specifically, it is assumed that both $\mathbf{S}$ and $\mathbf{X}$ satisfy Assumption 1 on page 88, i.e., they have finite differential entropy, finite differential entropy rate, and finite Shannon entropy of their quantized samples. From this assumption it immediately follows that the redundancy rate

$$\bar{R}(\mathbf{X}) := \lim_{n\to\infty} I(X_n; X_1^{n-1}) = h(X) - \bar{h}(\mathbf{X}) \tag{5.9}$$

exists and is finite. Moreover, Lemma 3.4 can be extended to a more meaningful version, essentially proving that linear filters do not change the information content of a signal, given that they are stable and causal:

Lemma 5.2 (Filters don't hurt)**.** *Let $\mathbf{S}$ and $\mathbf{X}$ be two jointly stationary stochastic processes satisfying Assumption 1, and let H be a stable, causal linear, time-invariant (LTI) filter. Then,*

for $\tilde{\mathbf{X}}$ *being the filtered version of* $\mathbf{X}$*,*

$$\bar{I}(\mathbf{S};\mathbf{X}) = \bar{I}(\mathbf{S};\tilde{\mathbf{X}}). \tag{5.10}$$

Proof. See Appendix D.1. □

The fact that the linear filter does not change the information content is closely related to the Paley-Wiener condition [114, p. 423]: A stable, causal LTI filter[25] satisfies

$$\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln|H(\mathrm{e}^{\jmath\theta})|^2 d\theta > -\infty \tag{5.11}$$

which is exactly the term by which the differential entropy rates change under linear filtering [114, p. 663]. Therefore, the relevant information content cannot be changed with stable, causal LTI filters.

Nonlinear systems, as it was extensively argued in this work, *can* change the information content of a signal. The influence on stationary stochastic processes is hard to compute in general; it is therefore beneficial to generalize the bound of Proposition 4.4 to processes as well:

Corollary 5.1 (Gaussian Upper Bound; Corollary to Proposition 4.4)**.** *Let* $\mathbf{S}$ *and* $\mathbf{X}$ *be jointly Gaussian, jointly stationary processes satisfying Assumption 1, and let* $\mathbf{Y}$ *be a process defined by* $Y_n := g(X_n)$*, where* g *is piecewise bijective. Let further* $\mathbf{Y}_G$ *(*$\mathbf{Y}_G, \mathbf{S}$*) be a Gaussian process with the same (joint) PSD as* $\mathbf{Y}$ *(*$\mathbf{Y}, \mathbf{S}$*). Then,*

$$\bar{L}_{\mathbf{S}}(\mathbf{X} \to \mathbf{Y}) \leq \bar{L}_{\mathbf{S}}(\mathbf{X} \to \mathbf{Y_G}). \tag{5.12}$$

The corollary follows from applying Proposition 4.4 to $\frac{1}{n}L_{S_1^n}(X_1^n \to Y_1^n)$ and taking the limit.

5.2 Application: Anti-Aliasing Filter Design

Multi-rate systems are ubiquitously used in digital systems to increase (upsample) or decrease (downsample) the rate at which a signal is processed. Especially downsampling is a critical operation since it can introduce aliasing, like sampling, and thus can cause information loss. Standard textbooks on signal processing deal with this issue by recommending an anti-aliasing filter prior to downsampling – resulting in a cascade which is commonly known as a decimator [111, Ch. 4.6]. In these books, this anti-aliasing filter is usually an ideal low-pass filter with a cut-off frequency of π/M, for an M-fold decimation system (cf. Fig. 5.1). Unser showed that this choice is optimal in terms of the mean-squared reconstruction error only if the input process is such that the pass-band portion of its power spectral density (PSD) exceeds all aliased components [149]. Similarly, it was shown by Tsatsanis and Giannakis [147], that the filter minimizing the mean-squared reconstruction error is piecewise constant, M-aliasing-free (i.e., the aliased components of the M-fold downsampled frequency response do not overlap), and has a pass-band depending on the PSD of the input process. Specifically, the filter which permits most of the energy to pass aliasing-free is optimal in the MSE sense.

In Section 3.5 it was argued that anti-aliasing filtering is futile from an information-theoretic point-of-view, since no filter can reduce the amount of information lost in a decimation system. The situation drastically changes if one has the signal model of Fig. 5.1 in mind: If one is not interested in minimizing the information loss rate in general, but only the information loss rate *relevant* w.r.t. some process $\mathbf{S}$, it can indeed be shown that an anti-aliasing filter can be helpful. This section is devoted to dealing with this topic, also for the practically relevant case where the input process $\mathbf{X}$ is the sum of a signal process $\mathbf{S}$ and a noise process $\mathbf{N}$ jointly stationary with $\mathbf{S}$. The notation, specifically for M-fold blocked processes, is adopted from Section 3.5.

[25] A filter with a band-limited frequency response (as, e.g., an ideal low-pass filter) cannot satisfy this condition (and thus, does not admit a causal implementation).

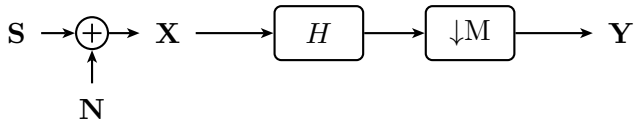


Figure 5.1: A simple decimation system consisting of a linear filter H and an M-fold downsampler. Note that now the process $\mathbf{X}$ at the input of the decimation system is the sum of two jointly stationary processes, where one of them represents the relevant information.

There is a tight connection to the problem of signal enhancement in Section 4.2: Anti-aliasing filters are used to remove spectral components of a signal that are either irrelevant for the user (e.g., out-of-band noise) or that would otherwise make a retrieval of the remaining information difficult (e.g., aliased components of the information signal). Hence, these filters simultaneously maximize loss of irrelevant, and minimize loss of relevant information. It was already shown in Theorem 3.3 together with Proposition 3.15 that $\bar{L}(\mathbf{X}^{(M)} \to \mathbf{Y}) = \infty$. Hence, if one can show that the relevant information loss rate $\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y})$ can be minimized (or even bounded) by proper filtering, this will suggest that anti-aliasing filters are a good solution to the signal enhancement problem. This is exactly the goal of this section.

Definition 5.3 (Energy Compaction Filter [150, Thm. 4])**.** The optimal energy compaction filter H for an M-fold downsampler and for a given PSD $S_X(\mathrm{e}^{\jmath\theta})$ satisfies, for all $k \in \{0, \ldots, M-1\}$,

$$H(\mathrm{e}^{\jmath\theta_l}) = \begin{cases} 1, & \text{if } l \text{ is the smallest natural number such that } S_X(\mathrm{e}^{\jmath\theta})|_{\theta=\theta_l} \geq S_X(\mathrm{e}^{\jmath\theta_k}) \\ 0, & \text{else} \end{cases} \tag{5.13}$$

where $\theta_k := \frac{\theta - 2k\pi}{M}$.

In essence, the construction of an energy compaction filter proceeds as follows: For each frequency $\theta = \theta_0 \in [0, 2\pi/M)$ the alias frequencies θ_k, $k = 1, \ldots, M$ are computed. Then, the PSD $S_X(\mathrm{e}^{\jmath\theta})$ is evaluated at these M frequencies; the index l corresponding to the maximum (or the first maximum, if there are multiple) indicates that θ_l is within the filter's pass-band, while all other indices $k \neq l$ are within the filter's stopband. According to [150], the thus designed filter is piecewise constant and M-aliasing-free, i.e., its pass-bands do not overlap during M-fold downsampling. The total bandwidth of the energy compaction filter is $2\pi/M$.

Theorem 5.1. *Assume $\mathbf{S}$ and $\mathbf{N}$ are jointly stationary, independent, Gaussian processes satisfying Assumption 1 and having sufficiently smooth PSDs $S_S(\mathrm{e}^{\jmath\theta})$ and $S_N(\mathrm{e}^{\jmath\theta})$, respectively. Consider the multirate system depicted in Fig. 5.1. The energy compaction filter for $S_S(\mathrm{e}^{\jmath\theta_k})/S_N(\mathrm{e}^{\jmath\theta_k})$ minimizes the relevant information loss rate.*

Proof. See Appendix D.2. □

Optimal energy compaction filters are also part of the theory of optimal filterbanks; in particular, the result presented here is strongly related to *principal component filter banks* (PCFBs) introduced by Tsatsanis and Giannakis[26] [147] (but see also [150]). Recently, Chen et al. analyzed the capacity of sub-Nyquist sampled, continuous-time additive Gaussian noise channels with frequency response $H_{\text{channel}}(f)$ in [18]. They showed that the capacity of the channel depends on the (continuous-time) anti-aliasing filter $H_c(f)$, and that the maximizing filter is the energy compaction filter for $|H_{\text{channel}}(f)|^2/S_N(f)$, where $S_N(f)$ is the PSD of the continuous-time noise process [18, Thm. 3].

[26] Interestingly, in the introduction of [147] another example of the common misconception of energy as information can be found; the authors claim: "The filter bank is designed so that *most of the signal's information* is optimally concentrated in the first few channels, by employing an appropriate 'energy compaction' criterion" (emphasis added).

As observed by Akkarakaran and Vaidyanathan in [4, Sec. IV], the optimal energy compaction filters for the ratio $S_S(\mathrm{e}^{\jmath\theta_k})/S_N(\mathrm{e}^{\jmath\theta_k})$ is the energy compaction filter for $S_S(\mathrm{e}^{\jmath\theta})$ whenever either $\mathbf{N}$ is white or when $S_N(\mathrm{e}^{\jmath\theta}) = cS_S(\mathrm{e}^{\jmath\theta})$ for some positive constant c. Moreover, the authors analyzed the optimality of the PCFB for noise suppression in the case of colored noise, under the assumption that both the input and the noise PSD are piecewise constant with all discontinuities lying at rational multiples of π. The PCFB for $\mathbf{N}$ was shown to minimize the total transmit power for discrete multitone modulation with colored noise in [151]. Filter banks were also analyzed from an information-theoretic point-of-view in, e.g., [133], where the authors derived the optimum FIR filters to maximize the mutual information rate for block transmission.

5.2.1 A Gaussian Bound for Non-Gaussian Processes

As Theorem 5.1 shows, under the assumption of a Gaussian signal model, the energy compaction filter minimizes the relevant information loss rate, or equivalently, maximizes the information rate. The analysis is now extended to a non-Gaussian signal superimposed with Gaussian noise. It can be shown that by modelling $\mathbf{S}$ as a Gaussian process and designing the filter H accordingly, not only an upper bound on the information rate $\bar{I}(\mathbf{S}^{(M)};\mathbf{Y})$ is maximized, but that also an upper bound on the relevant information loss rate is minimized. The principle underlying this proof is the maximum-entropy property of the Gaussian distribution. The inspiration for finding an upper bound on the relevant information loss rate came from Plumbley's work [118], which was extended here from finite-length vectors to stochastic processes.

Theorem 5.2 (Gaussian Bound on Relevant Information Loss Rate)**.** *Assume* $\mathbf{S}$ *and* $\mathbf{N}$ *are jointly stationary, independent processes satisfying Assumption 1 and having PSDs* $S_S(\mathrm{e}^{\jmath\theta})$ *and* $S_N(\mathrm{e}^{\jmath\theta})$*, respectively. Let* $\mathbf{N}$ *be Gaussian, and let* $\mathbf{S}_G$ *denote a stationary Gaussian process, independent from* $\mathbf{N}$*, with PSD* $S_S(\mathrm{e}^{\jmath\theta})$*. Consider the multirate system depicted in Fig. 5.1, where the LTI filter* H *is stable and causal. The relevant information loss rate is bounded by*

$$\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) \leq \bar{L}_{\mathbf{S}_G^{(M)}}(\mathbf{X}_G^{(M)} \to \mathbf{Y}_G) \tag{5.14}$$

where $X_{G,n} := S_{G,n} + N_n$*, and where* $\mathbf{Y}_G$ *is obtained by filtering* $\mathbf{X}_G$ *with* H *and downsampling by a factor of* M*.*

Proof. See Appendix D.3. □

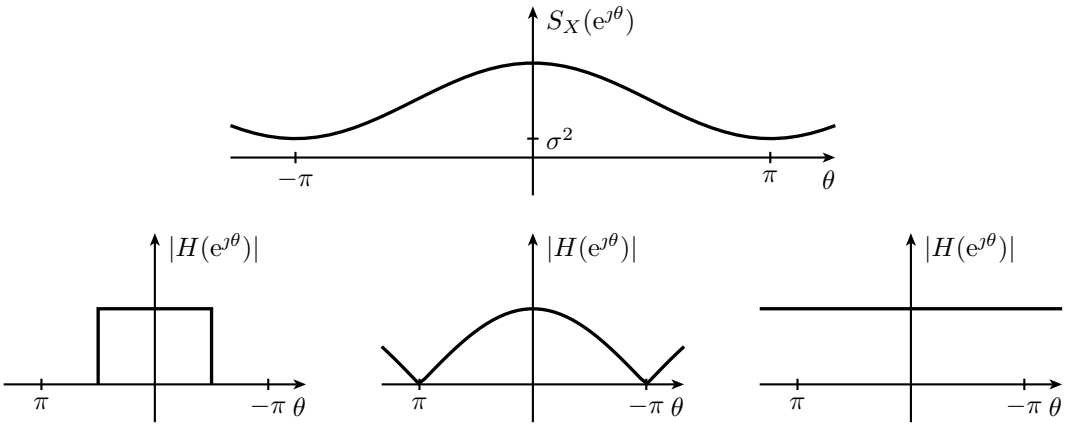
Note that the restriction that H is stable and causal cannot be dropped immediately: Applying Lemma D.1 as in the proof of Theorem 5.1 is not possible, since the sub-band processes may be dependent despite being uncorrelated.

The main statement of this result is that, under the assumption of a specific signal model, filter design according to energetic considerations might perform satisfactorily under information-theoretic cost functions. In particular, assuming a non-Gaussian signal superimposed by Gaussian noise, the theorem shows that the optimal energetic design minimizes an upper bound on the relevant information loss rate. Neither does this guarantee that the optimal energetic filter coincides with the information-theoretic optimum, nor is it certain that the obtained filter is "good" in the sense of destroying little relevant information (the difference between the upper bound and the true loss rate is in general not easy to compute). However, an energetic design grants a performance guarantee in the sense that the relevant information loss rate is bounded.

Generally, while this section treats the case of decimation systems, the results can be generalized to sampling of bandlimited, continuous-time processes by considering the equivalent decimation system for the critically sampled process. The author conjectures that the extension is also possible to non-bandlimited processes, since by the notion of relevance even in this case the information loss can remain finite. The extension to processes with PSDs vanishing on intervals shall be possible with an appropriate filterbank decomposition.

The extension to continuous-time processes (bandlimited or not) is connected to [18], where Chen et al. considered the capacity of an additive Gaussian noise channel (and hence did not specify the PSD $S_S(e^{\jmath\theta})$ of the input signal). Quite interestingly, Chen et al. showed that, depending on the channel transfer function, a filterbank sampling mechanism can have a strictly larger capacity than a single-channel sampling mechanism. In the light of this work, this suggests that one can further reduce the relevant information loss in the downsampler by replacing the filter H by a filterbank.

Example 35: Anti-Aliasing Filters are not that useless.



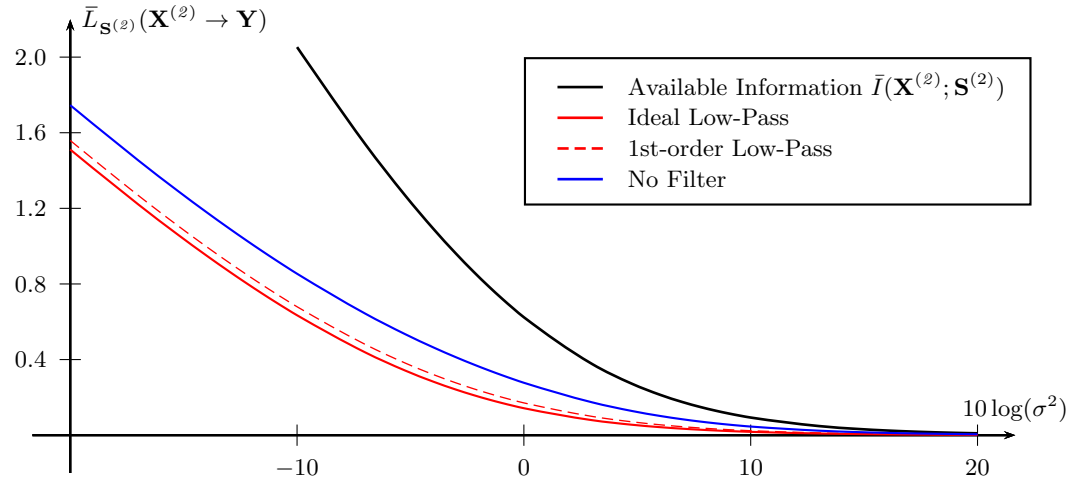
Let $\mathbf{X}$ be a stationary Gaussian process with PSD $S_X(e^{\jmath\theta}) = \sigma^2+1+\cos(\theta)$ depicted above for $\sigma^2 = 0.25$. This time it is assumed that $\mathbf{X}$ is the sum of two independent Gaussian processes $\mathbf{S}$ and $\mathbf{N}$ with PSDs $S_S(e^{\jmath\theta}) = 1+\cos(\theta)$ and $S_N(e^{\jmath\theta}) = \sigma^2$, respectively. Specifically, let $\mathbf{S}$ be a signal relevant to the observer and $\mathbf{N}$ a white-noise process corrupting the signal. The noisy observation $\mathbf{X}$ is fed through a filter H and a downsampling device with $M = 2$ (see Fig. 5.1). To minimize the *relevant* information loss rate across the downsampler, H can be any of the following three options (see figure above):

(a) An ideal anti-aliasing low-pass filter with cut-off frequency $\theta_c = \frac{1}{2}$

(b) An FIR low-pass filter satisfying the Paley-Wiener conditions

(c) A direct connection, i.e., $H(e^{\jmath\theta}) = 1$ for all θ.

By Theorem 5.1, option (a) is the optimal choice, since the resulting energy compaction filter minimizes the relevant information loss rate among all filter options. It can be shown that the first-order low-pass filter with impulse response $h[n] = \delta[n] + \delta[n-1]$ (shown as the second option above) is optimal among all first-order filters by minimizing the relevant information loss rate. Omitting the filter completely yields the worst result among the three options, leading to significant information loss (see figure below).



It is interesting to observe that the first-order low-pass filter performs not much worse than the ideal filter, suggesting that already a simple filter can gain a significant improvement compared to downsampling without filtering (which in this case is probably due to the strong decay of $S_S(\mathrm{e}^{\jmath\theta})$ towards $\theta = \pm\pi$). Moreover, as the figure suggests, the higher the noise variance σ^2 is, the smaller is the relevant information loss rate – simply because there is less information available to lose.

While this example was elaborated for a Gaussian signal model, Theorem 5.2 shows that the obtained loss rates act as upper bounds for the case of a non-Gaussian signal process $\mathbf{S}$ superimposed by Gaussian noise $\mathbf{N}$.

5.2.2 FIR Solutions for Information Maximization

While the ideal filter with unconstrained order is simple to obtain (cf. Theorem 5.1), the practically more relevant case of finite-order filters is much more difficult to solve even in the purely Gaussian case: The problem of maximizing (cf. Lemma D.1 in Appendix D.2)

$$\bar{I}(\mathbf{S}_G^{(M)};\mathbf{Y}_G) = \bar{I}(\tilde{\mathbf{S}}_G^{(M)};\mathbf{Y}_G) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{\sum_{k=0}^{M-1} S_S(\mathrm{e}^{\jmath\theta_k})|H(\mathrm{e}^{\jmath\theta_k})|^2}{\sum_{k=0}^{M-1} S_N(\mathrm{e}^{\jmath\theta_k})|H(\mathrm{e}^{\jmath\theta_k})|^2}\right) d\theta \tag{5.15}$$

does, except in particularly simple cases (see Example 35), not permit a closed-form solution, nor is it necessarily convex.

The situation simplifies when the noise is white, i.e., when $S_N(\mathrm{e}^{\jmath\theta}) = \sigma_N^2$, and with the restriction that the energy compaction filter satisfies the Nyquist-M condition [90, 150]

$$\frac{1}{M}\sum_{k=0}^{M-1} \left|H(\mathrm{e}^{\jmath\theta_k})\right|^2 = 1. \tag{5.16}$$

This restriction is meaningful, e.g., when the energy compaction filter is part of an orthonormal filter bank or a principal component filter bank.

Employing these restrictions and applying Jensen's inequality in (5.15) yields an upper bound on the information rate

$$\bar{I}(\tilde{\mathbf{S}}_G^{(M)};\mathbf{Y}_G) \leq \frac{1}{2}\ln\left(1 + \frac{1}{2\pi\sigma_N^2}\int_{-\pi}^{\pi}\frac{1}{M}\sum_{k=0}^{M-1} S_S(\mathrm{e}^{\jmath\theta_k})|H(\mathrm{e}^{\jmath\theta_k})|^2 d\theta\right) \tag{5.17}$$

$$\overset{(a)}{=} \frac{1}{2}\ln\left(1 + \frac{1}{2\pi\sigma_N^2}\int_{-\pi}^{\pi} S_S(\mathrm{e}^{\jmath\theta})|H(\mathrm{e}^{\jmath\theta})|^2 d\theta\right) \tag{5.18}$$

$$= \frac{1}{2}\ln\left(1 + \frac{\sigma_{\tilde{S}}^2}{\sigma_N^2}\right) = \frac{1}{2}\ln\left(1 + \frac{\sigma_{\tilde{S}}^2}{\sigma_{\tilde{N}}^2}\right) \tag{5.19}$$

where (a) is because the variance of a stationary process does not change during downsampling and where $\sigma_{\tilde{S}}^2$ ($\sigma_{\tilde{N}}^2$) is the variance of $\tilde{\mathbf{S}}$ ($\tilde{\mathbf{N}}$), the output of H to the input process $\mathbf{S}$ ($\mathbf{N}$).

Maximizing an upper bound on the information rate thus amounts to maximizing the signal-to-noise ratio

$$\mathsf{SNR} = \frac{\sigma_{\tilde{S}}^2}{\sigma_{\tilde{N}}^2} = \frac{\int_{-\pi}^{\pi} S_S(\mathrm{e}^{\jmath\theta})|H(\mathrm{e}^{\jmath\theta})|^2 d\theta}{\int_{-\pi}^{\pi} S_N(\mathrm{e}^{\jmath\theta})|H(\mathrm{e}^{\jmath\theta})|^2 d\theta} \tag{5.20}$$

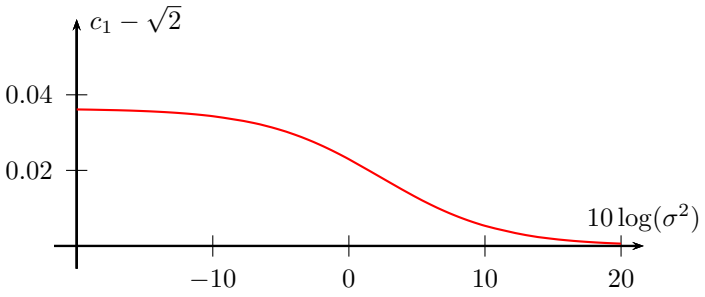
or, equivalently, the signal power, at the output of the downsampler or filter. This is exactly the objective of optimum FIR compaction filters for $S_S(\mathrm{e}^{\jmath\theta})$, which have been investigated in [90] and the references therein. The solution for filter orders strictly smaller than the downsampling factor M is the eigenvector associated with the largest eigenvalue of the autocorrelation matrix [90];

for larger filter orders, various analytical and numerical methods exist; see [148] for an overview. As a particular example, note that the first-order FIR filter minimizing the relevant information loss rate in Example 35 also maximizes the filter output SNR for the given signal PSD.

Obviously, for a given PSD $S_S(\mathrm{e}^{\jmath\theta})$ the upper bound is the better the larger the noise variance σ_N^2 is, because the non-constant terms have less influence on the argument of the logarithm. Hence, energetic design considerations will succeed especially in cases where the Gaussian noise is white and has a large variance; see also Example 36. One has to keep in mind, however, that the problem of FIR filters is solved only sub-optimally, since FIR energy compaction filters only *maximize an upper bound on the information rate*; the desired result would be, however, either a lower bound on the information rate or an upper bound on the relevant information loss rate.

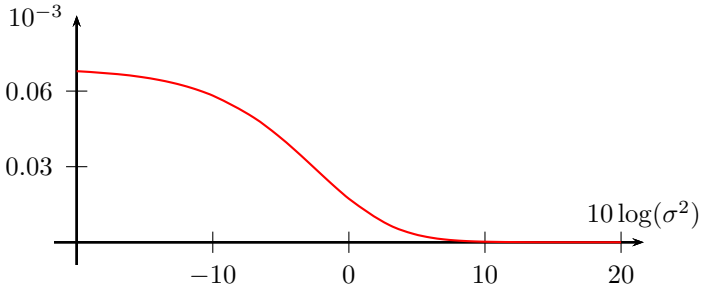
Example 36: A Three-Fold Downsampler and its Optimal FIR Filter.

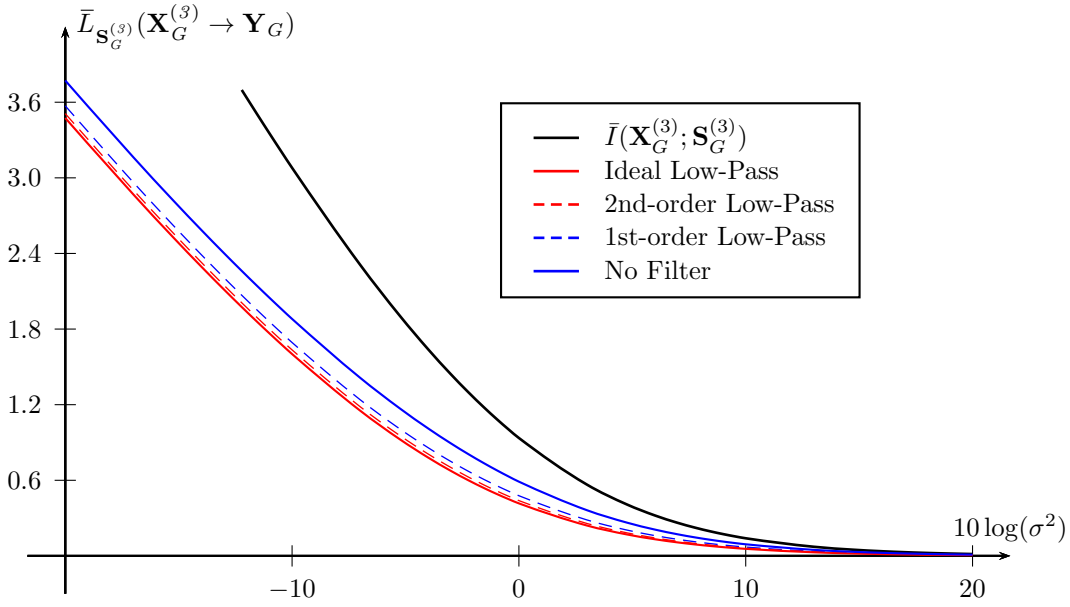
The analysis of Example 35 is repeated here with the same PSDs but with a three-fold downsampler, i.e., for $M = 3$. For the Gaussian assumption, the first-order FIR filter with impulse response $h[n] = \delta[n] + \delta[n-1]$ again proved optimal. Here, however, also the optimal filter coefficients for a second-order FIR filter with impulse response $h[n] = \delta[n] + c_1\delta[n-1] + c_2\delta[n-2]$ were determined numerically. Remarkably, for all considered variances, the optimum satisfies $c_2 = 1$. The optimal value for c_1 depends on the variance σ of the noise process, as indicated in figure below.



The filter coefficient is close to $\sqrt{2}$, which corresponds to the impulse response vector being equal to the eigenvector associated with the largest eigenvalue of the input process' autocorrelation matrix, and hence to the solution maximizing the filter output signal-to-noise ratio (see above). While the difference diminishes for large noise variance, for strong signals the coefficient is significantly different. This clearly illustrates that energetic and information-theoretic designs are inherently different, and one can hope to have similar solutions for both cost functions only in few, specialized scenarios. Knowing whether such a scenario applies or not is of prime importance for the system designer, since it could admit simple energetic design approaches to circumvent the need for non-linear, non-convex optimization to achieve the information-theoretic optimum. Moreover, in such a scenario the actual choice of the cost function – energetic or information-theoretic – is immaterial, since the solution will anyway be the same (or at least similar) in both cases. This is especially interesting in applications where the best cost function for a given task is not (yet) uncontroversially agreed upon, such as, e.g., in speech enhancement.

The additional loss induced by replacing the ideal coefficient c_1 by $\sqrt{2}$ is shown below; as can be seen, the additional loss is negligible, which – in this case – justifies energetic design considerations from an information-theoretic point-of-view.





The figure above shows the (upper bound on the) relevant information loss rate in nats as a function of the noise variance σ^2 for various filter options. For comparison, the available information rate $\bar{I}(\mathbf{X}_G^{(3)};\mathbf{S}_G^{(3)}) = 3\bar{I}(\mathbf{X}_G;\mathbf{S}_G)$ is plotted. One can observe that the loss is greater than for two-fold downsampling. Note that the curve for the optimal second-order FIR filter almost falls together with the curve of the ideal low-pass filter.

5.3 Application: Filtering Prior to Quantization

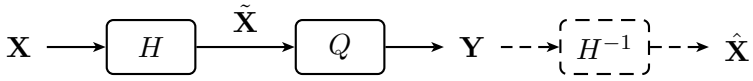


Figure 5.2: Simple quantization system consisting of a linear filter H and an entropy-R constrained uniform quantizer Q.

This section considers the problem of linear filtering prior to quantization, specifically, filter design using an information-theoretic cost function. Clearly, such a design is beneficial for digital communication systems, where the goal is to increase information rates or to decrease bit error rates. However, also in digital transmission of speech such a design may be justified: There, the information sink – the human listener – expects an analog signal which, consequently, has to be reconstructed from the digital one. Interestingly, the filter minimizing the reconstruction error variance has a poor perceptual performance [152, Ch. 8]. Apparently, there are other, non-energetic aspects of a signal which are relevant for speech quality. Among these aspects is, e.g., the spectral shape of the reconstruction error, which allows to exploit the masking property of the human hearing system.

The problem of this section – quantization – is well-treated in the literature, both from an energetic and an information-theoretic perspective. While most commonly the quantizer is designed to minimize the mean-squared reconstruction error (MSRE) [101], also the mutual information rate has been considered for quantizer design [142, 174]. Linear pre-processing prior to quantization has been investigated, e.g., in [63]. Joint optimization w.r.t. both MSRE and information rate is considered in rate-distortion theory [21, Ch. 10], [88, 89, 173].

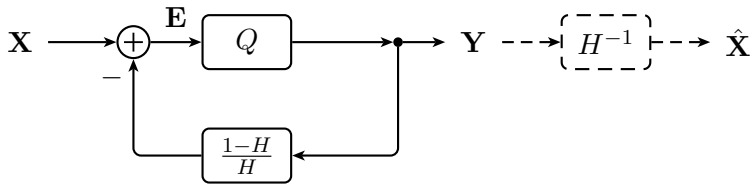


Figure 5.3: Closed-loop Prediction.

Talking about reconstruction, another interesting aspect of information-theoretic cost functions comes into view: Information, once lost, cannot be recovered; it is, therefore, sufficient to maximize the information rate over the quantizer, i.e., over the nonlinear device. To optimize the linear prefilter, it is not necessary to provide a method for reconstructing the source signal, as it usually is for energetic measures like the MSRE. In some cases reconstruction may not even be necessary in practice: Sticking to the example of digital speech transmission, an automatic speech recognition system could, theoretically, work equally well (or even better?) on the output of the digital transmission channel as on the reconstructed signal, provided it is designed accordingly.

Now consider the problem of this section: filtering prior to quantization as illustrated in Fig. 5.2. The amount of information lost in the quantizer is infinite, even 100% (cf. Example 1). However, the information rate between the source and the quantizer output is finite this time. It is, therefore, possible to maximize it by optimizing the filter accordingly. Although this chapter is devoted to the relevant information loss rate, parts of this section will be formulated using the information rate only. Moreover, for the sake of simplicity of the presented results, instead of the binary logarithm the natural logarithm will be used throughout this section.

Let us assume that the source process is Gaussian. Modeling also the quantization noise as a white Gaussian process not only simplifies the analysis, but also provides a lower bound on the information rate for high-resolution quantizers:

Lemma 5.3. *Consider the uniform quantization system depicted in Fig. 5.2. If $\tilde{\mathbf{X}}$ is a stationary Gaussian process with PSD $S_X(\mathrm{e}^{\jmath\theta}) \geq \epsilon > 0$ and satisfies Assumption 1, for a sufficiently small quantizer bin size Δ the information rate is bounded from below by*

$$\bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}) \geq \bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}_G) \tag{5.21}$$

where $\mathbf{Y}_G$ is the sum of $\tilde{\mathbf{X}}$ and independent, Gaussian noise with variance $\Delta^2/12$.

Proof. See Appendix D.4. □

Of course, the optimality results shown below depend again strongly on the validity of the signal model. In particular, one cannot expect that prefiltering a non-Gaussian process with nonlinear dependence structure will always increase the information rate at the quantizer output.

5.3.1 Optimal Prefilters for Quantization

Let us now investigate the effect of prefiltering on the information rate over a uniform quantizer (see Fig. 5.2). To this end, assume the quantizer is entropy-constrained and satisfies the high-rate assumption, i.e., that it can be modeled by an uncorrelated, white noise source with a variance depending linearly on the variance $\sigma^2_{\tilde{X}}$ of $\tilde{\mathbf{X}}$, cf. [65].

It can be shown using elementary information theory that the information rate $\bar{I}(\mathbf{X}; \mathbf{Y})$ equals the entropy rate $\overline{H}(\mathbf{Y})$ of the discrete-valued output process. For a fixed marginal distribution of

$\mathbf{Y}$ (or $\tilde{\mathbf{X}}$, respectively), the entropy rate is maximized if the process is a sequence of iid random variables; in the Gaussian case, the maximum is achieved if $\tilde{\mathbf{X}}$ is white. This suggests that the optimal filter H is a linear predictor, a *whitening filter*.

To make this precise, let us model the quantizer as an additive white Gaussian noise channel. While this assumption is uncommon, it holds for high-dimensional vector quantizers, cf. [173]. In addition to that, this assumption provides a lower bound on the information rate for a Gaussian input process $\mathbf{X}$ for high quantizer resolutions and an input PSD satisfying $S_X(\mathrm{e}^{\jmath\theta}) \geq \epsilon > 0$ a.e., as shown in Lemma 5.3. Thus, let $\mathbf{X}$ be a stationary, zero-mean Gaussian process with PSD $S_X(\mathrm{e}^{\jmath\theta}) > 0$ a.e. and variance σ_X^2. Note that the PSDs of $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ are given by $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) = S_X(\mathrm{e}^{\jmath\theta})|H(\mathrm{e}^{\jmath\theta})|^2$ and $S_Y(\mathrm{e}^{\jmath\theta}) = S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) + \sigma_{\tilde{X}}^2\gamma$, respectively, where $\gamma = 1/(e^{2R}-1)$ and R is the entropy constraint of the quantizer. The information rate between $\mathbf{X}$ and $\mathbf{Y}$ evaluates to [116]

$$\bar{I}(\mathbf{X};\mathbf{Y}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})}{\sigma_{\tilde{X}}^2\gamma}\right) d\theta. \tag{5.22}$$

Applying Jensen's inequality yields

$$\bar{I}(\mathbf{X};\mathbf{Y}) \leq \frac{1}{2}\ln\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} 1 + \frac{S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})}{\sigma_{\tilde{X}}^2\gamma} d\theta\right) = \frac{1}{2}\ln\left(1+\frac{1}{\gamma}\right) = R \tag{5.23}$$

with equality if and only if the argument of the logarithm is constant. In other words, the maximum information rate is achieved if H is a perfect whitening filter, or an *infinite-order linear open-loop predictor*. Note that such a linear predictor not only ensures that $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})$ is constant, but also that the mean-squared prediction error $\sigma_{\tilde{X}}^2$ is minimized. Then, $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) = \sigma_\infty^2$, where σ_∞^2 is the prediction error of the infinite-order predictor, or the entropy power of $\mathbf{X}$.

This result is interesting when compared to other aspects of linear, open-loop prediction. Specifically, assume that the quantizer output $\mathbf{Y}$ is filtered by H^{-1}, and that the resulting process $\hat{\mathbf{X}}$ shall approximate $\mathbf{X}$. When investigating the MSRE or the signal-to-quantization-noise ratio, [152, Ch. 8] shows that the infinite-order linear predictor cannot improve these two quantities compared to omitting the filter ($H \equiv 1$). If these energetic measures are taken as cost functions, the optimal filter turns out to be *half-whitening* [108]. Naturally, this filter performs sub-optimally in terms of the information rate, since the input to the quantizer is not white. Consequently, with this open-loop prediction scheme, energetic and information-theoretic cost functions in general cannot be optimized simultaneously.

The question remains whether it is possible to meet both design goals with closed-loop prediction, i.e., when the predictor operates on the quantized signal as shown in Fig. 5.3. To this end, recall that with Lemma 5.2, $\bar{I}(\mathbf{X};\mathbf{Y}) = \bar{I}(\mathbf{X};\hat{\mathbf{X}})$ if $\hat{\mathbf{X}}$ is obtained by filtering $\mathbf{Y}$ with a stable and causal filter H^{-1}. To calculate the information rate for closed-loop prediction, one can thus apply the error identity stating that $\hat{\mathbf{X}} = \mathbf{X} + \mathbf{Q}$, where $\mathbf{Q}$ is the white Gaussian quantization noise process. Since the quantizer is now in the prediction loop, prediction is based on noisy samples and the prediction error σ_E^2 satisfies $\sigma_E^2 > \sigma_{\tilde{X}}^2$. The difference between these variances becomes small for high quantizer resolutions [108, pp. 1505]. Consequently, the variance of $\mathbf{Q}$ is $\sigma_E^2\gamma$ and the information rate is

$$\bar{I}(\mathbf{X};\hat{\mathbf{X}}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{S_X(\mathrm{e}^{\jmath\theta})}{\sigma_E^2\gamma}\right) d\theta \tag{5.24}$$

where σ_E^2 is the only term that depends on the filter H. Minimizing the prediction error σ_E^2 hence not only minimizes the MSRE, but also maximizes the information rate.

That closed-loop prediction is superior to open-loop prediction in terms of MSRE is well-known [108,152]; but also its information-theoretic properties have been investigated. The focus of the relevant works, however, is mainly on comparing it to the rate-distortion function, a

theoretical lower bound on the number of bits one must send over a channel to reconstruct the input process with a given distortion (see, e.g., [21, Ch. 10]). It is known, for example, that an autoregressive Gaussian process and its innovation process have the same MSRE rate-distortion function. Kim and Berger showed that open-loop prediction cannot *achieve* this function [88] and they quantified the additional rate necessary to achieve the desired distortion for first-order autoregressive processes [89]. Quite contrarily, with proper pre- and post-filtering, the rate-distortion function is achievable by closed-loop prediction [173].

5.3.2 FIR Prefilters for Quantization

While the case is relatively simple for optimal infinite-order filters, the problem is less clear for finite impulse response (FIR) filters. Specifically, is the filter minimizing the prediction error also optimal in terms of information rates? While (5.24) gives an affirmative answer for the closed-loop predictor, for open-loop prediction the optimal solution will be different (except, of course, in the important case where $\mathbf{X}$ is an autoregressive process of the same or smaller order as the filter).

A non-trivial exception is the case of a high entropy constraint R, which leads to an emphasis of the fraction $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})/\sigma^2_{\tilde{X}}$ in the argument of the log in (5.22). Thus,

$$\bar{I}(\mathbf{X};\mathbf{Y}) \approx \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(\frac{S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})}{\sigma^2_{\tilde{X}}\gamma}\right) d\theta = \frac{1}{2}\ln\left(\frac{\sigma^2_{\infty}}{\sigma^2_{\tilde{X}}\gamma}\right). \tag{5.25}$$

Maximizing the information rate for large R hence is equivalent to minimizing $\sigma^2_{\tilde{X}}$, the prediction error.

This result is interesting from a practical point-of-view: While the FIR filter minimizing the prediction error variance is obtained from the autocorrelation properties of $\mathbf{X}$ [70, Ch. 3], the filter maximizing the information rate involves nonlinear optimization. Closed-form solutions are, if at all, only possible for particularly simple examples. The result now implies that, at least for high-resolution quantization, this nonlinear optimization will not lead to significant performance gains compared to just minimizing the prediction error variance.

That, in general, the energetically optimal FIR filter is different from the information-theoretic solution reminds of the design of anti-aliasing filters for downsampling systems in Section 5.2, in particular Example 36. That the same effect appears to be present here highlights again the difference between information-theoretic and energetic cost functions. It is important to understand in which cases these two design procedures lead to similar or completely different results.

In case both designs agree at least to some extent, the choice of the cost function has little influence on the performance of the system. If the resulting designs are significantly different, however, one has to take care which cost function one chooses for a specific application. Naturally, the author is clearly in favor of information-theoretic cost functions, simply due to the fact that in most cases one wishes to transmit *information*. In any case, it is important to know in which cases the information-theoretic cost function really provides an advantage, or more importantly, in which case the energetic cost function completely fails. At least for quantization (and decimation in the last section), some preliminary answers are now put on solid ground.

Example 37: Filtering prior to Quantization.

Let $\mathbf{X}$ be a first-order moving-average process with the process generating difference equation $X_n = W_n + W_{n-1}$, where $\mathbf{W}$ is a unit-variance, zero-mean, white Gaussian innovation process. The PSD of $\mathbf{X}$ is thus $S_X(\mathrm{e}^{\jmath\theta}) = 4\cos^2(\theta/2)$, and $\sigma^2_X = 2$.

Assume further that H is either first or second order, i.e., its impulse response in vector notation is either $h^{(1)} = [h_0^{(1)},\ h_1^{(1)}]^T$ or $h^{(2)} = [h_0^{(2)},\ h_1^{(2)},\ h_2^{(2)}]^T$. The filter coefficients

minimizing the mean-squared prediction error (PE) can be computed to [70, Ch. 3]:

$$h_{\mathrm{PE}}^{(1)} = [1,\ -1/2]^T \text{ with } \sigma_1^2 = 1.5 \tag{5.26a}$$

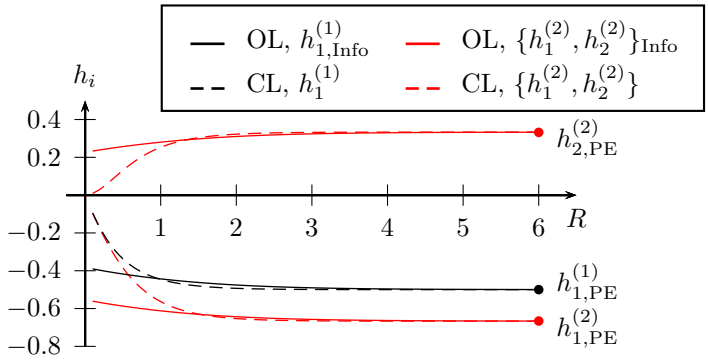
and

$$h_{\mathrm{PE}}^{(2)} = [1,\ -2/3,\ 1/3]^T \text{ with } \sigma_2^2 = 1.333 \tag{5.26b}$$

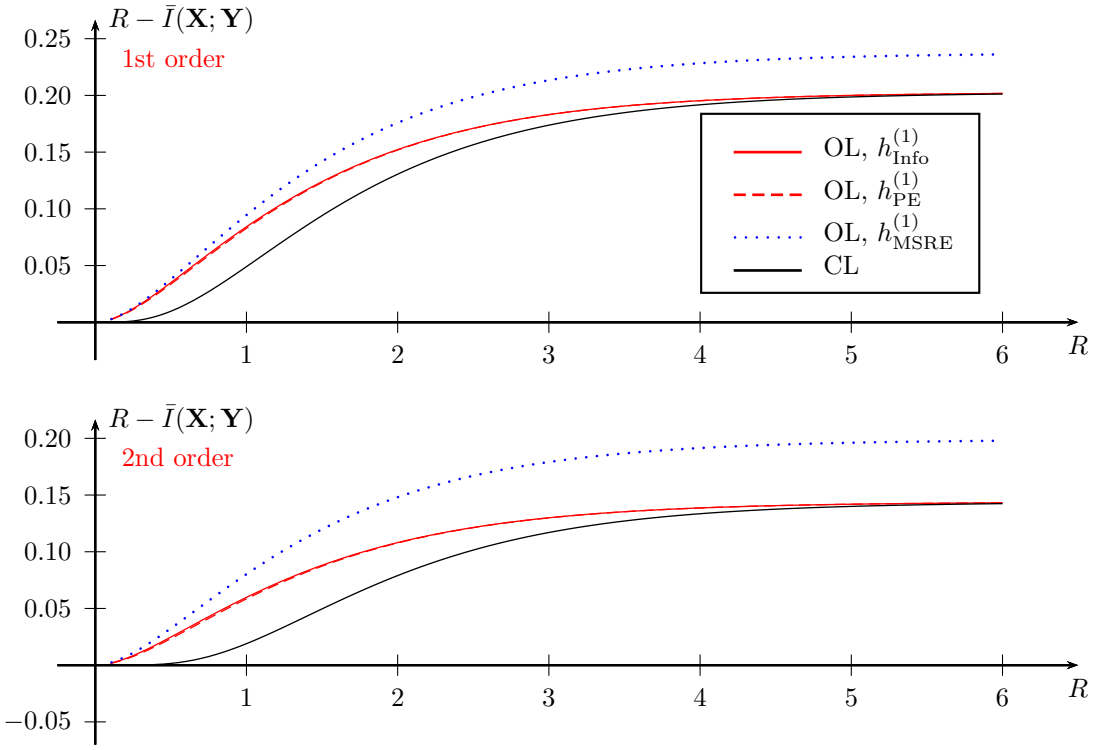
where σ_L^2 denotes the PE of the L-th order predictor. Clearly, for the infinite-order predictor, $\sigma_\infty^2 = 1$, the variance of the innovations process.

The information rates for first- and second-order open-loop predictors were designed with different cost functions: maximizing information rate, minimizing MSRE, and minimizing PE. For the closed-loop predictor the filter coefficients were chosen to minimize the PE σ_E^2, hence are optimal for each of the three cost functions, cf. Sections 5.3.1 and 5.3.2.

The quantizer entropy constraint was varied from $R = 0.1$ to $R = 6$. For each value of R, the optimal filter coefficients were obtained numerically. W.l.o.g., the first coefficient was set to unity, since a simple gain has no influence on the information rate (the quantization noise variance depends on the quantizer input variance). The figure below shows the coefficients as a function of R. It can be seen that for large R the information-theoretic optimum approaches the PE-optimal predictor coefficients in (5.26). Thus, minimizing the prediction error also maximizes the information rate. The coefficients for the closed-loop predictor also approach the PE-optimal coefficients of the open-loop predictor for large R, since then $\sigma_E^2 \approx \sigma_{\tilde{X}}^2$. For small R, the coefficients are close to zero to reduce the noise gain of the prediction filter.



The two figures below show the difference between the entropy constraint R and the information rate $\bar{I}(\mathbf{X};\mathbf{Y})$ for first- and second-order filters designed according to different cost functions. Second-order filters clearly outperform the first-order filters, and for the infinite-order open-loop predictors $h_{\mathrm{Info}}^{(\infty)}$ and $h_{\mathrm{PE}}^{(\infty)}$ one gets $R - \bar{I}(\mathbf{X};\mathbf{Y}) \equiv 0$. One can observe that the difference in information rates between the PE and the information-theoretic optimum is negligible. This is due to the filter coefficients being close to the PE solution even for small entropy constraints, and suggests that the energetic solution is a good approximation of the information-theoretic one. In contrary to that, designing H in Fig. 5.2 in order to minimize the MSRE, h_{MSRE}, leads to a performance loss for all rates R. Finally, it can be seen that while closed-loop prediction outperforms open-loop prediction, it does not provide significant performance gains for large R: Compare (5.22) and (5.24) with small γ, and consider that the entropy powers of $S_X(\mathrm{e}^{\jmath\theta})$ and of $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})$ are both equal to σ_∞^2.



The next example probably best represents the characteristic of this work: An analysis of the interplay between signal processing and information theory, an analysis of the similarities and dissimilarities between mutual information and the mean-squared reconstruction error. The following example gives an analysis of the Wiener filter in information-theoretic terms:

Example 38: The Wiener Filter.

Consider the case where $\mathbf{X}$ is the sum of a signal process $\mathbf{S}$ with PSD $S_S(\mathrm{e}^{\jmath\theta}) = 4\cos^2(\theta/2)$ and an independent, white Gaussian noise process $\mathbf{N}$ with variance σ_N^2. The goal is to design a first-order filter H with impulse response vector $h = [h_0,\ h_1]^T$ in order to maximize the information rate between $\mathbf{S}$ and the quantizer output, $\bar{I}(\mathbf{S};\mathbf{Y})$. In other words, the filter shall minimize the information loss rate relevant w.r.t. $\mathbf{S}$. The solution shall be compared to the filter minimizing the variance of $\mathbf{S} - \mathbf{Y}$. The latter filter can be shown to have coefficients [70, Ch. 2]

$$h_{\text{Wiener}} = \frac{1}{(1+\gamma)(\sigma_N^4 + 4\sigma_N^2 + 3)}[3 + 2\sigma_N^2,\ \sigma_N^2]^T. \tag{5.27}$$

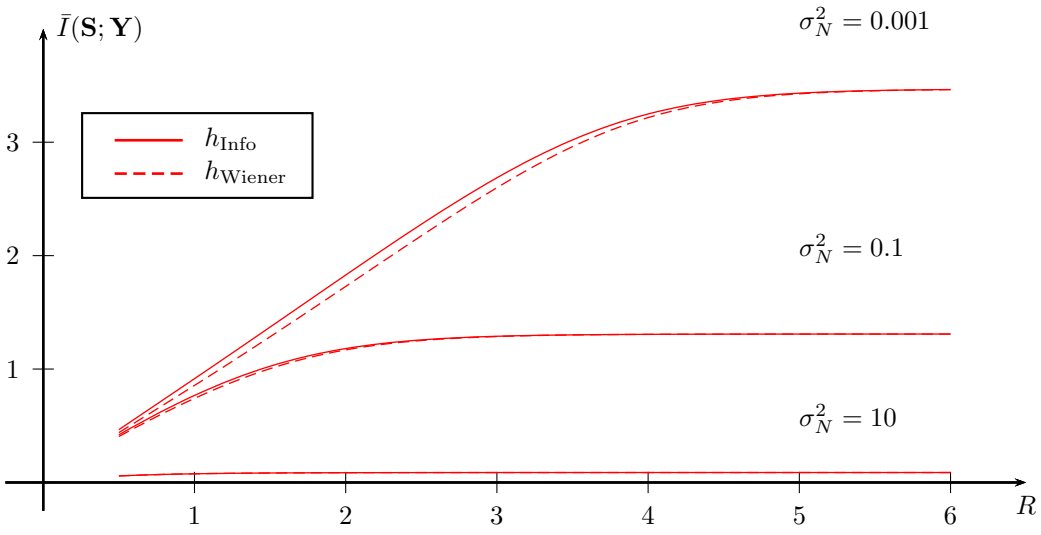
Note that for $R \to \infty$ and $\gamma \to 0$ the coefficient vector h_{Wiener} corresponds to the Wiener solution, hence the subscript. The Wiener filter would also be optimal if the quantization noise variance would not depend on the quantizer input.

To maximize the relevant information rate $\bar{I}(\mathbf{S};\mathbf{Y})$, one can again normalize the first filter coefficient to unity. Hence, with $H(\mathrm{e}^{\jmath\theta})$ being the frequency response of the FIR filter $h = [1,\ h_1]^T$, one strives to maximize

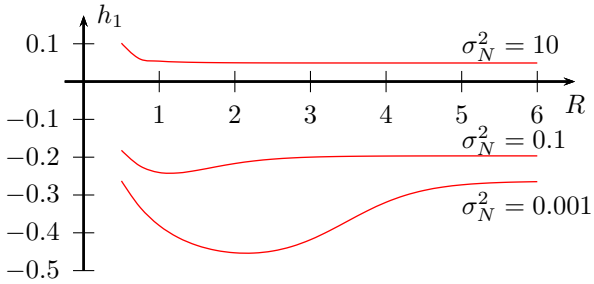
$$\bar{I}(\mathbf{S};\mathbf{Y}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{S_S(\mathrm{e}^{\jmath\theta})|H(\mathrm{e}^{\jmath\theta})|^2}{\sigma_N^2|H(\mathrm{e}^{\jmath\theta})|^2 + \gamma\sigma_{\tilde{X}}^2}\right) d\theta. \tag{5.28}$$

Note that for $\sigma_N^2 = 0$ this degenerates to (5.22). Furthermore, for $\gamma \to 0$ filtering becomes futile, because in this case no quantizer is present.

The cost function was optimized numerically, and the results for the filter coefficient h_1 and the resulting information rate are shown below. The variance of the noise process $\mathbf{N}$ was varied in $\sigma_N^2 = \{0.001, 0.1, 10\}$.



It can be seen that, for large entropy constraints R, the information rates saturate, a consequence of the noise added at the input of the filter. This saturation naturally occurs at a lower rate $\bar{I}(\mathbf{S};\mathbf{Y})$ if the variance σ_N^2 is large. Moreover, while for small noise variances there is a significant difference between the Wiener filter and the information-maximizing one, this difference diminishes for large noise variances – the bottleneck is now the noise source prior to quantization, rather than the quantizer.



Looking at the coefficient h_1 optimal in the information-theoretic sense in the figure above, one can see that the dependence on the entropy constraint R is much more emphasized compared to Example 37. First of all, for small noise variances, the filter is still predicting due to the negativity of h_1. However, h_1 is now larger (i.e., closer to zero) compared to the coefficient in Example 37. The reason is that in (5.28) there are three conflicting goals: equalizing $S_S(\mathrm{e}^{\jmath\theta})$, minimizing $\sigma_N^2|H(\mathrm{e}^{\jmath\theta})|^2$, and minimizing the quantizer input variance $\sigma_{\tilde{X}}^2$. For small R, the last goal is emphasized, while for large σ_N^2 minimizing $\sigma_N^2|H(\mathrm{e}^{\jmath\theta})|^2$ is important. This is also evident for the coefficient for $\sigma_N^2 = 10$. Here, not only is the coefficient close to zero, but it is also positive: In addition to having a small noise gain, the filter also averages noise and emphasizes parts of the input spectrum $S_S(\mathrm{e}^{\jmath\theta})$.

Maximizing the information rate over a quantizer is more difficult than minimizing the MSRE, since it involves nonlinear (and even non-convex) optimization of the filter coefficients. This is in stark contrast to the Wiener filter, whose coefficients are given in closed-form. However, the Wiener filter, useful as it may be for spectral shaping, cannot be assumed to perform well in information-theoretic terms (with h_1/h_0 lying between 0 and 0.5 for σ_N between 0 and ∞, respectively). Energetic design is thus, in general, sub-optimal from an information-theoretic point-of-view. Only in specific cases, such as those presented in Sections 5.3.1 and 5.3.2, both cost functions may lead to similar designs.

5.4 Open Questions

Extending the notion of relevant information loss to stochastic processes allowed its application to filter design, both for anti-aliasing filters and filters prior to quantization. This clearly marks the success and importance of the introduced measures: They can be applied to *make systems better*. However, the processes considered here as still restricted to be non-bandlimited (in discrete time) and absolutely continuous, restrictions which should be overcome in the future. Particularly, the signal process containing the relevant information should be allowed to be discrete-valued or non-stationary: This would eventually permit the analysis of digital communication systems with discrete-valued data signals. More difficult will be the extension to letting $\mathbf{S}$ be a sequence of continuous-valued pulses representing finitely many symbols: $\mathbf{S}$ will neither be stationary, nor will the joint distribution of any of its samples be absolutely continuous.

A logical extension of the anti-aliasing filtering problem is of course the case where the relevant information is not additive in the observed process. For example, the relevant information might be the sign of the observation, or it enters the observation via a noisy, nonlinear function. A first step into this direction has been made in [53]

A different application of the relevant information loss rate may be clustering of *hidden* Markov models: The relevant process can either be the discrete-valued state process (clustering the observation process) or the discrete- or continuous-valued observation process (clustering the state process). This might be of importance for, e.g., speech models, where, for different states the observation, or for different observations the underlying state may be the same.

6 Discussion

The reader will by now agree with the author's statements in the preface: The contents of this work are more diverse and less deep than those of other theses. And in hindsight, the main contributions of this work are threefold:

1. **Establishing a foothold for an information theory for deterministic systems.** Deterministic systems can only destroy, at best preserve, information, hence information loss is an adequate measure for the information-processing capabilities of such systems. This work introduces such measures for memoryless systems fed with either RVs or stationary stochastic processes, both without and with considering a relevant aspect of the input signal. Especially the last option is important for practical systems, since the notion of relevance allows to cast the signal enhancement problem in information-theoretic terms.

2. **Pointing at the difference between energetic and information-theoretic design.** Throughout the literature, various design problems are solved with energetic cost functions in mind, mainly because of their simplicity. This work shows, however, that without a specific signal model, many of these design practices fail from an information-theoretic perspective: E.g., neither PCA nor anti-aliasing filtering can prevent or reduce information loss. Contrarily, both PCA and anti-aliasing filters can be justified information-theoretically under specific assumption on the signal model and by introducing the notion of relevance. Hence, if one knows which aspect of a signal is relevant, deterministic processing can be applied successfully to retrieve as much of this relevant information as possible. Similarily, for filters prior to quantization it was shown that the information-theoretic optimum coincides with the energetic one, at least under specific assumptions. System design benefits from knowing in which situations these two different design objectives lead to different, or to similar solutions.

3. **Providing information-theoretic methods for model complexity reduction for Markov chains.** There are two ways of reducing the state space of a Markov chain by deterministic functions: First, by defining a new Markov chain on groups of states of the original chain. This method is called aggregation in this work and was shown to be connected to the notion of relevant information loss. The information bottleneck method was, to the best of the author's knowledge, applied the first time for aggregation of Markov chains. The second method relies on defining a new, possibly non-Markov process by projecting (random) realizations of sequences coordinate-wise through a non-injective function, again essentially grouping states together. This is called lumping in

this work, and a large body of literture analyzes the conditions under which this new process is Markov (at least of higher order). In this work, not only is this latter property of lumpability characterized with information theory, but also the information loss rate is used to determine classes of information-preserving lumpings. This result is closely related to lossless symbol-by-symbol compression, and the author assumes that certain classes of such information-preserving lumpings can be characterized with the asymptotic equipartition property.

It is also interesting to compare the present work with the five PhD theses mentioned in the introduction. For example, **Dr. Wu's** work [168, 169] had great influence on the present thesis, at least on its methodological aspects. Information dimension, which Wu employed for "almost" lossless analog compression, is a fundamental quantity in the proposed measures of relative information loss (rate). The present work, however, is not focused on lossless compression, but permits information loss and, thus, strictly positive reconstruction error. The link between information loss and error probability made in Proposition 2.12 allows to present a converse to analog compression, as outlined in Conjecture 2.3. Thus, the theory presented here is general enough to also apply to compressed sensing, at least to some aspects of it. What is missing is a more detailed analysis of information dimension: Does it satisfy a chain rule, how is it affected by more general, non-Lipschitz functions, and in which cases does the relation to relative information transfer mentioned in Conjecture 2.1 hold? These are questions which should be answered in future work.

The present work extends the one of **Dr. Evans** [37, 38] by defining relative information transfer not only for binary, but for general, even continuous-valued random variables (cf. Definition 2.9), and to connect this measure to the reconstruction error probability in Proposition 2.12. That most information is lost in the first elements of a cascade was also shown implicitly in Conjecture 2.2, which holds for discrete RVs and continuous RVs processed by projections. The work of Evans also suggests that the analysis of noisy systems is highly important. What is, for example, the relative information loss in an amplifier with an unknown gain offset? Assuming continuous inputs and a continuous distribution of the gain offset, the relative information loss will be 1 for a single measurement, but $1/N$ for N consecutive measurements, since the only uncertain element is a single, scalar RV. Thus, especially the case of noise with a lower-dimensional distribution is of interest and shall be considered in future work.

Comparing the author's work to the one of **Dr. Deng** [26, 27], the significant contribution lies not only in shedding light on the lifting method proposed there, but also in providing an alternative, which is optimal and, thus, better in a well-defined sense. Moreover, that the newly introduced **P**-lifting admits solving the aggregation problem sub-optimally by using the information bottleneck method illustrates its practical significance. But this comparison also reveals several new lines of investigation: First of all, it should be investigated if there is a connection between **P**-lifting and spectral theory of Markov chains. Secondly, the aggregation of hidden Markov models is of great interest, both when it comes to reducing the state and the observation space. The author has the feeling that the relevant information loss rate might be a more suitable cost function, compared to the Kullback-Leibler divergence rate proposed by Dr. Deng. On a more application- oriented level, it would of course be interesting if aggregating the thermal model of a building by employing the information bottleneck method has any advantages in terms of energy efficient control for heating, ventilation, and air conditioning.

The work of **Dr. Sánchez-Montañés** [131, 132] suggested that Shannon entropy can be exchanged by a different uncertainty measure, such as Bayes' error, which would consequently yield different solutions to the signal enhancement problem. Instead of investigating these different uncertainty measures, the author suggests to apply the theory to different application scenarios and design problems. This has been done in this work, e.g., in justifying PCA and anti-aliasing filters from an information-theoretic perspective, and in Example 27, where the receiver of a digital communication system is analyzed. These examples provide evidence for

the author's claim that the trade-off between maximizing irrelevant information loss (removal of spurious information) and minimizing relevant information loss (preservation of relevant information) is equivalent to the problem of signal enhancement; not only for biological systems such as those in [131, 132], but for all signal processing systems. However, there are many other (adaptive) systems which could also benefit from an information-centered design: beamformers, frequency estimators, channel equalizers (cf. [43]), etc. Lossy compression of images could be a further example, which would require knowledge about which part of the image is irrelevant for the human observer. Yet another would be speech enhancement: From an information-theoretic point-of-view, the most suitable objective function for a speech enhancement system would be the word error rate (or sentence error rate?) of an automatic speech recognition system working exactly as the intended sink: the human[27]. That information-theoretic concepts slowly enter speech processing recently (e.g. [79] and the references therein) justifies further investigation of this claim.

Dr. Henter also shares the author's opinion that signal enhancement and feature extraction are essentially a trade-off between complexity of the representation and its accuracy, cf. [72]. His work on four-gram word models in [73] is complemented by the author's work on bi-gram letter models in Section 3.2.6, although the cost function was quite different: Henter minimized the entropy rate of the model while preserving its state space and keeping the Kullback-Leibler divergence rate to the original chain small; in this work, the model is information-preserving on a smaller state space, but with a higher Markov order. The trade-off between accuracy and complexity for Markov model reduction is attractive, and future work could investigate whether a variational formulation of the results in Sections 3.2 or 4.3 is advantageous. Henter also considered complexity reduction of Gaussian processes, something which has not been analyzed in this work. Interestingly, he was able to show that under a specific signal-plus-noise-model, his variational formulation of entropy rate and Kullback-Leibler divergence rate to the signal process has the Wiener filter as solution. Whether the relevant information loss rate of Chapter 5 can also justify the Wiener filter in specific scenarios is within the scope of future work, too.

It can be seen as another contribution of this work, directly related to the first one listed above, that it posed more questions than it answered. Aside from the many open issues hinted at in the last paragraphs, or in the last section of each chapter, the author has the feeling that at least two or three PhD theses could continue where he had to stop: Systems with memory, for example, are almost completely neglected in this work, despite their great importance. At least for a restricted class of systems with memory, such as Volterra systems or affine-input systems, results should be obtainable, albeit with a different definition of information loss. Secondly, as mentioned before, hidden Markov models deserve more attention: Reducing the state and/or the observation space should be possible also by employing information-theoretic cost functions, drawing heavily from the main contribution listed above. Eventually, specifically for speech processing, also continuous observations should be considered. And finally, most importantly, information-theoretic system design should be analyzed more deeply. Maybe a proper analysis will reveal that the information bottleneck method and its variants can be used for system design as well, bringing signal processing and machine learning much closer together. Maybe this will turn over long-established design principles, surpassing the second contribution of this work.

There is (at least) one chapter missing: One interpreting the information lost in a system in a broader context. What does it mean, if "information is lost"? Where does it go? If information is not matter or energy, *what is it?* If information is lost somewhere, *where did it come from*?

[27] It might be the case that this speech recognition system requires a high redundancy in the incoming signal, to allow easy retrieval of the information contained in the signal ("listeing comfort"). A proper model of such a system needs to include the fact that humans are prone to fatigue and, hence, the word or sentence error rate increases stronger if the enhanced speech signal does not sound natural or has lots of artifacts.

These questions need not be impossible to answer, as a look into the literature reveals: Landauer [94], for example, claimed that information loss can be interpreted as a conversion of information-theoretic entropy to thermodynamical entropy. This interpretation could also remove the conflict between the second law of thermodynamics and gedankenexperiments like Maxwell's demon or Szilard's engine (e.g., [175] and the references therin). And if this conversion can go in one direction, can it also go into the other? Can one interpret *information sources* as systems in which some form of free energy is converted into information-theoretic entropy? And what are these sources? For example, it is known that chaotic systems reveal information about their initial state in samples of time series; systems are chaotic if they have a positive Kolmogorov-Sinaï entropy. Moreover, noise is a source of information, although one of little practical use. Most importantly, the human being is a source – and a sink – of information. Schrödinger wrote that the human being feeds on "negative entropy" to sustain its life [136] – does this mean that we feed on information? And is the creation of information within ourselves just a conversion from some other form of energy? Is information equivalent (but not identical) to matter and energy, so that it can be the principle of a monism?

These questions are not only philosophical, but also part of the realms of physics, information theory, and mathematics in general. It is not an easy task to give satisfactory answers – but it's an interesting one!

A
Proofs from Chapter 2

A.1 Proof of Theorem 2.2

Recall that Definition 2.2 states

$$L(X \to Y) := \lim_{n\to\infty} \left(I(\hat{X}^{(n)}; X) - I(\hat{X}^{(n)}; Y) \right). \tag{A.1}$$

Let $\hat{X}^{(n)} = \hat{x}_k$ if $x \in \hat{\mathcal{X}}_k^{(n)}$. The conditional probability measure satisfies $P_{X|\hat{X}^{(n)}=\hat{x}_k} \ll \lambda^N$ and thus possesses a density

$$f_{X|\hat{X}^{(n)}}(x, \hat{x}_k) = \begin{cases} \frac{f_X(x)}{p(\hat{x}_k)}, & \text{if } x \in \hat{\mathcal{X}}_k^{(n)} \\ 0, & \text{else} \end{cases} \tag{A.2}$$

where $p(\hat{x}_k) = P_X(\hat{\mathcal{X}}_k^{(n)}) > 0$. Also, $P_{Y|\hat{X}^{(n)}=\hat{x}_k} \ll \lambda^N$, and its PDF if given by the method of transformation. With [21, Ch. 8.5],

$$\begin{aligned} L(X \to Y) &= \lim_{n\to\infty} \left(h(X) - h(X|\hat{X}^{(n)}) - h(Y) + h(Y|\hat{X}^{(n)}) \right) \\ &= h(X) - h(Y) + \lim_{n\to\infty} \left(h(Y|\hat{X}^{(n)}) - h(X|\hat{X}^{(n)}) \right). \end{aligned} \tag{A.3}$$

The latter difference can be written as

$$h(Y|\hat{X}^{(n)}) - h(X|\hat{X}^{(n)}) = \sum_{\hat{x}_k} p(\hat{x}_k) \left(h(Y|\hat{X}^{(n)} = \hat{x}_k) - h(X|\hat{X}^{(n)} = \hat{x}_k) \right). \tag{A.4}$$

Since [114, Thm. 5-1]

$$h(Y|\hat{X}^{(n)} = \hat{x}_k) = -\int_{\mathcal{Y}} f_{Y|\hat{X}^{(n)}}(y, \hat{x}_k) \log f_{Y|\hat{X}^{(n)}}(y, \hat{x}_k) dy \tag{A.5}$$

$$= -\int_{\mathcal{X}} f_{X|\hat{X}^{(n)}}(x, \hat{x}_k) \log f_{Y|\hat{X}^{(n)}}(g(x), \hat{x}_k) dx \tag{A.6}$$

$$= -\frac{1}{p(\hat{x}_k)} \int_{\hat{\mathcal{X}}_k^{(n)}} f_X(x) \log f_{Y|\hat{X}^{(n)}}(g(x), \hat{x}_k) dx \tag{A.7}$$

one obtains

$$h(Y|\hat{X}^{(n)}) - h(X|\hat{X}^{(n)}) = \sum_k \int_{\hat{\mathcal{X}}_k^{(n)}} f_X(x) \log \frac{f_{X|\hat{X}^{(n)}}(x, \hat{x}_k)}{f_{Y|\hat{X}^{(n)}}(g(x), \hat{x}_k)} dx$$
$$= \int_{\mathcal{X}} f_X(x) \log \frac{f_{X|\hat{X}^{(n)}}(x, q^{(n)}(x))}{f_{Y|\hat{X}^{(n)}}(g(x), q^{(n)}(x))} dx \quad \text{(A.8)}$$

where $q^{(n)}(x) = \hat{x}_k$ if $x \in \hat{\mathcal{X}}_{\hat{k}}^{(n)}$ and where, by the method of transformation,

$$f_{Y|\hat{X}^{(n)}}(g(x), q^{(n)}(x)) = \sum_{x_i \in g^{-1}[g(x)]} \frac{f_{X|\hat{X}^{(n)}}(x_i, q^{(n)}(x))}{|\det \mathcal{J}_g(x_i)|}. \quad \text{(A.9)}$$

Since the preimage of $g(x)$ is a set separated by neighborhoods[28], there exists an n_0 such that

$$\forall n \geq n_0 \text{: } k = \{\hat{k} : x \in \hat{\mathcal{X}}_{\hat{k}}^{(n)}\} \text{: } g^{-1}[g(x)] \cap \hat{\mathcal{X}}_k^{(n)} = x \quad \text{(A.10)}$$

i.e., such that from this index on, the element of the partition under consideration, $\hat{\mathcal{X}}_k^{(n)}$, contains just a single element of the preimage, x. Since $f_{X|\hat{X}^{(n)}}$ is non-zero only for arguments in $\hat{\mathcal{X}}_k^{(n)}$, in this case (A.9) degenerates to

$$f_{Y|\hat{X}^{(n)}}(g(x), q^{(n)}(x)) = \frac{f_{X|\hat{X}^{(n)}}(x, q^{(n)}(x))}{|\det \mathcal{J}_g(x)|}. \quad \text{(A.11)}$$

Consequently, the ratio

$$\frac{f_{X|\hat{X}^{(n)}}(x, q^{(n)}(x))}{f_{Y|\hat{X}^{(n)}}(g(x), q^{(n)}(x))} \nearrow |\det \mathcal{J}_g(x)| \quad \text{(A.12)}$$

monotonically (the number of positive terms in the sum in the denominator reduces with n). Applying the monotone convergence theorem, e.g., [128, pp. 21], yields

$$\lim_{n\to\infty} h(Y|\hat{X}^{(n)}) - h(X|\hat{X}^{(n)}) = \int_{\mathcal{X}} f_X(x) \log |\det \mathcal{J}_g(x)| dx = \mathbb{E}\left(\log |\det \mathcal{J}_g(X)|\right). \quad \text{(A.13)}$$

This completes the proof. □

A.2 Proof of Proposition 2.3

Note that

$$H(W|Y) = \int_{\mathcal{Y}} H(W|Y = y) dP_Y(y) = -\int_{\mathcal{Y}} \sum_i p(i|y) \log p(i|y) f_Y(y) dy \quad \text{(A.14)}$$

where $p(i|y) = \Pr(W = i|Y = y) = P_{X|Y=y}(\mathcal{X}_i)$. For the sake of simplicity, permit the Dirac delta distribution δ as a PDF for discrete (atomic) probability measures. Following [19], one can write for the conditional PDF of Y given $X = x$,

$$f_{Y|X}(x, y) = \delta(y - g(x)) = \sum_{x_i \in g^{-1}[y]} \frac{\delta(x - x_i)}{|\det \mathcal{J}_g(x_i)|}. \quad \text{(A.15)}$$

[28] The space $\mathbb{R}^N$ is Hausdorff, so any two distinct points are separated by neighborhoods.

Applying Bayes' theorem for densities yields

$$p(i|y) = \int_{\mathcal{X}_i} f_{X|Y}(x,y)dx \tag{A.16}$$

$$= \int_{\mathcal{X}_i} \frac{f_{Y|X}(x,y)f_X(x)}{f_Y(y)}dx \tag{A.17}$$

$$= \frac{1}{f_Y(y)} \int_{\mathcal{X}_i} \sum_{x_k \in g^{-1}[y]} \frac{\delta(x-x_k)}{|\mathrm{det}\mathcal{J}_g(x_k)|} f_X(x)dx \tag{A.18}$$

$$= \begin{cases} \frac{f_X(g_i^{-1}[y])}{|\mathrm{det}\mathcal{J}_g(g_i^{-1}[y])|f_Y(y)}, & \text{if } y \in \mathcal{Y}_i \\ 0, & \text{if } y \notin \mathcal{Y}_i \end{cases} \tag{A.19}$$

by the properties of the delta distribution (e.g., [112]) and since, by Definition 2.3, at most one element of the preimage of y lies in $\mathcal{X}_i$.

One can rewrite (A.14) as

$$H(W|Y) \;=\; -\sum_i \int_{\mathcal{Y}_i} p(i|y) \log p(i|y) f_Y(y)dy \tag{A.20}$$

after exchanging the order of summation and integration with the help of Tonelli's theorem [42, Thm. 2.37] and by noticing that $p(i|y) = 0$ if $y \notin \mathcal{Y}_i$. Inserting the expression for $p(i|y)$ and changing the integration variables by substituting $x = g_i^{-1}[y]$ in each integral yields

$$H(W|Y)$$

$$= -\sum_i \int_{\mathcal{Y}_i} \frac{f_X(g_i^{-1}[y])}{|\mathrm{det}\mathcal{J}_g(g_i^{-1}[y])|} \log \frac{f_X(g_i^{-1}[y])}{|\mathrm{det}\mathcal{J}_g(g_i^{-1}[y])|f_Y(y)}dy \tag{A.21}$$

$$= -\sum_i \int_{\mathcal{X}_i} f_X(x) \log \frac{f_X(x)}{|\mathrm{det}\mathcal{J}_g(x)|f_Y(g(x))}dx \tag{A.22}$$

$$= -\int_{\mathcal{X}} f_X(x) \log \frac{f_X(x)}{|\mathrm{det}\mathcal{J}_g(x)|f_Y(g(x))}dx \tag{A.23}$$

$$\overset{(a)}{=} h(X) - h(Y) + \mathbb{E}\left(\log|\mathrm{det}\mathcal{J}_g(X)|\right) \tag{A.24}$$

$$= L(X \to Y) \tag{A.25}$$

where (a) is due to splitting the logarithm and applying [114, Thm. 5-1]. □

A.3 Proof of Proposition 2.4

The proof depends in parts on the proof of Proposition 2.3, where it is shown that

$$L(X \to Y) = \int_{\mathcal{Y}} H(W|Y=y) f_Y(y)dy. \tag{A.26}$$

The first inequality (2.38) is due to the maximum entropy property of the uniform distribution, i.e., $H(W|Y=y) \leq \log \mathrm{card}(g^{-1}[y])$ with equality if and only if $p(i|y) = 1/\mathrm{card}(g^{-1}[y])$ for all i for which $g_i^{-1}[y] \neq \emptyset$. But this translates to

$$\mathrm{card}(g^{-1}[y]) = \frac{|\mathrm{det}\mathcal{J}_g(g_i^{-1}[y])|f_Y(y)}{f_X(g_i^{-1}[y])}. \tag{A.27}$$

Inserting the expression for f_Y and substituting x for $g_i^{-1}[y]$ (it is immaterial which i is chosen, as long as the preimage of y is not the empty set) one obtains

$$\operatorname{card}(g^{-1}[g(x)]) = \sum_{x_k \in g^{-1}[g(x)]} \frac{f_X(x_k)}{|\det \mathcal{J}_g(x_k)|} \frac{|\det \mathcal{J}_g(x)|}{f_X(x)}. \tag{A.28}$$

The second inequality (2.39) is due to Jensen [21, 2.6.2],

$$\begin{aligned} \mathbb{E}\left(\log \operatorname{card}(g^{-1}[Y])\right) \leq \log \mathbb{E}\left(\operatorname{card}(g^{-1}[Y])\right) &= \log \int_{\mathcal{Y}} \operatorname{card}(g^{-1}[y]) dP_Y(y) \\ &= \log \int_{\mathcal{Y}} \sum_i \operatorname{card}(g_i^{-1}[y]) dP_Y(y) = \log \sum_i \int_{\mathcal{Y}_i} dP_Y(y) \end{aligned} \tag{A.29}$$

since $\operatorname{card}(g_i^{-1}[y]) = 1$ if $y \in \mathcal{Y}_i$ and zero otherwise. Equality is achieved if and only if $\operatorname{card}(g^{-1}[y])$ is constant P_Y-a.s. In this case also the third inequality (2.40) is tight, which is obtained by replacing the expected value of the cardinality of the preimage by its essential supremum.

Finally, the cardinality of the preimage cannot be larger than the cardinality of the partition used in Definition 2.3, which yields the last inequality (2.41). Equality holds if and only if, assuming that all previous requirements for equality in the other bounds are fulfilled, $P_Y(\mathcal{Y}_i) = 1$ for all i. This completes the proof. □

A.4 Proof of Theorem 2.3

The proof follows closely the proof of Fano's inequality [21, pp. 38], which starts by applying the chain rule

$$H(X|Y) = H(E|Y) + H(X|E,Y). \tag{A.30}$$

The first term, $H(E|Y)$ can be upper bounded by $H(E) = H_2(P_e)$, as in Fano's inequality. However, also

$$H(E|Y) = \int_{\mathcal{Y}} H_2(P_e(y)) dP_Y(y) = \int_{\mathcal{Y}\setminus\mathcal{Y}_b} H_2(P_e(y)) dP_Y(y) \leq \int_{\mathcal{Y}\setminus\mathcal{Y}_b} dP_Y(y) = 1 - P_b \tag{A.31}$$

since $H_2(P_e(y)) = P_e(y) = 0$ if $y \in \mathcal{Y}_b$ and since $H_2(P_e(y)) \leq 1$ otherwise. Thus,

$$H(E|Y) \leq \min\{H_2(P_e), 1 - P_b\}. \tag{A.32}$$

For the second part note that $H(X|E=0, Y=y) = 0$, hence one gets

$$H(X|E,Y) = \int_{\mathcal{Y}} H(X|E=1, Y=y) P_e(y) dP_Y(y). \tag{A.33}$$

Upper bounding the entropy by $\log\left(\operatorname{card}(g^{-1}[y]) - 1\right)$ yields

$$H(X|E,Y) \leq P_e \int_{\mathcal{Y}} \log\left(\operatorname{card}(g^{-1}[y]) - 1\right) \frac{P_e(y)}{P_e} dP_Y(y) \tag{A.34}$$

$$\overset{(a)}{\leq} P_e \log\left(\int_{\mathcal{Y}} \left(\operatorname{card}(g^{-1}[y]) - 1\right) \frac{P_e(y)}{P_e} dP_Y(y)\right) \tag{A.35}$$

$$\overset{(b)}{\leq} P_e \log\left(\int_{\mathcal{Y}} \left(\operatorname{card}(g^{-1}[y]) - 1\right) dP_Y(y)\right) + P_e \log \frac{1}{P_e} \tag{A.36}$$

where (a) is Jensen's inequality ($P_e(y)/P_e$ acts as a PDF) and (b) holds since $P_e(y) \leq 1$ and due to splitting the logarithm. This completes the proof. □

A.5 Proof of Proposition 2.7

By construction, $r_{\text{sub}}(y) = x$ whenever $x \in \mathcal{X}_k \cup \mathcal{X}_b$, and conversely, $r_{\text{sub}}(y) \neq x$ whenever $x \notin \mathcal{X}_k \cup \mathcal{X}_b$. This yields $\hat{P}_e = 1 - P_X(\mathcal{X}_k \cup \mathcal{X}_b)$.

For the Fano-type bound, notice again that

$$H(X|Y) = H(E|Y) + H(X|E,Y). \tag{A.37}$$

The first term can be written as

$$H(E|Y) = \int_{\mathcal{Y}} H_2(\hat{P}_e(y)) dP_Y(y) = \int_{\mathcal{Y}_k \setminus \mathcal{Y}_b} H_2(\hat{P}_e(y)) dP_Y(y) \leq P_Y(\mathcal{Y}_k \setminus \mathcal{Y}_b) \tag{A.38}$$

since $\hat{P}_e(y) = 0$ for $y \in \mathcal{Y}_b$ and $\hat{P}_e(y) = 1$ for $y \in \mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)$.

For the second term one can write

$$H(X|E,Y) = \int_{\mathcal{Y}} H(X|Y=y, E=1)\hat{P}_e(y) dP_Y(y) \tag{A.39}$$

$$\leq \int_{\mathcal{Y}_k \setminus \mathcal{Y}_b} \log(\overline{K}-1)\hat{P}_e(y) dP_Y(y) + \int_{\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)} \log \overline{K} dP_Y(y) \tag{A.40}$$

Now note that

$$\hat{P}_e = P_Y(\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)) + \int_{\mathcal{Y}_k \setminus \mathcal{Y}_b} \hat{P}_e(y) dP_Y(y) \tag{A.41}$$

which can be used above to get

$$H(X|E,Y) \leq \left(\hat{P}_e - P_Y(\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b))\right) \log(\overline{K}-1) + P_Y(\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)) \log \overline{K}. \tag{A.42}$$

Rearranging and using

$$P_b + P_Y(\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)) + P_Y(\mathcal{Y}_k \setminus \mathcal{Y}_b) = 1 \tag{A.43}$$

yields

$$H(X|Y) \leq 1 - P_b + \hat{P}_e \log(\overline{K}-1) + P_Y(\mathcal{Y} \setminus (\mathcal{Y}_k \cup \mathcal{Y}_b)) \left(\log \frac{\overline{K}}{\overline{K}-1} - 1\right). \tag{A.44}$$

The fact $0 \leq \log \frac{\overline{K}}{\overline{K}-1} \leq 1$ completes the proof. □

A.6 Proof of Proposition 2.10

By assumption, $g|_{\mathcal{X}_i}\colon \mathcal{X}_i \to \mathcal{Y}_i$ is a projection, which preserves exactly M_i of the N original coordinates. Assume, w.l.o.g., that the first M_i coordinates are preserved, and that the remaining $N - M_i$ coordinates are dropped. It follows that $\mathcal{Y}_i = \mathcal{X}_i \cap (\mathbb{R}^{M_i} \times \emptyset^{N-M_i})$, which is an M_i-dimensional set. Moreover, the preimage of a set $A \subset \mathcal{Y}_i$ is given by

$$g|_{\mathcal{X}_i}^{-1}[A] = \mathcal{X}_i \cap (A \times \mathbb{R}^{N-M_i}) = A \times (\mathcal{X}_i \cap (\emptyset^M \times \mathbb{R}^{N-M_i})). \tag{A.45}$$

From $P_{X|X\in\mathcal{X}_i} \ll \lambda^N$ immediately follows that $P_{Y|X\in\mathcal{X}_i} \ll \lambda^{M_i}$, since the distribution of X

possesses a PDF and the PDF of Y is obtained by marginalization. It needs to be shown that the distribution on the preimage of a singleton is absolutely continuous w.r.t. the $(N - M_i)$-dimensional Lebesgue measure.

To this end, take $A \subseteq \mathcal{Y}_i$ such that $P_{Y|X\in\mathcal{X}_i}(A) > 0$, thus $\lambda^{M_i}(A) > 0$. Then, assume that there exists a $B \subseteq \mathcal{X}_i \cap (\emptyset^M \times \mathbb{R}^{N-M_i})$ with $\lambda^{N-M_i}(B) = 0$ and, $P_{Y|X\in\mathcal{X}_i}$-a.s.,

$$\forall y \in A : P_{X|Y=y,X\in\mathcal{X}_i}(B) > 0. \tag{A.46}$$

Now, one can compute

$$P_{X|X\in\mathcal{X}_i}(A \times B) = \int_A P_{X|Y=y,X\in\mathcal{X}_i}(B) dP_{Y|X\in\mathcal{X}_i}(y) > 0 \tag{A.47}$$

while $\lambda^N(A \times B) = \lambda^{M_i}(A)\lambda^{N-M_i}(B) = 0$, which contradicts the assumption that $P_{X|X\in\mathcal{X}_i} \ll \lambda^N$. Consequently, $d(X|Y = y, X \in \mathcal{X}_i) = N - M_i$, $P_{Y|X\in\mathcal{X}_i}$-a.s.

With [144, 169],

$$d(X|Y = y) = \sum_{i=1}^{K} d(X|Y = y, X \in \mathcal{X}_i) P_{X|Y=y}(\mathcal{X}_i) = \sum_{i=1}^{K} (N - M_i) P_{X|Y=y}(\mathcal{X}_i) \tag{A.48}$$

and

$$d(X|Y) = \sum_{i=1}^{K} (N - M_i) \int_{\mathcal{Y}} P_{X|Y=y}(\mathcal{X}_i) dP_Y(y) = \sum_{i=1}^{K} (N - M_i) P_X(\mathcal{X}_i). \tag{A.49}$$

The fact $d(X) = N$ completes the proof.

For Corollary 2.6 note that $d(Y|X \in \mathcal{X}_i) = M_i$, and

$$d(X|Y) = \sum_{i=1}^{K} (N - M_i) P_X(\mathcal{X}_i) = N - \sum_{i=1}^{K} M_i P_X(\mathcal{X}_i) = d(X) - d(Y). \tag{A.50}$$

The author believes that a generalization to submersions is possible (but not necessary for this work): By the submersion theorem [95, Cor. 5.25] for every point $y \in \mathcal{Y} = \bigcup_{i=1}^{K} \mathcal{Y}_i$, the preimage under $g|_{\mathcal{X}_i}$ is either the empty set (if $y \notin \mathcal{Y}_i$) or an $(N - M_i)$-dimensional embedded submanifold of $\mathcal{X}_i$. Moreover, preimages of λ^{M_i}-null sets are λ^N-null sets [119], thus $P_{Y|X\in\mathcal{X}_i} \ll \lambda^{M_i}$, and $d(Y|X \in \mathcal{X}_i) = M_i$. Finally, with [140, Thm. 8.1] it follows that the conditional probability measures supported on the preimages of y under $g|_{\mathcal{X}_i}$ are smooth and possess a density. □

A.7 A Sketch for Conjecture 2.1

Lemma A.1. *Let X and Y be the input and output of a Lipschitz function g: $\mathcal{X} \to \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^N$, $\mathcal{Y} \subseteq \mathbb{R}^M$. Then,*

$$\lim_{n\to\infty} \frac{H(\hat{Y}^{(n)}|\hat{X}^{(n)})}{n} = 0. \tag{A.51}$$

Proof. By showing that $H(\hat{Y}^{(n)}|\hat{X}^{(n)} = \hat{x}_k)$ is uniformly bounded for all n and for all $\hat{x}_k$, it immediately follows that

$$\lim_{n\to\infty} \frac{H(\hat{Y}^{(n)}|\hat{X}^{(n)})}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_k H(\hat{Y}^{(n)}|\hat{X}^{(n)} = \hat{x}_k) P_X(\hat{\mathcal{X}}_k^{(n)}) = 0. \tag{A.52}$$

To this end, note that the conditional probability measure $P_{X|\hat{X}^{(n)}=\hat{x}_k}$ is supported on $\hat{\mathcal{X}}_k^{(n)}$, and that, thus, $P_{Y|\hat{X}^{(n)}=\hat{x}_k}$ is supported on $g(\hat{\mathcal{X}}_k^{(n)})$. Since g is Lipschitz, there exists a constant λ such that for all $a, b \in \hat{\mathcal{X}}_k^{(n)}$

$$|g(a) - g(b)| \leq \lambda |a - b|. \tag{A.53}$$

Choose a and b such that the term on the left is maximized, i.e.,

$$\sup_{a,b} |g(a) - g(b)| = |g(a^\circ) - g(b^\circ)| \leq \lambda |a^\circ - b^\circ| \leq \lambda \sup_{a,b} |a - b| \tag{A.54}$$

or, in other words,

$$\operatorname{diam}(g(\hat{\mathcal{X}}_k^{(n)})) \leq \lambda \operatorname{diam}(\hat{\mathcal{X}}_k^{(n)}) \leq \frac{\lambda\sqrt{N}}{2^n}. \tag{A.55}$$

The latter inequality follows since $\hat{\mathcal{X}}_k^{(n)}$ is inside an N-dimensional hypercube of side length $\frac{1}{2^n}$ (one may have equality in the last statement if $\hat{\mathcal{X}}_k^{(n)}$ is a subset of $\mathcal{X}$).

The support of $P_{Y|\hat{X}^{(n)}=\hat{x}_k}$ can be covered by an M-dimensional hypercube of side length $\operatorname{diam}(g(\hat{\mathcal{X}}_k^{(n)}))$, which can again be covered by

$$\left\lceil 2^n \operatorname{diam}(g(\hat{\mathcal{X}}_k^{(n)})) + 1 \right\rceil^M \tag{A.56}$$

M-dimensional hypercubes of side length $\frac{1}{2^n}$. By the maximum entropy property of the uniform distribution,

$$\begin{aligned} H(\hat{Y}^{(n)}|\hat{X}^{(n)} = \hat{x}_k) &\leq \log \left\lceil 2^n \operatorname{diam}(g(\hat{\mathcal{X}}_k^{(n)})) + 1 \right\rceil^M \\ &\leq M \log \left\lceil 2^n \frac{\lambda\sqrt{N}}{2^n} + 1 \right\rceil = M \log \left\lceil \lambda\sqrt{N} + 1 \right\rceil < \infty. \end{aligned} \tag{A.57}$$

This completes the proof. □

For the conjecture[29], note that by expanding the relative information transfer as

$$t(X \to Y) = \lim_{n\to\infty} \frac{I(\hat{X}^{(n)}; \hat{Y}^{(n)}) + I(\hat{X}^{(n)}; Y|\hat{Y}^{(n)})}{H(\hat{X}^{(n)})} \tag{A.58}$$

$$= \lim_{n\to\infty} \frac{I(\hat{X}^{(n)}; \hat{Y}^{(n)})/n + I(\hat{X}^{(n)}; Y|\hat{Y}^{(n)})/n}{H(\hat{X}^{(n)})/n} \tag{A.59}$$

$$= \lim_{n\to\infty} \frac{H(\hat{Y}^{(n)})/n + I(\hat{X}^{(n)}; Y|\hat{Y}^{(n)})/n}{H(\hat{X}^{(n)})/n} \tag{A.60}$$

due to Lemma A.1 the proof boils down to showing that

$$\frac{1}{n} I(\hat{X}^{(n)}; Y|\hat{Y}^{(n)}) \to 0. \tag{A.61}$$

Writing $\hat{X}^{(n)} = \{\hat{X}_1, \ldots, \hat{X}_n\}$ and $Y = \{\hat{Y}_1, \ldots, \hat{Y}_\infty\}$ one gets

$$\frac{1}{n} I(\hat{X}^{(n)}; Y|\hat{Y}^{(n)}) = \frac{1}{n} I(\hat{X}_1, \ldots, \hat{X}_n; \hat{Y}_1, \ldots, \hat{Y}_\infty | \hat{Y}_1, \ldots, \hat{Y}_n) \tag{A.62}$$

[29] The author is indebted to Siu-Wai Ho and, particularly, Tobias Koch for fruitful discussions about this problem.

$$= \frac{1}{n} I(\hat{X}_1, \ldots, \hat{X}_n; \hat{Y}_{n+1}, \ldots, \hat{Y}_\infty | \hat{Y}_1, \ldots, \hat{Y}_n) \tag{A.63}$$

$$= \frac{1}{n} \sum_{k=1}^{n} \underbrace{I(\hat{X}_k; \hat{Y}_{n+1}, \ldots, \hat{Y}_\infty | \hat{Y}_1, \ldots, \hat{Y}_n, \hat{X}_1, \ldots, \hat{X}_{n-1})}_{=:a_{k,n}} \tag{A.64}$$

Obviously, $a_{k,n} \to 0$ for $n \to \infty$, since one can express it again via the chain rule of a conditional mutual information. The latter is bounded by one, thus the remainder of the series needs to approach 0. Consequently, also $a_{n,n} \to 0$. But neither of these two facts is sufficient to show that the Césaro sum converges to 0, too: While $a_{k,n}$ is bounded for all k, it is not sure whether it is *uniformly* bounded. This is the missing part of the proof.

A.8 PCA: Information Dimension of the Input Matrix given the Output Matrix

It was already shown that $P_{\hat{\mathbf{W}}} \ll \lambda^{\frac{N(N-1)}{2}}$. Furthermore, since the sample covariance matrix of $\mathbf{Y}$ is a diagonal matrix, the corresponding equations

$$1 \leq i < j \leq N: \quad (\hat{\mathbf{C}}_Y)_{ij} = \frac{1}{n} \sum_{k=1}^{n} Y_{ik} Y_{jk} = 0 \tag{A.65}$$

restrict the possible values of $\mathbf{Y}$ from an nN-dimensional to an M-dimensional subspace with

$$M = nN - \frac{N(N-1)}{2}. \tag{A.66}$$

In fact, it can be shown that M elements of $\mathbf{Y}$ are random while the remaining $(nN - M)$ depend on these in a deterministic manner, i.e., one can determine Y_{ij} from the equation for $(\hat{\mathbf{C}}_Y)_{ij}$.

Since $\mathbf{X} = \hat{\mathbf{W}}\mathbf{Y}$, $\mathbf{X}$ is a smooth function from $\mathbb{R}^{\frac{N(N-1)}{2}} \times \mathbb{R}^M$ (the ranges of the values of $\hat{\mathbf{W}}$ and $\mathbf{Y}$) to $\mathbb{R}^{nN}$ (the range of values of $\mathbf{X}$), thus from $\mathbb{R}^{nN}$ to $\mathbb{R}^{nN}$. Since this smooth mapping maps null-sets to null-sets [95, Lem. 6.2],

$$P_{(\hat{\mathbf{W}}, \mathbf{Y})} \ll \lambda^{nN}. \tag{A.67}$$

(Note that $\hat{\mathbf{W}}$ and $\mathbf{Y}$ together have more than nN entries, but only nN of those can be chosen freely. In other words, the graph of the functions defining the remaining entries of $\mathbf{Y}$ and $\hat{\mathbf{W}}$ is an nN-dimensional submanifold of $\mathbb{R}^{N(n+N)}$ [95, Lem. 5.9].) The joint distribution of $(\hat{\mathbf{W}}, \mathbf{Y})$ thus possesses a density, and, by marginalizing and conditioning, so does $\hat{\mathbf{W}}|\mathbf{Y} = \mathbf{y}$. As a consequence,

$$P_{\hat{\mathbf{W}}|\mathbf{Y}=\mathbf{y}} \ll \lambda^{\frac{N(N-1)}{2}}. \tag{A.68}$$

Note further that this does not imply that $\hat{\mathbf{W}}$ is independent of $\mathbf{Y}$ – it only implies that these two quantities are at least not related deterministically.

The final step is taken by recognizing that if one knows $\mathbf{Y} = \mathbf{y}$, then $\mathbf{X}|\mathbf{Y} = \mathbf{y}$ is a linear function of $\hat{\mathbf{W}}|\mathbf{Y} = \mathbf{y}$. Since $\mathbf{Y}$ has full rank, the linear function maps the N^2-dimensional space of $\hat{\mathbf{W}}$ (on which the probability mass is concentrated on an $N(N-1)/2$-dimensional subspace) to the N^2-dimensional linear subspace of $\mathbb{R}^{nN}$. With [172, Remark 28.9] this transform is bi-Lipschitz and preserves the information dimension. Thus,

$$d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = \frac{N(N-1)}{2}. \tag{A.69}$$

B Proofs from Chapter 3

B.1 Proof of Theorem 3.1

The theorem follows from the mutually exclusive and collectively exhaustive implications

$$\mathcal{K} < \infty \Rightarrow \bar{L}(\mathbf{X} \to \mathbf{Y}) > 0\,, \tag{B.1a}$$
$$\mathcal{K} = \infty \Rightarrow \bar{L}(\mathbf{X} \to \mathbf{Y}) = 0 \tag{B.1b}$$

and

$$\mathcal{K} < \infty \Rightarrow \exists C > 0 : \Pr(\liminf_{n\to\infty} \sqrt[n]{T_n} \geq C) = 1\,, \tag{B.2a}$$
$$\mathcal{K} = \infty \Rightarrow \exists C > 0 : \sup_{n\to\infty} T_n \leq C\,. \tag{B.2b}$$

The proofs of implications (B.1b) and (B.2b) and of Proposition 3.1 are in Section B.1.1 and the proofs of implications (B.1a) and (B.2a) are sketched in Section B.1.2.

B.1.1 The information-preserving case

The definition of $\mathcal{K}$ in (3.14) implies that lumped trajectories of length n less than $\mathcal{K}+2$ have a unique preimage contingent on the endpoints, i.e., if $n-2 < \mathcal{K}$, then $\forall \check{x}, \hat{x} \in \mathcal{X}, y_1^n \in \mathcal{Y}^n$:

$$\begin{aligned}&\Pr(X_1 = \check{x}, Y_1^n = y_1^n, X_n = \hat{x}) > 0\\ &\qquad \Rightarrow \exists! x_2^{n-1} \in \mathcal{X}^{n-2} : \Pr(X_2^{n-1} = x_2^{n-1} | X_1 = \check{x}, Y_1^n = y_1^n, X_n = \hat{x}) = 1\,.\end{aligned} \tag{B.3}$$

Proof of Proposition 3.1. Assume $n-2 < \mathcal{K}$. The unique preimage (B.3) implies that the conditional entropy of the interior of a block, given its lumped image and the states at its ends, is zero:

$$\begin{aligned}&H(X_2^{n-1} | X_1, X_n, Y_1^n)\\ &\quad = \sum_{\substack{y_1^n \in \mathcal{Y}^n\\ \check{x},\hat{x}\in\mathcal{X}}} \Pr(X_1 = \check{x}, X_n = \hat{x}, Y_1^n = y_1^n) \underbrace{H(X_2^{n-1} | X_1 = \check{x}, X_n = \hat{x}, Y_1^n = y_1^n)}_{=0 \text{ by (B.3)}} = 0\,.\end{aligned} \tag{B.4}$$

Apply the chain rule of entropy to decompose the conditional block entropy into its interior and its boundary. The interior vanishes by (B.4) and the entropy at the endpoints is maximal for

the uniform distribution:

$$\begin{aligned} L(X_1^n \to Y_1^n) &= H(X_2^{n-1}|X_1, X_n, Y_1^n) + H(X_1, X_n|Y_1^n) \\ &\leq 0 + H(X_1, X_n|Y_1, Y_n) \\ &\leq H(X_1|Y_1) + H(X_n|Y_n) \\ &\leq 2\max\{\log \operatorname{card}(g^{-1}[y]) : y \in \mathcal{Y}\} \\ &\leq 2\log(\operatorname{card}(\mathcal{X}) - \operatorname{card}(\mathcal{Y}) + 1)\,. \end{aligned}$$

□

Proof of (B.1b). As $\mathcal{K} = \infty$, the bound from Proposition 3.1 holds uniformly for all n. Thus

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n} L(X_1^n \to Y_1^n) \leq \lim_{n\to\infty} \frac{2\log(\operatorname{card}(\mathcal{X}) - \operatorname{card}(\mathcal{Y}) + 1)}{n} = 0\,.$$

□

Proof of (B.2b). Let t_n be a realization of the preimage count T_n (i.e., for $Y_1^n = y_1^n$). Then,

$$\begin{aligned} t_n &= \sum_{x_1^n \in g^{-1}[y_1^n]} \lceil \Pr(X_1^n = x_1^n) \rceil \\ &= \sum_{x_1^n \in \mathcal{X}^n} \lceil \Pr(X_1^n = x_1^n | Y_1^n = y_1^n) \rceil \\ &= \sum_{x_1^n \in \mathcal{X}^n} \lceil \Pr(X_1 = x_1, X_n = x_n | Y_1^n = y_1^n) \rceil \\ &\quad \times \lceil \Pr(X_2^{n-1} = x_2^{n-1} | Y_1^n = y_1^n, X_1 = x_1, X_n = x_n) \rceil \\ &\overset{(a)}{=} \sum_{\substack{x_1 \in g^{-1}[y_1] \\ x_n \in g^{-1}[y_n]}} \lceil \Pr(X_1 = x_1, X_n = x_n | Y_1^n = y_1^n) \rceil \\ &\leq \operatorname{card}(\mathcal{X})^2 < \infty \end{aligned}$$

where (a) is due to (B.3). While this holds for all y_1^n with positive probability, the realization of the corresponding preimage count for y_1^n with zero probability is trivially zero. Since this holds for all n, the proof is completed. □

B.1.2 The lossy case

Let us start with some derivations common to the proof of (B.1a) and (B.2a). For $\mathcal{K} < \infty$, (3.14) is equivalent to the existence of $\check{x}, \hat{x} \in \mathcal{X}, y_1^{\mathcal{K}} \in \mathcal{Y}^{\mathcal{K}}, x_1^{\mathcal{K}} \in g^{-1}[y_1^{\mathcal{K}}]$ with

$$0 < \Pr(X_0 = \check{x}, X_1^{\mathcal{K}} = x_1^{\mathcal{K}}, X_{\mathcal{K}+1} = \hat{x}) < \Pr(X_0 = \check{x}, Y_1^{\mathcal{K}} = y_1^{\mathcal{K}}, X_{\mathcal{K}+1} = \hat{x})\,. \tag{B.5}$$

The *unreconstructable set of trajectories* $\mathcal{H}$ is

$$\mathcal{H} := \{\check{x}\} \times g^{-1}[y_1^{\mathcal{K}}] \times \{\hat{x}\}\,. \tag{B.6}$$

Equation (3.14) implies that $\mathcal{H}$ contains at least two elements with positive probability. Passing through $\mathcal{H}$ incurs an information loss[30] L:

$$L := H(X_1^{\mathcal{K}} | X_0^{\mathcal{K}+1} \in \mathcal{H}) > 0\,. \tag{B.7}$$

Let $\mathcal{I}$ be the random set of indices marking the start of non-overlapping runs of X_1^n through $\mathcal{H}$, that is

$$\mathcal{I} := \left\{ i \in \{1, \dots, n-\mathcal{K}-1\} : \begin{array}{c} X_i^{i+\mathcal{K}+1} \in \mathcal{H} \\ \text{and} \\ \forall j \in \{i+1, \dots, i+\mathcal{K}+1\} : X_j^{j+\mathcal{K}+1} \notin \mathcal{H} \end{array} \right\}, \tag{B.8}$$

where lower indices are selected greedily.

The proof of (B.1a) and (B.2a) is based on the fact that the random set $\mathcal{I}$ almost-surely increases linearly with n, in other words, there exists an $\alpha > 0$ such that

$$\lim_{n\to\infty} \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) = 1. \tag{B.9}$$

The proof that such an α exists is technically complicated and is a direct consequence of the ergodic theorem for Markov chains [167, Thm 3.55, p. 69]. The interested reader is referred to [48].

Sketch of the proof of (B.1a). Clearly,

$$\begin{aligned} L(X_1^n \to Y_1^n) &\geq \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) H(X_1^n | Y_1^n, \operatorname{card}(\mathcal{I}) \geq \alpha n) && \text{(B.10)} \\ &= \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) \sum_{\substack{I \subseteq \{1,\dots,n\} \\ \operatorname{card}(I) \geq \alpha n}} \Pr(\mathcal{I} = I | \operatorname{card}(\mathcal{I}) \geq \alpha n) H(X_1^n | Y_1^n, \mathcal{I} = I). && \text{(B.11)} \end{aligned}$$

It now rests to analyze $H(X_1^n | Y_1^n, \mathcal{I} = I)$. Again, the analysis is technically cumbersome and can be found in [48]. However, the intuition behind the result is quite simple: Given the indices where the non-overlapping runs through $\mathcal{H}$ start, with the help of the Markov property and the fact that this property is retained for Cartesian conditioning, one can ensure that

$$H(X_1^n | Y_1^n, \mathcal{I} = I) \geq L \operatorname{card}(I) \tag{B.12}$$

where L is from (B.7). Thus,

$$\begin{aligned} L(X_1^n \to Y_1^n) &\geq \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) \sum_{\substack{I \subseteq \{1,\dots,n\} \\ \operatorname{card}(I) \geq \alpha n}} \Pr(\mathcal{I} = I | \operatorname{card}(\mathcal{I}) \geq \alpha n) L \operatorname{card}(I) && \text{(B.13)} \\ &\geq \alpha n L \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) \sum_{\substack{I \subseteq \{1,\dots,n\} \\ \operatorname{card}(I) \geq \alpha n}} \Pr(\mathcal{I} = I | \operatorname{card}(\mathcal{I}) \geq \alpha n) && \text{(B.14)} \\ &= \alpha n L \Pr(\operatorname{card}(\mathcal{I}) \geq \alpha n) && \text{(B.15)} \end{aligned}$$

Since, by the ergodic theorem, in the limit $n \to \infty$ the probability evaluates to one, the information loss rate is bounded from below by $\alpha L > 0$. □

[30] A similar quantity, namely the entropy of a Markov trajectory given its start- and endpoints, was also considered in [83]. The difference is, however, that the length of the trajectory is defined in this work, while in [83] *only* start- and endpoints are defined. Loosely speaking, while Kafsi et al. consider the randomness of paths from $\check{x}$ to $\hat{x}$, this work considers only such paths of length n.

Proof of (B.2a). Since for each run through $\mathcal{H}$ there are at least two realizable elements in the preimage, it follws that

$$T_n \geq 2^{\mathrm{card}(\mathcal{I})}\,. \tag{B.16}$$

Thus, (B.16) and the ergodic theorem imply that there exists an $\alpha > 0$ such that

$$\begin{aligned}\liminf_{n\to\infty} \sqrt[n]{T_n} &\geq \liminf_{n\to\infty} \exp((\log 2)\frac{1}{n}\mathrm{card}(\mathcal{I})) \\ &= \exp((\log 2)\liminf_{n\to\infty} \frac{1}{n}\mathrm{card}(\mathcal{I})) \overset{\Pr-a.s.}{\geq} \exp((\log 2)\alpha) = 2^\alpha > 1\,. \end{aligned} \tag{B.17}$$

□

B.2 Proof of Proposition 3.2

The proof employs elementary results from graph theory: Let $\mathbf{A}$ denote the adjacency matrix of the Markov chain, i.e., $A_{i,j} = \lceil P_{i,j} \rceil$. The number of closed walks of length k on the graph determined by $\mathbf{A}$ is given as [23, p. 24]

$$\sum_{i=1}^{N} \lambda_i^k \tag{B.18}$$

where $\{\lambda_i\}_{i=1}^N$ is the set of eigenvalues of $\mathbf{A}$.

Let t_X^k denote the number of sequences $x_1^k \in \mathcal{X}^k$ of $\mathbf{X}$ with positive probability, i.e.,

$$t_X^k = \sum_{x_1^k \in \mathcal{X}^k} \lceil \Pr(X_1^k = x_1^k) \rceil\,. \tag{B.19}$$

Clearly, $t_X^k \geq \sum_{i=1}^N \lambda_i^k$. Furthermore, defining t_Y^k similarily one gets $t_Y^k \leq \mathrm{card}(\mathcal{Y})^k$. With $\lambda_{\max}$ denoting the largest eigenvalue of $\mathbf{A}$,

$$\frac{t_X^k}{t_Y^k} \geq \frac{\sum_{i=1}^N \lambda_i^k}{\mathrm{card}(\mathcal{Y})^k} \geq \left(\frac{\lambda_{\max}}{\mathrm{card}(\mathcal{Y})}\right)^k\,. \tag{B.20}$$

If $\lambda_{\max} > \mathrm{card}(\mathcal{Y})$, then the ratio of possible length-k sequences of $\mathbf{X}$ to those of $\mathbf{Y}$ increases exponentially. Then, the *pigeon-hole-principle* implies that also the preimage count T_n is unbounded. Thus,

$$\overline{L}(\mathbf{X} \to \mathbf{Y}) = 0 \Rightarrow \mathrm{card}(\mathcal{Y}) \geq \lambda_{\max}\,. \tag{B.21}$$

Finally, the *Perron-Frobenius theorem* for non-negative matrices [76, Cor. 8.3.3] bounds the largest eigenvalue of $\mathbf{A}$ from below by the minimum out-degree of the associated transition graph. □

B.3 Proof of Proposition 3.3

The r.h.s. of (3.21) is equivalent to

$$\begin{aligned} 0 &= H(Z_k|Z_0^{k-1}) - \overline{H}(\mathbf{Z}) \\ &= H(Z_k|Z_0^{k-1}) - \lim_{n\to\infty} H(Z_n|Z_0^{n-1}) \\ &= \lim_{n\to\infty} \left(H(Z_n|Z_{n-k}^{n-1}) - H(Z_n|Z_0^{n-1})\right) \end{aligned}$$

$$= \lim_{n\to\infty} I(Z_n; Z_0^{n-k-1}|Z_{n-k}^{n-1})\,.$$

By stationarity, the sequence in the last limit increases monotonically in n. A limit value of zero is equivalent to, for all $n \in \mathbb{N}$:

$$\begin{aligned}
&p_{Z_n|Z_{n-k}^{n-1}}(\cdot|\mathbf{z}) p_{Z_0^{n-k-1}|Z_{n-k}^{n-1}}(\cdot|\mathbf{z}) \\
&= p_{Z_n, Z_0^{n-k-1}|Z_{n-k}^{n-1}}(\cdot|\mathbf{z}) \\
&= p_{Z_n|Z_0^{n-1}}(\cdot|\cdot, \mathbf{z}) p_{Z_0^{n-k-1}|Z_{n-k}^{n-1}}(\cdot|\mathbf{z})\,,
\end{aligned}$$

where the first equality holds $p_{Z_{n-k}^{n-1}}$-a.s. The equality between the first and last line is equivalent to the higher-order Markov property (3.20). □

B.4 Proof of Proposition 3.14

A possible definition of Markovity is

Definition B.1 (Markov Process [30, II.6, p. 80])**.** A process $\mathbf{X}$ is a Markov process iff for all $i \in \mathbb{N}$, $a \in \mathbb{R}$, and integers $n_1 < n_2 < \cdots < n_i < n$, with probability one,

$$\Pr(X_n \le a | X_{n_1} = x_{n_1}, \ldots, X_{n_i} = x_{n_i}) = \Pr(X_n \le a | X_{n_i} = x_{n_i}). \tag{B.22}$$

Clearly, a process is Markov if, for all n,

$$f_{X_n|X_1^{n-1}} \overset{a.e.}{=} f_{X_n|X_{n-1}} \tag{B.23}$$

holds $P_{X_1^{n-1}}$-a.s. because (B.22) results from integrating the densities over $(-\infty, a]$.

The proof of the proposition follows along the same lines as the proof for Markov chains given in [48], and is built on the following Lemma, which is an extension of [21, Thm. 4.5.1]:

Lemma B.1 (Bounds on the differential entropy rate)**.** *Let* $\mathbf{X}$ *be a stationary Markov process with differential entropy rate* $\bar{h}(\mathbf{X}) = h(X_2|X_1)$ *and let* $\mathbf{Y}$ *be a stationary process derived from* $\mathbf{X}$ *by* $Y_n := g(X_n)$*, where* g *is piecewise bijective. Then,*

$$h(Y_n|Y_2^{n-1}, X_1) \le \bar{h}(\mathbf{Y}) \le h(Y_n|Y_1^{n-1}). \tag{B.24}$$

Proof. The upper bound follows from the fact that conditioning reduces entropy, so only the lower bound has to be shown. By Markovity of $\mathbf{X}$,

$$h(Y_n|Y_2^{n-1}, X_1) = h(Y_n|Y_2^{n-1}, X_k^1) \tag{B.25}$$

for all $k < 1$. Let $U_k = (Y_2^{n-1}, X_k^1)$ and $V_k = Y_k^{n-1}$. Obviously, there exists a function f such that $V_k = f(U_k)$, namely the function which is the identity function on the last $n-2$, and the function g on the first $2-k$ elements. By showing that

$$h(Y_n|U_k) \le h(Y_n|V_k) \tag{B.26}$$

the lower bound is proved by [114, Thm. 14.7]

$$h(Y_n|Y_2^{n-1}, X_1) = \lim_{k\to-\infty} h(Y_n|U_k) \le \lim_{k\to-\infty} h(Y_n|V_k) = \bar{h}(\mathbf{Y}). \tag{B.27}$$

To show the inequality,

$$h(Y_n|V_k) - h(Y_n|U_k) = h(Y_n, V_k) - h(V_k) - h(Y_n, U_k) + h(U_k) \tag{B.28}$$

$$\begin{aligned}
&\overset{(a)}{=} H(U_k|V_k) - \mathbb{E}\left(\log|\det \mathcal{J}_f(U_k)|\right) \\
&\quad - H(U_k, Y_n|V_k, Y_n) + \mathbb{E}\left(\log|\det \mathcal{J}_f(U_k)|\right) && \text{(B.29)}\\
&= H(U_k|V_k) - H(U_k|V_k, Y_n) && \text{(B.30)}\\
&\geq 0 && \text{(B.31)}
\end{aligned}$$

where (a) is due to Theorem 2.2 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. □

Proof of Proposition 3.14. Note that the assumption implies that

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} f_{Y_2,X_1}(y,x)\log\left(\frac{f_{Y_2|X_1}(y|x)}{f_{Y_2|Y_1}(y|g(x))}\right)dydx = h(Y_2|Y_1) - h(Y_2|X_1) = 0 \tag{B.32}$$

which renders the upper bounds of Lemma B.1 equal for $n = 2$. Thus, $\bar{h}(\mathbf{Y}) = h(Y_n|Y_1^{n-1}) = h(Y_2|Y_1)$ for all n. By stationarity,

$$\begin{aligned}
0 &= h(Y_n|Y_{n-1}) - h(Y_n|Y_1^{n-1}) && \text{(B.33)}\\
&= I(Y_n; Y_1^{n-2}|Y_{n-1}) && \text{(B.34)}\\
&= \mathbb{E}\left(\log\left(\frac{f_{Y_n,Y_1^{n-2}|Y_{n-1}}(Y_1^n)}{f_{Y_n|Y_{n-1}}(Y_{n-1}^n) f_{Y_1^{n-2}|Y_{n-1}}(Y_1^{n-1})}\right)\right) && \text{(B.35)}\\
&= \mathbb{E}\left(\log\left(\frac{f_{Y_n|Y_1^{n-1}}(Y_1^n)}{f_{Y_n|Y_{n-1}}(Y_{n-1}^n)}\right)\right) && \text{(B.36)}\\
&= \mathbb{E}\left(D(f_{Y_n|Y_1^{n-1}}(\cdot, Y_1^{n-1})||f_{Y_n|Y_{n-1}}(\cdot, Y_{n-1}))\right) && \text{(B.37)}
\end{aligned}$$

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

$$f_{Y_n|Y_1^{n-1}} \overset{a.e.}{=} f_{Y_n|Y_{n-1}} \tag{B.38}$$

$P_{Y_1^{n-1}}$-a.s. But this implies Markovity by Definition B.1 (cf. (B.23)) and completes the proof. □

B.5 Proof of Corollary 3.3

Note that (3.51) 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

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{1}{f_Y(y_1)}\sum_{x_1\in g^{-1}[y_1]}\frac{f_{Y_2|X_1}(y_2|x_1)f_X(x_1)}{|g'(x_1)|}. \tag{B.39}$$

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

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{1}{f_Y(y_1)}\sum_{x_1\in g_+^{-1}[y_1]}\frac{f_{Y_2|X_1}(y_2|x_1)f_X(x_1)}{|g'(x_1)|} \tag{B.40}$$

$$\stackrel{(a)}{=} \frac{f_{Y_2|X_1}(y_2|\hat{x})}{f_Y(y_1)} \sum_{x_1 \in g_+^{-1}[y_1]} \frac{f_X(x_1)}{|g'(x_1)|} \tag{B.41}$$

$$= f_{Y_2|X_1}(y_2|\hat{x}) \tag{B.42}$$

where (a) is due to (3.51). Since $f_{Y_2,X_1} = f_{Y_2|X_1} f_X$, one can apply Proposition 3.14 to complete the first part of the proof.

For the second part, note that with Proposition 3.14

$$\bar{h}(\mathbf{Y}) = h(Y_2|X_1) \tag{B.43}$$

and thus, with Proposition 3.8 and Theorem 2.2,

$$\bar{L}(\mathbf{X} \to \mathbf{Y}) = h(X_2|X_1) - h(Y_2|X_1) + \mathbb{E}\left(\log|g'(X)|\right) = H(X_2|X_1, Y_2). \tag{B.44}$$

It remains to show that (3.52) implies equality in (3.48) in the proof of Proposition 3.13. To this end, observe that

$$H(W_2|X_1) - H(W_2|X_1, Y_2) = I(W_2; Y_2|X_1) \tag{B.45}$$

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 $\Pr(W_2 = w|X_1 = x) > 0$,

$$f_{Y_2|X_1}(y|x) = f_{Y_2|W_2,X_1}(y|w, x). \tag{B.46}$$

But

$$f_{Y_2|W_2,X_1}(y|w, x) = \frac{f_{X_2|X_1}(g_w^{-1}(y)|x)}{p(w|x)|g'(g_w^{-1}(y))|} \tag{B.47}$$

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

$$f_{Y_2|X_1}(y|x) = \sum_w p(w|x) f_{Y_2|W_2,X_1}(y|w, x) \tag{B.48}$$

$$= \sum_w \frac{f_{X_2|X_1}(g_w^{-1}(y)|x)}{|g'(g_w^{-1}(y))|} \tag{B.49}$$

$$\stackrel{(a)}{=} \frac{f_{X_2|X_1}(g_{\hat{w}}^{-1}(y)|x)}{|g'(g_{\hat{w}}^{-1}(y))|} \sum_w [p(w|x) > 0] \tag{B.50}$$

$$= \frac{f_{X_2|X_1}(g_{\hat{w}}^{-1}(y)|x)}{|g'(g_{\hat{w}}^{-1}(y))|} \operatorname{card}(\{w : p(w|x) > 0\}) \tag{B.51}$$

$$\stackrel{(b)}{=} \frac{f_{X_2|X_1}(g_{\hat{w}}^{-1}(y)|x)}{|g'(g_{\hat{w}}^{-1}(y))|} \frac{1}{p(\hat{w}|x)} \tag{B.52}$$

$$= f_{Y_2|W_2,X_1}(y|\hat{w}, x) \tag{B.53}$$

where (a) is due to (3.52a) and (b) is due to (3.52b). □

B.6 Proof of Proposition 3.16

That $l(X \to Y) = P_X(\mathcal{X}_c)$ follows from Proposition 2.10 or Corollary 2.5.

Now take a finite sequence X_1^n obtained from the stochastic process $\mathbf{X}$ and look at the relative information loss incurred in g. Similarly as in the proof of Proposition 3.12, g^n induces a

finite partition of $\mathcal{X}^n$. Moreover, for every element of this partition, g^n is a composition of a bi-Lipschitz function (which does not influence the information dimension) and, possibly, a projection. One can thus apply Proposition 2.10 which leads to

$$\begin{aligned}
&l(X_1^n \to Y_1^n) = P_{X_1^n}(\mathcal{X}_c^n) \\
&+ \frac{n-1}{n} P_{X_1^n}(\mathcal{X}_c^{n-1} \times \overline{\mathcal{X}_c}) + \frac{n-1}{n} P_{X_1^n}(\mathcal{X}_c^{n-2} \times \overline{\mathcal{X}_c} \times \mathcal{X}_c) + \cdots + \frac{n-1}{n} P_{X_1^n}(\overline{\mathcal{X}_c} \times \mathcal{X}_c^{n-1}) \\
&\vdots \\
&+ \frac{1}{n} P_{X_1^n}(\mathcal{X}_c \times \overline{\mathcal{X}_c}^{n-1}) + \frac{1}{n} P_{X_1^n}(\overline{\mathcal{X}_c} \times \mathcal{X}_c \times \overline{\mathcal{X}_c}^{n-2}) + \cdots + \frac{1}{n} P_{X_1^n}(\overline{\mathcal{X}_c}^{n-1} \times \mathcal{X}_c) \\
&+ \frac{0}{n} P_{X_1^n}(\overline{\mathcal{X}_c}^{n})
\end{aligned} \tag{B.54}$$

where $\overline{\mathcal{X}_c} = \mathcal{X} \setminus \mathcal{X}_c$. Compactly written this yields

$$l(X_1^n \to Y_1^n) = \frac{1}{n} \sum_{i=1}^{n} i \Pr(\operatorname{card}(\{X_j \in X_1^n : X_j \in \mathcal{X}_c\}) = i). \tag{B.55}$$

Defining

$$V_n := \begin{cases} 1, & \text{if } X_n \in \mathcal{X}_c \\ 0, & \text{else} \end{cases} \tag{B.56}$$

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

$$\begin{aligned}
l(X_1^n \to Y_1^n) &= \frac{1}{n} \sum_{i=1}^{n} i \Pr(\sum_{j=1}^{n} V_j = i) && \text{(B.57)} \\
&= \frac{1}{n} \sum_{i=1}^{n} i \Pr(Z_n = i) && \text{(B.58)} \\
&= \frac{1}{n} \mathbb{E}(Z_n) && \text{(B.59)} \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}(V_j) && \text{(B.60)} \\
&\overset{(a)}{=} \mathbb{E}(V) && \text{(B.61)} \\
&= P_X(\mathcal{X}_c) && \text{(B.62)}
\end{aligned}$$

where (a) is due to stationarity of $\mathbf{X}$. This completes the proof. □

B.7 Proof of Lemma 3.3

By assumption, $|h(X)| < \infty$ and $|\bar{h}(\mathbf{X})| < \infty$. But since $\bar{h}(\mathbf{X}) = \lim_{n\to\infty} h(X_n|X_1^{n-1})$ it also follows from conditioning [21, Cor. to Thm. 8.6.1, p. 253] that

$$-\infty < \bar{h}(\mathbf{X}) \leq h(X_n|X_1^{n-1}) \leq h(X) < \infty \tag{B.63}$$

and similarly, by the chain rule of differential entropy [21, Thm. 8.6.2, p. 253]

$$-\infty < n\bar{h}(\mathbf{X}) \leq h(X_1^n) \leq nh(X) < \infty. \tag{B.64}$$

Now take $X_{\mathbb{J}} := \{X_j : j \in \mathbb{J}\}$. By conditioning and the chain rule,

$$-\infty < \mathrm{card}(\mathbb{J})\bar{h}(\mathbf{X}) \leq h(X_{\mathbb{J}}) \leq \mathrm{card}(\mathbb{J})h(X) < \infty. \tag{B.65}$$

By the assumption that $H(\hat{X}^{(0)}) < \infty$ it follows that the information dimension of X exists (and similarly for any finite collection $X_{\mathbb{J}}$ of samples). But $h(X_{\mathbb{J}})$ is the $\mathrm{card}(\mathbb{J})$-dimensional entropy of $X_{\mathbb{J}}$, which can only be finite[31] if $d(X_{\mathbb{J}}) = \mathrm{card}(\mathbb{J})$ [123]. This completes the proof. □

B.8 Proof of Lemma 3.4

If the filter is stable (i.e., the impulse response is absolutely summable [111, Ch. 2.4, p. 59] and, thus, square summable) and causal, it follows by the Paley-Wiener theorem [112, p. 215] that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\mathrm{e}^{\jmath\theta})|d\theta > -\infty. \tag{B.66}$$

Moreover, since the filter is stable, one has by Jensen's inequality:

$$\begin{aligned}\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln|H(\mathrm{e}^{\jmath\theta})|^2 d\theta &= \frac{1}{2}\mathbb{E}\left(\ln|H(\mathrm{e}^{\jmath\theta})|^2\right) &\text{(B.67)}\\ &\leq \frac{1}{2}\ln\left(\mathbb{E}\left(|H(\mathrm{e}^{\jmath\theta})|^2\right)\right) &\text{(B.68)}\\ &= \frac{1}{2}\ln G < \infty &\text{(B.69)}\end{aligned}$$

where the expectation is taken assuming the frequency variable θ is uniformly distributed on $[-\pi, \pi]$ and where the noise gain G is

$$G := \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\mathrm{e}^{\jmath\theta})|^2 d\theta. \tag{B.70}$$

The last (strict) inequality follows by assuming stability (square summability of the impulse response and Parseval's theorem [111, Tab. 2.2, p. 86]).

According to [114, p. 663], the differential entropy rate at the output of the filter H is given by

$$\bar{h}(\tilde{\mathbf{X}}) = \bar{h}(\mathbf{X}) + \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\mathrm{e}^{\jmath\theta})|d\theta \tag{B.71}$$

and thus, by assumption, finite.

With [114, (9-190), p. 421], one has for the autocorrelation function of the output of a linear filter

$$r_{\tilde{X}\tilde{X}}[m] = (r_{XX} * \rho)[m] \tag{B.72}$$

where

$$\rho[m] = \sum_k h[m+k]h^*[k]. \tag{B.73}$$

[31] If $d(X) < d$, the d-dimensional entropy of X will be $-\infty$; if $d(X) > d$, the d-dimensional entropy of X will be ∞.

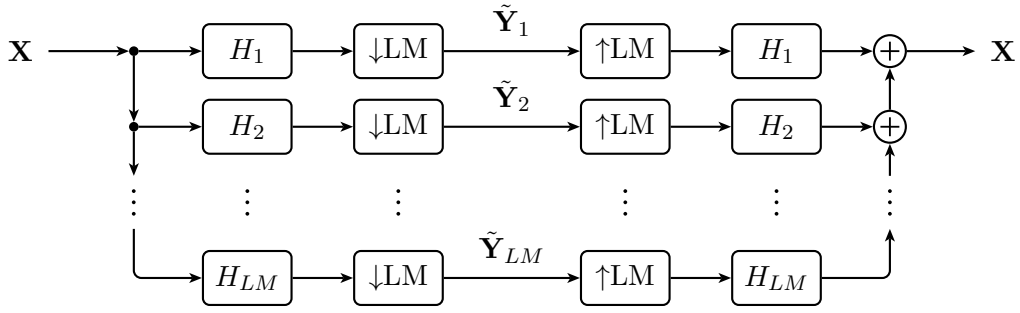


Figure B.1: Filterbank decomposition of the input process $\mathbf{X}$

Thus, by the fact that for a zero-mean process the variance satisfies $\sigma^2 = r_{XX}[0] \geq r_{XX}[m]$,

$$
\begin{aligned}
r_{\tilde{X}\tilde{X}}[0] &\leq \sigma^2 \sum_m \rho[m] && \text{(B.74)}\\
&\leq \sigma^2 \sum_m \sum_k |h[m+k]h^*[k]| && \text{(B.75)}\\
&= \sigma^2 \sum_k \left(|h^*[k]| \sum_m |h[m+k]| \right) && \text{(B.76)}\\
&\leq \sigma^2 \sum_k \left(|h^*[k]| C \right) && \text{(B.77)}\\
&\leq \sigma^2 C^2 < \infty && \text{(B.78)}
\end{aligned}
$$

where the last two lines follow from stability of H (the impulse response is absolutely summable) and by the assumption that $\mathbf{X}$ has finite variance. Thus, by conditioning and the maximum-entropy property of the Gaussian distribution,

$$-\infty < \bar{h}(\tilde{\mathbf{X}}) \leq h(\tilde{X}) \leq \frac{1}{2} \ln(2\pi \mathrm{e} \sigma^2 C^2) < \infty.$$

It remains to show that $H(\hat{\tilde{X}}^{(0)}) < \infty$. To this end, note that the variance of $\tilde{\mathbf{X}}$ is finite. Thus, with [169, Prop. 1], the desired result follows. This completes the proof. □

B.9 Proof of Theorem 3.3

The case $H \equiv 1$ and the case of a stable and causal H have already been dealt with. Thus, assume that H is piecewise constant with $H(\mathrm{e}^{\jmath\theta})$ being either one or zero. This assumption is unproblematic, since H can always be written as a cascade of a filter satisfying this assumption and a set of filters satisfying the Paley-Wiener condition. The latter filters can be omitted as made clear by Lemma 3.4.

Next, assume that the pass-band and stop-band intervals have rational endpoints. In other words, and since there are only finitely many such intervals, there exists an even integer L large enough such that the pass-band interval endpoints are integer multiples of $1/L$. With this in mind, observe Fig. B.1 which illustrates the filterbank decomposition of $\mathbf{X}$ [111, Ch. 4.7.6, p. 230]. There, H_i is an ideal brick-wall filter for the i-th frequency band, i.e.,

$$H_i(\mathrm{e}^{\jmath\theta}) = \begin{cases} 1, & \text{if } \frac{(i-1)\pi}{LM} < |\theta - 2k\pi| \leq \frac{i\pi}{LM}, \quad k \in \mathbb{Z} \\ 0, & \text{else} \end{cases}. \qquad \text{(B.79)}$$

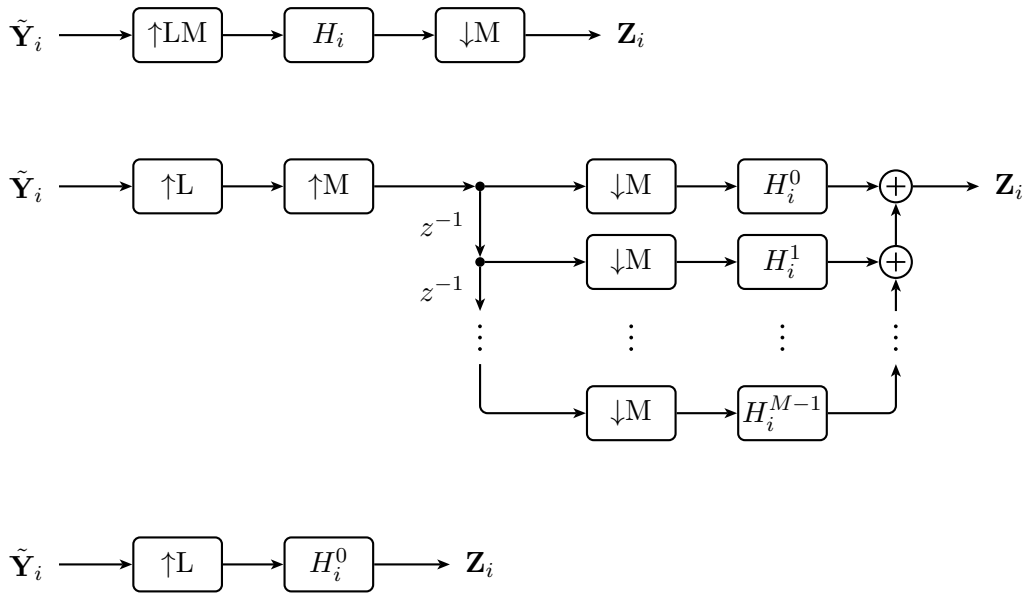


Figure B.2: All three systems are equivalent. The first equivalence is due to the polyphase decomposition of the filter H_i followed by the M-fold downsampler. The second equivalence is due to the fact that all but the first branch have a signal identical to zero. H_i^0 is the M-fold downsampled filter H_i, i.e., it has impulse response $h_i[nM]$. By linearity, $\mathbf{Y}$ is the sum of the processes $\mathbf{Z}_i$, $i = 1, \ldots, LM$.

Since $\mathbf{X}$ is a Gaussian process [114, p. 663],

$$\bar{h}(\mathbf{X}) = \frac{1}{2}\ln(2\pi\mathrm{e}) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln S_X(\mathrm{e}^{\jmath\theta})d\theta, \tag{B.80}$$

and from $|\bar{h}(\mathbf{X})| < \infty$ follows that $S_X(\mathrm{e}^{\jmath\theta}) > 0$ a.e. It also follows that, for all $i = 1, \ldots, LM$, $S_{\tilde{Y}_i}(\mathrm{e}^{\jmath\theta}) > 0$ a.e., where $\tilde{\mathbf{Y}}_i$ is Gaussian too.

The variance of the i-th downsampled process $\tilde{\mathbf{Y}}_i$ is positive (since its PSD is positive a.e.) and finite (since it is upper bounded by LM times the variance of $\mathbf{X}$). Thus, $|h(\tilde{Y}_i)| < \infty$, and $\bar{h}(\tilde{\mathbf{Y}}_i) < \infty$. The differential entropy rates of $\tilde{\mathbf{Y}}_i$ are obtained by splitting the integral in (B.80) into LM parts; the sum of these LM parts is $-\infty$ if at least one of its parts is $-\infty$ (since none of these parts can be ∞ by the fact that $|h(\tilde{Y}_i)| < \infty$). Thus, $|\bar{h}(\tilde{\mathbf{Y}}_i)| < \infty$, and with $H(\hat{\tilde{Y}}_i^{(0)}) < \infty$ (from finite variance), it follows that

$$d((\tilde{Y}_i)_1^n) = n \tag{B.81}$$

for all i. This is intuitive, since the collection $\tilde{\mathbf{Y}} := \{\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{LM}\}$ is equivalent to $\mathbf{X}$, in the sense that perfect reconstruction is possible. By this equivalence, one can abuse notation to get

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) = l(\tilde{\mathbf{Y}} \to \mathbf{Y}^{(L)}). \tag{B.82}$$

One now employs the linearity of the system to move the filter H next to the reconstruction filters H_i. By the assumption made about the pass-bands of H, the cascade of H and H_i either equals H_i or is identical to zero. The filter H thus amounts to eliminating some of the sub-band processes $\tilde{\mathbf{Y}}_i$; a simple projection. What remains to be analyzed is the effect of the M-fold downsampler, which can also be moved next to the reconstruction filters due to linearity. Notice that with the polyphase decomposition of decimation systems (cf. [111, Ch. 4.7.4, p. 228]), the

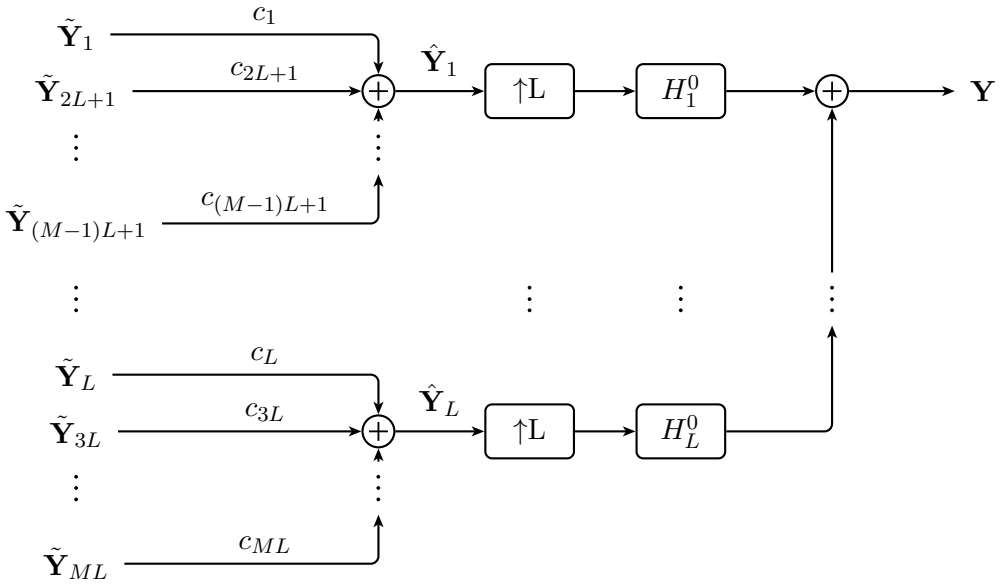


Figure B.3: Equivalent system for a decimation filter with a piecewise constant H (pass-band intervals with rational endpoints). The constants c_i indicate whether or not the sub-band process is eliminated by H, i.e., $c_i \in \{0,1\}$. Note that with (B.84) *the interpolator outputs can be added without information loss. Thus, information loss only occurs by eliminating and/or adding sub-band processes – a cascade of an intertible linear map and a projection. The system is shown for M odd. Note that $\hat{Y}_i$ depends on $\{\tilde{Y}_i, \tilde{Y}_{2L-i+1}, \tilde{Y}_{2L+i}, \tilde{Y}_{4L-i+1}, \dots\}$.*

i-th branch of the filterbank can be rearranged as in Fig. B.2. Due to the cascade of up- and downsampling only the filter H_i^0 is relevant, while all other filters H_i^l will have an all-zero input. In particular, while H_i is given by (B.79), one gets for the filter H_i^0, with impulse response $h_i[nM]$,

$$H_i^0(\mathrm{e}^{\jmath\theta}) = \frac{1}{M}\sum_{m=1}^{M} H_i(\mathrm{e}^{\jmath\frac{\theta-2m\pi}{M}}) \tag{B.83}$$

$$= \begin{cases} \frac{1}{M}, & \text{if } \frac{(i-1)\pi}{L} < |\theta - 2k\pi| \leq \frac{i\pi}{L}, \quad k \in \mathbb{Z} \\ 0, & \text{else} \end{cases} \tag{B.84}$$

where the last line follows from the fact that downsampling the filter impulse response causes no aliasing (H_i have bandwidths $1/LM$ and fall in exact one of the bands with width $1/M$).

By the 2π-periodicity of the transfer functions, the sequence of filters is periodic with $2L$, i.e., $H_i^0 = H_{i+2L}^0$. Moreover, $H_{L+k}^0 = H_{L-k+1}^0$, $k = 1, \dots, L$, by the symmetry of the filter. Therefore, there are exactly L different filters, each occurring M times.

Combining the last system from Fig. B.2 with Fig. B.1 and (B.84) the schematic in Fig. B.3 is obtained. Note that since the filters H_i^0 are orthogonal and L-aliasfree (i.e., the frequency response of the filter does not overlap after L-fold downsampling and can thus be reconstructed perfectly), adding the reconstruction filter outputs does not incur information loss. One can thus again abuse notation and write

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) = l(\tilde{\mathbf{Y}} \to \mathbf{Y}^{(L)}) = l(\tilde{\mathbf{Y}} \to \hat{\mathbf{Y}}) \tag{B.85}$$

where $\hat{\mathbf{Y}} := \{\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_L\}$. The transform from $\tilde{\mathbf{Y}}$ to $\hat{\mathbf{Y}}$ is linear, specifically, the cascade of an

intertible linear map and a projection. One can therefore apply Proposition 2.10 and gets

$$l(\tilde{Y}_1^n \to \hat{Y}_1^n) = 1 - \frac{d(\hat{Y}_1^n)}{d(\tilde{Y}_1^n)} = 1 - \frac{d(\hat{Y}_1^n)}{nML}. \tag{B.86}$$

The information dimension of $\hat{Y}_1^n =: \{(\hat{Y}_1)_1^n, \ldots, (\hat{Y}_L)_1^n\}$ is bounded from above by the number of its scalar components, which is nL. This completes the proof for filters H with rational endpoints of the pass-band intervals.

Assume now that one of the interval endpoints is an irrational a_i. Then, for a fixed L, there exists $A_i \in \mathbb{N}_0$ such that $A_i/L < a_i < (A_i + 1)/L$. Obviously, the filter with the irrational endpoint replaced by either of these two rational endpoints destroys either more or less information (either the corresponding coefficient c_m in Fig. B.3 is zero or one). For both of these filters, however, the information dimension of $\hat{Y}_1^n$ cannot exceed nL, and above analysis holds. This completes the proof. □

Note also that the proof suggests how to measure the exact relative information loss rate for the decimation system by evaluating the information dimension of $\hat{Y}_1^n$. For rational endpoints of the pass-band intervals this is simple since $d((\hat{Y}_i)_1^n) \in \{0, n\}$. For irrational endpoints one can always wedge the filter H between one which destroys more and one which destroys less information; for L sufficiently large, the resulting difference in the relative information loss rate will be small, and eventually vanish in the limit.

Moreover, the result should also hold for non-Gaussian processes satisfying Assumption 1. The intuition behind this is due to a bottleneck consideration: Since the filterbank decomposition is perfectly invertible, the information dimension of the input and output process need to be identical for all time windows $\{1, \ldots, n\}$. The Gaussian assumption was needed to show that the information dimension (for a given time window) of each sub-band process is related to the information dimension of the input process (in the same time window) and the number of filterbank channels. The author believes that the Gaussian assumption can be removed by the fact that all operations in the model are Lipschitz, and that therefore the information dimension cannot increase. As a consequence, it is not possible that the information dimension of the sub-band processes is smaller than in the Gaussian case, since in this case the information dimension of the (reconstructed) output would be smaller than the information dimension of the input – a contradiction.

C Proofs from Chapter 4

C.1 Proof of Proposition 4.5

For the first property, note that with $\mathbf{P}' = \mathbf{VQU}^{\pi}$ the condition

$$\mathbf{VUP}'\mathbf{V} = \mathbf{P}'\mathbf{V} \tag{C.1}$$

from Lemma 4.2 can be written as

$$\mathbf{VUVQU}^{\pi}\mathbf{V} = \mathbf{VQU}^{\pi}\mathbf{V}. \tag{C.2}$$

Since $\mathbf{U}^{\pi}\mathbf{V} = \mathbf{I}$ for all positive probability vectors π, equality is achieved and the first result is proved. By the same line of argument follows that $\mathbf{UP}'\mathbf{V} = \mathbf{Q}$. Note in passing that this actually proved *strong* lumpability of $\mathbf{X}'^{\pi}_{g}$ (cf. [87]).

For the second property (cf. [27, Thm. 2, Property 3]) note that with $\mathbf{P}' = \mathbf{VQU}^{\mu}$,

$$\mu^T\mathbf{P}' = \mu^T\mathbf{VQU}^{\mu} = \nu^T\mathbf{QU}^{\mu} = \nu^T\mathbf{U}^{\mu} = \mu^T \tag{C.3}$$

where the second and third equality are due to Lemma 4.2; the last one follows from the definition of $\mathbf{U}^{\mu}$ in (4.68).

For the third property the reader is referred to [27, Thm. 1].

Next, observe that $Q_{g(i)g(j)} = 0$ implies $P_{ij} = 0$; see (4.82), combined with the fact that μ is positive [87, Thm. 4.1.4]. The entries of the lifted matrix are given as

$$P'_{ij} = \frac{\pi_j}{\sum_{k\in g^{-1}[g(j)]}\pi_k} Q_{g(i)g(j)}. \tag{C.4}$$

Since π is positive, it follows that $P'_{ij} = 0$ implies $P_{ij} = 0$, or equivalently, $\mathbf{P}' \gg \mathbf{P}$.

The last property immediately follows from Lemma 4.3; the lemma can be applied because $\mathbf{X}'^{\mu}_{g}$ is lumpable to $\mathbf{Y}'_{g}$ (property 1) and $\mathbf{P}' \gg \mathbf{P}$ (property 4). This completes the proof. □

C.2 Proof of Theorem 4.1

For the proof note that another condition for lumpability is given by the entries of the matrix $\mathbf{R} = \mathbf{PV}$. In particular, if and only if for all $h, l \in \mathcal{Y}$ the elements

$$R_{il} := \sum_{j \in g^{-1}[l]} P_{ij} \tag{C.5}$$

are the same for all $i \in g^{-1}[h]$, the chain is strongly lumpable w.r.t. g [87]. Using this with (4.86) one gets

$$\hat{R}_{il} = \sum_{j \in g^{-1}[l]} \hat{P}_{ij} = Q_{hl}. \tag{C.6}$$

Clearly, $\hat{R}_{il}$ assumes the same values for all $i \in g^{-1}[h]$, as required. This completes the proof of the first statement[32]. Again, strong lumpability in the sense of [87] was proved.

The second statement is obvious from the definition of $\mathbf{P}$-lifting and from the proof of property 4 of Proposition 4.5.

For the third statement, introduce an arbitrary lifting

$$\tilde{P}_{ij} = b_{ij} Q_{g(i)g(j)} \tag{C.7}$$

subject to $\sum_{j \in g^{-1}[l]} b_{ij} = 1$ for all $l \in \mathcal{Y}$ and all $i \in \mathcal{X}$. With (C.5), this condition is necessary and sufficient for strong lumpability of the lifted chain $\tilde{\mathbf{X}}$ with transition matrix $\tilde{\mathbf{P}}$. Now write for the KLDR

$$\bar{D}(\mathbf{X}||\tilde{\mathbf{X}}) = \sum_{i,j \in \mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{\tilde{P}_{ij}} \tag{C.8}$$

$$= \sum_{i,j \in \mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{b_{ij} Q_{g(i)g(j)}} \tag{C.9}$$

$$= \overline{H}(\mathbf{Y}'_g) - \overline{H}(\mathbf{X}) + \sum_{i,j \in \mathcal{X}} \mu_i P_{ij} \log \frac{1}{b_{ij}}. \tag{C.10}$$

The last sum can be written as

$$\sum_{i,j \in \mathcal{X}} \mu_i P_{ij} \log \frac{1}{b_{ij}} = \sum_{i \in \mathcal{X}} \mu_i \sum_{l \in \mathcal{Y}} R_{il} \sum_{j \in g^{-1}[l]} \frac{P_{ij}}{R_{il}} \log \frac{1}{b_{ij}}. \tag{C.11}$$

Here, the last term on the right is a cross-entropy, since both b_{ij} and $\frac{P_{ij}}{R_{il}}$ are probability vectors on $g^{-1}[l]$. The cross-entropy is minimized[33] if and only if for all $j \in g^{-1}[l]$

$$b_{ij} = \frac{P_{ij}}{R_{il}} = \frac{P_{ij}}{\sum_{k \in g^{-1}[l]} P_{ik}}. \tag{C.12}$$

Since the sums over i and l are expecations, the minimum is achieved if and only if above condition holds also for all $i \in \mathcal{X}$ and all $l \in \mathcal{Y}$ for which $\mu_i R_{il} > 0$.

To show that the $\mathbf{P}$-lifting indeed yields a better bound, observe that with $H(Y_{g,n}|Y_{g,n-1}) =$

[32] Note that this statement holds for all stochastic matrices used for lifting, i.e., the lifting matrix does not have to be equal to the transition matrix of the original chain.

[33] This is a direct consequence of the fact that the Kullback-Leibler divergence vanishes if and only if the considered probability mass functions are equal [21, pp. 31].

$\overline{H}(\mathbf{Y}'_g)$ and with (4.93)

$$\begin{aligned}
\bar{D}(\mathbf{X}||\mathbf{X}'^{\mu}_g) - \bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) &= H(X) - \overline{H}(\mathbf{X}) - H(Y'_g) + H(Y_{g,n}|X_{n-1}) && \text{(C.13)}\\
&= H(X_n) - H(X_n|X_{n-1}) - H(Y_{g,n}) + H(Y_{g,n}|X_{n-1}) && \text{(C.14)}\\
&= I(X_n; X_{n-1}) - I(Y_{g,n}; X_{n-1}) && \text{(C.15)}\\
&\geq 0 && \text{(C.16)}
\end{aligned}$$

by the data processing inequality. $\bar{D}(\mathbf{X}||\mathbf{X}'^{\mathbf{P}}_g) \geq \bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$ is obtained by Lemma 4.3, which can be applied by properties 1 and 2.

For the fifth property, note that the sufficient and necessary condition for strong lumpability (4.69), namely that

$$R_{il} = \sum_{j \in g^{-1}[l]} P_{ij} = Q_{hl} \tag{C.17}$$

is the same for all $i \in g^{-1}[h]$, can be used in the definition of $\hat{\mathbf{P}}$:

$$\hat{P}_{ij} = \frac{P_{ij}}{\sum_{k \in \mathcal{S}_j} P_{ik}} Q_{g(i)g(j)} = \frac{P_{ij}}{\sum_{k \in \mathcal{S}_j} P_{ik}} R_{ig(j)} = P_{ij} \tag{C.18}$$

This proves the "⇒" part. The "⇐" part follows from Lemma 4.3. This completes the proof. □

C.3 Proof of Theorem 4.2

By Lemma 4.5 and Definition 4.9,

$$\begin{aligned}
L_S(X \to Y_1^M) &= h(Y_{M+1}^N|Y_1^M) - h(\tilde{N}_{M+1}^N|\tilde{N}_M) && \text{(C.19)}\\
&= h((Y_{M+1}^N)_G|(Y_1^M)_G) - \tilde{J}(Y_{M+1}^N|Y_1^M) && \text{(C.20)}\\
&\quad - h((\tilde{N}_{M+1}^N)_G|(\tilde{N}_1^M)_G) + \tilde{J}(\tilde{N}_{M+1}^N|\tilde{N}_1^M) && \text{(C.21)}\\
&\leq h((Y_{M+1}^N)_G|(Y_1^M)_G) - h((\tilde{N}_{M+1}^N)_G|(\tilde{N}_1^M)_G) && \text{(C.22)}\\
&\overset{(a)}{=} h((Y_{M+1}^N)_G) - h((\tilde{N}_{M+1}^N)_G|(\tilde{N}_1^M)_G) && \text{(C.23)}\\
&\overset{(b)}{=} h((Y_{M+1}^N)_G) - h((\tilde{N}_{M+1}^N)_G) && \text{(C.24)}\\
&= \frac{1}{2} \ln \left(\prod_{i=M+1}^{N} \frac{\lambda_i}{\mu} \right). && \text{(C.25)}
\end{aligned}$$

Here, (a) is due to the fact that the PCA decorrelates the output data Y and thus leads to independence of $(Y_{M+1}^N)_G$ and $(Y_1^M)_G$ (in the sense of Definition 4.9). By similar reasons (b) follows from the fact that N is iid ($\mathbf{C}_N$ is a scaled identity matrix, with all eigenvalues being equal $\mu_i = \mu$). Since the PCA ensures that the product contains the $N - M$ smalles eigenvalues λ_i of $\mathbf{C}_Y$, the last line (obtained with [21, Thm. 8.4.1] and [10, Fact 5.10.14]) represents the smallest Gaussian upper bound and completes the proof. □

C.4 Proof of Theorem 4.3

Recapitulate (C.23) from the proof of Theorem 4.2:

$$\begin{aligned}
L_S(X \to Y_1^M) &\leq h((Y_{M+1}^N)_G) - h((\tilde{N}_{M+1}^N)_G|(\tilde{N}_1^M)_G) && \text{(C.26)}\\
&= h((Y_{M+1}^N)_G) - h(\tilde{N}_G) + h((\tilde{N}_1^M)_G). && \text{(C.27)}
\end{aligned}$$

With [21, Thm. 8.4.1] and [10, Fact 5.10.14],

$$h(\tilde{N}_G) = \frac{1}{2}\ln\left((2\pi\mathrm{e})^N \prod_{i=1}^{N} \mu_i\right) \tag{C.28}$$

and

$$h((Y_{M+1}^N)_G) = \frac{1}{2}\ln\left((2\pi\mathrm{e})^{N-M} \prod_{i=M+1}^{N} \lambda_i\right). \tag{C.29}$$

If $\mathbf{C}_{\tilde{N}}$ denotes the $(M \times M)$-covariance matrix of $\tilde{N}_1^M$ (and, thus, of $(\tilde{N}_1^M)_G$) and $\{\tilde{\mu}_i\}$ the set of eigenvalues of $\mathbf{C}_{\tilde{N}}$, one will obtain

$$L_S(X \to Y_1^M) \leq \frac{1}{2}\ln\left(\frac{\prod_{i=M+1}^{N}\lambda_i \prod_{i=1}^{M}\tilde{\mu}_i}{\prod_{i=1}^{N}\mu_i}\right). \tag{C.30}$$

The proof is completed by providing upper bounds on the eigenvalues in the numerator. It is easy to verify that $\mathbf{C}_{\tilde{N}}$ is the top left principal submatrix of $\mathbf{W}^T\mathbf{C}_N\mathbf{W}$ (which, by the orthogonality of $\mathbf{W}$, has the same eigenvalues as $\mathbf{C}_N$). As a consequence, one can employ Cauchy's interlacing inequality [10, Thm. 8.4.5]:

$$\mu_{i+N-M} \leq \tilde{\mu}_i \leq \mu_i \tag{C.31}$$

The second bound, $\lambda_i \leq \mu_1$, is derived from Weyl's inequality [10, Thm. 8.4.11]

$$\lambda_i \leq \nu_i + \mu_1 \tag{C.32}$$

and by noticing that $\nu_j = 0$ for all $j > M$. Combining this yields an upper bound on the information loss

$$L_S(X \to Y_1^M) \leq \frac{1}{2}\ln\left(\frac{\prod_{i=M+1}^{N}\lambda_i \prod_{i=1}^{M}\tilde{\mu}_i}{\prod_{i=1}^{N}\mu_i}\right) \tag{C.33}$$

$$\leq \frac{1}{2}\ln\left(\frac{\prod_{i=M+1}^{N}\mu_1 \prod_{i=1}^{M}\mu_i}{\prod_{i=1}^{N}\mu_i}\right) \tag{C.34}$$

$$= \frac{1}{2}\ln\left(\prod_{i=M+1}^{N}\frac{\mu_1}{\mu_i}\right). \tag{C.35}$$

□

D Proofs from Chapter 5

D.1 Proof of Lemma 5.2

The proof is provided for jointly Gaussian processes $\mathbf{S}$ and $\mathbf{X}$ only; since the effect of linear filters is independent of the process statistics (cf. [114, p. 663]), the result can be extended to the general case.

First, note that a stable, causal LTI filter satisfies the Paley-Wiener condition (cf. Section 3.5), and that thus $H(\mathrm{e}^{\jmath\theta}) > 0$ a.e. From [114, Cor. to Thm. 9-4, p. 412] one gets $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) = |H(\mathrm{e}^{\jmath\theta})|^2 S_X(\mathrm{e}^{\jmath\theta})$. Since $\mathbf{X}$ has a finite entropy rate, $S_X(\mathrm{e}^{\jmath\theta}) > 0$ a.e., and, thus, $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) > 0$ a.e. That for the cross-power spectral density $S_{\tilde{X}S}(\mathrm{e}^{\jmath\theta}) = H(\mathrm{e}^{\jmath\theta}) S_{XS}(\mathrm{e}^{\jmath\theta})$ holds can be shown easily.

From [116, Thm. 10.2.1, p. 175],

$$\bar{I}(\mathbf{X};\mathbf{S}) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(1 - |\rho_{XS}(\mathrm{e}^{\jmath\theta})|^2\right) d\theta \tag{D.1}$$

where

$$|\rho_{XS}(\mathrm{e}^{\jmath\theta})|^2 = \begin{cases} \frac{|S_{XS}(\mathrm{e}^{\jmath\theta})|^2}{S_X(\mathrm{e}^{\jmath\theta}) S_S(\mathrm{e}^{\jmath\theta})}, & \text{if } S_{XS}(\mathrm{e}^{\jmath\theta}) \neq 0 \\ 0, & \text{else} \end{cases}. \tag{D.2}$$

With above reasoning one gets

$$|\rho_{\tilde{X}S}(\mathrm{e}^{\jmath\theta})|^2 = \begin{cases} \frac{|H(\mathrm{e}^{\jmath\theta})|^2 |S_{XS}(\mathrm{e}^{\jmath\theta})|^2}{|H(\mathrm{e}^{\jmath\theta})|^2 S_X(\mathrm{e}^{\jmath\theta}) S_S(\mathrm{e}^{\jmath\theta})}, & \text{if } H(\mathrm{e}^{\jmath\theta}) S_{XS}(\mathrm{e}^{\jmath\theta}) \neq 0 \\ 0, & \text{else} \end{cases} \tag{D.3}$$

which is a.e. equal to $|\rho_{XS}(\mathrm{e}^{\jmath\theta})|^2$ since $H(\mathrm{e}^{\jmath\theta}) > 0$ a.e. This completes the proof. □

D.2 Proof of Theorem 5.1

The proof requires a lemma similar to Lemma 5.2, but for the specific problem at hand. In particular, let H be the (yet-to-be-designed) anti-aliasing filter and let $\tilde{\mathbf{S}}$ be the Gaussian signal process $\mathbf{S}$ filtered by H. Then,

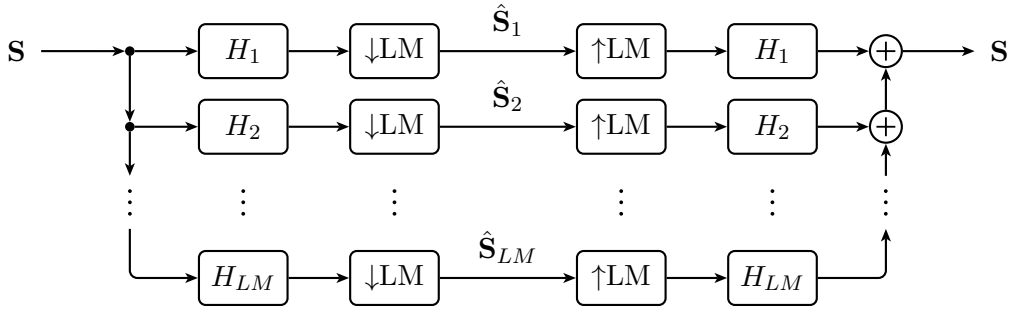


Figure D.1: Filterbank decomposition of the input process $\mathbf{S}$*. The filter responses of* H_i *are as in the proof of Theorem 3.3 (see Appendix B.9).*

Lemma D.1. *If* H *has at most finitely many stop-band intervals and is such that its* M*-fold downsampled frequency response is non-zero a.e.,*

$$\bar{I}(\mathbf{S}^{(M)};\mathbf{Y}) = \bar{I}(\tilde{\mathbf{S}}^{(M)};\mathbf{Y}). \tag{D.4}$$

Proof. For H being such that $|H(\mathrm{e}^{\jmath\theta})| > 0$ a.e., the result follows from Lemma 5.2. The more general case can be shown along the lines of Theorem 3.3: To this end, assume first that H has pass-band intervals with rational endpoints only. One can thus design a filterbank decomposition of $\mathbf{S}$ into LM bands, where L is such that the rational endpoints are integer multiples of $1/LM$ (see Fig. D.1). Since $\mathbf{S}$ is equivalent to $\hat{\mathbf{S}} := \{\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_{LM}\}$, one has[34]

$$\bar{I}(\mathbf{S}^{(M)};\mathbf{Y}) = \frac{1}{L}\bar{I}(\hat{\mathbf{S}};\mathbf{Y}^{(L)}). \tag{D.5}$$

Note that these information rates exist because both $\mathbf{S}$ and the LM-dimensional process $\hat{\mathbf{S}}$ satisfy Assumption 1. Because the sub-band processes are Gaussian and mutually uncorrelated, they are mutually independent. The filter H eliminates some of the sub-band processes, thus makes $\mathbf{Y}$ independent of these. Hence,

$$\bar{I}(\hat{\mathbf{S}};\mathbf{Y}^{(L)}) = \bar{I}(\hat{\mathbf{S}}_{\subseteq};\mathbf{Y}^{(L)}) \tag{D.6}$$

where $\hat{\mathbf{S}}_{\subseteq}^{(M)}$ is the subset of sub-band processes on which $\mathbf{Y}$ depends. Note that the information rate on the right-hand side exists because the filter H is such that $\mathbf{Y}$ has a PSD greater than zero a.e.; differential entropy and differential entropy rate of $\mathbf{Y}$ are finite, and so are those of $\mathbf{Y}^{(L)}$.

Since it is immaterial if the sub-band process is eliminated at the front or the end of the filterbank, one can see that $\hat{\mathbf{S}}_{\subseteq}$ consists exactly of the non-vanishing sub-band processes of the filterbank decomposition of $\tilde{\mathbf{S}}$. Therefore

$$\bar{I}(\tilde{\mathbf{S}}^{(M)};\mathbf{Y}) = \frac{1}{L}\bar{I}(\hat{\mathbf{S}}_{\subseteq};\mathbf{Y}^{(L)}). \tag{D.7}$$

If the pass-band intervals have irrational endpoints, one can always fix L and get upper and lower bounds on the information rates by either keeping or eliminating the according sub-band processes. Since by assumption the PSD are smooth, the bounds become better with increasing L. □

[34] $\mathbf{S}$ runs at M times the rate of $\mathbf{Y}$, thus one needs to block $\mathbf{S}$ in the rate. $\mathbf{Y}$ runs at L times the rate of $\hat{\mathbf{S}}_i$, thus one needs to block $\mathbf{Y}$; $\hat{\mathbf{S}}$ is just a vector process running at the same rate as its elements. Since $\mathbf{Y}^{(L)}$ contains L times as much information as $\mathbf{Y}$ per considered (blocked) sample, one has to divide by this number.

Proof of Theorem 5.1. Instead of minimizing $\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y})$ one can maximize $\bar{I}(\mathbf{S}^{(M)}; \mathbf{Y})$, since this is the only component depending on H.

Assume w.l.o.g. that H is such that its M-fold downsampled frequency response is nonzero a.e.: Were it not so, then one can always split H into a filter H' which is, and a filter H'' following the downsampling device which is piecewise constant; i.e., $H = H' \cdot H''$. Let $\mathbf{Y}$ be the process obtained by using H' as filter, and let $\tilde{\mathbf{Y}}$ be the process obtained by filtering $\mathbf{Y}$ with H''. By data processing (combine [116, Thm. 7.4.2, p. 110] with [116, Thm. 7.2.1, (7), p. 95] and the fact that $\mathbf{S}$ is such that the information rate with a jointly stationary process exists),

$$\bar{I}(\mathbf{S}^{(M)}; \mathbf{Y}) \geq \bar{I}(\mathbf{S}^{(M)}; \tilde{\mathbf{Y}}). \tag{D.8}$$

Thus, it is safe to assume that H is such that its M-fold downsampled frequency response is nonzero a.e. In addition, assume that H has finitely many pass-bands. By Lemma D.1 it follows that it suffices to maximize $\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y})$. Thus,

$$\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n} I(\tilde{S}_1^{nM}; Y_1^n) = \lim_{n\to\infty} \frac{1}{n} \left(h(Y_1^n) - h(Y_1^n | \tilde{S}_1^{nM}) \right) \tag{D.9}$$

The first term converges to $\bar{h}(\mathbf{Y})$. For the latter, note that by linearity, $Y_n = \tilde{X}_{nM} = \tilde{S}_{nM} + \tilde{N}_{nM}$, where $\tilde{\mathbf{N}}$ is the noise process filtered by H. Consequently, and by the fact that $\mathbf{S}$ and $\mathbf{N}$ are independent, $h(Y_1^n | S_1^{nM}) = h(\tilde{N}_M, \ldots, \tilde{N}_{nM})$.

Since $\mathbf{S}$ and $\mathbf{N}$ are Gaussian processes, one obtains

$$\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left(\frac{\sum_{k=0}^{M-1} S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2 + S_S(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2}{\sum_{k=0}^{M-1} S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2} \right) d\theta \tag{D.10}$$

where $\theta_k := \frac{\theta - 2k\pi}{M}$. This can be rearranged to

$$\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left(1 + \frac{\sum_{k=0}^{M-1} S_S(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2}{\sum_{k=0}^{M-1} S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2} \right) d\theta \tag{D.11}$$

$$= \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left(1 + \frac{\sum_{k=0}^{M-1} \frac{S_S(\mathrm{e}^{\jmath\theta_k})}{S_N(\mathrm{e}^{\jmath\theta_k})} S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2}{\sum_{k=0}^{M-1} S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2} \right) d\theta \tag{D.12}$$

$$= \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left(1 + \frac{\sum_{k=0}^{M-1} \frac{S_S(\mathrm{e}^{\jmath\theta_k})}{S_N(\mathrm{e}^{\jmath\theta_k})} \tilde{H}_k(\mathrm{e}^{\jmath\theta})}{\sum_{k=0}^{M-1} \tilde{H}_k(\mathrm{e}^{\jmath\theta})} \right) d\theta \tag{D.13}$$

where $\tilde{H}_k(\mathrm{e}^{\jmath\theta}) := S_N(\mathrm{e}^{\jmath\theta_k}) |H(\mathrm{e}^{\jmath\theta_k})|^2$. Maximizing the integral amounts to maximizing the fraction inside the logarithm for every value of θ. Since the fraction is a weighted average of the ratio $S_S(\mathrm{e}^{\jmath\theta_k}) / S_N(\mathrm{e}^{\jmath\theta_k})$, the weights satisfying, for all $k \in \{0, \ldots, M-1\}$,

$$\tilde{H}_l(\mathrm{e}^{\jmath\theta}) = \begin{cases} 1, & \text{if } l \text{ is the smallest natural number such that } \frac{S_S(\mathrm{e}^{\jmath\theta})}{S_N(\mathrm{e}^{\jmath\theta})}\big|_{\theta=\theta_l} \geq \frac{S_S(\mathrm{e}^{\jmath\theta_k})}{S_N(\mathrm{e}^{\jmath\theta_k})} \\ 0, & \text{else} \end{cases} \tag{D.14}$$

maximize the integral.

H is thus related to the piecewise constant functions $\tilde{H}_l(\mathrm{e}^{\jmath\theta})$ via

$$|H(\mathrm{e}^{\jmath\theta_k})|^2 = \frac{1}{S_N(\mathrm{e}^{\jmath\theta_k})} \tilde{H}_k(\mathrm{e}^{\jmath\theta}) \tag{D.15}$$

where the relation has to be satisfied for all $k = 0, \ldots, M-1$. Since, by assumption, $S_N(\mathrm{e}^{\jmath\theta})$ is such that its square root corresponds to the magnitude response of a causal, stable filter (the

differential entropy rate $\bar{h}(\mathbf{N})$ is finite), and since Lemma 5.2 holds, one can choose H to be piecewise constant. The assumption that the PSDs are smooth guarantees that the resulting filter H has finitely many pass-bands.

That H is identical to the optimal energy compaction filter for $S_S(\mathrm{e}^{\jmath\theta})/S_N(\mathrm{e}^{\jmath\theta})$ is evident from Definition 5.3. □

D.3 Proof of Theorem 5.2

For the proof note that, with Lemma 5.2,

$$\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) = \bar{I}(\mathbf{S}^{(M)}; \tilde{\mathbf{X}}^{(M)}) - \bar{I}(\mathbf{S}^{(M)}; \mathbf{Y}) = \bar{I}(\tilde{\mathbf{S}}^{(M)}; \tilde{\mathbf{X}}^{(M)}) - \bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) \quad \text{(D.16)}$$

where $\tilde{\mathbf{X}}$ is obtained by filtering $\mathbf{X}$ with H. Since $\tilde{X}_n = \tilde{S}_n + \tilde{N}_n$, and since $Y_n = \tilde{X}_{nM}$, one gets

$$\begin{aligned}
\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) &= \lim_{n\to\infty} \frac{1}{n} \left(h(\tilde{X}_1^{nM}) - h(\tilde{X}_1^{nM} | \tilde{S}_1^{nM}) - h(Y_1^n) + h(Y_1^n | \tilde{S}_1^{nM}) \right) && \text{(D.17)} \\
&= \lim_{n\to\infty} \frac{1}{n} \left(h(\tilde{X}_1^{nM}) - h(\tilde{N}_1^{nM}) - h(\tilde{X}_M, \ldots, \tilde{X}_{nM}) + h(\tilde{N}_M, \ldots, \tilde{N}_{nM}) \right) && \text{(D.18)} \\
&= \lim_{n\to\infty} \frac{1}{n} h(\tilde{X}_1^{M-1}, \tilde{X}_{M+1}^{2M-1}, \ldots, \tilde{X}_{(n-1)M+1}^{nM-1} | \tilde{X}_M, \tilde{X}_{2M}, \ldots, \tilde{X}_{nM}) \\
&\quad - \lim_{n\to\infty} \frac{1}{n} h(\tilde{N}_1^{nM}) + \lim_{n\to\infty} \frac{1}{n} h(\tilde{N}_M, \ldots, \tilde{N}_{nM}). && \text{(D.19)}
\end{aligned}$$

The first conditional differential entropy is always upper bounded by the corresponding expression for Gaussian RVs $\tilde{X}_{G,1}, \ldots, \tilde{X}_{G,mM}$ with the same joint first and second order moments as the original RVs (cf. [21, Thm. 8.6.5, p. 254]). Replacing $\mathbf{S}$ by $\mathbf{S}_G$ yields $\mathbf{X}_G$ and $\mathbf{Y}_G$ Gaussian (by Gaussianity of $\mathbf{N}$) and achieves this upper bound with equality. Hence, with $\tilde{X}_{n,G} = \tilde{S}_{n,g} + \tilde{N}_n$ and $Y_{n,G} = \tilde{X}_{nM,G}$,

$$\begin{aligned}
&\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) \\
&\leq \lim_{n\to\infty} \frac{1}{n} h((\tilde{X}_1^{M-1}, \tilde{X}_{M+1}^{2M-1}, \ldots, \tilde{X}_{(n-1)M+1}^{nM-1})_G | (\tilde{X}_M, \tilde{X}_{2M}, \ldots, \tilde{X}_{nM})_G) \\
&\quad - \lim_{n\to\infty} \frac{1}{n} h(\tilde{N}_1^{nM}) + \lim_{n\to\infty} \frac{1}{n} h(\tilde{N}_M, \ldots, \tilde{N}_{nM}) && \text{(D.20)} \\
&= \lim_{n\to\infty} \frac{1}{n} \left(h((\tilde{X}_1^{nM})_G) - h((\tilde{X}_1^{nM})_G | (\tilde{S}_1^{nM})_G) - h((Y_1^n)_G) + h((Y_1^n)_G | (\tilde{S}_1^{nM})_G) \right) && \text{(D.21)} \\
&= \bar{I}(\tilde{\mathbf{S}}_G^{(M)}; \tilde{\mathbf{X}}_G^{(M)}) - \bar{I}(\tilde{\mathbf{S}}_G^{(M)}; \mathbf{Y}_G) && \text{(D.22)} \\
&= \bar{L}_{\mathbf{S}_G^{(M)}}(\mathbf{X}_G^{(M)} \to \mathbf{Y}_G). && \text{(D.23)}
\end{aligned}$$

□

D.4 Proof of Lemma 5.3

The mutual information rate is

$$\begin{aligned}
\bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}) &\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n} I(\tilde{X}_1^n; Y_1^n) && \text{(D.24)} \\
&= \lim_{n\to\infty} \frac{1}{n} \left(h(\tilde{X}_1^n) - h(\tilde{X}_1^n | Y_1^n) \right) && \text{(D.25)} \\
&\overset{(b)}{=} \bar{h}(\tilde{\mathbf{X}}) - \lim_{n\to\infty} \frac{1}{n} h(\tilde{X}_1^n | Y_1^n) && \text{(D.26)}
\end{aligned}$$

$$\stackrel{(c)}{=} \frac{1}{2}\ln(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})d\theta - \lim_{n\to\infty}\frac{1}{n}h(\tilde{X}_1^n|Y_1^n) \tag{D.27}$$

$$\stackrel{(d)}{=} \frac{1}{2}\ln(2\pi e\sigma_\infty^2) - \lim_{n\to\infty}\frac{1}{n}h(\tilde{X}_1^n|Y_1^n) \tag{D.28}$$

$$\stackrel{(e)}{\geq} \frac{1}{2}\ln(2\pi e\sigma_\infty^2) - h(\tilde{X}_1|Y_1) \tag{D.29}$$

$$\stackrel{(f)}{\geq} \frac{1}{2}\ln(2\pi e\sigma_\infty^2) - \ln\Delta \tag{D.30}$$

$$= \frac{1}{2}\ln(2\pi e\sigma_\infty^2/\Delta^2) \tag{D.31}$$

where (a) and (b) are the definitions of information rate and differential entropy rate, respectively, (c) is the differential entropy rate of a Gaussian process [114, p. 663], and (d) is the definition of entropy power $\sigma_i^2 nfty$ [21, Ch. 12.5]. Inequality (e) follows from the chain rule of differential entropy, conditioning, and stationarity:

$$h(\tilde{X}_1^n|Y_1^n) = \sum_{i=1}^{n} h(\tilde{X}_i|\tilde{X}_1^{i-1}, Y_1^n) \leq \sum_{i=1}^{n} h(\tilde{X}_i|Y_i) = nh(\tilde{X}_1|Y_1) \tag{D.32}$$

Finally, the differential entropy of a random variable supported on a finite interval is bounded by the entropy of the uniform distribution over this interval. Therefore, since $\tilde{X}_1|Y_1 = y$ is supported on an interval of length Δ for all y, it follows that $h(\tilde{X}_1|Y_1) \leq \ln\Delta$ in (f).

Next, consider the case of high-rate quantization, where the quantization noise is modeled as uncorrelated noise with variance $\Delta^2/12$. Assuming that $\tilde{\mathbf{X}}$ and $\mathbf{Y}_G$ are jointly Gaussian with the same joint first and second moments as $\tilde{\mathbf{X}}$ and $\mathbf{Y}$, one gets

$$\bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}_G) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})}{\Delta^2/12}\right) d\theta. \tag{D.33}$$

For high resolutions (i.e., for sufficiently small Δ) and if $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})$ is bounded away from zero, the one in the argument becomes negligible, and one has

$$\bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}_G) \approx \frac{1}{2}\ln\left(12\sigma_\infty^2/\Delta^2\right). \tag{D.34}$$

Subtracting this from (D.31) yields

$$\bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}) - \bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}_G) \gtrapprox \frac{1}{2}\ln\left(\frac{2\pi e}{12}\right) > 0. \tag{D.35}$$

□

Bibliography

[1] (2012, Oct.) Principle component analysis. [Online]. Available: http://en.wikipedia.org/wiki/Principle_components_analysis

[2] A. M. Abdel-Moneim and F. W. Leysieffer, "Weak lumpability in finite Markov chains," *J. Appl. Prob.*, vol. 19, no. 3, pp. 685–691, Sep. 1982.

[3] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. New York, NY: Dover Publications, 1972.

[4] S. Akkarakaran and P. P. Vaidyanathan, "Results on principal component filter banks: colored noise suppression and existence issues," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1003–1020, 2001.

[5] R. W. Aldhaheri and H. K. Khalil, "Aggregation of the policy iteration method for nearly completely decomposable Markov chains," *IEEE Trans. Autom. Control*, vol. 36, no. 2, pp. 178–187, Feb. 1991.

[6] M. Baer, "A simple countable infinite-entropy distribution," 2008. [Online]. Available: https://hkn.eecs.berkeley.edu/~calbear/research/Hinf.pdf

[7] J. C. Baez, T. Fritz, and T. Leinster, "A characterization of entropy in terms of information loss," *Entropy*, vol. 13, no. 11, pp. 1945–1957, Nov. 2011.

[8] H.-P. Bernhard, "The mutual information function and its application to signal processing," Ph.D. dissertation, Technische Universität Wien, Dec. 1997.

[9] ——, "Tight upper bound on the gain of linear and nonlinear predictors," *IEEE Trans. Signal Process.*, vol. 46, no. 11, pp. 2909–2917, Nov. 1998.

[10] D. S. Bernstein, *Matrix Mathematics*. Princeton, NJ: Princeton University Press, 2005.

[11] J. J. Birch, "Approximation for the entropy for functions of Markov chains," *Ann. Math. Statist.*, vol. 33, pp. 930–938, 1962.

[12] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Transactions of the first Prague conference on information theory, Statistical decision functions, random processes held at Liblice near Prague from November 28 to 30, 1956*. Prague: Publishing House of the Czechoslovak Academy of Sciences, 1957, pp. 13–20.

[13] R. A. Brualdi, *Introductory Combinatorics*, 5th ed. Upper Saddle River, NJ: Pearson Education, 2010.

[14] P. Buchholz, "Exact and ordinary lumpability in finite Markov chains," *J. Appl. Prob.*, vol. 31, pp. 59–75, 1994.

[15] C. J. Burke and M. Rosenblatt, "A Markovian function of a Markov chain," *Ann. Math. Statist.*, vol. 29, pp. 1112–1122, 1958.

[16] J. W. Carlyle, "Identification of state-calculable functions of finite Markov chains," *Ann. Math. Statist.*, vol. 38, pp. 201–205, 1967.

[17] G. Chechik and N. Tishby, "Extracting relevant structures with side information," in *Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 857–864.

[18] Y. Chen, Y. C. Eldar, and A. J. Goldsmith, "Shannon meets Nyquist: Capacity of sampled Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4889–4914, Aug. 2013.

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

[20] J. T. Chu and J. C. Chueh, "Inequalities between information measures and error probability," *J. Franklin Inst.*, vol. 282, pp. 121–125, Aug. 1966.

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

[22] C. D. Cutler, "Computing pointwise fractal dimension by conditioning in multivariate distributions and time series," *Bernoulli*, vol. 6, no. 3, pp. 381–399, Jun. 2000.

[23] D. M. Cvetković, P. Rowlinson, and S. Simić, *Eigenspaces of graphs*, ser. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press, 1997.

[24] X.-H. Dang and J. Bailey, "A hierarchical information theoretic technique for the discovery of non linear alternative clusterings," in *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, Washington, D.C., Jul. 2010, pp. 573–582.

[25] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*. New York, NY: Springer, 1996.

[26] K. Deng, "Model reduction of Markov chains with applications to building systems," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2012.

[27] K. Deng, P. G. Mehta, and S. P. Meyn, "Optimal Kullback-Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2793–2808, Dec. 2011.

[28] S. W. Dharmadhikari, "Sufficient conditions for a stationary process to be a function of a finite Markov chain," *Ann. Math. Stat.*, vol. 34, pp. 1033–1041, 1963.

[29] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing: System Analysis and Design*, 2nd ed. Cambridge: Cambridge University Press, 2010.

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

[31] T. Downarowicz, *Entropy in Dynamical Systems*. Cambridge: Cambridge University Press, 2011.

[32] 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.

[33] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1569, 2002, special issue on Shannon theory: perspective, trends, and applications. [Online]. Available: http://dx.doi.org/10.1109/TIT.2002.1003838

[34] M. C. Er, "A fast algorithm for generating set partitions," *The Computer Journal*, vol. 31, no. 3, pp. 283–284, 1988.

[35] D. Erdogmus, R. Agrawal, and J. C. Principe, "A mutual information extension to the matched filter," *Signal Processing*, vol. 85, no. 5, pp. 927–935, May 2005.

[36] D. Erdogmus and J. C. Principe, "From linear adaptive filtering to nonlinear information processing," *IEEE Signal Process. Mag.*, vol. 23, no. 6, pp. 14–33, Nov. 2006.

[37] W. S. Evans, "Information theory and noisy computation," Ph.D. dissertation, Univ. Calif. Berkeley, Nov. 1994.

[38] W. S. Evans and L. J. Schulman, "Signal propagation and noisy circuits," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2367–2373, Nov. 1999.

[39] J. D. Farmer, E. Ott, and J. A. Yorke, "The dimension of chaotic attractors," *Physica D*, vol. 7, pp. 153–180, May 1983.

[40] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.

[41] J. Feret, T. Henzinger, H. Koeppl, and T. Petrov, "Lumpability abstractions of rule-based systems," *Electronic Proceedings in Theoretical Computer Science*, vol. 40, pp. 142–161, Aug. 2010.

[42] G. B. Folland, *Real Analysis. Modern Techniques and Their Applications*, 2nd ed. New York, NY: Wiley Interscience, 1999.

[43] G. D. Forney, Jr., "Shannon meets Wiener: On MMSE estimation in successive decoding schemes," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, Sep. 2004, pp. 923–932.

[44] A. Friedman and J. Goldberger, "Information theoretic pairwise clustering," in *SIMBAD*, ser. LNCS, E. Hancock and M. Pelillo, Eds. Berlin: Springer, 2013, vol. 7953, pp. 106–119.

[45] J. Galambos, *Advanced Probability Theory*, 2nd ed., ser. Probability: Pure and Applied. New York, NY: Marcel Dekker, Inc., 1995.

[46] B. C. Geiger and G. Kubin, "Information measures for deterministic input-output systems," Mar. 2013, in preparation; preprint available: `arXiv:1303.6409 [cs.IT]`.

[47] B. C. Geiger, T. Petrov, G. Kubin, and H. Koeppl, "Optimal Kullback-Leibler aggregation via information bottleneck," Apr. 2013, submitted to IEEE Trans. Autom. Control; preprint available: `arXiv:1304.6603 [cs.SY]`.

[48] B. C. Geiger and C. Temmel, "Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability," Dec. 2012, accepted in J. Appl. Prob.; preprint available: `arXiv:1212.4375 [cs.IT]`.

[49] 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, pp. 799–803, extended version available: `arXiv:1102.4794 [cs.IT]`.

[50] B. C. Geiger and G. Kubin, "Some results on the information loss in dynamical systems," in *Proc. IEEE Int. Sym. Wireless Communication Systems (ISWSC)*, Aachen, Nov. 2011, pp. 794–798, extended version available: `arXiv:1106.2404 [cs.IT]`.

[51] ——, "On the information loss in memoryless systems: The multivariate case," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Zurich, Feb. 2012, pp. 32–35, extended version available: `arXiv:1109.4856 [cs.IT]`.

[52] ——, "Relative information loss in the PCA," in *Proc. IEEE Information Theory Workshop (ITW)*, Lausanne, Sep. 2012, pp. 562–566, extended version available: `arXiv:1204.0429 [cs.IT]`.

[53] ——, "Information loss and anti-aliasing filters in multirate systems," Oct. 2013, in preparation; preprint available: `arXiv:1310.8487 [cs.IT]`.

[54] ——, "On the rate of information loss in memoryless systems," Apr. 2013, `arXiv:1304.5057 [cs.IT]`.

[55] ——, "Signal enhancement as minimization of relevant information loss," in *Proc. ITG Conf. on Systems, Communication and Coding*, Munich, Jan. 2013, pp. 1–6, extended version available: `arXiv:1205.6935 [cs.IT]`.

[56] ——, "Information loss and anti-aliasing filters in multirate systems," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Zürich, Feb. 2014, pp. 148–151.

[57] ——, "Information-maximizing prefilters for quantization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, May 2014, pp. 5001–5005.

[58] B. C. Geiger and C. Temmel, "Information-preserving Markov aggregation," in *Proc. IEEE Information Theory Workshop (ITW)*, Seville, Sep. 2013, pp. 258–262, extended version: `arXiv:1304.0920 [cs.IT]`.

[59] E. J. Gilbert, "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Stat.*, vol. 30, pp. 688–697, 1959.

[60] D. T. Gillespie, "Stochastic Simulation of Chemical Kinetics," *Annual Review of Physical Chemistry*, vol. 58, no. 1, pp. 35–55, 2007.

[61] D. Gondek and T. Hofmann, "Conditional information bottleneck clustering," in *IEEE Int. Conf. on Data Mining (ICDM), Workshop on Clustering Large Data Sets*, Melbourne, FL, Nov. 2003, pp. 36–42.

[62] ——, "Non-redundant data clustering," in *Proc. IEEE Int. Conf. on Data Mining (ICDM)*, Nov. 2004, pp. 75 – 82.

[63] V. K. Goyal, J. Zhuang, and M. Vetterli, "Transform coding with backward adaptive updates," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1623–1633, Jul. 2000.

[64] P. Grassberger, "Generalized dimensions of strange attractors," *Physics Letters*, vol. 97A, no. 6, pp. 227–230, Sep. 1983.

[65] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[66] R. M. Gray, *Entropy and Information Theory*. New York, NY: Springer, 1990.

[67] D. Guo, S. S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[68] L. Gurvits and J. Ledoux, "Markov property for a function of a Markov chain: a linear algebra approach," *Linear Algebra Appl.*, vol. 404, pp. 85–117, 2005. [Online]. Available: http://dx.doi.org/10.1016/j.laa.2005.02.007

[69] B. Hayes, "First links in the Markov chain," *America Scientist*, vol. 101, pp. 92–97, 2013.

[70] S. Haykin, *Adaptive Filter Theory*, 5th ed. Upper Saddle River, NJ: Pearson, 2014.

[71] A. Heller, "On stochastic processes derived from Markov chains," *Ann. Math. Stat.*, vol. 36, no. 4, pp. 1286–1291, Aug. 1965.

[72] G. E. Henter, "Probabilistic sequence models with speech and language applications," Ph.D. dissertation, KTH Royal Institute of Technology, Dec. 2013.

[73] G. E. Henter and W. B. Kleijn, "Minimum entropy rate simplification of stochastic processes," submitted to *IEEE Trans. Pattern Anal.*

[74] S.-W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5930–5942, Dec. 2010.

[75] S.-W. Ho and R. Yeung, "On the discontinuity of the Shannon information measures," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5362 –5374, Dec. 2009.

[76] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 2013.

[77] M. Hotz and C. Vogel, "Linearization of time-varying nonlinear systems using a modified linear iterative method," Nov. 2013, submitted to IEEE Trans. Sig. Proc.

[78] E. T. Jaynes, "Information theory and statistical mechanics," in *Statistical Physics*, ser. Brandeis University Summer Institute Lectures in Theoretical Physics, K. W. Ford, Ed. New York, NY: W. A. Benjamin, Inc., 1963, vol. 3, pp. 181–218.

[79] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[80] Q.-S. Jia, "On state aggregation to approximate complex value functions in large-scale Markov decision processes," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 333–344, Feb. 2011.

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

[82] J. Jost, *Dynamical Systems: Examples of Complex Behavior*. New York, NY: Springer, 2005.

[83] M. Kafsi, M. Grossglauser, and P. Thiran, "The entropy of conditional Markov trajectories," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5577–5583, Sep. 2013.

[84] A. Kaiser and T. Schreiber, "Information transfer in continuous processes," *Physica D*, vol. 166, pp. 43–62, Jun. 2002.

[85] E. Karapistoli, F.-N. Pavlidou, I. Gragopoulos, and I. Tsetsinas, "An overview of the IEEE 802.15.4a standard," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 47–53, Jan. 2010.

[86] M. A. Katsoulakis and J. Trashorras, "Information loss in coarse-graining of stochastic particle dynamics," *J. Stat. Phys.*, vol. 122, no. 1, pp. 115–135, 2006.

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

[88] K. T. Kim and T. Berger, "Sending a lossy version of the innovations process is suboptimal in QG rate-distortion," in *Proc. IEEE Int. Sym. on Information Theory (ISIT)*, 2005, pp. 209–213.

[89] ——, "The degree of suboptimality of sending a lossy version of the innovations process in Gauss-Markov rate-distortion," in *Proc. IEEE Int. Sym. on Information Theory (ISIT)*, 2006, pp. 808–812.

[90] A. Kirac and P. Vaidyanathan, "Theory and design of optimum FIR compaction filters," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 903–919, Apr. 1998.

[91] 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.

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

[93] R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent, "How to build the best macroscopic description of your multi-agent system?" Laboratoire d'Informatique de Grenoble, Tech. Rep., Jan. 2013. [Online]. Available: http://rr.liglab.fr/research_report/RR-LIG-035_orig.pdf

[94] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

[95] J. M. Lee, *Introduction to Smooth Manifolds*, ser. Graduate Texts in Mathematics. New York, NY: Springer, 2003.

[96] B. Lev, "The aggregation problem in financial statements: An informational approach," *Journal of Accounting Research*, vol. 6, no. 2, pp. 247–261, Autumn 1968.

[97] X. S. Liang and R. Kleeman, "Information transfer between dynamical system components," *Physical Review Letters*, vol. 95, pp. 244 101–1–244 101–4, Dec. 2005.

[98] B. Lindqvist, "How fast does a Markov chain forget the initial state? A decision theoretical approach," *Scandinavian Journal of Statistics*, vol. 4, no. 4, pp. pp. 145–152, 1977. [Online]. Available: http://www.jstor.org/stable/4615670

[99] ——, "On the loss of information incurred by lumping states of a Markov chain," *Scandinavian Journal of Statistics*, vol. 5, no. 2, pp. 92–98, 1978. [Online]. Available: http://www.jstor.org/stable/4615693

[100] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, Mar. 1988.

[101] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[102] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, 2nd ed. Cambridge, MA: MIT Press, 2000.

[103] K. Matsumoto and I. Tsuda, "Calculation of information flow rate from mutual information," *J. Phys. A*, vol. 21, pp. 1405–1414, 1988.

[104] M. Meilă and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, Nov. 2000, pp. 1–7.

[105] P. Meissner and K. Witrisal, "Analysis of a noncoherent UWB receiver for multichannel signals," in *Proc. IEEE Vehicular Technology Conf. (VTC-Spring)*, Taipei, May 2010, pp. 1–5.

[106] R. Moddemeijer. (2010, June) Matlab library. [Online]. Available: http://www.cs.rug.nl/~rudy/matlab/

[107] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, ser. Wiley Series in Probability and Mathematical Statistics. Hoboken, NJ: Wiley Interscience, 1982.

[108] P. Noll, "On predictive quantization schemes," *Bell System Technical Journal*, vol. 57, no. 5, pp. 1499–1532, May 1978.

[109] J. R. Norris, *Markov chains*. Cambridge university press, 1998, no. 2008.

[110] R. D. Nowak and B. D. V. Veen, "Volterra filter equalization: A fixed point approach," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 377–387, Feb. 1997.

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

[112] A. Papoulis, *The Fourier Integral and its Applications*. McGraw Hill, 1962.

[113] ——, "Maximum entropy and spectral estimation: A review," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1176–1186, Dec. 1981.

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

[115] A. Pedross and K. Witrisal, "Analysis of nonideal multipliers for multichannel autocorrelation UWB receivers," in *Proc. IEEE Int. Conf. on Ultra-Wideband (ICUWB)*, Sep. 2012, pp. 140–144.

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

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

[118] M. Plumbley, "Information theory and unsupervised neural networks," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR. 78, 1991.

[119] S. P. Ponomarev, "Submersions and preimages of sets of measure zero," *Siberian Mathematical Journal*, vol. 28, no. 1, pp. 153–163, Jan. 1987.

[120] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, ser. Information Science and Statistics. New York, NY: Springer, 2010.

[121] Z. Rached, F. Alajaji, and L. L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 917–921, May 2004.

[122] A. Raj and C. H. Wiggins, "An information-theoretic derivation of min-cut-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 988–995, Jun. 2010.

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

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

[125] M. Rosenblatt, "Functions of a Markov process that are Markovian," in *Selected Works of Murray Rosenblatt*, ser. Selected Works in Probability and Statistics, R. A. Davis, K.-S. Lii, and D. N. Politis, Eds. Springer, 2011, pp. 134–146.

[126] G. Rubino and B. Sericola, "On weak lumpability in Markov chains," *J. Appl. Prob.*, vol. 26, no. 3, pp. 446–457, Sep. 1989.

[127] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., ser. Int. Series in Pure and Applied Mathematics. New York, NY: McGraw Hill, 1976.

[128] ——, *Real and Complex Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1987.

[129] D. Ruelle, "Positivity of entropy production in nonequilibrium statistical mechanics," *J. Stat. Phys.*, vol. 85, pp. 1–23, 1996.

[130] T. Runolfsson and Y. Ma, "Model reduction of nonreversible Markov chains," in *Proc. IEEE Conf. on Decision and Control (CDC)*, New Orleans, LA, Dec. 2007, pp. 3739–3744.

[131] M. A. Sánchez-Montañés, "A theory of information processing for adaptive systems: Inspiration from biology, formal analysis and application to artificial systems," Ph.D. dissertation, Universidad Autónoma de Madrid, Jun. 2003.

[132] M. A. Sánchez-Montañés and F. J. Corbacho, "A new information processing measure for adaptive complex systems," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 917–927, Jul. 2004.

[133] A. Scaglione, S. Barbarossa, and G. B. Giannakis, "Filterbank transceivers optimizing information rate in block transmissions over dispersive channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 1019–1032, Apr. 1999.

[134] M. Schetzen, *Linear Time-Invariant Systems*. Piscataway, NJ: IEEE Press, 2003.

[135] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, Jul. 2000.

[136] E. Schrödinger, *What is Life? with Mind and Matter and Autobiographical Sketches*. Cambridge University Press, 1992.

[137] I. Semba, "An efficient algorithm for generating all partitions of the set $\{1, 2, \ldots, n\}$," *Journal of information processing*, vol. 7, no. 1, pp. 41–42, Mar. 1984.

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

[139] Y. S. Shmaliy, *Continuous-Time Systems*. Dordrecht: Springer, 2007.

[140] D. Simmons, "Conditional measures and conditional expectation; Rohlin's disintegration theorem," *Discrete and Continuous Dynamical Systems - Series A*, vol. 32, no. 7, pp. 2565–2582, Jul. 2012.

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

[142] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," vol. 57, no. 12, pp. 3629–3639, Dec. 2009.

[143] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 1999, pp. 617–623.

[144] 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.

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

[146] N. Tishby and N. Slonim, "Data clustering by Markovian relaxation and the information bottleneck method," in *Advances in Neural Information Processing Systems (NIPS)*, 2001. [Online]. Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.3488

[147] M. Tsatsanis and G. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1766–1777, Aug. 1995.

[148] J. Tuqan and P. Vaidyanathan, "A state space approach to the design of globally optimal FIR energy compaction filters," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2822–2838, 2000.

[149] M. Unser, "On the optimality of ideal filters for pyramid and wavelet signal approximation," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3591–3596, Dec. 1993.

[150] P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1528–1543, 1998.

[151] P. Vaidyanathan, Y.-P. Lin, S. Akkarakaran, and S.-M. Phoong, "Discrete multitone modulation with principal component filter banks," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 49, no. 10, pp. 1397–1412, 2002.

[152] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. Chichester: John Wiley & Sons, 2006.

[153] J. A. Vastano and H. L. Swinney, "Information transport in spatiotemporal systems," *Physical Review Letters*, vol. 60, no. 18, pp. 1773–1776, May 1988.

[154] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[155] S. Verdú, "Information measures and estimation theory," Plenary Talk at IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), May 2013.

[156] A. C. Verdugo Lazo and P. N. Rathie, "On the entropy of continuous probability distributions," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 1, pp. 120–122, Jan. 1978.

[157] M. Vidyasagar, "Kullback-Leibler divergence rate between probability distributions on sets of different cardinalities," in *Proc. of the IEEE Conf. on Decision and Control*, Atlanta, GA, Dec. 2010, pp. 948–953.

[158] ——, "Reduced-order modeling of Markov and hidden Markov processes via aggregation," in *Proc. IEEE Conf. on Decision and Control (CDC)*, Atlanta, GA, Dec. 2010, pp. 1810–1815.

[159] ——, "A metric between probability distributions on finite sets of different cardinalities and applications to order reduction," *IEEE Trans. Autom. Control*, vol. 57, no. 10, pp. 2464–2477, Oct. 2012.

[160] G. J. Wallinger, "Information loss and chaotic systems," Master's Thesis, Graz University of Technology, Nov. 2013.

[161] 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.

[162] E. Weinan, L. Tiejun, and E. Vanden-Eijnden, "Optimal partition and effective dynamics of complex networks," *PNAS*, vol. 105, no. 23, pp. 7907–7912, Jun. 2008.

[163] L. B. White, R. Mahony, and G. D. Brushe, "Lumpable hidden Markov models–model reduction and reduced complexity filtering," *IEEE Trans. Autom. Control*, vol. 45, no. 12, pp. 2297–2306, Dec. 2000.

[164] W. Wiegerinck and H. Tennekes, "On the information flow for one-dimensional maps," *Physics Letters A*, vol. 144, no. 3, pp. 145–152, Feb. 1990.

[165] D. Wilkinson, *Stochastic Modelling for Systems Biology*, ser. Chapman & Hall/CRC Mathematical & Computational Biology. Boca Raton, FL: Taylor & Francis, 2006.

[166] K. Witrisal, "Noncoherent autocorrelation detection of orthogonal multicarrier UWB signals," in *IEEE Int. Conf. on Ultra-Wideband (ICUWB)*, Hannover, Sep. 2008, pp. 161–164.

[167] W. Woess, *Denumerable Markov chains*, ser. EMS Textbooks in Mathematics. European Mathematical Society (EMS), Zürich, 2009, generating functions, boundary theory, random walks on trees. [Online]. Available: http://dx.doi.org/10.4171/071

[168] Y. Wu, "Shannon theory for compressed sensing," Ph.D. dissertation, Princeton University, 2011.

[169] 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.

[170] ——, "MMSE dimension," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4857–4879, Aug. 2011.

[171] ——, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6241–6263, Oct. 2012.

[172] J. Yeh, *Lectures on real analysis*. Singapore: World Scientific Publishing, 2000.

[173] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, 2008.

[174] G. Zeitler, A. C. Singer, and G. Kramer, "Low-precision A/D conversion for maximum information rate in channels with memory," vol. 60, no. 9, pp. 2511–2521, Sep. 2012.

[175] W. H. Zurek, "Algorithmic information content, Church-Turing thesis, physical entropy, and Maxwell's demon," in *Proc. Information Dynamics*, H. Atmanspacher and H. Scheingraber, Eds., Irsee, 1991, pp. 245–259.