Information Loss and Anti-Aliasing Filters in Multirate Systems

Bernhard C. Geiger, Student Member, IEEE, and Gernot Kubin, Member, IEEE

Abstract—This work investigates the information loss in a decimation system, i.e., in a downsampler preceded by an anti-aliasing filter. It is shown that, without a specific signal model in mind, the anti-aliasing filter cannot reduce information loss, while, e.g., for a simple signal-plus-noise model it can. For the Gaussian case, the optimal anti-aliasing filter is shown to coincide with the one obtained from energetic considerations. For a non-Gaussian signal corrupted by Gaussian noise, the Gaussian assumption yields an upper bound on the information loss, justifying filter design principles based on second-order statistics from an information-theoretic point-of-view.

I. INTRODUCTION

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 [1, Ch. 4.6]. In these books, this anti-aliasing filter is usually an ideal low-pass filter with a cut-off frequency of $\pi/M$, for an $M$-fold decimation system (cf. Fig. 1). Unser [2] showed that this choice is optimal in terms of the mean-squared reconstruction error (MSE) only if the input process is such that the passband portion of its power spectral density (PSD) exceeds all aliased components. Similarly, as it was shown by Tsatsanis and Giannakis [3], the filter minimizing the MSE is piecewise constant, $M$-aliasing-free (i.e., the aliased components of the $M$-fold downsampled frequency response do not overlap), and has a passband 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 this paper we consider a design objective vastly different from the MSE: information. The fact that information, compared to energy, can yield more successful system designs has long been recognized, e.g., for (non-linear) adaptive filters [4] or for state estimation using linear filters [5]. Mutual information has been used as a design objective for transceiver filter design, too: In [6], Al-Dhahir et al. derived a sub-optimal block transmission filter whose output approximates the optimal input statistics of a dispersive, noisy channel. Scaglione et al. [7] later showed that the resulting non-stationarity of the channel input can be achieved by an FIR filter bank, both for FIR and ARMA channels. Recently, Chen et al. [8] derived the capacity of sub-Nyquist sampled additive Gaussian noise channels for various sampling mechanisms: sampling after a filter, a filterbank, and a modulated filterbank. They showed that the capacity-maximizing sampling filter is piecewise constant and both maximizes the signal-to-noise ratio and minimizes the MSE of the reconstructed signal, thus building a bridge between information-theoretic and energetic filter design. All these works, however, consider either a signal-plus-noise model or assume that all processes are Gaussian.

We extend the existing literature in three different aspects: First, we present results for the case where no signal model is available, other than the PSD of the Gaussian input process (Section III). Second, we derive the optimal filter for a signal model in which the Gaussian filter input is correlated with a Gaussian signal process representing relevant information (Section IV). And finally, we consider a signal-plus-Gaussian-noise scenario, where we assume that the signal is non-Gaussian (Section V).

Our first result is surprising: Given mild assumptions on the input process of the decimation system, the information loss can be bounded independently of the anti-aliasing filter (see Section III). The reason is that, without a specific signal model, every bit of the input process is treated equivalently, regardless of the amount of energy by which it is represented. In order to remedy this counter-intuitivity, Section IV considers Gaussian processes with a specific signal model in mind: The input to the decimation system is correlated with a relevant data signal. A data signal corrupted by Gaussian noise is a special case of this scenario, thus connecting to the analysis of additive Gaussian noise channels in [8]. The optimal filter is shown to be piecewise constant and conceptually similar to those derived in [3, 8]. Since in most cases the Gaussian assumption is too restrictive, in Section V we let the decimator input be an arbitrarily distributed data signal corrupted by Gaussian noise. Following the approach of Plumbley in [9], we prove that the Gaussian assumption for the signal process yields an upper bound on the information loss in the general case. In other
words, designing a filter based on the PSDs of the signal and noise processes guarantees a bounded information loss in the decimation system. This justifies filter design based on second-order statistics, i.e., on energetic considerations, also from an information-theoretic perspective. In Section VII we illustrate our results in a simple toy example. The problem of designing optimal FIR filters for the decimation system, as it appears in the toy example, is briefly discussed in Section VII which also contains an outlook to future work.

II. PRELIMINARIES AND NOTATION

Throughout this work we adopt the following notation: Z is a real-valued random process, whose n-th sample is the random variable (RV) Z_n. Let Z_j := {Z_i : i ∈ J} and we abbreviate Z^n_j := {Z_i, Z_{i+1}, . . . , Z_j}. The differential entropy [10, Ch. 8] and the Rényi information dimension [11] of Z_j are h(Z_j) and d(Z_j), respectively, provided these quantities exist and are finite. Finally, we define the M-fold blocking Z(M) of Z as the sequence of M-dimensional RVs Z^n_1 := Z^n_1, Z^n_2 := Z^n_1∩Z^n_2+1, and so on. Since the systems used in this work are static, i.e., described by a function, we abuse notation and understand g(Z_j) and g(Z) as the function applied coordinate-wise or sample-wise, respectively.

In this work, we often consider a process Z satisfying

Assumption 1. Z is stationary, has finite variance, finite marginal differential entropy h(Z^n), and finite differential entropy rate

h(Z) := \lim_{n→∞} \frac{1}{n} h(Z^n) = \lim_{n→∞} h(Z^n) = \lim_{n→∞} h(Z^n_1 − 1). \tag{1}

Lemma 1 (Finite differential entropy (rate) and information dimension). Let Z be a stationary stochastic process satisfying Assumption 1. Then, for every finite set J, d(Z_j) = card(J).

Proof: See Appendix D-A.

As another direct consequence of Assumption 1, the mutual information rate with a process W jointly stationary with Z exists and equals [12] Thm. 8.3

I(Z; W) := \lim_{n→∞} \frac{1}{n} I(Z^n_1; W^n). \tag{2}

To measure the rate of information loss in a deterministic system, we introduce

Definition 1 (Relative Information Loss Rate). The relative information loss rate induced by the function g is

\hat{l}(Z → g(Z)) := \lim_{n→∞} \frac{d(Z^n_1 | g(Z^n_1))}{d(Z^n_1)} \tag{3}

provided the quantity on the right exists.

This definition is an extension of the relative information loss l(Z → g(Z)), as defined in [13], to stochastic processes. Roughly speaking, l(Z → g(Z)) captures the percentage of information lost by applying the function g to the RV Z. That the relative information loss is related to the (conditional) information dimension was also observed in [13], where the second equality in (3) was proved.

Clearly, in an invertible system no information is lost. One drawback of Definition 1 is that it is only defined for static systems. Hence, it will be necessary to abuse notation by presenting

Assumption 2. Let Z be the input process and \hat{Z} the output process of a static system. If \hat{Z} is another process which is equivalent to Z in the sense that there exists a (not necessarily static) invertible system which converts one to the other, then

l(\hat{Z} → Z) = l(\hat{Z} → Z). \tag{4a}

Likewise, if \hat{Z} is equivalent to Z, then

l(Z → \hat{Z}) = l(Z → Z). \tag{4b}

In particular, a polyphase decomposition or a perfect reconstruction filterbank decomposition of a process is equivalent to the original process in above sense.

In some cases not the total information lost in the system is of interest, but only the portion W which is relevant to the user. Hence, in [14], the notion of relevant information loss was introduced as the difference between mutual informations:

L_W(Z → g(Z)) := I(W; Z) − I(W; g(Z)) \tag{5}

This notion is extended to stochastic processes in

Definition 2 (Relevant Information Loss Rate). Let W be a process jointly stationary with Z, representing the relevant information content of Z. Then, the information loss rate relevant w.r.t. W induced by processing Z to a jointly stationary process \hat{Z} is

I_W(Z → \hat{Z}) := I(W; Z) − I(\hat{W}; \hat{Z}) \tag{6}

provided the quantities exist.

As a specific example, W might be the sign of Z, or W might be a noisy observation of W. Note further that in Definition 2 it is not necessary to assume that the system is static; if it was, then Z = g(Z), as in Definition 1.

Considering the scenario depicted in Fig. 1 we will be concerned with linear filters and their effect on the information-carrying processes. If the filter is stable and causal, its magnitude response satisfies the Paley-Wiener condition (cf. [15, p. 215] and [16, p. 423]):

1 \over 2\pi \int_{-\pi}^{\pi} \ln |H(e^{i\theta})|d\theta > -\infty. \tag{7}

For such filters, the following two lemmas can be presented:

Lemma 2. Let Z be a stochastic process satisfying Assumption 1 and let H be a stable, causal linear filter with input Z. Then, the output process \hat{Z} of the filter satisfies Assumption 1.

Proof: See Appendix D-B.

Lemma 3. Let W and Z be two jointly stationary stochastic processes satisfying Assumption 1 and let H be a stable,
causal linear filter with input $Z$. Then, for $\tilde{Z}$ being the output process of the filter,
\[ \tilde{T}(W; Z) = \tilde{T}(W; \tilde{Z}). \] (8)

**Proof:** See Appendix D-C

As the previous lemma suggests, the output process of a stable, causal linear filter is equivalent to its input process in the sense of Assumption 2. Moreover, combining this lemma with Definition 2 one can see that filtering a process with a stable, causal filter does not destroy information:
\[ I_W(Z \rightarrow \tilde{Z}) = I(W; Z) - I(W; \tilde{Z}) = 0. \] (9)

III. RELATIVE INFORMATION LOSS IN A DOWNSAMPLER

Consider the scenario depicted in Fig. I where $X$ satisfies Assumption 1. If the filter $H$ is stable and causal, so does $X$. To analyze the information loss rate in the downsampling device, we employ the relative information loss rate,
\[ I(X^{(M)} \rightarrow Y) = \lim_{n \to \infty} I((X^{(M)})_n \rightarrow Y^n) \] (10)

where we applied $M$-fold blocking to ensure that the mapping between $(X^{(M)})_1^n$ and $Y^n$ is static. Downsampling, $Y_n := X_{nM}$, is now a projection to a single coordinate, hence [13]
\[ I((X^{(M)})_1^n \rightarrow Y^n) = \frac{d((X^{(M)})_1^n)Y^n)}{d(X^{(M)})_1^n} = \frac{n(M-1)}{nM}. \] (11)

If the filter $H$ is stable and causal and, thus, has no influence on the information content of the stochastic process, we can use Assumption 2 in (a) below and combine (10) with (11) to
\[ I(X^{(M)} \rightarrow Y) \overset{(a)}{=} I(X^{(M)} \rightarrow Y) = \frac{M-1}{M}. \] (12)

The amount of information lost in the decimation system in Fig. I is the same for all stable, causal filters $H$.

The question remains whether an ideal anti-aliasing filter can prevent information loss, since it guarantees that the downsampling operation is invertible. To show that the answer to this question is negative, take, for example, the ideal low-pass filter recommended in standard textbooks [11] Ch. 4.6:
\[ H(e^{j\theta}) = \begin{cases} 1, & \text{if } |\theta| < \frac{\pi}{M} \\ 0, & \text{else} \end{cases} \] (13)

We decompose $X$ in an $M$-channel filterbank: The $k$-th channel is characterized by analysis and synthesis filters being constant in the frequency band $(k-1)/M \leq |\theta| < k/M$ and zero elsewhere. Let $Y_k$ be the ($M$-fold downsampled) process in the $k$-th channel — clearly, $Y_k \equiv Y_1$. It can be shown that every $Y_k$ satisfies Assumption 1 if $X$ is Gaussian (cf. proof of Theorem 1). Thus we obtain
\[ I(X^{(M)} \rightarrow Y) \overset{(a)}{=} I(Y_1, \ldots, Y_M \rightarrow Y_1) = \frac{M-1}{M}. \] (14)

where the information is again lost in a projection and where (a) is due to Assumption 2 since the filterbank decomposition is invertible. The ideal anti-aliasing low-pass filter prevents information from being lost in the downsampler by destroying information itself.

If the filter $H$ is a cascade of a causal, stable filter and of one with a piecewise-constant transfer function (with less trivial intervals as pass-bands), the analysis still holds; Information is either lost in the filter or in the downsampler:

**Theorem 1.** For a Gaussian process $X$ satisfying Assumption 1, the relative information loss rate in the decimation system depicted in Fig. I satisfies
\[ I(X^{(M)} \rightarrow Y) \geq \frac{M-1}{M} \] (15)

for every anti-aliasing filter $H$ with finitely many pass-band intervals.

**Proof:** See Appendix A

The reason for this seemingly counter-intuitive result is that, without a specific signal model, the amount of information is not necessarily proportional to the amount of energy by which it is represented: There is no reason to prefer a specific frequency band over another. This in some sense parallels our result on the relative information loss in principal components analysis (PCA), where we showed that PCA cannot reduce the amount of information being lost in reducing the dimensionality of the data [13].

IV. RELEVANT INFORMATION LOSS: GAUSSIAN CASE

To remove the counter-intuitivity of the previous section, we adapt the signal model: Let $S$ and $X$ be jointly stationary Gaussian processes with PSDs $S_S(e^{j\theta})$ and $S_X(e^{j\theta})$, respectively, with cross PSD $S_{SX}(e^{j\theta})$, and which satisfy Assumption 1. The information loss rate relevant w.r.t. $S$ is given by
\[ I(S^{(M)} \rightarrow Y) = I(S^{(M)}; X^{(M)}) - I(S^{(M)}; Y) \] (16)

and measures how much of the information $X$ conveys about $S$ is lost for each output sample due to downsampling.

While in the general case the filter which minimizes $I_S(S^{(M)} \rightarrow Y)$ is hard to find, for this Gaussian signal model the solution is surprisingly intuitive:

**Definition 3** (Optimal Energy Compaction Filter [17], Thm. 4)). The optimal energy compaction filter $H$ for an $M$-fold downsampler and for a given PSD $S_X(e^{j\theta})$ satisfies
\[ H(e^{j\theta}) = \begin{cases} 1, & \text{for smallest } l \text{ s.t. } \forall k : S_X(e^{j\theta_k}) \geq S_X(e^{j\theta_k}) \\ 0, & \text{else} \end{cases} \] (17)

where $\theta_k := \frac{\theta - 2k\pi}{M}$.

The energy compaction filter for a given PSD can be constructed easily: The $M$-fold downsampled PSD consists of $M$ aliased components; for each frequency point $\theta \in [-\pi/M, \pi/M]$, at least one of them is maximal. The pass-bands of the energy compaction filter correspond to exactly these maximal components [2], [3], [17].

**Theorem 2.** Let $S$ and $X$ be jointly stationary Gaussian processes satisfying Assumption 1 and having PSDs $S_S(e^{j\theta})$ and $S_X(e^{j\theta})$. Let further the cross PSD $S_{SX}(e^{j\theta})$ be such that
\[ \int_{-\pi}^{\pi} \ln (S_S(e^{j\theta})S_X(e^{j\theta}) - |S_{SX}(e^{j\theta})|^2) d\theta > -\infty. \] (18)
Then, the $M$-aliasing-free energy compaction filter for
\[
\frac{|S_{X}(e^{j\theta})|^2}{S_S(e^{j\theta})S_X(e^{j\theta}) - |S_{S}(e^{j\theta})|^2}
\]
minimizes the information loss rate relevant w.r.t. $S$ in the
decimation system depicted in Fig. 1.

**Proof:** See Appendix B.

The condition imposed on the cross PSD ensures that the
two-dimensional process $(S, X)$ is regular in the sense of (18),
in particular, it excludes Gaussian processes being linearly
dependent, e.g., where $X$ is obtained by filtering $S$.

The presented theorem admits an interesting

**Corollary 1.** Let $S$ and $N$ be independent, jointly stationary
Gaussian processes satisfying Assumption 7 and having PSDs
$S_S(e^{j\theta})$ and $S_N(e^{j\theta})$. Let $X_n = S_n + N_n$. Then, the $M$-
aliasing-free energy compaction filter for $S_S(e^{j\theta})/S_N(e^{j\theta})$
minimizes the information loss rate relevant w.r.t. $S$ in the
decimation system depicted in Fig. 1.

**Proof:** Due to independence, $S_{XX}(e^{j\theta}) = S_S(e^{j\theta})$ and
$S_X(e^{j\theta}) = S_S(e^{j\theta}) + S_N(e^{j\theta})$.

The energy compaction filter minimizing the relevant information
loss rate thus maximizes the SNR at each frequency.
In particular, since for white Gaussian noise $N$ the energy
compaction filter for $S_S(e^{j\theta})/S_N(e^{j\theta})$ coincides with the
energy compaction filter for $S_S(e^{j\theta})$, the filter that lets most of
the signal’s energy pass aliasing-free is also optimal in terms of
information.

Also the energy compaction filter of Theorem 2 in some
sense maximizes the SNR, if one interprets the numerator
of (19) as the signal, and the denominator as the noise
component.

Corollary 1 also connects tightly to (8), in which Chen et
al. analyzed the capacity of sub-Nyquist sampled, continuous-
time additive Gaussian noise channels with frequency response
$H_{\text{channel}}(f)$. 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 [8 Thm. 3].

**V. RELEVANT INFORMATION LOSS: NON-GAUSSIAN SIGNAL PLUS GAUSSIAN NOISE**

Although the result for Gaussian processes is interesting
due to its closed form, it is of little practical relevance. In
many cases, at least the relevant part of $X$, the data signal
process $S$, is non-Gaussian. We thus drop the restriction that
$S$ is Gaussian. For the result presented below, we have to
assume a signal-plus-Gaussian-noise model, i.e., we assume
that $X$ is the sum of $S$ and an independent Gaussian noise
process $N$.

One can expect that in this case a closed-form solution for $H$
will not be available. Assuming that $S$ is Gaussian yields an
upper bound on the information rate $I(S^{(M)}; Y)$. While this
upper bound is of little use for filter design (it does not make
sense to maximize an upper bound on the information rate),
it can also be shown that the Gaussian assumption provides
an upper bound on the relevant information loss rate. To this
end, we employ the approach of Plumbley [9], who showed
that, with a specific signal model, PCA can be justified from
an information-theoretic perspective (cf. also [14]).

**Theorem 3.** Let $H$ be stable and causal, let $S$ and $N$ be independent,
jointly stationary and satisfy Assumption 7 and
let $X_n = S_n + N_n$. $N$ is Gaussian, and $S_G$ is Gaussian with the
same PSD as $S$. Let $X_{G,n} = S_{G,n} + N_n$, and let $Y_G$ be
the corresponding output process of the decimation system,
respectively. Then,
\[
\mathbb{T}_{S^{(M)}_G}(X^{(M)}_G \rightarrow Y_G) \leq \mathbb{T}_{S^{(M)}_G}(X^{(M)}_G \rightarrow Y_G).
\]

**Proof:** See Appendix C.

A consequence of this theorem is that filter design by
energetic considerations, i.e., by considering the PSDs of the
signals only, has performance guarantees also in information-
theoretic terms. In particular, while the theorem is restricted
to stable and causal filters, intuition suggests that a high-order
filter in some way should approximate the energy compaction
filter from Corollary 1. One has to consider, though, that the
filter $H$ optimal in the sense of the upper bound might not
coincide with the optimal filter w.r.t. $T_{S^{(M)}_G}(X^{(M)}_G \rightarrow Y_G)$.

Note that, to the best of our knowledge, the statements of
Theorem 3 cannot be generalized to arbitrary correlations
between $S$ and $X$, as in Theorem 2. The reason is that applying
Plumbley’s idea requires an independent, additive Gaussian
noise component. At best, a generalization to non-Gaussian
noise is possible, if the noise is more Gaussian than the signal
in a well-defined sense (cf. [13]). This generalization, however,
is within the scope of future work.

**VI. EXAMPLES**

We now illustrate our results with an example: Let the
PSD of $S$ be given by $S_S(e^{j\theta}) = 1 + \cos \theta$ and let $N$
be independent white Gaussian noise with variance $\sigma^2$, i.e.,
$S_N(e^{j\theta}) = \sigma^2$. The PSD of $X$ is depicted in Fig. 2. We
consider downsampling by a factor of $M = 2$. Were $S$ Gaussian
too, the optimal filter would be an ideal low-pass filter with
cut-off frequency $\pi/2$ (cf. Corollary 1).

If we assume that $S$ is non-Gaussian, Theorem 3 allows us
to design a finite-order filter which minimizes an upper
bound on the relevant information loss rate. In particular,
it can be shown that among all first-order FIR filters with
impulse response $h[n] = \delta[n] + c\delta[n - 1]$, the filter with $c = 1$
minimizes the Gaussian bound (see also Section VII).

Fig. 3 shows the upper bound on the relevant information
loss rate as a function of the noise variance $\sigma^2$ for the ideal
low-pass filter and the optimal first-order FIR filter compared to the case where no filter is used. In addition, the available information \( I(X_G^{(2)}; S_G^{(2)}) = 2I(X_G; S_G) \) is plotted, which decreases with increasing noise variance. Indeed, filtering can reduce the relevant information loss rate compared to omitting the filter. This is in stark contrast with the results of Section \( \text{III} \) in which we showed that the relative information loss rate equals 1/2 regardless of the filter. The reason is that in Section \( \text{III} \) we did not have a signal model in mind, treating every bit of information equally. As soon as one knows which aspect of a stochastic process is relevant, one can successfully apply signal processing methods to retrieve as much information as possible (or to remove as much of the irrelevant information as possible, cf. [14]).

Interestingly, as Fig. 3 shows, the improvement of a first-order FIR filter over direct downsampling is significant. Using low-order filters is beneficial also from a computational perspective: To the best of our knowledge, the optimization problem does not permit a closed-form solution for the filter coefficients in general. Thus, numerical procedures will benefit from the fact that the number of coefficients can be kept small.

We repeated the experiment 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, we also determined numerically the optimal filter coefficients for a second-order FIR filter with impulse response \( h[n] = \delta[n] + c_1 \delta[n-1] + c_2 \delta[n-2] \). Remarkably, for all considered variances, the optimal value for \( c_2 \) is equal to one. The optimal value for \( c_1 \), however, depends on the variance \( \sigma^2 \) of the noise process, as indicated in Fig. 5. The filter coefficient is close to \( \sqrt{2} \), which yields the impulse response vector equal to the maximal eigenvector of the input process’ autocorrelation matrix, and hence to the solution maximizing the filter output signal-to-noise ratio (see Section \( \text{VII} \)). 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 to 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.

Comparing the relevant information loss rates depicted in Fig. 3 and Fig. 4, one can observe that the loss is greater than for two-fold downsampling. For comparison, again the available information rate \( I(X_G^{(3)}; S_G^{(3)}) = 3I(X_G; S_G) \) is plotted. Finally, Fig. 4 shows the additional loss induced by replacing the ideal coefficient \( c_1 \) by \( \sqrt{2} \), the coefficient yielding a maximum output signal-to-noise ratio. As can be seen, the additional loss is negligible, which justifies energetic design considerations from an information-theoretic point-of-

Fig. 3. Upper bounds on the relevant information loss rate in nats as a function of the noise variance \( \sigma^2 \) for various filter options (\( M = 2 \)).

Fig. 4. Upper bounds on the relevant information loss rate in nats as a function of the noise variance \( \sigma^2 \) for various filter options (\( M = 3 \)).
view, at least in this example and for strong additive white Gaussian noise.

VII. DISCUSSION AND OUTLOOK

In our opinion, the present work has two important messages worth repeating: First of all, we showed that, with no signal model in mind, anti-aliasing filtering is futile. Assuming that it is sensible to preserve as much energy aliasing-free as possible is guesswork and expresses the common misconception that energy and information behave similarly. In truth, the relevant information may be transported mainly in small signal components, or even in the sum of the alias terms. Hence, it might even turn out that anti-aliasing filtering, as it is proposed by standard textbooks on signal processing, does more harm than good. In this sense, the analysis of anti-aliasing filtering parallels our previous analysis of PCA in [13], where the conclusion was similar.

The second important message is that with a specific signal model in mind, anti-aliasing filtering can indeed reduce the information loss in the downsampler. In particular, for a Gaussian signal-plus-noise model it does make sense to preserve the signal components with the largest signal-to-noise ratio aliasing-free. Then, the information-theoretic optimum coincides with the energetic one, and filter design based on second-order statistics is well-justified.

One of the most important aims of future work is the design of finite-order filters with information-theoretic cost functions. While the information-maximizing filter with unconstrained order is simple to obtain (cf. Theorem 2), the practically more relevant case of finite-order filters is much more difficult even in the purely Gaussian signal-plus-noise case: The problem of maximizing (cf. Lemma 3)

$$\mathcal{I}(S_G^M; Y_G) = \mathcal{I}(S_G^M; Y_G)$$

which does, except in particularly simple cases (see Section VII), not permit a closed-form solution, nor is it necessarily convex.

The situation simplifies when the noise is white, i.e., when \( S_N(e^{j\theta}) = \sigma_N^2 \), and with the restriction that the filter satisfies the Nyquist-M condition [17], [19]

$$\frac{1}{M} \sum_{k=0}^{M-1} |H(e^{j\theta_k})|^2 = 1.$$  \hspace{1cm} (22)

This restriction is meaningful, e.g., when the filter is part of an orthonormal filter bank or a principal component filter bank.

Employing these restrictions and applying Jensen’s inequality to (21) yields an upper bound on the information rate

$$\mathcal{I}(S_G^M; Y_G) \leq \frac{1}{2} \ln \left( 1 + \frac{\sigma_N^2}{\sigma_N^2} \right) = \frac{1}{2} \ln \left( 1 + \frac{\sigma_N^2}{\sigma_N^2} \right)$$  \hspace{1cm} (23)

where \((\sigma_N^2)\) is because the variance of a stationary process does not change during downsampling and where \(\sigma_N^2\) is the variance of \(\hat{S}(\tilde{N})\), the output of \(H\) to the input process \(S\) (N).

Maximizing an upper bound on the information rate thus amounts to maximizing the signal-to-noise ratio, or equivalently, the signal power, at the output of the downsampler or filter. This is exactly the objective of optimum FIR compaction filters for \(S_G(e^{j\theta})\), which have been investigated in [19] and the references therein. The solution for filter orders strictly smaller than the downsampling factor \(M\) is the maximal eigenvector of the autocorrelation matrix [19]. For larger filter orders, various analytical and numerical methods exist; see [20] for an overview. All these represent a sub-optimal solution to the original problem of designing information-maximizing FIR filters; the problem of designing finite-order filters with, e.g., rational transfer functions, remains elusive.

Obviously, the upper bound (23) is the better the larger the noise variance \(\sigma_N^2\). Hence, energetic design considerations will succeed especially in cases where the Gaussian noise is white and has a large variance; see also Fig. 6. One has to keep in mind, however, that even 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, however, is either a lower bound on the information rate or an upper bound on the relevant information loss rate. Future work shall deal with this issue.

The extension of this work’s results to sampling of continuous-time processes is also of great interest. The extension in terms of relevant information loss rate has been made partly in [3], presenting a result similar to our Corollary 1. The authors of [3] furthermore showed that a filterbank sampling mechanism can have a strictly larger capacity than a single-channel sampling mechanism, suggesting that one can further reduce the relevant information loss rate in the downsampler by replacing the filter \(H\) by a filterbank.

In terms of relative information loss rates, the extension to continuous-time processes is immediate via employing a sampling expansion (Nyquist rate) and successive downsampling, at least for bandlimited processes. If the input process is not bandlimited and has a positive PSD a.e., we conjecture that the relative information loss rate will approach unity, i.e., that 100% of the available information is lost.

Finally, the generalization of our Theorems 1 and 3 to non-Gaussian processes and general filters \(H\), respectively, is the goal of future work. While the former is already sketched in Appendix A, the latter requires deeper investigation.

APPENDIX A

PROOF OF THEOREM 1

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 \(H(e^{j\theta})\) being either one or zero. This assumption is unproblematic, since \(H\) can always be split into a filter satisfying this assumption and a set of filters satisfying the
Paley-Wiener condition \([7]\). The latter filters can be omitted as made clear above.

Next, assume that the pass-band and stop-band intervals have rational endpoints. In other words, 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. 7 which illustrates the filterbank decomposition of \(X\) [1 Ch. 4.7.4, p. 228]. There, \(H_i\) is an ideal brick-wall filter for the \(i\)-th frequency band, i.e.,

\[
H_i(e^{j\theta}) = \begin{cases} 1, & \text{if } \frac{(i-1)\pi}{LM} < |\theta - 2k\pi| \leq \frac{i\pi}{LM}, \ k \in \mathbb{Z} \\ 0, & \text{else} \end{cases}
\]  

(24)

Since \(X\) is a Gaussian process [16 p. 663],

\[
\overline{r}(X) = \frac{1}{2} \ln(2\pi e) + \frac{1}{4} \int_{-\pi}^{\pi} \ln S_X(e^{j\theta}) d\theta,
\]  

(25)

and from \(|\overline{r}(X)| < \infty\) it follows that \(S_X(e^{j\theta}) > 0\) a.e. It naturally follows that, for all \(i = 1, \ldots, LM\), \(S_{\tilde{Y}_i}(e^{j\theta}) > 0\) a.e., where of course \(\tilde{Y}_i\) is Gaussian too.

Clearly, the variance of the \(i\)-th downsampled process \(\tilde{Y}_i\) is positive (since its PSD is positive a.e.) and finite (since it is upper bounded by \(LM\) times the variance of \(X\)). Thus, \(|\overline{r}(\tilde{Y}_i)| < \infty\), and \(\overline{r}(\tilde{Y}_i) < \infty\). The differential entropy rates of \(\tilde{Y}_i\) are obtained by splitting the integral in (25) 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 \(-\infty\) by the fact that \(|\overline{r}(\tilde{Y}_i)| \leq 0\)). Thus, \(|\overline{r}(\tilde{Y}_i)| < \infty\), and with Lemma 1 it follows that

\[
dl((\tilde{Y}_i)^n) = n
\]  

(26)

for all \(i\). Moreover, since the downsampled processes \(\tilde{Y}_i\) are mutually independent (they are uncorrelated and jointly Gaussian), it follows that

\[
dl((\tilde{Y}_1)^n, \ldots, (\tilde{Y}_{LM})^n) = nLM.
\]  

(27)

Since the collection \(\tilde{Y} := \{\tilde{Y}_1, \ldots, \tilde{Y}_{LM}\}\) is equivalent to \(X\), in the sense that perfect reconstruction is possible, by Assumption 2

\[
l(M) = l(Y \rightarrow Y^{(L)}).
\]  

(28)

We employ 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 \(\tilde{Y}_i\); a simple projection. What remains to be analyzed is the effect of the \(M\)-fold down sampler, which can also be moved next to the reconstruction filters due to linearity. Notice that with the polyphase decomposition of decimation systems (cf. [1 Ch. 4.7.4, p. 228]), the \(i\)-th branch of the filterbank can be rearranged as in Fig. 8. Due to the cascade of up- and downsampling, only the filter \(H_0\) is relevant, while all other filters \(H_{i}^d\) will have vanishing input. In particular, while \(H_i\) is given by (24), one gets for the filter \(H_0^d\) with impulse response \(h_i[nM]\),

\[
H_0^d(e^{j\theta}) = \frac{1}{M} \sum_{m=0}^{M-1} H_i(e^{j\frac{2\pi m}{M}})
\]  

(29)

\[
= \begin{cases} \frac{1}{M}, & \text{if } \frac{(i-1)\pi}{LM} < |\theta - 2k\pi| \leq \frac{i\pi}{LM}, \ k \in \mathbb{Z} \\ 0, & \text{else} \end{cases}
\]  

(30)

where the last line follows from the fact that \(H_i\) is \(LM\)-aliasing-free and, hence, \(M\)-aliasing free (\(H_i\) have bandwidths \(1/LM\) and fall in exactly one of the bands with width \(1/M\)).

By the \(2\pi\)-periodicity of the transfer functions it follows that the sequence of filters is periodic with \(2L\), i.e., \(H_0 = H_{2L-1}^d\). Moreover, \(H_{L+k}^d = H_{L-k}^d\), \(k = 1, \ldots, L\), by the symmetry of the filter. Therefore, there are exactly \(L\) different filters, each occurring \(M\) times.

Combining the last system from Fig. 8 with Fig. 7 and (30), the schematic in Fig. 9 is obtained. Note that since the filters \(H_0\) are orthogonal and \(L\)-aliasing-free (i.e., the frequency response of the filter does not overlap by \(L\)-fold downsampling and can thus be reconstructed perfectly), adding the reconstruction filter outputs does not incur information loss. We thus again use Assumption 2 and write

\[
l(X^{(M)} \rightarrow Y) = l(\tilde{Y} \rightarrow Y^{(L)}) = l(\tilde{Y} \rightarrow \tilde{Y})
\]  

(31)

where \(\tilde{Y} := \{\tilde{Y}_1, \ldots, \tilde{Y}_L\}\). But the transformation from \(\tilde{Y}\) to \(\tilde{Y}\) is linear, specifically, the cascade of an invertible linear
map and a projection. We therefore apply [13 Cor. 1] and get
\[
\begin{align*}
\ell((\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L) &\rightarrow (\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L) \\
= 1 - \frac{d((\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L)}{d((\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L)} &= 1 - \frac{d((\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L)}{nML}.
\end{align*}
\]
(32)

The information dimension of \{((\hat{Y}_1)^n_1, \ldots, (\hat{Y}_L)^n_L)\} 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 \(A_i \in \mathbb{Z}\) 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. 9 is zero or one). For both of these filters, however, the information dimension of \(\hat{Y}_i^n\) cannot exceed \(nL\), and above analysis holds. This completes the proof.

Note that the proof suggests how to measure the exact relative information loss rate for the decimation system by evaluating the information dimension of \(\hat{Y}_i^n\). For rational endpoints of the pass-band intervals this is simple since \(d((\hat{Y}_i)^n_i) \in \{0, n\}\). For irrational endpoints one can always wedge the filter \(H\) between one with destroys more and one which destroys less information; for \(L\) sufficiently large, the resulting difference in the relative information loss rates will be small, and eventually vanish in the limit \(L \rightarrow \infty\).

Moreover, the result should also hold for non-Gaussian processes satisfying Assumption 1. The intuition behind this is a bottleneck consideration: Since the filterbank decomposition is perfectly invertible, the information dimensions of the input and output processes need to be identical for all time windows \(\{1, \ldots, n\}\). The Gaussian assumption was required 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. We believe 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, cf. [21]. As a consequence, it is not possible that the information dimension of the sub-band processes is smaller than in the Gaussian case, since then the information dimension of the (reconstructed) output would be smaller than the information dimension of the input – a contradiction.

**APPENDIX B**

**PROOF OF THEOREM 1**

The goal is to maximize the information rate between \(Y\) and the \(M\)-dimensional input process \(S_i^{(M)}\), i.e., \(\mathcal{T}(S_i^{(M)}; Y)\), because it is the only component of \(\mathcal{T}_{S_i^{(M)}}(X^{(M)} \rightarrow Y)\) depending on \(H\).

The \(l\)-th coordinate of \(S_i^{(M)}\) shall be the process \(S_l\) with samples \(S_1, S_{M+1}, S_{2M+1}, \ldots\), where \(l = 1, \ldots, M\). We note in passing that the processes \(S_l\) constitute the polyphase decomposition of \(S\). The PSD of the \(l\)-th coordinate is given as

\[S_l(e^{j\theta}) := S_{S_l}(e^{j\theta}) = \frac{1}{M} \sum_{k=0}^{M-1} S_S(e^{jk\theta}); \quad (33)\]

It is immaterial whether the sum runs from 0 to \(M - 1\) or from 1 to \(M\). We will make repeated use of this fact below.
the cross PSD between the l-th and the m-th coordinate is
\[ S_{lm}(e^{i\theta}) := S_{SlSm}(e^{i\theta}) = \frac{1}{M} \sum_{k=0}^{M-1} S_k(e^{i\theta_k}) e^{i(l-m)\theta_k}. \]  
(34)

Note that \( S_{lm}(e^{i\theta}) = S_{ml}^*(e^{i\theta}) \).

With \( S_X(e^{i\theta}) = S_X(e^{i\theta})|H(e^{i\theta})|^2 \) and \( Y_n := \tilde{X}_{nm} \) the PSD of \( Y \) is given as
\[ S_Y(e^{i\theta}) = \frac{1}{M} \sum_{k=0}^{M-1} S_X(e^{i\theta_k}) |H(e^{i\theta_k})|^2. \]  
(35)

Finally, if \( S_{SX}(e^{i\theta}) \) is the cross PSD between \( S \) and \( X \), one has \( S_{SX}(e^{i\theta}) = S_X(e^{i\theta}) H^*(e^{i\theta}) \) \[ 10 \] Ch. 9-4 and for the cross-PSD between \( S_l \) and \( Y \),
\[ S_{SY}(e^{i\theta}) := S_{SLY}(e^{i\theta}) = \frac{1}{M} \sum_{k=0}^{M-1} S_X(e^{i\theta_k}) H^*(e^{i\theta_k}) e^{i\theta_k}. \]  
(36)

Again, \( S_{LY}(e^{i\theta}) = S_{Yl}(e^{i\theta}) \).

Let \( A_S \) be the \( M \times M \) PSD matrix containing the elements \( S_{lm}(e^{i\theta}) \), let \( y \) be a column vector with elements \( S_{ly}(e^{i\theta}) \), and let
\[ A_{SY} = \begin{bmatrix} A_S & S_Y \\ s^H_Y & S_Y(e^{i\theta}) \end{bmatrix} \]  
(37)

where \( H \) is the Hermitian transposition. Then, if
\[ \int_{-\pi}^{\pi} \ln |\det A_S| d\theta > -\infty \]  
(38)

the information rate \( I(S^{(M)}; Y) \) equals \[ 18 \] Thm. 10.4.1
\[ I(S^{(M)}; Y) = \frac{1}{4\pi} \int_{-\pi}^{\pi} S_Y(e^{i\theta}) \ln |\det A_S| d\theta. \]  
(39)

To verify that condition \( (38) \) holds, note that \( A_S = WAW^H \), where the \((l,m)\)-th element of \( W \) is \( e^{i\theta_{lm}}/\sqrt{M} \) and where \( A \) is a diagonal matrix with \( S_S(e^{i\theta}) \) in its \( l \)-th position. Hence, \( W \) is unitary, \( \det A_S = \det A = \prod_{k=0}^{M-1} S_S(e^{i\theta_k}) \), and, doing some calculus,
\[ \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln |\det A_S| d\theta = \ln |\det A_S|. \]  
(40)

We now consider the fraction \( \det A_{SY} / \det A_S \), where \( A_{SY} = \det A_{SY} / \det A_S \).

Since \( A_S \) is non-singular a.e. by Assumption \[ 11 \] we can write for the adjugate \( \det A_{SY} = A_S^{-1} \det A_{SY} \).

Hence, \( \det A_{SY} = S_Y(e^{i\theta}) - S_Y^H A_S^{-1} S_Y \).

With \( A_S^{-1} = W \Lambda W^H \), we can write for
\[ (A_S^{-1} S_Y)_l = \sum_{n=1}^{M} (A_S^{-1})_{ln} S_n Y(e^{i\theta}) \]  
(43)

\[ = \sum_{n,k,m=1}^{M} S_{SX}(e^{i\theta_{nm}}) H^*(e^{i\theta_{mk}}) e^{i(\theta_{nl} - \theta_{lk})} e^{i\theta_{mk}}. \]  
(44)

But
\[ \sum_{n=0}^{M-1} e^{i\theta_{nm} - \theta_{lk}} = \sum_{n=0}^{M-1} e^{i\theta_{n(m-k)/2}} \]  
(45)

vanishes if \( m \neq k \) and evaluates to \( M \) otherwise. Thus,
\[ (A_S^{-1} S_Y)_l = \sum_{k=1}^{M} S_{SX}(e^{i\theta_{lk}}) H^*(e^{i\theta_{lk}}) e^{i\theta_{lk}}. \]  
(46)
Finally,

\[ s_y^M A_S^{-1} s_y = \sum_{k=0}^{M-1} S_S(\theta_k) \left( H(\theta_k) \right) \]

\[ = \sum_{k,l,m=1}^{M} S_X(\theta_k) H^*(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

\[ = \sum_{k,l,m=1}^{M} S_X(\theta_k) H^*(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

The second term in (40) is a weighted average with weights

\[ \sum_{k,l,m=1}^{M} S_X(\theta_k) H^*(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

\[ \sum_{k,l,m=1}^{M} S_X(\theta_k) H^*(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

Summing \( e^{\theta_k - \theta_m} \) over index \( l \) again vanishes for \( k \neq m \) and evaluates to \( M \) otherwise. Hence,

\[ s_y^M A_S^{-1} s_y = \frac{1}{M} \sum_{k=0}^{M-1} S_S(\theta_k) \left[ H(\theta_k) \right]^2. \]

Maximizing the information rate amounts to pointwise maximizing the argument of the logarithm of (39), i.e.,

\[ \sum_{k=0}^{M-1} S_X(\theta_k) H(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

\[ = 1 + \frac{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]}{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]} \]

\[ \alpha(\theta_k) := \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} \]

The second term in (50) is a weighted average with weights

\[ w_k(\theta) := \alpha(\theta_k) H(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

\[ \sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right] \]

\[ = 1 + \frac{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]}{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]} \]

\[ \alpha(\theta_k) := \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} \]

where we inserted (55) for \( S_Y(\theta^m) \) and where

\[ \alpha(\theta_k) := \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} \]

The second term in (50) is a weighted average with weights

\[ w_k(\theta) := \alpha(\theta_k) H(\theta_k) \left( e^{\theta_k} S_S^{-1} e^{\theta_m} \right) H(\theta_l) \delta(\theta_k - \theta_m). \]

\[ = \frac{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]}{\sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right]} \]

\[ \alpha(\theta_k) := \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} \]

Evidently, all other weights have to be set to zero.

The filter \( H \) is thus related to the piecewise constant functions \( w_k(\theta) \) via

\[ |H(\theta_k)|^2 = \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} w_k(\theta) \]

where the relation has to be fulfilled for all \( k = 0, \ldots, M-1 \).

By assumption, \( |h(\theta)| \leq \infty \) and \( |\tilde{h}(\theta)| \leq \infty \). But since \( \tilde{h}(\theta) = \lim_{\theta \to \infty} h(\theta) \) it also follows from condition-

\[ \sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right] \]

\[ \sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right] \]

\[ \sum_{k=0}^{M-1} \left[ S_X(\theta_k) H(\theta_k) \delta(\theta_k - \theta_m) \right] \]

Evidently, all other weights have to be set to zero.

The filter \( H \) is thus related to the piecewise constant functions \( w_k(\theta) \) via

\[ |H(\theta_k)|^2 = \frac{S_S(\theta_k) S_X(\theta_k) - \left| S_X(\theta_k) \right|^2}{S_S(\theta_k)} w_k(\theta) \]

where the relation has to be fulfilled for all \( k = 0, \ldots, M-1 \).

By assumption, \( h(\theta) \) is continuous and, thus, square summable and causal, the Paley-Wiener theorem [15] p. 215] states that

\[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |H(\theta)| d\theta > -\infty. \]
Moreover, since the filter is stable, one has by Jensen’s inequality:

$$\frac{1}{4\pi} \int_{-\pi}^{\pi} \ln |H(e^{j\theta})|^2 d\theta = \frac{1}{2} \mathbb{E} \left( \ln |H(e^{j\theta})|^2 \right) \leq \frac{1}{2} \mathbb{E} \left( |H(e^{j\theta})|^2 \right) = \frac{1}{2} \ln G < \infty$$

(61)

where the expectation is taken assuming the frequency variable $\theta$ is uniformly distributed on $[-\pi, \pi]$ and where the noise gain $G$ is

$$G := \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\theta})|^2 d\theta.$$  

(64)

The last (strict) inequality follows by assuming stability (square summability of the impulse response and Parseval’s theorem [11, Tab. 2.2, p. 86]).

According to [16] p. 663, the differential entropy rate at the output of the filter $H$ is given by

$$\tilde{h}(Z) = \tilde{h}(Z(0)) + \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |H(e^{j\theta})| d\theta$$

(65)

and thus, by assumption, finite.

With [16] (9-190), p. 421, one has for the autocorrelation function of the output of a linear filter

$$r_{ZZ}[m] = (r_{Zz} * \rho)[m]$$

where

$$\rho[m] = \sum_k s_k[m+k]s_k^*[k].$$

(67)

Thus, by the fact that for a zero-mean process the variance satisfies $\sigma^2 = r_{ZZ}[0] \geq |r_{ZZ}[m]|$,

$$r_{ZZ}[0] \leq \sigma^2 \sum_m |\rho[m]|$$

(68)

$$\leq \sigma^2 \sum_k \sum_m |s_k[m+k]|s_k^*[k]$$

(69)

$$= \sigma^2 \sum_k \left( |s_k[k]| \sum_m |s_k[m+k]| \right)$$

(70)

$$\leq \sigma^2 \sum_k (|s_k[k]|C)$$

(71)

$$\leq \sigma^2 C^2 < \infty$$

(72)

where the last two lines follow from stability of $H$ (the impulse response is absolutely summable) and by the assumption that $Z$ has finite variance. Thus, by conditioning and the maximum-entropy property of the Gaussian distribution,

$$-\infty < \tilde{h}(Z) \leq \tilde{h}(Z) \leq \frac{1}{2} \ln(2\pi e\sigma^2 C^2) < \infty.$$

This completes the proof.

---

**C. Proof of Lemma 3**

The proof is provided for jointly Gaussian processes $W$ and $Z$, only; since the effect of linear filters on the differential entropy rate is independent of the process statistics (cf. [16] p. 663), the result can be extended to the general case.

First, note that a stable, causal filter satisfies the Paley-Wiener condition, and thus $H(e^{j\theta}) > 0$ a.e. From [16] Cor. to Thm. 9-4, p. 412 one gets $S_Z(e^{j\theta}) = |H(e^{j\theta})|^2 S_Z(e^{j\theta})$. Since $Z$ has a finite entropy rate, $S_Z(e^{j\theta}) > 0$ a.e., and thus, $S_Z(e^{j\theta}) > 0$ a.e. That for the cross PSD $S_{ZW}(e^{j\theta}) = H(e^{j\theta})S_{ZW}(e^{j\theta})$ holds can be shown easily.

From [13] Thm. 10.2.1, p. 175,

$$\tilde{I}(Z; W) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \ln (1 - |\rho_{ZW}(e^{j\theta})|^2)$$

(73)

where

$$|\rho_{ZW}(e^{j\theta})|^2 = \begin{cases} \frac{S_{ZW}(e^{j\theta})^2}{S_Z(e^{j\theta})S_W(e^{j\theta})}, & \text{if } S_{ZW}(e^{j\theta}) \neq 0 \\ 0, & \text{else} \end{cases}.$$  

(74)

With above reasoning one gets

$$|\rho_{ZW}(e^{j\theta})|^2$$

(75)

which is a.e. equal to $|\rho_{ZW}(e^{j\theta})|^2$ since $H(e^{j\theta}) > 0$ a.e. This completes the proof.

---

**References**

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

[2] 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.

[3] 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.

[4] J. C. Principe, *Information Theoretic Learning: Rényi’s Entropy and Kernel Perspectives*, ser. Information Science and Statistics. New York, NY: Springer, 2010.

[5] J. I. Galdós and D. E. Gustafson, “Information and distortion in reduced-order filter design,” IEEE Trans. Inf. Theory, vol. 23, no. 2, pp. 183–194, Mar. 1977.

[6] N. Al-Dhahir and J. M. Cioffi, “Block transmission over dispersive channels: transmit filter optimization and realization, and MMSE-DFE receiver performance,” IEEE Trans. Inf. Theory, vol. 42, no. 1, pp. 137–160, Jan. 1996.

[7] 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.

[8] 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.

[9] M. Plumbley, “Information theory and unsupervised neural networks,” Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR. 78, 1991.

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

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

[12] R. M. Gray, *Entropy and Information Theory*. New York, NY: Springer, 1990.
[13] B. C. Geiger and G. Kubin, “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].

[14] —— , “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].

[15] A. Papoulis, *The Fourier Integral and its Applications*. McGraw Hill, 1962.

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

[17] P. Vaidyanathan, “Theory of optimal orthonormal subband coders,” *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1528–1543, 1998.

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

[19] 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.

[20] 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.

[21] 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.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 2013.