2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)

INFORMATION-MAXIMIZING PREFILTERS FOR QUANTIZATION

Bernhard C. Geiger and Gernot Kubin

Signal Processing and Speech Communication Lab
Graz University of Technology, Austria
{geiger,g.kubin}@ieee.org

ABSTRACT

This work discusses open-loop and closed-loop prediction from an information-theoretic point-of-view. It is shown that the open-loop predictor which minimizes the mean-squared prediction error differs from the filter maximizing the information rate, but that this difference vanishes for high quantizer resolutions. The filter minimizing the mean-squared reconstruction error performs worse for all quantizer resolutions. For the closed-loop predictor, which is shown to be superior only at low quantizer resolutions, the filters maximizing the information rate and minimizing the mean-squared reconstruction error coincide.

We illustrate these results with a simple example and discuss similarities with the information-theoretic aspects of principal components analysis and anti-aliasing filtering. Furthermore, we briefly discuss the classical Wiener filter followed by a quantizer.

Index Terms— Information rate, quantization, prediction, Wiener filter

1. INTRODUCTION

A stable, causal linear filter cannot change the information content of a signal — this result is as old as information theory [1] and made its way into textbooks on probability theory, e.g, [2, p. 663]. It has not made its way, unfortunately, into textbooks on signal processing, in which linear filters and energetic measures (such as the mean-squared error) are still regarded as the most common tools for processing and measuring information. Only recently, signal processing has incorporated information-theoretic methods for its purpose: New approaches to adaptive filter design [3] are just one example.

The problem of this work – 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) [4], also the mutual information rate has been considered for quantizer design [5, 6]. Linear pre-processing prior to quantization has been investigated, e.g., in [7]. Joint optimization w.r.t. both MSRE and information rate is considered in rate-distortion theory [8, Ch. 10], [9–11].

What, to the best of our knowledge, is currently lacking is an information-theoretic analysis of energetic designs: For the specific problem of filtering prior to quantization, in which cases is a prediction filter minimizing some error variance optimal also in terms of information rate? We answer this question by highlighting similarities and differences between energetic and information-theoretic cost functions for linear pre-processing systems in Section 2. Then, for the specific problem of quantization, we determine the information-maximizing prefilters both for open- and closed-loop schemes, and compare the results to the energetic optima (Section 3). After briefly considering the restriction to FIR filters in Section 4, we illustrate our results with numerical examples.

2. ENERGETIC VS. INFORMATION-THEORETIC COST FUNCTIONS

An elementary result from information theory states that information, once lost in a nonlinear deterministic system, cannot be recovered [8, Ch. 2.8]. However, while post-processing cannot prevent information loss, pre-processing by, e.g., linear filtering, can. As a consequence, information-theoretic filter design only requires to maximize the information rate between the signal source and the output of the nonlinear device – the quantizer, in this case. It is not necessary to provide a method for reconstructing the source signal[1], as it usually is for energetic measures like the MSRE. In this sense, information-theoretic cost functions have an inherent advantage over energetic ones.

Regarding similarities and differences between these cost functions, we would like to start with a particularly intuitive example: that of principal components analysis (PCA) prior to dimensionality reduction. It is well-known that PCA preserves the subspace with the largest variance and, hence, minimizes the reconstruction error variance (e.g. [12]). The information-theoretic analysis is somewhat more difficult: Given that the input vector has a continuous joint distribution, an infinite amount of information is both lost and preserved[2], regardless which (full-rank) matrix is used for rotating or translating the vector prior to dimensionality reduction. Even more surprisingly, also the *percentage* of information lost or preserved is independent from the choice of the transformation matrix, as we made precise recently [13]: PCA cannot reduce the information lost in dimensionality reduction.

Things look totally different if the input vector is modeled as the sum of a signal and an independent, identically distributed (iid) noise vector, where the information *relevant* to the information sink is contained in the signal vector only. In case both signal and noise are Gaussian, Linsker showed that PCA maximizes the information transfer [14]. For Gaussian noise but a non-Gaussian signal it can be shown that PCA at least minimizes an upper bound on the information lost in dimensionality reduction [12, 15, 16]. We want to stress, however, that these optimality results hold only under specific assumptions on the signal model; in more general cases, PCA completely fails in information-theoretic terms [16, Sec. VI.C].

[1] In particular, we argue that in some cases reconstruction is not even necessary in practice: For example, an appropriately designed automatic speech recognition system could, theoretically, work equally well on the output of the transmission channel as on the reconstructed signal.

[2] A continuous-valued random variable is described by an infinite number of bits.

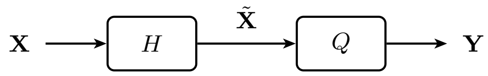


Fig. 1: Quantization system consisting of a linear filter H and an entropy-R constrained uniform quantizer Q.

A very similar picture appears for anti-aliasing filtering prior to downsampling of continuous-valued, discrete-time, non-bandlimited processes. The filter maximizing the energy transfer and minimizing the MSRE is known to be an aliasing-free energy compaction filter, whose passband depends on the power spectral density of the input process [17, 18]. From an information-theoretic perspective the downsampled signal shares infinite information with the input process, and at the same time suffers from infinite information loss. Specifically, the percentage of information lost during decimation is independent of the filter [19]: It is either lost in the downsampler or in the filter (e.g., if the filter is an ideal low-pass).

This counter-intuitivity can again be resolved by assuming that the input process is a signal process superimposed by noise. In case both processes are Gaussian and the noise is white, the energy compaction filter from [17, 18] maximizes the information rate over the downsampler. In case the signal process is non-Gaussian, at least an upper bound on the information loss rate can be minimized [19]. Again, optimality of linear filtering strongly depends on the assumptions of the signal model.

Finally, consider the problem of this work: filtering prior to quantization as illustrated in Fig. 1. While the amount of information lost in the quantizer is still infinite (even 100%, cf. [13]), the information rate between the source and the quantizer output is finite this time. It is, therefore, possible to maximize it by adapting the filter accordingly. We will stick to our previous modeling assumption and 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.

The optimality results we will show in the following depend again strongly on the validity of the signal model. Specifically, one cannot expect that prefiltering a non-Gaussian process with non-linear dependence structure will always increase the information rate at the quantizer output.

3. OPTIMAL PREFILTERS FOR QUANTIZATION

We now investigate the effect of prefiltering on the information rate over a uniform quantizer (see Fig. 1). We 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 proportional to the variance $\sigma^2_{\tilde{X}}$ of $\tilde{\mathbf{X}}$, cf. [20].

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 statement precise, we model the quantizer as an additive white Gaussian noise channel. While this assumption is uncommon, it holds for vector quantizers, cf. [11], and provides a lower bound on the information rate for a Gaussian input process $\mathbf{X}$ for high quantizer resolutions. Thus, let $\mathbf{X}$ be a stationary Gaussian process with positive power spectral density (PSD) $S_X(\mathrm{e}^{\jmath\theta})$ and variance σ^2_X. 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 $S_Y(\mathrm{e}^{\jmath\theta}) = S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) + \sigma^2_{\tilde{X}}\gamma$, 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 [21]

$$\bar{I}(\mathbf{X};\mathbf{Y}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1+\frac{S_{\tilde{X}}(\mathrm{e}^{\jmath\theta})}{\sigma^2_{\tilde{X}}\gamma}\right) d\theta. \tag{1}$$

Via Jensen's inequality, this expression can be bounded from above by R, 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 prediction error $\sigma^2_e = \sigma^2_{\tilde{X}}$ is minimized. Then, $S_{\tilde{X}}(\mathrm{e}^{\jmath\theta}) = \sigma^2_{\tilde{X}} = \sigma^2_\infty$, 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, [22, Ch. 8] shows that the infinite-order linear predictor cannot improve these two quantities compared to omitting the filter ($H \equiv 1$). If the MSRE is taken as cost function, the optimal filter turns out to be *half-whitening* [23]. Naturally, this half-whitening 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. Hence, one has to take care in choosing the cost function appropriate for the application.

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. To this end, recall that a causal, stable linear filter H^{-1} does not influence the information rate, i.e., that $\bar{I}(\mathbf{X};\mathbf{Y}) = \bar{I}(\mathbf{X};\hat{\mathbf{X}})$ if $\hat{\mathbf{X}}$ is obtained by filtering $\mathbf{Y}$ with H^{-1}. To calculate the information rate for closed-loop prediction, we 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, the prediction is based on noisy samples and the prediction error σ^2_e satisfies $\sigma^2_e > \sigma^2_{\tilde{X}}$, where the difference becomes small for high quantizer resolutions [23, pp. 1505]. Consequently, the variance of $\mathbf{Q}$ is $\sigma^2_e\gamma$ and the information rate is

$$\bar{I}(\mathbf{X};\hat{\mathbf{X}}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1+\frac{S_X(\mathrm{e}^{\jmath\theta})}{\sigma^2_e\gamma}\right) d\theta \tag{2}$$

where σ^2_e is the only term that depends on the filter H. Minimizing the prediction error σ^2_e hence not only maximizes the information rate, but also minimizes the MSRE.

That closed-loop prediction is superior to open-loop prediction in terms of MSRE is well-known [22, 23]; 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., [8, 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 [10] and they quantified the additional rate necessary to achieve the desired distortion for first-order autoregressive processes [9]. Quite contrarily, with proper pre- and post-filtering, the rate-distortion function is achievable by closed-loop prediction [11].

4. 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 (2) gives an affirmative answer for the closed-loop predictor, for open-loop prediction the optimal solution will be different (except 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 $S_{\tilde{X}}(e^{\jmath\theta})/\sigma^2_{\tilde{X}}$ in the argument of the logarithm in (1), thus

$$\bar{I}(\mathbf{X};\mathbf{Y}) \approx \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(\frac{S_{\tilde{X}}(e^{\jmath\theta})}{\sigma^2_{\tilde{X}}\gamma}\right) d\theta = \frac{1}{2}\ln\left(\frac{\sigma^2_\infty}{\sigma^2_{\tilde{X}}\gamma}\right). \quad (3)$$

Maximizing the information rate for large R is then 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 $\mathbf{X}$ [24, Ch. 3], the filter maximizing the information rate involves nonlinear optimization. Closed-form solutions are, if at all, only available for particularly simple examples. Above argumentation shows that, at least for high-resolution quantization, nonlinear optimization will not lead to significant performance gains compared to the closed-form energetic optima. Hence, while in general the energetically optimal FIR filter is different from the information-theoretic solution, in some cases design based on energetic cost functions can be justified also in information-theoretic terms.

5. EXAMPLES

We illustrate our results with a simple example: 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 $S_X(e^{\jmath\theta}) = 4\cos^2(\theta/2)$, 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 $\mathbf{h}^1 = [h_0,\ h_1]^T$ or $\mathbf{h}^2 = [h_0,\ h_1,\ h_2]^T$. The filter coefficients minimizing the mean-squared prediction error (PE) are [24, Ch. 3]:

$$\mathbf{h}^1_{\mathrm{PE}} = [1,\ -1/2]^T \text{ with } \sigma^2_1 = 1.5 \quad (4a)$$

and

$$\mathbf{h}^2_{\mathrm{PE}} = [1,\ -2/3,\ 1/3]^T \text{ with } \sigma^2_2 = 1.333 \quad (4b)$$

where σ^2_L denotes the PE of the L-th order predictor. Clearly, for the infinite-order predictor we have $\sigma^2_\infty = 1$, the variance of the innovation process.

We computed the information rates for first- and second-order open-loop predictors designed with different cost functions: maximizing information rate, minimizing MSRE, and minimizing PE. We also computed the information rate of the closed-loop predictor according to (2). The filter coefficients were chosen to minimize the PE σ^2_e, hence are optimal for each of the three cost functions.

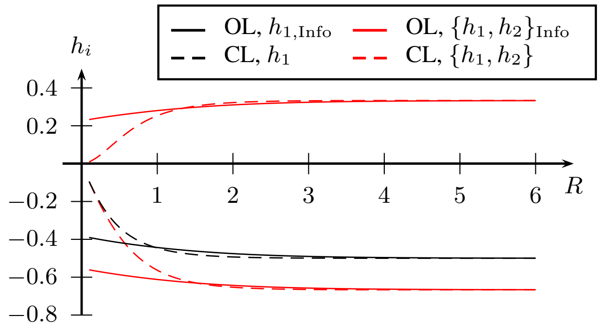


Fig. 2: Filter coefficients maximizing the information rate as a function of the entropy constraint R.

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[3]; 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). Fig. 2 shows the coefficients as a function of R. It can be seen that for high rates the information-theoretic optimum approaches the PE-optimal predictor coefficients in (4). Thus, for large R, 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 $\sigma^2_e \approx \sigma^2_{\tilde{X}}$. For small R, the coefficients are close to zero to reduce the noise gain of the prediction filter.

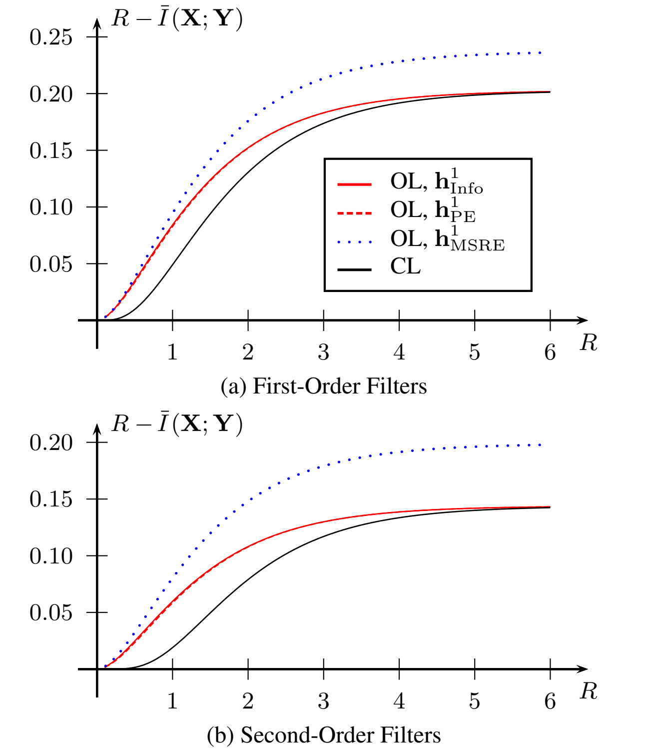
Fig. 3 shows the difference between the entropy constraint R and the information rate $\bar{I}(\mathbf{X};\mathbf{Y})$ for first- and second-order filters designed according to different cost functions. Second-order filters clearly outperform the first-order filters, and for the infinite-order open-loop predictors $\mathbf{h}^\infty_{\mathrm{Info}}$ and $\mathbf{h}^\infty_{\mathrm{PE}}$ 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. 1 in order to minimize the MSRE, $\mathbf{h}_{\mathrm{MSRE}}$, 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 (1) and (2) with small γ, and consider that the entropy powers of $S_X(e^{\jmath\theta})$ and of $S_{\tilde{X}}(e^{\jmath\theta})$ are both equal to σ^2_∞.

We present an additional example where $\mathbf{X}$ is the sum of a Gaussian signal process $\mathbf{S}$ with PSD $S_S(e^{\jmath\theta}) = 4\cos^2(\theta/2)$ and an independent, white Gaussian noise process $\mathbf{N}$ with variance σ^2_N. We design a first-order filter H with impulse response vector $\mathbf{h}^1 = [h_0,\ h_1]^T$ in order to maximize the information rate between $\mathbf{S}$ and the quantizer output, $\bar{I}(\mathbf{S};\mathbf{Y})$, and compare the solution with the filter minimizing the variance of $\mathbf{S} - \mathbf{Y}$. The latter filter can be shown to have coefficients [24, Ch. 2]

$$\mathbf{h}^1_{\text{Wiener}} = \frac{1}{(1+\gamma)(\sigma^4_N + 4\sigma^2_N + 3)}[3 + 2\sigma^2_N,\ \sigma^2_N]^T. \quad (5)$$

Note that for $R \to \infty$ and $\gamma \to 0$ the coefficient vector $\mathbf{h}^1_{\text{Wiener}}$ corresponds to the Wiener solution, hence the subscript. The Wiener

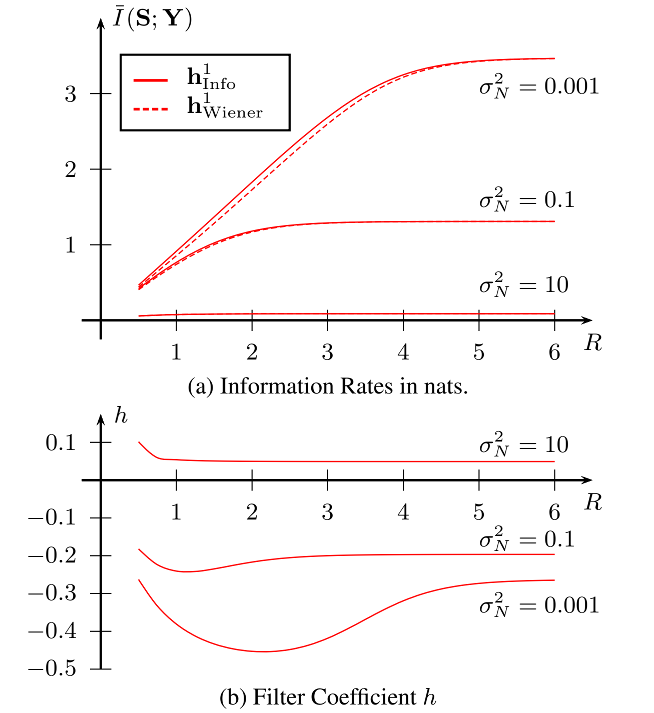
[3] Numerical results have been obtained with Wolfram Mathematica® 8; the notebooks and produced .dat-files for all figures are available for download at http://www.spsc.tugraz.at/biblio/geiger99982844.



(a) First-Order Filters

(b) Second-Order Filters

Fig. 3: Difference between entropy constraint R and information rate $\bar{I}(\mathbf{X};\mathbf{Y})$ in nats for different filters H for open- (OL) and closed-loop (CL) predictors. Note that the rates for $\mathbf{h}_{\text{Info}}$ and $\mathbf{h}_{\text{PE}}$ almost coincide.



(a) Information Rates in nats.

(b) Filter Coefficient h

Fig. 4: Analysis of the signal-plus-noise example. (a) shows the information rate as a function of the entropy constraint R for both the Wiener filter and the information-theoretically optimal filter, while (b) shows the filter coefficient of the latter.

filter would also be optimal if the quantization noise variance would not depend on the quantizer input.

To maximize the relevant information rate $\bar{I}(\mathbf{S};\mathbf{Y})$, we can again normalize the first filter coefficient to unity. Hence, with $H(e^{\jmath\theta})$ being the frequency response of the FIR filter $\mathbf{h}^1 = [1,\ h]^T$, we strive to maximize

$$\bar{I}(\mathbf{S};\mathbf{Y}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\left(1 + \frac{S_S(e^{\jmath\theta})|H(e^{\jmath\theta})|^2}{\sigma_N^2|H(e^{\jmath\theta})|^2 + \gamma\sigma_{\tilde{X}}^2}\right) d\theta. \quad (6)$$

Note that for $\sigma_N^2 = 0$ this degenerates to (1). For $\gamma \to 0$ filtering becomes futile, because in this case no quantizer is present.

We optimized this cost function numerically, and the results for the filter coefficient h and the resulting information rate are shown in Fig. 4. The variance of the noise process $\mathbf{N}$ was varied in $\sigma_N^2 = \{0.001, 0.1, 10\}$. It can be seen from Fig. 4a, that for large entropy constraints R the information rates saturate, a consequence of the noise added at the input of the quantizer. 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 optimal in the information-theoretic sense in Fig. 4b, one can see that the dependence on the entropy constraint R is now much more emphasized. First of all, for small noise variances, the filter is still predicting due to the negativity of h. However, h is now larger (i.e., closer to zero) compared to the coefficient in Fig. 2. The reason is that maximizing (6) involves three conflicting goals: equalizing $S_S(e^{\jmath\theta})$, minimizing $\sigma_N^2|H(e^{\jmath\theta})|^2$, 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 $\sigma_N^2|H(e^{\jmath\theta})|^2$ is important. This is also evident from Fig. 4b 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 $S_S(e^{\jmath\theta})$.

This example shows that, in general, energetic filter design is sub-optimal from an information-theoretic point-of-view. Being informed about those special cases in which these two cost functions coincide is of great practical importance, since energetic design often admits closed-form solutions, while information-theoretic cost functions require nonlinear optimization. Investigations, such as the present one, can hence justify – or reject – simplified design procedures from an information-theoretic perspective.

6. CONCLUSION

In this work, we highlighted the importance of information-theoretic cost functions: Information is inherently different from energy, thus a filter designed according to an energetic cost function generally fails to maximize information rates. Whether this is beneficial or not strongly depends on the application; in many cases, however, one wants to transmit *information*.

We then justified energetically optimal open-loop prediction from an information-theoretic point-of-view: While in general the filter maximizing the information rate over the quantizer is different from the filter minimizing the prediction error, the difference becomes negligible for high-resolution quantizers. Moreover, in information-theoretic terms, the performance of the open-loop predictor is similar to the closed-loop predictor's for high quantizer resolutions.

7. REFERENCES

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

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

[3] 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.

[4] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[5] G. Zeitler, A. C. Singer, and G. Kramer, "Low-precision A/D conversion for maximum information rate in channels with memory," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2511–2521, Sept. 2012.

[6] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, pp. 3629–3639, Dec. 2009.

[7] 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, July 2000.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, Hoboken, NJ, 2 edition, 2006.

[9] K. T. Kim and T. Berger, "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)*, Seattle, WA, July 2006, pp. 808–812.

[10] 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)*, Adelaide, Sept. 2005, pp. 209–213.

[11] 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, July 2008.

[12] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*, Springer, New York, NY, 1996.

[13] B. C. Geiger and G. Kubin, "Relative information loss in the PCA," in *Proc. IEEE Information Theory Workshop (ITW)*, Lausanne, Sept. 2012, pp. 562–566, extended version available: `arXiv:1204.0429 [cs.IT]`.

[14] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, Mar. 1988.

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

[16] B. C. Geiger and G. Kubin, "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]`.

[17] 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.

[18] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1766–1777, Aug. 1995.

[19] B. C. Geiger and G. Kubin, "Information loss and anti-aliasing filters in multirate systems," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Zürich, Feb. 2014, pp. 148–151, extended version available: `arXiv:1310.8487 [cs.IT]`.

[20] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

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

[22] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, Chichester, 2006.

[23] P. Noll, "On predictive quantization schemes," *Bell System Technical Journal*, vol. 57, no. 5, pp. 1499–1532, May 1978.

[24] S. Haykin, *Adaptive Filter Theory*, Pearson, Upper Saddle River, NJ, 5 edition, 2014.