ON THE RELATION BETWEEN THE GAUSSIAN INFORMATION BOTTLENECK AND MSE-OPTIMAL RATE-DISTORTION QUANTIZATION

Michael Meidlinger, Andreas Winkelbauer, and Gerald Matz

Institute of Telecommunications, Vienna University of Technology
Gusshausstrasse 25/389, 1040 Vienna, Austria
email: {michael.meidlinger,andreas.winkelbauer,gerald.matz}@nt.tuwien.ac.at

ABSTRACT

We use the Gaussian information bottleneck (GIB) to investigate the optimal rate-information trade-off for signal compression in linear Gaussian models and we provide a novel interpretation of the GIB in terms of the eigendecomposition of the Wiener filter. We further study mean-square-error-optimal rate-distortion compression preceded by a linear filter. Choosing this filter as square root of the Wiener filter is shown to be rate-information optimal. Finally, we extend our results to jointly stationary Gaussian random processes.

Index Terms— Gaussian information bottleneck, rate-distortion theory, Wiener filtering, channel output compression

1. INTRODUCTION

Rate-distortion (RD) theory characterizes the ultimate trade-off between compression and distortion in source coding. A different approach is taken by the information bottleneck method (IBM) [1], which replaces signal distortion as fidelity measure with the mutual information between the compressed source and a relevance variable. The IBM has been successfully applied to various problems in machine learning [2], computer vision [3], biomedical signal processing [4], and communications [5,6]. It is also inherently better suited than RD quantization for channel output compression in a communication system [7].

In this paper, we study the rate-information trade-off obtained with IBM and RD quantization for the case of jointly Gaussian vectors. More specifically, we show that the Gaussian information bottleneck (GIB) [8], which achieves the optimal trade-off, is closely related to minimum mean-square error (MSE) estimation. Furthermore, we show that the optimal GIB trade-off can also be accomplished by linear filtering followed by MSE-optimal source coding. Somewhat surprisingly, the optimal linear filter here is given by the square root of the Wiener filter. This is in contrast to the result of Sakrison [9], who showed that for noisy Gaussian source coding problems with MSE distortion the optimal filter is a Wiener filter. Our results also explain why direct MSE-optimal source coding (i.e., without filtering) in general does not achieve the optimal rate-information trade-off (as observed in [7]). Finally, we extend our results to the case of jointly stationary Gaussian random processes.

The equivalence of the GIB and MSE-optimal source coding with prefiltering is practically important because it implies that RD coding theorems directly apply to the GIB.Furthermore, this equivalence allows existing quantizer designs for MSE-optimal quantization to be reused for rate-information-optimal quantization.

The remainder of this paper is organized as follows. Section 2 introduces the problem setup considered in this work. In Section 3, we review the IBM and the closed-form solution for the GIB. Section 4 explores the relation between the GIB and MSE-optimal quantization. In Section 5, we generalize our results to stationary Gaussian processes. Conclusions are provided in Section 6.

Notation: We use boldface uppercase and lowercase letters for matrices and vectors, respectively, and upright sans-serif letters for random quantities. We denote expectation by $\mathcal{E}\{\cdot\}$ and the identity matrix by $\boldsymbol{I}$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ is shorthand for a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$. We use $[x]^+ \triangleq \max\{0, x\}$, $\log^+ x \triangleq [\log x]^+$, and we denote an $N \times N$ diagonal matrix with diagonal elements a_i by $\mathrm{diag}\{a_i\}_{i=1}^N$. All logarithms are to base 2.

Funding by WWTF Grant ICT12-054.

2. PROBLEM SETUP

We consider the linear model

$$\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ is a Gaussian random vector distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\mathbf{x}})$ and $\boldsymbol{H} \in \mathbb{R}^{M \times N}$ is a deterministic matrix. Furthermore, $\mathbf{n} \in \mathbb{R}^M$ is independent of $\mathbf{x}$ with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\mathbf{n}})$. Thus, $\boldsymbol{C}_{\mathbf{y}} = \boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C}_{\mathbf{n}}$. Our interest in (1) is rooted in communications (where $\boldsymbol{H}$ and $\mathbf{n}$ represent channel and additive Gaussian noise, respectively); however, due to [10, Theorem 4.5.5], any two zero-mean, jointly Gaussian random vectors $\mathbf{x}$ and $\mathbf{y}$ can be represented as in (1). Thus, all results presented in this paper hold for this general case.

Our goal is to find the optimum compression $\mathbf{z}$ of $\mathbf{y}$, characterized by the conditional distribution $p(\mathbf{z}|\mathbf{y})$, which has minimum compression rate while preserving as much information about $\mathbf{x}$ as possible. This trade-off is characterized by the information-rate function or, equivalently, by its inverse, the rate-information function. In the following, $I(\mathbf{x};\mathbf{z})$ denotes the mutual information of $\mathbf{x}$ and $\mathbf{z}$ [11].

Definition 1 *Let $\mathbf{x} - \mathbf{y} - \mathbf{z}$ be a Markov chain. The information-rate function $I\colon \mathbb{R}_+ \to \big[0, I(\mathbf{x};\mathbf{y})\big]$ is defined as*

$$I(R) \triangleq \max_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{x};\mathbf{z}) \quad \textit{subject to} \quad I(\mathbf{y};\mathbf{z}) \leq R; \tag{2}$$

the rate-information function $R\colon \big[0, I(\mathbf{x};\mathbf{y})\big] \to \mathbb{R}_+$ is defined as

$$R(I) \triangleq \min_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{y};\mathbf{z}) \quad \textit{subject to} \quad I(\mathbf{x};\mathbf{z}) \geq I. \tag{3}$$

3. OPTIMAL RATE-INFORMATION TRADE-OFF

3.1. IBM and GIB

The IBM considers the Markov chain $\mathsf{x} - \mathsf{y} - \mathsf{z}$, where x is the relevance variable, y is an observation, and z is a compressed representation of y. The joint statistics between x and y are assumed to be known. The method then solves the variational problem

$$\min_{p(\mathsf{z}|\mathsf{y})} I(\mathsf{y};\mathsf{z}) - \beta I(\mathsf{x},\mathsf{z}) \tag{4}$$

over all stochastic mappings $p(\mathsf{z}|\mathsf{y})$ of y to z. The parameter β in (4) trades compression rate $I(\mathsf{y};\mathsf{z})$ against relevant information $I(\mathsf{x},\mathsf{z})$. Initially, the IBM was considered only for discrete random variables [1]; here, a solution to (4) can be obtained only numerically via an iterative algorithm. In [8], a closed-form solution for the case where the relevance variable $\mathbf{x} \in \mathbb{R}^N$ and the observation $\mathbf{y} \in \mathbb{R}^M$ are jointly Gaussian random vectors was derived. The key observation here is that the optimal mapping is of the form

$$\mathbf{z} = \boldsymbol{A}\mathbf{y} + \boldsymbol{\xi}\,, \tag{5}$$

where $\boldsymbol{A}$ is a deterministic matrix and $\boldsymbol{\xi}$ is an $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$-distributed random vector that is independent of $\mathbf{x}$ and $\mathbf{y}$. This implies that the compressed random vector $\mathbf{z}$ is again jointly Gaussian with $\mathbf{x}$ and $\mathbf{y}$. The matrix $\boldsymbol{A}$ is completely determined by the auto- and cross-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\boldsymbol{C}_{\mathbf{x}}$, $\boldsymbol{C}_{\mathbf{y}}$, and $\boldsymbol{C}_{\mathbf{x},\mathbf{y}}$, respectively. For prescribed β, $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \operatorname{diag}\{\alpha_i\}_{i=1}^M \boldsymbol{V}^{\mathrm{T}}, \quad \alpha_i = \sqrt{\frac{[\beta(1-\lambda_i) - 1]^+}{\lambda_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{C}_{\mathbf{y}} \boldsymbol{v}_i}}. \tag{6}$$

Here, $\boldsymbol{V} = \big[\boldsymbol{v}_1 \cdots \boldsymbol{v}_M\big]$, and $\boldsymbol{v}_i^{\mathrm{T}}$ and $\lambda_i \geq 0$ are the left eigenvectors and corresponding eigenvalues of the matrix

$$\overline{\boldsymbol{W}} = \boldsymbol{C}_{\mathbf{y}|\mathbf{x}} \boldsymbol{C}_{\mathbf{y}}^{-1} = \boldsymbol{I} - \boldsymbol{C}_{\mathbf{x},\mathbf{y}}^{\mathrm{T}} \boldsymbol{C}_{\mathbf{x}}^{-1} \boldsymbol{C}_{\mathbf{x},\mathbf{y}} \boldsymbol{C}_{\mathbf{y}}^{-1}. \tag{7}$$

3.2. GIB and Wiener Filter

We next provide a novel reformulation and interpretation of the GIB for the linear model (1). Since here $\boldsymbol{C}_{\mathbf{y}|\mathbf{x}} = \boldsymbol{C}_{\mathbf{n}}$, the matrix $\overline{\boldsymbol{W}}$ in (7) can be shown to equal $\overline{\boldsymbol{W}} = \boldsymbol{C}_{\mathbf{n}}\big(\boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C}_{\mathbf{n}}\big)^{-1}$, which is seen to be the MSE-optimal Wiener filter for estimating $\mathbf{n}$ from $\mathbf{y}$. Furthermore, $\overline{\boldsymbol{W}} = \boldsymbol{I} - \boldsymbol{W}$, where

$$\boldsymbol{W} = \boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}}\big(\boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C}_{\mathbf{n}}\big)^{-1}$$

is the Wiener filter for estimating $\boldsymbol{H}\mathbf{x}$ from $\mathbf{y}$, i.e., it minimizes the MSE $\mathcal{E}\{\|\boldsymbol{W}\mathbf{y} - \boldsymbol{H}\mathbf{x}\|^2\}$. Note that $\boldsymbol{W}$ has the same left eigenvectors $\boldsymbol{v}_i^{\mathrm{T}}$ as $\overline{\boldsymbol{W}}$ and its eigenvalues are given by $\mu_i = 1 - \lambda_i$. The fact that the GIB matrix $\boldsymbol{A}$ in (6) involves the square root of λ_i and μ_i already hints at the relevance of the square-root Wiener filter in this context.

We next calculate the information-rate function for (1). In order to simplify the analytical treatment, we whiten and decorrelate the observation $\mathbf{y}$. The whitened vector $\tilde{\mathbf{y}} = \boldsymbol{C}_{\mathbf{n}}^{-1/2}\mathbf{y}$ has covariance $\boldsymbol{C}_{\tilde{\mathbf{y}}} = \boldsymbol{S} + \boldsymbol{I}$ with the signal-to-noise (SNR) matrix

$$\boldsymbol{S} = \boldsymbol{C}_{\mathbf{n}}^{-1/2} \boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}} \boldsymbol{C}_{\mathbf{n}}^{-1/2}. \tag{8}$$

Using the eigendecomposition

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^{\mathrm{T}} \quad \text{with } \boldsymbol{\Gamma} = \operatorname{diag}\{\gamma_i\}_{i=1}^M\,,$$

it follows that the elements of

$$\mathbf{y}' = \boldsymbol{U}^{\mathrm{T}}\tilde{\mathbf{y}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{n}}^{-1/2}\mathbf{y} \tag{9}$$

are uncorrelated with covariance $\boldsymbol{C}_{\mathbf{y}'} = \boldsymbol{\Gamma} + \boldsymbol{I}$. Note that $\boldsymbol{S}$ is symmetric and positive semi-definite and hence $\boldsymbol{U}$ is orthonormal and $\gamma_i \geq 0$, i.e., the mode SNRs are nonnegative.

We next derive the optimum rate-information trade-off in terms of the whitened and decorrelated vector $\mathbf{y}'$. This exploits the fact that $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{n}}^{-1/2}$ is invertible and hence the whitening and decorrelation has no effect on the mutual information, i.e., $I(\mathbf{y};\mathbf{z}) = I(\mathbf{y}';\mathbf{z})$. The Wiener filters in the whitened domain read

$$\begin{aligned}\widetilde{\overline{\boldsymbol{W}}} &= \boldsymbol{C}_{\mathbf{n}}^{-1/2}\overline{\boldsymbol{W}}\boldsymbol{C}_{\mathbf{n}}^{1/2} = (\boldsymbol{S} + \boldsymbol{I})^{-1} \\ &= \boldsymbol{U}\big(\boldsymbol{\Gamma} + \boldsymbol{I}\big)^{-1}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}\operatorname{diag}\{\lambda_i\}_{i=1}^M \boldsymbol{U}^{\mathrm{T}},\end{aligned}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{W}} &= \boldsymbol{C}_{\mathbf{n}}^{-1/2}\boldsymbol{W}\boldsymbol{C}_{\mathbf{n}}^{1/2} = \boldsymbol{S}(\boldsymbol{S} + \boldsymbol{I})^{-1} \\ &= \boldsymbol{U}\boldsymbol{\Gamma}(\boldsymbol{\Gamma} + \boldsymbol{I})^{-1}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}\operatorname{diag}\{\mu_i\}_{i=1}^M \boldsymbol{U}^{\mathrm{T}},\end{aligned} \tag{10}$$

where we used $\lambda_i = 1/(\gamma_i + 1)$ and $\mu_i = \gamma_i/(\gamma_i + 1)$. Furthermore, these expressions reveal that $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{n}}^{-1/2}$. It follows that

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{y}}\boldsymbol{V} = \boldsymbol{\Gamma} + \boldsymbol{I} = \operatorname{diag}\big\{\lambda_i^{-1}\big\}_{i=1}^M$$

and hence (cf. (6))

$$\boldsymbol{A} = \operatorname{diag}\{\alpha_i\}_{i=1}^M \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{n}}^{-1/2}, \quad \alpha_i = \sqrt{[\beta\mu_i - 1]^+}. \tag{11}$$

The parameter β in the variational problem (4) thus restricts the active modes to those with mode SNR $\gamma_i > 1/(\beta - 1)$ (equivalently, $\mu_i > 1/\beta$). We can now formulate the following result.

Theorem 1 *The optimum rate-information trade-off for* (1) *is characterized by the parametric equations*

$$I(\beta) = \frac{1}{2}\sum_{i=1}^M \log^+\!\left(\frac{\beta - 1}{\beta}\big(1 + \gamma_i\big)\right), \tag{12}$$

$$R(\beta) = \frac{1}{2}\sum_{i=1}^M \log^+\!\big((\beta - 1)\gamma_i\big), \tag{13}$$

where each choice of the parameter $\beta \in (1, \infty)$ *corresponds to a point on the rate-information and information-rate function.*

Proof: Due to the joint Gaussianity of all vectors, we have [11]

$$I(\mathbf{y};\mathbf{z}) = I(\mathbf{y}';\mathbf{z}) = \frac{1}{2}\log\det \boldsymbol{C}_{\mathbf{z}}\boldsymbol{C}_{\mathbf{z}|\mathbf{y}'}^{-1}, \tag{14}$$

$$I(\mathbf{x};\mathbf{z}) = \frac{1}{2}\log\det \boldsymbol{C}_{\mathbf{z}}\boldsymbol{C}_{\mathbf{z}|\mathbf{x}}^{-1}. \tag{15}$$

The result follows by inserting into these expressions the covariance matrices

$$\begin{aligned}\boldsymbol{C}_{\mathbf{z}} &= \boldsymbol{A}\boldsymbol{C}_{\mathbf{y}}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{I} = \operatorname{diag}\{\alpha_i\}_{i=1}^M (\boldsymbol{\Gamma} + \boldsymbol{I})\operatorname{diag}\{\alpha_i\}_{i=1}^M + \boldsymbol{I}, \\ &= \operatorname{diag}\big\{\alpha_i^2(\gamma_i + 1) + 1\big\}_{i=1}^M\,,\end{aligned}$$

$\boldsymbol{C}_{\mathbf{z}|\mathbf{y}'} = \boldsymbol{I}$, and $\boldsymbol{C}_{\mathbf{z}|\mathbf{x}} = \boldsymbol{A}\boldsymbol{C}_{\mathbf{n}}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{I} = \operatorname{diag}\big\{\alpha_i^2 + 1\big\}_{i=1}^M$. ■

4. GIB VERSUS RD-OPTIMAL COMPRESSION

4.1. Linear Filtering and RD Quantization

It has been observed in [7] that MSE-optimal RD quantization in general does not achieve the optimal rate-information trade-off. This can be explained by the fact that the GIB exploits the joint statistics of $\mathbf{x}$ and $\mathbf{y}$, whereas RD-optimal quantization uses only the statistics of $\mathbf{y}$. We demonstrate below that extracting the part of $\mathbf{y}$ most relevant for $\mathbf{x}$ requires linear filtering prior to RD quantization. We note that [9] showed that noisy source coding, i.e., minimizing the compression rate subject to a constraint on the MSE between the source and the quantizer output,

$$\min_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{y};\mathbf{z}) \quad \text{subject to} \quad \mathcal{E}\{\|\mathbf{z}-\mathbf{x}\|^2\} \leq D,$$

leads to MSE-optimal RD quantization of the Wiener filter output.

We next investigate MSE-optimal RD quantization preceded by a filter in the whitened and decorrelated domain, i.e., we consider (cf. (9))

$$\mathbf{w} = \boldsymbol{F}\mathbf{y}' \sim \mathcal{N}\Big(\mathbf{0}, \operatorname{diag}\{f_i^2(1+\gamma_i)\}_{i=1}^M\Big) \tag{16}$$

where $\boldsymbol{F} = \operatorname{diag}\{f_i\}_{i=1}^M$, and we solve

$$\min_{p(\mathbf{z}|\mathbf{w})} I(\mathbf{w};\mathbf{z}) \quad \text{subject to} \quad \mathcal{E}\{\|\mathbf{z}-\mathbf{w}\|^2\} \leq D.$$

While the RD trade-off for MSE-optimal source coding is well understood, we next assess the associated rate-information trade-off (recall that the relevant information equals $I(\mathbf{x};\mathbf{z})$).

Theorem 2 *The rate-information trade-off for MSE-optimal quantization of the filtered vector* $\mathbf{w}$ *is characterized by*

$$I(\vartheta,\boldsymbol{F}) = \frac{1}{2}\sum_{i=1}^M \log^+\!\left(\frac{1+\gamma_i}{1+\vartheta\frac{\gamma_i}{f_i^2(1+\gamma_i)}}\right), \tag{17}$$

$$R(\vartheta,\boldsymbol{F}) = \frac{1}{2}\sum_{i=1}^M \log^+\!\left(\frac{f_i^2(1+\gamma_i)}{\vartheta}\right). \tag{18}$$

Here, the waterlevel parameter $\vartheta \in [0,\infty)$ *is determined by the distortion* D.

Proof: The expression (18) for the rate $R(\vartheta,\boldsymbol{F}) = I(\mathbf{w};\mathbf{z})$ follows from the inverse waterfilling argument [11, Section 13.3] applied to the filtered vector $\mathbf{w}$. The relevant information $I(\vartheta,\boldsymbol{F}) = I(\mathbf{x};\mathbf{z})$ in (17) is calculated similarly as in (15), except that the mapping (5), which is required to compute the covariance matrices, is replaced by the "forward quantization channel" in [10, p.101]. ■

Eliminating the waterlevel ϑ from (17) and (18) yields an explicit relation between relevant information $I(\vartheta,\boldsymbol{F})$ and compression rate $R(\vartheta,\boldsymbol{F})$. Assuming that the variances $\omega_i \triangleq f_i^2(1+\gamma_i)$, $i=1,\ldots,M$, are sorted in descending order, we obtain

$$I_{\boldsymbol{F}}(R) = \frac{1}{2}\sum_{i=1}^M \log\frac{1+\gamma_i}{1+2^{-2R_i(R,\boldsymbol{F})}\gamma_i},$$

where the rate allocated to mode i is given by

$$R_i(R,\boldsymbol{F}) = \left[\frac{R}{l(R,\boldsymbol{F})} + \frac{1}{2}\log\frac{\omega_i}{\overline{\omega}_{l(R,\boldsymbol{F})}}\right]^+.$$

Here, $\overline{\omega}_l \triangleq \prod_{i=1}^l \omega_i^{1/l}$ is the geometric mean of $\omega_1,\ldots,\omega_l$ and $l(R,\boldsymbol{F}) = \max\{i : R_{\mathrm{c},i}(\boldsymbol{F}) \leq R\}$ denotes the number of active modes, which increases at the critical rates

$$R_{\mathrm{c},i}(\boldsymbol{F}) = \frac{1}{2}\sum_{k=1}^i \log\frac{\omega_k}{\omega_i}. \tag{19}$$

Direct MSE-optimal quantization of $\mathbf{y}$ corresponds to $\boldsymbol{F} = \boldsymbol{I}$ (i.e., no filtering) and noisy source coding [9] corresponds to $\boldsymbol{F} = \boldsymbol{F}_{\mathrm{W}} = \boldsymbol{\Gamma}(\boldsymbol{I}+\boldsymbol{\Gamma})^{-1}$ (i.e., Wiener filtering). Surprisingly, these two approaches in general are suboptimal in terms of rate-information trade-off. We next identify the uniformly rate-information optimum filter $\boldsymbol{F}_\star$ that satisfies $I_{\boldsymbol{F}_\star}(R) \geq I_{\boldsymbol{F}}(R)$ for all $\boldsymbol{F}$ and any R.

Theorem 3 *The optimum filter* $\boldsymbol{F}_\star$ *is given by the square root of the Wiener filter (cf.* (10)*),*

$$\boldsymbol{F}_\star = \boldsymbol{F}_{\mathrm{W}}^{1/2} = \boldsymbol{\Gamma}^{1/2}(\boldsymbol{I}+\boldsymbol{\Gamma})^{-1/2} = \operatorname{diag}\{\sqrt{\mu_i}\}_{i=1}^M \tag{20}$$

and achieves the same rate-information trade-off as the GIB.

Proof: The claim follows from observing that with $\boldsymbol{F} = \boldsymbol{F}_\star$ and $\vartheta = 1/(\beta-1)$, (17) and (18) coincide with the optimal GIB trade-off (12) and (13), respectively (recall that $\mu_i = \gamma_i/(\gamma_i+1)$). ■

Lemma 1 *The number of active modes satisfies*

$$l(R,\boldsymbol{I}) \,\geq\, l(R,\boldsymbol{F}_\star) \,\geq\, l(R,\boldsymbol{F}_{\mathrm{W}}), \tag{21}$$

which in turn is equivalent to

$$R_{\mathrm{c},i}(\boldsymbol{I}) \,\leq\, R_{\mathrm{c},i}(\boldsymbol{F}_\star) \,\leq\, R_{\mathrm{c},i}(\boldsymbol{F}_{\mathrm{W}}). \tag{22}$$

The critical rates are furthermore related as

$$R_{\mathrm{c},i}(\boldsymbol{F}_\star) = \frac{R_{\mathrm{c},i}(\boldsymbol{I}) + R_{\mathrm{c},i}(\boldsymbol{F}_{\mathrm{W}})}{2}. \tag{23}$$

Proof: The expression (23) can be verified directly from (19). The left-hand side inequality in (22) follows from [7, Lemma 9] which together with (23) implies the righ-hand side inequality. The double inequality (21) follows from the definition of $l(R,\boldsymbol{F})$ in terms of the critical rates. ■

4.2. Discussion and Illustration

We note that any scaled version of $\boldsymbol{F}_\star$ is also rate-information optimal. If the nonzero mode SNRs are identical, i.e., if $\gamma_i \in \{\gamma, 0\}$, then we have $\boldsymbol{F}_{\mathrm{W}} = \sqrt{\gamma/(\gamma+1)}\boldsymbol{F}_\star$ and hence in this case MSE-optimal noisy source coding is rate-information optimal. However, for widely different mode SNRs γ_i, $\boldsymbol{F}_{\mathrm{W}}$ and other suboptimum filters perform substantially worse. In particular, the performance loss

$$\Delta I_{\boldsymbol{F}}(R) \triangleq I_{\boldsymbol{F}_\star}(R) - I_{\boldsymbol{F}}(R) = \frac{1}{2}\sum_{i=1}^M \log\frac{1+2^{-2R_i(R,\boldsymbol{F})}\gamma_i}{1+2^{-2R_i(R,\boldsymbol{F}_\star)}\gamma_i}$$

of any filter $\boldsymbol{F}$ can be bounded as

$$\Delta I_{\boldsymbol{F}}(R) \leq \frac{1}{2}\sum_{i=1}^M \log(1+\gamma_i) - \frac{1}{2}\log\frac{f_1^2(1+\gamma_1)^2}{f_1^2(1+\gamma_1) + f_2^2\gamma_1(1+\gamma_2)}.$$

We next consider the filters $\boldsymbol{F}(n) = \boldsymbol{F}_{\mathrm{W}}^n = \operatorname{diag}\{\mu_i^n\}_{i=1}^M$ to illustrate the transition from the unfiltered case ($n=0$) to rate-information optimal filtering ($n=1/2$) and Wiener filtering ($n=1$). We assume $M=10$ and mode SNRs $\gamma_i = 2^{-ci}$, $i=1,\ldots,M$, with c chosen such that $C = \frac{1}{2}\sum_{i=1}^M \log(1+\gamma_i) = 1$. Fig. 1 shows the information-rate curve $I_{\boldsymbol{F}(n)}(R)$ for various n. Direct quantization without filtering ($n=0$) is seen to perform worst among the curves shown because it uses too many modes and allocates too little rate to the strongest modes (cf. Lemma 1). As n increases, the information-rate trade-off improves and is identical to the GIB optimum for $n=1/2$. Increasing n beyond $1/2$ deteriorates the rate-information performance. Noisy source coding with Wiener filtering ($n=1$) performs slightly poorer than the optimal solution since according to Lemma 1 too few modes are used, i.e., too much rate is allocated to the strongest modes. Interestingly, the information-rate curve is no longer concave for $n > 1/2$. Finally, we note that in general, the relative order in terms of information-rate performance for various n depends on the distribution of the mode SNRs γ_i.

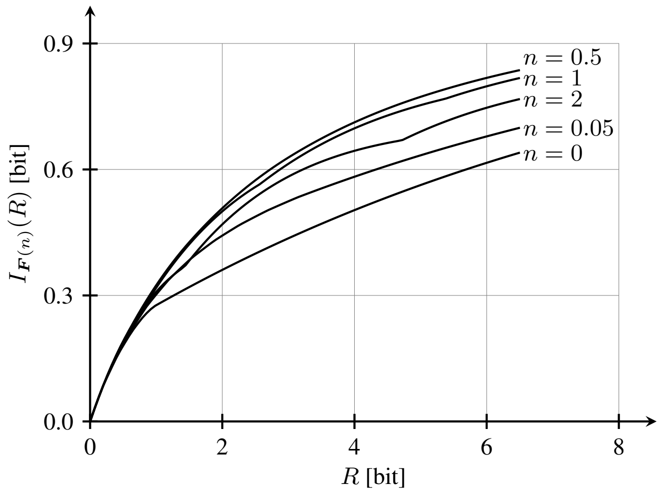

Fig. 1. Information-rate curve $I_{\boldsymbol{F}(n)}(R)$ for various n.

5. EXTENSION TO STATIONARY RANDOM PROCESSES

We next briefly outline the extension of our results to the case where $\mathsf{x}[k]$ and $\mathsf{n}[k]$ are independent stationary Gaussian processes with power spectral densities (PSDs) $S_{\mathsf{x}}(\theta)$ and $S_{\mathsf{n}}(\theta)$ and $\mathsf{y}[k] = \sum_{k'=-\infty}^{\infty} h[k']\,\mathsf{x}[k-k'] + \mathsf{n}[k]$, with $h[k]$ the impulse response of a linear time-invariant filter. For a finite time interval of duration N, this model reduces to (1) with $\boldsymbol{H}$ a Toeplitz matrix induced by $h[k]$ and all covariance matrices being Toeplitz as well.

We can then obtain asymptotic frequency-domain versions of all results derived above by using mutual information rate $\mathcal{I}(\mathsf{x}, \mathsf{y}) \triangleq \lim_{N\to\infty} \frac{1}{N} I(\mathbf{x}; \mathbf{y})$ and by invoking the following Lemma, whose proof is along the lines of [12, Corollary 4.1] but is omitted due to lack of space.

Lemma 2 *Consider a series of $N \times N$ Wiener-type Toeplitz matrices whose eigenvalues $\lambda_{N,k}$ have asymptotic eigenvalue spectrum $S(\theta)$ with $S(\theta) = \vartheta$ only on a set of measure zero and let $g(\cdot)$ be a continuous positive function. Then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{\lambda_{N,k}:\lambda_{N,k}>\vartheta} g(\lambda_{N,k}) = \frac{1}{2\pi} \int\limits_{\theta:S(\theta)>\vartheta} g\big(S(\theta)\big)\, \mathrm{d}\theta.$$

In particular, MSE-optimal source coding of the filtered observation $\mathsf{w}[k] = \sum_{k'=-\infty}^{\infty} f[k']\,\mathsf{y}[k-k']$ with PSD

$$S_{\mathsf{w}}(\theta) = \big|F(\theta)\big|^2 \left(\big|H(\theta)\big|^2 S_{\mathsf{x}}(\theta) + S_{\mathsf{n}}(\theta)\right),$$

leads to the rate-information trade-off (cf. (17), (18))

$$\mathcal{I}(\vartheta, F) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log^{+}\!\left(\frac{1 + \Gamma(\theta)}{1 + \vartheta \frac{\Gamma(\theta)}{|F(\theta)|^2 (1+\Gamma(\theta))}} \right) \mathrm{d}\theta,$$

$$\mathcal{R}(\vartheta, F) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log^{+}\!\left(\frac{\big|F(\theta)\big|^2 \big(1 + \Gamma(\theta)\big)}{\vartheta} \right) \mathrm{d}\theta.$$

Here, $F(\theta)$ and $H(\theta)$ denote the frequency responses of the filter $f[k]$ and the channel $h[k]$ and we used the SNR spectrum

$$\Gamma(\theta) = \big|H(\theta)\big|^2 S_{\mathsf{x}}(\theta) / S_{\mathsf{n}}(\theta).$$

The optimal filter is given by

$$F_{\star}(\theta) = \sqrt{\frac{\Gamma(\theta)}{1 + \Gamma(\theta)}}.$$

6. CONCLUSION

In this work, we established the link between MSE-optimal RD compression and the GIB, proving that linearly pre-filtered RD compression is equivalent to the GIB provided that a square-root Wiener filter is used. We derived closed form expressions for calculating the ultimate Gaussian rate-information trade-off, both for random vectors and stationary processes. Our results are practically useful since they allow MSE-optimal quantizers to be used for rate-information-optimal quantization. All results presented in this work can easily be extended to the complex case.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” in *Proc. 37th Allerton Conf. Communication, Control, and Computing*, Monticello, Illinois, USA, Sept. 1999, pp. 368–377.

[2] N. Slonim and N. Tishby, “Document clustering using word clusters via the information bottleneck method,” in *Proc. 23rd Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Athens, Greece, July 2000, pp. 208–215.

[3] S. Gordon, H. Greenspan, and J. Goldberger, “Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations,” in *Proc. 9th IEEE Int. Conf. Computer Vision*, Nice, France, Oct. 2003, pp. 370–377.

[4] E. Schneidman, N. Slonim, N. Tishby, R. de Ruyter van Steveninck, and W. Bialek, “Analyzing neural codes using the information bottleneck method,” *The Hebrew University*, 2002.

[5] G. Zeitler, R. Kötter, G. Bauch, and J. Widmer, “On quantizer design for soft values in the multiple-access relay channel,” in *Proc. IEEE ICC 2009*, Dresden, Germany, June 2009.

[6] A. Winkelbauer and G. Matz, “Joint network-channel coding for the asymmetric multiple-access relay channel,” in *Proc. IEEE ICC 2012*, Ottawa, Canada, June 2012.

[7] A. Winkelbauer, S. Farthofer, and G. Matz, “The rate-information trade-off for Gaussian vector channels,” submitted to *IEEE Int. Symp. Information Theory*, 2014.

[8] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for Gaussian variables,” *J. Machine Learning Res.*, vol. 6, no. 1, pp. 165–188, 2005.

[9] D. Sakrison, “Source encoding in the presence of random disturbance,” *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 165–167, Jan. 1968.

[10] T. Berger, *Rate Distortion Theory*. Englewood Cliffs (NJ): Prentice Hall, 1971.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[12] R. Gray, *Toeplitz and Circulant Matrices: A Review*. Foundations and Trends in Technology. Now Publishers, 2006, vol. 2, pp. 155–239.