The Rate-Information Trade-off for Gaussian Vector Channels

Andreas Winkelbauer, Stefan Farthofer, and Gerald Matz

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

***Abstract*—We consider lossy compression of the output of a Gaussian vector channel. This is relevant for quantizer design in rate-limited feedback and in receiver front-ends. In particular, we study the trade-off between compression rate and mutual information between channel input and compressed channel output. Using the Gaussian information bottleneck we provide closed-form expressions for the information-rate function and the rate-information function. We prove that optimal compression in the rate distortion sense with squared-error distortion does not achieve the optimal rate-information trade-off. The suboptimality of the rate distortion approach is quantified and we give an upper bound on the gap to the optimal rate-information trade-off. Finally, our results are corroborated by numerical examples.**

***Index Terms*—quantization, source coding, rate distortion theory, information bottleneck, data compression**

I. Introduction

We study the performance limits of Gaussian vector channels under channel output compression. This extends the scalar case previously studied in [1]. Our work is motivated by communications applications like quantizer design for receiver front-ends and for rate-limited feedback. A rate distortion (RD) approach is not suited for this kind of problems since we are interested in maximizing the data rate rather than in representing the received signal with small distortion.

In this paper, we study the *information-rate function* and the *rate-information function*, i.e., the optimal trade-off between compression rate on the one hand and mutual information of channel input and compressed channel output on the other hand. We show that for Gaussian vector channels the solution to the *Gaussian information bottleneck* (GIB) [2] is rate-information optimal and we provide a reverse waterfilling interpretation for the rate allocation. Finally, we show that RD theory with squared-error distortion [3] does not achieve the optimal rate-information trade-off (even though all distributions are Gaussian) and we analyze the corresponding gap. Due to space constraints, we state all results without proof.

The remainder of this paper is organized as follows. Section II provides the required background and definitions and introduces the system model. In Section III, we review the main results from [1]. The optimal rate-information trade-off and its properties are presented in Section IV. In Section V, we prove and quantify the suboptimality of squared-error distortion-optimal compression. Conclusions are provided in Section VI.

Funding by WWTF Grant ICT12-054 and by the EU FP7 project NEWCOM# (GA 318306).

Notation: We use boldface letters for column vectors and upright sans-serif letters for random variables. The expectation operator is denoted by $\mathbb{E}\{\cdot\}$ and we follow the notation of [4] for mutual information $I(\cdot\,;\cdot)$. The identity matrix is denoted by $\boldsymbol{I}$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ denotes a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. Furthermore, $[x]^+ \triangleq \max\{0, x\}$ and $\log^+ x \triangleq [\log x]^+$. All logarithms are to base 2.

II. Background and Definitions

A. Information Bottleneck Method

The *information bottleneck method* (IBM) [5] and its variants like the GIB have received limited attention outside of the machine learning community. Therefore, we next give a brief overview of the IBM and we discuss the GIB solution, which is used to characterize rate-information optimal compression in Section IV-B. We note that for discrete random variables, [6] considers a problem that is closely related to the IBM but formulated in terms of conditional entropy.

Let $\mathsf{x} - \mathsf{y} - \mathsf{z}$ be a Markov chain, where z is a compressed representation of y and the joint distribution of x and y is known. The IBM solves the variational problem

$$\min_{p(z|y)} I(\mathsf{y};\mathsf{z}) - \beta I(\mathsf{x};\mathsf{z}). \tag{1}$$

In the context of the IBM, x is called the *relevance variable* and the mutual information $I(\mathsf{x};\mathsf{z})$ is called *relevant information*. The trade-off between compression rate $I(\mathsf{y};\mathsf{z})$ and relevant information is determined by the positive parameter β. For the case of discrete random variables, an iterative algorithm that finds a locally optimal solution of the nonconvex problem (1) is described in [5].

We next consider the case of jointly Gaussian zero-mean random vectors $\mathsf{x} \in \mathbb{R}^m$, $\mathsf{y} \in \mathbb{R}^n$, which we assume to have full rank covariance matrices. Here, the original IBM algorithm cannot be used. However, it has been shown in [7] that in this case the optimal z is jointly Gaussian with y and can be written as

$$\mathsf{z} = \boldsymbol{A}\mathsf{y} + \boldsymbol{\xi}, \tag{2}$$

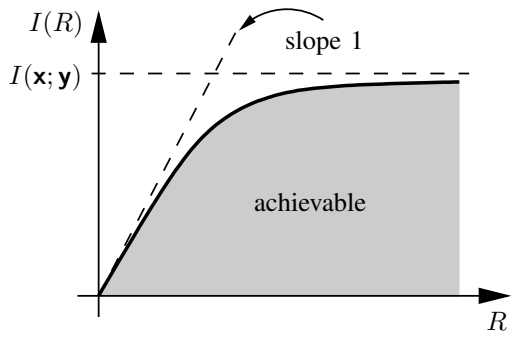


Figure 1: Illustration of the information-rate function $I(R)$.

where $\boldsymbol{A}$ is a matrix and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\xi}})$ is independent of y. Using (2), the problem in (1) can be rewritten as

$$\min_{\boldsymbol{A}, \boldsymbol{C}_{\boldsymbol{\xi}}} I(\mathsf{y};\mathsf{z}) - \beta I(\mathsf{x};\mathsf{z}). \tag{3}$$

Denote by $\boldsymbol{v}_k$ and λ_k, $k = 1, \ldots, n$, the left eigenvectors and associated eigenvalues of $\boldsymbol{C}_{\mathsf{y}|\mathsf{x}}\boldsymbol{C}_{\mathsf{y}}^{-1}$, where $\boldsymbol{C}_{\mathsf{y}} = \mathbb{E}\{\mathsf{y}\mathsf{y}^{\mathrm{T}}\}$ and $\boldsymbol{C}_{\mathsf{y}|\mathsf{x}} = \mathbb{E}\{\mathsf{y}\mathsf{y}^{\mathrm{T}}|\mathsf{x}\}$ are, respectively, the unconditional and the conditional covariance matrix of y. An optimal solution of (3) is then given by [2, Theorem 3.1]

$$\boldsymbol{A} = \mathrm{diag}\{\alpha_k\}_{k=1}^{n} \boldsymbol{V}^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{\xi}} = \boldsymbol{I}, \tag{4}$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_n]$ and

$$\alpha_k = \sqrt{\frac{[\beta(1-\lambda_k) - 1]^+}{\lambda_k \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{C}_{\mathsf{y}} \boldsymbol{v}_k}}, \qquad k = 1, \ldots, n. \tag{5}$$

Using (4) and (5), the rate-information trade-off reads

$$I(\mathsf{x};\mathsf{z}) = I(\mathsf{y};\mathsf{z}) - \frac{1}{2}\sum_{k=1}^{n} \log^+ \beta(1-\lambda_k). \tag{6}$$

B. Information-Rate Function and Rate-Information Function

We next formalize the trade-off between compression rate and relevant information.

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

$$I(R) \triangleq \max_{p(\mathsf{z}|\mathsf{y})} I(\mathsf{x};\mathsf{z}) \quad \text{subject to } I(\mathsf{y};\mathsf{z}) \le R, \tag{7}$$

and the rate-information function $R : [0, I(\mathsf{x};\mathsf{y})] \to \mathbb{R}_+$ is defined by

$$R(I) \triangleq \min_{p(\mathsf{z}|\mathsf{y})} I(\mathsf{y};\mathsf{z}) \quad \text{subject to } I(\mathsf{x};\mathsf{z}) \ge I. \tag{8}$$

$I(R)$ allows us to quantify the maximum of the relevant information that can be preserved when the compression rate is at most R. Conversely, $R(I)$ quantifies the minimum compression rate required when the retained relevant information must be at least I. The definitions (7) and (8) are structurally similar to the distortion-rate function and the rate-distortion function [3], respectively, only that the minimization (lower bound) of distortion is replaced with a maximization (upper bound) of the relevant information. We emphasize that, contrary to RD theory, the (average) "distance" between y and z is irrelevant for $I(R)$ and $R(I)$.

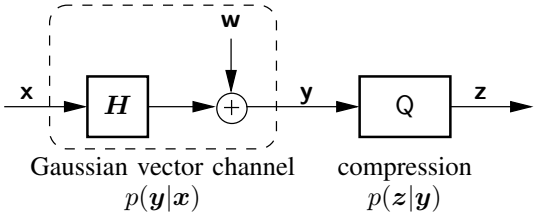


Figure 2: Gaussian vector channel with output compression.

By the data processing inequality, $I(R)$ is bounded as

$$I(R) \le \min\{R, I(\mathsf{x};\mathsf{y})\}. \tag{9}$$

Fig. 1 illustrates the information-rate function $I(R)$ (solid line) and the upper bound (9) (dashed lines). The shaded region in Fig. 1 corresponds to the achievable rate-information pairs (see [8] for details).

C. System Model

In what follows, we consider Gaussian vector channels as depicted in Fig. 2. We assume that the channel input x is zero-mean Gaussian, $\mathsf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\mathsf{x}})$. The channel output equals

$$\mathsf{y} = \boldsymbol{H}\mathsf{x} + \mathsf{w}, \tag{10}$$

where $\boldsymbol{H} \in \mathbb{R}^{n\times m}$ and the additive noise $\mathsf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ is independent of x. Hence, x, y are jointly Gaussian and

$$\boldsymbol{C}_{\mathsf{y}} = \boldsymbol{H}\boldsymbol{C}_{\mathsf{x}}\boldsymbol{H}^{\mathrm{T}} + \sigma^2\boldsymbol{I}. \tag{11}$$

Let z be a compressed representation of y. Since $\mathsf{x} - \mathsf{y} - \mathsf{z}$ is a Markov chain we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \int_{\mathbb{R}^n} p(\boldsymbol{z}|\boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\boldsymbol{y}, \tag{12}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the transition pdf of the Gaussian vector channel and $p(\boldsymbol{z}|\boldsymbol{y})$ describes the compression mapping Q. Due to the GIB, the optimal z is jointly Gaussian with y (cf. (2)). Hence, $\mathsf{x}, \mathsf{y}, \mathsf{z}$ are jointly Gaussian and $p(\boldsymbol{z}|\boldsymbol{x})$ in (12) is also a Gaussian distribution. Throughout, we fix the input distribution $p(\boldsymbol{x})$ and optimize $p(\boldsymbol{z}|\boldsymbol{y})$. Jointly optimizing $p(\boldsymbol{x})$ and $p(\boldsymbol{z}|\boldsymbol{y})$ is a harder problem and yields non-Gaussian $p(\boldsymbol{x})$.

III. The Scalar Case

Before we consider the rate-information trade-off for the vector case, we briefly summarize the main results for the scalar case $n = m = 1$ from [1]. Let $\mathsf{y} = h\,\mathsf{x}+\mathsf{w}$ be a Gaussian channel with signal-to-noise ratio (SNR) $\rho = h^2\sigma_{\mathsf{x}}^2/\sigma^2$. Here, the information-rate function equals [1, Theorem 2]

$$I(R) = R - \frac{1}{2}\log\frac{2^{2R} + \rho}{1+\rho} \tag{13}$$

$$= C(\rho) - C(2^{-2R}\rho), \tag{14}$$

where

$$C(\rho) \triangleq \frac{1}{2}\log(1+\rho) \tag{15}$$

is the capacity of the uncompressed channel. Due to the data processing inequality, $I(R) \le \min\{R, C(\rho)\}$. For small R,

the second term in (13) is small and $I(R) \approx R$. For large R, the second term in (14) is small and $I(R) \approx C(\rho)$. Since the overall channel $p(z|x)$ is Gaussian, too, we can write $I(R) = C(\hat{\rho})$ where

$$\hat{\rho} = \rho \frac{1 - 2^{-2R}}{1 + 2^{-2R}\rho} \leq \rho \tag{16}$$

is the equivalent SNR of the channel $p(z|x)$. For the rate-information function we have [1, Theorem 5]

$$R(I) = \frac{1}{2} \log \frac{\rho}{2^{-2I}(1+\rho) - 1}. \tag{17}$$

Finally, we note that RD-optimal compression of y with distortion $(z-y)^2$ achieves the rate-information trade-off.

IV. The Vector Case

A. Preliminaries

Let $\boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^{\mathrm{T}}$ be the eigendecomposition of $\boldsymbol{H}\boldsymbol{C}_{\mathsf{x}}\boldsymbol{H}^{\mathrm{T}}$ with orthonormal eigenvectors $\boldsymbol{u}_k$ and nonnegative eigenvalues γ_k, i.e., $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]$ and $\boldsymbol{\Gamma} = \operatorname{diag}\{\gamma_k\}_{k=1}^n$. We thus have $\boldsymbol{C}_{\mathsf{y}} = \boldsymbol{U}(\boldsymbol{\Gamma} + \sigma^2 \boldsymbol{I})\boldsymbol{U}^{\mathrm{T}}$ and we let

$$\widetilde{\mathbf{y}} = \boldsymbol{U}^{\mathrm{T}}\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma} + \sigma^2 \boldsymbol{I}). \tag{18}$$

Note that applying the orthogonal matrix $\boldsymbol{U}^{\mathrm{T}}$ to $\mathbf{y}$ changes neither the mutual information nor the squared-error distortion. Therefore, we work with $\widetilde{\mathbf{y}}$ instead of $\mathbf{y}$ in the following. Due to (18), the channel $p(\widetilde{\boldsymbol{y}}|\boldsymbol{x})$ decomposes into n parallel scalar channels (henceforth called *modes*) with SNRs

$$\rho_k \triangleq \frac{\gamma_k}{\sigma^2}. \tag{19}$$

Without loss of generality we assume in what follows that the eigenvalues and the SNRs are sorted in descending order, $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$.

B. Optimal Rate-Information Trade-off

In this subsection we use the GIB to find the optimal rate-information trade-off. The following theorem states a closed-form expression for the information-rate function and discusses its properties.

Theorem 2. *The information-rate function of a Gaussian vector channel with sorted mode SNRs is given by*

$$I(R) = R - \frac{1}{2} \sum_{k=1}^{n} \log \frac{2^{2R_k(R)} + \rho_k}{1 + \rho_k} \tag{20}$$

$$= \sum_{k=1}^{n} \big(C(\rho_k) - C(2^{-2R_k(R)}\rho_k) \big), \tag{21}$$

where the rate allocated to the kth mode is

$$R_k(R) = \left[\frac{R}{\ell(R)} + \frac{1}{2} \log \frac{\rho_k}{\prod_{i=1}^{\ell(R)} \rho_i^{1/\ell(R)}} \right]^+ \tag{22}$$

and the number of active modes equals

$$\ell(R) = \max \left\{ k \colon \frac{1}{2} \sum_{i=1}^{k} \log \frac{\rho_i}{\rho_k} \leq R \right\}. \tag{23}$$

$I(R)$ has the following properties:

1) *$I(R)$ is strictly concave on $\mathbb{R}_+$.*
2) *$I(R)$ is strictly increasing in R.*
3) *$I(R) \leq \min\{R, \sum_{k=1}^n C(\rho_k)\}$.*
4) *$I(0) = 0$ and $\lim_{R\to\infty} I(R) = \sum_{k=1}^n C(\rho_k)$.*
5) *$\frac{\mathrm{d}}{\mathrm{d}R} I(R) = \frac{1}{\ell(R)} \sum_{k=1}^{\ell(R)} (1 + 2^{2R_k(R)}\rho_k^{-1})^{-1}$.*

The information-rate function (20) can also be written as

$$I(R) = \sum_{k=1}^{n} \left[R_k(R) - \frac{1}{2} \log \frac{2^{2R_k(R)} + \rho_k}{1 + \rho_k} \right], \tag{24}$$

which we identify as the sum of n information-rate functions of scalar Gaussian channels with SNRs ρ_k (cf. (13)). The following corollaries are consequences of Theorem 2.

Corollary 3. *We can rewrite $I(R)$ in (20) as*

$$I(R) = \sum_{k=1}^{n} C(\hat{\rho}_k), \tag{25}$$

where the SNR of the kth mode after compression equals

$$\hat{\rho}_k = \rho_k \frac{1 - 2^{-2R_k(R)}}{1 + 2^{-2R_k(R)}\rho_k} \leq \rho_k. \tag{26}$$

Corollary 4. *Optimal channel output compression can equivalently be modeled using an additive Gaussian noise term $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\mathbf{n}})$ with*

$$\boldsymbol{C}_{\mathbf{n}} = \operatorname{diag}\left\{ \sigma^2 \frac{1 + \rho_k}{2^{2R_k(R)} - 1} \right\}_{k=1}^{n}. \tag{27}$$

We next consider the rate-information function.

Theorem 5. *The rate-information function of a Gaussian vector channel with sorted mode SNRs is given by*

$$R(I) = \sum_{k=1}^{n} \frac{1}{2} \log \frac{\rho_k}{2^{-2I_k(I)}(1+\rho_k) - 1}, \tag{28}$$

where the relevant information of the kth mode is

$$I_k(I) = \left[\frac{I}{\ell(I)} + \frac{1}{2} \log \frac{1 + \rho_k}{\prod_{i=1}^{\ell(I)} (1+\rho_i)^{1/\ell(I)}} \right]^+ \tag{29}$$

and the number of active modes equals

$$\ell(I) = \max \left\{ k \colon \frac{1}{2} \sum_{i=1}^{k} \log \frac{1+\rho_i}{1+\rho_k} \leq I \right\}. \tag{30}$$

$R(I)$ has the following properties:

1) *$R(I)$ is strictly convex on $\big[0, \sum_{k=1}^n C(\rho_k)\big]$.*
2) *$R(I)$ is strictly increasing in I.*
3) *$R(I) \geq I$.*
4) *$R(0) = 0$ and $\lim_{I \to \sum_{k=1}^n C(\rho_k)} R(I) = \infty$.*
5) *$\frac{\mathrm{d}}{\mathrm{d}I} R(I) = \frac{1}{\ell(I)} \sum_{k=1}^{\ell(I)} (1+\rho_k)/(1+\rho_k - 2^{2I_k(I)})$.*

The rate-information function (28) can be identified as the sum of n rate-information functions of scalar Gaussian channels with SNRs ρ_k (cf. (17)).

Corollary 6. *The rate-information function (28) is the inverse of the information-rate function (20), i.e., for* $\tilde{I} \in \left[0, \sum_{k=1}^{n} C(\rho_k)\right]$ *and* $\tilde{R} \in \mathbb{R}_+$ *we have*

$$I(R(\tilde{I})) = \tilde{I} \qquad \text{and} \qquad R(I(\tilde{R})) = \tilde{R}. \tag{31}$$

Therefore the derivatives of $I(R)$ *and* $R(I)$ *are related as*

$$I'(\tilde{R}) = \frac{1}{R'\big(I(\tilde{R})\big)} \qquad \text{and} \qquad R'(\tilde{I}) = \frac{1}{I'\big(R(\tilde{I})\big)}. \tag{32}$$

Corollary 7. *Let* $\varepsilon > 0$*. The minimum rate* R_ε *required to achieve* $I(R) \geq I_\varepsilon$ *with* $I_\varepsilon = (1-\varepsilon)\sum_{k=1}^{n} C(\rho_k)$ *is given by*

$$R_\varepsilon = \frac{1}{2}\sum_{k=1}^{\ell(I_\varepsilon)} \log \frac{\rho_k}{\prod\limits_{k=1}^{\ell(I_\varepsilon)} (1+\rho_k)^{\frac{\varepsilon}{\ell(I_\varepsilon)}} \prod\limits_{k=\ell(I_\varepsilon)+1}^{n} (1+\rho_k)^{-\frac{1-\varepsilon}{\ell(I_\varepsilon)}} - 1}. \tag{33}$$

C. Rate-Information Trade-off using RD-Optimal Compression

In this subsection we quantify the rate-information trade-off $I^{\mathrm{RD}}(R)$ of RD-optimal compression with squared-error distortion $\|\boldsymbol{z}-\boldsymbol{y}\|^2$.

Theorem 8. *For a Gaussian vector channel with sorted mode SNRs* ρ_k*, the rate-information trade-off of RD-optimal channel output compression with squared-error distortion is given by*

$$I^{\mathrm{RD}}(R) = \frac{1}{2}\sum_{k=1}^{n} \log \frac{1+\rho_k}{1+2^{-2R_k^{\mathrm{RD}}(R)}\rho_k}, \tag{34}$$

where the rate allocated to the k*th mode is*

$$R_k^{\mathrm{RD}}(R) = \left[\frac{R}{l(R)} + \frac{1}{2}\log \frac{1+\rho_k}{\prod_{i=1}^{l(R)}(1+\rho_i)^{1/l(R)}}\right]^+ \tag{35}$$

and the number of active modes equals

$$l(R) = \max\left\{k\colon \frac{1}{2}\sum_{i=1}^{k}\log\frac{1+\rho_i}{1+\rho_k} \leq R\right\}. \tag{36}$$

We note that (20) can be written in the same form as (34) with the only difference being the rate allocation. Specifically, we can show that the number of active modes in (23) is never larger than (36).

Lemma 9. *The number of active modes with RD-optimal channel output compression satisfies* $l(R) \geq \ell(R)$ *for all* $R \in \mathbb{R}_+$.

D. Remarks

The results presented in this paper can be generalized to the case of correlated noise, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\mathbf{w}})$, by whitening the noise in $\mathbf{y}$. If we let $\boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^{\mathrm{T}}$ be the eigendecomposition of $\boldsymbol{C}_{\mathbf{w}}^{-1/2}\boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{w}}^{-1/2}$, then we have

$$\widetilde{\mathbf{y}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{w}}^{-1/2}\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}+\boldsymbol{I}) \tag{37}$$

which corresponds to (18) with $\sigma^2 = 1$. Applying $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}_{\mathbf{w}}^{-1/2}$ to $\mathbf{y}$ will leave the rate-information trade-off unchanged. However, performing RD-optimal compression of $\widetilde{\mathbf{y}}$ with squared-error distortion yields a compression of $\mathbf{y}$ with respect to a *weighted squared-error distortion*.

Furthermore, our results can be extended to the case of jointly complex Gaussian random vectors $\mathbf{x}, \mathbf{y}$ by writing the respective covariances in real form. That is, a complete statistical description of a complex Gaussian random vector $\boldsymbol{\zeta} = \boldsymbol{\zeta}_R + \sqrt{-1}\boldsymbol{\zeta}_I \in \mathbb{C}^n$ is given by the mean vector and the covariance matrix of $\widetilde{\boldsymbol{\zeta}} = \left[\boldsymbol{\zeta}_R^{\mathrm{T}}\ \boldsymbol{\zeta}_I^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{2n}$.

V. Analysis of $I(R) - I^{\mathrm{RD}}(R)$

A. $I(R) \geq I^{\mathrm{RD}}(R)$

The only difference between (20) and (34) is the rate allocation. Hence, to show $I(R) \geq I^{\mathrm{RD}}(R)$ it suffices to show that the rate allocation in (22) is optimal.

Lemma 10. *The rate allocation* R_k*,* $k = 1, \ldots, n$*, maximizing*

$$\frac{1}{2}\sum_{k=1}^{n}\log\frac{1+\rho_k}{1+2^{-2R_k}\rho_k} \tag{38}$$

subject to $\sum_{k=1}^{n} R_k = R$ *and* $R_k \geq 0$ *is given by (22).*

The proof of Lemma 10 reveals that the optimal rate allocation is given by the following reverse waterfilling solution:

$$R_k(\nu) = \frac{1}{2}\log^+\frac{\rho_k}{\nu}, \tag{39}$$

where the waterlevel ν is chosen such that $\sum_{k=1}^{n} R_k(\nu) = R$.

Theorem 11. *RD-optimal compression with squared-error distortion of the output of a Gaussian vector channel is suboptimal in the sense that*

$$I^{\mathrm{RD}}(R) \leq I(R), \tag{40}$$

with equality iff all mode SNRs ρ_k *are identical.*

Having identical mode SNRs entails $\gamma_1 = \ldots = \gamma_n$. Therefore, we can expect that the difference $I(R) - I^{\mathrm{RD}}(R)$ increases with increasing eigenvalue spread of $\boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}}$. The examples in the following subsections confirm this intuition.

Corollary 12. *The difference* $I(R) - I^{\mathrm{RD}}(R)$ *is given by*

$$\delta I(R) \triangleq I(R) - I^{\mathrm{RD}}(R) = \frac{1}{2}\sum_{k=1}^{n}\log\frac{1+2^{-2R_k^{\mathrm{RD}}}\rho_k}{1+2^{-2R_k}\rho_k} \geq 0. \tag{41}$$

We can upper bound $\delta I(R)$ for all $R \in \mathbb{R}_+$ as follows:

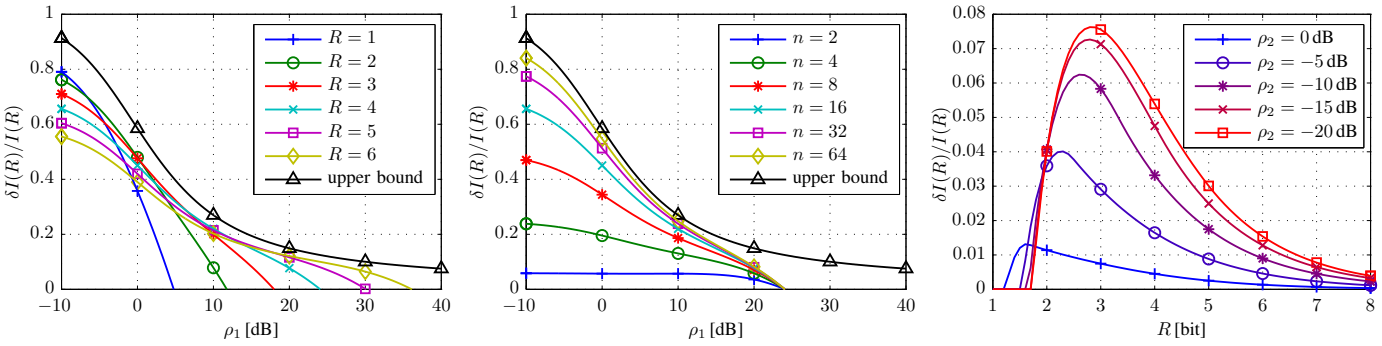
$$\delta I(R) \leq I(\mathbf{x};\mathbf{y}) - I^{\mathrm{RD}}\left(R_{c,2}^{\mathrm{RD}}\right) \tag{42}$$

$$= \sum_{k=1}^{n} C(\rho_k) - \frac{1}{2}\log\frac{(1+\rho_1)^2}{1+2\rho_1+\rho_1\rho_2}, \tag{43}$$

where $R_{c,2}^{\mathrm{RD}} = \frac{1}{2}\log\frac{1+\rho_1}{1+\rho_2}$ is the first nonzero critical rate in (36). The inequality in (42) holds because $I(R) \leq I(\mathbf{x};\mathbf{y})$ and $I(R)$, $I^{\mathrm{RD}}(R)$ are strictly increasing in R. Note that for $R < R_{c,2}^{\mathrm{RD}}$ we have $l(R) = 1$ and thus $\delta I(R) = 0$.

B. The SISO and MISO cases

In the SISO case (i.e, $\boldsymbol{H} = h \in \mathbb{R}$) and in the MISO case (i.e., $\boldsymbol{H} = \boldsymbol{h}^{\mathrm{T}} \in \mathbb{R}^{1\times m}$) we have only a single mode, since $\boldsymbol{H}\boldsymbol{C}_{\mathbf{x}}\boldsymbol{H}^{\mathrm{T}}$ is a scalar. We thus have $\delta I(R) = 0$.



(a) $n \times 1$ SIMO case (b) $n \times 1$ SIMO case (c) $2 \times m$ MIMO case

Figure 3: (a) $\delta I(R)/I(R)$ vs. ρ_1 for different R and $n = 16$. (b) $\delta I(R)/I(R)$ vs. ρ_1 for different n and $R = 4$. (c) $\delta I(R)/I(R)$ vs. R for different $\rho_2 \leq \rho_1 = 10\,\text{dB}$.

C. The SIMO case

In the SIMO case (i.e., $\boldsymbol{H} = \boldsymbol{h} \in \mathbb{R}^{n\times 1}$), $\boldsymbol{H}\boldsymbol{C}_{\mathsf{x}}\boldsymbol{H}^{\mathrm{T}}$ is a rank-1 matrix and therefore has only one nonzero eigenvalue (assuming $\boldsymbol{h} \neq \boldsymbol{0}$). Therefore we have $\rho_1 > 0$ and $\rho_2 = \ldots = \rho_n = 0$, i.e., $n-1$ modes contain noise only. However, for $R \geq R_{c,2}^{\mathrm{RD}} = C(\rho_1)$, RD-optimal compression allocates rate to *all* modes. We have

$$\delta I(R) = \begin{cases} 0, & R < R_{c,2}^{\mathrm{RD}} \\ \frac{1}{2}\log\frac{1+2^{-2R/n}\rho_1(1+\rho_1)^{1/n-1}}{1+2^{-2R}\rho_1}, & R \geq R_{c,2}^{\mathrm{RD}} \end{cases}, \quad (44)$$

and the upper bound (43) becomes

$$\delta I(R) \leq C\left(\frac{\rho_1}{1+\rho_1}\right). \quad (45)$$

Fig. 3a shows (44) and (45) normalized by $I(R)$ versus ρ_1 for different rates and $n = 16$. We observe that at low SNR $\delta I(R)/I(R)$ decreases with increasing R. Fig. 3b shows (44) and (45) normalized by $I(R)$ versus ρ_1 for different n and $R = 4\,\text{bit}$. We can see that the upper bound (44) gets tighter as n increases. Moreover, we have $\delta I(R)/I(R) = 0$ when ρ_1 is such that $R_{c,2}^{\mathrm{RD}} > R$.

D. The MIMO case

In the MIMO case (i.e., $\boldsymbol{H} \in \mathbb{R}^{n\times m}$), we have n modes with SNRs ρ_k, $k = 1,\ldots,n$. Using (41) we can compute $\delta I(R)$. However, expressions for $\delta I(R)$ involving only the mode SNRs and R become unwieldy for increasing n due to the large number of case distinctions. For $n = 2$ we have

$$\delta I(R) = \begin{cases} 0, & R < R_{c,2}^{\mathrm{RD}} \\ \frac{1}{2}\log\frac{f(\rho_1,\rho_2,R)}{(1+2^{-2R}\rho_1)(1+\rho_2)}, & R_{c,2}^{\mathrm{RD}} \leq R < R_{c,2} \\ \frac{1}{2}\log\frac{f(\rho_1,\rho_2,R)}{(1+2^{-R}\sqrt{\rho_1\rho_2})^2}, & R \geq R_{c,2} \end{cases}, \quad (46)$$

where $R_{c,2}^{\mathrm{RD}} = \frac{1}{2}\log\frac{1+\rho_1}{1+\rho_2}$, $R_{c,2} = \frac{1}{2}\log\frac{\rho_1}{\rho_2}$, and

$$f(\rho_1,\rho_2,R) \triangleq 1+2^{-R}\left(\sqrt{\tfrac{1+\rho_2}{1+\rho_1}}\rho_1 + \sqrt{\tfrac{1+\rho_1}{1+\rho_2}}\rho_2\right)+2^{-2R}\rho_1\rho_2. \quad (47)$$

We note that (46) simplifies to (44) (with $n = 2$) for $\rho_2 = 0$.

Fig. 3c shows (46) normalized by $I(R)$ versus R for different $\rho_2 \leq \rho_1 = 10\,\text{dB}$. The eigenvalue spread is proportional to $\rho_1 - \rho_2$ and we observe that $\delta I(R)/I(R)$ grows with increasing $\rho_1 - \rho_2$.

VI. Conclusions

We have considered channel output compression for Gaussian vector channels. We have provided closed-form expressions for the optimal rate-information trade-off and it turned out that the optimal rate allocation has a reverse waterfilling interpretation. Contrary to the scalar case, RD-optimal compression with squared-error distortion does not achieve the optimal rate-information trade-off in Gaussian vector channels. The suboptimality of the RD approach was quantified and essentially depends on the spread of the mode SNRs. Recently, [9] has shown that suitable linear pre-processing suffices to achieve the optimal rate-information trade-off using RD-optimal compression. Finally, we note that our results straightforwardly generalize to the complex-valued case with correlated noise.

References

[1] A. Winkelbauer and G. Matz, "Rate-information-optimal Gaussian channel output compression," in *Proc. 48th Annual Conference on Information Sciences and Systems (CISS 2014)*, March 2014.

[2] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, pp. 165–188, Jan. 2005.

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

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

[5] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Allerton Conf. on Communication, Control, and Computing*, Sept. 1999, pp. 368–377.

[6] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sept. 1975.

[7] A. Globerson and N. Tishby, "On the optimality of the Gaussian information bottleneck curve," The Hebrew University of Jerusalem, Tech. Rep., Feb. 2004.

[8] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2777, pp. 595–609.

[9] M. Meidlinger, A. Winkelbauer, and G. Matz, "On the relation between the Gaussian information bottleneck and MSE-optimal rate-distortion quantization," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP 2014)*, June 2014.