On Zero-Delay Source-Channel Coding

Emrah Akyol, *Member, IEEE*, Kumar B. Viswanatha, *Member, IEEE*,
Kenneth Rose, *Fellow, IEEE*, and Tor A. Ramstad

***Abstract*— This paper studies the zero-delay source-channel coding problem, and specifically the problem of obtaining the vector transformations that optimally map between the m-dimensional source space and k-dimensional channel space, under a given transmission power constraint and for the mean square error distortion. The functional properties of the cost are studied and the necessary conditions for the optimality of the encoder and decoder mappings are derived. An optimization algorithm that imposes these conditions iteratively, in conjunction with the noisy channel relaxation method to mitigate poor local minima, is proposed. The numerical results show strict improvement over prior methods. The numerical approach is extended to the scenario of source-channel coding with decoder side information. The resulting encoding mappings are shown to be continuous relatives of, and in fact subsume as special case, the Wyner–Ziv mappings encountered in digital distributed source coding systems. A well-known result in information theory pertains to the linearity of optimal encoding and decoding mappings in the scalar Gaussian source and channel setting, at all channel signal-to-noise ratios (CSNRs). In this paper, the linearity of optimal coding, beyond the Gaussian source and channel, is considered and the necessary and sufficient condition for linearity of optimal mappings, given a noise (or source) distribution, and a specified a total power constraint are derived. It is shown that the Gaussian source-channel pair is unique in the sense that it is the only source-channel pair for which the optimal mappings are linear at more than one CSNR values. Moreover, the asymptotic linearity of optimal mappings is shown for low CSNR if the channel is Gaussian regardless of the source and, at the other extreme, for high CSNR if the source is Gaussian, regardless of the channel. The extension to the vector settings is also considered where besides the conditions inherited from the scalar case, additional constraints must be satisfied to ensure linearity of the optimal mappings.**

Manuscript received February 11, 2013; revised July 14, 2014; accepted September 15, 2014. Date of publication October 3, 2014; date of current version November 18, 2014. This work was supported by the National Science Foundation through the Division of Computing and Communication Foundations under Grant CCF-0728986, Grant CCF-1016861, and Grant CCF-1118075. This paper was presented in part at the 2010 IEEE Information Theory Workshop, 2010 IEEE Data Compression Conference, and 2013 IEEE International Symposium on Information Theory.

E. Akyol was with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA. He is now with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL 61801 USA (e-mail: akyol@illinois.edu).

K. B. Viswanatha was with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA. He is now with the Corporate Research and Development Division, Qualcomm Technologies Inc., San Diego, CA 92121 USA (e-mail: kumar@ece.ucsb.edu).

K. Rose is with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: rose@ece.ucsb.edu).

T. A. Ramstad is with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim 7491, Norway (e-mail: ramstad@iet.ntnu.no).

Communicated by Y. Oohama, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2361532

***Index Terms*— Joint source channel coding, analog communications, estimation, distributed coding.**

I. Introduction

A FASCINATING result in information theory is that uncoded transmission of Gaussian samples, over a channel with additive white Gaussian noise (AWGN), is optimal in the sense that it yields the minimum achievable mean square error (MSE) between source and reconstruction [1]. This result demonstrates the potential of joint source-channel coding: Such a simple scheme, at no delay, provides the performance of the asymptotically optimal separate source and channel coding system, without recourse to complex compression and channel coding schemes that require asymptotically long delays. However, it is understood that the best source channel coding system at fixed finite delay may not, in general, achieve Shannon's asymptotic coding bound (see [2, Th. 21] or [3]).

The problem of obtaining the optimal scheme for a given finite delay is an important open problem with considerable practical implications. There are two main approaches to the practical problem of transmitting a discrete time continuous alphabet source over a discrete time additive noise channel: "analog communication" via direct amplitude modulation, and "digital communication" which typically consists of quantization, error control coding and digital modulation. The main advantage (and hence proliferation) of digital over analog communication is due to advanced quantization and error control techniques, as well as the prevalence of digital processors. However, there are two notable shortcomings: First, error control coding (and to some extent also source coding) usually incurs substantial delay to achieve good performance. The other problem involves limited robustness of digital systems against varying channel conditions, due to underlying quantization or error protection assumptions. The performance saturates due to quantization as the channel signal to noise ratio (CSNR) increases beyond the regime for which the system was designed. Also, it is difficult to obtain "graceful degradation" with decreasing CSNR, when it falls below the minimum requirement of the error correction code in use. Further, such threshold effects become more pronounced as the system performance approaches the theoretical optimum. Analog systems offer the potential to avoid these problems. As an important example, in applications where significant delay is acceptable, a hybrid approach (i.e., vector quantization + analog mapping) was proposed and analyzed [4], [5] to circumvent the impact of CSNR mismatch, wherein linear mappings were used and hence no optimality claims were made. Perhaps more importantly, in many applications delay is a paramount consideration. Analog coding schemes are

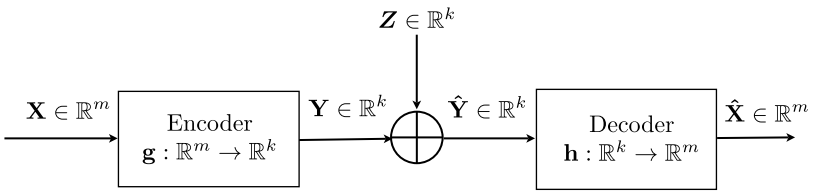


Fig. 1. A general block-based point-to-point communication system.

low complexity alternatives to digital methods, providing a "zero-delay" transmission which is suitable for such applications.

There are no known explicit methods to obtain such analog mappings for general sources and channels, nor is the optimal mapping known, in closed form, for other than the trivial case of the scalar Gaussian source-channel pair. Among the few practical analog coding schemes that have appeared in the literature are those based on the use of space-filling curves for bandwidth compression, originally proposed more than 50 years ago by Shannon [6] and Kotelnikov [7]. These were extended in the work of Fuldseth and Ramstad [8], Chung [9], Vaishampayan and Costa [10], Ramstad [11], and Hekland et.al. [12], where spiral-like curves were explored for transmission of Gaussian sources over AWGN channels for bandwidth compression ($m > k$) and expansion ($m < k$). There exist two main approaches to numerical optimization of the mappings: i) optimization of the parameter set of a structured mapping [11]–[14]. The performance of this approach is limited to the parametric form (structure) assumed. ii) Design based on power constrained channel optimized vector quantization (PCCOVQ) where a "discretized version" of the problem is tackled using tools developed for vector quantization [8], [15], [16].

A similar problem was solved in [17] and [18] albeit under the stringent constraint that both encoder and decoder be linear. A related problem, formulated in the pure context of digital systems, was studied by Fine [19]. Properties of the optimal mappings have been considered, over the years, in [6], [20], and [21]. Shannon's arguments [6] are based on the topological impossibility to map between regions in a "one-to-one", continuous manner, unless they have the same dimensionality. On this basis, he explained the threshold effect common to various communication systems. Moreover, Ziv [20] showed that for a Gaussian source transmitted over an AWGN channel, no single practical modulation scheme can achieve optimal performance at all noise levels, if the channel rate is greater than the source rate (i.e., bandwidth expansion). It has been conjectured that this property holds whenever the source rate differs from the channel rate [21]. Our own preliminary results appeared in [22]–[24]. The existence of optimal real time encoders has been studied in [25]–[28] for encoding a Markov source with zero-delay. Along these lines, for similar set of problems, [26] demonstrated the existence of optimal causal encoders using dynamic programming, its results are recently extended to partially observed Markov sources and multiterminal settings in [29]. The problem we consider is intrinsically connected to problems in stochastic control where the controllers must operate at zero delay. A control problem, similar to the zero-delay source channel coding problem here, is the Witsenhausen's well known counterexample [30] (see [31] for a comprehensive review) where a similar functional optimization problem is studied and it is shown that nonlinear controllers can outperform linear ones in decentralized control settings even under Gaussianity and MSE assumptions.

In this paper, we investigate the problem of obtaining vector transformations that optimally map between the m-dimensional source space and the k-dimensional channel space, under a given transmission power constraint, and where optimality is in the sense of minimum mean square reconstruction error. We provide necessary conditions for the optimality of the mappings used at the encoder and the decoder. It is important to note that virtually any source-channel communication system (including digital communication) is a special case of the general mappings shown in Figure 1. A typical digital system, including quantization, error correction and modulation, boils down to a specific mapping from the source space $\mathbb{R}^m$ to the channel space $\mathbb{R}^k$ and back to reconstruction space $\mathbb{R}^m$ at the receiver. Hence the derived optimality conditions are generally valid and subsume digital communications as an extreme special case. Based on the optimality conditions, we propose an iterative algorithm to optimize the mappings for any given m, k (i.e., for both bandwidth expansion or compression) and for any given source-channel statistics. We provide examples of such $m : k$ mappings for source-channel pairs and construct the corresponding source-channel coding systems that outperform the mappings obtained in [8]–[12]. We next study the functional properties of the point-to-point problem. Specifically, first we show that MSE is a concave functional of the source density, given a fixed noise density, and of the noise density given a fixed source. Secondly, MSE is a convex functional of the channel input density. The convexity result makes the optimal encoding mapping *essentially unique*.[1] Next, we derive the necessary and sufficient conditions for linearity of optimal mappings in terms of the source, channel densities and the power constraint. We study the CSNR asymptotics and particularly show that given a Gaussian source, optimal mappings are asymptotically linear at high CSNR, irrespective of the channel. Similarly, for a Gaussian channel, optimal mappings are asymptotically linear at low CSNR regardless of the source. We next extend our analysis to higher dimensional spaces and study the

[1]The optimal mapping is not strictly unique, in the sense that multiple trivially "equivalent" mappings can be used to obtain the same channel input density. For example, a scalar unit variance Gaussian source and scalar Gaussian channel with power constraint P, can be optimally encoded by either $y = \sqrt{P}x$ or $y = -\sqrt{P}x$.

implications of linearity conditions. The last part of the paper extends the numerical approach to the scenario of source-channel coding with decoder side information (i.e., the decoder has access to side information that is correlated with the source). This setting, in the context of pure source coding, goes back to the pioneering work of Slepian and Wolf [32] and Wyner and Ziv [33]. The derivation of the optimality conditions for the decoder side information setting is a direct extension of the point-to-point case, but the distributed nature of this setting results in highly nontrivial mappings. Straightforward numerical optimization of such mappings is susceptible to get trapped in numerous poor local minima that riddle the cost functional. Note, in particular, that in the case of jointly Gaussian source and side information and a Gaussian channel, linear encoders and decoder (automatically) satisfy the necessary conditions of optimality while, as we will see, careful optimization obtains considerably better mappings that are far from linear.

The paper is organized as follows: We formulate the problem in Section II. We analyze the functional properties of the problem, derive the necessary conditions for the optimality of the mappings and provide an iterative algorithm based on these conditions in Section III. We analyze the linearity of encoding and/or decoding mappings in Section IV. We provide example of mappings and comparative numerical results in Section V. Discussion and future work are presented in Section VI.

II. Problem Formulation

A. Preliminaries and Problem Definitions

Let $\mathbb{R}$, $\mathbb{N}$, $\mathbb{R}^+$, and $\mathbb{C}$ denote the respective sets of real numbers, natural numbers, positive real numbers and complex numbers. In general, lowercase letters (e.g., x) denote scalars, boldface lowercase (e.g., $\boldsymbol{x}$) column vectors, upper-case (e.g., C, X) matrices and random variables, and boldface uppercase (e.g., $\boldsymbol{X}$) random column vectors. I denotes the identity matrix. Unless otherwise specified, vectors and random vectors have length m, and matrices have size $m \times m$. The k^{th} element of vector $\boldsymbol{x}$ is denoted by $[\boldsymbol{x}]_k$ and the $(i, j) - th$ element and the k^{th} column of the matrix U by $[U]_{ij}$ and $[U]_k$ respectively. U^{-T} denotes $(U^T)^{-1}$. R_X and R_{XZ} denote the auto-covariance of $\boldsymbol{X}$ and cross covariance of $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively. A^T denotes the transpose of matrix (vector) A. ∇ denotes the gradient. Let $\mathbb{E}(\cdot)$, $\mathbb{P}(\cdot)$ and $\| \cdot \|$ denote the expectation, probability and l_2 norm operators, respectively. Let $f^{'}(x) = \frac{df(x)}{dx}$ denote the first-order derivative of the continuously differentiable function $f(\cdot)$. The Gaussian density with mean μ and variance σ^2 is denoted as $\mathcal{N}(\mu, \sigma^2)$. All logarithms in the paper are natural logarithms and may in general be complex, and the integrals are, in general, Lebesgue integrals. Throughout the paper, "almost everywhere" is denoted as *a.e.*.

We assume that the source $\boldsymbol{X}$ is an m-dimensional zero mean vector[2] and covariance R_X. The channel noise $\boldsymbol{Z}$ is additive, k-dimensional mean zero and covariance R_Z and is independent of the source $\boldsymbol{X}$. The m-fold source density is denoted $f_X(\cdot)$ and the k-fold noise density is $f_Z(\cdot)$ with characteristic functions $F_X(\boldsymbol{\omega})$ and $F_Z(\boldsymbol{\omega})$, respectively.

Let $\mathcal{S}_m^k$ denote the set of Borel measurable functions $\{\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^k\}$ with $\mathbb{E}\{||\boldsymbol{f}(\boldsymbol{X})||^2\} < \infty$ and $\mathcal{S}_m^+ \subset \mathcal{S}_m^k$ be the subset of *monotone* $\{\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m\}$ functions in $\mathcal{S}_m^k$. Monotonicity simplifies to "monotone increasing" in the scalar $\mathbb{R} \to \mathbb{R}$ case, while in higher dimensional settings, it is equivalent to the condition $(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) > 0$ a.e. in $\boldsymbol{x}$ and $\boldsymbol{y}$.

1) Point to Point: We consider the communication system with a block diagram shown in Figure 1. A vector source $\boldsymbol{X} \in \mathbb{R}^m$ is mapped onto $\boldsymbol{Y} \in \mathbb{R}^k$ by a function $\boldsymbol{g} \in \mathcal{S}_m^k$, and transmitted over an additive noise channel. The received vector $\hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{Z}$ is mapped by the decoder to the estimate $\hat{\boldsymbol{X}}$ via a function $\boldsymbol{h} \in \mathcal{S}_k^m$. The objective is to minimize the MSE

$$D(\boldsymbol{g}, \boldsymbol{h}) = \mathbb{E}\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2\}, \tag{1}$$

over the choice of encoder $\boldsymbol{g}(\cdot) \in \mathcal{S}_m^k$ and decoder $\boldsymbol{h}(\cdot) \in \mathcal{S}_k^m$, subject to the average power constraint,

$$P(\boldsymbol{g}) = \mathbb{E}\{||\boldsymbol{g}(\boldsymbol{X})||^2\} \leq P_T, \tag{2}$$

where P_T is the specified transmission power level. To impose the power constraint, we minimize the Lagrangian cost functional:

$$J(\boldsymbol{g}, \boldsymbol{h}) = D(\boldsymbol{g}, \boldsymbol{h}) + \lambda P(\boldsymbol{g}). \tag{3}$$

Note that this a well known relaxation in convex optimization and there is no duality gap since the distortion is a convex function of power [34]. Bandwidth compression-expansion is determined by the source and channel dimensions, k/m. The power constraint limits the choice of encoder function $\boldsymbol{g}(\cdot)$. Note that without a power constraint on $\boldsymbol{g}(\cdot)$, the CSNR is unbounded and the channel can be made effectively noise free. Let $\boldsymbol{g}^*$ and $\boldsymbol{h}^*$ denote the optimal mappings, i.e.,

$$J(\boldsymbol{g}^*, \boldsymbol{h}^*) \leq J(\boldsymbol{g}, \boldsymbol{h}), \tag{4}$$

for any $\boldsymbol{g} \in \mathcal{S}_m^k$ and $\boldsymbol{h} \in \mathcal{S}_k^m$.

2) Decoder Side Information: As shown in Figure 2, there are two correlated vector sources $\boldsymbol{X}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{m_2}$ with a joint density $f_{X_1,X_2}(\cdot, \cdot)$. The side information $\boldsymbol{X}_2$ is available only to the decoder, while $\boldsymbol{X}_1$ is mapped to $\boldsymbol{Y} \in \mathbb{R}^k$ by an encoding function $\boldsymbol{g} \in \mathcal{S}_{m_1}^k$ and transmitted over the channel with additive noise $\boldsymbol{Z} \in \mathbb{R}^k$, with a density $f_Z(\cdot)$, independent of $\boldsymbol{X}_1, \boldsymbol{X}_2$. The received channel output $\hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{Z}$ is mapped to the estimate $\hat{\boldsymbol{X}}_1$ by a decoding function $\boldsymbol{h} : \mathbb{R}^k \times \mathbb{R}^{m_2} \to \mathbb{R}^{m_1}$. The objective is to find optimal mapping functions $\boldsymbol{g}(\cdot), \boldsymbol{h}(\cdot)$ that minimize MSE

$$D(\boldsymbol{g}, \boldsymbol{h}) = \mathbb{E}\{||\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1||^2\}, \tag{5}$$

subject to $P(\mathbf{g}) \leq P_T$ where

$$P(\mathbf{g}) = \mathbb{E}\{||\mathbf{g}(\mathbf{X_1})||^2\}. \tag{6}$$

[2]The zero mean assumption is not necessary, but it considerably simplifies the notation. Therefore, it is made throughout the paper.

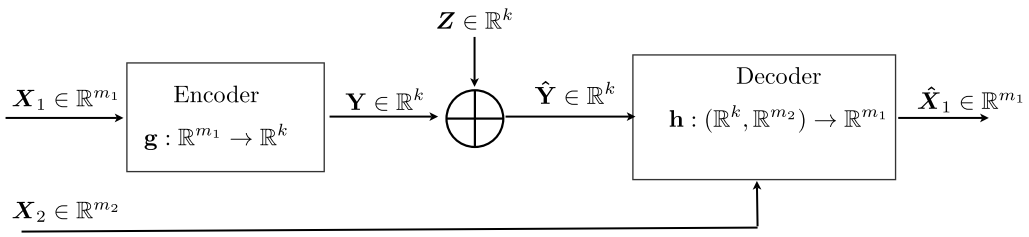


Fig. 2. Source-channel coding with decoder side information.

B. Asymptotic Bounds for Gaussian Source and Channel

Although the problem we consider is delay limited, it is insightful to consider asymptotic bounds obtained at infinite delay. Shannon's source and channel coding theorems imply that, asymptotically, the source can be compressed to $R(D)$ bits (per source sample) at distortion level D, and that $C(P_T)$ bits can be transmitted over the channel (per channel use) with arbitrarily low probability of error, where $R(D)$ is the source rate-distortion function, and $C(P_T)$ is the channel capacity at power P_T (see [35]). The asymptotically optimal coding scheme is the tandem combination of the optimal source and channel coding schemes, hence $mR(D) \leq kC(P_T)$ must hold. By setting

$$R(D) = \frac{k}{m} C(P_T), \tag{7}$$

one obtains a lower bound on the distortion of any source-channel coding scheme. Next, we specialize to Gaussian sources and channels, which we will use in the numerical results as benchmark. The rate-distortion function for the memoryless Gaussian scalar source of variance σ_X^2, under MSE is given by

$$R(D) = \max\left(0, \frac{1}{2}\log\frac{\sigma_X^2}{D}\right), \tag{8}$$

for any distortion value $D \geq 0$. The capacity of the additive, scalar, memoryless Gaussian channel is given by

$$C(P_T) = \frac{1}{2}\log\left(1 + \frac{P_T}{\sigma_Z^2}\right), \tag{9}$$

where σ_Z^2 is the noise variance. Plugging (8) and (9) in (7), we obtain the optimal performance theoretically attainable (OPTA):

$$D_{OPTA} = \frac{\sigma_X^2}{\left(1 + \frac{P_T}{\sigma_Z^2}\right)^{\frac{k}{m}}}. \tag{10}$$

For source-channel coding with decoder side information, OPTA can be obtained by equating Wyner-Ziv rate distortion function [33] to the channel capacity. The Wyner-Ziv rate distortion function of X_1 when X_2 serves as side information and $(X_1, X_2) \sim \mathcal{N}(\mathbf{0}, R_{X_1,X_2})$, where $R_{X_1,X_2} = \sigma_X^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ with $|\rho| \leq 1$ is:

$$R(D) = \max\left(0, \frac{1}{2}\log\frac{(1-\rho^2)\sigma_X^2}{D}\right), \tag{11}$$

We plug (11) and (9) in (7) to obtain

$$D_{OPTA} = \frac{(1-\rho^2)\sigma_X^2}{\left(1 + \frac{P_T}{\sigma_Z^2}\right)^{\frac{k}{m}}}. \tag{12}$$

Note that D_{OPTA} is derived without any delay constraints and may not be achievable by a delay-constrained coding scheme. No achievable theoretical bound is known for joint source-channel coding at zero-delay, although there are recent results that tighten the outer bound, see [36]–[38].

III. Functional Properties of Zero-Delay Source-Channel Coding Problem

In this section, we study the functional properties of the optimal zero-delay source-channel coding problem. These properties are not only important in their own right, but also enable the derivation of several subsequent results. Particularly concavity properties of J_m play an important role in jamming problems where the worst case additive noise distribution (maximizer of J_m) is explored (see [39] for details). Convexity properties enable the sufficiency of the linearity results presented later. Let us restate the Lagrangian cost (3), as $J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{h})$ which makes explicit its dependence on the source and channel noise $\boldsymbol{X}$ and $\boldsymbol{Z}$, beside the deterministic mappings $\boldsymbol{g}(\cdot)$ and $\boldsymbol{h}(\cdot)$ as:

$$J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{h}) = \mathbb{E}\left\{||\boldsymbol{X} - \boldsymbol{h}(\boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{Z})||^2\right\} + \lambda \mathbb{E}\left\{||\boldsymbol{g}(\boldsymbol{X})||^2\right\}. \tag{13}$$

The minimum cost is

$$J_m(\boldsymbol{X}, \boldsymbol{Z}) \triangleq \inf_{\boldsymbol{g}, \boldsymbol{h}} J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{h}). \tag{14}$$

A. Concavity of J_m in $f_X(\cdot)$ and $f_Z(\cdot)$

In this section, we show the concavity of J_m in $f_X(\cdot)$ and in $f_Z(\cdot)$. Similar results were derived for the related but different setting of MMSE estimation in [40], where a scalar estimation problem not involving communication and encoding was studied. We start with the following simple lemma which states the impact of conditioning on the overall cost. Conditioned on another random variable $\boldsymbol{U}$, $J_m(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{U})$ denotes $J_m(\boldsymbol{X}, \boldsymbol{Z})$ when $\boldsymbol{U}$ is available to both encoder and decoder.

Lemma 1: Conditioning cannot increase the overall cost, J_m i.e., $J_m(\boldsymbol{X}, \boldsymbol{Z}) \geq J_m(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{U})$ for any $\boldsymbol{U}$.

Proof: The knowledge of $\boldsymbol{U}$ cannot increase the total cost, since we can always ignore $\boldsymbol{U}$ and use the $\boldsymbol{g}(\cdot), \boldsymbol{h}(\cdot)$ pair that is optimal for $J_m(\boldsymbol{X}, \boldsymbol{Z})$. Hence, $J_m(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{U}) \leq J_m(\boldsymbol{X}, \boldsymbol{Z})$. □

Next, we have the following theorem.

Theorem 1: J_m is concave in $f_X(\cdot)$ and $f_Z(\cdot)$.

Proof: Let $\boldsymbol{X}$ be distributed according to $f_X = pf_{X_1} + (1-p)f_{X_2}$, where f_{X_1} and f_{X_2} respectively denote the densities of random variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Next, $\boldsymbol{X}$ can be expressed, in terms of a time sharing random variable U which takes values in the alphabet $\{1, 2\}$, with $\mathbb{P}\{U = 1\} = p : \boldsymbol{X} = \boldsymbol{X}_U$.

$$J_m(\boldsymbol{X}, \boldsymbol{Z}) \geq J_m(\boldsymbol{X}, \boldsymbol{Z}|U) \tag{15}$$

$$= pJ_m(\boldsymbol{X}_1, \boldsymbol{Z}) + (1-p)J_m(\boldsymbol{X}_2, \boldsymbol{Z}), \tag{16}$$

which proves the concavity of $J_m(\boldsymbol{X}, \boldsymbol{Z})$ for fixed f_Z. Similar arguments on $\boldsymbol{Z}$ prove that $J_m(\boldsymbol{X}, \boldsymbol{Z})$ is concave in f_Z for fixed f_X. □

B. Convexity of Overall Cost in $f_Y(\cdot)$

In this section, we study the convexity of $J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{h})$ in the channel input density $f_Y(\cdot)$ of $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$, when $\boldsymbol{h}(\cdot)$ is optimized for $\boldsymbol{g}(\cdot)$. An important distinction to make is that convexity in $\boldsymbol{g}(\cdot)$ is not implied in general. A trivial example to demonstrate non-convexity in $\boldsymbol{g}(\cdot)$ in the general setting is the scalar Gaussian source and channel setting, where both $Y = \sqrt{\frac{P_T}{\sigma_X^2}}X$ and $Y = -\sqrt{\frac{P_T}{\sigma_X^2}}X$ are optimal (when used in conjunction with their respective optimal decoders). This example leads to the intuition that the cost functional may be *essentially* convex (e.g., convex when $g(\cdot)$ is limited to be in the set $\mathcal{S}_1^+$ in the scalar case) although it is clearly not convex in the strict sense. It turns out that this intuition is correct: total cost is convex in $f_Y(\cdot)$ which implies that the cost is convex in $\boldsymbol{g}(\cdot)$ **when $\boldsymbol{g}(\cdot)$ is limited to be in the set $\mathcal{S}_m^+$.**

We first reformulate the mapping problem by allowing random mappings, i.e., we relax the mappings from deterministic functions $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ and $\hat{\boldsymbol{X}} = \boldsymbol{h}(\hat{\boldsymbol{Y}})$ to probabilistic transformations, expressed as $f_{Y|X}$ and $f_{\hat{X}|\hat{Y}}$. Note that similar relaxations have been used in the literature, e.g., recently in [41]. We define the *generalized* mapping problem as: minimize $J_{gen}(\boldsymbol{X}, \boldsymbol{Z}, f_{Y|X}, f_{\hat{X}|\hat{Y}})$ over the conditional densities $f_{Y|X}$ and $f_{\hat{X}|\hat{Y}}$ where the cost functional J_{gen} is defined as

$$J_{gen}(\boldsymbol{X}, \boldsymbol{Z}, f_{Y|X}, f_{\hat{X}|\hat{Y}}) \triangleq \mathbb{E}\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2\} + \lambda\mathbb{E}\{||\boldsymbol{Y}||^2\}, \tag{17}$$

where $\boldsymbol{Y}$ and $\hat{\boldsymbol{X}}$ are random transformations of $\boldsymbol{X}$ and $\hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{Z}$ through $f_{Y|X}$ and $f_{\hat{X}|\hat{Y}}$ respectively. We first show that this relaxation does not change the solution space, via the following lemma.

Lemma 2: The optimal $f_{Y|X}$ and $f_{\hat{X}|\hat{Y}}$ are deterministic, i.e., $\boldsymbol{Y}$ and $\hat{\boldsymbol{X}}$ which minimize (17) are deterministic functions of $\boldsymbol{X}$ and $\hat{\boldsymbol{Y}}$ respectively. Hence, $J_m(\boldsymbol{X}, \boldsymbol{Z}) = \inf_{\boldsymbol{g},\boldsymbol{h}} J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{h}) = \inf_{f_{Y|X}, f_{\hat{X}|\hat{Y}}} J_{gen}(\boldsymbol{X}, \boldsymbol{Z}, f_{Y|X}, f_{\hat{X}|\hat{Y}})$.

Proof: First, we observe that optimal $f_{\hat{X}|\hat{Y}}$ is deterministic since $\boldsymbol{h}(\boldsymbol{Y}) = \mathbb{E}\{\boldsymbol{X}|\hat{\boldsymbol{Y}}\}$ minimizes MSE. Next, consider the following:

$$\inf_{f_{Y|X}} J_{gen}(\boldsymbol{X}, \boldsymbol{Z}, f_{Y|X}, f_{\hat{X}|\hat{Y}}) = \inf_{\boldsymbol{h}} \int f_X(\boldsymbol{x}) \inf_{f_{Y|X}} \left\{ \int G_Z(\boldsymbol{x}, \boldsymbol{y}) f_{Y|X}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right\} \mathrm{d}\boldsymbol{x}. \tag{18}$$

where

$$G_Z(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \int \left(\|\boldsymbol{X} - \boldsymbol{h}(\boldsymbol{Y} + \boldsymbol{Z})\|^2 + \lambda \|\boldsymbol{Y}\|^2 \right) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}.$$

The minimization in (18) can be done, for a fixed $\boldsymbol{h}(\cdot)$, by choosing the $\boldsymbol{Y} = \boldsymbol{y}$ that minimizes $G_Z(\boldsymbol{x}, \boldsymbol{y})$ for each $\boldsymbol{X} = \boldsymbol{x}$. Using the optimal $\boldsymbol{h}(\cdot)$ as the fixed $\boldsymbol{h}(\cdot)$ in (18), it follows that the optimal $\boldsymbol{Y}$ is a deterministic function of $\boldsymbol{X}$.

(Alternatively, note that J_{gen} is affine in $f_{Y|X}$. The pointwise infimum of this functional with respect to $\boldsymbol{h}(\cdot)$ is concave in $f_{Y|X}$ and the minima of a concave functional occur in the boundary which corresponds to deterministic mappings in our problem.) □

Next, we investigate the essential convexity of the overall cost in the encoder mapping when the decoder is optimized, in the randomized setting with matched source-channel dimensions, i.e., $m = k$. To this aim, we define $J_r(f_Y)$ as the infimum of J_{gen} as a function of $f_Y(\boldsymbol{y})$, where infimum is taken over all conditional distributions $f_{\hat{X}|\hat{Y}}$, under the condition that the following is satisfied

$$f_{\hat{X}}(\hat{\boldsymbol{x}}) = \int f_Y(\boldsymbol{y}) f_{\hat{X}|\hat{Y}}(\hat{\boldsymbol{x}}, \boldsymbol{y} + \boldsymbol{z}) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z}, \tag{19}$$

i.e., we have (20), as shown at the bottom of this page.

Theorem 2: $J_r(f_Y)$ is convex in $f_Y(\cdot)$ in the setting of $m = k$.

Proof: We first note that $\mathbb{E}\{||\boldsymbol{Y}||^2\}$ does not depend on $f_{\hat{X}|\hat{Y}}$, hence can be taken out of the infimum. $\mathbb{E}\{||\boldsymbol{Y}||^2\}$ is a linear (and hence convex) function f_Y. Hence, the convexity of J_r is determined by the first term given in (21), as shown at the bottom of this page, which can be expressed as in (22), as shown at the bottom of this page,

$$J_r(f_Y) = \inf_{f_{\hat{X}|\hat{Y}}} \left\{ \mathbb{E}\left\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2\right\} + \lambda\mathbb{E}\left\{||\boldsymbol{Y}||^2\right\} \middle| f_{\hat{X}}(\hat{\boldsymbol{x}}) = \int f_Y(\boldsymbol{y}) f_{\hat{X}|\hat{Y}}(\hat{\boldsymbol{x}}, \boldsymbol{y} + \boldsymbol{z}) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z} \right\}. \tag{20}$$

$$J_1(f_Y) = \inf_{f_{\hat{X}|\hat{Y}}} \left\{ \mathbb{E}\left\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2\right\} \middle| f_{\hat{X}}(\hat{\boldsymbol{x}}) = \int f_Y(\boldsymbol{y}) f_{\hat{X}|\hat{Y}}(\hat{\boldsymbol{x}}, \boldsymbol{y} + \boldsymbol{z}) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z} \right\} \tag{21}$$

$$J_1(f_Y) = \int \inf_{f_{\hat{X}|Y,\boldsymbol{Z}=z}} \left\{ \mathbb{E}\left\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2 \middle| \boldsymbol{Z} = \boldsymbol{z}\right\} \middle| f_{\hat{X}}(\hat{\boldsymbol{x}}) = \int f_Y(\boldsymbol{y}) f_{\hat{X}|\hat{Y}}(\hat{\boldsymbol{x}}, \boldsymbol{y} + \boldsymbol{z}) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{z} \right\} f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \tag{22}$$

noting that $\hat{\boldsymbol{X}}$ depends on $\boldsymbol{X}$ only through $\boldsymbol{Y}$, i.e., $\hat{\boldsymbol{X}} - \boldsymbol{Y} - \boldsymbol{X}$ forms a Markov chain in this order. We also note that $\mathbb{E}\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 | \boldsymbol{Z} = \boldsymbol{z}\}$ is a linear functional of $f_{\hat{X},X|\boldsymbol{Z}=z}$, and hence of $f_{\hat{X},Y|\boldsymbol{Z}=z}$. The overall function is convex in $f_{\hat{X},Y}$, since any weighted sum of convex functionals is convex under affine constraints over the variables, i.e., in this case, under the condition that $f_{\hat{X},Y}(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \int f_Z(\boldsymbol{z}) f_{\hat{X},Y|\boldsymbol{Z}=z}(\hat{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}$.

This implies that we have a jointly convex functional of $f_{\hat{X}|Y}$ and f_Y where (19) must be satisfied. It is known that in general $\Xi(b) = \inf_{a \in \mathcal{C}} \Phi(a, b)$ is a convex function of b under one-to-one affine constraints on a, b when $\mathcal{C}$ is a convex set, $\Phi(a, b) > -\infty$, and $\Phi(a, b)$ is convex in the product space of a, b (see [42, Th. 5.7, p. 38] or [34, Sec. 3.2.5, Example 3.17]). In our problem, the set of conditional probabilities $f_{\hat{X}|Y}$ is convex, and $J_1(f_Y) > 0$; hence $J_1(f_Y)$ is convex in $f_Y(\cdot)$, which implies that $J_r(f_Y)$ is convex in $f_Y(\cdot)$. □

A practically important consequence of Theorem 2 is stated in the following corollary. Let $J_r(\boldsymbol{g})$ denote J_r when the deterministic encoding mapping $\boldsymbol{g}(\cdot)$ is used in conjunction with its optimal decoder.

Corollary 1: $J_r(\boldsymbol{g})$ is convex in $\boldsymbol{g}(\cdot)$ *in the set of* $\mathcal{S}_m^+$.

Proof: There is one-to-one mapping between $\boldsymbol{Y}$ and the encoder $\boldsymbol{g}(\cdot) \in \mathcal{S}_m^+$ as $\mathcal{F}_X(\boldsymbol{X}) = \mathcal{F}_Y(\boldsymbol{g}(\boldsymbol{X}))$ where $\mathcal{F}_X$ and $\mathcal{F}_Y$ denote the cumulative distribution functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. It follows from Theorem 2 that for any f_{Y_1} and f_{Y_2} and for $1 \geq \alpha \geq 0$ we have

$$\alpha J_r(f_{Y_1}) + (1-\alpha) J_r(f_{Y_2}) \geq J_r(\alpha f_{Y_1} + (1-\alpha) f_{Y_2}). \quad (23)$$

The fact $J_r(f_Y)$ is achieved by a unique $\boldsymbol{g}(\cdot) \in \mathcal{S}_m^+$ implies that

$$\alpha J_r(\boldsymbol{g}_1) + (1-\alpha) J_r(\boldsymbol{g}_2) \geq J_r(\alpha \boldsymbol{g}_1 + (1-\alpha)\boldsymbol{g}_2), \quad (24)$$

which shows the convexity of J_r in $\boldsymbol{g}(\cdot)$, in $\mathcal{S}_m^+$. □

C. Optimality Conditions

We proceed to develop the necessary conditions for the optimality of the encoder and decoder subject to the average power constraint (2), in the general setting of $m, k \in \mathbb{N}$. While the optimality conditions follow from standard arguments, they have not been explicitly reported in the literature, hence are presented in the following theorem.[3]

Theorem 3: Given source and noise densities, a coding scheme $(\boldsymbol{g}(\cdot), \boldsymbol{h}(\cdot))$ *is optimal only if*

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{1}{\lambda} \int \boldsymbol{h}'(\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{z})[\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{z})] f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \quad (25)$$

$$\boldsymbol{h}(\hat{\boldsymbol{y}}) = \frac{\int \boldsymbol{x}\, f_X(\boldsymbol{x})\, f_Z(\hat{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}}{\int f_X(\boldsymbol{x})\, f_Z\left(\hat{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}}, \quad (26)$$

where varying λ *provides solutions at different levels of power constraint* P_T. *In fact,* λ *is the slope of the distortion-power curve:* $\lambda = -\frac{dD}{dP_T}$.

Proof: See Appendix A. □

Corollary 2: In the setting of $m = k$, *(25) and (26) are sufficient for optimality.*

Proof: The proof follows from Theorem 2. □

The following auxiliary result will be used in the next section.

Corollary 3: There exist linear mappings $\boldsymbol{g}(\boldsymbol{x}) = K_e \boldsymbol{X}$ *and* $\boldsymbol{h}(\boldsymbol{Y}) = K_d \boldsymbol{Y}$ *for some* $K_e \in \mathbb{R}^{m \times k}$, $K_d \in \mathbb{R}^{k \times m}$ *that satisfy (25) for any* $m, k \in \mathbb{N}$, *regardless of the source and channel densities.*

Proof: Let us plug $\boldsymbol{h}(\boldsymbol{Y}) = K_d \boldsymbol{Y}$ in (25). Noting that $\boldsymbol{h}'(\boldsymbol{Y}) = K_d$ *a.e.* in Y, we have

$$\lambda \boldsymbol{g}(\boldsymbol{X}) = K_d \int (\boldsymbol{X} - K_d \boldsymbol{g}(\boldsymbol{X}) - K_d \boldsymbol{z}) f_Z(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \quad (27)$$

a.e. in $\boldsymbol{X}$. Evaluating the integral and noting that $\mathbb{E}\{\boldsymbol{Z}\} = \boldsymbol{0}$, we have

$$\lambda \boldsymbol{g}(\boldsymbol{X}) = K_d(\boldsymbol{X} - K_d \boldsymbol{g}(\boldsymbol{X})) \quad (28)$$

a.e. in $\boldsymbol{X}$ and hence $\boldsymbol{g}(\boldsymbol{X}) = K_e \boldsymbol{X}$. □

The necessary conditions for optimality in Theorem 3 are not sufficient in general settings, as is demonstrated in particular by the following corollary.

Corollary 4: For a Gaussian source $\boldsymbol{X}$ *and a Gaussian channel* $\boldsymbol{Z}$, *(25) and (26) are satisfied by linear mappings* $\boldsymbol{g}(\boldsymbol{X}) = K_e \boldsymbol{X}$ *and* $\boldsymbol{h}(\boldsymbol{Y}) = K_d \boldsymbol{Y}$ *for some* $K_e \in \mathbb{R}^{m \times k}$, $K_d \in \mathbb{R}^{k \times m}$ *for any* $m, k \in \mathbb{N}$.

Remark 1: Although linear mappings satisfy the necessary conditions of optimality for the Gaussian case for any m and k, they are highly suboptimal when $m \neq k$, *see [18].*

Proof: Linear mappings satisfy the first necessary condition, (25) due to Corollary 3. Optimal decoder is linear in the Gaussian source-channel setting, satisfying (26). □

1) Extension to Distributed Settings: Optimality conditions for the setting of decoder side information can be obtained by following similar steps (see Appendix B). We note, in particular, that for these settings a similar result as in Corollary 4 holds, i.e., for Gaussian sources and channels, linear mappings satisfy the necessary conditions. Perhaps surprisingly, even in the matched bandwidth case, e.g., scalar source, channel and side information, linear mappings are strictly suboptimal. This observation highlights the need for powerful numerical optimization tools.

D. Algorithm Design

A basic approach is to iteratively alternate between the imposition of individual necessary conditions for optimality, and thereby successively decrease the total Lagrangian cost. Imposing optimality condition for the decoder is straightforward, since the decoder can be expressed as closed form functional of known quantities, $\boldsymbol{g}(\cdot)$, $f_X(\cdot)$ and $f_Z(\cdot)$. Since (25) is not in closed form, we perform steepest descent search in the direction of the functional derivative of the Lagrangian with respect to the encoder mapping $\boldsymbol{g}(\cdot)$ as:

$$\boldsymbol{g}_{i+1}(\boldsymbol{x}) = \boldsymbol{g}_i(\boldsymbol{x}) - \mu \nabla \boldsymbol{J}(\boldsymbol{g}, \boldsymbol{h}), \quad (29)$$

where i is the iteration index, $\nabla \boldsymbol{J}(\boldsymbol{g}, \boldsymbol{h})$ is the directional derivative in the direction of mapping $\boldsymbol{g}(\cdot)$ whose exact expression is provided in Appendix A, in (44), and μ is the step size. At each iteration i, the total cost decreases monotonically

[3] To simplify the expressions of the optimality conditions, we assume that $\boldsymbol{h}(\cdot)$ is differentiable *a.e.*, noting that this assumption is not essential.

and iterations are continued until convergence to a local minimum. Note that there is no guarantee that an iterative descent algorithms of this type will converge to the globally optimal solution. A low complexity approach to mitigate the poor local minima problem, is to embed in the solution the noisy relaxation method of [43] and [44]. We initialize the encoding mapping with random initial conditions and run the algorithm at very low CSNR (high Lagrangian parameter λ). We gradually increase the CSNR (decrease λ) while tracking the minimum until we reach the prescribed CSNR (or power P_T for a given channel noise level). The numerical results of this algorithm is presented in Section V.

IV. On Linearity of Optimal Mappings

In this section, we address the problem of *linearity* of the optimal encoding and decoding mappings. Our approach builds on [45], where conditions for linearity of optimal estimation are derived, and on Theorems 2 and 3 of the previous section. Throughout this section, we only study matched source and channel dimension settings, i.e., $m = k$. Let us first provide some background that is particularly relevant to linearity in vector settings. The source and the channel noise have covariance matrices R_X and R_Z, which allow the eigen-decomposition

$$R_X = Q_X \Lambda_X Q_X^T, \text{ and } R_Z = Q_Z \Lambda_Z Q_Z^T, \tag{30}$$

where Q_X and Q_Z are unitary matrices, i.e., they satisfy $Q_X Q_X^T = Q_Z Q_Z^T = I$, and Λ_X and Λ_Z are diagonal matrices whose entries are $\boldsymbol{\lambda}_X = [\lambda_X(1), \ldots, \lambda_X(m)]^T$ and $\boldsymbol{\lambda}_Z = [\lambda_Z(1), \ldots, \lambda_Z(m)]^T$, respectively. We assume[4] that $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_Z$ are inversely ordered, i.e., $\boldsymbol{\lambda}_X(i) \geq \boldsymbol{\lambda}_X(j)$ and $\boldsymbol{\lambda}_Z(j) \geq \boldsymbol{\lambda}_Z(i)$ for all $i < j$.

A. Optimal Linear Mappings

We first briefly revisit the optimal linear encoder and decoder. In the scalar case, $g(X) = k_e X$ and $h(Y) = k_d Y$ where k_e and k_d are given by

$$k_e = \sqrt{\frac{P_T}{\sigma_X^2}}, \quad k_d = \frac{1}{k_e}\left(\frac{P_T}{P_T + \sigma_Z^2}\right). \tag{31}$$

It is well known that if X and Z are scalar Gaussian, $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Z \sim \mathcal{N}(0, \sigma_Z^2)$, $g^*(X) = k_e X$ and $h^*(Y) = k_d Y$ are unbeatable even by coding at asymptotically high delay. In vector settings, derivation of the optimal linear mappings is not as straightforward as the scalar case. Here, we reproduce the classical result due to [17] (see also [18], [46], [47] for alternative derivations of this result).

Theorem 4 ([17]): The encoding-decoding linear transforms that minimize the MSE distortion subject to the total power constraint P_T is

$$K_e = Q_Z \Sigma Q_X^T, \tag{32}$$

and

$$K_d = R_X K_e^T (K_e R_X K_e^T + R_Z)^{-1}, \tag{33}$$

where Σ is a diagonal power allocation matrix that depends on P_T.

B. On Simultaneous Linearity of Optimal Encoder and Decoder

We next show that optimality requires that mappings either both be linear or both nonlinear. In other words, a linear encoder with a nonlinear decoder, or a nonlinear encoder in conjunction with a linear decoder, are both strictly suboptimal. We show this in two steps in the following lemmas.

Lemma 3: $\boldsymbol{g}^*(\boldsymbol{X}) = K_e \boldsymbol{X}$ a.e. in $\boldsymbol{X}$ if $\boldsymbol{h}^*(\boldsymbol{Y}) = K_d \boldsymbol{Y}$.

Proof: Follows directly from Corollaries 2 and 3. □

Lemma 4: $\boldsymbol{h}^*(\boldsymbol{Y}) = K_d \boldsymbol{Y}$ a.e. in $\boldsymbol{Y}$ if $\boldsymbol{g}^*(\boldsymbol{X}) = K_e \boldsymbol{X}$.

Proof: See Appendix C. □

Next, we summarize our main result pertaining to the simultaneous linearity of optimal encoder and decoder.

Theorem 5: The optimal mappings are either both linear or they are both nonlinear.

Proof: The proof directly follows from Lemma 3 and Lemma 4. □

C. Conditions for Linearity of Optimal Mappings: Scalar Settings

In this section, we study the conditions for linearity of optimal encoder and/or decoder. We first focus on the scalar case, $m = k = 1$, and next extend to higher dimensional spaces ($m = k > 1$). The following theorem presents the necessary and sufficient condition for linearity of optimal encoder and decoder mappings.

Theorem 6: For a given power limit P_T, noise Z with variance σ_Z^2 and characteristic function $F_Z(\omega)$, source X with variance σ_X^2 and characteristic function $F_X(\omega)$, the optimal encoding and decoding mappings are linear if and only if

$$F_X(\alpha\omega) = F_Z^{\gamma}(\omega), \tag{34}$$

where $\gamma = \frac{P_T}{\sigma_Z^2}$ and $\alpha = \sqrt{\frac{P_T}{\sigma_X^2}}$.

Proof: See Appendix D. □

We next explore some special cases obtained by varying CNSR (i.e., γ) and utilizing the matching conditions for linearity of optimal mappings given in Theorem 6. We start with a simple but perhaps surprising result.

Theorem 7: Given a source and noise of the same variance, equal to the power limit ($\sigma_X^2 = \sigma_Z^2 = P_T$), the optimal mappings are linear if and only if *the noise and source distributions are identical, i.e., $f_X(x) = f_Z(x)$, a.e. and in which case, the optimal encoder is $g^*(X) = X$ and the optimal decoder is $h^*(\hat{Y}) = \frac{1}{2}\hat{Y}$.*

Proof: It is straightforward to see from (34) that, at $\gamma = 1$, the characteristic functions must be identical. Since the characteristic function uniquely determines the distribution [48], $f_X(x) = f_Z(x)$, *a.e.*. □

[4]This assumption is made only for simplicity in the presentation of the vector linearity conditions, in order to avoid permutation matrices in related expressions.

Remark 2: Note that Theorem 7 holds, albeit at a specific power constraint and second order statistics of the source and the channel, irrespective of the source (and channel) density. This example demonstrates the departure from the well known example of scalar Gaussian source and channel.

Next, we investigate the asymptotic behavior of optimal encoding and decoding functions at low and high CSNR. The results of our asymptotic analysis are of practical importance since they justify, under certain conditions, the use of linear mappings without recourse to complexity arguments at asymptotically high or low CSNR regimes.

Theorem 8: In the limit $\gamma \to 0$, the optimal encoding and decoding functions are asymptotically linear if the channel is Gaussian, regardless of the source. Similarly, as $\gamma \to \infty$, the optimal mappings are asymptotically linear if the source is Gaussian, regardless of the channel.

Proof: The proof follows from applying the central limit theorem [48] to the matching condition (34). The central limit theorem states that as $\gamma \to \infty$, for any finite variance noise Z, the characteristic function of the matching source $F_X(\omega) = F_Z^{\gamma}(\omega/k_e)$ converges to the Gaussian characteristic function. Hence, at asymptotically high CSNR, any noise distribution is matched by the Gaussian source. Similarly, as $\gamma \to 0$ and for any $F_X(k_e\omega)$, $F_X^{\frac{1}{\gamma}}(k_e\omega)$ converges to the Gaussian characteristic function and hence the optimal mappings are asymptotically linear if the channel is Gaussian. □

Let us next consider a setup with given source and noise variables and a power which may be scaled to vary the CSNR, γ. Can the optimal mappings be linear at multiple values of γ ? This question is motivated by the practical setting where γ is not known in advance or may vary (e.g., in the design stage of a communication system). It is well-known that the Gaussian source-Gaussian noise pair makes the optimal mappings linear at all γ levels. Below, we show that this is the only source-channel pair for which the optimal mappings are linear at more than one CSNR value.

Theorem 9: Given X and Z, let power P_T be scaled to vary CSNR, γ. The optimal mappings $g^*(\cdot)$ and $h^*(\cdot)$ are linear at two different power levels P_1 and P_2 if and only if *source and noise are both Gaussian.*

Remark 3: This theorem also holds for the setting where X or Z is scaled to change CSNR for a given power P_T.

Proof: See Appendix E. □

Having discovered the necessary and sufficient condition as answer to the question of *when optimal zero-delay encoding and decoding mappings are linear*, we next focus on the question: *when can we find a matching source (or noise) for a given noise (source)?* Given a valid characteristic function $F_Z(\omega)$, and for some $\gamma \in \mathbb{R}^+$, the function $F_Z^{\gamma}(\omega)$ may or may not be a valid characteristic function, which determines the existence of a matching source. For example, matching is guaranteed for integer γ and it is also guaranteed for infinitely divisible Z. Conditions on γ and $F_Z(\omega)$ for $F_Z^{\gamma}(\omega)$ to be a valid characteristic function were studied in detail in [45], to which we refer for brevity and to avoid repetition.

D. Conditions for Linearity of Communication Mappings: Vector Settings

For a source $\boldsymbol{X} \in \mathbb{R}^m$ with covariance R_X and a channel noise $\boldsymbol{Z} \in \mathbb{R}^m$ with covariance R_Z, we derive the necessary and sufficient condition for simultaneous linearity of optimal encoder and decoder. Similar to the scalar case, we will only investigate the conditions for linearity of optimal decoder given that the encoder is linear due to Theorem 5.

Theorem 10: Let the characteristic functions of the transformed source and channel noise ($\Sigma Q_X^T \boldsymbol{X}$ and $Q_Z^T \boldsymbol{Z}$) be $F_{\Sigma Q_X^T X}(\boldsymbol{\omega})$ and $F_{Q_Z^T Z}(\boldsymbol{\omega})$, respectively. The necessary and sufficient condition for linearity of optimal mappings is:

$$\frac{\partial \log F_{\Sigma Q_X^T X}(\boldsymbol{\omega})}{\partial \omega_i} = S_i \frac{\partial \log F_{Q_Z^T Z}(\boldsymbol{\omega})}{\partial \omega_i}, \quad 1 \le i \le m, \quad (35)$$

where S_i are the elements of the diagonal matrix $S = \Sigma \Lambda_X \Sigma \Lambda_Z^{-1}$, Σ is a diagonal power allocation matrix, Λ_X and Λ_Z are diagonal matrices whose entries are ordered eigenvalues of R_X and R_Z.

Proof: See Appendix F. □

Further insight into the above necessary and sufficient condition is provided via the following corollaries. The first one states that the scalar matching condition, necessary and sufficient for linearity of optimal mappings, is also a necessary condition for each source and channel component in the transform domain, where the transforms render the source and channel components uncorrelated (note that Q_X^T and Q_Z^T are the eigen-transforms of $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively, and Σ is a diagonal matrix.).

Corollary 5: Let $F_{[\Sigma Q_X^T X]_i}(\omega)$ and $F_{[Q_Z^T Z]_i}(\omega)$ be the marginal characteristic functions of the transform coefficients $[\Sigma Q_X^T \boldsymbol{X}]_i$ and $[Q_Z^T \boldsymbol{Z}]_i$, respectively. A necessary condition for linearity of optimal mappings is:

$$F_{[\Sigma Q_X^T X]_i}(\omega) = F^{S_i}_{[Q_Z^T Z]_i}(\omega), 1 \le i \le m. \quad (36)$$

Proof: See Appendix G. □

Another set of necessary conditions is presented in the following corollary.

Corollary 6: A necessary condition for linearity of optimal mappings is that one of the following holds for every pair i, j, $1 \le i, j \le m$:

- i) $S_i = S_j$
- ii) $[Q_X^T \boldsymbol{X}]_i$ *is independent of* $[Q_X^T \boldsymbol{X}]_j$ *and* $[Q_Z^T \boldsymbol{Z}]_i$ *is independent of* $[Q_Z^T \boldsymbol{Z}]_j$.

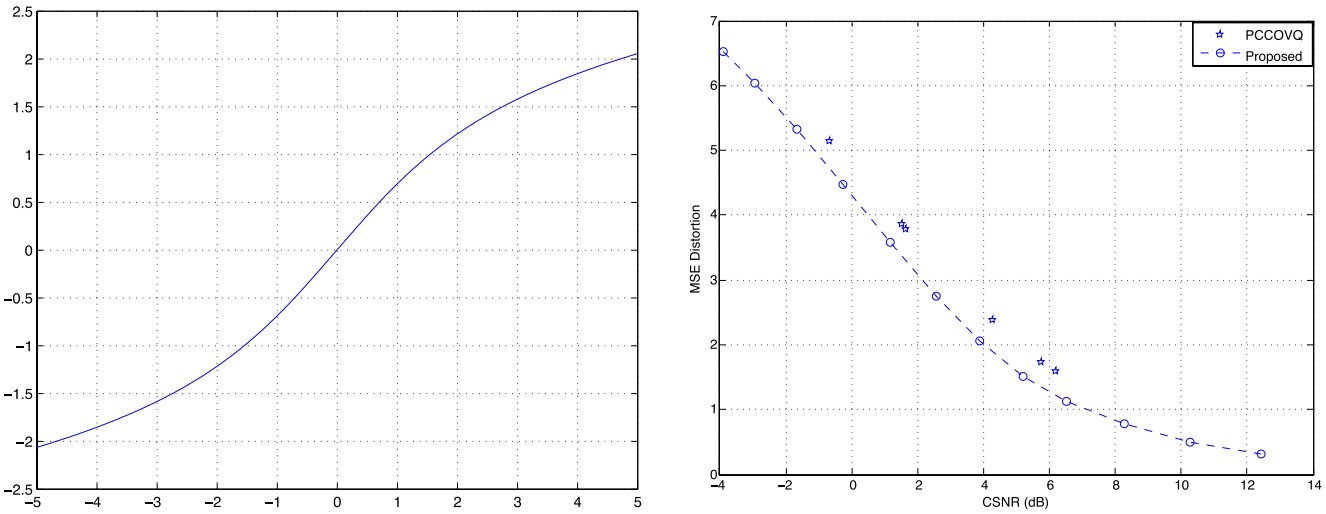
Proof: See Appendix H. □

Note that we only presented necessary conditions so far. In the following, we present a sufficient condition.

Corollary 7: If the necessary condition of Corollary 5 is satisfied, the second condition of the Corollary 6 is sufficient for linearity of optimal mappings.

Proof: Independence of the transform coefficients implies that the joint characteristic function is the product of the marginals:

$$F_{\Sigma Q_X^T}(\boldsymbol{\omega}) = \prod_{i=1}^{m} F_{[\Sigma Q_X^T X]_i}(w_i), \;\; F_{Q_Z^T Z}(\boldsymbol{\omega}) = \prod_{i=1}^{m} F_{[Q_Z^T Z]_i}(w_i). \quad (37)$$



(a) (b)

Fig. 3. Example encoder mapping and comparative results. (a) Encoder mapping for bi-modal scalar GMM source, modes at 3 and -3 as in (38) and a scalar Gaussian channel. (b) Comparative results: PQCOVQ vs. the proposed method.

Plugging (37) into the necessary and sufficient condition (35) of Theorem 10, it is straightforward to show that (36), the necessary condition of Corollary 5, is now both necessary and sufficient. □

Remark 4: While the condition in Corollary 7 requires independence of transform coefficients, the weaker property of uncorrelatedness is already guaranteed by the use of eigen-transformations.

Corollary 8: For Gaussian $\boldsymbol{X}$ and $\boldsymbol{Z}$, linear mappings are optimal, irrespective of R_X, R_Z and P_T.

Proof: Gaussian $\boldsymbol{X}$ and $\boldsymbol{Z}$ satisfy (36) for any S, Q_X and Q_Z. As any linear transform preserves joint Gaussianity in the transform domain, Q_X^T and Q_Z^T generates jointly Gaussian and uncorrelated coefficients which are therefore independent, satisfying the conditions of Corollary 7. □

Remark 5: Although linear mappings are optimal for a Gaussian vector source and channel pair in the zero-delay setting; they may not be, in general, optimal from an information theoretic point of view (asymptotically high delay settings), see [35]. This is in contrast with the scalar Gaussian setting where linear encoding-decoding mappings are optimal even from an information theoretic perspective.

V. Numerical Results

We implement the proposed algorithm by numerically calculating the derived integrals. For that purpose, we sample the source and noise distributions on a uniform grid with a step size $\Delta = 0.01$, i.e., to obtain the numerical results, we approximated the integrals as Riemann sums. We impose bounded support (-5σ to $+5\sigma$) i.e., neglect tails of infinite support distributions in the examples.

A. Scalar Mappings ($m = 1, k = 1$), Gaussian Mixture Source and Gaussian Channel

We consider a Gaussian mixture source with distribution

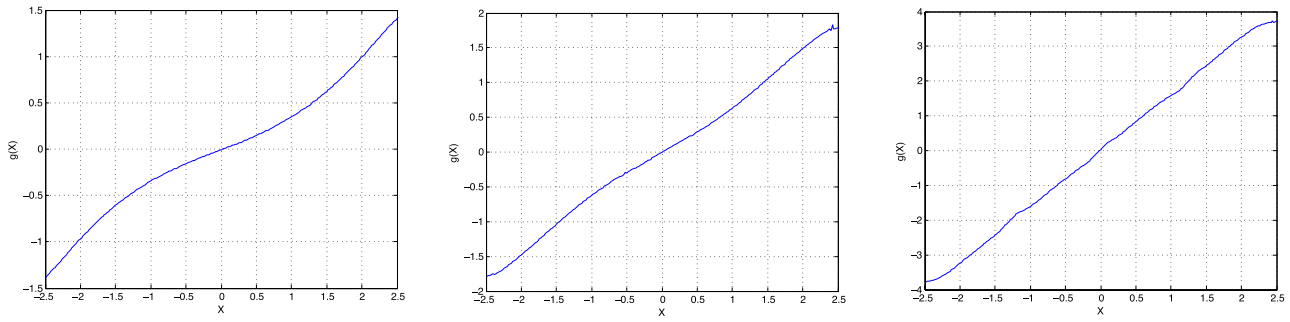
$$f_x(x) = \frac{1}{2\sqrt{2\pi}} \left\{ e^{\frac{-(x-3)^2}{2}} + e^{\frac{-(x+3)^2}{2}} \right\}, \tag{38}$$

and unit variance Gaussian noise. The encoder and decoder mappings for this source-channel setting are numerically obtained as shown in Figure 3(a). As intuitively expected, since the two modes of the Gaussian mixture are well separated, each mode locally behaves as Gaussian. Hence the curve can be approximated as piece-wise linear, deviating significantly from a truly linear mapping. This illustrates the importance of nonlinear mappings for general distributions that diverge from the pure Gaussian.

B. A Numerical Comparison With Vector Quantizer Based Design

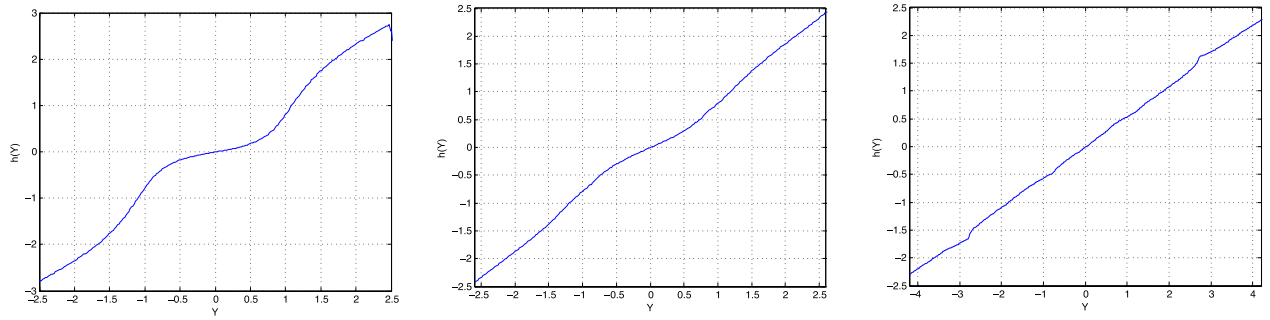
In the following, we compare the proposed approach to the power constrained channel optimized vector quantization (PQCOVQ) based approach which first discretizes the problem, numerically solves the discrete problem and next interpolates between the selected points linearly (see [16]). The main difference between our approach and PQCOVQ based approaches is that we derive the necessary conditions of optimality in the original, "analog" domain without any discretization. This allows not only a theoretical analysis of the problem but also enables a completely different numerical method which iteratively imposes the optimality conditions of the "original problem".

On the other hand, in PQCOVQ, the necessary conditions are derived for the discrete problem, which may not correspond to the original problem if the discretization points are not dense enough. Moreover, it is well known that there are problems with closed form solutions which become NP hard once they are discretized [49]. Indeed, it is straightforward to show that the discrete version of the analog mapping problem can be converted to the discretized Withsenhausen's counterexample in stochastic control, which is known to be NP hard, in polynomial number of steps, and hence the "discretized" analog mapping problem is also NP hard. Admittedly, the analog approach proposed in this paper does not necessarily have lower computational complexity than the discretized



(a) (b) (c)

Fig. 4. This figure shows the optimal encoder at various CSNR values when $X \sim \mathcal{N}(0,1)$ and Z is distributed uniformly on the interval $[-1,1]$ and CSNR is varied by changing power, P. Observe that the optimal encoder converges to linear as CSNR increases. (a) CSNR $= -5.70$dB. (b) CSNR $= -1.69$dB. (c) CSNR $= 5.69$dB.



(a) (b) (c)

Fig. 5. This figure shows the optimal decoder (estimator) at various CSNR (in dB) values. Observe that the optimal decoder, similar to the optimal encoder in Figure 4, converges to linear as CSNR increases. (a) CSNR $= -5.70$dB. (b) CSNR $= -1.69$dB. (c) CSNR $= 5.69$dB.

one (at the same sampling resolution), however the proposed approach allows the discovery of closed form solutions, if they exist in addition to the analysis of the functional properties of the problem.

To compare our method to PQCOVQ, we consider our running example of Gaussian mixture source and Gaussian channel. For both methods, we use 10 sampling points for the encoder mapping. The main difference is due to two facts: i) The proposed method is based on the necessary condition derived in the "original" analog domain, and discretization is merely used to perform the ultimate numerical operations. On the other hand, PQCOVQ defines a "discretized" version of the problem from the outset, with the implicit assumption that the discretized problem, at sufficiently high resolution, approximates well the original problem. Hence, although both methods eventually optimize and interpolate a discrete set of points, the proposed algorithm finds the values of these points while accounting for the fact that they will eventually be (linearly) interpolated. PQCOVQ does not account for eventual interpolation and merely solves the discrete problem. ii) Since we consider the problem in its original domain, we naturally use the optimal decoder, namely, conditional expectation. The PQCOVQ method uses the standard maximum likelihood method for decoding, see [16].

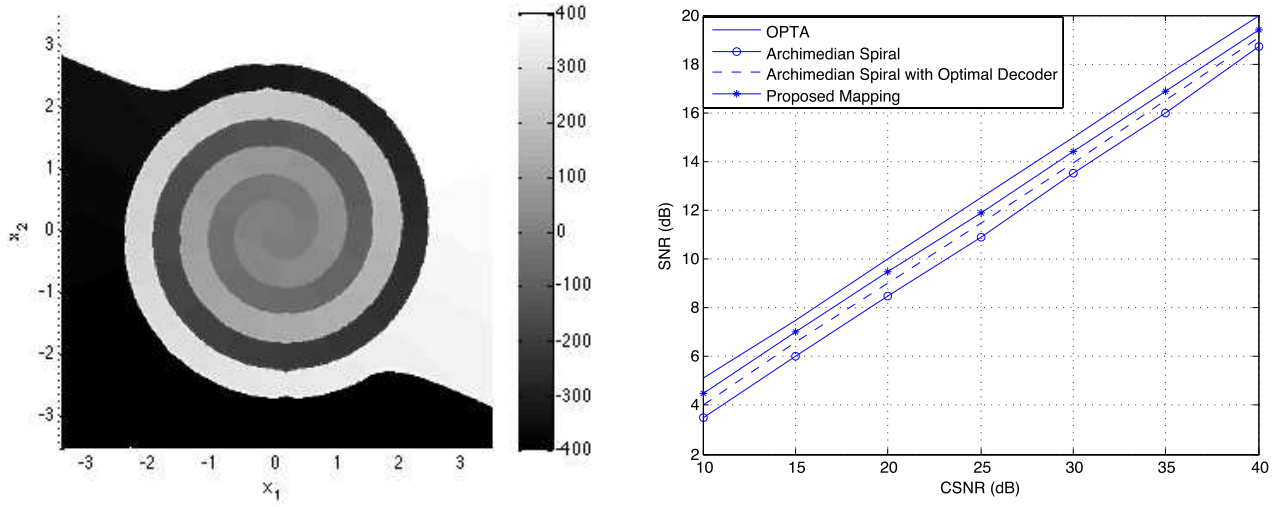
The numerical comparisons are shown in Figure 3(b). As expected, the proposed method outperforms PQCOVQ for the entire range of CNSRs in this resolution constrained setting of 10 samples. We note that the performance difference diminishes at higher sampling resolution. The purpose of this comparison is to demonstrate the conceptual difference between these two approaches at finite resolution while acknowledging that the proposed method does not provide gains at asymptotically high resolution.

C. A Numerical Example for Theorem 8

Let us consider a numerical example that illustrates the findings in Theorem 8. Consider a setting where the channel noise Z is uniform over the interval $[-1,1]$, and the source X is Gaussian with unit variance, i.e., $X \sim \mathcal{N}(0,1)$. We change γ (CSNR) by varying allowed power P_T, and observe how the optimal mappings behave for different γ. Figures 4 and 5 respectively show how the optimal encoder and decoder mappings converge to linear as CSNR increases. Note that at $\gamma = -5.70$, optimal mappings are both highly nonlinear while at $\gamma = 5.69$, they practically converge to linear, as theoretically anticipated from Theorem 8.

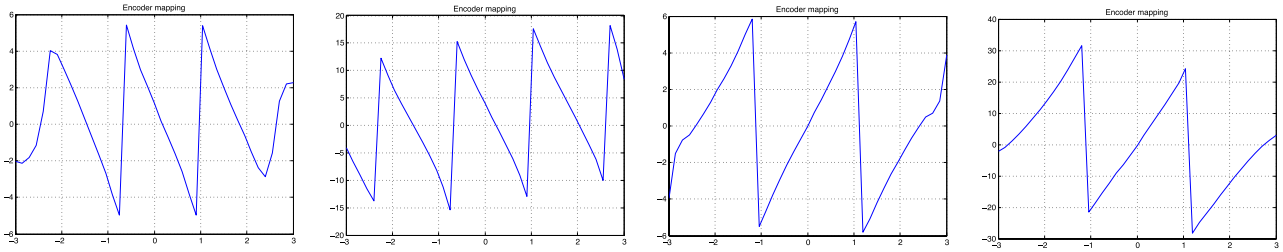
D. $(m = 2, k = 1)$ Gaussian Source-Channel Mapping

In this section, we present a bandwidth compression example with 2:1 mappings for Gaussian vector source of size two (source samples are assumed to be independent and identically distributed with unit variance) and scalar Gaussian channel, to demonstrate the effectiveness of our algorithm in differing source and channel dimensions. We compare the proposed mapping to the asymptotic bound (OPTA) and prior work [50]. We also compare the optimal encoder-decoder pair to the setting where only the decoder is optimized and the



(a) (b)

Fig. 6. Example encoder mapping and comparative results. (a) Encoder 2:1 mapping for unit variance Gaussian source and channel, at CSNR=40dB, SNR=19.41dB. The axes show the two dimensional input ($\boldsymbol{x}$) and the function value ($\boldsymbol{g}(\boldsymbol{x})$) is reflected in the intensity level. (b) Comparative results for Gaussian source-channel, 2:1 mapping.



(a) (b) (c) (d)

Fig. 7. Encoder mappings for Gaussian scalar source, channel and side information at different CSNR and correlation levels. (a) CSNR = 10dB, $\rho = 0.97$. (b) CSNR = 22dB, $\rho = 0.97$. (c) CSNR = 10dB, $\rho = 0.9$. (d) CSNR = 23dB, $\rho = 0.9$.

encoder is fixed. In prior work [9], [11], [50], the Archimedian spiral is found to perform well for Gaussian 2:1 mappings, and used for encoding and decoding with maximum likelihood criteria.

The obtained encoder mapping is shown in Figure 6(a). While the mapping produced by our algorithm resembles a spiral, it nevertheless differs from the Archimedian spiral, as will be evident from the performance results. Note further that the encoding scheme differs from prior work in that we continuously map the source to the channel signal, where the two dimensional source is mapped to the nearest point on the space filling spiral. The comparative performance results are shown in Figure 6(b). The proposed mapping outperforms the Archimedian spiral [50] over the entire range of CSNR values. It is notable that the "intermediate" option of only optimizing the decoder captures a significant portion of the gains.

E. Source-Channel Coding With Decoder Side Information

In this section, we demonstrate the use of the proposed algorithm by focusing on the specific scenario of Figure 2. While the proposed algorithm is general and directly applicable to any choice of source and channel dimensions and distributions, for conciseness of the results section, we assume that sources are jointly Gaussian scalars with correlation coefficient ρ, and the channel is scalar Gaussian as described in Section II.B.

Figure 7 presents a sample of encoding mappings obtained by varying the correlation coefficient and CSNR. Interestingly, the analog mapping captures the central characteristic observed in digital Wyner-Ziv mappings, in the sense of many-to-one mappings, where multiple source intervals are mapped to the same channel interval, which will potentially be resolved by the decoder given the side information. To see the effect of correlation on the encoding mappings, we lower the correlation from $\rho = 0.97$ to $\rho = 0.9$. As intuitively expected, the side information is less reliable and source points that are mapped to the same channel representation grow further apart from each other. Comparative results in Figure 8 show that the proposed mapping outperforms linear mapping over the entire range of CSNR values. We note that this characteristic of the encoding mappings was also noted in experiments with the PCCOVQ approach in [16], and was implemented in [51], for optimizing hybrid (digital + analog) mappings.

VI. Discussion and Future Work

In this paper, we studied the zero-delay source-channel coding problem. First, we derived the necessary conditions

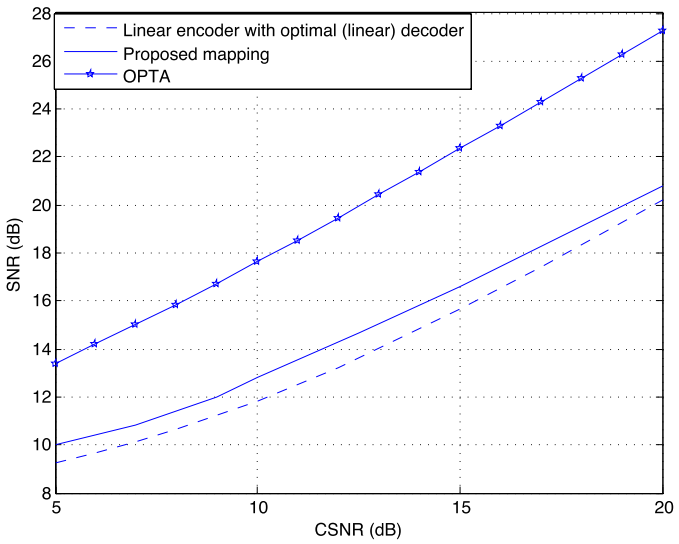


Fig. 8. Comparative results for correlation coefficient $\rho = 0.9$, Gaussian scalar source, channel and side information

for the optimality of the encoding and decoding mappings for a given source-channel system. Based on the necessary conditions, we proposed an iterative algorithm which generates locally optimal encoder and decoder mappings. Comparative results and example mappings are provided and it is shown that the proposed method improves upon the results of prior work. Moreover, we investigated the functional properties of the zero-delay source-channel coding problem. Using these functional properties and the necessary conditions of optimality we had derived, we obtained the necessary and sufficient condition for linearity of optimal mappings. We studied the implications of this matching condition and particularly showed that the optimal mappings converge to linear at asymptotically high CSNR for a Gaussian source, irrespective of the channel density and similarly for a Gaussian channel, at asymptotically low CSNR, irrespective of the source. We also extended our analysis to vector spaces.

The numerical algorithm presented in this paper is feasible for relatively low source and channel dimensions (m, k). For high dimensional vector spaces, the numerical approach should be supported by imposing a tractable structure to the mappings, to mitigate the problem of the dimensionality. A set of preliminary results in this direction appeared in [47], where a linear transformation followed by scalar non-linear mappings were utilized for the decoder side information setting. The purely linear solution had been investigated in [52], where numerical algorithms are proposed to find the optimal bandwidth compression transforms in network settings. The analysis in this paper, specifically conditions for linearity (and generalizations to other structural forms) of optimal mappings, as well as the numerical approach, can be extended to well known control problems such as the optimal jamming problem [53] and Witsenhausen's counterexample [30], [31]. See [39], [54] for preliminary results in these directions.

An interesting question pertains to the existence of structure in the optimal mappings in some fundamental scenarios. For instance, in [51], a hybrid digital-analog encoding was employed for the problem of zero-delay source-channel coding with decoder side information, where the source, the side information and the channel noise are all scalar and Gaussian. The reported performance results are very close to the performance of the optimal unconstrained mappings. In contrast, in [16], a sawtooth-like structure was assumed and its parameters were optimized as well as PCCOVQ was employed to obtain the non-structured mappings, where non-negligible performance difference between these approaches was reported. Hence, this fundamental question, on whether the optimal zero-delay mappings are structured for the scalar Gaussian side information setting, is currently open. This problem can numerically be approached by employing a powerful non-convex optimization tool, such as deterministic annealing [55]. Preliminary results in this direction appeared in [56] and [57].

Appendix A
Proof of Theorem 3

Let $\boldsymbol{g}(\cdot)$ be fixed. The optimal decoder is the MMSE estimator of $\boldsymbol{X}$ given $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}}$ which can be written, using Bayes' rule and noting that $f_{\hat{Y}|X}(\boldsymbol{x}, \hat{\boldsymbol{y}}) = f_Z(\hat{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{x}))$ as

$$\boldsymbol{h}(\hat{\boldsymbol{y}}) = \frac{\int \boldsymbol{x}\, f_X(\boldsymbol{x})\, f_Z(\hat{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}}{\int f_X(\boldsymbol{x})\, f_Z(\hat{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}. \tag{39}$$

Let $\boldsymbol{h}(\cdot)$ be fixed. Applying the standard method in variational calculus [58] to the cost function defined in (3), we have

$$\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} J\left(\boldsymbol{g}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta}(\boldsymbol{x}), \boldsymbol{h}\right) = 0, \tag{40}$$

i.e., we perturb the cost functional for all admissible[5] variation functions $\boldsymbol{\eta}(\boldsymbol{x})$. Since the power constraint is accounted for in the cost function, the variation function $\boldsymbol{\eta}(\cdot)$ needs not be restricted to satisfy the power constraint (all measurable functions $\boldsymbol{\eta} : \mathbb{R}^m \to \mathbb{R}^k$ are admissible). Applying (40), as shown at the top of the next page. we get (41). Evaluating (41) at $\epsilon = 0$, we have (42), as shown at the top of the next page, where $\boldsymbol{h}'(\cdot)$ denotes the Jacobian of the vector valued function $\boldsymbol{h}(\cdot)$. Equality for all admissible variation functions, $\boldsymbol{\eta}(\cdot)$, requires the expression in braces to be identically zero (more formally the functional derivative [58] vanishes at an extremum point of the functional). Hence, we have

$$\nabla J(\boldsymbol{g}, \boldsymbol{h}) = 0, \tag{43}$$

where

$$\nabla J(\boldsymbol{g}, \boldsymbol{h}) = \lambda f_X(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x}) - f_X(\boldsymbol{x}) \int \boldsymbol{h}'(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{z}))f_Z(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\,. \tag{44}$$

[5]Our admissibility definition does not need to be very restrictive since it is used to derive a necessary condition. Hence, the only condition required for the admissible functions is to be (Borel) measurable, that the integrals exist, and that we can change the order of integration and differentiation.

$$\left\{\int\left\{\lambda\left(\boldsymbol{g}(\boldsymbol{x})+\epsilon\boldsymbol{\eta}(\boldsymbol{x})\right)-\int \boldsymbol{h}'(\boldsymbol{g}(\boldsymbol{x})+\epsilon\boldsymbol{\eta}(\boldsymbol{x})+\boldsymbol{z})\left(\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x})+\epsilon\boldsymbol{\eta}(\boldsymbol{x})+\boldsymbol{z})\right) f_Z(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\right\}\boldsymbol{\eta}(\boldsymbol{x}) f_X(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\}\Big|_{\epsilon=0}=0, \quad (41)$$

$$\int\left\{\lambda\boldsymbol{g}(\boldsymbol{x})-\int \boldsymbol{h}'(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{z})\left(\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{z})\right) f_Z(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\right\}\boldsymbol{\eta}(\boldsymbol{x}) f_X(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0 \quad (42)$$

$$\nabla J(\boldsymbol{x}_1,\boldsymbol{x}_2)=\lambda f_{X_1,X_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{g}(\boldsymbol{x}_1)-\int \boldsymbol{h}'(\boldsymbol{g}(\boldsymbol{x}_1)+\boldsymbol{z},\boldsymbol{x}_2)\left(\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{x}_1)+\boldsymbol{z},\boldsymbol{x}_2)\right) f_Z(\boldsymbol{z}) f_{X_1,X_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{z}. \quad (50)$$

APPENDIX B
OPTIMALITY CONDITIONS FOR CODING WITH DECODER SIDE INFORMATION

Let the encoder $\boldsymbol{g}(\cdot)$ be fixed. The optimal decoder is

$$\boldsymbol{h}(\hat{\boldsymbol{y}},\boldsymbol{x}_2)=\mathbb{E}\{\boldsymbol{X}_1|\hat{\boldsymbol{y}},\boldsymbol{x}_2\} \quad (45)$$

$$=\frac{\int \boldsymbol{x}_1 f_{X_1,X_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) f_Z(\hat{\boldsymbol{y}}-\boldsymbol{g}(\boldsymbol{x}_1))\,\mathrm{d}\boldsymbol{x}_1}{\int f_{X_1,X_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) f_Z(\hat{\boldsymbol{y}}-\boldsymbol{g}(\boldsymbol{x}_1))\,\mathrm{d}\boldsymbol{x}_1}. \quad (46)$$

We next consider the distortion functional

$$D=\mathbb{E}\{||\boldsymbol{X}_1-\boldsymbol{h}(\boldsymbol{g}(\boldsymbol{X}_1)+\boldsymbol{Z},\boldsymbol{X}_2)||^2\}, \quad (47)$$

and the Lagrangian cost to minimize:

$$J=D+\lambda P. \quad (48)$$

Assuming the decoder $\boldsymbol{h}(\cdot)$ is fixed, we have

$$\nabla J(\boldsymbol{x}_1,\boldsymbol{x}_2)=0,\ \forall \boldsymbol{x}_1,\ \boldsymbol{x}_2, \quad (49)$$

where $\nabla J(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is given in (50), as shown at the top of this page.

APPENDIX C
PROOF OF LEMMA 4

We prove this lemma in two steps. First, we focus on the scalar setting and prove the lemma for $m=1$. Next, we extend the analysis to vector setting. Plugging $g(x)=k_e x$ in (25), we obtain

$$\lambda k_e x=\int (x-h(k_e x+z))h'(k_e x+z) f_Z(z)\mathrm{d}z, \quad (51)$$

a.e. in x. Since $h(\cdot)$ is a continuous[6] function from $\mathbb{R}\to\mathbb{R}$, the Weierstrass theorem [59] guarantees that there is a sequence of real valued polynomials that uniformly converges to it:

$$h(y)=\lim_{i\to\infty}\sum_{r=0}^{\infty}\alpha_r(i)y^r, \quad (52)$$

where $\alpha_r(i)\in\mathbb{R}$ is the r^{th} polynomial coefficient of the i^{th} polynomial. Since Weierstrass convergence is uniform in y, we can interchange the limit and summation and hence,

$$h(y)=\sum_{r=0}^{\infty}\alpha_r y^r, \quad (53)$$

a.e. in y, where $\alpha_r=\lim_{i\to\infty}\alpha_r(i)$. Plugging (53) in (51) we obtain

$$\lambda k_e x=\int\left(x-\sum_{i=0}^{\infty}\alpha_i(k_e x+z)^i\right)\times\left(\sum_{i=0}^{\infty} i\alpha_i(k_e x+z)^{i-1}\right) f_Z(z)\mathrm{d}z. \quad (54)$$

Interchanging the summation and integration,[7] we have

$$-\lambda k_e x+x-\sum_{i=0}^{\infty} i\alpha_i\int(k_e x+z)^{i-1} f_Z(z)\mathrm{d}z=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i\alpha_i\alpha_j\int(k_e x+z)^{i-1}(k_e x+z)^j f_Z(z)\mathrm{d}z. \quad (55)$$

Note that the above equation must hold *a.e.* in x, hence the coefficients of x^r must be identical for all $r\in\mathbb{N}$. Expanding the expressions $(k_e x+z)^{i-1}$ and $(k_e x+z)^j$ via binomial expansion, we have the following set of equations

$$\sum_{i=r+1}^{\infty} i\binom{i-1}{r}\alpha_i\mathbb{E}\{Z^{i-1-r}\}=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{i-1}\sum_{p=r-l+1}^{j-1}\binom{j}{p}\binom{i-1}{l} i\alpha_i\alpha_j\mathbb{E}\{Z^{i+j-1-p-l}\}, \quad (56)$$

which must hold for all $r\geq 2$.

We note that every equation introduces a new variable α_r, so each new equation is linearly independent of its predecessors. Next, we solve these equations recursively, starting from $r=1$. At each r, we have one unknown (α_r) which is related "linearly" to known constants. Since the number of linearly independent equations is equal to the number of unknowns for each r, there must exist a unique solution. We know that $\alpha_r=0$, for all $r\geq 2$ is a solution to (56), so it is the only solution.

Having proved the scalar version of the proposition, we extend it to vector spaces by contradiction. Let us assume $\boldsymbol{g}^*(\boldsymbol{X})=K_e\boldsymbol{X}$ and $\boldsymbol{h}^*(\boldsymbol{Y})\neq K_d\boldsymbol{Y}$. It can be shown, using (25) and (26), that $g^*(X_i)=\mathbb{E}\{\boldsymbol{g}^*(\boldsymbol{X})\}$ and $h^*(Y_i)=E\{\boldsymbol{h}(\boldsymbol{Y})\}$ where the expectations are taken over the joint distributions $\boldsymbol{X}-\{X_i\}$ and $\boldsymbol{Y}-\{Y_i\}$, respectively for all i. Noting that

[6]We assume continuous $h(\cdot)$, however this assumption is not essential since continious functions are dense in $\mathbb{R}^2$, i.e., we can approximate any function with bounded second order moments, as the limit of a sequence of continuous functions with arbitrarily small l_2 error [59].

[7]Since the polynomials $\sum_{i=0}^{\infty}\alpha_i(k_e x+z)^i$ and $\sum_{i=0}^{\infty} i\alpha_i(k_e x+z)^{i-1}$ respectively converge to $h(k_e x+z)$ and $h'(k_e x+z)$ uniformly in x and z, and hence both upper bounded in magnitude, we can use Lebesgue's dominated convergence theorem to interchange the summation and the integration.

there exists an i such that $\boldsymbol{g}^*(X_i) = \mathbb{E}\{\boldsymbol{g}^*(\boldsymbol{X})\} = k_e X_i$ and $\boldsymbol{h}^*(Y_i) = \mathbb{E}\{\boldsymbol{g}^*(\boldsymbol{X})\} \neq k_d Y_i$ since $\boldsymbol{g}^*(\boldsymbol{X}) = K_e \boldsymbol{X}$ and $\boldsymbol{h}^*(\boldsymbol{Y}) \neq K_d \boldsymbol{Y}$. This contradicts with the scalar version of this proposition and hence if $\boldsymbol{g}^*(\boldsymbol{X}) = K_e \boldsymbol{X}$ then $\boldsymbol{h}^*(\boldsymbol{Y}) = K_d \boldsymbol{Y}$ for any m.

Appendix D
Proof of Theorem 6

Theorem 5 states that the optimal $\boldsymbol{g}(\cdot)$ is linear *if and only if* the optimal $\boldsymbol{h}(\cdot)$ is linear. Hence, we only focus on the case where encoder and decoder are simultaneously linear. Given Corollary 3, only the second necessary condition, (26) remains to be verified. Plugging $g(X) = k_e X$ and $h(\hat{Y}) = k_d \hat{Y}$ in (26), we have

$$k_d \hat{y} = \frac{\int x \, f_X(x) \, f_Z(\hat{y} - k_e x) \mathrm{d}x}{\int \, f_X(x) \, f_Z(\hat{y} - k_e x) \, \mathrm{d}x}. \tag{57}$$

Opening up (57), we obtain

$$k_d \hat{y} \int \, f_X(x) \, f_Z(\hat{y} - k_e x) \, \mathrm{d}x = \int x \, f_X(x) \, f_Z\left(\hat{y} - k_e x\right) \, \mathrm{d}x. \tag{58}$$

Taking the Fourier transform of both sides and via change of variables $u \triangleq \hat{y} - k_e x$, we have

$$\begin{aligned} &\int \int k_d (u + k_e x) f_X(x) f_Z(u) \exp(-j\omega(u + k_e x)) \mathrm{d}x \mathrm{d}u \\ &\quad = \int \int x f_X(x) f_Z(u) \exp(-j\omega(u + k_e x)) \mathrm{d}x \mathrm{d}u, \end{aligned} \tag{59}$$

and rearranging the terms, we obtain

$$\left(\frac{1 - k_e k_d}{k_e k_d}\right) F_Z(\omega) F_X'(k_e \omega) = F_X(k_e \omega) F_Z'(\omega). \tag{60}$$

Noting that

$$\gamma = \frac{P_T}{\sigma_Z^2} = \frac{k_e k_d}{1 - k_e k_d}, \tag{61}$$

we have

$$\frac{F_X'(k_e \omega)}{F_X(k_e \omega)} = \gamma \, \frac{F_Z'(\omega)}{F_Z(\omega)}, \tag{62}$$

which implies

$$(\log F_X(k_e \omega))' = (\log F_Z^{\gamma}(\omega))'. \tag{63}$$

The solution to this differential equation is

$$\log F_X(k_e \omega) = \log F_Z^{\gamma}(\omega) + \theta, \tag{64}$$

where θ is constant. Noting that $F_X(0) = F_Z(0) = 1$, we determine $\theta = 0$ and hence

$$F_X(k_e \omega) = F_Z^{\gamma}(\omega). \tag{65}$$

Since the solution is essentially unique, due to Corollary 1, (65) is not only a necessary but also the sufficient condition for linearity of optimal mappings.

Appendix E
Proof of Theorem 9

Let γ_1 and γ_2 denote two CSNR levels, $g_1(X) = k_{e_1} X$ and $g_2(X) = k_{e_2} X$ denote encoding mappings. Let the power be scaled by α^2 ($\alpha \in \mathbb{R}^+$), i.e., $P_2 = \alpha^2 P_1$ which yields

$$\gamma_2 = \alpha^2 \gamma_1, \;\; k_{e_2} = \alpha k_{e_1}. \tag{66}$$

Using (34), we have

$$F_X(k_{e_1} \omega) = F_Z^{\gamma_1}(\omega), \, F_X(k_{e_2} \omega) = F_Z^{\gamma_2}(\omega). \tag{67}$$

Hence,

$$F_Z^{\gamma_1}(\omega) = F_Z^{\gamma_2}(\alpha \omega). \tag{68}$$

Taking the logarithm on both sides of (68), applying (66) and rearranging terms, we obtain

$$\alpha^2 = \frac{\log F_Z(\alpha \omega)}{\log F_Z(\omega)}. \tag{69}$$

Note that (69) should be satisfied for both α and $-\alpha$ since they yield the same γ. Hence, $F_Z(\alpha\omega) = F_Z(-\alpha\omega)$ for all $\alpha \in \mathbb{R}$, which implies $F_Z(\omega) = F_Z(-\omega)$, *a.e.* in $\omega \in \mathbb{R}$. Using the fact that the characteristic function is conjugate symmetric (i.e., $F_Z(-\omega) = F_Z^*(\omega)$), we get $F_Z(\omega) \in \mathbb{R}$, *a.e.* in ω. As $\log F_Z(\omega)$ is a function from $\mathbb{R} \to \mathbb{C}$, the Weierstrass theorem [59] guarantees that we can uniformly approximate $\log F_Z(\omega)$ arbitrarily closely by a polynomial $\sum_{i=0}^{\infty} k_i \omega^i$, where $k_i \in \mathbb{C}$. Hence, by (69) we obtain:

$$\alpha^2 = \frac{\sum_{i=0}^{\infty} k_i (\omega\alpha)^i}{\sum_{i=0}^{\infty} k_i \omega^i}, \tag{70}$$

a.e. in ω only if all coefficients k_i vanish, except for k_2, i.e., $\log F_Z(\omega) = k_2 \omega^2$, or $\log F_Z(\omega) = 0$ *a.e.* in $\omega \in \mathbb{R}$ (the solution $\alpha = 1$ is of no interest). The latter is not a characteristic function, and the former is the Gaussian characteristic function, $F_Z(\omega) = e^{k_2 \omega^2}$, where we use the established fact that $F_Z(\omega) \in \mathbb{R}$. Since a characteristic function determines the distribution uniquely, the Gaussian source and noise must be the only allowable pair.

Appendix F
Proof of Theorem 10

First, we derive the necessary and sufficient conditions for linearity of optimal decoder given a linear encoder $\boldsymbol{g}(\boldsymbol{X}) = K_e \boldsymbol{X}$. Let us rewrite the MSE optimal decoder, $\boldsymbol{h}(\boldsymbol{y}) = \mathbb{E}\{\boldsymbol{X}|\boldsymbol{y}\}$ using Bayes' rule and independence of $\boldsymbol{X}$ and $\boldsymbol{Z}$:

$$\boldsymbol{h}(\boldsymbol{y}) = \frac{\int \boldsymbol{x} \, f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - K_e \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{\int f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - K_e \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}. \tag{71}$$

Plugging $\boldsymbol{h}(\boldsymbol{y}) = K_d \boldsymbol{y}$ in (71) we obtain,

$$K_d \boldsymbol{y} \int f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - K_e \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int \boldsymbol{x} \, f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - K_e \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{72}$$

Taking the Fourier transform of both sides, we have

$$jK_d\nabla\left[F_X(K_e\boldsymbol{\omega})F_Z(\boldsymbol{\omega})\right] = jK_e^{-1}[\nabla F_X(K_e\boldsymbol{\omega})]F_Z(\boldsymbol{\omega}). \quad (73)$$

Rearranging terms, we get

$$\left(K_e^{-1}-K_d\right)\frac{1}{F_X(K_e\boldsymbol{\omega})}\nabla F_X(K_e\boldsymbol{\omega}) = K_d\frac{1}{F_Z(\boldsymbol{\omega})}\nabla F_Z(\boldsymbol{\omega}). \quad (74)$$

Using $\nabla\log F_X(K_e\boldsymbol{\omega}) = \frac{1}{F_X(K_e\boldsymbol{\omega})}\nabla F_X(K_e\boldsymbol{\omega})$,

$$\nabla\log F_X(K_e\boldsymbol{\omega}) = (K_e^{-1}-K_d)^{-1}K_d\nabla\log F_Z(\boldsymbol{\omega}). \quad (75)$$

Let us define $\zeta \triangleq (K_e^{-1}-K_d)^{-1}K_d$ where K_d is given in (33). We have

$$\begin{aligned}\zeta &= (K_e^{-1}-R_XK_e^T(K_eR_XK_e^T+R_Z)^{-1})^{-1}\\ &\quad\times R_XK_e^T(K_eR_XK_e^T+R_Z)^{-1}. \end{aligned}\quad (76)$$

It is straightforward to see that

$$\zeta^{-1} = R_Z(K_eR_XK_e^T)^{-1}, \quad (77)$$

and hence

$$\zeta = K_eR_XK_e^TR_Z^{-1}. \quad (78)$$

Plugging (32) in (78), we have

$$\zeta = Q_Z\Sigma\Lambda_X\Sigma\Lambda_Z^{-1}Q_Z^T, \quad (79)$$

and hence

$$\nabla\log F_X(K_e\boldsymbol{\omega}) = Q_Z(\Sigma\Lambda_X\Sigma\Lambda_Z^{-1})Q_Z^T\nabla\log F_Z(\boldsymbol{\omega}). \quad (80)$$

Defining $S \triangleq \Sigma\Lambda_X\Sigma\Lambda_Z^{-1}$, we have

$$Q_Z^T\nabla v_1(\boldsymbol{\omega}) = SQ_Z^T\nabla\log F_Z(\boldsymbol{\omega}), \quad (81)$$

where $v_1(\boldsymbol{\omega}) = \log F_X(K_e\boldsymbol{\omega})$. Let us define $\tilde{\boldsymbol{\omega}} \triangleq Q_Z^T\boldsymbol{\omega}$, hence $\boldsymbol{\omega} = Q_Z\tilde{\boldsymbol{\omega}}$. Plugging this in (81), we have

$$Q_Z^T\nabla v_1(Q_Z\tilde{\boldsymbol{\omega}}) = SQ_Z^T\nabla\log F_Z(Q_Z\tilde{\boldsymbol{\omega}}). \quad (82)$$

At this point, we need the following auxiliary lemma, whose proof appeared in the Appendix E of [45]:

Lemma 5 ([45]): Given a function $f:\mathbb{R}^n\to\mathbb{R}$, matrix $A\in\mathbb{R}^{n\times m}$, vector $\boldsymbol{x}\in\mathbb{R}^m$, and $v(\boldsymbol{x}) = f(A\boldsymbol{x})$.

$$\nabla v(\boldsymbol{x}) = A^T\nabla f(A\boldsymbol{x}). \quad (83)$$

Using Lemma 5, we can rewrite (82) as

$$\nabla v_2(\tilde{\boldsymbol{\omega}}) = S\nabla v_3(\tilde{\boldsymbol{\omega}}), \quad (84)$$

where $v_2(\boldsymbol{\omega}) = \log F_X(K_e^TQ_Z\boldsymbol{\omega})$ and $v_3(\boldsymbol{\omega}) = \log F_Z(Q_Z\boldsymbol{\omega})$. Noting that $K_e^TQ_Z = Q_X\Sigma$ and the characteristic functions of the source and noise after transformation can be written in terms of the known characteristic functions $F_X(\boldsymbol{\omega})$ and $F_Z(\boldsymbol{\omega})$, specifically $F_{\Sigma Q_X^TX}(\boldsymbol{\omega}) = F_X(\Sigma Q_X^T\boldsymbol{\omega})$ and $F_{Q_Z^TZ}(\boldsymbol{\omega}) = F_Z(Q_Z\boldsymbol{\omega})$, we have

$$\nabla\log F_{\Sigma Q_X^TX}(\tilde{\boldsymbol{\omega}}) = S\nabla\log F_{Q_Z^TZ}(\tilde{\boldsymbol{\omega}}). \quad (85)$$

Using the fact that S is diagonal, we convert (85) to the set of m scalar differential equations of (35). Converse can be shown by retracing the steps in the derivation of the necessary condition. Note that none of these steps, (71)-(85), introduce any loss of generality, hence retracing back from (85) to (71), we show that (35) implies that the optimal decoder is linear if the encoder is linear. The dual part, i.e., the proof of linearity of optimal encoder, given the optimality of a linear decoder follows from Theorem 5.

Appendix G
Proof of Corollary 5

The marginal characteristic functions of $[\Sigma Q_X^T\boldsymbol{X}]_i$ and $[Q_Z^T\boldsymbol{Z}]_i$ are obtained by setting $\omega_k = 0, \forall k\neq i$ in $F_{\Sigma Q_X^TX}(\boldsymbol{\omega})$ and $F_{Q_Z^TZ}(\boldsymbol{\omega})$ respectively. By setting $\omega_k = 0, \forall k\neq i$ in both sides of (35), we have

$$\frac{\partial\log F_{[\Sigma Q_X^TX]_i}(\omega)}{\partial\omega} = S_i\frac{\partial\log F_{[Q_Z^TZ]_i}(\omega)}{\partial\omega}, \quad 1\le i\le m \quad (86)$$

The solution to this differential equation is:

$$\log F_{[\Sigma Q_X^TX]_i}(\omega) = S_i\log F_{[Q_Z^TZ]_i}(\omega)+\theta, \quad (87)$$

where θ is a constant. Imposing $F_{[Q_Z^TZ]_i}(0) = F_{[U\Sigma Q_X^T]_i}(0) = 1$, we obtain $\theta = 0$, which implies:

$$F_{[\Sigma Q_X^TX]_i}(\omega) = F_{[Q_Z^TZ]_i}^{S_i}(\omega), 1\le i\le m. \quad (88)$$

Appendix H
Proof of Corollary 6

Let us rewrite (35) explicitly for the i^{th} and j^{th} coefficients:

$$\frac{\partial\log F_{\Sigma Q_X^TX}(\boldsymbol{\omega})}{\partial\omega_i} = S_i\frac{\partial\log F_{Q_Z^TZ}(\boldsymbol{\omega})}{\partial\omega_i}. \quad (89)$$

$$\frac{\partial\log F_{\Sigma Q_X^TX}(\boldsymbol{\omega})}{\partial\omega_j} = S_j\frac{\partial\log F_{Q_Z^TZ}(\boldsymbol{\omega})}{\partial\omega_j}. \quad (90)$$

Taking the partial derivatives of both sides of (89) with respect to ω_j, and both sides of (90) with respect to ω_i, we obtain the following:

$$\frac{\partial^2\log F_{\Sigma Q_X^TX}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j} = S_i\frac{\partial^2\log F_{Q_Z^TZ}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j}, \quad (91)$$

$$\frac{\partial^2\log F_{\Sigma Q_X^TX}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j} = S_j\frac{\partial^2\log F_{Q_Z^TZ}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j}. \quad (92)$$

There are only two ways to simultaneously satisfy (91) and (92): i) $S_i = S_j$, ii) the second-order derivatives vanish, i.e.,

$$\frac{\partial^2\log F_{\Sigma Q_X^TX}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j} = 0. \quad (93)$$

$$\frac{\partial^2\log F_{Q_Z^TZ}(\boldsymbol{\omega})}{\partial\omega_i\partial\omega_j} = 0. \quad (94)$$

Let us focus on $\boldsymbol{X}$ i.e., (93); derivation for $\boldsymbol{Z}$ follows similarly. $F_{[\Sigma Q_X^TX]_{ij}}(\omega_i,\omega_j)$, i.e., the marginal characteristic function of the pair $([\Sigma Q_X^T\boldsymbol{X}]_i, [\Sigma Q_X^T\boldsymbol{X}]_j)$ is obtained by setting $\omega_k = 0$, $\forall k\neq i, j$. Note that (93) implies

$$\frac{\partial^2\log F_{[\Sigma Q_X^TX]_{ij}}(\omega_i,\omega_j)}{\partial\omega_i\partial\omega_j} = 0, \quad (95)$$

which implies that

$$\log F_{[\Sigma Q_X^T X]_{ij}}(\omega_i, \omega_j) = A(\omega_i) + B(\omega_j), \tag{96}$$

for some functions A and B, i.e., $\log F_{[\Sigma Q_X^T X]_{ij}}(\omega_i, \omega_j)$ is additively separable in terms of ω_i and ω_j. This implies that

$$F_{[\Sigma Q_X^T X]_{ij}}(\omega_i, \omega_j) = C(\omega_i) D(\omega_j), \tag{97}$$

for some functions C and D. But (97) implies independence of the i^{th} and j^{th} transform coefficients of source $\boldsymbol{X}$. The independence of the i^{th} and j^{th} transform coefficients of the noise $\boldsymbol{Z}$ follows from similar arguments. By the fact that Σ is merely a diagonal scaling matrix, not effecting independence, we obtain Corollary 6.

REFERENCES

[1] T. Goblick, Jr., "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inf. Theory*, vol. 11, no. 4, pp. 558–567, Oct. 1965.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 1, pp. 379–423, 1948.

[3] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[4] U. Mittal and N. Phamdo, "Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1082–1102, May 2002.

[5] M. Skoglund, N. Phamdo, and F. Alajaji, "Hybrid digital-analog source-channel coding for bandwidth compression/expansion," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3757–3763, Aug. 2006.

[6] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[7] V. A. Kotel'nikov, *The Theory of Optimum Noise Immunity*. New York, NY, USA: McGraw-Hill, 1959.

[8] A. Fuldseth and T. A. Ramstad, "Bandwidth compression for continuous amplitude channels based on vector approximation to a continuous subset of the source signal space," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4. Apr. 1997, pp. 3093–3096.

[9] S.-Y. Chung, "On the construction of some capacity approaching coding schemes," Ph.D. dissertation, Dept. Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2000.

[10] V. A. Vaishampayan and S. I. R. Costa, "Curves on a sphere, shift-map dynamics, and error control for continuous alphabet sources," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1658–1672, Jul. 2003.

[11] T. A. Ramstad, "Shannon mappings for robust communication," *Telektronikk*, vol. 98, no. 1, pp. 114–128, 2002.

[12] F. Hekland, P. A. Floor, and T. A. Ramstad, "Shannon–Kotel'nikov mappings in joint source-channel coding," *IEEE Trans. Commun.*, vol. 57, no. 1, pp. 94–105, Jan. 2009.

[13] Y. Hu, J. Garcia-Frias, and M. Lamarca, "Analog joint source-channel coding using non-linear curves and MMSE decoding," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3016–3026, Nov. 2011.

[14] N. Wernersson, M. Skoglund, and T. Ramstad, "Polynomial based analog source-channel codes—[transactions papers]," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2600–2606, Sep. 2009.

[15] P. A. Floor, T. A. Ramstad, and N. Wernersson, "Power constrained channel optimized vector quantizers used for bandwidth expansion," in *Proc. 4th Int. Symp. Wireless Commun. Syst.*, Oct. 2007, pp. 667–671.

[16] J. Karlsson and M. Skoglund, "Optimized low-delay source-channel-relay mappings," *IEEE Trans. Commun.*, vol. 58, no. 5, pp. 1397–1404, May 2010.

[17] K.-H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. 24, no. 12, pp. 1283–1290, Dec. 1976.

[18] T. Başar, B. Sankur, and H. Abut, "Performance bounds and optimal linear coding for discrete-time multichannel communication systems (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 26, no. 2, pp. 212–217, Mar. 1980.

[19] T. Fine, "Properties of an optimum digital system and applications," *IEEE Trans. Inf. Theory*, vol. 10, no. 4, pp. 287–296, Oct. 1964.

[20] J. Ziv, "The behavior of analog communication systems," *IEEE Trans. Inf. Theory*, vol. 16, no. 5, pp. 587–594, Sep. 1970.

[21] M. D. Trott, "Unequal error protection codes: Theory and practice," in *Proc. IEEE Inf. Theory Workshop*, Jun. 1996, p. 11.

[22] E. Akyol, K. Rose, and T. Ramstad, "Optimal mappings for joint source channel coding," in *Proc. IEEE Inf. Theory Workshop*, Jan. 2010, pp. 1–5.

[23] E. Akyol, K. Rose, and T. Ramstad, "Optimized analog mappings for distributed source-channel coding," in *Proc. IEEE Data Compress. Conf.*, Mar. 2010, pp. 159–168.

[24] E. Akyol, K. Viswanatha, K. Rose, and T. Ramstad, "On zero delay source-channel coding: Functional properties and linearity conditions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 6–10.

[25] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, no. 6, pp. 1437–1451, 1979.

[26] J. Walrand and P. Varaiya, "Optimal causal coding—Decoding problems," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 814–820, Nov. 1983.

[27] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4017–4035, Sep. 2006.

[28] S. Matloub and T. Weissman, "Universal zero-delay joint source–channel coding," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5240–5250, Dec. 2006.

[29] S. Yuksel, "On optimal causal coding of partially observed Markov sources in single and multiterminal settings," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 424–437, Jan. 2013.

[30] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM J. Control*, vol. 6, no. 1, pp. 131–147, 1968.

[31] T. Başar, "Variations on the theme of the Witsenhausen counterexample," in *Proc. 47th IEEE Conf. Decision Control*, Dec. 2008, pp. 1614–1619.

[32] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[33] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

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

[36] A. Ingber, I. Leibowitz, R. Zamir, and M. Feder, "Distortion lower bounds for finite dimensional joint source-channel coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1183–1187.

[37] S. Tridenski and R. Zamir, "Bounds for joint source-channel coding at high SNR," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul./Aug. 2011, pp. 771–775.

[38] A. Reani and N. Merhav, "Data-processing bounds for scalar lossy source codes with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4057–4070, Jul. 2013.

[39] E. Akyol, K. Rose, and T. Başar, "On optimal jamming over an additive noise channel," in *Proc. IEEE 52nd Annu. Conf. Decision Control*, Dec. 2013, pp. 3079–3084.

[40] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1289–1301, Mar. 2012.

[41] S. Yüksel and T. Linder, "Optimization and convergence of observation channels in stochastic control," *SIAM J. Control Optim.*, vol. 50, no. 2, pp. 864–887, 2012.

[42] R. T. Rockafellar, *Convex Analysis*, vol. 28. Princeton, NJ, USA: Princeton Univ. Press, 1997.

[43] S. Gadkari and K. Rose, "Robust vector quantizer design by noisy channel relaxation," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1113–1116, Aug. 1999.

[44] P. Knagenhjelm, "A recursive design method for robust vector quantization," in *Proc. Int. Conf. Signal Process. Appl. Technol.*, 1992, pp. 948–954.

[45] E. Akyol, K. B. Viswanatha, and K. Rose, "On conditions for linearity of optimal estimation," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3497–3508, Jun. 2012.

[46] T. Başar, "A trace minimization problem with applications in joint estimation and control under nonclassical information," *J. Optim. Theory Appl.*, vol. 31, no. 3, pp. 343–359, 1980.

[47] E. Akyol and K. Rose, "On linear transforms in zero-delay Gaussian source channel coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2012, pp. 1543–1547.

[48] P. Billingsley, *Probability and Measure*. New York, NY, USA: Wiley, 2008.

[49] C. H. Papadimitriou and J. Tsitsiklis, "Intractable problems in control theory," *SIAM J. Control Optim.*, vol. 24, no. 4, pp. 639–654, 1986.
[50] F. Hekland, G. E. Oien, and T. A. Ramstad, "Using 2:1 Shannon mapping for joint source-channel coding," in *Proc. IEEE Data Compress. Conf.*, Mar. 2005, pp. 223–232.
[51] X. Chen and E. Tuncel, "Zero-delay joint source-channel coding using hybrid digital-analog schemes in the Wyner–Ziv setting," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 726–735, Feb. 2014.
[52] N. Goela and M. Gastpar, "Reduced-dimension linear transform coding of correlated signals in networks," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 3174–3187, Jun. 2012.
[53] T. Başar, "The Gaussian test channel with an intelligent jammer," *IEEE Trans. Inf. Theory*, vol. 29, no. 1, pp. 152–157, Jan. 1983.
[54] M. Mehmetoglu, E. Akyol, and K. Rose, "A deterministic annealing approach to Witsenhausen's counterexample," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Feb. 2014, pp. 3032–3036.
[55] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov. 1998.
[56] M. S. Mehmetoglu, E. Akyol, and K. Rose, "A deterministic annealing approach to optimization of zero-delay source-channel codes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.
[57] M. S. Mehmetoglu, E. Akyol, and K. Rose, "Optimization of zero-delay mappings for distributed coding by deterministic annealing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4259–4263.
[58] D. G. Luenberger, *Optimization by Vector Space Methods*. New York, NY, USA: Wiley, 1969.
[59] R. M. Dudley, *Real Analysis and Probability*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

Emrah Akyol (S'03–M'12) received the Ph.D. degree in 2011 from the University of California at Santa Barbara. From 2006 to 2007, he held positions at Hewlett-Packard Laboratories and NTT Docomo Laboratories, both in Palo Alto, where he worked on topics in video compression. From 2013 to 2014, Dr. Akyol was a postdoctoral researcher in the Electrical Engineering Department, University of Southern California.

Currently, Dr. Akyol is a postdoctoral research associate in the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. His current research is on the interplay of networked information theory, game theory, communications and control. Dr. Akyol received the 2010 UCSB Dissertation Fellowship and the 2014 USC Postdoctoral Training Award.

Kumar B. Viswanatha (S'08–M'14) received his PhD degree in 2013 in the Electrical and Computer Engineering department from University of California at Santa Barbara (UCSB), USA. He is now with Qualcomm Technologies Inc., Corporate Research and Development Division, San Diego, CA. Prior to joining Qualcomm, he was an intern associate in the equity volatility desk at Goldman Sachs Co., New York, USA. His research interests include multi-user information theory, wireless communications, joint compression and routing for networks and distributed compression for large scale sensor networks.

Kenneth Rose (S'85–M'91–SM'01–F'03) received the Ph.D. degree in 1991 from the California Institute of Technology, Pasadena.

He then joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, and include rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and nonconvex optimization. He is interested in the relations between information theory, estimation theory, and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

Dr. Rose was co-recipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society, as well as the 2004 and 2007 IEEE Signal Processing Society Best Paper Awards.

Tor A. Ramstad received his Siv.Ing. and Dr.Ing. degrees in 1968 and 1971, respectively, both from the Norwegian Institute of Technology (NTH) (now part of the Norwegian University of Science and Technology (NTNU)). He has held various positions at the Faculty of Electrical Engineering at the same university, and became a full professor of Telecommunications in 1983. From 2012 he is Professor Emeritus at NTNU. He has spent sabbatical leaves at University of California, Santa Barbara (1982-83, 1997-98, 2008-09), at Georgia Institute of Technology (1989-90), and at EURECOM, France (2003-04, spring 2009). He was associate editor of IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, and a member of the IEEE DSP Technical Committee. He has chaired the Doctoral Committee NTH. Dr. Ramstad is a member of the Norwegian Academy of Technical Sciences. His research interests are in the fields of multirate signal processing, speech and image processing with emphasis on compression, and joint source-channel coding.