Quantizer Design to Exploit Common Information in Layered and Scalable Coding

Mehdi Salehifar, *Member, IEEE*, Tejaswi Nanjundaswamy, *Member, IEEE*, and Kenneth Rose, *Fellow, IEEE*

***Abstract*—This paper considers a layered coding framework with a relaxed hierarchical structure, tailored to serve content at multiple quality levels, where a key challenge is the conflict between coding optimality at each layer and efficient use of storage and networking resources. The prevalent approach of storing and transmitting independent copies for each quality level, is highly wasteful in resources. The alternative of conventional scalable coding incurs the notorious "scalability penalty" at the enhancement layers, due to its rigid structure. The approaches pursued in this work involve a layered coding framework, wherein information common to one or more subsets of the quality levels is first extracted and transmitted, and then complemented by individual (quality level specific) bit streams. This framework ensures that no redundant or irrelevant information is sent to any decoder, enables achieving all intermediate operating points between the two extremes of conventional scalable coding versus independent coding, and hence mitigates the layered coding penalty. Joint design of common and individual layers ensures that all extracted common information is fully usable by the target decoders, as needed to approach optimality. Simulation results for practically important sources, confirm the superiority of the proposed framework.**

***Index Terms*—Common information, rate-distortion theory, layered and scalable coding, video and audio compression.**

I. Introduction

TECHNOLOGICAL advances ranging from multigigabit high-speed Internet to wireless communication and mobile, limited resource receivers, have created an extremely heterogeneous network scenario with data consumption devices of highly diverse decoding and display capabilities, all accessing the same content over networks of time varying bandwidth and latency. The primary challenge is to maintain optimal signal quality for a wide variety of users, while ensuring efficient use of resources for storage and transmission across the network.

Manuscript received November 2, 2019; revised October 8, 2020 and January 9, 2021; accepted February 25, 2021. Date of publication March 8, 2021; date of current version March 31, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gwo Giun Lee. This work was supported by National Science Foundation under Grants CCF-1016861 and CCF-1320599. *(Corresponding author: Mehdi Salehifar.)*

Mehdi Salehifar 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 LG Electronics, San Diego, CA 92131 USA (e-mail: mehdi.salehifar@ieee.org).

Tejaswi Nanjundaswamy 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 Apple Inc., Cupertino, CA 95014 USA (e-mail: tejaswi@ece.ucsb.edu).

Kenneth 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).

Digital Object Identifier 10.1109/TSP.2021.3064281

The simplest (and common) solution is to store and transmit independent copies of the signal for every type of user the provider serves. This solution is highly wasteful in network resources. The main alternative, namely, conventional scalable coding [1], [2], generates layered bit-streams, with a base layer offering coarse quality reconstruction and successive enhancement layers to refine the quality. Depending on the network and user constraints, a suitable number of layers are received and decoded, yielding a prescribed quality level. However, there is a well documented inherent loss due to scalable coding, which incurs higher distortion than independent (non-scalable) coding at the same overall received rate [3]–[5], as most sources are not successively refinable at finite delay under common distortion measures. Moreover, at fixed receive rates (over the last link to user devices), non-scalable coding and conventional scalable coding require the highest and lowest total transmit (or storage) rate, respectively. Thus, non-scalable coding and conventional scalable coding represent two extreme points in terms of the tradeoff between total transmit rate and decoder distortion, for prescribed receive rates.

Recent work (from our lab) proposed a novel layered coding paradigm for multiple quality levels [6], inspired by the information-theoretic concept of *common information* of correlated random variables [7]–[9], wherein only a subset of the information needed to reconstruct at a lower quality level is shared with the higher quality level receiver. This flexibility enables efficient extraction of common information between quality levels and achieves intermediate operating points in the tradeoff between total transmit rate and decoder distortion, in effect controlling the layered coding penalty. The information-theoretic foundation of the framework was established in [6], [9], and preliminary implementation with a standard audio codec [10] demonstrated its potential gains. This paper focuses on the important problem of quantizer design for this layered coding framework.

We first consider the simple setting of two quality levels with fixed receive rates. This setting requires the design of *three* quantizers: one for the common layer, whose output is sent to both decoders, and two level-specific quantizers to refine the common layer information for the respective quality levels, each sent solely to the respective decoder. First, joint design of quantizers across all layers is proposed for this setting. Specifically, an iterative approach is developed for designing the three quantizers, wherein at each iteration step one quantizer is updated to minimize the overall cost function while the others are fixed, and the iterations are performed repeatedly over all quantizers until convergence. We further derive ("Lloyd algorithm style") optimal update rules, wherein the more challenging derivation is for the common layer quantizer to minimize the overall cost while accounting for the effects of the individual layer quantizers. A complementary contribution develops a low complexity variant of the joint quantizer design approach. The approach is then specialized to the practically important case of Laplacian sources. Early results appeared in a conference paper [11] with focus on low complexity quantizer design for Laplacian sources based on applying a so-called dead-zone quantizer (DZQ) to the common layer and a mix of DZQ and uniform quantizers to the individual layers, as explained in Section IV-B.

Finally we propose an iterative technique for joint design of vector quantizers for all layers of this framework. For simplicity, we first explain the approach for the setting of two quality levels. Then we explain the extended approach to other relaxed hierarchical structures. We develop a cost function which explicitly controls the tradeoff between distortions, receive rates, and total transmit rate. We then propose an iterative approach for jointly designing vector quantizers for all the layers, wherein we estimate optimal quantizer partitions at all the layers, given reconstruction codebooks, and optimal reconstruction codebooks for all quality levels, given quantizer partitions, iteratively, until convergence.

Experimental evaluation results for Laplacian and multivariate normal distributions substantiate the usefulness of the proposed technique.

The rest of this paper is organized as follows. Section II provides background and preliminaries. Sections III and IV present the proposed methods. Experimental results are summarized in Section V, with conclusion in Section VI.

II. Background and Preliminaries

Few mathematical results have had as much impact on the foundation of the information age as Shannon's 1948 point-to-point communication theorems [12]. However, the communication model assumed in these seminal contributions is inadequate for the realities of modern networks. Extensions of the theory to multi-terminal settings have proven difficult and, despite several spectacular advances, many questions remain only partially answered, and particularly lacking is significant progress on the conversion of available insights into practical approaches.

A. Successive Refinement and Scalable Coding

Rate-distortion theory is a major branch of information theory, focused on the theoretical foundation for lossy data compression. The fundamental theorem of rate-distortion theory [12] is that the minimum bit rate (per sample) required to convey a sequence of independent random variables, each drawn from the probability distribution of a generic random variable X, so that the sequence can be reconstructed at average distortion of at most D, is given by:

$$R(D) = \min_{p(\hat{x}|x): E\{d(x,\hat{x})\} \leq D} I(X;\hat{X}), \tag{1}$$

where $E\{d(x,\hat{x})\}$ evaluates the distortion between the two random variables. The result states that the rate-distortion function, indicating the minimum achievable rate for prescribed distortion, is given by minimizing the mutual information $I(X;\hat{X})$ between X and reconstruction variable $\hat{X}$, over all random encoders satisfying the distortion constraint. It led to extensive research efforts devoted to finding rate-distortion bounds for various new settings [7], [13]–[16], numerical evaluation of the rate distortion function for generic sources and distortion measures [17]–[20] and to practical scalar/vector quantizer design and analysis [21]–[23].

Rate-distortion theory considers scalable coding, a special case of multiple descriptions coding, through the concept of successive refinement of information, and specifically determines conditions under which scalable coding incurs no rate-distortion performance penalty [24]–[31]. In scalable coding, the encoder generates two layers of information, namely, the base layer at rate R_{12}, and the enhancement layer at rate R_2 (the subscripts specify to which decoders the information is routed). The base layer provides a coarse reconstruction of the source (at rate R_{12}), while the enhancement layer is used to 'refine' the reconstruction beyond the base layer (at an overall rate of $R_2 + R_{12}$). The base and enhancement layer distortions are D_1 and D_2, respectively, where $D_2 < D_1$.

A source-distortion pair for which it is possible to achieve rate-distortion optimality simultaneously, at both the layers, is called successively refinable in the literature. The necessary and sufficient condition for such optimality to be achieved at distortion levels D_1 and D_2 is that there exists a conditional probability distribution $p(\hat{x}_1, \hat{x}_2|x)$ such that:

$$\begin{aligned} &E\{d(X,\hat{X}_1)\} \leq D_1, \qquad E\{d(X,\hat{X}_2)\} \leq D_2, \\ &I(X;\hat{X}_1) = R(D_1), \qquad I(X;\hat{X}_2) = R(D_2), \\ &\qquad\qquad\qquad X \leftrightarrow \hat{X}_2 \leftrightarrow \hat{X}_1, \end{aligned} \tag{2}$$

where $X \leftrightarrow \hat{X}_2 \leftrightarrow \hat{X}_1$ denotes the requirement that the three variables form a Markov chain. Thus, if a source-distortion pair is not successively refinable, it is impossible to maintain optimality at both layers simultaneously within the scalable coding framework.

B. The Gray-Wyner Network and Common Information

The Gray-Wyner (GW) network, consists of an encoder that transmits two correlated sources to two receivers using three channels: a common channel linking the encoder to both receivers, and two private channels linking it to individual receivers. The channels are assumed to be noiseless and each has a specified per bit communication cost (Fig. 1).

The objective is to minimize the communication cost while maintaining decoder distortion at or below the prescribed levels. Rate tradeoffs are central here as we seek the optimal rate triplet (R_{12}, R_1, R_2). GW derived the asymptotic minimum cost achievable for this network.

There are several concepts associated with the notion of the common information (CI) of two random variables. One important CI definition is due to Wyner [8] who characterized

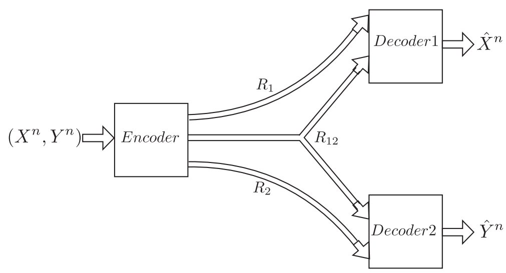


Fig. 1. The Gray-Wyner network.

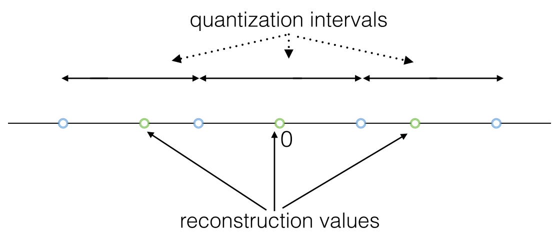


Fig. 2. The optimal scalar dead-zone quantizer for Laplacian sources with nearly uniform reconstruction rule [11].

it in terms of the GW network: the minimum achievable rate R_{12} on the shared branch of the lossless GW network, when the sum transmit rate is set to its minimum (the joint entropy), i.e., $R_0 + R_1 + R_2 = H(X,Y)$. A different concept of CI was proposed by Gács and Körner [32], and is extremely relevant to the contribution in this paper. Ahlswede and K ö rner [33] gave an alternative characterization of this CI that directly ties it to the GW network: the maximum achievable rate R_{12} on the shared branch of the lossless GW network, when the two receive rates are set to their minimum levels, i.e., $R_{12} + R_1 = H(X)$ and $R_{12} + R_2 = H(Y)$.

C. Laplacian Sources

The Laplacian distribution is a common model for multimedia sources in many practical applications:

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \tag{3}$$

with parameter $\lambda > 0$.

1) Scalar Quantization of Laplacian Sources: In [34], the optimal entropy constrained quantizer for the Laplacian source was derived and shown to be the so-called dead-zone quantizer (DZQ) depicted in Fig. 2, which exhibits a uniform step-size for all intervals except the dead-zone wider interval about the origin.

2) Scalable Coding of Laplacian Sources: Current scalable coding standards including scalable HEVC [35] for video, and scalable AAC [36] for audio, employ DZQ at the base layer for quantizing the source, and at the enhancement layer, a scaled version of the base layer DZQ to quantize the base layer reconstruction error. However, a significantly improved approach to scalable coding of Laplacian sources, called conditional enhancement layer quantization (CELQ) [5]:

- The base layer employs a DZQ.

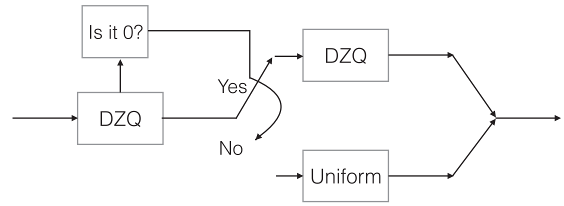


Fig. 3. Conditional enhancement layer quantizer for Laplacian sources [5], [11]. Based on the base layer DZQ interval, the enhancement layer quantizer is chosen.

- The enhancement layer quantizers are conditioned on the base layer quantization interval: Use DZQ if a dead zone interval was established by the base layer, and use a uniform quantizer otherwise (as illustrated in Fig. 3).

While CELQ offers considerable gains and approaches optimality within the classical scalable coding framework, it nevertheless suffers from the notorious scalable coding penalty, which motivates the approach pursued in this paper.

III. Overview of the Proposed Framework

The proposed layered coding framework is motivated by the recent lossy generalization of common information [6], [9]. We define the lossy generalization of the Gács-Körner CI at (D_1, D_2), denoted $C_{GK}(X, Y; D_1, D_2)$, as follows: Let $R_{GK}(D_1, D_2)$ be the set of rate triplets (R_{12}, R_1, R_2) such that for any $\epsilon > 0$,

$$R_{12} + R_1 \leq R_X(D_1) + \epsilon, \qquad R_{12} + R_2 \leq R_Y(D_2) + \epsilon. \tag{4}$$

Then,

$$C_{GK}(X, Y; D_1, D_2) = \sup_{R_{GK}(D_1, D_2)} R_{12}. \tag{5}$$

The lossless $C_{GK}(X, Y; 0, 0)$ was considered a negative result by Gács-Körner, as it is typically much smaller than the mutual information of X and Y and often zero. But the seemingly degenerate special case, where $X = Y$ in the lossy setting, is a highly significant quantity. Here, C_{GK} reflects the CI between *reconstructions of a given source at different quality levels*, which is directly applicable to layered coding. Specifically, [6], [9] derived the single letter characterization for the lossy C_{GK} for layered coding as: *sup* $I(X; U)$ where U is an auxiliary random variable and the supremum is over all conditional distributions $P(U, \hat{X}_1, \hat{X}_2|X)$ such that $\hat{X}_1$ and $\hat{X}_2$ achieve RD optimality at D_1 and D_2, respectively, and the following Markov chains hold:

$$X \leftrightarrow \hat{X}_1 \leftrightarrow U, \quad X \leftrightarrow \hat{X}_2 \leftrightarrow U. \tag{6}$$

In this section we illustrate both the novel layered coding paradigm [6] and the significance of effective extraction of common information, via quantizer design for a toy example involving a simple uniform distribution. For a uniformly distributed random variable, the optimal entropy constrained scalar quantizer (ECSQ), at rate $\log(N)$, where N is an integer, is a uniform quantizer with N levels [37]. Fig. 4 shows, for a source distributed uniformly over the interval [0,6], the optimal

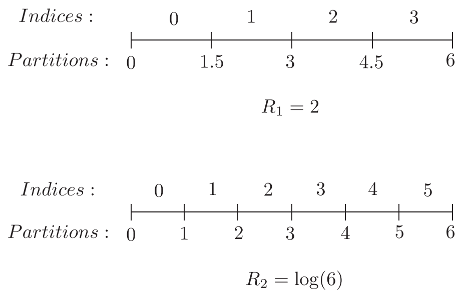


Fig. 4. Optimal quantizer partitions for a source uniformly distributed in [0, 6], at rates $R_1 = 2$ and $R_2 = \log(6)$.

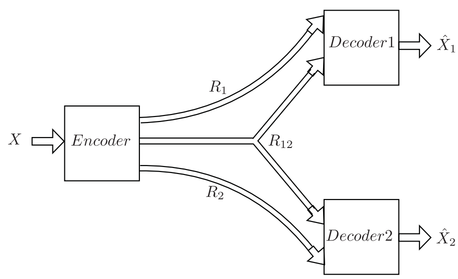


Fig. 5. Common information based layered coding paradigm: rate R_{12} sent to both decoders, and individual rates R_1 and R_2 sent to respective decoders.

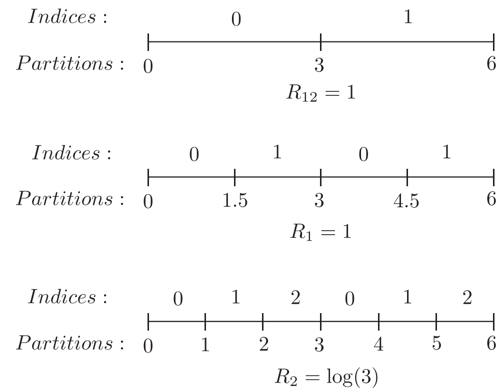


Fig. 6. Quantizer partitions for common information based layered coding, achieving successive refinement for the example of Fig. 4.

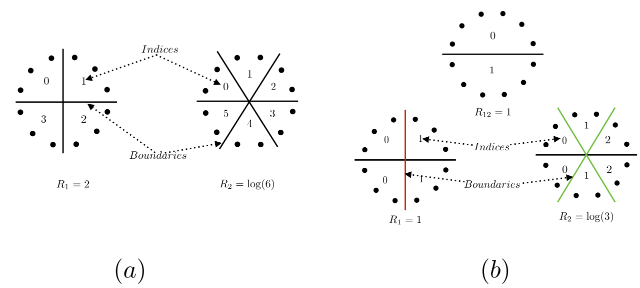


Fig. 7. A 2D example with training points shown as black dots. (a) Individual coding partitions at rates $R_1 = 2$ and $R_2 = \log(6)$. (b) Common information based layered coding partitions, with common rate $R_{12} = 1$, and individual rates $R_1 = 1$ (red partitions) and $R_2 = \log(3)$ (green partitions) sent to respective decoders.

quantizer partitions at rates $R_1 = 2$ and $R_2 = \log(6)$, resulting in distortion D_1 and D_2, respectively. Note that not all boundary points of quantizer 1 align with boundary points of quantizer 2. This implies that the source is not successively refinable at these rates, and scalable coding yields enhancement layer performance that is worse than independent quantization at rate $\log(6)$. Clearly, not all the information required to achieve D_1 is useful to achieve D_2, hence the suboptimality. On the other hand independent coding is wasteful as there is still considerable overlap in information conveyed to the two decoders.

These observations strongly motivate the common information based layered coding framework to ensure that each decoder receives only information necessary for its reconstruction. This layered coding paradigm is illustrated in Fig. 5. The encoder generates three packets, one at rate R_1 sent exclusively to the base decoder reconstructing $\hat{X}_1$, one at rate R_2 sent exclusively to the decoder reconstructing $\hat{X}_2$, and the "common" packet, at rate R_{12} sent to both decoders.

This paradigm subsumes, as special cases, conventional scalable coding (let $R_1 = 0$) and non-scalable coding (let $R_{12} = 0$). Moreover, this framework provides an extra degree of freedom such that rate-distortion optimality at both layers can potentially be achieved at a lower total transmit rate than non-scalable coding. Returning to our toy example of a uniformly distributed source reconstructed at receive rates of 2 and $\log(6)$, let us employ the proposed layered coding paradigm, with rates $R_{12} = 1$, sent to both decoders; $R_1 = 1$ and $R_2 = \log(3)$ each sent to the respective decoder (as per Fig. 6). Note that the framework implements the same quantization partitions as independent coding, ensuring rate-distortion optimality at the decoders, but with a 22% reduction in total transmit rate. This example illustrates how appropriately designed quantizers for the layered coding paradigm can efficiently extract information common to different quality levels. Fig. 7 provides similar illustration for a vector source. Such quantizer design methods are derived next.

IV. Joint Design of Quantizers for Common Information Based Layered Coding

The objective is to design entropy-constrained scalar/vector quantizers for specified received rates, $R_{r_1} = R_{12} + R_1 = c_1$ and $R_{r_2} = R_{12} + R_2 = c_2$, at decoder 1 and 2, respectively. Optimization under this constraint reflects the fundamental tradeoff between total transmit rate, $R_t = R_{12} + R_1 + R_2 = c_1 + c_2 - R_{12}$ (or equivalently R_{12}), and the distortion levels D_1 and D_2. The two extremes are: i) non-scalable coding, incurring the maximum $R_t = c_1 + c_2$ (lowest $R_{12} = 0$), and minimum distortion $D^*(c_1) + D^*(c_2)$, where $D^*(\cdot)$ is the optimal distortion at a given rate; and ii) conventional scalable coding, with the minimum $R_t = c_2$ (highest $R_{12} = c_1$), but at high distortion at the enhancement layer due to the scalable coding penalty. Let us formulate the cost function that captures this tradeoff, subject to the prescribed receive rates, as the Lagrangian:

$$J = a_1 D_1 + a_2 D_2 + \lambda_1 (R_1 + R_{12}) + \lambda_2 (R_2 + R_{12}) + \lambda_{12} R_{12}, \quad (7)$$

where a_1 and a_2 control the relative importance of D_1 and D_2, and λ_1, λ_2, and λ_{12} constrain R_{r_1}, R_{r_2} and R_{12}, respectively.

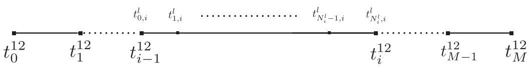


Fig. 8. Quantizer partition boundaries for subdividing a common layer quantization interval. The decision points of the common layer are shown below the line, and above it for individual layer l.

Note that we maintain a slightly redundant notation for consistency with existing literature. Clearly, one of the five weights, $a_1, a_2, \lambda_1, \lambda_2$, and λ_{12} in (7) is redundant. Quantizers designed to minimize the cost function (7) achieve an operating point on the optimal curve implementing the tradeoff between weighted sum of distortions, transmit rate and receive rates.

A. Quantizer Design for Scalar Sources:

We design the quantizers iteratively, with one quantizer updated in each iteration to minimize the overall cost function while the others are fixed, until convergence. In the following subsections, the superscripts ^{1}, ^{2} and ^{12} refer to parameters of individual layers 1, 2, and common layer, respectively.

1) Individual Layer Quantizer Design: Given a common layer quantizer with M intervals and partition boundaries $t_i^{12},\ i = 0, 1, \ldots, M$, we need to design ECSQ for each interval, $(t_{i-1}^{12}, t_i^{12}),\ i = 1, 2, \ldots, M$ and each individual layer, $l = 1, 2$. Note that, given the common layer, the individual quantizers of layers 1 and 2 have no effect on each other's distortion and rate and can be optimized independently. Hence, the cost function for optimization of individual quantizers for each of the layer simplifies to $J_l = a_l D_l + \lambda_l R_l$ for $l = 1, 2$.

We employ the well known iterative ECSQ design technique for each interval, $(t_{i-1}^{12}, t_i^{12}),\ i = 1, 2, \ldots, M$ and each individual layer, $l = 1, 2$. Let N_i^l be the number of subintervals for layer l at common layer interval (t_{i-1}^{12}, t_i^{12}). Fig. 8 depicts partition boundaries for layer l at interval (t_{i-1}^{12}, t_i^{12}) as $t_{q,i}^l,\ q = 0, 1, 2, \ldots, N_i^l$. (Note the end points $t_{0,i}^l = t_{i-1}^{12}$ and $t_{N_i^l,i}^l = t_i^{12}$.) The iterative ECSQ algorithm employed is summarized below:

i) Initialize the partition boundaries, $t_{q,i}^l,\ q = 1, 2, \ldots, N_i^l - 1$.

ii) Calculate N_i^l representative levels $x_{q,i}^l$ and subinterval probabilities $p_{q,i}^l$ for $q = 1, 2, \ldots, N_i^l$:

$$x_{q,i}^l = \frac{\int_{t_{q-1,i}^l}^{t_{q,i}^l} x f(x) dx}{\int_{t_{q-1,i}^l}^{t_{q,i}^l} f(x) dx}, \tag{8}$$

$$p_{q,i}^l = \int_{t_{q-1,i}^l}^{t_{q,i}^l} f(x) dx, \quad q = 1, 2, \ldots, N_i^l, \tag{9}$$

where $f(x)$ denotes the source distribution.

iii) Calculate partition boundaries, $t_{q,i}^l;\ q = 1, 2, \ldots, N_i^l - 1$:

$$t_{q,i}^l = \frac{x_{q,i}^l + x_{q+1,i}^l}{2} - \frac{\lambda_l}{a_l} * \frac{\log_2 p_{q,i}^l - \log_2 p_{q+1,i}^l}{2(x_{q,i}^l - x_{q+1,i}^l)}, \tag{10}$$

iv) Repeat (8), (9) and (10) until there is no further reduction in cost (or a prescribed stopping criterion is met).

2) Common Layer Quantizer Design: Unlike individual layer quantizer design, any partition change in the common layer quantizer impacts all distortions and rates, hence we cannot simply apply the standard ECSQ update rules. Given both individual layers' quantizers, the optimal update for common layer decision points, $t_i^{12},\ i = 1, 2, \ldots, M - 1$ is (see Appendix A):

$$\begin{aligned} t_i^{12} &= \frac{(a_1(x_{1,i+1}^1)^2 + a_2(x_{1,i+1}^2)^2) - (a_1(x_{N_i^1,i}^1)^2 + a_2(x_{N_i^2,i}^2)^2)}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))} \\ &\quad - \frac{\lambda_1(\log_2 p_{1,i+1}^1 - \log_2 p_{N_i^1,i}^1) + \lambda_2(\log_2 p_{1,i+1}^2 - \log_2 p_{N_i^2,i}^2)}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))} \\ &\quad - \frac{\lambda_{12}(\log_2 p_{i+1}^{12} - \log_2 p_i^{12})}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))}, \end{aligned} \tag{11}$$

where centroids $p_i^{12} = \int_{t_{i-1}^{12}}^{t_i^{12}} f(x) dx$ are calculated using previous iteration values of t_i^{12}, for $i = 1, 2, \ldots, M$.

3) Joint Design of Quantizers: The overall algorithm for joint design of quantizers for all layers is:

i) Initialize the common layer partition boundaries, $t_i^{12},\ i = 1, 2, \ldots, M - 1$.

ii) Update the individual layer quantizers using the iterative steps (8), (9) and (10) to calculate $x_{q,i}^l,\ p_{q,i}^l$ for $q = 1, 2, \ldots, N_i^l$ and $t_{q,i}^l$ for $q = 1, 2, \ldots, N_i^l - 1$ for both $l = 1, 2$ and all $i = 1, 2, \ldots, M$.

iii) Update the common layer quantizer using (11) to calculate $t_i^{12},\ i = 1, 2, \ldots, M - 1$.

iv) Repeat steps (ii) and (iii) until there is no further reduction in cost (or a prescribed stopping criterion is met).

Note that during the ECSQ design at common or individual layers, the number of partitions, i.e., M and N_i^l are not known. To circumvent this, we simply initialize our algorithm with a large number of partitions, and in each iteration, based on the given $a_1, a_2, \lambda_1, \lambda_2$, and λ_{12}, the algorithm reduces the number of partitions, as necessary.

B. Low Complexity Quantizer Design for Laplacian Sources

For the practically important case of Laplacian source distribution we propose a low complexity alternative design:

i) For the common layer, we estimate the best step size for the DZQ at a given rate, R_{12}.

ii) For the two individual layers, we design optimal entropy constrained quantizers for each common layer quantizer interval, at their corresponding rates of R_1 and R_2. Specifically, we iteratively optimize the quantizer interval partitions and reconstruction points to minimize the entropy constrained distortion, with smart initializations of,

- A DZQ for the dead zone interval, and
- A uniform quantizer for other intervals,

of the common layer quantizer.

iii) We then numerically estimate the optimal common layer rate, by trying multiple allowed common rates and selecting the one that results in minimum cost J.

Since the dead zone interval contains a truncated Laplacian distribution and other intervals contain a truncated exponential distribution, we select the initializations in step (ii) above to be the optimal entropy constrained quantizers of their corresponding non-truncated distributions.

Note that we can achieve non-zero common rate at negligible ΔD, if the DZQ at rate R_{12} is such that all its partition points align closely with partition points of DZQ at both rates c_1 and c_2. Conditions for such an alignment of partitions between two DZQ were derived in [38]: the dead-zone of the coarser DZQ has to be divided into $2n+1$ intervals, and other intervals of this DZQ have to be divided into $m+1$ intervals, with $2n/m = z$, where, n and m are integers, and z is the ratio of the dead-zone to the uniform interval lengths. Our design technique numerically estimates the common layer DZQ which closely satisfies these conditions with DZQ at both rate c_1 and c_2.

Note that the proposed design technique does not ensure joint optimality of the quantizers, since we independently optimize the common layer quantizer (e.g., DZQ for Laplacian) without considering its effect on other layers. Nevertheless the approach achieves considerable performance gains.

C. Quantizer Design for Vector Sources:

We design the quantizers iteratively by alternating between the steps of optimal partitioning and optimal codebook estimation, until convergence, similar to the generalized Lloyd algorithm [39].

We design an M-codebook ECVQ for the common layer, and for each common layer region, $i = 1, 2, \ldots, M$ we design an ECVQ for each individual layer, $l = 1, 2$. Let N_i^l be the number of subregions for layer l at common layer region i. Let $\mathbf{c}_{q,i}^l$ and $p_{q,i}^l$ be the representative levels, and subregion probabilities, respectively, for $q = 1, 2, \ldots, N_i^l$, $i = 1, 2, \ldots, M$, and $l = 1, 2$. Finally, let p_i^{12} be the common layer regions probabilities for $i = 1, 2, \ldots, M$. Following is the overall iterative algorithm:

i) Guess an initial set of representative levels $\mathbf{c}_{q,i}^l$ and their corresponding probabilities $p_{q,i}^l$ for $q = 1, 2, \ldots, N_i^l$, $i = 1, 2, \ldots, M$, and $l = 1, 2$.

ii) Assign each sample $\mathbf{x}_t$ in training set S to common layer region i and subregions' representatives $\mathbf{c}_{q_1,i}^1$ and $\mathbf{c}_{q_2,i}^2$, to minimize the Lagrangian cost:

$$J_{\mathbf{x}_t}(\mathbf{c}_{q_1,i}^1, \mathbf{c}_{q_2,i}^2) = (a_1||\mathbf{x}_t - \mathbf{c}_{q_1,i}^1||^2 + a_2||\mathbf{x}_t - \mathbf{c}_{q_2,i}^2||^2) - (\lambda_1 log_2 p_{q_1,i}^1 + \lambda_2 log_2 p_{q_2,i}^2 + \lambda_{12} log_2 p_i^{12}). \tag{12}$$

iii) Find subregion $B_{q,i}^l$:
$B_{q,i}^l = \{\mathbf{x} \in S : \mathbf{x} \text{ is assigned to } \mathbf{c}_{q,i}^l\}$
for $q = 1, 2, \ldots, N_i^l$, $i = 1, 2, \ldots, M$, and $l = 1, 2$.

iv) Calculate new representative levels and probabilities:

$$\mathbf{c}_{q,i}^l = \frac{1}{||B_{q,i}^l||} \sum_{\mathbf{x} \in B_{q,i}^l} \mathbf{x}, \tag{13}$$

$$p_{q,i}^l = \frac{||B_{q,i}^l||}{||S||}, \tag{14}$$

for $q = 1, 2, \ldots, N_i^l$, $i = 1, 2, \ldots, M$, and $l = 1, 2$. Also $p_i^{12} = \sum_{q=1}^{N_i^1} p_{q,i}^1$ for $i = 1, 2, \ldots, M$.

v) Repeat steps (ii), (iii) and (iv) until there is no further reduction in overall cost: $J = \sum_{\mathbf{x}_t} J_{\mathbf{x}_t}$.

Notes on the designed layered coding quantizers: The common layer quantizer does not reconstruct the source, leading to flexibility in structure of the quantizer partition. That is, we could potentially have irregular quantizers in the common layer alone, or even in both common and individual layers, as long as they combine together to result in an overall regular quantizer. (This phenomenon is key to distributed quantization, for example). Perhaps surprisingly, despite this flexibility, we show in the following lemma that optimality implies regularity of the common layer quantizer.

Lemma 1: If the overall quantizer partitions are regular, then so is the optimal common layer quantizer partition.

Proof: Assume common layer's cell j is not convex. Without loss of generality, let us assume it consists of two regular sub-cells j_1 and j_2. Let us define a new common layer quantizer partition by dividing cell j into two cells j_1 and j_2. Since we assumed the overall quantization partitions are regular, sub-cells of j in individual layers are forced to be wholly contained within either cell j_1 or j_2. Which implies that dividing common layer cell j into two cells j_1 and j_2, does not alter the overall quantization cells. Since the distortions at the decoder depend only on the overall quantization cells, they are not altered by the above common layer cell division. Also note that, receive rate at decoder l is:

$$R_l + R_{12} = -\sum_{i=1}^{M} \sum_{q=1}^{N_i^l} p_{q,i}^l \log_2 p_{q,i}^l, \quad \text{for } l = 1, 2. \tag{15}$$

Since the overall quantization regions are not altered, rate contributions from subregions of common layer cell j (via $p_{q,j}^l$) get redistributed to rate contributions from subregions of common layer cell j_1 (via p_{q,j_1}^l) and cell j_2 (via p_{q,j_2}^l), and thus $R_l + R_{12}$ for $l = 1, 2$, remain unchanged. However, there is increase in $R_{12} = -\sum_{i=1}^{M} p_i^{12} \log_2 p_i^{12}$ due to the subdivision of cell j. That is, we obtain a new set of quantizers with the same distortion levels and received rates, but a reduced transmit rate (equivalently increased R_{12}), which implies that cost J is reduced. Hence, the optimal common layer quantizer must be regular. ■

D. Layered Coding With Multiple Quality Levels

It is not straightforward to extend the concept of common information to more than two quality levels. The main challenge emerges when one notes that shared information can exist between any of the levels. Clearly, just one common layer, common to all quality levels cannot capture all the redundancies present. In fact, common information can exist between any subset of quality levels, and the number of bit-stream would grow combinatorially with the number of quality levels. A recent result from our lab proves that "combinatorial message sharing" can be

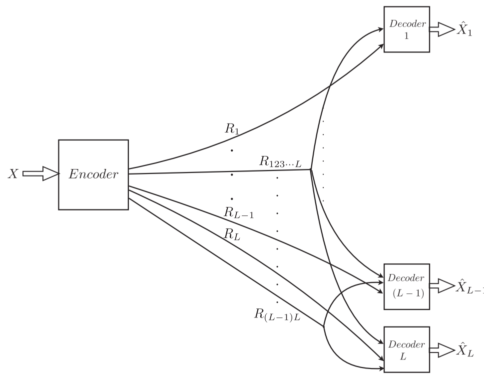


Fig. 9. Common information based layered coding: the multi-layer scenario.

used to strictly improve the theoretically achievable region for the closely related problem of multiple descriptions coding [29], [31]. While this approach is useful to obtain asymptotic bounds, it is obviously impractical for real world applications. We hence propose employing a linearly growing rate-splitting approach where each layer receives an individual packet for itself and all the common packets received by lower layers (layers with higher distortion constraints). Specifically, with L decoders, there are $2L-1$ packets consisting of: L individual packets at rates $R_i, i = 1, \ldots, L$; and $L-1$ common packets at rates $R_{123\cdots L}, R_{23\cdots L}, \ldots, R_{(L-1)L}$. The overall set of packets is indexed by the destination decoders and the index set is

$$Q = \{1, 2, \ldots, L-1, L; 123\cdots L, 23\cdots L, \ldots, (L-1)L\},$$

where, for example, index $23\cdots L$ means the packet is sent to decoders $2, 3, \ldots, L$. Fig. 9 depicts this scenario.

Similar to previous part we could define the cost as

$$J = \sum_{l=1}^{L} a_l D_l + \sum_{q \in Q} \lambda_q R_q, \tag{16}$$

where the first and second terms in the cost J, represent the distortion and rate penalties respectively. Finally using the approach explained in Section IV-C we design the best entropy constrained vector quantizers. For the specific example of 3 quality level coding, step (ii) of the algorithm will assign each training sample $\mathbf{x}_t$ to cell i of the common layer shared by all three levels (that is sent at rate R_{123}), subcell (q_1, i) within i for the private layer of quality level 1 (that is sent at rate R_1), subcell (q_{23}, i) within i of the common layer shared by quality levels 2 and 3 (that is sent at rate R_{23}), subcell (q_2, q_{23}, i) within (q_{23}, i) for the private layer of quality level 2 (that is sent at rate R_2) and subcell (q_3, q_{23}, i) within (q_{23}, i) for private layer of quality level 3 (that is sent at rate R_3), to minimize the cost in (16). Other steps are extended similarly.

V. Experimental Results

A. *Joint Design of Scalar Quantizers*

Simulations were performed on a Laplacian source with $\lambda = 1$. In the first experiment, the receive rates were set to $c_1 = 2$

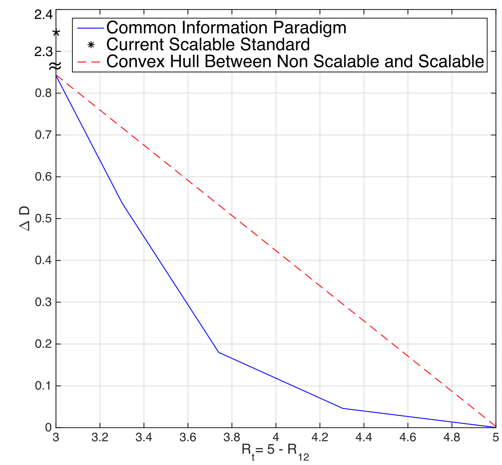


Fig. 10. Excess distortion versus total transmit rate.

TABLE I
Transmit Rate at Negligible Excess Distortion: Independent Coding vs Common Information Based (Joint Design) Approach

Required receive rates (R_{r_1}, R_{r_2})	Independent coding transmit rate $R_1 + R_2$	CI-based transmit rate $R_{12} + R_1 + R_2$	Rate reduction
$(2, 4)$	$2 + 4 = 6$	$0.8 + 1.2 + 3.2 = 5.2$	13%
$(2.1, 3.2)$	$2.1 + 3.2 = 5.3$	$0.8 + 1.3 + 2.4 = 4.5$	15%
$(2.5, 4.3)$	$2.5 + 4.3 = 6.8$	$1.1 + 1.4 + 3.2 = 5.7$	16%
$(2.5, 3.2)$	$2.5 + 3.2 = 5.7$	$1.2 + 1.3 + 2 = 4.5$	21%
$(3, 3.6)$	$3 + 3.6 = 6.6$	$1.5 + 1.5 + 2.1 = 5.1$	23%
$(2.7, 3.4)$	$2.7 + 3.4 = 6.1$	$1.4 + 1.3 + 2 = 4.7$	23%
$(2.4, 3)$	$2.4 + 3 = 5.4$	$1.5 + 0.9 + 1.5 = 3.9$	28%
$(3.1, 3.6)$	$3.1 + 3.6 = 6.7$	$2 + 1.1 + 1.6 = 4.7$	30%

and $c_2 = 3$. Fig. 10 shows the total transmit/storage rate R_t versus the excess distortion $\Delta D = D_1 + D_2 - D^*(c_1) - D^*(c_2)$, obtained by quantizers designed by the proposed iterative technique at common layer rates (R_{12}) ranging from 0 (independent coding) to 2 bits (scalable coding). It also shows the convex hull between independent and scalable coding, which can be obtained without recourse to common information via time sharing. Note that scalable coding as employed by current standards incurs about 1.5 dB excess distortion over the efficient scalable coding point obtained as a special case of our approach, which itself incurs about 0.8 dB excess distortion over non-scalable coding. The results clearly demonstrate that the concept of common information enables operating at all intermediate tradeoff points, at considerably better performance than the convex hull between independent and scalable coding.

An interesting observation in Fig. 10 is the existence of an operating point at $R_t = 4.3$ (or $R_{12} = 0.7$), with distortion close to non-scalable coding, yet with a 14% reduction in total transmit rate. We thus conducted more experiments under various receive rate constraints and tabulated in Table I the transmit rate savings achievable at negligible excess distortion. Significant transmit rate savings are observed, ranging from 13% to 30%, and demonstrate the capability of the proposed technique to efficiently extract the common information across quality levels. These savings will translate to significant operating cost reduction at data centers for storage, transmission to, and caching at, intermediate nodes for content providers who currently default to generating independently coded copies at different quality levels.

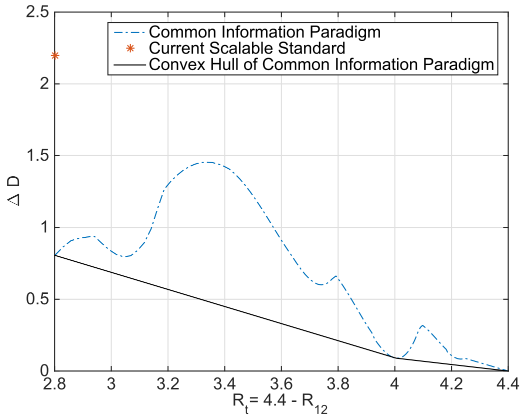


Fig. 11. Excess distortion versus total transmit rate (low complexity) as presented earlier in [11].

TABLE II
TRANSMIT RATE AT NEGLIGIBLE EXCESS DISTORTION: INDEPENDENT CODING VS COMMON INFORMATION BASED (LOW COMPLEXITY) APPROACH

Required receive rates (R_{r_1}, R_{r_2})	Independent coding transmit rate $R_1 + R_2$	CI-based transmit rate $R_{12} + R_1 + R_2$	Rate reduction
$(1.6, 2.8)$	$1.6 + 2.8 = 4.4$	$0.4 + 1.2 + 2.4 = 4$	9%
$(1.5, 2.3)$	$1.5 + 2.3 = 3.8$	$0.3 + 1.2 + 2 = 3.5$	8%
$(1.4, 2)$	$1.4 + 2 = 3.4$	$0.3 + 1.1 + 1.7 = 3.1$	9%

B. Low Complexity Quantizer Design for Laplacian Sources:

This section reports on experiments with the low complexity variant derived for sources that are modeled as Laplacian, wherein some coding optimality is sacrificed for low complexity. We test performance on a Laplacian source with $\lambda = 1$. The receive rates were first set to $c_1 = 1.6$ and $c_2 = 2.8$. Fig. 11 depicts the excess distortion ΔD versus the transmit rate R_t, obtained by quantizers designed at common layer rates (R_{12}) ranging from 0 (independent coding) to 1.6 bits (scalable coding using CELQ), as well as the convex hull for common information coding, obtained by time sharing. Note again that the scalable coding employed by current standards incurs about 1.5 dB excess distortion over our own scalable coding solution, whose excess distortion is 0.8 dB over non-scalable coding. The proposed technique can operate at all points along the convex hull and at considerably better performance compared to the scalable coding of current standards.

In this experiment we also note an interesting operating point in Fig. 11, at $R_t = 4$ or equivalently $R_{12} = 0.4$, where distortion close to non-scalable coding is maintained, but at a 9% reduction in total transmit rate. We experimented with various receive rate constraints to obtain similar operating points where rate reduction is achieved at negligible excess distortion cost. These are tabulated in Table II and demonstrate the capability of the approach to efficiently extract the common information across quality levels.

In summary, comparing the results for joint versus low complexity design, we observe that joint design offers more extensive coding gains due to better extraction of commoin information, while the low complexity approach offers more modest gains but at obvious practical implementation benefits.

TABLE III
TRANSMIT RATE AT NEGLIGIBLE EXCESS DISTORTION: INDEPENDENT CODING VS COMMON INFORMATION BASED APPROACH, MULTIVARIATE NORMAL SOURCE

Required receive rates (R_{r_1}, R_{r_2})	Independent coding transmit rate $R_1 + R_2$	CI-based transmit rate $R_{12} + R_1 + R_2$	Rate reduction
$(2.3, 3.5)$	$2.3 + 3.5 = 5.8$	$0.5 + 1.8 + 3 = 5.3$	8%
$(1.5, 3.6)$	$1.5 + 3.6 = 5.1$	$1.1 + 0.4 + 2.5 = 4$	21%
$(4.0, 5.3)$	$4.0 + 5.3 = 9.3$	$2.1 + 1.9 + 3.2 = 7.2$	22%
$(3.1, 3.8)$	$3.1 + 3.8 = 6.9$	$1.6 + 1.5 + 2.2 = 5.3$	23%
$(3.3, 4.3)$	$3.3 + 4.3 = 7.6$	$2.1 + 1.2 + 2.2 = 5.5$	27%
$(2.6, 4.8)$	$2.6 + 4.8 = 7.4$	$2 + 0.6 + 2.8 = 5.4$	27%
$(3.7, 5.4)$	$3.7 + 5.4 = 9.1$	$2.5 + 1.2 + 2.9 = 6.6$	27%

TABLE IV
THREE QUALITY LEVELS SETTING: TRANSMIT RATE AT NEGLIGIBLE EXCESS DISTORTION: SEPARATE CODING VS COMMON INFORMATION BASED APPROACH, MULTIVARIATE NORMAL SOURCE

Required receive rates $(R_{r_1}, R_{r_2}, R_{r_3})$	Independent coding transmit rate $R_1 + R_2 + R_3$	CI-based transmit rate $R_1 + R_2 + R_3 + R_{123} + R_{23}$	Rate reduction
$(1.2, 1.5, 2.0)$	$1.2 + 1.5 + 2.0 = 4.7$	$0.2 + 0.2 + 0.7 + 1 + 0.3 = 2.4$	49%
$(1.2, 2.0, 2.3)$	$1.2 + 2.0 + 2.3 = 5.5$	$0.2 + 0.4 + 0.7 + 1 + 0.6 = 2.9$	47%
$(0.8, 2.6, 3.3)$	$0.8 + 2.6 + 3.3 = 6.7$	$0.2 + 2.0 + 2.7 + 0.6 + 0 = 5.5$	22%
$(2.2, 2.7, 2.85)$	$2.2 + 2.7 + 2.85 = 7.75$	$1 + 0.55 + 0.7 + 1.2 + 0.95 = 4.4$	43%
$(3.15, 3.25, 3.55)$	$3.15 + 3.25 + 3.55 = 9.95$	$0.85 + 0.35 + 0.65 + 2.3 + 0.6 = 4.75$	52%
$(3.3, 3.4, 3.6)$	$3.3 + 3.4 + 3.6 = 10.3$	$1 + 0.7 + 0.9 + 2.3 + 0.4 = 5.3$	48%

C. Joint Design of Vector Quantizers

Consider a vector source drawn from multivariate normal distribution $x \sim N(\mu, \Sigma)$ with $\mu = (0, 1)$ and $\Sigma = [\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}]$. For independent (non-scalable) coding, we employ the generalized Lloyd algorithm [39] to design the ECVQ. Note that while Gaussian sources are successively refinable asymptotically (in delay/block-length), this is not true at finite delay, which suggests potential benefits to our proposed common information paradigm. As shown in Table III, similar to the scalar case we have significant reduction in total transmit rate compared to independent coding at negligible cost in excess distortion. These results further support the effectiveness of the proposed approach in both scalar and vector quantizer design.

We further observe that even though the design algorithm does not impose regularity on the codebook initialization (step 1 of the algorithm), the ultimate quantizers were always regular, as indeed anticipated by the lemma.

D. Design for Multiple Quality Levels

Here we experiment with the same multivariate Gaussian source which is encoded to three quality levels $L = 3$. In this case, the encoder generates five packets at rates R_1, R_2, R_3, R_{23}, and R_{123}. Decoders 1, 2, and 3 receive packets at rates $R_{r_1} = R_1 + R_{123}$, $R_{r_2} = R_2 + R_{23} + R_{123}$, and $R_{r_3} = R_3 + R_{23} + R_{123}$, respectively. The total transmit rate in the common information paradigm is $R_t = R_1 + R_2 + R_3 + R_{23} + R_{123}$.

The results summarized in Table IV show significant reduction in total transmit rate compared to non-scalable coding, at negligible cost in excess distortion. If we compare the two quality level ECVQ case and the multi-level case, we observe that, on average, more significant coding gains in multi-level design, with roughly 50% reduction in total transmit rate compared to 30% reduction in the two layer case. This is intuitively reasonable, as the multi-level scenario offers more opportunity to exploit inter-layer common information.

Note that the results presented in Tables I, II, III, and IV, are operating points (no loss compare to the independent coding of the sources, with lower rate) which the other state of the arts approaches such as scalable coding cannot achieve.

VI. Conclusion

A novel fundamental design technique is proposed for the common information based layered coding framework, wherein joint design of quantizers in all common and individual layers, overcomes the limitations of conventional scalable coding and independent (non-scalable) coding. The approach ensures that each decoder receives only information useful to its reconstruction. It further offers the flexibility to achieve better operating points on the optimal tradeoff curve between the extremes of scalable or independent coding. The iterative scalar and vector quantizer design techniques optimize all the quantizers jointly to minimize the overall cost at each iteration. They enable common information extraction between different quality levels at negligible distortion penalty. A low complexity variant was derived for Laplacian sources. Simulation results for scalar Laplacian and Gaussian vector sources provide evidence for the benefits of the proposed approach with up to 50% reduction in total transmit rate compared to independent coding.

Appendix

To obtain the optimal partition for the common layer quantizer, we minimize the Lagrangian cost

$$J = a_1 D_1 + a_2 D_2 + \lambda_1 (R_1 + R_{12}) + \lambda_2 (R_2 + R_{12}) + \lambda_{12} R_{12}, \tag{17}$$

over the decision points t_i^{12}. The receive rate at decoder l is

$$R_l + R_{12} = -\sum_{i=1}^{M} \sum_{q=1}^{N_i^l} p_{q,i}^l \log_2 p_{q,i}^l, \qquad l = 1, 2; \tag{18}$$

the common rate is

$$R_{12} = -\sum_{i=1}^{M} p_i^{12} \log_2 p_i^{12}, \tag{19}$$

and the distortion at decoder l is

$$D_l = \sum_{i=1}^{M} \sum_{q=1}^{N_i^l} \int_{t_{q-1,i}^l}^{t_{q,i}^l} (x - x_{q,i}^l)^2 f(x) dx, \qquad l = 1, 2. \tag{20}$$

Also note that $t_{N_i^l,i}^l = t_i^{12}$ and $t_{0,i+1}^l = t_i^{12}$. Differentiating J with respect to t_i^{12}:

$$\frac{\partial J}{\partial t_i^{12}} = a_1 \frac{\partial D_1}{\partial t_i^{12}} + a_2 \frac{\partial D_2}{\partial t_i^{12}} + \lambda_1 \frac{\partial (R_1 + R_{12})}{\partial t_i^{12}} + \lambda_2 \frac{\partial (R_2 + R_{12})}{\partial t_i^{12}} + \lambda_{12} \frac{\partial R_{12}}{\partial t_i^{12}}, \tag{21}$$

where

$$\begin{aligned}\frac{\partial D_l}{\partial t_i^{12}} &= \frac{\partial \int_{t_{N_i^l-1,i}^l}^{t_{N_i^l,i}^l} (x - x_{N_i^l,i}^l)^2 f(x) dx}{\partial t_i^{12}} + \frac{\partial \int_{t_{0,i+1}^l}^{t_{1,i+1}^l} (x - x_{1,i+1}^l)^2 f(x) dx}{\partial t_i^{12}} \\ &= (t_i^{12} - x_{N_i^l,i}^l)^2 f(t_i^{12}) - (t_i^{12} - x_{1,i+1}^l)^2 f(t_i^{12}) \\ &= (((x_{N_i^l,i}^l)^2 - (x_{1,i+1}^l)^2) - 2\, t_i^{12} (x_{N_i^l,i}^l - x_{1,i+1}^l)) f(t_i^{12}),\end{aligned} \tag{22}$$

and

$$\begin{aligned}-\frac{\partial (R_l + R_{12})}{\partial t_i^{12}} &= \frac{\partial (p_{N_i^l,i}^l \log_2 p_{N_i^l,i}^l)}{\partial t_i^{12}} + \frac{\partial (p_{1,i+1}^l \log_2 p_{1,i+1}^l)}{\partial t_i^{12}} \\ &= \frac{d (p_{N_i^l,i}^l \log_2 p_{N_i^l,i}^l)}{d p_{N_i^l,i}^l} * \frac{\partial p_{N_i^l,i}^l}{\partial t_i^{12}} + \frac{d (p_{1,i+1}^l \log_2 p_{1,i+1}^l)}{d p_{1,i+1}^l} * \frac{\partial p_{1,i+1}^l}{\partial t_i^{12}} \\ &= (\log_2 e + \log_2 p_{N_i^l,i}^l) * f(t_i^{12}) - (log_2 e + \log_2 p_{1,i+1}^l) * f(t_i^{12}) \\ &= (\log_2 p_{N_i^l,i}^l - \log_2 p_{1,i+1}^l) * f(t_i^{12}),\end{aligned} \tag{23}$$

and similarly,

$$-\frac{\partial (R_{12})}{\partial t_i^{12}} = (\log_2 p_i^{12} - \log_2 p_{i+1}^{12}) * f(t_i^{12}). \tag{24}$$

By setting $\frac{\partial J}{\partial t_i^{12}}$ to zero and reducing $f(t_i^{12})$ from all terms we finally obtain:

$$\begin{aligned}t_i^{12} &= \frac{(a_1 (x_{1,i+1}^1)^2 + a_2 (x_{1,i+1}^2)^2) - (a_1 (x_{N_i^1,i}^1)^2 + a_2 (x_{N_i^2,i}^2)^2)}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))} \\ &\quad - \frac{\lambda_1 (\log_2 p_{1,i+1}^1 - \log_2 p_{N_i^1,i}^1) + \lambda_2 (\log_2 p_{1,i+1}^2 - \log_2 p_{N_i^2,i}^2)}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))} \\ &\quad - \frac{\lambda_{12} (\log_2 p_{i+1}^{12} - \log_2 p_i^{12})}{2((a_1 x_{1,i+1}^1 + a_2 x_{1,i+1}^2) - (a_1 x_{N_i^1,i}^1 + a_2 x_{N_i^2,i}^2))}.\end{aligned} \tag{25}$$

■

References

[1] J. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

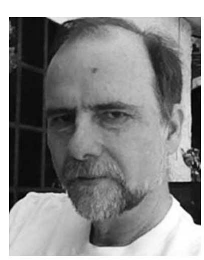
[3] J. Chow and T. Berger, "Failure of successive refinement for symmetric Gaussian mixtures," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 350–352, Jan. 1997.

[4] F. Wu, S. Li, and Y. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 332–344, Mar. 2001.

[5] A. Aggarwal, S. Regunathan, and K. Rose, "Efficient bit-rate scalability for weighted squared error optimization in audio coding," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1313–1327, Jul. 2006.
[6] K. Viswanatha, E. Akyol, T. Nanjundaswamy, and K. Rose, "On common information and the encoding of sources that are not successively refinable," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2012, pp. 129–133.
[7] R. Gray and A. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J.*, vol. 53, no. 9, pp. 1681–1721, Nov. 1974.
[8] A. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.
[9] K. Viswanatha, E. Akyol, and K. Rose, "The lossy common information of correlated sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3238–3253, Jun. 2014.
[10] T. Nanjundaswamy, K. Viswanatha, and K. Rose, "On relaxing the strict hierarchical constraints in layered coding of audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 3680–3684.
[11] M. Salehifar, T. Nanjundaswamy, and K. Rose, "Quantizer design for exploiting common information in layered coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 4463–4467.
[12] C. Shanon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[13] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.
[14] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
[15] T. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 3, pp. 277–288, May 1980.
[16] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.
[17] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.
[18] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.
[19] I. Csiszar, "On the computation of rate-distortion functions (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 1, pp. 122–124, Jan. 1974.
[20] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1939–1952, Nov. 1994.
[21] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[22] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-6, no. 1, pp. 7–12, Mar. 1960.
[23] M. Sabin and R. Gray, "Global convergence and empirical consistency of the generalized lloyd algorithm," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 2, pp. 148–155, Mar. 1986.
[24] W. Equitz and T. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.
[25] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.
[26] A. Gamal and T. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.
[27] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 502–521, Jul. 1987.
[28] R. Venkataramani, G. Kramer, and V. Goyal, "Multiple description coding with many channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2106–2114, Sep. 2003.
[29] K. Viswanatha, E. Akyol, and K. Rose, "Combinatorial message sharing for a refined multiple descriptions achievable region," in *Proc. IEEE Int. Symp. Inf. Theory*, Aug. 2011, pp. 1312–1316.
[30] E. Tuncel and K. Rose, "Additive successive refinement," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1983–1991, Aug. 2003.
[31] K. Viswanatha, E. Akyol, and K. Rose, "Combinatorial message sharing and a new achievable region for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 769–792, Feb. 2016.
[32] P. Gacs and J. Korner, "Common information is far less than mutual information," *Problems Control Inf. Theory*, vol. 2, no. 2, pp. 149–162, 1973.
[33] R. Ahlswede and J. Korner, "On common information and related characteristics of correlated information sources," in *Proc. 7th Prague Conf. Inf. Theory*, Sep. 1974, republished in *General Theory of Information Transfer and Combinatorics* (Lecture Notes in Computer Science), R. Ahlswede *et al.*, Eds., Berlin: Springer, vol. 4123, 2006, pp. 664–677.
[34] G. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1365–1374, Sep. 1996.
[35] Y. Yan and P. Andrivon, "The scalable extensions of HEVC for ultra-high-definition video delivery," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 58–64, Sep. 2014.
[36] *Information Technology - Coding of Audio-Visual Objects - Part 3: Audio - Subpart 4: General Audio Coding (GA)*, ISO/IEC Std. ISO/IEC JTC1/SC29 14496-3:2005, 2005.
[37] A. György and T. Linder, "Optimal entropy-constrained scalar quantization of a uniform source," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2704–2711, Nov. 2000.
[38] G. Sullivan, "On embedded scalar quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* , vol. 4, May 2004, pp. 605–608.
[39] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

Mehdi Salehifar (Member, IEEE) received the B.Sc. degree in electrical and computer engineering from the University of Tehran, Tehran, Iran, in 2012, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, Santa Barbara, CA, USA, in 2014 and 2017, respectively. He is currently with LG Electronics as a Staff Research Engineer. His research interests include signal processing, general quantization theory, information theory, and video or audio coding. He was with LG Electronics as a Senior Research Engineer, Video Researcher Intern, and with Qualcomm as an Audio Researcher. He was the recipient of several titles in national and international mathematic competitions.

Tejaswi Nanjundaswamy (Member, IEEE) received the B.E. degree in electronics and communications engineering from the National Institute of Technology Karnataka Surathkal, Mangalore, India, in 2004, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara (UCSB), Santa Barbara, CA, USA, in 2009 and 2013, respectively. He is currently an Audio Codec Engineer with Apple Inc., Cupertino, CA, USA. He was with Signal Compression Lab, UCSB, Infocoding Labs Inc., Mayachitra Inc., and Ittiam Systems, where he focused on audio/video compression, processing, and related technologies. He was the recipient of the Student Technical Paper Award at the AES 129th Convention.



Kenneth Rose (Fellow, IEEE) received the Ph.D. degree from the California Institute of Technology, Pasadena, CA, USA, in 1991. He then joined the Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, USA, where he is currently a Distinguished Professor. His main research interests include information theory and signal processing, and include rate-distortion theory, source and source-channel coding, audio-video coding and networking, pattern recognition, and nonconvex optimization. He is interested in the relations between information theory, estimation theory, statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines. He was the senior co-author of a paper for which his students received the 2015 IEEE Signal Processing Society Young Author Best Paper Award. He was the recipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society and the 2004 and 2007 IEEE Signal Processing Society Best Paper Awards.