Robust Distributed Compression with Learned Heegard–Berger Scheme

Eyyüp Taşçı
Dept. of Electrical and Electronics Engineering
Bogazici University, Istanbul, TURKEY
eyyup.tasci@std.bogazici.edu.tr

Ezgi Özyılkan, Oğuzhan Kubilay Ülger, Elza Erkip
Dept. of Electrical and Computer Engineering
New York University, New York, USA
{ezgi.ozyilkan, kubi, elza}@nyu.edu

***Abstract*—We consider lossy compression of an information source when decoder-only side information may be absent. This setup, also referred to as the Heegard–Berger or Kaspi problem, is a special case of robust distributed source coding. Building upon previous works on neural network-based distributed compressors developed for the decoder-only side information (Wyner–Ziv) case, we propose learning-based schemes that are amenable to the availability of side information. We find that our learned compressors mimic the achievability part of the Heegard–Berger theorem and yield interpretable results operating close to information-theoretic bounds. Depending on the availability of the side information, our neural compressors recover characteristics of the point-to-point (i.e., with no side information) and the Wyner–Ziv coding strategies that include *binning* in the source space, although no structure exploiting knowledge of the source and side information was imposed into the design.**

I. Introduction

Imagine a distributed sensor network consisting of individual cameras positioned across various locations within a city, each independently capturing images of its surroundings. In this scenario, each sensor node compresses and sends its correlated image to a central processing unit, which then combines them to generate a comprehensive visual map of the city. However, direct communication among sensors is often infeasible, and unreliable channels (e.g., due to fading) may further hinder communication. This poses a key question: How can we leverage the correlation among sensor data in a robust manner, preventing a system failure when some nodes cannot transmit their observations? In this work, we take a first step on addressing the link failure scenario, where the decoder may not receive some correlated sensor data.

Distributed source coding (DSC) refers to the task of efficiently compressing information from physically separated encoders. Wyner and Ziv (WZ) [1] examined a simple lossy distributed compression case where the decoder has access to a correlated source, known as the *side information*, losslessly. WZ theorem relies on joint typicality and *random binning* arguments, and is non-constructive. Although the theory of DSC predicts substantial improvements in compression efficiency compared to the point-to-point setups [2], developing practical distributed compressors operating in the finite blocklength regime remains a challenging open problem to date. Recent work on learned Wyner–Ziv compressors [3], [4] showed binning-like behavior for the quadratic-Gaussian sources. Additionally, practical models employing neural networks to compress high-dimensional source data for distributed stereo image compression [5], [6] and for distributed task-aware image compression have been proposed [7].

This work was in part done during E. Taşçı's summer internship at NYU.

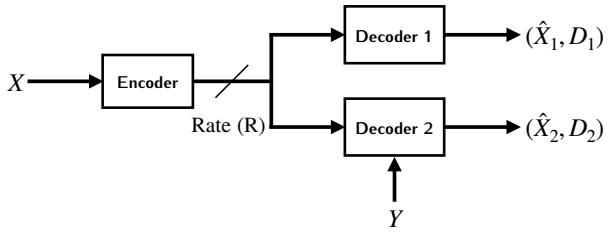


Fig. 1: Lossy source coding when side information may be absent, also known as the *Heegard–Berger* or the *Kaspi* problem.

An interesting question is whether it is still possible to achieve a reconstruction with non-trivial distortion when side information fails to reach the decoder in the WZ setup. An equivalent formulation of this problem containing two decoders, one with side information and the other without, is illustrated in Fig. 1. Heegard and Berger (HB) [8] fully characterized the asymptotic rate–distortion (R-D) function for the quadratic-Gaussian case of Fig. 1, and also considered scenarios when different side information is available at several decoders. In [9], Kaspi also established the R-D function when side information is available at both encoder and decoder. Intuitively, one can improve the robustness of a WZ coding system against the absence of side information by reducing the number of codewords in each bin. Such an approach yields a trade-off between compression efficiency and system robustness, encapsulating the essence of the robust DSC scheme developed by Ishwar *et al.* [10].

Recent studies in [3], [4], [11] have demonstrated that learning-based WZ compressors can recover different types of interpretable random binning mechanisms without any specific structure being imposed onto the design. These results offer empirical evidence that learned distributed compressors can achieve competitive constructive solutions in the non-asymptotic blocklength, closely resembling those of hand-crafted frameworks such as DISCUS [12], without requiring a priori knowledge of source statistics. In these learned WZ compressors, the encoder sends the bin index, and the decoder infers the quantization index with the help of the side informa-

tion. Finally, the decoder reconstructs the source by utilizing the deduced quantization index and the side information, according to the distortion criterion. In their current form, it is not clear whether these learned WZ compressors can effectively adapt to the unavailability of side information.

Expanding on the learned WZ compressors [3], [4], [11], in this paper we find constructive solutions for the non-asymptotic regime of the HB problem where side information may be absent, by leveraging universal function approximation capabilities of artificial neural networks (ANNs) [13], [14]. Similar to [3], [4], [11], our focus is on the one-shot regime, where each source realization is compressed one at a time, as in popular ANN-based image compressors [15], [16]. We present three unique solutions for the HB problem, where we either jointly address the quantization and binning components (Figs. 2a and 2b) or alternatively, adopt a two-step approach that involves a learned quantizer coupled with an ideal Slepian-Wolf (SW) coder (Fig. 2c).

The paper is organized as follows. We first give an overview of HB theorem, detailing the achievability result (Section II-A) and the quadratic-Gaussian setup we consider (Section II-B). To define training objectives for the proposed robust schemes, we minimize upper bounds on mutual information. These are formulated using one of three probabilistic models employing ANNs, which can also interpreted as operational robust distributed schemes (Section III). Finally, we discuss empirical results (Section IV) and conclude the paper (Section V).

II. Heegard–Berger Problem Setup

A. General setting

Let (X, Y) be a pair of random variables with joint distribution $p(x, y)$ defined on the product alphabet $\mathcal{X} \times \mathcal{Y}$, where X and Y represent the source and decoder-only side information, respectively. As shown in Fig. 1, the encoder wishes to describe the source X to two decoders, an *uninformed* and an *informed* one, with the latter having access to the side information Y. The first decoder reconstructs $\hat{X}_1$ in a point-to-point fashion while the second decoder reconstructs $\hat{X}_2$ with the help of the side information Y, under distortion metrics $d_i : \mathcal{X} \times \hat{\mathcal{X}}_i \to \mathbb{R}_{\geq 0}$ where $\hat{\mathcal{X}}_i$ are the reconstruction alphabets for $i \in \{1, 2\}$. The goal is to find the minimum achievable rate under two expected distortion constraints: $\mathbb{E}[d_i(x, \hat{x}_i)] \leq D_i$ for some $D_i \geq 0$. In the asymptotic blocklength regime, where we consider joint compression of n i.i.d. source samples as $n \to \infty$, the following theorem by Heegard and Berger [8] characterizes the optimal R-D function.

Theorem 1. *(Heegard–Berger Theorem [8]) Let (X, Y) be a pair of random variables with joint distribution $p(x, y)$, representing the source and the correlated side information respectively, and $d_i : \mathcal{X} \times \hat{\mathcal{X}}_i \to \mathbb{R}_{\geq 0}$ be single letter distortion measures for $i \in \{1, 2\}$. The R-D function for X with side information Y available only at one of the decoders is:*

$$R(D_1, D_2) = \min_{p(w,u|x)} (I(X; W) + I(X; U|Y, W)), \quad (1)$$

where the minimization is over all conditional probability distributions $p(w, u|x)$ such that there exists functions $g_1(W) = \hat{X}_1$ and $g_2(W, U, Y) = \hat{X}_2$ satisfying the distortion constraints

$$\mathbb{E}[d_1(X, \hat{X}_1)] \leq D_1, \;\; \mathbb{E}[d_2(X, \hat{X}_2)] \leq D_2, \quad (2)$$

and $(W, U) - X - Y$ is a Markov chain.

We remark that the informed decoder encounters a WZ problem [1] (cf. second mutual information term in Eq. (1)), while the uninformed decoder is subjected to an ordinary point-to-point R-D problem (cf. first mutual information term in Eq. (1)).

Note that the HB theorem considers two distortion constraints as in Eq. (2). Alternatively, one can opt for a combined weighted sum distortion constraint given as:

$$\mathbb{E}\Big[\beta d_1(X, \hat{X}_1) + (1 - \beta) d_2(X, \hat{X}_2)\Big] \leq \boldsymbol{D}, \quad (3)$$

for some $\beta \in [0, 1]$. In this case, the asymptotically minimum weighted distortion for a fixed rate is given as [17]:

$$\boldsymbol{D}^*(R) = \min_{D_1, D_2 : R(D_1, D_2) \leq R} \beta D_1 + (1 - \beta) D_2, \quad (4)$$

where $R(D_1, D_2)$ is given in Theorem 1. This weighted distortion measure can also be conceptualized as having a β probability of receiving no side information (see the discussion in [17]). As will be seen in Section III, considering a combined distortion constraint, such as the one in Eq. (4), simplifies the learning procedure of HB compressors, and yet, still offers valuable insights about how close they operate with respect to the optimum.

B. Quadratic-Gaussian case

Similarly to the WZ R-D function [1], the HB formula in Eq. (1) has a closed-form expression only in few special cases [8]. To compare the performance of our learned distributed compressors to the asymptotic HB R-D function, we consider a quadratic-Gaussian setup: Suppose $Y = X + N$, where $N \sim \mathrm{N}(0, \sigma_n^2)$ is independent from $X \sim \mathrm{N}(0, \sigma_x^2)$, and the distortion metrics are $d_i(x, \hat{x}_i) = (x - \hat{x}_i)^2$ for $i \in \{1, 2\}$. The analytical expression of $R(D_1, D_2)$ for this case is given as [8]:

$$R(D_1, D_2) = \frac{1}{2} \log\left(\frac{\sigma_x^2}{\Delta_1}\right) + \frac{1}{2} \log\left(\frac{\sigma_n^2 \Delta_1}{\Delta_2(\Delta_1 + \sigma_n^2)}\right), \quad (5)$$

where $\Delta_1 = \min(\sigma_x^2, D_1)$, $\Delta_2 = \min([1/\Delta_1 + 1/\sigma_n^2]^{-1}, D_2)$. The two terms of Eq. (5) respectively correspond to the two mutual information terms given in Theorem 1. The first and second expressions in Eq. (5) represent the rate required to describe W and U under the distortion constraints of D_1 and D_2, respectively. There are four operationally different R-D regions depending on the value of the pair (D_1, D_2). When $D_1 \leq \sigma_x^2$ and $D_2 \geq \frac{D_1 \sigma_n^2}{D_1 + \sigma_n^2}$, the second term in Eq. (5) vanishes, and the problem degenerates into a standard point-to-point lossy compression problem i.e., having only the first decoder in Fig. 1. Similarly, when $D_1 \geq \sigma_x^2$, and $D_2 \leq \frac{D_1 \sigma_n^2}{D_1 + \sigma_n^2}$ the first term in Eq. (5) vanishes, and the problem degenerates to the WZ coding problem i.e., having only the second decoder in Fig. 1. When $D_1 \geq \sigma_x^2$, and $D_2 \geq \frac{D_1 \sigma_n^2}{D_1 + \sigma_n^2}$, both terms

vanish, and the distortion constraints can be trivially satisfied with zero rate. The most operationally interesting region arises when $D_1 \leq \sigma_x^2$ and $D_2 \leq \frac{D_1 \sigma_n^2}{D_1 + \sigma_n^2}$. Here, both rate terms are non-zero, forcing the encoder to allocate its total rate budget between the point-to-point R-D problem of describing W and the WZ R-D problem of describing U. Optimal weighted rate-distortion function in Eq. (4) for the quadratic-Gaussian setting can be found in [18].

III. Operational Neural Heegard–Berger Schemes

For our learned compressors, we first consider the system model in Fig. 2a, which corresponds to a straightforward parametrization of the HB setup depicted in Fig. 1. Note that in this case, a joint description is sent to both uninformed and informed decoders, the latter having access to the side information Y. By introducing the auxiliary variable $V = (W, U)$, we first assume that the encoder in the achievability proof of HB theorem can be represented by a probability model $p_{\boldsymbol{\theta}}(v|x)$ with parameters $\boldsymbol{\theta}$. This formulation, which we name as *joint*, yields the following variational upper bound:

$$I(X;W) + I(X;U|Y,W) \leq I(X;U,W) = I(X;V), \tag{6}$$

$$\leq \mathbb{E}_{\substack{p(x) \\ p_{\boldsymbol{\theta}}(v|x)}} \left[\log \frac{p_{\boldsymbol{\theta}}(v|x)}{q_{\boldsymbol{\eta}}(v)} \right], \tag{7}$$

$$= \mathbb{E}_{p(x)} \, \mathrm{r_j}(x), \tag{8}$$

where $q_{\boldsymbol{\eta}}(v)$, with parameters $\boldsymbol{\eta}$, is a model of the distribution $p(v) \triangleq p(w,u)$, which is generally not known in closed form. $\mathrm{r_j}(x)$ refers to the expectation in Eq. (7) with respect to $p_{\boldsymbol{\theta}}(v|x)$. The upper bound in Eq. (8) follows from cross-entropy [19] being larger or equal to entropy [2].

Next, we consider the system models illustrated in Figs. 2b and 2c. In these schemes, we opt for a layered encoding approach, aligned more closely with the HB theorem, where we separately encode the auxiliary variables W and U with probability models $p_{\boldsymbol{\omega}}(w|x)$ and $p_{\boldsymbol{\gamma}}(u|w,x)$ (with parameters $\boldsymbol{\omega}$ and $\boldsymbol{\gamma}$), respectively. For our objective functions, building onto Theorem 1, we will consider either a *marginal* or a *conditional* formulation [3] whose variational upper bounds respectively are:

$$I(X;W) + I(X;U|Y,W) \leq \mathbb{E}_{\substack{p(x) \\ p_{\boldsymbol{\omega}}(w|x) \\ p_{\boldsymbol{\gamma}}(u|w,x)}} \left[\log \frac{p_{\boldsymbol{\omega}}(w|x)}{q_{\boldsymbol{\zeta}}(w)} + \log \frac{p_{\boldsymbol{\gamma}}(u|w,x)}{q_{\boldsymbol{\psi}}(u|w)} \right] = \mathbb{E}_{p(x)} \, \mathrm{r_m}(x), \tag{9}$$

$$I(X;W) + I(X;U|Y,W) \leq \mathbb{E}_{\substack{p(x,y) \\ p_{\boldsymbol{\omega}}(w|x) \\ p_{\boldsymbol{\gamma}}(u|w,x)}} \left[\log \frac{p_{\boldsymbol{\omega}}(w|x)}{q_{\boldsymbol{\zeta}}(w)} + \log \frac{p_{\boldsymbol{\gamma}}(u|w,x)}{q_{\boldsymbol{\mu}}(u|w,y)} \right] = \mathbb{E}_{p(x,y)} \, \mathrm{r_c}(x,y). \tag{10}$$

where $\mathrm{r_m}(x)$ and $\mathrm{r_c}(x,y)$ refer to respective expectations in Eqs. (9) and (10) with respect to $p_{\boldsymbol{\omega}}(w|x)$ and $p_{\boldsymbol{\gamma}}(u|w,x)$. Here $q_{\boldsymbol{\zeta}}(w)$ (with parameters $\boldsymbol{\zeta}$) is a model of the distribution $p(w)$, and $q_{\boldsymbol{\psi}}(u|w)$ and $q_{\boldsymbol{\mu}}(u|w,y)$ (with parameters $\boldsymbol{\psi}$ and $\boldsymbol{\mu}$, respectively) are two different models of the distribution $p(u|w,y)$.

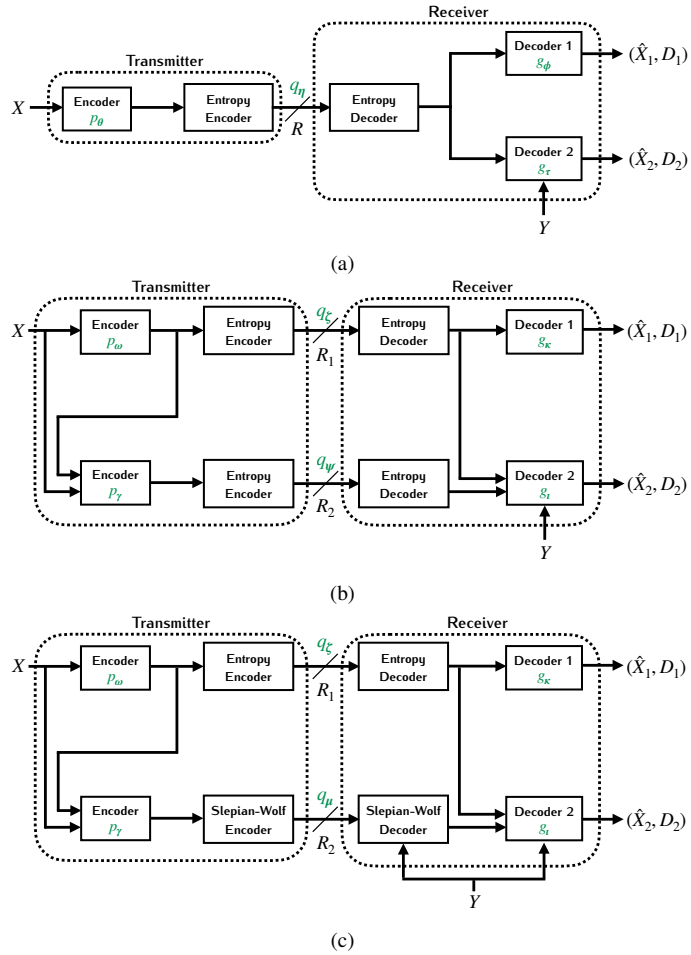


Fig. 2: The three lossy compression systems that we consider: (a) learned compressor sending a joint description for both decoders using a classic entropy coder (i.e., the *joint* formulation, see Eq. (13)), (b) learned compressors sending individual descriptions to both decoders using classic entropy coders (i.e., the *marginal* formulation, see Eq. (14)), and (c) using a combination of a classic entropy coder and an ideal Slepian–Wolf coder (i.e., the *conditional* formulation, see Eq. (15)). The learned parameters are indicated in green.

We will utilize these probabilistic models (i.e., $p_{\boldsymbol{\theta}}(v|x), p_{\boldsymbol{\omega}}(w|x)$ and $p_{\boldsymbol{\gamma}}(u|w,x)$) to aid in the learning process of respective encoders. We set the encoder outputs in a deterministic way, similar to [3], e.g., $v = \arg\max_h p_{\boldsymbol{\theta}}(h|x)$. To actualize a practice-oriented compression setting, we also set all encoder outputs (i.e., V, W, and U) as discrete. Similarly to [3], [4], [11], without loss of generality, we define all probabilistic models as discrete distributions with probabilities $P_k = \frac{\exp \alpha_k}{\sum_{i=1}^{K} \exp \alpha_i}$ for $k \in \{1, \ldots, K\}$, where K is a model parameter. The unnormalized log-probabilities (*logits*) are either directly treated as learnable parameters (i.e., $q_{\boldsymbol{\eta}}(v)$ and $q_{\boldsymbol{\zeta}}(w)$), or computed by ANNs as functions of the conditioning variable (e.g., w and y for $q_{\boldsymbol{\mu}}(u|w,y)$). These design choices keep the parametric families as general as possible without imposing any structure. Specifically, this allows the encoders to learn, when needed, quantization schemes that may incorporate discontiguous bins in the

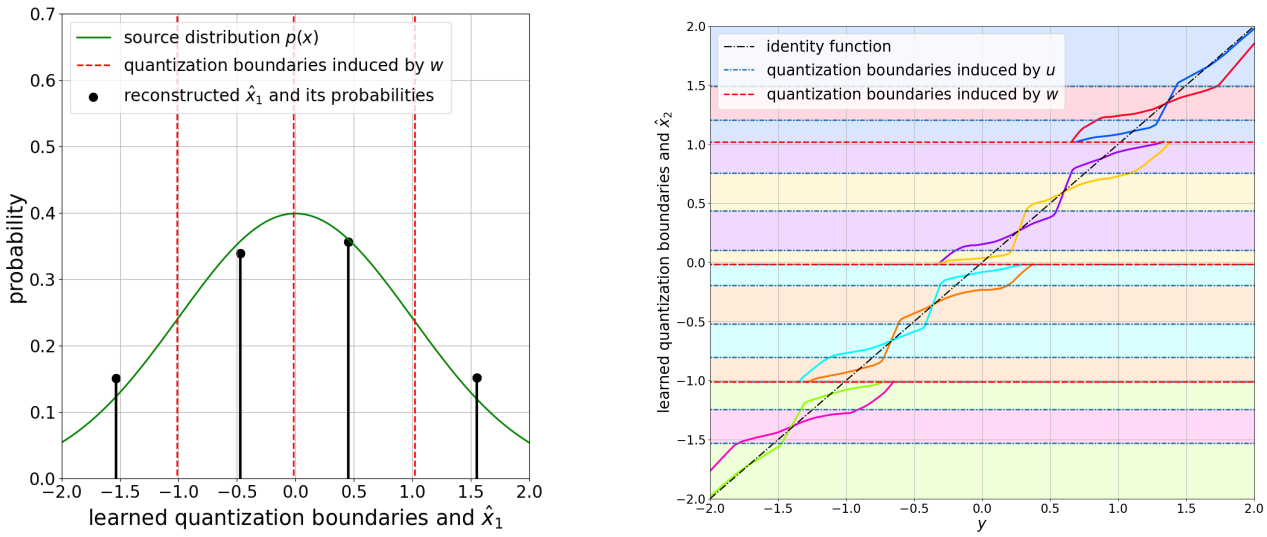


Fig. 3: Visualizations of the learned optimized encoders, $w = \arg\max_h p_{\boldsymbol{\omega}}(h|x)$ and $u = \arg\max_l p_{\boldsymbol{\gamma}}(l|w,x)$ in Eq. (9), and of the decoders, $\hat{x}_1 = g_{\boldsymbol{\kappa}}(w)$ and $\hat{x}_2 = g_{\boldsymbol{\iota}}(w,u,y)$ in Eq (12), for the marginal formulation (see Fig. 2b and Eq. (14)). **Left:** The dashed vertical red lines are quantization boundaries induced by w, and the codebook points represent the outputs of the decoder $g_{\boldsymbol{\kappa}}$ to which all source values within the corresponding quantization region are mapped. The height of each codebook stalk represents the likelihood of that code vector under the entropy model $q_{\boldsymbol{\zeta}}$ (see Eq. (9)). **Right:** The dashed horizontal red lines coincide with those depicted in the left panel induced by w, while the dotted-dashed horizontal blue lines are quantization boundaries induced by u. The colors between each boundary represent a unique (w,u) pair. We depict the decoding function learned by $g_{\boldsymbol{\iota}}$ with the solid lines, each representing a different pair of (w,u) as inputs within its respective quantization region, distinguished by unique colors. The experimental setup parameters (see Section II-B) are configured as $\sigma_x^2 = 1.00$, $\sigma_n^2 = 0.01$ and $\beta = 0.20$. This model attains -15.55 dB at a rate of 2.85 bits

source space, resembling the random binning operation in the achievability part of the HB and WZ theorems. Previously, it was demonstrated in [3], [4], [11] that a similarly parametrized neural distributed compressor can learn different types of interpretable binning mechanisms for the WZ problem, such as periodic-like mappings for the quadratic-Gaussian case.

Since we focus on a weighted distortion metric (see Eq. (4)), instead of having individual distortion constraints as is the case for the HB theorem (see Section II), we define the following distortion functions for the cases where there is either joint (see Fig. 2a) or separate descriptions (see Fig. 2b and Fig. 2c) to the decoders, respectively, as:

$$\mathrm{d_j}(x,y) = \beta d(x, g_{\boldsymbol{\phi}}(v)) + (1-\beta) d(x, g_{\boldsymbol{\tau}}(v,y)), \tag{11}$$

$$\mathrm{d_s}(x,y) = \beta d(x, g_{\boldsymbol{\kappa}}(w)) + (1-\beta) d(x, g_{\boldsymbol{\iota}}(u,w,y)), \tag{12}$$

where $g_{\boldsymbol{\phi}}(v)$ and $g_{\boldsymbol{\kappa}}(w)$ denote the uninformed decoders, represented by ANNs with parameters $\boldsymbol{\phi}$ and $\boldsymbol{\kappa}$ respectively, which yield $\hat{x}_1$. Similarly, $g_{\boldsymbol{\tau}}(v,y)$ and $g_{\boldsymbol{\iota}}(u,w,y)$ correspond to the informed decoders, represented by ANNs with parameters $\boldsymbol{\tau}$ and $\boldsymbol{\iota}$ respectively, which output $\hat{x}_2$. Building on the upper bounds developed in Eqs. (8), (9) and (10) in tandem with the distortion metrics defined in Eqs. (11) and (12), this yields the loss functions for three different variants we consider:

$$L_{\mathrm{j}}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\tau}) = \mathbb{E}[\mathrm{r_j}(x) + \lambda \cdot \mathrm{d_j}(x,y)], \tag{13}$$

$$L_{\mathrm{m}}(\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\zeta},\boldsymbol{\psi},\boldsymbol{\kappa},\boldsymbol{\iota}) = \mathbb{E}[\mathrm{r_m}(x) + \lambda \cdot \mathrm{d_s}(x,y)], \tag{14}$$

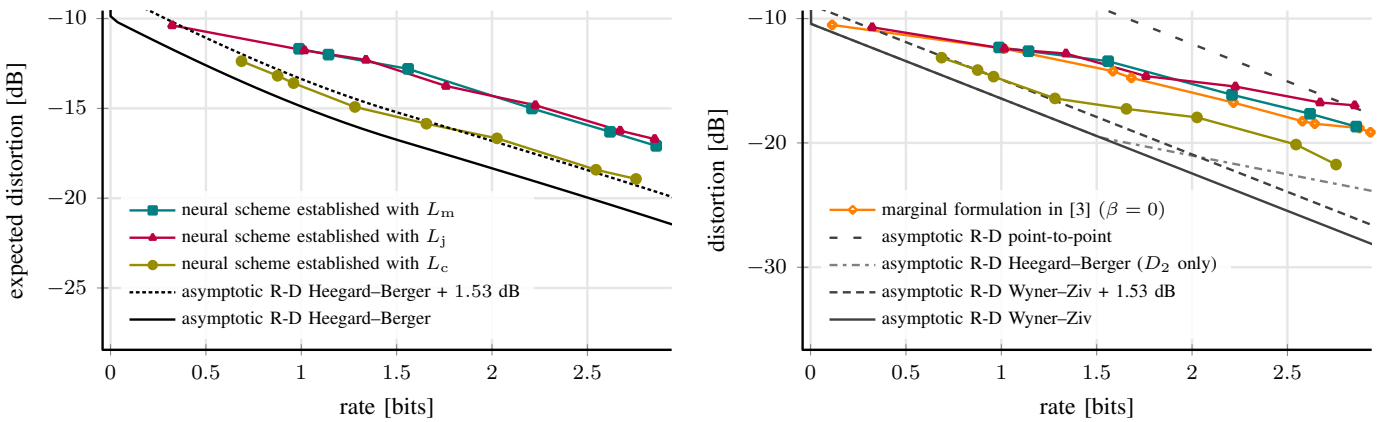
$$L_{\mathrm{c}}(\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\zeta},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\iota}) = \mathbb{E}[\mathrm{r_c}(x,y) + \lambda \cdot \mathrm{d_s}(x,y)], \tag{15}$$

where we relax the constrained formulation of the HB theorem to unconstrained ones using Lagrange multipliers. Here, $\{\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\tau},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\zeta},\boldsymbol{\psi},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\iota}\}$ are optimization parameters, and $\lambda > 0$ controls the trade-off. We can obtain different points in the achievable R-D region induced by the weighted distortion constraint by simply varying the parameter λ. The optimized $p_{\boldsymbol{\theta}}(v|x), p_{\boldsymbol{\omega}}(w|x)$ and $p_{\boldsymbol{\gamma}}(u|w,x)$ models yield the ANN-based encoders depicted in Fig. 2. Similarly, the optimized $\big(g_{\boldsymbol{\phi}}(v), g_{\boldsymbol{\tau}}(v,y)\big)$ and $\big(g_{\boldsymbol{\kappa}}(w), g_{\boldsymbol{\iota}}(u,w,y)\big)$ correspond to a pair of uninformed-informed decoder components for joint and separate description cases, respectively.

Following the popular class of neural compressors [15], [16], [21], we use stochastic gradient descent (SGD) to jointly optimize all learnable parameters. Since SGD relies on a Monte Carlo approximation of the expectations in the loss functions, we use the Gumbel-max technique, as in previous work [3], for sampling from discrete distributions. Similarly, we also leverage Concrete distributions [22] to aid stochastic optimization. Discussion on the operational meaning of the different schemes and additional details about our experimental setup are provided in [18].

IV. Discussion

We first consider a setting where the proposed neural compressors clearly recover some elements of the optimal theoretical solution for the lossy source coding problem where side information may be absent. The visualization of the learned compressor obtained with the marginal scheme (see Fig. 2b) is provided in Fig. 3. As seen in the left panel of the figure, the neural encoder $p_{\boldsymbol{\omega}}$ quantizes the source in a manner similar to standard point-to-point lossy compression. Looking at the right panel of the figure, we remark that the discontiguous quantization bins are learned as the neural compressor



(a) (b)

Fig. 4: Rate–distortion (R-D) performances obtained with joint, marginal and conditional formulations (see Eqs. (13), (14) and (15)), where experimental setup parameters (see Section II-B) are set as $\sigma_x^2 = 1.00$, $\sigma_n^2 = 0.10$ and $\beta = 0.01$. In both panels, we plot the empirical results versus the asymptotic bounds. We provide the expected distortion achieved by two decoders (see Fig. 2) in the left panel, while we only plot the distortion attained by the informed decoder, which has access to side information, in the right panel. The 1.53 dB refers to the mean-squared error gap that the entropy-constrained one-shot lattice quantizer is subjected to in high-rate regime [20].

exhibits periodic-like mappings in the source space, akin to the binning-like behavior recovered by the neural WZ compressor proposed in [3], [4], [11]. We note that this grouping behavior is aligned with the achievability part of the HB theorem, which yields a rate discount similar to the random binning argument invoked in the WZ proof. The figure demonstrates that this learned distributed compressor, formulated based on the proposed marginal approach, exhibits a greater adaptability to robust scenarios compared to the WZ compression case previously explored in [3], [4], [11]. This is evident from the model's ability to recover both WZ and point-to-point coding strategies, despite not imposing any explicit structure exploiting the source knowledge onto the design. Although we noticed similar behaviors persisting in different experimental configurations other than the one considered in Fig. 3 for the marginal variant, we did not observe such consistent binning-like behavior in the visualizations of the compressors obtained by the joint and conditional formulations (not shown). For the conditional scheme, similar to the explanation in [3], we posit that this is due to having an ideal SW coder in the system design (see Fig. 2c), unlike the marginal variant (see Fig. 2b). We speculate that this choice of entropy coding scheme incentivizes the model to delegate the task of binning solely to the ideal SW code, rather than to the learned encoder.

In Fig. 4a, we provide R-D performances achieved with these three different formulations. The joint and marginal variants achieve similar performances, whereas the conditional model outperforms them, approaching the asymptotic HB bound. We attribute the comparable performances of the marginal and joint schemes to the operational equivalence between having two layered descriptions managed by separate classical entropy coders (see Eq. (9)) and having a single classical entropy coder with the input being the unified description of these two representations (see Eq. (8)). We explain the improved R-D performance of the conditional formulation as follows. The SW code that this variant employs, which may leverage high-dimensional channel codes (e.g., as in DISCUS [12]), enables binning over long sequences, i.e., in a multi-shot fashion. This type of *compress-bin* [23] is much more efficient than the one that could be attained by the learned encoder, which could only bin a one-shot manner as it compresses each source realization one at a time. We refer the reader to [18] for a discussion on an additional set of experiments.

In Fig.4b, we illustrate the trade-off between system robustness and compression efficiency by evaluating distortion attained only with the informed decoders of all three proposed schemes, along with the learned WZ compressor proposed in [3] (which operationally coincides with having $\beta = 0$ in Eq. (4), that is decoder-only side information is always assumed to be available). Recalling that all three proposed learned compressors are now also formulated to accommodate scenarios where side information may be absent, which is not the case for the learned WZ compressor studied in [3], a decline in distortion is expected. Interestingly, the conditional variant again surpasses all other learned schemes considered, underscoring the efficiency of high-dimensional binning capability facilitated by the SW coder.

V. Conclusion

In this work, we have proposed three learning-based solutions to the problem of lossy source coding when side information may be absent, for which the optimal theoretical solution is asymptotic and non-constructive. By explicitly visualizing the behavior of the learned encoders and decoders, we demonstrated that they recover schemes aligned with the characteristics of the achievability of the HB theorem, that is WZ and standard lossy source coding. Future research directions include analyzing the robustness in fully distributed compression scenarios. Another interesting direction is to explore the performance of the models for more complex, high-dimensional sources such as images.

REFERENCES

[1] A. Wyner and J. Ziv, "The rate–distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1 – 10, 1976.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[3] E. Özyılkan, J. Ballé, and E. Erkip, "Learned Wyner–Ziv compressors recover binning," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 701–706.

[4] E. Ozyilkan, J. Ballé, and E. Erkip, "Neural distributed compressor discovers binning," *IEEE Journal on Selected Areas in Information Theory*, pp. 1–1, 2024.

[5] N. Mital, E. Özyılkan, A. Garjani, and D. Gündüz, "Neural distributed image compression using common information," in *2022 Data Compression Conference (DCC)*, 2022, pp. 182–191.

[6] ——, "Neural distributed image compression with cross-attention feature alignment," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2497–2506.

[7] P. han Li, S. K. Ankireddy, R. Zhao, H. N. Mahjoub, E. M. Pari, ufuk topcu, S. P. Chinchali, and H. Kim, "Task-aware distributed source coding under dynamic bandwidth," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=1A4ZqTmnye

[8] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Transactions on Information Theory*, vol. 31, no. 6, pp. 727–734, 1985.

[9] A. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2031–2034, 1994.

[10] P. Ishwar, R. Puri, S. Pradhan, and K. Ramchandran, "On compression for robust estimation in sensor networks," in *IEEE International Symposium on Information Theory, 2003. Proceedings.*, 2003, pp. 193–.

[11] E. Ozyilkan, J. Ballé, and E. Erkip, "Neural distributed compressor does binning," in *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023.

[12] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[13] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, p. 359–366, jul 1989.

[14] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Networks*, vol. 6, no. 6, pp. 861–867, Jan. 1993. [Online]. Available: https://doi.org/10.1016/s0893-6080(05)80131-5

[15] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations (ICLR)*, 2017.

[16] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018.

[17] C. T. K. Ng, C. Tian, A. J. Goldsmith, and S. Shamai, "Minimum expected distortion in gaussian source coding with fading side information," *IEEE Transactions on Information Theory*, vol. 58, no. 9, pp. 5725–5739, 2012.

[18] E. Tasci, E. Ozyilkan, O. K. Ulger, and E. Erkip, "Robust distributed compression with learned Heegard-Berger scheme," 2024.

[19] S. Kullback, *Information Theory and Statistics*, ser. A Wiley Publication in Mathematical Statistics. Dover Publications, 1997.

[20] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[21] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2021.

[22] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=S1jE5L5gl

[23] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. USA: Cambridge University Press, 2012.