

Neural Compress-and-Forward for the Relay Channel

Ezgi Özyılkan*, Fabrizio Carpi*, Siddharth Garg, Elza Erkip

Department of Electrical and Computer Engineering, New York University, Brooklyn, NY
 {ezgi.ozyilkan, fabrizio.carpi, siddharth.garg, elza}@nyu.edu

Abstract—The relay channel, consisting of a source-destination pair and a relay, is a fundamental component of cooperative communications. While the capacity of a general relay channel remains unknown, various relaying strategies, including compress-and-forward (CF), have been proposed. For CF, given the correlated signals at the relay and destination, *distributed compression* techniques, such as Wyner–Ziv coding, can be harnessed to utilize the relay-to-destination link more efficiently. In light of the recent advancements in neural network-based distributed compression, we revisit the relay channel problem, where we integrate a learned one-shot Wyner–Ziv compressor into a *primitive relay channel* with a finite-capacity and orthogonal (or out-of-band) relay-to-destination link. The resulting neural CF scheme demonstrates that our task-oriented compressor recovers *binning* of the quantized indices at the relay, mimicking the optimal asymptotic CF strategy, although no structure exploiting the knowledge of source statistics was imposed into the design. We show that the proposed neural CF scheme, employing finite order modulation, operates closely to the capacity of a primitive relay channel that assumes a Gaussian codebook. Our learned compressor provides the first proof-of-concept work toward a practical neural CF relaying scheme.

Index Terms—relay channel, Wyner–Ziv source coding, decoder-only side information, task-aware compression, binning.

I. INTRODUCTION

The relay channel, as introduced by van der Meulen [2], is a building block of multi-user communications. In this model, a relay facilitates communication between a source and a destination by forwarding its “overheard” received signal to the destination. As such, the relay channel comprises a *broadcast channel*, from the source to both the relay and the destination, and also a *multiple access channel*, from both the source and the relay to the destination. The relay channel also forms the foundation of cooperative networking, which has been shown to be effective in mitigating fading [3], [4], increasing data rates [5], and managing interference [6]. With the advent of 6G, new forms of relaying and cooperation are envisioned for communicating in highly dynamic settings [7].

Despite decades of research, the capacity of the general relay channel is still unknown to this day. Cover and El Gamal [8] provided upper and lower bounds for the general relay channel by invoking information theoretic achievability and converse arguments. These bounds coincide only in a few special cases, such as the physically degraded Gaussian relay channel. Even though optimum relaying strategies are not known in general, various effective relaying techniques have been proposed, which can be broadly categorized into two main classes: *decode-and-forward* (DF) and *compress-and-forward* (CF); see [8] for a detailed analysis of DF,

CF, their variations and combinations. While DF is known to be efficient in certain scenarios [5], its achievable rate is bounded by the capacity of the source-to-relay channel since the relay is required to perfectly decode the source information. On the other hand, in CF, the relay refrains from directly decoding the source and instead, compresses its received signal to send to the destination. Upon reception of the compression index, the destination combines it with its own received signal to decode the source information. Given that the received signals at the relay and destination are correlated, the relay can leverage *distributed compression* techniques to reduce the compression rate without requiring explicit knowledge of the received signal at the destination. As such, it can utilize Wyner–Ziv (WZ) source coding [9], also known as source coding with decoder-only side information, to efficiently describe its received signal. Unlike DF, CF relaying consistently outperforms direct transmission since the relay always aids in communication, even when the source-to-relay channel is subpar. For additional discussion on scenarios where CF has been proven to be optimal, we direct readers to [10]. Despite its benefits, the limitations of WZ implementations operating in the finite blocklength regime have hampered the widespread use of CF.

In this paper, drawing on recent advances in neural distributed compression [11], [12], we revisit practical CF relaying, and illustrate the potentials of learning for reaping the benefits of CF. We focus on the *primitive relay channel* (PRC) [13], depicted in Fig. 1, where there is an orthogonal (out-of-band) noiseless link of rate R connecting the relay to the destination. Our motivation for considering the PRC is two-fold. First, the PRC provides the simplest setting in which the compressed relay signal can be readily transmitted to the destination. In addition, it is known that CF is optimal for the PRC if the relay is unaware of the source codebook, also known as *oblivious* relaying [14]. The oblivious setting is well-suited to the learning framework, in which the relay is not explicitly informed about the transmission strategy used by the source; rather it trains its compressor based on samples of its channel output. Focusing on a fixed modulation scheme, we train the system consisting of the compressor at the relay and the demodulator at the destination in an end-to-end fashion to optimize the trade-off between the out-of-band rate R and the overall source-to-destination achievable rate.

There is limited literature addressing practical CF designs, e.g., [15], [16]. Both of these works proposed entropy-constrained scalar quantizer designs with BPSK modulation for the half-duplex Gaussian relay channel, with [15] considering Slepian–Wolf coded nested quantization as a practical form of WZ compression, and [16] not taking

* denotes equal contribution. This work was supported in part by the NYU WIRELESS Industrial Affiliates Program, and by the NSF grants #1925079 and #2003182. A longer version of this work is available [1].

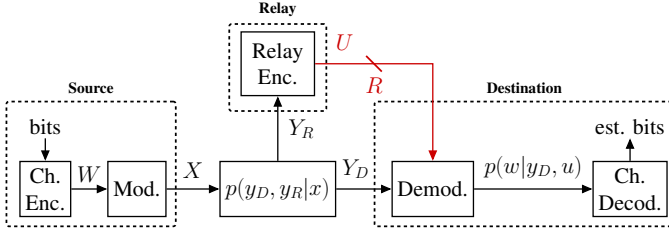


Fig. 1: The *primitive* relay channel (PRC) under consideration. The red link denotes out-of-band relaying between relay and destination.

into account the side information at the destination while quantizing at the relay. In addition, these works relied on handcrafted and analytical solutions, thereby constraining their applicability to more complex communication settings. Recent learning approaches for the relay channel [17], [18] considered a joint source-channel setting, where the former focused on image transmission via joint source-channel coding, while the latter one targeted text communications utilizing attention-based transformer architectures. Our paper, in contrast, concentrates only on the channel part and addresses an important open problem in the cooperative communications literature, namely how to make CF practical.

II. SYSTEM MODEL

A. Primitive Relay Channel

We consider the PRC setup [13], illustrated in Fig. 1. The Gaussian PRC, which we study in this paper, is given by:

$$Y_R = X + N_R, \quad Y_D = X + N_D, \quad (1)$$

where X denotes the signal transmitted by the source, Y_R and Y_D denote the received signals at the relay and the destination, respectively. The corresponding noise components, $N_R \sim \mathcal{N}(0, \sigma_R^2)$ and $N_D \sim \mathcal{N}(0, \sigma_D^2)$, are independent. Note that by allowing for arbitrary (σ_R^2, σ_D^2) , we can incorporate the effect of different channel gains for the source-to-relay and source-to-destination links. As customary, we consider communication over a blocklength of n , with n asymptotically large, and i.i.d. noise. For brevity, we omit the time index in (1). The out-of-band relay-to-destination channel is represented by a link of capacity R bits/channel use.

For a general PRC $p(y_D, y_R|x)$ with an oblivious relay, where the relay is unaware of the codebook shared by source and destination, it has been shown that the capacity can be attained by the CF strategy with time sharing [14]. Without time-sharing, the following rate C is achievable [14]:

$$C = \max I(X; Y_D, U), \quad (2)$$

$$\text{s.t. } R \geq I(Y_R; U | Y_D), \quad (3)$$

where maximization is with respect to the distribution $p(x)p(u|y_R)$. Here, U corresponds to the relay's compressed description of Y_R , and the rate constraint in (3) coincides with the one that emerges in WZ rate-distortion function [9]. Recall that in CF, the relay regards its received signal Y_R as an unstructured random process jointly distributed with the signal received at the destination Y_D . This enables the relay to exploit WZ compression [9], to efficiently describe its received signal. We note that the capacity of the PRC without oblivious

relaying constraint is still not fully characterized [14].

For the Gaussian PRC in (1) with $\sigma^2 = \sigma_R^2 = \sigma_D^2$, the following CF rate is achieved with Gaussian input under power constraint P [14]:

$$C_{CF} = \frac{1}{2} \log_2 \left(1 + \gamma + \frac{\gamma}{1 + \frac{1+2\gamma}{(2^{2R}-1)(\gamma+1)}} \right), \quad (4)$$

where $\gamma = P/\sigma^2$. It is shown in [14] that while the Gaussian input is not necessarily optimal, the rate in (4) is at most 1/2 bit away from the capacity of the Gaussian PRC, even if the relay is not oblivious. Hence, we will use (4) as a benchmark for our learned CF communication rates.

B. Performance Criterion

For our learning-based CF framework, we assume a finite order modulation such that an index $W \in \{1, \dots, |\mathcal{X}|\}$, which represents the output of the channel encoder, is mapped to a symbol $X \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}$ is a constellation of cardinality $|\mathcal{X}|$. We consider a fixed modulation scheme with equally likely symbols, and do not optimize over the constellation $\mathcal{X}$ nor the distribution $p(x)$. Incorporating the learned probabilistic and geometric constellation shaping [19] into our neural CF framework is beyond the scope of this work. Our goal is to jointly learn the *relay encoder*, which outputs a compressed description U , and the (soft) *demodulator* at the destination, which outputs a probability distribution on W (Fig. 1) that maximize the mutual information $I(X; Y_D, U)$ subject to the rate constraint R , as in (2) and (3). We assume the availability of good channel codes to be used in conjunction with the modulation scheme, and as such the mutual information $I(X; Y_D, U)$ can be viewed as a CF achievable rate. In Section III-B, we will discuss how this performance criterion is incorporated into the loss function used in the learning.

III. NEURAL COMPRESS-AND-FORWARD (CF) SCHEMES

In this section, we propose three neural CF schemes to be employed in the PRC shown in Fig. 1. As discussed in Section II-B, the modulation scheme is fixed. On the other hand, the relay's encoder, employing CF strategy, and the demodulator will be parametrized via artificial neural networks (ANNs) that will be jointly optimized in an end-to-end fashion.

A. Neural CF Architectures

Building onto neural distributed compressors proposed in [11], we consider learning-based CF schemes at the relay that include neural one-shot WZ compressors (with side information Y_D at the destination), paired with either a classic entropy coder (EC) or a Slepian-Wolf (SW) coder. We will name these two variants as *marginal* (marg.) and *conditional* (cond.) formulations, respectively. As a benchmark, we also consider a neural one-shot *point-to-point* (p2p) compressor coupled with a classic EC. All of these learned compressors are combined with a neural demodulator available at the destination, which has access to the side information Y_D .

The overall proposed neural CF architectures are illustrated in Fig. 2. The relay's encoder ANN is denoted by $e_\theta(\cdot)$, where θ represents its parameters; the probability distribution

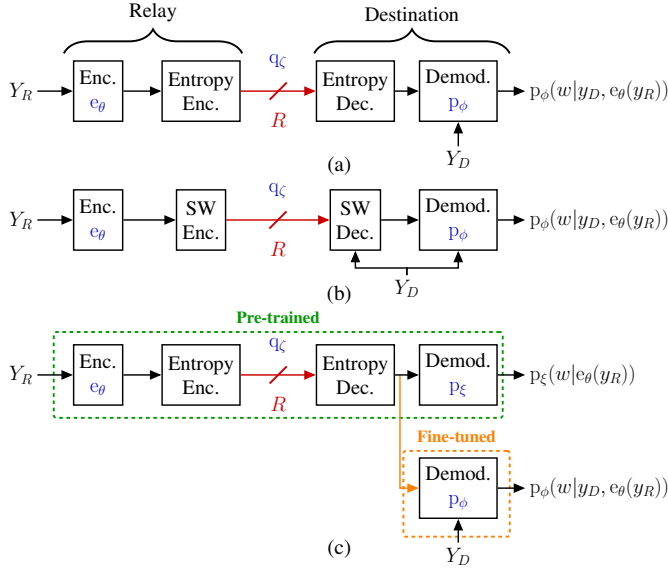


Fig. 2: The three proposed detection-oriented neural compress-and-forward relay schemes: (a) and (b) are based on marginal and conditional formulations respectively; (c) is the point-to-point scheme. The learned parameters are indicated in blue. Note that the schemes in (a) and (b) operationally correspond to task-aware neural Wyner-Ziv compressors, since the encoder can exploit the side information Y_D at the receiver side.

of the relay's encoder output (which is then used by the EC or SW coder) is modeled with q_ζ , parameterized by ζ ; the demodulator's ANN is $p_\phi(w|y_D, e_\theta(y_R))$, where ϕ denotes its parameters. The mapping defined by the demodulator represents the posterior probability on the alphabet $\{1, \dots, |\mathcal{X}|\}$ (soft decision), which serves as an approximation of the true posterior distribution $p(w|y_D, e_\theta(y_R))$. In the learning process of a p2p compressor, as shown in Fig. 2c, we initially train a demodulator $p_\xi(w|e_\theta(y_R))$ to prevent this neural compressor from utilizing the side information Y_D during training. The pre-trained p2p neural compressor as such (highlighted in green) is then used as input for fine-tuning the demodulator $p_\phi(w|y_D, e_\theta(y_R))$, which incorporates side information (highlighted in orange). Note that, in this p2p scenario, e_θ is not able to compress with side information Y_D and therefore, cannot exhibit *binning* (grouping) in the source space.

Note that we use a deterministic encoder in our proposed schemes. Specifically, we set the encoder as $U \triangleq e_\theta(Y_R)$ as in Fig. 1, where $e_\theta(Y_R)$ is discrete. Similar to [11], the probabilistic models, $e_\theta(Y_R)$ and q_ζ , are defined, without loss of generality, as discrete distributions with probabilities $P_k = \frac{\exp \alpha_k}{\sum_{i=1}^K \exp \alpha_i}$ for $k \in \{1, \dots, K\}$, where K is a model parameter. The unnormalized log-probabilities (*logits*) α_i are either directly treated as learnable parameters or computed by ANNs as functions of the conditioning variable. The lossless compression rates induced by the models q_ζ are attainable with high-order classic EC or SW coder, operating on discrete values [20]. As in the popular class of neural compressors [21], we use stochastic gradient descent to optimize all learnable parameters jointly, which relies on

Monte Carlo approximation for the expectations in the loss function. Following the approach in [11], we employ the widely recognized Gumbel-max technique [22] to generate samples from discrete distributions. Additionally, we utilize Concrete distributions [23] to aid in stochastic optimization. During training, to align with the distribution of samples from e_θ , we also opt for Concrete distribution for the models q_ζ .

B. Loss Function

In contrast to prior works on neural distributed compression [11], which focus on minimizing the *distortion* in the reconstruction of the input source in tandem with variable rate entropy coding, our goal in this work is to optimize the operational trade-off between relay-to-destination compression rate and source-to-destination communication rate in the PRC setup.

For our objective function, building onto the compression rate in (3), we first consider the following upper bound:

$$I(Y_R; U | Y_D) \leq H(U | Y_D), \quad (5)$$

$$\leq \mathbb{E}[-\log_2 q_\zeta(e_\theta(y_R))] \triangleq \tilde{R}, \quad (6)$$

where $\tilde{R}$ represents an operational upper bound on the relay's *compression* rate, which is upper bounded by R as in (3). The inequality in (6) is due to the fact that the cross-entropy is larger or equal to entropy [24, Theorem 5.4.3]. Here, $\tilde{R}$ encapsulates the compression rate of a relay quantizer having a one-shot encoder coupled with high-order entropy coder over large blocks of the quantized source.

Similarly, we also establish a lower bound based on the communication rate term in (2) as follows:

$$I(X; Y_D, U) = H(W) - H(W | Y_D, U), \quad (7)$$

$$\geq \log(|\mathcal{X}|) - \tilde{D}, \quad (8)$$

where $\tilde{D} \triangleq \mathbb{E}[-\log(p_\phi(x|y_D, e_\theta(y_R)))]$ and (8) is a lower bound on the source-to-destination *communication* rate C from (2). Here, (7) follows from X being a one-to-one deterministic function of W , and (8) is again due to cross-entropy being larger or equal to entropy. Since we have a fixed modulation scheme and do not perform any probabilistic shaping, in (8) we have $H(W) = H(X) = \log(|\mathcal{X}|)$.

For a demodulator taking hard decisions as

$$\hat{W} = \arg \max_{w \in \{1, \dots, |\mathcal{X}|\}} p_\phi(w|y_D, e_\theta(y_R)), \quad (9)$$

the corresponding symbol error rate (SER) would be $\text{SER} = P(W \neq \hat{W})$. Since minimizing the cross-entropy $\tilde{D}$ is known to be a surrogate for maximizing the accuracy of classification (i.e., symbol detection) [25], minimizing $\tilde{D}$ also operationally corresponds to minimizing SER.

Building onto the above bounds, the training objective of all the proposed neural CF relaying schemes depicted in Fig. 2 can be described by the following loss function:

$$L(\theta, \phi, \zeta) = \tilde{R} + \lambda \cdot \tilde{D}, \quad (10)$$

where $\tilde{R}$ and $\tilde{D}$ are from (6) and (8) respectively, and $\lambda > 0$ controls the trade-off. The optimized e_θ , q_ζ and p_ϕ models, parameterized by θ , ζ and ϕ , yield the ANN-based encoder,

EC or SW coder, and demodulator component, respectively.

Note that in spite of considering specific modulation schemes in training, we do not assume *a priori* knowledge of modulation symbols by the relay in our neural CF schemes. The parameters $\{\theta, \phi, \zeta\}$ are learned solely in a data-driven fashion from samples, through the proposed loss function in (10). Further improvement in the performance may be obtained by also optimizing the probabilistic ($p(x)$ in optimization (2)-(3)) and geometric shaping (constellation $\mathcal{X}$) of the modulation [19]. Consistent with findings in [15], [16], we empirically confirmed that minimizing mean squared distortion at the quantizers may not always maximize the source-to-destination communication rate.

IV. RESULTS AND DISCUSSION

While our framework can be adapted to different modulation schemes and PRC setups, we adopt the following system configuration to showcase numerical results. We consider BPSK and 4-PAM modulations, having constellations $\mathcal{X} = \{\pm 1\}$ and $\mathcal{X} = \{\pm 1, \pm 3\}$, respectively, and equally likely symbols, $p(x) = 1/|\mathcal{X}|$. We assume that the noise variances on both the direct and the relay paths are equal, i.e., $\sigma_R^2 = \sigma_D^2 = \sigma^2$. The signal-to-noise ratio (SNR) is defined as $\gamma = P/\sigma^2$, where $P = \mathbb{E}[|X|^2]$.

For the parametrization of e_θ and p_ϕ , we use ANNs of three dense layers, with 100 units each, except the last one, and leaky rectified linear unit as the activation function. All neural CF schemes are trained for 500 epochs with randomly initialized network weights. Initially, we set the model parameter $K = 32$. The output dimension of p_ϕ is set to be $|\mathcal{X}|$, since this probabilistic model represents the posterior over the transmitted constellation.

We evaluate our learned CF relaying schemes in terms of the trade-off between the relay rate R (using the proxy $\tilde{R}$ in (6)), and two metrics: (i) the communication rate $I(X; Y_D, U)$, for which we use the lower bound (hence, a pessimistic estimate) in (8), and (ii) the SER = $P(W \neq \hat{W})$ (see (9)). All empirical estimates of compression rates, communication rates, and SER are obtained by averaging over at least 10^6 realizations.

A. Baselines

The regimes where $R = 0$ and $R \rightarrow \infty$ are referred to as *without relay* and *perfect relay* scenario, respectively. In particular, in the perfect relay regime, the demodulator has full access to Y_R , and it optimally combines (Y_D, Y_R) . This corresponds to an increased SNR, γ , with respect to the point-to-point scenario. When the variances $\sigma^2 = \sigma_D^2 = \sigma_R^2$ are equal, the perfect relay setting has double the SNR in comparison with the one without relay. In these two regimes, mutual information and SER can be numerically computed for BPSK and 4-PAM modulations as a function of γ .

When $0 < R < \infty$, we consider C_{CF} from (4) as a benchmark on the communication rate for our learned CF schemes with discrete modulations. Increasing the modulation order, $|\mathcal{X}|$, gives more degrees of freedom for the end-to-end learned communication system to approach the rate of a PRC that assumes Gaussian inputs, that is C_{CF} in (4).

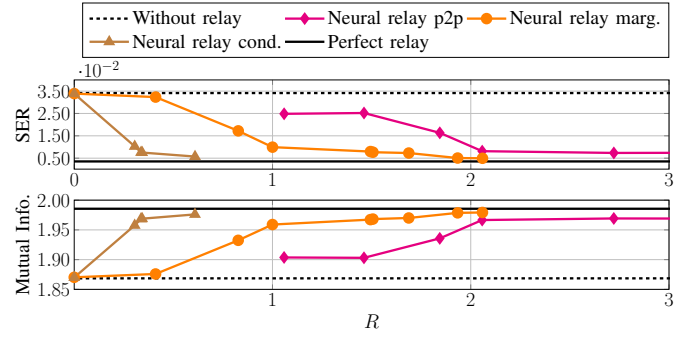


Fig. 3: SER and mutual information results as a function of the relay-to-destination rate R , for the 4-PAM modulation with $\gamma = 13$ dB. The colored lines represent the performance of three neural CF relay architectures (Fig. 2), where each marker corresponds to a unique model trained for a particular value of λ in (10). The horizontal lines provide baseline results without relaying ($R = 0$) and with perfect relaying ($R \rightarrow \infty$).

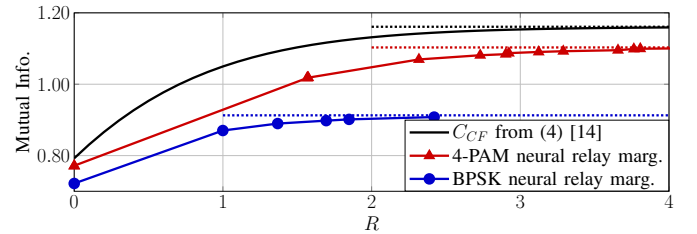


Fig. 4: Mutual information results for the marginal model (Fig. 2a) in case of BPSK and 4-PAM modulations with $\gamma = 3$ dB. The solid line represents C_{CF} in (4) [14], obtained for Gaussian inputs. The dotted lines represent the perfect relay ($R \rightarrow \infty$) bounds for the respective curves, similar to Fig. 3.

B. Numerical Results for the Learned CF Schemes

Fig. 3 shows the SER and mutual information for the 4-PAM modulation for $\gamma = 13$ dB. In this case, Y_R and Y_D , are highly correlated. We observe that the three models exhibit different trade-offs. The conditional model yields the best performance as the side information is also exploited within the SW coder, which operationally executes binning over long sequences i.e., in a multi-shot fashion. The marginal model surpasses the p2p model mainly due to the learned one-shot binning behavior in the source space (see Fig. 5), yielding rate reduction.

Fig. 4 compares C_{CF} from (4) with the mutual information obtained with the marginal formulation for the BPSK and 4-PAM modulations. Here, the SNR for all the considered schemes is $\gamma = 3$ dB, suggesting a lower correlation between Y_R and Y_D compared to the one illustrated in Fig. 3. As expected, increasing the modulation order reduces the gap with the bound in (4), at higher rates meeting the performance of the CF relaying strategy that assumes Gaussian inputs.

Fig. 5 illustrates the learned CF strategy and hard demodulation decision regions (see (9)) for the 4-PAM with $\gamma = 13$ dB. The vertical axis and horizontal axis show Y_R and Y_D , respectively. The colors represent the transmitted indices $e_\theta(Y_R)$ by the relay, and the horizontal lines are the corresponding quantization boundaries. Note that the neural CF architecture exhibits binning (grouping) since non-adjacent intervals are assigned to the same index (same color). The

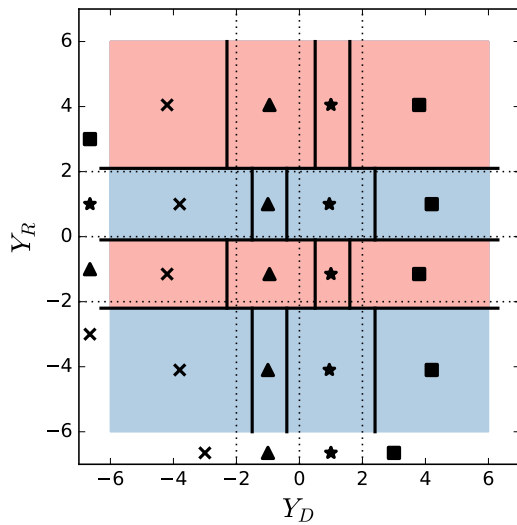


Fig. 5: Visualization (best viewed in color) of the learned CF strategy (marginal scheme in Fig. 2a) and demodulation decisions for the 4-PAM modulation with $\gamma = 13$ and relay rate $R \approx 1$. The horizontal lines denote the quantization boundaries on Y_R , and the colors designate the transmitted index $e_{\theta}(Y_R)$. The vertical lines denote the hard decision boundaries for the demodulator, and the markers represent the decisions. The transmitted symbols (denoted by cross, triangle, star, square) are also reported near the axis for reference.

vertical lines denote the hard decision boundaries, where the markers denote the decisions $\hat{W}$. We observe that the lines are shifted with respect to the midpoints between transmitted symbols (optimal boundaries without relaying). This highlights the interpretability of our neural CF relaying scheme. For example, when *cross* or *star* are transmitted, the index *blue* will be the (most likely) relayed index. In this case, the decision regions for *cross* and *star* at the destination are larger than the others symbols.

Moreover, Fig. 5 can be used as a look-up table for a direct deployment of the resulting CF relaying strategy. Although ANN-based architectures (Fig. 2) were used to minimize the loss function in (10), the actual hard demodulator implementation at test time relies only on the learned threshold values shown in Fig. 5.

V. CONCLUSION

We revisit CF relaying in the context of learned distributed compression and incorporate a task-oriented neural WZ compressor into a PRC setup as a practical form of CF relaying mechanism. Our proposed framework represents the first proof-of-concept work for an interpretable learned CF relaying scheme, where both the compressor and demodulator components are parameterized with lightweight ANNs. Such a design choice also enables us to provide post-hoc interpretations of these learned components by explicitly visualizing their behaviors (see Fig. 5). This reveals that the learned CF scheme exhibits characteristics of optimal asymptotic relaying strategy, such as binning of the quantized indices at the relay, while its performance is close to the one of a PRC model that assumes continuous Gaussian inputs.

Extending our framework to a general relay channel,

in which the destination does successive decoding of the compressed relay index and the source information, would be possible. Additional design constraints arising from incorporating learned CF in full-duplex and half-duplex relay channels, as well as more complex channel models would be interesting future research directions.

REFERENCES

- [1] E. Ozyilkan, F. Carpi, S. Garg, and E. Erkip, "Learning-based compress-and-forward schemes for the relay channel," *arxiv preprint 2405.09534*, 2024.
- [2] E. C. van der Meulen, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, no. 1, p. 120–154, 1971.
- [3] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part I: System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, 2003.
- [4] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.
- [5] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.
- [6] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, 2010.
- [7] M. Najla, Z. Becvar, P. Mach, and D. Gesbert, "Integrating UAVs as transparent relays into mobile networks: A deep learning approach," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–6.
- [8] T. Cover and A. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 572–584, 1979.
- [9] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1 – 10, 1976.
- [10] W. Kang and S. Ulukus, "Capacity of a class of diamond channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4955–4960, 2011.
- [11] E. Özyilkan, J. Ballé, and E. Erkip, "Learned Wyner–Ziv compressors recover binning," in *2023 IEEE ISIT*, 2023, pp. 701–706.
- [12] E. Ozyilkan and E. Erkip, "Distributed compression in the era of machine learning: A review of recent advances," to appear in *58th Annual CISS*, 2024.
- [13] Y.-H. Kim, "Coding techniques for primitive relay channels," in *Forty-Fifth Annual Allerton Conference*, 2007.
- [14] O. Simeone, E. Erkip, and S. Shamai, "On codebook information for interference relay channels with out-of-band relaying," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2880–2888, 2011.
- [15] M. Uppal, Z. Liu, V. Stankovic, and Z. Xiong, "Compress-forward coding with BPSK modulation for the half-duplex Gaussian relay channel," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4467–4481, 2009.
- [16] A. Chakrabarti, A. Sabharwal, and B. Aazhang, "Practical quantizer design for half-duplex estimate-and-forward relaying," *IEEE Trans. Commun.*, vol. 59, no. 1, pp. 74–83, 2011.
- [17] C. Bian, Y. Shao, H. Wu, and D. Gunduz, "Deep joint source-channel coding over cooperative relay networks," to appear in *2024 IEEE ICMLCN*, 2024.
- [18] E. Arda, E. Kutay, and A. Yener, "Semantic forwarding for next generation relay networks," to appear in *58th Annual CISS*, 2024.
- [19] M. Stark, F. Ait Aoudia, and J. Hoydis, "Joint learning of geometric and probabilistic constellation shaping," in *2019 IEEE Globecom Workshops*.
- [20] J. Rissanen and G. Langdon, "Universal modeling and coding," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 12–23, 1981.
- [21] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE J. Sel. Topics in Signal Process.*, vol. 15, no. 2, pp. 339–353, 2021.
- [22] E. J. Gumbel, "Statistical theory of extreme values and some practical applications : A series of lectures," 1954.
- [23] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR*, 2017.
- [24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2006.
- [25] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2006.