Fundamental Limits to Exploiting Side Information for CSI Feedback in Wireless Systems

Heasung Kim, *Member, IEEE*, Gustavo de Veciana, *Fellow, IEEE*, and Hyeji Kim, *Senior Member, IEEE*

***Abstract*—In modern wireless systems, the feedback of DownLink (DL) Channel State Information (CSI) from User Equipment (UE) to Base Stations (BS) may require substantial computational and feedback bandwidth overheads. A promising approach to improve feedback efficiency is to leverage side information which is correlated to DL CSI. Despite potential of doing so, critical aspects remain underexplored in current research, particularly the quantification of the benefits and the inherent limitations of utilizing side information. This paper addresses these gaps by introducing a novel algorithm to compute the rate-distortion function for general compression scenarios incorporating side information. We apply this algorithm to the DL CSI feedback problem having UL CSI as the side information and generate rate-distortion functions. Using the estimated rate-distortion functions, we measure the gain of side information over diverse feedback rates and UE mobility profiles. The results reveal that the benefits of leveraging side information are particularly significant for UEs characterized by high mobility and constrained to operate at low feedback overheads.**

***Index Terms*—Channel state information, compression, coding, feedback, rate-distortion, side information, MIMO, FDD.**

I. Introduction

EFFICIENT Downlink (DL) Channel State Information (CSI) feedback from User Equipment (UE) to Base Station (BS) in wireless systems has emerged as a critical problem, given that CSI is essential to enabling wireless networks to effectively minimize interference and maximize data throughput. The primary challenge lies in the potentially high feedback and computational overheads required, especially for Multiple Input Multiple Output (MIMO) communications. To enhance feedback efficiency, such systems can exploit side information correlated with the DL CSI. This approach can be viewed as a problem of *compressing with side information*, where the UE compresses the DL CSI and transmits it to the BS. The BS then uses such side information to decode the transmitted data, which aids in designing DL precoders or directly recovering the DL CSI.

Despite extensive research into this problem, several fundamental aspects of compression with side information remain underexplored. Key questions include: *How can the value of side information be quantified?* and *Under what conditions is side information more or less valuable?* The answers to these questions are crucial to drive real-world system design decisions, in particular in deciding whether to incorporate additional features (side information) to enhance performance, especially when constrained by computation resources.

The rate-distortion function can be used to address these questions by characterizing the minimum achievable distortion for a given rate, enabling one to quantify the value of side information and observing the trends in the associated gains. In the context of CSI feedback, *distortion* refers to the discrepancy between the desired and estimated outputs of the decoder, while *rate* pertains to the feedback rate. However, deriving closed-form expressions for rate-distortion functions proves challenging, and computational complexity persists even with iterative methods, particularly when the distributions of the input and side information are unknown, and the domain extends over continuous or large discrete sets. Furthermore, exploring rate-distortion functions within a generalized compression framework with side information, aimed at computing a desired output, remains an underexplored area despite its increasing relevance in contemporary research.

To bridge these gaps, we introduce a novel algorithm to estimate the rate-distortion function which is applicable to generalized compression tasks taking side information into account. Using this approach, we generate rate-distortion curves for CSI feedback problems, exploit these findings to quantify the value of side information. The estimated rate-distortion functions provide not only theoretical benchmarks for assessing the performance of practical compression algorithms but also suggest potential system design principles. Specifically, our contributions can be outlined as follows.

A. Contributions

Algorithm design. First, we present a generalized framework for estimating the rate-distortion function for computing desired output with side information. We formulate a Lagrangian loss function where the minimization of this function is achieved through specific encoding and decoding schemes that can achieve point(s) on the rate-distortion function. Our algorithm focuses on minimizing the loss by parameterizing the conditional distributions of the codewords for a given source and side information, as well as the decoder. The algorithm is designed to alternatively update

Received 15 May 2024; revised 15 December 2024; accepted 14 January 2025. Date of publication 9 April 2025; date of current version 19 June 2025. This work was supported in part by the National Science Foundation (NSF) under Grant 2148224 and Grant CNS-2008824; in part by the Office of the Under Secretary of Defense for Research and Engineering (OUSD R&E), National Institute of Standards and Technology (NIST), and Industry Partners as Specified in the Resilient and Intelligent NextG Systems (RINGS) Program; in part by the Office of Naval Research (ONR) under Award N000142412542; in part by InterDigital, Inc., through 6G@UT Center within the Wireless Networking and Communications Group (WNCG) at The University of Texas at Austin; and in part by the Army Research Office (ARO) under Grant W911NF2310062. An earlier version of this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT) 2024 [DOI: 10.1109/ISIT57864.2024.10619642]. *(Corresponding author: Heasung Kim.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: heasung.kim@utexas.edu; deveciana@utexas.edu; hyeji.kim@austin.utexas.edu).

Digital Object Identifier 10.1109/JSAC.2025.3559165

these parameters to efficiently minimize the Lagrangian loss. Notably, our approach offers accurate rate-distortion estimates for challenging high-dimensional or continuous input sources, where conventional Blahut-Arimoto type algorithms do not perform well.

Applications to theoretical and practical problems. We validate our algorithm through both theoretical examples and practical simulations. Initially, we confirm its efficacy by considering scenarios characterized by known closed-form rate-distortion functions, such as where the source and side information have correlated Gaussian distributions. The results show the numerical accuracy of our approach in estimating the rate-distortion function. Furthermore, we extend our methodology to address the DL CSI feedback problem incorporating uplink (UL) side information, thereby providing insights into the potential gains of exploiting side information.

Simulation results. Our further analysis delves into the CSI feedback problems in diverse environmental conditions and considering UE mobility speeds going beyond static DL CSI feedback scenarios prevalent in learning-based CSI compression research. The results reveal that the benefits of side information are particularly significant in low-rate regime or high-mobility scenarios. Moreover, we discuss potential system design principles informed by the results, paving the way for enhanced efficiency and performance in practical applications.

In the remaining sections, we discuss related work in Section II, formalize the system model in Section III, and describe our rate-distortion estimation algorithm in Section IV. We validate our algorithm for the Gaussian source compression in Section V and apply it to the CSI feedback problem in Section VI. We conclude in Section VII, with proofs and detailed simulation configurations provided in Appendices.

II. Related Work

A. CSI Feedback in Wireless Communications

CSI feedback schemes are often inherently structured in terms of a compression or packing problem, driven by the high volume of data transmitted from UE to BS and constraints of limited bandwidth overheads and computational resources.

1) Conventional Implicit CSI Feedback: Modern communication systems frequently employ an implicit CSI feedback approach, wherein the UE communicates channel quality via predefined indicators such as the Rank Indicator (RI), Precoding Matrix Indicator (PMI), and Channel Quality Indicator (CQI) [2]. Notably, the PMI feedback utilizes a codebook-based method [3], [4], with significant research focusing on the optimal design of these codebooks [5]. Solutions have explored frameworks such as Grassmannian packing [6], [7], [8] and Random Vector Quantization [9], [10].

Conventional implicit feedback methods often face limitations in fully exploiting correlations across numerous subcarriers in advanced systems, particularly with the advent of MIMO technologies, which greatly increase the channel information per UE. These challenges highlight the need for more efficient feedback mechanisms to accommodate the growing complexity of modern communication systems.

2) Compressive Sensing and Learning-Based CSI Feedback: To enhance feedback efficiency, recent advancements have introduced the use of compressive sensing [11], [12] and deep learning-based methods [13]. Notably, with significant advancements in learning-based image compression technologies, deep learning techniques have been applied to explicitly compress CSI, aiming to minimize the distortion between the input source, CSI, for the encoder of the UE and the decoder's output at the BS. These approaches predominantly utilize parameterized deep neural networks and update the trainable parameters in the direction of minimizing the distortion. Pioneering work in this area [14] employed convolutional and fully connected layers, achieving superior performance over traditional compression methods. Further innovations have incorporated architectures inspired by the Inception block [15] and attention mechanisms [16], with recent research focusing on quantization and variable compression ratios to enhance practicality [17], [18], [19], [20], [21], [22], [23].

3) Advanced Feedback Frameworks: Other advancements have extended the feedback framework to enhance performance by incorporating distributed encoding techniques [24], [25], where feedback from multiple correlated UEs is collectively processed. Moreover, a multioutput autoencoder framework has been proposed, featuring a global encoder capable of generating codewords adaptable to various channel environments, which are then decoded by multiple decoders [26]. Additionally, to further refine the efficiency of communication systems, frameworks for joint denoising-compression [27] and joint source-channel coding [28] have been developed, demonstrating the potential for integrated processing modules to improve system performance.

4) CSI Feedback With Side Information: Despite the substantial advancements in compression techniques and frameworks, there remains a significant computational and resource burden on the devices. To design more efficient feedback frameworks, recent research has proposed utilizing side information at the decoder. Methods based on recurrent neural networks leverage temporally correlated CSI as side information to recover desired CSI [29], [30], [31]. Furthermore, UL CSI, typically acquired via pilot transmissions from the UE to the BS and correlated with DL CSI due to frequency-invariant characteristics [32], [33], has been utilized. In [34], the magnitude of UL CSI is leveraged at the BS to improve compression performance, while [35] explores the partial reciprocity of DL and UL channels.

Despite these advancements, the theoretical perspective quantifying the performance gains from such side information remains underexplored. Current research largely focuses on empirical improvements using advanced CSI frameworks, yet a comprehensive theoretical understanding of the achievable performance and the exact benefits of incorporating side information is lacking. This gap highlights the need for a more rigorous exploration of how these enhancements translate across different wireless system environments. In response to this gap, our study proposes an estimation of the rate-distortion function specific to the CSI feedback scenarios.

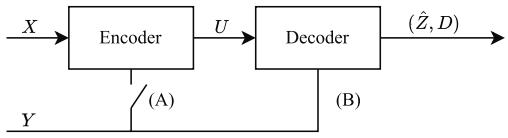

Fig. 1. Coding for Computing with Side Information. We consider a configuration where switch (A) remains open while switch (B) is closed, permitting access to side information at the decoder. The decoder aims to compute a function $Z = g(X, Y)$; we let $\hat{Z}$ and D denote the decoder's output and distortion, respectively.

B. Computing Rate-Distortion Functions

The rate-distortion function, which characterizes the optimal rate-distortion tradeoff, serves as an important theoretical benchmark for assessing the effectiveness of compression algorithms, as highlighted in recent studies [36], [37], [38], [39]. However, the accurate computation of Shannon's information measures, such as entropy and mutual information, which form the basis of the rate-distortion function, is notably challenging. This is particularly true in scenarios where one must rely solely on samples from real-world distributions without prior knowledge of the distributions or in cases involving high-dimensional input sources, which make data distribution approximation significantly more complex. Indeed, closed-form solutions for these measures are generally limited to specific circumstances, e.g., Gaussian sources.

An approach to numerically computing the rate-distortion functions and associated information measures for general distributions has been devised based on iterative algorithms in 1972 by Blahut [40] and Arimoto [41]. Known collectively as the Blahut-Arimoto algorithms, they have been adapted to address multiterminal source coding settings [42]. However, these conventional iterative approaches face limitations, especially when applied to high-dimensional or continuous sources [37].

To overcome these challenges, recent studies have explored solutions utilizing neural networks or advanced optimization techniques. The Restricted Boltzmann Machine is integrated with neural networks to estimate rate-distortion functions [36]. In [37], the rate-distortion function duality concept, e.g., [43], is utilized in estimation methods. A sandwich bound for rate-distortion function is introduced in [38] through distribution parameterization with neural networks. In [44], neural estimation methods with a generative model framework are utilized. Notably, the Wasserstein gradient descent algorithm proposed in [39] has demonstrated state-of-the-art performance for rate-distortion estimation without relying on neural networks.

C. Rate-Distortion Function for Computing With Side Information

The rate-distortion function concept can be extended to encompass scenarios where side information, correlated with the input source, is available at the decoder, or at both the encoder and decoder, as illustrated in Figure 1. This adaptation is referred to as the Wyner-Ziv rate-distortion function [45]. Additionally, the notion of the rate-distortion is further broadened by considering communication systems where the goal is to compute a function of the source. Such applications of the Wyner-Ziv rate-distortion function are commonly referred to as *Coding for Computing* [46].

Such a broader perspective of the rate-distortion function is crucial for evaluating compression algorithms in practical scenarios and understanding side information's role in various contexts. It also offers a way to quantify the relevance of different types of side information for various sources.

While recent studies have increasingly incorporated side information with specific objectives in constructive compression algorithms [47], [48], [49], research on rate-distortion estimation with side information, especially for continuous or large-dimensional distributions, remains limited. Prior studies have focused on discrete sources and side information [50], extending the Blahut-Arimoto Algorithm, but often struggle with high-dimensional distributions, particularly when there is no a priori knowledge of the distributions. Recent contributions in [51] have begun to employ neural networks for estimating the rate-distortion function with side information. However, this work primarily concentrates on discrete codewords and focuses on estimating variational upper bounds of the Wyner-Ziv rate-distortion function.

We address these gaps through a neural network-based direct estimation method for the rate-distortion function for computing with side information, along with applicable methodologies.

III. System Model

Consider the communication system model depicted in Figure 1 where switch (A) remains open and switch (B) is closed. This system model focuses on minimizing the distortion between a target output Z and its estimated counterpart $\hat{Z}$ through the optimized encoder and decoder modules. The encoder receives an input source X and compresses it into a codeword U. The decoder, utilizing this codeword along with side information Y, produces the estimated output $\hat{Z}$. Notably, this framework permits the target output Z to differ from the input source X, allowing Z to be any functional output $g(X, Y)$ tailored to specific system requirements. For example, consider the application in the DL CSI feedback systems from a UE to a BS. Rather than reconstructing the original CSI X, the BS leverages the codeword U to generate precoding vectors, which are essential for efficient DL transmission. In this context, Z would represent the precoding vectors.

The source and side information pair (X, Y) is independently and identically distributed (i.i.d.), following the joint distribution $p_{X,Y}(x, y)$, where x and y are realizations from the respective domains $\mathcal{X}$ and $\mathcal{Y}$. The codeword U resides within the domain $\mathcal{U}$, and the output from the decoder, $\hat{Z}$, belongs to $\mathcal{Z}$. The distortion level is denoted by D, with a distortion measure d defined as $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$. To facilitate clarity in cases where the original information X is to be reconstructed, i.e., if $g(X, Y) = X$, the output of the decoder will be denoted as $\hat{X}$.

In this setting, the corresponding rate-distortion function determines the minimum necessary rate to compute $g(X, Y)$ within a given distortion threshold D. The rate-distortion function, denoted R_{D}, is given as follows [52].

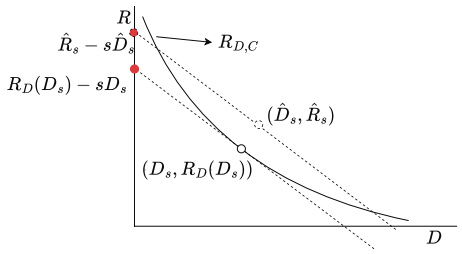

Fig. 2. R_{D} is convex with respect to distortion D. For a given slope s, minimizing the y-intercept of a line originating from an achievable point $(\hat{D}_s, \hat{R}_s)$ in the rate-distortion region leads to a new y-intercept, which corresponds to a line that is tangent to the R_{D} curve at point(s) with the same slope s.

Definition 1: Rate-distortion function for computing with side information

$$R_{\mathrm{D}}(D) = \min_{q_{U|X}(u|x), f(u,y):\ \mathbb{E}[d(Z,\hat{Z})]\le D} I(X;U|Y), \quad (1)$$

where $q_{U|X}(u|x)$ is a conditional probability distribution of U given X. Z and $\hat{Z}$ are the desired function output and the decoder output, respectively. f is a decoder taking u and side information y as an input pair as $f(u, y) = \hat{z}$. Note that the system model implies that the Markov chain $U-X-Y$ holds.

IV. Rate-Distortion Estimation Algorithm

In this section, we focus on developing a method to estimate $R_{\mathrm{D}}(D)$ for a general function, particularly in scenarios where the joint distribution $p_{X,Y}(x, y)$ is unknown and a dataset of N data points $(x_i, y_i)_{i=1}^{B}$ that are sampled from $p_{X,Y}(x, y)$ is available. This setup is typical in real-world contexts, where the exact distribution underlying a dataset is often not known.

We start with a Lagrangian formulation to address the optimization problem defined in (1) by exploiting convexity and non-increasing property of $R_{\mathrm{D}}(D)$ with respect to the distortion D. We can formulate an optimization problem for finding the vertical intercept of the tangent with slope $s(\le 0)$ to the rate-distortion curve as follows.

$$R_{\mathrm{D}}(D_s) - sD_s = \min_{q_{U|X}, f} \{I(X;U|Y) - s\mathbb{E}[d(Z,\hat{Z})]\}. \quad (2)$$

For a given slope s and a corresponding achievable (distortion, rate) pair, $(\hat{D}_s, \hat{R}_s)$ illustrated in Fig. 2, the y-intercept at this line is $\hat{R}_s - s\hat{D}_s$. This intercept is equivalent to $I(X;U|Y) - s\mathbb{E}[d(Z,\hat{Z})]$, attained by the specific encoding and decoding schemes associated with $q_{U|X}$, f corresponding to $(\hat{D}_s, \hat{R}_s)$. This y-intercept can be minimized through optimization, adjusting the encoding and decoding schemes accordingly.

Due to the convexity of R_{D}, the lowest achievable value of the vertical intercept corresponds to $R_{\mathrm{D}}(D_s) - sD_s$ where the distortion D_s and rate $R_{\mathrm{D}}(D_s)$ is a point that lies on the R_{D} curve itself. By determining a point on the R_{D} curve for each slope s and then varying s, we can estimate the R_{D} curve.

To facilitate estimation using a given dataset, we reformulate the optimization term in (2) as detailed in Appendix A:

$$\min_{q_{U|X}, f} \left\{ \mathbb{E}_{X,Y,U}\left[\log \frac{q_{U|X}(U|X)}{q_{U|Y}(U|Y)}\right] - s\mathbb{E}[d(Z,\hat{Z})] \right\}, \quad (3)$$

where $q_{U|Y}(u|y) = \sum_{x\in\mathcal{X}} p_{X|Y}(x|y) q_{U|X}(u|x)$ (when X is a discrete random variable) and $\hat{Z} = f(U, Y)$. This formulation enables the computation of expectation terms using Monte Carlo estimation with data points following the distribution $q_{X,Y,U}(x, y, u)$. Here, we use the notation q to represent a probability distribution influenced by $q_{U|X}$ and f, while p has been used to denote distributions independent of $q_{U|X}$ and f.

To proceed with this approach, we parameterize the key components of the optimization problem using a neural network-based model. First, we represent the conditional distribution $q_{U|X}(u|x)$ as $q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$ where $\boldsymbol{\theta}_{\mathrm{po}}$ denotes a set of parameters for $q_{U|X}$. Similarly, we parameterize the decoding function with a set of parameters $\boldsymbol{\theta}_{\mathrm{dec}}$ as $f(u, y; \boldsymbol{\theta}_{\mathrm{dec}})$.

It should be noted that the parameterization of $q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$ directly determines the related marginal and joint distributions, such as $q_{U|X,Y}$, $q_{U|Y}$, and $q_{X,Y,U}$ under the fixed $p_{X,Y}$. These distributions, governed by the parameter set $\boldsymbol{\theta}_{\mathrm{po}}$, are thus denoted as $q_{U|X,Y;\boldsymbol{\theta}_{\mathrm{po}}}$, $q_{U|Y;\boldsymbol{\theta}_{\mathrm{po}}}$, and $q_{X,Y,U;\boldsymbol{\theta}_{\mathrm{po}}}$.

In addressing the problem (3), however, a major challenge arises from this parametric approach: as we parameterize $q_{U|X}$, the distribution $q_{U|Y}$ deterministically depends on both the parameterized $q_{U|X}$ and the joint distribution of the input source and side information, $p_{X,Y}$, and its computation presents difficulties; we further elaborate on these challenges and our approaches to addressing them in Sec. IV-A. In summary, we begin with the Lagrangian optimization problem for the rate-distortion function, also known as the supporting hyperplane method. We employ neural networks to parameterize the key components of our loss function in (3). This optimization strategy draws parallels with the conventional Blahut-Arimoto algorithms [40], [41] in terms of formulating the Lagrangian loss, while also drawing inspiration from recent works [38], which has achieved state-of-the-art results in rate-distortion estimation through neural networks. In the following subsections, we delve into a comprehensive explanation of our proposed algorithm, detailing the steps and techniques involved.

A. Algorithm

The proposed method is detailed in Algorithm 1. This algorithm iteratively computes the gradient of the loss function (3) over T training iterations and updates the relevant parameters to minimize the loss.

Line 3. Specifically, in each iteration, a minibatch with size b, $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{b}$, is sampled. To estimate the expectation $\mathbb{E}_{X,Y,U}[\log q_{U|X}(U|X) - \log q_{U|Y}(U|Y)]$, it is necessary to generate data point triples (x_i, y_i, u_i) following the distribution $p_{X,Y}(x, y) q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$. For each sampled pair (x_i, y_i), a corresponding u_i is drawn from the distribution $q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$. This sampling results in triples (x_i, y_i, u_i) that adhere to the joint distribution $q_{X,Y,U}(x, y, u) = p_{X,Y}(x, y) q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$.

Utilizing these samples, we compute the average gradient of the loss function, which involves the computation of the expected value of $\log \frac{q_{U|X}(U|X)}{q_{U|Y}(U|Y)}$. This corresponds to

Algorithm 1 Estimation of Rate-Distortion Function for Computing With Side Information at Decoder

1: **Input:** Slope s, dataset $\{x_i, y_i\}_{i=1}^{B}$, initialized sets of parameters $\boldsymbol{\theta}_{\text{po}}$, $\boldsymbol{\theta}_{\text{pr}}$, $\boldsymbol{\theta}_{\text{dec}}$
2: **for** $\tau = 0$ to T **do**
3: Sample minibatch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{b}$ and sample $\{u_i\}_{i=1}^{b}$ from $\{q_{U|X=x_i}\}_{i=1}^{b}$
4: Compute $\nabla L_1 = \nabla \frac{1}{b} \sum_{i=1}^{b} [\log(q_{U|X}(u_i|x_i;\boldsymbol{\theta}_{\text{po}}) - \log q_{U|Y}(u_i|y_i;\boldsymbol{\theta}_{\text{pr}}) - s[d(g(x_i,y_i), f(u_i,y_i;\boldsymbol{\theta}_{\text{dec}}))]$
5: Update $\boldsymbol{\theta}_{\text{po}} \leftarrow \boldsymbol{\theta}_{\text{po}} - \nabla_{\boldsymbol{\theta}_{\text{po}}} L_1$ and $\boldsymbol{\theta}_{\text{dec}} \leftarrow \boldsymbol{\theta}_{\text{dec}} - \nabla_{\boldsymbol{\theta}_{\text{dec}}} L_1$
6: **for** $\tau' = 0$ to T' **do**
7: Sample minibatch $\mathcal{B}' = \{(x_i, y_i)\}_{i=1}^{b}$ and sample $\{u_i\}_{i=1}^{b}$ from $\{q_{U|X=x_i}\}_{i=1}^{b}$
8: Compute $\nabla L_2 = \nabla \frac{1}{b} \sum_{i=1}^{b} [\log(q_{U|X}(u_i|x_i;\boldsymbol{\theta}_{\text{po}}) - \log q_{U|Y}(u_i|y_i;\boldsymbol{\theta}_{\text{pr}})]$
9: Update $\boldsymbol{\theta}_{\text{pr}} \leftarrow \boldsymbol{\theta}_{\text{pr}} - \nabla_{\boldsymbol{\theta}_{\text{pr}}} L_2$

$\log \frac{q_{U|X}(U|X;\boldsymbol{\theta}_{\text{po}})}{q_{U|Y;\boldsymbol{\theta}_{\text{po}}}(U|Y)}$ based on the parameterization where $q_{U|Y;\boldsymbol{\theta}_{\text{po}}}$ is formulated as

$$\begin{aligned} q_{U|Y;\boldsymbol{\theta}_{\text{po}}}(u|y) &= \sum_{x\in\mathcal{X}} p_{X|Y}(x|y) q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|x,y) \\ &= \sum_{x\in\mathcal{X}} p_{X|Y}(x|y) q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}}). \end{aligned} \tag{4}$$

The efficient computation of $q_{U|Y;\boldsymbol{\theta}_{\text{po}}}$ is critical, as it needs to be executed for multiple instances to obtain the average of the log probability. However, this computation of (4) presents a substantial challenge due to the unknown nature of the distribution $p_{X|Y}$, with only sample-based access available. Furthermore, using sampling approaches for the estimation of the sum over $\mathcal{X}$ is non-trivial when domain $\mathcal{X}$ is a high-dimensional space and the data instances are limited. To address this issue, we leverage the following lemma, with its proof detailed in Appendix B.

Lemma 1: Consider a fixed set of parameters $\boldsymbol{\theta}_{po}$ and the scenario where the side information Y is available only at the decoder. Then we have

$$\arg\min_{\hat{q}_{U|Y}} \mathbb{E}_{X,Y,U}\left[\log \frac{q_{U|X}(U|X;\boldsymbol{\theta}_{po})}{\hat{q}_{U|Y}(U|Y)}\right] = q_{U|Y;\boldsymbol{\theta}_{po}}. \tag{5}$$

Based on this lemma, we conclude that instead of executing the summation in (4) to derive $q_{U|Y;\boldsymbol{\theta}_{\text{po}}}$ for a given $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$, we can model the distribution of U given Y as $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$ where $\boldsymbol{\theta}_{\text{pr}}$ denotes a set of free parameters and then use the parameterized distribution $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$ as an argument for the problem (5). The solution of (5) will lead to $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}}) = q_{U|Y;\boldsymbol{\theta}_{\text{po}}}(u|y)$ as long as the parametrization of $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$ is expressive enough.

Lines 4-5. By using the parametrized functions $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$, $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$, and $f(u,y;\boldsymbol{\theta}_{\text{dec}})$, in Line 4, we compute the gradient of (3). Subsequently, in Line 5, the parameters $\boldsymbol{\theta}_{\text{po}}$ and $\boldsymbol{\theta}_{\text{dec}}$ are updated to minimize the loss.

Lines 6-9. At the end of each iteration, we update $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$ by solving (5) based on the newly updated $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$ to correctly compute the main loss function (3) in the subsequent iteration. Problem (5) can be solved through gradient descent updates of the set of parameters $\boldsymbol{\theta}_{\text{pr}}$ as described in Lines 7-9 of Algorithm 1. More specifically, for each inner-iteration (occurring T' times), we sample a minibatch and obtain pairs $\{(x_i, y_i, u_i)\}_{i=1}^{b}$. We then update $\boldsymbol{\theta}_{\text{pr}}$ to minimize the objective in (5). Practically, we have found that setting $T' = 1$ and reusing the same minibatch $\mathcal{B}$ for $\mathcal{B}'$ not only offers computational efficiency but also recovers the optimal result for Gaussian settings (as detailed in Sec. V).

B. Parameterization

In Algorithm 1, we utilize three distinct parameterized models: $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$, $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$, and $f(u,y;\boldsymbol{\theta}_{\text{dec}})$. Various parameterization setups exist, including Gaussian, uniform distribution-based parameterizations, and more sophisticated forms relevant to modern machine learning research [53]. In our study, we opt for Gaussian distributions for parameterization. For example, in sampling from the distribution $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$, the random variable U is assumed to follow a Gaussian distribution characterized by mean $\mu(x;\boldsymbol{\theta}_{\text{po}})$ and variance $\Sigma(x;\boldsymbol{\theta}_{\text{po}})$, both of which depend on the given realization x. The functions μ and Σ can be designed in various ways, depending on the specifics of the problem, where they take x as input and output the corresponding mean and variance. We provide more details on the implementation in Sec. V.

The choice of parameterization and the construction of these functions yield a point that represents an upper bound on the rate-distortion curve. This is because the variable spaces for the minimization problem in (3) are constrained by the assumptions inherent in the chosen distribution models. We can easily show the following:

Proposition 1: Consider any sets of parameters $\{\boldsymbol{\theta}_{\text{po}}, \boldsymbol{\theta}_{\text{pr}}, \boldsymbol{\theta}_{\text{dec}}\}$ and corresponding parameterized distributions and functions $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$, $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$, and $f(u,y;\boldsymbol{\theta}_{\text{dec}})$. Let $D = \mathbb{E}[d(Z, f(U,Y;\boldsymbol{\theta}_{\text{dec}}))]$ denote its distortion. Then we have $R_{\text{D}}(D) \le \mathbb{E}_{X,Y,U}\left[\log \frac{q_{U|X}(U|X;\boldsymbol{\theta}_{\text{po}})}{q_{U|Y}(U|Y;\boldsymbol{\theta}_{\text{pr}})}\right]$.

Proposition 1 shows that the rate-distortion point estimated by our algorithm is on or above the true rate-distortion curve R_{D}. This implies that if a compression algorithm achieves a rate-distortion point above the estimated curve produced by our method, it is assured that such a point does not lie on the true rate-distortion curve, and there is a gap between the theoretical limit and the constructive algorithm. In practical scenarios where the joint distribution (X, Y, Z) is unknown, the expectation in Proposition 1 can be approximated using the dataset $\{x_i, y_i\}_{i=1}^{B}$. Although the estimated mean obtained through the Monte Carlo method may not strictly satisfy the inequality, for sufficiently large B, the algorithm's expected behavior can still be effectively tracked.

V. Rate-Distortion Function Estimation for 2-Component White Gaussian Noise

To assess the effectiveness of our algorithm, we initially focus on scenarios where the true rate-distortion function with side information is known as a closed-form. By comparing this known rate-distortion function to the estimates produced by our method, we can reliably measure the accuracy of our algorithm.

A. Problem Description

We adopt a scenario from [46, Sec. 11.3], featuring a 2-component White Gaussian Noise (2-WGN(P, ρ)) source, where (X, Y) forms pairs of i.i.d. jointly Gaussian random variables. Each pair in the sequence (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ is correlated by $Y = X + W$ where the distributions of X and W have zero mean, $\mathbb{E}[X] = \mathbb{E}[W] = 0$, and variance $\mathbb{E}[X^2] = P$ and $\mathbb{E}[W^2] = N$. With a squared error distortion measure d, the rate-distortion function R_{D} is given as follows [45].

$$R_{\mathrm{D}}(D) = \max\left\{\frac{1}{2}\log\left(\frac{PN}{(P+N)D}\right), 0\right\}. \tag{6}$$

Our objective is to apply our algorithm to estimate rate-distortion points and assess its accuracy in mapping these points on the true rate-distortion curve (6).

B. Configuration

To implement our approach, we employed a multi-layer perceptron (MLP) to model $q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$, $q_{U|Y}(u|y;\boldsymbol{\theta}_{\mathrm{pr}})$, and $f(u, y; \boldsymbol{\theta}_{\mathrm{dec}})$. Specifically, for $q_{U|X}(u|x;\boldsymbol{\theta}_{\mathrm{po}})$ and $q_{U|Y}(u|y;\boldsymbol{\theta}_{\mathrm{pr}})$, we use a single-layer MLP that takes an n-dimensional input and outputs a $2n$-dimensional vector, half for mean and half for variance, to model an n-dimensional independent multivariate Gaussian distribution. For $f(u, y; \boldsymbol{\theta}_{\mathrm{dec}})$, we used a 2-layer MLP with leaky ReLU activation, which takes (u, y) as an input and outputs an n-dimensional $\hat{z}$.

C. Results

In Fig. 3, we set $P = 1, n = 10$, and provide simulation results for various N values in $\{0.5, 1.0, 1.5, 2.0\}$. Each subplot displays the R_{D} curves, alongside four rate-distortion points estimated by our algorithm for different slopes $s \in \{-1, -2, -4, -8\}$. The autoencoder model processes the sequence (X, Y) of length n for compression, and the element-wise distortion-rate values are evaluated for each scenario. These distortion-rate pairs are expected to lie within the single-letter rate-distortion function (6). We also plot $R(D)$ curves, which correspond to the true rate-distortion function without side information. y-axis has natural units (Nats) and x-axis represents mean squared error distortion. The dashed lines associated with each of the estimated points correspond to the learning trajectory, i.e., the achieved (distortion, rate) points during the training process.

In Figure 3, our algorithm demonstrates a consistent ability to estimate points on the rate-distortion curve R_{D}. A decrease in the value of s leads to points on the left side of the curve, indicative of higher rates and lower distortion. Notably, an increase in noise N on the side information results in a higher rate for a given distortion, causing $R_{\mathrm{D}}(D)$ to converge towards $R(D)$. Our method effectively estimates points on R_{D} across various scenarios. In Figure 4, as we increase P while keeping

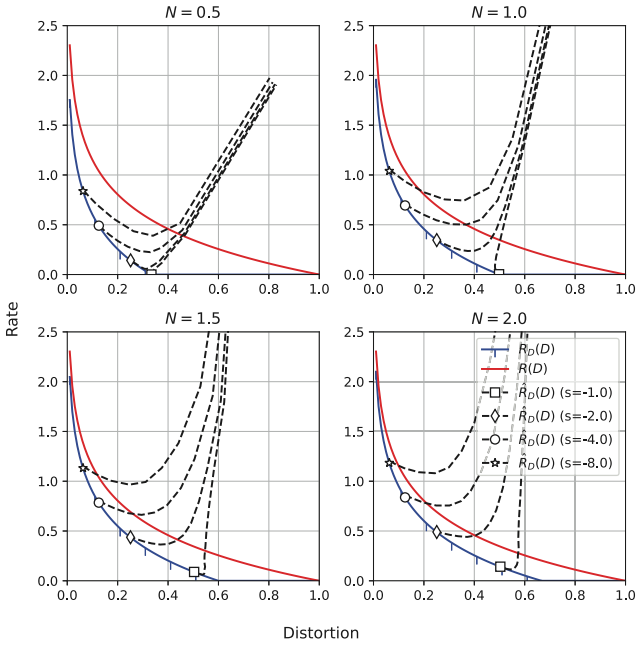

Fig. 3. Compression of Gaussian sources with varying N: The solid lines represent the known rate-distortion functions with and without decoder side information, $R_{\mathrm{D}}(D)$ and $R(D)$, respectively. The estimated rate-distortion points $(D_s, \hat{R}_{\mathrm{D}}(D_s))$ for four values of s obtained through our algorithm are also shown, along with their trajectory during the optimization phase. We observe that the estimated rate-distortion points forming $\hat{R}_{\mathrm{D}}(D)$ align closely with the true rate-distortion function $R_{\mathrm{D}}(D)$, exhibiting a negligible gap.

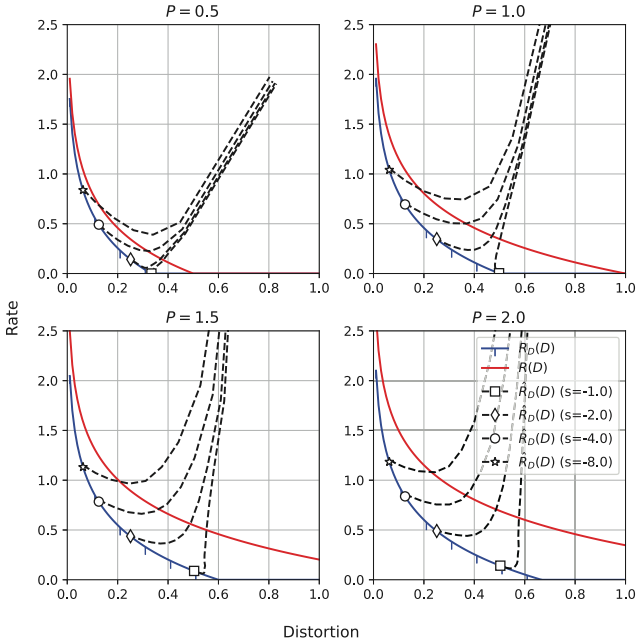

Fig. 4. Compression of Gaussian sources with varying P: Similar to the results in Figure 3, our estimated rate-distortion points $\hat{R}_D(D)$ align closely with the true rate-distortion function with side information $R_D(D)$.

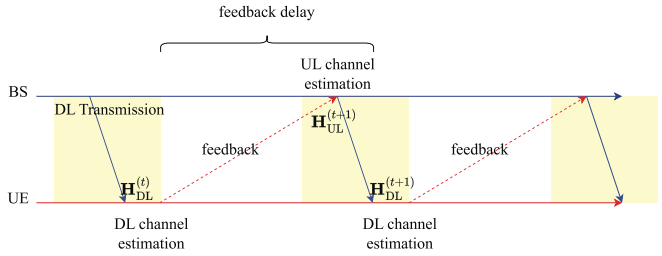

Fig. 5. Considered feedback system model.

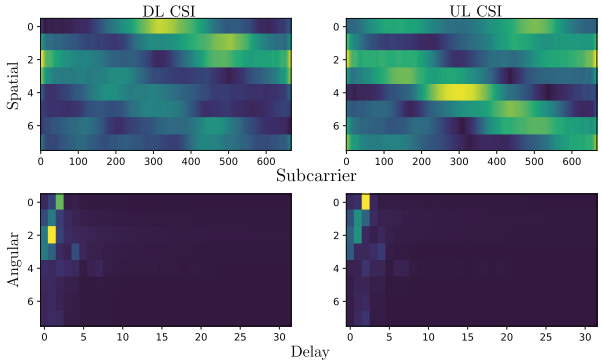

Fig. 6. A pair of samples is depicted, with DL CSI on the left-hand side and UL CSI on the right-hand side. In the compression problem, the objective is to reconstruct DL CSI while leveraging UL CSI, which serves as side information. Both DL CSI and UL CSI possess shared frequency-invariant characteristics, making the side information beneficial for reconstructing DL CSI.

$N = 1$, we again observe that our estimated rate-distortion points are closely aligned with the true rate-distortion function with side information $R_D(D)$. We also note that our learning trajectory reflects the fact that our algorithm acts as an upper-bound, with no points estimated below the true rate-distortion function $R_{\mathrm{D}}(D)$.

VI. Rate-Distortion Function Estimation for FDD CSI Compression With Side Information

In this section, we will employ our rate-distortion estimation algorithm to investigate rate-distortion functions specific to the Frequency Division Duplexing (FDD) CSI feedback problem. Our primary objective is to explore fundamental questions regarding the use of side information in wireless CSI feedback systems, guided by the key inquiries regarding: *(1) the application of Rate-Distortion Estimation in quantifying the gain from side information; (2) evaluating the performance of constructive algorithms relative to rate-distortion predictions; (3) the enhancement provided by side information in scenarios of limited feedback; (4) the influence of client mobility on the achievable rate-distortion function; and, (5) the corresponding gains associated with exploiting of side information.*

We note that these questions are crucial as they elucidate the quantifiable gains from utilizing side information and its effectiveness in enhancing wireless network performance across diverse operational environments. In the following subsection, we elaborate on the CSI feedback framework for FDD MIMO communication systems, exploring the problems and benefits of compression across varying mobile UE at different speeds.

A. Problem Description

We consider practical CSI feedback scenarios including feedback delays and apply our proposed algorithm to derive rate-distortion curves. We examine MIMO systems as illustrated in Figure 5, where the UE and BS communicate across n_{sc} subcarriers employing n_{tx} BS transmit antennas with a single UE antenna. Each feedback interval involves the UE estimating the DL CSI, $\mathbf{H}_{\mathrm{DL}}^{(t)}$, from the BS's downlink transmission, which includes CSI Reference Signals (CSI-RS). This information is then compressed and fed back to the BS. The BS simultaneously estimates UL CSI, $\mathbf{H}_{\mathrm{UL}}^{(t+1)}$, and receives the compressed DL CSI feedback from the UE. The ultimate goal for the BS is to estimate $\mathbf{H}_{\mathrm{DL}}^{(t+1)}$. The superscript $\cdot^{(t)}$ denotes the time slot, highlighting the changes in the CSI over time.

Predicting future CSI corresponding to the target DL data transmission, $\mathbf{H}_{\mathrm{DL}}^{(t+1)}$, is a critical challenge in MIMO systems due to the adverse impacts of feedback delays on system performance. This challenge is closely related to the concept of channel aging [54], [55], [56], which describes the modelling of variation in channel characteristics from the time they are estimated to when they are actually utilized [57].

In this setting, we will consider the Normalized Mean Squared Error (NMSE) as a distortion metric, defined as $\mathbb{E}[\|Z - \hat{Z}\|_2^2/\|Z\|_2^2]$, where $\|\cdot\|_2$ is elementwise square norm. We measure the distortion over the cropped angular-delay domain, as detailed in Appendix C. Then, the desired computing output Z corresponds to $\mathbf{H}_{\mathrm{DL}}^{(t+1)}$, X corresponds to $\mathbf{H}_{\mathrm{DL}}^{(t)}$, and Y corresponds to $\mathbf{H}_{\mathrm{UL}}^{(t+1)}$.

It is important to note that the desired output Z is not necessarily a deterministic function of the input source X and the side information Y. This is because even with perfect knowledge of the previous downlink CSI, $\mathbf{H}_{\mathrm{DL}}^{(t)}$, and the side information, $\mathbf{H}_{\mathrm{UL}}^{(t+1)}$, the future downlink CSI cannot be fully recovered due to inherent randomness. Formally, this scenario can be modeled as a *noisy Wyner-Ziv Coding* problem [58], where the desired output is correlated with the input source and side information but includes noise, and the noiseless input source is not available at the encoder. In this context, $\mathbf{H}_{\mathrm{DL}}^{(t)}$ can be viewed as a noisy observation of $\mathbf{H}_{\mathrm{DL}}^{(t+1)}$, which is correlated with $\mathbf{H}_{\mathrm{UL}}^{(t+1)}$. The encoder designs a codeword based solely on the noisy observation $\mathbf{H}_{\mathrm{DL}}^{(t)}$, while the decoder has access to the side information $\mathbf{H}_{\mathrm{UL}}^{(t+1)}$. Under this system model, the corresponding rate-distortion function $R_{\mathrm{D}}^{\mathrm{NWZ}}(D)$ is defined as follows [58], [59].

$$R_{\mathrm{D}}^{\mathrm{NWZ}}(D) = \min_{q_{U|X}, f:\mathbb{E}[d(Z,\hat{Z})]\leq D} I(X;U|Y) \tag{7}$$

where a Markov chain $(Z,Y) - X - U$ holds. It is evident that our proposed approach can be directly applied even within this noisy Wyner-Ziv compression framework without requiring any modifications, as the rate-distortion function can be obtained through the equal optimization problem under the different Markov chain.

The use of UL CSI as side information is based on the insight that UL CSI, generally obtained through pilot transmissions from the UE to the BS, shares a correlation

with DL CSI due to frequency-invariant characteristics [32], [33]. Figure 6 illustrates DL and UL CSI samples captured on the same time slot but represented in two distinct domains. The first row displays the normalized DL and UL CSI in the Spatial-Frequency domain, with the y-axis representing the number of antennas and the x-axis representing the subcarrier count. In the second row, both CSI instances are represented in the Angular-Delay domain, derived by applying the Inverse Fast Fourier Transform (IFFT) to the spatial-frequency domain instances and subsequently truncating the higher delay components. Notably, in the Angular-Delay domain, the DL and UL CSI demonstrate partial overlap in the dominant elements, underscoring the frequency-invariant features shared between them.

B. Configuration

Following the 3GPP specification, the feedback delay is fixed to 5ms [60], [61]. A CSI instance X characterized by the setting $n_{\text{tx}} = 8$ transmit (Tx) antennas and $n_{\text{sc}} = 667$ subcarriers. UL CSI also has the same parameters. The BS antennas are vertically polarized, as specified in [62]. In our numerical evaluation, we adopt the CDL-C channel model outlined in [63]. Specifically, for the implementation of UL and DL CSI, we select center frequencies of 1.9 GHz and 2.1 GHz, respectively, with a subcarrier spacing of 15 kHz. To account for a 5 ms feedback delay, we first generate time-correlated channel and delay coefficients with a delay spread of 300ns. These coefficients are then used to derive spatial-frequency domain representations. Subsequently, the spatial-frequency domain channel is divided based on the DL and UL bandwidth configurations described earlier.

In Appendices VII-A and VII-B, detailed configurations for parameterizing distributions, data preprocessing methods, and neural network training methods are provided.

C. Constructive Neural CSI Compression Algorithms

For the sake of comparison, we consider a fixed-rate compression task for the same system model and implement CSI compression algorithms utilizing the same processing architecture detailed in Sec. VII-B. It is pertinent to recall that for rate-distortion estimation, our approach entails an encoder that yields the mean and variance for the codeword probability distribution $q_{U|X}$, a core element of the Lagrangian loss function. This formulation ensures that the autoencoder architecture avoids an information bottleneck, as the output dimension of the encoder is not reduced.

In contrast to this architecture, for deterministic fixed codeword generation, we directly enforce a constraint on the encoder's output size. Specifically, consider a N_{cl}-bit compression of the given DL CSI. To implement this, we configure the encoder output size to be $\frac{N_{\text{cl}} N_{\text{Ebd}}}{\log_2(B)}$, where N_{Ebd} denotes the dimension of the embedding vector and B is a base of the codeword. By setting B to 16 (greater than 2), we employ a hexadecimal code to reduce the encoder output size. The output is then reshaped into a $\frac{N_{\text{cl}}}{\log_2(B)} \times N_{\text{Ebd}}$-size matrix, forming a total of $\frac{N_{\text{cl}}}{\log_2(B)}$ embedding vectors. Subsequently, each embedding vector is quantized by replacing it with the nearest embedding vector in a trainable codebook. Following this, the indices of the embedding vectors become the codewords, as depicted in Figure 7. Additional details can be found in Appendix VII-C.

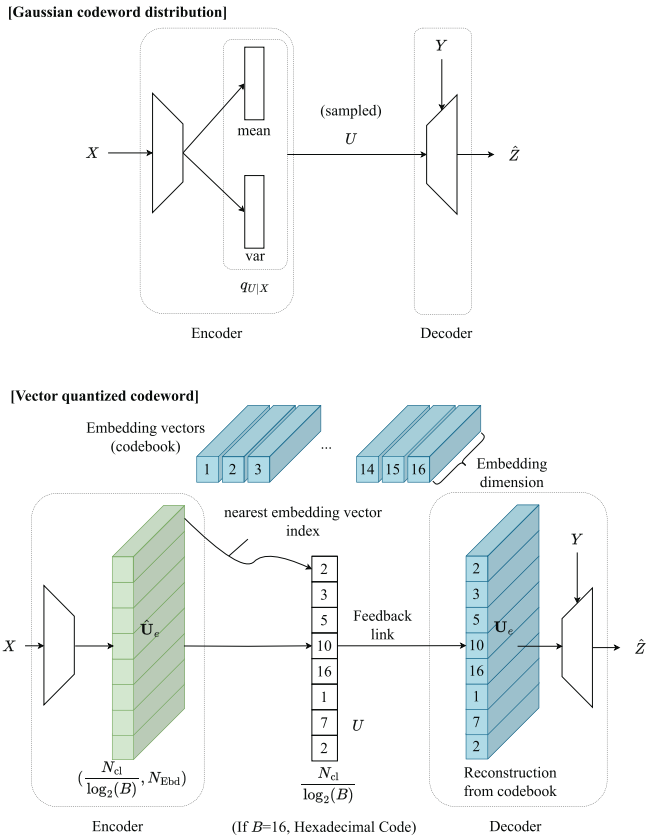

Fig. 7. Autoencoder architectures. (Top) For rate-distortion estimation, the autoencoder uses a Gaussian codeword distribution parameterized by a neural network that outputs the mean and variance. (Bottom) For the fixed-rate compression tasks, the Vector Quantized-Variational AutoEncoder (VQ-VAE) structure is applied where the encoder's output is quantized using embedding vectors, generating codewords through their indices.

We note that the comparison between our proposed rate-distortion estimation and the performance of the fixed-rate compression algorithm offers insights into whether the fixed-rate compression approach, which constitutes the majority of standardized CSI feedback, performs comparably with the rate-distortion benchmark.

D. Results

1) Rate-Distortion Framework Utilization: Can rate-distortion estimation be utilized to assess the advantages of incorporating side information in wireless systems? In Figure 8, four distinct curves are presented: the estimated rate-distortion curve with side information, $\hat{R}_{\text{D}}$, the one without side information, $\hat{R}$, the rate-distortion curve derived from the constructive compression algorithm with side information (compression with SI), and that without side information (compression). The five points plotted on $\hat{R}_{\text{D}}$ and $\hat{R}$ denote distinct estimated rate-distortion points, with their positions corresponding to specific s values: $-100, -10, -1, -0.1$, and -0.01 arranged from left to right. $\hat{R}(D)$, the estimated $R(D)$, is obtained by using the approach of [38] and using the same

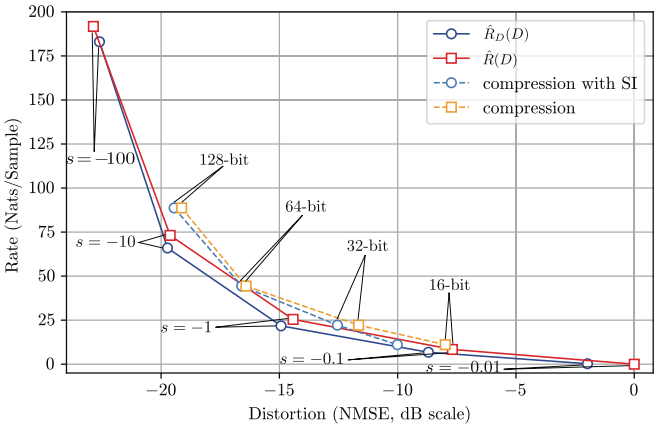

Fig. 8. CSI feedback for static UEs: Comparison of the estimated rate-distortion function, estimated rate-distortion function with side information, and (distortion, rate) points achieved by the neural compression algorithms.

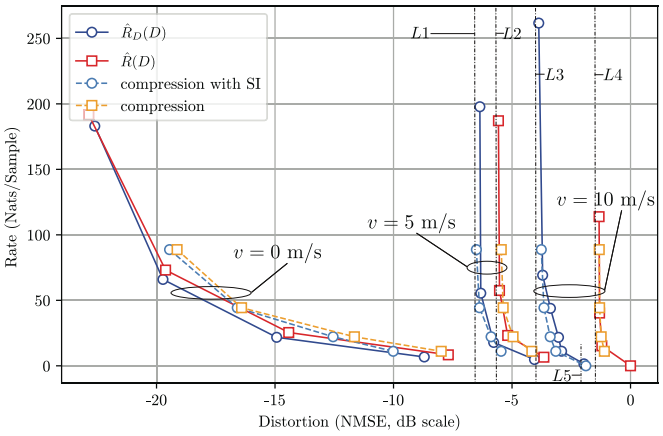

Fig. 9. Estimated rate-distortion functions and performance of the constructive CSI compression algorithms with different UE speeds. Unlike the static UE case ($v = 0$ m/s), the estimated rate-distortion functions corresponding to $v = 5, 10$ m/s show asymptotic achievable distortion limits ($L1$-$L4$). Moreover, as UE speed v increases, the gain of the side information—indicated by the gap between $\hat{R}_{\mathrm{D}}(D)$ and $\hat{R}(D)$-becomes greater.

neural architectures but which ignores the side information. By adjusting s values, we explore distortion levels from 0 dB to approximately -24 dB, connecting these points linearly to serve as an upper bound for the estimated rate-distortion curves. The rate-distortion curves from the constructive algorithms are generated by varying the compression bit rates as $N_{\mathrm{cl}} \in \{16, 32, 64, 128\}$ and connecting the points.

It is observed that introducing UL CSI for DL CSI compression is beneficial as, for all non-negative D, $\hat{R}_{\mathrm{D}}(D) \leq \hat{R}(D)$. Using this result, one can quantify the estimated gain from side information either by assessing the improvement in achievable distortion at a specific feedback rate or by evaluating the reduction in the required rate to achieve a target distortion. Our subsequent analysis will show that there is often a non-negligible gap between the minimum distortion achievable with and without side information, highlighting the crucial role side information may play in enhancing system performance.

2) Utilization of the Estimated Rate-Distortion Function as a Benchmark: How close is the achievable performance of constructive algorithms versus the estimated rate-distortion functions? The estimated rate-distortion functions can serve as benchmarks for evaluating the effectiveness of constructive compression algorithms. By varying the feedback rate of these algorithms, which operate at a fixed rate, one can measure the resulting distortion and compare the performance against the established rate-distortion functions. Recall that the proposed rate-distortion estimation algorithm provides an upper bound for the true rate-distortion function, setting a performance target that constructive algorithms are expected to meet or exceed. If the performance of a constructive algorithm falls below this benchmark, it clearly indicates an area for improvement.

The neural compression algorithm incorporating side information, achieved a rate-distortion curve that establishes an upper bound of $\hat{R}_{\mathrm{D}}$, and the discrepancy between $\hat{R}_{\mathrm{D}}$ and the constructive CSI compression algorithm's performance signals room for improvement. For example, in the case of around 80 Nats/sample, we may anticipate a potential improvement of about 1 dB. Notably, this gap is less pronounced in scenarios with lower rates, as shown in Figure 8.

3) Side Information Gains Under Limited Feedback: In wireless systems operating with limited CSI feedback, how are the benefits of side information affected?

The impact of side information on performance, both in terms of the rate and distortion, varies across different feedback regimes. The benefits of side information are particularly notable at lower CSI feedback rates, as illustrated in Figure 8. For instance, in regions where the rate is near zero, a gain of about 2.5 dB is expected from UL side information. This advantage diminishes with increased feedback resources; for example, at 150 Nats/Sample, the gain is observed to be near zero.

However, this trend, i.e., the diminishing relative side information gain as the feedback rate increases, does not hold if the UE exhibits high mobility, a topic we will explore further below.

4) Impact of UE Mobility on Rate-Distortion Trends: What trends emerge under varying mobility conditions?

Now we extend our analysis to include feedback scenarios for UE with mobility. Figure 9 illustrates the estimated rate-distortion functions across various UE speeds (0m/s, 5m/s, and 10m/s), with prior rate-distortion estimation results at $v = 0$m/s depicted in Figure 8 for reference. Each set of lines, enclosed within ellipses, corresponds to results observed within the same UE speed environment. Each vertical lines labeled $L1$ to $L4$ have points of lowest achievable distortion attained through rate-distortion tradeoffs from UEs with mobility.

As UE mobility increases, the achievable distortion for a given rate without side information deteriorates, reaching a plateau despite increasing the rate. Specifically, when $v = 0$m/s, distortion continually decreases with increasing rate. However, at $v = 5$m/s and $v = 10$m/s, distortion diminishes with rate escalation, converging to approximately -6 dB and -2 dB (see $L2$ and $L4$), respectively, even at high compression rates exceeding 100 Nats. This phenomenon arises from the inability to perfectly predict future CSI solely based on outdated information. The stochastic nature of channel

variations over time, compounded by mobility, renders complete modeling challenging. This trend still persists even when side information is available. In contrast to the $v = 0$m/s scenario, the asymptotes $L1$ and $L3$ suggest that achieving distortions of approximately -7 dB and -4 dB represents fundamental limits achievable through rate increase.

It is important to note that the trend of diminishing side information gain with increased feedback does not apply in these mobility scenarios. For instance, in the case of moderate mobility ($v = 5$m/s), even at high feedback rates exceeding 150 Nats/Sample, side information is expected to still contribute over 1 dB gain. Similarly, at high mobility ($v = 10$m/s) with feedback rates around or above 100 Nats/Sample, a substantial gain of about 3dB from side information is still anticipated. These observations highlight that in scenarios with higher mobility, side information retains its value, providing benefits that cannot be compensated for by merely increasing feedback rates.

5) Impact of UE Mobility on the Benefit of Side Information: How does UE mobility influence the relative benefit of side information?

The advantage of including side information becomes more pronounced with higher UE speeds. As the speed increases, the utility of the older CSI is increasingly ineffective. In contrast, when UL CSI is available as side information, it provides the decoder with more current and reliable information, enhancing its ability to accurately predict the DL CSI, thereby widening the performance gap between scenarios with and without side information.

In instances where the feedback codeword from the UE stems from significantly outdated CSI—typically associated with high UE speeds—the correlation between the desired output and side information can surpass that of the outdated CSI. This observation is evident in comparison with the distortion performance $L5$ and $L4$, where $L5$ represents distortion achieved solely based on side information without feedback, and $L4$ denotes the rate-distortion point achieved at high rates, believed to converge to the lowest achievable distortion. Notably, $L5 < L4$, indicating that side information conveys more meaningful insights for the desired output. This trend is also observed in the fixed-rate compression schemes with side information. Specifically, compression with side information at $v = 10$ m/s and zero feedback rate yields better performance than 128-bit compression schemes without side information. At $v = 0$m/s and $v = 5$m/s, on the other hand, possessing large information of the input source proves more informative than relying solely on side information.

E. Discussion

Our key findings can be summarized as follows. First, we observe that leveraging side information in a low-rate regime is advantageous, particularly when the channel remains relatively stable, as depicted in Figure 8. This is especially relevant in scenarios where it is critical to conserve the UE's power by limiting the feedback link. In such scenarios, it may be advantageous for the BS to utilize side information to realize those benefits.

Furthermore, for high-mobility UEs, employing side information is significantly beneficial. The simulation reveals that performance with more than 100 Nats without side information deteriorates to about -2 dB, which is significantly poorer than the performance achievable with side information even at nearly zero rates. In such cases, the benefits of investing in the processing of UL CSI may be expected to outweigh the costs, leading to notable performance enhancements.

The estimated rate-distortion functions incorporating side information for UEs with mobility indicate the existence of sweet spots achieving near the lowest feasible distortion at a reduced rate—specifically, 50 Nats at -6 dB NMSE for 5 m/s UEs, and 40 Nats for -4 dB NMSE for 10 m/s UEs. These findings suggest that there might be no benefit in increasing the feedback rate beyond a certain threshold for moving UEs and underscore the necessity for a mobility-dependent feedback rate control to enhance system efficiency further.

The recovered CSI obtained by minimizing distortion has wide-ranging applications, including interference mitigation at the BS or as a key enabler for precoding strategies. In the subsequent subsection, we explore the gain of side information in DL transmission, particularly its impact on Bit-Error-Rate (BER) performance. This analysis seeks to translate the observed gain of side information in rate-distortion tradeoffs into benefits for downstream tasks in wireless communication systems.

1) Applications to DL Transmission: The recovered CSI can be employed for precoding in DL transmission. Naturally, more accurately recovered CSI is expected to result in reduced BER. To assess the impact of UL side information in DL transmission scenarios, we evaluate constructive CSI compression algorithms—both with and without side information—by comparing their respective BER performance across various feedback rates and UE mobility conditions.

To achieve this, the recovered CSI of size 8×667 complex matrix is transposed, with each of the eight-dimensional vectors corresponding to a precoding vector for one of the 667 subcarriers. Each vector is normalized to have a Euclidean norm of one. BER measurements are then performed across a range of Signal-to-Noise Ratios (SNRs) from -2 dB to 3 dB, based on the implemented precoding strategy.

The results shown in Fig. 10 reveal consistent trends regarding the relationship between UE mobility and the benefits of side information. For static UEs with $v = 0$ m/s, side information yields notable performance gains in the low-rate feedback region, where limited knowledge of the original source data exists. As the feedback rate increases, the side information's impact diminishes, aligning with observations from Section VI-D-3. The middle subplot of Fig. 10 presents BER-SNR curves for UEs moving at $v = 5$ m/s. Similar to the findings in Section VI-D-4, UE mobility introduces a notable gain in the high-feedback-rate region, contrasting with the behavior observed for static UEs. The bottom subplot provides the results for high-mobility UEs ($v = 10$ m/s). In this scenario, side information continues to provide positive gains in high rate region. It is also observed that for schemes utilizing fewer than 64 feedback bits with side information and ones without side information, the BER performance

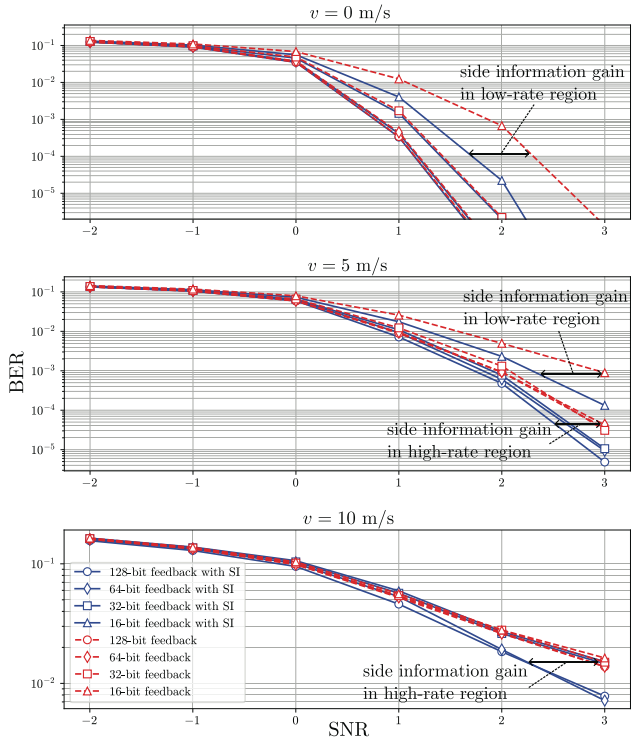

Fig. 10. BER vs. SNR curves for scenarios with $v = 0$, 5, and 10 m/s (top to bottom). The value of UL side information is more pronounced in the low-rate feedback region for static UEs. In contrast, high-mobility UEs can benefit from side information at high-rate region, as it provides more relevant and updated information compared to the outdated CSI.

are similar potentially due to the reduced accuracy of CSI recovery, as these schemes exhibit distortion levels exceeding -3 dB (see Fig. 9).

In summary, these additional analysis shows that side information is particularly valuable in low-rate feedback scenarios for low-mobility UEs. Conversely, for high-mobility scenarios, side information can provide benefits in the high-feedback-rate region.

F. Limitations, Potential Extensions, and Future Directions

1) Neural Network Models and Underlying Codeword Distribution: The proposed rate-distortion estimation method relies on the parameterizing distributions via neural networks. In this study, we utilized the model from [15] for the angular-delay domain representation of CSI. Our methodology for estimating the rate-distortion function with side information, along with the fixed-rate compression scheme utilizing UL side information, offers high flexibility, allowing the use of various existing parameterized models for processing CSI feedback.

However, as discussed in IV-B, neural network-based optimization may yield suboptimal performance if the set of representable distributions defined by the parameters does not include the true distribution or if the optimization process itself is suboptimal.

Potential improvements in the estimation accuracy could be achieved by employing more advanced architectures, such as the Transformer [64], known for its superior representational capabilities, or by exploring alternatives to the parameterized Gaussian distribution to potentially achieve a tighter rate-distortion curve. Additionally, increasing the number of data samples and employing advanced optimization techniques to solve the Lagrangian minimization problem more accurately can help narrow the estimation gap. That said, these approaches would introduce trade-offs, as larger and more complex models may increase computational challenges in optimization.

2) Side Information at Encoder and Decoder: The proposed rate-distortion estimation method can be readily extended to cases where side information is provided at the encoder as well, corresponding to the scenario where, in Figure 1, the switch (A) is closed. For instance, when the decoder's objective is to recover input source X, and side information is available at both the encoder and decoder, the rate-distortion function $R_{\mathrm{ED}}(D)$ is given as follows [46].

$$R_{\mathrm{ED}}(D) = \min_{q_{U|X,Y}(u|x,y),\hat{x}=f(u,y):\ \mathbb{E}[d(X,\hat{X})]\le D} I(X;U|Y). \quad (8)$$

For a given negative slope value s, a corresponding point on $R_{\mathrm{ED}}(D)$ is achieved by solving the following problem:

$$\min_{q_{U|X,Y},f} \left\{ \mathbb{E}_{X,Y,U}\left[\log \frac{q_{U|X,Y}(U|X,Y)}{q_{U|Y}(U|Y)}\right] - s\mathbb{E}[d(X,\hat{X})] \right\}. \quad (9)$$

Similarly, it is expected that the parameterization of $q_{U|X,Y}(U|X,Y)$, $q_{U|Y}(U|Y)$, and f can be adapted to solve the problem. This scenario, where side information is available at the encoder, is anticipated to model situations where the CSI feedback encoder incorporates some UE information, such as the UE's speed, with the BS being aware of this side information (i.e., mobility-aware feedback), or other available information.

3) Different Types of Side Information, Distortion Measures, and Objectives: There are different types of useful side information for signal design including historical uplink/downlink CSI and non-Radio Frequency (RF) data, such as GPS and LiDAR. Integrating this information into RF tasks like beamforming shows promise for advancing wireless systems [65], [66]. Moreover, ray-tracing techniques in modern digital twin systems can provide highly accurate channel estimates that closely correlate with true CSI [67], [68], [69]. Analysis using our estimation method to quantify the benefits of integrating those information in downstream tasks is expected to provide useful insights for future wireless networks. Additionally, applying our approach to scenarios with different objectives and distortions, such as the average cosine similarity across channel vectors or the BER in the DL communication channel, would also be an interesting direction for future research.

VII. Conclusion

In this paper, we propose a new algorithm for estimating the generalized rate-distortion function, with a specific emphasis on the rate-distortion function for computing with side information. We apply this algorithm to the DL CSI feedback problem having UL CSI as the side information and generate rate-distortion functions. Using the estimated rate-distortion

functions, we measure the gain of side information across different feedback rates and UE mobility profiles. The results signal that the benefits of side information depend significantly on environmental factors and are most pronounced for UEs with high mobility and low feedback overheads. We expect that such observation can aid practical system design, helping to decide whether to incorporate side information to enhance performance at an extra cost.

Appendix A
Problem Formulation (3)

The definition of conditional mutual information leads to the following formulation.

$$\min_{q_{U|X}(u|x),f}\Big\{I(X;U|Y)-s\mathbb{E}[d(Z,\hat{Z})]\Big\}$$
$$=\min_{q_{U|X}(u|x),f}\Big\{\sum_{\mathcal{X},\mathcal{Y},\mathcal{U}} q_{X,Y,U}(x,y,u) \tag{10}$$
$$\times\log\frac{p_Y(y)q_{X,Y,U}(x,y,u)}{p_{X,Y}(x,y)q_{Y,U}(y,u)}-s\mathbb{E}[d(Z,\hat{Z})]\Big\}. \tag{11}$$

This can be equivalently expressed as

$$\min_{q_{U|X}(u|x),f}\left\{\mathbb{E}\left[\log\frac{q_{U|X,Y}(U|X,Y)}{q_{U|Y}(U|Y)}\right]-s\mathbb{E}[d(Z,\hat{Z})]\right\}. \tag{12}$$

Here, the expectation is taken with respect to the joint distribution of (X,Y,U). For given realizations of X and Y, the conditional distribution of the codeword U is determined solely by X based on the communication model that we deal with. This restriction arises from the system model, which does not allow for the codeword to be controlled based on side information. Consequently, this simplifies to $q_{U|X,Y}(U|X,Y)=q(U|X)$, thereby completing the proof.

Appendix B
Proof of Lemma 1

We start with the following equation:

$$q_{U|X}(U|X;\boldsymbol{\theta}_{\text{po}})=q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(U|X,Y). \tag{13}$$

This equation stems from the premise that the distribution of codeword U is deterministic on X when it is given. Following this, we have

$$\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{X,Y,U}\left[\log\frac{q_{U|X}(U|X;\boldsymbol{\theta}_{\text{po}})}{\hat{q}_{U|Y}(U|Y)}\right]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{X,Y}[\mathrm{KL}(q_{U|X}(U|X;\boldsymbol{\theta}_{\text{po}})\|\hat{q}_{U|Y}(U|Y))]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{X,Y}[\mathrm{KL}(q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(U|X,Y)\|\hat{q}_{U|Y}(U|Y))]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{X,Y}\Big[\sum_{\mathcal{U}}q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|X,Y)$$
$$\big(\log q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|X,Y)-\log\hat{q}_{U|Y}(u|Y)\big)\Big]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{Y}\Big[\sum_{\mathcal{X},\mathcal{U}}p_{X|Y}(x|Y)q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|x,Y)$$
$$\big(\log q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|x,Y)-\log\hat{q}_{U|Y}(u|Y)\big)\Big]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{Y}\Big[\sum_{\mathcal{X},\mathcal{U}}q_{U,X|Y;\boldsymbol{\theta}_{\text{po}}}(u,x|Y)$$
$$\big(\log q_{U|X,Y;\boldsymbol{\theta}_{\text{po}}}(u|x,Y)-\log\hat{q}_{U|Y}(u|Y)\big)\Big]$$
$$=\arg\min_{\hat{q}_{U|Y}}\mathbb{E}_{Y}\Big[\sum_{\mathcal{X},\mathcal{U}}q_{U,X|Y;\boldsymbol{\theta}_{\text{po}}}(u,x|Y)\big(\log q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})-\log\hat{q}_{U|Y}(u|Y)\big)\Big]$$
$$=\arg\max_{\hat{q}_{U|Y}}\mathbb{E}_{Y}\left[\sum_{\mathcal{X},\mathcal{U}}q_{U,X|Y;\boldsymbol{\theta}_{\text{po}}}(u,x|Y)\log\hat{q}_{U|Y}(u|Y)\right]$$
$$=\arg\max_{\hat{q}_{U|Y}}\mathbb{E}_{Y}\left[\sum_{\mathcal{U}}q_{U|Y;\boldsymbol{\theta}_{\text{po}}}(u|Y)\log\hat{q}_{U|Y}(u|Y)\right]. \tag{14}$$

As both $q_{U|Y;\boldsymbol{\theta}_{\text{po}}}$ and $\hat{q}_{U|Y}$ are probability distributions, applying Gibbs' inequality completes the proof.

Appendix C
Simulation Configuration in Sec. VI

A. Training Configuration

We utilize the Adam optimizer [70] with a cosine annealing learning rate scheduler [71], where the learning rate ranges from 5e-4 to 1e-6, alongside a minibatch size of 100. We set $T=2\times10^6$ and $T'=1$.

B. Parameterization of Distributions

CSI instances are preprocessed by converting to the angular-delay domain via Inverse Fast Fourier Transform (IFFT) and trimming high-delay near-zero regions, following existing CSI preprocessing methods [14], [29]. This results in an 8×32 complex-valued matrix, with 8 angular and 32 cropped delay components. We utilize inception block-based [72] encoding and decoding schemes [15] for the distribution parameterization. The encoding module processes the input source using the specified encoding scheme, producing an output of size $2\times8\times32$, which corresponds to a real-valued 8×32 complex matrix. This output is then expanded to a 1024-dimensional representation via a linear layer. The expanded representation is subsequently divided into two components, each with dimensions $2\times8\times32$, representing the component-wise mean and variance for $q_{U|X}(u|x;\boldsymbol{\theta}_{\text{po}})$. The same structure is used for $q_{U|Y}(u|y;\boldsymbol{\theta}_{\text{pr}})$. The decoder, $f(u,y;\boldsymbol{\theta}_{\text{dec}})$, takes two 8×32 complex-valued matrices (codeword and side information) as input. Initially, the codeword is processed through a linear layer and then concatenated with the side information. Subsequently, this concatenated information is fed into an inception block-based decoder [15], outputting an 8×32 complex matrix.

C. A Constructive Neural CSI Compression Algorithm

In this subsection, we describe the implementation of the constructive CSI compression algorithm. By adjusting the output of the encoder while maintaining the same processing architectures of both the encoder and decoder, a deterministic N_{cl}-bit compression algorithm can be devised. Consider N_{cl} sized binary codeword for the compression. For efficient implementation, we take $B=16$ as a new base and consider codewords of length $l=N_{\text{cl}}/\log_2(16)$ to maintain cardinality. The encoder outputs a vector $\hat{\mathbf{U}}_e\in\mathbb{R}^{\frac{N_{\text{cl}}}{\log_2(16)}\times N_{\text{Ebd}}}$ where N_{Ebd}

is an embedding dimension, which is achieved by adjusting the size of the linear transformation layer of the encoding module in [15]. This vector is then quantized using a trainable codebook of 16 different N_{Ebd}-dimensional vectors, based on [73]. The quantized output is denoted by $\mathbf{U}_e$. The encoder transmits indices of these vectors via a wireless link to the BS, forming codeword U. The BS reconstructs $\mathbf{U}_e$ using these indices and the corresponding vectors in the codebook. The same decoder modules are then applied. We have $N_{\text{Ebd}} = 8$ and set the distortion function as the NMSE.

The codebook containing the embedding vectors is also trainable. This is achieved by penalizing the loss function such that the quantized output of the encoder is encouraged to be close to the corresponding embedding vectors, while simultaneously guiding the embedding vectors to be close to the encoder output. Specifically, for given instances z, $\hat{z}$, and u, we use the penalized loss function [73] as

$$d(z,\hat{z}) + \beta(\|\mathtt{sg}[\mathbf{U}_e] - u\|_{\text{F}}^2 + \|\mathbf{U}_e - \mathtt{sg}[u]\|_{\text{F}}^2), \tag{15}$$

where $\mathtt{sg}$ represents the stop-gradient operator. This operator treats its input as a constant, effectively blocking the gradient propagation through the input. We set β equal to the elementwise variance of $\mathbf{H}_{\text{DL}}$ and update all trainable parameters in the direction of minimizing the penalized loss.

REFERENCES

[1] H. Kim, H. Kim, and G. De Veciana, "Estimation of Rate- distortion function for computing with decoder side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 422–427, doi: 10.1109/ISIT57864.2024.10619642.

[2] Evolved Universal Terrestrial Radio Access (E-UTRA) Physical Layer Procedures, Standard 36.213, 3rd Gener. Partnership Project (3GPP), Tech. Specification Group Radio Access Netw., 2017.

[3] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. Hoboken, NJ, USA: Wiley, 2005.

[4] Q. Li et al., "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[5] D. Love, R. Heath, V. N. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[6] T. Strohmer andR. W. Heath Jr., "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[7] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[8] D. J. LoveandR. W. Heath Jr., "Grassmannian beamforming on correlated MIMO channels," in *Proc. IEEE Global Telecommun. Conf., GLOBECOM*, vol. 1, Nov. 2004, pp. 106–110.

[9] W. Santipach and M. L. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," in *Proc. IEEE MILCOM*, Jul. 2004, pp. 141–146.

[10] C. Au-Yeung and D. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, Feb. 2007.

[11] P. Cheng and Z. Chen, "Multidimensional compressive sensing based analog CSI feedback for massive MIMO-OFDM systems," in *Proc. IEEE VTC-Fall*, Sep. 2014, pp. 1–6.

[12] R. Ahmed, E. Visotsky, and T. Wild, "Explicit CSI feedback design for 5G new radio phase II," in *Proc. WSA; 22nd Int. ITG Workshop Smart Antennas*, Mar. 2018, pp. 1–5.

[13] J. Guo, C. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 8017–8045, Dec. 2022.

[14] C. Wen, W. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[15] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI feedback with deep learning in massive MIMO system," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[16] Q. Cai, C. Dong, and K. Niu, "Attention model for massive MIMO CSI compression feedback and recovery," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–5.

[17] J. Shin, Y. Kang, and Y.-S. Jeon, "Vector quantization for deep-learning-based CSI feedback in massive MIMO systems," 2024, *arXiv:2403.07355*.

[18] S. Ravula and S. Jain, "Deep autoencoder-based massive MIMO CSI feedback with quantization and entropy coding," in *Proc. IEEE GLOBECOM*, Dec. 2021, pp. 1–6.

[19] J. So and H. Kwon, "Universal auto-encoder framework for MIMO CSI feedback," in *Proc. IEEE Global Commun. Conf.*, Dec. 2023, pp. 1–7.

[20] B. Park, H. Do, and N. Lee, "Multi-rate variable-length CSI compression for FDD massive MIMO," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 7715–7719.

[21] M. Nerini, V. Rizzello, M. Joham, W. Utschick, and B. Clerckx, "Machine learning-based CSI feedback with variable length in FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 2886–2900, May 2023.

[22] X. Liang, H. Chang, H. Li, X. Gu, and L. Zhang, "Changeable rate and novel quantization for CSI feedback based on deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10100–10114, Dec. 2022.

[23] H. Kim, H. Kim, and G. De Veciana, "Learning variable-rate codes for CSI feedback," in *Proc. IEEE Global Commun. Conf.*, Dec. 2022, pp. 1435–1441.

[24] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4044–4057, Jul. 2021.

[25] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2621–2633, Dec. 2020.

[26] X. Li, J. Guo, C.-K. Wen, S. Jin, S. Han, and X. Wang, "Multi-task learning-based CSI feedback design in multiple scenarios," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7039–7055, Dec. 2023.

[27] Y. Sun, W. Xu, L. Fan, G. Y. Li, and G. K. Karagiannidis, "AnciNet: An efficient deep learning approach for feedback compression of estimated CSI in massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 12, pp. 2192–2196, Dec. 2020.

[28] J. Xu, B. Ai, N. Wang, and W. Chen, "Deep joint source-channel coding for CSI feedback: An end-to-end approach," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 260–273, Jan. 2023.

[29] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 416–419, Apr. 2019.

[30] X. Li and H. Wu, "Spatio-temporal representation with deep neural recurrent network in MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 653–657, May 2020.

[31] Q. Li, A. Zhang, P. Liu, J. Li, and C. Li, "A novel CSI feedback approach for massive MIMO using LSTM-attention CNN," *IEEE Access*, vol. 8, pp. 7295–7302, 2020.

[32] D. Vasisht, S. Kumar, H. Rahul, and D. Katabi, "Eliminating channel feedback in next-generation cellular networks," in *Proc. ACM SIGCOMM Conf.*, Aug. 2016, pp. 398–411.

[33] D. Han, J. Park, and N. Lee, "FDD massive MIMO without CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4518–4530, May 2024.

[34] Y.-C. Lin, T.-S. Lee, and Z. Ding, "Exploiting partial FDD reciprocity for beam-based pilot precoding and CSI feedback in deep learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1474–1488, Feb. 2024, doi: 10.1109/TWC.2023.3289929.

[35] Y. Liu and O. Simeone, "HyperRNN: Deep learning-aided downlink CSI acquisition via partial channel reciprocity for FDD massive MIMO," in *Proc. IEEE SPAWC*, Sep. 2021, pp. 31–35.

[36] Q. Li and Y. Chen, "Rate distortion via deep learning," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 456–465, Jan. 2020.

[37] E. Lei, H. Hassani, and S. Saeedi Bidokhti, "Neural estimation of the rate-distortion function with applications to operational source coding," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 4, pp. 674–686, Dec. 2022.

[38] Y. Yang and S. Mandt, "Towards empirical sandwich bounds on the rate-distortion function," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2021, pp. 1–21.

[39] Y. Yang, S. Eckstein, M. Nutz, and S. Mandt, "Estimating the rate-distortion function by Wasserstein gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 1–12.

[40] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.
[41] 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.
[42] Y. Uuğur, I. E. Aguerri, and A. Zaidi, "A generalization of blahut-arimoto algorithm to compute rate-distortion regions of multiterminal source coding under logarithmic loss," in *Proc. IEEE ITW*, Nov. 2017, pp. 349–353.
[43] A. Dembo and L. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.
[44] D. Tsur, B. Huleihel, and H. Permuter, "Rate distortion via constrained estimated mutual information minimization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2023, pp. 695–700.
[45] A. D. 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.
[46] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[47] S. Ayzik and S. Avidan, "Deep image compression using decoder side information," in *Proc. 16th ECCV*. Cham, Switzerland: Springer, Jan. 2020, pp. 699–714.
[48] N. Mital, E. Özyilkan, A. Garjani, and D. Gündüz, "Neural distributed image compression using common information," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2022, pp. 182–191.
[49] Y. Huang et al., "Learned distributed image compression with multi-scale patch matching in feature domain," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 4, pp. 4322–4329.
[50] F. Dupuis, W. Yu, and F. M. J. Willems, "Blahut-arimoto algorithms for computing channel capacity and rate-distortion with side information," in *Proc. Int. Symp. Inf. Theory (ISIT)*, 2004, p. 181.
[51] E. Özyılkan, J. Ballé, and E. Erkip, "Learned wyner-ziv compressors recover binning," in *Proc. IEEE ISIT*, Jan. 2023, pp. 701–706.
[52] H. Yamamoto, "Wyner-Ziv theory for a general function of the correlated sources (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 803–807, Sep. 1982.
[53] J. Ballé, D. Minnen, S. Singh, S. Jin Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.
[54] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, 2013.
[55] Y. Han, Q. Liu, C.-K. Wen, M. Matthaiou, and X. Ma, "Tracking FDD massive MIMO downlink channels by exploiting delay and angular reciprocity," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1062–1076, Sep. 2019.
[56] M. Arnold, S. Dörner, S. Cammerer, S. Yan, J. Hoydis, and S. T. Brink, "Enabling FDD massive MIMO through deep learning-based channel prediction," 2019, *arXiv:1901.03664*.
[57] Y. Zhang, A. Alkhateeb, P. Madadi, J. Jeon, J. Cho, and C. Zhang, "Predicting future CSI feedback for highly-mobile massive MIMO systems," 2022, *arXiv:2202.02492*.
[58] D. Rebollo-Monedero and B. Girod, "A generalization of the rate-distortion function for wyner-ziv coding of noisy sources in the quadratic-Gaussian case," in *Proc. Data Compress. Conf.*, 2005, pp. 23–32.
[59] H. Yamamoto and K. Itoh, "Source coding theory for multiterminal communication systems with a remote source," *IEICE Trans.*, vol. E-63, no. 10, pp. 700–706, Oct. 1980.
[60] Discussion Summary for CSI Enhancements MTRP and FR1 FDD Reciprocity, Standard TSG RAN WG1 Meeting 102-e, 3rd Gener. Partnership Project (3GPP), 2020.
[61] L. Thiele, M. Olbrich, M. Kurras, and B. Matthiesen, "Channel aging effects in CoMP transmission: Gains from linear channel prediction," in *Proc. Conf. Rec. Forty 5th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 1924–1928.
[62] Study on Channel Model for Frequencies From 0.5 To 100 GHz, Standard TR 38.901, 3rd Gener. Partnership Project (3GPP), 2019.
[63] Study on Channel Model for Frequencies From 0.5 To 100 Ghz, Standard G. T. 38.901, 2017.
[64] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.
[65] D. Roy et al., "Going beyond RF: A survey on how AI-enabled multimodal beamforming will shape the NextG standard," *Comput. Netw.*, vol. 228, Jun. 2023, Art. no. 109729.
[66] B. Salehihikouei et al., "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7639–7655, Jul. 2022.
[67] T. Orekondy, P. Kumar, S. Kadambi, H. Ye, J. Soriaga, and A. Behboodi, "Winert: Towards neural ray tracing for wireless channel modelling and differentiable simulations," in *Proc. The 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–21.
[68] H. Choi, J. Oh, J. Chung, G. C. Alexandropoulos, and J. Choi, "WiThRay: A versatile ray-tracing simulator for smart wireless environments," *IEEE Access*, vol. 11, pp. 56822–56845, 2023.
[69] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong, and T. Kürner, "The design and applications of high-performance ray-tracing simulation platform for 5G and beyond wireless communications: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 10–27, 1st Quart., 2019.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[71] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[72] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[73] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 1–12.

Heasung Kim (Member, IEEE) received the B.S. degree in Computer and Communication Engineering from Korea University, Seoul, South Korea, in 2017, and the M.S. degree from the Department of Electrical and Computer Engineering, Seoul National University, Seoul, in 2019. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering with The University of Texas at Austin. From 2019 to 2021, he was a Machine Learning Engineer with Samsung Network Business and Samsung Electronics. His research interests include wireless networks, information theory, and machine learning algorithms.

Gustavo de Veciana (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993, respectively. He was the Director and the Associate Director of the Wireless Networking and Communications Group (WNCG), The University of Texas at Austin, from 2003 to 2007. He is currently a Professor and the Associate Chair of the Department of Electrical and Computer Engineering. His research focuses on the design, analysis and control networks, information theory, and applied probability. His current research interests include measurement, modeling, and performance evaluation; wireless and sensor networks; architectures and algorithms to design reliable computing and network systems. In 2009, he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He was a recipient of the Cockrell Family Regents Chair in Engineering and the National Science Foundation CAREER Award 1996. He was a co-recipient of six best paper awards, including the IEEE William McCalla Best ICCAD Paper Award for 2000; and the Best Paper in *ACM Transactions on Design Automation of Electronic Systems*, from January 2002 to 2004. He has been an Editor of IEEE/ACM TRANSACTIONS ON NETWORKING.

Hyeji Kim (Senior Member, IEEE) received the B.S. degree in electrical engineering from KAIST in 2011 and the Ph.D. degree in electrical engineering from Stanford University in 2016. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of Texas at Austin (UT Austin). Before joining UT Austin, she was a Post-Doctoral Research Associate with the University of Illinois at Urbana-Champaign and a Researcher with Samsung AI Research Cambridge, U.K. Her research interests include the intersection of information theory, machine learning, and wireless communications. She was a recipient of the NSF CAREER Award.