Practical Short-Length Coding Schemes for Binary Distributed Hypothesis Testing

Ismaila Salihou Adamou, Elsa Dupraz, *Senior Member, IEEE*, Reza Asvadi, *Senior Member, IEEE*, and Tadashi Matsumoto, *Life Fellow, IEEE*

***Abstract*—This paper addresses the design of practical short-length coding schemes for Distributed Hypothesis Testing (DHT). While most prior work on DHT has focused on information-theoretic analyses—deriving bounds on Type-II error exponents via achievability schemes based on quantization and quantize-binning—the practical implementation of DHT coding schemes has remained largely unexplored. Moreover, existing practical coding solutions for quantization and quantize-binning approaches were developed for source reconstruction tasks considering very long code lengths, and they are not directly applicable to DHT. In this context, this paper introduces efficient short-length implementations of quantization and quantize-binning schemes for DHT, constructed from short binary linear block codes. Numerical results show the efficiency of the proposed coding schemes compared to uncoded cases and to existing schemes initially developed for data reconstruction. In addition to practical code design, the paper derives exact analytical expressions for the Type-I and Type-II error probabilities associated with each proposed scheme. The provided analytical expressions are shown to predict accurately the practical performance measured from Monte Carlo simulations of the proposed schemes. These theoretical results are novel and offer a useful framework for optimizing and comparing practical DHT schemes across a wide range of source and code parameters.**

***Index Terms*—Distributed Hypothesis Testing, short-length codes, binary quantization, quantize-binning scheme, linear block codes, Neyman-Pearson.**

I. Introduction

In the era of 5th generation (5G) wireless communications systems and beyond, the paradigm of communication systems is shifting to address emerging challenges and requirements. Historically, these systems were designed to ensure reliable transmission, mostly focusing on minimizing error probability or distortion between original and reconstructed data [1], [2]. However, modern communication systems are now increasingly dedicated to specific tasks that require an optimized design. Especially in the emerging field of goal-oriented communications [3]–[5], the objective is no longer to reconstruct the data but rather to apply specific tasks directly upon the received data. While examples of such tasks include regression, classification, and semantic analysis, this paper focuses on the important case of decision-making. Examples of applications include sensors embedded in the human body for healthcare disease detection [6], underwater activity monitoring [7], or traffic jam detection from route planning of autonomous vehicles [8].

In Information Theory, the problem of decision-making over coded data is formalized as distributed hypothesis testing (DHT), first introduced in [9]. DHT considers a scenario with two separate terminals: one observing a source X, and the other observing a source Y. These sources are jointly distributed according to the joint distribution P_{XY}, which depends on one of the two hypotheses, $\mathcal{H}_0$ or $\mathcal{H}_1$ [9]–[11]. The terminals transmit encoded messages under given rate constraints, R_1 for X and R_2 for Y. The receiver then makes a decision between $\mathcal{H}_0$ and $\mathcal{H}_1$, by applying a hypothesis test over the received coded data. In this work, we focus on two distinct setups. The first setup, referred to as the asymmetric setup, assumes that the receiver has access to a coded version of X, while Y is losslessly observed as side information at the receiver. This setup is similar to the Wyner-Ziv setup for lossy source coding [1]. The second setup, known as the symmetric setup, assumes that both X and Y are encoded.

In both setups, the hypothesis testing performance is characterized by two types of error probabilities: Type-I error probability, denoted by α_n, and Type-II error probability, denoted by β_n. A Type-I error occurs when $\mathcal{H}_1$ is chosen while the true hypothesis is $\mathcal{H}_0$, whereas a Type-II error arises when $\mathcal{H}_0$ is selected while $\mathcal{H}_1$ is true. The central question in the framework of DHT is: how can one design coding schemes so as to satisfy the rate constraints while ensuring optimal decision-making at the receiver? Here, optimality is defined as minimizing the Type-II error probability, β_n, under a given constraint on the Type-I error probability α_n. Addressing this question requires tackling both the information-theoretic limits of the problem and the practical design of coding schemes.

A. Prior Works in Information Theory

In the information-theoretic analysis of DHT, the primary objective is to characterize the achievable error exponent of the Type-II error probability, while the Type-I error probability is

Ismaila Salihou Adamou and Elsa Dupraz are with the Department of Mathematical and Electrical Engineering, IMT Atlantique, 29238 Brest, France

Reza Asvadi is with the Department of Telecommunications, Faculty of Electrical Engineering, Shahid Beheshti University, Tehran 1983969411, Iran

Tadashi Matsumoto is with the Department of Mathematical and Electrical Engineering, IMT Atlantique, 29238 Brest, France, and also with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland. He is also Professor Emeritus of Japan Advanced Institute of Science and Technology, Japan

This work has received French government support granted to the Cominlabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01.

Part of the content of this paper was published in: Elsa Dupraz, Ismaila Salihou Adamou, Reza Asvadi, Tadashi Matsumoto, Practical Short-Length Coding Schemes for Binary Distributed Hypothesis Testing, *International Symposium on Information Theory (ISIT)*, 2024. We extend the analysis to encompass the symmetric two encoders hypothesis testing setup.

kept below a prescribed threshold [9], [10]. The asymmetric setup has been extensively studied, with several achievable coding schemes proposed to refine lower bounds on the error exponent. Ahlswede and Csiszár first introduced the quantization scheme [9], which is optimal for a special case known as testing against independence. Han enhanced the quantization scheme by incorporating a joint typicality check between the source and its quantized version [10]. However, these approaches do not fully exploit the correlation between sources X and Y.

To address this limitation, Shimokawa et al. proposed the Shimokawa-Han-Amari (SHA) scheme, also known as the quantize-binning scheme, which performs random binning after the quantization [12]. This approach originates from the Wyner-Ziv coding scheme [1] and achieves tighter lower bounds on the error exponent. Further refinements to the SHA scheme addressed the trade-off between binning errors and hypothesis testing errors, as investigated for i.i.d. sources in [13] and for non-i.i.d. sources in [14]. Most recently, Kochman and Wang enhanced the SHA scheme by refining the entropy check introduced by Shimokawa et al. [15]. Finally, the SHA scheme has been generalized to more complex communication scenarios, including discrete memoryless channels [16], multiple-access channels [17], and two-hop relay networks [18].

The symmetric setup has received much less attention, except in the specific case of zero-rate compression where one or both coding rates asymptotically approach zero [10], [19]–[22]. The zero-rate case has limited relevance for conventional lossless or lossy compression, but it has important applications in statistical hypothesis testing [10]. In this case, research has focused on designing optimal testing schemes [22], [23] and characterizing achievable error exponents [10], [19]–[23].

B. Prior works on practical coding schemes

While the information-theoretic performance of DHT has been extensively studied, the design of practical coding schemes for this setup has received considerably less attention. In fact, information-theoretic schemes are not directly implementable and largely rely on impractical assumptions, such as infinite block lengths, which are incompatible with the finite and typically short sequences encountered in practical decision-making scenarios. This gap in the literature serves as the primary motivation of this paper, with a focus on binary sources.

For data reconstruction, several practical coding techniques have been proposed in the literature. They include binary quantizers [24], binning schemes [25]–[28], and quantize-binning schemes [29], all constructed using linear block codes. For instance, in [26], [27], binning schemes based on Low-Density Parity-Check (LDPC) codes are proposed, achieving lossless compression near the Slepian-Wolf limit for correlated binary sources. Similarly, binary quantization schemes have been developed using LDPC codes [30], or Low-Density Generator Matrix (LDGM) codes of which decoding is conducted by Bias Propagation (BiP) algorithms [24], [31], [32]. Furthermore, it was shown in [29], [33] that compound LDGM/LDPC constructions for quantize-binning schemes can achieve the Wyner-Ziv rate-distortion function for binary sources.

While these coding schemes are effective for source reconstruction, they are less suitable for DHT due to their use of belief propagation decoding algorithms which are designed for very long sequences, often exceeding 10^3 to 10^4 bits. In contrast, DHT involves short-length sequences, where only a few dozen bits may suffice for accurate decision-making. This raises an important question on how to design efficient quantizers and quantize-binning schemes with short-length codes. Another key issue is the design of effective hypothesis tests over coded data, given that the methods proposed in information-theoretic proofs of DHT are impractical. In this paper, we address these issues and propose efficient short-length coding schemes for DHT.

C. Contributions

In this paper, we introduce a first practical short-length coding scheme for DHT in the asymmetric setup, where side information Y is fully available at the decoder. We then extend this scheme to the symmetric setup, where both the source X and side information Y are compressed by independent encoders. We propose practical quantization and quantize-binning schemes for DHT in both asymmetric and symmetric setups.

Based on the principles of practical Wyner-Ziv coding, our schemes employ linear block codes designed specifically for short block lengths ($n < 100$ bits). We first describe the construction of the coding scheme and provide the hypothesis test expression. We then derive exact analytical expressions for the Type-I and Type-II error probabilities for the given code. These analytical tools are novel, and enable the optimization and comparison of the proposed schemes across a broad range of source and code parameters.

The major contributions of this paper are summarized as follows.

- We discuss and compare two uncoded schemes for DHT. In the first scheme, called the truncation scheme, the encoders transmit only the first $l < n$ bits of their source sequences to the receiver (Section III-A). In the second scheme, called the separate scheme, each encoder independently makes a local decision based on its observation and transmits a single bit to the receiver (Section III-B). While these schemes are not novel, they provide reference points for evaluating the proposed quantizer and quantize-binning schemes. They also allow us to compare decide-and-compress versus compress-and-decide strategies, in line with previous information-theoretic works that investigated estimate-and-compress versus compress-and-estimate setups [34], [35].
- For both symmetric and asymmetric setups, we introduce quantizer-alone (Section IV) and quantize-binning schemes (Section V) for DHT, constructed with short-length linear block codes ($n < 100$). Simulation results show that the proposed constructions outperform solutions initially proposed for long block lengths in [24], [31], [32]. The simulation results also demonstrate the superiority of the quantize-binning scheme over the quantizer-alone and uncoded truncation schemes, particularly when code parameters are optimized.

- We derive exact analytical expressions for the Type-I and Type-II error probabilities of the quantizer-alone and quantize-binning schemes in the asymmetric setup (Propositions 1 and 2). Numerical results validate the accuracy of the analytical error probability expressions by comparison with Monte Carlo simulations.

D. Outline

The remainder of this paper is organized as follows. Section II presents the DHT setup. Section III introduces the uncoded schemes. Section IV presents the quantization scheme and provides the analytical expressions for Type-I and Type-II error probabilities. Section V describes the quantize-binning scheme and provides the analytical expressions for Type-I and Type-II error probabilities. Section VI presents the numerical results.

II. Distributed Hypothesis Testing

This section introduces the considered DHT setup and presents existing information-theoretic coding schemes for this problem.

A. Notation

Let $[\![1, M]\!]$ denote the set of integers from 1 to M. Random variables are represented by uppercase letters, *e.g.*, X, while their realizations are in lowercase, *e.g.*, x. Boldface letters, *e.g.*, $\mathbf{X}^n$ denote vectors of length n. The Hamming weight of a vector $\mathbf{x}^n$ is denoted as $w(\mathbf{x}^n)$, and the Hamming distance between two vectors $\mathbf{x}^n$ and $\mathbf{y}^n$ is $d(\mathbf{x}^n, \mathbf{y}^n)$. The binomial coefficient of two integers n, k, with $k \leq n$, is expressed as $\binom{n}{k}$.

B. Hypothesis Testing with binary sources

We consider two source vectors $\mathbf{X}^n$ and $\mathbf{Y}^n$ of length n. As in the conventional DHT setup [10], we assume that the components of $\mathbf{X}^n$ and $\mathbf{Y}^n$ are i.i.d., drawn according to random variables X and Y, respectively. The pair (X, Y) follows one of two possible joint distributions:

$$\begin{aligned} \mathcal{H}_0 &: (X, Y) \sim \mathbb{P}_0, \\ \mathcal{H}_1 &: (X, Y) \sim \mathbb{P}_1. \end{aligned} \tag{1}$$

In what follows, with a slight abuse of notation, we always denote $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) the random variable or vector probability distribution under $\mathcal{H}_0$ (resp. $\mathcal{H}_1$). For instance, $\mathbb{P}_0(\mathbf{x}^n)$ is the probability of vector $\mathbf{x}^n$ under $\mathcal{H}_0$. We consider a general case where the marginal distributions of X under $\mathcal{H}_0$ and $\mathcal{H}_1$ are not necessarily identical. Therefore, the marginal distributions of Y under $\mathcal{H}_0$ and $\mathcal{H}_1$ are not necessarily identical either.

In this work, we focus on binary sources, where X and Y take values in the alphabet $\{0, 1\}$ and follow the model $Y = X \oplus Z$, with Z being a binary random variable independent of X. Let $p = \mathbb{P}(X = 1)$ and $c = \mathbb{P}(Z = 1)$, where $0 < c \leq 1/2$. The hypotheses provided in (1) can then be expressed as:

$$\begin{aligned} \mathcal{H}_0 &: (p = p_0, c = c_0) \\ \mathcal{H}_1 &: (p = p_1, c = c_1). \end{aligned} \tag{2}$$

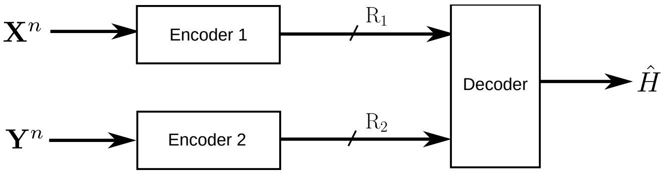

Fig. 1: Distributed hypothesis testing scheme.

For convenience, we assume that $0 < p_0 \leq 0.5$, $p_0 \leq p_1$, and $c_0 \leq c_1$. This model has been studied from an information-theoretic perspective in [23], [36]. Notably, when $p_0 = p_1$ and $c_1 = 1/2$, the problem reduces to testing against independence [9].

C. Coding schemes

Figure 1 illustrates the DHT setup. Encoder 1 and Encoder 2 send coded representations of $\mathbf{X}^n$ and $\mathbf{Y}^n$ at rates R_1 and R_2, respectively. The decoder uses the received coded information to make a decision between $\mathcal{H}_0$ and $\mathcal{H}_1$. In this paper, two setups are considered:

- *Asymmetric setup*: $\mathbf{X}^n$ is encoded at rate R_1, while $\mathbf{Y}^n$ is fully available at the decoder.
- *Symmetric setup*: both $\mathbf{X}^n$ and $\mathbf{Y}^n$ are encoded at rates R_1 and R_2, respectively.

For a given block length n, the encoding functions are defined as:

$$f_1^{(n)} : \{0, 1\}^n \to [\![1, 2^{nR_1}]\!], \tag{3}$$

$$f_2^{(n)} : \{0, 1\}^n \to [\![1, 2^{nR_2}]\!], \tag{4}$$

and the decision function as:

$$g^{(n)} : [\![1, 2^{nR_1}]\!] \times [\![1, 2^{nR_2}]\!] \to \{0, 1\}. \tag{5}$$

For given encoding and decision functions $(f_1^{(n)}, f_2^{(n)}, g^{(n)})$, we define Type-I error probability α_n and Type-II error probability β_n as [23]

$$\alpha_n = \mathbb{P}_0\left(g^{(n)}(f_1^{(n)}(\mathbf{X}^n), f_2^{(n)}(\mathbf{Y}^n)) = 1\right), \tag{6}$$

$$\beta_n = \mathbb{P}_1\left(g^{(n)}(f_1^{(n)}(\mathbf{X}^n), f_2^{(n)}(\mathbf{Y}^n)) = 0\right). \tag{7}$$

In the information-theoretic analysis, it is shown that the Type-II error probability decays exponentially fast with the blocklength n, characterized by an error exponent θ. This regime is commonly referred to as the Stein regime [23]. For a given value $\epsilon \in (0, 1)$ such that $\alpha_n < \epsilon$, the error exponent θ is defined as [23]

$$\lim_{n\to\infty} \sup \frac{1}{n} \log \frac{1}{\beta_n} \geq \theta. \tag{8}$$

D. Short-length nature of DHT

Existing lower bounds on the error exponent θ are derived in the asymptotic regime, *i.e.*, by considering $n \to \infty$ [12], [23], [37], [38]. Therefore, they provide a scaling law $\beta_n \approx e^{-n\theta}$, which characterizes the exponential decay of the Type-II error probability. Here, we use this approximation to investigate the block length n needed to achieve sufficiently low Type-II error probability. Among existing lower bounds on the error-exponent for the asymmetric setup, the one provided in [36] is based on a quantize-binning achievable coding scheme and is simpler to evaluate than the one of the SHA scheme in [12]. Specifying this bound for the DHT problem with binary sources defined in (2) leads to

$$\theta \geq \sup_{\delta \in [0,1]} \min \Bigg\{ R_1 - [H_2(p_0 * \delta) - H_2(\delta)], \tag{9}$$
$$(p_0 * \delta) \log \frac{p_0 * \delta}{p_1 * \delta} + (1 - (p_0 * \delta)) \log \frac{1 - (p_0 * \delta)}{1 - (p_1 * \delta)} \Bigg\}.$$

Here, H_2 is the binary entropy function, and $*$ is the binary convolution operator defined as $x * y = (1-x)y + (1-y)x$, with $0 \leq x, y \leq 1$. In addition, δ is a parameter of the information-theoretic coding scheme. Consider parameters $p_0 = 0.05$, $p_1 = 0.5$, $\delta = 0.1$, $r = 0.4$. We observe that β_n drops from 10^{-6} with $n = 50$ to 10^{-12} with $n = 100$. This highlights the importance of focusing on small values of n ($n < 100$), motivating our study of short-length coding schemes. Note that the approximation $\beta_n \approx e^{-n\theta}$ provides a lower bound on the DHT performance for a given n, but θ was obtained under the conditions that (i) $n \to \infty$, (ii) Type-I error probability α_n also converges to 0 as $n \to \infty$. In this paper, we will provide more accurate analytical expressions of the Type-I and Type-II error probabilities for finite n, for the proposed practical coding schemes.

E. Information-theoretic coding scheme

We now review the SHA quantize-binning achievable scheme which have been proposed in the literature of DHT for the asymmetric setup [12]. This will allow us to identify the main steps of a practical coding for this problem. For brevity, we do not provide here the corresponding lower bound on the error exponent. Indeed, the bound has a quite complex expression and is not essential to convey the main message of this part. The quantize-binning scheme of [12], [37] operates as follows.

- **Codebook generation**: Construct a random codebook
$$\mathcal{C} = \{\mathbf{u}^n(m, \ell) \colon m \in [\![1, \lfloor 2^{nR_1} \rfloor]\!], \ell \in [\![1, \lfloor 2^{nr_1} \rfloor]\!]\} \tag{10}$$
by drawing all entries of all codewords i.i.d. according to a chosen distribution $P_{U|X}$
- **Encoding**: To encode $\mathbf{x}^n$, the encoder looks for indices (m, ℓ) such that $\mathbf{u}^n(m, \ell)$ and $\mathbf{x}^n$ are jointly typical [12], in the sense that $\forall (a, b) \in \{0, 1\}$,
$$\left| \frac{\mathbb{N}((a,b)}{n} |(\mathbf{x}^n, \mathbf{u}^n(m, \ell))) - P_{XU}(a, b) \right| \leq \mu P_{XU}(a, b), \tag{11}$$
where μ is a parameter and $\mathbb{N}((a,b)|(\mathbf{x}^n, \mathbf{u}^n)$ is the number of occurrences of the pair (a, b) in the pair of sequences $(\mathbf{x}^n, \mathbf{u}^n)$. The codeword $\mathbf{u}^n$ can be interpreted as a quantized version of $\mathbf{x}^n$. If successful, the encoder picks one of these indices uniformly at random and sends the index m to the receiver. Otherwise, it sends $m = 0$.
- **Decoder**: The decoder receives the message m from the encoder, and it observes $\mathbf{y}^n$. If $m = 0$, the decoder declares $\hat{\mathcal{H}} = \mathcal{H}_1$. Otherwise, it first performs an empirical entropy decoding step [12] to identify the most likely sequence within the received bin, and then decides between the two hypotheses using a typicality check. Specifically, the decoder searches for an index $\hat{\ell} \in [\![1, \lfloor 2^{nr_1} \rfloor]\!]$ such that $\mathbf{u}^n(m, \hat{\ell})$ has the lowest empirical entropy $H_e(\mathbf{u}^n|\mathbf{y}^n)$ defined as
$$H_e(\mathbf{u}^n|\mathbf{y}^n) = -\frac{1}{n} \sum_{k=1}^{n} \log P_{U|Y}(u_k|y_k), \tag{12}$$
where $P_{U|Y}$ is the conditional probability distribution of U given Y under $\mathcal{H}_0$. The receiver then declares $\mathcal{H}_0$ if the extracted sequence $\mathbf{u}^n(m, \hat{\ell})$ and the side information $\mathbf{y}^n$ are jointly typical. in the sense of (11), where $\mathbf{x}^n$ is replaced by $\mathbf{y}^n$.

The information-theoretic coding scheme described above relies on quantization and binning steps, which we aim to implement in a practical form. In this work, we will rely on linear block codes for both steps. In addition, the previous criteria of joint typicality and empirical entropy in (11) and (12) allow some error probability terms to vanish asymptotically in the information-theoretic proof. In this work, we consider the finite-length regime. Consequently, we adopt the Maximum Likelihood (ML) estimator and the Neyman-Pearson (NP) test, which are standard techniques in signal processing and are known to be optimal under specific conditions, as detailed later in the paper.

III. Uncoded schemes

Before introducing the proposed practical quantization and quantize-binning schemes, in this section we describe two schemes that do not require coding. These schemes will serve as baselines when evaluating the performance of our proposed coding schemes. In this section, we also introduce the key elements of the NP theory which will be essential in our practical coding schemes.

A. Truncation scheme

The truncation scheme consists of sending the first $l \leq n$ symbols of the source vector $\mathbf{x}^n$ and $\mathbf{y}^n$ at the coding rate $R_1 = R_2 = l/n$ at the decoder. In this part, we consider equal rates for simplicity. Indeed, considering $R_1 \neq R_2$ in the truncation scheme would result in sending a different number of symbols from $\mathbf{x}^n$ and $\mathbf{y}^n$, which may not be efficiently exploited by the NP test since the sources are i.i.d. Note that in the upcoming quantization and quantize-binning schemes, we will consider the general case $R_1 \neq R_2$.

The decoder can then perform a standard NP test on the pair $(\mathbf{x}^l, \mathbf{y}^l)$. Under the constraint $\alpha_n^{(t)} < \epsilon$ on Type-I error probability for the truncation scheme, the NP lemma [39] provides an optimal decision rule. Specifically, for a given value $\mu \in \mathbb{R}$, the following NP test:

$$\mathbb{P}_1(\mathbf{x}^l, \mathbf{y}^l) < \mu \mathbb{P}_0(\mathbf{x}^l, \mathbf{y}^l), \tag{13}$$

where $\mathcal{H}_0$ is decided if the inequality is satisfied and $\mathcal{H}_1$ is decided otherwise, minimizes the Type-II error probability $\beta_n^{(t)}$ under the constraint $\alpha_n^{(t)} < \epsilon$. In (13), the threshold value μ is chosen to satisfy this Type-I error constraint. Given that $p_0 \leq p_1$ and $c_0 \leq c_1$, by expressing the joint probabilities $\mathbb{P}_0(\mathbf{x}^l, \mathbf{y}^l)$ and $\mathbb{P}_1(\mathbf{x}^l, \mathbf{y}^l)$ under each hypothesis and taking the logarithm in both sides of (13), the Neyman–Pearson test (13) simplifies to

$$w(\mathbf{x}^l) \log \frac{p_1(1-p_0)}{p_0(1-p_1)} + w(\mathbf{z}^l) \log \frac{c_1(1-c_0)}{c_0(1-c_1)} < \tau_t, \quad (14)$$

where $\mathbf{z}^l = \mathbf{x}^l \oplus \mathbf{y}^l$, and τ_t is a threshold value chosen so as to satisfy the Type-I error constraint. The threshold τ_t in (14) can be expressed from the threshold μ in (13). However, in NP tests, this expression is in general not provided nor used: the value of τ_t is directly set so as to satisfy the Type-I error probability constraint $\alpha_n^{(t)} < \epsilon$ from the test (14), without need to come back to (13).

In the NP theory, the threshold value τ_t is often set from analytical expressions of Type-I and Type-II error probabilities of the considered test. For the truncation scheme, given that $p_0 < p_1$ and $c_0 < c_1$, the analytical expressions of Type-I and Type-II errors are given by

$$\alpha_n^{(t)} = \sum_{\substack{(\lambda,j):\\ T_{\lambda,j} \geq \tau_t}} \binom{l}{\lambda} p_0^{\lambda}(1-p_0)^{l-\lambda} \binom{l}{j} c_0^{j}(1-c_0)^{l-j} \quad (15)$$

$$\beta_n^{(t)} = \sum_{\substack{(\lambda,j):\\ T_{\lambda,j} \leq \tau_t}} \binom{l}{\lambda} p_1^{\lambda}(1-p_1)^{l-\lambda} \binom{l}{j} c_1^{j}(1-c_1)^{l-j} \quad (16)$$

where the exponent (t) refers to the truncation scheme, and $T_{\lambda,j} = \lambda \log_2 \frac{p_1(1-p_0)}{p_0(1-p_1)} + j \log_2 \frac{c_1(1-c_0)}{c_0(1-c_1)}$.

In all the coding schemes introduced in the paper, we will use the NP lemma to derive optimal tests that minimize Type-II error probability under a constraint on Type-I error probability. We will also aim to derive analytical expressions of Type-I and Type-II error probability.

B. *Separate scheme (local decisions)*

When the marginal distributions of X and Y depend on the hypothesis $\mathcal{H}_0$ or $\mathcal{H}_1$ (that is $p_0 \neq p_1$), each encoder can locally perform a Neyman–Pearson (NP) test [39] based on its own observation $\mathbf{x}^n$, or $\mathbf{y}^n$. In this case, each encoder sends one bit indicating its local decision, leading to coding rates $R_1 = R_2 = 1/n$. For given threshold values $\mu_1, \mu_2 \in \mathbb{R}$ chosen to satisfy the constraints $\alpha_n^{(1)} < \epsilon$ and $\alpha_n^{(2)} < \epsilon$ on the Type-I error probabilities for Encoder 1 and Encoder 2, respectively, we consider the following NP tests:

$$\mathbb{P}_1(\mathbf{x}^n) < \mu_1 \mathbb{P}_0(\mathbf{x}^n), \quad (17)$$

$$\mathbb{P}_1(\mathbf{y}^n) < \mu_2 \mathbb{P}_0(\mathbf{y}^n), \quad (18)$$

where in each case $\mathcal{H}_0$ is decided if the inequality is satisfied. Given that $p_0 \leq p_1$, and $c_0 \leq c_1$, the tests described by (17) and (18) are equivalent, respectively, to the conditions:

$$w(\mathbf{x}^n) < \lambda_1, \quad (19)$$

$$w(\mathbf{y}^n) < \lambda_2, \quad (20)$$

where $\lambda_1, \lambda_2 \in \mathbb{N}$ are threshold values.

The decoder decides hypothesis $\mathcal{H}_i$ if both encoders agree on i, and otherwise follows Encoder 1's decision. This choice is motivated by the fact that according to the model defined in (2), Y is a noisy version of X, so that $\mathbb{P}(Y = 1) > \mathbb{P}(X = 1)$ under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Note that other strategies may be considered depending on the values of p_0, p_1, c_0, c_1. The Type-I and Type-II error probabilities of the separate scheme can be derived similarly to the truncation scheme by following the standard NP arguments. This setup offers the advantage of achieving very low communication rates by transmitting just one bit of information to the decoder. This constitutes a special case of the estimate-and-compress setup introduced in [34] in the context of parameter estimation.

IV. Quantization scheme

In their seminal work [9], Ahlswede and Csiszár introduced the first information-theoretic DHT scheme based only on a quantizer. Here, we present a practical implementation of this scheme for short-length sequences by utilizing linear block codes.

A. *Code construction for the symmetric setup*

First, in this section, we consider the special case where $R_1 = R_2$. To practically implement binary quantization for the symmetric setup, we follow the approach of [24], [29], [32] and consider a generator matrix G_q with dimension $n \times k$ of a linear block code. According to the ML rule, for given source sequences $\mathbf{x}^n$ and $\mathbf{y}^n$, the encoders produce vectors $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ of length k bits as [40]

$$\mathbf{u}_q^k = \arg \min_{\mathbf{u}^k \in \{0,1\}^k} d\left(G_q \mathbf{u}^k, \mathbf{x}^n\right), \quad (21)$$

$$\mathbf{t}_q^k = \arg \min_{\mathbf{t}^k \in \{0,1\}^k} d\left(G_q \mathbf{t}^k, \mathbf{y}^n\right). \quad (22)$$

The vectors $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ are the compressed versions of $\mathbf{x}^n$ and $\mathbf{y}^n$, respectively. We further denote $\mathbf{x}_q^n = G_q \mathbf{u}_q^k$, and $\mathbf{y}_q^n = G_q \mathbf{t}_q^k$.

In (21), and (22), the key difficulty lies in determining the quantized vectors $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ that achieve the minimum Hamming distance. In [24], [29], [32], the matrix G_q is constructed as an LDGM code, which enables the use of a low complexity message-passing algorithm called BiP to solve (21), (22). The schemes introduced in [24], [29], [32] consider long codes (more than 10^3 bits). But at short length, LDGM codes are penalized by their low minimum distances between codewords. Instead, we opt to consider well-known short linear block codes such as BCH, Reed-Muller, Polar codes, since they have good minimum distance properties. Unfortunately, their generator matrices are not sparse, which prevents the use of the BiP algorithm.

To find $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ we reformulate problems (21) and (22) as

$$\mathbf{x}_q^n = \arg\min_{\mathbf{x}_q^n} d(\mathbf{x}_q^n, \mathbf{x}^n) \text{ s.t. } H_q \mathbf{x}_q^n = \mathbf{0}^m, \quad (23)$$

$$\mathbf{y}_q^n = \arg\min_{\mathbf{y}_q^n} d(\mathbf{y}_q^n, \mathbf{y}^n) \text{ s.t. } H_q \mathbf{y}_q^n = \mathbf{0}^m, \quad (24)$$

where H_q is a parity check matrix of size $m \times n$ of the code defined by G_q, with $m = n - k$. Once the closest valid codewords $\mathbf{x}_q^n$ and $\mathbf{y}_q^n$ are identified, the corresponding compressed vectors $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ can be retrieved by solving the linear systems $G_q \mathbf{u}_q^k = \mathbf{x}_q^n$ and $G_q \mathbf{t}_q^k = \mathbf{y}_q^n$, typically via Gaussian elimination.

Problems (23) and (24), which are equivalent to (21) and (22), respectively, are a standard channel coding formulation. In this sense, a natural approach might be to apply Belief Propagation (BP) decoding [41] to (23) and (24), as commonly done for LDPC codes in channel coding. However, BP decoders perform poorly for binary quantization, especially at short block lengths. This arises because, unlike in channel decoding where the received vector is a noisy version of a transmitted codeword, in binary quantization the source vectors $\mathbf{x}^n$ and $\mathbf{y}^n$ are typically far from any valid codeword. Consequently, BP often fails to converge to a valid solution in this context.

In this work, we do not rely on BP decoders, which are not well-suited to short codes or to binary quantization. Instead, we propose solving problems (23) and (24) exactly using Maximum Likelihood (ML) decoders tailored to the considered codes. Specifically, we use generic syndrome-based decoders with complexity in $O(2^m)$ [42, Section 3.1.4], which can be applied to a broad range of short block codes, including BCH and Reed-Muller codes. When applicable, code-specific decoders such as the Berlekamp-Massey algorithm for BCH codes [42, Section 3.2.2], or generic near-ML decoders such as Linear-Programming (LP) decoders [43], can also be employed to further reduce complexity see Figures 5 and 8).

B. Hypothesis test for the symmetric setup

The codewords $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$ are transmitted to the decoder at code rates $R_1 = R_2 = k/n$. The decoder first reconstructs the quantized vectors $\mathbf{x}_q^n = G_q \mathbf{u}_q^k$ and $\mathbf{y}_q^n = G_q \mathbf{t}_q^k$. It then applies a NP test to the pair $(\mathbf{x}_q^n, \mathbf{y}_q^n)$. For a given threshold value $\mu_q \in \mathbb{R}$ set under the constraint $\alpha_n^{(q)} < \epsilon$ on Type-I error probability for the quantization scheme, the NP test can be written as

$$\mathbb{P}_1\left(\mathbf{x}_q^n, \mathbf{y}_q^n\right) \leq \mu_q \mathbb{P}_0\left(\mathbf{x}_q^n, \mathbf{y}_q^n\right), \quad (25)$$

where $\mathcal{H}_0$ is decided if the inequality is satisfied. Moreover, while condition (12) arises in an asymptotic setting, the NP test remains optimal for any finite n.

However, computing the joint distributions $\mathbb{P}_0(\mathbf{x}_q^n, \mathbf{y}_q^n)$ and $\mathbb{P}_1(\mathbf{x}_q^n, \mathbf{y}_q^n)$ in (25) is non-trivial, as the underlying code G_q introduces statistical dependencies within the vectors $\mathbf{x}_q^n$ and $\mathbf{y}_q^n$. To simplify the analysis, we model the vectors $\mathbf{x}_q^n$ and $\mathbf{v}_q^n = \mathbf{x}_q^n \oplus \mathbf{y}_q^n$ as realizations of i.i.d. Bernoulli random variables: $X_q \sim \text{Bern}(\hat{p})$ with $\hat{p} = \hat{p}_0$ under $\mathcal{H}_0$ and $\hat{p} = \hat{p}_1$ under $\mathcal{H}_1$, and $V_q \sim \text{Bern}(\hat{c})$ with $\hat{c} = \hat{c}_0$ under $\mathcal{H}_0$ and $\hat{c} = \hat{c}_1$ under $\mathcal{H}_1$.

Under these assumptions, the test (25) can be reformulated as

$$w(\mathbf{x}_q^n) \log_2 \frac{\hat{p}_1(1-\hat{p}_0)}{\hat{p}_0(1-\hat{p}_1)} + w(\mathbf{v}_q^n) \log_2 \frac{\hat{c}_1(1-\hat{c}_0)}{\hat{c}_0(1-\hat{c}_1)} \leq \tau_q, \quad (26)$$

where τ_q is the test threshold chosen to meet the Type-I error probability requirement. The parameters $(\hat{p}_0, \hat{c}_0)$ and $(\hat{p}_1, \hat{c}_1)$ are estimated through Monte Carlo simulations. In practice, the numerical results presented later on in the paper demonstrate that the proposed test provides accurate decisions under this approximation (Figures 4 and Figure 5).

Remark: Note that when $p_0 \neq p_1$, the NP test in (26) cannot be simplified further. In this case, deriving closed-form expressions of the Type-I and Type-II error probabilities is very difficult, due to the log coefficients. Therefore, in the theoretical analysis, we focus on symmetric source models ($p_0 = p_1$), for which the test (26) reduces to comparing $w(\mathbf{v}_q^n)$ to a threshold value.

Finally, note that the coding scheme and the hypothesis test introduced in this section can be straightforwardly generalized to the case where $R_1 \neq R_2$, by considering two generator matrices $G_q^{(1)}$, $G_q^{(2)}$ along with their parity check matrices $H_q^{(1)}$, $H_q^{(2)}$. Therefore, the only formula that need to be changed are the quantization rules (21) and (22). However the hypothesis test itself does not change since it is applied over vectors $\mathbf{x}_q^n$ and $\mathbf{y}_q^n$ of the same length.

C. Code construction and hypothesis test for the asymmetric setup

In the asymmetric setup, the coding scheme is obtained in a straightforward manner from the symmetric setup. In this case, only $\mathbf{x}^n$ is quantized and transmitted at rate $R_1 = k/n$, while $\mathbf{y}^n$ is available at the decoder. Therefore, the decoder first computes the quantized vector $\mathbf{x}_q^n = G_q \mathbf{u}_q^k$. Then the NP test remains the same as (26), except that $\mathbf{y}_q^n$ is replaced by $\mathbf{y}^n$ in the expression of $\mathbf{v}_q^n$. Note that in the special case where $p_0 = p_1$, the NP test (25) reduces to

$$\sum_{i=1}^{n} (x_{q,i} \oplus y_i) < \lambda_q, \quad (27)$$

where λ_q is the threshold chosen so as to satisfy the Type-I error probability constraint.

D. Theoretical analysis of the quantization scheme

In this section, we provide a theoretical analysis of the Type-I and Type-II error probabilities for the proposed practical quantization scheme under the asymmetric setup, focusing on the special case where $p_0 = p_1 = 1/2$. We derive exact analytical expressions for the error probabilities of the scheme considering the generator matrix G_q. Extensions to the symmetric setup and to cases where $p_0 \neq p_1$ involve significantly more complex analyses and are therefore left for future work.

We begin by introducing the notation associated with the code defined by the generator matrix G_q. Throughout the analysis, we assume that the all-zero codeword $\mathbf{x}_q^n$ is transmitted. Due to symmetry, the quantization error probability is independent of the transmitted codeword [41].

Let $\mathcal{C}_0^{(q)}$ denote the decision region, or coset, corresponding to $\mathbf{x}_q^n = \mathbf{0}^n$, defined as

$$\mathcal{C}_0^{(q)} := \left\{ \mathbf{x}^n \in \{0,1\}^n : \arg\min_{\mathbf{u}^k} d(G_q \mathbf{u}^k, \mathbf{x}^n) = \mathbf{0}^k \right\}.$$

In other words, $\mathbf{x}^n \in \mathcal{C}_0^{(q)}$ implies that the solution of (23) for $\mathbf{x}_q^n$ is $\mathbf{0}^n$. Next consider the set of integers $\{E_\gamma^{(q)}\}_{\gamma \in [\![0, d_{\max}^{(q)}]\!]}$ where $E_\gamma^{(q)}$ is the number of sequences $\mathbf{x}^n$ of Hamming weight γ, also referred to as number of coset leaders of weight γ, that belong to the coset $\mathcal{C}_0^{(q)}$. In addition, $d_{\max}^{(q)}$ is the maximum possible weight in $\mathcal{C}_0^{(q)}$. The total number of coset leaders is then given by $N_0^{(q)} = \sum_{\gamma=0}^{d_{\max}^{(q)}} E_\gamma^{(q)}$. The reader is referred to [42, Section 3.1.4] for more details about these concepts which were originally introduced in the context of channel coding.

With these definitions in place, we proceed to derive the theoretical expressions for the Type-I and Type-II error probabilities of the proposed quantization scheme under the asymmetric setup.

Proposition 1: Consider the quantization scheme in the asymmetric setup, with $p_0 = p_1 = 0.5$, and a fixed threshold value λ_q. Type-I and Type-II error probabilities of this scheme are given by

$$\alpha_n^{(q)} = 1 - \frac{1}{N_0^{(q)}} \sum_{\lambda=0}^{\lambda_q} \sum_{\gamma=0}^{d_{\max}^{(q)}} \sum_{j=0}^{n} E_\gamma^{(q)} \Delta_{\lambda,j,\gamma} \binom{n}{j} c_0^j (1-c_0)^{n-j}, \tag{28}$$

$$\beta_n^{(q)} = \frac{1}{N_0^{(q)}} \sum_{\lambda=0}^{\lambda_q} \sum_{\gamma=0}^{d_{\max}^{(q)}} \sum_{j=0}^{n} E_\gamma^{(q)} \Delta_{\lambda,j,\gamma} \binom{n}{j} c_1^j (1-c_1)^{n-j}, \tag{29}$$

where

$$\Delta_{\lambda,j,\gamma} = \frac{\Gamma_{\lambda,j,\gamma}}{\sum_{i=0}^{\max(\lambda,j)} \binom{\lambda}{i}\binom{n-\lambda}{j-i}}, \tag{30}$$

and, for $\gamma = j + \lambda - 2u$ and $0 \leq u \leq \min(\lambda, j) \leq n$,

$$\Gamma_{\lambda,j,\gamma} = \binom{\lambda}{u}\binom{n-\lambda}{j-u}. \tag{31}$$

The terms $\Delta_{\lambda,j,\gamma}$ in (28) and (29) represent the probabilities of binary sequences $\mathbf{x}^n$ and $\mathbf{y}^n$ of Hamming weights λ and γ, respectively, with Hamming distance j between $\mathbf{x}^n$ and $\mathbf{y}^n$. These terms appear due to the fact that a decision error can occur if the side information vector $\mathbf{y}^n$ is too far from $\mathbf{x}_q^n$, but $\mathbf{y}^n$ is generated from a conditional distribution $P(\mathbf{y}^n|\mathbf{x}^n)$, and the distribution $P(\mathbf{x}^n|\mathbf{x}_q^n)$ is not i.i.d..

Proof: By symmetry due to the linear block code, the quantizer error probability is independent of the transmitted codeword [41]. Therefore, it is sufficient to consider the all-zero codeword $\mathbf{x}_q^n = \mathbf{0}^n$. From (27), we develop

$$\begin{aligned}
\alpha_n^{(q)} &= 1 - \sum_{\lambda=0}^{\lambda_q} \mathbb{P}_0(w(\mathbf{Y}^n) = \lambda) \qquad (32)\\
&= 1 - \sum_{\lambda=0}^{\lambda_q} \sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \mathbb{P}_0(w(\mathbf{Y}^n) = \lambda | w(\mathbf{X}^n) = \gamma)\\
&= 1 - \sum_{\lambda=0}^{\lambda_q} \sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \sum_{j=0}^{n} \mathbb{P}_0(d(\mathbf{X}^n, \mathbf{Y}^n) = j) \Delta_{\lambda,j,\gamma}\\
&= 1 - \sum_{\lambda=0}^{\lambda_q} \sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \sum_{j=0}^{n} \binom{n}{j} c_0^j (1-c_0)^{n-j} \Delta_{\lambda,j,\gamma}.
\end{aligned}$$

This leads to (28). To obtain (29), we remark that $\beta_n^{(q)} = \sum_{\lambda=0}^{\lambda_q} \mathbb{P}_1(w(\mathbf{Y}^n) = \lambda)$. Following the same steps as in (32), and by replacing c_0 by c_1, the proof is completed. □

To the best of the authors' knowledge, these theoretical results are novel and differ from both the classical information-theoretic analysis of DHT and existing results in channel coding. Especially, while error probability expressions are well-established for linear block codes in the context of channel coding, such analytical characterizations had not been derived for the DHT problem, where the threshold parameter λ_q impacts both Type-I and Type-II error probabilities. These new analytical expressions enable the evaluation of decision performance without relying on computationally intensive Monte Carlo simulations. Consequently, they facilitate code design by only considering the code parameters such as $E_\gamma^{(q)}, N_0^{(q)}$, and $d_{\max}^{(q)}$, as demonstrated in [44].

V. Quantize-binning scheme

In their seminal work [12], Shimokawa et al. introduced the quantize-binning scheme for DHT. This scheme leverages the correlation between the sources X and Y to reduce the code rate after compression. We now introduce a practical short-length implementation of this scheme by using linear block codes.

A. Code construction for the symmetric setup

In this part as well, we first assume that $R_1 = R_2$ for simplicity. To practically implement the quantize-binning scheme, we again consider the generator matrix G_q of size $n \times k$ of a linear block code. Additionally, we consider the parity-check matrix H_b of size $\ell \times k$ of *another* linear block code. In the symmetric setup, given the source vectors $\mathbf{x}^n$ and $\mathbf{y}^n$, the encoders apply the quantization method described by (21) and (22) to obtain the sequences $\mathbf{u}_q^k$ and $\mathbf{t}_q^k$, respectively. Then, the encoders use the parity check matrix H_b to compute the syndromes

$$\mathbf{r}^\ell = H_b \mathbf{u}_q^k, \tag{33}$$

and

$$\mathbf{s}^\ell = H_b \mathbf{t}_q^k, \tag{34}$$

both of length ℓ. This follows the same approach as in [25]–[28], where binning is performed using the parity-check matrix of an LDPC code. In contrast, and consistent with our earlier discussion, we employ efficient short-length linear block codes such as BCH, Polar codes, and Reed–Muller codes.

B. Hypothesis test for the symmetric setup

The syndromes $\mathbf{r}^\ell$ and $\mathbf{s}^\ell$ are then transmitted to the decoder at rates $R_1 = R_2 = \ell/n$. At the decoder, as discussed in Section IV, we avoid using message-passing algorithms such as BP decoders since they do not perform well with short-length codes. Instead, we directly consider the ML rule. Therefore, the decoder first identifies by exhaustive search, vectors $\hat{\mathbf{u}}_q^k$ and $\hat{\mathbf{t}}_q^k$ as

$$(\hat{\mathbf{u}}_q^k, \hat{\mathbf{t}}_q^k) = \arg\min_{\mathbf{u}^k,\mathbf{t}^k} d\left(G_q\mathbf{u}^k, G_q\mathbf{t}^k\right) \text{ s.t.} \qquad (35)$$
$$H_b\mathbf{u}^k = \mathbf{r}^\ell, \text{ and } H_b\mathbf{t}^k = \mathbf{s}^\ell.$$

We then compute $\hat{\mathbf{x}}_q^n = G_q\hat{\mathbf{u}}_q^k$ and $\hat{\mathbf{y}}_q^n = G_q\hat{\mathbf{t}}_q^k$, and, apply the following NP test for a given threshold $\mu_{q,b} \in \mathbb{R}$,

$$\mathbb{P}_1\left(\hat{\mathbf{x}}_q^n, \hat{\mathbf{y}}_q^n\right) \le \mu_{q,b}\mathbb{P}_0\left(\hat{\mathbf{x}}_q^n, \hat{\mathbf{y}}_q^n\right), \qquad (36)$$

where $\mathcal{H}_0$ is decided if the inequality is satisfied.

As in Section IV-A, we note that computing the joint distributions $\mathbb{P}_0(\hat{\mathbf{x}}_q^n, \hat{\mathbf{y}}_q^n)$ and $\mathbb{P}_1(\hat{\mathbf{x}}_q^n, \hat{\mathbf{y}}_q^n)$ in (36) is difficult. Therefore, we adopt the same assumptions as in Section IV-A. Specifically, we assume that $\hat{\mathbf{x}}_q^n$ and $\hat{\mathbf{v}}_q^n = \hat{\mathbf{x}}_q^n \oplus \hat{\mathbf{y}}_q^n$ are the realizations of i.i.d. random variables $\hat{X}_q \sim \text{Bern}(\hat{p}_b)$ (with $\hat{p}_b = \hat{p}_{0,b}$ under $\mathcal{H}_0$, and $\hat{p}_b = \hat{p}_{1,b}$, under $\mathcal{H}_1$), and $\hat{V}_q \sim \text{Bern}(\hat{c}_b)$ (with $\hat{c}_b = \hat{c}_{0,b}$ under $\mathcal{H}_0$, and $\hat{c}_b = \hat{c}_{1,b}$, under $\mathcal{H}_1$). The NP test (36) can then be rewritten as

$$w(\hat{\mathbf{x}}_q^n)\log_2\frac{\hat{p}_{1,b}(1-\hat{p}_{0,b})}{\hat{p}_{0,b}(1-\hat{p}_{1,b})} + w(\hat{\mathbf{v}}_q^n)\log_2\frac{\hat{c}_{0,b}(1-\hat{c}_{0,b})}{\hat{c}_{1,b}(1-\hat{c}_{1,b})} \le \tau_{q,b}, \qquad (37)$$

where $\tau_{q,b}$ is the test threshold chosen so as to satisfy the Type-I error probability. The values $\hat{p}_{0,b}$, $\hat{c}_{0,b}$, $\hat{p}_{1,b}$, and $\hat{c}_{1,b}$ are estimated through Monte Carlo simulations.

Remark: Note that the quantize-binning scheme can be straightforwardly generalized to the case where $R_1 \neq R_2$, by considering two different generator matrices $G_q^{(1)}$, $G_q^{(2)}$ in (21) and (22) and two different parity check matrices $H_b^{(1)}$, $H_b^{(2)}$ in (33) and (34).

C. Code construction and hypothesis test for the asymmetric setup

In the asymmetric case, only $\mathbf{x}^n$ is quantized and binned into $\mathbf{r}^\ell$, according to (33). The vector $\mathbf{r}^\ell$ is then transmitted at rate $R_1 = \ell/n$, while $\mathbf{y}^n$ serves as side information at the decoder. At the receiver, the vector $\hat{\mathbf{u}}^k$ is identified by solving the ML rule

$$\hat{\mathbf{u}}_q^k = \arg\min_{\mathbf{u}^k} \; d(G_q\mathbf{u}^k, \mathbf{y}^n) \text{ s.t. } H_b\mathbf{u}^k = \mathbf{r}^\ell. \qquad (38)$$

through an exhaustive search. Next, in the asymmetric setup, the NP test is the same as (36), except that $\hat{\mathbf{y}}_q^n$ is replaced by $\mathbf{y}^n$ in the derivation. In the special case where $p_0 = p_1$, this test reduces to

$$\sum_{i=1}^{n}(\hat{x}_{q,i} \oplus y_i) < \lambda_{qb}, \qquad (39)$$

where $\hat{\mathbf{x}}_q^n = G_q\hat{\mathbf{u}}^k$, and λ_{qb} is the test threshold.

D. Theoretical analysis of the quantize-binning scheme

We now provide exact analytical expressions for the Type-I and Type-II error probabilities of the quantize-binning scheme, under the asymmetric setup and in the particular case where $p_0 = p_1 = 1/2$. Extensions to the symmetric setup and the general case where $p_0 \neq p_1$ are considerably more complex because in this case, the hypothesis test (37) does not simplify further. These extensions are therefore left for future works.

As in Section IV-D, we assume that the all-zero codeword is transmitted. We use the definitions introduced in Section IV-D for $\mathcal{C}_0^{(q)}$ and $E_\gamma^{(q)}$ associated with the generator matrix G_q. Additional definitions are necessary for the quantize-binning scheme, which corresponds to the concatenation of the two codes used for quantization and binning.

We first define the decision region, or coset, $\mathcal{C}_0^{(qb)}$ as

$$\mathcal{C}_0^{(qb)} := \left\{\mathbf{y}^n \in \{0,1\}^n : \arg\min_{\mathbf{u}^k} d(G_q\mathbf{u}^k, \mathbf{y}^n) = \mathbf{0}^k\right\}, \qquad (40)$$

for the all-zero codeword of the quantize-binning scheme. Importantly, the condition $H_b\mathbf{u}^k = \mathbf{r}^\ell$ on the syndrome is not explicitly included in (40), as it is automatically satisfied for $\mathbf{u}^k = \mathbf{0}^k$ due to the linearity of the code. Especially, a side information vector $\mathbf{y}^n$ belongs to $\mathcal{C}_0^{(qb)}$ if the solution of (38) for this vector is $\hat{\mathbf{u}}_q^k = \mathbf{0}^k$.

Next, we define the set of integers $\{E_\nu^{(qb)}\}_{\nu\in[\![0,d_{\max}^{(qb)}]\!]}$, where $E_\nu^{(qb)}$ denotes the number of sequences $\mathbf{y}^n$ with a Hamming weight ν that belong to the decision region $\mathcal{C}_0^{(qb)}$. These quantities are also referred to as the number of coset leaders of weight ν. Here, $d_{\max}^{(qb)}$ denotes the maximum possible weight in $\mathcal{C}_0^{(qb)}$. Additionally, we define the set of integers $\{A_t^{(qb)}\}_{t\in[\![0,n]\!]}$, where $A_t^{(qb)}$ denotes the number of sequences $\mathbf{x}_q^n$ with a Hamming weight t that can be expressed as $\mathbf{x}_q^n = G_q\mathbf{u}_q^k$, for some $\mathbf{u}_q^k$ satisfying $H_b\mathbf{u}_q^k = \mathbf{0}^\ell$. Thus, the set $\{A_t^{(qb)}\}_{t\in[\![0,n]\!]}$ corresponds to the code weight distribution [42, Section 3.1.3] of the concatenated code.

Proposition 2: For the quantize-binning scheme considering $p_0 = p_1 = 1/2$, and for a fixed threshold value λ_{qb}, Type-I and Type-II error probabilities are given by

$$\alpha_n^{(qb)} = 1 - \mathbb{P}_B(c_0) - \mathbb{P}_{\bar{B}}(c_0), \qquad (41)$$
$$\beta_n^{(qb)} = \mathbb{P}_B(c_1) + \mathbb{P}_{\bar{B}}(c_1), \qquad (42)$$

where

$$\mathbb{P}_B(\delta) = \sum_{\nu=0}^{\min(d_{\max}^{(qb)},\lambda_{qb})} \frac{E_\nu^{(qb)}}{\binom{n}{\nu}} \sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \sum_{j=0}^{n} \Gamma_{\nu,j,\gamma}\delta^j(1-\delta)^{n-j}, \tag{43}$$

$$\mathbb{P}_{\bar{B}}(\delta) = \sum_{i=0}^{n} \left[\left(\sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \sum_{j=0}^{n} \Gamma_{i,j,\gamma}\delta^j(1-\delta)^{n-j} \right) \right. \tag{44}$$

$$\left. \times \left(\sum_{t=1}^{n} \sum_{\nu=0}^{\lambda_{qb}} \frac{E_\nu^{(qb)}}{\binom{n}{\nu}} \frac{A_t^{(qb)}\Gamma_{i,\nu,t}}{\binom{n}{i}} \right) \right],$$

where $\Gamma_{\nu,j,\gamma}$ is defined in (31).

Proof: We consider the all-zero codeword $\mathbf{x}_q^n = \mathbf{0}$. Under the hypothesis $\mathcal{H}_0$, we express

$$\alpha_n^{(qb)} = 1 - \mathbb{P}_0(\widehat{\mathcal{H}}_0, B) - \mathbb{P}_0(\widehat{\mathcal{H}}_0, \bar{B}), \tag{45}$$

where B is the event that the correct sequence $\hat{\mathbf{x}}_q^n = \mathbf{x}_q^n$ is retrieved by the decoder, while $\bar{B}$ is the event that an incorrect sequence $\hat{\mathbf{x}}_q^n \neq \mathbf{x}_q^n$ is output by the decoder. In addition, $\widehat{\mathcal{H}}_0$ is the event that the hypothesis $\mathcal{H}_0$ is chosen by the decoder. We further denote $\mathbb{P}_B(p_0) = \mathbb{P}_0(\widehat{\mathcal{H}}_0, B)$ and $\mathbb{P}_{\bar{B}}(p_0) = \mathbb{P}_0(\widehat{\mathcal{H}}_0, \bar{B})$. We then express

$$\mathbb{P}_B(p_0) = \sum_{\nu=0}^{n} \mathbb{P}_0(w(\mathbf{Y}^n) = \nu)\mathbb{P}_0(\widehat{\mathcal{H}}_0, B | w(\mathbf{Y}^n) = \nu) \tag{46}$$

$$= \sum_{\nu=0}^{\min(d_{\max}^{(qb)},\lambda_{qb})} \mathbb{P}_0(w(\mathbf{Y}^n) = \nu) \frac{E_\nu^{(qb)}}{\binom{n}{\nu}}. \tag{47}$$

By following the same steps as in the proof of Proposition 1, we can show that

$$\mathbb{P}_0(w(\mathbf{Y}^n) = \nu) = \sum_{\gamma=0}^{d_{\max}^{(q)}} \frac{E_\gamma^{(q)}}{N_0^{(q)}} \sum_{j=0}^{n} \Gamma_{\nu,j,\gamma} c_0^j (1-c_0)^{n-j}, \tag{48}$$

which provides (43). We then write

$$\mathbb{P}_{\bar{B}}(p_0) = \sum_{i=0}^{n} \mathbb{P}_0(w(\mathbf{Y}^n) = i)\mathbb{P}_0(\widehat{\mathcal{H}}_0, \bar{B} | w(\mathbf{Y}^n) = i), \tag{49}$$

where $\mathbb{P}_0(w(\mathbf{Y}^n) = i)$ is given by (48). We develop

$$\mathbb{P}_0(\widehat{\mathcal{H}}_0, \bar{B} | w(\mathbf{Y}^n) = i)$$

$$= \sum_{t=1}^{n} \sum_{\nu=0}^{\lambda_{qb}} \mathbb{P}_0(w(\hat{\mathbf{X}}_q^n) = t, d(\hat{\mathbf{X}}_q^n, \mathbf{Y}^n) = \nu | w(\mathbf{Y}^n) = i) \tag{50}$$

$$= \sum_{t=1}^{n} \sum_{\nu=0}^{\lambda_{qb}} \frac{E_\nu^{(qb)}}{\binom{n}{\nu}} \frac{A_t^{(qb)}\Gamma_{i,\nu,t}}{\binom{n}{i}}. \tag{51}$$

This provides the expression of $\mathbb{P}_{\bar{B}}(c_0)$ in (44).

We can derive the expression of $\beta_n^{(qb)}$ in (42) by noticing that $\beta_n^{(qb)} = \mathbb{P}_B(c_1) + \mathbb{P}_{\bar{B}}(c_1)$ and using the same derivation as above. This ends the proof. □

In this derivation, one key difficulty compared to the quantization scheme lies in that under both hypothesis, even when the sequence $\mathbf{x}_q^n$ is decoded *incorrectly*, that is $\hat{\mathbf{x}}_q^n \neq \mathbf{x}_q^n$, the pair $(\hat{\mathbf{x}}_q^n, \mathbf{y}^n)$ can still pass the NP test to decide $\mathcal{H}_0$. This is because the decoding targets to minimize the distance between $\hat{\mathbf{x}}_q^n$ and $\mathbf{y}^n$, distance which is then used for comparison in the NP test. This introduces the sum over the distribution of codewords characterized by the set of integers $\{A_t^{(qb)}\}_{t\in[\![0,n]\!]}$.

To the best of the authors' knowledge, these theoretical results are new. It is important to note that the derivations have been conducted exclusively for the asymmetric setup, as extending them to the symmetric case appears to be prohibitively complex. These analytical expressions are expected to facilitate the optimization and comparison of practical schemes across a wide range of source and code parameters. For instance, optimizing $E_\gamma^{(q)}$, $E_\nu^{(qb)}$, and $A_t^{(qb)}$, could lead to an optimal quantize-binning scheme for DHT. We leave such optimization as future work.

VI. Numerical Results

We now evaluate and compare the decision performance of the four proposed coding schemes: separate coding, truncation, quantization, and quantize-binning. To this end, we provide Receiver Operating Characteristic (ROC) curves which plot the Type-II error probability versus the Type-I error probability for each scheme. ROC curves are a standard tool for evaluating the performance of hypothesis tests and provide a clear visualization of the trade-offs between Type-I and Type-II errors.

A. Separate scheme versus truncation scheme

We begin by comparing the two uncoded schemes: the separate scheme and the truncation scheme. The source parameters are set to $c_0 = 0.1$, $p_1 = 0.5$, $c_1 = 0.35$, while several values of $p_0 \in \{0.08, 0.2, 0.3\}$ are considered. We fix the source sequence length to $n = 30$ bits, and the truncation length to $l = 15$ bits.

Figure 2 shows the ROC curves of the two uncoded schemes for different values of p_0. It can be observed that when p_0 is small, the separate scheme outperforms the truncation scheme. This is because Encoder 1 has access to the full n-bits sequence to decide between the two hypotheses characterized by significantly different values, $p_0 = 0.08$ and $p_1 = 0.5$. However, as p_0 increases and approaches p_1, the truncation scheme becomes advantageous. Indeed, in this regime, it becomes increasingly difficult for Encoder 1 in the separate scheme to accurately distinguish between $\mathcal{H}_0$ and $\mathcal{H}_1$ based solely on its local observations. In contrast, the truncation scheme benefits from additional information through the values of c_0 and c_1, which improves the decision-making process.

In summary, while the separate scheme is rate-efficient, transmitting only one bit of information, its decision accuracy deteriorates as p_0 approaches p_1. Conversely, the truncation scheme, though less rate-efficient due to the transmission of truncated sequences at rates $R_1 = R_2 = R = l/n$, achieves better decision accuracy via joint decoding as p_0 increases. These observations motivate the development of coding schemes aimed at further improving both Type-I and Type-II error probabilities beyond the performance of the truncation scheme.

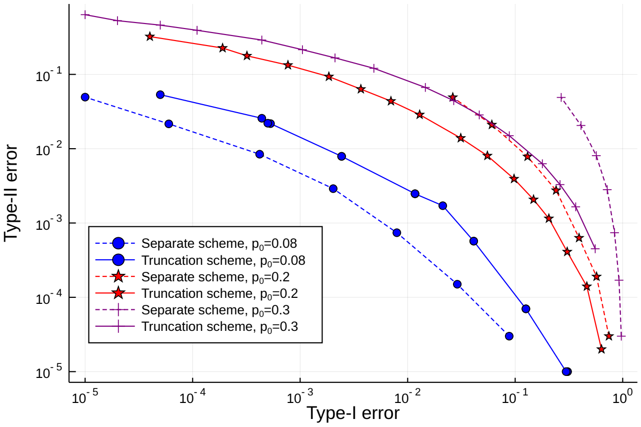

Fig. 2: ROC curves for separate and truncation schemes. Source parameters are set to $c_0 = 0.1$, $p_1 = 0.5$ and $c_1 = 0.35$, $p_0 \in \{0.08, 0.2, 0.3\}$, $n = 30$, $l = 15$.

B. Quantization scheme in the asymmetric setup

We now investigate the performance of the quantization scheme in the asymmetric setup. We first assess the accuracy of the theoretical analysis provided in Proposition 1. The source parameters are set to $p_0 = 0.5$, $p_1 = 0.5$, $c_1 = 0.5$, and $c_0 \in \{0.07, 0.1\}$. We first consider a (31,16) BCH code with minimum distance $d_{\min} = 7$, which gives a source length $n = 31$ and $k = 16$ coded bits. For comparison, we also consider the $(32, 16)$ LDPC CCSDS code, and the $(32, 16)$ Polar code from [45], both with parameters $n = 32$ and $k = 16$. For the three codes, the quantization operation (23), is performed from ML decoding.

Figure 3 presents the ROC curves for both schemes with the previous parameters and the three considered codes. Theoretical curves are obtained via numerical evaluation of the expressions in Proposition 1, while practical results are measured from Monte Carlo simulations averaged over $100,000$ trials per point. For all codes, the results indicate a close match between theoretical and empirical Type-I and Type-II error probabilities. This is because the error probability expressions take into account the considered code through the terms $E_\gamma^{(q)}$. Therefore, the theoretical expressions are found useful for the design of DHT curves. Moreover, Figure 3 shows that the BCH code outperforms the considered LDPC and Polar code in the asymmetric setup for the considered values of c_0.

C. Quantization scheme in the symmetric setup

We now investigate the symmetric setup, aiming to compare the proposed quantization scheme against the truncation scheme and existing quantization schemes based on LDGM codes with BiP algorithms [24], [29], [32]. We also aim to investigate the effect of different rates $R_1 \neq R_2$. We first consider a source length $n = 31$, and source parameters $p_0 \in \{0.05, 0.08\}$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$. In order to consider unequal rates $R_1 \neq R_2$, we use k_x (resp. k_y) to denote the number of coded bits for source X (resp. Y). We consider three different code constructions for the sources X and Y, all such that $\frac{R_1+R_2}{2} = 0.51$ bits/symbol.

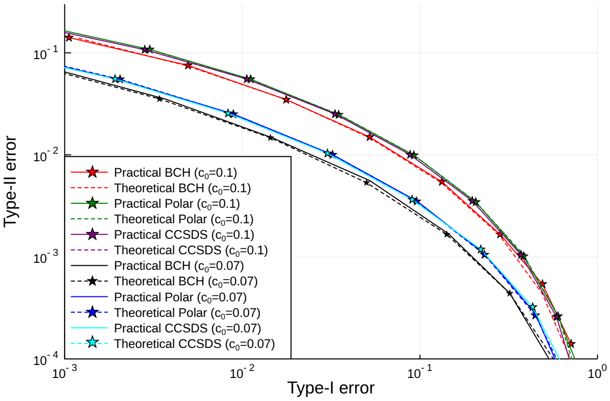

Fig. 3: ROC curves obtained from theoretical analysis (dashed lines) and Monte Carlo simulations (plain lines), for the $(31, 16)$ BCH code, the $(32, 16)$ CCSDS code, and the $(32, 16)$ Polar code, each used as quantizers in the assymetric setup. Source parameters are set to $p_0 = 0.5$, $p_1 = 0.5$, $c_1 = 0.5$, and $c_0 \in \{0.07, 0.1\}$.

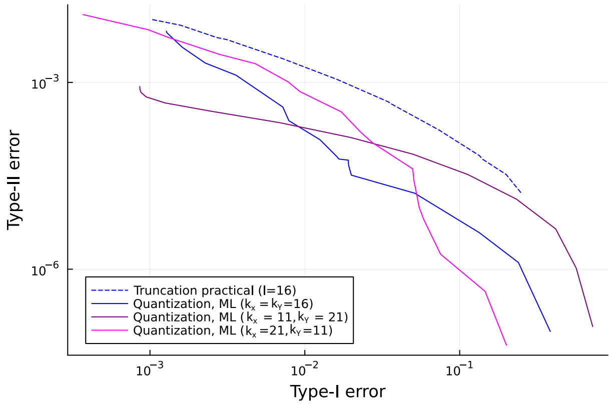

Fig. 4: ROC curves for quantization from $n = 31$ bits, compared to the truncation scheme in the symmetric setup. Source parameters are set to $p_0 = 0.05$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$. We consider three different BCH code constructions for the sources X and Y, each verifying $\frac{R_1+R_2}{2} = 0.51$ bits/symbol.

1) Both sources are quantized with the $(31, 16)$ BCH code, resulting in $k_x = k_y = 16$ bits
2) The source X is quantized with the $(31, 11)$ BCH code, and the source Y is quantized with the $(31, 21)$ BCH code, resulting in $k_x = 11$ bits, $k_y = 21$ bits.
3) The source X is quantized with the $(31, 21)$ BCH code, and the source Y is quantized with the $(31, 11)$ BCH code, resulting in $k_x = 21$ bits, $k_y = 11$ bits.

Figure 4 shows the ROC curves for the truncation scheme with $l = 16$, and for the previous three code constructions under ML decoding. For the three constructions, a clear advantage is observed for the quantization scheme decoder over the truncation scheme. Interestingly, we also observe that each of the code constructions provides a different tradeoff in terms of Type-I versus Type-II error probability. We also observe that the curves for the quantization scheme in Figure 4,

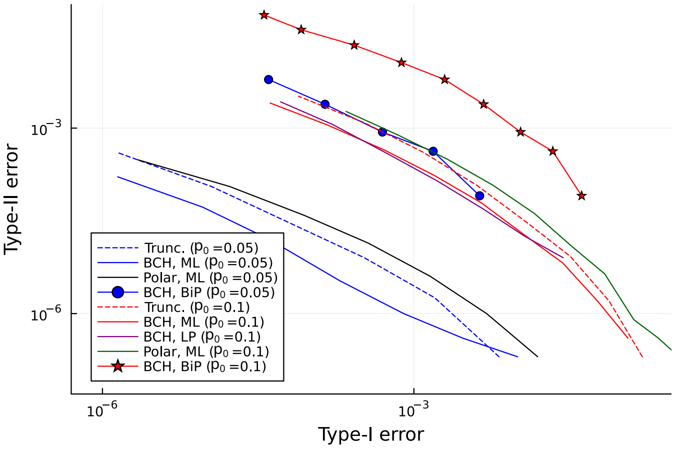

Fig. 5: ROC curve for the $(63, 45)$ BCH code and the $(64, 43)$ Polar code used as quantizers, compared to the truncation scheme in the symmetric setup. Source parameters are set to $p_0 \in \{0.05, 0.1\}$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$.

although averaged over 10^8 simulations, have some points of irregularity. This comes from the test in equation (26) which involves integer values $w(\mathbf{x}_q^n)$ and $w(\mathbf{v}_q^n)$, and from the low source length $n = 31$, which limits the range of these values.

We next investigate the impact of code length in the symmetric setup. We consider a longer source sequence of $n = 63$ bits and consider the $(63, 36)$ BCH code with minimum distance 11. The source parameters are set to $p_0 \in \{0.05, 0.1\}$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$. For comparison, we also consider the $(64, 43)$ Polar code from [45]. For each code, the quantization (23) is performed from ML decoding. For the $(63, 36)$ BCH code, we also make a comparison with LP decoding from [43], and with BiP decoder with decimation of [32].The LP decoder is implemented from the open-source Julia optimization library SCIP. The ROC curves are shown in Figure 5. As before, a clear performance gain is observed for the quantization scheme with ML decoding over the truncation scheme. For the $(63, 36)$ BCH code, we observe that the LP decoder provides the same performance as the syndrome-based ML decoder, while the BiP decoder exhibits a significant performance degradation compared to the other two decoders. This performance gap is attributed to the fact that the BCH code is not sparse, which lowers the efficiency of message-passing algorithms such as BiP. Although LDGM codes could be considered to address this issue, their typically low minimum distance at short block lengths can severely degrade performance. We also observe that the Polar code shows lower performance not only compared to the BCH code, but also compared to the truncation scheme. This comes from the fact that the truncation scheme already performs well for DHT, given that the purpose is not on reconstructing the source but on making an accurate decision. This makes the DHT code design problem challenging.

D. *Quantize-binning scheme in the asymmetric setup*

We now evaluate the performance of the quantize-binning scheme in the asymmetric setup. As with the quantization scheme, we first investigate the accuracy of the theoretical analysis of Proposition 2. The source parameters are set to $p_0 = 0.5$, $p_1 = 0.5$, $c_0 \in \{0.01, 0.03\}$, $c_1 = 0.35$. We consider a source length $n = 31$, using the BCH $(31, 16)$-code for quantization and the $(16, 5)$ Reed-Muller code for binning. We also make a comparison with a quantizer alone, realized from the $(31, 11)$ BCH code. As a result, the final codeword length is $\ell = 11$ bits for the quantize-binning scheme, $k = 11$ bits for the quantizer scheme, and $l = 11$ bits for the truncation scheme. In all setups, the quantization (23) and the debinning are realized with ML decoding.

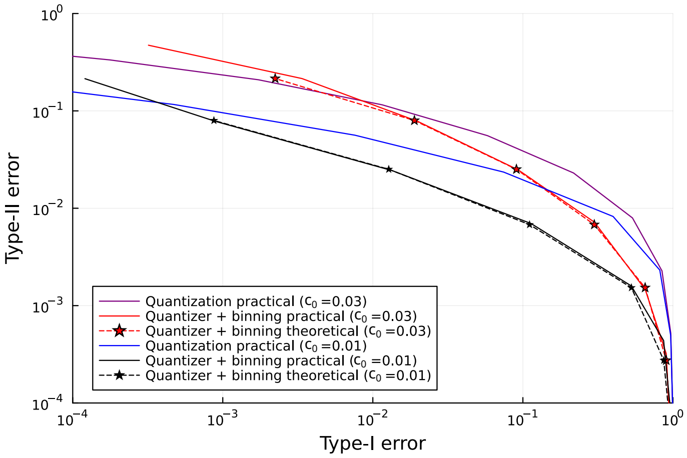

Fig. 6: ROC curves in the asymmetric case for the quantize-binning scheme built from the $(31, 16)$ BCH code for quantization and the $(16, 5)$ Reed-Muller code for binning, compared with the quantizer scheme with the $(31, 11)$ BCH code. Source parameters are set to $p_0 = 0.5$, $p_1 = 0.5$, $c_0 \in \{0.01, 0.03\}$, $c_1 = 0.35$.

Figure 6 shows the ROC curves for both schemes. Theoretical curves are obtained by numerical evaluation of the expressions in Proposition 2, while practical performance is measured via Monte Carlo simulations. The results show that the quantize-binning scheme provides a significant performance improvement over the quantization scheme. Furthermore, the theoretical Type-I and Type-II error probabilities closely match the empirical results, validating the accuracy of the theoretical analysis. In summary, the quantize-binning scheme effectively improves the decision performance in the asymmetric setup while maintaining a low transmission rate.

E. *Quantize-binning scheme in the symmetric setup*

In the symmetric setup, the source parameters are set to $p_0 \in \{0.05, 0.08\}$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$. We consider again the $(31, 16)$ BCH code for the quantizers and the $(16, 5)$-Reed-Muller code for the binning. We also make a comparison with the truncation scheme with $l = 11$ transmitted bits. Figure 7 shows the ROC curves obtained from Monte-Carlo simulations, averaged over 10^5 trials for each point, for both the quantization scheme and the quantize-binning scheme. As before, we observe that the quantize-binning scheme outperforms the truncation scheme. These

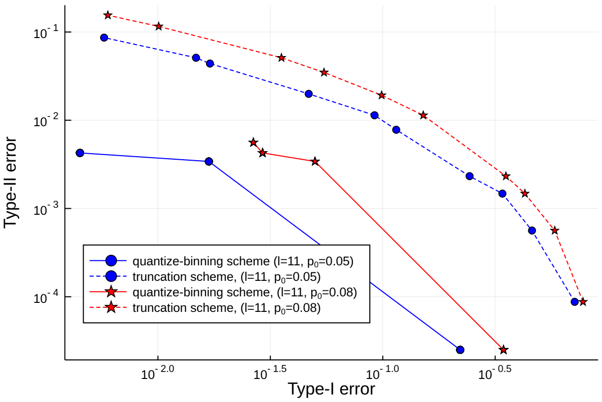

Fig. 7: ROC curves in the asymmetric case for the quantize-binning scheme built from the $(31,16)$ BCH code for quantization and the $(16,5)$ Reed-Muller code for binning. Source parameters are set to $p_0 \in \{0.05, 0.08\}$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$.

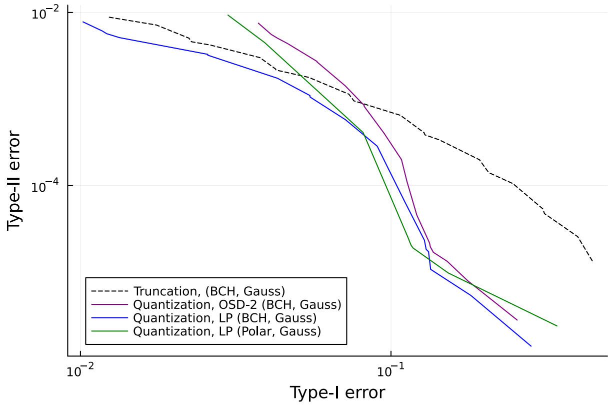

Fig. 8: ROC curve for Gaussian sources, for the quantization scheme built from the BCH code $(31,16)$ and from the Polar $(32,16)$ code. The source parameters are set to $p_0 = 0.05$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$, $\sigma = 0.5$.

results confirm the relevance of practical quantization-binning schemes for DHT.

F. Extension to a Gaussian source model

We now show that the DHT coding schemes proposed in this paper can be extended to Gaussian source models. To describe this model, we first consider three sources X_p, Y_p, Z_p, taking values in the binary alphabet $\{-1, +1\}$, with $Y_p = X_p Z_p$, $\mathbb{P}(X_p = -1) = p$, $\mathbb{P}(Z_p = -1) = c$. Next, the encoders observe $X = X_p + N_1$, $Y = Y_p + N_2$, where both N_1 and N_2 are independent random variables following the same Gaussian distribution $\mathcal{N}(0, \sigma^2)$. In this case, the hypothesis test is still given by (2), as in the binary case.

For Monte Carlo simulations, we consider the symmetric setup, with source parameters $p_0 = 0.05$, $p_1 = 0.5$, $c_0 = 0.1$, $c_1 = 0.35$, $\sigma = 0.5$, and we evaluate the $(31,16)$ BCH code and the $(32,16)$ Polar code for quantization. For Gaussian sources, the quantization is still realized from equations (23) and (24), where the Hamming distance is replaced by the Euclidean distance. We use either the LP decoder from [43] or the Ordered-Statistic decoder (OSD) of order 2 to implement the quantization. We also make a comparison with the truncation scheme where $l = 16$ symbols of X and Y are one-bit quantized and transmitted. The ROC curves are shown in Figure 8. For the BCH code, we first observe that the quantization scheme shows better performance than the truncation scheme. We also observe that the LP decoder, which implements ML decoding, shows the best performance, while the OSD provides a good tradeoff between complexity and performance. Interestingly, the Polar code shows performance close to the BCH code. This opens interesting perspectives for other source models for DHT, and for investigating the tradeoff between DHT coding scheme performance and complexity.

VII. Conclusion

In this paper, we proposed practical short-length coding schemes for binary DHT in both asymmetric and symmetric setups. We introduced two coding schemes, one built with a binary quantizer, and the other built as a quantize-binning scheme. Both schemes were constructed using short linear block codes. For each considered scheme, in addition to practical constructions, we derived theoretical expressions of Type-I and Type-II error probabilities in the asymmetric case. Simulation results demonstrated that the proposed quantization and quantize-binning schemes outperform the baseline truncation scheme, and further confirmed the accuracy of the theoretical analyses. Future work will focus on leveraging the theoretical error probability expressions to optimize the code for DHT. In that purpose, we will investigate methods derived from channel coding, such as for instance on the use of genetic algorithms to build generator matrices or parity check matrices that optimize a certain tradeoff between Type-I and Type-II error probability. We will also investigate intermediate block length between 100 and 1000 bits, aiming to develop efficient low-complexity decoding solutions for DHT. For this, we may consider the use of LDPC and Polar codes which may become more competitive in this regime. Finally, we also aim to consider more complex source models, and in particular to generalize the results of Proposition 1 and 2 to asymmetric sources where $p_0 \neq p_1$.

References

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

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

[3] E. C. Strinati and S. Barbarossa, "6g networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[4] P. A. Stavrou and M. Kountouris, "A rate distortion approach to goal-oriented communication," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 590–595.

[5] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian, and H. V. Poor, "Goal-oriented quantization: Analysis, design, and application to resource allocation," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 42–54, 2022.

[6] M. Lindén and M. Björkman, "Embedded sensor systems for health–a step towards personalized health," in *pHealth 2018*. IOS Press, 2018, pp. 69–74.

[7] I. Lior, A. Sladen, D. Rivet, J.-P. Ampuero, Y. Hello, C. Becerril, H. F. Martins, P. Lamare, C. Jestin, S. Tsagkli *et al.*, “On the detection capabilities of underwater distributed acoustic sensing,” *Journal of Geophysical Research: Solid Earth*, vol. 126, no. 3, p. e2020JB020925, 2021.

[8] J. A. Sánchez, D. Melendi, R. García, X. G. Pañeda, V. Corcoba, and D. García, “Distributed and collaborative system to improve traffic conditions using fuzzy logic and v2x communications,” *Vehicular Communications*, vol. 47, p. 100746, 2024.

[9] R. Ahlswede and I. Csiszár, “Hypothesis testing with communication constraints,” *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[10] T. S. Han, “Hypothesis Testing with Multiterminal Data Compression,” *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[11] T. S. Han and S. Amari, “Statistical inference under multiterminal data compression,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.

[12] H. Shimokawa, T. S. Han, and S. Amari, “Error bound of hypothesis testing with data compression,” in *IEEE International Symposium on Information Theory (ISIT)*, 1994, p. 114.

[13] G. Katz, P. Piantanida, R. Couillet, and M. Debbah, “On the necessity of binning for the distributed hypothesis testing problem,” *IEEE International Symposium on Information Theory (ISIT)*, pp. 2797–2801, 2015.

[14] I. Salihou Adamou, E. Dupraz, and T. Matsumoto, “An information-spectrum approach to distributed hypothesis testing for general sources,” in *International Zurich Seminar on Information and Communication (IZS). Proceedings*. ETH Zürich, 2024, pp. 144–148.

[15] Y. Kochman and L. Wang, “Improved random-binning exponent for distributed hypothesis testing,” *arXiv preprint arXiv:2306.14499*, 2023.

[16] S. Sreekumar and D. Gündüz, “Distributed hypothesis testing over discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2044–2066, 2019.

[17] S. Salehkalaibar and M. Wigger, “Distributed hypothesis testing over multi-access channels,” in *IEEE Global Communications Conference (Globecom)*, 2018, pp. 1–6.

[18] S. Salehkalaibar, M. Wigger, and L. Wang, “Hypothesis testing over the two-hop relay network,” *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4411–4433, 2019.

[19] H. M. Shalaby and A. Papamarcou, “Multiterminal detection with zero-rate data compression,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 254–267, 1992.

[20] T. S. Han and K. Kobayashi, “Exponential-type error probabilities for multiterminal hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 2–14, 2006.

[21] S.-I. Amari and T. S. Han, “Statistical inference under multiterminal rate restrictions: A differential geometric approach,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 217–227, 1989.

[22] S. Watanabe, “Neyman–pearson test for zero-rate multiterminal hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4923–4939, 2017.

[23] E. Haim and Y. Kochman, “On Binary Distributed Hypothesis Testing,” pp. 1–37, 2017. [Online]. Available: http://arxiv.org/abs/1801.00310

[24] J. Fridrich and T. Filler, “Binary quantization using belief propagation with decimation over factor graphs of LDGM codes,” in *Proceedings of the 45th Allerton Conference on Coding, Communication, and Control*, 2007, pp. 495–501.

[25] Z. Xiong, A. D. Liveris, and S. Cheng, “Distributed source coding for sensor networks,” *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, 2004.

[26] A. D. Liveris, Z. Xiong, and C. N. Georghiades, “Compression of binary sources with side information at the decoder using LDPC codes,” *IEEE Communications Letters*, vol. 6, no. 10, pp. 440–442, 2002.

[27] A. Savard and C. Weidmann, “Improved decoding for binary source coding with coded side information,” in *IEEE Information Theory Workshop (ITW)*, 2013, pp. 1–5.

[28] F. Ye, E. Dupraz, Z. Mheich, and K. Amis, “Optimized rate-adaptive protograph-based LDPC codes for source coding with side information,” *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 3879–3889, 2019.

[29] M. J. Wainwright and E. Martinian, “Low-density graph codes that are optimal for binning and coding with side information,” *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 1061–1079, 2009.

[30] Y. Matsunaga and H. Yamamoto, “A coding theorem for lossy data compression by LDPC codes,” *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2225–2229, 2003.

[31] M. J. Wainwright, E. Maneva, and E. Martinian, “Lossy source compression using low-density generator matrix codes: Analysis and algorithms,” *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1351–1368, 2010.

[32] D. Castanheira and A. Gameiro, “Lossy source coding using belief propagation and soft-decimation over LDGM codes,” in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2010, pp. 431–436.

[33] M. Nangir, M. Ahmadian-Attari, and R. Asvadi, “Binary Wyner–Ziv code design based on compound LDGM–LDPC structures,” *IET Communications*, vol. 12, no. 4, pp. 375–383, 2018.

[34] J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, 1970.

[35] J. Liu, W. Zhang, and H. V. Poor, “A rate-distortion framework for characterizing semantic information,” in *IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2894–2899.

[36] G. Katz, P. Piantanida, and M. Debbah, “Distributed Binary Detection with Lossy Data Compression,” *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5207–5227, 2017.

[37] M. S. Rahman and A. B. Wagner, “On the optimality of binning for distributed hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6282–6303, 2012.

[38] S. Watanabe, “On sub-optimality of random binning for distributed hypothesis testing,” in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 2708–2713.

[39] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*. Springer, 2005, vol. 3.

[40] V. Chandar, E. Martinian, and G. W. Wornell, “Information embedding codes on graphs with iterative encoding and decoding,” in *IEEE International Symposium on Information Theory*, 2006, pp. 866–870.

[41] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge university press, 2008.

[42] W. Ryan and S. Lin, *Channel Codes: Classical and Modern*. Cambridge university press, 2009.

[43] S. Scholl, F. Kienle, M. Helmling, and S. Ruzika, “Integer programming as a tool for analysis of channel codes,” in *SCC 2013; 9th International ITG Conference on Systems, Communication and Coding*. VDE, 2013, pp. 1–6.

[44] F. Khaledian, R. Asvadi, E. Dupraz, and T. Matsumoto, “Covering codes as near-optimal quantizers for distributed hypothesis testing against independence,” in *IEEE Information Theory Workshop (ITW)*, 2024, pp. 67–72.

[45] M. Helmling, S. Scholl, F. Gensheimer, T. Dietz, K. Kraft, O. Griebel, S. Ruzika, and N. Wehn, “Database of Channel Codes and ML Simulation Results,” www.rptu.de/channel-codes, 2025.