Distributed Source Coding for Parametric and Non-Parametric Regression

Jiahui Wei, *Student Member, IEEE*, Elsa Dupraz, *Member, IEEE*, Philippe Mary, *Member, IEEE*

Abstract

The design of communication systems dedicated to machine learning tasks is one key aspect of goal-oriented communications. In this framework, this article investigates the interplay between data reconstruction and learning from the same compressed observations, particularly focusing on the regression problem. We establish achievable rate-generalization error regions for both parametric and non-parametric regression, where the generalization error measures the regression performance on previously unseen data. The analysis covers both asymptotic and finite block-length regimes, providing fundamental results and practical insights for the design of coding schemes dedicated to regression. The asymptotic analysis relies on conventional Wyner-Ziv coding schemes which we extend to study the convergence of the generalization error. The finite-length analysis uses the notions of information density and dispersion with additional term for the generalization error. We further investigate the trade-off between reconstruction and regression in both asymptotic and non-asymptotic regimes. Contrary to the existing literature which focused on other learning tasks, our results state that in the case of regression, there is no trade-off between data reconstruction and regression in the asymptotic regime. We also observe the same absence of trade-off for the considered achievable scheme in the finite-length regime, by analyzing correlation between distortion and generalization error.

Index Terms

J. Wei and P. Mary are with Univ Rennes, INSA Rennes, CNRS, IETR-UMR 6164, F-35000 Rennes, France.

J. Wei and E. Dupraz are with IMT Atlantique, CNRS UMR 6285, Lab-STICC, Brest, France

This work has received a 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. This work was also funded by the Brittany region.

Part of the content of this paper has been published in Eusipco 2023 and IZS 2024. We extend the analysis to the non-trivial case of kernel methods and investigate the non-asymptotic trade-off between data reconstruction and regression.

Source coding, Wyner-Ziv coding, Goal-oriented communications, Regression, Finite block length.

I. Introduction

A. Context and problem

The interaction of machine learning and communications is presently a vibrant area of research with numerous works dedicated to machine learning for communications. But the reciprocal relationship is also of significant interest and lies in the design of communication systems dedicated to machine learning tasks. This paradigm falls into the emerging area of goal-oriented communications [1]. In this case, the primary objective of the communication system shifts towards extracting and transmitting relevant information for the targeted learning task, encompassing methods such as hypothesis testing [2], regression [3], or classification [4]. When engineers tackle machine learning over a rate-limited channel, the following key questions emerge: do the optimal encoder and decoder design for the learning task align with those used in conventional communication systems? Or is there an inherent trade-off between the learning task and data reconstruction? In this article, we address these questions in the context of regression.

Regression is one of the most popular supervised machine learning tasks and despite its apparent simplicity, it is used in many practical signal processing and telecommunication problems. For instance, non-parametric regression is used to reconstruct electrocardiograms in [5]. In [6], the base station relies on the users signal feedback to estimate a radio map with a semi-parametric regression. In [7], frequency hopping parameters are inferred from a sparse linear regression. The work in [8] deals with the identification of non-linearities in Wiener systems from a semi-parametric regression technique. In [9], a parametric regression technique is proposed to estimate a field function through distributed and noisy sensor measurements sent to a fusion center. However, none of these works considers the problem of compressing and communicating the data so as to perform the regression task at the remote server.

In this paper, we formulate the distributed regression problem as follows. As in standard regression, we consider a pair of real-valued random variables (X, Y). We assume that the statistical relation between X and Y is described by a function $f : \mathbb{R} \to \mathbb{R}$ such that $X = f(Y) + N$, where f is unknown and N is a Gaussian noise. Like in the conventional Wyner-Ziv setup [10], we consider that X acts as the source to be encoded, while Y serves as side information available only at the decoder. But in our case, the objective of the decoder is to

infer the function f from Y and from the coded version of X. This regression is performed by minimizing the Mean-Squared Error (MSE) $\mathbb{E}\left[(\hat{f}(Y)-X)^2\right]$ with respect to $\hat{f}$.

In cases where there is prior knowledge about the structure of the function f (linear, polynomial, etc.) and f depends on a finite number of parameters, the problem is termed as parametric regression. In this context, the ordinary least squares (OLS) estimator is known for providing the best unbiased estimator [11]. On the other hand, non-parametric regression does not make any assumption about the structure of the underlying function f. In this case, various methods, such as kernel methods, K-Nearest Neighbors (KNN), or modeling involving a local or global averaging over the training set are applicable [12]. In this paper, we investigate both parametric regression and non-parametric kernel regression.

As learning performance criterion, we consider the regression generalization error, defined as the MSE evaluated on test samples different from the training samples. Our first objective is to provide achievable rates under constraints on the generalization error, for parametric and non-parametric regressions. Our second objective is to investigate the trade-off between reconstruction and regression, whenever the coding scheme is required to satisfy not only the constraint on the generalization error, but also another constraint on the distortion on X.

B. Related works

Coding for computation has been a long-standing area of research, extending the Wyner-Ziv setup to cases where the decoder aims to compute a function of the source and side information. Studies such as [13], [14], building upon earlier works [15], [16], have investigated the theoretical limits of coding data specifically for computation purposes. These studies typically focus on decoding one output value $f(X_k, Y_k)$ for each sample pair (X_k, Y_k), using a predefined function f. However, this approach differs from our regression problem, where the goal is to infer the function f itself across an entire length-n sequence of sample pairs $\{(X_k, Y_k)\}_{k=1}^n$. Additionally, the theoretical frameworks in the previous works rely on the entropy of a characteristic graph, a measure suitable only for functions with finite support, rendering it less appropriate for addressing regression.

Regarding coding schemes dedicated to learning, [17], [18] state that the optimal performance is achieved through an estimate-and-compress strategy. However, practical limitations due to hardware constraints or computational capabilities at the encoder, as well as the distributed nature of data across networks, often make this strategy impractical. In such cases, a compress-

and-estimate scheme [19]–[22] is more relevant. For example, a rate-distortion framework has been introduced in [19] specifically for semantic communications involving continuous sources, where each source is subject to its own distortion constraints: one for the information observed at the encoder and another for the hidden semantic source. This approach was further extended to discrete sources in [20]. Moreover, the information bottleneck framework [23]–[25] utilizes mutual information to measure the relevance of information extracted from the source. Despite these advancements, there remains a significant challenge in linking distortion or mutual information metrics directly to the performance of the considered learning tasks.

Some other works have considered coding with performance metrics specific to the considered learning task. Especially, the study conducted in [21] demonstrated that, under the criterion of the variance of unbiased estimator, the rate necessary for estimating a parameter θ of the joint distribution P_{XY} is lower than that required for source reconstruction. Distributed hypothesis testing has also been widely explored recently. For instance, [22], [26], [27] provided Type-II error exponents under constraints on the Type-I error for various hypothesis testing problems including testing against independence. In addition, Raginsky has established lower and upper bounds on the learning generalization error of coding schemes dedicated to a range of distributed learning problems involving two sources X and Y [3]. However, we demonstrated in [28], [29] that the upper bound in [3] is loose for both linear and polynomial regression. Building on these findings, this paper aims to investigate regression more broadly, addressing both parametric and non-parametric regression.

In the context of parameter estimation [21], hypothesis testing [22], as well as in the rate-distortion framework for semantic compression [19], [20], [30], research has consistently demonstrated an inherent trade-off between data reconstruction and the specific learning task being considered. A similar trade-off was showcased for the problem of visual perception versus data reconstruction [31], where visual perception was measured from a divergence term, and also in the context of data identification and reconstruction in a noisy database [32]. In this paper, we also investigate this trade-off for the considered regression problems.

C. Contributions

In this paper, we provide rate-generalization error regions for both parametric regression and kernel regression, across both asymptotic and finite block-length regimes, for the source coding setup with side information.

We first investigate the asymptotically achievable rates under generalization error constraints. We utilize standard methods from asymptotic information theory, *e.g.*, [10], [33], which we extend to address the regression problem and analyze the generalization error. More specifically, we consider the achievable coding scheme of Draper [33], originally proposed for Wyner-Ziv coding when the joint distribution P_{XY} is unknown. Within this scheme, our novel contribution lies in analyzing the convergence of the generalization error for regression, instead of the distortion. Our results demonstrate that the minimum expected regression generalization error can be achieved at any positive rate, thus closing the gap between the lower and upper bounds on the generalization error, and improving upon the upper bound established by Raginsky [3].

Furthermore, we extend these results beyond the asymptotic regime by employing finite block-length tools from [34]–[36]. Especially, while the original information density vector for the Wyner-Ziv problem comprised three components, two for the rate and one for the distortion [36], we introduce an additional term accounting for the regression generalization error. This allows us to provide an achievable finite block-length rate-generalization error region for regression.

Finally, in both the asymptotic and the finite block-length regime, we investigate the trade-off in terms of coding rate between regression and data reconstruction. Interestingly, our asymptotic analysis reveals a noteworthy outcome: contrary to findings in existing literature, there appears to be no trade-off between data reconstruction and regression in our investigated context. This result comes from the fact that asymptotically the generalization error upper bound matches the lower bound. At finite-length, we propose a novel method to analyze the trade-off by investigating the correlation between the distortion and generalization-error. We show that for the proposed achievability scheme, there is again no trade-off between the two constraints at finite-length. This analysis of the correlation could be easily extended to other achievability schemes which may be derived for the regression-reconstruction problem in the future.

The remaining of the paper is organized as follows. Section II presents the source model and describes the considered regression methods. Section III defines the generalization error as well as the coding scheme for regression. Section IV and Section V provide the asymptotic and non-asymptotic rate-generalization error regions, respectively. Section VI provides numerical results in non-asymptotic regime for several regression problems.

II. Regression problem

In this section, after providing the notation we will use throughout the paper, we define the statistical model we consider for the sources X and Y. We then present the two regression methods we investigate in this paper: parametric regression with OLS, and non-parametric regression from kernel methods.

A. Notation

A random variable X is denoted with a capital letter, while a realization of the random variable is denoted with lower-case letter x. Let $\mathbb{E}[X]$ and $\mathbb{V}[X]$ be the mean and variance of the random variable X. Random vectors of length n are denoted in bold, *e.g*, $\boldsymbol{X} = [X_1, ..., X_n]^T$, and $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{C}[\boldsymbol{X}]$ are the mean vector and the covariance matrix of $\boldsymbol{X}$, respectively. Next, we use bold letters with underlines, e.g., $\underline{\boldsymbol{X}}$ to denote matrices. When $\underline{\boldsymbol{X}}$ is a squared matrix, we use Tr($\underline{\boldsymbol{X}}$) to denote its trace, while $\lambda_{\max}(\underline{\boldsymbol{X}})$ and $\lambda_{\min}(\underline{\boldsymbol{X}})$ are the maximum and minimum eigenvalues of $\underline{\boldsymbol{X}}$, respectively. We further denote $||\underline{\boldsymbol{X}}||$ as the norm-2 of a matrix $\underline{\boldsymbol{X}}$. Sets are denoted with calligraphic fonts, and if $f : \mathcal{X} \to \mathcal{Y}$ is a mapping then $|f|$ is the cardinality of $\mathcal{Y}$. In addition, $\log(\cdot)$ denotes the base-2 logarithm. Moreover, the indicator function is defined as $\mathbf{1}[x \in \mathcal{A}] = 1$ if $x \in \mathcal{A}$ and 0 otherwise.

Let us consider the measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$, where $\mathscr{B}(\mathcal{X})$ is the Borel σ-algebra on the set $\mathcal{X}$. The probability measure P_X over $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is the distribution of X. The notation $\mathbb{P}[\cdot]$ is used for the probability of an event over the underlying probability space. When $\mathcal{X} = \mathbb{R}$ and the Radon-Nykodym derivative of P_X with respect to the Lebesgue measure λ exists, then it is denoted p_X and is called the probability density function of the random variable X. When $\mathcal{X}$ is countable or finite and the Radon-Nykodym derivative with respect to the counting measure μ exists, p_X is referred to as the the probability mass function of this random variable.

Being given a random vector $\boldsymbol{X}$, the probability measure $P_{\boldsymbol{X}}$ on the measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ admits a joint probability density function $p_{\boldsymbol{X}}$ if for all $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$ in $\mathbb{R}^n$ we have:

$$P_{\boldsymbol{X}}(]-\infty, x_1] \times \cdots \times]-\infty, x_n]) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_{\boldsymbol{X}}(u_1, \cdots, u_n)\, du_1 \cdots du_n. \tag{1}$$

Moreover, for a joint probability measure P_{XY} on $\mathcal{X} \times \mathcal{Y}$, the information density is denoted as [37]

$$\iota(x, y) := \log \frac{dP_{Y|X=x}}{dP_Y}(y), \tag{2}$$

where the ratio above is the Radon-Nykodym derivative of the conditional measure $P_{Y|X=x}$ with respect to the measure P_Y, in y. Given a pair (X, Y) on the measurable space $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$ induced by the joint probability measure P_{XY}, then the function

$$\begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^+ \\ (x,y) & \longmapsto & \frac{p_{XY}(x,y)}{p_X(x)} \end{array} \tag{3}$$

is called the conditional probability density function of Y given X and is denoted $p_{Y|X}(y \mid x)$.

When X is discrete and Y is continuous, let us consider the measurable space $(\mathbb{N} \times \mathbb{R}, \mathcal{P}(\mathbb{N}) \otimes \mathscr{B}(\mathbb{R}))$, where $\mathcal{P}(\mathbb{N})$ is a partition of the set of integers. We define the joint probability measure P_{XY} such as, for all $A \in \mathcal{P}(\mathbb{N})$ and $B \in \mathscr{B}(\mathbb{R})$, we have

$$P_{XY}(A \times B) \triangleq \int_{A \times B} p_X(x) p_{Y|X}(y|x) d\mu(x) d\lambda(y) = \sum_{x \in A} p_X(x) \int_B p_{Y|X}(y|x) dy. \tag{4}$$

B. Source definitions

Let $(X, Y) \sim P_{XY}$ be a pair of jointly distributed real-valued random variables, where X is the source to be encoded and Y is the side information only available at the decoder. We assume that there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that

$$X = f(Y) + N, \tag{5}$$

where $N \sim \mathcal{N}(0, \sigma^2)$ follows a Gaussian distribution with mean 0 and variance σ^2. We further suppose that N is independent from Y. Without loss of generality but for simplicity, we consider $\mathbb{E}[Y] = 0$. We do not make any further assumption on the distribution of Y, except for kernel regression where the distribution support of source Y has to be bounded. Therefore, our theoretical results apply to a wide range of distributions for X and Y. In addition, the function f between X and Y is deterministic, and we consider that it is unknown. The purpose of regression is to infer the function f from a set of observations represented by n independent and identically distributed (i.i.d.) sample pairs $\{(X_k, Y_k)\}_{k=1}^n$. There exists different types of regression, depending on what prior knowledge is available on the structure of the function f.

C. Parametric regression

In the case of parametric regression, (5) can be rewritten as [11]

$$X = \sum_{i=0}^{k-1} \beta_i h_i(Y) + N, \tag{6}$$

where k is the order of the regression, and the functions $h_i : \mathbb{R} \to \mathbb{R}$ are fixed and known in advance, while the parameters β_i are unknown, for all $i \in [\![0, k-1]\!]$. Therefore, parametric regression reduces to estimating the parameter vector $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_{k-1}]$.

Here, we consider the OLS estimator, known for being the unbiased estimator with the minimal variance [11, Chapter 7]. Let us define $\boldsymbol{Y_j^\star} = [h_0(Y_j), ..., h_{k-1}(Y_j)]^T \in \mathbb{R}^k$, and $\underline{\boldsymbol{Y}}^\star = [\boldsymbol{Y_1^\star}, ..., \boldsymbol{Y_n^\star}] \in \mathbb{R}^{k \times n}$. For given vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, the OLS estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}\right)^{-1} \underline{\boldsymbol{Y}}^\star \boldsymbol{X}. \tag{7}$$

According to the properties of OLS estimators, we have [11, Chapter 7]:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \quad \text{and} \quad \mathbb{C}\left[\hat{\boldsymbol{\beta}} | \boldsymbol{Y}\right] = \sigma^2_{X|Y} \left(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}\right)^{-1}, \tag{8}$$

where $\mathbb{C}\left[\hat{\boldsymbol{\beta}} | \boldsymbol{Y}\right]$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{Y}$ and $\sigma^2_{X|Y}$ is the conditional variance of X given Y.

D. Non-parametric regression

In the case of non-parametric regression, no prior assumption on the form of the function f is made, and we typically resort to various local or global smoothing techniques to estimate the regression function $\mathbb{E}[X|Y = y]$ [12]. In this paper, we consider the widely used kernel regression technique as an example, and we leave the extension to other non-parametric regression techniques for future works.

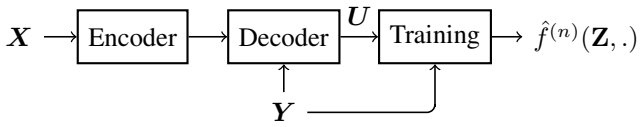
A one-dimensional kernel is any smooth, symmetric function $K : \mathbb{R} \to \mathbb{R}$ such that $\forall x \in \mathbb{R}, K(x) \geq 0$, and the following relations hold [38]

$$\int_{\mathbb{R}} K(x)dx = 1, \quad \int_{\mathbb{R}} xK(x)dx = 0, \quad \text{and} \quad 0 \leq \int_{\mathbb{R}} x^2 K(x)dx \leq \infty. \tag{9}$$

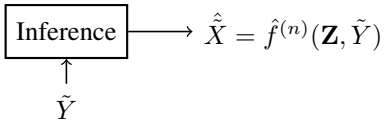
The Nadaraya-Watson Kernel regression over $(\boldsymbol{X}, \boldsymbol{Y})$ is defined as [12]:

$$\hat{f}(Y) = \frac{\sum_{j=1}^n K\left(\frac{Y - Y_j}{h}\right) X_j}{\sum_{j=1}^n K\left(\frac{Y - Y_j}{h}\right)}. \tag{10}$$

Here h is a positive number referred to as the bandwidth. Essentially, $\hat{f}$ represents a local average of $\boldsymbol{Y}$ based on the kernel K. The choice of the kernel and of the parameter h have been the subject of extensive research in statistical learning. It has been shown that estimators using different kernels have similar performance in terms of estimation loss $\mathbb{E}\left[(X - \hat{f}(Y))^2\right]$, while the choice of the bandwidth, which controls the smoothing, is of greater significance [39]. But



(a) Training phase



(b) Inference phase

Figure 1: Coding scheme for regression, with one training phase (a) over the learning sequence $\mathbf{Z} = (\mathbf{U}, \mathbf{Y})$ which provides a predictor $\hat{f}^{(n)}(\mathbf{Z}, .)$, and one inference phase (b) which consists of applying the predictor on new samples $\tilde{Y}$.

our theoretical results are generic and will apply to different kernels K and to a range of values for h.

III. Coding scheme for regression

In this section, we describe the coding setup we consider for regression, with one training phase and one inference phase. We then introduce the generalization error used to evaluate the regression performance and provide formal definitions of the considered coding scheme.

A. Training and inference phases

Regression, as a standard supervised learning problem, comprises one training phase and one inference phase, as shown in Figure 1. A training sequence $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) \in \mathcal{Z}^n$ of length n is built using the available side information $\boldsymbol{Y}$ and a coded representation of $\boldsymbol{X}$, denoted as $\boldsymbol{U}$. The training phase aims to estimate the function f on $\mathbf{Z}$ from either parametric or non-parametric regression. It provides a sequence of functions, called predictors, denoted as $\hat{f}^{(n)} : \mathcal{Z}^n \times \mathbb{R} \to \mathbb{R}$. It is important to mention that the training sequence involves the coded sequence $\boldsymbol{U}$ because the decoder does not have direct access to $\boldsymbol{X}$. Therefore, (7) and (10) need to be updated so as to account for $\boldsymbol{U}$, as will be described in Section IV.

Next, we use $\tilde{X}$ and $\tilde{Y}$ to denote random variables from the inference phase, where the pair $(\tilde{X}, \tilde{Y})$ follows the same probability distribution P_{XY} of the pair (X, Y) while being independent from it. At the inference phase, the decoder uses the predictor $\hat{f}^{(n)}$ to produce estimates of $\tilde{X}$

as $\hat{\tilde{X}} = \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y})$. It is worth noting that this does not require any data transmission, since the side information $\tilde{Y}$ is already available to the decoder. Therefore, this paper investigates the coding scheme for the training phase only, while the performance of this scheme is evaluated over the inference phase with the generalization error.

B. Generalization error

Usually, the performance of a lossy source coding scheme is evaluated from a distortion measure. Since here the objective of the receiver is also to learn a regression function, we need to consider additional metrics relevant for the regression problem. For that purpose, we use the notions of expected loss and generalization error already considered in [3].

A quadratic loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(x, \hat{x}) = (x - \hat{x})^2$ is considered. The expected loss L is defined as:

$$L(f) = \mathbb{E}\left[\ell(X, f(Y))\right]. \tag{11}$$

For a given regression problem, let $\mathcal{F}$ represents the set of regression functions of the form $f : \mathbb{R} \to \mathbb{R}$, with a predefined form (parametric regression) or free of specific assumptions (non-parametric regression). For instance, for polynomial regression, $\mathcal{F}$ is the set of all polynomial functions of a fixed order k. The minimum expected loss $L^\star$ is then given by:

$$L^\star(\mathcal{F}) = \inf_{f \in \mathcal{F}} L(f). \tag{12}$$

Next, the generalization error is defined as:

$$G(\hat{f}^{(n)}, \boldsymbol{Z}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) \mid \boldsymbol{Z}\right]. \tag{13}$$

The generalization error defined in (13) is a random variable due to the conditioning on $\boldsymbol{Z}$. Therefore, in what follows, we will also resort to the expected generalization error $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\hat{f}^{(n)}, \boldsymbol{Z}\right)\right]$.

The minimum expected loss defined in (12) is reached for the function $f^\star$ that minimizes the quantity $\mathbb{E}\left[\ell(X, f^\star(Y))\right]$ over the space of functions $\mathcal{F}$. However, there is no guarantee that this optimal function $f^\star$ can be estimated exactly from the training sequence $\mathbf{Z}$. Therefore, the generalization error measures the average quadratic loss which can be achieved for a specific training sequence $\boldsymbol{Z}$ and for a given predictor $\hat{f}^{(n)}(\boldsymbol{Z}, \cdot)$. The generalization error is evaluated as the expectation over the distribution $P_{\tilde{X}\tilde{Y}}$ of the MSE between the symbol $\tilde{X}$ and the estimated symbol $\hat{\tilde{X}} = \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})$. In our case, we assume that $(\tilde{X}, \tilde{Y})$ follows the same distribution as (X, Y), but (13) would also apply otherwise.

In addition, by bias-variance decomposition, it can be shown that the expected generalization error $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\hat{f}^{(n)}, \boldsymbol{Z}\right)\right]$ is lower bounded as

$$L^{\star}(\mathcal{F}) \leq \mathbb{E}_{\boldsymbol{Z}}\left[G\left(\hat{f}^{(n)}, \boldsymbol{Z}\right)\right]. \tag{14}$$

Therefore, the difference $\delta = \mathbb{E}_{\boldsymbol{Z}}\left[G\left(\hat{f}^{(n)}, \boldsymbol{Z}\right)\right] - L^{\star}(\mathcal{F})$ is a crucial quantity for characterizing the performance of a regression coding scheme. This is why our rate-generalization error regions defined in the next section will be expressed with this quantity.

C. Coding scheme

In [28], we introduced a coding scheme which was initially dedicated to the linear regression problem, by adapting definitions from [3]. But this scheme is actually generic enough to be adopted for any type of parametric regression, and for non-parametric regression as well. This is why we restate it here.

Definition 1. *A regression scheme at rate R is defined by a sequence $\{(e_n, d_n, R, t_n)\}$ with an encoder $e_n : \mathcal{X}^n \to [\![1, M_n]\!]$, a decoder $d_n : \mathcal{Y}^n \times [\![1, M_n]\!] \to \mathcal{U}^n$, and a learner $t_n : \mathcal{Y}^n \times \mathcal{U}^n \to \mathcal{F}$, such that*

$$\limsup_{n\to\infty} \frac{\log M_n}{n} \leq R.$$

Definition 2. *An (n, M, G) code for the sequence $\{(e_n, d_n, R, t_n)\}$ is a code with $|e_n| = M_n$ such that*

$$\mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \leq G \text{ and } \limsup_{n\to\infty} \frac{\log M_n}{n} \leq R. \tag{15}$$

Definition 3. *A pair (R, δ) is said to be achievable if an $(n, M, G)-$code exists such that*

$$\limsup_{n\to\infty} \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \leq L^{*}(\mathcal{F}) + \delta. \tag{16}$$

As discussed in the previous section and similar to the definition used in [3], the achievable region is defined in terms of the gap between $\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)})\right]$ and $L^{*}(\mathcal{F})$.

In this paper, we also consider the case where the decoder may either want to reconstruct the source X, or perform regression. In this case, the reconstruction task is evaluated with the standard quadratic distortion measure $d(x, \hat{x}) = (x - \hat{x})^2$, where $\hat{x}$ is the reconstruction of x at the decoder. We further define the coding scheme with both reconstruction and regression constraints as follows.

Definition 4. *An (n, M, D, G) code for the sequence $\{(e_n, d_n, R, t_n)\}$ is a code with $|e_n| = M$ such that*

$$\mathbb{E}\left[d(X, \hat{X})\right] \leq D \ , \ \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \leq G \ , \ \textit{and} \ \frac{\log M_n}{n} \leq R. \tag{17}$$

We now provide definitions for the finite-length analysis of the coding schemes.

Definition 5. *An (n, M, G, ε) code for the sequence $\{(e_n, d_n, R, t_n)\}$ and $\varepsilon \in (0, 1)$ is a code with $|e_n| = M_n$ such that*

$$\mathbb{P}\left[G(\hat{f}^{(n)}, \boldsymbol{Z}) \geq G\right] \leq \varepsilon \ \textit{and} \ \frac{\log M}{n} \leq R. \tag{18}$$

Definition 6. *For fixed G and block-length n, the finite block-length rate-loss function with excess loss ε is defined by:*

$$R(n, G, \varepsilon) = \inf_{R}\{\exists \ \ (n, M, G, \varepsilon) \ code\}. \tag{19}$$

Definition 7. *An $(n, M, D, G, \varepsilon)$ code for the sequence $\{(e_n, d_n, R, t_n)\}$ and $\varepsilon \in (0, 1)$ is a code with $|e_n| = M$ such that*

$$\mathbb{P}\left[\left\{d(X, \hat{X}) \geq D\right\} \ \cup \ \left\{G(\hat{f}^{(n)}, \boldsymbol{Z}) \geq G\right\}\right] \leq \varepsilon \tag{20a}$$

$$\frac{\log M}{n} \leq R. \tag{20b}$$

Definition 8. *For fixed D, G and block-length n, the finite block-length rate-distortion-generalization error functions with excess loss ε is defined by:*

$$R(n, D, G, \varepsilon) = \inf_{R}\{\exists \ \ (n, M, D, G, \varepsilon) \ code\}. \tag{21}$$

Definitions 1 to 3 will be used for the asymptotic analysis of Section IV, and are similar to what was initially introduced in [3] and later considered in [28]. On the other hand, Definitions 4 to 8 did not appear in [3], and will serve to investigate the trade-off between data reconstruction and regression, as well as for the finite block-length analysis in Section V.

IV. Asymptotic Analysis

In [3, Theorem 3.3], it is shown that, when considering a quadratic loss function, the expected generalization error can be lower and upper bounded as follows:

$$L^{\star}(\mathcal{F})^{\frac{1}{2}} \leq \limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})^{\frac{1}{2}}\right] \leq L^{\star}(\mathcal{F})^{\frac{1}{2}} + 2\mathbb{D}_{X|Y}(R)^{1/2}, \tag{22}$$

where $\mathbb{D}_{X|Y}(R)$ represents the conditional distortion-rate function [40]. It is worth noting that for regression problems considering the quadratic loss function, we have

$$L(\hat{f}) = \sigma^2 + \mathbb{E}\left[\left(\hat{f}(Y) - f(Y)\right)^2\right] \geq \sigma^2, \tag{23}$$

with equality iff $\hat{f} = f$. So the minimum expected loss defined in (12) is given by $L^\star(\mathcal{F}) = \sigma^2$. In this section, we propose a coding scheme which improves over the upper bound in (22), both for parametric and non-parametric regression.

A. Achievable rate-generalization error regions

The next two Theorems provide the rate-generalization error regions which can achieved for both parametric regression and kernel regression.

Theorem 1 (Parametric regression)**.** *Given any rate $R > 0$, the pair $(R, 0)$ is achievable for parametric regression with quadratic loss, for sources (X, Y) following the model in* (6)*.*

This Theorem generalizes results we obtained in [28], [29] (linear and polynomial regression), to any type of parametric regression described by (6). It states that the minimum generalization error $L^\star(\mathcal{F})$ can be achieved with arbitrary rate R, as long as the length n of the training sequence is large enough. Therefore, Theorem 1 improves over the result of [3] by eliminating the term $\mathbb{D}_{X|Y}(R)$ in the upper bound in (22). This makes our result tight in the sense that the upper bound equates the lower bound $L^\star(\mathcal{F})$ for any rate $R > 0$.

Theorem 2 (Kernel regression)**.** *Under the following conditions:*

i. Y is bounded almost surely

ii. the probability density function p_Y is continuously differentiable and positively lower bounded,

iii. the regression function f is twice continuously differentiable, i.e. f', and f'' exist,

iv. $h = h_n$ is a deterministic sequence such that when $n \to \infty$, the bandwidth h satisfies $h \to 0$ and $nh \to \infty$,

given any rate $R > 0$, the pair $(R, 0)$ is asymptotically achievable for the kernel regression with quadratic loss.

Like for parametric regression, the previous Theorem states that kernel regression over the pair $(\boldsymbol{U}, \boldsymbol{Y})$ can asymptotically achieve the same performance as when applied on original data $(\boldsymbol{X}, \boldsymbol{Y})$. In fact, in the case of kernel regression, we will show that the generalization error

can be divided into three parts: a first part for the intrinsic noise N given by the term σ^2, a second and third part related to the bias and the variance of the estimator. We show that the last two terms go to 0 as n goes to infinity because of condition iv in Theorem 2, in particular. In addition, the proof for the rate-generalization error region for kernel regression differs from the case of parametric regression, given that in the later case no prior assumption on the regression function is considered. But the conclusion is still that the gap $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\hat{f}^{(n)}, \boldsymbol{Z}\right)\right] - L^{\star}(\mathcal{F})$ tends to 0 as n goes to infinity. We leave for future works the investigation of other methods, like local polynomial regression, which could further reduce the bias for finite n.

B. Proof of Theorem 1 and Theorem 2

We now briefly describe the achievability scheme that is considered in the proofs of Theorem 1 and Theorem 2. We then provide expressions as well as convergence analysis of the generalization error for both parametric and kernel regression, since those constitute our technical contribution for the asymptotic case.

In our considered achievability scheme, we make use of a Gaussian test channel described by $U = \alpha(X + \Phi)$, where $\Phi \sim \mathcal{N}(0, \sigma_{\Phi}^2)$ is independent of X. The parameters α and σ_{Φ} are constant and depend on the distributions of X and Y. We then consider the achievability scheme proposed by Draper in [33] for Wyner-Ziv coding in the case where the distribution P_{XY} is unknown. This scheme is based on quantization and binning, and provides a criterion on empirical information density for debinning. We also use this criterion, but do not consider the incremental coding strategy of [33] which is not necessary here as the coding rate is fixed given that σ^2 is known. This scheme is described into details in Appendix B. The results of [33] show that the sequence $\boldsymbol{U}$ can be reconstructed by the decoder with vanishing error probability as n tends to infinity. We next demonstrate that the Gaussian test channel allows us to achieve the optimal rate-generalization error region for both parametric regression and kernel regression, by expressing the generalization error in both cases.

1) Convergence analysis of the generalization error in Theorem 1: For the parametric regression model described in (6), the OLS estimator applied over the pair $(\boldsymbol{U}, \boldsymbol{Y})$ is given by

$$\hat{\boldsymbol{\beta}} = \alpha^{-1}(\underline{\boldsymbol{Y}}^{\star}\underline{\boldsymbol{Y}}^{\star T})^{-1}\underline{\boldsymbol{Y}}^{\star}\boldsymbol{U}, \tag{24}$$

and it has the following properties:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \quad \text{and} \quad \mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right] = \frac{1}{\alpha^2}\sigma_{U|Y}^2(\underline{\boldsymbol{Y}}^{\star}\underline{\boldsymbol{Y}}^{\star T})^{-1}. \tag{25}$$

Note that this differs from what was defined in (7) and (8) since the decoder has no direct access to $\boldsymbol{X}$. From (13) and (24), the generalization error can be expressed as

$$\begin{aligned} G(\hat{f}^{(n)}, \boldsymbol{Z}) &= \mathbb{E}_{\tilde{X}\tilde{Y}}\left[[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T} [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + N^2 | \boldsymbol{Z}\right] \\ &= [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T}\right] [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \sigma^2, \end{aligned} \tag{26}$$

where $\tilde{\boldsymbol{Y}}^\star = [h_0(\tilde{Y}), ..., h_{k-1}(\tilde{Y})]^T$ refers to the vector composed of $h_i(\tilde{Y}), \forall i \in [\![0, k-1]\!]$ independent from Y. By defining $\underline{\tilde{\boldsymbol{\Sigma}}} = \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T}\right]$ and $\underline{\boldsymbol{\Sigma}} = \frac{1}{n}\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}$, the expected generalization error can be expressed as

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \\ &= \sigma^2 + \mathbb{E}\left[\frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))^T \underline{\tilde{\boldsymbol{\Sigma}}} \frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))\right] \\ &= \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}\mathbb{E}\left[\mathrm{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)\right] &(27) \\ &\leq \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}\mathbb{E}\left[\frac{k\lambda_{\max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{\min}(\underline{\tilde{\boldsymbol{\Sigma}}}) - ||\underline{\tilde{\boldsymbol{\Sigma}}} - \underline{\boldsymbol{\Sigma}}||}\right] &(28) \\ &\leq \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}\mathbb{E}\left[\frac{k\lambda_{\max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{\min}(\underline{\tilde{\boldsymbol{\Sigma}}})}\right] &(29) \\ &\leq \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}kC, &(30) \end{aligned}$$

where $C = \frac{\lambda_{max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{min}(\underline{\tilde{\boldsymbol{\Sigma}}})}$ is a constant. When $n \to \infty$, the generalization error $\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right]$ converges to σ^2, which completes the convergence analysis of Theorem 1.

2) Convergence analysis of the generalization error in Theorem 2: Given the fact that the kernel regression is applied on the pair $(\boldsymbol{U}, \boldsymbol{Y})$, and according to the Gaussian test channel $U = \alpha(X + \Phi)$, (10) can be rewritten as:

$$\hat{f}(y) = \frac{\sum_{i=1}^n K(\frac{y - y_i}{h})\frac{u_i}{\alpha}}{\sum_{i=1}^n K(\frac{y - y_i}{h})}. \tag{31}$$

We now provide the main key steps of the convergence analysis of the generalization error. The details of the derivation are provided in Appendix C. For a given pair $(\tilde{x}, \tilde{y})$, the so-called test error can be expressed as:

$$\mathbb{E}_{\boldsymbol{Z}}\left[(\hat{f}^{(n)}(\tilde{y}, \boldsymbol{Z}) - f(\tilde{y}))^2\right] = b_n^2(\tilde{y}) + V_n(\tilde{y}), \tag{32}$$

where $b_n(\tilde{y}) = \mathbb{E}\left[\hat{f}^{(n)}(\tilde{y}, \boldsymbol{Z}) - f(\tilde{y})\right]$ is the bias and $V_n(\tilde{y}) = \mathbb{V}\left[\hat{f}^{(n)}(\tilde{y}, \boldsymbol{Z})\right]$ is the variance of the estimator $\hat{f}^{(n)}$ with respect to the training sequence $\mathbf{Z}$. The expected generalization error is then given by

$$\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] = \mathbb{E}_{\tilde{X}\tilde{Y}\boldsymbol{Z}}\left[(\hat{f}^{(n)}(\tilde{Y}, \boldsymbol{Z}) - \tilde{X})^2\right] \tag{33}$$

$$= \sigma^2 + \int b_n^2(\tilde{y}) p_Y(\tilde{y}) d\tilde{y} + \int V_n(\tilde{y}) p_Y(\tilde{y}) d\tilde{y}. \tag{34}$$

For a given $\tilde{y}$, by analyzing the convergence of the numerator and the denominator of $\hat{f}(y)$ in (31), it is shown in Appendix C that

$$b_n(\tilde{y}) = \frac{h^2}{2}\left(2\frac{f'(\tilde{y})p'_Y(\tilde{y})}{p_Y(\tilde{y})} + f''(\tilde{y})\right)\int_{\mathbb{R}} u^2 K(u) du + o(h^2), \tag{35}$$

$$V_n(\tilde{y}) = \frac{(\sigma^2 + \sigma_\Phi^2)}{p_Y(\tilde{y}) nh}\int_{\mathbb{R}} K^2(u) du + o\left(\frac{1}{nh}\right). \tag{36}$$

Finally, as $n \to \infty, h \to 0$ and $nh \to 0$ (by condition iv in Theorem 2), $\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right]$ in (33) tends to σ^2.

C. Comparison with existing works

First, we point out that authors in [41] also proposed a Gaussian approximation of the quantization error under which the MSE of an estimator (not dedicated to regression) applied to compressed data is equivalent to the same estimator when applied to a corrupted version of data by a Gaussian noise. This is in line with our achievable scheme.

We now comment on our improvement of the upper bound of Raginsky in [3]. First of all, the results of [3] are stated for a generic loss function ℓ which can then be specified for various learning problems including regression or classification. In [3], the empirical loss $\hat{L}_{\boldsymbol{X},\boldsymbol{Y}}(f)$ for a certain function f is defined as

$$\hat{L}_{\boldsymbol{X},\boldsymbol{Y}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i), \tag{37}$$

and the difference between $\hat{L}_{\boldsymbol{X},\boldsymbol{Y}}(f)$ and $\hat{L}_{\boldsymbol{U},\boldsymbol{Y}}(f)$ is upper bounded as

$$\sup_f |\hat{L}_{\boldsymbol{X},\boldsymbol{Y}}(f) - \hat{L}_{\boldsymbol{U},\boldsymbol{Y}}(f)| \leq \eta(d(\boldsymbol{U}, \boldsymbol{X})), \tag{38}$$

where d is a distortion measure and η is a concave continuous function. Taking the expectation of (38) as well as further mathematical manipulation lead to the upper bound on the generalization error in (22), especially given that $\mathbb{E}[d(\boldsymbol{U}, \boldsymbol{X})] = \mathbb{D}_{X|Y}$. But in our case, expressing for instance

the generalization error for parametric regression with quadratic loss and with the OLS estimator defined in (24) gives that the term $\mathbb{E}\left[d(\boldsymbol{U},\boldsymbol{X})\right]$ (which is σ_Φ^2 in (30)) is multiplied by a factor $1/n$ and therefore vanishes as n goes to infinity. This is why we get that the generalization error converges to the minimum expected loss σ^2. As a result, the upper bound of [3] is not tight in our setup, but on the other hand it applies to a larger range of learning problems.

In addition, consider an alternative regression problem that is to infer a function g such that $U = \alpha(g(Y)+N)+\Phi$, with $\Phi \sim \mathcal{N}(0,\mathbb{D}_{X|Y})$. For this alternative problem, the minimum expected loss (12) expressed for $L(g)=\mathbb{E}[\ell(g(U),Y)]$ would be given by $\sigma^2+\mathbb{D}_{X|Y}$, and hence we would retrieve the upper bound of (22). But here, since the target is to estimate the function f such that $Y=f(X)+N$, it turns out that the minimum expected loss σ^2 can be achieved, despite applying regression on the pair $(\boldsymbol{U},\boldsymbol{Y})$.

D. Regression-reconstruction trade-off

Consider the achievability scheme described in Section IV-B for conventional Wyner-Ziv coding for reconstruction. This scheme achieves the Wyner-Ziv rate-distortion function provided in [42], that is $R_{WZ}(D)=\inf I(X;U|Y)$ where the $\inf$ is taken over $p_{U|X}(u|x)$ and is such that the Markov chain $X \leftrightarrow U \leftrightarrow Y$ holds. Given that we consider this same achievability scheme in our proofs of Theorem 1 and Theorem 2, we can formulate a rate-distortion-generalization error function as follows:

$$R\left(D,G\right) = \inf_{\substack{p(u\mid x):\\ \mathbb{E}\left[d(X,\hat{X})\right]\leq D\\ \mathbb{E}\left[G(\hat{f}^{(n)},\boldsymbol{Z})\right]\leq G}} I(X;U|Y). \tag{39}$$

The next Corollary investigates the trade-off in $R(D,G)$ between the two constraints $\mathbb{E}\left[d(X,\hat{X})\right]\leq D$ and $\mathbb{E}\left[G(\hat{f}^{(n)},\boldsymbol{Z})\right]\leq G$.

Corollary 1 (Asymptotic trade-off for parametric and kernel regression)**.** *For a pair of sources (X,Y) modeled from* (5)*, and for some non-negative constants D and $G\geq\sigma^2$, we have*

$$R(D,G)=R_{WZ}(D) \tag{40}$$

for both parametric and kernel regression.

Proof. See Appendix D. ■

This results shows that the previous achievability scheme which minimizes the generalization error for regression can also achieve the optimal Wyner-Ziv rate-distortion function for reconstruction. Therefore, asymptotically there is no trade-off in terms of coding rate between reconstruction and regression. Note that this result differs from existing ones in the literature, which show that there is a trade-off between data reconstruction and other specific tasks, such as in the distortion-perception problem [31], for semantic communications [19], [20], [30], for parameter estimation [21], and for hypothesis testing [22].

The next section provides finite block-length rate-generalization error regions, and also investigates the trade-off between regression and reconstruction at finite length.

V. Finite block-length analysis

The non-asymptotic source coding problem with a distortion constraint and without side information was first investigated in [34], [43] using the notions of information density and dispersion region. This non-asymptotic analysis was also extended to the case with side information at the decoder in [36]. The main idea behind these analysis is to approximate the distribution of error events by the Berry-Esséen Theorem and to bound the resulting approximation error. In this section, we extend these tools so as to also treat the regression problem. Especially, in finite block-length analysis, the excess probability ϵ which appears in Definition 5 plays a crucial role as not all the codewords satisfy the generalization error constraint.

In what follows, we directly address the source coding problem with the two objectives of data reconstruction and regression, and investigate the trade-off between these two tasks. The proposed analysis applies to both parametric and kernel regression.

A. Definitions

Let us consider the following four sets:

$$\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}}) := \{(u,y) : \iota\,(u,y) \geq \gamma_{\mathrm{p}}\}\,, \tag{41}$$

$$\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}) := \{(u,x) : \iota\,(u,x) \leq \gamma_{\mathrm{c}}\}\,, \tag{42}$$

$$\mathcal{T}_{d}(D) := \{(x,\hat{x}) : d(x,\hat{x}) \leq D\}\,, \tag{43}$$

$$\mathcal{T}_{g}(G) := \left\{(\boldsymbol{u},\boldsymbol{y}) : \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X},\hat{f}^{(n)}(\boldsymbol{z},\tilde{Y}))\right] \leq G\right\}, \tag{44}$$

where ι is the information density defined in (2), γ_p, γ_c are predefined thresholds, and G, D are the generalization error and distortion constraint, respectively. The first three sets already appeared in [36] for the conventional setup of data reconstruction with side information, while we introduce the last one specifically for the analysis of the generalization error of the regression problem.

Accordingly, we define the information-density-distortion-generalization error vector as follows:

$$\boldsymbol{i}(X, \boldsymbol{U}, \boldsymbol{Y}, \hat{X}) := \begin{bmatrix} -\iota\left(U, Y\right) \\ \iota(U, X) \\ d(X, \hat{X}) \\ \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right] \end{bmatrix}, \tag{45}$$

where $\boldsymbol{U}$ represents the full training sequence of length n, while U refers to one variable within this sequence. The same applies for $\boldsymbol{Y}$ and Y. Taking the expectation over the distribution $P_{X\boldsymbol{U}\boldsymbol{Y}\hat{X}}$ of this vector gives

$$\boldsymbol{J}(\boldsymbol{i}) := \mathbb{E}\left[\boldsymbol{i}(X, \boldsymbol{U}, \boldsymbol{Y}, \hat{X})\right] = \begin{bmatrix} -I(U;Y) \\ I(U;X) \\ \mathbb{E}\left[d(X, \hat{X})\right] \\ \mathbb{E}_{\boldsymbol{Z}\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right] \end{bmatrix}, \tag{46}$$

where the sum of the first two components provides the Wyner-Ziv coding rate. The covariance matrix of this vector is defined as

$$\underline{\boldsymbol{V}} = \mathbb{C}\left(\boldsymbol{i}(X, \boldsymbol{U}, \boldsymbol{Y}, \hat{X})\right). \tag{47}$$

Let $\underline{\boldsymbol{V}} \in \mathbb{R}^{4\times 4}$ be a positive-semi-definite matrix. Given a Gaussian random vector $\boldsymbol{B} \sim \mathcal{N}(0, \underline{\boldsymbol{V}})$, the dispersion region is defined with respect to the covariance matrix as [44]

$$\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon) := \{\boldsymbol{b} \in \mathbb{R}^4 : \Pr(\boldsymbol{B} \leq \boldsymbol{b}) \geq 1 - \varepsilon\}. \tag{48}$$

B. Non-asymptotic regions

The non-asymptotic achievability regions can be obtained by the method of channel resolvability [45], [46], or by mutual covering lemmas proposed in [47]. In our case, we consider the former analysis and its extension to the case with side information [36]. In our proofs, we adapt the analysis of [36] by further considering the regression problem through the generalization

error, and by taking into consideration the fourth set $\mathcal{T}_g(G)$ in (44). This led to the following two Theorems.

Theorem 3 (Non-asymptotic achievable code)**.** *For arbitrary constants $\gamma_p, \gamma_c, D, G \geq 0$, and positive integer N, there exists an $(n, M, D, G, \varepsilon)$ code satisfying*

$$\varepsilon \leq \mathbb{P}_{X\boldsymbol{U}\boldsymbol{Y}\hat{X}}\left[(u,y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c \cup (u,x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c \cup (x,\hat{x}) \in \mathcal{T}_{\mathrm{d}}(D)^c \cup (\boldsymbol{u},\boldsymbol{y}) \in \mathcal{T}_g(G)^c\right] + \frac{N}{2^{\gamma_{\mathrm{p}}}|\mathcal{M}|} + \frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{N}}. \tag{49}$$

Proof. The proof is provided in appendix E. ■

By choosing $\gamma_p = \log \frac{N}{|\mathcal{M}_n|} + \log n$ and $\gamma_c = \log N - \log n$, and by applying Theorem 3 together with the multidimensional Berry-Esséen Theorem, we derive the achievable second-order rate-distortion-generalization error region as follows.

Theorem 4 (Second-order coding rate)**.** *For every $0 < \varepsilon < 1$, and n sufficiently large, the (n, ε)-rate-distortion-generalization error function satisfies:*

$$R(n, \varepsilon, D, G) \leq \inf\left\{\boldsymbol{M}\left(\boldsymbol{J} + \frac{\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon)}{\sqrt{n}} + \frac{2\log n}{n}\mathbf{1}_4\right)\right\}, \tag{50}$$

with $\boldsymbol{M} = [1 \quad 1 \quad 0 \quad 0]$.

Proof. The proof is provided in appendix F. ■

The previous result is not a straightforward extension of the proofs in [36] as we introduce the generalization error term in the information-density-distortion-generalization error vector $\boldsymbol{i}$. Especially, the vector $\boldsymbol{i}$ depends on the full sequence $\mathbf{Z}$ (it is not single-letter anymore), because the generalization error depends on the full training sequence. Therefore, the Berry-Esséen Theorem needs to be applied by conditioning on the other $n-1$ training samples.

Finally, the bound in Theorem 4 is composed by two parts. The vector $\boldsymbol{J}$ corresponds to the asymptotic result of Section IV-D, while the other terms provide the non-asymptotic penalty introduced by the Gaussian approximation.

C. Non-asymptotic trade-off

As in the asymptotic regime, we now investigate the trade-off between distortion and generalization error. To proceed, we further investigate the dispersion region to explore the relation between regression and reconstruction.

Corollary 2 (Non-asymptotic trade-off for parametric and kernel regression)**.** *For a finite sequence* $(\boldsymbol{X}, \boldsymbol{Y})$ *following the model in* (5)*,* $0 < \varepsilon < 1$ *and* n *sufficiently large, there exists an achievable rate-distortion-generalization error region such that*

$$R_b(n, G, D, \varepsilon) > \max\{R_b(n, G, \varepsilon), R_b(n, D, \varepsilon)\}, \tag{51}$$

where $R_b(\cdot)$ *denotes the minimum achievable rate introduced by the right hand side of* (50)*.*

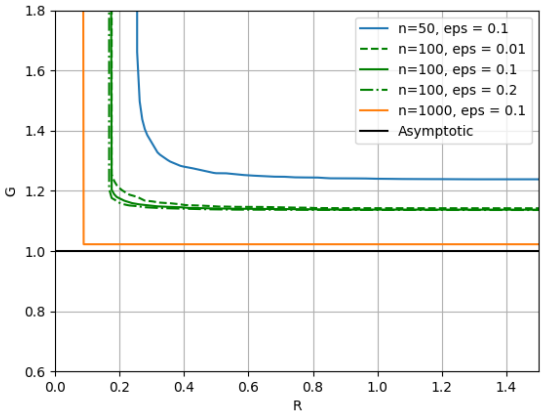
Proof. The proof is provided in Appendix G. ■

The previous result shows that for the considered achievability scheme, there is no trade-off in terms of coding rate between the distortion and the generalization error. In order to prove Corollary 2, we first showed that the terms $d(X, \hat{X})$ and $\mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right]$ in (45) are decorrelated, which turns into conditional independence with respect to the first two terms of the matrix, due to the Gaussian approximation from the Berry-Esséen Theorem. However, there is no guarantee that we can achieve the minimum for both criterion in finite block-length because of the excess probability constraint. Note also that this result is specific to the considered achievable coding scheme. However, the approach of analyzing the correlation between terms in the information-density vector could be applied to other achievability schemes. Finally, the analysis in Appendix G highly depends on the Gaussian test channel we have chosen: the assumption that quantization error Φ and the system noise N are independent from the source X and Y plays a vital role in the calculation of the correlation.

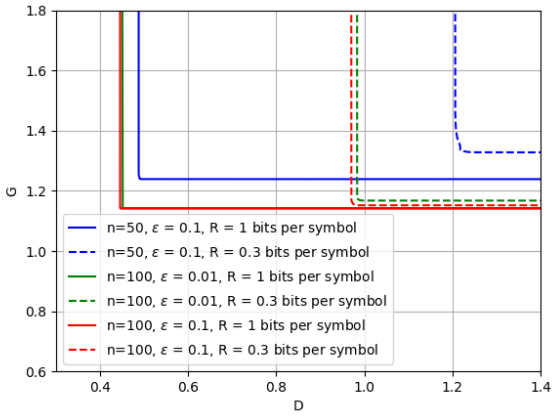
VI. Numerical results

In this section, we provide numerical evaluations of the finite-length achievable rate-distortion-generalization error regions provided in Section V. As a particular case, we consider that X and Y follow a polynomial relation defined by $X = \boldsymbol{\beta}^T \boldsymbol{Y}^* + N$, where $\boldsymbol{\beta} = [2, 1, 1]^T$, $\boldsymbol{Y}^* = [Y^0, Y^1, Y^2]$, $\sigma = 1$, and Y is uniformly distributed over $[-1, 1]$. We provide the finite-length achievable regions obtained from Theorem 4 for both parametric and non-parametric kernel regression. In the latter case, we consider a Gaussian kernel, and the bandwidth h_n for different block-length n is set as $h_n = \left(\frac{(\sigma^2+\sigma_\Phi^2)C_2}{C_1} n\right)^{-\frac{1}{5}}$. This value is known to be optimal in the asymptotic regime [12], but it might be sub-optimal for minimizing the expected generalization error at finite length.

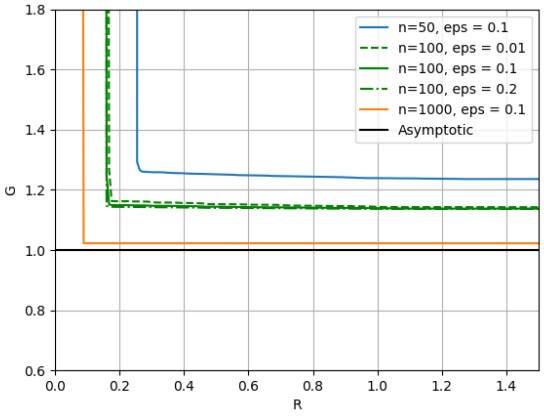
In both parametric and kernel regressions, the covariance matrix $\underline{\boldsymbol{V}}$ defined in equation (47) has to be estimated. To do so, we sample information-density-distortion-generalization error vectors



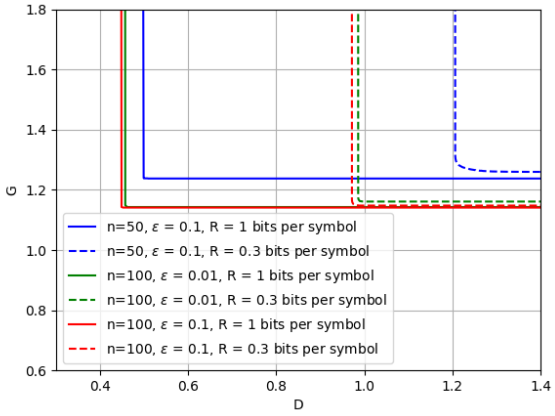
(a) Rate-generalization error region for polynomial regression labeled on the block-length n and the excess loss probability ε.



(b) Distortion-generalization error region for polynomial regression on the block-length n, the excess loss probability ε and rate R.



(c) Rate-generalization error region for kernel regression labeled on the block-length n and the excess loss probability ε.



(d) Distortion-generalization error region for kernel regression on the block-length n, the excess loss probability ε and rate R.

Figure 2: Non-asymptotic rate-distortion-generalization error region

$\boldsymbol{i}$ in (45) by first generating n samples of X and Y, and then calculating the four components of the vector. In order to do so, we need the following prior results:

The probability density function of U: by the Theorem of variable change, for $\beta_2 > 0$ and $\beta_1^2 + 4\beta_2(w - \beta_0) \geq 0$, we can show that the distribution of $W = \boldsymbol{\beta}^T \boldsymbol{Y}^\star$ is given by:

$$P_W(w) = \begin{cases} \frac{1}{\sqrt{\beta_1^2+4\beta_2(w-\beta_0)}} & |y_1(w)| \leq 1 \text{ and } |y_2(w)| \leq 1, \\ \frac{1}{2\sqrt{\beta_1^2+4\beta_2(w-\beta_0)}} & |y_1(w)| \leq 1 \text{ or } |y_2(w)| \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

where $y_1 = \frac{-\beta_1 - \sqrt{\beta_1^2+4\beta_2(w-\beta_0)}}{2\beta_2}, y_2 = \frac{-\beta_1 + \sqrt{\beta_1^2+4\beta_2(w-\beta_0)}}{2\beta_2}$. The probability density function of $U = \alpha(W + N + \Phi)$ can then be expressed as

$$P_U(u) = \frac{1}{\alpha\sqrt{2\pi(\sigma^2 + \sigma_\Phi^2)}} \int_{-\infty}^{\infty} P_W(w) e^{-\frac{(\frac{u}{\alpha} - w)^2}{2(\sigma^2 + \sigma_\Phi^2)}} dw, \tag{52}$$

which can be evaluated numerically;

The conditional distribution of $(U|X)$ and $(U|Y)$: by the test channel defined in Section IV-B, we have $(U|Y) \sim \mathcal{N}(0, \alpha^2(\sigma^2 + \sigma_\Phi^2))$ and $(U|X) \sim \mathcal{N}(0, \alpha^2\sigma_\Phi^2)$.

Then, the main steps of the numerical evaluation of the covariance matrix $\underline{\boldsymbol{V}}$ are as follows:

1) Generate n samples of X, Y and U, according to the Gaussian test channel defined in Section IV-B
2) For each sample (u, x, y), the information densities $\iota(x; u)$ and $\iota(u; y)$ are obtained with (2);
3) The distortion is calculated by:

$$d(x, \hat{x}) = (\hat{x} - x)^2 . \tag{53}$$

4) The generalization error $G(\hat{f}^{(n)}, \boldsymbol{Z})$ is given by (13), where the expectation is estimated with $N^\star = 500$ samples for kernel regression, and is directly calculated by (26) for parametric regression.
5) Repeat steps 1) to 4) $N^\star = 500$ times to get numerical estimation of the covariance matrix (47).

Then the achievable region is obtained by Theorem 4.

In addition, Figures 2a and 2c show the boundaries of the rate-generalization error regions for polynomial and kernel regressions, considering different block-length n and excess probability ε. In both cases, we observe that the achievable regions converges to the asymptotic one as n increases, and we also observe that lower rates can be achieved if higher excess probabilities

are allowed. Figures 2b and 2d illustrate the distortion-generalization error region for coding rates $R = 0.3$ bit/symbol and $R = 1$ bit/symbol. The regions are consistent with our Corollary 2 which states that the decorelation between the distortion and the generalization error results in the absence of trade-off between the two criterions. In addition, we observe that both distortion and generalization error decrease with the coding rate R.

Finally, for fixed rate R and excess probability ϵ, we see that the generalization error of OLS estimator converges faster than the generalization error of kernel estimators, which is consistent with the different convergence rates of these two types of regression. Especially it is shown in [12] that for kernel regression, the optimal h is of order $O\left(n^{-\frac{1}{5}}\right)$ and that the expected generalization error decreases to the minimum expected loss σ^2 at rate $O\left(n^{-\frac{4}{5}}\right)$. On the opposite, in parametric methods, the generalization error decreases to σ^2 at rate $O\left(n^{-1}\right)$. The slower rate $O\left(n^{-\frac{4}{5}}\right)$ is the price of using non-parametric methods.

VII. Conclusion

In this article, we investigated regression under the generalization error criterion within the framework of goal-oriented communications. Our information-theoretic analysis provided rate-generalization error regions for parametric regression and kernel regression in both asymptotic and non-asymptotic regimes. We improved upon existing bounds [3] in the asymptotic regime, demonstrating convergence of generalization error to the minimum expected loss. In the non-asymptotic regime, we relied on the finite-length tools introduced in [36] and extended these tools to our regression problems. We further investigated the trade-off between regression and reconstruction, and as a key finding of our research, we showed that, in both cases (asymptotic and non-asymptotic), there is no trade-off between reconstruction and regression. A posterior remark of this result is that for both reconstruction and regression, we used the same test-channel. The established optimality of this test channel in infinite block-length further solidified our findings. The converse in the non-asymptotic regime remains an open question, inviting further exploration in future works.

Appendix A
Preliminary Theorems

Here we restate the channel resovability problem [48, Chapitre 6] and related definitions used in [36]. The statements of [36] apply for discrete source, and this appendix generalizes them to

arbitrary distributions.

A. Smoothing of a distribution

Denote $\mathscr{P}(\mathcal{X})$ as the set of all probability distributions on a measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$, and let $\mathscr{P}'(\mathcal{X})$ be the set of all sub-normalized non-negative functions (not necessarily a probability measure). Note that if $P \in \mathscr{P}'(\mathcal{X})$ is normalized then $P \in \mathscr{P}(\mathcal{X})$. For a subset $\mathcal{T} \subset \mathcal{X}$, the smoothed sub-normalized function $\bar{P}_X$ of P_X is defined as, $\forall A \in \mathscr{B}(\mathcal{X})$

$$\bar{P}_X(A) = \int_A \mathbf{1}[x \in \mathcal{T}] p_X(x) dx. \tag{54}$$

For two functions $P, Q \in \mathscr{P}'(\mathcal{X})$, the variational distance between P and Q is:

$$d_{TV}(P, Q) = \sup_{A \in \mathscr{B}(\mathcal{X})} |P(A) - Q(A)|, \tag{55}$$

and it has the following property.

Lemma 1 (Property of variational distance [36])**.** *For a distribution $P \in \mathscr{P}(\mathcal{X})$ and a sub-normalized measure $Q \in \mathscr{P}'(\mathcal{X})$, and any subset Γ of $\mathcal{X}$,*

$$P(\Gamma) \leq Q(\Gamma) + d_{TV}(P, Q) + \frac{1 - Q(\mathcal{X})}{2}. \tag{56}$$

Hence the variational distance between the original distribution and a smoothed one is

$$d_{TV}(P, \bar{P}) = \frac{P(\mathcal{T}^c)}{2}, \tag{57}$$

where $\mathcal{T}^c$ stands for the complementary set of $\mathcal{T}$. For a channel $P_{U|X} : \mathcal{X} \to \mathcal{U}$ a subset $\mathcal{T} \subset \mathcal{X} \times \mathcal{U}$, and the event $B \in \mathscr{B}(\mathcal{U})$ and $x \in \mathcal{X}$, the smoothed conditional function $\bar{P}_{U|X}$ is defined by

$$\bar{P}_{U|X}(B|X = x) = \int_B \mathbf{1}[(u, x) \in \mathcal{T}] p_{U|X}(u|x) du. \tag{58}$$

B. Channel resolvability and identification code

Let us consider a channel $P_{U|X}$ and an input distribution P_X. In the channel resolvability problem, we choose M elements in the input set $\mathcal{X}$, *i.e.*, a codebook $\mathcal{C} = \{x_1, x_2, ..., x_M\}$, such that the output distribution P_U expressed from the input distribution P_X as

$$P_U(B) = \int_B \int_{\mathcal{X}} p_X(x) p_{U|X}(u|x) du dx \tag{59}$$

is close enough to the output distribution $P_{U'}(B)$ obtained when the input is assumed to be uniformly distributed [45], [46], i.e.

$$P_{U'}(B) = \int_B \sum_{i=1}^{M} \frac{\mathbf{1}[x = x_i]}{M} p_{U|X}(u|x) du, \tag{60}$$

where we suppose the channel $P_{U|X}$ is absolutely continuous. By the soft covering lemma from [46, Corollary 7.2] and [45, Lemma 2], the following result states that

Corollary 3 (Lemma 25 of [36])**.** *Let $\mathcal{T} = \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})$ defined in* (42)*, for any $\gamma_{\mathrm{c}} \geq 0$, we have*

$$\mathbb{E}_{\mathcal{C}}\left[d_{TV}(\bar{P}_U, \bar{P}_{U'})\right] \leq \frac{\Delta(\gamma_{\mathrm{c}}, P_{UX})}{2\sqrt{M}}, \tag{61}$$

with $\Delta(\gamma_{\mathrm{c}}, P_{UX}) = \sqrt{\mathbb{E}_{UX}\left[\frac{dP_{U|X}(u|x)}{dP_U(u)} \mathbf{1}[(u,x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})]\right]}$.

APPENDIX B
GAUSSIAN TEST CHANNEL AND CODING SCHEME

Consider the test channel $U = \alpha(X + \Phi)$ defined in Section IV-B. Since we assume that the function f and the joint distribution P_{XY} are unknown in both the encoder and the decoder, we employ the achievable scheme proposed in [33] based on the method of types and binning. However, compared to [33], no incremental transmission is needed since we suppose that the noise variance σ^2 is known. Therefore, the test channel parameters as well as the binning rate are fixed. In fact, by setting

$$\alpha = \frac{\sigma^2 - D}{\sigma^2} \text{ and } \sigma_\Phi^2 = \frac{D\sigma^2}{\sigma^2 - D} \tag{62}$$

the distortion constraint $\mathbb{E}\left[d(X, \hat{X})\right] \leq D$ can be achieved for Gaussian source [42]. Thus the variable-rate scheme in [33] becomes a fixed rate coding scheme. However, we need to keep the prefix transmission of types applied by Draper [33] since the joint distribution P_{XY} is unknown.

This scheme works as follows:

1) The codebook is formed by generating randomly 2^{nR_1} sequences $\boldsymbol{u}$, which are uniformly distributed into 2^{nR} bins, with $R_1 > R$.
2) The encoder identifies a sequence $\boldsymbol{u}$ which is typical with $\boldsymbol{x}$, and transmits the index of the bin to which $\boldsymbol{u}$ belongs.
3) At the decoder, a typicality test is performed between the side information $\boldsymbol{y}$ and all the sequences in the bins, allowing a sequence $\hat{\boldsymbol{u}}$ to be extracted from the bin.

Draper shows in [33] that the probability of debinning error can be made as small as desired if the block length n is large enough. In addition, given that $D < \sigma_x^2$ and $(X \mid Y)$ is Gaussian, we will show that the rate-distortion function $R_b(D) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\sigma_\Phi^2}\right)$ is achievable for $\mathbb{E}_{XU}\left[d(X,U)\right] \leq D$, where D is a function of σ_Φ^2. Next, in our proof, we need to express the generalization error for regression, and to analyze its convergence with respect to n. In our proofs, this analysis is specific to the considered regression problem, parametric or non-parametric. For parametric regression, this analysis is provided in Section IV-B. For kernel regression, the analysis is provided in the next Appendix.

APPENDIX C
PROOF OF THEOREM 2

Consider the definition of $\hat{f}(y)$ in equation (31). For a given $\tilde{y}$, for $\forall i \in [\![1, n]\!]$, note that

$$\frac{U_i}{\alpha} = f(Y_i) + N_i + \Phi_i = f(\tilde{y}) + (f(Y_i) - f(\tilde{y})) + (N_i + \Phi_i). \tag{63}$$

Therefore,

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)\frac{U_i}{\alpha} &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right) f(\tilde{y}) + \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)(f(Y_i) - f(\tilde{y})) \\ &\quad + \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)(N_i + \Phi_i) \\ &= \hat{p}_Y(\tilde{y}) f(\tilde{y}) + \hat{m}_1(\tilde{y}) + \hat{m}_2(\tilde{y}) \end{aligned} \tag{64}$$

where $\hat{p}_Y(\tilde{y}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)$ is the kernel density estimation of $\tilde{y}$ from observation $\boldsymbol{Y}$ [12], $\hat{m}_1(\tilde{y}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)(f(Y_i) - f(\tilde{y}))$ and $\hat{m}_2(\tilde{y}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)(N_i + \Phi_i)$. The analysis of the asymptotic distribution of the kernel estimator $\hat{f}(\tilde{y})$ is based on [38], which relies on the analysis of the numerator and the denominator of (31).

First, for $\hat{m}_2(\tilde{y})$, we have that $\mathbb{E}_{Y\Phi N}\left[\hat{m}_2(\tilde{y})\right] = 0$, and its variance can be expressed as:

$$\mathbb{V}\left[\hat{m}_2(\tilde{y})\right] = \mathbb{V}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tilde{y}-Y_i}{h}\right)(N_i + \Phi_i)\right] \tag{65}$$

$$= \frac{\sigma^2 + \sigma_\Phi^2}{nh^2}\int K^2\left(\frac{\tilde{y}-y}{h}\right) p_Y(y)dy \tag{66}$$

$$= \frac{\sigma^2 + \sigma_\Phi^2}{nh}\int K^2(u) p_Y(\tilde{y} + uh)du \tag{67}$$

$$= \frac{\sigma^2 + \sigma_\Phi^2}{nh}\int K^2(u)\left(p_Y(\tilde{y}) + p_Y'(\tilde{y})uh + o(h)\right)du \tag{68}$$

$$= \frac{(\sigma^2 + \sigma_\Phi^2) p(\tilde{y})}{nh} \int K^2(u) du + o\left(\frac{1}{nh}\right) \tag{69}$$

where (67) follows from a change of variable, and (68) comes from a Taylor approximation of $p_Y(\tilde{y} + uh)$ when $h \to 0$.

Also, following the same derivation as in [38], for $\hat{m}_1(\tilde{y})$, we show that

$$\mathbb{E}\left[\hat{m}_1(\tilde{y})\right] = \frac{h^2}{2} \left(2f'(\tilde{y}) p'_Y(\tilde{y}) + f''(\tilde{y}) p_Y(\tilde{y})\right) \int u^2 K(u) du + o(h^2) \tag{70}$$

and $\mathbb{V}\left[\hat{m}_1(\tilde{y})\right] = O(\frac{h}{n})$, which is negligible compared to the variance of $\hat{m}_2(\tilde{y})$. From (64) to (70), the central limit theorem is applied to obtain that as $h \to 0$ and $nh \to \infty$, we have

$$\hat{m}_1(\tilde{y}) \xrightarrow{p} \frac{h^2}{2} \left(2f'(\tilde{y}) p'_Y(\tilde{y}) + f''(\tilde{y}) p_Y(\tilde{y})\right) \int u^2 K(u) du \tag{71}$$

$$\hat{m}_2(\tilde{y}) \xrightarrow{d} \mathcal{N}\left(0, \frac{(\sigma^2 + \sigma_\Phi^2) p_Y(\tilde{y})}{nh} \int K^2(u) du\right) \tag{72}$$

where $\xrightarrow{p}$ denotes the convergence in probability and $\xrightarrow{d}$ denotes the convergence in distribution. By the property of kernel density estimation [12], it can be shown that $\hat{p}_Y \xrightarrow{p} p_Y$. Next, the kernel function (10) can be expressed as :

$$\hat{f}(\tilde{y}) = f(\tilde{y}) + \frac{\hat{m}_1(\tilde{y})}{\hat{p}_Y(\tilde{y})} + \frac{\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})} \tag{73}$$

By equation (71) to (73), we have the bias and variance in (35) and (36). By Slutsky's theorem [49], we have :

$$\frac{\hat{m}_1(\tilde{y}) + \hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})} \xrightarrow{d} \mathcal{N}\left(\frac{\mathbb{E}\left[\hat{m}_1(\tilde{y}\right]}{p_Y(\tilde{y})}, \frac{\mathbb{V}\left[\hat{m}_2(\tilde{y}\right]}{p_Y(\tilde{y})}\right). \tag{74}$$

Hence

$$\hat{f}(\tilde{y}) - f(\tilde{y}) \xrightarrow{d} \mathcal{N}\left(b_n(\tilde{y}), V_n(\tilde{y})^2\right). \tag{75}$$

where $b_n(\tilde{y}), V_n(\tilde{y})$ are defined in (35), (36). According to (33), the generalisation error can be expressed as:

$$\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] = \sigma^2 + \frac{h^4}{4} C_1 + \frac{\sigma^2 + \sigma_\Phi^2}{nh} C_2 + o\left(\frac{1}{nh}\right) + o(h^4) \tag{76}$$

where $C_1 = \int \left(2\frac{f'(y)p'_Y(y)}{p_Y(y)} + f''(y)\right)^2 dy \left(\int u^2 K(u) du\right)^2$ and $C_2 = \int \frac{1}{p_Y(y)} dy \int K^2(u) du$. Recall that the asymptotic generalization error with uncompressed observations $(\boldsymbol{X}, \boldsymbol{Y})$ is [12]

$$\mathbb{E}_{\boldsymbol{XY}}\left[G(\hat{f}^{(n)}, \boldsymbol{XY})\right] = \sigma^2 + \frac{h^4}{4} C_1 + \frac{\sigma^2}{nh} C_2 + o\left(\frac{1}{nh}\right) + o(h^4) \tag{77}$$

which indicates that asymptotically, the regression performed on coded data can achieve the same performance than that the one performed on original data.

APPENDIX D
PROOF OF COROLLARY 1

We first consider the conditional setup where the side information Y is available to both the encoder and the decoder. In this case, for $(X|Y) \sim \mathcal{N}(0, \sigma^2)$, the optimal conditional rate-distortion function is [40]

$$R_{X|Y}(D) = \frac{1}{2} \log \left(\frac{\sigma^2}{D} \right), \tag{78}$$

We now show that when the same relation between X and Y holds, i.e. $(X|Y) \sim \mathcal{N}(0, \sigma^2)$, but the side information is only available at the decoder, the Wyner-Ziv rate-distortion function $R_{WZ}(D)$ is equal to the conditional one $R_{X|Y}(D)$. Using the test channel provided in previous section $U = \alpha(X + \Phi)$ with $\alpha = \frac{\sigma^2 - D}{\sigma^2}$ and $\sigma_\Phi^2 = \frac{D\sigma^2}{\sigma^2 - D}$, together with the scheme proposed in [33], the intermediate variable U can be recovered without error as the block-length tends to infinity. By considering the minimum mean square error estimator for the reconstruction $\hat{X}$ of X, it can be show that $\mathbb{E}\left[(X - \hat{X})^2\right] = (\alpha - 1)^2\sigma^2 + \alpha^2\sigma_\Phi^2$. Replacing α and σ_Φ by their expression leads to $\mathbb{E}\left[(X - \hat{X})^2\right] = D$.

Second, the binning rate in [33] can be expressed as

$$\begin{aligned} I(X;U) - I(Y;U) &= h(U|Y) - h(U|X) \\ &= \frac{1}{2} \log \left(\frac{\sigma^2 + \sigma_\Phi^2}{\sigma_\Phi^2} \right) \end{aligned} \tag{79}$$

where $h(\cdot)$ denotes the differential entropy and the last equality comes from the fact that under the previous test channel both $(U|Y)$ and $(U|X)$ follow a Gaussian distribution. By replacing σ_Φ^2 in the previous expression, we obtain $R_{WZ}(D) = R_{X|Y}(D)$, the optimal rate for reconstruction. Then by Theorem 1, for any positive $R_{WZ}(D)$, the pair $(R_{WZ}(D), 0)$ is also achievable for the generalization error using the same scheme. In other words, it states that $R(D, G) = R_{WZ}(D)$ for all $G \geq \sigma^2$. This indicates that the Gaussian test channel is also optimal in our regression setup as long as the conditional distribution of $(X|Y)$ is Gaussian, whatever the distribution of Y.

APPENDIX E
PROOF OF THEOREM 3

This proof follows the channel resolvability type code of [36] for the Wyner-Ziv problem. We adapt it so as to also account for the generalization error in the information-density vector

defined in (45). Some preliminary results for channel resolvability and identification code were provided in Appendix A.

Code construction: The encoder uses a stochastic map $P_{I|X} : \mathcal{X} \to \mathcal{I}$. It generates $i \in \mathcal{I}$ according to $P_{I|X}$. Then the encoder sends i by random binning $\kappa : \mathcal{I} \to \mathcal{M}$. It means for every $i \in \mathcal{I}$, it is independently and uniformly assigned to a random bin $m \in \mathcal{M}$. For given $m \in \mathcal{M}$ and $y \in \mathcal{Y}$, the decoder finds the unique index $i \in \mathcal{I}$ such that $\kappa(i) = m$, and

$$(u_i, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}}). \tag{80}$$

As mentioned in the channel resolvability problem in Appendix A, the stochastic map $P_{I|X}$ is constructed such that the joint distribution $P_{\hat{I}X}$ is indistinguishable from $P_{IX'}$, where for any measurable $E \in \mathcal{P}(\mathcal{I}) \times \mathscr{B}(\mathcal{X})$

$$P_{IX'}(E) = \int_E \sum_i^{|\mathcal{I}|} \frac{\mathbf{1}[u_i = u]}{|\mathcal{I}|} p_{X|U}(x|u)dx, \tag{81}$$

and $P_{IX'}$ is the joint distribution of the couple (I, X'), where X' is the random variable that induces the probability measure $P_{X|U}$ when U is chosen uniformly in the codebook $\{u_1, \cdots u_{|\mathcal{I}|}\}$. Then the decoder has two objectives: i) performing the regression according to $\boldsymbol{U}$ and $\boldsymbol{Y}$ to obtain a function $\hat{f}$, and ii) reconstruction of X. In both cases, a decoder error occurs if there is no i satisfying (80) or if there is more than one such i satisfying (80).

Let $\hat{I}$ be random index chosen by the encoder via the stochastic map $P_{I|X}$. The joint distribution of $(\hat{I}, X)$ is given by

$$P_{\hat{I}X} = P_X P_{I|X}. \tag{82}$$

And the joint distribution of $(\hat{I}, X, Y, \hat{X}, G)$ is given by

$$P_{\hat{I}XY\hat{X}} = P_{\hat{I}X} P_{Y|X} P_{\hat{X}|UY}. \tag{83}$$

Note that here the generalization error is fully determined by the sequence $\boldsymbol{U}$ and $\boldsymbol{Y}$. The smoothed versions $\bar{P}_{\hat{I}X}$ and $\bar{P}_{\hat{I}XY\hat{X}}$ are given by

$$\bar{P}_{\hat{I}X}(E) = \int_E \sum_i^{|\mathcal{I}|} \mathbf{1}[(u_i, x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})] p_X(x) p_{I|X}(i|x)dx \tag{84}$$

$$\bar{P}_{\hat{I}XY\hat{X}} = \bar{P}_{\hat{I}X} P_{Y|X} P_{\hat{X}|UY}. \tag{85}$$

$P_{e,n}(D, G) = \mathbb{P}\left[\left\{d(X, \hat{X}) \geq D\right\} \cup \left\{G(\hat{f}^{(n)}, \boldsymbol{Z}) \geq G\right\}\right]$ from Definition 7. This error probability is bounded away from 0 if at least of one of the following error events occurs:

$$\mathcal{E}_0 := \{(\boldsymbol{u}, \boldsymbol{y}) \notin \mathcal{T}_g(G)\}, \tag{86}$$

$$\mathcal{E}_1 := \{(x, \hat{x}) \notin \mathcal{T}_d(D)\}, \tag{87}$$

$$\mathcal{E}_2 := \{(u_i, y) \notin \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})\}, \tag{88}$$

$$\mathcal{E}_3 := \{\exists\, i' \neq i \text{ s.t. } \kappa(i') = \kappa(i), (u_{i'}, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})\}. \tag{89}$$

For fixed codebook $\mathcal{C}$, following the same step as in [36], the excess probability is thus upper bounded by :

$$\begin{aligned}
&P_{\hat{I}XY\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) \\
\leq &\bar{P}_{IX'Y\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2) + \bar{P}_{IX'Y\hat{X}}(\mathcal{E}_3) + d_{TV}(P_{\hat{I}X}, \bar{P}_{IX'}) \\
&+ \frac{1 - \bar{P}_{IX'Y\hat{X}}(\mathcal{I} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})}{2}
\end{aligned} \tag{90}$$

Next, the excess probability $P_{e,n}(D, G)$ is averaged over the random coding function κ and the random codebook $\mathcal{C}$ can be upper bounded as

$$\begin{aligned}
&\mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}[P_{e,n}(D, G)] \\
\leq &\mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}\left[P_{\hat{I}XY\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3)\right] \\
\leq &\mathbb{E}_{\mathcal{C}}\left[\bar{P}_{IX'Y\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)\right] + \mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}\left[\bar{P}_{IX'Y\hat{X}}(\mathcal{E}_3)\right] \\
&+ \mathbb{E}_{\mathcal{C}}\left[d_{TV}(P_{\hat{I}X}, \bar{P}_{IX'})\right] + \mathbb{E}_{\mathcal{C}}\left[\frac{1 - \bar{P}_{IX'Y\hat{X}}(\mathcal{I} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})}{2}\right].
\end{aligned} \tag{91}$$

The expectation of the first term can be expressed as:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\left[\mathbb{P}_{IX'Y\hat{X}}\left[(u_i, x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}) \cap \left(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2\right)\right]\right] \\
&= \mathbb{P}_{X\boldsymbol{U}\boldsymbol{Y}\hat{X}}\left[(u, x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}) \cap \left(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2\right)\right].
\end{aligned} \tag{92}$$

For the three last terms as proved in [36], we can prove that the second term in (91) is upper bounded as:

$$\begin{aligned}
&\mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}\left[\bar{P}_{IX'Y}(\mathcal{E}_3)\right] \\
=&\mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}\Bigg[\int_{\mathcal{U}\mathcal{X}\mathcal{Y}} \sum_i \frac{\mathbf{1}\left[(u_i, x) \in \mathcal{T}_c(\gamma_c)\right]}{|\mathcal{I}|} \times \mathbf{1}[\exists i' \neq i \ s.t. \kappa(i') = \kappa(i), (u_{i'}, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})] \\
&\quad p_{X|U}(x|u) p_{Y|X}(y|x) dx dy\Bigg] \\
\leq &\mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}}\Bigg[\int_{\mathcal{U}\mathcal{X}\mathcal{Y}} \sum_i \frac{\mathbf{1}\left[(u_i, x) \in \mathcal{T}_c(\gamma_c)\right]}{|\mathcal{I}|} |\mathcal{I}| \sum_{i' \neq i} \frac{\mathbf{1}[\kappa(i') = \kappa(i)] \times \mathbf{1}[(u_{i'}, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})]}{|\mathcal{I}|}
\end{aligned} \tag{93}$$

$$p_{X|U}(x|u)p_{Y|X}(y|x)dxdy\Big] \tag{94}$$

$$\leq \frac{|\mathcal{I}|}{|\mathcal{M}|}\int_{\mathcal{U}\mathcal{X}\mathcal{Y}} \mathbf{1}\left[(u,x)\in\mathcal{T}_c(\gamma_c)\right] p_{XYU}(x,y,u)dxdydu \int_{\mathcal{U}} \mathbf{1}[(u,y)\in\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})]p_U(u)du \tag{95}$$

$$\leq \frac{|\mathcal{I}|}{|\mathcal{M}|}\int_{\mathcal{U}\mathcal{Y}} \mathbf{1}[(u,y)\in\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})]p_U(u)p_Y(y)dudy \tag{96}$$

$$\leq \frac{|\mathcal{I}|}{2^{\gamma_p}|\mathcal{M}|}, \tag{97}$$

where (95) comes from the fact that $\mathbb{E}_\kappa\left[\mathbf{1}[\kappa(i')=\kappa(i)]\right] \leq \frac{1}{|\mathcal{M}|}$. The third term can be upper bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}[d_{TV}(P_{\hat{I}X},\bar{P}_{IX'})]\\
\leq&\mathbb{E}_{\mathcal{C}}[d_{TV}(P_{\hat{I}X},\bar{P}_{\hat{I}X})] + \mathbb{E}_{\mathcal{C}}[d_{TV}(\bar{P}_{\hat{I}X},\bar{P}_{IX'})] &(98)\\
\leq&\frac{P_{UX}((u,x)\notin\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}))}{2} + \frac{\Delta(\gamma_{\mathrm{c}},P_{UX})}{2\sqrt{|\mathcal{I}|}}\\
\leq&\frac{P_{UX}((u,x)\notin\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}))}{2} + \frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{|\mathcal{I}|}}. &(99)
\end{aligned}$$

The expectation of the last term can be evaluated as :

$$\mathbb{E}_{\mathcal{C}}\left[1-\bar{P}_{IX'Y\hat{X}}(\mathcal{I}\times\mathcal{X}\times\mathcal{Y}\times\hat{\mathcal{X}})\right] \tag{100}$$

$$=1-\mathbb{E}_{\mathcal{C}}\Bigg[\int_{\mathcal{X}\mathcal{Y}\hat{\mathcal{X}}\mathcal{U}}\sum_i \frac{\mathbf{1}[u_i=u]\times\mathbf{1}\left[(u,x)\in\mathcal{T}_c(\gamma_c)\right]}{|\mathcal{I}|}\cdot$$

$$p_{X|U}(x|u)p_{Y|X}(y|x)p_{\hat{X}|UY}(\hat{x}|u,y)dxdyd\hat{x}du\Bigg] \tag{101}$$

$$=\mathbb{P}_{UX}\left[(u,x)\notin\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})\right]. \tag{102}$$

Combining the results above provides the upper bound

$$\begin{aligned}
\varepsilon \leq\ &\mathbb{P}_{X\boldsymbol{U}\boldsymbol{Y}\hat{X}}\left[(u,y)\in\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c\cup(u,x)\in\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c\right.\\
&\left.\cup(x,\hat{x})\in\mathcal{T}_{\mathrm{d}}(D)^c\cup(\boldsymbol{u},\boldsymbol{y})\in\mathcal{T}_g(G)^c\right] + \frac{N}{2^{\gamma_{\mathrm{p}}}|\mathcal{M}|} + \frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{N}}. &(103)
\end{aligned}$$

APPENDIX F
PROOF OF THEOREM 4

The proof is based on a Gaussian approximation using the following multi-dimensional Berry-Esséen theorem.

Theorem 5 (Multidimensional Berry-Esséen theorem [50])**.** *Let $\boldsymbol{U}_1, \boldsymbol{U}_2, ..., \boldsymbol{U}_n$ be independent random vectors in $\mathbb{R}^k$ with zero mean. Let $\boldsymbol{S}_n = \frac{1}{\sqrt{n}}(\boldsymbol{U}_1 + ... + \boldsymbol{U}_n)$ and $\mathbb{C}\left[\boldsymbol{S}_n\right] = \underline{\boldsymbol{V}} > \mathbf{0}$. Consider a Gaussian random vector $\boldsymbol{B} \sim \mathcal{N}(\mathbf{0}, \underline{\boldsymbol{V}})$, then for all $n \in \mathbb{N}$, we have*

$$\sup_{C \in \mathscr{C}_k} \left| \mathbb{P}_{\boldsymbol{S}_n}\left[C\right] - \mathbb{P}_{\boldsymbol{B}_n}\left[C\right] \right| \leq O\left(\frac{1}{\sqrt{n}}\right) \tag{104}$$

where $\mathscr{C}_k$ is the family of all convex Borel measurable subsets of $\mathbb{R}^k$.

As mentioned earlier, the components of the information-density-distortion-generalization error vector defined in (45) are not independent, since the generalization error $\mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y}))\right]$ depends on the full sequence $\mathbf{Z}$, while the other components only depend on a random occurrence of U, Y. Therefore, let us consider the conditional information-density-distortion-generalization error vector rewritten as follow:

$$\boldsymbol{\jmath}_i(U_i, X_i, Y_i, \hat{X}_i | \boldsymbol{Z}_{-\boldsymbol{i}}) = \begin{bmatrix} -\iota(u_i, y_i) \\ \iota(x_i, u_i) \\ d(x_i, \hat{x}_i) \\ \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(u_i, y_i, \tilde{Y}))|\boldsymbol{Z}_{-\boldsymbol{i}}\right] \end{bmatrix} \tag{105}$$

where $\boldsymbol{Z}_{-\boldsymbol{i}} = [\boldsymbol{u}^{i-1}, \boldsymbol{u}_{i+1}^n, \boldsymbol{y}^{i-1}, \boldsymbol{y}_{i+1}^n]$. Given that U and Y are independent random variables, let $\boldsymbol{Z}^\star = \boldsymbol{Z}_{-\boldsymbol{i}} \sim P_{\boldsymbol{U}^{n-1}\boldsymbol{Y}^{n-1}}$. Its expectation $\boldsymbol{J}(\boldsymbol{Z}^\star)$ can be expressed as

$$\boldsymbol{J}(\boldsymbol{Z}^\star) = \mathbb{E}[\mathbf{j}_i(U_i, X_i, Y_i, \hat{X}_i | \boldsymbol{Z}_{-i})] = \begin{bmatrix} -I(U;Y) \\ I(U;X) \\ \mathbb{E}_{X\hat{X}}\left[d(X, \hat{X})\right] \\ \mathbb{E}_{\boldsymbol{Z}\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{U}, \boldsymbol{Y}, \tilde{Y})|\boldsymbol{Z}_{-i}\right] \end{bmatrix}. \tag{106}$$

Let $\gamma_p = \log \frac{|\mathcal{I}_n|}{|\mathcal{M}_n|} + \log n$, where $\mathcal{M}_n = \{1, \cdots, M_n\}$, and $\gamma_c = \log |\mathcal{I}_n| - \log n$, using (103) allows us to show that there exists a code such that

$$\begin{aligned} P_{e,n}(G, D) &\leq \mathbb{P}_{X\boldsymbol{U}\boldsymbol{Y}\hat{X}}\left[(u, y) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c \cup (u, x) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c \right. \\ &\left. \cup (x, \hat{x}) \in \mathcal{T}_{\mathrm{d}}(D)^c \cup (\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{T}_g(G)^c\right] + \frac{1}{n} + \frac{1}{2\sqrt{n}} \end{aligned} \tag{107}$$

$$\leq \mathbb{E}_{\boldsymbol{Z}^*}\left[\mathbb{P}\left[\sum_i^n \begin{bmatrix} -\iota(u_i, y_i) \\ \iota(x_i, u_i) \\ d(x_i, \hat{x}_i) \\ \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(u_i, y_i, \tilde{Y}))|\boldsymbol{Z}_{-\boldsymbol{i}}\right] \end{bmatrix} \geq \begin{bmatrix} \log \frac{|\mathcal{M}_n|}{|\mathcal{I}_n|} \\ \log |\mathcal{I}_n| \\ nD \\ nL \end{bmatrix} - \mathbf{\log n} \right]\right] + \frac{1}{n} + \frac{1}{2\sqrt{n}} \tag{108}$$

$$= \mathbb{E}_{\boldsymbol{Z}^*}\left[\mathbb{P}\left[\sum_i^n [\boldsymbol{j}_i - \boldsymbol{J}(\boldsymbol{Z}^\star)] \geq \begin{bmatrix}\log\frac{|\mathcal{M}_n|}{|\mathcal{I}_n|}\\ \log|\mathcal{I}_n|\\ nD\\ nL\end{bmatrix} - n\boldsymbol{J}(\boldsymbol{Z}^\star) - \log \boldsymbol{n}\right]\right] + \frac{1}{n} + \frac{1}{2\sqrt{n}}. \tag{109}$$

Let

$$\tilde{\boldsymbol{b}}|\boldsymbol{Z}^* = \sqrt{n}\left[\begin{bmatrix}\frac{1}{n}\log\frac{|\mathcal{M}_n|}{|\mathcal{I}_n|}\\ \frac{1}{n}\log|\mathcal{I}_n|\\ D\\ L\end{bmatrix} - \boldsymbol{J}(\boldsymbol{Z}^\star) - \frac{\mathbf{2}\log \boldsymbol{n}}{\boldsymbol{n}}\right], \tag{110}$$

where $\frac{\log \boldsymbol{n}}{\boldsymbol{n}}$ denotes the vector $\frac{\log n}{n}\mathbf{1}_4$ with same size as vector $\boldsymbol{J}$. We have

$$1 - P_e(n, L, D) \tag{111}$$

$$\geq \mathbb{P}_{\boldsymbol{Z}^*}\left[\mathbb{P}\left[\frac{1}{\sqrt{n}}\sum_i^n [\boldsymbol{j}_i - \boldsymbol{J}(\boldsymbol{Z}^\star)] \leq \tilde{\boldsymbol{b}}|\boldsymbol{Z}^* + \frac{\log \boldsymbol{n}}{\sqrt{\boldsymbol{n}}}\right]\right] - \frac{1}{n} - \frac{1}{2\sqrt{n}} \tag{112}$$

$$= \mathbb{E}_{\boldsymbol{Z}^*}\left[\mathbb{P}_{\boldsymbol{J_i}|\boldsymbol{Z}^*}\left[\frac{1}{\sqrt{n}}\sum_i^n [\boldsymbol{j}_i - \boldsymbol{J}] \leq \tilde{\boldsymbol{b}}|\boldsymbol{Z}^* + \frac{\log \boldsymbol{n}}{\sqrt{\boldsymbol{n}}}\right]\right] - \frac{1}{n} - \frac{1}{2\sqrt{n}} \tag{113}$$

$$\geq \mathbb{P}_{\boldsymbol{Z}^*}\left[\mathbb{P}_{\boldsymbol{B}|\boldsymbol{Z}^*}\left[\boldsymbol{B}|\boldsymbol{Z}^* \leq \tilde{\boldsymbol{b}}|\boldsymbol{Z}^* + \frac{\log \boldsymbol{n}}{\sqrt{\boldsymbol{n}}}\right]\right] - O\left(\frac{1}{\sqrt{n}}\right) \tag{114}$$

$$= \mathbb{E}_{\boldsymbol{Z}^*}\left[\mathbb{P}_{\boldsymbol{B}|\boldsymbol{Z}^*}\left[\boldsymbol{B}|\boldsymbol{Z}^* \leq \tilde{\boldsymbol{b}}|\boldsymbol{Z}^*\right]\right] + O\left(\frac{\log \boldsymbol{n}}{\sqrt{\boldsymbol{n}}}\right) \tag{115}$$

$$= \mathbb{E}_{\boldsymbol{B}}\left[\boldsymbol{B} \leq \tilde{\boldsymbol{b}}\right] + O\left(\frac{\log \boldsymbol{n}}{\sqrt{\boldsymbol{n}}}\right) \tag{116}$$

$$\geq 1 - \epsilon, \tag{117}$$

which indicates that

$$\tilde{\boldsymbol{b}} = \mathbb{E}_{\boldsymbol{Z}^*}\left[\tilde{\boldsymbol{b}}|\boldsymbol{Z}^*\right] = \sqrt{n}\left[\begin{bmatrix}\frac{1}{n}\log\frac{|\mathcal{M}_n|}{|\mathcal{I}_n|}\\ \frac{1}{n}\log|\mathcal{I}_n|\\ D\\ L\end{bmatrix} - \mathbb{E}_{\boldsymbol{Z}^*}[\boldsymbol{J}] - \frac{\mathbf{2}\log \boldsymbol{n}}{\boldsymbol{n}}\right] \in \mathscr{S}(\underline{\boldsymbol{V}}, \epsilon) \tag{118}$$

where $\mathscr{S}(\underline{\boldsymbol{V}}, \epsilon)$ is the dispersion region defined in (48). This completes the proof.

APPENDIX G
PROOF OF COROLLARY 2

As shown in Theorem 4, the bound for the second order coding rate is mainly affected by the dispersion region of $\boldsymbol{i}(X, \boldsymbol{U}, \boldsymbol{Y}, \hat{X})$, and especially by its covariance matrix. In what follows, we aim to show that

$$\text{Cov}\left(d(X, \hat{X}), G(\hat{f}^{(n)}, \boldsymbol{Z})\right) = 0, \tag{119}$$

which means that the distortion and the generalization error are uncorrelated. Here we provide the proof for both parametric regression with OLS estimator and kernel regression. Since $X, \hat{X}$ are i.i.d., without loss of generality, we consider $d(X, \hat{X}) = d(X_1, \hat{X}_1)$.

A. Parametric regression

The covariance term (119) can be expressed as:

$$\text{Cov}\left(d(X_1, \hat{X}_1), G(\hat{f}^{(n)}, \boldsymbol{Z})\right) = \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z}) d(X_1, \hat{X}_1)\right] - \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \mathbb{E}\left[d(X_1, \hat{X}_1)\right] \tag{120}$$

with

$$\mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \mathbb{E}\left[d(X_1, \hat{X}_1)\right] = \left(\sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n} \mathbb{E}\left[\text{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}} \underline{\boldsymbol{\Sigma}}^{-1}\right)\right]\right) \mathbb{E}\left[d(X_1, \hat{X}_1)\right] \tag{121}$$

and

$$\mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z}) d(X_1, \hat{X}_1)\right] \tag{122}$$

$$= \sigma^2 \mathbb{E}\left[d(X_1, \hat{X}_1)\right] + \mathbb{E}\left[[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}} \tilde{\boldsymbol{Y}}^T\right] [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] d(X_1, \hat{X}_1)\right] \tag{123}$$

$$= \sigma^2 \mathbb{E}\left[d(X_1, \hat{X}_1)\right] + \frac{1}{n^2} \mathbb{E}_Y\left[\text{tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}} \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{Y}}\ \mathbb{E}_{N\Phi}\left[(\boldsymbol{N} + \boldsymbol{\Phi}) d(X_1, \hat{X}_1) (\boldsymbol{N} + \boldsymbol{\Phi})^T\right] \underline{\boldsymbol{Y}}^T \underline{\boldsymbol{\Sigma}}^{-1}\right)\right] \tag{124}$$

$$= \sigma^2 \mathbb{E}\left[d(X_1, \hat{X}_1)\right] + \frac{\sigma^2 \sigma_\phi^2}{n^2} \mathbb{E}\left[\text{tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}} \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{Y}} \underline{\boldsymbol{Y}}^T \underline{\boldsymbol{\Sigma}}^{-1}\right)\right] \tag{125}$$

$$= \sigma^2 \mathbb{E}\left[d(X_1, \hat{X}_1)\right] + \frac{\sigma^2 \sigma_\phi^2}{n} \mathbb{E}\left[\text{tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}} \underline{\boldsymbol{\Sigma}}^{-1}\right)\right] \tag{126}$$

where (125) follows from the fact that $\mathbb{E}_{N\Phi}\left[(\boldsymbol{N} + \boldsymbol{\Phi}) d(X_1, \hat{X}_1) (\boldsymbol{N} + \boldsymbol{\Phi})^T\right]$ remains the same for all $i \in [\![1, n]\!]$. Using the fact that $\sigma^2 \sigma_\phi^2 = (\sigma^2 + \sigma_\phi^2) \mathbb{E}\left[d(X_1, \hat{X}_1)\right]$ completes the proof for the parametric case.

B. Kernel regression

For kernel regression, we express:

$$\begin{aligned}
&\mathbb{E}_{\tilde{X}\tilde{Y}\boldsymbol{Z}}\left[G(\hat{f}^{(n)},\boldsymbol{Z})\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right] \qquad (127)\\
&=\left(\sigma^2+\mathbb{E}_{\tilde{Y}}\left[f^2\left(\tilde{Y}\right)\right]+\mathbb{E}_{\tilde{Y}\boldsymbol{Z}}\left[\hat{f}^{(n)^2}(\tilde{Y})\right]\right.\\
&\quad\left.-2\mathbb{E}_{\tilde{Y}}\left[f\left(\tilde{y}\right)\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)}\left(\tilde{y}\right)\right]\middle|\tilde{Y}=\tilde{y}\right]\right)\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\tilde{X}\tilde{Y}\boldsymbol{Z}}\left[G(\hat{f}^{(n)},\boldsymbol{Z})d(X_1,\hat{X}_1)\right] \qquad (128)\\
&=\sigma^2\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right]+\mathbb{E}_{\tilde{Y}}\left[f^2(\tilde{Y})\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right]\\
&\quad+\mathbb{E}_{\tilde{Y}}\left[\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)^2}(\tilde{y})d(X_1,\hat{X}_1)\right]\middle|\tilde{Y}=\tilde{y}\right] \qquad (129)\\
&\quad-2\mathbb{E}_{\tilde{Y}}\left[f(\tilde{y})\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)}(\tilde{y})d(X_1,\hat{X}_1)\right]\middle|\tilde{Y}=\tilde{y}\right] \qquad (130)
\end{aligned}$$

For the last term, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)}(\tilde{y})d(X_1,\hat{X}_1)\right] \qquad (131)\\
&=f(\tilde{y})\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right]+\mathbb{E}_{\boldsymbol{Y}}\left[\frac{\hat{m}_1(\tilde{y})}{\hat{p}(\tilde{y})}\right]\mathbb{E}_{\boldsymbol{N\Phi}}\left[d(X_1,\hat{X}_1)\right]+\mathbb{E}_{\boldsymbol{Y N\Phi}}\left[\frac{\hat{m}_2(\tilde{y})d(X_1,\hat{X}_1)}{\hat{p}_Y(\tilde{y})}\right] \qquad (132)\\
&=f(\tilde{y})\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right]+\mathbb{E}_{\boldsymbol{Y}}\left[\frac{\hat{m}_1(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right]\mathbb{E}_{\boldsymbol{N\Phi}}\left[d(X_1,\hat{X}_1)\right]\\
&\quad+\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{Y}}\left[\frac{K(\frac{\tilde{y}-Y_i}{h})}{nh\hat{p}(\tilde{y})}\right]\mathbb{E}_{\boldsymbol{N\Phi}}\left[(N_i+\Phi_i))d(X_1,\hat{X}_1)\right] \qquad (133)\\
&=f(\tilde{y})\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right]+\mathbb{E}_{\boldsymbol{Y}}\left[\frac{\hat{m}_1(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right]\mathbb{E}_{\boldsymbol{N\Phi}}\left[d(X_1,\hat{X}_1)\right] \qquad (134)\\
&=\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)}(\tilde{y})\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right], \qquad (135)
\end{aligned}$$

where (132) comes from the fact that $d(X_1,\hat{X}_1)$ is independent from $\hat{p}_Y(\tilde{y})$ since N and Φ are independent from Y. In addition, (134) is because for all $i\in[\![1,n]\!]$, $\mathbb{E}_{\boldsymbol{N\Phi}}\left[(N_i+\Phi_i))d(X_1,\hat{X}_1)\right]=0$.

Then for the third term (129), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)^2}(\tilde{y})d(X_1,\hat{X}_1)\right] \qquad (136)\\
&=\mathbb{E}\left[f^2(\tilde{y})+\left(\frac{\hat{m}_1(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right)^2+2\frac{f(\tilde{y})\hat{m}_1(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right]\mathbb{E}\left[d(X_1,\hat{X}_1)\right] \qquad (137)
\end{aligned}$$

$$+\mathbb{E}\left[\left(2\frac{f(\tilde{y})\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})}+2\frac{\hat{m}_1(\tilde{y})\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right)d(X_1,\hat{X}_1)\right] \tag{138}$$

$$+\mathbb{E}\left[\left(\frac{\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right)^2 d(X_1,\hat{X}_1)\right] \tag{139}$$

where (137) is obtained from the same arguments as for (132), and the second part equals to zero because of equation (134). The last part (139) can be developped as

$$\mathbb{E}_{\boldsymbol{Z}}\left[\left(\frac{\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right)^2 d(X_1,\hat{X}_1)\right] \tag{140}$$

$$=\frac{1}{n^2h^2}\mathbb{E}_{\boldsymbol{YN\Phi}}\left[\sum_{i=1}^{n}\left(\frac{K(\frac{\tilde{y}-Y_i}{h})(N_i+\Phi_i)}{\hat{p}_Y(\tilde{y})}\right)^2 d(X_1,\hat{X}_1)\right] \tag{141}$$

$$=\frac{1}{n^2h^2}\mathbb{E}_{\boldsymbol{Y}}\left[\sum_{i=1}^{n}\left(\frac{K(\frac{\tilde{y}-Y_i}{h})}{\hat{p}_Y(\tilde{y})}\right)^2\right]\mathbb{E}_{\boldsymbol{N\Phi}}\left[(N_i+\Phi_i)^2\left((\alpha-1)N_j+\alpha\Phi_j\right)^2\right] \tag{142}$$

$$=\frac{\sigma^2+\sigma_\Phi^2}{n^2h^2}\mathbb{E}_{\boldsymbol{Y}}\left[\sum_{i=1}^{n}\left(\frac{K(\frac{\tilde{y}-Y_i}{h})}{\hat{p}_Y(\tilde{y})}\right)^2\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right] \tag{143}$$

$$=\mathbb{E}_{\boldsymbol{Z}}\left[\left(\frac{\hat{m}_2(\tilde{y})}{\hat{p}_Y(\tilde{y})}\right)^2\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right] \tag{144}$$

with (143) is obtained from the same arguments as for (125).

This gives

$$\begin{aligned}&\mathbb{E}_{\tilde{Y}}\left[f(\tilde{y})\mathbb{E}_{\boldsymbol{Z}}\left[\hat{f}^{(n)}(\tilde{y})d(X_1,\hat{X}_1)\,\middle|\,\tilde{Y}=\tilde{y}\right]\right]\\ =&\mathbb{E}_{\tilde{Y}\boldsymbol{Z}}\left[f(\tilde{y})\hat{f}^{(n)}(\tilde{y})\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X_1,\hat{X}_1)\right],\end{aligned} \tag{145}$$

therefore

$$\begin{aligned}&\mathbb{E}_{\tilde{X}\tilde{Y}\boldsymbol{Z}}\left[G(\hat{f}^{(n)},\boldsymbol{Z})d(X,\hat{X})\right]\\ =&\mathbb{E}_{\tilde{X}\tilde{Y}\boldsymbol{Z}}\left[G(\hat{f}^{(n)},\boldsymbol{Z})\right]\mathbb{E}_{\boldsymbol{Z}}\left[d(X,\hat{X})\right],\end{aligned} \tag{146}$$

and $\mathrm{Cov}\left(d(X,\hat{X}),G(\hat{f}^{(n)},\boldsymbol{Z})\right)=0$.

Denote $R_b(n,D,G,\varepsilon)$ as the infimum introduced by Theorem 4 for the rate-distortion-generalization error function, $R_b(n,D,\varepsilon)$ and $R_b(n,G,\varepsilon)$ for the rate-distortion function and rate-generalization error function, respectively. Consider the vector $\boldsymbol{B}=[B_1,B_2,B_3,B_4]^T$ defined in (48), by the achivability of $R_b(n,D,\varepsilon)$ and $R_b(n,G,\varepsilon)$, we have :

$$\mathbb{P}_{B_1B_2B_3}\left[B_1\le b_1,B_2\le b_2,B_3\le D\right]=1-\varepsilon, \tag{147}$$

$$\mathbb{P}_{B_1B_2B_4}\left[B_1 \leq b_1', B_2 \leq b_2', B_4 \leq G\right] = 1-\varepsilon \tag{148}$$

with $R_b(n,D,\varepsilon) = I(X;U) - I(U;Y) + b_1 + b_2 + O\left(\frac{\log n}{n}\right)$ and $R_b(n,G,\varepsilon) = I(X;U) - I(U;Y) + b_1' + b_2' + O\left(\frac{\log n}{n}\right)$.

Consider firstly $R_b(n,D,\varepsilon) \geq R_b(n,G,\varepsilon)$ and $R_b = \max\{R_b(n,D,\varepsilon), R_b(n,G,\varepsilon)\}$, we have

$$\mathbb{P}_{\boldsymbol{B}}\left[B_1 \leq b_1, B_2 \leq b_2, B_3 \leq D, B_4 \leq G\right] \tag{149}$$

$$=\mathbb{P}\left[B_1 \leq b_1, B_2 \leq b_2\right]\mathbb{P}\left[B_3 \leq D | B_1 \leq b_1, B_2 \leq b_2\right]\mathbb{P}\left[B_4 \leq G | B_1 \leq b_1, B_2 \leq b_2\right] \tag{150}$$

$$=\mathbb{P}\left[B_1 \leq b_1, B_2 \leq b_2, B_3 \leq D\right]\mathbb{P}\left[B_4 \leq G | B_1 \leq b_1, B_2 \leq b_2\right] \tag{151}$$

$$=(1-\varepsilon)\mathbb{P}\left[B_4 \leq G | B_1 \leq b_1, B_2 \leq b_2\right] \tag{152}$$

$$<1-\varepsilon, \tag{153}$$

where (150) follows the fact that uncorrelation of Gaussian variables indicates independence, (153) is because the cumulative density function of Gaussian source is smaller than 1.

The same analysis applies for the case $R_b(n,G,\varepsilon) \leq R_b(n,D,\varepsilon)$. It implies that in order to ensure the same excess probability, a higher rate $R_b(n,D,G,\varepsilon)$ is necessary by the approximation of Berry-Esséen Theorem, that is

$$R_b(n,D,G,\varepsilon) > \max\{R_b(n,D,\varepsilon), R_b(n,G,\varepsilon)\} \tag{154}$$

This completes the proof.

REFERENCES

[1] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6g: Vision, principles, and technologies," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 78–85, 2023.

[2] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[3] M. Raginsky, "Learning from compressed observations," in *2007 IEEE Information Theory Workshop*, 2007, pp. 420–425.

[4] M. Ehrlich and L. S. Davis, "Deep residual learning in the jpeg transform domain," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3484–3493.

[5] O. N. Onak, T. Erenler, and Y. S. Dogrusoz, "A novel data-adaptive regression framework based on multivariate adaptive regression splines for electrocardiographic imaging," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 963–974, 2022.

[6] N. Dal Fabbro, M. Rossi, G. Pillonetto, L. Schenato, and G. Piro, "Model-free radio map estimation in massive mimo systems via semi-parametric gaussian regression," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 473–477, 2022.

[7] D. Angelosante, G. B. Giannakis, and N. D. Sidiropoulos, "Estimating multiple frequency-hopping signal parameters via sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5044–5056, 2010.

[8] M. Pawlak, Z. Hasiewicz, and P. Wachel, "On nonparametric identification of wiener systems," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 482–492, 2007.

[9] Y. Wang and P. Ishwar, "Distributed field estimation with randomly deployed, noisy, binary sensors," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1177–1189, 2009.

[10] 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.

[11] A. C. Rencher and G. B. Schaalje, *Linear models in statistics*. John Wiley & Sons, 2008.

[12] L. Wasserman, *All of Nonparametric Statistics (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[13] P. Koulgi, E. Tuncel, S. L. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 99–111, 2003.

[14] D. Malak and M. Médard, "Hyper binning for distributed function coding," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[15] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 903–917, 2001.

[16] A. C.-C. Yao, "Some complexity questions related to distributive computing (preliminary report)," in *Proceedings of the eleventh annual ACM symposium on Theory of computing*, 1979.

[17] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.

[18] 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.

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

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

[21] M. El Gamal and L. Lai, "Are Slepian-Wolf rates necessary for distributed parameter estimation?" in *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 1249–1255.

[22] 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.

[23] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[24] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*. IEEE, 2015, pp. 1–5.

[25] F. Pezone, S. Barbarossa, and P. Di Lorenzo, "Goal-oriented communication for edge learning based on the information bottleneck," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8832–8836.

[26] 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.

[27] S. Sreekumar and D. Gündüz, "Distributed hypothesis testing over discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2044–2066, 2020.

[28] J. Wei, E. Dupraz, and P. Mary, "Asymptotic and non-asymptotic rate-loss bounds for linear regression with side information," in *31st European Signal Processing Conference, EUSIPCO*, 2023.

[29] J. Wei, P. Mary, and E. Dupraz, "Rate-loss regions for polynomial regression with side information," in *Submitted to the 2024 International Zurich Seminar on Information and Communication*, 2023.

[30] P. A. Stavrou and M. Kountouris, "The role of fidelity in goal-oriented semantic communication: A rate distortion approach," *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3918–3931, 2023.

[31] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[32] E. Tuncel and D. Gündüz, "Identification and lossy reconstruction in noisy databases," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 822–831, 2014.

[33] S. C. Draper, "Universal incremental Slepian-Wolf coding," in *42nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Citeseer, 2004, pp. 1332–1341.

[34] V. Kostina and S. Verdu, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.

[35] V. Kostina and S. Verdú, "A new converse in rate-distortion theory," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, 2012, pp. 1–6.

[36] S. Watanabe, S. Kuzuoka, and V. Y. Tan, "Nonasymptotic and second-order achievability bounds for coding with side-information," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1574–1605, 2015.

[37] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[38] W. Härdle, M. Müller, S. Sperlich, A. Werwatz *et al.*, *Nonparametric and semiparametric models*. Springer, 2004, vol. 1.

[39] T. J. Hastie, "Generalized additive models," in *Statistical models in S*. Routledge, 2017, pp. 249–307.

[40] R. M. Gray, "Conditional rate-distortion theory," Stanford Univ CA Stanford Electronic Labs, Tech. Rep., 1972.

[41] A. Kipnis and G. Reeves, "Gaussian approximation of quantization error for estimation from compressed data," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2029–2033.

[42] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-ii: General sources," *Information and control*, vol. 38, pp. 60–80, 1978.

[43] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *2011 Data Compression Conference*, 2011, pp. 53–62.

[44] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 881–903, 2014.

[45] M. Hayashi, "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1562–1575, 2006.

[46] P. Cuff, "Distributed channel synthesis," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7071–7096, 2013.

[47] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1–8.

[48] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer Publishing Company, Incorporated, 2014.

[49] A. S. Goldberger, *Econometric theory*. New York, Wiley, 1964.

[50] F. Gotze, "On the Rate of Convergence in the Multivariate CLT," *The Annals of Probability*, vol. 19, no. 2, pp. 724 – 739, 1991. [Online]. Available: https://doi.org/10.1214/aop/1176990448