Rate-Loss Regions for Polynomial Regression with Side Information

Jiahui Wei, Philippe Mary, Elsa Dupraz

Rate-Loss Regions for Polynomial Regression with Side Information

Jiahui Wei[1,2], Philippe Mary[2], and Elsa Dupraz[1]
[1] IMT Atlantique, CNRS UMR 6285, Lab-STICC, Brest, France
[2] Univ. Rennes, INSA, IETR, UMR CNRS, Rennes, France

***Abstract*—In the context of goal-oriented communications, this paper addresses the achievable rate versus generalization error region of a learning task applied on compressed data. The study focuses on the distributed setup where a source is compressed and transmitted through a noiseless channel to a receiver performing polynomial regression, aided by side information available at the decoder. The paper provides the asymptotic rate generalization error region, and extends the analysis to the non-asymptotic regime. Additionally, it investigates the asymptotic trade-off between polynomial regression and data reconstruction under communication constraints. The proposed achievable scheme is shown to achieve the minimum generalization error as well as the optimal rate-distortion region.**

***Index Terms*—Information theory, source coding, statistical learning, rate-distortion theory, generalization error**

I. Introduction

Learning under communication constraints has received increased attention recently, for instance for distributed learning and sensor networks applications [1]. When considering a rate-limited channel, one key question is whether the design principles for the encoder and decoder for a learning task still align with those of traditional communication systems, where the main goal is data reconstruction.

To address this issue, researchers have explored simple distributed learning problems involving two correlated sources X and Y, where X is the source to be encoded and Y serves as side information at the decoder. Distributed hypothesis testing has been extensively studied for specific hypothesis tests on the joint distribution P_{XY}, and asymptotic limits on the Type-II error exponent have been determined in [2]–[4]. Furthermore, [5] demonstrated that the rate required for estimating a parameter θ from the joint distribution P_{XY} is less than the rate necessary for reconstructing the source. Finally, [6] developed a universal achievable bound on the learning generalization error, applicable to a wide range of distributed learning problems involving two sources. However, it was later shown in [7] that this bound is quite loose when applied to linear regression. Building upon [7], this paper focuses on the wider problem of polynomial regression and aims to establish achievable generalization error bounds that improve over the ones presented in [6]. Despite its simplicity, polynomial regression, captures essential learning theory concepts and is widely applied in signal and image processing, *e.g.*, [8], [9].

Morever, this paper investigates a secondary, yet significant concern, which is the trade-off between data reconstruction and learning under communication constraints. In this matter, [10] demonstrated that there indeed exists a tradeoff between data reconstruction and visual perception. Similar tradeoffs have been observed for other problems, such as hypothesis testing in [2], or identifying noisy data in a database in [11]. All previous works utilize distortion as the figure of merit for data reconstruction and employ distinct measures for the learning aspect; like a divergence between two distributions in [10], and the type-II error exponent in [2]. Unfortunately, none of these metrics are applicable to polynomial regression, underlining the need for a different analysis in our case.

Least squares regression, a fundamental statistical prediction problem, has been extensively investigated in literature. The ordinary least squares (OLS) estimator is a popular regression method, and its generalization error with k predictors and n samples is known to scale as $\frac{k}{n-k+1}$ [12]. However, this result does not take into account the communication constraint, which is an important consideration in many practical scenarios. In the context of polynomial regression, this paper determines the minimum achievable source coding rate under a constraint on the generalization error for both asymptotic and non-asymptotic regimes. The regions are derived using both standard asymptotic information theory tools [13], [14] and finite-length tools [15], and they improve over the bounds established by [6]. Additionally, the analysis reveals that no trade-off exists between data reconstruction and polynomial regression in terms of coding rate.

The outline of the paper is as follows. Section II defines the problem of coding for polynomial regression. Section III introduces the asymptotic rate-loss bounds. Section IV provides the rate-loss bounds in finite blocklength. Section V shows numerical results.

II. Problem statement

A. Notation

Throughout this article, random variables and their realizations are denoted with capital and lower-case letters, respectively, *e.g.*, X and x. Random vectors of length n are denoted $\boldsymbol{X} = [X_1, ..., X_n]^T$, and $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{C}[\boldsymbol{X}]$ are the expected value and the covariance matrix of $\boldsymbol{X}$, respectively.

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.

Next, $\underline{\boldsymbol{X}} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p]$ is a matrix gathering a p-length sequence of random vectors $\boldsymbol{X}_i$, $i \in [\![1, p]\!]$. We use $\mathrm{Tr}(\underline{\boldsymbol{X}})$ to denote the trace of matrix $\underline{\boldsymbol{X}}$, while $\lambda_{\max}(\underline{\boldsymbol{X}})$ and $\lambda_{\min}(\underline{\boldsymbol{X}})$ are the maximum and minimum eigenvalues of matrix $\underline{\boldsymbol{X}}$, 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 $|f|$ denotes the cardinality of $\mathcal{Y}$. Finally $\log(\cdot)$ denotes the base-2 logarithm.

B. Source definitions

Let $(X, Y) \sim P_{XY}$ be a pair of jointly distributed random variables, where X is the source to be encoded and Y is the side information only available at the decoder, see Figure 1. For simplicity and without loss of generality, we consider $\mathbb{E}[Y] = 0$. We define $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{k-1}]^T \in \mathbb{R}^k$, and $\boldsymbol{Y}^\star = [Y^0, Y^1, ..., Y^{k-1}]^T \in \mathbb{R}^k$, where Y^i is the variable Y raised to power i. We assume that X follows a polynomial model of order k defined as

$$X = \sum_{i=0}^{k-1} \beta_i Y^i + N = \boldsymbol{\beta}^T \boldsymbol{Y}^\star + N, \tag{1}$$

where $N \sim \mathcal{N}(0, \sigma^2)$ follows a Gaussian distribution with mean 0 and variance σ^2. The vector $\boldsymbol{\beta}$ is constant and unknown at the transmitter.

C. Polynomial Regression

Polynomial regression aims at estimating the parameter vector $\hat{\boldsymbol{\beta}}$ from realizations, or noisy realizations, of X and Y. As a standard supervised learning problem, polynomial regression consists of two phases. We use X, Y to denote symbols generated at the training phase, and $\tilde{X}$, $\tilde{Y}$ for symbols generated at the inference phase. The training phase consists of estimating $\boldsymbol{\beta}$ from a training sequence composed by the available side information $\boldsymbol{Y}$ and by a coded version of $\boldsymbol{X}$ which is denoted $\boldsymbol{U}$. The inference phase consists of calculating estimates of the symbols $\tilde{X}$ as $\hat{X} = \hat{\boldsymbol{\beta}} \tilde{\boldsymbol{Y}}^\star$, where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ from the training phase. Note that the inference phase does not need any data transmission, since the side information $\tilde{Y}$ is directly available to the decoder.

Following the notation introduced by Raginsky in [6], we next formalize the problem as follows. Let $\mathcal{F}$ be the set of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of the form $f(y) = \boldsymbol{\alpha}^T \boldsymbol{y}^\star$, where $\boldsymbol{\alpha} \in \mathbb{R}^k$. Polynomial regression outputs a sequence of functions $\widehat{f}^{(n)} \in \mathcal{F}$, called predictors, such that $\widehat{f}^{(n)} : \mathcal{Z}^n \times \mathbb{R} \to \mathbb{R}$, where $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) \in \mathcal{Z}^n$ is a training sequence in which $\boldsymbol{U}$ and $\boldsymbol{Y}$ are sequences of length n. Given that $\widehat{f}^{(n)} \in \mathcal{F}$, we can equivalently write

$$\widehat{f}^{(n)}(\boldsymbol{Z}, y) = \boldsymbol{\alpha}(\boldsymbol{Z})^T \boldsymbol{y}^\star, \tag{2}$$

where $\boldsymbol{\alpha} : \mathcal{Z}^n \to \mathbb{R}^k$.

Consider the quadratic loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(x, \hat{x}) = (x - \hat{x})^2$. The minimum expected loss is defined as in [6], [7] as[1]

$$L^\star(\mathcal{F}, \boldsymbol{\beta}) = \inf_{f \in \mathcal{F}} \mathbb{E}\left[\ell(X, f(Y))\right]. \tag{3}$$

[1]One may also define a loss over a sequence. However, since the samples from the training and inference phases are i.i.d. it does not change the analysis.

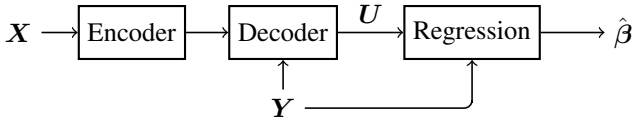

Fig. 1. Coding scheme for regression

The generalization error is defined as

$$G(\widehat{f}^{(n)}, \boldsymbol{\beta}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) | \boldsymbol{Z}\right]. \tag{4}$$

where $(\tilde{X}, \tilde{Y}) \sim P_{XY}$ is independent from $\boldsymbol{Z}$, the training sequence. The generalization error being a random variable due to the conditioning on $\boldsymbol{Z}$, the quantity $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right]$ is referred to as the expected generalization error.

In the previous expressions, the minimum expected loss (3) simply expresses the average gap between X and $f(Y)$, for the function f that minimizes the quantity $\mathbb{E}[\ell(X, f(Y))]$ over the space of polynomial functions $\mathcal{F}$. However, there is no guarantee that this optimal function f can be obtained from training. On the opposite, the generalization error measures the learning performance as the expected loss for a certain training sequence $\boldsymbol{Z}$. This training sequence allows to produce an estimated function $\widehat{f}^{(n)}(\boldsymbol{Z}, \cdot)$ which can then be used to evaluate new samples $\hat{\tilde{X}} = \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})$ at the inference phase. Especially, it is easy to show that $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right] \geq L^\star(\mathcal{F}, \boldsymbol{\beta})$. Therefore, the gap $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right] - L^\star(\mathcal{F}, \boldsymbol{\beta})$ is a key quantity to characterize the performance of a coding scheme dedicated to learning, and this is why our rate-learning regions will be expressed from this quantity.

D. Coding scheme

The coding scheme is analogue to the one for linear regression in [7]. However, the theoretical analysis differs and becomes more complex, as will be described in the next sections.

Definition 1. *A polynomial regression scheme at rate R is defined by a sequence $\{(e_n, d_n, R, \hat{f}^{(n)}))\}$ 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 the 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, l, ε) code for the sequence $\{(e_n, d_n, R, \hat{f}^{(n)})\}$ and $\varepsilon \in (0, 1)$ is a code with $|e_n| = M$ such that*

$$\mathbb{P}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta}) \geq l\right] \leq \varepsilon \text{ and } \frac{\log M}{n} \leq R. \tag{5}$$

Definition 3. *For fixed l and blocklength n, the finite blocklength rate-loss functions with excess loss ε is defined by:*

$$R(n, l, \varepsilon) = \inf_R \{\exists \ \ (n, M, l, \varepsilon) \text{ code}\} \tag{6}$$

Definition 4. *A pair (R, δ) is said to be achievable if there exists a sequence $\{(e_n, d_n, R, \hat{f}^{(n)})\}$ such that*

$$\limsup_{n\to\infty} \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})\right] \leq L^*(\mathcal{F}, \boldsymbol{\beta}) + \delta \tag{7}$$

As discussed in Section II-C, the achievable region is defined in terms of gap between $\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})\right]$ and $L^*(\mathcal{F}, \boldsymbol{\beta})$. Although the regions defined in this section pertain to rate-generalization error regions, for the sake of simplicity and with a minor deviation in terminology, we refer to them as rate-loss regions in the subsequent discussions.

III. Asymptotic Bound on the Rate-Loss function

In [6, Theorem 3.3], it is shown that, for a quadratic loss function, the generalization error can be bounded as:

$$\begin{aligned} L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{\beta}) \leq \limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})^{\frac{1}{2}}\right] &\leq L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{\beta}) \\ &\quad + 2\mathbb{D}_{X|Y}(R)^{1/2} \end{aligned} \tag{8}$$

where $\mathbb{D}_{X|Y}(R)$ is the conditional distortion-rate function. It can be shown that for the polynomial regression, the minimum expected loss in (3) is $L^\star(\mathcal{F}, \boldsymbol{\beta}) = \sigma^2$. In this section, we build a coding scheme that allows to improve the upper bound in (8) for the polynomial model.

A. Rate-loss region

Theorem 1. *Given any rate $R > 0$, the pair $(R, 0)$ is achievable for the polynomial regression scheme with squared loss, for sources (X, Y) following the polynomial model* (1)*.*

This result states that the minimum generalization error which is given by the loss function $L^\star(\mathcal{F}, \boldsymbol{\beta})$ in (8) can be achieved with any arbitrary rate R, as long as the training sequence is long enough. The proof of the Theorem is based on an achievability scheme built on a Gaussian test channel. This test channel is known for being optimal for joint Gaussian sources when considering data reconstruction [16], although it may be suboptimal for other models like the one we consider in this paper. However, in our case, we show that this test channel achieves the optimal rate-loss region $(R, 0)$ for polynomial regression, and we further discuss its optimality for data reconstruction in Section III-C.

B. Proof of Theorem 1 : Achievability scheme

Let us consider the test channel $U = \alpha(X + \Phi)$, where $\Phi \sim \mathcal{N}(0, \sigma_\Phi^2)$ is independent of X, and α and σ_Φ^2 are two parameters which depend on the distribution of X and Y.

The parameters $\boldsymbol{\beta}$ and the joint distribution P_{XY} are unknown to the encoder and decoder but the noise variance of the model, i.e. σ^2, is assumed to be known at the encoder. Hence, the transmission rate is perfectly known at the encoder and the variable-rate scheme in [14] becomes a fixed-rate coding scheme in our setup. The same idea of binning is used and the de-binning is performed based on the empirical mutual information between $\boldsymbol{x}$ and $\boldsymbol{u}$ evaluated thanks to the type of $\boldsymbol{x}$ transmitted in a prefix transmission [14]. Given that $D < \sigma_x^2$ and $(X \mid Y)$ is Gaussian, we 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.

Then, for a training sequences $(\boldsymbol{y}, \boldsymbol{u})$, the OLS estimator $\hat{\boldsymbol{\beta}}$ is given by [17, Chapter 7]

$$\hat{\boldsymbol{\beta}} = \alpha^{-1}(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T})^{-1}\underline{\boldsymbol{Y}}^\star \boldsymbol{u}. \tag{9}$$

where $\underline{\boldsymbol{Y}}^\star = [\boldsymbol{Y_1^\star}, ..., \boldsymbol{Y_n^\star}] \in \mathbb{R}^{k\times n}$ and this estimator has the following statistical properties :

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \;\text{ and }\; \mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right] = \frac{1}{\alpha^2}\sigma^2_{U|Y}(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T})^{-1} \tag{10}$$

where $\mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right]$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{Y}$. Hence, the generalization error (4) can be rewritten as

$$\begin{aligned} G(\widehat{f}^{(n)}, \boldsymbol{\beta}) &= \mathbb{E}_{\tilde{X}\tilde{Y}}\left[[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{N}^T\boldsymbol{N}|\boldsymbol{Z}\right] \\ &= [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T}\right][\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \sigma^2. \end{aligned} \tag{11}$$

Let $\underline{\tilde{\boldsymbol{\Sigma}}} = \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star T}\right]$ and $\underline{\boldsymbol{\Sigma}} = \frac{1}{n}\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}$. Then, the expected generalization error is

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \\ &= \sigma^2 + \mathbb{E}\left[\frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))^T \underline{\tilde{\boldsymbol{\Sigma}}}\frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))\right] \\ &= \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}\mathbb{E}\left[\mathrm{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)\right]. \end{aligned} \tag{12}$$

The next step is to show that $\mathbb{E}\left[\mathrm{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)\right]$ is bounded by some constant C for n large enough. The following proposition bounds the trace of a product of two matrices by their eigenvalues.

Proposition 1. *[18, p 340] (Ruhe's trace inequality). If $\underline{\boldsymbol{U}}$ and $\underline{\boldsymbol{V}}$ are $k \times k$ positive semidefinite Hermitian matrices with eigenvalues $\lambda_i(\underline{\boldsymbol{U}}), \lambda_i(\underline{\boldsymbol{V}})$, $i \in \{1, \cdots, k\}$ then*

$$\mathrm{Tr}(\underline{\boldsymbol{U}}\underline{\boldsymbol{V}}) \leq \sum_{i=1}^{k} \lambda_i(\underline{\boldsymbol{U}})\lambda_i(\underline{\boldsymbol{V}}) \tag{13}$$

Lemma 1. *If $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{B}}$ are real symmetric matrices, then:*

$$\lambda_{\min}(\underline{\boldsymbol{A}}) \geq \lambda_{\min}(\underline{\boldsymbol{B}}) - ||\underline{\boldsymbol{A}} - \underline{\boldsymbol{B}}|| \tag{14}$$

Proof: Let $\boldsymbol{x}$ be a vector such that $||x||_2 = 1$, by Cauchy-Schwartz inequality, for a real symmetric matrix $\underline{\boldsymbol{M}}$, we have

$$-||\underline{\boldsymbol{M}}|| \leq \boldsymbol{x}^T\underline{\boldsymbol{M}}\boldsymbol{x} \leq ||\underline{\boldsymbol{M}}||. \tag{15}$$

With the properties of eigenvalues, we have

$$\lambda_{\min}(\underline{\boldsymbol{M}}) \leq \boldsymbol{x}^T\underline{\boldsymbol{M}}\boldsymbol{x} \leq \lambda_{\max}(\underline{\boldsymbol{M}}). \tag{16}$$

For real symmetric matrices $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{B}}$, we have

$$\boldsymbol{x}^T\underline{\boldsymbol{A}}\boldsymbol{x} = \boldsymbol{x}^T\underline{\boldsymbol{B}}\boldsymbol{x} + \boldsymbol{x}^T(\underline{\boldsymbol{A}} - \underline{\boldsymbol{B}})\boldsymbol{x}. \tag{17}$$

Applying the above inequalities shows the desired result. ■

We remark that $\underline{\boldsymbol{\Sigma}}$ is an estimator of the covariance matrix of $\boldsymbol{Y}$. Then, from Proposition 1 and Lemma 1, for n large enough, $\mathrm{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)$ is bounded almost surely by:

$$\mathrm{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right) \leq k\frac{\lambda_{\max}(\tilde{\boldsymbol{\Sigma}})}{\lambda_{\min}(\tilde{\boldsymbol{\Sigma}}) - ||\tilde{\boldsymbol{\Sigma}} - \underline{\boldsymbol{\Sigma}}||}. \tag{18}$$

Substituting this into (12) with some constant $C = \frac{\lambda_{\max}(\tilde{\boldsymbol{\Sigma}})}{\lambda_{\min}(\tilde{\boldsymbol{\Sigma}})}$ and the fact that $||\tilde{\boldsymbol{\Sigma}} - \underline{\boldsymbol{\Sigma}}|| \to 0$ almost surely, shows that the expected generalization error is upper bounded by

$$\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \leq \sigma^2 + \frac{(\sigma^2 + \sigma_\Phi^2)}{n}kC \tag{19}$$

Thus $\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \to \sigma^2$ as $n \to \infty$, which completes the proof.

Our result closes the gap between the lower bound and the upper bound from [6] (see equation (8)). In order to provide a bound applicable to a wide range of problems, the upper bound from [6] considered both the observation noise between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the distortion between $\boldsymbol{X}$ and $\boldsymbol{U}$. While in our result, by the Gaussian test channel and OLS estimation from $\boldsymbol{U}$ and $\boldsymbol{Y}$, we show that the quantification error term in (19), and hence the distortion term, is vanishing with the block-length n.

C. Trade-off between data reconstruction and polynomial regression

In this section, we show that the previous achievability scheme considered for polynomial regression also achieves the optimal Wyner-Ziv rate-distortion function for data reconstruction, for sources modeled by (1).

Corollary 1. *For a pair of sources (X, Y) modeled from* (1)*, there is no trade-off in terms of coding rate between distortion and polynomial regression generalization error.*

Proof: We first investigate the conditional setup in which the side information Y is also available at the encoder. Since the random variable $(X|Y) \sim \mathcal{N}(0, \sigma^2)$, the following conditional rate-distortion function can be achieved [19]

$$R_{X|Y}(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right), \tag{20}$$

where $D = \mathbb{E}\left[(X - \hat{X})^2\right]$ is the distortion. We now show that in the Wyner-Ziv setup where Y is only available at the decoder, the rate-distortion function $R_{\mathrm{WZ}}(D)$ is equal to $R_{X|Y}(D)$ when considering the same test channel $U = \alpha(X + \Phi)$ as in the proof of Theorem 1, with $\alpha = \frac{\sigma^2 - D}{\sigma^2}$, and $\sigma_\Phi^2 = \frac{D\sigma^2}{\sigma^2 - D}$. By using the proposed achievability scheme, the random variable U can be recovered perfectly at the decoder, and then produces $\hat{X} = U + (1-\alpha)\boldsymbol{\beta}^T\boldsymbol{Y}^\star$. This allows us to evaluate $\mathbb{E}\left[(X - \hat{X})^2\right] = (\alpha - 1)^2\sigma^2 + \alpha^2\sigma_\Phi^2$. Replacing α and σ_Φ^2 by their expressions leads to $\mathbb{E}\left[(X - \hat{X})^2\right] = D$. Second, the Wyner-Ziv rate-distortion function has expression [16]

$$I(X;U) - I(Y;U) = \frac{1}{2}\log_2\left(\frac{\sigma^2 + \sigma_\Phi^2}{\sigma_\Phi^2}\right)$$

where the equality comes from the fact that N and Φ are Gaussian random variables. Replacing σ_Φ^2 by its expression gives that $R_{\mathrm{WZ}}(D) = R_{X|Y}(D)$ in (20), which shows that the Gaussian test channel is optimal when considering our polynomial source model. Note that in the previous derivation, we considered that $\boldsymbol{\beta}$ is perfectly known. If this is not the case, $\hat{X}$ is computed from $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$, and following the same derivation as for the generalization error permits to show that $\mathbb{E}\left[(X - \hat{X})^2\right] \to D$ as $n \to \infty$.

This result differs from the other ones in literature that show that there is a tradeoff between reconstruction and learning, such as for the hypothesis testing problem for instance [2]. ■

IV. Rate-Loss non-asymptotic bound

In the finite-blocklength regime, not all codewords satisfy the generalization error constraint, and hence the excess probability, defined in Definition 2, has to be taken into account. The characterization of the non-asymptotic achievable bound for the rate-generalization error region is built from the rate-distortion problem in finite blocklength regime, studied in [15]. Similarly, we define the information-loss density vector as follows:

$$\boldsymbol{i}(U, X, Y, \tilde{X}, \tilde{Y}) := \begin{bmatrix} -\log\dfrac{P_{U|Y}(U|Y)}{P_U(U)} \\ \log\dfrac{P_{U|X}(U|X)}{P_U(U)} \\ \ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})) \end{bmatrix} \tag{21}$$

where the third term is specific to our non-linear regression problem. The expectation of $\boldsymbol{i}$ over the distribution $P_{UXY\tilde{X}\tilde{Y}}$ is $\boldsymbol{J} = \left[-I(U;Y), I(U;X), \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^n, \boldsymbol{\beta})\right]\right]^T$, where the sum of the first two components gives the Wyner-Ziv coding rate. The covariance matrix of (21) is

$$\boldsymbol{V} = \mathbb{C}\left[\boldsymbol{i}(U, X, Y, \tilde{X}, \tilde{Y})\right]. \tag{22}$$

Let k be a positive integer and $\boldsymbol{V} \in \mathcal{R}^{k\times k}$ be a positive-semi-definite matrix. Given a Gaussian random vector $\boldsymbol{B} \sim \mathcal{N}(0, \boldsymbol{V})$, the dispersion region is [20]

$$\mathscr{S}(\boldsymbol{V}, \varepsilon) := \{\boldsymbol{b} \in \mathbb{R}^k : \Pr(\boldsymbol{B} \leq \boldsymbol{b}) \geq 1 - \varepsilon\}. \tag{23}$$

By replacing the distortion measure by the generalization error, and adapting some steps of the analysis, we can obtain a similar result as Theorem 2 in [15]. Finally, by applying this theorem in conjunction with the multidimensional Berry-Esséen Theorem, we show that for all $0 < \varepsilon < 1$ and n sufficiently large, the (n, ε)-rate-generalization error function satisfies:

$$R_{\mathrm{b}}(n, \varepsilon, l) \leq \inf\left\{\boldsymbol{M}^T\left(\boldsymbol{J} + \frac{\mathscr{S}(\boldsymbol{V}, \varepsilon)}{\sqrt{n}} + \frac{2\log n}{n}\mathbf{1}_3\right)\right\} \tag{24}$$

with $\boldsymbol{M} = [1 \quad 1 \quad 0]^T$ and $\mathbf{1}_3 = [1 \quad 1 \quad 1]^T$.

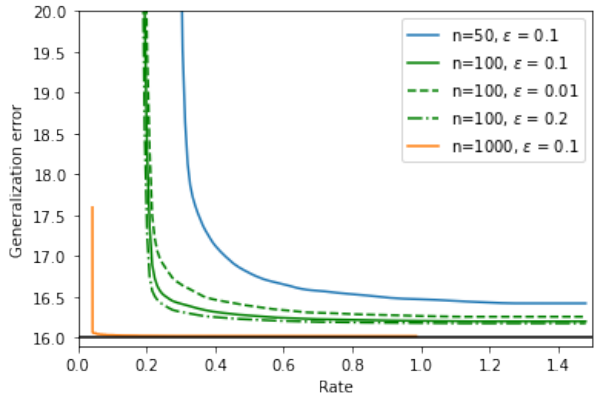

Fig. 2. Non-asymptotic rate-generalization error region labeled on the blocklength n and the excess loss probability ε.

V. Numerical results

Let us consider $X = \beta_0 + \beta_1 Y + \beta_2 Y^2 + N$, and assume that Y is uniform over $[-1, 1]$. We also set $\boldsymbol{\beta} = [2, 3, 1]^T$ and $\sigma^2 = 16$. From the theorem of change variable, for $\beta_2 > 0$ and $\beta_1^2 + 4\beta_2(v - \beta_0) \geq 0$, the distribution of $V = \boldsymbol{\beta}^T \boldsymbol{Y}^\star$ is:

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

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

$$P_U(u) = \frac{1}{\alpha\sqrt{2\pi(\sigma^2 + \sigma_\Phi^2)}} \int_{-\infty}^{\infty} P_V(v) e^{-\frac{(\frac{u}{\alpha} - v)^2}{2(\sigma^2 + \sigma_\Phi^2)}} dv \quad (25)$$

which can be evaluated numerically. Using (24) with $(U|Y) \sim \mathcal{N}(0, \alpha^2(\sigma^2 + \sigma_\Phi^2))$ and $(U|X) \sim \mathcal{N}(0, \alpha^2\sigma_\Phi^2)$, we can estimate the information-density-loss vector by generating a large number of samples, and thus estimate the dispersion region in (23). Figure 2 shows the boundaries of the achievable rate-loss region for different parameters n and ε. The black line represents the best achievable generalization error, i.e. σ^2. We observe that the achievable region enlarges when the source size, n, or the excess probability increases. Indeed, when the excess probability is larger, the proportion of codewords which exceeds the generalization error constraint is larger, and this situation occurs for smaller rate. Moreover, for a fixed excess probability, increasing n allows to reduce the rate since the poorly reconstructed U is compensated by the large number of samples for estimating the regression parameters. These results do not deal with an outer bound at finite blocklength, i.e. a rate-loss region that *cannot* be exceeded, and the region outside the boundary needs further investigation.

VI. Conclusion

This paper provided achievable rate-generalization error regions for the polynomial regression problem in both asymptotic and non-asymptotic regimes. An important result of our study states that asymptotically there is no trade-off between data reconstruction and polynomial regression under communication constraints. The characterization of the outer bound (converse) for the rate-generalization error region is also of great interest and would allow to refine the analysis. The developed framework could be extended to more complex learning taks, such as non-parametric estimation, in the future.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] E. Siggiridou and D. Kugiumtzis, "Dimension reduction of polynomial regression models for the estimation of granger causality in high-dimensional time series," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5638–5650, 2021.

[9] G. D. Finlayson, M. Mackiewicz, and A. Hurlbert, "Color correction using root-polynomial regression," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1460–1470, 2015.

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

[11] 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.

[12] J. Mourtada, "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices," *The Annals of Statistics*, vol. 50, no. 4, 2022.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*, 2nd ed. Springer, 2011, vol. 143.

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

[20] 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.