Asymptotic and non-Asymptotic Rate-Loss Bounds for Linear Regression with Side Information

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

***Abstract*—In the framework of goal-oriented communications, this paper investigates the fundamental achievable rate-loss function of a learning task performed on compressed data. It considers the setup where the data, collected remotely, are compressed and sent over a noiseless channel to a server that aims at applying linear regression on compressed data and side information. The paper contributions are threefold: i) the rate-loss region is first derived in the asymptotic regime, *i.e.*, when the length of the source tends to infinity, (ii) the tradeoff between data reconstruction and linear regression is investigated from the asymptotic rate-loss region, and iii) the approach is extended to the finite blocklength regime.**

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

I. Introduction

The problem of learning under communication constraints arises frequently in various contexts, including distributed learning and estimation in sensor networks. Often, the agents that gather the data are spatially separated from the location where the learning takes place, which requires to establish a communication link over a rate-limited channel. Of recent and increasing interest to the scientific community is the question of whether the code and decoder for a learning task should be designed in the same manner as in classical communication, wherein the primary objective is data reconstruction.

This question was first investigated for some simple distributed learning problems that consider two correlated sources X and Y, where X is the source to be encoded and Y is available as side information at the decoder. In [1], it was shown that the rate required for estimating a parameter θ related to the joint probability distribution P_{XY} is lower than the Slepian-Wolf rate needed for lossless coding of the two sources. Distributed hypothesis testing was also widely investigated, and in [2]–[4], asymptotic bounds on the error exponent of Type-II errors were derived for some hypothesis tests on the joint distribution P_{XY}. In an effort to gain a broader understanding of the issue, [5] derived a generic achievable bound on the generalization error, and showed that this bound is applicable to a wide range of distributed learning problems with two sources.

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.

Rather than solely establishing some achievable bounds for the learning problem, an alternative approach consists of investigating the trade-off between data reconstruction and learning in terms of coding rate. In [6], this problem was modeled using classical rate-distortion theory with an additional constraint on the visual perception, represented by a divergence measure between two distributions. It was shown that the achievable rate-distortion-perception region is reduced compared to the usual Shannon rate-distortion lower bound, thus highlighting a trade-off between the two criterion. This trade-off was also evidenced in [2] for hypothesis testing, and in [7], [8] for noisy data identification in a database, both versus data reconstruction.

In this paper, we consider the simple problem of linear regression between the source X and the side information Y. We first address the fundamental issue of determining the minimum achievable source coding rate at any given blocklength n under a constraint on the generalization error. We provide the resulting rate-generalization error region, both in the asymptotic regime and in the non-asymptotic regime, by relying on standard information theory tools [9], [10] and on finite-length tools [11]–[13], i.e. information density, dispersion matrix and excess probability, respectively. It is worth mentioning that the finite-length tools in the works above are dedicated to data reconstruction and had not yet been applied to the communication for learning problems. We further utilize the obtained regions to explore the trade-off between linear regression and data reconstruction.

Despite its apparent simplicity, linear regression is still of significant interest in supervised machine learning, as well as various other fields such as economics or biology. Additionally, previous findings suggested that there is always a trade-off between distortion and learning constraints [2], [7], [8], while for linear regression we show that there is no trade-off. Moreover, for the specific case of linear regression, we improve the achievable rate-generalization error bound introduced in [5], which was rather loose. Finally, using the simple problem of linear regression as a starting point allows us to develop a framework for analyzing the asymptotic and non-asymptotic performance of learning schemes under communication constraints, which could be applied to more complex learning problems in the future.

The outline of the paper is as follows. Section II defines the problem of coding for linear regression. Section III provides the asymptotic rate-loss bounds while Section IV gives

the related results in finite blocklength. Section V presents numerical results.

II. Problem statement

Throughout this article, random variables X are denoted by upper-case letters, and their realizations x by lower-case letters. Vectors $\boldsymbol{X} = (X_1, ..., X_n)$ of length n are denoted by upper-case bold-face letters and their realizations $\boldsymbol{x} = (x_1, ..., x_n)$ are denoted by lower-case bold-face letters. $\mathcal{X}$ is a set with cardinality $|\mathcal{X}|$. $\mathbb{E}[X]$ and $\mathbb{V}[X]$ are the expected value and the variance of X, respectively, and $\mathbb{C}\text{ov}(X, Y)$ is the covariance of the random variables X and Y. Finally $\log(\cdot)$ denotes base-2 logarithm.

A. Source definitions

Let X and Y be jointly distributed random variables, where Y is the side information only available at the decoder. We assume that Y follows a Gaussian distribution $Y \sim \mathcal{N}(0, \sigma_Y^2)$. The source X is defined from a linear model as:

$$X = \beta_0 + \beta_1 Y + N, \tag{1}$$

where $N \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise, and $\beta_0, \beta_1 \in \mathbb{R}$ are constant parameters. Therefore, $X \sim \mathcal{N}(\beta_0, \sigma_X^2)$, where $\sigma_X^2 = \beta_1^2 \sigma_Y^2 + \sigma^2$. We further assume that all parameters σ_Y^2, σ_X^2, σ^2, β_0, β_1 are unknown, both by the encoder and decoder. We define $\mathcal{S} = \{\sigma_x^2, \sigma_y^2, \sigma^2, \beta_0, \beta_1\}$ as the set of source parameters which fully defines the joint Gaussian probability distribution P_{XY}.

B. Linear regression

Instead of reconstructing the source X at the decoder, we aim to realize a linear regression, that is to estimate β_0 and β_1 from some source sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ of length n, as illustrated in Figure 1. Following the notation introduced by Raginsky in [5], we formalize the problem as follows.

Let $\mathcal{F}$ be the set of linear functions $f : \mathbb{R} \to \mathbb{R}$ of the form $f(y) = \alpha_0 + \alpha_1 y$, where $\alpha_0, \alpha_1 \in \mathbb{R}$. Linear regression outputs a sequence of functions $\widehat{f}^{(n)} \in \mathcal{F}$, called predictors, such that $\widehat{f}^{(n)} : \boldsymbol{Z}^n \times \mathbb{R} \to \mathbb{R}$, from a training sequence $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) \in \mathcal{Z}^n$, where $\boldsymbol{U}$ is a random sequence. Given that linear regression estimates the coefficients α_0 and α_1 from $\boldsymbol{Z}$, we hence have

$$\widehat{f}^{(n)}(\boldsymbol{Z}, y) = \alpha_0(\boldsymbol{Z}) + \alpha_1(\boldsymbol{Z}) y. \tag{2}$$

Consider the quadratic loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(x, \hat{x}) = (x - \hat{x})^2$. For a certain function $f \in \mathcal{F}$, the expected loss is defined as[1]

$$L(f, \mathcal{S}) = \mathbb{E}\left[\ell(X, f(Y))\right] \tag{3}$$

and the minimum expected loss is defined as

$$L^\star(\mathcal{F}, \mathcal{S}) = \inf_{f \in \mathcal{F}} L(f, \mathcal{S}). \tag{4}$$

[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 change the analysis.

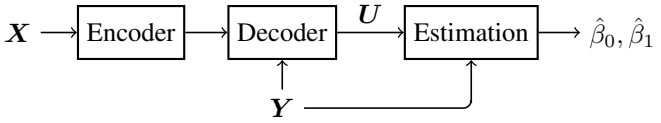

Fig. 1. Coding scheme for linear regression

The generalization error is defined as

$$L(\widehat{f}^{(n)}, \mathcal{S}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) | \boldsymbol{Z}\right]. \tag{5}$$

where the pair $(\tilde{X}, \tilde{Y})$ is distributed according to P_{XY} and it is independent from $\boldsymbol{Z}$, the training sequence. The generalization error is a random variable that depends on $\boldsymbol{Z}$, and we refer to as the expected generalization error the quantity $\mathbb{E}_{\boldsymbol{Z}}\left[L\left(\widehat{f}^{(n)}, \mathcal{S}\right)\right]$.

C. Coding scheme for linear regression

Definition 1. *A learning scheme at rate R is defined by a sequence $\{(e_n, d_n, R, \mathcal{L}_n)\}$ with an encoder*

$$e_n : \mathcal{X}^n \to \{1, 2, ..., M_n\}$$

a decoder

$$d_n : \mathcal{Y}^n \times \{1, 2, ..., M_n\} \to \mathcal{U}^n$$

and the learner

$$\mathcal{L}_n : \mathcal{Y}^n \times \mathcal{U}^n \to \mathcal{F}$$

such that

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

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

$$\mathbb{E}_{\boldsymbol{Z}}\left[L(\widehat{f}^{(n)}, \mathcal{S})\right] \leq l \tag{6}$$

and

$$\frac{\log M}{n} \leq R.$$

Definition 3. *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*

$$\mathbb{P}\left[L(\widehat{f}^{(n)}, \mathcal{S}) \geq l\right] \leq \varepsilon \tag{7}$$

and

$$\frac{\log M}{n} \leq R$$

Definition 4. *For fixed l and blocklength n, the finite blocklength rate-loss functions with average loss and with excess loss are respectively defined by:*

$$\begin{aligned} R(n, l) &= \inf_R \{\exists \ \ (n, M, l) \ code\} \\ R(n, l, \varepsilon) &= \inf_R \{\exists \ \ (n, M, l, \varepsilon) \ code\} \end{aligned} \tag{8}$$

Definition 5. *A pair (R,δ) is said to be achievable if an $(n,M,l)-$code exists such as*

$$\limsup_{n\to\infty}\mathbb{E}_{\boldsymbol{Z}}\left[L(\hat{f}^{(n)},\mathcal{S})\right]\leq L^*(\mathcal{F},\mathcal{S})+\delta \tag{9}$$

Even though the regions defined in this section correspond to rate-generalization error regions, in what follows, we often refer to them as rate-loss regions for simplicity and by a slight abuse of language.

III. Asymptotic Bound on the Rate-Loss function

In [5], it is shown that the generalization error can be bounded as

$$\sigma\leq\limsup_{n\to\infty}\mathbb{E}\left[L(\hat{f}^{(n)},P)^{\frac{1}{2}}\right]\leq\sigma(1+2^{-R+1}), \tag{10}$$

where P is the distribution of (X,Y) in [5], and σ^2 is the minimum expected loss for linear regression, i.e. $L^\star(\mathcal{F},\mathcal{S})=\sigma^2$. Our next result provides an achievable rate-loss region which tightens the upper bound in (10) for Gaussian sources.

A. Main result

Theorem 1. *Given any rate $R>0$, the pair $(R,0)$ is achievable for a linear regression learner for Gaussian sources with squared loss.*

This result shows that the minimum expected generalization error $L^\star(\mathcal{F},\mathcal{S})=\sigma^2$ can be achieved even with a very small rate R, as long as the length of the training sequence is large enough. This also provides a refined upper bound σ in (10), which does not depend on R and is actually tight with the lower bound. The result of Theorem 1 comes from an achievability scheme which we now describe.

B. Proof of Theorem 1: Achievability scheme

Define U as the output of a test channel $P_{U|X}$ which satisfies the Markov chain $U-X-Y$ such as:

$$U=\alpha(X+\Phi) \tag{11}$$

where $\alpha=\frac{\sigma_X^2-D}{\sigma_X^2}$, $\Phi\sim\mathcal{N}\left(0,\sigma_\phi^2\right)$, $\sigma_\phi^2=\frac{D\sigma_X^2}{\sigma_X^2-D}$ and $D>0$ is a parameter. Since we assume that the parameters $\mathcal{S}$ of the joint Gaussian distribution P_{XY} are unknown to the encoder and decoder, we resort to the achievability scheme proposed in [10].

This scheme relies on a prefix transmission to estimate σ_x^2 so that it is known by the encoder and by the decoder. This completely defines the test channel (11). The codebook is constructed by randomly generating 2^{nR_1} sequences $\boldsymbol{u}$, which are then randomly and uniformly assigned to one of 2^{nR} bins, where $R_1>R$. For a given source sequence $\boldsymbol{x}$, the encoder picks a sequence $\boldsymbol{u}$ which is typical with $\boldsymbol{x}$, and incrementally sends to the decoder the index s to which $\boldsymbol{u}$ belongs. Universal de-binning is applied to retrieve the sequence $\boldsymbol{u}$ from the bin index and from the side information sequence $\boldsymbol{y}$. As soon as the empirical mutual information of the sequences pair satisfies a time-varying threshold θ_k, the decoder declares that the source input has been correctly decoded, otherwise the transmitter stops if $nI(X;U)$ bits have been transmitted. It is shown in [10] that the de-binning error probability can be made as small as desired, as long as the time-varying threshold is appropriately chosen and by letting the observation length n be sufficiently large.

The original achievability scheme of [10] then relies on an universal estimation method to reconstruct the sequence $\boldsymbol{x}$ as $\widehat{\boldsymbol{x}}$ such that $\mathbb{E}\left[d(X,\widehat{X})\right]<D$, where $d(\cdot,\cdot)$ is a squared distortion measure. Here, instead, we apply the linear regression directly over the intermediate vector $\boldsymbol{u}$ after de-binning. In this case, for $D<\sigma_x^2$, the results of [10] show that the rate

$$R_b(D)=\frac{1}{2}\log\left(1+\frac{\sigma^2}{\sigma_\phi^2}\right) \tag{12}$$

is achievable for $\mathbb{E}_{XU}\left[d(X,U)\right]\leq D$.

Then, according to (1) and (11), the conventional least square estimation of β_0 and β_1 from $\boldsymbol{u}$ and $\boldsymbol{y}$ is [14, Chapter 6]

$$\begin{aligned}\hat{\beta}_1&=\frac{1}{\alpha}\frac{\sum_{i=1}^n u_iy_i-n\bar{u}\bar{y}}{\sum_{i=1}^n y_i^2-n\bar{y}^2},\\ \hat{\beta}_0&=\frac{1}{\alpha}(\bar{u}-\bar{y}\hat{\beta}_1)\end{aligned} \tag{13}$$

where $\bar{y}$ and $\bar{u}$ are the empirical means of vectors $\boldsymbol{y}$ and $\boldsymbol{u}$, respectively. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Let us denote B_0 and B_1 the random variables representing $\hat{\beta}_0$ and $\hat{\beta}_1$. The generalization error defined in (5) is:

$$\begin{aligned}L(\widehat{f}^{(n)},\mathcal{S})&=\mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X},\widehat{f}^{(n)}(\boldsymbol{Z},\tilde{Y})\right)|\boldsymbol{Z}\right] \\ &=\mathbb{E}_{\tilde{X}\tilde{Y}}\left[\left((B_0-\beta_0)+(B_1-\beta_1)\tilde{Y}-N\right)^2\mid\boldsymbol{Z}\right]\\ &=(B_0(\boldsymbol{Z})-\beta_0)^2+\mathbb{E}[\tilde{Y}^2](B_1(\boldsymbol{Z})-\beta_1)^2+\sigma^2\end{aligned} \tag{14}$$

since $\mathbb{E}\left[N\right]=0$ and $\mathbb{E}\left[\tilde{Y}\right]=0$. We then need to express

$$\mathbb{E}_{\mathbf{Z}}\left[L(\widehat{f}^{(n)},\mathcal{S})\right]=\mathbb{E}_{\mathbf{Y}}\left[\mathbb{E}_{\mathbf{U}}[L(\widehat{f}^{(n)},\mathcal{S})|\boldsymbol{Y}=\boldsymbol{y}]\right]. \tag{15}$$

By defining $S_{yy}=\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})^2$ and given that $\mathbb{E}[B_1]=\beta_1$, $\mathbb{E}[B_0]=\beta_0$, and $\mathbb{V}\left[(U|\boldsymbol{Y}=\boldsymbol{y})\right]=\alpha^2(\sigma^2+\sigma_\phi^2)$, we have

$$\begin{aligned}\mathbb{V}\left[B_1|\boldsymbol{Y}=\boldsymbol{y}\right]&=\mathbb{V}\left[\left(\frac{\sum_{i=1}^n U_iy_i-n\bar{U}\bar{y}}{\alpha\sum_{i=1}^n y_i^2-n\bar{y}^2}|\boldsymbol{Y}=\boldsymbol{y}\right)\right]\\ &=\left(\frac{1}{n\alpha S_{yy}}\right)^2\sum_{i=1}^n(y_i-\bar{y})^2\mathbb{V}\left[U_i|\boldsymbol{Y}=\boldsymbol{y}\right]\\ &=\frac{\sigma^2+\sigma_\phi^2}{nS_{yy}},\end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathbb{V}\left[B_0|\boldsymbol{Y}=\boldsymbol{y}\right]&=\mathbb{V}\left[\frac{1}{\alpha}(\bar{U}-\bar{y}B_1)|\boldsymbol{Y}=\boldsymbol{y}\right]\\ &=\frac{1}{\alpha^2}\left(\mathbb{V}\left[\bar{U}|\boldsymbol{Y}=\boldsymbol{y}\right]+\bar{y}^2\mathbb{V}\left[B_1|\boldsymbol{Y}=\boldsymbol{y}\right]\right.\\ &\quad\left.-2\bar{y}\mathbb{C}\mathrm{ov}(\bar{U},B_1|\boldsymbol{Y}=\boldsymbol{y})\right)\\ &=\frac{\sigma^2+\sigma_\phi^2}{n}\left(1+\frac{\bar{y}^2}{\alpha^2S_{yy}}\right).\end{aligned} \tag{17}$$

The expected generalization error in (15) can be expressed as:

$$\begin{aligned}&\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)},\mathcal{S})]\\&=\sigma^2+\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{V}\left[B_0|\boldsymbol{Y}=\boldsymbol{y}\right]\right]+\sigma_Y^2\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{V}\left[B_1|\boldsymbol{Y}=\boldsymbol{y}\right]\right]\\&=\sigma^2+\frac{\sigma^2+\sigma_\phi^2}{n}\left(1+\frac{1}{\alpha^2}\mathbb{E}\left[\frac{\bar{y}^2}{S_{yy}}\right]+\sigma_Y^2\mathbb{E}\left[\frac{1}{S_{yy}}\right]\right).\end{aligned}\tag{18}$$

First, $\frac{nS_{yy}}{\sigma_Y^2}\sim\chi^2(n-1)$, $\frac{n\bar{y}^2}{\sigma_Y^2}\sim\chi^2(1)$, where $\chi^2(n)$ is a Chi-squared distribution with n degrees of freedom [14, Chapter 5], and $\frac{(n-1)\bar{y}^2}{S_{yy}}\sim\mathcal{F}(1,n-1)$, where $\mathcal{F}$ is the Fisher–Snedecor distribution [14, Chapter 5]. By the properties of the Chi-squared and Fischer-Snedecor distributions, we get [14, Chapter 5]

$$\mathbb{E}\left[\frac{1}{S_{yy}}\right]=\frac{n}{(n-3)\sigma_Y^2},\quad\mathbb{E}\left[\frac{\bar{y}^2}{S_{yy}}\right]=\frac{1}{n-3},\tag{19}$$

which gives

$$\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)},\mathcal{S})]=\sigma^2+\frac{(\sigma^2+\sigma_\phi^2)(1+2n\alpha^2-3\alpha^2)}{n(n-3)\alpha^2}.\tag{20}$$

Then, $\lim_{n\to\infty}\mathbb{E}_{\mathbf{Z}}[L(\widehat{f}^{(n)},\mathcal{S})]=\sigma^2$. Therefore, as $n\to\infty$, $L(\hat{f}^{(n)},\mathcal{S})\to L^*(\mathcal{F},\mathcal{S})=\sigma^2$, which completes the proof of Theorem 1.

C. On the trade-off between loss and distortion

We now discuss the tradeoff between linear regression and data reconstruction. The proof of Theorem 1 shows that the considered scheme permits to achieve the minimum expected loss σ^2 for any R, as long as the training sequence is long enough. In [10], it is shown that the same scheme allows to achieve the Wyner-Ziv rate-distortion function [9] for joint Gaussian sources. In [10], the reconstruction of the source X is realized by universal estimation from U after debinning, while we consider least square estimation of β_0 and β_1 instead. This allows us to state the following result.

Corollary 1. *For joint Gaussian sources, there is no trade-off in terms of coding rate between distortion and the linear regression generalization error.*

IV. Non-asymptotic bound on the Rate-Loss function

Theorem 1 provides the asymptotic rate-generalization function for linear regression in the setup of Figure 1. We now aim to investigate the finite-blocklength regime. In the classical problem of rate-distortion with side information in finite blocklength, the probability of excess distortion plays an important role, since not all the codewords can satisfy the distortion constraint. This problem has been well studied recently, by using the notion of distortion dispersion [11], [12], [15]. In the following, based on the work in [12], we derive a non-asymptotic achievability bound for the rate-generalization error region.

A. Definitions

Let us consider the following three sets, similar to those defined in [12]:

$$\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}}):=\left\{(u,y):\log\frac{P_{Y|U}(y|u)}{P_Y(y)}\geq\gamma_{\mathrm{p}}\right\},\tag{21}$$

$$\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}):=\left\{(u,x):\log\frac{P_{X|U}(x|u)}{P_X(x)}\leq\gamma_{\mathrm{c}}\right\},\tag{22}$$

$$\mathcal{T}_e(l):=\left\{(\tilde{x},\tilde{y},u,y):\ell(\tilde{x},\hat{f}^{(n)}(\mathbf{z},\tilde{y}))\leq l\right\},\tag{23}$$

where γ_p, γ_c are predefined thresholds, and l is the target generalization error. The first two sets already appeared in [12] for the conventional setup of data reconstruction with side information, while the third one is specific to our linear regression problem. Accordingly, define the information-density-loss vector as

$$\mathbf{i}(U,X,Y,\tilde{X},\tilde{Y}):=\begin{bmatrix}-\log\dfrac{P_{Y|U}(Y|U)}{P_Y(Y)}\\\log\dfrac{P_{X|U}(X|U)}{P_X(X)}\\\ell(\tilde{X},\hat{f}^{(n)}(\mathbf{Z},\tilde{Y}))\end{bmatrix}.\tag{24}$$

Taking the expectation over the distribution $P_{UXY\tilde{X}\tilde{Y}}$ of this vector gives

$$\mathbf{J}(\mathbf{i}):=\mathbb{E}[\mathbf{i}(U,X,Y,\tilde{X},\tilde{Y})]\tag{25}$$

$$=\begin{bmatrix}-I(U;Y)\\I(U;X)\\\mathbb{E}_{\boldsymbol{Z}\tilde{X}\tilde{Y}}\left[\ell(\tilde{X},\hat{f}^{(n)}(\mathbf{Z},\tilde{Y})\right]\end{bmatrix}.\tag{26}$$

The sum of the first two components gives the Wyner-Ziv coding rate defined in equation (12). The covariance matrix of this vector is

$$\mathbf{V}=\mathbb{C}\mathrm{ov}(\mathbf{i}(U,X,Y,\tilde{X},\tilde{Y})).\tag{27}$$

Let k be a positive integer and $\mathbf{V}\in\mathcal{R}^{k\times k}$ be a positive-semi-definite matrix. Given a Gaussian random vector $\mathbf{B}\sim\mathcal{N}(0,\mathbf{V})$, the dispersion term is defined w.r.t. the covariance matrix as [13]

$$\mathscr{S}(\mathbf{V},\varepsilon):=\{\mathbf{b}\in\mathbb{R}^k:\Pr(\mathbf{B}\leq\mathbf{b})\geq1-\varepsilon\}.\tag{28}$$

B. Main result

By replacing the excess distortion measure in [12], by the generalization error, and adapting some steps of the analysis, we obtain the following result.

Theorem 2. *For arbitrary constants $\gamma_p,\gamma_c,l\geq0$, and positive integer N, there exists an (n,M,l,ε) code satisfying*

$$\begin{aligned}\varepsilon\leq&P_{UXY\hat{X}}[(u,y)\in\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c\cup(u,x)\in\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c\\&\cup(\tilde{x},\tilde{y},u,y)\in\mathcal{T}_e(l)^c]\\&+\frac{N}{2^{\gamma_{\mathrm{p}}}|\mathcal{M}|}+\frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{N}}.\end{aligned}\tag{29}$$

Proof: The proof is omitted due to the lack of space but follows the same steps as in [12]. ■

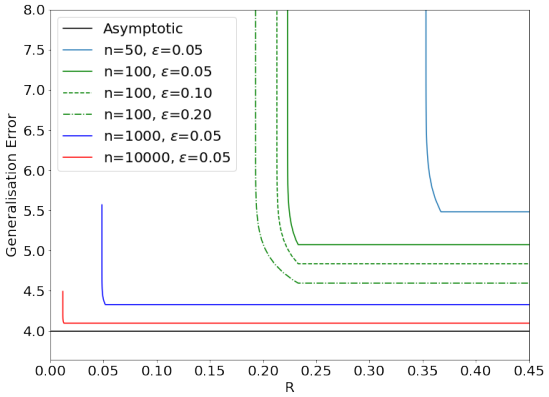

Fig. 2. Non-asymptotic rate-generalization error region labeled on the blocklength n and the excess loss probability ε.

By choosing $\gamma_p = \log \frac{N}{|\mathcal{M}_n|} + \log n$ and $\gamma_c = \log N - \log n$, and by applying Theorem 2 together with the multidimensional Berry-Esséen Theorem, we derive the following achievable second-order coding region as follows.

Theorem 3. *For every $0 < \varepsilon < 1$, and n sufficiently large, the (n, ε)-rate-generalization error function satisfies:*

$$R_{\mathrm{b}}(n, \varepsilon, l) \leq \inf \left\{ \mathbf{M} \left(\mathbf{J} + \frac{\mathscr{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_3 \right) \right\} \quad (30)$$

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

Proof: The proof is omitted due to the lack of space but follows the same steps as in [12]. ∎

V. Numerical results

In this section, we consider $\beta_0 = 2$, $\beta_1 = 5$, $\sigma^2 = 4$, $\sigma_y^2 = 16$, and $\sigma_x^2 = \sigma^2 + \beta_1^2 \sigma_y^2 = 404$. We use Theorem 3 to plot the rate-generalization error region in finite blocklength, where the dispersion term $\mathscr{S}(\mathbf{V}, \varepsilon)$ is estimated by generating samples from the known joint distribution P_{UXY}. The information density $i(x; u|y)$ is then estimated from these samples, and the dispersion is estimated from (28).

Figure 2 plots the boundaries of finite blocklength achievable regions for various values of n and ε. It illustrates that as n increases, whatever ε is, the achievable region gets closer to the asymptotic one shown in black. Similarly, for a fixed blocklength n, when the constraint on the excess loss becomes less stringent, the achievable region enlarges. This implies that for a fixed generalization error and blocklength n, a lower rate is possible if a higher excess probability is allowed, since only the average error is taken into account. These results do not tell us what happens for a rate-generalization error pair outside the region, as our findings are achievability results, and the characterization of the outer bound remains an open problem.

VI. Conclusion

In this paper, we investigated achievable rate-generalization error regions for the linear regression problem. Achievable region have been provided in both non-asymptotic and asymptotic regimes, i.e. with and without the excess loss probability. As an important outcome, our study shows that there is no trade-off between the distortion for data reconstruction and the generalization error for linear regression, given that the blocklength n is sufficiently large. The characterization of the outer bound (converse) for the rate-loss region is also of great interest and would allow to refine the analysis. Future works also include the extension of this approach to more complex learning tasks.

References

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

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

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

[7] E. Tuncel, "Capacity/storage tradeoff in high-dimensional identification systems," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2097–2106, 2009.

[8] 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.

[9] 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.

[10] S. C. Draper, "Universal incremental slepian-wolf coding," in *Proc. 42nd Allerton Conf. on Communication, Control and Computing*. Citeseer, 2004, pp. 1332–1341.

[11] 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.

[12] 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.

[13] 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.

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

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