Information theoretic analysis and coding scheme for learning : regression and more

Jiahui WEI, Philippe MARY and Elsa DUPRAZ

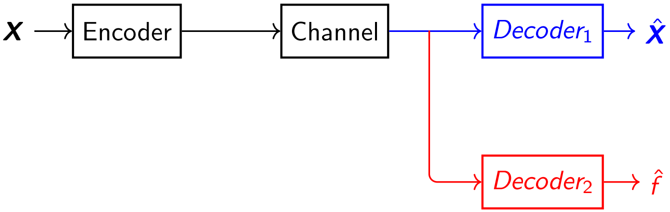
INSA Rennes, IMT Atlantique

June 7, 2024

Context

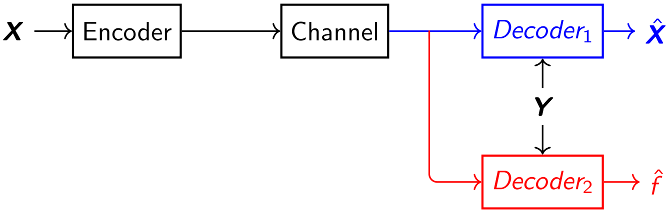
- Conventional communication : reconstruct the source even with distortion => Rate-Distortion theory

Context



- Conventional communication : reconstruct the source even with distortion => Rate-Distortion theory
- Goal-oriented communication : construct the coding scheme to address some machine learning problems

Context

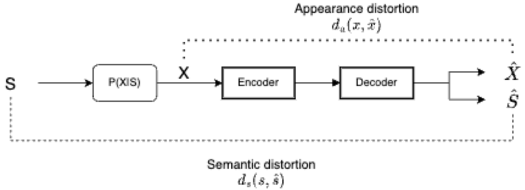


- Conventional communication : reconstruct the source even with distortion => Rate-Distortion theory
- Goal-oriented communication : construct the coding scheme to address some machine learning problems

- **Question : Do we need the same method for coding?**

State of the art

- Existing works : **Rate-distortion framework with semantic and appearance distortion**[1]; Rate for parameter estimation[2]; Hypothesis testing[3]; Rate-distortion-perception trade-off[4].



[1] 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.

[2] 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

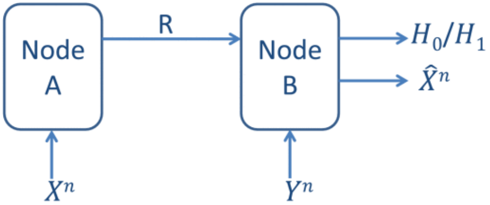
[3] G. Katz, P. Piantanida and M. Debbah, "Distributed Binary Detection With Lossy Data Compression," in IEEE Transactions on Information Theory. Aug. 2017

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

[5] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

State of the art

- Existing works : Rate-distortion framework with semantic and appearance distortion[1]; Rate for parameter estimation[2]; **Hypothesis testing**[3]; Rate-distortion-perception trade-off[4].



[1] 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.

[2] 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

[3] G. Katz, P. Piantanida and M. Debbah, "Distributed Binary Detection With Lossy Data Compression," in IEEE Transactions on Information Theory. Aug. 2017

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

[5] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

State of the art

- Existing works : Rate-distortion framework with semantic and appearance distortion[1]; Rate for parameter estimation[2]; Hypothesis testing[3]; Rate-distortion-perception trade-off[4].
    - Rate-distortion + **Goal**;
    - **Trade-off** between the task and data reconstruction

[1] 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.

[2] 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

[3] G. Katz, P. Piantanida and M. Debbah, "Distributed Binary Detection With Lossy Data Compression," in IEEE Transactions on Information Theory. Aug. 2017

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

[5] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

State of the art

- Existing works : Rate-distortion framework with semantic and appearance distortion[1]; Rate for parameter estimation[2]; Hypothesis testing[3]; Rate-distortion-perception trade-off[4].
    - Rate-distortion + **Goal**;
    - **Trade-off** between the task and data reconstruction

- Regression
    - A fundamental statistical method;
    - A rate-loss bound for general regression with side information is provided by Raginsky[5];
    - This bound is **loose** and the **trade-off** is not investigated.

[1] 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.

[2] 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

[3] G. Katz, P. Piantanida and M. Debbah, "Distributed Binary Detection With Lossy Data Compression," in IEEE Transactions on Information Theory. Aug. 2017

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

[5] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

Outline

Outline

Problem statement

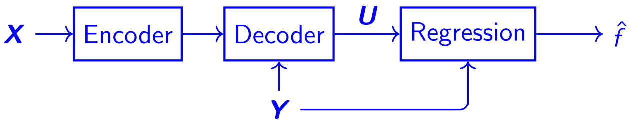


Figure 1: Coding scheme for regression

Regression model for X and Y:

$$X = f(Y) + N, \quad (1)$$

where $N \sim \mathcal{N}(0, \sigma^2)$ independent from X and Y.

1. Training phase with sequence $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) => \hat{f}$;

Problem statement

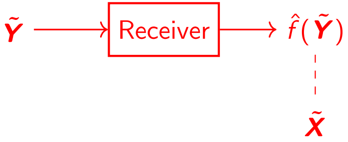


Figure 1: Inference phase

Regression model for X and Y:

$$X = f(Y) + N, \tag{1}$$

where $N \sim \mathcal{N}(0, \sigma^2)$ independent from X and Y.

1. Training phase with sequence $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) => \hat{f}$;
2. Inference phase $=>$ Generalization error.

Parametric regression : OLS

$$X = \boldsymbol{\beta}^T \boldsymbol{Y}^\star + N = \sum_{i=0}^{k-1} \beta_i h_i(Y) + N, \tag{2}$$

where $\boldsymbol{Y_j^\star} = [h_0(Y_j), ..., h_{k-1}(Y_j)]^T$ and $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_{k-1}]^T$ is unknown. Further define $\underline{\boldsymbol{Y}}^\star = [\boldsymbol{Y_1^\star}, ..., \boldsymbol{Y_n^\star}] \in \mathbb{R}^{k\times n}$, we have

- OLS estimation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by[6]

$$\hat{\boldsymbol{\beta}} = \left(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}\right)^{-1} \underline{\boldsymbol{Y}}^\star \boldsymbol{X}. \tag{3}$$

- Properties :

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \quad \text{and} \quad \mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right] = \sigma^2_{X|Y}\left(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}\right)^{-1}, \tag{4}$$

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.

[6] Chapter 7, A. C. Rencher and G. B. Schaalje, Linear models in statistics. John Wiley & Sons, 2008

Non-parametric regression : kernel regression

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

without any prior knowledge of the regression form.

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

$$\int_{\mathbb{R}} K(x)dx = 1, \quad \int_{\mathbb{R}} xK(x)dx = 0, \quad \text{and} \quad 0 \leq \int_{\mathbb{R}} x^2 K(x)dx \leq \infty. \tag{6}$$

- The Nadaraya-Watson Kernel regression over $(\boldsymbol{X}, \boldsymbol{Y})$ is defined as [7]:

$$\hat{f}(Y) = \frac{\sum_{j=1}^{n} K\left(\frac{Y - Y_j}{h}\right) X_j}{\sum_{j=1}^{n} K\left(\frac{Y - Y_j}{h}\right)}. \tag{7}$$

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

Non-parametric regression : kernel regression

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

without any prior knowledge of the regression form.

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

$$\int_{\mathbb{R}} K(x)dx = 1, \quad \int_{\mathbb{R}} xK(x)dx = 0, \quad \text{and} \quad 0 \leq \int_{\mathbb{R}} x^2 K(x)dx \leq \infty. \tag{6}$$

- The Nadaraya-Watson Kernel regression over $(\boldsymbol{X}, \boldsymbol{Y})$ is defined as [7]:

$$\hat{f}(Y) = \frac{\sum_{j=1}^{n} K\left(\frac{Y - Y_j}{h}\right) X_j}{\sum_{j=1}^{n} K\left(\frac{Y - Y_j}{h}\right)}. \tag{7}$$

Attention : X is not available in our setup so the regression needs to be processed with the compressed observation U.

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

Definitions

Definition

A regression scheme at rate R is defined by a sequence $\{(e_n, d_n, R, \mathcal{L}_n)\}$ with

$$\begin{aligned} \textit{an encoder} \quad & e_n : \mathcal{X}^n \to \{1, 2, ..., M_n\}, \\ \textit{a decoder} \quad & d_n : \mathcal{Y}^n \times \{1, 2, ..., M_n\} \to \mathcal{U}^n, \\ \textit{and } \textbf{the learner} \quad & \mathcal{L}_n : \mathcal{Y}^n \times \mathcal{U}^n \to \mathcal{F}, \end{aligned}$$

such that

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

Loss function $\ell(x, \hat{x}) = (x - \hat{x})^2$.
For a fixed function f, the **expected loss** is defined as

$$L(f) = \mathbb{E}\left[\ell(X, f(Y))\right] \tag{8}$$

and the **minimum expected loss** is defined as the

$$L^\star(\mathcal{F}) = \inf_{f \in \mathcal{F}} L(f) = \sigma^2 \tag{9}$$

Generalization error

The **generalization error** is defined as

$$G(\widehat{f}^{(n)}, \boldsymbol{Z}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) | \boldsymbol{Z}\right]. \quad (10)$$

where $(\tilde{X}, \tilde{Y}) \sim P_{XY}$ is independent from $\boldsymbol{Z}$ (i.e. $\left(\boldsymbol{U}^n, \boldsymbol{Y}^n\right)$).

Objective

Derive the rate-generalization error regions

2 methods of regression : **Parametric regression using OLS estimator** and **Kernel regression**.

Generalization error

The **generalization error** is defined as

$$G(\widehat{f}^{(n)}, \boldsymbol{Z}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right)|\boldsymbol{Z}\right]. \quad (10)$$

where $(\tilde{X}, \tilde{Y}) \sim P_{XY}$ is independent from $\boldsymbol{Z}$ (i.e. $\left(\boldsymbol{U}^n, \boldsymbol{Y}^n\right)$).

Objective

Derive the rate-generalization error regions

2 methods of regression : **Parametric regression using OLS estimator** and **Kernel regression**.

Our contributions :

- The rate-generalization error regions in both asymptotic and non-asymptotic regime;
- Improvement of the upper bound provided by Raginsky;
- Investigation of reconstruction vs regression trade-off.

Outline

Asymptotic Rate-generalization error regions

An $(\mathbf{n}, \mathbf{M}, \mathbf{G})$ **code** for regression is a code with $|e_n| = M$ such that $\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{Z})\right] \leq G$.

Definition

A pair (R, δ) is said to be achievable if an $(n, M, G)-$code exists such as

$$\limsup_{n\to\infty} \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \leq L^*(\mathcal{F}, \boldsymbol{Z}) + \delta \quad (11)$$

Recall the result of Raginsky[8]

$$L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{Z}) \leq \limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})^{\frac{1}{2}}\right] \leq L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{Z}) + 2\mathbb{D}_{X|Y}(R)^{1/2} \quad (12)$$

[8] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

Asymptotic Rate-generalization error regions

An $(\mathbf{n}, \mathbf{M}, \mathbf{G})$ **code** for regression is a code with $|e_n| = M$ such that $\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{Z})\right] \leq G$.

Definition

A pair (R, δ) is said to be achievable if an $(n, M, G)-$code exists such as

$$\limsup_{n\to\infty} \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] \leq L^*(\mathcal{F}, \boldsymbol{Z}) + \delta \tag{11}$$

Recall the result of Raginsky[8]

$$L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{Z}) \leq \limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})^{\frac{1}{2}}\right] \leq L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{Z}) + 2\mathbb{D}_{X|Y}(R)^{1/2} \tag{12}$$

Theorem (Parametric & Kernel regression)

Given any rate $R > 0$, the pair $(\mathbf{R}, \mathbf{0})$ is achievable for the **parametric regression** *and* **kernel regression** *with squared loss, for sources (X, Y) following the regression model, that is*

$$\limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] = L^*(\mathcal{F}, \boldsymbol{Z}) \quad \text{and} \quad \delta = 0. \tag{13}$$

[8] M. Raginsky, "Learning from compressed observations," in IEEE Information Theory Workshop, 2007

Sketch of proof : Parametric regression

- Idea : **quantization + binning**. Consider a **Gaussian test channel**

$$U = \alpha(X + \Phi), \quad \text{with } \Phi \sim \mathcal{N}(0, \sigma_\Phi^2)$$

Remark : The method is based on **prefix transmission of types**[9] of observations + **binning** of conventional WZ coding.

[9] S. C. Draper, "Universal incremental slepian-wolf coding," in Proc. 42nd Allerton Conf. on Communication, Control and Computing. Citeseer, 2004

Sketch of proof : Parametric regression

- Idea : **quantization + binning**. Consider a **Gaussian test channel**

$$U = \alpha(X + \Phi), \quad \text{with } \Phi \sim \mathcal{N}(0, \sigma_\Phi^2)$$

- For a training sequences $(\boldsymbol{y}, \boldsymbol{u})$, the OLS estimator $\hat{\beta}$ becomes

$$\hat{\beta} = \alpha^{-1}(\underline{\boldsymbol{Y}\boldsymbol{Y}}^T)^{-1}\underline{\boldsymbol{Y}}\boldsymbol{u}. \tag{14}$$

where $\underline{\mathbf{Y}} = \begin{bmatrix} y_1^0 & \dots & y_n^0 \\ \dots & \dots & \dots \\ y_1^{k-1} & \dots & y_n^{k-1} \end{bmatrix}$.

- The generalization error can be rewritten as

$$G(\widehat{f}^{(n)}, \boldsymbol{Z}) = \sigma^2 + [\beta - \hat{\beta}]^T \mathbb{E}_{\tilde{Y}} \left[\tilde{\boldsymbol{Y}} \tilde{\boldsymbol{Y}}^T \right] [\beta - \hat{\beta}] \tag{15}$$

- Let $\underline{\tilde{\boldsymbol{\Sigma}}} = \mathbb{E}_{\tilde{Y}} \left[\underline{\tilde{\boldsymbol{Y}} \tilde{\boldsymbol{Y}}}^T \right]$, $\underline{\boldsymbol{\Sigma}} = \frac{1}{n}\underline{\boldsymbol{Y}\boldsymbol{Y}}^T$ and $C = \frac{\lambda_{max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{min}(\underline{\tilde{\boldsymbol{\Sigma}}})}$, the **expected generalization error** :

$$\mathbb{E}_{\boldsymbol{Z}} \left[G(\widehat{f}^{(n)}, \boldsymbol{Z}) \right] = \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(\underline{\tilde{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}}^{-1} \right) \right] \leq \sigma^2 + \frac{(\sigma^2 + \sigma_\Phi^2)}{n} kC \tag{16}$$

Sketch of proof : Kernel regression

- The same test channel, we suppose the following conditions for kernel regression :
  Y **bounded**; p_Y continuously differentiable and positively bounded; $\exists f', f''$;
  $\mathbf{n} \to \infty, \mathbf{h} \to \mathbf{0}$ and $\mathbf{nh} \to \infty$.

Sketch of proof : Kernel regression

- The same test channel, we suppose the following conditions for kernel regression : Y **bounded**; p_Y continuously differentiable and positively bounded; $\exists f', f''$; $\mathbf{n} \to \infty, \mathbf{h} \to \mathbf{0}$ and $\mathbf{nh} \to \infty$.
- The kernel regression between the sequence $(\boldsymbol{y}, \boldsymbol{u})$ becomes

$$\hat{f}(y) = \frac{\sum_{i=1}^{n} K(\frac{y-y_i}{h})\frac{u_i}{\alpha}}{\sum_{i=1}^{n} K(\frac{y-y_i}{h})}. \tag{17}$$

- The generalization error can be rewritten as

$$\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{Z})\right] = \sigma^2 + \int b_n^2(\tilde{y})p_Y(\tilde{y})d\tilde{y} + \int V_n(\tilde{y})p_Y(\tilde{y})d\tilde{y}. \tag{18}$$

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 $\mathbf{Z}$.
- We proved that

$$b_n(\tilde{y}) = \frac{h^2}{2}\left(2\frac{f'(\tilde{y})p_Y'(\tilde{y})}{p_Y(\tilde{y})} + f''(\tilde{y})\right)\int_{\mathbb{R}} u^2 K(u)du + o(h^2), \tag{19}$$

$$V_n(\tilde{y}) = \frac{(\sigma^2 + \sigma_\Phi^2)}{p_Y(\tilde{y})nh}\int_{\mathbb{R}} K^2(u)du + o\left(\frac{1}{nh}\right). \tag{20}$$

Regression-Reconstruction trade-off

Corollary

There is **no trade-off** *in terms of coding rate between distortion and regression generalization error. That is*

$$R(G, D) = R(D) \tag{21}$$

where $R(G, D)$ is the communication rate under reconstruction and regression constraints.

Regression-Reconstruction trade-off

Corollary

There is **no trade-off** *in terms of coding rate between distortion and regression generalization error. That is*

$$R(G, D) = R(D) \tag{21}$$

where $R(G, D)$ *is the communication rate under reconstruction and regression constraints.*

- Idea of the proof : show that the scheme provided for regression achieves the optimal rate-distortion region for reconstruction.

Regression-Reconstruction trade-off

Corollary

There is **no trade-off** *in terms of coding rate between distortion and regression generalization error. That is*

$$R(G, D) = R(D) \tag{21}$$

where $R(G, D)$ is the communication rate under reconstruction and regression constraints.

- Idea of the proof : show that the scheme provided for regression achieves the optimal rate-distortion region for reconstruction.
- For $X|Y \sim \mathcal{N}$, by replacing $\sigma_\Phi^2 = \frac{D\sigma^2}{\sigma^2 - D}$, we obtain

$$R_b(D) = \frac{1}{2}\log\left(\frac{\sigma^2 + \sigma_\Phi^2}{\sigma_\Phi^2}\right) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) \tag{22}$$

Outline

Non-asymptotic rate-distortion-generalization error regions

In finite block-length n, **the excess error probability** is $\mathbb{P}\left[G(\widehat{f}^{(n)}, \beta) \geq G \text{ or } d(X, \hat{X}) \geq D\right]$.

Non-asymptotic rate-distortion-generalization error regions

In finite block-length n, **the excess error probability** is $\mathbb{P}\left[G(\widehat{f}^{(n)}, \beta) \geq G \text{ or } d(X, \hat{X}) \geq D\right]$.

Definition

An $(n, M, G, D, \varepsilon)$ 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

$$\mathbb{P}\left[G(\widehat{f}^{(n)}, \beta) \geq G \text{ or } d(X, \hat{X}) \geq D\right] \leq \varepsilon \text{ and } \frac{\log M}{n} \leq R. \quad (23)$$

Definition

For fixed G and block-length n, the finite block-length rate-loss functions with excess loss is defined by:

$$R(n, G, D, \varepsilon) = \inf_{R}\{\exists \ (n, M, G, D, \varepsilon) \text{ code}\} \quad (24)$$

Non-asymptotic achievable regions

Define the **loss-information density** $\boldsymbol{i}$ as
$$\boldsymbol{i}(U,X,Y,\hat{X}) := \begin{bmatrix} -\log \dfrac{P_{U|Y}(U|Y)}{P_U(U)} \\ \log \dfrac{P_{U|X}(U|X)}{P_U(U)} \\ \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y}))\right] \\ d(X,\hat{X}) \end{bmatrix},$$

Let $\mathbf{J} = \mathbb{E}_{UXY\hat{X}}[\boldsymbol{i}]$ and $\underline{\boldsymbol{V}} = \mathbb{C}\left[\boldsymbol{i}(U,X,Y,\hat{X})\right]$.
The **dispersion region** is defined as $\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon) := \{\mathbf{b} \in \mathbb{R}^k : \Pr(\mathbf{B} \leq \mathbf{b}) \geq 1-\varepsilon\}$ with $\mathbf{B} \sim \mathcal{N}(0, \underline{\boldsymbol{V}})$.

Non-asymptotic achievable regions

Define the **loss-information density** $\boldsymbol{i}$ as
$$\boldsymbol{i}(U, X, Y, \hat{X}) := \begin{bmatrix} -\log \dfrac{P_{U|Y}(U|Y)}{P_U(U)} \\ \log \dfrac{P_{U|X}(U|X)}{P_U(U)} \\ \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y}))\right] \\ d(X, \hat{X}) \end{bmatrix},$$

Let $\mathbf{J} = \mathbb{E}_{UXY\hat{X}}[\boldsymbol{i}]$ and $\underline{\boldsymbol{V}} = \mathbb{C}\left[\boldsymbol{i}(U, X, Y, \hat{X})\right]$.

The **dispersion region** is defined as $\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon) := \{\mathbf{b} \in \mathbb{R}^k : \Pr(\mathbf{B} \leq \mathbf{b}) \geq 1 - \varepsilon\}$ with $\mathbf{B} \sim \mathcal{N}(0, \underline{\boldsymbol{V}})$.

Theorem

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

$$R_{\mathrm{b}}(n, \varepsilon, G, D) \leq \inf\left\{\mathbf{M}\left(\boldsymbol{J} + \frac{\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon)}{\sqrt{n}} + \frac{2\log n}{n}\mathbf{1}_4\right)\right\} \tag{25}$$

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

Sketch of proof

Consider the following sets similar to that of [WKT15][9]

$$\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}}) := \left\{(u,y) : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_{\mathrm{p}}\right\}, \tag{26}$$

$$\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}) := \left\{(u,x) : \log \frac{P_{X|U}(x|u)}{P_X(x)} \leq \gamma_{\mathrm{c}}\right\}, \tag{27}$$

$$\mathcal{T}_d(D) := \{(x,\hat{x}) : d(x,\hat{x}) \leq D\}, \tag{28}$$

$$\mathcal{T}_g(G) := \left\{(\boldsymbol{u},\boldsymbol{y}) : \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{z}, \tilde{Y}))\right] \leq G\right\}. \tag{29}$$

[9] 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.

Sketch of proof

Consider the following sets similar to that of [WKT15][9]

$$\mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}}) := \left\{(u, y) : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_{\mathrm{p}}\right\}, \tag{26}$$

$$\mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}}) := \left\{(u, x) : \log \frac{P_{X|U}(x|u)}{P_X(x)} \leq \gamma_{\mathrm{c}}\right\}, \tag{27}$$

$$\mathcal{T}_d(D) := \{(x, \hat{x}) : d(x, \hat{x}) \leq D\}, \tag{28}$$

$$\mathcal{T}_g(G) := \left\{(\boldsymbol{u}, \boldsymbol{y}) : \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{z}, \tilde{Y}))\right] \leq G\right\}. \tag{29}$$

Theorem

For arbitrary constants $\gamma_p, \gamma_c, G, D \geq 0$, and positive integer N, there exists an $(n, M, G, D, \varepsilon)$ code satisfying

$$\begin{aligned}\varepsilon \leq & P_{UXY\hat{X}}[(\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{T}_{\mathrm{p}}(\gamma_{\mathrm{p}})^c \cup (\boldsymbol{u}, \boldsymbol{x}) \in \mathcal{T}_{\mathrm{c}}(\gamma_{\mathrm{c}})^c \cup (\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{T}_g(G)^c \cup (\boldsymbol{x}, \hat{\boldsymbol{x}}) \in \mathcal{T}_d(D)^c] \\ & + \frac{N}{2^{\gamma_{\mathrm{p}}}|\mathcal{M}|} + \frac{1}{2}\sqrt{\frac{2^{\gamma_{\mathrm{c}}}}{N}}.\end{aligned} \tag{30}$$

[9] 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.

Numerical results

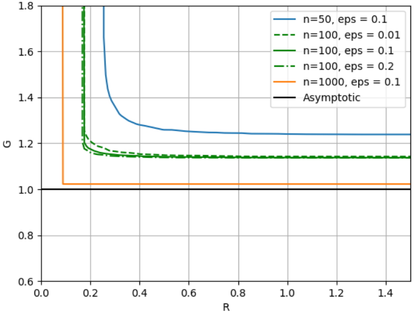


Figure 2: Rate-generalization error region for polynomial regression labeled on the block-length n and the excess probability ε.

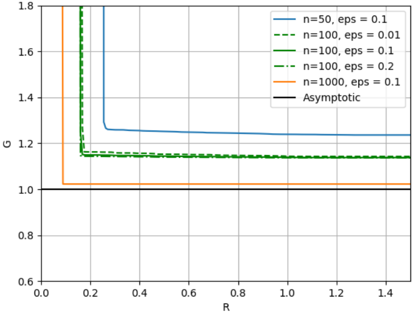


Figure 3: Rate-generalization error region for kernel regression labeled on the block-length n and the excess loss probability ε.

Non-asymptotic trade-off

Corollary

For $0 < \varepsilon < 1$ and n sufficiently large, there exists an achievable rate-distortion-generalization error region such that

$$R_b(n, G, D, \varepsilon) > \max\{R_b(n, G, \varepsilon), R_b(n, D, \varepsilon)\}. \tag{31}$$

And there is **no trade-off** *between generalization error of regression and reconstruction.*

By Gaussian approximation, the dispersion matrix $\mathscr{S}(\underline{\boldsymbol{V}}, \varepsilon)$ is determined by the correlation matrix $\underline{\boldsymbol{V}}$. We show that

$$\mathrm{Cov}\left(\mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell(\tilde{X}, \hat{f}^{(n)}(\mathbf{Z}, \tilde{Y}))\right], d(\boldsymbol{X}, \hat{\boldsymbol{X}})|\boldsymbol{Z} = \boldsymbol{z}\right) = 0 \tag{32}$$

Distortion-generalization error region

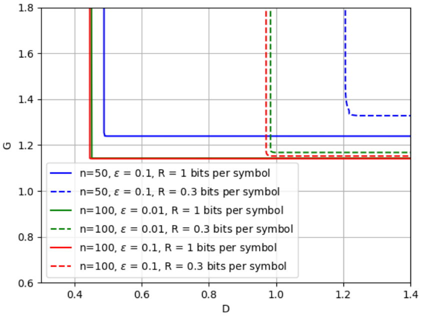


Figure 4: Distortion-generalization error region for polynomial regression on the block-length n, the excess loss probability ε and rate R.

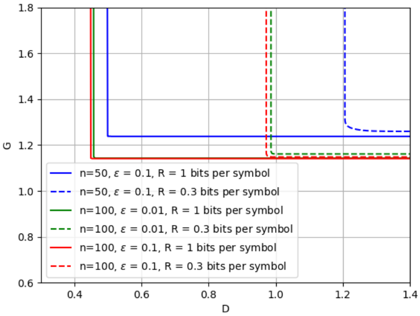


Figure 5: Distortion-generalization error region for kernel regression on the block-length n, the excess loss probability ε and rate R.

Outline

Practical coding scheme for parametric regression

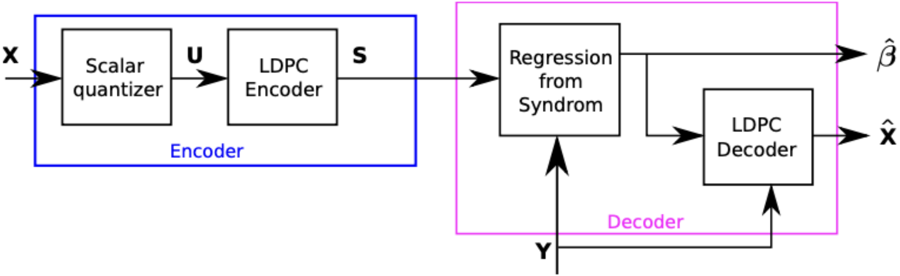


Figure 6: Practical coding scheme using LDPC codes

- Encoder
  - Scalar quantizer over 2^q levels (Lloyd-max or uniform)
  - LDPC encoder : $\boldsymbol{s} = \underline{\boldsymbol{H}}\boldsymbol{u}$ with LDPC codes in $GF(q)$
- Decoder
  - Maximum Likelihood parameter estimation over the syndrom $\boldsymbol{s}$: $\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} \mathbb{P}(\mathbf{s}|\mathbf{y})$
  - LDPC decoder from estimated $\hat{\beta}$

Numerical results for polynomial regression

- $Y \sim \mathcal{U}[-1,1], X = \beta_0 + \beta_1 Y + \beta_2 Y^2 + N$
- $n = 100$, regular (3, 6)-LDPC code in GF(4), with rate $r = \frac{1}{2}$
- $R = 1$ bit/symbol
- Lower bound : σ^2; Upper bound : $\sigma^2 + \frac{kC(\sigma^2+\sigma_\Phi^2)}{n}$

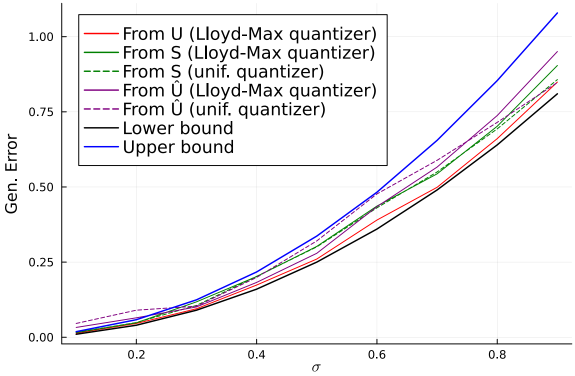


Figure 7: Generalization error with respect to σ for polynomial regression of order 3 with rate $R = 1$ bit/symbol

Numerical results for logistic regression

- $Y \sim \mathcal{U}[-1,1], X = \beta_0 + \frac{\beta_1}{1+e^{-2Y}} + \frac{\beta_2}{1+e^{-4Y}} + N$
- $n = 100$, regular (3, 6)-LDPC code in GF(4), with rate $r = \frac{1}{2}$
- $R = 1$ bit/symbol

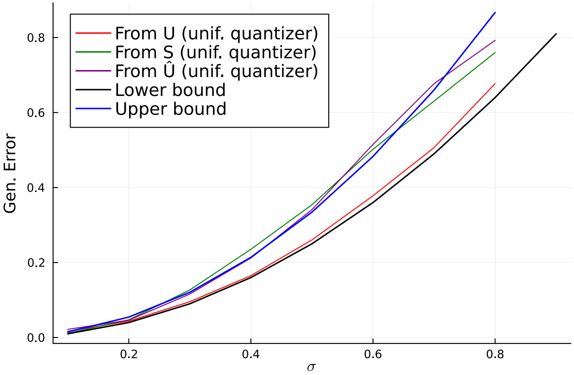


Figure 8: Generalization error with respect to σ for logistic regression of order 3 with rate $R = 1$ bit/symbol

Outline

Conclusion

- Conclusions :
  - In asymptotic regime: $R(D, G) = R(D)$;
  - In non-asymptotic regime: an achievable region with excess probability is provided;
  - **No trade-off** between regression and reconstruction;
  - Learning over compressed data without any prior decompression is possible;
  - **Same coding method can be used for regression**.

- Ongoing work :
  - Semantic communication with decoder's side information;
  - Classification as an example.

Current work : Classification

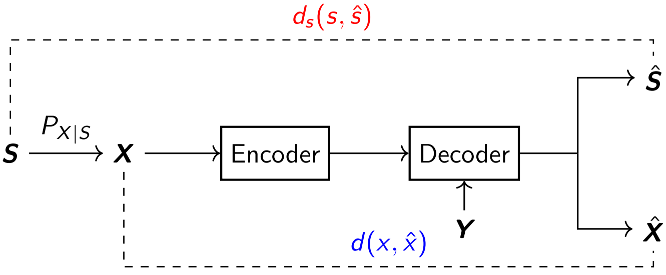


Figure 9: Coding scheme for semantic WZ coding

We consider the following problem

$$R(D, D_s) = \inf_{\substack{p(u \mid x):\\ \mathbb{E}\left[d(X,\hat{X})\right] \leq D \\ \mathbb{E}\left[d'(X,\hat{S})\right] \leq D_s}} I(X;U|Y). \tag{33}$$

with $d'(x,\hat{s}) = \frac{1}{p(x)} \sum_{s\in\mathcal{S}} p(x,s) d_s(s,\hat{s})$, $d_s(\boldsymbol{s^n},\boldsymbol{\hat{s}^n}) = \frac{1}{n}\sum_{i=1}^{n} d_s(s_i,\hat{s}_i)$ and $d(\boldsymbol{x^n},\boldsymbol{\hat{x}^n}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i,\hat{x}_i)$.

Thank You!

J. Wei, E. Dupraz, and P. Mary, "Distributed source coding for parametric and non-parametric regression," arXiv preprint arXiv:2404.18688, 2024