CoLearn

IMT Atlantique
Elsa Dupraz
Ahcen Aliouat
François-Xavier Socheleau

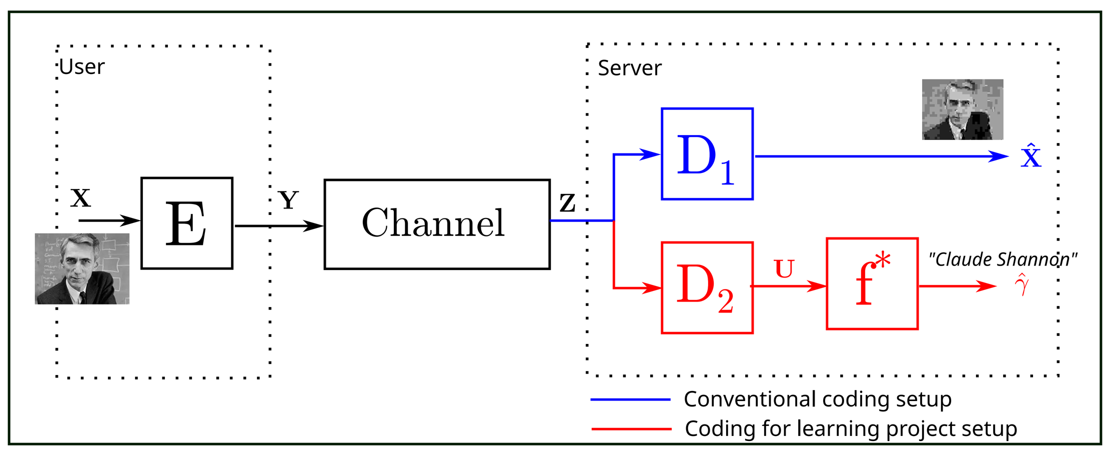
INSA Rennes
Philippe Mary
Jiahui Wei

INRIA Rennes
Aline Roumy
Thomas Maugey
Rémi Piau

September 2021 - December 2025

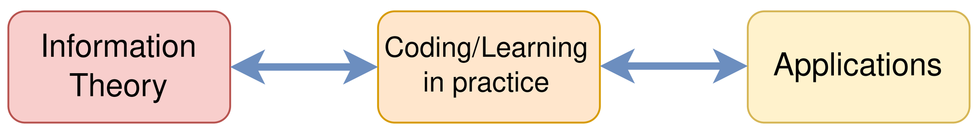
Project Objectives

Context: Huge mass of data (images, video, etc.) need to be sorted, processed, stored, recommended to users, etc.



Objective: **Learning and data reconstruction over coded data**

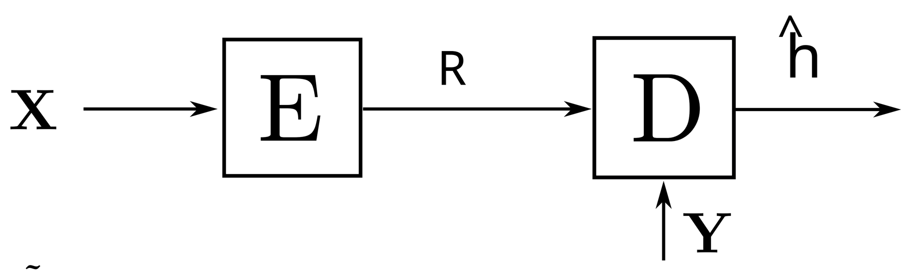
Key questions:
– Is there a tradeoff between the data reconstruction and learning objectives?
– Can one perform learning without prior decoding?
– Does the source-channel separation principle still hold?



Regression

Problem addressed:
• Few is known about Information-Theoretic limits of communication-for-learning schemes
• We consider **regression** as a first yet simple learning problem

$$X = \sum_{k=1}^{K} \alpha_k h_k(Y) + \epsilon$$



Training sequence $(\mathbf{X},\mathbf{Y})$, Test sequence $(\tilde{\mathbf{X}},\tilde{\mathbf{Y}})$

Expected Generalization error (GE): $G^{(k)}(\hat f) = E_{\mathbf{X},\mathbf{Y}}\left[E\left[(\tilde X - \hat f(\tilde Y))^2 | \mathbf{X},\mathbf{Y}\right]^k\right]$

Finite-length Rate-Distortion-GE region:
Distortion-Loss-Information Density Vector:

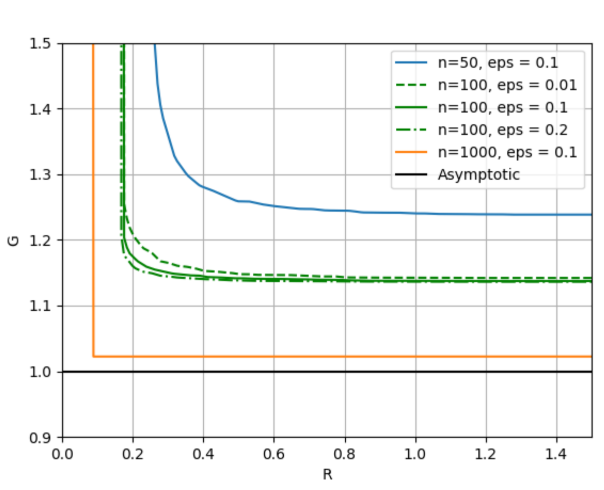
$$\mathbf{i} := \begin{bmatrix} -\log \dfrac{P_{U|Y}(U|Y)}{P_U(U)} \\ \log \dfrac{P_{U|X}(U|X)}{P_U(U)} \\ \ell(\tilde X, \hat f^{(n)}(\mathbf{Z},\tilde Y)) \\ d(X,\hat X) \end{bmatrix}$$

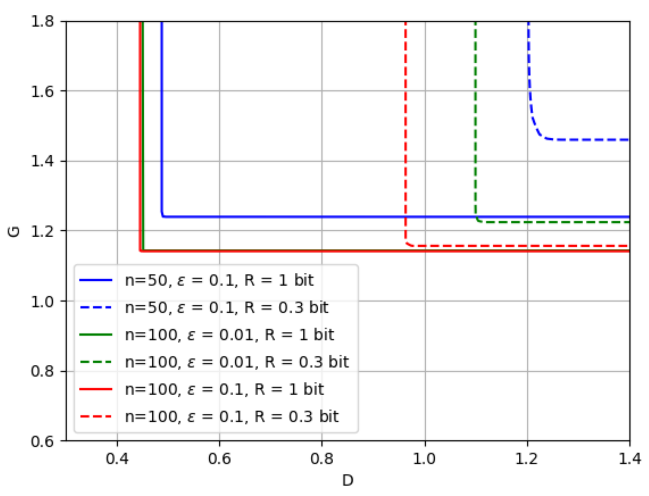
Region for a certain excess probability ε:

$$R(n,\varepsilon,g,d) \quad \inf\left\{\mathbf{M}\left(\mathbb{E}[\mathbf{i}] + \frac{\mathcal{S}(\mathbb{C}[\mathbf{i}],\varepsilon)}{\sqrt{n}} + \frac{2\log n}{n}\mathbf{I}_3\right)\right\}$$

Our bounds hold for both parametric and non-parametric regression
We showed that there is no tradeoff between data reconstruction and regression performance

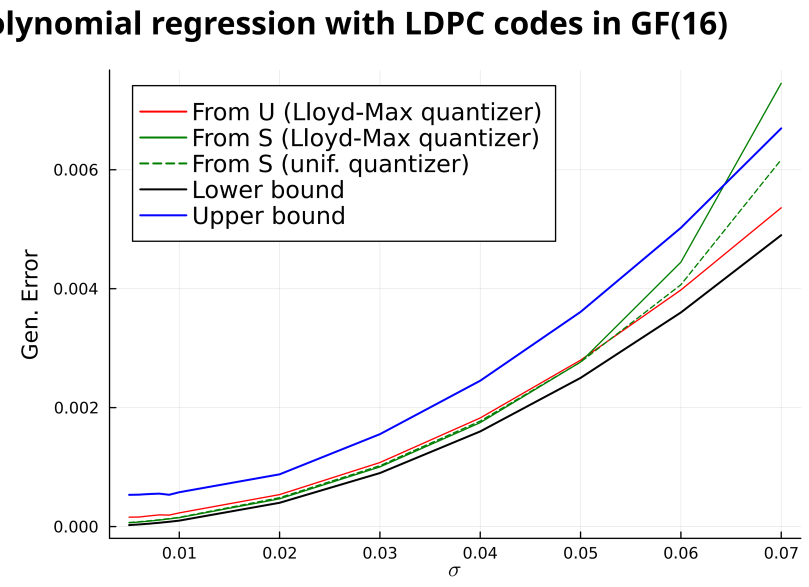
Numerical evaluation of the regions:





Practical coding scheme for regression:
- We proposed practical coding schemes for parametric regression
- After quantization, source vectors are encoded with LDPC codes as syndroms **s=Hx**
- We proposed a method to apply parametric regression over the syndrom, without need for prior LDPC decoding

Polynomial regression with LDPC codes in GF(16)



Logistic regression with LDPC codes in GF(4)

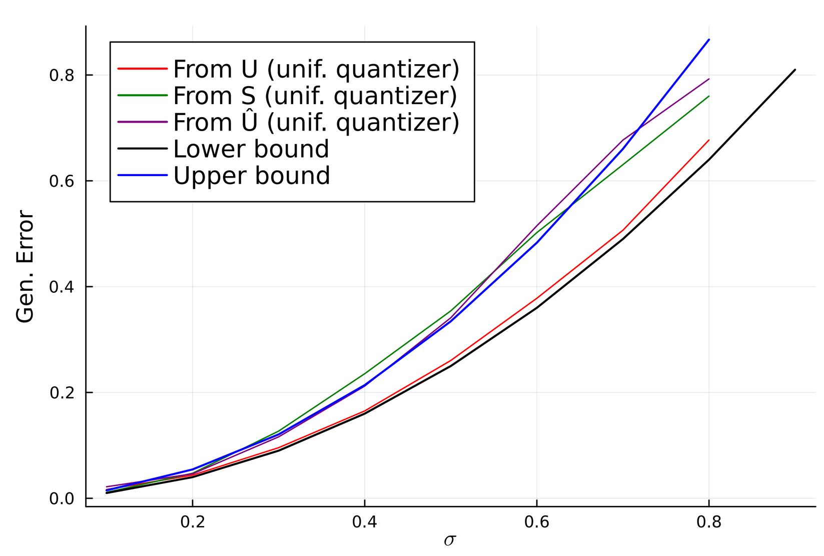
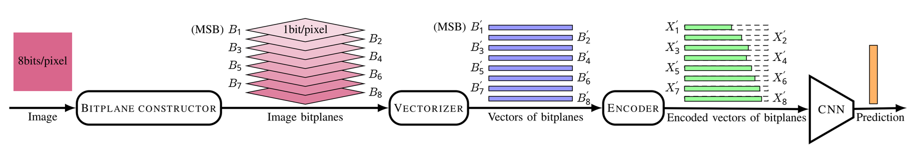


Image classification

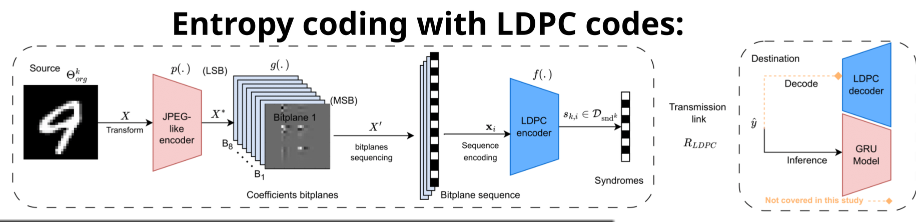
Problem addressed:
• Entropy-coding breaks the data structure
• Can we do **image classification** over compressed data without any prior decoding?

Considered setups:
• We first considered full compliance with JPEG standards, with **Huffman** as entropy coding technique
•We next proposed to modify the entropy coder for improved learning performance
• We consider entropy coding with **LDPC codes** (see CoMet for details)

Entropy coding with Huffman:

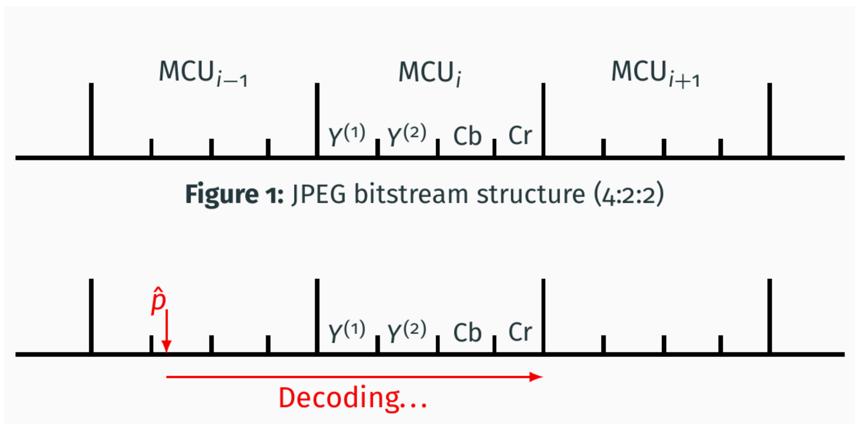


Entropy coding with LDPC codes:

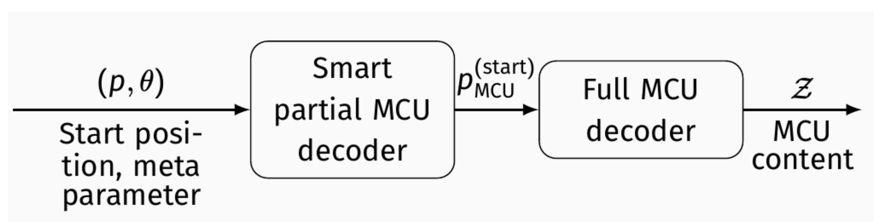


Dataset	Model	Without coding		On Original data (Setup1)		
		None	None MSB	Huff[1]	Arith[1]	LDPC
MNIST	GRU12(proposed)	0.9439	0.8842	-	-	0.8192
	GRU32(proposed)	0.9799	0.9154	-	-	0.8556
	UVGG11 [1]	0.9891	-	0.8323	0.6313	-
	URESNET18 [1]	0.9875	-	0.7450	0.5949	-
	FullyConn [2]	0.9200	-	-	-	-
Fashion -MNIST	GRU12	0.8616	0.8052	-	-	0.8166
	GRU32	0.8750	0.8314	-	-	0.8306
	UVGG11 [1]	0.9018	-	0.7634	0.6898	-
	URESNET18 [1]	0.8497	-	0.6862	0.6116	-
YCIFAR -10	GRU12	0.3127	0.3249	-	-	0.4070
	GRU32	0.3596	0.3560	-	-	0.4171
	UVGG11 [1]	0.5657	-	0.3606	0.2976	-
	URESNET18 [1]	0.3836	-	0.2591	0.2432	-
	FullyConn [2]	0.3800	-	-	-	-

Random access to JPEG coefficients without decoding
• Problem: finding structure in coded bitstream is hard



Proposed algorithm: resynchronize using "decodability"



Learning accuracy subject to specific error types (Imagenette 10 classes):

Errors			
Sampling	Position	Decoding	Accuracy
✓			0.87269
✓	✓		0.7364
✓		✓	0.80137
✓	✓	✓	0.68516

Learning accuracy subject to specific error types (Imagenette 10 classes):

Errors			Accuracy	Accuracy	Δ
Sampling	Position	Decoding	without MLE	with MLE	Accuracy
✓		✓	75.911%	80.137%	+4.2260%
✓	✓	✓	63.497%	68.516%	+5.0190%

On-going works and Perspectives

-**Consider other applications:** classification of underwater acoustic signals over compressed data (very low-rate communication link)
- Develop **universal** practical coding schemes for learning over compressed data, that can tolerate different learning tasks over the same coded data
- Consider **more complex learning tasks** such as image retrieval over compressed data
- Investigate **information-theoretic limits** of classification over coded data, and unsupervised learning over coded data
- Study the effect of the **channel** onto the learning performance

Publications:

•Jiahui Wei, Elsa Dupraz, Philippe Mary, Practical Coding Schemes based on LDPC Codes for Distributed Parametric Regression, accepted at the Information Theory Workshop (ITW) 2024
•Ahcen Aliouat, Elsa Dupraz, Learning on JPEG-LDPC Compressed Images: Classifying with Syndromes, accepted at EUSIPCO 2024
•Jiahui Wei, Philippe Mary, Elsa Dupraz, Rate-Loss Regions for Polynomial Regression with Side Information, accepted at the International Zurich Seminar on Information and Communication (IZS) 2024
•Rémi Piau, Thomas Maugey, Aline Roumy, Predicting CNN learning accuracy using chaos measurement, ICASSP 2023
• Jiahui Wei, Elsa Dupraz, Philippe Mary, Régions atteignables pour la régression linéaire sur données compressées avec information adjacente, GRETSI 2023
• Rémi Piau, Thomas Maugey, Aline Roumy, Prédiction de la précision d'apprentissage des réseaux de neurones convolutifs par mesure du chaos, GRETSI 2023
• Jiahui Wei, Elsa Dupraz, Philippe Mary, Asymptotic and non-asymptotic rate-loss bounds for linear regression with side information, EUSIPCO 2023
• Remi Piau, Aline Roumy, and Thomas Maugey, Learning on entropy coded images with CNN, ICASSP 2023
• Alireza Tasdighi and Elsa Dupraz, An End-to-End Scheme for Learning over Compressed Data Transmitted Through a Noisy Channel, IEEE Access, 2023