Journal of Machine Learning Research 25 (2024) 1-71 Submitted 12/21; Revised 4/24; Published 9/24

Studying the Interplay between Information Loss and Operation Loss in Representations for Classification

Jorge F. Silva JORGESIL.EDU@GMAIL.COM
Information and Decision System Group
Universidad de Chile
Santiago, 8370448, Chile

Felipe Tobar F.TOBAR@IMPERIAL.AC.UK
Department of Mathematics & I-X
Imperial College London
London, W12 0BZ, UK

Mario Vicuña MARIO.VICUNA@ING.UCHILE.CL
Information and Decision System Group
Universidad de Chile
Santiago, 8370448, Chile

Felipe Cordova FELIPECORDOVA@UG.UCHILE.CL
Information and Decision System Group
Universidad de Chile
Santiago, 8370448, Chile

Editor: Aurelien Garivier

Abstract

Information-theoretic measures have been widely adopted for machine learning (ML) feature design. Inspired by this, we look at the relationship between information loss in the Shannon sense and the operation loss in the minimum probability of error (MPE) sense when considering a family of lossy representations. Our first result offers a lower bound on a weak form of information loss as a function of its respective operation loss when adopting a discrete encoder. When considering a general family of lossy continuous representations, we show that a form of vanishing information loss (a weak informational sufficiency (WIS)) implies a vanishing MPE loss. Our findings support the observation that selecting/designing representations that capture informational sufficiency is appropriate for learning. However, this selection is rather conservative if the intended goal is achieving MPE in classification. Supporting this, we show that it is possible to adopt an alternative notion of informational sufficiency (strictly weaker than pure sufficiency in the mutual information sense) to achieve operational sufficiency in learning. Furthermore, our new WIS condition is used to demonstrate the expressive power of digital encoders and the capacity of two existing compression-based algorithms to achieve lossless prediction in ML.

Keywords: Representation learning, feature design, sufficiency, invariant models, digitalization, data-driven partitions, info-max learning, information bottleneck (IB), lossy compression for lossless prediction, mutual information estimation

1. Introduction

Given a continuous random object X, the problem of representation learning formalizes the task of finding lossy descriptions (or features) of X, denoted by U, that are sufficient (in some sense) to discriminate a target discrete variable of interest Y (e.g., a class or concept). In numerous contexts, the raw observation X lives in a finite-dimensional continuous space $\mathbb{R}^d$. In this mixed *continuous-discrete setting*, a reasonable assumption is that the raw X is redundant, i.e., many explanatory factors interact in the expression of X beyond Y and, consequently, a lossy description (aka coding) U has the potential to capture almost all, or ideally all, the information that X offers to discriminate Y (Bengio et al., 2013). Supporting this idea, it has been shown that under some structural conditions (Bloem-Reddy and Teh, 2020; Dubois et al., 2021), there is a lossy description $U = g(X)$ that is information sufficient in the sense that $I(X;Y) = I(U;Y)$, where $I(X;Y)$ denotes the mutual information (MI) between X and Y (Cover and Thomas, 2006). From the data-processing inequality (Cover and Thomas, 2006; Gray, 1990b), informational sufficiency implies that $I(X;Y|U) = 0$, meaning that X and Y are conditionally independent given U. A relevant context where this strong latent structure arises is problems with probabilistic symmetries or invariances with respect to (w.r.t.) a group of transformations (Bloem-Reddy and Teh, 2020; Dubois et al., 2021).

In practice, lossy descriptions have been instrumental in learning problems because they regularize the hypothesis space by reducing the complexity/dimensionality of the features, thus providing better generalization from training to unseen testing conditions, which is arguably the cornerstone of the learning problem (Xu and Mannor, 2012; Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Devroye et al., 1996; Bousquet et al., 2004). There is a large body of work that addresses the design of lossy representations from data. Many of these approaches rely on the use of information-theoretic measures to quantify the predictive relationship between X and Y, using, for instance, MI $I(X;Y)$, or conditional entropy $H(Y|X)$ or other approaches (Achille and Soatto, 2018a; Amjad and Geiger, 2019; Alemi et al., 2017; Achille and Soatto, 2018b). Along the same lines of learning a minimal (or compressed) sufficient representation from X, the Information Bottleneck (IB) method has been adopted in learning and decision (Amjad and Geiger, 2019; Alemi et al., 2017; Achille and Soatto, 2018b; Tishby et al., 1999) to optimize a tradeoff between relevance $I(U;Y)$ and compression $I(U;X)$ over a collection of probabilistic mappings from X to a (latent) variable U (Zaidi et al., 2020). There is also a deterministic version of the IB method where the objective is to find the optimal tradeoff between $I(Y,U)$ and $H(U)$ where U is generated through a family of finite alphabet mappings (or vector quantizations) of X (Tishby et al., 1999; Strouse and Schwab, 2017; Tegmark and Wu, 2019).

In the context of learning representation as outlined above, the concept of asymptotic *sufficiency* can be introduced: an infinite collection of lossy descriptions $U_1, U_2, ...$ of X is said to be *information sufficient* (IS) if $\lim_{i\to\infty} I(U_i;Y) = I(X;Y)$. In contrast, a collection $U_1, U_2, ...$ is said to be *operationally sufficient* (OS) if the performance of classifying Y from U_i, in the minimum probability of error (MPE) sense, achieves—as i tends to infinity—the performance of the optimal MPE classifier that uses X losslessly to predict Y. Then, a natural question is: If a method designs a collection of IS descriptions, is this collection also OS? More generally, is there a strictly weaker notion of IS that implies OS?

To address both questions and, in particular, whether there is a strictly weak notion of IS that implies OS, in this paper, we focus on studying the interplay between a weak form of information loss and the operation loss over a family of problems (models) induced by lossy continuous representations of X. In particular, we consider a model (X, Y) with joint distribution $\mu_{X,Y}$ and a family of lossy representations (encoders) $\{U_i\}_{i\geq 1}$ of X, where $U_i = g_i(X)$ is a continuous mapping and $\mu_{U_i,Y}$ denotes the joint distribution of (U_i, Y). In this context, we look at the following weak form of information loss[1] $I((\tilde{r}(X), U_i); Y) - I(U_i; Y) \geq 0$ where $\tilde{r}(X)$ denotes the *MPE decision rule*[2].

1.1 Contributions

For the case of discrete representations (i.e., U_i is induced by a vector quantizer (VQ)), Theorem 10 presents a lower bound for $I((\tilde{r}(X), U_i); Y) - I(U_i; Y) \geq 0$ as a function of its respective operation loss (OL) $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) \geq 0$ where $\ell(\mu_{U_i,Y})$ and $\ell(\mu_{X,Y})$ are the MPE associated to the model $\mu_{U_i,Y}$ and $\mu_{X,Y}$, respectively. OL is the loss attributed to using U_i instead of X in classifying Y. Using this bound, our first main result (Theorem 12) shows that if $\{U_i\}_{i\geq 1}$ is weakly information sufficient (WIS), in the sense that $\lim_{i\to\infty} [I((\tilde{r}(X), U_i); Y) - I(U_i; Y)] = 0$ then $\{U_i\}_{i\geq 1}$ is operationally sufficient (OS) to discriminate Y in the sense that $\lim_{i\to\infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y})$. In other words, a form of sufficiency that is strictly weaker than IS implies a vanishing operation loss when $\{U_i\}_{i\geq 1}$ is a family of general continuous representations of X.

On the technical contribution, we obtain Theorem 12 using the bound in Theorem 10. In particular, the argument to prove this result goes from discrete (or VQ) to continuous representations. We demonstrate first the scenario of discrete representations in Theorem 35, Section 3.2, to prove then Theorem 12 in the general continuous case in Section 10.2.3. The proofs of Theorems 12 and 10 rely on two important information-theoretic results: The first by Ho and Verdú (2010) that characterizes, using a specific rate-distortion function, a tight upper bound for the conditional entropy given an error probability and the second from Liese et al. (2006) on asymptotic sufficient partitions for mutual information. Regarding the optimality of the WIS condition for $\{U_i\}_{i\geq 1}$ in Theorem 12, we show in Theorem 14 that when the MPE rule is unique almost surely w.r.t. to the model $\mu_{X,Y}$, then "*WIS is equivalent to OS*" for a general class of continuous representations of X. This result offers a context where WIS is tight and optimal, in the sense that no weaker expressivity condition on $\{U_i\}_{i\geq 1}$ could be found to guarantee OS.

In the second part of this paper, we work on applications and extensions of Theorem 12 in machine learning (ML). Applying Theorem 12 for the task of evaluating representations or encoders of X, we discover that discrete encoders (in the form of a VQ) have the expressive power to be operationally sufficient (OS) for any model $\mu_{X,Y}$ in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. We present two versions of this expressiveness result: Theorem 15 for non-adaptive partitions and Theorem 18 for adaptive (data-driven) partitions. These results are used to demonstrate that three specific schemes are OS distribution-free in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$: a uniform partition of the space (Liese et al., 2006), a data-driven statistically equivalent partition (Devroye et al., 1996; Gessaman, 1970) and a data-driven tree-structured partition (Devroye et al., 1996;

1. This weak information loss (WIL) is formally introduced in Definition 3 - Section 2.4.
2. $\tilde{r}(\cdot)$ is formally introduced in Eq.(10) - Section 2.4.

Silva and Narayanan, 2010a). These results, obtained from Theorem 12, shed light on two interesting aspects of representation learning: the expressive power that can be achieved using partitions (a digitalization of the problem) and the capacity of non-supervised data-driven methods to achieve informational and operational sufficiency.

Studying how Theorem 12 (WIS $\Rightarrow$ OS) can be extended in an ML setting, we introduce a setting where the true (unknown) model belongs to class $\Lambda \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ or collection of learning tasks. In this context, we introduce a measure-theoretic characterization for the structure of Λ. That characterization is used to determine a lossy mapping $\eta_\Lambda(\cdot)$ that is operationally sufficient (OS) for Λ, in the sense that $\ell(\mu_{X,Y}) = \ell(\mu_{\eta_\Lambda(X),Y})$ for any $\mu_{X,Y} \in \Lambda$. This property facilitates the following non-oracle extension of Theorem 12: if $I(\eta_\Lambda(X); Y|U_i) \to 0$ (non-oracle WIS) then $\{U_i\}_{i\geq 1}$ is OS. Importantly, this new result (stated in Theorem 24) replaces $\tilde{r}(X)$, which is an oracle object, by $\eta_\Lambda(X)$ that is non-oracle and used as a prior knowledge of the learning setting (inductive bias). On the application of Theorem 24, we highlight the class of invariant models to the action of a compact group[3]. For this class, there is a well-established lossy mapping $\eta_\Lambda(\cdot)$, which is maximal invariant w.r.t. a group of transformations on $\mathcal{X}$, that precisely meets the mentioned operational requirement, i.e., $\ell(\mu_{X,Y}) = \ell(\mu_{\eta_\Lambda(X),Y})$ for any $\mu_{X,Y} \in \Lambda$. More generally, we study many classes of models and examples where Theorem 24 can be applied meaningfully. In some of these contexts, we demonstrate that our new condition $I(\eta_\Lambda(X); Y|U_i) \to 0$ is strictly weaker than IS. Finally, Theorem 24 (non-oracle WIS $\Rightarrow$ OS) is used to explain the expressive power of two existing compression-based learning algorithms: two variations of the information bottleneck (IB) method (Tishby et al., 1999; Tishby and Zaslavsky, 2015; Chechik et al., 2005) and a version of the lossy compression for lossless prediction (LCLP) method by Dubois et al. (2021).

1.2 Related Work

Our analyses relate fundamentally to the interplay between (minimum) probability of error and conditional entropy (or equivocation entropy) that has been studied systematically in information theory (Feder and Merhav, 1994; Ho and Verdú, 2010; Prasad, 2015). One of the most recognized results in this area is *Fano's inequality*[4] that offers a lower bound for the probability of error as a function of the entropy (Cover and Thomas, 2006). A refined analysis between conditional entropy and minimum error probability was presented by Feder and Merhav (1994). They explored the relationship between these quantities, providing tight (achievable) lower and upper bounds for the conditional entropy given a minimum error probability restriction. Refining this analysis, Ho and Verdú (2010) studied a more specific problem that is relevant in the Bayesian treatment of classification: given the prior distribution μ_Y of Y, they were interested in the interplay between the error probability of predicting Y from an observation X and the conditional entropy of Y given X when X is a discrete (finite-alphabet) observation. They provided a closed-form expression for the maximal conditional entropy that can be achieved as a function of the prior distribution μ_Y and the minimum probability error ϵ in the non-trivial regime when $\epsilon \leq (1 - \max_{y\in\mathcal{Y}} \mu_y(y))$.

3. An excellent exposition for this family is presented in (Bloem-Reddy and Teh, 2020).
4. $H(Y|X) \leq h(\ell(\mu_{X,Y})) + \ell(\mu_{X,Y}) \log(|\mathcal{Y}| - 1)$ where $h(r) = -r\log(r) - (1-r)\log(1-r)$ is the binary entropy (Cover and Thomas, 2006).

These results offer tight bounds between conditional entropy and error probability, thus providing refined and specialized versions of *Fano's type of bounds* (Ho and Verdú, 2010). These bounds were extended to countably infinite alphabets, a regime for which Fano's original inequality has not been defined. A relevant corollary of these bounds says that a vanishing probability of error implies a vanishing conditional entropy. The converse result is also true under some conditions (Feder and Merhav, 1994). Then, when the classification task is almost perfect or degenerate (zero probability of error), the relationship between error probability and conditional entropy is rather evident (zero error $\Leftrightarrow$ zero conditional entropy). However, this connection is less evident for most cases that deviate from this highly discriminative context as it is clearly presented in (Feder and Merhav, 1994; Ho and Verdú, 2010).

The focus of our work in this paper is different from the results mentioned in this subsection as we are interested in the interplay between a form of information loss and its respective operation loss over a family of problems induced by lossy continuous representations (encoders) of X.

1.3 Organization

The rest of the paper is organized as follows. Sections 2 formalizes our main question and introduces notations, required concepts, and preliminary results. Section 3 presents the statement and interpretations of the main asymptotic result (Theorem 12). Section 3 also covers a result for a finite alphabet representation of X (Theorem 10), which is essential to extend the analysis to continuous representations. Section 4 discusses about the optimality of the WIS condition (Theorem 14). The application of Theorem 12 to demonstrate the expressive power of digital representations is presented in Section 5. Section 6 extends the concept of weak informational sufficiency into a learning setting (Theorem 24). The expressive analysis of two existing compression-based representation learning algorithms is presented in Section 7. Concluding remarks and discussions are presented in Section 9. The proofs of the main results of this work are presented in Section 10. Finally, complementing numerical examples are presented in Section 8, and the supporting technical material of this work is organized in the Appendix.

2. Preliminaries

Let us consider a decision problem expressed in terms of the joint probability[5] $\mu_{X,Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ of a random vector (r.v.) (X,Y) where Y takes values in a finite set $\mathcal{Y}=\{1,..,M\}$ (e.g., a *class label*) and X takes values in a finite dimensional space $\mathcal{X}=\mathbb{R}^d$. On the operational side, the *minimum probability of error* (MPE) of predicting Y using X as an observation, given the model $\mu_{X,Y}$, is expressed by the following task

$$\begin{aligned}\ell(\mu_{X,Y}) &\equiv \min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{P}(f(X)\neq Y)\\ &= \int_{\mathcal{X}} (1-\max_{y\in\mathcal{Y}} \mu_{Y|X}(y|x))\partial\mu_X(x), \end{aligned} \tag{1}$$

5. $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ denotes the collection of probabilities in the measurable space $(\mathcal{X}\times\mathcal{Y}, \mathcal{B}(\mathcal{X}\times\mathcal{Y}))$ where $\mathcal{B}(\mathcal{X}\times\mathcal{Y})$ denotes the product sigma field of $\mathbb{R}^d\times\{1,..,M\}$ (Breiman, 1968).

where $\mu_{Y|X}(\cdot|x)$ denotes the conditional *probability mass function* (pmf) of Y given the event $\{X = x\}$ and μ_X denotes the marginal probability of X in $\mathcal{X}$. On the information side, the conditional entropy of Y given X — also known as the equivocation entropy (EE) in this decision context (Feder and Merhav, 1994; Ho and Verdú, 2010) — is

$$H(Y|X) \equiv \int_{\mathcal{X}} \mathcal{H}(\mu_{Y|X}(\cdot|x))\partial\mu_X(x), \tag{2}$$

where

$$\mathcal{H}(\mu_{Y|X}(\cdot|x)) \equiv -\sum_{y\in\mathcal{Y}} \mu_{Y|X}(y|x)\log\mu_{Y|X}(y|x) \in [0, \log M]$$

is *the Shannon entropy* of $\mu_{Y|X}(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ (Gray, 1990b; Cover and Thomas, 2006). The mutual information (MI) of $\mu_{X,Y}$ is (Gray, 1990b; Cover and Thomas, 2006)

$$\mathcal{I}(\mu_{X,Y}) = I(X;Y) \equiv \mathcal{H}(\mu_Y) - H(Y|X) \geq 0. \tag{3}$$

The standard notation for MI is $I(X;Y)$, however we also use $\mathcal{I}(\mu_{X,Y})$ to emphasize in our analysis that MI is a functional of the joint distribution $\mu_{X,Y}$.

2.1 Representations of X

A representation of X is a measurable function $\eta : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathcal{U}, \mathcal{B}(\mathcal{U}))$ where $\mathcal{U}$ is the representation space with its respective sigma field denoted by $\mathcal{B}(\mathcal{U})$. In general, we are interested in the case of a lossy function $\eta(\cdot)$ where $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ is a continuous measurable space. However, attention will be given to the relevant case where $|\mathcal{U}| = K < \infty$, meaning that $\eta(\cdot)$ is a vector quantizer (VQ) of X.[6] This VQ induces the following finite partition of size K:

$$\pi_\eta \equiv \left\{\eta^{-1}(\{u\}), u \in \mathcal{U}\right\}, \tag{4}$$

where conversely we have that $\eta(x) = \sum_{u\in\mathcal{U}} u \cdot \mathbf{1}_{\eta^{-1}(\{u\})}(x)$.

We denote by $U \equiv \eta(X)$ the representation of X induced by $\eta(\cdot)$, and we denote by $\mu_{U,Y}$ the joint distribution of (U, Y) (induced by $\mu_{X,Y}$ and $\eta(\cdot)$) in $\mathcal{U} \times \mathcal{Y}$. As the expressions in (1) and (3) are functions of the model $\mu_{X,Y}$, they can be extended to $\mu_{U,Y}$, where i) $\ell(\mu_{U,Y})$ is the MPE of predicting Y from U, and ii) $\mathcal{I}(\mu_{U,Y})$ is the MI between U and Y.

2.2 Information Loss and Operation Loss

We are interested in the *information loss* (IL) of using U (instead of X) to resolve Y in the Shannon sense. This can be measured by[7]

$$\mathcal{I}(\mu_{X,Y}) - \mathcal{I}(\mu_{U,Y}) = I(X;Y|U), \tag{5}$$

where the conditional MI (CMI) of X and Y given U is

$$I(X;Y|U) \equiv \int_{\mathcal{U}} \mathcal{I}(\mu_{X,Y|U}(\cdot|u))\partial\mu_U(u) \geq 0. \tag{6}$$

6. The main result of this work is for continuous measurable transformations. However, studying the case of finite VQs is instrumental as elaborated in Sections 3.1 and Appendix 10.2.
7. The equality in (5) comes from the chain rule of MI and the definition of the conditional MI (Gray, 1990b; Cover and Thomas, 2006).

The main objective of this work is to understand how an information loss of the form in (5) relates to its respective *operation loss* (OL) of using U (instead of X) to classify Y in the MPE sense, i.e.,

$$\ell(\mu_{U,Y}) - \ell(\mu_{X,Y}) \geq 0. \tag{7}$$

2.3 Sufficiency

Let us consider a family of measurable functions (or encoders) $\eta_i : \mathcal{X} \to \mathcal{U}_i$, indexed by $i \in \mathbb{N}$, where $\mathcal{U}_i$ is a continuous space, for example a finite dimensional Euclidean space $\mathbb{R}^q$. Using $\eta_i(\cdot)$, we consider the representation variable $U_i = \eta_i(X)$ (e.g., a *feature*) and the respective joint distribution of (U_i, Y) characterized by $\mu_{U_i,Y}$ in $\mathcal{U}_i \times \mathcal{Y}$. At this point, we introduce the following definitions for informational and operational sufficiency, respectively.

Definition 1 *A sequence of representations $\{\eta_i(\cdot)\}_{i\geq 1}$ for X (and its respective representation variables $\{U_i\}_{i\geq 1}$) is said to be operationally sufficient (OS) for the model $\mu_{X,Y}$ (in the MPE sense) if*

$$\lim_{i\longrightarrow\infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y}). \tag{8}$$

Definition 2 *A sequence of representations $\{\eta_i(\cdot)\}_{i\geq 1}$ for X (and $\{U_i\}_{i\geq 1}$, respectively) is said to be information sufficient (IS) for $\mu_{X,Y}$ if*

$$\lim_{i\longrightarrow\infty} \mathcal{I}(\mu_{U_i,Y}) = \mathcal{I}(\mu_{X,Y}). \tag{9}$$

Let us introduce a weak version of IS for $\{\eta_i(\cdot)\}_{i\geq 1}$ w.r.t. to $\mu_{X,Y}$. For this, let us recall that the MPE rule (a sufficient statistic) is a quantizer of $\mathcal{X}$ of size $M = |\mathcal{Y}|$ given by[8]

$$\tilde{r}_{\mu_{X,Y}}(x) \equiv \arg\max_{y\in\mathcal{Y}} \mu_{Y|X}(y|x), \tag{10}$$

where $\ell(\mu_{X,Y}) = \mathbb{P}(\tilde{r}_{\mu_{X,Y}}(x) \neq Y)$. This rule induces a (distribution dependent) partition of $\mathcal{X}$ given by

$$\tilde{\pi} \equiv \left\{ \tilde{A}_y \equiv \tilde{r}^{-1}_{\mu_{X,Y}}(\{y\}), y \in \mathcal{Y} \right\}, \tag{11}$$

and a finite-alphabet (VQ) lossy representation of X given by $\tilde{U} \equiv \tilde{r}_{\mu_{X,Y}}(X) \in \mathcal{Y}$.

Definition 3 *A sequence of representations $\{\eta_i(\cdot)\}_{i\geq 1}$ for X (and $\{U_i\}_{i\geq 1}$) is said to be weakly information sufficient (WIS) for $\mu_{X,Y}$ if*

$$\lim_{i\longrightarrow\infty} \underbrace{\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y})}_{I(Y;\tilde{U}|U_i)\geq 0} = 0, \tag{12}$$

where $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X)$ and $I(Y;\tilde{U}|U_i)$ is the conditional MI between Y and $\tilde{U}$ given U_i (Cover and Thomas, 2006).

This weak IS definition (WIS) is introduced from the observation that $\tilde{U}$ is a sufficient statistic for the inference task in the operational MPE sense, see Eq.(10). Finally in (12), we have the following weak information loss $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = I(\tilde{U},U_i;Y) - I(U_i;Y) = I(\tilde{U};Y|U_i)$.

8. The optimal rule in (10) is not unique in general. If for some x several $y \in \mathcal{Y}$ achieve the minimum in (10), we select the smallest one to define $\tilde{r}_{\mu_{X,Y}}(x)$ in (10).

2.4 A Basic Analysis of the Losses Induced by a VQ

Let us consider the discrete and finite case where $\mathcal{U}_i = \{1, \dots, k_i\}$ for any $i \geq 1$. In this context, there is a finite measurable partition induced by $\eta_i(\cdot)$:

$$\pi_{\eta_i} \equiv \left\{ B_{i,j} \equiv \eta_i^{-1}(\{j\}), j \in \mathcal{U}_i = \{1, \dots, k_i\} \right\}, \tag{13}$$

where $\eta_i(x) = \sum_{j \in \mathcal{U}_i} \mathbf{1}_{B_{i,j}}(x) \cdot j$. In this VQ context, the following results present useful expressions for $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y})$ and $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y})$.

Proposition 4 *$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{B_{i,j} \in \pi_{\eta_i}} \mu_X(B_{i,j}) \cdot g(\mu_{X,Y}, B_{i,j})$, where*

$$g(\mu_{X,Y}, B_{i,j}) \equiv \left[1 - \max_{y \in \mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - \sum_{\tilde{A}_u \in \tilde{\pi}} \frac{\mu_X(\tilde{A}_u \cap B_{i,j})}{\mu_X(B_{i,j})} \left[1 - \max_{y \in \mathcal{Y}} \mu_{Y|X}(y|\tilde{A}_u \cap B_{i,j})\right]. \tag{14}$$

On the information loss side, instead of looking at $I(X;Y|U_i)$ in (4), we look at the MI loss of observing U_i with respect to a re-defined reference case $(\tilde{U}, U_i)$ with $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X)$ introduced in (12).

Proposition 5 *$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = \sum_{B_{i,j} \in \pi_i} \mu_X(B_{i,j}) \cdot I(\tilde{U}; Y|X \in B_{i,j})$, where*

$$I(\tilde{U}; Y|X \in B_{i,j}) \equiv \mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \sum_{\tilde{A}_u \in \tilde{\pi}} \frac{\mu_X(\tilde{A}_u \cap B_{i,j})}{\mu_X(B_{i,j})} \mathcal{H}(\mu_{Y|X}(|\tilde{A}_u \cap B_{i,j})) \tag{15}$$

is the MI between Y and $\tilde{U} = \sum_{u \in \mathcal{Y}} u \cdot \mathbf{1}_{\tilde{A}_u}(X)$ conditioning on the event $\{X \in B_{i,j}\}$.

Remark 6 *The term $g(\mu_{X,Y}, B_{i,j}) \geq 0$ in Proposition 4 can be interpreted as the gain in MPE from a "*prior scenario*" where the marginal distribution of Y follows $(\mu_{Y|X}(y|B_{i,j}))_{y \in \mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$ to a "*posterior scenario*" where we observe $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X)$ to classify Y under the joint conditional model $\left(\mu_{\tilde{U},Y|X}(u, y|B_{i,j}) \equiv \frac{\mu_{X,Y}(\tilde{A}_u \cap B_{i,j} \times \{y\})}{\mu_X(B_{i,j})}\right)_{(u,y) \in \mathcal{Y}^2}$ in $\mathcal{P}(\mathcal{Y} \times \mathcal{Y})$.*[9]

Remark 7 *There is a parallel connection between $g(\mu_{X,Y}, B_{i,j})$ in (14), prior minus posterior MPE loss condition on $\{X \in B_{i,j}\}$, and $I(\tilde{U}; Y|X \in B_{i,j})$ in (15), which is the prior minus the posterior Shannon entropy condition on $\{X \in B_{i,j}\}$.*

2.5 An Information-Theoretic Lower Bound

Finally, to establish a relationship between the two losses presented in Section 2.4, we consider the following result by Ho and Verdú (2010).

Lemma 8 *(Ho and Verdú, 2010, Th.4) Let us consider Y a random variable in $\mathcal{Y} = \{1, \dots, M\}$ and a finite observation space $\mathcal{X}$ such that $|\mathcal{X}| \geq M$. If we denote by $\mathcal{P}(\mathcal{X}|\mathcal{Y})$*

9. This Bayesian gain interpretation of the term $g(\mu_{X,Y}, B_{i,j})$ will be central for the results in Section 3.

the collection of conditional probabilities from $\mathcal{Y}$ to $\mathcal{X}$, then for any non-negative $\epsilon \leq \underbrace{(1-\max_{y\in\mathcal{Y}} \mu_Y(y))}_{\text{the prior error of } \mu_Y}$, it follows that[10]

$$f(\mu_Y, \epsilon) \equiv \min_{\rho_{X|Y}\in\mathcal{P}(\mathcal{X}|\mathcal{Y}) st. \ell(\rho_{X|Y}\cdot\mu_Y)=\epsilon} \mathcal{I}(\rho_{X|Y}\cdot\mu_Y) = \mathcal{H}(\mu_Y) - \mathcal{H}(\mathcal{R}(\mu_Y,\epsilon)) \geq 0, \tag{16}$$

where $\rho_{X|Y}\cdot\mu_Y$ is a joint probability in $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ and $\mathcal{R}(\mu_Y,\epsilon)\in\mathcal{P}(\mathcal{Y})$ is a well-defined probability function of μ_Y and ϵ.[11]

This Lemma offers a tight (achievable) lower bound on the minimum MI achieved by a family of joint discrete distributions in $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ that satisfy two conditions: i) they meet an MPE restriction parametrized by $\epsilon\in[0, 1-\max_{y\in\mathcal{Y}}\mu_Y(y)]$ and ii) they have a marginal distribution on Y given by $\mu_Y\in\mathcal{P}(\mathcal{Y})$.[12]

Remark 9 *Considering a discrete observation X such that $(X,Y)\sim\mu_{X,Y}$, Lemma 8 can be used directly to obtain a lower bound for $I(X;Y)=\mathcal{I}(\mu_{X,Y})$ as a function of $\ell(\mu_{X,Y})$. More precisely, from (16) we have that*

$$I(X;Y)=\mathcal{I}(\mu_{X,Y}) \geq f(\mu_Y, \ell(\mu_{X,Y})) = H(Y) - \mathcal{H}(\mathcal{R}(\mu_Y, \ell(\mu_{X,Y}))). \tag{17}$$

Importantly, the bound in (17) recovers the known fact that if $\ell(\mu_{X;Y}) < (1-\max_{y\in\mathcal{Y}}\mu_Y(y))$ then $I(X;Y)>0$, or, conversely, $I(X;Y)=0$ (zero information) implies $\ell(\mu_{X,Y})=(1-\max_{y\in\mathcal{Y}}\mu_Y(y))$, i.e., there is a zero gain in MPE when observing X.

In the following section, we present two results that express the interplay between the weak information loss (WIL) $I(Y;\tilde{U}|U_i)$ in (12) and the operation loss (OL) introduced in (7) for which Lemma 8 and Propositions 4 and 5 will be instrumental.

3. Interplay between Information Loss and Operation Loss

This section presents the main asymptotic result of this work: WIS implies OS (Theorem 12). This theorem is constructed from a series of results that analyze the interplay between WIL and OL from discrete (VQs) to continuous representations.

3.1 The Non-Asymptotic Bound for Vector Quantizers

To derive Theorem 12, it is essential to have a lower bound on the WIL (in Proposition 5) as a function of its respective OL (in Proposition 4). This result for VQ is the following:

10. The closed-form expression of the probability $\mathcal{R}(\mu_Y,\epsilon)\in\mathcal{P}(\mathcal{Y})$ is presented in (Ho and Verdú, 2010) and in Appendix I for completeness.
11. In information theory, the function $f(\mu_Y,\epsilon)$ in (16) is a special case of the celebrated rate-distortion function of a memoryless source (i.i.d.) with marginal distribution μ_Y and distortion function given by the hamming distance (or the 0-1 loss) (Gray, 1990a).
12. For the non-trivial case when $\epsilon < (1-\max_{y\in\mathcal{Y}}\mu_Y(y))$, Ho and Verdú (2010) showed that $\mathcal{H}(\mu_Y) > \mathcal{H}(\mathcal{R}(\mu_Y,\epsilon)) \Rightarrow f(\mu_Y,\epsilon)>0$, while for the trivial case when $\epsilon=(1-\max_{y\in\mathcal{Y}}\mu_Y(y))$ they showed that $\mathcal{R}(\mu_Y,\epsilon)=\mu_Y \Rightarrow f(\mu_Y,\epsilon)=0$ (Ho and Verdú, 2010).

Theorem 10 *Let us consider a model $\mu_{X,Y}$ and a finite alphabet representation U_i of X (induced by $\eta_i(\cdot)$), then*

$$\underbrace{\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y})}_{I(Y;\tilde{U}|U_i)} \geq \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot \left[\mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B_{i,j}), \epsilon_{i,j})\right] \geq 0, \tag{18}$$

where

$$\begin{aligned}\epsilon_{i,j} &\equiv \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - g(\mu_{X,Y}, B_{i,j}) \\ &= \sum_{\tilde{A}_u\in\tilde{\pi}} \frac{\mu_X(\tilde{A}_u \cap B_{i,j})}{\mu_X(B_{i,j})} \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|\tilde{A}_u \cap B_{i,j})\right]\end{aligned}$$

and $g(\mu_{X,Y}, B_{i,j})$ as in (14).

The proof is presented in Section 10.1.

Analysis of Theorem 10:

1. The lower bound on the WIL in (18) is an explicit function of the decomposition of the OL presented in Proposition 4.

2. On the proof of Theorem 10, the lower bound in (18) comes from writing the OL as (from Proposition 4)

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot g(\mu_{X,Y}, B_{i,j}), \tag{19}$$

   i.e., as the sum of some posterior minus prior MPE gains (see Remark 6) and the application of Lemma 8.

3. The inequality in (18) implies that

   Corollary 11 *If $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) > 0$ then $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) > 0$.*

   Therefore, a non-zero OL (on using U_i instead of X to discriminate Y) implies a respective non-zero WIL. Conversely, Corollary 11 states that if $U_i = \eta_i(X)$ is weakly IS in the sense that $I(\tilde{U}; Y|U_i) = 0$, then $\ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y})$, i.e., $\eta_i(\cdot)$ (and U_i) is OS for $\mu_{X,Y}$. The proof is presented in Appendix C.

4. It is worth noting that for a large class of models (continuos in nature), U_i being weakly IS, i.e., $I(\tilde{U}; Y|U_i) = 0$, is strictly weaker than asking that U_i is IS for $\mu_{X,Y}$ in the sense that $I(X; Y|U_i) = 0$. In fact, $I(X; Y|U_i) = 0$ implies that $I(\tilde{U}; Y|U_i) = 0$[13], but the converse result is not true in general.[14]

5. The difference between the information loss (IL), i.e., $I(X; Y|U_i)$, and the weak information loss (WIL), i.e., $I(\tilde{U}; Y|U_i)$, is further discussed in Section 3.4 and its non-zero discrepancy is illustrated by examples in Section 3.4.1 and Section 8.2.

13. $I(X; Y|U_i) = 0 \Rightarrow I(\tilde{U}; Y|U_i) = 0$ follows from the fact that $\tilde{U}$ is a deterministic function of X and the chain rule of the MI (Cover and Thomas, 2006).

14. Artificial construction for $\mu_{X,Y}$ can be made, which are discrete in nature, where $I(X; Y) = I(\tilde{U}; Y)$. In this discrete context, it is simple to verify that $I(X; Y|U_i) = I(\tilde{U}; Y|U_i)$ independent of U_i.

3.2 The Main Asymptotic Result

The following result shows that a family of weakly IS representations for $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is OS for $\mu_{X,Y}$. This result can be interpreted as a non-trivial extension of Corollary 11 (from Theorem 10) because it relaxes the zero WIL condition (considering instead a family of representations that achieves zero WIL asymptotically) and, more importantly, it is valid for any type of encoder (continuous and discrete).

Theorem 12 *Let $\{U_i\}_{i\geq 1}$ be a sequence of representations for X obtained from $\{\eta_i(\cdot)\}_{i\geq 1}$ following the setting introduced in Section 2.3. If $\{U_i\}_{i\geq 1}$ is WIS for $\mu_{X,Y}$ (Definition 3), then $\{U_i\}_{i\geq 1}$ is OS for $\mu_{X,Y}$ (Definition 1).*

Remarks on Theorem 12 and its interpretation:

1. WIS in (12) as a condition on $\{U_i\}_{i\geq 1}$ means that as i tends to infinity, U_i captures all the information (in the Shannon sense) that $\tilde{U}$ has to offer to resolve Y. As a corollary of this result, we obtain that pure IS (Definition 2) implies OS.

2. Importantly, for a large class of continuous models in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we have that $I(\tilde{U};Y) < I(X;Y)$ because $\tilde{U}$ is an M size quantized version of X (see Eq.11). Then, the WIS condition has the potential to be strictly weaker (for a large class of models and representations) than asking for pure IS (Definition 2). This important point is further analyzed in Section 3.4 and demonstrated with examples in Section 3.4.1 and Section 8.2.

3. Theorem 12 formalizes the intuition that achieving sufficiency in the Shannon MI sense is appropriate but a conservative criterion if the operational objective is classification, as a strictly weaker condition does exist (WIS) that guarantees OS. Indeed, if $\{U_i\}_{i\geq 1}$ is not OS for $\mu_{X,Y}$, Theorem 12 implies that $\liminf_{i\to\infty} \mathcal{I}(\mu_{U_i,Y}) < \mathcal{I}(\mu_{X,Y})$, meaning that the representations $\{U_i\}_{i\geq 1}$ do not achieve (eventually in i) the MI of the lossless variable X.

To conclude this part, it is worth mentioning that WIS, as a condition on $\{\eta_i(\cdot)\}_{i\geq 1}$, is theoretically relevant for the reasons mentioned in the previous points, but it is not appropriate as a criterion for feature design/selection in a learning setting. The reason is that the reference representation $\tilde{U}$ (used in Eq.12) is a function of the model $\mu_{X,Y}$, which is by nature unavailable in learning. This practical limitation motivates the extensions of Theorem 12 into a learning setting, which is the main focus of Section 6 and the applications presented in Section 7.

3.3 Proof of Theorem 12: from Discrete to Continuous Representations

The proof of Theorem 12 is presented in Section 10.2 and is divided in two main sections. The first part restricts the problem to the important case of finite alphabet representations, or VQs of X. In this discrete scenario, we use many results from information theory to prove that WIS implies OS: see Theorem 35 stated in Section 10.2.2. The decision to begin studying the case of finite alphabet representations was essential because it offers a path to adopt results on the interplay between probability of error and conditional entropy only

available for discrete random variables (see Theorem 10 in Section 3.1). In the second part, presented in Section 10.2.3, we make a connection between the discrete result in Theorem 35 and the unconstrained version of this result in Theorem 12. Importantly, the finite alphabet result, Theorem 35, is used as a building block to extend the proof argument to the continuous case. For this challenging task, results on sufficient partitions for mutual information approximation were extensively used (Liese et al., 2006).

3.4 Is WIS Strictly Weaker than IS?

On the significance of Theorem 12, it is important to confirm if WIS is strictly weaker than IS. A basic angle to address this question implies looking at the difference between the two introduced information losses, i.e., the difference between $I(X;Y)$ and $I((\tilde{U},U_i);Y)$. On this, we could say that:

- The difference IL-WIL $= I(X;Y) - I((\tilde{U},U_i);Y) \geq 0$ depends on the model $\mu_{X,Y}$ and the representation U_i. The WIL uses $\tilde{U}$ (a quantized version of X of size M) as a reference, while IL uses X, which is a continuous random variable (with infinite information) in the context of our general mixed model $\mu_{X,Y}$.
- It is known that $I(X;Y)$ is the supremum of $I(\eta(X);Y)$ over all possible finite-size measurable quantizers $\eta(\cdot)$ (Liese et al., 2006; Silva and Narayanan, 2010a; Vajda, 2002) (see the results presented in Section 5.1). Then, we say that a model $\mu_{X,Y}$ is continuous (from a MI point of view) if the MI is not achieved by any finite size representation of X (Liese et al., 2006; Silva and Narayanan, 2010a; Vajda, 2002). Conversely, a model $\mu_{X,Y}$ is discrete (from a MI point of view) if a quantized version of X achieves $I(X;Y)$, i.e., $\exists \eta(\cdot)$ a VQ such that $I(X;Y) = I(\eta(X),Y)$(Cover and Thomas, 2006). Then, WIL is strictly smaller than IL for the rich scenario where we have a continuous model and a finite alphabet representation[15].
- On the previous point, the continuous scenario for $\mu_{X,Y}$ is an important case study as we do not impose any discrete structural assumptions on the model. Furthermore, in many practical domains with continuous observations (images, audio, time series continuous signals, etc), it is reasonable to consider that any quantized (digital) version of X induces a non-zero mutual information loss about Y (see Corollary 11).

These (non-asymptotic) observations on the non-zero discrepancy of IL v.s. WIL motivate the construction presented next, where we confirm that WIS is strictly weaker than IS.

3.4.1 An illustrative Example

To illustrate the discrepancy between WIS and IS as a condition on $\{U_i\}_{i\geq 1}$ given a model $\mu_{X,Y}$, we design a simple construction:

- Y takes two values with $\mu_Y(1) = \mu_Y(2) = 1/2$.
- X given Y follows a Gaussian distribution: $X \sim Normal(K,\sigma)$ when $Y = 1$ and $X \sim Normal(-K,\sigma)$ when $Y = 2$. $K > 0$ and $\sigma > 0$ (the parameters).

15. If $\mu_{X,Y}$ is continuous, then for any VQ $\eta_i(\cdot)$: $I((\tilde{U},\eta_i(X));Y) < I(X;Y)$.

- The MPE decision is: $\tilde{U} = 1$ if $X \geq 0$ and $\tilde{U} = 2$ if $X < 0$.
- Let us consider the following collection of indexed partitions:

$$\pi_i = \left\{(-\infty, -1/2^i), [-1/2^i, 1/2^i], (1/2^i, \infty)\right\} \text{ for } i \geq 1.$$

- If we denote by A_i^1, A_i^2 and A_i^3 the cells of π_i, this produces a VQ of X determined by: $U_i = 1$ if $X \in A_i^1$, $U_i = 2$ if $X \in A_i^2$, and $U_i = 3$ if $X \in A_i^3$.
- It is simple to show that $I(U_i; Y) < I(X; Y)$ (as the model is continuous) (Liese et al., 2006; Silva and Narayanan, 2010a; Cover and Thomas, 2006) and furthermore that $\lim_{i\to\infty} I(U_i; Y) = I(\tilde{U}, Y) < I(X; Y)$.[16] In other words, the collection $\{U_i\}_{i\geq 1}$ is not IS: i.e, $I(X; Y) - I(U_i; Y) = I(X; Y|U_i)$ is not vanishing as i tends to infinity.
- In contrast, by the construction of this family, U_i determines $\tilde{U}$ in the limit (it follows that $\lim_{i\to\infty} H(\tilde{U}|U_i) = 0$) and, consequently, we have that $\lim_{i\to\infty} I(\tilde{U}; Y|U_i) = 0$ (Cover and Thomas, 2006). Therefore, this family of representations $\{U_i, i \geq 1\}$ is WIS.
- Finally, from Theorem 12, $\{U_i, i \geq 1\}$ is OS (Def. 1) but not IS (Def.2).

This example is important for our question "*is WIS strictly weaker than IS?*" as it illustrates a scenario where the difference between IL and WIL is strictly positive for any $i \geq 1$. Importantly, this discrepancy is non-vanishing when i grows: WIL tends to zero, but IL does not. Therefore, we conclude from this construction that WIS (Def. 3) is strictly weaker than IS (Def. 2) as a general condition. Furthermore in the context of this example, we observe that IL as fidelity indicator is blind on predicting the quality that the collection $\{U_i, i \geq 1\}$ has to achieve the MPE in (1). Supporting this finding, in Section 8.2 we show other constructions (a sequence of VQs) and (continuous) models where the same finding illustrated in this example is observed.

4. On the Optimality of WIS

On this section, we show that WIS is equivalent to OS if we consider a uniqueness condition of the Bayes rule in (10) of a given model $\mu_{X,Y}$. More precisely, we have that

Definition 13 *A model $\mu_{X,Y}$ is said to have a unique MPE decision rule, if for any rule $r : \mathcal{X} \to \mathcal{Y}$ such that $\mathbb{P}(r(X) \neq Y) = \ell(\mu_{X,Y})$ then $r(\cdot)$ is equal to the MAP rule $\tilde{r}_{\mu_{X,Y}}(\cdot)$ in (10) almost surely.*[17]

Theorem 14 *Let $\{U_i\}_{i\geq 1}$ be a sequence of representations for X and let us assume that $\mu_{X,Y}$ has a unique MPE decision rule (Def. 13). Then, $\{U_i\}_{i\geq 1}$ is OS for $\mu_{X,Y}$ if, and only if, $\{U_i\}_{i\geq 1}$ is WIS for $\mu_{X,Y}$.*

Therefore, under this uniqueness condition of the optimal rule for $\mu_{X,Y}$, WIS is the weakest condition on the representations needed to achieve OS. From that perspective, our result in Theorem 12 can be considered optimal for this family of models.

The proof of Theorem 14 is presented in Section 10.5.

16. We verify this strict inequality in Appendix D.
17. More precisely, $\mathbb{P}(r(X) \neq \tilde{r}_{\mu_{X,Y}}(X)) = \mu_X(\{x \in \mathcal{X} : r(x) \neq \tilde{r}_{\mu_{X,Y}}(x)\}) = 0$.

5. On the Expressiveness of Digital Representations

Important results in the literature of information theory (IT) show that a collection of measurable partitions is asymptotically sufficient for approximating the Kullback-Leibler (KL) divergence and MI (Berlinet and Vajda, 2005; Liese et al., 2006; Vajda, 2002). From these IT results and Theorem 12, we present conditions and specific constructions that demonstrate that finite-size representations (VQs) are expressive for classification, i.e., new expressiveness conditions and results for VQs (Theorems 15 and 18). Special attention is placed on the family of data-driven (stochastic) partitions (Silva and Narayanan, 2010a, 2007, 2010b; Vajda, 2002; Darbellay and Vajda, 1999; Gonzales et al., 2022). We show that these data-driven representations have the capacity to be IS with probability one and, consequently, are OS for classification with probability one from Theorem 12.

5.1 A Universal IS Representation

This section shows that digitalization (a VQ) of X offers a universal (distribution-free) representation scheme with the capacity to retain an arbitrary amount of the MI that X has about Y, i.e., IS in the sense of Def. 2. We present an approximation condition to guarantee IS for a sequence of embedded partitions and from that result a simple construction illustrating this IS capacity.

Theorem 15 *Let $\{\eta_i(\cdot), i \geq 1\}$ be a collection of discrete finite-size representations (VQs) equipped with its induced measurable partitions $\{\pi_{\eta_i}, i \geq 1\} \in \mathcal{B}(\mathbb{R}^d)$ in Eq.(4). If the collection is embedded in the sense that*[18] *$\sigma(\eta_1) \subset \sigma(\eta_2) \subset \sigma(\eta_3) \ldots$ and $\sigma(\pi_{\eta_1} \cup \pi_{\eta_2} \ldots) = \mathcal{B}(\mathbb{R}^d)$ then for any distribution $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$*

$$\lim_{i \to \infty} I(X; Y | \eta_i(X)) = 0, \tag{20}$$

and, consequently,

$$\lim_{i \to \infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y}) \tag{21}$$

where $U_i = \eta_i(X)$.

The proof is presented in Section 10.7.

Some observations about Theorem 15 are:

- The result establishes a universal (distribution-free) expressiveness condition on the quality of $\{\eta_i(\cdot), i \geq 1\}$ to make the collection of digital representations IS and OS for any model $\mu_{X,Y}$. The condition $\sigma(\pi_{\eta_1} \cup \pi_{\eta_2} \ldots) = \mathcal{B}(\mathbb{R}^d)$ expresses that the partitions approximate in the limit any measurable event in $\mathcal{B}(\mathbb{R}^d)$. The assumption that $\sigma(\eta_1) \subset \sigma(\eta_2) \subset \sigma(\eta_3) \ldots$ is a functional embedded condition meaning that $\eta_{i+1}(\cdot)$ is a refined version of $\eta_i(\cdot)$ for any $i \geq 1$.

- The proof of this result follows from the seminal work by Liese et al. (2006) on sufficient measurable partitions to approximate the KL divergence. The details on this connection are presented in the proof of Theorem 15 (Section 10.7).

18. $\sigma(\eta_i) \subset \mathcal{B}(\mathbb{R}^d)$ denotes the smallest sub-sigma field that makes $\eta_i(\cdot)$ measurable and $\sigma(\pi_{\eta_1} \cup \pi_{\eta_2} \ldots)$ denotes the smallest sub-sigma field that contains the collection of events $\bigcup_{i \geq 1} \pi_{\eta_i} \subset \mathcal{B}(\mathbb{R}^d)$ (Halmos, 1950; Varadhan, 2001; Breiman, 1968).

- On the feasibility of finding digital representations (VQs) that meet the conditions of Theorem 15, the following section offers a simple construction to demonstrate that digitalization has that universal (distribution-free) capacity.

5.1.1 The Universal (distribution-free) Construction

We present a universal partition scheme introduced by Liese et al. (2006) for the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ that is IS (distribution-free) from Theorem 15. The construction is the following:

$$\tilde{\pi}_m = \{B_{m,0}\} \cup \left\{B_{m,\bar{j}}, \bar{j} = (j_1, \ldots, j_d) \in \mathcal{J}_m\right\}, \tag{22}$$

where the index set is $\mathcal{J}_m = \{-m2^m, \ldots, m2^m - 1\}^d$ and

$$B_{m,0} = \mathbb{R}^d \setminus [-m, m)^d, \tag{23}$$

$$B_{m,j_1,\ldots,j_d} = \bigotimes_{k=1}^{d} \left[\frac{j_k}{2^m}, \frac{j_k + 1}{2^m}\right), \ \forall (j_1, \ldots, j_d) \in \mathcal{J}_m. \tag{24}$$

This construction is universal for the Borel sigma-field $\mathcal{B}(\mathbb{R}^d)$, as any interval in $\mathcal{B}(\mathbb{R}^d)$ can be approximated (arbitrarily closely) by the union of cells of $\tilde{\pi}_m$ as m goes to infinity (Liese et al., 2006). Consequently, we have that $\sigma(\cup_{m\geq 1}\tilde{\pi}_m) = \mathcal{B}(\mathbb{R}^d)$ (Liese et al., 2006). Then adopting Theorem 15, we have the following zero-information loss result:

$$\lim_{m\to\infty} I(X; Y|\eta_{\tilde{\pi}_m}(X)) = 0, \tag{25}$$

where the representation $\eta_{\tilde{\pi}_m}(\cdot)$ from $\mathbb{R}^d$ to $\{(m2^m, ..., m2^m)\} \cup \mathcal{J}_m$ of size $(m2^{m+1})^d + 1$ (the encoder induced by $\tilde{\pi}_m$) is given by:

$$\eta_{\tilde{\pi}_m}(x) = (m2^m, ..., m2^m) \cdot \mathbf{1}_{B_{m,0}}(x) + \sum_{\bar{j}\in\mathcal{J}_m} \bar{j} \cdot \mathbf{1}_{B_{m,\bar{j}}}(x). \tag{26}$$

5.1.2 Final Remarks

From (25), there is a collection of finite-size partitions (of size $(m2^{m+1})^d + 1$) that asymptotically captures all the information of (X, Y) in a distribution-free manner (independent of $\mu_{X,Y}$). Furthermore, Theorem 15 and Theorem 12 show that $\{\eta_{\tilde{\pi}_n}(\cdot), n \geq 1\}$ is OS distribution-free too: i.e., $\forall \mu_{X,Y}$, $\lim_{n\to\infty} \ell(\mu_{U_n,Y}) = \ell(\mu_{X,Y})$, where $U_n = \eta_{\tilde{\pi}_m}(X)$.

5.2 Data-Driven Partitions

The universal analysis presented in Section 5.1 offers unquestionable evidence about the power of digitalization; however, the analysis is limited to deterministic and data-independent representations. Complementing these findings, we present a result in a learning setting that shows that data-driven (stochastic) partitions can be IS (in a probabilistic sense) and, consequently, they have the potential to provide a better tradeoff between complexity (size of the partition) and information loss for a given model $\mu_{X,Y}$. We construct two data-driven representations that are IS (with probability one) and, consequently, they are OS (with probability one) in light of Theorem 12.

First, let us introduce the data-dependent (stochastic) partition concept. An n-sample partition rule $\pi_n(\cdot)$ is a mapping from the data space $(\mathbb{R}^d \times \mathcal{Y})^n$ to $\mathcal{Q}(\mathbb{R}^d)$, where $\mathcal{Q}(\mathbb{R}^d)$ denotes the space of finite size measurable partitions of $\mathbb{R}^d$. A special case of this family is when $\pi_n(\cdot)$ has the X-property (Devroye et al., 1996, Ch. 20.2) meaning that $\pi_n(\cdot)$ is a mapping from the unsupervised data space $\mathbb{R}^{dn}$ to $\mathcal{Q}(\mathbb{R}^d)$. Finally, a scheme $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), \dots\}$ is a collection of partition rules of different lengths.

We need to extend the notion of IS and OS to this stochastic representation setting:

Definition 16 *A partition scheme $\Pi = \{\pi_n, n \geq 1\}$ is said to be information sufficient (IS), if for any $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, it follows that*

$$\lim_{n\to\infty} I(\eta_{\pi_n(Z_1,\dots,Z_n)}(X); Y) = I(X;Y), \text{ with probability one} \tag{27}$$

where $Z_i \equiv (X_i, Y_i)$, $Z_1, \dots, Z_n$ are i.i.d. samples from $\mu_{X,Y}$, and $\eta_{\pi_n(Z_1,\dots,Z_n)}(\cdot)$ is the data-driven VQ induced by the partition $\pi_n(Z_1, \dots, Z_n)$.[19]

Definition 17 *A partition scheme Π is said to be operationally sufficient (OS), if for any $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$,*

$$\lim_{n\to\infty} \ell(\mu_{U_n,Y}) = \ell(\mu_{X,Y}), \text{ with probability one,} \tag{28}$$

where $U_n = \eta_{\pi_n(Z_1,\dots,Z_n)}(X)$.

5.2.1 A Shrinking-cell Condition for Data-Driven Sufficiency

Here, we present a data-driven condition on Π that guarantees that the scheme is IS (Def. 16). Before that, we need to introduce a few definitions. The diameter of an event $B \subset \mathcal{B}(\mathbb{R}^d)$ is

$$diam(B) \equiv \sup_{x,y\in B} ||x - y||, \tag{29}$$

where $||\cdot||$ is the Euclidian norm in $\mathbb{R}^d$. Considering a partition rule $\pi_n : \mathbb{R}^{dn} \to \mathcal{Q}(\mathbb{R}^d)$, a point $x \in \mathbb{R}^d$ and data-sequence $z_1^n \in (\mathbb{R}^d \times \mathcal{Y})^n$, we use $\pi_n(x|z_1^n)$ to denote the cell in $\pi_n(z_1^n)$ that contains x.

Theorem 18 *Let us consider $\mu_{X,Y}$ and $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), \dots\}$ a partition scheme driven by $Z_1, Z_2, \dots$ where $Z_i \sim \mu_{X,Y}$ for any $i \geq 1$. If μ_X has a density function*[20] *and Π satisfies that for any $\delta > 0$* [21]

$$\lim_{n\to\infty} \mu_X\left(\left\{x \in \mathbb{R}^d, diam(\pi_n(x|Z_1^n)) > \delta\right\}\right) = 0, \text{ with probability one,} \tag{30}$$

then

$$\lim_{n\to\infty} I(\eta_{\pi_n(Z_1,\dots,Z_n)}(X); Y) = I(X;Y), \text{ with probability one,} \tag{31}$$

and

$$\lim_{n\to\infty} \ell(\mu_{U_n,Y}) = \ell(\mu_{X,Y}), \text{ with probability one,} \tag{32}$$

where $U_n = \eta_{\pi_n(Z_1,..,Z_n)}(X)$.

19. It is worth noting that $\pi_n(Z_1, \dots, Z_n)$ is a random element in $\mathcal{Q}(\mathbb{R}^d)$, as it is a function of the r.v. $Z_1^n = (Z_1, \dots, Z_n)$.
20. μ_X is absolutely continuous with respect to the Lebesgue measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.
21. The probability one is with respect to the process distribution of $(Z_n)_{n\geq 1} = (X_n, Y_n)_{n\geq 1}$.

The proof is presented in Section 10.8.

Some observations about Theorem 18 are:

- Theorem 18 presents a *shrinking-cell condition* (in (30)) sufficient to make Π IS (Def. 16) and OS (Def. 17). In particular, if the diameter of the random partition $(\pi_n(X_1^n))_{n\geq 1}$ tends to zero almost surely w.r.t. the process distribution of $(Z_n)_{n\geq 1}$, then its stochastic representations $\left\{\eta_{\pi_n(Z_1,\ldots,Z_n)}(\cdot), n \geq 1\right\}$ are both IS and OS. Different flavors of this high-resolution condition have been presented in the statistical learning literature (Lugosi and Nobel, 1996; Devroye et al., 1996; Nobel, 1996). The one adopted in this work comes from results on histogram-based estimation for information measures (Silva and Narayanan, 2010a, 2012, 2010b).

- For the proof, we use (Silva and Narayanan, 2010a, Th. 2) to show that (30) implies (31) and then we use Theorem 12 to extend in this random setting that IS (Def. 16) implies OS (Def. 17).

- A special case of Theorem 18 applies when Π has the X-property, meaning that $\pi_n(\cdot)$ only depends on the unsupervised portion of the data, i.e., $(X_1, \ldots, X_n)$.

The following subsections present two constructions with the X-property (unsupervised representations) that meet the shrinking cell condition in (30).

5.2.2 Statistically Equivalent Blocks

Here, we present a scheme implementing the principle of statistically equivalent partitions (Devroye et al., 1996; Gessaman, 1970). Let $X_1, .., X_n$ be i.i.d. samples of $\mu_X \in \mathcal{P}(\mathcal{X})$ for which we assume that $\mu_X \ll \lambda$. The idea is to partition the space $\mathbb{R}^d$ by axis-parallel hyperplane in such a way that at the end of the process we have almost the same number of sample points per cell. For that, let $l_n > 0$ be the number of samples (a non zero integer) that ideally we want to have at the end of the process in every cell of π_n. The method chooses an arbitrary order of the axis-coordinate, let us say the order $(1, 2, .., d)$, and considers $T_n = \lfloor (n/l_n)^{1/d} \rfloor$ as the number of partitions to produce in every axis. The method goes as follows: choose the first coordinate and project the data in that direction $Z_1, .., Z_n \in \mathbb{R}$; compute the order statistics that we denote by $Z^{(1)} < .... < Z^{(n)}$. From this sequence, define the following axis-parallel partition of the real line:

$$\left\{I_i^1 : i = 1, .., T_n\right\} = \left\{(-\infty, Z^{(s_n)}], (Z^{(s_n)}, Z^{(2\cdot s_n)}], .., (Z^{((T_n-1)\cdot s_n)}, \infty)\right\} \subset \mathbb{R}, \tag{33}$$

where $s_n = \lfloor n/T_n \rfloor$. Then assigning the vector samples $X_1, .., X_n$ to the cells of $\pi_n^{(1)}(X_1^n) = \left\{I_i^1 \times \mathbb{R}^{d-1}, i = 1, .., , T_n\right\}$ concludes the first iteration. The second iteration applies the same principle (statistically equivalent partition with axis-parallel hyperplanes) over the cells of $\pi_n^{(1)}(X_1^n)$ but in the second coordinate, for which the original samples $X_1, .., X_n$ are assigned to each individual cell of $\pi_n^{(1)}(X_1^n)$, accordingly. At the end of the second iteration, we produce $\pi_n^{(2)}(X_1^n)$. Iterating this algorithm until the last coordinate, d, we obtain $\pi_n^{(d)}(X_1^n)$.[22] This data-driven assignment is critical to derive the following result:

22. It can be shown that $\forall A \in \pi_n^{(d)}(X_1^n)$ the number of training samples that belong to A is greater than or equal to l_n (Silva and Narayanan, 2010a).

Corollary 19 *Let μ_X be a probability in $\mathbb{R}^d$ such that $\mu_X \ll \lambda$ and let $(X_n)_{n\geq 1}$ be i.i.d. samples driven by μ_X. If (l_n) is $o(n)$,[23] it follows that for any $\delta > 0$*

$$\lim_{n\to\infty} \mu_X \left(\left\{ x \in \mathbb{R}^d, diam(\pi_n^{(d)}(x|X_1^n)) > \delta \right\} \right) = 0, \text{ with probability one,} \tag{34}$$

and then $\Pi = \left\{ \pi_n^{(d)}(\cdot), n \geq 1 \right\}$ is IS (Def. 16) and OS (Def 17).

Corollary 19 derives from (Silva and Narayanan, 2010a, Th.4) that proves that if (l_n) is $o(n)$ then (34) holds and from Theorem 18 follows the rest of the result.

5.2.3 Balanced Search Tree

Here we present a version of a balanced search tree (Devroye et al., 1996, Ch. 20), in particular the binary case studied in (Silva and Narayanan, 2010a, Sec.V.B). Given $X_1, ..., X_n$ i.i.d. samples (unsupervised data) driven by μ_X, the partition rule selects a coordinate of the set $\{1, .., d\}$ in a given sequential order, let us say the i-coordinate, and the axis-parallel hyperplane

$$H_i(X_1^n) = \left\{ x \in \mathbb{R}^d : x(i) \leq X^{(\lceil n/2 \rceil)}(i) \right\}, \tag{35}$$

where $X^{(1)}(i) < X^{(2)}(i) <, .., < X^{(n)}(i)$ is the order statistics of X_1^n projected over the coordinate i. Then, we create a binary partition of $\mathbb{R}^d$ given by $\bar{\pi}_n^{(1)}(X_1^n) = \{H_i(X_1^n), H_i(X_1^n)^c\}$. Assigning X_1^n to its respective cells in $\bar{\pi}_n^{(1)}(X_1^n)$, the process continues with the following coordinate, let us say the j-coordinate, in the mentioned sequential order applying the median statistically equivalent principle in (35) in each cell of $\bar{\pi}_n^{(1)}(X_1^n)$ by projecting the data over the j-coordinate. Importantly, the axis-parallel binary partition in (35) is conduced if the number of samples associated with the resulting cells is greater than or equal to $l_n > 0$ (a parameter of the method); otherwise, the algorithm stops the splitting process for this cell. l_n is designed to guarantee that the statistically equivalent splitting approach has a sufficient number of samples[24]. Therefore, after iterating this principle and meeting the stopping criterion in every resulting cell, we have a partition $\bar{\pi}_n^{(l_n)}(X_1^n)$ with a binary tree-structure that has almost the same number of samples in every cell (balanced). This stopping criterion and the statistically equivalent principle of this scheme are critical to prove the following:

Corollary 20 *Let μ_X be a probability on $\mathbb{R}^d$ such that $\mu_X \ll \lambda$ and let $(X_n)_{n\geq 1}$ be i.i.d. samples driven by μ_X. If (l_n) is $\mathcal{O}(n^p)$ with $p \in (0, 1)$, it follows that for any $\delta > 0$*

$$\lim_{n\to\infty} \mu_X \left(\left\{ x \in \mathbb{R}^d, diam(\bar{\pi}_n^{(l_n)}(x|X_1^n)) > \delta \right\} \right) = 0, \text{ with probability one,} \tag{36}$$

and then $\bar{\Pi} = \left\{ \tilde{\pi}_n^{(l_n)}(\cdot), n \geq 1 \right\}$ is IS (Def. 16) and OS (Def. 17).

Corollary 20 derives from (Silva and Narayanan, 2010a, Ths.5 and 6) that proves that if (l_n) is $\mathcal{O}(n^p)$ with $p \in (0, 1)$ then (36) holds and from Theorem 18 follows the rest of the result.

23. (a_n) being $o(b_n)$ means that $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

24. A systematic exposition of the statistical properties of this scheme and its implementation and use can be found in (Devroye et al., 1996) for pattern recognition and in (Silva and Narayanan, 2010a,b) for estimating information measures.

5.2.4 Final Remarks

To conclude, in Theorem 18 we show a quantitative condition to meet IS and OS for data-driven partitions (the shrinking cell condition in Eq.30) and two practical construction with the X-property that meet this requirement for a large class of models (see Corollaries 19 and 20). These results confirm that digitalization offers expressive representations in ML and that VQs learned from the principle of statistically equivalent division of the space can meet IS and OS.

6. Weak Informational Sufficiency (WIS) for a Class of Models

As we discussed in Section 3.2, the WIS condition on $\{\eta_i(\cdot)\}_{i\geq 1}$ used in Theorem 12 cannot be adopted as a criterion in an actual learning setting. The reason for this is that the reference representation $\tilde{U}$ used in Definition 3 is a function of the true model $\mu_{X,Y}$. Then, adopting this oracle WIS condition in learning, for instance, for learning representations from data, is not possible.

To give significance to the adoption of Theorem 12 in a learning setting, in this section we move towards considering a family of indexed models $\Lambda = \left\{\mu^{\theta}_{X,Y}, \theta \in \Theta\right\} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (or hypotheses) to formalize on Λ the structure of a learning task. More precisely, we assume that the unknown model $\mu_{X,Y}$ belongs to Λ, which can be seen as a form of prior knowledge. The objective here is to use this knowledge and Theorem 12 to develop a new IS condition that implies OS but is strictly weaker than pure IS (in Definition 2) for certain classes of models Λ.

6.1 Formalizing Operational Structure

For any $\mu^{\theta}_{X,Y} \in \Lambda$, let us consider its *MPE decision rule* given by[25]:

$$\tilde{r}_{\theta}(x) \equiv \arg\max_{y\in\mathcal{Y}} \mu^{\theta}_{Y|X}(y|x) \tag{37}$$

and its induced optimal partition:

$$\tilde{\pi}^{\theta} \equiv \left\{\tilde{r}^{-1}_{\theta}(\{y\}), y \in \mathcal{Y}\right\} \subset \mathcal{B}(\mathbb{R}^d). \tag{38}$$

Importantly, we can introduce

$$\sigma(\Lambda) \equiv \sigma\left(\bigcup_{\theta\in\Theta} \tilde{\pi}^{\theta}\right) \subset \mathcal{B}(\mathbb{R}^d) \tag{39}$$

to be the smallest sub-sigma field that makes all the decision rules $\{\tilde{r}_{\theta}(\cdot), \theta \in \Theta\}$ of Λ measurable from $\mathcal{X}$ to $\mathcal{Y}$ (Halmos, 1950). $\sigma(\Lambda)$ in (39) can be seen as an operational form of structure, in the sense that $\sigma(\Lambda)$ captures the smallest σ-field that makes all MPE rules in (37) measurable and, consequently, it is a function of both Λ and the operational problem in (1). In general, the smaller (or, the simpler) $\sigma(\Lambda)$ is relative to the Borel sigma field of $\mathbb{R}^d$, the more structural knowledge we have from assuming that our model belongs to Λ.

25. This rule is not unique. In general, we could select any function that is a solution of (37) and thus achieves the MPE.

6.2 Finite-Size Families

An important case to consider is when Λ has an intrinsic finite-size structure.

Definition 21 *Λ has a finite size operational structure if $\exists \pi = \{A_i, i = 1, .., K\} \subset \mathcal{B}(\mathbb{R}^d)$ such that $\sigma(\Lambda) = \sigma(\pi)$.*

The condition in Def. 21 holds, for instance, when $\bigcap_{\theta\in\Theta} \tilde{\pi}^\theta$ reduces to a measurable partition of finite size (let say $K > 0$) that we denote by $\pi = \{A_i, i = 1, .., K\} \subset \mathcal{B}(\mathbb{R}^d)$.[26] Under this finite-size assumption, we can construct a finite-size operationally sufficient representation (VQ) for the whole family Λ. In particular, we can choose a prototype $p_i \in A_i$ for every cell of π and the following VQ: $\forall x \in \mathcal{X}$

$$\eta_\Lambda(x) \equiv \sum_{j=1}^{K} \mathbf{1}_{A_j}(x) \cdot p_j \in \mathbb{R}^d. \tag{40}$$

Importantly, from (38), (39) and the construction $\eta_\Lambda(\cdot)$ in (40), it follows that for any $\theta \in \Theta$

$$\tilde{r}_\theta(x) = \tilde{r}_\theta(\eta_\Lambda(x)),\ \forall x \in \mathbb{R}^d. \tag{41}$$

Therefore, all the MPE rules on our class of models are insensitive to this lossy operator $\eta_\Lambda(\cdot)$. From the invariant (to the action of $\eta_\Lambda(\cdot)$) property presented in (41), we have the following representation result:

Proposition 22 *Let us consider a lossy mapping $\eta^* : \mathcal{X} \to \mathcal{X}$. If the family Λ is invariant to the action of $\eta^*(\cdot)$ in the sense of (41), then for any $\mu^\theta_{X,Y} \in \Lambda$*

$$\ell(\mu^\theta_{\eta^*(X),Y}) = \ell(\mu^\theta_{X,Y}). \tag{42}$$

Proof The proof derives from the observation that the condition in (41) implies that $\tilde{r}_\theta(\cdot)$ is a deterministic function of $\eta^*(\cdot)$ for any $\theta \in \Theta$, and the fact that by construction $\tilde{r}_\theta(\cdot)$ is operationally sufficient for $\mu^\theta_{X,Y}$ (i.e., $\ell(\mu^\theta_{\tilde{r}_\theta(X),Y}) = \ell(\mu^\theta_{X,Y})$). ■

Proposition 22 shows that the finite-size operational structure in Λ (Def. 21) reduces to a VQ of X ($\eta_\Lambda(\cdot)$ in Eq.40) that is operationally sufficient for all members of Λ. In general, the structure for Λ (beyond the finite-size case) is captured in $\sigma(\Lambda)$, as long as $\sigma(\Lambda)$ is strictly contained in $\mathcal{B}(\mathbb{R}^d)$. The objective of this section is to use this structure and, with that, the construction of a lossy mapping satisfying (42) to extend Theorem 12. This objective will be the focus of the following two subsections.

6.3 The Main Result

Here, we present a weak form of informational sufficiency for a family of models $\Lambda \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that implies operational sufficiency. Before that, let us introduce the following definition:

26. An example that meets this condition is presented in Section 6.4.

Definition 23 *We say that a lossy mapping $\eta^* : \mathcal{X} \to \mathcal{X}$ is operationally sufficient (OS) for a class $\Lambda = \left\{\mu^{\theta}_{X,Y}, \theta \in \Theta\right\} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ if $H(\tilde{r}_\theta(X)|\eta^*(X)) = 0$ when $X \sim \mu^{\theta}_X$ for any $\theta \in \Theta$.* [27]

At this point, we can state the following result:

Theorem 24 *Let $\{U_i\}_{i\geq 1}$ be a family of representations of X obtained from the functions $\{\eta_i(\cdot)\}_{i\geq 1}$, and let us assume that there is a transformation $\eta^* : \mathcal{X} \to \mathcal{X}$ that is OS for Λ (Def. 23). If for any models $\mu^{\theta}_{X,Y} \in \Lambda$*

$$\lim_{i\to\infty} \underbrace{\mathcal{I}(\mu^{\theta}_{(\eta^*(X),U_i),Y}) - \mathcal{I}(\mu^{\theta}_{U_i,Y})}_{I(\eta^*(X);Y|\eta_i(X)) \text{ when } (X,Y)\sim\mu^{\theta}_{X,Y}} = 0, \tag{43}$$

then

$$\lim_{i\to\infty} \ell(\mu^{\theta}_{\eta_i(X),Y}) = \ell(\mu^{\theta}_{X,Y}). \tag{44}$$

The proof is presented in Section 10.9.

Analysis and interpretation of Theorem 24:

1. Theorem 24 can be considered as a non-oracle extension of Theorem 12 as we do not need the true model to establish a sufficient condition to achieve OS in (44). Instead, we assume that the true model belongs to a class Λ with an operational structure represented by $\eta^*(\cdot)$.

2. The new sufficient condition for the representations $\{\eta_i(\cdot)\}_{i\geq 1}$ in (43) is in principle strictly weaker (if $\eta^*(\cdot)$ is a lossy mapping) than IS in Definition 2.

3. To make Theorem 24 useful, it is relevant to determine a lossy mapping $\eta^*(\cdot)$ that is OS for a class Λ. It would be ideal, although difficult, to find the simplest mapping of Λ that satisfies the condition in Definition 23. This last optimal representation problem for a given Λ is not evident and is not addressed in this work (Cover and Thomas, 2006, pp.37). However, on the existence of a lossy mapping that meets OS for Λ (Def. 23), Section 6.4 illustrates some relevant examples.

4. Complementing the previous point, one might note that the identity function from $\mathcal{X}$ to $\mathcal{X}$ is OS for any class of models Λ (Def. 23). With $\eta^*(\cdot) = id(\cdot)$, (43) reduces to the IS condition (Def. 2). Then in this trivial context, Theorem 24 recovers the known result that if $\{U_i\}_{i\geq 1}$ is IS for $\mu^{\theta}_{X,Y}$ then $\{U_i\}_{i\geq 1}$ is OS for $\mu^{\theta}_{X,Y}$.

5. Notably, the finite-size family introduced in Section 6.2 has a lossy function in (40) (a VQ) that meets the requirement of being OS (Def. 23) for Λ. Then, Theorem 24 applies in this case in a non-trivial way. More details about this class of models can be found in Section 6.4.1.

27. Definition 23 implies that $\tilde{r}_\theta(\cdot)$ is a deterministic function of $\eta^*(\cdot)$ for any $\theta \in \Theta$ (μ^{θ}_X almost surely) and, consequently, we have that $\ell(\mu^{\theta}_{\eta^*(X),Y}) = \ell(\mu^{\theta}_{X,Y})$, from the same argument used to prove Proposition 22 from (41).

6. There is a special representation scenario of the setting of Theorem 24 worth describing. The scenario is when $\eta_i(\cdot)$ is a deterministic function of $\eta^*(\cdot)$ for any $i \geq 1$, i.e., for any i, there is $\tilde{\eta}_i(\cdot)$ such that $\eta_i(\cdot) = \tilde{\eta}_i(\cdot) \circ \eta^*(\cdot)$. In this projected context, it is simple to verify that our weak IL $\mathcal{I}(\mu^{\theta}_{(\eta^*(X),U_i),Y}) - \mathcal{I}(\mu^{\theta}_{U_i,Y})$ in (43) is the pure IL induced by $\tilde{\eta}_i(\cdot)$ for the induced transform model $\mu^{\theta}_{\eta^*(X),Y}$. Then, in this setting, the WIS condition in (43) reduces to the pure IS condition for $\{\tilde{\eta}_i(\cdot)\}_{i\geq 1}$ on the transform (or projected) model $\mu^{\theta}_{\eta^*(X),Y}$. We present an example of this projected learning scenario in Section 7.1.

7. Finally, for the general (non-projected) setting stated in Theorem 24, the proof of this result derives from Theorem 12. In Section 7, we present two unsupervised representation learning algorithms operating in Theorem 24's general non-projected context.

6.4 Application of Theorem 24

At this point, we could use $\sigma(\Lambda)$ to determine a lossy mapping $\eta(\cdot)$ that meets Definition 23 for Λ. Along this line, we revisit the case of finite-size families and introduce a class of models that satisfies some invariant properties to illustrate two relevant contexts where Theorem 24 can be adopted.

6.4.1 Finite Size Families

In the special case when Λ has an intrinsic finite-size structure (see Section 6.2), we have a vector quantizer (VQ) $\eta(\cdot)$ in (40) that satisfies the conditions in (41) and, consequently, is OS for Λ (Definition 23). Therefore, Theorem 24 applies in this case using the finite-size VQ in (40) as the sufficient representation for the class Λ in (43). In this case, condition (43) is strictly weaker than pure IS and, consequently, the result in Theorem 24 is meaningful.

A simple example of this class of problems is when $\mu_{X,Y}$ belongs to a finite class of hypotheses $\Lambda = \left\{\mu^i_{X,Y}, i = 1, .., L\right\}$, where it is possible to show from (39) that $\sigma(\Lambda) = \sigma(\pi)$ where $\pi \equiv \bigcap_{i=1}^{L} \tilde{\pi}^i$ and $\tilde{\pi}^i$ is the M-size partition induced by the MPE decision rule of $\mu^i_{X,Y}$ in (38), and, consequently, Λ has an intrinsic finite size $K \leq M^L < \infty$.[28]

6.4.2 Invariant Models

Another interesting example is when Λ has some invariances and operational symmetries (Bloem-Reddy and Teh, 2020). We will consider the case where Λ is invariant w.r.t. the action of a compact group $\mathcal{G}$ of measurable transformations on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$[29]. For this purpose, we introduce a new form of operational invariances for Λ relative to a group $\mathcal{G}$.

28. The proof that $\sigma(\Lambda) = \sigma(\pi)$ for $\Lambda = \left\{\mu^i_{X,Y}, i = 1, .., L\right\}$ is presented in the Appendix E.
29. A compact group $\mathcal{G}$ acting on $\mathcal{X}$ is a collection of Borel measurable functions from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that: for any pair $g, h \in \mathcal{G}$, $g \circ h \in \mathcal{G}$; for any $g \in \mathcal{G}$, $g^{-1} \in \mathcal{G}$; and the identity mapping belongs to $\mathcal{G}$ (Eaton, 1989; Rotman, 1995).

Definition 25 *(Functional invariance) A measurable transformation* $f : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathcal{U}, \mathcal{B}(\mathcal{U}))$ *is* $\mathcal{G}$*-invariant if for any* $g \in \mathcal{G}$

$$f \circ g(x) = f(x), \ \forall x \in \mathcal{X}.$$

Definition 26 *(Operational invariance of a model) A model* $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ *is said to be operational invariant with respect to* $\mathcal{G}$ *(in short* $\mathcal{G}$*-invariant) if there is a MPE decision rule, solution of (37), which is* $\mathcal{G}$*-invariant (Def. 25).*

Definition 27 *(Operational invariance of a class) A class* $\Lambda = \left\{\mu^{\theta}_{X,Y}, \theta \in \Theta\right\} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ *is said to be operational invariant with respect to* $\mathcal{G}$ *(in short* $\mathcal{G}$*-invariant) if* $\mu^{\theta}_{X,Y}$ *is* $\mathcal{G}$*-invariant (Def. 26) for any* $\mu^{\theta}_{X,Y} \in \Lambda$*.*

A lossy mapping $\eta^*_{\mathcal{G}}(\cdot)$ **for** $\mathcal{G}$**:** Let us consider the orbit of $\mathcal{G}$ at $x \in \mathcal{X}$ given by

$$\mathcal{G}(x) \equiv \{g(x), g \in \mathcal{G}\} \in \mathcal{B}(\mathcal{X}).$$

We can induce an equivalent relationship in $\mathcal{X}$ where $x \longleftrightarrow y$ if $\mathcal{G}(x) = \mathcal{G}(y)$ and from this a measurable partition of $\mathcal{X}$ given by $\pi_{\mathcal{G}} \equiv \{\mathcal{G}(x), x \in \mathcal{X}\} \subset \mathcal{B}(\mathcal{X})$. $\pi_{\mathcal{G}}$ is the collection of orbits induced by the application of $\mathcal{G}$ in every point of $\mathcal{X}$. Importantly, there exists a measurable cross section $\mathcal{C} \subset \mathcal{X}$ of $\mathcal{G}$ (Eaton, 1989)[30], which is a collection of prototypes for every orbit of $\pi_{\mathcal{G}}$ satisfying that $\forall x \in \mathcal{X}$

$$\mathcal{C} \cap \mathcal{G}(x) \text{ is a singleton (i.e. } \mathcal{C} \text{ selects one prototype for every cell of } \pi_{\mathcal{G}}), \tag{45}$$

and we denote this element (selection) by $\mathcal{C}(x)$. Then, we can construct the following lossy mapping $\eta^*_{\mathcal{G}}(x) \equiv \mathcal{C}(x)$ from $\mathcal{X}$ to $\mathcal{C} \subset \mathcal{X}$. From construction, $\eta^*_{\mathcal{G}}(\cdot)$ is Borel measurable and $\mathcal{G}$-invariant (Def. 25). Importantly, $\eta^*_{\mathcal{G}}(\cdot)$ is also *maximal-invariant* in the sense that for any pair $x, y \in \mathcal{X}$ that belongs to different orbits of $\pi_{\mathcal{G}}$ (i.e., $\mathcal{G}(x) \cap \mathcal{G}(y) = \emptyset$) then $\eta^*_{\mathcal{G}}(x) \neq \eta^*_{\mathcal{G}}(y)$. This last discrimination property over disjoint orbits of $\pi_{\mathcal{G}}$ is central to show that $\eta^*_{\mathcal{G}}(\cdot)$ is OS (Def. 23) for the collection of $\mathcal{G}$-invariant models (Def. 27):

Proposition 28 *Let* $\Lambda = \left\{\mu^{\theta}_{X,Y}, \theta \in \Theta\right\}$ *be* $\mathcal{G}$*-invariant (Def. 27) w.r.t. a compact measurable group* $\mathcal{G}$ *acting on* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$*. Then* $\eta^*_{\mathcal{G}}(\cdot)$ *is OS for* Λ *(Def. 23) meaning that* $\forall \theta \in \Theta$ *there is* $\tilde{r}_{\theta}(\cdot)$*, solution of (37), such that* $H(\tilde{r}_{\theta}(X)|\eta^*_{\mathcal{G}}(X)) = 0$ *when* $X \sim \mu^{\theta}_X$ *and, consequently,* $\ell(\mu^{\theta}_{\eta^*_{\mathcal{G}}(X),Y}) = \ell(\mu^{\theta}_{X,Y})$*.*

The proof is presented in Appendix G. Then, we have the following:

Corollary 29 *Let* $\Lambda = \left\{\mu^{\theta}_{X,Y}, \theta \in \Theta\right\}$ *be operational* $\mathcal{G}$*-invariant (Def. 27), and let* $\{U_i\}_{i \geq 1}$ *be a family of representations of* X *obtained from* $\{\eta_i(\cdot)\}_{i \geq 1}$*. For any* $\mu^{\theta}_{X,Y} \in \Lambda$*, if*

$$\lim_{i \to \infty} \underbrace{\mathcal{I}(\mu^{\theta}_{(\eta^*_{\mathcal{G}}(X),U_i),Y}) - \mathcal{I}(\mu^{\theta}_{U_i,Y})}_{I(\eta^*_{\mathcal{G}}(X);Y|\eta_i(X)) \text{ when } (X,Y) \sim \mu^{\theta}_{X,Y}} = 0, \tag{46}$$

then $\lim_{i \to \infty} \ell(\mu^{\theta}_{\eta_i(X),Y}) = \ell(\mu^{\theta}_{X,Y})$*.*

30. A systematic exposition of this result can be found in (Bloem-Reddy and Teh, 2020) and references therein.

Proof The result follows from Proposition 28 and Theorem 24. ■

Remark 30 *It is important to mention that our definition of operational invariances (Def. 27) for Λ only implies that $\eta^*_{\mathcal{G}}(\cdot)$ is operationally sufficient (i.e., the requirement that $\ell(\mu^{\theta}_{\eta^*_{\mathcal{G}}(X),Y}) = \ell(\mu^{\theta}_{X,Y})$ for any $\mu^{\theta}_{X,Y} \in \Lambda$). This operational condition does not imply that $\eta^*_{\mathcal{G}}(\cdot)$ is information sufficient for all the models in Λ. Therefore, we could have that $\mathcal{I}(\mu^{\theta}_{X,Y}) > \mathcal{I}(\mu^{\theta}_{\eta^*_{\mathcal{G}}(X),Y})$ where $(X,Y) \sim \mu^{\theta}_{X,Y}$ for some model $\mu^{\theta}_{X,Y} \in \Lambda$. This last non-zero information loss condition makes Corollary 29 non-trivial and interesting because, under this scenario, (46) is strictly weaker than pure IS in Definition 2.*

6.4.3 $\mathbb{S}_d$-invariant models

An important example of Definition 27 is the collection of models invariant to the action of permutations of the coordinates of $\mathcal{X} = \mathbb{R}^d$.[31] In this case, the compact group $\mathcal{G}$ is denoted by $\mathbb{S}_d$ where for any $g \in \mathbb{S}_d$ there is a permutation of $[d] = \{1, , ., d\}$ $p : [d] \to [d]$ such that $g(\mathbf{x}) = (x_{d(1)}, x_{d(2)}, .., x_{d(d)})$ $\forall \mathbf{x} \in \mathbb{R}^d$. Therefore, if a function $f(\cdot)$ is $\mathbb{S}_d$-invariant (see Def. 25), it means that its output is invariant to the action of any permutation of $\mathbf{x} = (x_1, .., x_d)$, and therefore $f(\mathbf{x})$ depends on the set $\{x_1, .., x_d\} \subset \mathbb{R}$ induced by $\mathbf{x}$.[32]

Here we consider the family of operational $\mathbb{S}_d$-invariant models $\mathcal{P}^{\mathbb{S}_d} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ from Definition 27. For this group, it is well known that the *empirical distribution* $\mathcal{M} : \mathbb{R}^d \to \mathcal{P}(\mathbb{R})$ [33] is invariant to the actions of $\mathbb{S}_d$, but, more importantly (for the adoption of Corollary 29), $\mathcal{M}(\cdot)$ is *maximal-invariant* for $\mathbb{S}_d$ (Bloem-Reddy and Teh, 2020). Then, in the adoption of Corollary 29 for $\mathcal{P}^{\mathbb{S}_d}$, we could consider the lossy mapping $\eta^*_{\mathbb{S}_d}(\cdot) = \mathcal{M}(\cdot)$.

6.4.4 Other Invariant Examples

In Section 8, we illustrate other simple examples of models with an operational structure where Theorem 24 can be used. In particular, examples are presented with specific symmetries and operational invariant properties (to translation, rotation, and scale operations), as well as some expressive representations that show that the WIS condition in Theorem 24 is strictly weaker than IS and, as a consequence, the evident gap that might exist between OS and IS in some scenarios.

6.4.5 A Final Remark: Probabilistic Invariance

Bloem-Reddy and Teh (2020) studied a stronger notion of invariance for Λ under the action of a compact group $\mathcal{G}$. They consider the case where Λ is $\mathcal{G}$-invariant if for any $g \in \mathcal{G}$ and any model $\mu^{\theta}_{X,Y} \in \Lambda$, it follows that $(X,Y) = (g(X),Y)$ in distribution when $(X,Y) \sim \mu^{\theta}_{X,Y}$. This means that the complete joint model of (X,Y) is invariant to the actions of $\mathcal{G}$. Importantly, this model-based assumption on Λ is stronger than the operational $\mathcal{G}$-

31. This class is systematically studied in the excellent paper by Bloem-Reddy and Teh (2020).
32. A complete characterization of this family of permutation invariant functions is presented in (Zaheer et al., 2017) and revisited and extended for a family of probabilistic models in (Bloem-Reddy and Teh, 2020).
33. $\mathcal{M}(\mathbf{x}) = \frac{1}{d}\sum_{i=1}^{d} \delta_{x_i}(\cdot) \in \mathcal{P}(\mathbb{R})$ denotes the empirical distribution induced by $\mathbf{x}$ in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

invariant condition we used in Definition 27 to derive Corollary 29.[34] Furthermore, under the model-based invariance used in (Bloem-Reddy and Teh, 2020), the authors showed that any maximal invariant transformation $\eta^*(\cdot)$ of $\mathcal{G}$ offers a D-separation of the model $\mu^{\theta}_{X,Y} \in \Lambda$, in the sense that $I(X;Y|\eta^*(X)) = 0$ when $(X,Y) \sim \mu^{\theta}_{X,Y}$. Then $\eta^*(X)$ is an information sufficient representation of Y, which is strictly stronger than the concepts of OS (in Def. 23) used to derive Corollary 29 (from Theorem 24). Indeed, under this stronger model-based invariant assumption for Λ, it is simple to show that (46) reduces to the IS condition in Definition 2, and the application of Corollary 29 reduces to restate the result (from Theorem 12) that if $\{U_i\}_{i\geq 1}$ is IS for every model in Λ, then $\{U_i\}_{i\geq 1}$ is OS for every model in Λ.

7. Lossy Compression Algorithms

In this section, Theorem 24 (non-oracle WIS $\Rightarrow$ OS) is used to explain the operational quality of existing machine learning algorithms. We analyze two compression-based methods that select lossy representations (encoders) by implementing info-max optimization principles: a version of the information bottleneck (IB) method (Tishby et al., 1999; Tishby and Zaslavsky, 2015; Chechik et al., 2005) and a version of the recently introduced lossy compression for lossless prediction (LCLP) method (Dubois et al., 2021). For these analyses, we introduce a simple adaptation of the IB method to deal with the operational assumption studied in Section 6.

7.1 Information Bottleneck in a Projected OS Domain

Under the prior knowledge assumed in Theorem 24, there is a simple adaptation of the IB method that can be used if $\mu_{X,Y} \in \Lambda$ (prior knowledge) and there is a lossy mapping $\eta_\Lambda(\cdot)$ that is OS for Λ (Def. 23). For that objetive, let us use the OS variable $U_\Lambda \equiv \eta_\Lambda(X)$ as our new reference where $\eta_\Lambda : \mathcal{X} \to \mathcal{X}$. Then, we can adapt the IB method in the lossy representation U_Λ as the solution of:

$$\max_{U} I(U;Y) \; s.t. \; I(U_\Lambda = \eta_\Lambda(X); U) \leq B, \tag{47}$$

where $U = \eta(U_\Lambda)$ and $\eta(\cdot)$ represents a collection of lossy encoders (or latent variable) obtained from U_Λ, i.e., measurable functions acting on U_Λ (more details in Theorem 31). In (47), $B > 0$ parametrizes a compression constraint on U, i.e., the information bottleneck.

The IB method has been adopted as a learning principle in ML to learn compressed supervised latent variables (Zaidi et al., 2020). DNN has been used for solving a stochastic version of (47)[35] by inducing a family of expressive parametric encoders (Alemi et al., 2017; Amjad and Geiger, 2019; Achille and Soatto, 2018b). Indeed, it has been argued that the well-known functional expressive power of DNN offers the potential to create supervised representations U^B (solution of the regularized info-max problem in (47)) with the capacity to approximate the optimal tradeoff between compression $I(U_\Lambda;U)$ and information $I(U;Y)$ in (47).

34. In fact, it is simple to verify that if $\mu_{X,Y}$ is $\mathcal{G}$-invariant in the sense that $(X,Y) = (g(X),Y)$ in distribution for any $g \in \mathcal{G}$, then $\mu_{X,Y}$ is operational $\mathcal{G}$-invariant (Def. 26). For completeness, this is shown in Appendix H.

35. The called deep variational IB problem (Goldfeld and Polyanskiy, 2020, Sec. IV).

For any model $\mu_{X,Y} \in \Lambda$, it is expected that U^B, the solution of (47) for $B > 0$, achieves the maximum MI value $I(U_\Lambda; Y)$ of this projected setting, as we relax the compression constraint $B \to \infty$. Then, the IB algorithm in (47) has the potencial to satisfy that $I((U_\Lambda, U^B); Y) - I(U^B; Y) \to 0$ (i.e., the WIS condition in Eq.43) as B increases. Consequently, Theorem 24 can be used to justify the expressive power of the IB method in the MPE sense. Supporting this claim, we have the following result for the IB algorithm in (47):

Theorem 31 *Let $\Lambda \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and let us consider a mapping $\eta_\Lambda : \mathcal{X} \to \mathcal{X}$ that is OS for Λ (Def. 23). If $\mathcal{F}(X, \mathbb{R})$ represents the family of measurable functions from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $\eta^B(\cdot) \in \mathcal{F}(X, \mathbb{R})$ denotes the projected-based IB encoder solution of*

$$\arg \max_{\eta(\cdot) \in \mathcal{F}(X,\mathbb{R})} I(\eta(U_\Lambda); Y) \text{ s.t. } I(U_\Lambda; \eta(U_\Lambda)) \leq B, \tag{48}$$

then for any model $\mu_{X,Y} \in \Lambda$, we have that

$$\lim_{B \to \infty} I(U_\Lambda; Y | \eta^B(U_\Lambda)) = 0, \tag{49}$$

and, consequently,

$$\lim_{B \to \infty} \ell(\mu_{\eta^B(U_\Lambda), Y}) = \ell(\mu_{X,Y}). \tag{50}$$

The proof of Theorem 31 is presented in Section 10.10.

- Theorem 31 shows that the projected version of the IB algorithm introduced in (47) can produce expressive representations for classification, i.e., by finding the optimal tradeoff between information (minimizing the WIL in this case) and compression we obtain compressed representations that achieve lossless prediction in the MPE sense.
- The proof of Theorem 31 shows that it is feasible to construct a sequence of finite-information versions (in bits) of U_Λ that are the solution of the IB problem in (47). These compressed representations $\{U_m, m \geq 1\}$ meet only the WIS fidelity criterion: i.e., $\lim_{m \to \infty} I(U_\Lambda; Y | U_m) = 0$. Indeed, if U_Λ has a non-zero information loss for $\mu_{X,Y}$, i.e., $I(X; Y | U_\Lambda) > 0$ then $\{U_m, m \geq 1\}$ is not IS (from the proof of Theorem 31). Consequently, the adapted IB algorithm in (47) is exploiting the prior structural knowledge of Λ and only meeting the weaker WIS fidelity condition stated in Theorem 24.
- The proof of this result comes from three non-trivial technical elements elaborated in this work: the universal expressiveness of digital encoders studied in Section 5, Theorem 15 in Section 5.1, and Theorem 24 (non-oracle WIS $\Rightarrow$ OS).
- Finally, for the range of the encoder $\eta(\cdot)$, Theorem 31 can be extended to any finite-dimensional continuous space.

Remark 32 *It is worth noting that the identity function is OS for any model in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. In this trivial scenario, from the perspective of modeling prior knowledge for a learning task (see Section 6), Theorem 31 can be used. Our result shows that the original IB algorithm*

acting on X (Tishby et al., 1999; Tishby and Zaslavsky, 2015) offers IS representations for any model $\mu_{X,Y}$. Then, we have that $I(X;Y|U^B) \to 0$ and, consequently, $\ell(\mu_{U^B,Y}) \to \ell(\mu_{X,Y})$ from Theorem 31. This universal predicting capacity of the IB method is non-evident considering that X has infinite information (when μ_X has a density function (Renyi, 1959)) and the representation U^B obtained from (47) only preserves a finite bit description of X. Then, for any $B > 0$, the information loss of U^B relative to the information of X is unbounded. However, the information loss of predicting Y from U^B relative to the same prediction from X is vanishing as B increases. This exciting result verifies the expressive power of digitalization (in the form of compression) in conjunction with an info-max learning principle. To the best of our knowledge, this is the first time that the IB method's IS and OS lossless prediction capacity has been proven in the machine learning literature.

7.2 (Unsupervised) Lossy Compression for Lossless Prediction

A sufficient condition to meet that $I(\eta_\Lambda(X);Y|U_i) \to 0$ is asking for $H(\eta_\Lambda(X)|U_i) \to 0$. This condition is only applicable to discrete classes of models, i.e., when the range of $\eta_\Lambda(\cdot)$ is finite or countable with $H(\eta_\Lambda(X)) < \infty$ (see Section 6.2) and, consequently, the entropy and conditional entropy of $\eta_\Lambda(X)$ are well defined (Cover and Thomas, 2006; Renyi, 1959). In this discrete context, $H(\eta_\Lambda(X)|U_i) \to 0$ implies the WIS condition of Theorem 24.[36] This new information vanishing condition is non-supervised, a property useful for analyzing many representation learning methods (Bengio et al., 2013; Kingma and Welling, 2014).

Relevantly, this non-supervised sufficient condition can be adopted to evaluate the OS expressiveness of an existing compression-based algorithm. To illustrate this capacity, here we revisit the work by Dubois et al. (2021) on "*Lossy Compression for Lossless Prediction*"(LCLP). The authors proposed a compression-based learning algorithm designed to be minimax optimal over a collection of invariant models denoted by Λ. As in Section 6, Λ is used to model some prior knowledge about the learning task (the downstream tasks). Using the theory of rate-distortion for lossy compression (Gray, 1990b; Berger, 1971), the authors proposed an algorithm that solves the optimal tradeoff between *compression* $I(X;U)$ (in bits) and *fidelity* $H(\eta_\Lambda(X)|U)$[37]. Here, we analyze a version of this problem: finding U_δ (a compressed description of X) that is solution of the following non-supervised task:

$$\min_U I(X;U) \text{ s.t. } H(\eta_\Lambda(X)|U) \leq \delta. \tag{51}$$

Solving (51) offers a collection of compressed variables $\{U_{\delta_n}\}_{n\geq 1}$ where $H(\eta_\Lambda(X)|U_{\delta_n}) \leq \delta_n$ by the information constraint in (51) (see Dubois et al. 2021, Th.2). Then, if we design $\delta_n \to 0$ as n grows, $\{U_{\delta_n}\}_{n\geq 1}$ meets our WIS conditions in Eq.(43) because $I(\eta_\Lambda(X);Y|U_{\delta_n}) \leq H(\eta_\Lambda(X)|U_{\delta_n}) \to 0$ as (δ_n) is $o(1)$. Consequently, we obtain that the solutions $\{U_{\delta_n}\}_{n\geq 1}$ meet the OS fidelity criterion from Theorem 24. As in the case of the IB algorithm in (47), this LCLP algorithm has the expressive capacity to meet WIS, but it is not designed to meet IS when $\eta_\Lambda(\cdot)$ is a lossy information mapping, i.e., when $I(X;Y|\eta_\Lambda(X)) > 0$. Then, the representations obtained from (51) can meet the WIS condition of Theorem 24 and, consequently, achieve lossless prediction (OS) with no guarantee to be IS.

36. $H(\eta_\Lambda(X)|U_i) \to 0$ implying $I(\eta_\Lambda(X);Y|U_i) \to 0$ follows from the fact that $I(\eta_\Lambda(X);Y|U_i) \leq H(\eta_\Lambda(X)|U_i)$ in this discrete setting (Cover and Thomas, 2006).

37. Remarkably, under some regularity conditions on the family of invariant models Λ, $H(\eta_\Lambda(X)|U)$ is shown to be the worse-case information loss over Λ (Dubois et al., 2021, Proposition 1).

In summary, under the assumption used in (Dubois et al., 2021, Proposition 1) that $\mu_{X,Y} \in \Lambda$ and Λ is a discrete family, the unsupervised lossy compression algorithm proposed by Dubois et al. (2021) in (51) can be designed to extract all the MI that $\eta_\Lambda(X)$ has about Y distribution-free over Λ. In this context, Theorem 24 explains the expressive power of this algorithm in the MPE sense (i.e., OS), proving that these lossy representations can achieve lossless prediction (i.e., zero operational loss).[38]

Remark 33 *Alternatively, this new (unsupervised) condition $H(\eta_\Lambda(X)|U_i) \to 0$ can be used to modify the IB algorithm in (48) minimizing $H(U_\Lambda|\eta(U_\Lambda))$ instead of maximizing $I(\eta(U_\Lambda);Y)$, making this modified IB algorithm non-supervised. This algorithm is presented next.*

7.3 The Unsupervised IB Algorithm: A Compressed Auto-encoder

The following result is an extension of the IB method and shows the operational expressiveness of an unsupervised variation of the IB algorithm introduced in Eq.(48), under the discrete assumption that $H(\eta_\Lambda(X)) < \infty$.

Theorem 34 *Let $\Lambda \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and let us consider a mapping $\eta_\Lambda : \mathcal{X} \to \mathcal{X}$ that is OS for Λ (Def. 23) and discrete in the sense that $H(\eta_\Lambda(X)) < \infty$ for any $\mu_{X,Y} \in \Lambda$. If $\mathcal{F}(X, \mathbb{R})$ represents the family of measurable functions from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$, and $\tilde{\eta}^B(\cdot) \in \mathcal{F}(X, \mathbb{R})$ denotes the unsupervised IB encoder solution of*

$$\arg \min_{\eta(\cdot) \in \mathcal{F}(X,\mathbb{R})} H(U_\Lambda|\eta(X)) \text{ s.t. } I(U_\Lambda; \eta(X)) \leq B, \tag{52}$$

then for any model $\mu_{X,Y} \in \Lambda$, we have that

$$\lim_{B \to \infty} I(U_\Lambda; Y|\tilde{\eta}^B(X)) = 0, \tag{53}$$

and, consequently,

$$\lim_{B \to \infty} \ell(\mu_{\tilde{\eta}^B(X),Y}) = \ell(\mu_{X,Y}). \tag{54}$$

The proof of Theorem 34 is presented in Appendix 10.11.

8. Numerical Analysis

In this section, we design some simple examples (the model $\mu_{X,Y}$ and family of representations $\{U_i, i \geq 1\}$) to illustrate the interplay between information loss and operation loss studied in this work. For a given model $\mu_{X,Y}$, we consider different families of representations of X. We focus on discrete encoders (VQs). We consider universal IS partitions from the results in Section 5.1, and IS data-driven partitions from the results in Section 5.2. The idea is to have a diverse range of representations of X in terms of information losses and see how this diversity translates in the operation loss. From this, we analyze scenarios where WIS is strictly weaker than IS complementing the example presented in Section 3.4.1. We also evaluate if the order of representations obtained from the information loss is an adequate predictor of the order obtained with the operation loss for regimes where the information loss is non-zero.

38. The lossless prediction was mentioned as one of the intended properties of this method in (Dubois et al., 2021). Our result (Theorem 24) proves this result.

8.1 Settings and Experimental Design

We consider three classes of models. Each of them simple enough to approximate the true information losses and operation losses, and each of them expressing some interesting structure that could be used as a prior knowledge to obtain more effective representations for the task.

8.1.1 The Models

We consider a simple setting with $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{1, \dots, 4\}$ using a uniform marginal (i.e., $\mu_Y(\{y\}) = 1/4$). The probability of X given $Y = y$ (i.e., $\mu_{X|Y}(\cdot|y)$) follows a normal distribution ($\sim \mathcal{N}(\mathbf{m}_{X|Y}(y), \mathbf{K}_{X|Y}(y))$) with an isotropic covariance $\mathbf{K}_{X|Y}(y) = \sigma_y^2 \cdot \mathbf{I}_{2\times 2}$, where $\mathbf{I}_{2\times 2}$ denotes the identity matrix. The mean is denoted by $\mathbf{m}_{X|Y}(y) \equiv \mathbb{E}(X|Y = y) \in \mathbb{R}^2$. In this parametric context, we consider three scenarios (classes of models):

- *Scale Invariant Models*: In this family, the mean vector is given by: $\mathbf{m}_{X|Y}(1)^t = (\alpha, \alpha)$, $\mathbf{m}_{X|Y}(2)^t = (-\alpha, \alpha)$, $\mathbf{m}_{X|Y}(3)^t = (\alpha, -\alpha)$ and $\mathbf{m}_{X|Y}(4)^t = (-\alpha, -\alpha)$ and $\mathbf{K}_{X|Y}(y) = \sigma^2 \cdot \mathbf{I}_{2\times 2}$. Consequently, two scalar parameters (degrees of freedom) $\sigma > 0$ and $\alpha > 0$ determine the joint distribution of (X, Y), which we denote by $\mu^S_{X,Y}(\sigma, \alpha)$. Samples of the 4 conditional distributions $\left\{\mu^S_{X|Y}(\cdot|y), y \in \mathcal{Y}\right\}$ are illustrated in Fig.1. From the symmetry of the class $\left\{\mu^S_{X,Y}(\sigma, \alpha), \sigma > 0, \alpha > 0\right\} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, the MPE rule of $\mu^S_{X,Y}(\sigma, \alpha)$ is independent of α and σ, and induces the following OS partition

$$\tilde{\pi}_S = \left\{ \underbrace{[0,\infty) \times [0,\infty)}_{\tilde{A}_{S,1}}, \underbrace{(-\infty,0) \times [0,\infty)}_{\tilde{A}_{S,2}}, \underbrace{[0,\infty) \times (-\infty,0)}_{\tilde{A}_{S,3}}, \underbrace{(-\infty,0) \times (-\infty,0)}_{\tilde{A}_{S,4}} \right\}. \tag{55}$$

  The MPE rule is $\tilde{r}_S(x) = \sum_{j=1}^4 1_{\tilde{A}_{S,j}}(x) \cdot j$. For the following analysis, we use $\sigma^2 = 1$ and $\alpha = 1.5$.

- *Translation Invariant Models*: This 2D joint distribution $\mu_{X,Y}$ follows the same Gaussian parametric structure for $\mu_{X|Y}(\cdot|y)$ and $\mathbf{K}_{X|Y}(y)$ as the previous scenario but with $\mathcal{Y} = \{1, \dots, 5\}$. In this case, the mean vectors of the 5 equiprobable clases are oriented in one (1D) direction as illustrated in Fig. 7. This 1D linear disposition of the mean vectors makes the MPE rules invariant (and the $\mu_{X,Y}$, see Def. 26) to any translation in the direction that is orthogonal to the direction used to place the mean vectors of $\left\{\mu_{X|Y}(\cdot|y), y \in \mathcal{Y}\right\}$.

- *Rotation Invariant Models*: We use the same 2D Gaussian parametric structure of the previous two examples. In this scenario, $\left\{\mu_{X|Y}(\cdot|y), y \in \mathcal{Y}\right\}$ shares the same mean vector (the zero vector for simplicity), where σ_y in $\mathbf{K}_{X|Y}(y) = \sigma_y^2 \cdot \mathbf{I}_2$ is a function of the class $y \in \mathcal{Y}$.[39] This centered structure makes the model MPE rule invariant to any rotation of the space (see Defs. 25 and 26). Samples of this model are illustrated in Fig. 9.

39. In particular, we consider $|\mathcal{Y}| = 3$ and $\sigma_1^2 = 1$, $\sigma_2^2 = 3$ and $\sigma_3^2 = 10$.

8.1.2 The Partitions

For the three models presented in Section 8.1.1, we will use the following partitions of $\mathcal{X}$:

- *Product partition (PP)*: we consider $K > 0$ sufficiently large, and we produce a uniform quantization of the bounded space $[-K, K] \times [-K, K]$ following the product structure presented in Section 5.1.1. We denote these partitions by $\left\{\pi_n^P, n = 1 \dots N\right\} \subset \mathcal{Q}(\mathcal{X})$ with sizes $k_n = \left|\pi_n^P\right|$ and its representations in (26) (VQs) by $\left\{\eta_n^P(\cdot), n = 1 \dots N\right\}$. An illustration of this partition is presented in Fig.1a.

- *Gessaman partition (GP)*: Using i.i.d. samples from μ_X, i.e., $X_1 \dots X_m$, we implement the statistically equivalent partition presented in Section 5.2.2. Therefore, we have a family of data-driven partitions that we denote by $\left\{\pi_n^G, n = 1, .., N\right\} \subset \mathcal{Q}(\mathcal{X})$ with $k_n = \left|\pi_n^G\right|$. The different sizes (number of cells) were obtained fixing $X_1 \dots X_m$ and changing the threshold ℓ_n used to construct π_n^G (see Eq.33).[40] An illustration of this non-product data-driven partition is presented in Fig. 1b.

- *Tree-structured partition (TSP)*: Using i.i.d. samples from μ_X, i.e., $X_1 \dots X_m$, we implement the TSP in Section 5.2.3. This scheme produces a family of data-driven partitions that we denote by $\left\{\pi_n^T, n = 1 \dots N\right\} \subset \mathcal{Q}(\mathcal{X})$ with $k_n = \left|\pi_n^T\right|$ where the different sizes (number of cells) were obtained by changing the threshold ℓ_n.[41] An illustration of this data-driven partition is presented in Fig. 1c.

For each model $\mu_{X,Y}$, we use at least three partition schemes: $\left\{\pi_n^P\right\}$, $\left\{\pi_n^G\right\}$, and $\left\{\pi_n^T\right\}$. We designed them in increasing order of complexity, i.e., $|\pi_n| < |\pi_{n+1}|$, and covering similar sizes (number of cells) in the range $\{1, \dots, 1.000\}$. In addition, we include a partition scheme that uses some prior knowledge of the model class (scale invariant, translation invariant, and rotation invariant) to produce more effective representations.

8.1.3 Estimation of Information and Operation losses

For a model $\mu_{X,Y}$ and a partition $\pi_i \in \mathcal{Q}(\mathcal{X})$, we consider $\eta_{\pi_i}(\cdot)$, $U_i = \eta_{\pi_i}(X)$ and the induced distribution of (U_i, Y), i.e., $\mu_{U_i,Y}$. For the partition, we have two scenarios: non data-driven π_i and data-driven $\pi_i(x_1 \dots x_m)$, where this last object is a function of the unsupervised sample $x_1 \dots x_m \in \mathcal{X}^m$ that follow the true marginal μ_X. For data-dependent partitions, the dependency on $x_1 \dots x_m$ will be implicit. Then, the cells of a partition will be denoted by $\pi_i = \{A_j, j = 1, .., k_i\}$ in all cases. In addition to the (unsupervised) sample $x_1 \dots x_m \in \mathcal{X}^m$ used to create the data-driven partitions, we use an independent set of supervised i.i.d. realizations of $(X, Y) \sim \mu_{X,Y}$ to estimate $\mathcal{I}(\mu_{X,Y})$, $\mathcal{I}(\mu_{U_i,Y})$, $\ell(\mu_{X,Y})$ and $\ell(\mu_{U_i,Y})$ by strongly consistent estimators.

40. Using Eq.(33) and a fixed sample size $m = 10.000$, we hand-selected different values of ℓ_n (the threshold) such that we achieve a representative collection of values $k_n = \left|\pi_n^G\right|$ in the range $\{1, \dots, 1.000\}$.
41. For a fixed sample size $m = 10.000$, we hand-selected ℓ_n (the threshold of the method) to cover a rich collection of values $k_n = \left|\pi_n^T\right|$ in the range $\{1, \dots, 1.000\}$.

To estimate the IL of U_i, we know that $\mathcal{I}(\mu_{X,Y}) = \mathbb{E}_{(X,Y)}\left\{\log \frac{\mu_{Y|X}(Y|X)}{\mu_Y(Y)}\right\} < \infty$ and, consequently, a natural empirical estimator (assuming knowledge of the model) is

$$\hat{\mathcal{I}}_n(\mu_{X,Y}) \equiv \frac{1}{n}\sum_{j=1}^{n} \log \frac{\mu_{Y|X}(Y_j|X_j)}{\mu_Y(Y_j)}. \tag{56}$$

Concerning the discrete model $\mu_{U_i,Y} \in \mathcal{P}([k_i] \times \mathcal{Y})$, we have that

$$\mathcal{I}(\mu_{U_i,Y}) = I(U_i;Y) = \sum_{j=1}^{k_i}\sum_{y\in\mathcal{Y}} \mu_{X,Y}(A_j \times \{y\}) \log \frac{\mu_{X,Y}(A_j \times \{y\})}{\mu_X(A_j)\cdot \mu_Y(\{y\})}. \tag{57}$$

Then, we can use the empirical version (known as the plug-in estimator) of (57) by[42]

$$\mathcal{I}(\hat{\mu}^n_{U_i,Y}) \equiv \sum_{j=1}^{k_i}\sum_{y\in\mathcal{Y}} \hat{\mu}^n_{X,Y}(A_j \times \{y\}) \log \frac{\hat{\mu}^n_{X,Y}(A_j \times \{y\})}{\hat{\mu}^n_X(A_j)\cdot \hat{\mu}^n_Y(\{y\})}, \tag{58}$$

where

$$\hat{\mu}^n_{X,Y}(A \times \{y\}) \equiv \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{A\times\{y\}}(X_j, Y_j).$$

Finally, our empirical estimation of the information loss of U_i is

$$\hat{\mathcal{I}}_n(\mu_{X,Y}) - \mathcal{I}(\hat{\mu}^n_{U_i,Y}). \tag{59}$$

From the law of large numbers, $\lim_{n\to\infty} \hat{\mathcal{I}}_n(\mu_{X,Y}) = \mathcal{I}(\mu_{X,Y})$ and $\lim_{n\to\infty} \mathcal{I}(\hat{\mu}^n_{U_i,Y}) = \mathcal{I}(\mu_{U_i,Y})$ with probability one. Therefore, our empirical estimation of $\mathcal{I}(\mu_{X,Y}) - \mathcal{I}(\mu_{U_i,Y})$ in (59) is strongly consistent.

For the operation loss of U_i, we determine the MPE decision $\tilde{r}_{\mu_{X,Y}}(\cdot)$ and $\tilde{r}_{\mu_{U_i,Y}}(\cdot)$ analytically as we know $\mu_{X,Y}$ and $\mu_{U_i,Y}$. Using the supervised i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from $\mu_{X,Y}$, we use the empirical risks:

$$\hat{\ell}_n(\mu_{X,Y}) \equiv 1 - \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{Y_j\}}(\tilde{r}_{\mu_{X,Y}}(X_j)) \tag{60}$$

$$\hat{\ell}_n(\mu_{U_i,Y}) \equiv 1 - \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{Y_j\}}(\tilde{r}_{\mu_{U_i,Y}}(\eta_{\pi_i}(X_j)), \tag{61}$$

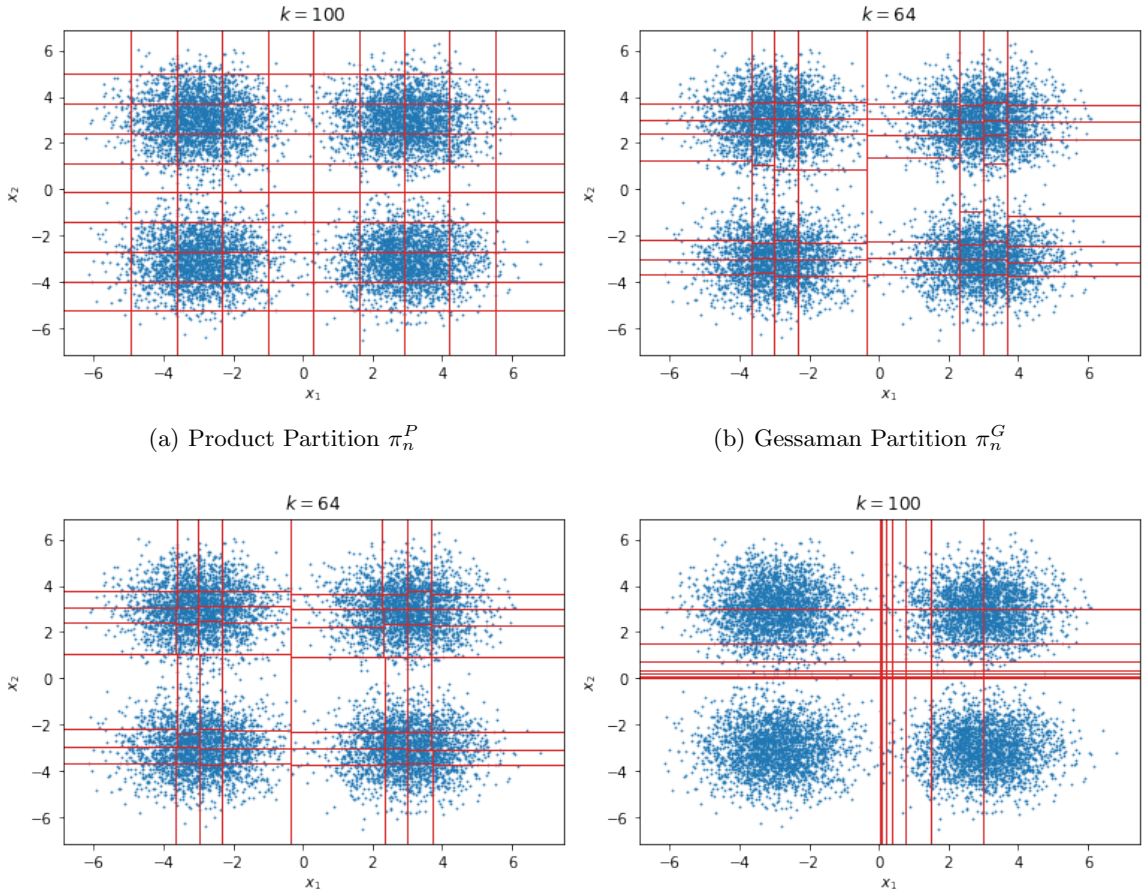
and the empirical operation loss of U_i is

$$\hat{\ell}_n(\mu_{U_i,Y}) - \hat{\ell}_n(\mu_{X,Y}). \tag{62}$$

As before, $\lim_{n\to\infty} \hat{\ell}_n(\mu_{U_i,Y}) - \hat{\ell}_n(\mu_{X,Y}) = \ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y})$ with probability one.

For the following sections, we use these consistent estimators to obtain precise indicators of the true losses. For that, we can select a sufficiently large sample size n for the computation of (59) and (62). For our analysis, we found that $n = 1.000.000$ realizations of $\mu_{X,Y}$ were sufficient to obtain accurate estimations of the information losses and operation losses in all the numerical examples described below.

42. In (58), we can update the classical plug-in estimador using the fact that we know that $\mu_Y(\{y\}) = 1/|\mathcal{Y}|$.



(a) Product Partition π_n^P

(b) Gessaman Partition π_n^G

(c) TSP Partition π_n^T

(d) Asymmetric Partition π_n^B

Figure 1: Illustration of the data-driven and non data-driven partitions used to obtain lossy representations (vector quantizers) of X. For the data-driven methods, we show the sample of μ_X used for this construction: the Scale Invariant Model in Section 8.1.1 with parameters $\alpha = 3$ and $\sigma = 1$.

8.2 Information Loss vs. Operation Loss

We begin with the scale invariant model. We consider the product partitions $\left\{\pi_n^P\right\}_n$, the Gessaman partitions $\left\{\pi_n^G\right\}_n$, and TSP $\left\{\pi_n^T\right\}_n$. In addition, we include a scheme that uses some prior knowledge of the task $\left\{\pi_n^B\right\}_n$. This scheme quantizes the space $\mathbb{R}^2$ with vertical and horizontal lines in an asymmetric way as illustrated in Fig.1d. These vertical and horizontal boundaries are added to increase the size of the partitions. In the limit (of infinite partition size) the added vertical and horizontal lines converge to the boundaries of the optimal OS partition $\tilde{\pi}_S$ presented in (55). For each of the four representation strategies, we produce a collection of partitions of different sizes (number of cells). Figure 1 illustrates how the i.i.d. samples for each of the four classes are distributed in the space and the four strategies adopted in this analysis.

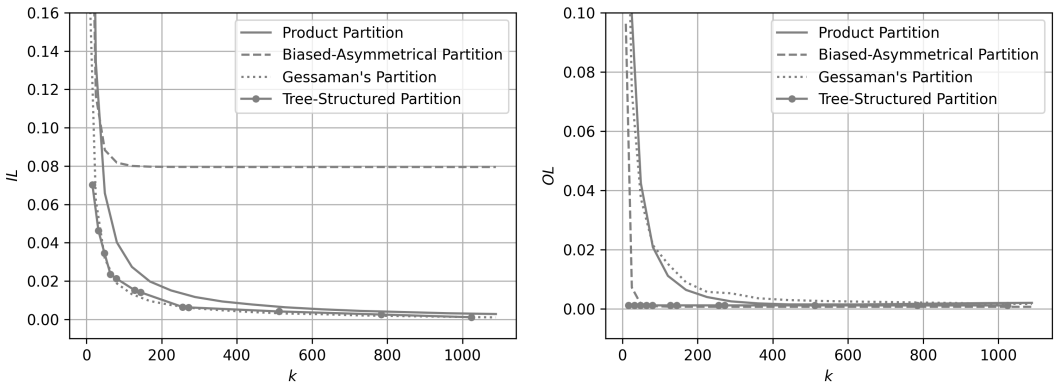


Figure 2: Information losses (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product (π_n^P), Gessaman (π_n^G), Tree-structured (π_n^T) and Biased-Asymmetrical (π_n^B).

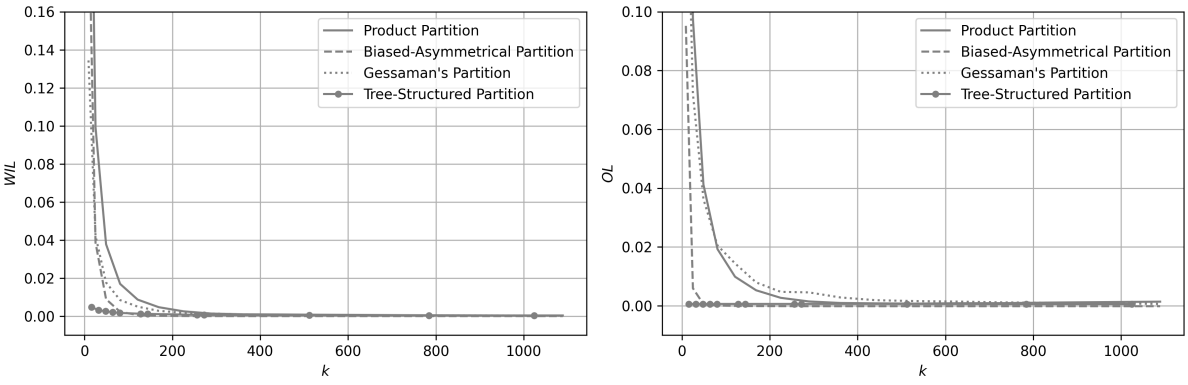


Figure 3: Weak information losses (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product (π_n^P), Gessaman (π_n^G), Tree-structured (π_n^T) and Biased-Asymmetrical (π_n^B).

Figure 2 presents the information losses (the left panel) and the operation losses (the right panel) across a range of partition sizes from 10 to $1,000$. The trends of all four curves follow the expected decreasing pattern: as the partition size increases so does the quality to represent the information that U_i has about Y. However, on the information loss side, there are two clear groups of curves. The first group shows a vanishing information loss (Product, Gessaman, and TSP) expressing the fact that these families are IS (Def. 2), which is consistent with the results elaborated in Section 5. For k in the range $[50 - 400]$, there are representation differences among these three cases. The data-driven schemes (TSP and Gessaman) offer better information losses than the product (non-adaptive) partition. This discrepancy is attributed the known flexibility and representation quality of data-driven partitions (Lugosi and Nobel, 1996; Silva and Narayanan, 2010b, 2012; Gonzales et al., 2022).

On the expressiveness of the Product, Gessaman, and TSP scheme in terms of operation loss (OL) (the right panel), the exact order observed in the information loss side is not preserved. Indeed, for partition size in the range $[50, 400]$, the product scheme (non-adaptive) shows marginally better OL than the Gessaman scheme (data-driven) in opposition of what is observed in the IL indicator. On the other hand, the TSP shows a clear advantage in OL compared with the other two methods (the Gessaman and the product partition) that is not explained by looking at the IL indicator. On explaining this last advantage, TSP construction captures with very few cells ($k < 10$) the optimal solution in (55), which is attributed to the symmetry observed in the unsupervised samples (see Fig.1) mimicking the symmetry of the optimal rule in (55).

Finally, if we look at the task-informed asymmetrical partitions $\left\{\pi_n^B\right\}$, the analysis is insightful. We observe that this family is not IS (left panel in Fig.2), however, on the operation loss (right panel of Fig.2) this family is OS. Furthermore, this scheme has an almost zero OL in all the size regimes (the only exception is when $k < 20$), which is clearly better than the Gessaman and the Product partitions in all regimes. This numerical finding is relevant considering that these last two schemes (Gessaman and Product) are IS. This analysis demonstrates how prior knowledge of the task (in this case the symmetry of the optimal solution) can be used to meet OS very efficiently without the need to achieve the conservative IS requirement. In addition, this asymmetric construction shows an scenario (based on prior knowledge of the task) where pure IL is not adequate as a predictor of the quality of representations if the objective is having a small predictive error.

Complementing this analysis, Fig.3 presents the WIL (i.e., $I(\tilde{U}; Y|U_i)$) vs. OL curves for the same four strategies (product, Gessaman, TSP, and asymmetrical) and settings used in Fig.2 for the IL vs. OL analysis.[43] From Figs. 2 and 3, the WILs are upper bounded by the ILs as expected. More importantly, the asymmetrical partitions $\left\{\pi_n^B\right\}$ meet the WIS condition (left panel of Fig.3), explaining their vanishing operation loss. This scenario shows again that WIS, as a condition, is strictly weaker than IS (a condition demonstrated in Section 3.4.1). Furthermore, the raking obtained in the WIL domain (left panel) is more consistent with the order followed by the OL indicator, especially in the regime where $k \leq 100$. In particular, this is observed in the relative order given to (π_n^P), (π_n^T) and (π_n^B) from one domain (WIL) to the other (OL), which is in clear contrast with what is observed

43. To compute the WILs in Fig.3, we use the OS rule $\tilde{r}_S(\cdot)$ induced by the partition in (55). Then, we estimate $I(\tilde{U}, U_i; Y) - I(U_i; Y)$ using the consistent MI estimation $\mathcal{I}(\hat{\mu}^n_{(\tilde{U},U_i),Y}) - \mathcal{I}(\hat{\mu}^n_{U_i,Y})$ from (58).

in Fig. 2 for the same three schemes. Overall, the WILs offer better predictions of the quality (prediction error) of the representations.

8.2.1 Rotated version of the Scale Invariant Model

To make the previous numerical setting a bit more challenging for the partitions schemes that are coordinate oriented, we consider a rotated version of the scale invariant model. The samples for this rotated model and the obtained partitions used for this analysis are illustrated in Fig. 4. We consider the same family of data-driven partitions (TSP and Gessaman) and non-data driven partitions (Product and Asymmetric). All partitions (with the exception of the biased-asymmetrical that used prior knowledge of the orientation of $\tilde{\pi}_S$) are produced in a coordinate based manner.

Fig. 5 shows the information loss and the operation loss curves. These curves show again two groups of decreasing curves on the information loss side: three schemes are IS (Gemmanan, TSP and Product), while the asymmetric scheme $\left\{\pi_n^B\right\}$ does not show a vanishing information loss. In contrast, not only $\left\{\pi_n^B\right\}$ has a vanishing operation loss but is the scheme with the best operational performance (see the right panel in Fig.5). Consequently, as in the previous example, the information loss is not adequate predictor of the ranking in operation loss among these four schemes. More critically, the IL indicator is blind in expressing the vanishing operation loss of $\left\{\pi_n^B\right\}$. In contrast, Fig.6 shows the counterpart of Fig.5 using instead the WIL indicator. As in the previous example, $\left\{\pi_n^B\right\}$ meet the WIS condition explaining their vanishing OL (from Theorem 24). This case is another example that demonstrates that WIS is strictly weaker than IS as a condition. Overall, the relative order provided by WIL indicators offers a very accurate prediction of the quality of the representations (seen in the right panel of Fig.6).

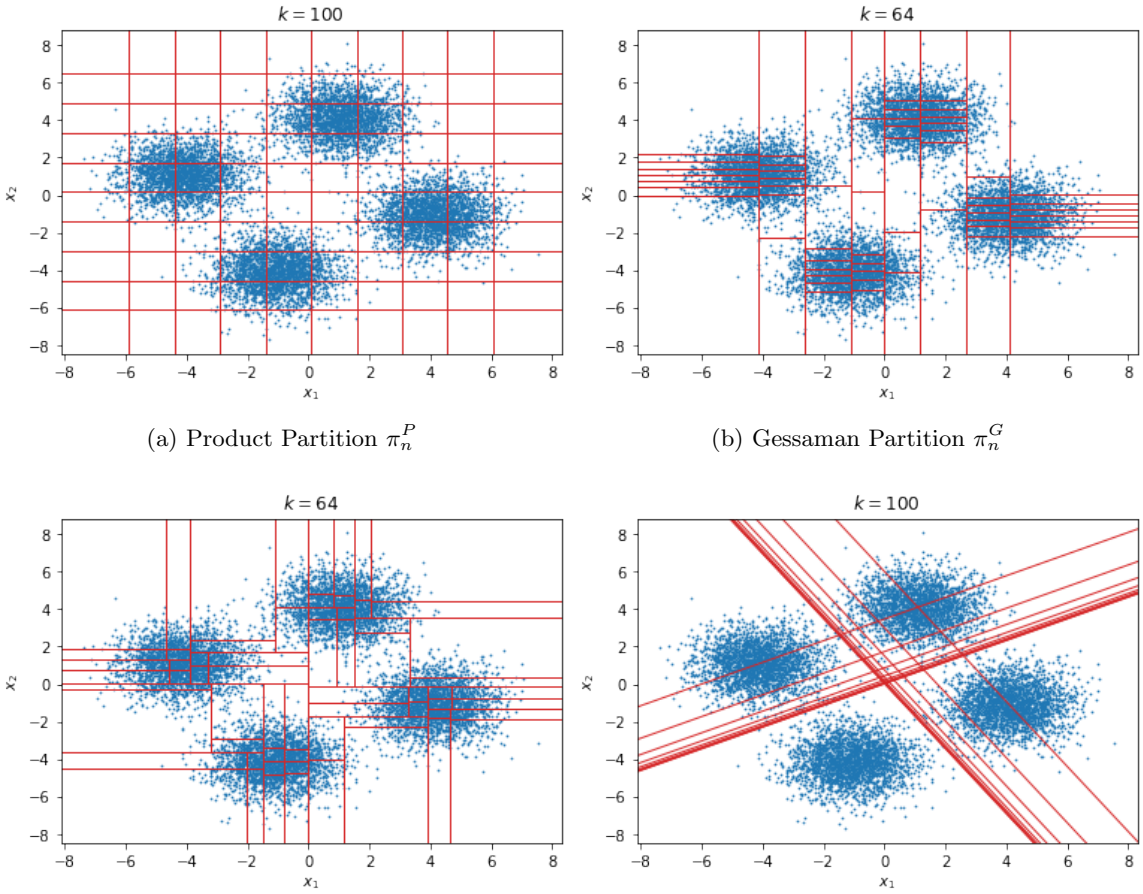
As a side comment, we observe that achieving a close to zero operation loss in this case is quite more difficult than what is observed on the previous non-rotated example as we anticipated. Indeed, the data-driven methods achieve close to zero operation loss only after $k > 800$ in clear contrast with what is presented in the right panel of Fig. 2.

8.2.2 Translation Invariant Model

In this case, we consider a rotated version of the *Translation Invariant Model* introduced in Section 8.1.1 to make the problem non-coordinate oriented and more challenging. The samples of this model are presented in Fig.7. In addition to the partitions presented in Section 8.1.2, we include a partition that project the data in the direction of symmetry of the problem (1D projection) to then perform a uniform quantization in this projected scalar domain. This informed (with prior knowledge of the task) representation of X is illustrated in Fig.7d and denoted by π_n^S.

As this problem is translation invariant, there is a linear lossy mapping[44] $\eta : \mathbb{R}^2 \to \mathbb{R}$ that is operationally sufficient for $\mu_{X,Y}$, in the sense that $\ell(\mu_{X,Y}) = \ell(\mu_{\eta(X),Y})$ (see Corollary 29). Because of this operational structure, we consider the projected information loss $I((\eta(U), U_i); Y) - I(U_i; Y) = I(\eta(U); Y|U_i)$ for our analysis (see Eq.43) to evaluate if this projected indicator (from Theorem 24) predicts the operation loss in this task.

44. $\eta(\bar{x}) = a_1 \cdot x_1 + a_2 \cdot x_2$, where the vector $\bar{a} = (a_1, a_2)^t \in \mathbb{R}^2$ is orthogonal to the direction used to locate the means of the conditional Gaussian distribution (i.e., $\mu_{X|Y}(\cdot|y)$, $y \in \mathcal{Y}$) of the model $\mu_{X,Y}$.



(a) Product Partition π_n^P

(b) Gessaman Partition π_n^G

(c) TSP Partition π_n^T

(d) Asymmetric Partiton π_n^B

Figure 4: Illustration of the data-driven and non data-driven partitions used to obtain lossy representations (vector quantizers) of X. For the data-driven methods, we show samples of the distribution used for this analysis: a rotated version of the Scale Invariant Model in Section 8.1.1 with parameters $\alpha = 3$ and $\sigma = 1$.

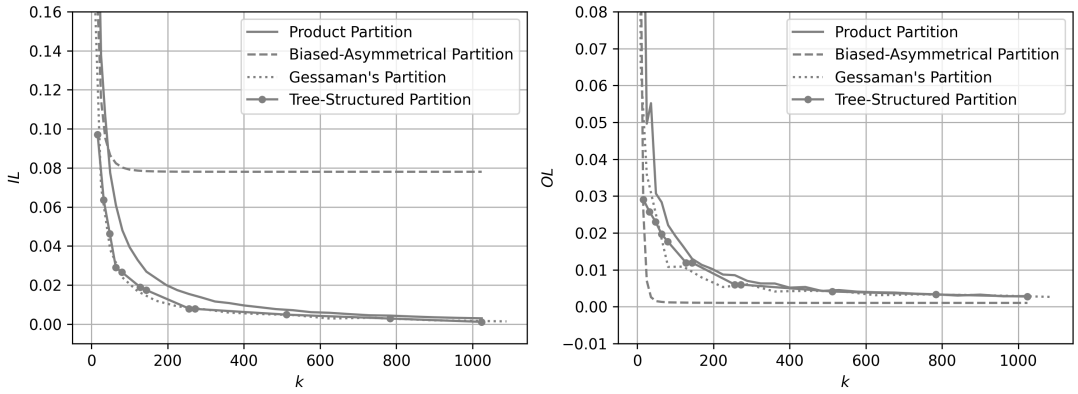


Figure 5: Information loss (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product (π_n^P), Gessaman (π_n^G), Tree-structured (π_n^T) and Biased-Asymmetrical (π_n^B).

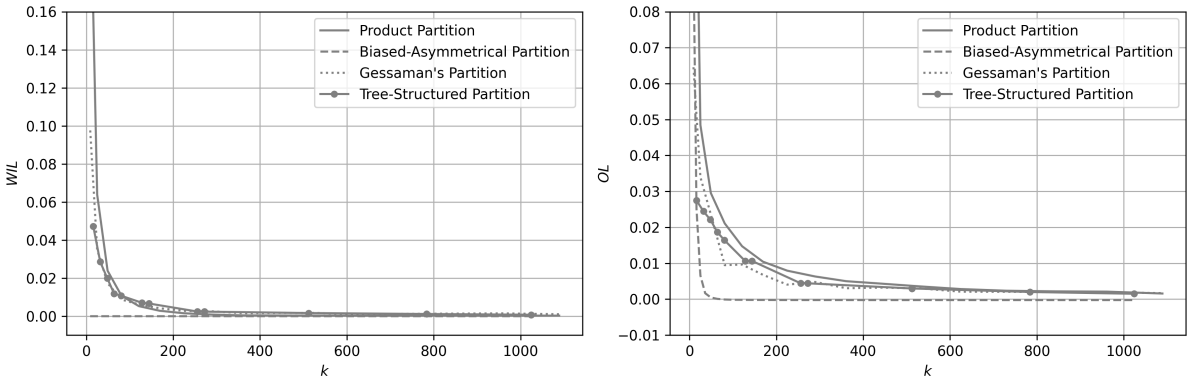
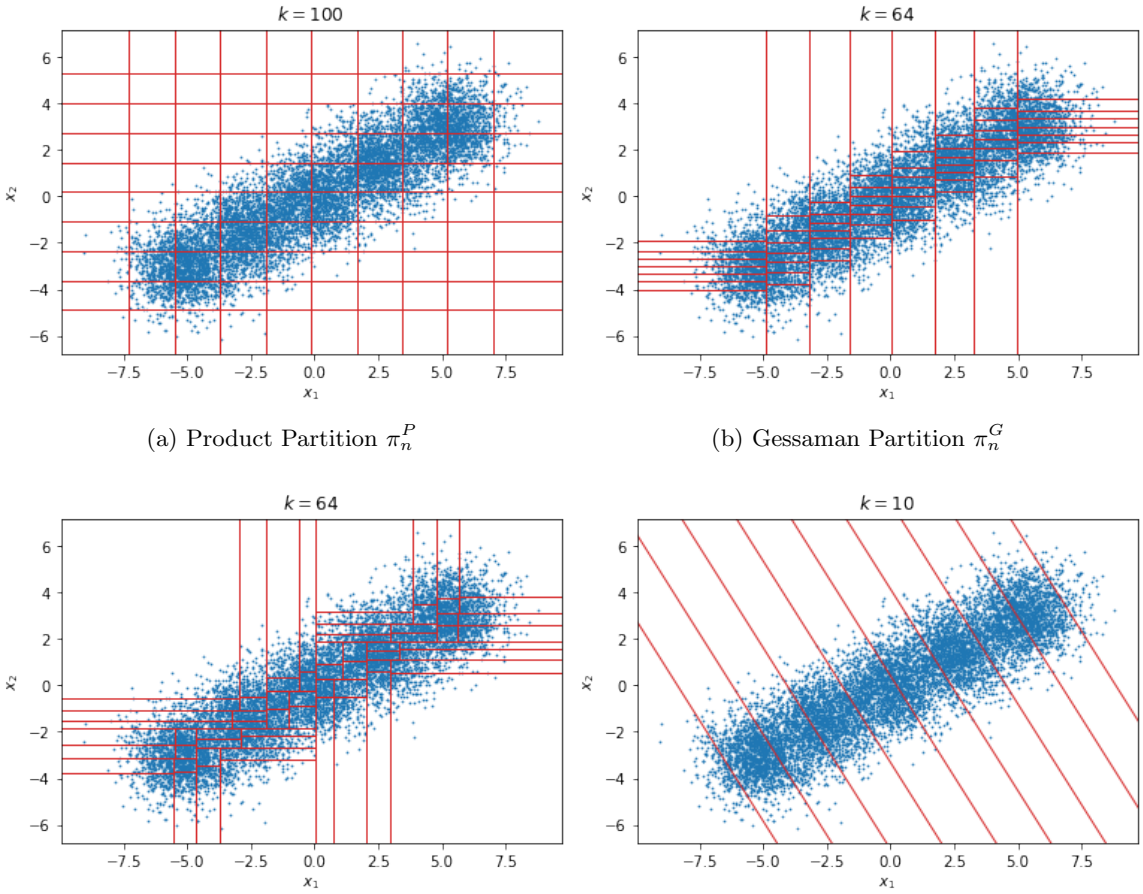


Figure 6: Weak Information loss (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product (π_n^P), Gessaman (π_n^G), Tree-structured (π_n^T) and Biased-Asymmetrical (π_n^B).



(a) Product Partition π_n^P

(b) Gessaman Partition π_n^G

(c) TSP Partition π_n^T

(d) 1D-Uniform Partiton π_n^S

Figure 7: Illustration of the data-driven and non data-driven partitions used to obtain lossy representations (vector quantizers) of X. For the data-driven methods, we show samples of the distribution used for this analysis: a rotated version of the Translation Invariant Model in Section 8.1.1.

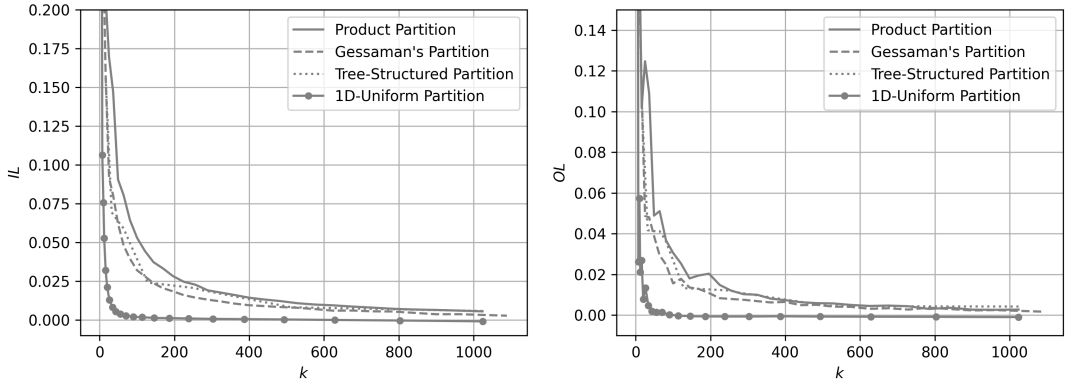


Figure 8: Projected (task-depedent) information loss (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product, Gessaman, Tree-structured and 1D-Uniform projection.

Fig.8 shows the task-dependent information loss $I(\eta(U);Y|U_i)$ and the operation loss for the four schemes. In the information loss, all schemes perform very well in capturing the information that $\eta(X)$ has about Y and they offer the expected monotonic and vanishing decreasing pattern. In particular, data-driven partitions are slightly better than the product one (that is consistent with their more efficient adaptation illustrated in Fig. 7). Importantly, the projected partition (that use the operationally sufficient representation $\eta(\cdot)$ in its construction) takes advantage of this prior knowledge, which is expressed in a clearly superior performance in the information loss. In fact, the decreasing trend of its loss curve is very drastic showing that this representation is very effective in capturing the latent information structure of $I(\eta(X);Y)$. For example, the informed scheme (1D-Uniform) achieves a loss with 10 cells that is obtained with the TSP and Gessaman (partitioning the whole space) with more than 500 cells. Therefore, the gain of using a prior knowledge of the learning task (i.e., projecting the problem to a smaller dimension to design π_n^S) is significant in terms of the projected information loss $I(\eta(U);Y|U_i)$. Importantly, these patterns are preserved in the operation loss domain (right panel of Fig.8). Here, the drastic difference in performance of the 1D-Uniform representation with respect to the rest is also observed. Also the order, or the ranking, from the less informative to the more informative representation is consistently observed in the operation side (right panel), showing for this example that $I(\eta(U);Y|U_i)$ provides a good prediction of the relative (operational) performance of the schemes.

8.2.3 Rotation Invariant Model

Finally, we consider the rotation invariant model introduced in Section 8.1.1 that is illustrated in Fig. 9. This problem is very challenging for the representations in Section 8.1.2 that are coordinate oriented. As in the previous example, this model has a 1D projection that is operationally sufficient[45] $\eta:\mathbb{R}^2\to\mathbb{R}$, meaning that $\ell(\mu_{X,Y})=\ell(\mu_{\eta(X),Y})$. Here, we

45. $\eta(\bar{x})=||\bar{x}||^2=x_1^2+x_2^2$, where $\bar{x}=(x_1,x_2)^t\in\mathbb{R}^2$.

also consider the projected information loss for our analysis, i.e. $I(\eta(U); Y|U_i)$, and a specific representation (1D-Uniform) induced by projecting X using $\eta(\cdot)$ and then performing a uniform quantization in this scalar domain (see Fig.9d). These collection of partitions including this last informed partition scheme (1D-Uniform) are illustrated in Fig.9.

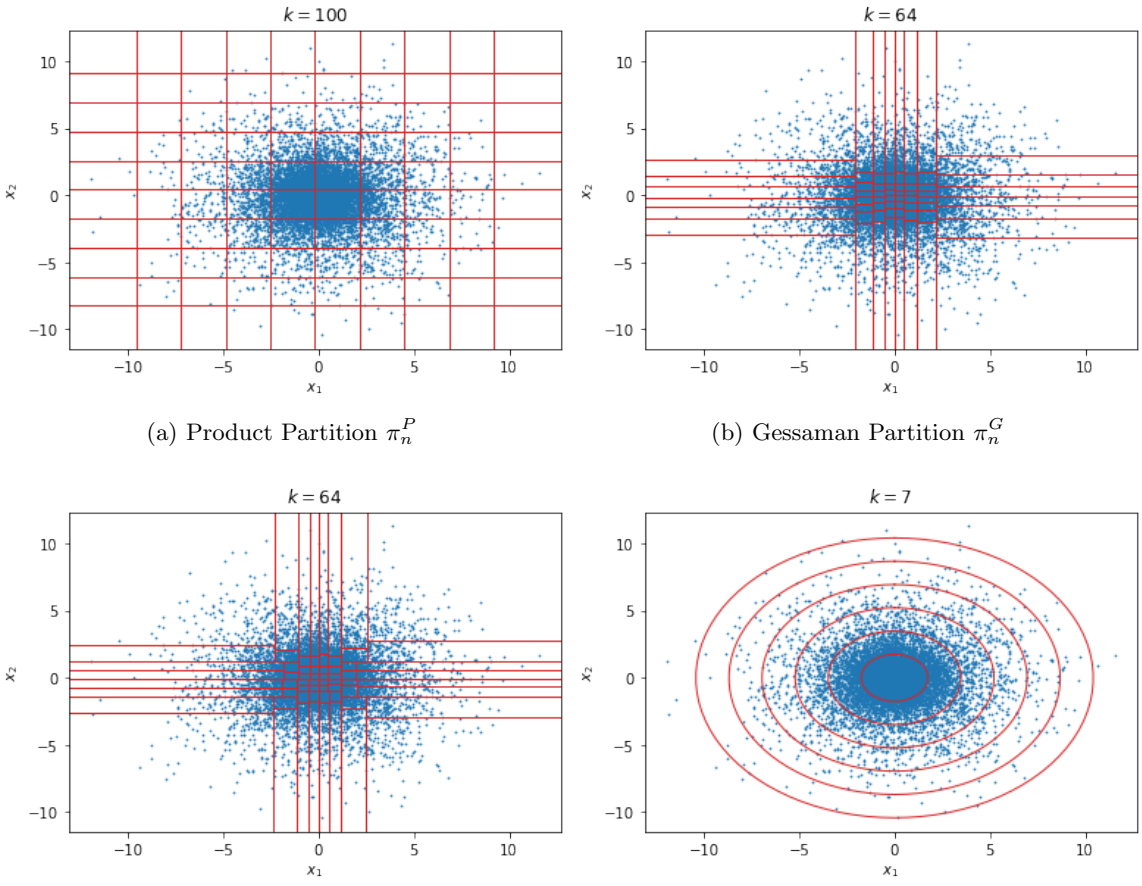
Fig.10 shows the curves associated to the projected information loss and the operation loss. As in the previous case, the difference in information loss for the projected scheme is significant, meaning that prior knowledge of the task translates in a significant boost in expressiveness. In this scenario, the discrepancy is even more drastic than the previous example in Section 8.2.2, where the 1D-Uniform representation has a projected information loss with 10 cells that is better than what all the other alternative partitions can achieve with $1,000$ cells, which is a very impressive difference. As in the previous example in Section 8.2.2, these differences translate in the operation loss showing a clear advantage of the 1D-Uniform with respect to the rest.

9. Final Discussion

This work offers new results that shed light on the interplay between information loss (in the Shannon sense) and operation loss (in the classical MPE sense) when considering a general family of lossy continuous representations of an observation vector X in $\mathbb{R}^d$. Our main asymptotic result (Theorem 12) supports the idea that creating a family of information sufficient representations is adequate in the sense that these representations have a vanishing residual error with respect to the best decision acting on X to classify Y. At the same time, Theorem 12 shows that pure IS (in the sense of Def.2) is a conservative criterion. Indeed, Theorems 24 demonstrates that a weaker notion of IS suffices to obtain the required operational result (OS).

We worked on a non-oracle extension of Theorem 12 adopting a learning setting where $\mu_{X,Y}$ belongs to a class of models $\Lambda \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. We studied how the structure of Λ can be modeled operationally using sub-sigma fields, and we connected this structure with the existence of a lossy OS transformation for the task. We use this operational structure as prior knowledge to propose a less conservative and non-oracle-weak form of informational sufficiency that implies OS. This new non-oracle result is stated in Theorem 24. In the application of Theorem 24, we look at two important families of models: the operationally invariant models (invariant to the action of a compact group) (Bloem-Reddy and Teh, 2020) and finite-size models (Xu and Mannor, 2012). In these two relevant contexts, it was possible to determine "*a non-oracle*" lossy surrogate of $\tilde{r}_{\mu_{X,Y}}(\cdot)$ (in Theorem 12) that extends our main result (WIS implies OS) in a realistic learning-like setting. Some applications of this result are discussed and presented in Section 6.4. To conclude, in Section 7, we demonstrate that two existing compression-based learning algorithms have the expressive power to achieve lossless prediction (OS) by applying Theorem 24 (WIS implies OS).

Complementing these results, our empirical analysis in Section 8 supports our theoretical findings, verifying with some concrete examples that the proposed WIS condition is strictly weaker than IS. In these examples, we validate that pure information loss is not always an adequate predictor of operation loss in classification and, therefore, that looking only at mutual information as a fidelity indicator could be misleading in some contexts. In this regard, our theoretical results and supported numerical evidence are consistent with



(a) Product Partition π_n^P

(b) Gessaman Partition π_n^G

(c) TSP Partition π_n^T

(d) 1D-Uniform Partiton π_n^R

Figure 9: Illustration of the data-driven and non data-driven partitions (with $k = 100$) used to obtain lossy representations (vector quantizers) of X. For the data-driven methods, we show samples of the distribution used for this analysis: the Rotation Invariant Model in Section 8.1.1.

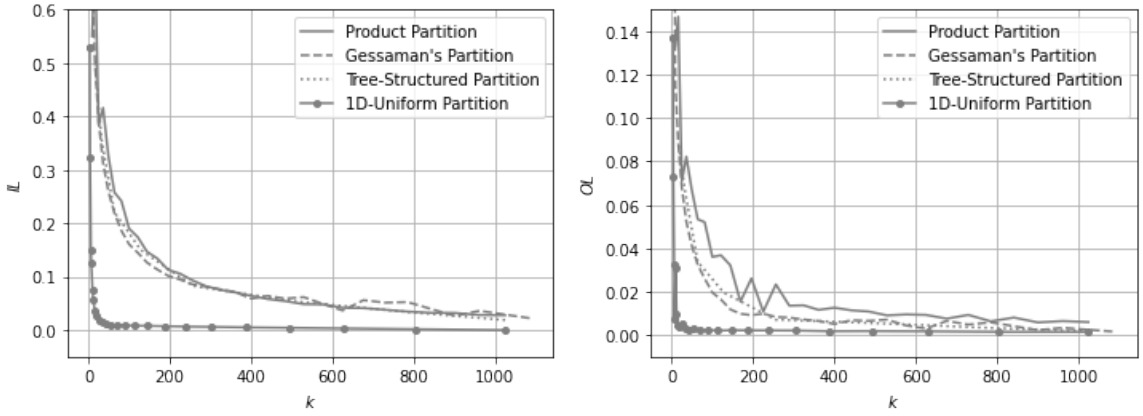


Figure 10: Projected information loss (the left panels) and operation losses (the right panel) curves for different number of cells (k) of the partitions: Product, Gessaman, Tree-structured and 1D-Uniform projection.

some of the findings made in feature selection (Frenay et al., 2013). Finally, our empirical findings show that in the presence of prior knowledge about the task, in the form studied in Section 6 and formalized in Theorem 24, this operational knowledge can be used to design representations that offer significantly better operational performances.

9.1 Applications in ML

Studying the interplay between vanishing information and operation loss was motivated by our desire to understand the role of lossy compression in machine learning. Our two main results (Theorems 12 and 24) shed light on this theoretical dimension. Furthermore, our theoretical findings do offer new insights for the design of ML methods (see Section 7.1) and provide formal arguments to explain the appropriateness and expressiveness of existing encoder strategies (see Section 5) and compression-based learning algorithms (see Section 7). On this last dimension, we can say that:

- Our results support the universality of approximating (or learning) compressed representations that capture the mutual information between X and Y, for example, via minimization of the conditional entropy $H(Y|U)$, or maximization of $I(U;Y)$ as illustrated in Section 7 for the IB method. This info-max principle is widely adopted in representation learning, in the form of maximizing empirical versions of the mutual information (info-max problems) or minimizing empirical versions of the conditional entropy over a family of encoders of X (Amjad and Geiger, 2019; Alemi et al., 2017; Achille and Soatto, 2018b; Strouse and Schwab, 2017; Tegmark and Wu, 2019).

- Our two main results (on WIS implies OS) also show that IS is a conservative criterion if the objective is designing expressive representations in the MPE sense. Indeed, we introduce a weaker (strictly weaker in some cases) information-sufficient condition that implies OS. Importantly, our non-oracle result in Theorem 24 (WIS implies OS) could be adopted for the analysis of learning problems as shown in Section 6.4 and Section

7. Indeed, two existing compression-based algorithms are studied in Section 7. These solutions can meet our *non-oracle WIS* criterion in Theorem 24 and, consequently, we demonstrate their expressive power to achieve lossless prediction (OS): Theorem 31. In light of these results, we believe that our findings, analyses, and theoretical results could be used to motivate new avenues of practical research in ML: for instance, designing new "non-oracle" losses inspired by WIS, weaker than pure IS that could be adopted to rank, select or optimize representations from data when some prior knowledge of the task is available beforehand.

- Finally, on the use of our raw oracle result in Theorem 12, which is purely theoretical, we show in Section 5 that this result facilitates an exciting connection with the problem of mutual information estimation. Not only that, but Theorem 12 is used to demonstrate the universal expressive power of digital representations (VQs) for classification (see Theorems 15 and 18). This is an important result for ML and representation learning in particular, saying that digitalization can be used with vanishing performance degradation, which justifies the adoption of lossy compression ideas and methods in ML (Strouse and Schwab, 2017; Tegmark and Wu, 2019; Goldfeld and Polyanskiy, 2020; Zaidi et al., 2020; Dubois et al., 2021). On this, it is worth emphasizing the role of data-driven partitions (Silva and Narayanan, 2010a, 2007, 2010b; Vajda, 2002; Darbellay and Vajda, 1999; Gonzales et al., 2022). We establish a condition that makes these stochastic encoders IS with probability one (Theorem 18). Then, Theorem 12 is used to prove that these data-driven VQs are OS (with probability one) for classification, as presented in Corollaries 19 and 20 for two widely used data-driven constructions.

10. Proofs of the Main Results

10.1 Proof of Theorem 10

Proof Let us first look at the definition of $g(\mu_{X,Y}, B)$ in (14). This is a function of the model $\mu_{X,Y}$, the partition $\tilde{\pi} = \left\{\tilde{A}_y, y \in \mathcal{Y}\right\}$ in (11) and a set $B \subset \mathcal{X}$. In particular, we have that

$$g(\mu_{X,Y}, B) = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B)\right] - \sum_{\tilde{A}_u \in \tilde{\pi}} \mu_X(\tilde{A}_u|B) \cdot \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|\tilde{A}_u \cap B)\right]. \quad (63)$$

The first term on the RHS of (63) can be seen as the prior minimum error probability of a random variable $\tilde{Y}$ in $\mathcal{Y}$ with marginal probability $(v_{\tilde{Y}}(y) \equiv \mu_{Y|X}(y|B))_{y\in\mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$. On the other hand, the second term on the RHS of (63) can be seen as the MPE of a joint vector $(\tilde{X}, \tilde{Y})$ in $\mathcal{Y} \times \mathcal{Y}$ with probability $v_{\tilde{X},\tilde{Y}}$ in $\mathcal{P}(\mathcal{Y} \times \mathcal{Y})$ defined by

$$v_{\tilde{X},\tilde{Y}}(u, y) \equiv \frac{\mu_{X,Y}(\tilde{A}_u \cap B \times \{y\})}{\mu_X(B)}, \ \forall (u, y) \in \mathcal{Y}^2. \quad (64)$$

The second term in (63) is precisely $\ell(v_{\tilde{X},\tilde{Y}})$. Adopting Lemma 8 in this context, we can use its corollary in (17) to obtain that

$$\begin{aligned}
\mathcal{I}(v_{\tilde{X},\tilde{Y}}) = I(\tilde{X};\tilde{Y}) &\geq H(\tilde{Y}) - \mathcal{H}(\mathcal{R}(v_{\tilde{Y}}, \ell(v_{\tilde{X},\tilde{Y}}))) \\
&= \mathcal{H}(v_{\tilde{Y}}) - \mathcal{H}(\mathcal{R}(v_{\tilde{Y}}, \ell(v_{\tilde{X},\tilde{Y}}))) \\
&= \mathcal{H}(\mu_{Y|X}(\cdot|B)) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B), \ell(\mu_{\tilde{X},\tilde{Y}}))), \quad (65)
\end{aligned}$$

where $\ell(v_{\tilde{X},\tilde{Y}}) = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B)\right] - g(\mu_{X,Y}, B)$ from (63) and the construction of $v_{\tilde{X},\tilde{Y}}$ in (64). The inequality in (65) is obtained as a function of $B \subset \mathcal{X}$, as it is used to construct $v_{\tilde{X},\tilde{Y}}$ in (64).

Returning to the main object of interest of this result, we have that

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = I(\tilde{U};Y|U_i) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot I(\tilde{U};Y|X = B_{i,j}). \quad (66)$$

The first equality is by the chain rule of MI and the second is by definition of the conditional MI (Cover and Thomas, 2006). Finally we recognize that $I(\tilde{U};Y|X = B) = \mathcal{I}(\mu_{\tilde{U};Y|X}(\cdot|B))$, where $\mu_{\tilde{U};Y|X}(\cdot|B)$ is precisely the distribution $v_{\tilde{X},\tilde{Y}}$ defined in (64). Consequently, applying (65) in each $B_{i,j} \in \pi_i$, we have that

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) \geq \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot \left[\mathcal{H}(\mu_{Y|X}(\cdot|B_{i.j})) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B_{i,j}), \epsilon_{i,j}))\right],$$

where $\epsilon_{i,j} = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - g(\mu_{X,Y}, B_{i,j})$. ■

10.2 Proof of Theorem 12: from Discrete to Continuous Representations

10.2.1 Organization of the Proof

The proof of Theorem 12 is divided in many stages. The first stage, presented in Section 10.2.2, restricts the analysis to the case of finite alphabet representations to prove Theorem 35 below. In the second stage, in Section 10.2.3, we make a connection between the discrete and the continuous version of this problem. To conclude, the finite alphabet result (Theorem 35) is used as a building block to prove Theorem 12.

10.2.2 Discrete version of Theorem 12

Theorem 35 *Let $\{U_i\}_{i\geq 1}$ be a sequence of representations for X obtained from $\{\eta_i(\cdot)\}_{i\geq 1}$ where $|\mathcal{U}_i| < \infty$ for any $i \geq 1$. If $\{U_i\}_{i\geq 1}$ is WIS for $\mu_{X,Y}$ then $\{U_i\}_{i\geq 1}$ is OS for $\mu_{X,Y}$.*

The proof of Theorem 35 is presented in Section 10.3.

Technical remarks about the proof of Theorem 35:

1. The proof of this result uses a sample-wise version of the inequality presented in (18) (Theorem 10) as a key element in the argument.

2. Another important technical element of the proof was characterizing and analyzing the following information object:

$$\mathcal{I}_{loss}(\epsilon, M) \equiv \min_{v\in\mathcal{P}^{\epsilon}([M])} \left\{\mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon))\right\}, \tag{67}$$

   where $\mathcal{P}^{\epsilon}([M]) \equiv \{v \in \mathcal{P}([M]), prior(v) \geq \epsilon\}$ and $M = |\mathcal{Y}|$. Indeed, a non-trivial part of this argument was to prove that $\mathcal{I}_{loss}(\epsilon, M) > 0$ for some values of $\epsilon > 0$ (see Theorem 37 and Appendix 10.4). To achieve this key result, we derived an explicit lower bound for $\mathcal{I}_{loss}(\epsilon, M)$ function of ϵ and M.

10.2.3 Proof of Theorem 12

Proof Without loss of generality, let us assume that $\eta_i : \mathcal{X} \to \mathcal{U}_i$ is such that $\mathcal{U}_i \subset \mathcal{U} = \mathbb{R}^{\mathfrak{q}}$ for some $q \geq 1$[46]. Here we use a result from the seminal work of Liese et al. (2006) on asymptotic sufficient partition for MI. In particular, in the context of our work we have the following:

Lemma 36 *(Liese et al., 2006) There is an infinite collection of finite-size embedded partitions $\pi_1 \ll \pi_2 \ldots \subset \mathcal{B}(\mathbb{R}^q)$ of $\mathcal{U} = \mathbb{R}^q$ such that for any model $\mu_{X,Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ and any measurable function $\eta : \mathcal{X} \to \mathcal{U}$ it follows that*

$$\lim_{i\to\infty} I(Y; m_{\pi_i}(\eta(X))) = I(Y; \eta(X)), \tag{68}$$

where

$$m_{\pi_i}(u) \equiv \sum_{A_l\in\pi_i} l \cdot \mathbf{1}_{A_l}(u) \in \{1, .., |\pi_i|\}, \ \forall u \in \mathcal{U} \tag{69}$$

denotes the lossy function (VQ) induced by the partition $\pi_i = \{A_i, i = 1, .., |\pi_i|\}$.

46. The general case derives directly from the argument presented for this case, and it only requires the introduction of additional notations that occludes the proof flow.

Lemma 36 is a remarkable implication of the work by Liese et al. (2006). This result shows the existence of a finite-size VQ family that approximates (universally) any well-defined MI on a continuous space in the sense presented in (68). More details of this result and the construction of $\{\pi_i\}_{i\geq 1}$ are presented in Section 5.1.1.

In the context of this argument, we can use the universal embedded quantization $\{\pi_i\}_{i\geq 1}$ of $\mathcal{U}$ stated in Lemma 36 to obtain as a direct corollary of Lemma 36 that for any $\eta_j : \mathcal{X} \to \mathcal{U}$

$$\lim_{i\to\infty} I((\tilde{U}, m_{\pi_i}(U_j)); Y) - I(m_{\pi_i}(U_j); Y) = I((\tilde{U}, U_j); Y) - I(U_j; Y) = I(\tilde{U}; Y|U_j) \geq 0, \quad (70)$$

where $U_j = \eta_j(X)$ and $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X) \in \mathcal{Y}$ (see Eq.10).

On the other hand, from the hypothesis that assumes that $\{\eta_j(\cdot)\}_{j\geq 1}$ is WIS, we have that

$$\lim_{j\to\infty} I(\tilde{U}; Y|U_j = \eta_j(X)) = 0. \quad (71)$$

Let us consider an arbitrary sequence $(\epsilon_n)_{n\geq 1} \in \mathbb{R}^+ \setminus \{0\}$ such that $\epsilon_n \to 0$ as n tends to infinity. Using (70), we have that for any $j \geq 1$ there exists $i_j^*(\epsilon_j, \eta_j) \geq 1$ sufficiently large such that[47]

$$I(\tilde{U}; Y|U_j) + \epsilon_j > \underbrace{I((\tilde{U}, m_{\pi_{i_j^*}}(U_j)); Y) - I(m_{\pi_{i_j^*}}(U_j); Y)}_{I(\tilde{U}; Y|m_{\pi_{i_j^*}}(U_j)))} > I(\tilde{U}; Y|U_j) - \epsilon_j. \quad (72)$$

In (72), it is worth noticing that $m_{\pi_{i_j^*}}(U_j) = m_{\pi_{i_j^*}} \circ \eta_j(X)$. Then, we can define

$$\tilde{\eta}_j \equiv m_{\pi_{i_j^*}} \circ \eta_j : \mathcal{X} \to \left\{1, .., \left|\pi_{i_j^*}\right| < \infty\right\}, \quad (73)$$

which is a finite alphabet representation (vector quantization) of X. Therefore using $\{\eta_j(\cdot)\}_{j\geq 1}$ and $(\epsilon_n)_{n\geq 1}$, we have constructed a family of finite alphabet lossy representations of X, which we denoted by $\{\tilde{\eta}_j(\cdot)\}_{j\geq 1}$ in (73), satisfying that

$$\lim_{j\to\infty} I(\tilde{U}; Y|\tilde{\eta}_j(X))) = 0, \quad (74)$$

from (72), (71), and the fact $(\epsilon_n)_{n\geq 1}$ is $o(1)$. Therefore, (74) means that $\{\tilde{\eta}_j(\cdot)\}_{j\geq 1}$ is *weakly information sufficient* (Def.3). Then, Theorem 35 implies that

$$\lim_{j\to\infty} \ell(\mu_{\tilde{\eta}_j(X),Y}) = \ell(\mu_{X,Y}). \quad (75)$$

Finally, by construction, we have that $\tilde{\eta}_j(X) = m_{\pi_{i_j^*}} \circ \eta_j(X)$. Then, $\tilde{\eta}_j(X)$ is indeed a deterministic function of $\eta_j(X)$ for any j. Therefore, from classical results on Bayes decision $\ell(\mu_{\tilde{\eta}_j(X),Y}) \geq \ell(\mu_{\eta_j(X),Y})$, which concludes the proof from (75). ■

47. For what follows, we omitted the dependency on ϵ_j, η_j in i_j^* to simplify the notation.

10.3 Proof of Theorem 35

Let us begin introducing some preliminaries that will be used in the main argument in Section 10.3.2.

10.3.1 Preliminaries

Let us consider a finite alphabet representation $\eta : \mathcal{X} \to \mathcal{U}$ where $|\mathcal{U}| < \infty$. Using the expressions presented in Propositions 4 and 5 and the interplay between them, determined in Theorem 10, we define the *information loss density* (ILD) and the *operation loss density* (OLD) associated with $\eta(\cdot)$ as follows:

$$\ell_\eta(x) \equiv \sum_{A \in \pi_\eta} \mathbf{1}_A(x) \cdot g(\mu_{X,Y}, A) \geq 0, \ \forall x \in \mathcal{X} \tag{76}$$

$$\mathcal{I}_\eta(x) \equiv \sum_{A \in \pi_\eta} \mathbf{1}_A(x) \cdot I(\tilde{U}; Y|X \in A) \geq 0, \ \forall x \in \mathcal{X}. \tag{77}$$

It is useful to denote by $\pi_\eta(x)$ the cell in π_η that contains $x \in \mathcal{X}$. Using this notation, we have that $\ell_\eta(x) = g(\mu_{X,Y}, \pi_\eta(x))$ and $\mathcal{I}_\eta(x) = I(\tilde{U}; Y|X \in \pi_\eta(x))$. The names of $\ell_\eta(\cdot)$ and $\mathcal{I}_\eta(\cdot)$ come from the observation that

$$\mathbb{E}_X \{\ell_\eta(X)\} = \ell(\mu_{U,Y}) - \ell(\mu_{X,Y}) \tag{78}$$

$$\mathbb{E}_X \{\mathcal{I}_\eta(X)\} = \mathcal{I}(\mu_{(\tilde{U},U),Y}) - \mathcal{I}(\mu_{U,Y}), \tag{79}$$

where $U = \eta(X)$.

From the proof of Theorem 10, we obtain the following sample-wise inequality: for any $A \in \mathcal{B}(\mathcal{X})$

$$I(\tilde{U}; Y|X \in A) \geq \mathcal{H}(\mu_{Y|X}(\cdot|A)) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|A), \mathit{prior}(\mu_{Y|X}(\cdot|A)) - g(\mu_{X,Y}, A))), \tag{80}$$

where $\mathit{prior}(\mu_Y) \equiv (1 - \max_{y \in \mathcal{Y}} \mu_Y(y))$ denotes the prior risk of a prior model $\mu_Y \in \mathcal{P}(\mathcal{Y})$. Adopting this inequality, it follows that for any $x \in \mathcal{X}$

$$\mathcal{I}_\eta(x) \geq \mathcal{H}(\mu_{Y|X}(\cdot|\pi_\eta(x))) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_\eta(x)), \mathit{prior}(\mu_{Y|X}(\cdot|\pi_\eta(x))) - \ell_\eta(x))). \tag{81}$$

Then the ILD $\mathcal{I}_\eta(x)$ is lower bounded by a function of the posterior model $\mu_{Y|X}(\cdot|\pi_\eta(x)) \in \mathcal{P}(\mathcal{Y})$ and the gain of observing $\tilde{U}$ when the prior distribution on $\mathcal{Y}$ is $\mu_{Y|X}(\cdot|\pi_\eta(x))$, i.e.,

$$\left[\mathit{prior}(\mu_{Y|X}(\cdot|\pi_\eta(x))) - \ell_\eta(x))\right] = \sum_{\tilde{A}_u \in \tilde{\pi}} \mu_X(\tilde{A}_u|A) \cdot \left[1 - \max_{y \in \mathcal{Y}} \mu_{Y|X}(y|\tilde{A}_u \cap A)\right] \geq 0.$$

Let us assume that we have a family of WIS representations for $\mu_{X,Y}$ (Definition 3) given by $\{\eta_i(\cdot)\}_{i \geq 1}$ where $\eta_i : \mathcal{X} \to \mathcal{U}_i$ and $|\mathcal{U}_i| < \infty$ for any i. Using the definition of the ILD in (77) and (79), it follows that WIS is equivalent to

$$\lim_{i \longrightarrow \infty} \mathbb{E}_X \{\mathcal{I}_{\eta_i}(X)\} = 0. \tag{82}$$

As $\mathcal{I}_{\eta_i}(x) \leq \log |\mathcal{Y}|$ (uniformly in i and x), the convergence in (82) is equivalent to the convergence in probability of $(\mathcal{I}_{\eta_i}(X))_{i\geq 1}$, i.e., $\forall \epsilon > 0$ it follows that $\lim_{i\to\infty} \mathbb{P}\left(\{\mathcal{I}_{\eta_i}(X) > \epsilon\}\right) = 0$.

Using again (78), the proof of Theorem 35 reduces to verify that

$$\lim_{i \longrightarrow \infty} \mathbb{E}_X \left\{\ell_{\eta_i}(X)\right\} = 0. \tag{83}$$

Again $\ell_{\eta_i}(\cdot)$ is uniformly bounded by 1, then the convergence in (83) is equivalent to the convergence in probability of the random sequence $(\ell_{\eta_i}(X))_{i\geq 1}$, i.e., for any $\epsilon > 0$

$$\lim_{i\to\infty} \mathbb{P}\left(\{\ell_{\eta_i}(X) > \epsilon\}\right) = 0. \tag{84}$$

10.3.2 Main Argument

Proof Let us prove the result by contradiction. Let us assume that $\{\eta_i(\cdot)\}_{i\geq 1}$ is not OS. Then from (84), there exists $\epsilon > 0$ such that $\limsup_{i\to\infty} \mu_X(B^i_\epsilon) > 0$ where

$$B^i_\epsilon \equiv \{x \in \mathcal{X}, \ell_{\eta_i}(x) > \epsilon\} \subset \mathcal{X}. \tag{85}$$

Then, we can pick $\delta > 0$ where $\forall N > 0\ \exists i \geq N$ such that (from the hypothesis that $\limsup_{i\to\infty} \mu_X(B^i_\epsilon) > 0$)

$$\mu_X(B^i_\epsilon) \geq \delta. \tag{86}$$

Using the definition of the function $\mathcal{R}(v, \epsilon)$ (see Appendix I), for any $v \in \mathcal{P}(\mathcal{Y})$, it follows — from the expression of $f(v, \epsilon)$ in (16) — that $\mathcal{H}(\mathcal{R}(v, \epsilon_1)) \geq \mathcal{H}(\mathcal{R}(v, \epsilon_2))$ when $\epsilon_1 \geq \epsilon_2$; therefore, from (81), if $x \in B^i_\epsilon$, we have that

$$\mathcal{I}_{\eta_i}(x) \geq \mathcal{H}(\mu_{Y|X}(\cdot|\pi_i(x))) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_i(x)), prior(\mu_{Y|X}(\cdot|\pi_i(x))) - \epsilon)) \tag{87}$$

where $\pi_i(x)$ is a shorthand for $\pi_{\eta_i}(x)$.

The bound in (87) will be central to prove the result: a lower bound on the information loss density function of the operation loss density that is lower bounded by $\epsilon > 0$. More precisely, given $\epsilon > 0$, we proceed by finding a uniform lower bound for

$$\mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon)) \tag{88}$$

over all models $v \in \mathcal{P}(\mathcal{Y})$ that are admissible in the sense that $prior(v) \geq \epsilon$.

In particular, we will consider the following general information v.s. operation loss problem:

$$\mathcal{I}_{loss}(\epsilon, M) \equiv \min_{v\in\mathcal{P}^\epsilon([M])} \left\{\mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon))\right\}, \tag{89}$$

where

$$\mathcal{P}^\epsilon([M]) \equiv \left\{v \in \mathcal{P}([M]), prior(v) \geq \epsilon\right\}. \tag{90}$$

In this notation, we use $\mathcal{Y} = [M] \equiv \{1, .., M\}$ to make explicit the role that the cardinality of $\mathcal{Y}$ plays in this analysis. Importantly, we have the following (information loss vs. operation loss) interplay result that shows that a non-zero operation loss ($\epsilon > 0$) implies a positive information loss for any $M \geq 1$:

Theorem 37 *$\forall M \geq 1$, and for any $\epsilon \in (0, 1 - 1/M]$, it follows that $\mathcal{I}_{loss}(\epsilon, M) > 0$.*

The proof of this result requires (non-trivial) technical elements that are presented in Section 10.4.

Returning to the main proof argument, by definition of the operation loss density in (76) , we have that $\ell_{n_i}(x) \leq prior(\mu_{Y|X}(\cdot|\pi_i(x)))$, which implies that $\mu_{Y|X}(\cdot|\pi_i(x)) \in \mathcal{P}^{\ell_{n_i}(x)}([M])$ in (90). Then using (87) and (89), for any $x \in B^i_\epsilon$ (considering that $\epsilon < \ell_{\eta_i}(x)$ if $x \in B^i_\epsilon$)

$$\begin{aligned}\mathcal{I}_{\eta_i}(x) &\geq \mathcal{H}(\mu_{Y|X}(\cdot|\pi_i(x))) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_i(x)), prior(\mu_{Y|X}(\cdot|\pi_i(x))) - \epsilon)) \\ &\geq \min_{v \in \mathcal{P}^\epsilon([M])} \{\mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon))\} = \mathcal{I}_{loss}(\epsilon, M), \end{aligned} \tag{91}$$

where the second inequality comes from the observation that $\mu_{Y|X}(\cdot|\pi_i(x)) \in \mathcal{P}^{\ell_{n_i}(x)}([M]) \subset \mathcal{P}^\epsilon([M])$ from (90).

Finally, we use Theorem 37 and (91) to obtain that

$$\forall x \in B^i_\epsilon,\ \mathcal{I}_{\eta_i}(x) \geq \mathcal{I}_{loss}(\epsilon, M) > 0. \tag{92}$$

In particular, we have that for any $\bar{\epsilon} \in (0, \mathcal{I}_{loss}(\epsilon, M))$, $B^i_\epsilon \subset A^i_{\bar{\epsilon}} \equiv \{x \in \mathcal{X}, \mathcal{I}_{\eta_i}(x) > \bar{\epsilon}\}$. Then using the hypothesis in (86), we have that for any $N > 0$ there exists $i \geq N$ such that $\mu_X(A^i_{\bar{\epsilon}}) \geq \mu_X(B^i_{\bar{\epsilon}}) \geq \delta > 0$: i.e., $\limsup_{i\to\infty} \mu_X(A^i_{\bar{\epsilon}}) > 0$. This last result is equivalent to say that $(\mathcal{I}_{\eta_i}(X))_{i\geq 1}$ does not converge to zero in probability. Then, from the argument presented in Section 10.3.1 (see Eq.82), this last result contradicts the fact that $\{\eta_i(\cdot)\}_{i\geq 1}$ is WIS. This concludes the proof of Theorem 35. ■

10.4 Proof of Theorem 37

Proof Given a probability $\mu \in \mathcal{P}^\epsilon([M])$, Ho and Verdú (2010) presented a closed-form analytical expression for $\mathcal{R}(\mu, prior(\mu) - \epsilon)$ (see details in Appendix I) appearing in the definition of $\mathcal{I}_{loss}(\epsilon, M)$ in (89). To present this induced distribution more clearly, we assume, without loss of generality, that $\mu(1) \geq \mu(2) \geq \ldots \geq \mu(M)$. Then $\mu^\epsilon \equiv \mathcal{R}(\mu, prior(\mu) - \epsilon)$ has the following structure:[48]

$$\mu^\epsilon(1) = \mu(1) + \epsilon \leq 1 \tag{93}$$
$$\mu^\epsilon(2) = \theta$$
$$\ldots$$
$$\mu^\epsilon(K) = \theta \tag{94}$$
$$\mu^\epsilon(K+1) = \mu(K+1)$$
$$\ldots$$
$$\mu^\epsilon(M) = \mu(M), \tag{95}$$

48. To simplify notation $\mu(j)$ denotes $\mu(\{j\})$, i.e., $\mu(j)$ is a short-hand of the probability mass function (pmf).

where both $K \in \{2, .., M\}$ and $\theta \in (0, \mu(1))$ are functions of μ and $\epsilon > 0$ satisfying the following condition:

$$\sum_{j=2}^{K} (\mu(i) - \theta) = \epsilon > 0, \tag{96}$$

which makes μ^ϵ a well-defined probability in $\mathcal{P}([M])$.[49]

Therefore, using (93), (94) and (95), we have that for any $\mu \in \mathcal{P}^\epsilon([M])$:

$$\begin{aligned}\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) = \mu(1) \log \frac{1}{\mu(1)} - (\mu(1) + \epsilon) \log \frac{1}{\mu(1) + \epsilon} \\ + \sum_{j=2}^{K(\mu,\epsilon)} \mu(j) \log \frac{1}{\mu(j)} - (K(\mu, \epsilon) - 1)\theta(\mu, \epsilon) \log \frac{1}{\theta(\mu, \epsilon)},\end{aligned} \tag{97}$$

where here we make explicit the dependency of K and θ on μ and ϵ. It is important to note that by construction $\theta(\mu, \epsilon) < \mu(K) \leq \mu(K-1) \ldots \leq \mu(1)$. At this point, we will use the following result:

Lemma 38 *$\forall \epsilon > 0$ and for any $\mu \in \mathcal{P}^\epsilon([M])$, it follows that*

$$\sum_{j=2}^{K(\mu,\epsilon)} \mu(j) \log \frac{1}{\mu(j)} \geq (\theta(\mu, \epsilon) + \epsilon) \log \frac{1}{\theta(\mu, \epsilon) + \epsilon} + (K(\mu, \epsilon) - 2)\theta(\mu, \epsilon) \log \frac{1}{\theta(\mu, \epsilon)}. \tag{98}$$

The proof is presented in Appendix F.

Remark 39 *The proof of Lemma 38 comes from the use of some information-theoretic inequalities, similar to the argument used to prove that the Shannon entropy (over a finite alphabet) is minimized with a degenerated distribution (Cover and Thomas, 2006; Gray, 1990b).*

Applying Lemma 38, we have that for all $\mu \in \mathcal{P}^\epsilon([M])$:

$$\begin{aligned}\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) \geq \mu(1) \log \frac{1}{\mu(1)} - (\mu(1) + \epsilon) \log \frac{1}{\mu(1) + \epsilon} \\ + \left[(\theta(\mu, \epsilon) + \epsilon) \log \frac{1}{\theta(\mu, \epsilon) + \epsilon} - \theta(\mu, \epsilon) \log \frac{1}{\theta(\mu, \epsilon)} \right].\end{aligned} \tag{99}$$

Using the fact that $\theta(\mu, \epsilon) < \mu(K) \leq \mu(K-1) \ldots \leq \mu(1)$, and that $\sum_{j=2}^{K(\mu,\epsilon)} (\mu(j) - \theta(\mu, \epsilon)) = \epsilon$, it is simple to verify that[50]

$$\mu(2) - \theta(\mu, \epsilon) \geq \frac{\epsilon}{K - 1}, \tag{100}$$

which implies that $\theta(\mu, \epsilon) \leq \mu(2) - \epsilon/(K-1)$.

49. Ho and Verdú (2010) show that for any $\epsilon \leq prior(\mu)$, $\exists \theta \in [0, \mu(1))$ and $K \in \{2, .., M\}$ that meet the condition in (96).

50. This because $\mu(2) - \theta(\mu, \epsilon) \geq \mu(3) - \theta(\mu, \epsilon) \geq \ldots \geq \mu(K) - \theta(\mu, \epsilon) > 0$.

On the other hand, if we consider the following function (used in Eq.99):

$$f_1(\theta,\epsilon) \equiv (\theta+\epsilon)\log\frac{1}{\theta+\epsilon} - \theta\log\frac{1}{\theta}, \tag{101}$$

where $\frac{\partial f_1(\theta,\epsilon)}{\partial\theta}(\theta,\epsilon) = \log\frac{\theta}{\theta+\epsilon} < 0$. Then, $f_1(\theta,\epsilon)$ is strictly decreasing in the domain $\theta > 0$, for any $\epsilon > 0$. Therefore from (100), we have that

$$f_1(\theta(\mu,\epsilon),\epsilon) \geq f_1(\mu(2) - \epsilon/(K-1),\epsilon).$$

Applying this last inequality in (99), we have that

$$\begin{aligned}\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) &\geq -f_1(\mu(1),\epsilon) + f_1(\theta(\mu,\epsilon),\epsilon)\\ &\geq -f_1(\mu(1),\epsilon) + f_1(\mu(2)-\epsilon/(K-1),\epsilon).\end{aligned} \tag{102}$$

Furthermore, we have that $\mu(2) - \epsilon/(K-1) \leq \mu(2) - \epsilon/(M-1)$, which is a bound that is independent of $K(\mu,\epsilon)$. Finally applying this bound in (102), we have that

$$\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) \geq -f_1(\mu(1),\epsilon) + f_1(\mu(2)-\epsilon/(M-1),\epsilon). \tag{103}$$

At this point, we return to our main problem:

$$\begin{aligned}\mathcal{I}_{loss}(\epsilon,M) &= \min_{\mu\in\mathcal{P}^\epsilon([M])} \mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon)\\ &\geq \min_{\mu(1)\in[1/M,1-\epsilon]} \left(-f_1(\mu(1),\epsilon) + \min_{\mu(2)\in[0,\min\{\mu(1),1-\mu(1)\}]} (f_1(\mu(2)-\epsilon/(M-1),\epsilon))\right),\end{aligned} \tag{104}$$

where the lower bound in (104) comes from (103) and the fact that $\mu(1) = \max\{\mu(j), j\in[M]\} \in [1/M, 1-\epsilon]$ if $\mu\in\mathcal{P}^\epsilon([M])$. For the rest of the proof, we concentrate on the analysis of the RHS of (104), where we recognize for the second optimization in (104) two scenarios.

Case 1 (the restriction $\mu(2) \leq \mu(1)$ is active in Eq.(104): If we restrict the second optimization problem in (104) to the case where $\mu(1) \leq 1-\mu(1)$, this scenario implies that $\mu(1) \leq \frac{1}{2}$. In addition, we have that $\mu(1) \geq 1/M$ (achieved for the case of a uniform distribution in $[M]$). Then under this hypothesis, it follows that

$$\mathcal{I}_{loss}(\epsilon,M) \geq \min_{\mu(1)\in[1/M,1/2]} -f_1(\mu(1),\epsilon) + f_1(\mu(1)-\epsilon/(M-1),\epsilon). \tag{105}$$

The last bound comes from (104) using that $f_1(x,\epsilon)$ is strictly decreasing for $x\in(0,\infty)$ for any $\epsilon > 0$. Let us define $\tilde{f}(x,\epsilon) \equiv -f_1(x,\epsilon) + f_1(x-\epsilon/(M-1),\epsilon)$. It is simple to verify that $\frac{\partial\tilde{f}(x,\epsilon)}{\partial x} < 0$ for any $x > 0$[51]. This implies that

$$\mathcal{I}_{loss}(\epsilon,M) \geq \tilde{f}(1/2,\epsilon) = f_1\left(1/2 - \frac{\epsilon}{M-1},\epsilon\right) - f_1\left(1/2,\epsilon\right) > 0, \tag{106}$$

as we know that $(f_1(x,\epsilon))_{x>0}$ is strictly decreasing for any $\epsilon > 0$.

Case 2 (the restriction $\mu(2) \leq 1-\mu(1)$ is active in Eq.104): If we restrict the second optimization problem in (104) to the case where $1-\mu(1) < \mu(1)$, this scenario implies that

51. $\frac{\partial\tilde{f}(x,\epsilon)}{\partial x} = \log\frac{\psi_\epsilon(x)}{\psi_\epsilon(x-\epsilon/(M-1))} < 0$ for any $x > 0$, where $\psi_\epsilon(x) \equiv (1+\epsilon/x)$.

$\mu(1) > \frac{1}{2}$. In addition, as $\mu \in \mathcal{P}^{\epsilon}([M])$, it follows that $\mu(1) \leq 1 - \epsilon$. Therefore, under this hypothesis,

$$\mathcal{I}_{loss}(\epsilon, M) \geq \min_{\mu(1)\in(1/2,1-\epsilon]} -f_1(\mu(1), \epsilon) + f_1((1-\mu(1)) - \epsilon/(M-1), \epsilon), \tag{107}$$

from (104) and using that $(f_1(x,\epsilon))_{x\in(0,\infty)}$ is strictly decreasing for any $\epsilon > 0$. In this case, we consider $\tilde{\phi}(x,\epsilon) \equiv -f_1(x,\epsilon) + f_1((1-x) - \epsilon/(M-1), \epsilon)$. It is simple to verify that $\frac{\partial\tilde{\phi}(x,\epsilon)}{\partial x} > 0$ for any $x > 0$. Consequently, we have that

$$\mathcal{I}_{loss}(\epsilon, M) \geq \tilde{\phi}(1/2, \epsilon) = f_1\left(1/2 - \frac{\epsilon}{M-1}, \epsilon\right) - f_1\left(1/2, \epsilon\right) > 0. \tag{108}$$

In (108) and (106), we arrived to the same positive lower bound for $\mathcal{I}_{loss}(\epsilon, M)$, which concludes the proof of Theorem 37. ■

10.5 Proof of Theorem 14

The proof of this result is divided in two stages. First, we show that under the uniqueness assumption of $\tilde{r}_{\mu_{X,Y}}(\cdot)$ (Def. 13), the OS condition implies that $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$, where $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X) \in \mathcal{Y}$ in the MPE predictor of Y. The second stages used a refined version of *Fano's inequality* stated in (Feder and Merhav, 1994, Th.1) to prove that $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$ implies that $\lim_{i\to\infty} I(\tilde{U}; Y|U_i) = 0$. Finally, the equivalence stated in Theorem 14 is obtained from our result in Theorem 12 (WIS $\Rightarrow$ OS).

10.5.1 Stage 1: $\lim_{i\to\infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y}) \Rightarrow \lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$

Proof For the MPE decision rule, we use the expression of $\tilde{r}_{\mu_{X,Y}}(\cdot)$ in (10) and its induced partition $\tilde{\pi} = \left\{\tilde{A}_y, y \in \mathcal{Y}\right\}$ (with M cells) in (11). On the same line, we can introduce:

$$\tilde{r}_{\mu_{U_i,Y}}(u) \equiv \arg\max_{y\in\mathcal{Y}} \mu_{Y|U_i}(y|u). \tag{109}$$

and

$$\pi^i \equiv \left\{A^i_y \equiv \eta_i^{-1}(\tilde{r}^{-1}_{\mu_{U_i,Y}}(\{y\})), y \in \mathcal{Y}\right\} \subset \mathcal{B}(\mathcal{X}). \tag{110}$$

From these, we have that $\ell(\mu_{X,Y}) = \sum_{j=1}^{M}(1 - \mu_{X|Y}(\tilde{A}_j|j))\mu_Y(j)$ and $\ell(\mu_{U_i,Y}) = \sum_{j=1}^{M}(1 - \mu_{X|Y}(A^i_j|j))\mu_Y(j)$. Then, the operation loss of U_i (or $\eta_i(\cdot)$) can be expressed by:

$$\begin{aligned}\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) &= \sum_{j=1}^{M}(\mu_{X|Y}(\tilde{A}_j|j) - \mu_{X|Y}(A^i_j|j)) \cdot \mu_Y(j) \\ &= \sum_{j=1}^{M}(\mu_{X,Y}(\tilde{A}_j \times \{j\}) - \mu_{X,Y}(A^i_j \times \{j\}) \end{aligned} \tag{111}$$

On the other hand, our object of interest is $\ell(\mu_{U_i,\tilde{U}})$, where we have that

$$\begin{aligned}\ell(\mu_{U_i,\tilde{U}}) &\leq \mathbb{P}(\tilde{r}_{\mu_{U_i,Y}}(U_i) \neq \tilde{U}) = \mathbb{P}(\tilde{r}_{\mu_{U_i,Y}}(\eta_i(X)) \neq \tilde{U}) \\ &= 1 - \mathbb{P}(\tilde{r}_{\mu_{U_i,Y}}(\eta_i(X)) = \tilde{r}_{\mu_{X,Y}}(X)) = 1 - \sum_{j=1}^{M} \mu_X(\tilde{A}_j \cap A_j^i) \\ &= \sum_{j=1}^{M} (\mu_X(\tilde{A}_j) - \mu_X(\tilde{A}_j \cap A_j^i)) = \sum_{j=1}^{M} \mu_X(\tilde{A}_j \setminus A_j^i), \end{aligned} \tag{112}$$

where the first inequality comes from the definition of the MPE rule and the third equality from the fact that $\tilde{r}_{\mu_{U_i,Y}}(\eta_i(x)) = \sum_{j=1}^{M} \mathbf{1}_{A_j^i}(x) \cdot j$ and $\tilde{r}_{\mu_{X,Y}}(x) = \sum_{j=1}^{M} \mathbf{1}_{\tilde{A}_j}(x) \cdot j$.

To upper bound (112), let us work with the information loss expression in (111). It follows that

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{j=1}^{M} \mu_{X,Y}(\tilde{A}_j \setminus A_j^i \times \{j\}) - \sum_{\tilde{j}=1}^{M} \mu_{X,Y}(A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \times \{\tilde{j}\}). \tag{113}$$

Using the fact that $\bigcup_{j=1}^{M} \tilde{A}_j \setminus A_j^i = \bigcup_{\tilde{j}=1}^{M} A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}}$, then for any $j \in \mathcal{Y}$, it follows that $A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} = (\bigcup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) \cap A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}}$. From this last identity, we have that:

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{j=1}^{M} \left[\mu_{X,Y}(\tilde{A}_j \setminus A_j^i \times \{j\}) - \sum_{\tilde{j}=1,\tilde{j}\neq j}^{M} \mu_{X,Y}(A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \cap \tilde{A}_j \setminus A_j^i \times \{\tilde{j}\}) \right] \tag{114}$$

Let us analize one of the terms in the RHS of (114),

$$\begin{aligned} &\mu_{X,Y}(\tilde{A}_j \setminus A_j^i \times \{j\}) - \sum_{\tilde{j}=1,\tilde{j}\neq j}^{M} \mu_{X,Y}(A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \cap \tilde{A}_j \setminus A_j^i \times \{\tilde{j}\}) \\ &= \int_{\tilde{A}_j \setminus A_j^i} f_{X,Y}(x,j)dx - \sum_{\tilde{j}=1,\tilde{j}\neq j}^{M} \int_{A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \cap \tilde{A}_j \setminus A_j^i} f_{X,Y}(x,\tilde{j})dx \\ &\geq \int_{\tilde{A}_j \setminus A_j^i} \left[\underbrace{f_{X,Y}(x,j) - \max_{\tilde{j}\in\mathcal{Y},\tilde{j}\neq j} f_{X,Y}(x,\tilde{j})}_{\geq 0} \right] dx \geq 0, \end{aligned} \tag{115}$$

where $f_{X,Y}(x,y)$ denotes the density of $\mu_{X,Y}$. The last inequality in (115) comes from the definition of the MPE, the fact that for any $x \in \tilde{A}_j \setminus A_j^i$, $f_{X,Y}(x,j) = \max_{y\in\mathcal{Y}} f_{X,Y}(x,y)$, the fact that for any $x \in A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \cap \tilde{A}_j \setminus A_j^i$, $f_{X,Y}(x,\tilde{j}) \leq \max_{y\in\mathcal{Y},y\neq j} f_{X,Y}(x,y)$, and the fact that $\tilde{A}_j \setminus A_j^i = \bigcup_{\tilde{j}=1}^{M} \tilde{A}_j \setminus A_j^i \cap A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}}$. Finally, using (115) in (114), we have that

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) \geq \sum_{j=1}^{M} \int_{\tilde{A}_j \setminus A_j^i} (f_{X,Y}(x,j) - \max_{\tilde{j}\in\mathcal{Y},\tilde{j}\neq j} f_{X,Y}(x,\tilde{j}))dx. \tag{116}$$

Proving the result under a strong condition on $\mu_{X,Y}$:
For simplicity and clarity, let us assume for the moment the following discrimination condition on $\mu_{X,Y}$: $\exists K > 0$ and $\exists A \subset \mathcal{X}$ s.t. $\forall x \in A$

$$\left[\max_{y\in\mathcal{Y}} f_{X,Y}(x,y) - \max_{y\in\mathcal{Y}, y\neq\tilde{r}_{\mu_{X,Y}}(x)} f_{X,Y}(x,y)\right] \geq K \cdot f_X(x) \tag{117}$$

where $f_X(x) = \sum_{y\in\mathcal{Y}} f_{X,Y}(x,y)$ is the marginal density of X and $\mu_X(A) = 1$. This strong discrimination assumption on $\mu_{X,Y}$ is instrumental to directly prove our result. Indeed, under this assumption, we have that

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) \geq \sum_{j=1}^{M} \int_{(\tilde{A}_j \setminus A_j^i)\cap A} (f_{X,Y}(x,j) - \max_{\tilde{j}\in\mathcal{Y}, \tilde{j}\neq j} f_{X,Y}(x,\tilde{j}))dx \tag{118}$$

$$\geq K \cdot \sum_{j=1}^{M} \int_{(\tilde{A}_j \setminus A_j^i)\cap A} f_X(x)dx = K \cdot \sum_{j=1}^{M} \mu_X((\tilde{A}_j \setminus A_j^i) \cap A) \tag{119}$$

$$= K \cdot \sum_{j=1}^{M} \mu_X(\tilde{A}_j \setminus A_j^i) \geq K \cdot \ell(\mu_{U_i,\tilde{U}}). \tag{120}$$

The first inequality comes from (116) and the fact that $(\tilde{A}_j \setminus A_j^i) \cap A \subset \tilde{A}_j \setminus A_j^i$ for every $j \in \mathcal{Y}$ and the observation that the function been integrated is non-negative. The second inequality comes from the discrimination condition stated in (117). The second equality comes from the assumption that $\mu_X(A) = 1$, and the last inequality from (112). Therefore, $\lim_{i\to\infty} \ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = 0$ implies that $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$ under the discrimination condition on $\mu_{X,Y}$ stated in (117).

Relaxing the discrimination condition in (117):
The argument used above to prove that $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$ can be extended when we relax the condition stated in (117). For that let us introduce the following set:

$$A^{\epsilon}_{\mu_{X,Y}} \equiv \left\{x \in \mathcal{X}, f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)}) > \epsilon \cdot f_X(x)\right\} \tag{121}$$

where $y_{(1)} \equiv \arg\max_{y\in\mathcal{Y}} f_{X,Y}(x,y)$ and $y_{(2)} \equiv \arg\max_{y\in\mathcal{Y}, y\neq y_{(1)}} f_{X,Y}(x,y)$.[52] Importantly, we have the following result:

$$\lim_{\epsilon\to 0} \mu_x(A^{\epsilon}_{\mu_{X,Y}}) = \lim_{n\to\infty} \mu_x(A^{1/n}_{\mu_{X,Y}}) = \mu_X(\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}}), \tag{122}$$

where the last equality is from the continuity of μ_X under a sequence of monotonic events (Varadhan, 2001). The following important result will be instrumental for our analysis:

Theorem 40 *The model $\mu_{X,Y}$ has a unique MPE decision rule (Def. 13) if, and only if,*

$$\mu_X(\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}}) = \mu_X(\left\{x \in \mathcal{X}, f_{X,Y}(x, y_{(1)}) > f_{X,Y}(x, y_{(2)})\right\}) = 1. \tag{123}$$

52. To simplify the notation, we omit the dependency of $y_{(1)}$ and $y_{(2)}$ on x.

This result offers a concrete characterization of models with a unique MPE decision rule. The proof is presented in Section 10.6.

Returning to the main argument, we have from (115) that for any $j \in \mathcal{Y}$

$$\begin{aligned}
&\mu_{X,Y}(\tilde{A}_j \setminus A_j^i \times \{j\}) - \sum_{\tilde{j}=1,\tilde{j}\neq j}^{M} \mu_{X,Y}(A_{\tilde{j}}^i \setminus \tilde{A}_{\tilde{j}} \cap \tilde{A}_j \setminus A_j^i \times \{\tilde{j}\}) \\
&\geq \int_{\tilde{A}_j \setminus A_j^i} \left[f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)})\right] dx \qquad (124)\\
&\geq \int_{\tilde{A}_j \setminus A_j^i \cap A^{\epsilon}_{\mu_{X,Y}}} \left[f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)})\right] dx \geq \epsilon \int_{\tilde{A}_j \setminus A_j^i \cap A^{\epsilon}_{\mu_{X,Y}}} f_X(x) dx \qquad (125)\\
&= \epsilon \cdot \mu_X(\tilde{A}_j \setminus A_j^i \cap A^{\epsilon}_{\mu_{X,Y}}), \qquad (126)
\end{aligned}$$

where the second inequality comes from the fact that by definition $f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)}) \geq 0$, and the third inequality from the definition of $A^{\epsilon}_{\mu_{X,Y}}$ in (121). Applying this last inequality in (114), it follows that for any $\epsilon > 0$

$$\begin{aligned}
\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) &\geq \epsilon \cdot \sum_{j=1}^{M} \mu_X(\tilde{A}_j \setminus A_j^i \cap A^{\epsilon}_{\mu_{X,Y}}) \\
&= \epsilon \cdot \mu_X((\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) \cap A^{\epsilon}_{\mu_{X,Y}}). \qquad (127)
\end{aligned}$$

Consequently, using the assumption that $\lim_{i\to\infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y})$, it follows that for any $\epsilon > 0$

$$\lim_{i\to\infty} \mu_X((\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) \cap A^{\epsilon}_{\mu_{X,Y}}) = 0. \qquad (128)$$

Finally, for any $n \geq 1$, we have from (128) that

$$\begin{aligned}
\lim_{i\to\infty} \mu_X(\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) &\leq \lim_{i\to\infty} \mu_X((\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) \cap A^{1/n}_{\mu_{X,Y}}) + \mu_X((A^{1/n}_{\mu_{X,Y}})^c) \\
&= 1 - \mu_X(A^{1/n}_{\mu_{X,Y}}). \qquad (129)
\end{aligned}$$

This last bound implies that

$$\begin{aligned}
\lim_{i\to\infty} \mu_X(\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) &\leq 1 - \lim_{n\to\infty} \mu_X(A^{1/n}_{\mu_{X,Y}}) \\
&= 1 - \mu_X(\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}}). \qquad (130)
\end{aligned}$$

At this point, we use the assumption that $\mu_{X,Y}$ has a unique MPE and Theorem 40 to obtain from (130) and (112) that

$$\lim_{i\to\infty} \mu_X(\cup_{j=1}^{M} \tilde{A}_j \setminus A_j^i) = 0 \Rightarrow \lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0. \qquad (131)$$

■

10.5.2 Stage 2: $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0 \Rightarrow$ WIS

Proof For this part, we use the following result:

Lemma 41 *(Feder and Merhav, 1994, Th.1)*[53] *For any model $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ in the mixed continuous-discrete setting introduced in Section 2, it follows that*

$$\phi(\ell(\mu_{X,Y})) \geq H(Y|X)$$

where $\phi(r) = h(r) + r \cdot \log(|\mathcal{Y}| - 1)$ and $h(r) = -r\log(r) - (1-r)\log(1-r)$ denotes the binary entropy *(Cover and Thomas, 2006).*

Applying Lemma 41 in our context, i.e. over the family of models $\left\{\mu_{U_i,\tilde{U}}\right\}_{i\geq 1}$, we have that for any $i \geq 1$

$$h(\ell(\mu_{U_i,\tilde{U}})) + \ell(\mu_{U_i,\tilde{U}}) \cdot \log(M-1) \geq H(\tilde{U}|U_i). \tag{132}$$

Using the hypothesis that $\lim_{i\to\infty} \ell(\mu_{U_i,\tilde{U}}) = 0$ in (132) and the fact that $\lim_{r\to 0} h(r) = h(0) = 0$ (the continuity of the binary entropy (Cover and Thomas, 2006)), we have from (132) that $\lim_{i\to\infty} H(\tilde{U}|U_i) = 0$. Finally, by definition of the conditional MI (Cover and Thomas, 2006), we have that $I(\tilde{U};Y|U_i) \leq H(\tilde{U}|U_i)$ which proves that $\lim_{i\to\infty} I(\tilde{U};Y|U_i) = 0$ (WIS). ■

10.6 Proof of Theorem 40

Proof First, it is simple to verify, from the definition of $A^{\epsilon}_{\mu_{X,Y}}$ in (121), that $\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}} = \left\{x \in \mathcal{X}, f_{X,Y}(x, y_{(1)}) > f_{X,Y}(x, y_{(2)})\right\}$.

Let us begin proving that if $\mu_{X,Y}$ has a unique MPE rule then $\mu_X(\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}}) = 1$. We prove this implication by contradiction by assuming that $\mu_X(\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}}) < 1$. Let us denote by

$$B \equiv \left\{x \in \mathcal{X}, f_{X,Y}(x, y_{(1)}) = f_{X,Y}(x, y_{(2)})\right\} = (\bigcup_{n\geq 1} A^{1/n}_{\mu_{X,Y}})^c,$$

which is non-empty by our assumption. From this set, we can construct two different optimal (MPE) decision rules:

$$r_1(x) = y_{(1)} = \tilde{r}_{\mu_{X,Y}}(x), \forall x \in \mathcal{X} \tag{133}$$

and

$$\begin{aligned} r_2(x) &= r_1(x), \forall x \in \mathcal{X} \setminus B \\ r_2(x) &= y_{(2)}, \forall x \in B. \end{aligned} \tag{134}$$

53. The result in (Feder and Merhav, 1994, Th.1) also offers a tight lower bound of the form $H(Y|X) \geq \psi(\ell(\mu_{X,Y}))$, where $\psi(\cdot)$ is presented in closed-form in (Feder and Merhav, 1994, Eq.14). Importantly, the result proves that both bounds are tight, i.e., they are achievable for some model $\mu_{X,Y}$ in the class $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

By the definition of $y_{(1)}$, $y_{(2)}$ and B, we have that $\mathbb{P}(r_1(X) \neq Y) = \mathbb{P}(r_2(X) \neq Y) = \ell(\mu_{X,Y})$ and $\mathbb{P}(r_1(X) \neq r_2(X)) = \mu_X(B) > 0$. Then, $\mu_{X,Y}$ does not have a unique MPE from Definition 13.

For the other implication, let us assume that $\mu_X(B) = 0$ and let us consider the optimal MAP rule $\tilde{r}_{\mu_{X,Y}}(x) = y_{(1)}$. In addition, let us asume $r : \mathcal{X} \to \mathcal{Y}$ s.t. $\mathbb{P}(r(X) \neq Y) = \ell(\mu_{X,Y})$. Both $\tilde{r}_{\mu_{X,Y}}(\cdot)$ and $r(\cdot)$ induce an M-cell partition of $\mathcal{X}$ given by $\tilde{\pi} = \left\{\tilde{A}_j, j = 1, .., M\right\}$ and $\pi = \left\{A_j = r^{-1}(\{j\}), j = 1, .., M\right\}$, respectively. Using the same arguments used in the proof of Theorem 14, we have that:[54]

$$\mathbb{P}(r(X) \neq Y) - \ell(\mu_{X,Y}) = \sum_{j=1}^{M} \left[\mu_{X,Y}(\tilde{A}_j \setminus A_j \times \{j\}) - \sum_{\tilde{j}=1, \tilde{j} \neq j}^{M} \mu_{X,Y}((A_{\tilde{j}} \setminus \tilde{A}_{\tilde{j}}) \cap (\tilde{A}_j \setminus A_j) \times \left\{\tilde{j}\right\}) \right] = 0, \tag{135}$$

We observe that every term of this last summation over j is non-negative from construction of the MAP rule. Consequently, to meet the zero condition in (135), it follows that for any $j = 1, .., M$

$$\mu_{X,Y}(\tilde{A}_j \setminus A_j \times \{j\}) - \sum_{\tilde{j}=1, \tilde{j} \neq j}^{M} \mu_{X,Y}((A_{\tilde{j}} \setminus \tilde{A}_{\tilde{j}}) \cap (\tilde{A}_j \setminus A_j) \times \left\{\tilde{j}\right\}) = 0 \tag{136}$$

Therefore, from the inequality presented in Eq.(115) and (136), for any j

$$\int_{\tilde{A}_j \setminus A_j} \left[f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)}) \right] dx = 0. \tag{137}$$

At this point, we consider the set $A^{\epsilon}_{\mu_{X,Y}}$ in (121), where we have that

$$\begin{aligned} \int_{\tilde{A}_j \setminus A_j} \left[f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)}) \right] dx &\geq \int_{\tilde{A}_j \setminus A_j \cap A^{1/n}_{\mu_{X,Y}}} \left[f_{X,Y}(x, y_{(1)}) - f_{X,Y}(x, y_{(2)}) \right] dx \\ &\geq \frac{1}{n} \int_{\tilde{A}_j \setminus A_j \cap A^{1/n}_{\mu_{X,Y}}} f_X(x) dx \\ &= \frac{1}{n} \cdot \mu_X(\tilde{A}_j \setminus A_j \cap A^{1/n}_{\mu_{X,Y}}) = 0, \end{aligned} \tag{138}$$

for any $j = 1, .., M$ and any $n \geq 1$. Consequently, from (138) and the additivity (Breiman, 1968) $\mu_X((\cup_{j=1}^{M} \mathcal{A}_j \setminus A_j) \cap A^{1/n}_{\mu_{X,Y}}) = 0$. Then, for any $n \geq 1$

$$\begin{aligned} \mu_X(\cup_{j=1}^{M} \mathcal{A}_j \setminus A_j) &\leq \underbrace{\mu_X((\cup_{j=1}^{M} \mathcal{A}_j \setminus A_j) \cap A^{1/n}_{\mu_{X,Y}})}_{=0} + (1 - \mu_X(A^{1/n}_{\mu_{X,Y}})) \\ &= 1 - \mu_X(A^{1/n}_{\mu_{X,Y}}). \end{aligned} \tag{139}$$

54. In particular, from Eq.(114).

Finally, taking the limit in n in (139), we have that

$$\mu_X(\cup_{j=1}^M \tilde{A}_j \setminus A_j) \leq 1 - \lim_{n\to\infty} \mu_X(A_{\mu_{X,Y}}^{1/n}) = 1 - \mu_X(\cup_{n\geq 1} A_{\mu_{X,Y}}^{1/n}) = 0, \tag{140}$$

the last equality from the assumption that $\mu_X(\cup_{n\geq 1} A_{\mu_{X,Y}}^{1/n}) = 1$. To conclude, it is simple to verify that [55]

$$\mathbb{P}(r(X) \neq \tilde{r}_{\mu_{X,Y}}(X)) \leq \mu_X(\cup_{j=1}^M \tilde{A}_j \setminus A_j) = 0,$$

which means that $r(\cdot)$ is equal to $\tilde{r}_{\mu_{X,Y}}(\cdot)$ almost surely. Therefore, the MPE rule associated to $\mu_{X,Y}$ is unique (Def.13) under the assumption that $\mu_X(\cup_{n\geq 1} A_{\mu_{X,Y}}^{1/n}) = 1$. ■

10.7 Proof of Theorem 15

Proof For the proof of Theorem 15, we use a result on asymptotic sufficient partitions to approximate the divergence (Kullback-Leibler divergence or information divergence) between two distributions in an abstract measurable space (Gray, 1990b; Csiszár and Shields, 2004).

Lemma 42 *(Liese et al., 2006) Let us consider P, Q probability measures in the measurable space $(\Omega, \mathcal{F})$ such that $D(P||Q) < \infty$. Let us consider $\{\pi_n, n \geq 1\}$ a family of embedded finite size measurable partitions of Ω, in the sense that $\sigma(\pi_1) \subset \sigma(\pi_2) \subset .....$ Let us denote $\mathcal{S} \equiv \sigma(\pi_1 \cup \pi_2, ...)$[56], then it follows that*

$$\lim_{n\to\infty} D_{\sigma(\pi_n)}(P||Q) = D_{\mathcal{S}}(P||Q) \leq D(P||Q). \tag{141}$$

In this result, $D_{\sigma(\pi_n)}(P||Q) \equiv \sum_{A\in\pi_n} P(A) \log \frac{P(A)}{Q(A)}$ denotes the KL divergence of P with respect to Q restricted over the sigma-field induced by π_n, and

$$D_{\mathcal{S}}(P||Q) \equiv \sup_{\pi\in\mathcal{Q}(\mathcal{S})} D_{\sigma(\pi)}(P||Q),$$

where $\mathcal{S}$ is a general sub-sigma field of Ω (i.e., $\mathcal{S} \subset \mathcal{F}$)(Gray, 2009; Breiman, 1968), $\mathcal{Q}(\mathcal{S})$ denotes the collection of measurable finite partitions in $\mathcal{S}$, and

$$D(P||Q) \equiv D_{\mathcal{F}}(P||Q).$$

Consequently, if there exists $\{\pi_n, n \geq 1\}$ such that that $\sigma(\pi_1 \cup \pi_2, ...) = \mathcal{F}$, Lemma 42 implies that this family of partitions is sufficient for the KL divergence in the sense that for any pair P, Q such that $D(P||Q) < \infty$,

$$\lim_{n\to\infty} D_{\sigma(\pi_n)}(P||Q) = D(P||Q). \tag{142}$$

The result in Lemma 42 can be adapted to our mixed continuous-discrete setting $\Omega = \mathbb{R}^d \times \mathcal{Y}$ with the MI to obtain the following[57]:

55. This inequality follows from the same step presented to derive (112).
56. $\sigma(\mathcal{A})$ denotes the smallest sigma field that contains $\mathcal{A} \subset \mathcal{F}$ (Gray, 2009).
57. It is well-known that $I(\mu_{X,Y}) = D(\mu_{X,Y}||\mu_X \cdot \mu_Y)$ (Cover and Thomas, 2006; Gray, 1990b).

Corollary 43 *Let us consider a joint random vector (X,Y) in mixed setting $\Omega = \mathbb{R}^d \times \mathcal{Y}$ with probability $\mu_{X,Y}$ and a collection of finite embedded partitions $\{\pi_n, n \geq 1\}$ in $\mathcal{B}(\mathbb{R}^d)$ such that $\sigma(\pi_1) \subset \sigma(\pi_2) \subset .....$ Then*

$$\lim_{n\to\infty} I_{\sigma(\pi_n)}(X;Y) = I_{\mathcal{S}}(X;Y), \tag{143}$$

where $\mathcal{S} \equiv \sigma(\pi_1 \cup \pi_2, ...) \subset \mathcal{B}(\mathbb{R}^d)$,

$$I_{\sigma(\pi_n)}(X;Y) \equiv \sum_{A\in\pi_n}\sum_{y\in\mathcal{Y}} \mu_{X,Y}(A\times\{y\}) \log \frac{\mu_{X,Y}(A\times\{y\})}{\mu_X(A)\cdot\mu_Y(\{y\})} \leq I(X;Y)$$

is the quantized (discrete) MI, and

$$I_{\mathcal{S}}(X;Y) \equiv \sup_{\pi\in\mathcal{Q}(\mathcal{S})} I_{\sigma(\pi)}(X;Y). \tag{144}$$

Furthermore, from Corollary 43 we have the following:

Corollary 44 *In the setting of Corollary 43, if $\sigma(\pi_1 \cup \pi_2, ...) = \mathcal{B}(\mathbb{R}^d)$ then*

$$\lim_{n\to\infty} I_{\sigma(\pi_n)}(X;Y) = I(X;Y) \tag{145}$$

where (Gray, 1990b)

$$I(X;Y) = I_{\mathcal{B}(\mathbb{R}^d)}(X;Y), \tag{146}$$

for any model $\mu_{X,Y}$ in $(\mathbb{R}^d \times \mathcal{Y}, \sigma(\mathcal{B}(\mathbb{R}^d) \times 2^{\mathcal{Y}}))$.

In the last two results, it is simple to verify that

$$I_{\sigma(\pi_n)}(X;Y) = \sum_{A\in\pi_n, y\in\mathcal{Y}} \mu_{X,Y}(A\times\{y\}) \log \frac{\mu_{X,Y}(A\times\{y\}}{\mu_X(A)\mu_Y(\{y\})}$$

is equal to $\mathcal{I}(\mu_{U_n,Y}) = I(U_n;Y)$ where $U_n = \eta_{\pi_n}(X)$ and $\eta_{\pi_n}(\cdot)$ is the following finite-size representation (VQ) (induced by π_n)

$$\eta_{\pi_n}(x) \equiv \sum_{A_l\in\pi_n} l \cdot \mathbf{1}_{A_l}(x) \in \{1, .., |\pi_n|\}\,,\ \forall x \in \mathbb{R}^d. \tag{147}$$

Then under the expressiveness condition of Corollary 44, (145) implies that the $\{\eta_{\pi_n}(\cdot), n \geq 1\}$ is IS distribution-free.

Returning to the statement of Theorem 15 and noting that $\sigma(\eta_i) = \sigma(\pi_{\eta_i})$, we can use Corollary 43 to obtain (20) from (145). To conclude, we use Theorem 12 (IS $\Rightarrow$ OS) to obtain (21), which concludes the proof. ■

10.8 Proof of Theorem 18

First, we use Theorem 12 in this random representation setting to prove the following:

Lemma 45 *Let $\Pi = \{\pi_n(\cdot), n \geq 1\}$ be a partition scheme driven by a random process $(Z_n)_{n\geq 1}$ with $Z_i \sim \mu_{X,Y}$. If Π is IS (Def. 16), then Π is OS (Def. 17).*

Proof Let us define the collection of typical sequences

$$\mathcal{Z} \equiv \left\{(z_n)_{n\geq 1} \in (\mathcal{X}\times\mathcal{Y})^{\mathbb{N}} : \lim_{n\to\infty} I(\eta_{\pi_n(z_1,..,z_n)}(X);Y) = I(X;Y)\right\}. \tag{148}$$

Then for any $(z_n)_{n\geq 1} \in \mathcal{Z}$, the induced sequence of representations $\left\{\eta_{\pi_n(z_1,..,z_n)}(\cdot) : n \geq 1\right\}$ is IS (Def. 2) from (148). Then applying Theorem 12 (IS $\Rightarrow$ OS), we have that for all $(z_n)_{n\geq 1} \in \mathcal{Z}$

$$\lim_{n\to\infty} \ell(\mu_{\eta_{\pi_n(z_1,..,z_n)}(X),Y}) = \ell(\mu_{X,Y}). \tag{149}$$

From (149), we conclude that all the representations indexed by a sequence in $\mathcal{Z}$ are OS (Def. 1). Considering that $\mathbb{P}(\mathcal{Z}) = 1$ (from the hypothesis that Π is IS), we conclude that Π is OS. ■

Second, we have the following result introduced by Silva and Narayanan (2010a)[58] for analysing the approximation error (bias) induced by data-driven partitions (schemes) in the problem of MI estimation.

Lemma 46 *(Silva and Narayanan, 2010a) Let $\mu_{X,Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ in our mixed continuous-discrete setting and $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), \dots\}$ be a partition scheme driven by $Z_1, Z_2, \dots$ where $Z_i \sim \mu_{X,Y}$ for any $i \geq 1$. If μ_X has a density and Π satisfies the shrinking cell condition in (30) then*

$$\lim_{n\to\infty} I(\eta_{\pi_n(Z_1,..,Z_n)}(X);Y) = I(X;Y), \text{ with probability one.} \tag{150}$$

Finally, the proof of Theorem 18 comes from Lemma 46 to obtain (31) from the hypothesis in (30) and then the application of Lemma 45 to obtain (32).

10.9 Proof of Theorem 24

Proof Let us assume that $(X,Y) \sim \mu^{\theta}_{X,Y}$ for some arbitrary $\theta \in \Theta$. By the hypothesis in (43), we know that

$$I(\eta^*(X);Y|\eta_i(X)) = I((\eta^*(X),\eta_i(X));Y) - I(\eta_i(X);Y) \longrightarrow 0 \tag{151}$$

as i tends to infinity (the expressiveness condition of $\{\eta_i\}_{i\geq 1}$). For the rest, we will use Theorem 12, for which we focus on $I((\tilde{r}_\theta(X),\eta_i(X));Y) - I((\eta_i(X));Y)$.

Using the fact that $H(\tilde{r}_\theta(X)|\eta^*(X)) = 0$, it is simple to verify that for any $i \geq 1$

$$I((\eta^*(X),\eta_i(X));Y) \geq I((\tilde{r}_\theta(X),\eta_i(X));Y). \tag{152}$$

58. This result derives from the proof of (Silva and Narayanan, 2010a, Th. 2).

Indeed $\forall i \geq 1$

$$\begin{aligned} & I((\eta^*(X), \eta_i(X)); Y) - I((\tilde{r}_\theta(X), \eta_i(X)); Y) \\ & = I((\eta^*(X), \eta_i(X), \tilde{r}_\theta(X)); Y) - I((\tilde{r}_\theta(X), \eta_i(X)); Y) \\ & = I(\eta^*(X); Y | \tilde{r}_\theta(X), \eta_i(X)) \geq 0, \end{aligned} \tag{153}$$

the first equality from the assumption that $H(\tilde{r}_\theta(X)|\eta^*(X)) = 0$ ($\eta^*(\cdot)$ is operationally sufficient for Λ) and the second from the fact that conditional MI is non-negative (Cover and Thomas, 2006).

Using (152), we have that for any $i \geq 1$

$$I((\eta^*(X), \eta_i(X)); Y) - I(\eta_i(X); Y) \geq I((\tilde{r}_\theta(X), \eta_i(X)); Y) - I(\eta_i(X); Y). \tag{154}$$

Consequently, the asymptotic condition in (151) implies that $(U_i = \eta_i(X))_{i \geq 1}$ is WIS (see Def. 3) for $\mu^\theta_{X,Y}$ and the application of Theorem 12 implies that (see Def. 1):

$$\lim_{i \to \infty} \ell(\mu^\theta_{\eta_i(X),Y}) = \ell(\mu^\theta_{X,Y}). \tag{155}$$

Finally, the presented argument is valid for any $\mu^\theta_{X,Y} \in \Lambda$, which concludes the proof. ■

10.10 Proof of Theorem 31

Proof First, we denote by $\eta^B(\cdot)$ a solution of the IB problem

$$\max_{\eta(\cdot) \in \mathcal{F}(X, \mathbb{R})} I(\eta(U_\Lambda); Y) \; s.t. \; I(U_\Lambda; \eta(U_\Lambda)) \leq B, \tag{156}$$

for any $B > 0$.

For the proof, we adopt some of the expressiveness result presented in Section 5. In particular, we use the digital universal (distribution-free) construction presented in Section 5.1.1. Let us consider the collection of finite measurable partitions $\{\tilde{\pi}_m, m \geq 1\}$ presented in (22), (23) and (24) with its associated notations for the sets $(B_{m,0}, B_{m,\bar{j}}, \mathcal{J}_m)$ and its induced measurable encoder (VQ) $\eta_{\tilde{\pi}_m}(\cdot) : \mathcal{X} \to \{(m2^m, ..., m2^m)\} \cup \mathcal{J}_m \subset \mathbb{Z}^d$ in (26). As the cardinality of $\tilde{\pi}_m$ is $(m2^{m+1})^d + 1 < \infty$, we can construct an injective scalar (one to one) function $f^m : \{(m2^m, ..., m2^m)\} \cup \mathcal{J}_m \to \left\{1, \ldots, (m2^{m+1})^d + 1\right\} \subset \mathbb{R}$ and with that the following new encoder $\tilde{\eta}_{\tilde{\pi}_m}(\cdot) \in \mathcal{F}(X, \mathbb{R})$ induced by $\tilde{\pi}_m$:

$$\tilde{\eta}_{\tilde{\pi}_m}(x) \equiv f^m(m2^m, ..., m2^m) \cdot \mathbf{1}_{B_{m,0}}(x) + \sum_{\bar{j} \in \mathcal{J}_m} f^m(\bar{j}) \cdot \mathbf{1}_{B_{m,\bar{j}}}(x) \in \mathbb{R}, \tag{157}$$

$\forall x \in \mathcal{X}$. Considering that $\tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda)$ is a deterministic function of U_Λ and a finite-alphabet variable, we have that (Cover and Thomas, 2006)

$$I(U_\Lambda; \tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda)) \leq H(\tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda)) \leq \log_2((m2^{m+1})^d + 1) < \underbrace{d\left[\log_2(m) + m + 1\right] + 1}_{\beta_m \equiv}. \tag{158}$$

Then, for any $B \geq \beta_m$ (scalar defined in Eq. 158) and $m \geq 1$, we have from the definition of $\eta^B(\cdot)$ in (156) that

$$I(\eta^B(U_\Lambda); Y) \geq I(\tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda); Y). \tag{159}$$

At this point, we consider $B_m = \beta_m$ for any $m \geq 1$. Using Theorem 15 and the fact that $\sigma(\cup_{m\geq 1}\tilde{\pi}_m) = \mathcal{B}(\mathbb{R}^d)$ (Liese et al., 2006) (see more details in Section 5.1.1) it follows that

$$\lim_{m\to\infty} I(\tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda); Y) = I(U_\Lambda; Y), \tag{160}$$

which using the inequality in (159) implies the WIS condition of Theorem 24, i.e.,

$$\lim_{m\to\infty} I(U_\Lambda; Y|\eta^{B_m}(U_\Lambda)) = 0. \tag{161}$$

At this point, we adopt Theorem 24 (WIS $\Rightarrow$ OS) to obtain that

$$\lim_{m\to\infty} \ell(\mu_{\eta^{B_m}(U_\Lambda),Y}) = \ell(\mu_{U_\Lambda,Y}) = \ell(\mu_{X,Y}). \tag{162}$$

The last equality in (161) comes from the hypothesis that $\eta_\Lambda(\cdot)$ is OS for Λ (see Def. 23).

To conclude the argument, we note that $B_m = \beta_m$ is $\mathcal{O}(m)$, then (161) and (162) imply (49) and (50), respectively. ■

10.11 Proof of Theorem 34

Proof The problem in Eq.(52) can be written as:

$$\tilde{\eta}^B(\cdot) = \arg \max_{\eta(\cdot)\in\mathcal{F}(X,\mathbb{R})} I(\eta(X); \tilde{Y}) \ s.t. \ I(U_\Lambda; \eta(X)) \leq B, \tag{163}$$

where we introduce the discrete target variable $\tilde{Y} \equiv U_\Lambda$.

Following the same argument and the collection of lossy encoders $(\tilde{\eta}_{\tilde{\pi}_m}(\eta_\Lambda(\cdot))$ in Eq. 157) used in the proof of Theorem 31 (Section 10.10), there is a monotonically increasing sequence of real numbers $(\beta_m)_{m\geq 1}$ (introduced in Eq. 158) where we have that[59]

$$\lim_{m\to\infty} I(\tilde{\eta}^{\beta_m}(X); \tilde{Y}) = \lim_{m\to\infty} I(\tilde{\eta}_{\tilde{\pi}_m}(U_\Lambda); \tilde{Y}) = I(U_\Lambda; \tilde{Y}) = H(\tilde{Y}). \tag{164}$$

This implies that

$$\lim_{m\to\infty} H(\tilde{Y}|\tilde{\eta}^{\beta_m}(X)) = H(\tilde{Y}) - \lim_{m\to\infty} I(\tilde{\eta}^{\beta_m}(X); \tilde{Y}) \tag{165}$$

$$= H(\tilde{Y}) - H(\tilde{Y}) = 0, \tag{166}$$

and, consequently, $\lim_{B\to\infty} H(U_\Lambda = \tilde{Y}|\tilde{\eta}^B(X)) = 0$ from (166) and the fact that $\beta_m = d[\log_2(m) + m + 1] + 1 < \infty$ (see the proof of Theorem 31).

59. To derive (164), we use the universal encoder in (157), Theorem 15, the distribution-free inequality in (158) and the definition of $\tilde{\eta}^B(\cdot)$ in Eq.(163). The details follows the same steps presented in the proof of Theorem 31.

Finally, as U_Λ is a discrete (a finite entropy) r.v., we have that for any $\mu_{X,Y} \in \Lambda$

$$I(U_\Lambda; Y|\tilde{\eta}^B(X)) \leq H(U_\Lambda|\tilde{\eta}^B(X)). \tag{167}$$

Then from (166) and (167), we have that

$$\lim_{B\to\infty} I(U_\Lambda; Y|\tilde{\eta}^B(X)) = 0, \tag{168}$$

and from Theorem 24 (non-oracle WIS $\Rightarrow$ OS)

$$\lim_{B\to\infty} \ell(\mu_{\tilde{\eta}^B(X),Y}) = \ell(\mu_{X,Y}). \tag{169}$$

■

Acknowledgments

The authors of this work acknowledge support from ANID Fondecyt-Regular 1210315 and the Advanced Center for Electrical and Electronic Engineering (AC3E) ANID-Basal project FB0008. In addition, Dr. Tobar acknowledges support from Google, and ANID grants Fondecyt-Regular 1210606 and the Center for Mathematical Modeling (CMM) Basal project FB210005. The authors thank Diane Greenstein for editing and proofreading this material. We thank the two anonymous reviewers for their constructive and rigorous analysis of this work. This work was developed when Dr. Tobar was with the Initiative for Data & Artificial Intelligence, Universidad de Chile.

Appendix A. Proof of Proposition 4

Proof From Bayes decision, it is known that $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X)$ is a sufficient statistic of X in the operational sense, i.e., $\ell(\mu_{\tilde{U},Y}) = \ell(\mu_{X,Y})$. For this analysis, it is useful to consider the augmented observation vector $(\tilde{U}, U_i)$, where its error $\ell(\mu_{(\tilde{U},U_i),Y})$ is at most the error achieved by $\tilde{U}$. Consequently, we have that $\ell(\mu_{(\tilde{U},U_i),Y}) = \ell(\mu_{X,Y})$. This identity helps us to express the loss in (7) conveniently:

$$\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \ell(\mu_{U_i,Y}) - \ell(\mu_{(\tilde{U},U_i),Y}) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - \sum_{\tilde{A}_u\in\tilde{\pi}} \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}\cap\tilde{A}_u) \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|\tilde{A}_u\cap B_{i,j})\right]. \tag{170}$$

Finally (14) follows directly from (170). ■

Appendix B. Proof of Proposition 5

Proof From the definition of MI and the discrete nature of the joint vector $(\tilde{U}, U_i)$ (Cover and Thomas, 2006), we have that

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) = H(Y) - \sum_{\tilde{A}_u\in\tilde{\pi}} \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}\cap\tilde{A}_u)\cdot\mathcal{H}(\mu_{Y|X}(\cdot|\tilde{A}_u\cap B_{i,j})). \tag{171}$$

On the other hand, $\mathcal{I}(\mu_{U_i,Y}) = H(Y) - \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j})\cdot\mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j}))$. The result in (15) derives directly from these expressions. ■

Appendix C. Proof of Corollary 11

Proof Assuming that (19) holds, this implies that at least one component j of the sum satisfies $g(\mu_{X,Y}, B_{i,j}) > 0 \Leftrightarrow \epsilon_{i,j} < \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right]$. Then Lemma 8 implies that $\mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B_{i,j}), \epsilon_{i,j}) > 0$. This last inequality suffices to show that

$$\mathcal{I}(\mu_{X,Y}) - \mathcal{I}(\mu_{U_i,Y}) \geq \mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) > 0. \tag{172}$$

The first inequality in (172) comes from the fact that $(\tilde{U}, U_i)$ is a deterministic function of X (and the chain rule of MI) and the second comes from (18). ■

Appendix D. The Inequality of Section 3.4.1

By construction, we have that $\lim_{i\to\infty} \eta_{\pi_i}(\cdot) = \eta_{\tilde{\pi}}(\cdot)$ point-wise, where

$$\tilde{\pi} = \{(-\infty, 0), \{0\}, (0, \infty)\} \subset \mathcal{B}(\mathcal{X})$$

and $\eta_{\tilde{\pi}}(\cdot) = \tilde{r}_{\mu_{X,Y}}(\cdot)$, μ_X-almost surely. From this, we have that $\lim_{i\to\infty} \mu_{U_i,Y} = \mu_{\tilde{U},Y}$ in total variation. Considering that $\{\mu_{U_i,Y}, i \geq 1\} \subset \mathcal{P}(\{1,2,3\} \times \mathcal{Y})$ and that the entropy and MI are a continuous functionals (w.r.t. the total variational distance) in finite-alphabet spaces (Cover and Thomas, 2006), we have that:

$$\lim_{i\to\infty} \mathcal{I}(\mu_{U_i,Y}) = \mathcal{I}(\mu_{\tilde{U},Y}) = I(\tilde{U};Y). \tag{173}$$

Finally, as the model $\mu_{X,Y}$ is continuous, we have that $I(\tilde{U};Y) < I(X;Y)$[60]. Then from (173), $I(X;Y|U_i) = I(X;Y) - I(U_i;Y)$ is non-vanishing when $i \to \infty$.

Appendix E. Finite-Size Families: $\sigma(\Lambda) = \sigma(\pi)$

Proposition 47 *Let us consider a finite collection of models $\Lambda = \left\{\mu^i_{X,Y}, i = 1,..,L\right\}$, then*

$$\sigma(\Lambda) = \sigma(\pi)$$

where $\pi \equiv \bigcap_{i=1}^{L} \pi^{*,i}$ and $\pi^{*,i}$ is the M-size partition induced by the MPE decision rule of $\mu^i_{X,Y}$ in (38).

Proof The elements of π can be denoted and indexed by

$$A_{j_1,\ldots,j_L} \equiv A^1_{j_1} \cap A^2_{j_2} \cap \ldots A^L_{j_L}, \forall (j_1,\ldots,j_L) \in [M]^L \tag{174}$$

where $\pi^{*,i} = \left\{A^i_j, j \in [M]\right\}$. For simplicity, let us assume that any $(j_1,\ldots,j_L) \in [M]^L$ indexes a unique and none-empty event in π,

Let us begin with the implication $\sigma(\Lambda) \subset \sigma(\pi)$: It is simple to verify that for any $A \in \bigcup_{i=1}^{L} \pi^{*,i}$, $\exists i \in [L], \exists j \in [M]$ such that $A = A^i_j$ and, consequently, from (174)

$$A = \bigcup_{j_1=1}^{M} \ldots \bigcup_{j_{i-1}=1}^{M} \bigcup_{j_{i+1}=1}^{M} \ldots \bigcup_{j_L=1}^{M} A_{j_1,\ldots,j_{i-1},j,j_{i+1}\ldots,j_L} \in \sigma(\pi). \tag{175}$$

The last condition in (175) is from the fact that union of events in π belongs to $\sigma(\pi)$. Therefore, $\bigcup_{i=1}^{L} \pi^{*,i} \subset \sigma(\pi)$, which implies that $\sigma(\Lambda) = \sigma(\bigcup_{i=1}^{L} \pi^{*,i}) \subset \sigma(\pi)$, considering $\sigma(\pi)$ is a sigma-field and $\sigma(\bigcup_{i=1}^{L} \pi^{*,i})$ is the smallest sigma-field that contains $\bigcup_{i=1}^{L} \pi^{*,i}$.

For the converse implication $\sigma(\pi) \subset \sigma(\Lambda)$: $A_{j_1,\ldots,j_L}$ in (174) is induced by a finite number of set operations of $\bigcup_{i=1}^{L} \pi^{*,i}$. This implies that $A_{j_1,\ldots,j_L} \in \sigma(\bigcup_{i=1}^{L} \pi^{*,i})$ and, consequently, $\pi \subset \sigma(\bigcup_{i=1}^{L} \pi^{*,i})$. Then $\sigma(\pi) \subset \sigma(\bigcup_{i=1}^{L} \pi^{*,i}) = \sigma(\Lambda)$. ■

Appendix F. Proof of Lemma 38

Proof Let us consider an arbitrary $\mu \in \mathcal{P}^{\epsilon}([M])$, where we have that $\mu(1) \geq \mu(2) \geq \ldots \mu(K) > \theta$ and that $\sum_{j=2}^{K}(\mu(j) - \theta) = \epsilon$. In this analysis, the dependency of K and θ on

60. This inequality can be verify numerically using the estimation strategy presented in Section 8.1.3.

μ and ϵ will be considered implicit. We consider the conditional probability $\tilde{\mu} \equiv \mu(\cdot|\beta) \in \mathcal{P}([M])$ for the set $\beta = \{2, \ldots, K\}$, i.e.,

$$\begin{aligned}
\tilde{\mu}(2) &= \frac{\mu(2)}{\theta(K-1)+\epsilon} \geq \tilde{\theta} \equiv \frac{\theta}{\theta(K-1)+\epsilon} > 0, \\
&\ldots \\
\tilde{\mu}(K) &= \frac{\mu(K)}{\theta(K-1)+\epsilon} \geq \tilde{\theta}.
\end{aligned} \tag{176}$$

In this context, it is instrumental to introduce the following family of admissible distributions $\{\bar{e}_2, \ldots, \bar{e}_K\} \subset \mathcal{P}([M])$ with support in β, where $\bar{e}_j$ is given by

$$\begin{aligned}
\bar{e}_j(2) &= \tilde{\theta}, \ldots, \\
\bar{e}_j(j-1) &= \tilde{\theta}, \\
\bar{e}_j(j) &= \tilde{\theta} + \frac{\epsilon}{\theta(K-1)+\epsilon}, \\
\bar{e}_j(j+1) &= \tilde{\theta}, \ldots, \\
\bar{e}_j(K) &= \tilde{\theta}.
\end{aligned} \tag{177}$$

Importantly, it is simple to verify that $\tilde{\mu}$ (in Eq.176) can be written as a convex combination of our admissible family $\{\bar{e}_2, \ldots, \bar{e}_K\}$, i.e., $\exists (w_2, .., w_K) \in [0,1]^{K-1}$ such that $\sum_{j=2}^{K} w_j = 1$ and

$$\tilde{\mu} = \sum_{j=2}^{K} w_j \cdot \bar{e}_j, \tag{178}$$

where $w_j = \frac{\tilde{\mu}(j) - \tilde{\theta}}{\tilde{\epsilon}}$ with $\tilde{\epsilon} \equiv \frac{\epsilon}{\theta(K-1)+\epsilon} > 0$.

Let us define two random variables Z and O such that Z takes values in $[M]$ and O takes values in $\{2, .., K\}$ and

$$P_{Z|O}(\cdot|k) = \bar{e}_k \in \mathcal{P}([M]), \text{ and } P_O(k) = w_k, \tag{179}$$

$\forall k \in \{2, .., K\}$. By construction, $P_Z = \sum_{j=2}^{K} w_j \cdot \bar{e}_j = \tilde{\mu}$. Therefore, we can use that $H(Z|O) \leq H(Z)$ (Cover and Thomas, 2006), which implies that $\sum_{j=2}^{K} w_j \cdot \mathcal{H}(\bar{e}_j) \leq \mathcal{H}(\tilde{\mu})$. Finally, by the invariant of the entropy to one-to-one permutations, $\mathcal{H}(\bar{e}_2) = \ldots = \mathcal{H}(\bar{e}_K)$, then we have that $\mathcal{H}(\bar{e}_2) \leq \mathcal{H}(\tilde{\mu})$, which implies that

$$(\tilde{\theta} + \tilde{\epsilon}) \log \frac{1}{\tilde{\theta} + \tilde{\epsilon}} + (K-2)\tilde{\theta} \log \frac{1}{\tilde{\theta}} \leq \mathcal{H}(\tilde{\mu}). \tag{180}$$

Returning to our original problem, we have that

$$\sum_{j=2}^{K} \mu(j) \log \frac{1}{\mu(j)} = \mu(\beta)\mathcal{H}(\tilde{\mu}) + \mu(\beta) \log \frac{1}{\mu(\beta)} \geq$$

$$(\theta(K-1)+\epsilon) \cdot \left[(\tilde{\theta}+\tilde{\epsilon}) \log \frac{1}{\tilde{\theta}+\tilde{\epsilon}} + (K-2)\tilde{\theta} \log \frac{1}{\tilde{\theta}} \right] + (\theta(K-1)+\epsilon) \log \frac{1}{(\theta(K-1)+\epsilon)}$$

$$= (\theta+\epsilon) \log \frac{(K-1)\theta+\epsilon}{\theta+\epsilon} + (K-2)\theta \log \frac{(K-1)\theta+\epsilon}{\theta} + (\theta(K-1)+\epsilon) \log \frac{1}{(\theta(K-1)+\epsilon)}$$

$$= (\theta+\epsilon) \log \frac{1}{\theta+\epsilon} + (K-2)\theta \log \frac{1}{\theta}, \tag{181}$$

where for the first inequality we use the lower bound in (180) and the fact that $\mu(\beta) = \theta(K-1)+\epsilon$, and for the first equality we use that $\tilde{\theta} = \theta/((K-1)\theta+\epsilon)$ and $\tilde{\epsilon} = \epsilon/((K-1)\theta+\epsilon)$. Finally, (181) proves the result in (98). ■

Appendix G. Proof of Proposition 28

Proof Let $\eta^*(\cdot)$ be maximal invariant for $\mathcal{G}$. This means that for any pair $(x,y) \in \mathcal{X}^2$ where $\mathcal{G}(x) \neq \mathcal{G}(y)$ then $\eta^*(x) \neq \eta^*(y)$. We have that for any $\mu^{\theta}_{X,Y} \in \Lambda$, $\tilde{r}_\theta(\cdot)$ solution of (37) is $\mathcal{G}$-invariant (see Def. 25), this invariant condition implies that $\tilde{r}_\theta(\cdot)$ is fully determined if we know the values of $\tilde{r}_\theta(\cdot)$ in every cell of $\pi_{\mathcal{G}} = \{\mathcal{G}(x), x \in \mathcal{X}\}$. Therefore, if we know $\tilde{r}_\theta(\eta^*(x))$ for any $x \in \mathcal{X}$, we fully determine $\tilde{r}_\theta(\cdot)$ using the fact that $\eta^*(\cdot)$ is maximal invariant. Indeed, we have that $\tilde{r}_\theta(x) = \tilde{r}_\theta(\eta^*(x))$, which implies that $H(\tilde{r}_\theta(X)|\eta^*(X)) = 0$ where $X \sim \mu^{\theta}_X$. Then, we use Proposition 22 to conclude the proof. ■

Appendix H. Proposition 48

Proposition 48 *If $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is model-based $\mathcal{G}$-invariant, in the sense that $(X,Y) = (g(X),Y)$ in distribution for any $g \in \mathcal{G}$, then $\mu_{X,Y}$ is operational $\mathcal{G}$-invariant (see Def.26).*

Proof If $\mu_{X,Y}$ is model-based $\mathcal{G}$-invariant, this implies that $\mu_{Y|X}(y|x) = \mu_{Y|X}(y|g(x))$ for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. From this property of the posterior, it is direct to show that there exists $r(\cdot)$, which is solution of (10) (for the joint model $\mu_{X,Y}$) that is functional $\mathcal{G}$-invariant (Def. 25). ■

Appendix I. The Construction of $\mathcal{R}(\mu,\epsilon)$

The model $\mathcal{R}(\mu,\epsilon)$ that solves the optimization problem in (16) is presented here for completeness. As this optimization problem is function of $\mu \in \mathcal{P}([M])$ and $\epsilon > 0$, the model $\mathcal{R}(\mu,\epsilon) \in \mathcal{P}([M])$ is an explicit function of these two elements. To find the simplest description of this object, we assume, without loss of generality, that $\mu(1) \geq \mu(2) \geq \dots \mu(M)$.

In this context, Ho and Verdú (2010) showed that the model $\mu^{\bar{\epsilon}} \equiv \mathcal{R}(\mu, prior(\mu) - \bar{\epsilon})$ for $\bar{\epsilon} \in [0, prior(\mu)]$ has the follows form:

$$\mu^{\bar{\epsilon}}(1) = \mu(1) + \bar{\epsilon} \leq 1 \tag{182}$$
$$\mu^{\bar{\epsilon}}(2) = \theta$$
$$\dots$$
$$\mu^{\bar{\epsilon}}(K) = \theta \tag{183}$$
$$\mu^{\bar{\epsilon}}(K+1) = \mu(K+1)$$
$$\dots$$
$$\mu^{\bar{\epsilon}}(M) = \mu(M), \tag{184}$$

where $K \in \{2, .., M\}$ and $\theta \in (0, \mu(1))$ are functions of μ and $\bar{\epsilon}$ meeting the following condition

$$\sum_{j=2}^{K} (\mu(i) - \theta) = \bar{\epsilon} > 0. \tag{185}$$

Importantly, Ho and Verdú (2010) showed that for any $\epsilon \leq prior(\mu)$, $\exists \theta \in (0, \mu(1))$ and $K \in \{2, .., M\}$ that meet the condition in (185) for $\bar{\epsilon} = prior(\mu) - \epsilon$. From the estructure of $\mu^{\bar{\epsilon}}$ in (182), it follows that if $\epsilon = prior(\mu)$ then $\mu^{\bar{\epsilon}} = \mu$. On the other hand, if we assume that $\epsilon < prior(\mu)$ then $\mu^{\bar{\epsilon}} \neq \mu$ and $H(\mu) > H(\mu^{\bar{\epsilon}})$, which implies that $f(\mu, \epsilon) > 0$ in (16) in this regime.

Finally, from the closed-from expressions in (182)-(184) we have that:

$$\begin{aligned} f(\mu, \epsilon = prior(\mu) - \bar{\epsilon}) &= H(\mu) - H(\mu^{\bar{\epsilon}}) \\ &= \mu(1) \log \frac{1}{\mu(1)} - (\mu(1) + \bar{\epsilon}) \log \frac{1}{\mu(1) + \bar{\epsilon}} \\ &+ \sum_{j=2}^{K(\mu,\epsilon)} \mu(j) \log \frac{1}{\mu(j)} - (K(\mu, \epsilon) - 1)\theta(\mu, \epsilon) \log \frac{1}{\theta(\mu, \epsilon)} \geq 0, \end{aligned} \tag{186}$$

where in this last expression we show the fact that the parameters K and θ are explicit function of μ and ϵ.

References

A. Achille and S. Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19:1–34, 2018a.

A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897 – 2905, January 2018b.

A. Alemi, I. Fisher, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *The Thirteenth International Conference on Learning Representations*, pages 24–26, April 2017.

R.A. Amjad and B. C. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10.1109/TPAMI.2019.2909031, 2019.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013.

T. Berger. *Rate Distortion Theory*. Prentice Hall, 1971.

A. Berlinet and I. Vajda. On asymtotic sufficiency and optimality of quantizations. *Journal of Statistical Planning and Inference*, 136:4217–4238, 2005.

B. Bloem-Reddy and Y. W. Teh. Probabilistic symmetry and invariant neural networks. *Journal of Machine Learning Research*, 21:1–61, 2020.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

O. Bousquet, S. Boucheron, and G. Lugosi. *Theory of Classification: A Survey of Recent Advances*. ESAIM: Probability and Statistics, URL:http:/www.emath.fr/ps, 2004.

L. Breiman. *Probability*. Addison-Wesley, 1968.

G. Chechik, A. Globersonand, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Interscience, New York, second edition, 2006.

I. Csiszár and P. C. Shields. *Information Theory and Statistics: A Tutorial*. Now Inc., 2004.

G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partition of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

Y. Dubois, B. Bloem-Reddy, K. Ullrich, and C. J. Maddison. Lossy compression for losless prediction. In *Neural compression workshop at ICLR*, pages 1–26, 2021.

M. L. Eaton. Group invariance in application in statistics. In *Regional Conference Series in Probability and Statistics Volume 1*. Instituto of Mathematical Statistics and American Statistical Asociation, 1989.

M. Feder and N. Merhav. Relationship between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, January 1994.

B. Frenay, G. Doquire, and M. Verleysen. Theoretical and empirical study on the potential inadequancy of mutual information for feature selection in classification. *Neurocomputing*, 112:64–78, 2013.

M. P. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *Ann. Math. Statist.*, 41:1344–1346, 1970.

Z. Goldfeld and Y. Polyanskiy. The information bottleneck method problem and its applications in machine learning. *IEEE Journal on Seleted Areas in Information Theory*, 1 (1):19–38, May 2020.

M. Gonzales, J. F. Silva, M. Videla, and M. E. Orchard. Data-driven representations for testing independence: Modeling, analysis and connection with mutual information estimation. *IEEE Transactions on Signal Processing*, 70:158–173, 2022.

R. M. Gray. *Source Coding Theory*. Norwell, MA: Kluwer Academic, 1990a.

R. M. Gray. *Entropy and Information Theory*. Springer - Verlag, New York, 1990b.

R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, second edition edition, 2009.

P. R. Halmos. *Measure Theory*. Van Nostrand, New York, 1950.

S. Ho and S. Verdú. On the interplay between conditional entropy and error probability. *IEEE Transactions on Information Theory*, 56(12):5930–5942, December 2010.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, pages 1–14, 2014.

F. Liese, D. Morales, and I. Vajda. Asymptotically sufficient partition and quantization. *IEEE Transactions on Information Theory*, 52(12):5599–5606, 2006.

G. Lugosi and A. B. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

A. B. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

S. Prasad. Bayesian error-based sequeneces of statistical information bounds. *IEEE Transactions on Information Theory*, 61(9):5052–5062, September 2015.

A. Renyi. On the dimension and entropy of probability distribution. *Acta Mathematica Hingarica*, 10(1-2), March 1959.

J. Rotman. *An Introduction to the Theory of Groups*, volume 148 of *Graduate Texts in Mathematics*. Springer - Verlag, New York, 4th edition, 1995.

S. Shalev-Shwartz, O. Shamir, and N. Srebro. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

J. F. Silva and S. Narayanan. Universal consistency of data-driven partitions for divergence estimation. In *IEEE International Symposium on Information Theory*. IEEE, 2007.

J. F. Silva and S. Narayanan. Non-product data-dependent partitions for mutual information estimation: Strong consistency and applications. *IEEE Transactions on Signal Processing*, 58(7):3497–3511, July 2010a.

J. F. Silva and S. Narayanan. Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11):3180 – 3198, November 2010b.

J. F. Silva and S. Narayanan. Complexity-regularized tree-structured partition for mutual information estimation. *IEEE Transactions on Information Theory*, 58(3):1940 – 1952, 2012.

D. J. Strouse and D.J. Schwab. The deterministic information bottleneck. *Mass. Inst. Tech. Neural Computing*, 26(1611-1630), 2017.

M. Tegmark and T. Wu. Pareto-optimal data compression for binary classifiaction tasks. *Entropy*, 22(7):1–27, 2019.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *Information Theory Workshop*, pages 1–5, 2015.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the Thirty-seventh Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, September 1999.

I. Vajda. On convergence of information contained in quantized observations. *IEEE Transactions on Information Theory*, 48(8):2163–2172, 2002.

S.R.S. Varadhan. *Probability Theory*. American Mathematical Society, 2001.

H. Xu and S. Mannor. Robustness and generalization. *Machine Learning*, 86(391–423), 2012.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *Advances in Neural Information Processing System 30*, pages 3391–3401, 2017.

A. Zaidi, I. Estella-Aguerri, and S. Shamai. On the information bottleneck problems: Models, connections, applications and information theoretic views. *Entropy*, 22(151):1–36, 2020.