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

Jorge F. Silva
Information and Decision System Group
Universidad de Chile

Felipe Tobar
Initiative for Data & Artificial Intelligence
Universidad de Chile

Abstract

Information-theoretic measures have been widely adopted in the design of features for learning and decision problems. Inspired by this, we look at the relationship between **i)** a weak form of information loss in the *Shannon* sense and **ii)** operational loss in the *minimum probability of error* (MPE) sense when considering a family of lossy continuous representations of an observation. 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 lossy representation (quantization) instead of the original raw observation. From this, our main result shows that a specific form of vanishing information loss (a weak notion of asymptotic informational sufficiency) implies a vanishing MPE loss (or asymptotic operational sufficiency) when considering a family of lossy continuous representations. Our theoretical findings support the observation that the selection of feature representations that attempt to capture informational sufficiency is appropriate for learning, but this design principle is a rather conservative if the intended goal is achieving MPE in classification. On this last point, we discuss about studying weak forms of informational sufficiencies to achieve operational sufficiency in learning settings.

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., there are many explanatory factors that interact in the expression of X beyond Y and, consequently, a lossy description (aka coding) U has the potential to fully 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, 1990a], 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 Markov separation structure arises is in problems with probabilisitic symmetries or invariances with respect to 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,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

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, Vera et al., 2018] to optimize a tradeoff between relevance $I(U;Y)$ and compression $I(U;X)$ over a collection of probabilistic mappings (or channels) from X to a (latent) variable U [Zaidi et al., 2020]. There is also a deterministic version of the IB problem 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) deterministic mappings (or 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 $\lim_{i\to\infty} I(U_i;Y) = I(X;Y)$. On the other hand, 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 the following: If a method designs a collection of IS descriptions, is this collection also OS? More generally, is there a strictly weak notion of IS that implies OS?

To address these questions, 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 introduce a weak form of information loss[1] $I((r^*(X), U_i);Y) - I(U_i;Y) \leq I(X;Y) - I(U_i;Y)$ where $r^*(X)$ denotes the *MPE decision rule* [2] (a finite-size representation of X).

1.1 Contributions

Justifying our weak information loss selection, for the case of discrete representation (i.e., U_i is induced by a vector quantizer (VQ)), Theorem 1 presents a lower bound for $I((r^*(X), U_i);Y) - I(U_i;Y)$ as a function of the operation loss $\int_{\mathcal{U}_i}(1 - \max_{y\in\mathcal{Y}} \mu_{Y|U_i}(y|u))d\mu_{U_i}(u) - \int_{\mathcal{X}}(1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|x))d\mu_X(x) \geq 0$ attributed to the use of U_i instead of X in classifying Y. Using this bound, our main result (Theorem 2) shows that if $\{U_i\}_{i\geq 1}$ is *weakly information sufficient* (WIS), in the sense that $\lim_{i\to\infty} [I((r^*(X), U_i);Y) - I(U_i;Y)] = \lim_{i\to\infty} I(r^*(X);Y|U_i) = 0$, then $\{U_i\}_{i\geq 1}$ is operationally sufficient (OS) to discriminate Y (i.e., U_i achieves the MPE of X in the limit). In other words, a form of informational sufficiency (strictly weaker than IS mentioned above) implies a vanishing operation loss when $\{U_i\}_{i\geq 1}$ is a family of general continuous representations of X.

On the technical side, we derive Theorem 2 using the bound presented in Theorem 1: i.e., the argument goes from discrete (VQ) to continuous representations. In particular, we build the argument from the scenario of discrete (finite alphabet) representations to prove Theorem 2 in the general continuous (in Section 4.2). The proofs of Theorems 2 and 1 rely on two important information theoretic results: The first by Ho and Verdú [2010] that characterizes, using a specific rate-distortion function [Cover and Thomas, 2006], a tight upper bound for the conditional entropy (equivocation entropy) given an error probability and the second by Liese et al. [2006] on asymptotic sufficient partitions for mutual information.

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* [3] 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 interplay 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 of Y (μ_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

[1] This loss is formally introduced in Definition 3.

[2] $r^*(\cdot)$ is formally introduced in (9).

[3] $H(Y|X) \leq h(\ell(\mu_{X,Y})) + \ell(\mu_{X,Y})\log(|\mathcal{Y}| - 1)$ where $h(r) = -r\log(r) - (1-r)\log(1-r)$ is the binary entropy [Cover and Thomas, 2006].

when $\epsilon \leq (1 - \max_{y \in \mathcal{Y}} \mu_y(y))$. These results offer tight (achievable) bounds between conditional entropy and error probability, thus providing refined and specialized versions of *Fano's type of bounds*[4] [Ho and Verdú, 2010]. A relevant corollary of these bounds is the fact 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, in cases when the classification task is almost perfect or degenerate (zero probability of error), the interplay between error probability and conditional entropy is rather evident (zero error ⇔ zero conditional entropy). This interplay, however, is less evident for the majority of 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 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 (models) induced by lossy representations of X.

1.3 Organization

The rest of the paper is organized as follows. Sections 2 and 3 formalize our main question and introduce notation, required concepts and preliminary results. Section 4 presents the statement and interpretations of the main asymptotic result (Theorem 2). The proof of Theorem 2 is sketched in Section D. Discussion, final remarks and extensions are elaborated in Sections 5 and 5.1. The proofs of the main two results of this paper and the presentation of supporting technical arguments are relegated to the Supplementary Material.

2 PRELIMINARIES

Let us consider a decision problem expressed in terms of the joint model (probability) $\mu_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ of a vector (X, Y) where Y takes values in a finite space $\mathcal{Y} = \{1, .., M\}$ (e.g., a *class label*) and X takes values in a continuous 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 the model $\mu_{X,Y}$, is expressed by

$$\ell(\mu_{X,Y}) \equiv \int_{\mathcal{X}} (1 - \max_{y \in \mathcal{Y}} \mu_{Y|X}(y|x)) d\mu_X(x), \qquad (1)$$

where $\mu_{Y|X}(\cdot|x)$ denotes the *probability mass function* (pmf) of Y conditional to 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) [Feder and Merhav, 1994, Ho and Verdú, 2010] — is

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

where

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

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

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

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

2.1 Representations (Encoder) 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 respected sigma field denoted by $\mathcal{B}(\mathcal{U})$. In general, we are interested in the case of a lossy mapping $\eta(\cdot)$. To begin our analysis, particular attention will be given to the relevant case where $|\mathcal{U}| = K < \infty$, meaning that $\eta(\cdot)$ is a *vector quantizer* (VQ) of X. This VQ induces the following finite partition on $\mathcal{X}$ of size K:[5]

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

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

In general, 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 naturally to $\mu_{U,Y}$, where i) $\ell(\mu_{U,Y})$ is the MPE of predicting Y from U, and ii) $\mathcal{I}(\mu_{U,Y}) = I(U; Y)$ 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 naturally by

$$\mathcal{I}(\mu_{X,Y}) - \mathcal{I}(\mu_{U,Y}) = I(X; Y|U) \geq 0, \qquad (5)$$

where the identity in (5) comes from the chain rule of MI and the definition of the conditional MI [Gray,

[4] Indeed, these bounds were extended to countably-infinite alphabets, a regime for which Fano's original inequality has not been defined.

[5] The main result of this work is for continuous representations. However, studying the case of finite VQs is instrumental as elaborated in Sections 4.1 and D.

1990a, 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.,

$$\ell(\mu_{U,Y}) - \ell(\mu_{X,Y}) \geq 0. \tag{6}$$

In the following sections, we will study the interplay between a relaxed (weak) information loss expression (of the form in (5)) and the operation loss introduced in (6) in different contexts.

3 FORMALIZATION AND BASIC RESULTS

Let us consider a family of mappings $\eta_i : \mathcal{X} \to \mathcal{U}_i$, 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 asymptotic definitions for informational and operational sufficiency, respectively.

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

$$\lim_{i\longrightarrow\infty} \ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y}). \tag{7}$$

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*

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

Let us introduce a weak version of IS for $\mu_{X,Y}$. For this, let us recall that the MPE rule (a sufficient statistics) is a quantizer of $\mathcal{X}$ of size $M = |\mathcal{Y}|$ given by[6]

$$\tilde{r}_{\mu_{X,Y}}(x) \equiv \arg\max_{y\in\mathcal{Y}} \mu_{Y|X}(y|x). \tag{9}$$

This rule induces both a (distribution dependent) partition of $\mathcal{X}$ given by[7]

$$\pi^* \equiv \left\{ A^*_y \equiv \tilde{r}^{-1}_{\mu_{X,Y}}(\{y\}), y \in \mathcal{Y} \right\}, \tag{10}$$

and a finite alphabet 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 $\{U_i\}_{i\geq 1}$, respectively) is said to be weakly information sufficient (WIS) for $\mu_{X,Y}$ if*

$$\lim_{i\longrightarrow\infty} \underbrace{\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y})}_{I(Y;\tilde{U}|U_i)\geq 0} = 0, \tag{11}$$

where $I(Y;\tilde{U}|U_i)$ is the conditional MI between Y and $\tilde{U}$ given U_i [Cover and Thomas, 2006].

3.1 Preliminary Analysis

Let us first consider the discrete case where $\mathcal{U}_i = \{1, .., k_i\}$ for any $i \geq 1$. In this context, we can elaborate expressions for $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y})$ and $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y})$. For this, let us consider the finite partition induced by the mapping $\eta_i(\cdot)$ which is given by

$$\pi_{\eta_i} \equiv \left\{ B_{i,j} \equiv \eta_i^{-1}(\{j\}), j \in \mathcal{U}_i = \{1, .., k_i\} \right\}, \tag{12}$$

where $\eta_i(x) = \sum_{j\in\mathcal{U}_i} \mathbf{1}_{B_{i,j}}(x) \cdot j$.

The following results present useful expressions for the losses in (6) and (5) in terms of the model $\mu_{X,Y}$ and the cells of π^* and π_{η_i}, respectively.

Proposition 1 [8] $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{B_{i,j}\in\pi_{\eta_i}} \mu_X(B_{i,j}) \cdot g(\mu_{X,Y}, B_{i,j})$ *where*

$$g(\mu_{X,Y}, B_{i,j}) \equiv \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - \sum_{A^*_u\in\pi^*} \frac{\mu_X(A^*_u \cap B_{i,j})}{\mu_X(B_{i,j})} \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|A^*_u \cap B_{i,j})\right] \geq 0. \tag{13}$$

The operation loss in Proposition 1 is expressed as the weighted sum of the terms $\{g(\mu_{X,Y}, B_{i,j})\}_{B_{i,j}\in\pi_{\eta_i}}$, each one of them associated with a non-negative contribution (in the loss) indexed by individual cells of π_{η_i}.

Remark 1 *The term $g(\mu_{X,Y}, B_{i,j}) \geq 0$ 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 $(\mu_{\tilde{U},Y|X}(u, y|B_{i,j}) \equiv \frac{\mu_{X,Y}(A^*_u\cap B_{i,j}\times\{y\})}{\mu_X(B_{i,j})})_{(u,y)\in\mathcal{Y}^2}$ in $\mathcal{P}(\mathcal{Y}\times\mathcal{Y})$.*[9]

On the information loss, instead of looking at $I(X;Y|U_i)$ in (5), we decided to consider the MI loss of observing U_i with respect to a re-defined reference

[6] The optimal rule in (9) is not unique. If for some x many y achieves the minimum in (9), we select the smallest one to define $\tilde{r}_{\mu_{X,Y}}(x)$.

[7] where $\tilde{r}_{\mu_{X,Y}}(x) = \sum_{y\in\mathcal{Y}} \mathbf{1}_{A^*_y}(x) \cdot y$.

[8] The proof is presented in the Supplementary Material.

[9] This Bayesian gain interpretation of the term $g(\mu_{X,Y}, B_{i,j})$ will be central for the results in Section 4.

case where we observe the pair $(\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X), U_i)$, which is a deterministic function of X. Intuitively, this choice is based on the observation that $\tilde{U}$ is a sufficient statistic for X in the operational MPE sense, see Eq.(9). Consequently, our re-defined information loss is

Proposition 2 [10] *$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot I(\tilde{U};Y|X \in B_{i,j})$, where*

$$\begin{aligned} I(\tilde{U};Y|X \in B_{i,j}) \equiv \mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \\ \sum_{A_u^* \in \pi^*} \frac{\mu_X(A_u^* \cap B_{i,j})}{\mu_X(B_{i,j})} \mathcal{H}(\mu_{Y|X}(\cdot|A_u^* \cap B_{i,j})) \end{aligned} \tag{14}$$

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

Alternatively, we have that $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = I(\tilde{U}, U_i; Y) - I(U_i;Y) = I(\tilde{U};Y|U_i)$.

Remark 2 *Based on Remark 1, it is worth noting the conceptual connection between $g(\mu_{X,Y}, B_{i,j})$ in (13), which is the prior risk minus the posterior risk in the MPE sense condition on $\{X \in B_{i,j}\}$, and $I(\tilde{U};Y|X \in B_{i,j})$ in (14), which is the prior minus the posterior Shannon entropy condition on the same event $\{X \in B_{i,j}\}$.*

4 MAIN RESULTS

Before presenting the main result of this work (Theorem 2), it is relevant to find a lower bound on the information loss expressed in Proposition 2 as a function of the operation loss expressed in Proposition 1. For that, we introduce the following instrumental lemma:

Lemma 1 *[Ho and Verdú, 2010, Th.4] Let us consider Y a random variable in $\mathcal{Y} = \{1,..,M\}$ and a finite observation space $\mathcal{X}$ such that $|\mathcal{X}| \geq M$. If we denote by $\mathcal{P}(\mathcal{X}|\mathcal{Y})$ the collection of conditional probabilities from $\mathcal{Y}$ to $\mathcal{X}$ (or channels), 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*

$$\begin{aligned} f(\mu_Y, \epsilon) &\equiv \min_{\rho_{X|Y}\in\mathcal{P}(\mathcal{X}|\mathcal{Y}) \ st. \ \ell(\rho_{X|Y}\mu_Y)=\epsilon} \mathcal{I}(\rho_{X|Y}\mu_Y) \\ &= \mathcal{H}(\mu_Y) - \mathcal{H}(\mathcal{R}(\mu_Y, \epsilon)) \geq 0, \end{aligned} \tag{15}$$

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 both μ_Y and ϵ. [11]

This result offers a tight (achievable) lower bound on the minimum MI achieved by a family of joint models in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that satisfies 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 a fixed model $\mu_Y \in \mathcal{P}(\mathcal{Y})$. Importantly, for the non-trivial case when $\epsilon < (1 - \max_{y\in\mathcal{Y}} \mu_Y(y))$, Ho and Verdú [2010] show 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 show that $\mathcal{R}(\mu_Y, \epsilon) = \mu_Y \Rightarrow f(\mu_Y, \epsilon) = 0$ [Ho and Verdú, 2010]. [12]

Remark 3 *Considering a discrete observation X such that $(X, Y) \sim \mu_{X,Y}$, Lemma 1 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 (15), we have that*

$$\begin{aligned} I(X;Y) = \mathcal{I}(\mu_{X,Y}) &\geq f(\mu_Y, \ell(\mu_{X,Y})) \\ &= H(Y) - \mathcal{H}(\mathcal{R}(\mu_Y, \ell(\mu_{X,Y}))). \end{aligned} \tag{16}$$

Importantly, the bound in (16) 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., a zero gain in MPE (from the prior) when observing X.

4.1 A Non-Asymptotic Result

Returning to our original mixed continuous-discrete setting (Section 2), the application of Lemma 1 in the context of our weak information loss vs. operation loss analysis offers the following result:

Theorem 1 *Let us consider our model $\mu_{X,Y}$ and a finite alphabet (vector quantizer) lossy representation U_i (induced by $\eta_i(\cdot)$) of X, then*

$$\begin{aligned} &\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) \geq \\ &\sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \times \\ &\left[\mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B_{i,j}), \epsilon_{i,j})\right] \end{aligned} \tag{17}$$

where

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

and $g(\mu_{X,Y}, B_{i,j})$ is in (13).

[10] The proof is presented in the Supplementary Material.

[11] The closed-form expression of the probability $\mathcal{R}(\mu_Y, \epsilon)$ is presented in [Ho and Verdú, 2010]. This expression is presented in the Supplementary Material (see Eqs.(53)-(55)).

[12] In information theory, the function $f(\mu_Y, \epsilon)$ in (15) 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, 1990b].

Remarks and implications of Theorem 1:

1. The lower bound on the information loss in (17) is an explicit function of the decomposition of the operation loss presented below in (18). On the proof, the expression in (17) comes from interpreting the operational loss in (18) as the sum of some posterior minus prior MPE gains (see Remark 1) and the application of Lemma 1 in this context (see the proof of this result in the Supplementary Material).

2. **Corollary 1** *Let us assume a positive operation loss, i.e.,* $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) =$
$$\sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j})\cdot g(\mu_{X,Y},B_{i,j}) > 0 \qquad (18)$$
then from Theorem 1 it follows that $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) > 0$.

   Proof of Corollary 1: Assuming that $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) > 0$ in (18), this implies that at least one component j of the sum satisfies that $g(\mu_{X,Y},B_{i,j}) > 0 \Leftrightarrow \epsilon_{i,j} < \left[1-\max_{y\in\mathcal{Y}}\mu_{Y|X}(y|B_{i,j})\right]$ (see (18)). Then Lemma 1 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 and (17) suffice to show that
$$\mathcal{I}(\mu_{X,Y}) - \mathcal{I}(\mu_{U_i,Y}) \geq \mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) > 0. \qquad (19)$$
The first inequality in (19) comes from the fact that $(\tilde{U},U_i)$ is a deterministic function of X and the second comes from (17). ■

   Therefore, a non-zero operation loss on using U_i instead of X (stated in (18)) implies a respective non-zero weak information loss as stated in (19).

3. Corollary 1 implies that if $U_i = \eta_i(X)$ (for some finite i) is weakly information sufficient in the sense that $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = I(\tilde{U};Y|U_i) = 0$, then $\ell(\mu_{U_i,Y}) = \ell(\mu_{X,Y})$, i.e., η_i (and U_i) is operational sufficient for $\mu_{X,Y}$.

4. It is worth noting that for a large clases of models (continuous in nature), U_i being weakly information sufficient, i.e., $I(\tilde{U};Y|U_i) = 0$, is strictly weaker than asking that U_i is information sufficient for $\mu_{X,Y}$, i.e., $I(X;Y|U_i) = 0$. In fact, $I(X;Y|U_i) = 0$ implies that $I(\tilde{U};Y|U_i) = 0$ from the observation that $\tilde{U}$ is a deterministic function of X and the chain rule of the MI [Cover and Thomas, 2006], but the converse result is not true in general.[13]

5. The difference between the pure 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 4.3 and its non-zero discrepancy is illustrated by an example in Section 4.4.

4.2 The Asymptotic Result

The following is the main asymptotic result of this work that shows that a family of weakly IS representations (continuous in general) for $\mu \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is operation sufficient for μ. This result can be interpreted as a non-trivial asymptotic extension of Corollary 1 (from Theorem 1).

Theorem 2 *Let* $\{U_i\}_{i\geq 1}$ *be a sequence of representations for* X *obtained from* $\{\eta_i(\cdot)\}_{i\geq 1}$*. 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 about the statement of Theorem 2 and its interpretation:

1. The proof of Theorem 2 (Section 4 in the Supplemental Material) shows that if a family of representations $\{U_i\}_{i\geq 1}$ is not operationally sufficient, i.e., $\liminf_{i\to\infty}\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) > 0$, then
$$\liminf_{i\to\infty}\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = \liminf_{i\to\infty} I(\tilde{U};Y|U_i) > 0. \qquad (20)$$

2. The condition (11) (WIS in Definition 3) 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 the uncertainty of Y. In general for continuos models, we have that $I(\tilde{U};Y) < I(X;Y)$ because $\tilde{U}$ is an M size quantized version of X (see Eq.(10)). Then, the sufficient condition stated in (11) is strictly weaker (for a large class of models) than asking for informational sufficiency (Definition 2).

3. The condition in (11) and Theorem 2 further emphasize the fact that achieving pure sufficiency in the Shannon sense is very conservative if the operational objective is classification, as a strictly weaker notion does exist that guarantees operational sufficiency. An example is presented in Section 4.4 that illustrates this point.

4. Complementing the previous point, it is evident that $\{U_i\}_{i\geq 1}$ being operationally sufficient for $\mu_{X,Y}$ does not imply that $\{U_i\}_{i\geq 1}$ is information sufficient for $\mu_{X,Y}$, in general. We illustrate this with an example in Section 4.4. Conversely, if $\{U_i\}_{i\geq 1}$ is not OS, then $\limsup_{i\to\infty}\mathcal{I}(\mu_{U_i,Y}) <$

[13] In contrast, examples for $\mu_{X,Y}$ can be constructed (discrete in nature) where $I(X;Y) = I(\tilde{U};Y)$. Here, $\tilde{U}$ (a discrete variable of size M) is IS for $\mu_{X,Y}$. In this trivial discrete context, $I(X;Y|U_i){=}I(\tilde{U};Y|U_i)$ independent of U_i.

$\mathcal{I}(\mu_{X,Y})$ considering that $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) \leq \mathcal{I}(\mu_{X,Y})$ for any $i \geq 1$ and (20).

5. Finally, it is worth mentioning that WIS, as a condition on $\{\eta_i(\cdot)\}_{i\geq 1}$, is theoretically interesting for the reasons mentioned in previous points, but it is unnatural in terms of its practical adoption for feature design in a learning context. The reason is that the reference representation $\tilde{U}$ (used in (11)) is a function of the model $\mu_{X,Y}$, which is by construction unavailable in learning. This limitation motivates further research on extensions of Theorem 2 into learning settings as discussed in Section 5.

4.3 How much weaker is WIS than IS?

On the significance of our main result (Theorem 2), a key aspect is to analyze and evaluate much weaker is the WIS condition used in Theorem 2 from the traditional IS. The analysis implies looking at the differences between the information losses, i.e., the difference between $I(X;Y)$ and $I((\tilde{U},U_i);Y)$. On this, we could say that:

- The analysis of 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 weak information loss uses $\tilde{U}$ (a quantized version of X of size M) as a reference, while IL uses X, which is a continuous random variable in the context of our general model $\mu_{X;Y}$.
- We know from information theory that $I(X;Y)$ is the supremum of the discrete MI between $\eta(X)$ and Y over all possible finite-size quantizers $\eta(\cdot)$ [Liese et al., 2006, Silva and Narayanan, 2010a, Vajda, 2002] (result in the Supplemental). The scenario where MI is not achieved by any finite size version of X makes the model $\mu_{X,Y}$ continuous from a MI point of view [Liese et al., 2006, Silva and Narayanan, 2010a, Vajda, 2002]. In contrast, a model where a quantized version of X achieves the MI between X and Y makes the model $\mu_{X,Y}$ discrete from a MI point of view [Cover and Thomas, 2006].
- Assuming the non-trivial case that $\mu_{X;Y}$ (the model) is continuous, i.e., $I(X;Y)$ is not achieved by any finite-alphabet function (or vector quantization) of X, we have that for any representation $\eta_i(\cdot)$ (U_i) that is a VQ: $I((\tilde{U},U_i);Y) < I(X;Y)$. Consequently, WIL is strictly smaller than IL for the rich case where we have a continuous model and finite alphabet representations.
- On the previous point, the continuous scenario for $\mu_{X;Y}$ is an essential case study for the analysis presented in this paper as we do not impose any structural assumptions on $\mu_{X,Y}$. Also, in practical domains of continuous observations, it is reasonable to consider that a quantized (digital) version of X induces a non-zero loss of mutual information about Y.

4.4 An illustrative Example

To illustrate the gap between WIS and IS and the potential significance of our result, here we present a simple construction to analyze the interplay between IS vs. OS and WIS vs. OS.

- Y takes two values in $\{1,2\}$ 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).
- 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: $\pi_1 = \{(-\infty,-1/2),[-1/2,1/2],(1/2,\infty)\}$; $\pi_2 = \{(-\infty,-1/4),[-1/4,1/4],(1/4,\infty)\}$; $\ldots$ $\pi_i = \big\{(-\infty,-1/2^i),[-1/2^i,1/2^i],(1/2^i,\infty)\big\}$,....
- If we denote by A_i^1, A_i^2 and A_i^3 the cells of π_i, these produce 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 that $\lim_{i\to\infty} I(U_i;Y) < I(X;Y)$. In other words, the collection $\{U_i\}_{i\geq 1}$ is not information sufficient: i.e, $I(X;Y) - I(U_i;Y)$ 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 2, $\{U_i\}_{i\geq 1}$ is OS (Def. 1) but not IS (Def.2).

This simple construction offers a scenario where the difference between IL and WIL is strictly positive and relevant for any i. This discrepancy is non-trivial when i grows: WIL tends to zero, but IL does not. Indeed, WIS (Def. 3) is strictly weaker than IS (Def.2) in this example. Furthermore, we observe that IL as a criterion is blind on predicting the quality of $\{U_i\}_{i\geq 1}$ to achieve the MPE in (1). In the extended version of this paper, we show other constructions (VQs) and models where the same finding illustrated in this example is experimentally observed [Silva et al., 2021].

4.5 Overview of the Proof of Theorem 2: from Discrete to Continuous Representations

The proof of Theorem 2 is divided in two main stages (details in Section D of the Supplementary Material). The first stage (in Section D.1, Suplemental) restricts the analysis to the important case of finite alphabet representations, or vector quantizers of X. In this discrete context, we use results from information theory to show that WIS implies OS (see Theorem 3 and its proof in Sections E and F of the Supplementary Material). The decision to begin studying the case of finite alphabet representations was essential because it offers a path to adopt concrete results on the interplay between probability of error and conditional entropy only available for discrete random variables (see Section 4.1 and Section D.1 of the Supplementary Material).

In the second stage of the proof (in Section D.2 of the Supplementary Material), we make a connection between the discrete (Theorem 3 in the Supplementary Material) and the continuous result (Theorem 2). Importantly, the finite alphabet result is used as a building block to extend the proof argument to the continuous case stated in Theorem 2. For this objective, results on information sufficient partitions for mutual information were instrumental [Liese et al., 2006] (details in Sections D.2 of the Supplemental Material).

5 SUMMARY, DISCUSSION AND EXTENSIONS

This work offers new results to understand the interplay between information loss (in the Shannon sense) and operational loss (in the classical MPE sense) when considering a general family of lossy representations of an observation vector X in $\mathbb{R}^d$. Our main asymptotic result (Theorem 2) supports the idea that creating a family of information sufficient representations is an adequate criterion in the sense that these representations have a vanishing residual error with respect to the MPE decision acting on X to classify (a class) Y. On the other hand, our result also shows that pure informational sufficiency (in the sense of Def. 2) is a conservative criterion. Indeed, Theorem 2 shows that a strictly weaker notion of informational sufficiency (in the form of Def. 3) suffices to obtain the required operational result.

To give practical significance to our main result, we have worked on extensions of Theorem 2 in a learning setting where $\mu_{X,Y}$ is unknown, but it belongs to a collection of models Λ (prior knowledge). In an extended version of this paper [Silva et al., 2021, Sec. VII], we studied how the structure of Λ can be used to determine less conservative (and non-oracle) weak forms of informational sufficiency that could be adopted by algorithms that select representations from data. We studied the case where Λ is the family of invariant models (invariant to the action of a compact group) [Bloem-Reddy and Teh, 2020, Dubois et al., 2021], where it is possible to determine "*a non-oracle*" surrogate of $r^*(\cdot)$ (in Theorem 2) that extends our result (WIS implies OS) in a learning scenario [Silva et al., 2021].

Finally, on the optimality of the WIS condition in Theorem 2, in the extended version of this paper [Silva et al., 2021, Sec.V], we proved that when the MAP rule in (9) is unique (almost surely w.r.t. to the model $\mu_{X,Y}$), then "*WIS is equivalent to OS*". This last result shows a context where WIS (as a condition) is tight and optimal, in the sense that no weaker representation condition on $\{U_i\}_{i\geq 1}$ could be found to guarantee OS.

5.1 A Broader Perspective on the Application of these Results

The analysis presented in this work about the interplay between vanishing information and operation loss offers relevant insight in learning settings. First, 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 $I(U;Y)$. This is a widely adopted criterion in representation learning, in the form of maximizing empirical versions of the mutual information, or alternatively, minimizing empirical versions of the conditional entropy over a family of compressed representations of X [Amjad and Geiger, 2019, Alemi et al., 2017, Achille and Soatto, 2018b, Vera et al., 2018, Strouse and Schwab, 2017, Tegmark and Wu, 2019].

On the other hand, the implication of Theorem 2 is relevant in the sense that we show evidence that pure IS could be very conservative as a condition to achieve OS in some scenarios. On this dimension, we introduce a weaker (strictly weaker in many cases) information sufficient condition that imply OS for classification.

We know that our main result (WIS=>OS) by itself does not offer a direct practical strategy for representation learning. However, our findings could motivate practical avenues of research: for instance, finding "*non-*

oracle" information losses, inspirited by WIS, weaker than IS (for some models or family of models) that could be adopted in practice to learn sufficient representation from data in the MPE sense. In this direction, we have worked on extending Theorem 2 to come out with a non-oracle WIS condition under some prior information ($\mu_{X,Y}$ is invariant to the action of a compact group of transformations). These results are presented in an extended version of this paper in [Silva et al., 2021].

Concerning the notion of informational sufficiency studied in this paper, there is an interesting connection with the analysis and results used for mutual information estimation. On the approximation error analysis of this estimation (learning) problem, there are numerous results in the literature guaranteeing that a collection of representations (partitions) are asymptotically sufficient for approximating the mutual information $I(X;Y)$ [Berlinet and Vajda, 2005, Liese et al., 2006, Vajda, 2002], or information sufficient in the language of this paper. These results offer concrete conditions and constructions for finite-size representations (VQ) that in light of Theorem 2 will be operationally sufficient for classification. In this context, it is worth noting the family of data-driven constructions studied in [Silva and Narayanan, 2010a, 2007, 2010b, Vajda, 2002, Darbellay and Vajda, 1999, Gonzales and Silva, 2020, Gonzales et al., 2022]. These are data-driven representations (or representations learned from data) that are information sufficient (with probability one) and, consequently, they are operationally sufficient (with probability one) as a corollary our main result in Theorem 2. The formalizations of these connections are presented in the extended version of this paper in [Silva et al., 2021, Sec. VI].

Finally, we emphasize that our findings are purely from an information-theoretic perspective of sufficiency and, if put in practice, the features resulting from the ideas laid out here need to be exhaustively examined. We are at a time when automatic feature discovery has reached an unprecedented efficiency in terms of performance metrics, however, when dealing with sensible human-centered scenarios and datasets, operational efficiency should not be the unique criterion and we must also consider societal implications, fairness, and interpretability.

Acknowledgements

We would like to acknowledge support for this project from ANID-Chile (Fondecyt Grant 1210315 and 1210606), the Advanced Center for Electrical and Electronic Engineering (Basal Project FB0008) and the Center for Mathematical Modeling (Basal Projects ACE210010 & FB210005). The authors thank Diane Greenstein for editing and proofreading part of this material and the anonymous reviewers for proving numerous constructive comments that help improve this work.

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 *ICLR*, 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.

A. Berlinet and Igor 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.

Oliver 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.

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

Georges A. Darbellay and Igor 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 *at ICLR 2021 neural compression workshop*, pages 1–26, 2021.

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

Mauricio Gonzales and Jorge F Silva. Data-driven representations for testing independence: A connection with mutual information estimation. In *IEEE International Symposium on Information Theory*. IEEE, 2020.

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

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

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

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

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

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

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.

Jorge F. Silva and Shrikanth 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.

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

Jorge F Silva, Felipe Tobar, Mario Vicuna, and Felipe Cordova. Studying the interplay between information loss and operation loss in representations for classification. https://arxiv.org/abs/2112.15238, December 2021.

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 classification tasks. *Entropy*, 22(7): 1–27, 2019.

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.

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

M. Vera, P. Piantanida, and L.R. Vega. The role of information complexity and randomization in representation learning. https://arxiv.org/abs/1802.05355, 2018.

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

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.

Supplementary Material: On the Interplay between Information Loss and Operation Loss in Representations for Classification

A PROPOSITION 1

Proposition 1 $\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \sum_{B_{i,j}\in\pi_{\eta_i}} \mu_X(B_{i,j}) \cdot g(\mu_{X,Y}, B_{i,j})$ where

$$g(\mu_{X,Y}, B_{i,j}) \equiv \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - \sum_{A_u^*\in\pi^*} \frac{\mu_X(A_u^*\cap B_{i,j})}{\mu_X(B_{i,j})} \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|A_u^*\cap B_{i,j})\right]. \tag{21}$$

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 (6) conveniently:

$$\begin{aligned}\ell(\mu_{U_i,Y}) - \ell(\mu_{X,Y}) = \ell(\mu_{U_i,Y}) - \ell(\mu_{(\tilde{U},U_i),Y}) &= \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] \\ &- \sum_{A_u^*\in\pi^*} \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}\cap A_u^*) \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|A_u^*\cap B_{i,j})\right].\end{aligned} \tag{22}$$

Finally (13) follows directly from (22). □

B PROPOSITION 2

Proposition 2 $\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot I(\tilde{U}; Y|X\in B_{i,j})$, where

$$I(\tilde{U}; Y|X\in B_{i,j}) \equiv \mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \sum_{A_u^*\in\pi^*} \frac{\mu_X(A_u^*\cap B_{i,j})}{\mu_X(B_{i,j})} \mathcal{H}(\mu_{Y|X}(\cdot|A_u^*\cap B_{i,j})) \tag{23}$$

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

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

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) = H(Y) - \sum_{A_u^*\in\pi^*} \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}\cap A_u^*) \cdot \mathcal{H}(\mu_{Y|X}(\cdot|A_u^*\cap B_{i,j})). \tag{24}$$

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 (14) derives directly from these expressions. □

C PROOF OF THEOREM 1

Theorem 1 Let us consider our model $\mu_{X,Y}$ and a finite alphabet lossy representation U_i (induced by $\eta_i(\cdot)$) of X, then

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) \geq \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \left[\mathcal{H}(\mu_{Y|X}(\cdot|B_{i,j})) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B_{i,j}), \epsilon_{i,j})\right] \geq 0, \tag{25}$$

where $\epsilon_{i,j} = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - g(\mu_{X,Y}, B_{i,j})$ and $g(\mu_{X,Y}, B_{i,j})$ is in Eq.(13) (of the main paper).

Proof: Let us first look at the definition of $g(\mu_{X,Y}, B)$ in (13). This is a function of the model $\mu_{X,Y}$, the partition $\pi^* = \left\{A^*_y, y \in \mathcal{Y}\right\}$ in (10) and a set $B \subset \mathcal{X}$. In particular, we have that

$$g(\mu_{X,Y}, B) = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B)\right] - \sum_{A^*_u \in \pi^*} \mu_X(A^*_u|B) \cdot \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|A^*_u \cap B)\right]. \tag{26}$$

The first term on the RHS of (26) 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 (26) 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

$$v_{\tilde{X},\tilde{Y}}(u, y) \equiv \frac{\mu_{X,Y}(A^*_u \cap B \times \{y\})}{\mu_X(B)},\ \forall(u, y) \in \mathcal{Y}^2. \tag{27}$$

The second term in (26) is precisely $\ell(v_{\tilde{X},\tilde{Y}})$. Adopting Lemma 1 in this context, we can use its corollary in (16) to obtain that

$$\begin{aligned}\mathcal{I}(v_{\tilde{X},\tilde{Y}}) = I(\tilde{X}; \tilde{Y}) &\geq H(\tilde{Y}) - \mathcal{H}(\mathcal{R}(v_{\tilde{Y}}, \ell(v_{\tilde{X},\tilde{Y}})))\\ &= \mathcal{H}(v_{\tilde{Y}}) - \mathcal{H}(\mathcal{R}(v_{\tilde{Y}}, \ell(v_{\tilde{X},\tilde{Y}})))\\ &= \mathcal{H}(\mu_{Y|X}(\cdot|B)) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|B), \ell(\mu_{\tilde{X},\tilde{Y}}))) \qquad (28)\end{aligned}$$

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 (26) and the construction of $v_{\tilde{X},\tilde{Y}}$ in (27). The inequality in (28) is obtained as a function of $B \subset \mathcal{X}$, as it is used to construct $v_{\tilde{X},\tilde{Y}}$ in (27).

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

$$\mathcal{I}(\mu_{(\tilde{U},U_i),Y}) - \mathcal{I}(\mu_{U_i,Y}) = I(\tilde{U}; Y|U_i) = \sum_{B_{i,j}\in\pi_i} \mu_X(B_{i,j}) \cdot I(\tilde{U}; Y|X = B_{i,j}). \tag{29}$$

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 (27). Consequently, applying (28) 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 $\epsilon_{i,j} = \left[1 - \max_{y\in\mathcal{Y}} \mu_{Y|X}(y|B_{i,j})\right] - g(\mu_{X,Y}, B_{i,j})$. □

D THEOREM 2: FORM DISCRETE TO CONTINUOUS REPRESENTATIONS

Theorem 2 Let $\{U_i\}_{i\geq 1}$ be a sequence of representations for X obtained from $\{\eta_i(\cdot)\}_{i\geq 1}$. 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).

D.1 Discrete version of Theorem 2

Theorem 3 *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 3 is presented in Section E.

Technical remarks about the proof of Theorem 3:

1. The proof of this result uses a sample-wise version of the inequality presented in (17) (in Theorem 1) as a key element in the argument.

2. Another important technical element of the proof was characterizing and analyzing the following information object:
$$\mathcal{I}_{loss}(\epsilon, M) \equiv \min_{v \in \mathcal{P}^{\epsilon}([M])} \left\{ \mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon)) \right\}, \tag{30}$$
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 4 and Appendix F). To achieve this result, we derived an explicit lower bound for $\mathcal{I}_{loss}(\epsilon, M)$ function of ϵ and M.

D.2 Proof of Theorem 2

Proof: Without loss of generality, let us assume that $\eta_i : \mathcal{X} \to \mathcal{U}_i$ is such that $\mathcal{U}_i \subset \mathcal{U} = \mathbb{R}^q$ for some $q \geq 1$[14]. Here we use a result from the seminal work of Liese, Morales and Vajda [Liese et al., 2006] on asymptotic sufficient partition for MI. In particular, in the context of our work we have the following:

Lemma 2 *[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*
$$\lim_{i \to \infty} I(Y; m_{\pi_i}(\eta(X))) = I(Y; \eta(X)), \tag{31}$$
where
$$m_{\pi_i}(u) \equiv \sum_{A_l \in \pi_i} l \cdot \mathbf{1}_{A_l}(u) \in \{1, .., |\pi_i|\}, \ \forall u \in \mathcal{U} \tag{32}$$
denotes the lossy function (vector quantizer) induced by the partition $\pi_i = \{A_l, l = 1, .., |\pi_i|\}$.

Lemma 2 is a remarkable implication of the work by Liese et al. [2006]. This result shows the existence of a finite-size quantization family that approximates (universally) any well-defined MI on a continuous space in the sense presented in (31).

In the context of this argument, we can use the universal embedded quantization $\{\pi_i\}_{i \geq 1}$ of $\mathcal{U}$ stated in Lemma 2 to obtain as a direct corollary of Lemma 2 that for any $\eta_j : \mathcal{X} \to \mathcal{U}$
$$\lim_{i \to \infty} I((\tilde{U}, m_{\pi_i}(U_j)); Y) - I(m_{\pi_i}(U_j); Y) = I((\tilde{U}, U_j); Y) - I(U_j; Y) = I(\tilde{U}; Y | U_j) \geq 0, \tag{33}$$
where $U_j = \eta_j(X)$ and $\tilde{U} = \tilde{r}_{\mu_{X,Y}}(X) \in \mathcal{Y}$ (see Eq.(9)).

On the other hand, from the hypothesis that assumes that $\{\eta_j(\cdot)\}_{j \geq 1}$ is WIS, we have that
$$\lim_{j \to \infty} I(\tilde{U}; Y | U_j = \eta_j(X)) = 0. \tag{34}$$

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 (33) we have that for any $j \geq 1$ there exists $i_j^*(\epsilon_j, \eta_j) \geq 1$ sufficiently large such that[15]
$$\epsilon_j + I(\tilde{U}; Y | U_j) > \underbrace{I((\tilde{U}, m_{\pi_{i_j^*}}(U_j)); Y) - I(m_{\pi_{i_j^*}}(U_j); Y)}_{I(\tilde{U}; Y | m_{\pi_{i_j^*}}(U_j)))} > I(\tilde{U}; Y | U_j) - \epsilon_j. \tag{35}$$

In (35), it is worth noticing that $m_{\pi_{i_j^*}}(U_j) = m_{\pi_{i_j^*}} \circ \eta_j(X)$. Then, we can define
$$\tilde{\eta}_j \equiv m_{\pi_{i_j^*}} \circ \eta_j : \mathcal{X} \to \left\{1, .., \left|\pi_{i_j^*}\right| < \infty \right\}, \tag{36}$$
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 (36), satisfying that
$$\lim_{j \to \infty} I(\tilde{U}; Y | \tilde{\eta}_j(X))) = 0, \tag{37}$$

[14] 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.

[15] For what follows, we omitted the dependency on ϵ_j and $\eta_j(\cdot)$ in i_j^* to simplify the notation.

from (35), (34), and the fact $(\epsilon_n)_{n\geq 1}$ is $o(1)$. Therefore, (37) means that $\{\tilde{\eta}_j(\cdot)\}_{j\geq 1}$ is *weakly information sufficient* (Def. 3). Then, Theorem 3 implies that

$$\lim_{j\to\infty}\left[\ell(\mu_{\tilde{\eta}_j(X),Y})-\ell(\mu_{X,Y})\right]=0. \tag{38}$$

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 (38). □

E PROOF OF THEOREM 3

Let us begin introducing some preliminaries that will be used in the main argument in Section E.2.

E.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 1 and 2 and the interplay between them, determined in Theorem 1, we define the *information loss density* (ILD) and the *operational loss density* (OLD) associated with $\eta(\cdot)$ as follows:

$$\ell_\eta(x)\equiv\sum_{A\in\pi_\eta}\mathbf{1}_A(x)\cdot g(\mu_{X,Y},A)\geq 0,\ \forall x\in\mathcal{X} \tag{39}$$

$$\mathcal{I}_\eta(x)\equiv\sum_{A\in\pi_\eta}\mathbf{1}_A(x)\cdot I(\tilde{U};Y|X\in A)\geq 0,\ \forall x\in\mathcal{X}. \tag{40}$$

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

$$\begin{aligned}\mathbb{E}_X\left\{\ell_\eta(X)\right\}&=\ell(\mu_{U,Y})-\ell(\mu_{X,Y})\\ \mathbb{E}_X\left\{\mathcal{I}_\eta(X)\right\}&=\mathcal{I}(\mu_{(\tilde{U},U),Y})-\mathcal{I}(\mu_{U,Y}),\end{aligned} \tag{41}$$

where $U=\eta(X)$.

From the proof of Theorem 1, we obtain the following sample-wise inequality: for any $A\in\mathcal{B}(\mathcal{X})$

$$I(\tilde{U};Y|X\in A)\geq\mathcal{H}(\mu_{Y|X}(\cdot|A))-\mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|A),\mathit{prior}(\mu_{Y|X}(\cdot|A))-g(\mu_{X,Y},A))), \tag{42}$$

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

$$\mathcal{I}_\eta(x)\geq\mathcal{H}(\mu_{Y|X}(\cdot|\pi_\eta(x)))-\mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_\eta(x)),\mathit{prior}(\mu_{Y|X}(\cdot|\pi_\eta(x)))-\ell_\eta(x))). \tag{43}$$

Then $\mathcal{I}_\eta(x)$ (the ILD) 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.,

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

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 (40) and (41), it follows that

$$\lim_{i\longrightarrow\infty}\mathbb{E}_X\left\{\mathcal{I}_{\eta_i}(X)\right\}=0. \tag{44}$$

As $\mathcal{I}_{\eta_i}(x)\leq\log|\mathcal{Y}|$ (uniformly in i and x), the convergence in (44) 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 (41), the proof reduces to verify that

$$\lim_{i \longrightarrow \infty} \mathbb{E}_X \left\{ \ell_{\eta_i}(X) \right\} = 0. \tag{45}$$

Again $\ell_{\eta_i}(x)$ is uniformly bounded by 1, then the convergence in (45) is equivalent to the convergence in probability of the random sequence $(\ell_{\eta_i}(X))_{i \geq 1}$, i.e., for any $\epsilon > 0$

$$\lim_{i \to \infty} \mathbb{P}\left(\{ \ell_{\eta_i}(X) > \epsilon \} \right) = 0. \tag{46}$$

E.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, there exists $\epsilon > 0$ such that $\liminf_{i \to \infty} \mu_X(B^i_\epsilon) > 0$ where $B^i_\epsilon \equiv (\{x, \ell_{\eta_i}(x) > \epsilon\}) \subset \mathcal{X}$. Then, we can pick $\delta > 0$ and $N > 0$, such that for any $i \geq N$,

$$\mu_X(B^i_\epsilon) \geq \delta. \tag{47}$$

Using the definition of the function $\mathcal{R}(v, \epsilon)$ in [Ho and Verdú, 2010] (see (15) in Lemma 1), for any $v \in \mathcal{P}(\mathcal{Y})$, it follows — from the expression of $f(v, \epsilon)$ in (15) — that $\mathcal{H}(\mathcal{R}(v, \epsilon_1)) \geq \mathcal{H}(\mathcal{R}(v, \epsilon_2))$ when $\epsilon_1 \geq \epsilon_2$; therefore, from (43), if $x \in B^i_\epsilon$, we have that

$$\mathcal{I}_{\eta_i}(x) \geq \mathcal{H}(\mu_{Y|X}(\cdot|\pi_i(x))) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_i(x)), prior(\mu_{Y|X}(\cdot|\pi_i(x))) - \epsilon)) \tag{48}$$

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

The bound in (48) 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

$$\mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon)) \tag{49}$$

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 vs. operation loss problem:

$$\mathcal{I}_{loss}(\epsilon, M) \equiv \min_{v \in \mathcal{P}^\epsilon([M])} \left\{ \mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon)) \right\}, \tag{50}$$

where

$$\mathcal{P}^\epsilon([M]) \equiv \left\{ v \in \mathcal{P}([M]), prior(v) \geq \epsilon \right\}. \tag{51}$$

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 4 *$\forall M \geq 1$, 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 F of this Suplemental.

Returning to the main proof argument, by definition of the operation loss density in (39), 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 (51). Then using (48) and (50), for any $x \in B^i_\epsilon$ (considering that $\epsilon < \ell_{\eta_i}(x)$ if $x \in B^i_\epsilon$)

$$\begin{aligned} \mathcal{I}_{\eta_i}(x) &\geq \mathcal{H}(\mu_{Y|X}(\cdot|\pi_i(x))) - \mathcal{H}(\mathcal{R}(\mu_{Y|X}(\cdot|\pi_i(x)), prior(\mu_{Y|X}(\cdot|\pi_i(x))) - \epsilon)) \\ &\geq \min_{v \in \mathcal{P}^\epsilon([M])} \left\{ \mathcal{H}(v) - \mathcal{H}(\mathcal{R}(v, prior(v) - \epsilon)) \right\} = \mathcal{I}_{loss}(\epsilon, M), \end{aligned} \tag{52}$$

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 (51).

At this point, we use Theorem 4: we have that for any $x \in B^i_\epsilon$, $\mathcal{I}_{\eta_i}(x) \geq \mathcal{I}_{loss}(\epsilon, M) > 0$. 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 (47), we have that for any $i \geq N$ $\mu_X(A^i_{\bar{\epsilon}}) \geq \mu_X(B^i_{\bar{\epsilon}}) \geq \delta > 0$. This implies that $(\mathcal{I}_{\eta_i}(X))_{i \geq 1}$ does not converge to zero in probability, which from the argument presented in Section E.1 contradicts the fact that $\{\eta_i(\cdot)\}_{i \geq 1}$ is WIS. This concludes the proof of Theorem 3. □

F PROOF OF THEOREM 4

Proof: Given a probability $\mu \in \mathcal{P}^{\epsilon}([M])$ (see Eq.(51)), Ho and Verdú [2010] presented a closed-form analytical expression for $\mathcal{R}(\mu, prior(\mu) - \epsilon)$ (the details are presented in [Ho and Verdú, 2010]) appearing in the definition of $\mathcal{I}_{loss}(\epsilon, M)$ in (50). 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: [16]

$$\begin{aligned}
&\mu^{\epsilon}(1) = \mu(1) + \epsilon \leq 1 && (53)\\
&\mu^{\epsilon}(2) = \theta \\
&\ldots \\
&\mu^{\epsilon}(K) = \theta && (54)\\
&\mu^{\epsilon}(K+1) = \mu(K+1) \\
&\ldots \\
&\mu^{\epsilon}(M) = \mu(M). && (55)
\end{aligned}$$

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

$$\sum_{j=2}^{K} (\mu(i) - \theta) = \epsilon > 0, \tag{56}$$

which makes μ^{ϵ} a well-defined probability in $\mathcal{P}([M])$.[17]

Therefore, using (53), (54) and (55), we have that for any $\mu \in \mathcal{P}^{\epsilon}([M])$:

$$\begin{aligned}
\mathcal{H}(\mu) - \mathcal{H}(\mu^{\epsilon}) = {} & \mu(1) \log \frac{1}{\mu(1)} - (\mu(1) + \epsilon) \log \frac{1}{\mu(1) + \epsilon} \\
& + \sum_{j=2}^{K(\mu,\epsilon)} \mu(j) \log \frac{1}{\mu(j)} - (K(\mu,\epsilon) - 1) \cdot \theta(\mu,\epsilon) \log \frac{1}{\theta(\mu,\epsilon)},
\end{aligned} \tag{57}$$

where here we make explicit the dependency of K and θ on μ and ϵ. From the construction of μ^{ϵ} (Eqs.(53), (54), (55) and the condition in (56)), it is important to note that $\theta(\mu,\epsilon) < \mu(K) \leq \mu(K-1) \ldots \leq \mu(1)$. At this point, we will use the following result:

Lemma 3 *$\forall \epsilon > 0$ and for any $\mu \in \mathcal{P}^{\epsilon}([M])$, it follows that*

$$\sum_{j=2}^{K(\mu,\epsilon)} \mu(j) \log \frac{1}{\mu(j)} \geq (\theta(\mu,\epsilon) + \epsilon) \log \frac{1}{\theta(\mu,\epsilon) + \epsilon} + (K(\mu,\epsilon) - 2)\theta(\mu,\epsilon) \log \frac{1}{\theta(\mu,\epsilon)}. \tag{58}$$

The proof is presented in Section G of this Suplemental.

Remark 4 *The proof of Lemma 3 comes from the use of some information-theoretic inequalities, similar to the arguments used to prove that the Shannon entropy over a finite alphabet is minimized with a degenerated distribution [Cover and Thomas, 2006, Gray, 1990a].*

Applying Lemma 3 in (57), we have that for all $\mu \in \mathcal{P}^{\epsilon}([M])$:

$$\begin{aligned}
\mathcal{H}(\mu) - \mathcal{H}(\mu^{\epsilon}) \geq {} & \mu(1) \log \frac{1}{\mu(1)} - (\mu(1) + \epsilon) \log \frac{1}{\mu(1) + \epsilon} \\
& + \left[(\theta(\mu,\epsilon) + \epsilon) \log \frac{1}{\theta(\mu,\epsilon) + \epsilon} - \theta(\mu,\epsilon) \log \frac{1}{\theta(\mu,\epsilon)} \right].
\end{aligned} \tag{59}$$

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

[17] 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 (56).

Using the fact that $\theta(\mu,\epsilon) < \mu(K) \leq \mu(K-1) \ldots \leq \mu(1)$, and that $\sum_{j=2}^{K(\mu,\epsilon)}(\mu(j)-\theta(\mu,\epsilon)) = \epsilon$ (see Eq.(56)), it is simple to verify that[18]

$$\mu(2) - \theta(\mu,\epsilon) \geq \frac{\epsilon}{K-1}, \tag{60}$$

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

On the other hand, if we consider the following function used in (59):

$$f_1(\theta,\epsilon) \equiv (\theta+\epsilon)\log\frac{1}{\theta+\epsilon} - \theta\log\frac{1}{\theta}, \tag{61}$$

$\frac{\partial f_1(\theta,\epsilon)}{\partial\theta}(\theta,\epsilon) = \log\frac{\theta}{\theta+\epsilon} < 0$, then $f_1(\theta,\epsilon)$ is strictly decreasing in the domain $\theta > 0$, for any $\epsilon > 0$. Therefore from (60), we have that $f_1(\theta(\mu,\epsilon),\epsilon) \geq f_1(\mu(2)-\epsilon/(K-1),\epsilon)$. Applying this last inequality in (59), we have that

$$\begin{aligned}\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) &\geq -f_1(\mu(1),\epsilon) + f_1(\theta(\mu,\epsilon),\epsilon)\\ &\geq -f_1(\mu(1),\epsilon) + f_1(\mu(2)-\epsilon/(K-1),\epsilon).\end{aligned} \tag{62}$$

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

$$\mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon) \geq -f_1(\mu(1),\epsilon) + f_1(\mu(2)-\epsilon/(M-1),\epsilon). \tag{63}$$

At this point, we return to our main problem:

$$\begin{aligned}\mathcal{I}_{loss}(\epsilon,M) &= \min_{\mu\in\mathcal{P}^\epsilon([M])} \mathcal{H}(\mu) - \mathcal{H}(\mu^\epsilon)\\ &\geq \min_{\mu(1)\in[1/M,1-\epsilon]}\left(-f_1(\mu(1),\epsilon) + \min_{\mu(2)\in[0,\min\{\mu(1),1-\mu(1)\}]}\left(f_1(\mu(2)-\epsilon/(M-1),\epsilon)\right)\right),\end{aligned} \tag{64}$$

where the lower bound in (64) comes from (63) 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 (64), where we recognize for the second optimization (from left to right) in (64) two scenarios.

Case 1 (the restriction $\mu(2) \leq \mu(1)$ is active in (64)): If we restrict the second optimization problem in (64) 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

$$\mathcal{I}_{loss}(\epsilon,M) \geq \min_{\mu(1)\in[1/M,1/2]} -f_1(\mu(1),\epsilon) + f_1(\mu(1)-\epsilon/(M-1),\epsilon), \tag{65}$$

the last bound from (64) using the fact 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 $x > 0$[19]. This implies that

$$\mathcal{I}_{loss}(\epsilon,M) \geq \tilde{f}(1/2,\epsilon) = f_1\left(1/2 - \frac{\epsilon}{M-1},\epsilon\right) - f_1\left(1/2,\epsilon\right) > 0, \tag{66}$$

using again 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 (64)): If we restrict the second optimization problem in (64) to the case where $1-\mu(1) < \mu(1)$, this scenario implies that $\mu(1) > \frac{1}{2}$. In addition, as $\mu\in\mathcal{P}^\epsilon([M])$, it follows that $\mu(1) \leq 1-\epsilon$. Therefore, under this hypothesis,

$$\mathcal{I}_{loss}(\epsilon,M) \geq \min_{\mu(1)\in(1/2,1-\epsilon]} -f_1(\mu(1),\epsilon) + f_1((1-\mu(1))-\epsilon/(M-1),\epsilon), \tag{67}$$

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

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

the last bound from (64) using the fact 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

$$\mathcal{I}_{loss}(\epsilon, M) \geq \tilde{\phi}(1/2, \epsilon) = f_1\left(1/2 - \frac{\epsilon}{M-1}, \epsilon\right) - f_1\left(1/2, \epsilon\right) > 0. \tag{68}$$

Interestingly in (68) and (66), we arrived to the same positive closed-form lower bound for $\mathcal{I}_{loss}(\epsilon, M)$, which concludes the proof of Theorem 4. □

G PROOF OF LEMMA 3

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 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.,

$$\begin{aligned}
\tilde{\mu}(2) &= \frac{\mu(2)}{\theta(K-1)+\epsilon} \geq \tilde{\theta} \equiv \frac{\theta}{\theta(K-1)+\epsilon} > 0, \\
&\cdots \\
\tilde{\mu}(K) &= \frac{\mu(K)}{\theta(K-1)+\epsilon} \geq \tilde{\theta}.
\end{aligned} \tag{69}$$

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

$$\begin{aligned}
\bar{e}_j(2) &= \tilde{\theta}, \ldots, \\
\bar{e}_j(j-1) &= \tilde{\theta}, \\
\bar{e}_j(j) &= \tilde{\theta} + \frac{\epsilon}{\theta(K-1)+\epsilon}, \\
\bar{e}_j(j+1) &= \tilde{\theta}, \ldots, \\
\bar{e}_j(K) &= \tilde{\theta}.
\end{aligned} \tag{70}$$

Importantly, it is simple to verify that $\tilde{\mu}$ (in (69)) 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

$$\tilde{\mu} = \sum_{j=2}^{K} w_j \cdot \bar{e}_j, \tag{71}$$

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

$$P_{Z|O}(\cdot|k) = \bar{e}_k \in \mathcal{P}([M]), \text{ and } P_O(k) = w_k, \tag{72}$$

$\forall k \in \{2, .., K\}$. 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 is equivalent to

$$(\tilde{\theta} + \tilde{\epsilon}) \log \frac{1}{\tilde{\theta} + \tilde{\epsilon}} + (K-2)\tilde{\theta} \log \frac{1}{\tilde{\theta}} \leq \mathcal{H}(\tilde{\mu}). \tag{73}$$

Returning to our original problem, we have that

$$\begin{aligned}
&\sum_{j=2}^{K} \mu(j) \log \frac{1}{\mu(j)} = \mu(\beta)\mathcal{H}(\tilde{\mu}) + \mu(\beta) \log \frac{1}{\mu(\beta)} \geq \\
&(\theta(K-1)+\epsilon) \cdot \left[(\tilde{\theta}+\tilde{\epsilon}) \log \frac{1}{\tilde{\theta}+\tilde{\epsilon}} + (K-2)\tilde{\theta} \log \frac{1}{\tilde{\theta}} \right] + (\theta(K-1)+\epsilon) \log \frac{1}{(\theta(K-1)+\epsilon)} \\
&= (\theta+\epsilon) \log \frac{(K-1)\theta+\epsilon}{\theta+\epsilon} + (K-2)\theta \log \frac{(K-1)\theta+\epsilon}{\theta} + (\theta(K-1)+\epsilon) \log \frac{1}{(\theta(K-1)+\epsilon)} \\
&= (\theta+\epsilon) \log \frac{1}{\theta+\epsilon} + (K-2)\theta \log \frac{1}{\theta}, \qquad (74)
\end{aligned}$$

where for the first inequality we use the lower bound in (73) 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, (74) proves the result. □