Enriching Disentanglement: Definitions to Metrics

Yivan Zhang [1 2] **Masashi Sugiyama** [2 1]

Abstract

A multitude of *metrics* for learning and evaluating *disentangled representations* has been proposed. However, it remains unclear what these metrics truly quantify and how to compare them. To solve this problem, we introduce a systematic approach for transforming an equational definition into a quantitative metric via *enriched category theory*. We show how to replace (i) equality with metric, (ii) logical connectives with order operations, (iii) universal quantifier with aggregation, and (iv) existential quantifier with the best approximation. Using this approach, we can derive useful metrics for measuring the modularity and informativeness of a disentangled representation extractor.

1. Introduction

In *supervised learning*, we often use a real-valued function $\ell : Y \times Y \to \mathbb{R}$ to measure how close a predicted output $f(x)$ of a function $f : X \to Y$ is to a target label y, i.e., $\ell(f(x), y)$, to quantify the cost of inaccurate prediction. Then, we can use the *total cost* over a collection of input-output pairs to measure the performance of a function. From a functional perspective, this operation induces a "*metric*" $L : [X, Y] \times [X, Y] \to \mathbb{R}$ between *functions*:

$$L(f, g) := \sum_x \ell(f(x), g(x)), \tag{1}$$

where g is a "*ground-truth function*" that maps each input x to its target label y. Measurements of this form can be used as both *learning objectives* and *evaluation metrics*. What does Eq. (1) measure? It measures how much two functions f and g are *equal*:

$$(f = g) := \forall x.\ (f(x) = g(x)). \tag{2}$$

However, as the learning problem becomes more complex, *measuring how good a function is becomes a non-trivial task*. The quality of a function cannot always be measured by how close it is to a fixed ground truth.

For example, in *representation learning* (Bengio et al., 2013), we may want a function to preserve *informative* factors in the data (Eastwood & Williams, 2018). Only when a representation extractor is informative, its output can be used in downstream tasks without any loss of distinguishability. This criterion could be formulated as follows: if two inputs x_1 and x_2 have different factors, $x_1 \neq x_2$, then their representations extracted by a function $f : X \to Z$ should be different too, $f(x_1) \neq f(x_2)$. This means that the representation extractor f should be *injective*, with regard to certain *equality* (Mazur, 2008) defined on the domain X and codomain Z.

There are a few equivalent ways to check if a function is injective. For example, an injective function $f : X \to Z$ is *left-cancellable*:

$$\forall g_1, g_2 : W \to X.\ (f \circ g_1 = f \circ g_2) \to (g_1 = g_2). \tag{3}$$

Moreover, an injective function has a *left-inverse*:

$$\exists g : Z \to X.\ g \circ f = \mathrm{id}_X\,. \tag{4}$$

To characterize an injective function f, we need to use (a) other functions g with different domain and codomain, (b) equality $=$ of functions, (c) function composition $\circ$, (d) identity functions id, (e) implication $\to$, (f) universal quantification $\forall$, and (g) existential quantification $\exists$.

Note that unlike the real-valued measurement given by Eq. (1), Eqs. (2) to (4) are based on predicate logic and can provide only binary "yes or no" answers. Since the total cost can be seen as a real-valued measurement for function equality, it raises the question of whether we can derive similar real-valued measurements for properties like *left-cancellability*, *left-invertibility*, and *injectivity*. Measuring the degree of injectivity of a function enables us to assess the informativeness of a representation extractor in a more fine-grained way.

In order to derive real-valued measurements for injectivity and other properties of a function, (A) we need a systematic way to *enrich* an equational definition of a desired property to a quantitative metric for this property such that the metric is compatible with the definition; (B) we need a consistent approach to incorporate composition, implication, quantification, and other operators; and (C) we also need to consider whether equivalent definitions lead to consistent metrics and how to compare and transform different metrics.

[1]The University of Tokyo [2]RIKEN AIP. Correspondence to: Yivan Zhang <yivanzhang@ms.k.u-tokyo.ac.jp>.

Presented at the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. *Copyright 2023 by the author(s).*

Disentangled representation learning (Bengio et al., 2013) is such a field where defining the desired properties of a function and measuring its performance quantitatively are not straightforward tasks (Zhang & Sugiyama, 2023). Existing approaches are based on various algebraic or statistical concepts, such as invariance (Higgins et al., 2017), independence (Chen et al., 2018), modularity (Ridgeway & Mozer, 2018), and equivariance (Cohen & Welling, 2014; Higgins et al., 2018). However, comparing different approaches theoretically has been a challenging task.

To address this, a recent work Zhang & Sugiyama (2023) proposed to use *category theory* to unify existing definitions and shed light on the essence of disentanglement. It was shown that the *cartesian/monoidal product* underlies many different approaches to disentanglement. In this paper, we further explore this direction and propose to use *enriched category theory* (Lawvere, 1973; Kelly, 1982) to study the relationship between the definitions and metrics of disentanglement.

Concretely, in this paper, we introduce a systematic method for quantifying the degree to which a function satisfies a *first-order equational predicate* based on quantale-enriched and metric-enriched categories (Section 2). Then, we apply this method to derive metrics for measuring two desired properties — modularity (Section 3) and informativeness (Section 4) — of a disentangled representation extractor. A demonstration of the proposed metrics is given in the end.

2. Enrichment: from equality to measurement

In this section, we briefly explain the proposed technique based on the following definition:

Definition 1 (Compatible functional metrics)**.** Let $\mathbf{Q}$ be a *quantale*, and $\mathbf{V}$ be the category of $\mathbf{Q}$-valued *premetrics*. Then, a collection of *compatible functional metrics* $\mathbf{C}$ is a $\mathbf{V}$-*enriched category*. If $\mathbf{C}$ has a monoidal structure, $\mathbf{C}$ should be a $\mathbf{V}$-*enriched monoidal category*.

2.1. Quantale

A quantale is a *preorder* with specified order operations. Let us explain this concept with two examples: binary truth values $(\{\bot, \top\}, \vdash)$ and non-negative extended real-values $([0, \infty], \geq)$. The order operations that a quantale support are listed in Table 1.

Let us explain why these operations are important. Consider the *conjunction* $a \wedge b$ of two logical values a and b, which has the property that $c \vdash a \wedge b$ if and only if $c \vdash a$ and $c \vdash b$. We can see its similarity with the *maximum* because $c \geq \max\{a, b\}$ if and only if $c \geq a$ and $c \geq b$. Also, as *true* $\top$ is the unit of the conjunction, $a \wedge \top \equiv a \equiv \top \wedge a$, *zero* is the unit of the maximum, $\max\{a, 0\} = a = \max\{0, a\}$.

Table 1: Quantale

$(Q, \preceq)$	$(\{\bot, \top\}, \vdash)$	$([0, \infty], \geq)$
top $\top$	true $\top$	zero 0
bottom $\bot$	false $\bot$	infinity ∞
meet $\wedge$	conjunction $\wedge$	maximum $\max$
join $\vee$	disjunction $\vee$	minimum $\min$
monoidal product $\oplus$	conjunction $\wedge$	addition $+$
internal hom $\multimap$	implication $\rightarrow$	subtraction $-$

Moreover, in logic, the *implication* introduction (deduction theorem) and elimination (modus ponens) rules ensure that $q \wedge s \vdash t$ and $q \vdash s \rightarrow t$ are equivalent (Abramsky & Tzevelekos, 2011). Then, we can find an operation similar to the implication for the *addition* of real-values as well: the (truncated) *subtraction* satisfies that $q + s \geq t$ if and only if $q \geq \max\{t - s, 0\}$.

These facts indicate that the concept of quantale can accommodate both logical and real-valued quantification, making it possible to analyze the relationship between equational definitions and quantitative metrics.

2.2. Premetric

Next, we can consider a real-valued *premetric* between functions from domain A to codomain B, which is a binary function $d_{[A,B]} : [A, B] \times [A, B] \rightarrow [0, \infty]$ satisfying $d_{[A,B]}(f, f) = 0$. We only consider premetrics because some distances commonly used in machine learning such as the *relative entropy* (Perrone, 2022) do not satisfy the symmetry nor the triangle inequality, but we still intend to incorporate them into the theoretical framework.

2.3. Enrichment

Since we need to consider functions with different domains and codomains to characterize properties like injectivity, we have some requirements on the premetrics based on the properties of the enriched monoidal category. For example:

The *composition* $\circ$, which combines functions in series, satisfies the following inequality:

$$d_{[B,C]}(g, g') + d_{[A,B]}(f, f') \geq d_{[A,C]}(g \circ f, g' \circ f'). \quad (5)$$

The *monoidal product* $\otimes$, which combines functions in parallel, satisfies the following inequality:

$$d_{[A,B]}(f, f') + d_{[C,D]}(h, h') \geq d_{[A \otimes C, B \otimes D]}(f \otimes h, f' \otimes h'). \quad (6)$$

3. Measuring modularity

We apply the theory above to derive metrics for measuring *modularity* (Ridgeway & Mozer, 2018), a desired property of a disentangled representation extractor.

First, we need to define modularity equationally. Zhang & Sugiyama (2023) revealed that this property can be defined via the *product morphisms* in a category. Concretely, given a data generator $g : Y \to X$ that maps a tuple $(y_1, \dots, y_N) =: y \in Y := Y_1 \times \cdots \times Y_N$ of *factors* to an *observation* $x \in X$ (images, texts, etc.), a good representation extractor $f : X \to Z$ should encode each factor separately using *codes* $Z := Z_1 \times \cdots \times Z_N$, such that *the composition $m = f \circ g$ is a product function*.

Then, we can derive a real-valued metric from this definition by replacing the equality $=$ with a (pre)metric d:

Metric 1. A function $m : Y \to Z$ is a *product function* $m = \prod_i m_{i,i}$ if

$$\forall i \in [1..N].\ \exists m_{i,i} : Y_i \to Z_i. \qquad m_i : Y \to Z_i := p_i \circ m = m_{i,i} \circ p_i, \tag{7}$$

where p_i is the *projection* ($p_i : Y \to Y_i$ or $p_i : Z \to Z_i$).

The degree to which a function m is a product function can be measured by the distance between m and its product function approximation:

$$\max_{i \in [1..N]} \min_{m_{i,i} : Y_i \to Z_i} \underbrace{\max_{y \in Y} d_{Z_i}(m_i(y), m_{i,i}(y_i))}_{d_{[Y,Z_i]}(m_i, m_{i,i} \circ p_i)}. \tag{8}$$

The corresponding algorithm can be described as follows:

Algorithm 1: Measuring product via approximation

for $i \in [1..N]$ **do**
- Collect a set of instances with the i-th factor y_i fixed and other factors $y_{\setminus i}$ varying;
- Calculate their i-th codes $m_i(y_i, y_{\setminus i})$;
- Construct an approximation $m_{i,i} : y_i \mapsto \arg\min_{z_i \in Z_i} \max_{y_{\setminus i} \in Y_{\setminus i}} d_{Z_i}(m_i(y_i, y_{\setminus i}), z_i)$;

Calculate the distance between the function m and its product function approximation $\prod_i m_{i,i}$.

If the codomain metric d_{Z_i} is the Euclidean distance, then the best approximation $m_{i,i}$ maps a factor y_i to the *center* of the *smallest bounding sphere* (Welzl, 1991) of the outputs $\{m_i(y_i, y_{\setminus i}) \mid y_{\setminus i} \in Y_{\setminus i}\}$, and the induced metric is the maximal *radius*.

If we do not strictly follow the theory, we may change the maximization to summation/mean and choose different distances. Then, we can obtain different approximations. For example, the *(geometric) median* minimizes the mean distance, and the induced metric is the *mean absolute deviation around the median* (Weiszfeld, 1937). The *mean* minimizes the mean squared distance, and the induced metric is the *variance*.

Alternatively, we can determine if a function is a product by examining whether its *exponential transpose* is *constant* (Zhang & Sugiyama, 2023):

Metric 2. A function $m : Y \to Z$ is a product function if the exponential transpose $\widehat{m_i} : Y_{\setminus i} \to [Y_i, Z_i]$ of its i-th component $m_i : Y \to Z_i$ is constant for all i in $[1..N]$.

The degree to which a function m is a product function can be measured by the maximal pairwise distance between the i-th outputs when the i-th input is fixed:

$$\max_{i \in [1..N]} \max_{y_{\setminus i}, y'_{\setminus i} \in Y_{\setminus i}} \underbrace{\max_{y_i \in Y_i} d_{Z_i}(m_i(y_i, y_{\setminus i}), m_i(y_i, y'_{\setminus i}))}_{d_{[Y_i, Z_i]}(\widehat{m_i}(y_{\setminus i}), \widehat{m_i}(y'_{\setminus i}))}. \tag{9}$$

The corresponding algorithm can be described as follows:

Algorithm 2: Measuring product via constancy

for $i \in [1..N]$ **do**
- Collect a set of pairs of instances with the i-th factor y_i fixed and other factors $y_{\setminus i}$ varying;
- Calculate their i-th codes $m_i(y_i, y_{\setminus i})$ and $m_i(y_i, y'_{\setminus i})$;
- Calculate all pairwise distances $d_{Z_i}(m_i(y_i, y_{\setminus i}), m_i(y_i, y'_{\setminus i}))$ as a constancy measurement;

Aggregate all constancy measurements.

4. Measuring informativeness

We point out that another desired property of a disentangled representation extractor — *informativeness* (Eastwood & Williams, 2018) — is related to the *(split) monomorphisms* in a category (Zhang & Sugiyama, 2023). Using the same technique, we can derive the following metrics based on two equivalent definitions:

Metric 3. A function $m : Y \to Z$ is *left-invertible* if

$$\exists h : Z \to Y.\ h \circ m = \mathrm{id}_Y. \tag{10}$$

The degree of left-invertibility of a function m can be measured by approximating its left-inverse:

$$\min_{h : Z \to Y} \underbrace{\max_{y \in Y} d_Y(h(m(y)), y)}_{d_{[Y,Y]}(h \circ m, \mathrm{id}_Y)}. \tag{11}$$

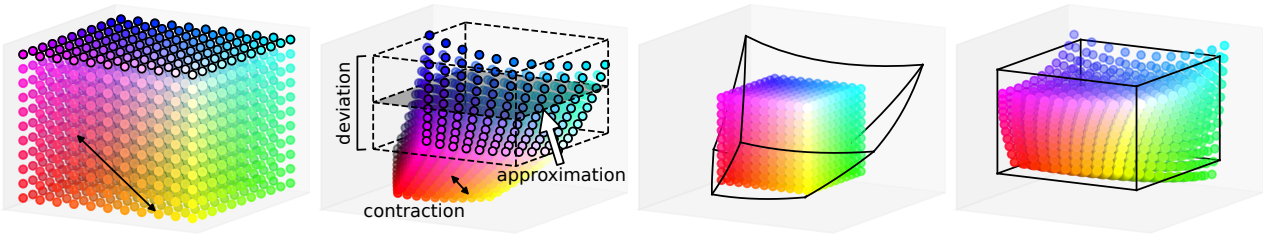
Metric 4. A function $m : Y \to Z$ is *injective* if

$$\forall y, y' \in Y.\ (m(y) = m(y')) \to (y = y'). \tag{12}$$

The degree of injectivity of a function m can be measured by the degree to which m contracts pairs of inputs:

$$\max_{y, y' \in Y} \max\{d_Y(y, y') - d_Z(m(y), m(y')), 0\}. \tag{13}$$

A demonstration of the proposed metrics is shown in Fig. 1.



(a) true factors Y (b) entangled codes Z (c) product approximation of an encoder $m : Y \to Z$ (d) linear approximation of its left-inverse $h : Z \to Y$

Figure 1: (a) a set of *factors* Y represented by the RGB color model; (b) a set of entangled *codes* Z extracted by an encoder $m : Y \to Z$; (c) a *product* function approximation ($m \approx \prod_i m_{i,i}$, Metric 1); and (d) a linear approximation of its *left-inverse* ($h \circ m \approx \mathrm{id}_Y$, Metric 3). Without approximation, we can still measure the *modularity* of an encoder by using a set of factors $\{(y_1, y_{2,3}) \mid y_{2,3} \in Y_2 \times Y_3\}$ with one factor y_1 fixed and other factors $y_{2,3}$ varying (black-edged dots in Fig. 1a) and their codes $\{m(y_1, y_{2,3}) \mid y_{2,3} \in Y_2 \times Y_3\}$ (dots and their bounding box in Fig. 1b) and measuring the deviation $d_{Z_i}(m_i(y_1, y_{2,3}), m_i(y_1, y'_{2,3}))$ (the height of the bounding box, Metric 2). We can also measure the *informativeness* of an encoder by measuring how much it contracts pairs of inputs (double-headed arrows in Figs. 1a and 1b, Metric 4).

References

Abramsky, S. and Tzevelekos, N. Introduction to categories and categorical logic. In *New Structures for Physics*, pp. 3–94. Springer, 2011. 2.1

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 1, 1

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Neural Information Processing Systems*, 2018. 1

Cohen, T. and Welling, M. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, 2014. 1

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. 1, 4

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 1

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 1

Kelly, M. Basic concepts of enriched category theory. In *London Mathematical Society Lecture Note Series*, volume 64. Cambridge University Press, 1982. 1

Lawvere, F. W. Metric spaces, generalized logic, and closed categories. *Rendiconti del seminario matématico e fisico di Milano*, 43:135–166, 1973. 1

Mazur, B. When is one thing equal to some other thing? In *Proof and Other Dilemmas: Mathematics and Philosophy*, pp. 221–242. Mathematical Association of America, 2008. 1

Perrone, P. Markov categories and entropy. *arXiv preprint arXiv:2212.11719*, 2022. 2.2

Ridgeway, K. and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss. In *Neural Information Processing Systems*, 2018. 1, 3

Weiszfeld, E. Sur le point pour lequel la somme des distances de n points donnés est minimum (on the point for which the sum of the distances to n given points is minimum). *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937. 3

Welzl, E. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science*, pp. 359–370. Springer, 1991. 3

Zhang, Y. and Sugiyama, M. A category-theoretical meta-analysis of definitions of disentanglement. In *International Conference on Machine Learning*, 2023. 1, 3, 3, 4