Enriching Disentanglement: Definitions to Metrics

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

[1]The University of Tokyo [2]RIKEN AIP

arXiv:2305.11512
https://yivan.xyz
yivanzhang@ms.k.u-tokyo.ac.jp

Can we measure injectivity?

- In **supervised learning**, we can use the *total cost* over a collection of input-output pairs to measure the performance of a function, which can be considered as 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. It measures *how much* two functions f and g are *equal*:

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

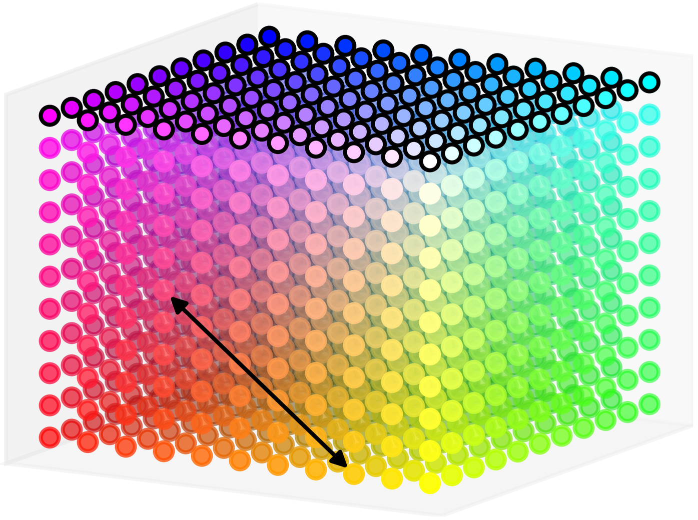
- In **representation learning** [Bengio et al., 2013], we may want a function to preserve *informative* factors in data: if two inputs x_1 and x_2 have different factors, $x_1 \neq x_2$, then their representations extracted by a function $f : X \to Z$ should be different too, $f(x_1) \neq f(x_2)$, which means that the representation extractor $f : X \to Z$ should be *injective*.
- An injective function $f : X \to Z$ is *left-cancellable*:

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

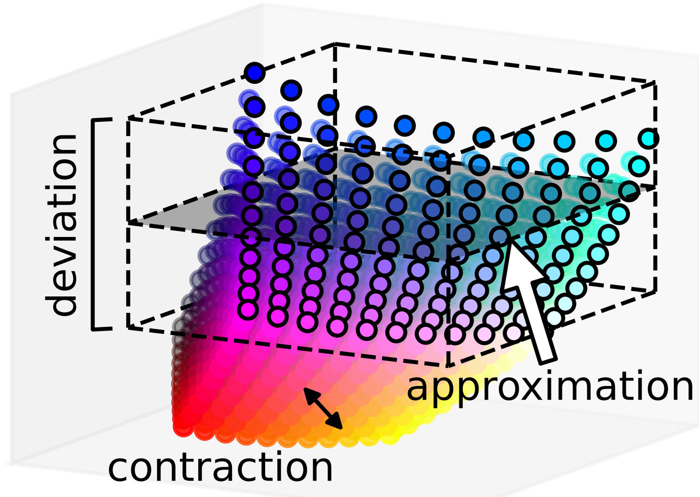
An injective function $f : X \to Z$ has a *left-inverse*:

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

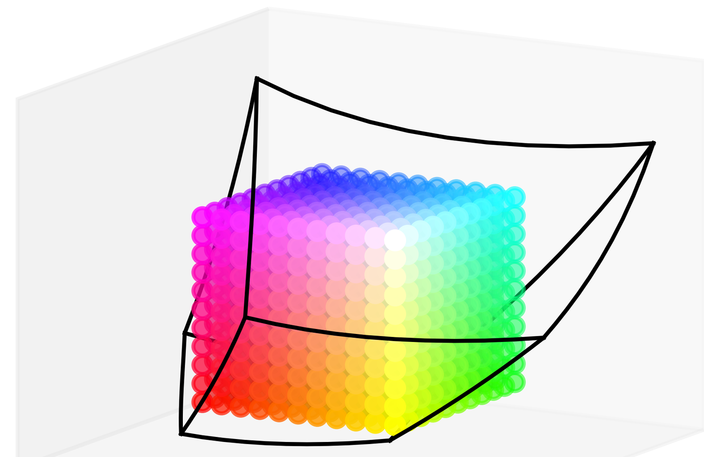
- Just as we can measure *function equality* in terms of the total cost, can we measure *injectivity*, *left-cancellability*, and *left-invertibility*?



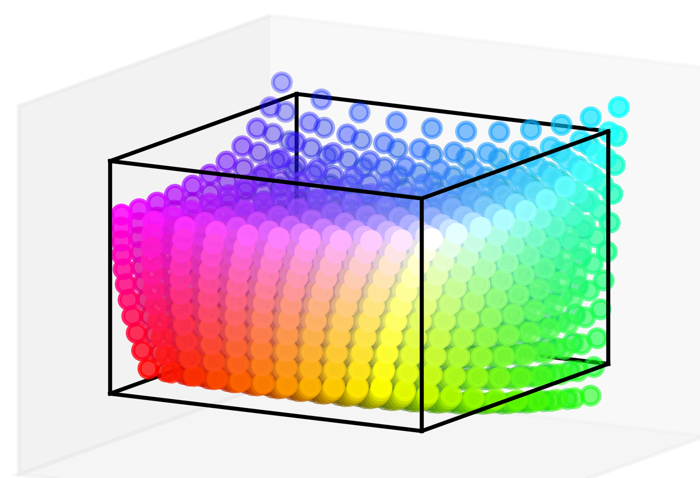
(a) true factors Y



(b) entangled codes Z



(c) product approximation of an encoder $m : Y \to Z$



(d) linear approximation of its left-inverse $h : Z \to Y$

Premetric-enriched monoidal category

A *monoidal category enriched in a category of premetric spaces*, in which there is a *premetric* $d_{[A,B]}$ from object A to object B, describes a collection of *premetrics between morphisms* that are compatible with composition and monoidal product.

- The *composition* $\circ$ combines morphisms in *series*.

$$d_{[B,C]}(g, g') \oplus d_{[A,B]}(f, f') \preceq d_{[A,C]}(g \circ f, g' \circ f').$$

- The *monoidal product* $\otimes$ combines morphisms in *parallel*.

$$d_{[A,B]}(f, f') \oplus d_{[C,D]}(h, h') \preceq d_{[A\otimes C, B\otimes D]}(f \otimes h, f' \otimes h').$$

$$\begin{tikzcd}
& B \arrow[dr, "g'"'] \arrow[dr, bend left=50, "g"] & \\
A \arrow[ur, "f'"] \arrow[ur, bend left=50, "f"] \arrow[ur, phantom, bend left=25, "d(f,f')"] \arrow[rr, "g'\circ f'"] \arrow[rr, bend right=50, "g\circ f"'] \arrow[rr, phantom, bend right=25, "d(g\circ f, g'\circ f')"] & & C \arrow[ul, phantom, bend right=25, "d(g,g')"]
\end{tikzcd}
\qquad
\begin{tikzcd}
A \arrow[r, bend left, "f"] \arrow[r, bend right, "f'"'] \arrow[r, phantom, "d(f,f')"] & B \\
A \otimes C \arrow[r, bend left, "f\otimes h"] \arrow[r, bend right, "f'\otimes h'"'] \arrow[r, phantom, "d(f\otimes h, f'\otimes h')"] & B \otimes D \\
C \arrow[r, bend left, "h"] \arrow[r, bend right, "h'"'] \arrow[r, phantom, "d(h,h')"] & D
\end{tikzcd}$$

Quantale-valued premetric

If the premetrics take value in a *quantale* (monoidal closed cocartesian thin category), there exist *order operations* that behave like *logical connectives* (e.g., conjunction, implication).

$(Q, \preceq)$	$(\{\bot, \top\}, \vdash)$	$([0, \infty], \geq)$
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 $\to$	subtraction $-$

We can use a monoidal category enriched in a category of quantale-valued premetrics to derive disentanglement metrics from disentanglement definitions [Zhang and Sugiyama, 2023]!

Modularity: a code encodes only one factor

- $m : Y \to Z$ is a *product function* $m = \prod_i m_{i,i}$:

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

- Product approximation:

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

- The *exponential transpose* $\widehat{m_i} : Y_{\setminus i} \to [Y_i, Z_i]$ is constant:

$$\forall i \in [1..N].\ \forall y_{\setminus i}, y'_{\setminus i} \in Y_{\setminus i}.\ \widehat{m_i}(y_{\setminus i}) = \widehat{m_i}(y'_{\setminus i}).$$

- The maximal pairwise distance between the i-th outputs when the i-th input is fixed:

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

Informativeness: codes encode factors faithfully

- $m : Y \to Z$ is *left-invertible*:

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

- Left-inverse approximation:

$$\min_{h:Z\to Y} \max_{y\in Y} d_Y(h(m(y)), y).$$

- $m : Y \to Z$ is *injective*:

$$\forall y, y' \in Y.\ (m(y) = m(y')) \to (y = y').$$

- Contraction:

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

Future research directions

- Can we use other aggregate functions (e.g., mean, median)?
- Can we optimize these metrics with minimal supervision?

References

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

Yivan Zhang and Masashi Sugiyama. A category-theoretical meta-analysis of definitions of disentanglement. In *ICML*, 2023.