TITLE: Enriching Disentanglement: From Logical Definitions to Quantitative Metrics

TL;DR: # Turn Logical Formulas into Loss Functions

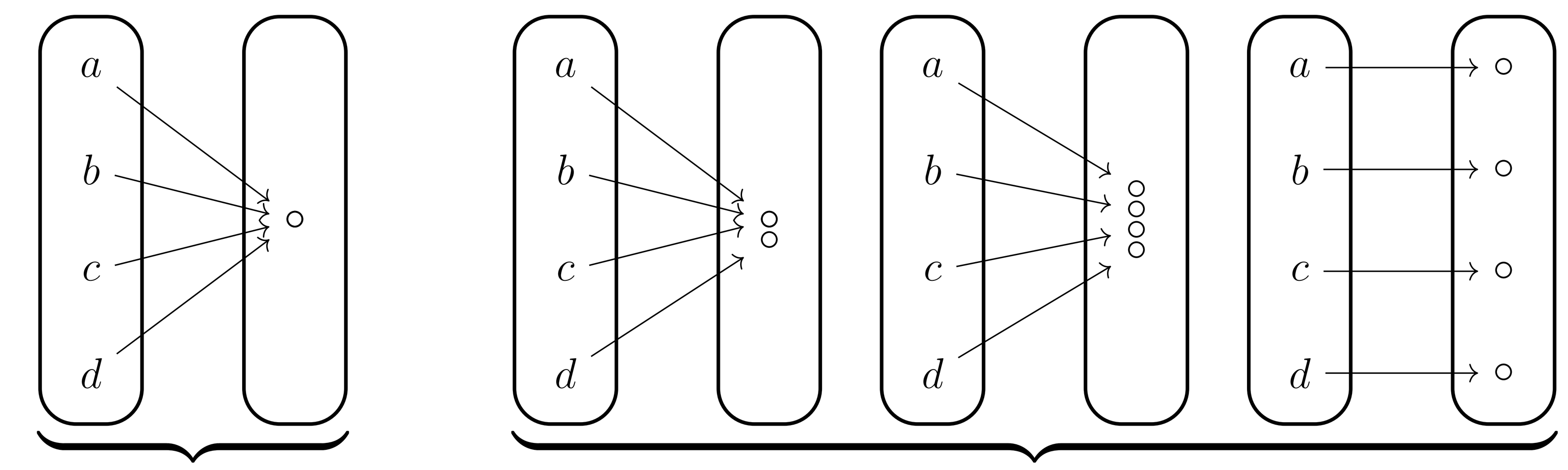
Yivan Zhang[1,2] Masashi Sugiyama[2,1]

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

arXiv:2305.11512
https://yivan.xyz
yivan.xyz@gmail.com

Motivation: measuring properties of functions

A *constant function* is a function that maps all inputs to the same output, but **"*how constant*" is a non-constant function?**

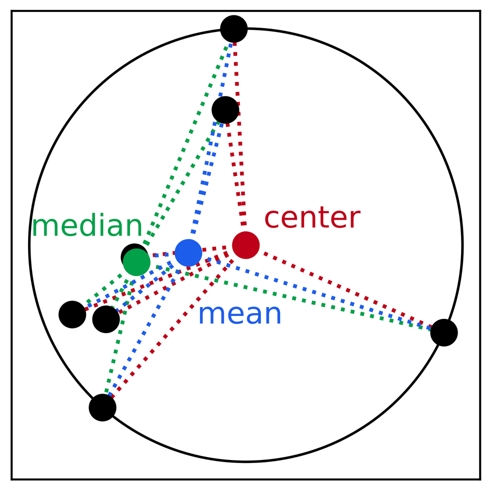


✓ a constant function ✗ non-constant functions

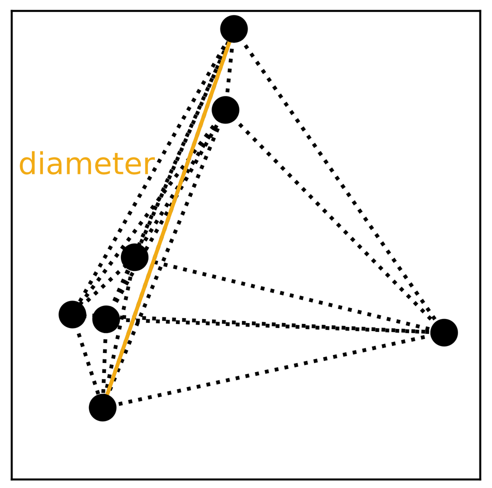
- Intuitively, we can measure how the outputs are distributed over the output space (constancy = 0 deviation).
- If we have a real-valued and preferably differentiable metric for the degree of constancy, we can *measure* the constancy of a function and *optimize* it using gradient descent.
- **How can we measure properties of functions?**

Idea: quantifying logical predicates

- **We can derive metrics from the definition of properties.**
- The constancy of a function $f : A \to B$ can be defined by (binary-valued) logical predicates:
  (a) $\exists b \in B.\ \forall a \in A.\ f(a) = b$ or (b) $\forall a, a' \in A.\ f(a) = f(a')$
- From these definitions, we can derive their corresponding (real-valued) quantitative metrics:
  (a) $\inf_{b \in B} \mathrm{agg}_{a \in A}\, d(f(a), b)$ or (b) $\mathrm{agg}_{a,a' \in A}\, d(f(a), f(a'))$



(a) how far the outputs are from a central point



(b) how far the outputs are from each other

We can use them as *learning objectives* or *evaluation criteria*. They differ in terms of *computation cost* and *differentiability*.

Background: logic and metric in machine learning

- Even in the *supervised learning* setting, we can view the total cost $L(f, g) := \sum_{x \in X} \ell(f(x), g(x))$ as a measure of the function equality $(f = g) := \forall x \in X.\ f(x) = g(x)$ between a learning model $f : X \to Y$ and the ground-truth $g : X \to Y$.
- Can we extend this parallel to *representation learning*?

Problem: disentangled representation learning

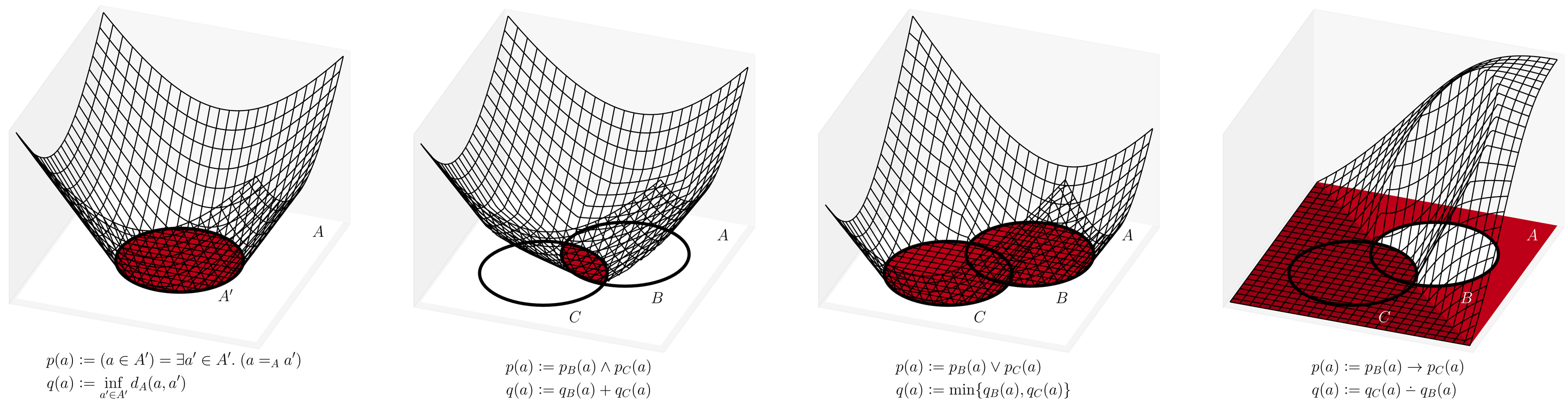
- No unified logical definition, many evaluation metrics [Carbonneau et al., 2022, Zhang and Sugiyama, 2023]
- Unclear what properties these metrics quantify
- Usually non-differentiable and computationally inefficient
- Unproven if a learning method can truly optimize a metric

Enrichment: from logic to metric

Like logic, we can construct metrics *compositionally*!

Logic		Metric	
truth values	$\{\top, \bot\}$	real values	$[0, \infty]$
predicate	$A \xrightarrow{p} \{\top, \bot\}$	quantity	$A \xrightarrow{q} [0, \infty]$
equality	$(a = a')$	strict premetric	$d(a, a')$
conjunction	$\wedge$	addition	$+$
disjunction	$\vee$	minimum	$\min$
implication	$\to$	subtraction*	$\dot{-}$
universal quantifier	$\forall$	aggregator**	∇
existential quantifier	$\exists$	infimum	$\inf$

* truncated subtraction: $b \dot{-} a := \max\{b - a, 0\}$
** e.g., maximum, sum, mean, and mean square



(a) quantity (b) conjunction (c) disjunction (d) implication

Theory: (sub)homomorphisms from metric to logic

- **Zero predicate**: $\zeta : [0, \infty] \to \{\top, \bot\} := x \mapsto (x = 0)$
- **Homomorphisms** from metric to logic via ζ:

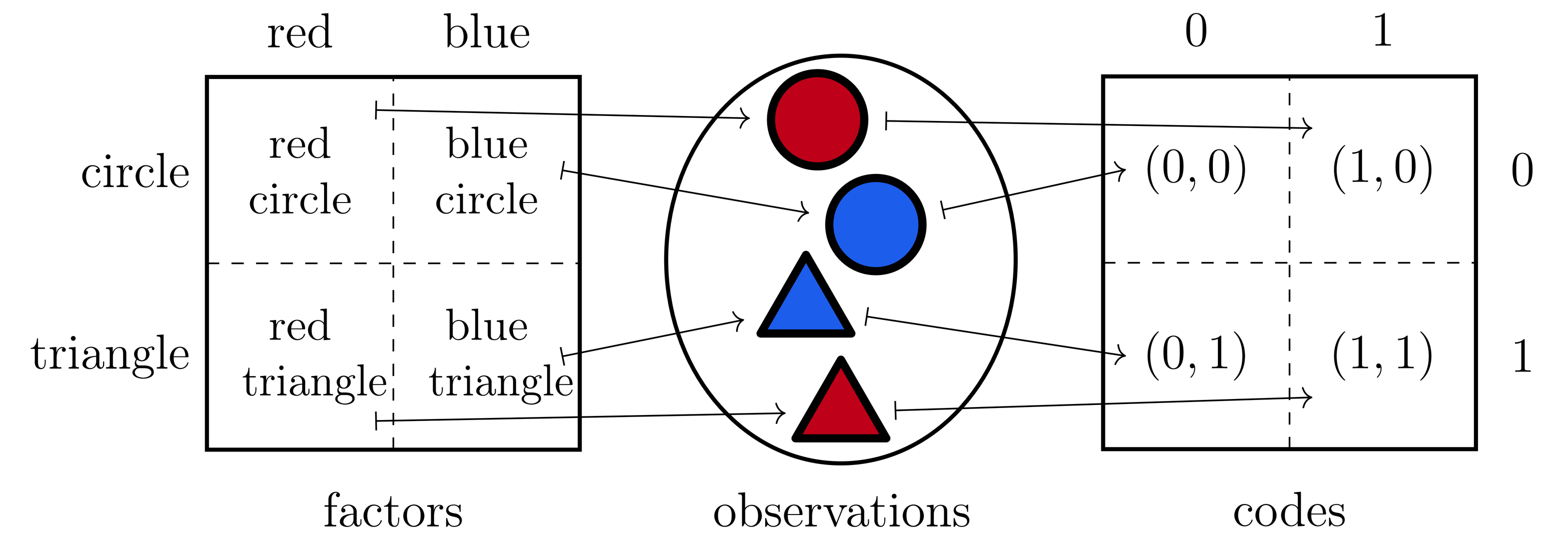
$$\begin{tikzcd} A \arrow[r, "\mathrm{id}_A"] \arrow[d, "q"'] & A \arrow[d, "p"] \\ {[0,\infty]} \arrow[r, "\zeta"'] & \{\top, \bot\} \end{tikzcd} \quad \begin{tikzcd} {[0,\infty]^n} \arrow[r, "\zeta^n"] \arrow[d, "\alpha"'] & \{\top, \bot\}^n \arrow[d, "\beta"] \\ {[0,\infty]} \arrow[r, "\zeta"'] & \{\top, \bot\} \end{tikzcd} \quad \begin{tikzcd} {[0,\infty]^A} \arrow[r, "\zeta^A"] \arrow[d, "\alpha_A"'] & \{\top, \bot\}^A \arrow[d, "\beta_A"] \\ {[0,\infty]} \arrow[r, "\zeta"'] & \{\top, \bot\} \end{tikzcd}$$

from quantity q to predicate p | from quantitative operation α to logical operation β | from aggregator α_A to quantifier β_A

- **Subhomomorphisms**: replace equality $=$ by implication $\to$
  Truncated subtraction is subhomomorphic to implication.
  No continuous operation is homomorphic to implication.
- **Main theorem**: *If the components are (sub)homomorphic, so is the compound*: $q(a) = 0$ equals (implies) $p(a) = \top$.
- **Benefits**: (1) **no failure modes**; (2) **no hyperparameters**; (3) **no stochastic components**; (4) some are **differentiable**.

Logical definitions of disentangled representations

Modularity: reconstruct the product structure



$$\begin{tikzcd} Y_{\text{color}} \times Y_{\text{shape}} \arrow[r, "g"] \arrow[rr, bend right] & X \arrow[r, "f"] & Z_{\text{color}} \times Z_{\text{shape}} \end{tikzcd}$$

Example: $m := f \circ g = m_{\text{color}} \times m_{\text{shape}}$

$$\begin{cases} (\textcolor{red}{\blacksquare}, \bigcirc) \mapsto (2, 4) \\ (\textcolor{red}{\blacksquare}, \triangle) \mapsto (2, 6) \\ (\textcolor{blue}{\blacksquare}, \bigcirc) \mapsto (6, 4) \\ (\textcolor{blue}{\blacksquare}, \triangle) \mapsto (6, 6) \end{cases} = \underbrace{\begin{cases} \textcolor{red}{\blacksquare} \mapsto 2 \\ \textcolor{blue}{\blacksquare} \mapsto 6 \end{cases}}_{m_{\text{color}}} \times \underbrace{\begin{cases} \bigcirc \mapsto 4 \\ \triangle \mapsto 6 \end{cases}}_{m_{\text{shape}}} \qquad \begin{cases} (\textcolor{red}{\blacksquare}, \bigcirc) \mapsto (1, 2) \\ (\textcolor{red}{\blacksquare}, \triangle) \mapsto (3, 4) \\ (\textcolor{blue}{\blacksquare}, \bigcirc) \mapsto (5, 6) \\ (\textcolor{blue}{\blacksquare}, \triangle) \mapsto (7, 8) \end{cases}$$

✓ a product function ✗ not a product function

Definition:

$$p_{\text{product}}(m) := \exists m_{1,1} : Y_1 \to Z_1.\ \exists m_{2,2} : Y_2 \to Z_2.\ m = m_{1,1} \times m_{2,2}$$

Quantitative metrics of disentangled representations

Metric:

$$q_{\text{product}}(m) := \inf_{m_{1,1} \in [Y_1, Z_1]} \inf_{m_{2,2} \in [Y_2, Z_2]} d_{[Y,Z]}(m, m_{1,1} \times m_{2,2})$$

Instantiations:

- $\mathrm{mean}_{y_1 \in Y_1} \mathrm{var}_{y_2 \in Y_2}\, m_1(y_1, y_2) + \mathrm{mean}_{y_2 \in Y_2} \mathrm{var}_{y_1 \in Y_1}\, m_2(y_1, y_2)$
- $\max_{y_1 \in Y_1} \mathrm{diam}_{y_2 \in Y_2}\, m_1(y_1, y_2) + \max_{y_2 \in Y_2} \mathrm{diam}_{y_1 \in Y_1}\, m_2(y_1, y_2)$

Results:

- We have derived fine-grained, efficient, and differentiable quantitative metrics for disentangled representations.
- **We can quantify any logically defined properties!**

	Modularity						Informativeness						Existing metrics					
		Product approx.		Constancy				Retraction approx.			Contraction		Pair		Info.	Regressor		
		Rad.	MAD	Var.	Diam.	MPD		ME	MAE	MSE	Max	Mean	Beta[a]	Factor[b]	MIG[c]	Dis.[d]	Com.[d]	Info.[d]
entanglement	✗	0.44	0.75	0.96	0.19	0.82	✓	0.76	0.96	0.99	0.44	0.78	0.89	0.83	0.18	0.28	0.28	1.00
rotation	✗	0.22	0.51	0.80	0.05	0.64	✓	1.00	1.00	1.00	1.00	1.00	0.96	0.34	0.17	0.40	0.40	1.00
duplicate	✗	0.24	0.43	0.67	0.06	0.56	✓	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59	1.00
complement	✗	0.12	0.28	0.55	0.01	0.42	✓	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.63	1.00
misalignment	✗	0.22	0.44	0.74	0.05	0.58	✓	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00
redundancy	✓	1.00	1.00	1.00	1.00	1.00	✓	1.00	1.00	1.00	1.00	1.00	1.00	0.33	1.00	1.00	0.93	1.00
contraction	✓	1.00	1.00	1.00	1.00	1.00	✓	1.00	1.00	1.00	0.18	0.49	1.00	1.00	1.00	1.00	1.00	1.00
nonlinear	✓	1.00	1.00	1.00	1.00	1.00	✓	0.79	0.93	0.99	0.65	0.95	1.00	1.00	0.88	1.00	1.00	1.00
constant	✓	1.00	1.00	1.00	1.00	1.00	✗	0.42	0.76	0.90	0.18	0.48	0.33	0.33	0.00	0.00	0.00	0.00
random	✗	0.22	0.48	0.78	0.05	0.61	✗	0.42	0.76	0.90	0.22	0.83	0.34	0.33	0.00	0.00	0.00	0.04

References

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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