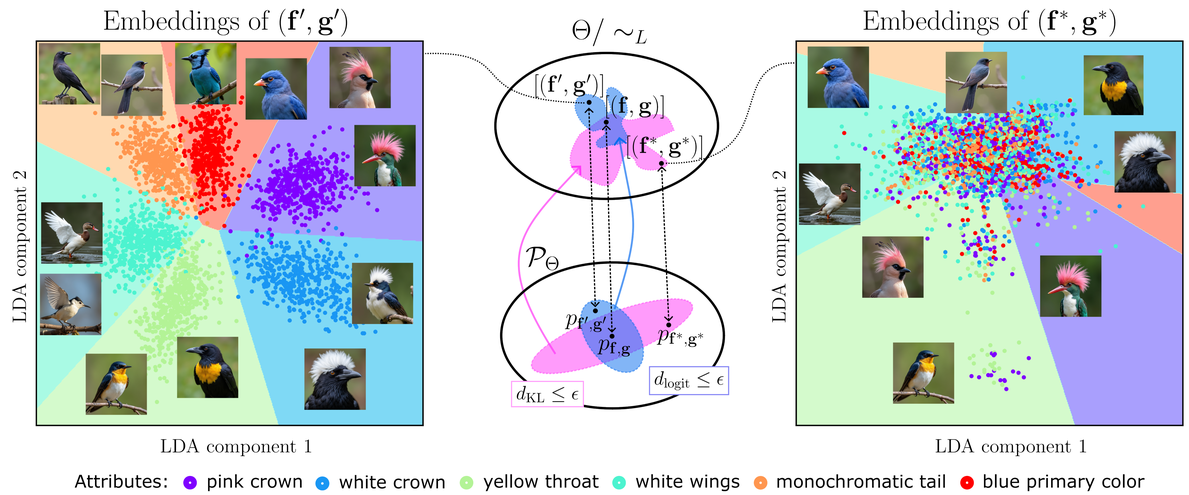
Identifiability:
Distributions are equal if and only if representations are linearly equivalent.

What if distributions are not equal?



Model Class:

$$p_{\mathbf{f},\mathbf{g}}(y|\mathbf{x}) \propto \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y))$$

Logits:

$$\mathbf{u}(\mathbf{x}) = (\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y_1), \ldots, \mathbf{f}(\mathbf{x})^\top \mathbf{g}(y_k))^\top$$

$$d^2_{\text{rep}}((\mathbf{f},\mathbf{g}),(\mathbf{f}',\mathbf{g}')) := C \sum_{\tilde{y}\in\mathcal{Y}} \sum_{\mathcal{J}\subseteq\mathcal{Y}\setminus\{\tilde{y}\}} \mathbb{E}_{\mathbf{x}\sim p_\mathbf{x}} \left\| \mathbf{f}(\mathbf{x}) - \tilde{\mathbf{A}}_\mathcal{J} \mathbf{f}'(\mathbf{x}) \right\|_2^2$$

Logit Distance Bounds Representational similarity

Representational similarity:
How close are models to being linearly equivalent?

- Linearly equivalent models will have maximal m_{CCA}, but maximal m_{CCA} does not guarantee equivalent models.
- d_{rep} is zero if and only if models are equivalent.

Closeness in distribution $\implies$ representational similarity?

- KL divergence does not give a practical bound;
- Logit distance does, where

$$d^2_{\text{logit}}(p_{\mathbf{f},\mathbf{g}}, p_{\mathbf{f}',\mathbf{g}'}) = \mathbb{E}_{\mathbf{x}\sim p_\mathbf{x}} \left\| \mathbf{u}(\mathbf{x}) - \mathbf{u}'(\mathbf{x}) \right\|_2^2$$

Implications (for distillation and interpretability):

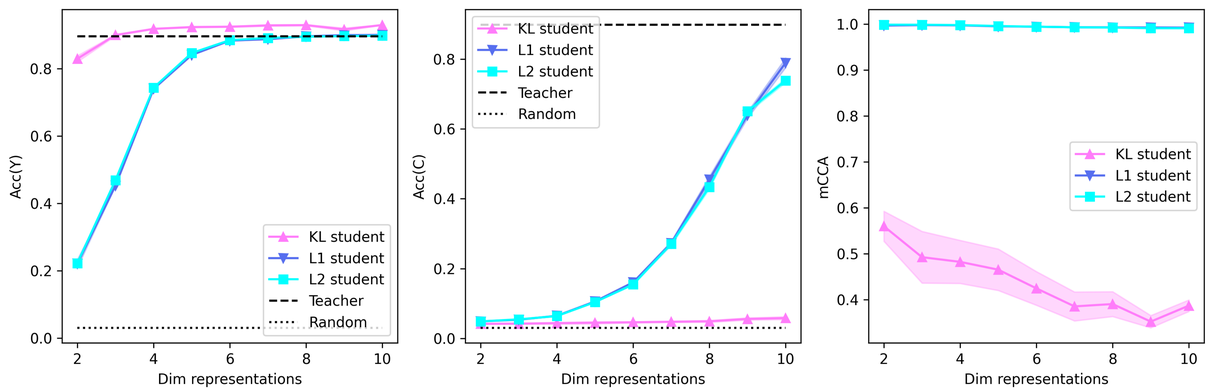
- KL-based distillation can match teacher predictions, but fails to preserve label rankings and linear concepts
- Logit distance distillation better preserves these concepts.

Results on SUB.

Teachers are in **gray**, KL-students in **pink**, L_1-students in **light blue**, and L_2-students in **blue**.

Acc(Y)(↑)	Acc(C)(↑)	d_{rep}(↓)	m_{CCA}(↑)
0.91 ± 0.01	0.92 ± 0.01	–	–
$\mathbf{0.93 \pm 0.01}$	0.06 ± 0.01	3100 ± 170	0.42 ± 0.01
0.92 ± 0.01	$\mathbf{0.75 \pm 0.01}$	10.0 ± 0.9	$\mathbf{0.99 \pm 0.01}$
0.92 ± 0.01	0.72 ± 0.01	$\mathbf{1.6 \pm 0.2}$	$\mathbf{0.99 \pm 0.01}$

SUB varying students representational dimension



Paper

Code

Beatrix M. G. Nielsen[1], Emanuele Marconato[1], Luigi Gresele[2], Andrea Dittadi[2], Simon Buchholz[2]
[1] Equal contribution, [2] Shared last author

TANGO HELMHOLTZ AI PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE IT UNIVERSITY OF CPH