What are the mathematics of abstraction?

Dovetail Research is offering a paid research fellowship for UK-based researchers interested in uncovering the mathematics behind agents. Here's one problem you could work on.

How can we characterise the minimal shared information between two random variables governed by a joint distribution $P(X, Y)$? One proposal by Wentworth and Lorell (2025) involves so-called *natural latent* variables labelled Λ and characterised by a conditional distribution $P(\Lambda|X, Y)$. To be a 'natural latent', these variables must satisfy three conditions. First, they must approximately *mediate* between X and Y, meaning that X and Y are (almost) independent when conditioned on Λ. Additionally, Λ must not contain any information solely about X which is not contained in Y (and vice versa). This means that Λ and X must be approximately independent, conditioned on Y (and Λ and Y must be approximately independent, conditioned on X). Mathematically, these three conditions are written:

$$I(X : Y|\Lambda) \leq \epsilon\,, \quad I(X : \Lambda|Y) \leq \epsilon\,, \quad I(Y : \Lambda|X) \leq \epsilon\,. \tag{1}$$

The larger the value of ϵ, the more 'approximate' the latent. If all three of these conditions are satisfied, we say that Λ is an approximate natural latent with error ϵ. We are interested in a variation of this problem. In particular, if there is exists a latent Λ which satisfies the conditions in (1), we want to ask whether there always exists a different latent Γ, characterised by $P(\Gamma|X, Y)$ which satisfies the following three conditions:

$$I(X : Y|\Gamma) \leq \delta, \quad H(\Gamma|X) \leq \delta, \quad H(\Gamma|Y) \leq \delta\,. \tag{2}$$

If a latent Γ satisfies these three conditions we say that it is a *deterministic natural latent* with error δ. This is because, if $H(\Gamma|X)$ is low, we can say that Γ is approximately a deterministic function of X (and similar for $H(\Gamma|Y)$). In particular, we want to know whether we can always construct a deterministic natural latent with error that is low compared to the original ie. we wish to characterise how low δ can be made, compared to ϵ.

Research Questions

- How small can δ be made compared to ϵ?
- Does this depend on the dimensions of X and Y?
- How do errors in the different inequalities trade off against each other?
- What implications would a result (positive or negative) of this kind have for 'ontology identification' in AI systems?

Relevance to AI Safety

If two latents Λ_1 and Λ_2 both satisfy the conditions (1), then we can prove that Λ_1 is approximately a function of Λ_2. If we interpret the two latents as being different 'models' of the data learned by two different agents, then this gives us a condition for when we can translate between different models of the world (eg. the model used by an AI and the model used by a human). Deterministic natural latents capture the degree to which this shared information is fully captured by either X or Y.

If you have experience in maths, physics, or computer science and are interested in working on problems like this, apply using the QR code. Rolling acceptance until application closes on May 17.

This job is part of an Advanced Research + Invention Agency-funded project.

To see more problems or apply scan here or visit dovetailresearch.org