Block 5: Identifying belief geometry in transformers

B.3 CompMech

Up to this point, we have focused on conceptual aspects of computational mechanics; now we will consider whether this framework applies to neural networks (e.g., transformers & RNNs). In this session, you will read the key sections of a few select papers, then discuss the findings in small groups.

1. ***Readings (45 minutes).*** *Follow the guide outlined in the Readings Section.*
2. ***Discussion (15 minutes).*** *Form groups of 3-4 based on how much you read. Discuss aspects of the readings you find interesting/confusing. See the Discussion Section for prompts you may wish to consider.*

Readings

This reading session consists of a main reading and an extension reading. Focus on the main reading, and proceed to the extension reading if you have time. If you can answer the 'Evaluation' question it is likely you have understood a key component of the reading. Feel free to skim the overview of the extension reading if you don't have time to read it.

[Main] Transformers represent belief state geometry in their residual stream

Reading suggestion:

- **Skim.** Sections 1, 2.1 & 2.2.
- **Read.** Sections 2.3, 3–5.
- **Evaluation.** Explain Figure 6 D.

Overview. This was the first paper to present evidence that transformers trained on HMM data admit linear(!) representations of the belief state in their residual stream. Evidence for this claim is presented for two processes Mess3 & random-random XOR (RRXOR). For each process linear regression is used to identify an affine map from residual stream activations to the corresponding belief state, which is then evaluated by computing the mean-squared error (MSE). MSE is evaluated across model checkpoints and is shown to decrease through training. A shuffle-control is used to illustrate that the low MSE from activations to beliefs is not simply explained by projecting from a high-dimensional space (the activations) to a low-dimensional space (beliefs).

[Extension] Neural networks leverage nominally quantum and post-quantum representations

Reading suggestion:

- **Skim.** Sections 1–3
- **Read.** Sections 4–8.
- **Evaluation.** Explain Figure 3 B. How are light-blue (orange) plots related to the blue (orange) plots?

Overview. This paper presents evidence that the Shai et al. result extends to generalised hidden Markov Models (GHMMs) – a strictly more expressive class of processes than HMMs, allowing for quantum and post-quantum representations. To illustrate, there exists stochastic processes that admit realisation in terms of a GHMM with finite-dimensional transition matrices, while the corresponding minimal HMM requires exponentially many hidden states.

Using techniques similar to Shai et al., it is shown that neural networks (transformers, LSTMs, RNNs, GRUs) trained on stochastic processes admitting distinct GHMM and HMM realisations learn belief state representations corresponding to the GHMM. This demonstrates the ability of neural networks to flexibly select the model class underlying their representations to capture exponential savings.

Discussion

Here are some discussion prompts to consider:

- What is the most convincing evidence that transformers represent belief state geometry?
- What is the least convincing?
- What additional experiments would you run to gain more evidence?
- Are there alternative hypotheses that explain these results?

(*Feel free to ignore these if you/your group is excited to explore other aspects of the readings.*)