Constrained Belief Updating and Geometric Structures in Transformer Representations

Mateusz Piotrowski Paul Riechers[1] Daniel Filan[2] Adam Shai[1]

[1]Simplex [2]MATS

Background

Previous work has shown that transformers represent belief states over the data generating process in their residual stream.
The mechanistic understanding of how transformers compute these beliefs remained unclear, as transformers compute in parallel while belief updating is recursive.

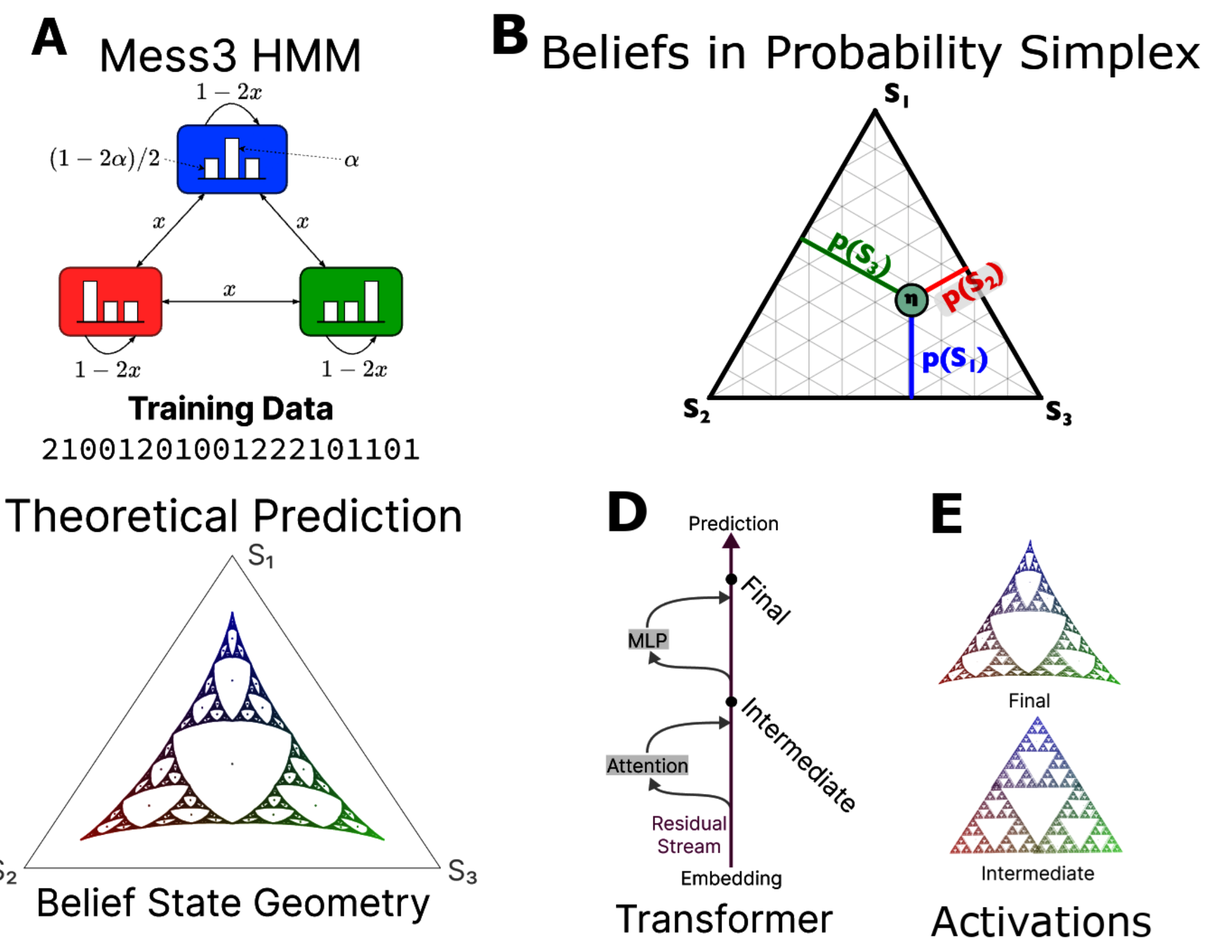
Core Question

How do transformers build belief state geometries from their architectural constraints?

Key Findings

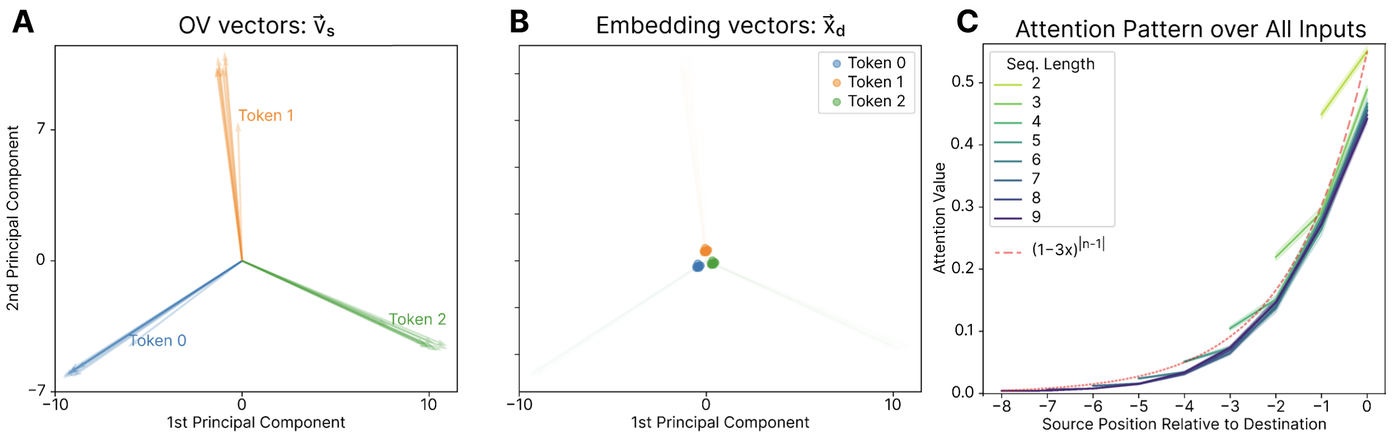
- Starting from the data-generating process, we find geometric structures in intermediate representations that are:
  - Fractal patterns distinct from final belief state geometry
  - Directly visible in PCA of model activations
  - Consistent across different process parameters
- We show how transformers construct these representations given architectural constraints:
  - Attention combines information through geometric vector addition
  - Parallel computation replaces traditional recursive updates
  - Theory predicts minimum architectural requirements (e.g., when multiple attention heads are needed)
- Theory bridges belief geometry and transformer mechanisms:
  - Explains construction of intermediate fractals
  - Shows how architectural constraints shape updates
  - Connects model structure to belief state geometry

Intermediate representations



We train a transformer to predict sequences from HMM (A). For optimal prediction, an observer should maintain beliefs - probability distributions over hidden states - that can be visualized on a probability simplex (B). Plotting all possible belief states reveals a fractal-like pattern (C). Remarkably, the transformer's internal representations organize into similar patterns in both its final and intermediate layers (D, E).

Intermediate representations are built by algorithms in the belief simplex

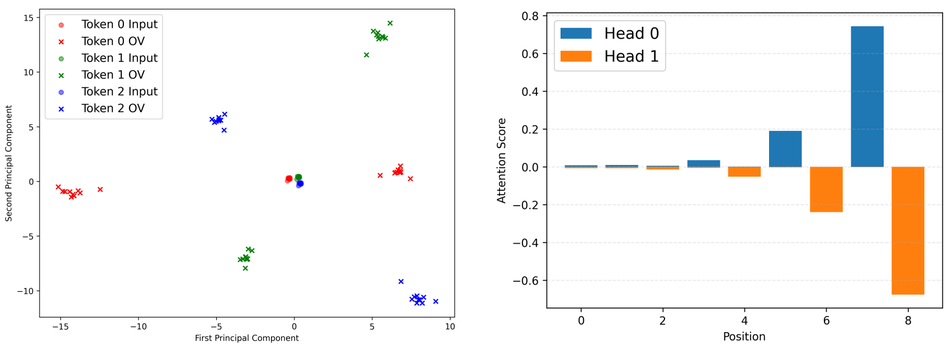


Construction of intermediate representations through attention. The OV circuit creates three update vectors (A, B) combined via decaying attention weights (C). (D) shows progressive vector addition for sequence "01120", with gray dots showing all possible states.

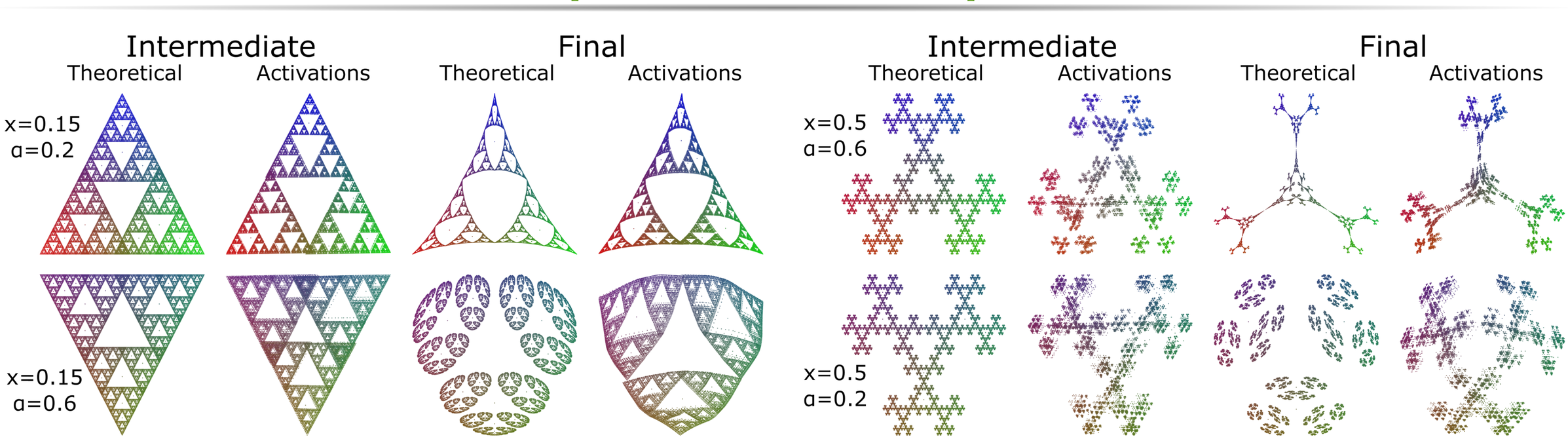
Constrained belief updating equation

$$r^{(L)} = \eta_\infty + \sum_{n=0}^{L-1} \left(\eta_\infty T^{|a_{L-n}} T^n - \eta_\infty \right)$$

This equation describes how transformers construct belief states and predicts the minimal architectural requirements needed. Let me show you how!



Model representation and theoretical predictions



Both intermediate and final representations (right) closely match theoretical predictions (left) across different HMM parameters, showing our theory generalizes across data-generating processes despite their distinct geometric patterns.