
Constrained Belief Updates and Geometric Structures in Transformer Representations for the RRXOR Process¹

Brianna Grado-White

With
PIBBSS, Timaeus, & Apart Research

Abstract

In this work, we seek to further understand the computational structure formed by transformers trained to predict the next token of a data set. In particular, we extend the work of (Piotrowski et al., 2025) by analyzing if and how these transformers implement constrained Bayesian updating, focusing on data generated by a Random-Random-XOR (RRXOR) process. Unlike the MESS3 processes studied originally in (Piotrowski et al., 2025), the RRXOR process leads to slightly more nuanced theoretical and practical models. Nevertheless, we verify some of the main claims of (Piotrowski et al., 2025). Namely, we find attention patterns that decay exponentially with distance. We further analyze the behavior of OV vectors.

Keywords: Physics-informed AI safety, Computational Mechanics, Mechanistic Interpretability

1. Introduction

With increasing rapidity over the last several years, AI has become ubiquitous – permeating major aspects of life by dispersing political information, automating insurance claims decisions, or acting as a pseudo companion and psychologist to people in mentally vulnerable states. No matter how positively transformative this technology has already proved to be and still promises to soon become, ceding vast amounts of decision power to a black box is fundamentally unwise. Understanding

¹ Research conducted at the AI Safety x Physics Grand Challenge, 2025

the world models and algorithms models use is thus of utmost importance to AI safety. (Shai et al., 2024), made important strides in this regard, as it used examples from computational mechanics (in which optimal prediction is well understood) to understand the world model formed by transformers, showing that the non-trivial structures dictated by optimal prediction can be found in models’ residual streams. The framework of computational mechanics draws on fundamental concepts in physics, to give concrete, theoretically understood mechanisms by which transformer-implemented algorithms can be benchmarked against. This in turn allows us to understand transformers not as arbitrary functions, but as systems building world models and attempting to implement optimal prediction. If this were a universal fact (even approximately), and belief states could be read off by searching for self-similar structures within model activations, this would mark an important foothold for interpretability (and therefore alignment).

As pointed out by (Piotrowski et al., 2025), however, there is a naive tension between the near-optimal next token prediction accuracy that transformers are able to achieve and the architectural constraints imposed by transformer: optimal prediction requires recursive Bayesian belief updating, while transformers rely on parallelized attention mechanisms. As (Piotrowski et al., 2025) show, transformers nevertheless form geometrically structured representations of an approximate Bayesian inference under these architectural constraints. In this work, we begin to extend that analysis to transformers trained on RRXOR data, which pose additional theoretical and practical subtleties. Unlike previous cases where belief geometry was localized to a single part of the residual stream, RRXOR-trained transformers distribute their representations across multiple layers. Further, the associated transition matrix has a different structure (in particular, the transition matrix, marginalized over token emissions, contains multiple non-zero eigenvalues) that further complicates the original theoretical framework.

Despite these differences, we demonstrate that a key insight from (Piotrowski et al., 2025) remains valid. Specifically, we verify that the constraints of the attention pattern scaling with token distance still hold theoretically, and can be found experimentally in the final layer. We also analyze the OV (output-value) circuit patterns in this setting, though the conclusions of this analysis are less transparently analogous to (Piotrowski et al., 2025).

2. Methods

Technical Implementation: We trained a 4-layer transformer with 4 attention heads (with no layer normalization) each on data generated by the Random-Random-XOR (RRXOR) process, depicted in Figure 1. Throughout, our methods closely mirror those of (Piotrowski et al., 2025), which focuses on the simpler MESS3 processes. Unlike with MESS3 processes, for RRXOR, distinct belief states can give the same next token predictions. As noted in (Shai et al., 2024), this degeneracy is correlated with the fact that the belief state geometry formed by the transformer can no longer be found at some single place in the residual stream. Instead, we extract the belief state geometry by concatenating the

residual stream after each attention block and regressing onto the theoretical predictions. Code for generating the data and training can be found at [RRXOR_training](#). Code for analyzing the transformer outputs can be found at [RRXOR_attention_analysis](#). As noted in these files, we borrow heavily from the [epsilon-transformer repository](#)². Additional architectural and training parameters are detailed in the Appendix. We verify the success of the training by the successful extraction of the belief state geometry in the transformers residual stream, verified visually (see Figure 2 below). We then analysed the attention patterns of these transformers, both empirically and theoretically. Finally, we study the OV vectors associated with the transformer.

Random-Random XOR

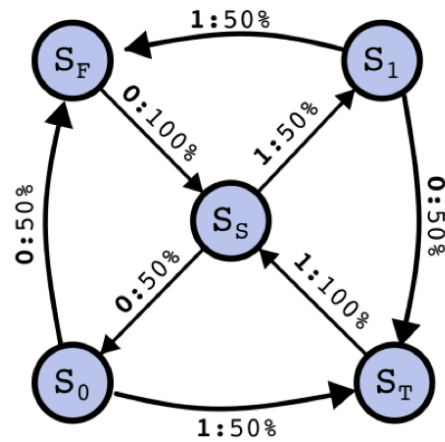


Figure 1: An image of the RRXOR process, stolen from (Shai et al., 2024).

²In addition, the first notebook also borrows from code used in a previous Apart Computational Mechanics Hackathon (in order to easily implement code in the epsilon-transformer repository). This is additionally noted in the file, in appropriate places

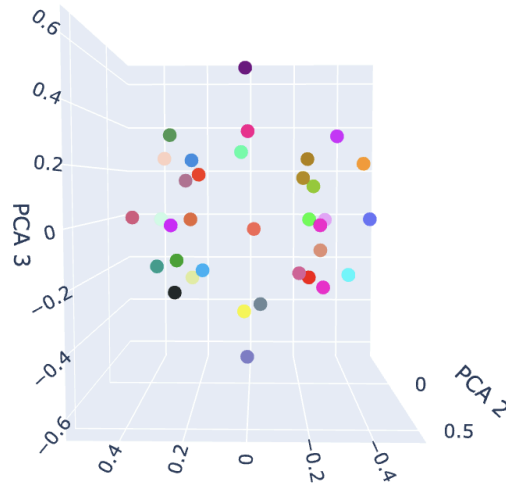


Figure 2. After concatenating activations in the residual stream at each layer, and doing a PCA analysis to find the optimal subspace, we find a representation of the belief simplex that matches the theoretical predictions. (For clarity in plotting, we bin an average points associated with the 36 belief states).

Theoretical Framework: We rely on the computational mechanics framework applied to transformers as developed in (Piotrowski et al., 2025; Shai et al., 2024). This framework dictates both the data generation, the form of belief state geometry to be found in the residual stream, and the approximate form of Bayesian updating that occurs because our models are constrained by the parallel nature of attention built into the transformer architecture. Though our main results are empirical, we make a small extension to circuit analysis of (Piotrowski et al., 2025), verifying that key elements still carry over in the RRXOR setting.

3. Results

- **Theoretical Insights:** As noted in the Methods Section, we make a small generalization to the results of (Piotrowski et al., 2025). In particular, we verify that the approximate Bayesian update process associated with constraints from the parallel nature of the transformer architecture continues to imply attention patterns to decay exponentially with distance between tokens in the RRXOR setting. In particular, this involves generalizing the setting of (Piotrowski et al., 2025) (which focused on the MESS3 setting), to the case where the transition matrix associated to data generation has more than one non-unity eigenvalue (restricting to the case of interest here, in which two complex-conjugate eigenvalues are present). The details of this derivation can be found in the Appendix.
- **Empirical Findings:** We verify that these theoretical constraints can still be found empirically in our trained transformer. First, we look for attention patterns with the expected exponential decay. This is found in the last

attention layer, as seen in Figure 3 below. Note that this is contrast to the results of (Piotrowski et al., 2025), which finds that (for MESS3 data, and in models with multiple layers) the first layer implements the constrained Bayesian updates (with associated decaying attention patterns), with later layers transforming this constrained belief geometry into the theoretically optimal one in later layers. Here, as mentioned above, the optimal belief state geometry cannot be found in any one layer. As a seeming consequence, the theoretically predicted attention pattern is not found until the final layer.

Additionally, we inspect the OV vectors, as in (Piotrowski et al., 2025). There, the theoretical analysis predicts that OV-vectors associated with the same token are necessarily parallel, so that each update to the belief state pushes it towards the direction of a corner of the belief state simplex. We note that this structure does not follow (at least at first glance) in the case of RRXOR data. Nevertheless, we empirically attempt to find such structure in our trained transformer. Unfortunately, to go along with our theoretical shortcomings, we find no such analogous structure in the transformer (we leave understanding whether this is due to errors in analysis or fundamental theoretical differences to future work).

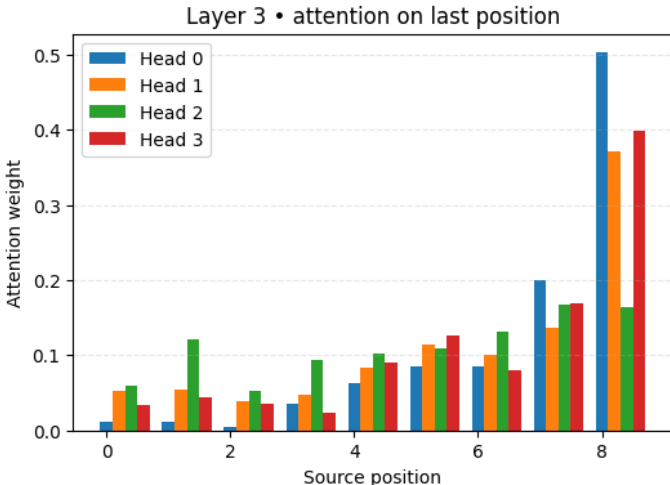


Figure 3. Attention patterns in the last layer are determined by positional distance between the destination and source tokens. These appear to follow an exponential decay (with coefficient determined by the real part of the non-zero, non-unity eigenvalue of the RRXOR transition matrix). Due to constraints on time, we do not actually fit this curve to verify the exponential nature, instead relying on visual inspection.

4. Discussion and Conclusion

Theory-Practice Bridge: Following the methodology of (Piotrowski et al., 2025), we continue to study the intersection of physics-inspired computational mechanics problems with the practicalities of mechanistic interpretability, showing that even in slightly more complicated settings, transformers still seem to manage implement parts of recursive Bayesian updates, in spite of being constrained to parallel computations due to attention. This setting provides settings in which theoretical, optimal/constrained Bayesian predictions can be concretely realized in transformer settings, giving a top-down explanation for the structure of computations and representations built by the transformer.

Limitations & Future Directions: Obvious limitations and future directions for these results arise from the incomplete analysis of the OV vectors, for which we did not find simple theoretical or empirical results for. This can likely be resolved with more time, and we leave this work to future projects. With an eye towards more ambitious goals, we note that though these toy models offer the ability to see how a transformer can build an accurate world model of the data it is trained on, we recognize the obvious limitation that we are analyzing small models trained on highly artificial data. In the natural language settings of most interest to safety, the data will have more complex structure, there will be additional bottlenecks to the representations (as transformers lack sufficient parameters to make exactly optimal predictions in messy real-world data), and the model architectures will correspondingly more complex (e.g. with the inclusion of layer norms throughout). Additionally, it is known that not all networks with doing optimal prediction represent belief state geometries linearly in their residual stream (see e.g. (Goldstein, 2024)). Understanding which aspects of architecture and training lead to solutions with this property remains an important future direction.

Impact Assessment: These results suggest that some aspects of the general framework developed by (Piotrowski et al., 2025) continue to hold outside the simplest settings. In particular, even when the belief state geometry is distributed across multiple layers in the residual stream, remnants of the predicted constrained Bayesian updating process can still be found. We hope these results will ultimately give insight beyond the toy-model tasks studied here, leading to a deeper understanding of how language models form and use accurate world models in their predictions.

References:

Goldstein, R. (2024). *Handcrafting a Network to Predict Next Token Probabilities for the Random-Random-XOR Process*.

<https://apartresearch.com/project/handcrafting-a-network-to-predict-next->

token-probabilities-for-the-random-random-xor-process

Piotrowski, M., Riechers, P. M., Filan, D., & Shai, A. S. (2025). *Constrained belief updates explain geometric structures in transformer representations* (No. arXiv:2502.01954). arXiv. <https://doi.org/10.48550/arXiv.2502.01954>

Shai, A. S., Marzen, S. E., Teixeira, L., Oldenziel, A. G., & Riechers, P. M. (2024). *Transformers represent belief state geometry in their residual stream* (No. arXiv:2405.15943; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2405.15943>

5. Appendix

Note, that for the RRXOR process, the transition matrices are defined by³

$$T^{(0)} = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad T^{(1)} = \begin{bmatrix} 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Summing over the potential emitted tokens (here, adding together each row) yields the state transition matrix

$$T = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

³ We apologize for sloppy formatting throughout this Section. The initial document template was given as a Google Docs (not Latex) format, and we do not enjoy equation manipulation here.

The eigenvalues for this matrix are $\{1, -0.5 + 0.87i, -0.5 - 0.87i, 0\}$, (where the zero eigenvalue has degeneracy 2). The corresponding eigenvectors are

$$[0.71, 0.71, 0.71, 0, 0],$$

$$[-0.18-0.31i, 0.18 + 0.31i, 0.35, -0.5, -0.5],$$

$$[-0.18 - 0.31i, -0.18+0.31i, 0.35, 0.5, 0.5],$$

$$[-0.18 + 0.31i, -0.18-0.31i, 0.35, 0.5, -0.5],$$

$$[-0.18 + 0.31i, -0.18-0.31i, 0.35, -0.5, +0.5],$$

(unnormalized, and not orthogonal, here) respectively. Note that these vectors are linearly independent, and that Equation 11 of (Piotrowski et al., 2025) still holds. Here, however, we have two non-zero, non-unity eigenvalues: the complex conjugates $-0.5 + 0.87i$ and $-0.5 - 0.87i$. Equation 11 thus simplifies to

$$A_{d,s}f(\vec{v}) = \xi^{d-s} \pi T^{|z_s} T_\lambda + (\xi^*)^{d-s} (\pi T^{|z_s} T_\lambda^*),$$

where we note that the projectors associated to the complex-conjugate eigenvalues are also complex conjugate. This simplifies to

$$A_{d,s}f(\vec{v}) = \pi T^{|z_s} (2\text{Re}(\xi^{d-s} T_\lambda)).$$

Though this cannot be factored as cleanly as in the MESS3 case (which only had one, non-zero, non-unity eigenvalue), we can still conclude that

$$|A_{d+m,s}| \sim |\zeta^m| |A_{d,s}|,$$

so that $A_{d+m,s}$ decays with ζ (up to additional oscillatory behavior, due to the complex phase). The presence of the complex eigenvalue pairs complicates finding an exact, general expression for f (the analogue of equations 14 and 15 of (Piotrowski et al., 2025)), and we leave such general, analytic expressions to future works with longer timescales.

The configuration settings used for training are:

n_ctx: 10

act_fn: relu

d_head: 16

d_model: 64
d_vocab: 5
n_heads: 4
n_layers: 4
optimizer: adamw
batch_size: 128
num_epochs: 500
weight_decay: 0.0
attention_dir: causal
learning_rate: 0.0003
normalization_type: LN
fixed_train_dataset: False
num_train_samples: 100000
fixed_valid_dataset: True
num_valid_samples: 2000