ν-Wave State Space Models: Traveling waves as a biologically plausible context

T. Anderson Keller – *The Kempner Institute at Harvard University*

Abstract: Mounting evidence suggests that traveling waves are a key dynamical motif in biological neural information processing systems [Muller et al. (2018)], yet their precise computational role remains under investigation. While the Transformer architecture [Vaswani et al. (2023)] has achieved state-of-the-art performance across various data modalities, it relies heavily on an explicit 'context' – processing the entire input sequence simultaneously. This approach is both biologically implausible and computationally inefficient, resulting in quadratic complexity relative to sequence length. Recent interest has shifted toward State Space Models (SSMs) [Gu et al. (2022)] as a potential 'context-free' alternative to Transformers, demonstrating comparable language modeling performance [Gu & Dao (2024)]. In this work, we demonstrate that these SSMs inherently implement a form of traveling wave dynamics with a fixed velocity ($\nu = 1 \frac{neuron}{timestep}$). However, this fixed wave velocity limits their capacity to approximate the Transformer's context mechanism, leading to poor performance on context-dependent tasks like sequence copying [Jelassi et al. (2024)]. Building on this insight, we introduce the ν-Wave SSM – a framework that generalizes SSMs by incorporating variable wave velocities ($\nu \neq 1$). By allowing for adjustable wave dynamics, our model effectively enhances the approximation of an explicit context. Empirically, we demonstrate that our model learns exponentially faster and achieves significantly lower error rates on large-scale memory-dependent tasks, matching the performance of Transformers previously thought unattainable by SSMs; showing that Transformer-like 'context' can be effectively implemented in a more biologically plausible manner through wave dynamics. Our findings bridge the gap between artificial and biological neural processing, offering new insights into the role of traveling waves in neural information processing and memory. This work underscores the importance of further investigating traveling waves in natural neural systems and explains their adoption in state-of-the-art recurrent neural network language models [Beck et al. (2024)].

Additional Detail: In the field of machine learning, Transformer architectures [Vaswani et al. (2023)] have reached state-of-the-art broadly across many data modalities; surpassing the recurrent neural networks which previously held their place on tasks such as translation and language modeling. Many have attributed this superior performance to the ability of transformers to processes entire sequences in parallel through the use of a 'context'. In settings such as language generation, this context explicitly stores the entire sequential history of preceding words, facilitating the ability of the network make comparisons and learn relationships over long distances. This maximal flexibility has arguably led to the strong performance of transformer models on a range of tasks, but also comes with computational drawbacks. Primarily, this all-to-all attention is costly to compute, scaling quadratically in sequence length. Additionally, this explicit context is biologically implausible, as biological neural networks are clearly highly recurrent without any known mechanisms to maintain an explicit copy of the entire past history.

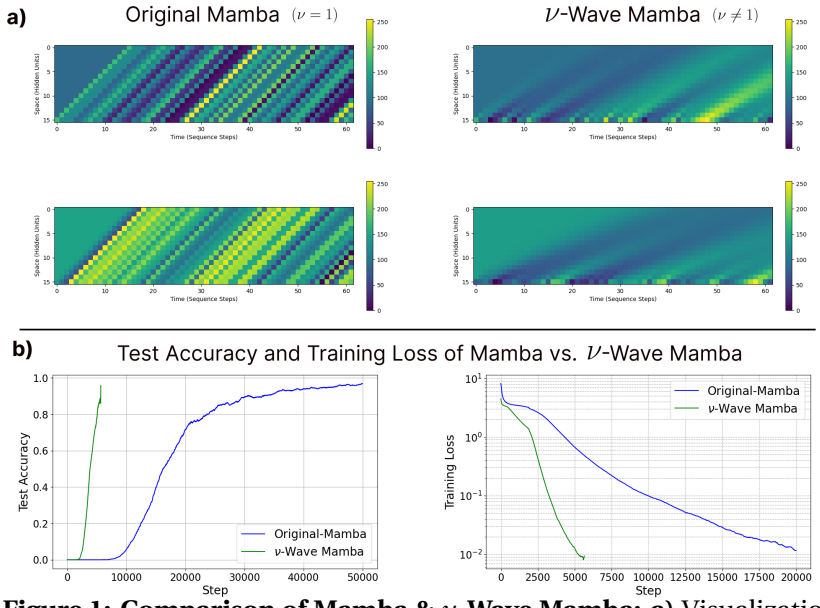

Figure 1: Comparison of Mamba & ν-Wave Mamba: a) Visualization of hidden state (vertical axis) over time (horizontal axis) for the original Mamba model (left), and our proposed ν-Wave Mamba (right). We see that traveling waves are visible in both cases as bands of diagonal activity, while the ν-Wave Mamba model has waves which propagate with a variable velocity ($\nu \neq 1$). **b)** Empirically (bottom) we show this simple modification yields dramatic improvement of learning speed on a length 20 copy task by enhancing the 'context' of the SSM.

Recently, State Space Models (SSMs) [Gu et al. (2022)] have gained attention as a potential alternative to Transformers, and crucially operate as recurrent neural networks, like their biological counterparts. Such

models have demonstrated impressive performance at long sequence modeling, and language modeling, yet still lag behind Transformer models on specific tasks which are adversarially designed to require a long explicit context. For example, recent work has shown that on a simple copy task where the model is provided with a sequence of up to 300 characters that it must repeat precisely, transformers learn significantly faster and generalize to significantly longer sequences than comparable SSMs [Jelassi et al. (2024)]. The authors argue, and show theoretically, that this difference in performance is due to the fact that the memory (or context) of SSMs is inherently limited.

In the neuroscience community, interest in spatiotemporal neural dynamics has been growing in recent years. A specific spatiotemporal pattern, namely traveling waves, have been observed across the brain at a diversity of spatial and temporal scales, spanning brain regions from hippocampus to primary visual cortex [Muller et al. (2018)]. Following these observations, a suite of hypotheses for their potential function have been proposed. Recent perspectives [Muller et al. (2024)] have argued that traveling waves may be used as a means of storing information about the recent past, and recent empirical work has demonstrated the effectiveness of such a method in simple recurrent neural network models [Keller et al. (2024)].

Interestingly, with the correct viewpoint, all modern state space models can be seen to actually already incorporate traveling wave dynamics to form a type of context. Explicitly, most modern state space models can be seen to be based on the 'H3 block' from [Fu et al. (2023)], meant to mimic a linear form of self-attention. This block is composed of a 'shift-SSM' and a diagonal SSM which interact multiplicativity to be able to perform token comparisons across time. With careful interpretation, the shift-SSM can be viewed equivalently as the finite difference approximation to the one-way one-dimensional wave equation: $\frac{\partial f}{\partial t} = \nu \frac{\partial f}{\partial x}$. Explicitly, it is constructed as a standard state space model ($\boldsymbol{h}_{t+1} = \boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{B}\boldsymbol{x}_t$, $\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{h}_t$), with the constraints that $\boldsymbol{A}$ is given as a shift operator (a zero matrix with ones along the first upper diagonal), and $\boldsymbol{B}$ is given by a canonical basis vector ($\boldsymbol{e}_0$). As described in prior work [Keller et al. (2024)], this shift matrix is exactly derived from a discretization of the above wave equation with $\nu = 1$. In Figure 1 (left), we plot the hidden state of the Shift-SSM of Mamba [Gu & Dao (2024)], a state of the art SSM, and visualize concretely that there exist traveling waves of activation propagating through the hidden state.

Inspired by this finding, we investigated whether the greater flexibility afforded by having variable velocity waves may improve the 'context' capabilities of state space models, and thereby serve as evidence for traveling waves as a general biological implementation of 'context', as put forth by Muller, Churchland and Sejnowski [Muller et al. (2024)]. In Figure 2 we compare the performance of such a modification (denoted ν-Wave Mamba), compared with the original Mamba, and two transformer architectures with different position encodings (alibi and rope) [Press et al. (2022); Su et al. (2023)]. We see that indeed, the variable velocity Mamba performs on-par with the transformer models, and learns exponentially faster than the $\nu = 1$ counterpart, effectively solving the copy task that was previously deemed impossible for SSMs due to their limited context capabilities. Overall, we present this work as evidence that Transformer-like context can be effectively implemented in a more biologically plausible manner through wave dynamics, and hope that it serves to encourage further investigation into traveling waves in natural neural systems.

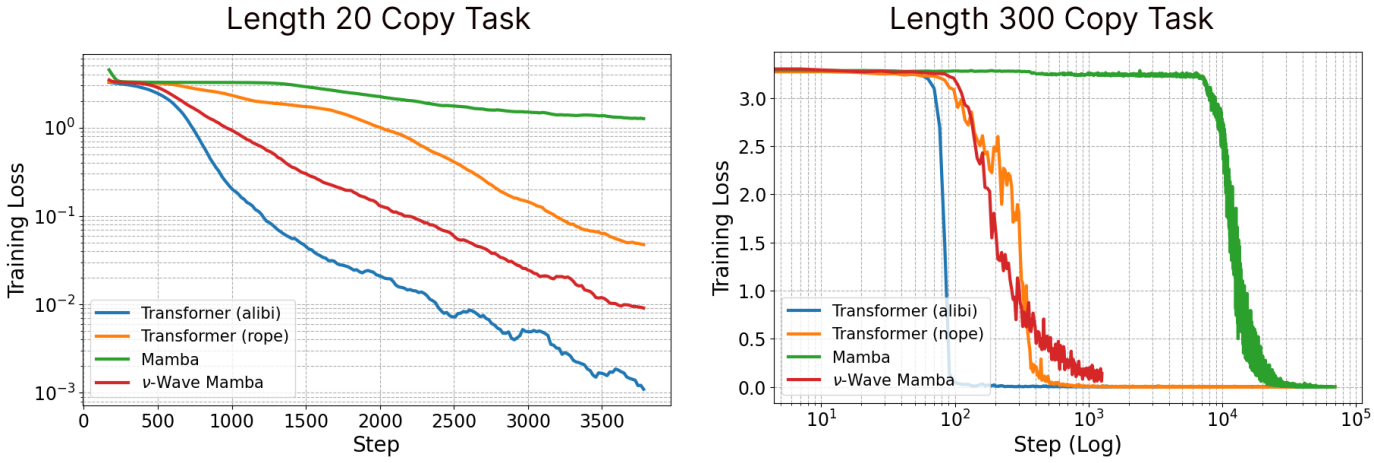

Figure 2: Comparison of training loss of multiple transformer models (blue, orange), the original Mamba architecture (green), and the newly proposed ν-Wave Mamba with variable velocity waves for copy tasks of length 20 (left), and 300 (right), as tested in [Jelassi et al. (2024)]. We see by re-interpreting the hidden state in terms of traveling wave dynamics, the SSM now performs on par with the transformer models on the copy task, learning exponentially faster.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., … Hochreiter, S. (2024). *xlstm: Extended long short-term memory.* Retrieved from https://arxiv.org/abs/2405.04517

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., & Ré, C. (2023). *Hungry hungry hippos: Towards language modeling with state space models.* Retrieved from https://arxiv.org/abs/2212.14052

Gu, A., & Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces.* Retrieved from https://arxiv.org/abs/2312.00752

Gu, A., Goel, K., & Ré, C. (2022). *Efficiently modeling long sequences with structured state spaces.* Retrieved from https://arxiv.org/abs/2111.00396

Jelassi, S., Brandfonbrener, D., Kakade, S. M., & Malach, E. (2024). *Repeat after me: Transformers are better than state space models at copying.* Retrieved from https://arxiv.org/abs/2402.01032

Keller, T. A., Muller, L., Sejnowski, T., & Welling, M. (2024). Traveling waves encode the recent past and enhance sequence learning. In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=p4S5Z6Sah4

Muller, L., Chavane, F., Reynolds, J., & Sejnowski, T. J. (2018, March). Cortical travelling waves: mechanisms and computational principles. *Nature Reviews Neuroscience*, *19*(5), 255–268. Retrieved from https://doi.org/10.1038/nrn.2018.20 doi: 10.1038/nrn.2018.20

Muller, L., Churchland, P. S., & Sejnowski, T. J. (2024). *Transformers and cortical waves: Encoders for pulling in context across time.* Retrieved from https://arxiv.org/abs/2401.14267

Press, O., Smith, N. A., & Lewis, M. (2022). *Train short, test long: Attention with linear biases enables input length extrapolation.* Retrieved from https://arxiv.org/abs/2108.12409

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). *Roformer: Enhanced transformer with rotary position embedding.* Retrieved from https://arxiv.org/abs/2104.09864

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2023). *Attention is all you need.* Retrieved from https://arxiv.org/abs/1706.03762