Appendix: Emergence of the Primacy Effect in Structured State-Space Models

Takashi Morita TMORITA@ALUM.MIT.EDU
Academy of Emerging Sciences, Chubu University, Japan
Institute for Advanced Research, Nagoya University, Japan

Editors: Hung-yi Lee and Tongliang Liu

Appendix A. Distributions of Δt across Training Runs

Figure A.1 reports the distribution of the Δt parameters before and after each of the ten training runs (decomposing the leftmost panel in Figure 5).

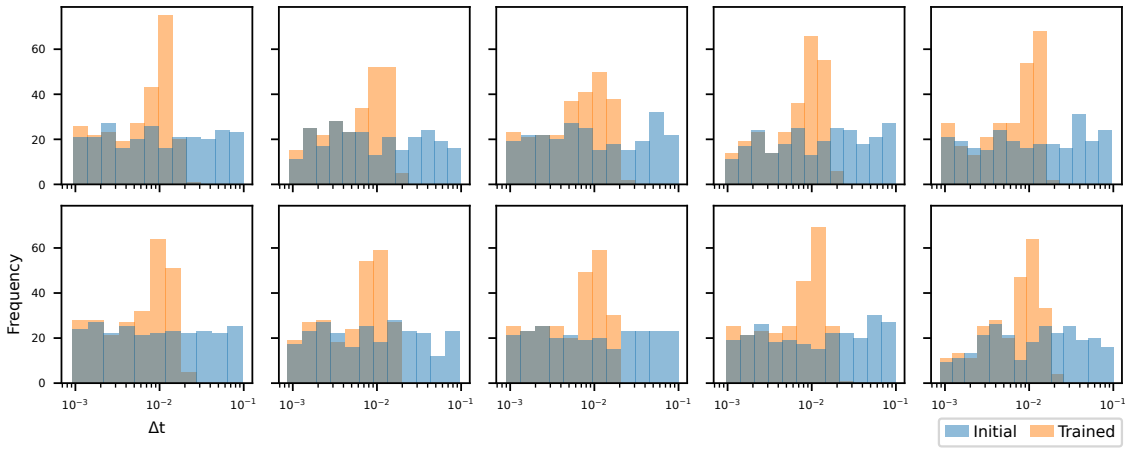

Figure A.1: Histograms displaying the initial (blue) and final (orange) values of Δt. Each panel presents the distributions obtained from one of the ten training runs with different random seeds.

Appendix B. Associative Recall

This section investigates the primacy effect in a more advanced task—*associative recall* (Fu et al., 2023)—compared to the binary memory verification. Similar to the memory verification setup, the associative recall task first presents a sequence of random tokens (study items) to the model, followed by a shuffled version of the same sequence (excluding the final token; Figure B.1).[1] Unlike the verification task, however, none of the shuffled

1. In the original formulation of the associative recall task, the model is presented with a single query token sampled from the study items. This study extends the task in a more empirical setting, requiring the model to recall multiple pairs of precedent and successor tokens within a single sequence.

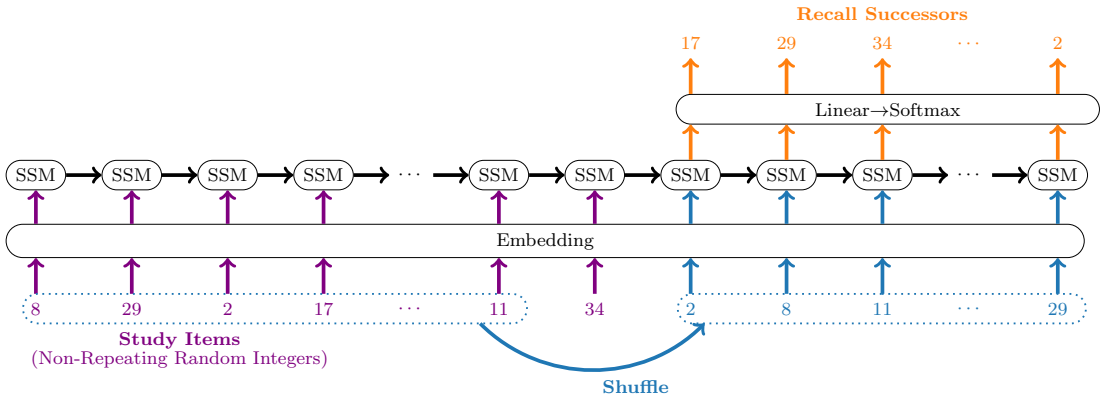

Figure B.1: Schematic illustration of the (extended) associative recall task.

query tokens are replaced with distractors; instead, the model is required to return the *immediate successor* of each query token. For example, given study items $(8, 29, 2, 17)$ and queries $(2, 8, 29)$, the correct outputs are $(17, 29, 2)$.

Despite its apparent simplicity, previous work has shown that the ability to recall the successor of previously presented input tokens is a meaningful indicator of language models' capacity (Olsson et al., 2022; Fu et al., 2023; Gu and Dao, 2024). As such, investigating associative recall helps bridge the primacy effect observed in SSMs with real-world applications, particularly in language modeling.

Model and task hyperparameters were manually tuned to ensure that overall accuracy remained suboptimal. Specifically, the length of the study items was set to $L = 96$, and the vocabulary size to $K = 128$. The number of latent channels in the SSM (also equal to the dimensionality of the input embeddings) was set to 1024. All other hyperparameters were identical to those used in the memory verification task (see §3.1).

Figure B.2 shows the SSM's accuracy on the associative recall task. The model exhibited higher accuracy for initial study items compared to terminal ones (vertical differences), despite overall retention being less persistent than in the memory verification task (horizontal declines). Besides this general trend, the model also appears to have learned delayed reconstructions of input sequences, as evidenced by the diagonal patterns observed in the heatmaps. This behavior may reflect a statistical property of the adopted task; input-output distances of length L occurred more frequently than either shorter or longer dependencies during training (see Footnote 6 in the main text).

References

Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024.

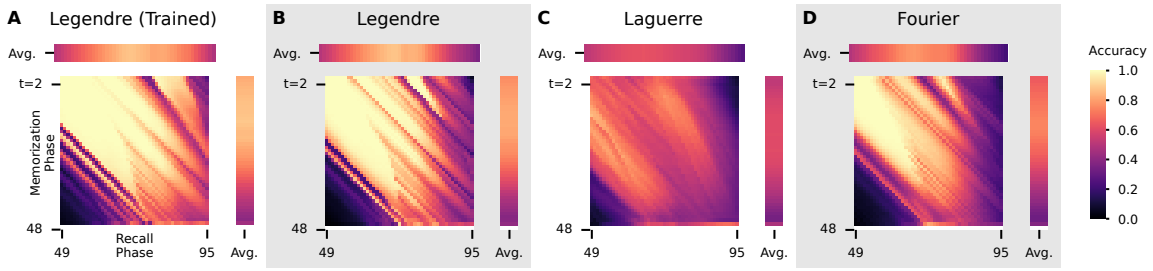

Figure B.2: Accuracy of the associative recall task performed by the SSM (S4) under different parameter configurations. Each cell in the square heatmaps represents the accuracy (or the recall score) for study items (target successors) that were presented at the time indexed by the corresponding row and queried at the time indexed by the corresponding column. The top separate row and the right separate column display the average accuracy across memorization times (rows) and recall times (columns), respectively. The state and input matrices of the SSM were initialized to approximate the latent dynamics of input sequences using the Legendre (A,B), Laguerre (C), and Fourier (D) polynomials. The state and input matrices were optimized for the task in panel A, and remained fixed at their initial values elsewhere. The discretization step size Δt was initialized in the range $0.001 \leq \Delta t \leq 0.1$. The length of study items was set to $L = 96$, and the vocabulary size to $K = 128$.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.