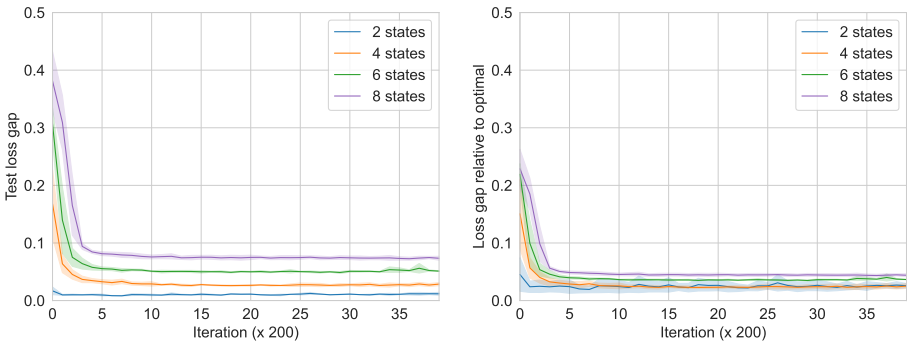
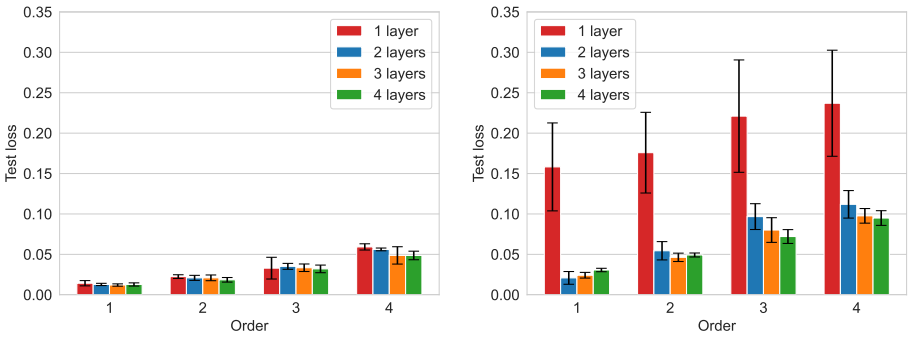


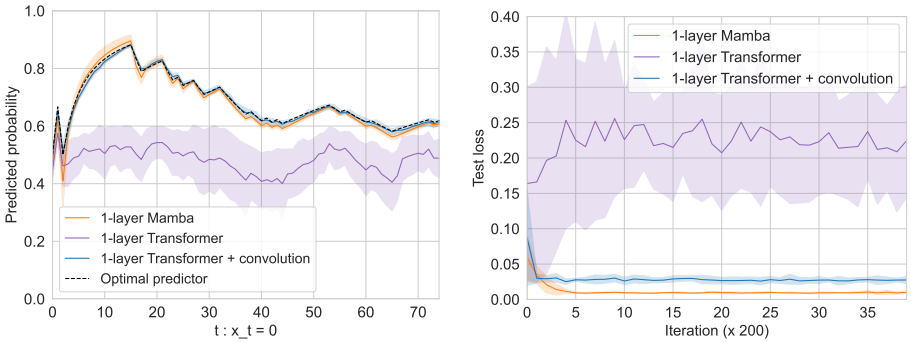
(a) Absolute test loss gap (b) Relative test loss gap

Figure 1: Test loss gap from the optimal for 1-layer Mamba and first-order Markov data, for different number of states. (a) shows the absolute gap $L(\theta) - L^*$; (b) shows the relative gap $(L(\theta) - L^*)/L^*$. The loss gap is consistently small for all state sizes.



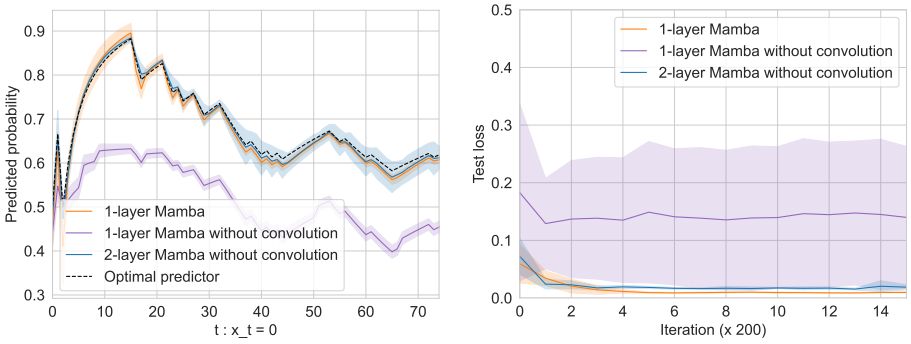
(a) Mamba (b) Transformers

Figure 2: Test loss gap from the optimal for Mamba and Transformers, for different number of layers and Markov orders. Mamba has a smaller loss gap than transformers across all orders. Furthermore, adding more layers to Mamba does not significantly improve performance. As expected, 1-layer Transformers cannot solve the Markov task, while 1-layer Mamba does.



(a) Predicted probability (b) Test loss

Figure 3: Predicted probability and test loss for Transformers with and without convolution. Adding convolution to the K, Q, V matrices of transformers makes the models succeed in learning the Markov task, similarly to 1-layer Mamba.



(a) Predicted probability (b) Test loss

Figure 4: Predicted probability and test loss for the full 1-layer Mamba and a 2-layer Mamba without convolution. Similarly to transformers, Mamba needs two layers to solve the Markov task when convolution is removed.

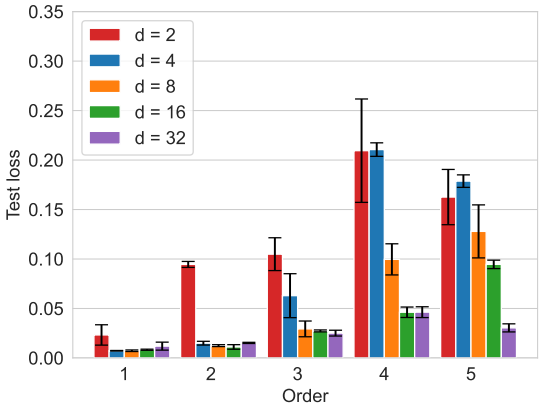


Figure 5: Relation between the Markov order k and the hidden dimension d of the 1-layer Mamba model. The plot shows that $d = 2^k$ is sufficient for the model to learn the k-th order Markov task. This corroborates the fact that Theorem 2 in the main paper has the correct order dependency.