Tel Aviv University[1]
Google[2]

NEURAL INFORMATION PROCESSING SYSTEMS

Provable Benefits of Complex Parameterizations for Structured State Space Models

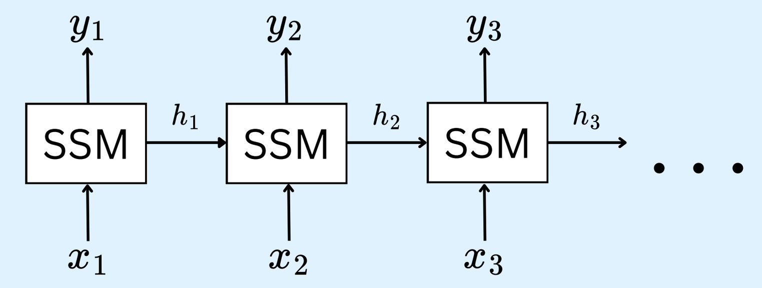
Yuval Ran-Milo[1], Eden Lumbroso[1], Edo Cohen-Karlik[1], Raja Giryes[1], Amir Globerson[1,2], Nadav Cohen[1]

We establish **formal gaps** between **real** and **complex SSMs** in terms of **expressiveness** and **practical learnability**

Structured State Space Models (SSMs)

An SSM is a linear RNN (linear dynamical system)

$$\mathbf{h}_t = A\mathbf{h}_{t-1} + Bx_t$$
$$y_t = C\mathbf{h}_t$$
$$A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times 1},\ C \in \mathbb{R}^{1\times n}$$



SSMs serve as the backbone for prominent neural networks like S4, Mamba, LRU

Real vs. Complex Parameterizations

SSMs can have real or complex parameterizations

$$\mathbf{h}_t = A\mathbf{h}_{t-1} + Bx_t$$
$$y_t = \Re(C\mathbf{h}_t)$$
$$A \in \mathbb{C}^{n\times n},\ B \in \mathbb{C}^{n\times 1},\ C \in \mathbb{C}^{1\times n}$$

Q: To what extent do complex parameterizations benefit SSMs?

Evidence in the literature is mixed
(Gu et al. 2022, Orvieto et al. 2023, Gu and Dao 2023, Ma et al. 2023)

Conjecture: Complex parameterizations are beneficial for continuous data, not for discrete data
(Gu and Dao 2023)

Theoretical Result I: Separation in Expressiveness

Denote by n_R and n_C the dimensions of real and complex SSMs

Denote by t the length of the input sequence.

A Complex SSM can precisely express any mapping realizable by a real SSM if $n_C \geq n_R$

The converse is not true:

Theorem (informally stated)

Many mappings realizable by a 1-dimensional complex SSM cannot be approximately expressed up to time t by a real SSM unless $n_R \approx t$

Theoretical Result II: Separation in Practical Learnability

What if n_R and n_C are large enough to realize a mapping up to time t?

Theorem (informally stated)

If a mapping satisfies a **mild condition**, then a real SSM requires $\exp(t)$ parameter values, which cannot be learned via GD

In contrast,

Theorem (informally stated)

Any mapping can be realized by a complex SSM with parameter values linear in t

This **mild condition** is satisfied by natural mappings:

Canonical Copy | Basic oscillatory | Random (generic)

Experiments

Errors in Theoretical Setting

Parameterization	Copy (↓)	Random (↓)	Oscillatory (↓)
Real	7.7×10^{-1}	5.3×10^{-1}	8.1×10^{-1}
Complex	$\mathbf{1.6 \times 10^{-5}}$	$\mathbf{6.3 \times 10^{-5}}$	$\mathbf{1.6 \times 10^{-4}}$

Accuracies in Real-World Setting

Parameterization	CIFAR-10 (↑)
Real	78.27%
Complex	**89.10%**

Complex parameterizations for SSMs significantly improve performance

Experiments with Selectivity

Selectivity is an SSM-based architecture yielding SotA performance.

Benefits of complex parametrization in selective SSMs are more nuanced:
We see benefits on some tasks, not on others

Accuracies in different tasks

Parameterization	Copy	Induction-Head
Real (highest)	80.17%	**98.35%**
Complex (highest)	**95.27%**	97.64%

Future Work: Theory Accounting for Selectivity

Without Selectivity: Real SSMs struggle to express oscillations unlike complex SSMs

With Selectivity: Experiments suggest real SSMs are as good as complex on some tasks

Hypothesis: Selectivity allows importing oscillations from the input

Hypothesis aligns with the conjecture of Gu and Dao as Continuous data contains only low frequencies, discrete contains all
(Gu et al. 2022, Orvieto et al. 2023, Gu and Dao 2023, Ma et al. 2023)

This May fully delineate benefits of complex parameterizations for SSMs