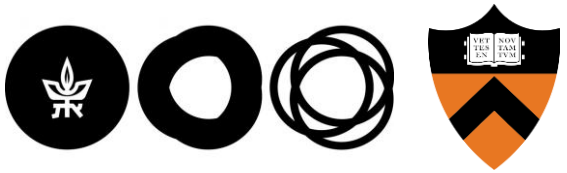
The Implicit Bias of Structured State Space Models Can Be Poisoned with Clean Labels

Yonatan Slutzky, Yotam Alexander, Noam Razin, Nadav Cohen

Tel Aviv University, Foundations of Deep Learning Lab

Princeton University



NeurIPS 2025 Spotlight Presentation

November 2025

Implicit Bias of Gradient Descent in SSMs

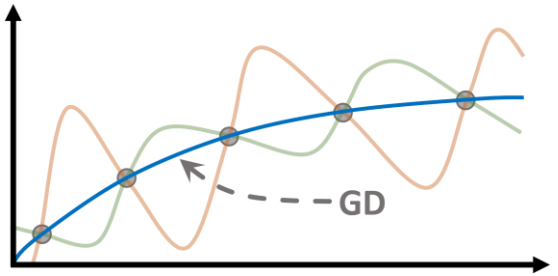
Phenomenon

With various DL models, optimizing via GD often leads to **generalization**, even when:

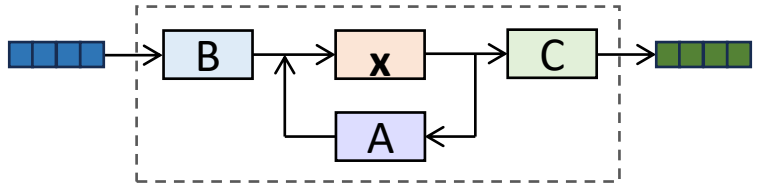
- Model size ≫ training set size
- There is no explicit regularization

Conventional Wisdom

GD induces **implicit bias** towards generalizing mappings



Structured state space models (**SSMs**) are sequence-to-sequence models underlying prominent neural networks, e.g., S4, Mamba



Goal: Theoretically analyze the implicit bias of GD with **SSMs**

Basic SSM

Single-in single-out linear dynamical system with diagonal transition matrix:

- **Parameters**

$$\mathbf{A} \in \mathbb{R}^{D \times D} \ , \ \mathbf{B} \in \mathbb{R}^{D \times 1} \ , \ \mathbf{C} \in \mathbb{R}^{1 \times D}$$

Diagonal transition matrix — Input matrix — Output matrix

- **Dynamics**

$$\text{For } h = 0,1,2, \ldots : \quad \mathbf{x}_{h+1} = \mathbf{A}\mathbf{x}_h + \mathbf{B}u_h \quad , \quad y_h = \mathbf{C}\mathbf{x}_h$$

Next state — Current state — Input — Output

Teacher-Student Setting

Consider (unknown) **teacher** SSM $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ with dim D^*

Given

Pre-recorded training set of **horizon** H: $\mathcal{T} = \{(\mathbf{u}^{(1)}, y^{*(1)}), \ldots, (\mathbf{u}^{(N)}, y^{*(N)})\}$

Input sequence $(u_0^{(1)}, u_1^{(1)}, \ldots, u_{H-1}^{(1)})$

Output at time H of teacher SSM $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ under input sequence $\mathbf{u}^{(1)}$

Goal

Learn mapping that fits teacher SSM up to **any horizon**

Method

Overparameterized student SSM $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ with dim $D \gg \max\{D^*, H\}$ trained via **GD** over:

$$\mathcal{L}_H(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \frac{1}{N} \sum_{n=1}^{N} \left(y_H^{(n)} - y^{*(n)} \right)^2$$

Output at time H of student SSM $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ under input sequence $\mathbf{u}^{(n)}$

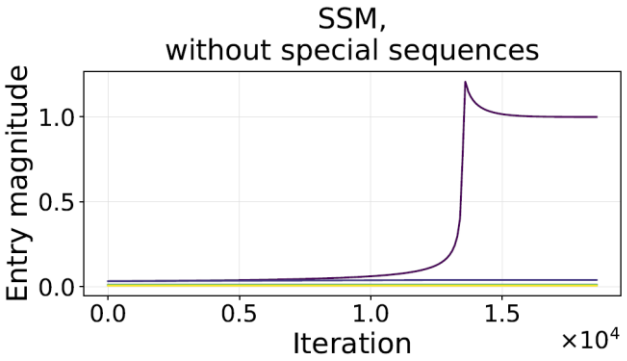
Dynamical Analysis: Greedy Low Rank Learning

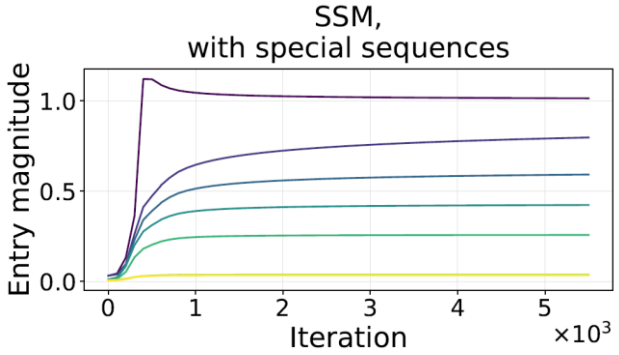
Proposition (informal)

In learning student SSM $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ via GD , if all $\mathbf{u}^{(n)}$ are **not "special"** then learned $\mathbf{A}$ exhibits **greedy low rank learning**

Sufficient condition for generalization (ground truth is low dim)

Experiment





Implicit Bias Can Be Disrupted by Special Training Examples

Theorem (informal)

Under technical conditions, $\forall H' > H + 1$, there exist:

- a training set $\mathcal{T}$ without special input sequences
- a special input sequence $\mathbf{u}^\dagger$ with label $y^\dagger$ generated by teacher SSM $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$

Clean label

s.t., when learning student SSM $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ via GD, generalization to horizon H':

- takes place if $\mathcal{T}$ is used on its own
- does not take place if $(\mathbf{u}^\dagger, y^\dagger)$ is appended to $\mathcal{T}$

Clean-label poisoning

Experiment

Setting	Without special sequences	With special sequences
SSM per Theorem	1.34×10^{-3}	4.1×10^{-2}
SSM beyond Theorem	1.94×10^{-1}	16.61
SSM in non-linear neural network	1.61×10^{-3}	5.39×10^{-2}

Recap

SSMs are an emerging, efficient alternative to Transformers

Typically, many weight settings achieve low training loss, only some generalize

Implicit Bias of GD

With low dim teacher, it provably:

- Leads to generalization in most cases
- Can be disrupted by special training examples (susceptible to clean-label poisoning)

Future Work

- Analyze the implicit biases of more complicated SSMs (e.g., Mamba)
- Use theoretical insights to derive practical defenses against clean-label poisoning