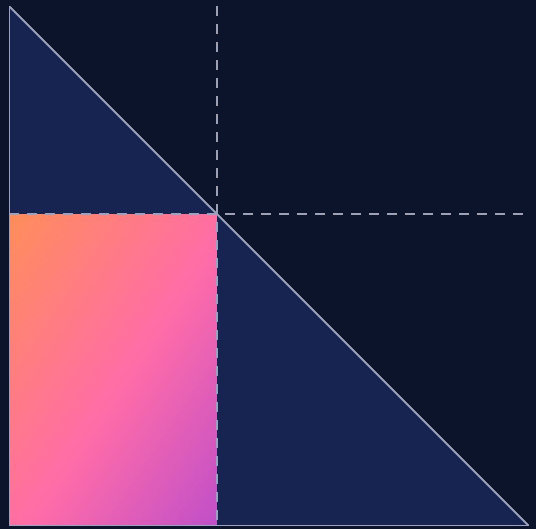
teads

NEURIPS 2026

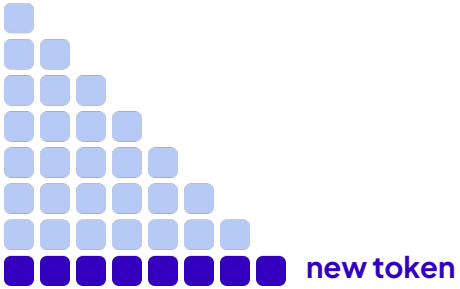
A Spectral Rejection Test for Value-Linear Attention Surrogates



Blaž Škrlj · Ivan Can Arisoy · Solal Vernier

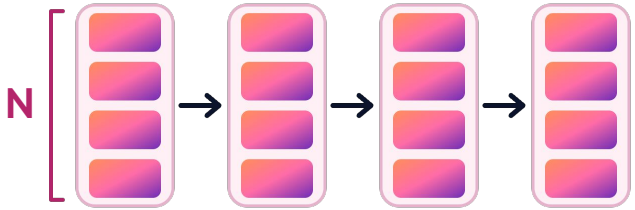
How much memory does a recurrent layer need?

Attention



Looks back at every earlier token, so cost grows with length

Recurrent layer

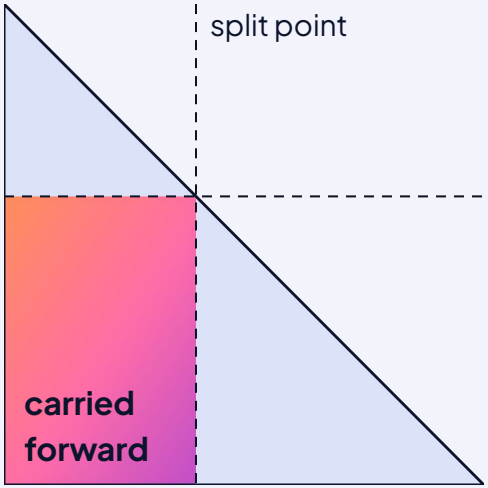


Carries a memory of fixed size N, so cost per token stays constant

We compute a lower bound on N from the attention pattern, before any training.

teads

A memory lower bound from the attention pattern



One head, one prompt

Split the text anywhere. Everything carried forward passes through a memory of size N, which keeps at most N independent directions.

$$\|A - M_N\| \geq \sigma_{N+1}(A_{split})$$

Left: worst-case error over all unit-length value vectors (spectral norm)

Right: (N+1)-th singular value of the pink block, the first direction a memory of size N cannot keep

Applies to value-linear recurrent layers (linear attention, state-space models), not to arbitrary recurrent architectures.

No training needed. Minutes, even at 8B.

teads

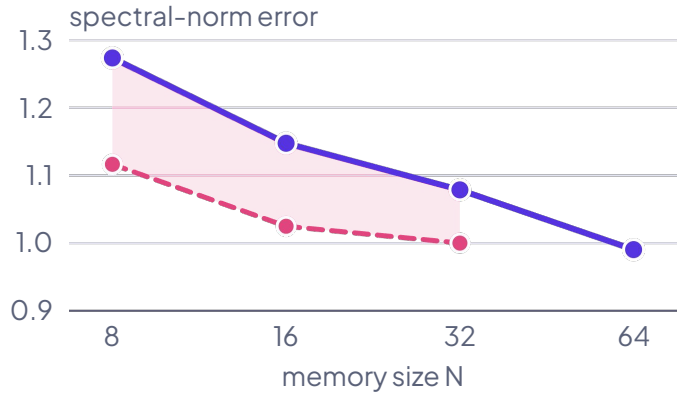
Tight on retrieval, loose on smooth attention

Trained replacement Our lower bound

Synthetic attention, rows sum to 1 · error not normalized

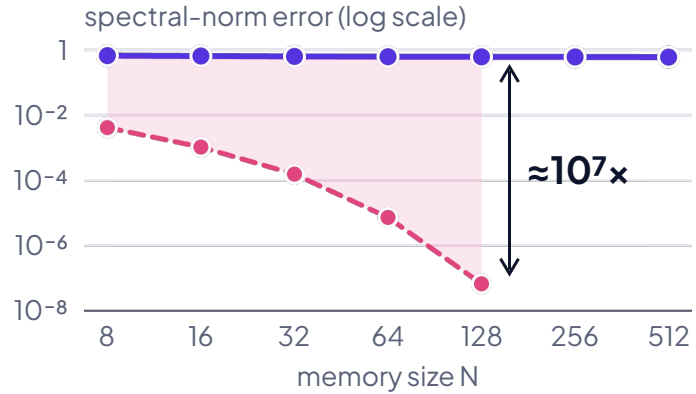
Retrieval-like attention

Training lands close to the bound



Smooth attention

Training stays up to 10^7× above it



The bound is informative for retrieval-like attention but can be extremely loose for smooth attention.

teads

Real attention heads exhibit low spectral memory lower bounds

Necessary lower-bound estimates: 95th-percentile spectral rank at 10% error, over heads and prompts



Memory below the bar value minus one is ruled out; reaching it guarantees nothing · Random: causal attention of the same shape

teads

A check to run before training

1

Measure

Each head's memory lower bound on typical prompts

2

Compare

With the memory size N you plan to use

3

Decide

Lower bound exceeds N on over 5% of prompts? Reject this memory size.

What it does not tell you

Whether training will succeed

How to split memory across heads

Anything beyond linear-attention and state-space layers

teads

Takeaways

Rule out insufficient memory sizes before training—not guarantee that sufficient ones will train successfully.

- The attention pattern gives a lower bound on memory size for value-linear recurrent layers.
- Tight on retrieval-like heads, loose on smooth ones.
- Small memory sizes cannot be ruled out for most pretrained heads.

Code for all experiments

teads