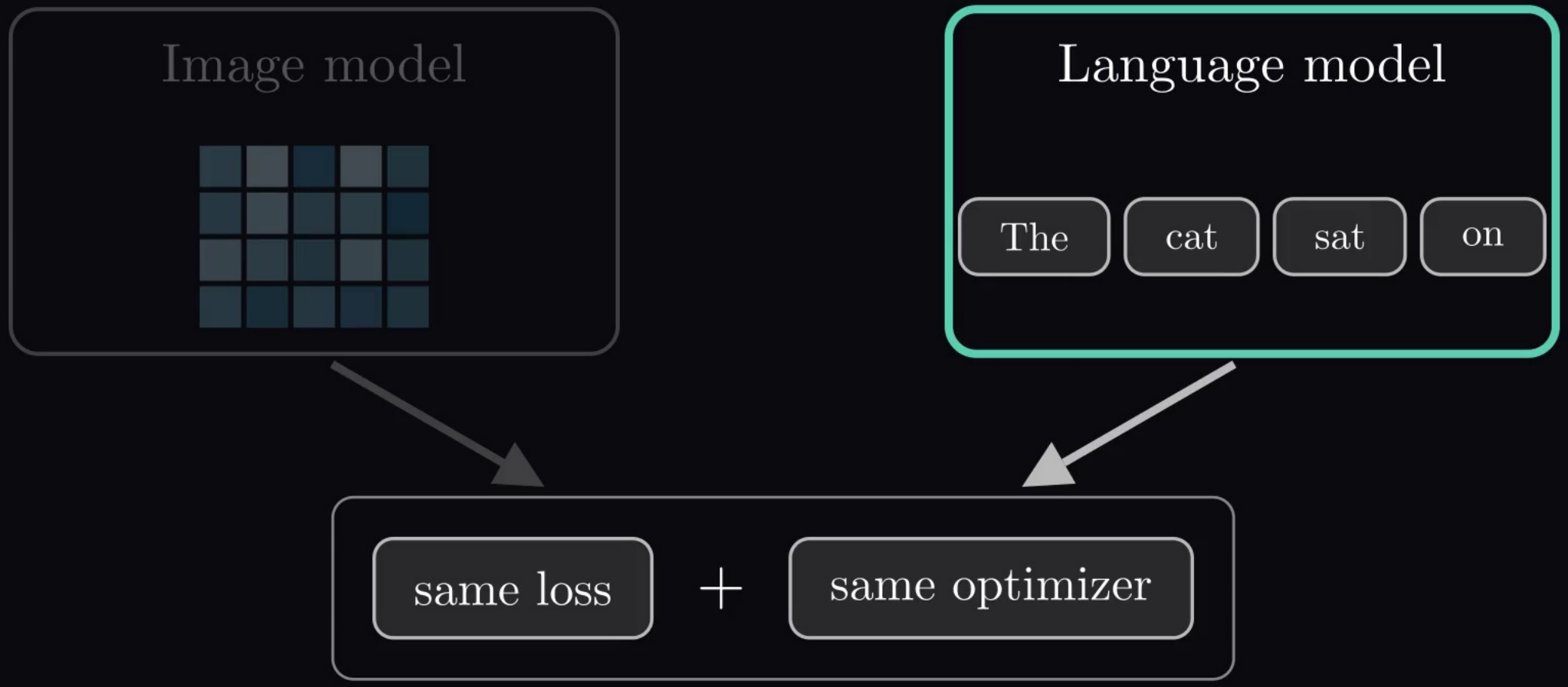

Image model
Language model
The
cat
sat
on
same loss
+
same optimizer

Can we use it?

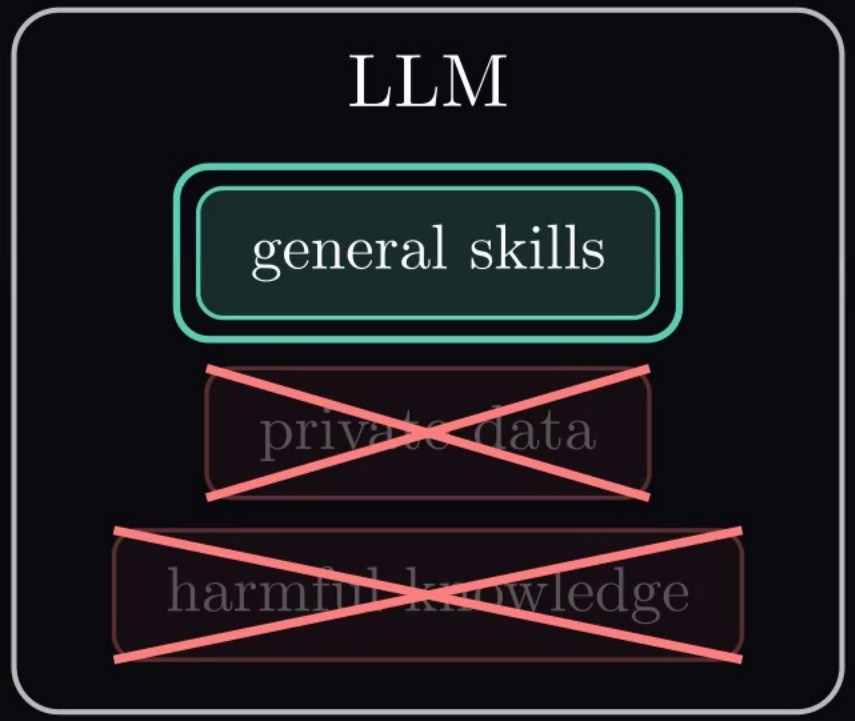
Spectral Unlearning

Transformer Structure-Preserving Updates for Language Models

Sung Il Choi Junhao Cai Dohun Kim Changhee Joo

Korea University · NeurIPS 2026

Unlearning has two jobs



2. Keep

1. Forget

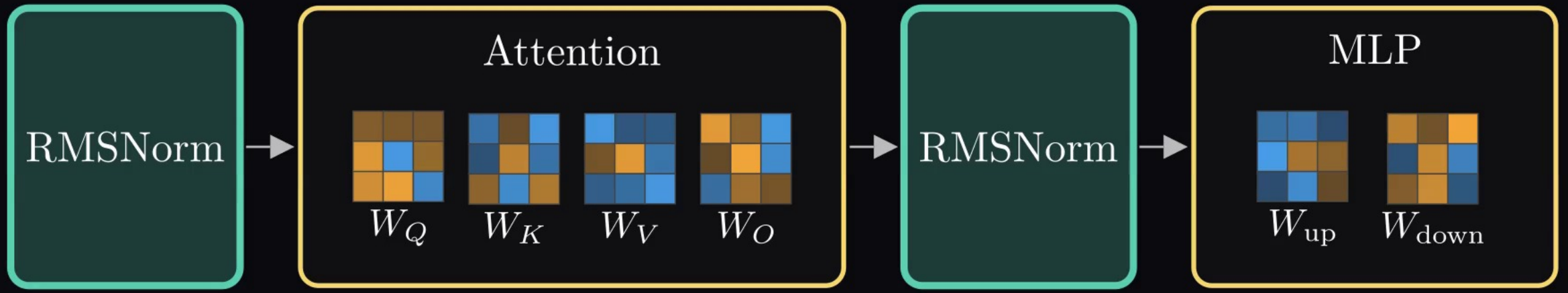
Recent work designs the loss

$$\min_{w} \mathcal{L} = \mathcal{L}_{\text{forget}} + \alpha \mathcal{L}_{\text{retain}}$$

How do we get there?



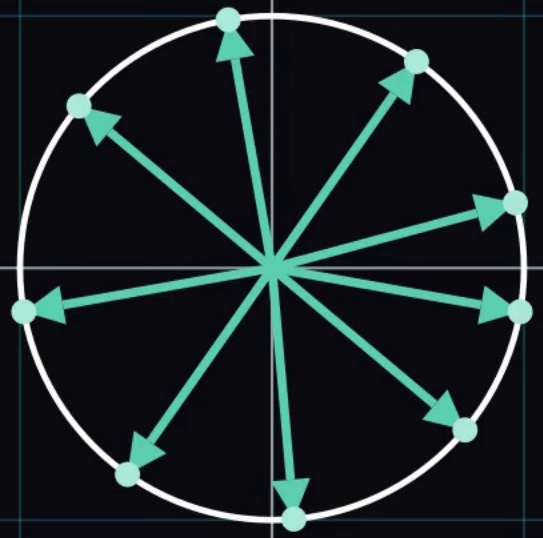
Before each matrix: RMSNorm



weight matrices

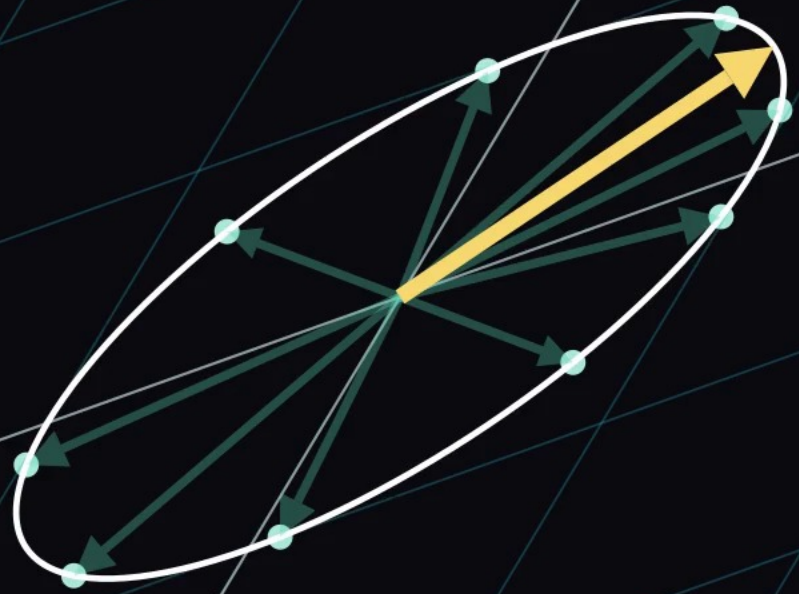
Inputs = points on a circle

$$W = \begin{bmatrix} 1.4 & 0.6 \\ 0.5 & 1.0 \end{bmatrix}$$



It tells how strong the matrix is

$$W = \begin{bmatrix} 1.4 & 0.6 \\ 0.5 & 1.0 \end{bmatrix}$$

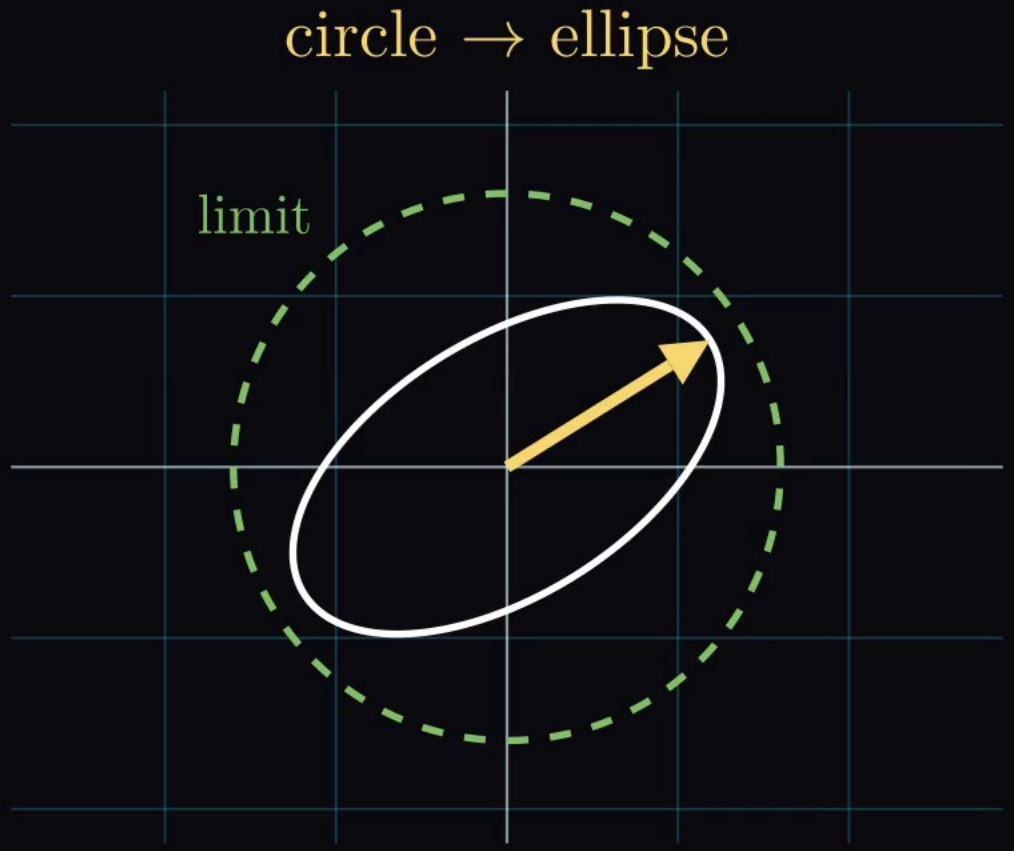


$$\|W\|_2 = \sigma_{\max}$$

A good update: no direction stretched too much

$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$
$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$

ΔW



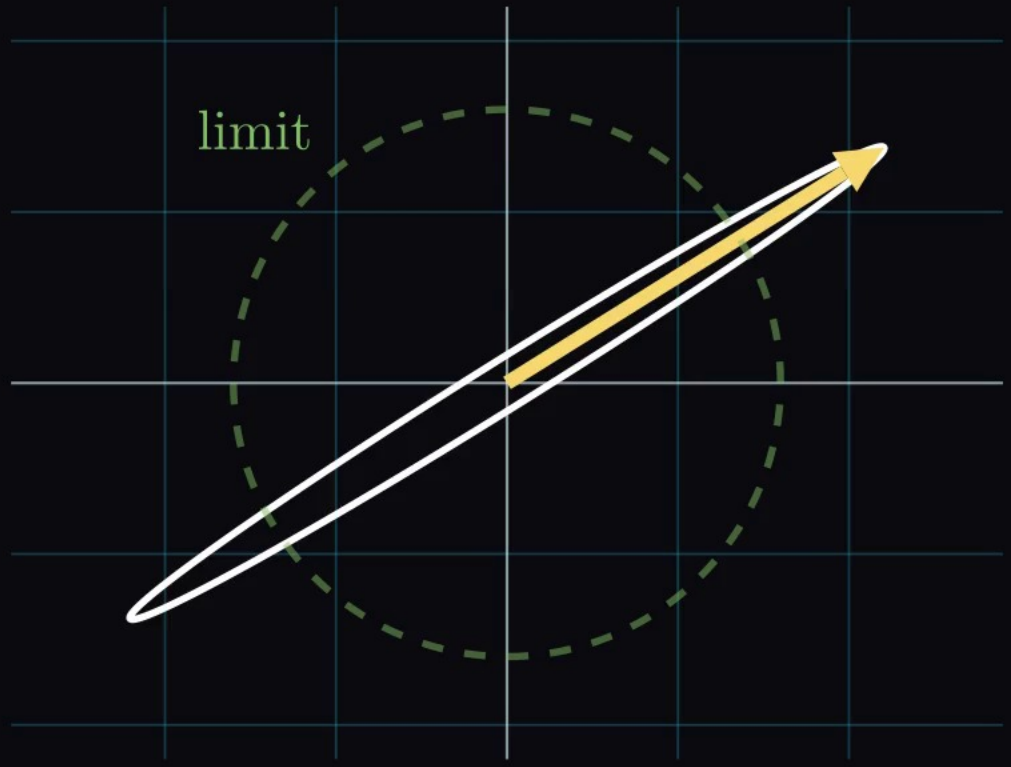
The layer's output changes a lot

$|\Delta W_{ij}| = \varepsilon$ each number moves a little

$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$
$+\varepsilon$	$-\varepsilon$	$+\varepsilon$	$+\varepsilon$	$-\varepsilon$	$+\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$
$-\varepsilon$	$+\varepsilon$	$-\varepsilon$	$-\varepsilon$	$+\varepsilon$	$-\varepsilon$

ΔW

long and thin



one direction changes far too much

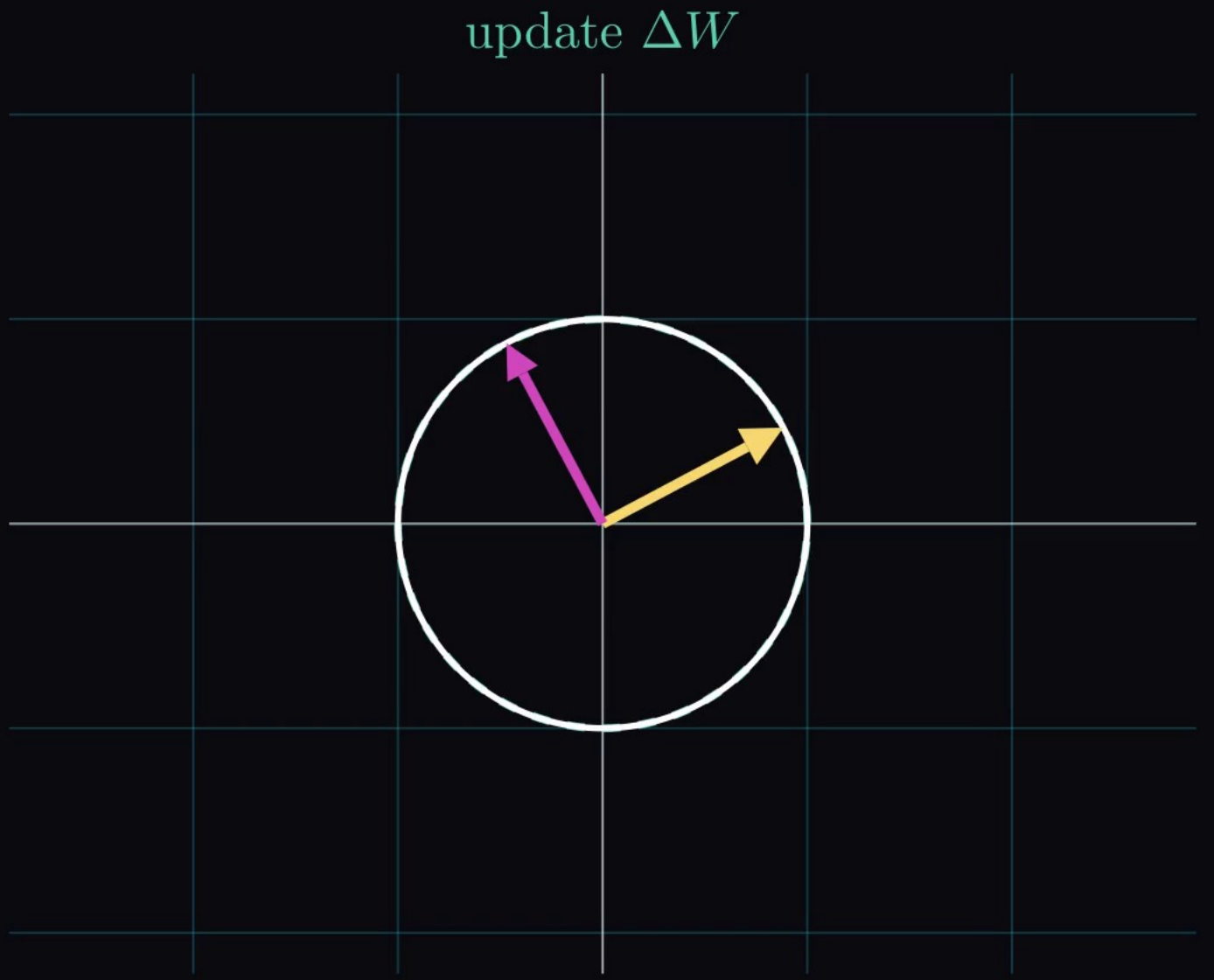
The whole matrix moves together

$$\|\Delta W\|_2 \leq \eta$$

$$\min_{\Delta W} \langle G,\ \Delta W \rangle$$

$$G = U \begin{bmatrix} 1.8 & 0.0 \\ 0.0 & 0.4 \end{bmatrix} V^\top$$

$$\boxed{\Delta W = -\eta\, U V^\top}$$



Three optimizers approximate it

Muon

Newton–Schulz iteration

KLShampoo

$L^{-1/4} G R^{-1/4}$

SOAP

Adam in Shampoo's eigenbasis

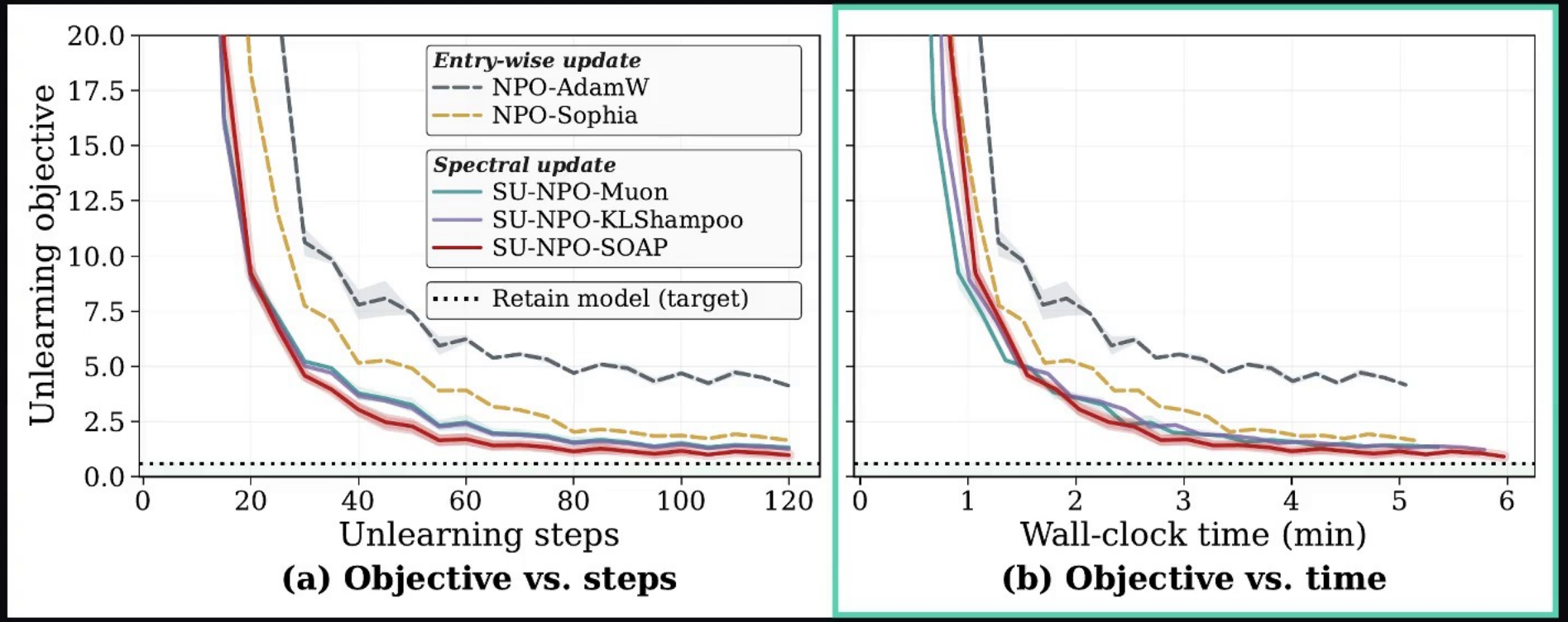
$$\Delta W \approx -\eta\, U V^{\top}$$



close to a circle

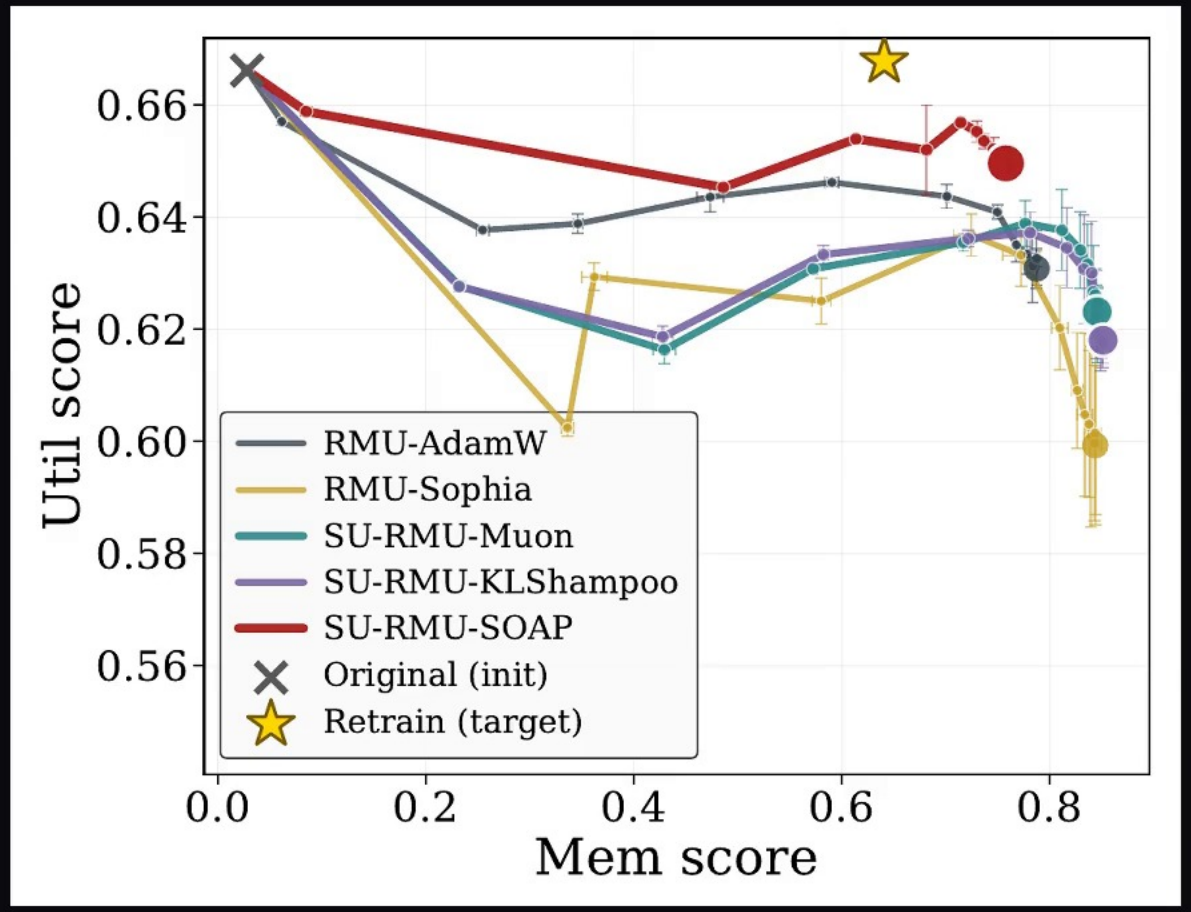
The loss stays the same: GD, NPO, RMU

1. Speed



- unlearning loss over steps (NPO)
- AdamW, Sophia
- ours (spectral)
- target (retrained)
- Ours reach the target first
- also in wall-clock time

2. Trade-off



right: more forgetting

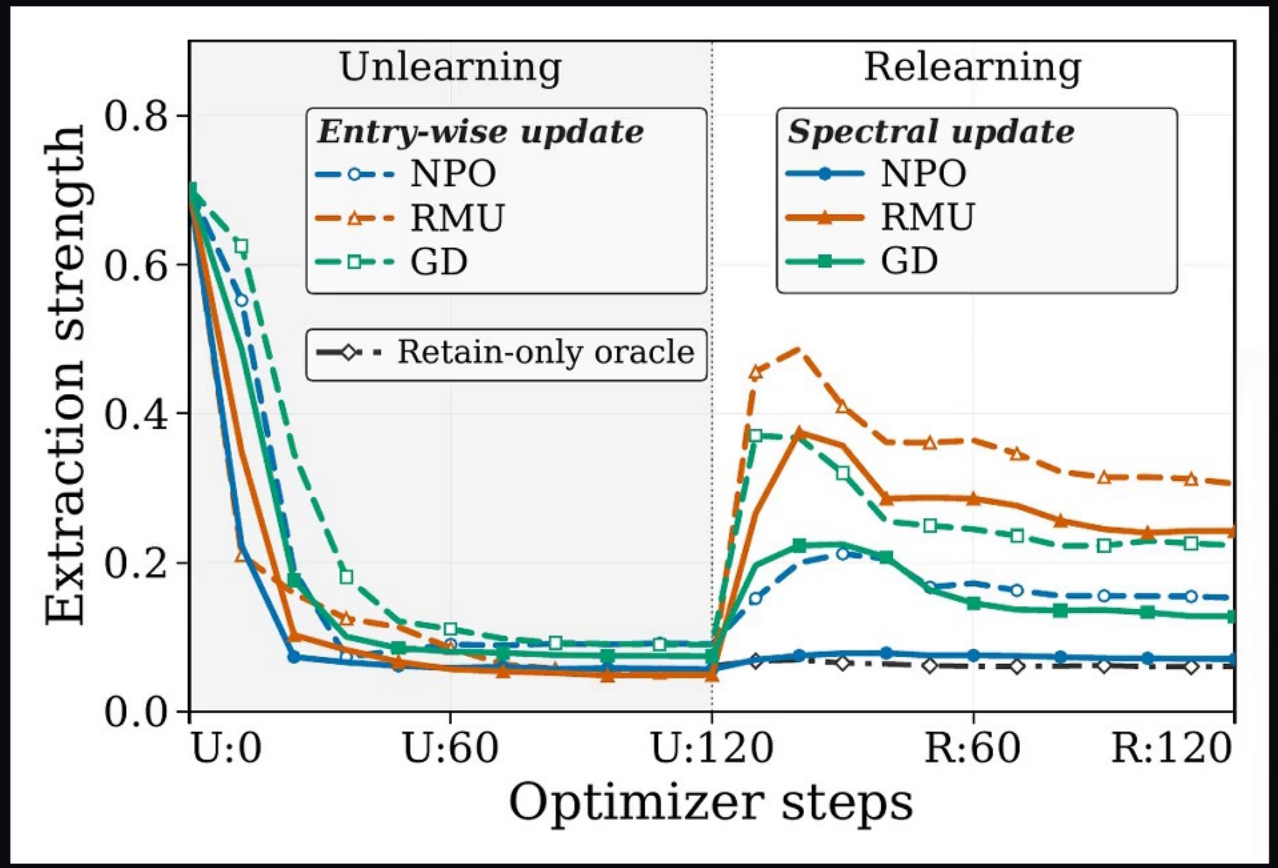
down: less utility

the retrained model

so the star is our goal

SOAP ends closest to the star

3. Relearning attacks



- attacker fine-tunes on a few forget examples
- spectral (solid): less comes back
- Why?
- forget examples moved farther
- harder to bring back

Takeaway

An LLM is a Transformer.

Want to erase something from it?

Respect the Transformer's structure.

Keep your loss. Update each matrix as a matrix.

$$\Delta W = -\eta\, U V^{\top}$$

Thank you

Spectral Unlearning

Transformer Structure-Preserving Updates for Language Models

Sung Il Choi Junhao Cai Dohun Kim Changhee Joo

github.com/unlearn-ml/spectral-unlearning