Spectral Unlearning

Transformer Structure-Preserving Updates for Language Models

Paper: Sung Il Choi, Junhao Cai, Dohun Kim, and Changhee Joo

Reading seminar | arXiv:2611.11902v1
12 November 2026

Update geometry is a second design axis for unlearning

Takeaway

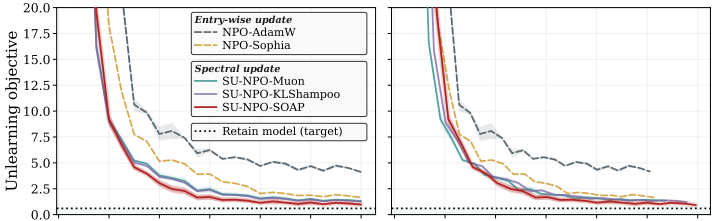
Keep the unlearning objective; change how matrix gradients become updates.

- Ideal spectral direction: $-UV^\top$ for $G = U\Sigma V^\top$.
- Practical realizations: Muon, KLShampoo, and SOAP.
- SOAP often improves balanced forgetting; gains have compute costs and exceptions.

arXiv v1: 9 November 2026.
Accepted at NeurIPS 2026, per authors' lab and arXiv.
Official code. Status checked 12 November 2026.

Paper, Abstract and Secs. 1, 3–5, PDF pp. 1–10. Acceptance is distinguished from formal publication.

The same NPO objective follows different optimization paths



(a) Objective vs. steps **(b) Objective vs. time**

- Left: unlearning steps. Right: wall-clock time. Lower objective is preferred.
- Solid: spectral variants; dashed: entry-wise AdamW and Sophia.
- Loss convergence motivates the method; forgetting quality needs separate metrics.

Paper, Fig. 1, PDF p. 2. Llama-3.2-1B on TOFU; NPO unlearning.

GD and NPO generate different forgetting signals

Gradient Difference (GD)

$$\mathcal{L}_{\mathrm{GD}} = \mathbb{E}_u[\log \pi_w(y \mid x)] + \alpha \mathbb{E}_r[-\log \pi_w(y \mid x)]$$

Negative Preference Optimization (NPO): forget term

$$\mathcal{L}_{u,\mathrm{NPO}} = \frac{2}{\beta}\mathbb{E}_u \log\left[1 + \left(\frac{\pi_w(y \mid x)}{\pi_{w_0}(y \mid x)}\right)^{\beta}\right]$$

- GD's forget term is unbounded below; NPO's is lower-bounded by zero.
- Experiments add retain NLL to NPO, with $\beta = 0.1$ and unit loss weights.

Paper, Eqs. (1)–(3), PDF p. 3; App. C.3, PDF p. 20. u/r: forget/retain data; w_0: frozen reference.

RMU changes representations while anchoring retained examples

$$\mathcal{L}_{\mathrm{RMU}} = \mathbb{E}_u \| h_w^{(\ell)}(x) - cu \|_2^2 + \alpha \mathbb{E}_r \| h_w^{(\ell)}(x) - h_{w_0}^{(\ell)}(x) \|_2^2.$$

- u: fixed random unit direction; c: steering magnitude.
- Forget representations move to a random target; retained ones stay near the reference.
- TOFU uses block 7 and $c = 2$, with all transformer parameters trainable.
- The spectral update is applied to this loss without replacing it.

Paper, Eq. (4), PDF p. 3; App. C.3, PDF p. 20. TOFU setup differs from layer-restricted WMDP RMU.

A norm decides which change counts as a large step

$$\Delta w^* \in \arg\min_{\Delta w} \left[\langle g, \Delta w \rangle + \frac{\lambda}{2} \|\Delta w\|^2 \right]$$

$$\Delta w^* = -\frac{\|g\|_*}{\lambda}\, t^*(g), \quad t^*(g) \in \arg\max_{\|t\|=1} \langle g, t \rangle$$

- Direction: maximize first-order loss change per unit step.
- Magnitude: dual norm divided by local curvature scale λ.
- Vector ℓ_∞ geometry gives $\operatorname{sign}(g)$.
- The AdamW/sign connection removes moving averages and decay; it is an idealization.

Paper, Eqs. (5)–(8), PDF pp. 3–4; App. A.1, PDF p. 14.

The spectral-norm budget gives the polar-factor direction

$$G = U\Sigma V^\top, \qquad \|W\|_2 = \sup_{\|x\|_2=1} \|Wx\|_2$$

$$\max_{\|T\|_2\le 1} \langle G, T\rangle = \sum_i \sigma_i, \qquad T^* = UV^\top$$

$$\boxed{\Delta W^* = -\frac{\operatorname{tr}\Sigma}{\lambda} UV^\top} \quad \Longrightarrow \quad W_{t+1} = W_t - \eta UV^\top$$

- Keep singular directions and flatten their update amplitudes.
- Scheduled learning rate η replaces the exact local-model magnitude.
- At rank deficiency, the optimal direction need not be unique.

Paper, Eqs. (9)–(12), PDF p. 4; App. A.3, PDF p. 15.

Spectral unlearning reshapes gradients, not the weight spectrum

Teaching example: a diagonal gradient

$$G = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}, \qquad UV^\top = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

- Euclidean gradient descent: update amplitude ratio $10 : 1$.
- Spectral direction: ratio $1 : 1$ on these singular modes.
- Weaker modes receive more relative weight.
- No low-rank truncation or permanent orthogonality constraint is imposed on W.

Teaching illustration of Paper Eq. (12), PDF p. 4; not an experimental result.

RMS geometry connects normalized features to spectral norms

$$\|x\|_{\mathrm{RMS}} = \frac{\|x\|_2}{\sqrt{d}}, \qquad \boxed{\|W\|_{\mathrm{RMS}\to\mathrm{RMS}} = \sqrt{\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}}\|W\|_2}$$

- RMSNorm in pre-norm blocks motivates measuring feature scale this way.
- A selected network norm takes the maximum of the layerwise RMS-to-RMS norms.
- Its local steepest-descent direction is $-U_l V_l^\top$, with dimension-dependent step sizes.
- This is a geometric derivation; it is not a certified-unlearning guarantee.

Paper, Lemma 1/Theorem 2, Eqs. (16)–(17), PDF pp. 6–7. Multi-layer derivation in backup.

Three matrix optimizers realize related, distinct updates

Muon	Five Newton–Schulz iterations approximate the polar factor of a momentum matrix.
KLShampoo	Running row/column second moments; update proportional to $-L_t^{-1/4} G_t R_t^{-1/4}$.
SOAP	Shampoo-style eigenbases, with Adam-style adaptive updates in the rotated coordinates.

- Momentum, EMA, damping, and basis refreshes distinguish these from exact instantaneous SVD.
- Attention/MLP matrices use matrix updates; embeddings use AdamW.

Paper, Secs. 3.2–3.3, PDF pp. 5–6; App. C.4, PDF pp. 20–21.

One-sided polar identities have specific conditions

$$G \text{ tall, full column rank}: \quad G(G^\top G)^{-1/2} = UV^\top,$$
$$G \text{ wide, full row rank}: \quad (GG^\top)^{-1/2} G = UV^\top.$$

- MLP up/gate: precondition on the residual input side; down: on the residual output side.
- SOAP uses `max_precond_dim`$= d$; larger intermediate axes are skipped.
- Exact identities use instantaneous inverse square roots and full rank.
- EMA statistics and SOAP's inner Adam remain practical approximations.

Paper, App. B, Table 4 and Eqs. (46)–(48), PDF pp. 17–19. Not every projection directly receives RMSNorm output.

The experiments compare objectives and optimizers separately

TOFU	Llama-3.2-1B/3B; forget 1%, 5%, 10%
RWKU	Llama-3-8B-Instruct; real-person knowledge
WMDP	Zephyr-7B-β; Bio/Cyber, with MMLU retention
Main combinations	3 losses $\times$ 5 optimizers = 15 methods
Training	10 epochs; first epoch warmup; LR sweep per cell
Seeds	3 trajectories; initial checkpoint and data split fixed

- Original: full-data fine-tuned checkpoint. Retrain: retain-only fine-tuning reference.
- The oracle concerns this fine-tuning task, not deletion from all pretraining data.

Paper, Sec. 5.1 and App. C, PDF pp. 7, 19–22. Extra baselines use their canonical training settings.

Balanced evaluation needs forgetting, privacy, and utility

$$\mathrm{Agg} = \mathrm{HM}(\mathrm{Mem}, \mathrm{Priv}, \mathrm{Util}), \qquad \mathrm{HM}(x_1, \ldots, x_n) = \frac{n}{\sum_i 1/\max(x_i, 10^{-12})}$$

- **Mem** ↑: HM of inverted extraction, token agreement, forget probability, and a symmetric truth ratio.
- **Priv** ↑: similarity of four MIA AUCs to the retain-only reference.
- **Util** ↑: ten retention/knowledge/fluency sub-metrics.
- Higher Agg can reflect a better balance while one component declines.

Paper, App. D, Eqs. (49)–(56), PDF pp. 22–23. Privacy is an empirical score, not a formal guarantee.

SOAP improves the main TOFU aggregate comparisons

Setting	GD	NPO	RMU
1B / 10%	.5704 → .6043	.6876 → .6926	.6888 → .6982
1B / 5%	.5653 → .6075	.6599 → .6711	.6801 → .6862
3B / 10%	.5839 → .6091	.7243 → .7262	.7038 → .7165
3B / 5%	.5737 → .6455	.7062 → .7211	.7317 → .7360

- Each cell: best entry-wise mean → same-loss SOAP mean. Agg ↑.
- SOAP wins these 12 comparisons; margins vary substantially.
- Other spectral variants can lose: 3B/10% GD-Muon .5016 vs. Sophia .5839.

Paper, Table 1, PDF p. 7; precise values from Table 12, PDF p. 36. No significance claim from small mean gaps.

A similar aggregate score does not mean the oracle is matched

3B / 5%, RMU	Agg	Mem	Priv	Util
AdamW	.7151	.8440	.6856	.6444
Sophia	.7317	.6920	.9010	.6474
SOAP	.7360	.7160	.8468	.6673
Retrain	.7402	.6362	.9999	.6753

- Both SOAP and Retrain round to .74 in the main table.
- Precise aggregate gap: .0042; privacy behavior still differs visibly.
- SOAP improves balance vs. AdamW, even though its Mem score falls.

Paper, Tables 12/16, PDF pp. 36, 38. All scores higher is better; comparisons use reported means.

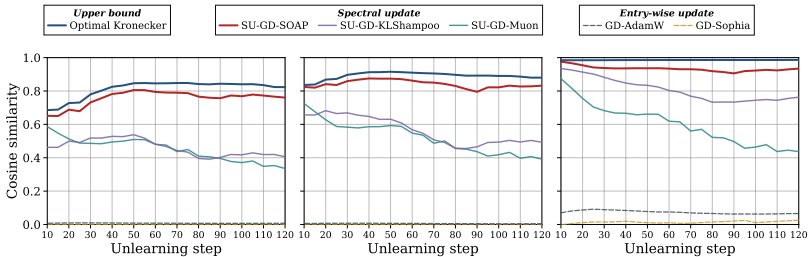
The 1% split breaks a uniform optimizer ranking

Setting	SOAP Agg	Comparator Agg
1B / NPO / 1%	.6445 ± .0128	Sophia: .6486 ± .0042
3B / RMU / 1%	.7139 ± .0033	AdamW: .7169 ± .0006
3B / GD / 1%	.6217 ± .0190	Muon: .6392 ± .0040

- Across all 18 loss/model/split cells, SOAP beats the best entry-wise mean in 16.
- The small split gives few updates: about 13 steps at 1B, 25 at 3B.
- A strong overall optimizer is not the winner in every configuration.

Paper, Table 12, PDF p. 36; App. C.2, PDF p. 19. The 16/18 count is calculated from table means.

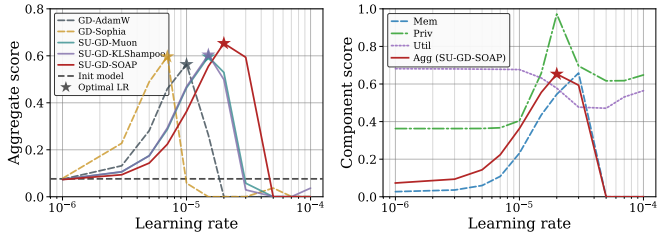
Matrix statistics capture gradient correlations missed by diagonal ones



- Output, MLP up, MLP down projections; cosine similarity to measured covariance.
- SOAP tracks the optimal Kronecker reference most closely here.
- This measures covariance alignment, not exact update alignment with $UV^\top$.

Paper, Fig. 2, PDF p. 5; depth/loss extensions in Figs. 7–8, PDF pp. 24–25.

A broader useful learning-rate range still has a failure threshold

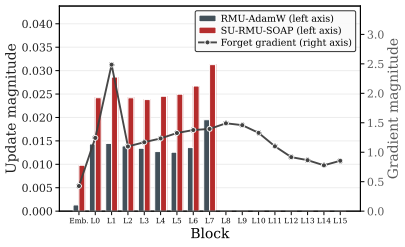


(a) Aggregate score vs. LR **(b) Component score vs. LR**

- GD, 3B, forget 5%; left: Agg, right: Mem/Priv/Util decomposition.
- SOAP's peak is higher; its effective LR region is wider.
- At extreme LR, truth-ratio and retention failures can collapse Agg.

Paper, Fig. 4, PDF p. 8; normalized sensitivity/JSD and sub-metrics in Figs. 9–10, PDF pp. 26–27.

RMU updates become more selective across blocks



(a) Block-wise update magnitude **(b) Per-block activation drift**

- Left: mean update magnitude by block; right: hidden-state drift.
- SOAP places more update mass early/mid-stack, especially at L1.
- Forget drift increases more than retain drift, supporting selective change.

Paper, Fig. 5, PDF p. 9. The reference line is called forget gradient in the figure and gradient disagreement in the text.

Better optimization carries substantial time and memory costs

3B / TOFU 10%	Time (min)	Peak (GiB)	Agg
GD-AdamW	8.79	14.64	.3813
GD-SOAP	18.08	36.76	.6091
NPO-AdamW	15.69	20.55	.7243
NPO-SOAP	24.32	42.75	.7262

- GD-SOAP: 2.06× total time, 2.51× peak memory at equal epochs.
- NPO: higher cost buys a much smaller score gain in this setting.
- Matched-time trajectories can favor SOAP; per-step computation is not cheaper.

Paper, Table 14, PDF p. 37; Agg from Table 12, PDF p. 36. Ratios calculated from reported means.

RWKU gains can come from retaining more neighboring knowledge

Loss	Update	Forget ↓	Neighbor ↑	MMLU ↑	Gain ↑
GD	AdamW	.1127	.5657	.6392	.1117
GD	SOAP	.1460	.6539	.6376	.1295
NPO	AdamW	.3250	.7520	.6292	.0997
NPO	SOAP	.2034	.7334	.6281	.1336

- Llama-3-8B-Instruct; real-person knowledge removal.
- GD-SOAP has higher joint Gain, although remaining target knowledge rises.
- Gain averages signed changes from Original across all three metrics.

Paper, Table 3 and Sec. 5.5, PDF pp. 9–10.

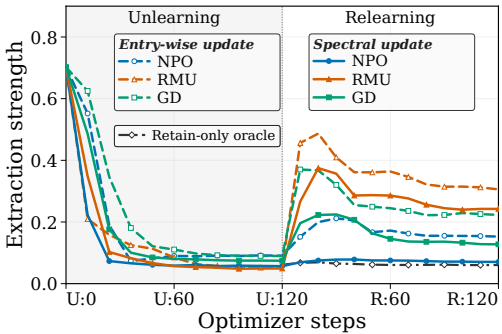
WMDP improves the forgetting–retention balance

Loss	Update	Bio ↓	Cyber ↓	MMLU ↑	Gain ↑
GD	AdamW	.2916	.2574	.4527	.1271
GD	SOAP	.2789	.2473	.5084	.1533
RMU	AdamW	.3079	.2723	.5387	.1454
RMU	SOAP	.3028	.2581	.5476	.1548

- Zephyr-7B-β; four-way multiple-choice Bio/Cyber tests.
- Chance-level reference is .25; it does not test every form of hazardous capability.
- SOAP improves joint Gain for GD, NPO, and RMU.

Paper, Table 3, PDF p. 10; three-seed standard deviations in Table 13, PDF p. 36.

Relearning recovers less content after spectral updates



- Forget 10% of all data; relearn on 10% of that set (1% of all data).
- Solid SOAP curves end below dashed AdamW curves for each loss.
- NPO stays closest to the retain-only oracle.
- Recovery still occurs; this is empirical resistance under a limited attack.

Paper, Fig. 6 and Sec. 5.6, PDF p. 10. Extraction strength is lower-is-better.

The contribution is useful geometry, with concrete boundaries

1. **Mechanism:** flatten gradient singular amplitudes while retaining matrix directions.
2. **Evidence:** improved multi-axis results, useful LR ranges, and reduced relearning recovery.
3. **Conditions:** practical optimizers approximate ideal directions; small splits have counterexamples.
4. **Trade-off:** extra time/memory, with objective-dependent gains.

Open questions: distinguish general optimization gains from unlearning-specific gains, and compare alternative layerwise norms.

Paper, Sec. 6 and App. H, PDF pp. 10, 43; numerical boundaries from Appendix G.

References and official implementation

- Choi, Cai, Kim, and Joo.
  Spectral Unlearning: Transformer Structure-Preserving Updates for Language Models.
  arXiv:2611.11902v1, 9 November 2026.
- Accepted at NeurIPS 2026, per authors' laboratory and arXiv comments.
- Official repository: implementations, configurations, and experiment scripts.

Figures are reused from the authors' arXiv v1 source. Calculated counts and ratios are labeled. This is a reading seminar, not an author presentation.

Publication/code checked 12 November 2026. Formal proceedings publication was not independently established.

Backup: the layerwise maximum norm couples step magnitudes

$$\|\Delta W\|_{\mathrm{net}} = \max_l \|\Delta W_l\|_{\mathrm{RMS}\to\mathrm{RMS}}, \qquad a_l = \sqrt{d_{\mathrm{out,l}}/d_{\mathrm{in,l}}}$$

$$s_l = a_l \operatorname{tr}\Sigma_l, \qquad c^* = \frac{1}{\lambda}\sum_k s_k, \qquad \Delta W_l^* = -c^* a_l U_l V_l^\top$$

- Fix the maximum step budget: every nonzero-gradient layer can use it without increasing the penalty.
- All such layers take the same RMS-to-RMS magnitude c^*.
- Square matrices have $a_l = 1$; rectangular matrices retain dimensional factors.
- The empirical advantage of this max aggregation over layerwise sums is open.

Paper, Theorem 2, Eq. (17), PDF pp. 6–7; proof in App. A.4, PDF pp. 16–17; App. H, PDF p. 43.

Backup: the memorization score penalizes overconfident wrong answers

$$\mathrm{Mem} = \mathrm{HM}(1-\mathrm{ES}, 1-\mathrm{EM}, 1-\mathrm{F.Prob}, \mathrm{F.TR})$$

$$R_z = \frac{p_z^{\mathrm{pert}}}{p_z^{\mathrm{para}} + 10^{-10}}, \qquad \mathrm{F.TR} = \mathbb{E}_z \min(R_z, (R_z + 10^{-10})^{-1})$$

- ES: contiguous suffix extraction; EM: token-level greedy agreement.
- F.Prob: length-normalized likelihood on original forget prompts.
- F.TR peaks when paraphrased correct and perturbed wrong answers have comparable likelihood.
- Paraphrase probability is reported separately, but excluded from this Mem aggregate.

Paper, App. D.1, Eqs. (51)–(53), PDF pp. 22–23. EM here is not whole-answer exact match.

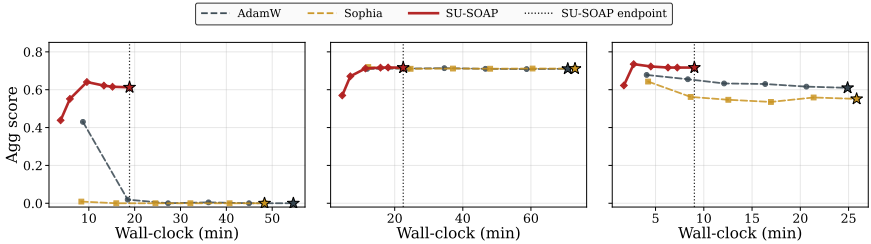
Backup: privacy targets the retain oracle, not minimum AUC

$$s_k = 1 - |\mathrm{AUC}_k - \mathrm{AUC}_k^{\mathrm{retain}}|, \qquad \mathrm{Priv} = \mathrm{HM}(s_{\mathrm{LOSS}}, s_{\mathrm{ZLib}}, s_{\mathrm{MinK}}, s_{\mathrm{MinK++}})$$

- LOSS, compression-adjusted loss, bottom-40% token scores, and standardized bottom-token scores.
- AUC below .5 does not automatically mean better privacy under this protocol.
- Util gives equal weight to ten sub-metrics: three measures on each of retain/real-author/world-facts, plus fluency.
- A non-gibberish classifier may reward fluent repetition; examine outputs as well as scores.

Paper, App. D.2–D.3, Eqs. (54)–(56), PDF p. 23; high-LR failure discussed in Sec. 5.3, PDF p. 9.

Backup: matched-time curves reveal objective-dependent behavior

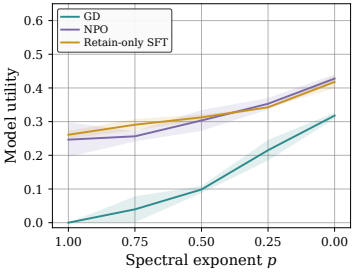


(g) 3B/F10 — GD **(h) 3B/F10 — NPO** **(i) 3B/F10 — RMU**

- Fig. 13 panels (g)–(i): 3B, forget 10%; GD, NPO, RMU.
- Baselines run past SOAP's endpoint; NPO can approach a similar plateau.
- The full figure covers 5%/10%, not 1%; NPO at 3B/5% even shows a slightly higher late AdamW score.

Paper, Fig. 13, PDF p. 29. Shared legend and selected original panels composed without changing data.

Backup: flattening also benefits retain-only fine-tuning

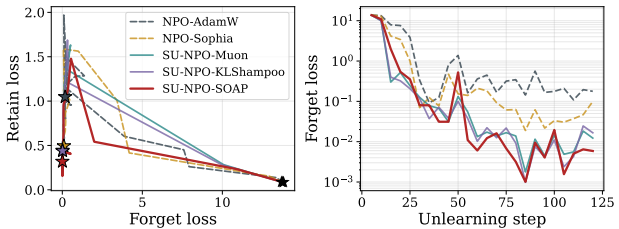


$$\Delta W_p = -\eta \frac{U\Sigma^p V^\top}{\mathrm{RMS}(U\Sigma^p V^\top)}$$

- $p = 1$: momentum direction; $p = 0$: flattened spectrum.
- Lower p improves mean utility for GD, NPO, and retain SFT.
- General optimization gains may contribute to the unlearning improvement.

Paper, App. E.7, Fig. 14, PDF p. 30. Shading: seed variation after averaging over learning rates.

Backup: parameter trajectories support the optimization account



(a) Forget–retain trajectory **(b) Forget loss vs. step**

- NPO, 1B, forget 10%: joint objective space and forget objective over steps.
- SOAP reaches low forget-objective values with competitive retain loss.
- Forget loss is the NPO objective. Fig. 12 adds PCA paths; these visuals do not isolate direction from all optimizer factors.

Paper, App. E.5, Fig. 11 panels (a)–(b), PDF p. 28; Fig. 12 provides PCA parameter paths.

Backup: qualitative examples expose failure modes hidden by scores

Forget side	GD can reproduce the target; NPO can produce fluent unrelated hallucinations; RMU can produce degenerate sequences.
Retain side	NPO/RMU with SOAP often preserve details better in the selected examples.
Persistent failure	GD variants still answer some general-knowledge questions incorrectly, including the SOAP variant.

- Changing an answer is not enough to establish useful unlearning.
- Five selected examples illustrate mechanisms; they do not estimate failure frequencies.

Paper, App. F, Tables 7–11, PDF pp. 31–34. Generations are truncated in the source tables.