Retain-Neutral Surrogates for Min–Max Unlearning

Junhao Cai
Korea University
junhochae@korea.ac.kr

Dohun Kim
Korea University
dohunkim@korea.ac.kr

Dowon Kim
Korea University
dowonkim@korea.ac.kr

Sung Il Choi
Korea University
sungchoi@korea.ac.kr

Chengjun Jin
Korea University
chengjunjin2001@korea.ac.kr

Juhyun Park
Korea University
juhyunpark@korea.ac.kr

Changhee Joo*
Korea University
changhee@korea.ac.kr

Abstract

Machine unlearning aims to make a model forget designated training data while retaining performance on the rest. In min–max unlearning, an inner adversary perturbs the model to maximize the forget loss, and the model is updated to minimize the retain loss at the resulting surrogate point. When the forget and retain gradients are positively coupled, the standard perturbation also moves toward higher retain loss, and the retain step begins by recovering that loss. We propose **R**etain-**O**rthogonal **S**urrogate **U**nlearning (ROSU), which redefines the adversary to admit only perturbations orthogonal to the retain gradient. The one that raises the forget loss fastest has a closed form, and ROSU uses it to set the surrogate point and amplifies it into the forgetting step. Under positive coupling the ROSU surrogate avoids the retain cost that coupling induces in the standard surrogate, and the difference shrinks as the gradients approach orthogonality. ROSU has the smallest accuracy gap to retraining in every vision setting (CIFAR-10, CIFAR-100, Tiny-ImageNet) and is competitive with the baselines on language benchmarks (TOFU, WMDP). The redefinition helps most under high positive coupling. Code is available at `https://github.com/unlearn-ml/machine-unlearning`.

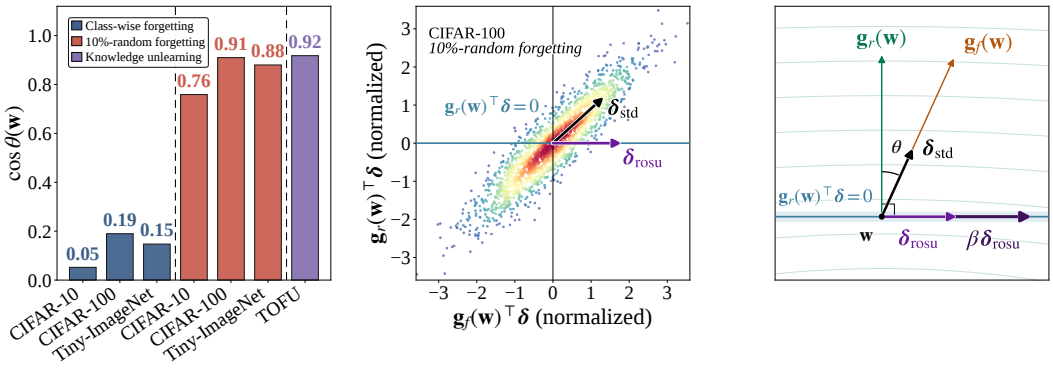
1 Introduction

Machine unlearning aims to remove the influence of a designated forget set from a trained model while preserving performance on the remaining retain set [1, 2]. Retraining without the forget data is the reference, but it is prohibitive at modern scale [3, 4], so approximate methods edit the trained model instead [5, 6]. The two goals conflict because the same parameters serve both and because the forget and retain sets are statistically dependent, which can drive unconstrained ascent away from the retraining reference [7]. Moving along the forget gradient is the most direct way to forget, but it works against retention whenever the forget and retain gradients point in similar directions. Figure 1a shows where this happens. The coupling between the two gradients is positive in every setting we measure. It is low in class-wise forgetting [8] but high in 10% random forgetting [9] and in knowledge unlearning on TOFU (Task of Fictitious Unlearning) [10, 11].

Min–max unlearning addresses this conflict by assigning the two goals to different levels rather than one combined loss. An inner adversary perturbs the model to raise the forget loss, and a retain

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Coupling diagnosis. (b) Retain-orthogonal surrogate. (c) Retain-neutral amplification.

Figure 1: **(a)** The coupling $\cos\theta(\mathbf{w})$ between the forget and retain gradients $\mathbf{g}_f(\mathbf{w})$ and $\mathbf{g}_r(\mathbf{w})$ is low in class-wise forgetting and high in random forgetting and TOFU. **(b)** In a highly coupled setting, $\boldsymbol{\delta}_{\mathrm{std}}$ moves toward higher forget and retain losses, while $\boldsymbol{\delta}_{\mathrm{rosu}}$ moves toward higher forget loss on the retain-neutral hyperplane $\mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta} = 0$. The arrows indicate directions. **(c)** The same direction $\boldsymbol{\delta}_{\mathrm{rosu}}$ sets the surrogate point, and its amplification $\beta\boldsymbol{\delta}_{\mathrm{rosu}}$ is the forgetting step, which stays retain-neutral.

step updates the model to minimize the retain loss at the resulting surrogate point. Since retention belongs to the retain step, the adversary should not disturb it. In the prior min–max method [8], the adversary is free to move the model in any direction to maximize the forget loss. Its perturbation follows the forget gradient, so under positive coupling it also moves toward higher retain loss, and the retain step begins by recovering that loss (Figure 1b). The retain cost comes from the directions the adversary is allowed to take. This raises a question. *Which perturbations should the inner adversary use?*

We admit only retain-neutral perturbations, which are orthogonal to the retain gradient at the current weights. Among them we take the one that raises the forget loss fastest. Unless the two gradients are parallel, this choice is unique and has a closed form. **R**etain-**O**rthogonal **S**urrogate **U**nlearning (ROSU) uses this one direction twice (Figure 1c). First, the direction sets the surrogate point, and the retain gradient there is transported back to the current weights as the retain step. Second, the amplified direction is the forgetting step, which is retain-neutral and leaves retention to the retain step. Among retain-neutral steps of the same length, it raises the forget loss fastest.

The contributions of our work are as follows:

- We identify how the retain cost that coupling induces enters min–max unlearning. When the forget and retain gradients are positively coupled, a step along the forget gradient moves toward higher retain loss. An unconstrained inner adversary takes such a step before the retain step, and this cost grows with the coupling and vanishes at orthogonality.
- We propose ROSU, a min–max unlearning method whose inner adversary admits only perturbations orthogonal to the retain gradient. These perturbations avoid the retain cost that coupling induces in an unconstrained inner adversary. Among them, the one that raises the forget loss fastest has a closed form. This perturbation sets the surrogate point of the retain step, and its amplification is a retain-neutral forgetting step.
- We evaluate ROSU on vision benchmarks (CIFAR-10, CIFAR-100, Tiny-ImageNet) and language benchmarks (TOFU, WMDP). It has the smallest accuracy gap to retraining in both class-wise and random forgetting on every vision benchmark and is competitive on both language benchmarks.

2 Related work

Exact unlearning restructures training so that a deletion request updates a few summations or retrains one affected sub-model [1, 2, 12]. Formal guarantees for approximate deletion were obtained for

linear and logistic models [3, 13]. For deep networks, retraining remains the reference, and its cost motivates approximate methods [4, 14]. Baselines fine-tune on the retained data or ascend the forget loss, and Fisher forgetting adds noise shaped by the retained data [5]. Later methods distill from the original model, unlearn on a sparsified model, or update only salient or forget-specific weights [9, 15–17]. Large language model (LLM) unlearning adapted gradient ascent and preference objectives and introduced benchmarks for fictitious and hazardous knowledge [10, 18–20]. Across these methods the forgetting objective varies, while any retain term is evaluated at the current weights.

Min–max methods evaluate the objective at a surrogate point instead. Sharpness-aware minimization perturbs the weights toward higher loss and descends with the gradient taken there, which favors neighborhoods of uniformly low loss [21, 22]. Unlearning-Aware Minimization (UAM) carries this structure to unlearning, perturbing the model to maximize the forget loss and minimizing the retain loss with the gradient at that point [8]. Projection methods instead orthogonalize the final update, from gradient surgery in multi-task learning [23] to null-space and retain-orthogonal updates in unlearning [24–28]. Both lines leave the inner adversary unconstrained or absent. ROSU redefines the adversary to admit only retain-neutral perturbations, so the surrogate point, the transported retain step, and the forgetting step come from one solution. Appendix A gives an extended comparison.

3 Surrogate geometry in min–max unlearning

3.1 Notation

Let $\mathbf{w} \in \mathbb{R}^p$ denote the full parameter vector. All coordinates are optimized unless a trainable subset is specified. The training data are partitioned into a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$. Write $L(\mathbf{w}; \mathcal{D})$ for the empirical loss on dataset $\mathcal{D}$. Set $L_f(\mathbf{w}) := L(\mathbf{w}; \mathcal{D}_f)$ and $L_r(\mathbf{w}) := L(\mathbf{w}; \mathcal{D}_r)$, with gradients $\mathbf{g}_f(\mathbf{w}) := \nabla_{\mathbf{w}} L_f(\mathbf{w})$ and $\mathbf{g}_r(\mathbf{w}) := \nabla_{\mathbf{w}} L_r(\mathbf{w})$.

For nonempty mini-batches $\mathcal{B}_f \subseteq \mathcal{D}_f$ and $\mathcal{B}_r \subseteq \mathcal{D}_r$, write $L_f(\mathbf{w}; \mathcal{B}_f) := L(\mathbf{w}; \mathcal{B}_f)$ and $L_r(\mathbf{w}; \mathcal{B}_r) := L(\mathbf{w}; \mathcal{B}_r)$. We use $\mathbf{I}$ for the identity matrix, $\|\cdot\|_2$ for the Euclidean norm, and $\|\cdot\|_{\mathrm{op}}$ for the induced operator norm. When both gradients are nonzero, define their coupling by

$$\cos\theta(\mathbf{w}) := \frac{\mathbf{g}_f(\mathbf{w})^\top \mathbf{g}_r(\mathbf{w})}{\|\mathbf{g}_f(\mathbf{w})\|_2 \, \|\mathbf{g}_r(\mathbf{w})\|_2}, \qquad \theta(\mathbf{w}) \in [0, \pi].$$

3.2 Min–max unlearning

Fine-tuning (FT) [5, 29] minimizes L_r, while negative gradient (NG) [5] maximizes L_f. We study updates that evaluate retention at a point selected by a forget-seeking perturbation [8]. To isolate surrogate selection, let $\rho > 0$ be a perturbation radius and write the bilevel template as follows:

$$\min_{\mathbf{w}} \; L_r\big(\mathbf{w} + \boldsymbol{\delta}^\star(\mathbf{w})\big) \quad \text{s.t.} \quad \boldsymbol{\delta}^\star(\mathbf{w}) \in \arg\max_{\|\boldsymbol{\delta}\|_2 \le \rho} \; L_f(\mathbf{w} + \boldsymbol{\delta}).$$

Here $\boldsymbol{\delta}^\star(\mathbf{w})$ denotes a selected inner maximizer, and $\mathbf{w} + \boldsymbol{\delta}^\star(\mathbf{w})$ is the surrogate point. The template keeps only the part of min–max unlearning [8] that chooses the surrogate point. For a perturbation $\boldsymbol{\delta}$, we call $\mathbf{g}_f(\mathbf{w})^\top \boldsymbol{\delta}$ its forget gain, the first-order increase of the forget loss. For $\mathbf{g}_f(\mathbf{w}) \neq \mathbf{0}$, maximizing the forget gain over the fixed Euclidean ball gives the standard perturbation

$$\boldsymbol{\delta}_{\mathrm{std}}(\mathbf{w}) = \rho \, \frac{\mathbf{g}_f(\mathbf{w})}{\|\mathbf{g}_f(\mathbf{w})\|_2}.$$

We use $\boldsymbol{\delta}_{\mathrm{std}}$ as the fixed-budget comparator. For $\mathbf{g}_r(\mathbf{w}) \neq \mathbf{0}$, its surrogate retain objective and first-order retain effect are

$$F_{\mathrm{std}}(\mathbf{w}) := L_r\big(\mathbf{w} + \boldsymbol{\delta}_{\mathrm{std}}(\mathbf{w})\big), \qquad \mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta}_{\mathrm{std}}(\mathbf{w}) = \rho \, \|\mathbf{g}_r(\mathbf{w})\|_2 \cos\theta(\mathbf{w}). \tag{1}$$

Positive coupling makes this term positive, so the surrogate point incurs a first-order retain cost before the retain step. The inner perturbation should therefore satisfy $\mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta} = 0$ while retaining as much forget gain as possible under the same budget.

3.3 Retain-neutral inner problem

Definition 3.1 (Retain-neutral inner problem)**.** Let $\mathbf{g}_r(\mathbf{w}) \neq \mathbf{0}$. A perturbation $\boldsymbol{\delta}$ is retain-neutral at $\mathbf{w}$ if $\mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta} = 0$. Among retain-neutral perturbations of norm at most ρ, the retain-neutral inner problem selects one with the largest forget gain,

$$\max_{\boldsymbol{\delta}} \ \mathbf{g}_f(\mathbf{w})^\top \boldsymbol{\delta} \quad \text{s.t.} \quad \|\boldsymbol{\delta}\|_2 \leq \rho, \quad \mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta} = 0. \tag{2}$$

The feasible set is the intersection of the perturbation ball and the retain-neutral hyperplane. Write

$$\mathbf{u}_r(\mathbf{w}) := \frac{\mathbf{g}_r(\mathbf{w})}{\|\mathbf{g}_r(\mathbf{w})\|_2}, \qquad \mathbf{P}_r(\mathbf{w}) := \mathbf{u}_r(\mathbf{w})\mathbf{u}_r(\mathbf{w})^\top, \qquad \mathbf{q}(\mathbf{w}) := (\mathbf{I} - \mathbf{P}_r(\mathbf{w}))\mathbf{g}_f(\mathbf{w}).$$

Here $\mathbf{u}_r$ is the unit retain direction, $\mathbf{P}_r$ projects onto its span, and $\mathbf{q}$ is the forget component orthogonal to that span. The next proposition shows that the constrained inner problem has a unique closed-form solution. The proof is in Appendix D.1.

Proposition 3.2 (Closed-form solution of the inner problem)**.** *Assume* $\mathbf{g}_r(\mathbf{w}) \neq \mathbf{0}$ *and* $\mathbf{q}(\mathbf{w}) \neq \mathbf{0}$*. Then the problem in* (2) *admits the unique maximizer*

$$\boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w}) = \rho \, \frac{\mathbf{q}(\mathbf{w})}{\|\mathbf{q}(\mathbf{w})\|_2}, \tag{3}$$

with optimal value $\rho\|\mathbf{q}(\mathbf{w})\|_2$.

The maximizer removes from the forget gradient its component along the retain gradient and spends the whole budget on the remainder. When $\mathbf{q}(\mathbf{w}) = \mathbf{0}$, every feasible perturbation has zero forget gain, and the normalized construction is undefined.

Retain neutrality scales the optimal forget gain by a factor that depends only on the angle between the two gradients. The constrained gain $\rho\|\mathbf{q}\|_2$ equals the unconstrained gain $\rho\|\mathbf{g}_f\|_2$ times $\sin\theta$ (Appendix D.1). The relative loss is negligible near orthogonality and grows as $|\cos\theta(\mathbf{w})|$ approaches one, while the standard perturbation incurs a positive first-order retain cost whenever $\cos\theta(\mathbf{w}) > 0$.

3.4 Retain-loss guarantees

A retain-neutral perturbation can still change the retain loss through curvature. To quantify this effect, define the surrogate retain objective

$$F_{\mathrm{rosu}}(\mathbf{w}) := L_r\big(\mathbf{w} + \boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w})\big).$$

This objective supplies the retain component of the full update in Section 4. The next theorem connects its value to the retain loss at the current weights $\mathbf{w}$ and compares it with the standard surrogate. Its proof expands both surrogates to second order under a Hessian bound on the two perturbation segments (Appendix D.1). The ideal construction gives these bounds, and Appendices D.3 and C.4 quantify the residuals introduced by regularization and mini-batch estimation.

Theorem 3.3 (Retain properties of the surrogate)**.** *Let* $\mathbf{g}_f(\mathbf{w}) \neq \mathbf{0}$*,* $\mathbf{g}_r(\mathbf{w}) \neq \mathbf{0}$*, and* $\mathbf{q}(\mathbf{w}) \neq \mathbf{0}$*. Assume that* L_r *is* C^2 *on an open neighborhood of the two segments* $\{\mathbf{w} + \zeta\boldsymbol{\delta}_{\mathrm{std}}(\mathbf{w}) : \zeta \in [0,1]\}$ *and* $\{\mathbf{w} + \zeta\boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w}) : \zeta \in [0,1]\}$*, and that* $\|\nabla^2 L_r(\mathbf{z})\|_{\mathrm{op}} \leq M_r$ *for every* $\mathbf{z}$ *on these segments. Then*

(i) $\mathbf{g}_r(\mathbf{w})^\top \boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w}) = 0$*, and*

$$\big|F_{\mathrm{rosu}}(\mathbf{w}) - L_r(\mathbf{w})\big| \leq \frac{M_r}{2}\rho^2.$$

(ii) The surrogate retain objectives of the two perturbations satisfy

$$F_{\mathrm{std}}(\mathbf{w}) - F_{\mathrm{rosu}}(\mathbf{w}) \geq \rho\,\|\mathbf{g}_r(\mathbf{w})\|_2 \cos\theta(\mathbf{w}) - M_r\rho^2.$$

By part (i) of Theorem 3.3, the retain loss at the surrogate point departs from the retain loss at $\mathbf{w}$ only through curvature, because the constraint removes the linear term. In part (ii), the first term is the retain cost that coupling induces in the standard surrogate, and the second bounds the difference between the curvature terms of both surrogates. The comparison between the two surrogates depends on the coupling measured in Figure 1a. Under positive coupling the first term favors ROSU, and Appendix D.1 states when it outweighs the second. Under negative coupling the standard perturbation has a negative first-order retain effect. Near orthogonality the two perturbations are close.

Algorithm 1 Retain-Orthogonal Surrogate Unlearning (ROSU)

Input: weights $\mathbf{w}$, nonempty forget mini-batch $\mathcal{B}_f$ and retain mini-batch $\mathcal{B}_r$
Parameters: learning rate $\eta > 0$, perturbation radius $\rho > 0$, amplification coefficient $\beta \geq 0$, stabilizer $\tau > 0$, threshold $\varepsilon_q > 0$
Output: updated weights $\mathbf{w}$

$\mathbf{g}_f \leftarrow \nabla L_f(\mathbf{w}; \mathcal{B}_f), \quad \mathbf{g}_r \leftarrow \nabla L_r(\mathbf{w}; \mathcal{B}_r)$
$\mathbf{q}_\tau \leftarrow \mathbf{g}_f - \frac{\mathbf{g}_f^\top \mathbf{g}_r}{\|\mathbf{g}_r\|_2^2 + \tau} \mathbf{g}_r$
if $\|\mathbf{q}_\tau\|_2 \leq \varepsilon_q$ **then**
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \mathbf{g}_r$ ▷ Degenerate fallback, Section 4.3
else
    $\mathbf{u}_\tau \leftarrow \frac{\mathbf{q}_\tau}{\|\mathbf{q}_\tau\|_2}, \quad \boldsymbol{\delta}_\tau \leftarrow \rho \mathbf{u}_\tau, \quad \alpha_\tau \leftarrow \frac{\rho}{\|\mathbf{q}_\tau\|_2}$
    $\widetilde{\mathbf{g}}_{r,\tau} \leftarrow \nabla_{\mathbf{z}} L_r(\mathbf{z}; \mathcal{B}_r)\big|_{\mathbf{z}=\mathbf{w}+\boldsymbol{\delta}_\tau}$
    $\mathbf{v} \leftarrow \widetilde{\mathbf{g}}_{r,\tau} + \alpha_\tau \Big(\widetilde{\mathbf{g}}_{r,\tau} - \frac{\mathbf{g}_r^\top \widetilde{\mathbf{g}}_{r,\tau}}{\|\mathbf{g}_r\|_2^2 + \tau} \mathbf{g}_r - (\mathbf{u}_\tau^\top \widetilde{\mathbf{g}}_{r,\tau}) \mathbf{u}_\tau \Big)$
    $\mathbf{w} \leftarrow \mathbf{w} + \beta \boldsymbol{\delta}_\tau - \eta \mathbf{v}$
end if

4 Retain-Orthogonal Surrogate Unlearning

Proposition 3.2 identifies the direction of maximal forget gain under retain neutrality. Theorem 3.3 bounds the retain loss at the resulting surrogate point. Let $\mathbf{w}^+$ denote the next iterate, $\eta > 0$ the learning rate, $\beta \geq 0$ the amplification coefficient, and $\mathbf{v}(\mathbf{w})$ the retain correction defined below. ROSU uses the same inner solution for retain correction and retain-neutral amplification and combines them in one update,

$$\mathbf{w}^+ = \mathbf{w} + \beta \boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w}) - \eta \mathbf{v}(\mathbf{w}).$$

The correction $\mathbf{v}(\mathbf{w})$ transports the retain gradient from the surrogate point back to $\mathbf{w}$. The retain-neutral amplification term $\beta\boldsymbol{\delta}_{\mathrm{rosu}}$ supplies the largest forget gain under its retain-neutral budget (Section 3.3).

4.1 Computing the retain correction

The gradient of the surrogate retain objective must account for the dependence of the surrogate point on $\mathbf{w}$. For $L_f, L_r \in C^2$, $\mathbf{g}_r \neq \mathbf{0}$, and $\mathbf{q} \neq \mathbf{0}$, set $\widetilde{\mathbf{g}}_r(\mathbf{w}) := \nabla_{\mathbf{z}} L_r(\mathbf{z})|_{\mathbf{z}=\mathbf{w}+\boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w})}$. Let $\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}$ denote the Jacobian of the perturbation map with ρ fixed. The chain rule gives

$$\nabla F_{\mathrm{rosu}}(\mathbf{w}) = \big(\mathbf{I} + \mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w})^\top\big)\widetilde{\mathbf{g}}_r(\mathbf{w}).$$

Here $\widetilde{\mathbf{g}}_r$ is the retain gradient at the ideal surrogate point. This Jacobian contains the forget Hessian and the differential of $\mathbf{P}_r$ (Appendix D.2). Both the surrogate point and its Jacobian depend on the forget data through $\boldsymbol{\delta}_{\mathrm{rosu}}$. The retain-neutral amplification term $\beta\boldsymbol{\delta}_{\mathrm{rosu}}$ scales this $\boldsymbol{\delta}_{\mathrm{rosu}}$, so the retain correction and the amplification direction come from the same inner solution.

Relaxations A1 and A2 make this correction tractable. A1 drops the projector differential from the derivative computation. A2 replaces the forget Hessian $\nabla^2 L_f(\mathbf{w})$ by $\mathbf{I}$, using an identity-Hessian approximation previously considered in surrogate-based unlearning [8]. The retain-neutral constraint is unchanged, and the projector is recomputed at each step. Set $\mathbf{u}_\perp := \frac{\mathbf{q}}{\|\mathbf{q}\|_2}$ and $\mathbf{P}_\perp := \mathbf{u}_\perp \mathbf{u}_\perp^\top$, with all quantities evaluated at $\mathbf{w}$. The relaxed Jacobian and correction are

$$\widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w}) = \frac{\rho}{\|\mathbf{q}(\mathbf{w})\|_2}\big(\mathbf{I} - \mathbf{P}_r(\mathbf{w}) - \mathbf{P}_\perp(\mathbf{w})\big), \qquad \mathbf{v} := \big(\mathbf{I} + \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w})\big)\widetilde{\mathbf{g}}_r(\mathbf{w}). \tag{4}$$

The Jacobian correction projects $\widetilde{\mathbf{g}}_r$ onto the complement of the retain and perturbation directions and scales it by $\frac{\rho}{\|\mathbf{q}\|_2}$ before adding it to $\widetilde{\mathbf{g}}_r$. The relaxed Jacobian is symmetric, so $\mathbf{v}$ is the chain-rule gradient with $\widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}$ in place of $\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}$. For this ideal unregularized map, the added term is orthogonal to both gradients at the current weights, so it leaves their first-order projections unchanged.

The next proposition bounds the deviation of the ideal unregularized A1/A2 correction from the exact chain-rule gradient. The bound depends on how far the forget Hessian is from the identity and on the projector differential $\mathrm{D}\mathbf{P}_r(\mathbf{w})[\boldsymbol{\xi}]$ along a direction $\boldsymbol{\xi}$. The proof is in Appendix D.2.

Proposition 4.1 (Deviation of the relaxed correction)**.** *Assume* $L_f, L_r \in C^2$*,* $\mathbf{g}_r(\mathbf{w}) \neq \mathbf{0}$*, and* $\mathbf{q}(\mathbf{w}) \neq \mathbf{0}$*, and let* $\varepsilon_H, \varepsilon_P \geq 0$ *satisfy*

$$\|\nabla^2 L_f(\mathbf{w}) - \mathbf{I}\|_{\mathrm{op}} \leq \varepsilon_H, \qquad \sup_{\|\boldsymbol{\xi}\|_2=1} \|\mathrm{D}\mathbf{P}_r(\mathbf{w})[\boldsymbol{\xi}]\|_{\mathrm{op}} \leq \varepsilon_P.$$

Then

$$\left\|\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w}) - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w})\right\|_{\mathrm{op}} \leq \frac{\rho}{\|\mathbf{q}(\mathbf{w})\|_2}\big(\varepsilon_H + \varepsilon_P \|\mathbf{g}_f(\mathbf{w})\|_2\big), \tag{5a}$$

$$\left\|\nabla F_{\mathrm{rosu}}(\mathbf{w}) - \mathbf{v}\right\|_2 \leq \frac{\rho\,\|\widetilde{\mathbf{g}}_r(\mathbf{w})\|_2}{\|\mathbf{q}(\mathbf{w})\|_2}\big(\varepsilon_H + \varepsilon_P \|\mathbf{g}_f(\mathbf{w})\|_2\big). \tag{5b}$$

The bound shrinks as the forget Hessian approaches the identity and as the projector varies slowly. When the deviation is smaller than $\|\mathbf{v}\|_2$, $-\mathbf{v}$ is a descent direction of F_{rosu} (Appendix D.2).

4.2 Retain-neutral amplification

Let $\Delta\mathbf{w} := \beta\boldsymbol{\delta}_{\mathrm{rosu}} - \eta\mathbf{v}$ be the ideal nondegenerate displacement. The identities $\mathbf{g}_r^\top \mathbf{q} = 0$ and $\mathbf{g}_f^\top \mathbf{q} = \|\mathbf{q}\|_2^2$ give

$$\mathbf{g}_r(\mathbf{w})^\top \Delta\mathbf{w} = -\eta\, \mathbf{g}_r(\mathbf{w})^\top \mathbf{v}, \qquad \mathbf{g}_f(\mathbf{w})^\top \Delta\mathbf{w} = \beta\rho\, \|\mathbf{q}(\mathbf{w})\|_2 - \eta\, \mathbf{g}_f(\mathbf{w})^\top \mathbf{v}. \tag{6}$$

For the ideal displacement in (6), retain-neutral amplification adds no first-order retain change. Its forget contribution $\beta\rho\|\mathbf{q}\|_2$ is optimal among retain-neutral displacements of norm at most $\beta\rho$. For $\beta > 0$, this follows from Proposition 3.2. For $\beta = 0$, the only such displacement is zero.

The forgetting step $\beta\boldsymbol{\delta}_{\mathrm{rosu}}$ has length $\beta\rho$, so the radius ρ also scales it. At iteration t, we therefore write the amplification level as $\beta_t = \frac{c_t\eta_t}{\rho}$, where η_t is the learning rate and $c_t \geq 0$ is a scale factor. On ideal nondegenerate steps the forgetting step is then $\beta_t\boldsymbol{\delta}_{\mathrm{rosu}} = c_t\eta_t \frac{\mathbf{q}}{\|\mathbf{q}\|_2}$, so its length $c_t\eta_t$ is independent of ρ, while the surrogate point and correction still depend on ρ. The update then reads $\mathbf{w}^+ = \mathbf{w} - \eta_t\big(\mathbf{v} - c_t \frac{\mathbf{q}}{\|\mathbf{q}\|_2}\big)$, so c_t weights the forgetting step against the retain correction. In our experiments, c_t is chosen so that the step length is either held fixed or largest in the early epochs and smaller afterwards. The schedule and level are selected together with ρ and η (Appendix B.4).

4.3 Regularization and fallback

Algorithms 1–2 give the update and its fallback with mini-batch gradients and a regularized retain operator. In the algorithms, $\mathbf{g}_f$ and $\mathbf{g}_r$ denote gradients on the indicated mini-batches, and the optimizer settings are in Appendix B.5.

For a stabilizer $\tau > 0$, Algorithm 1 uses $\mathbf{P}_r^{(\tau)} := \frac{\mathbf{g}_r\mathbf{g}_r^\top}{\|\mathbf{g}_r\|_2^2+\tau}$ and $\mathbf{q}_\tau = (\mathbf{I} - \mathbf{P}_r^{(\tau)})\mathbf{g}_f$. For $\mathbf{q}_\tau \neq \mathbf{0}$, set $\mathbf{u}_\tau := \frac{\mathbf{q}_\tau}{\|\mathbf{q}_\tau\|_2}$, $\boldsymbol{\delta}_\tau := \rho\mathbf{u}_\tau$, and $\alpha_\tau := \frac{\rho}{\|\mathbf{q}_\tau\|_2}$.

Algorithm 1 tests $\|\mathbf{q}_\tau\|_2 \leq \varepsilon_q$, where $\varepsilon_q > 0$ is the degeneracy threshold. For $\mathbf{g}_r \neq \mathbf{0}$, orthogonal decomposition gives $\|\mathbf{q}\|_2 \leq \|\mathbf{q}_\tau\|_2$. The test therefore bounds the optimal forget gain in (2) by $\rho\varepsilon_q$. The fallback takes a retain-gradient step and forgoes at most this gain.

The algorithm evaluates $\widetilde{\mathbf{g}}_{r,\tau} := \nabla_{\mathbf{z}} L_r(\mathbf{z}; \mathcal{B}_r)|_{\mathbf{z}=\mathbf{w}+\boldsymbol{\delta}_\tau}$ on steps with $\|\mathbf{q}_\tau\|_2 > \varepsilon_q$ and uses the symmetric relaxed Jacobian of (4) at the regularized direction. Appendix D.3 bounds the resulting differences from the ideal construction, which are $O(\tau)$ as $\tau \to 0$ at any $\mathbf{w}$ with $\mathbf{g}_r \neq \mathbf{0}$ and $\mathbf{q} \neq \mathbf{0}$. All experiments fix $\tau = 10^{-8}$ and $\varepsilon_q = 10^{-6}$.

5 Experiments

We ask whether ROSU comes closest to retraining on vision benchmarks (Section 5.1) and is competitive with the baselines on LLM unlearning (Section 5.2), and which components matter (Section 5.3). Our main benchmarks are CIFAR-10, CIFAR-100 [30], and Tiny-ImageNet [31] for vision, and TOFU [10] and WMDP [20] for language. Tables 1–3 and Table 6 in Appendix C.1 report the main results, and Figures 2 and 3 show qualitative TOFU examples and an ablation of the A1/A2

Table 1: **Unlearning on CIFAR-10.** Values in parentheses are absolute differences from Retrain, and standard deviations are over ten forgotten classes for class-wise forgetting and three seeds for random forgetting. In each protocol, bold and underline mark the smallest and second-smallest ΔAcc among the unlearning methods.

Method	RA	FA	TA	ΔAcc.↓	MIA-Eff.	Time (min)
Class-wise forgetting						
Retrain	$99.99_{\pm 0.00}$	$0.00_{\pm 0.00}$	$95.15_{\pm 0.56}$	0.00	$100.00_{\pm 0.00}$	33.8
FT	$98.22_{\pm 0.54}$ (1.77)	$24.94_{\pm 6.99}$ (24.94)	$92.68_{\pm 1.07}$ (2.47)	29.18	$99.13_{\pm 0.98}$ (0.87)	1.63
NG	$88.03_{\pm 2.92}$ (11.96)	$0.84_{\pm 1.21}$ (0.84)	$83.00_{\pm 2.89}$ (12.15)	24.95	$99.28_{\pm 1.10}$ (0.72)	1.64
FF	$98.95_{\pm 0.12}$ (1.04)	$10.98_{\pm 11.29}$ (10.98)	$93.26_{\pm 0.62}$ (1.89)	13.91	$100.00_{\pm 0.00}$ (0.00)	1.07
IU	$90.86_{\pm 9.85}$ (9.13)	$10.28_{\pm 15.62}$ (10.28)	$85.86_{\pm 8.80}$ (9.29)	28.70	$98.27_{\pm 3.09}$ (1.73)	0.59
ℓ_1-sparse	$97.18_{\pm 0.46}$ (2.81)	$0.00_{\pm 0.00}$ (0.00)	$92.07_{\pm 0.76}$ (3.08)	5.89	$100.00_{\pm 0.00}$ (0.00)	1.73
GU	$94.20_{\pm 2.36}$ (5.79)	$6.48_{\pm 3.72}$ (6.48)	$88.72_{\pm 2.19}$ (6.43)	18.70	$95.19_{\pm 3.10}$ (4.81)	7.59
OrthoGrad	$98.59_{\pm 0.72}$ (1.40)	$0.46_{\pm 0.77}$ (0.46)	$92.31_{\pm 0.92}$ (2.84)	4.70	$99.66_{\pm 0.66}$ (0.34)	7.62
PCGrad	$98.58_{\pm 0.32}$ (1.41)	$0.07_{\pm 0.12}$ (0.07)	$92.93_{\pm 0.78}$ (2.22)	3.70	$100.00_{\pm 0.00}$ (0.00)	2.10
UAM	$99.91_{\pm 0.05}$ (0.08)	$0.02_{\pm 0.03}$ (0.02)	$94.42_{\pm 0.68}$ (0.73)	<u>0.83</u>	$100.00_{\pm 0.00}$ (0.00)	2.47
ROSU	$99.96_{\pm 0.01}$ (0.03)	$0.00_{\pm 0.00}$ (0.00)	$95.01_{\pm 0.53}$ (0.14)	**0.17**	$100.00_{\pm 0.00}$ (0.00)	2.98
Random forgetting						
Retrain	$99.99_{\pm 0.00}$	$95.11_{\pm 0.37}$	$94.38_{\pm 0.23}$	0.00	$11.61_{\pm 0.68}$	33.8
FT	$99.29_{\pm 0.06}$ (0.70)	$99.06_{\pm 0.08}$ (3.95)	$93.28_{\pm 0.15}$ (1.10)	5.75	$4.07_{\pm 0.30}$ (7.54)	1.64
NG	$64.30_{\pm 45.24}$ (35.69)	$64.18_{\pm 45.23}$ (30.93)	$60.74_{\pm 42.32}$ (33.64)	100.26	$4.27_{\pm 3.88}$ (7.34)	1.63
FF	$79.72_{\pm 7.37}$ (20.27)	$80.02_{\pm 7.12}$ (15.09)	$77.22_{\pm 5.93}$ (17.16)	52.52	$40.24_{\pm 18.03}$ (28.63)	1.16
IU	$95.72_{\pm 5.57}$ (4.27)	$95.67_{\pm 5.51}$ (0.56)	$89.89_{\pm 5.37}$ (4.49)	9.32	$6.87_{\pm 7.00}$ (4.74)	0.59
ℓ_1-sparse	$98.22_{\pm 0.51}$ (1.77)	$96.43_{\pm 0.85}$ (1.32)	$92.13_{\pm 0.52}$ (2.25)	5.34	$8.20_{\pm 0.48}$ (3.41)	1.77
GU	$98.60_{\pm 0.29}$ (1.39)	$98.77_{\pm 0.49}$ (3.66)	$92.96_{\pm 0.31}$ (1.42)	6.47	$3.78_{\pm 0.87}$ (7.83)	7.59
OrthoGrad	$99.97_{\pm 0.02}$ (0.02)	$94.81_{\pm 0.82}$ (0.30)	$92.65_{\pm 0.31}$ (1.73)	<u>2.05</u>	$7.81_{\pm 1.12}$ (3.80)	7.60
PCGrad	$98.48_{\pm 0.30}$ (1.51)	$95.57_{\pm 0.41}$ (0.46)	$92.27_{\pm 0.41}$ (2.11)	4.08	$8.85_{\pm 0.50}$ (2.76)	1.99
UAM	$99.79_{\pm 0.05}$ (0.20)	$96.25_{\pm 0.22}$ (1.14)	$93.29_{\pm 0.28}$ (1.09)	2.43	$8.02_{\pm 0.22}$ (3.59)	2.44
ROSU	$99.84_{\pm 0.03}$ (0.15)	$95.13_{\pm 0.23}$ (0.02)	$94.29_{\pm 0.09}$ (0.09)	**0.26**	$7.88_{\pm 0.16}$ (3.73)	2.92

Table 2: **Unlearning on CIFAR-100.** Same notation as Table 1.

Method	RA	FA	TA	ΔAcc.↓	MIA-Eff.	Time (min)
Class-wise forgetting						
Retrain	$99.98_{\pm 0.00}$	$0.00_{\pm 0.00}$	$77.80_{\pm 0.19}$	0.00	$100.00_{\pm 0.00}$	36.8
FT	$99.98_{\pm 0.00}$ (0.00)	$84.44_{\pm 5.72}$ (84.44)	$78.34_{\pm 0.22}$ (0.54)	84.98	$99.42_{\pm 0.47}$ (0.58)	1.79
NG	$53.58_{\pm 14.54}$ (46.40)	$10.12_{\pm 29.96}$ (10.12)	$38.83_{\pm 15.37}$ (38.97)	95.49	$89.78_{\pm 30.26}$ (10.22)	1.82
FF	$99.98_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.00}$ (0.00)	$77.53_{\pm 0.15}$ (0.27)	<u>0.27</u>	$100.00_{\pm 0.00}$ (0.00)	11.24
IU	$99.08_{\pm 0.74}$ (0.90)	$10.28_{\pm 16.95}$ (10.28)	$73.83_{\pm 1.53}$ (3.97)	15.15	$99.34_{\pm 1.30}$ (0.66)	0.57
ℓ_1-sparse	$99.98_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.00}$ (0.00)	$74.92_{\pm 0.38}$ (2.88)	2.88	$100.00_{\pm 0.00}$ (0.00)	1.99
GU	$90.34_{\pm 4.81}$ (9.64)	$1.28_{\pm 1.18}$ (1.28)	$66.46_{\pm 3.22}$ (11.34)	22.26	$98.90_{\pm 0.91}$ (1.10)	8.34
OrthoGrad	$97.30_{\pm 2.53}$ (2.68)	$1.68_{\pm 1.88}$ (1.68)	$71.70_{\pm 2.44}$ (6.10)	10.46	$98.74_{\pm 1.17}$ (1.26)	8.18
PCGrad	$99.24_{\pm 0.69}$ (0.74)	$0.99_{\pm 0.87}$ (0.99)	$74.06_{\pm 1.66}$ (3.74)	5.47	$100.00_{\pm 0.00}$ (0.00)	2.43
UAM	$99.97_{\pm 0.01}$ (0.01)	$0.18_{\pm 0.25}$ (0.18)	$76.63_{\pm 0.73}$ (1.17)	1.36	$100.00_{\pm 0.00}$ (0.00)	2.78
ROSU	$99.98_{\pm 0.00}$ (0.00)	$0.12_{\pm 0.26}$ (0.12)	$77.78_{\pm 0.13}$ (0.02)	**0.14**	$100.00_{\pm 0.00}$ (0.00)	3.22
Random forgetting						
Retrain	$99.98_{\pm 0.00}$	$77.70_{\pm 0.26}$	$76.66_{\pm 0.18}$	0.00	$47.39_{\pm 0.35}$	33.9
FT	$99.98_{\pm 0.00}$ (0.00)	$99.96_{\pm 0.03}$ (22.26)	$78.38_{\pm 0.09}$ (1.72)	23.98	$18.20_{\pm 0.62}$ (29.19)	1.65
NG	$32.71_{\pm 2.02}$ (67.27)	$31.53_{\pm 2.04}$ (46.17)	$25.79_{\pm 1.28}$ (50.87)	164.31	$65.20_{\pm 1.81}$ (17.81)	1.67
FF	$1.08_{\pm 0.09}$ (98.90)	$0.99_{\pm 0.09}$ (76.71)	$1.06_{\pm 0.12}$ (75.60)	251.21	$33.69_{\pm 46.21}$ (13.70)	11.17
IU	$97.51_{\pm 0.41}$ (2.47)	$96.59_{\pm 0.32}$ (18.89)	$71.28_{\pm 0.40}$ (5.38)	26.74	$10.46_{\pm 0.81}$ (36.93)	0.59
ℓ_1-sparse	$99.93_{\pm 0.03}$ (0.05)	$79.45_{\pm 0.42}$ (1.75)	$73.02_{\pm 0.27}$ (3.64)	5.44	$50.67_{\pm 0.20}$ (3.28)	1.79
GU	$96.48_{\pm 0.20}$ (3.50)	$96.45_{\pm 0.32}$ (18.75)	$73.05_{\pm 0.10}$ (3.61)	25.86	$9.91_{\pm 0.35}$ (37.48)	7.46
OrthoGrad	$99.64_{\pm 0.02}$ (0.34)	$76.72_{\pm 1.16}$ (0.98)	$69.05_{\pm 0.09}$ (7.61)	8.93	$32.99_{\pm 1.69}$ (14.40)	7.51
PCGrad	$95.88_{\pm 0.59}$ (4.10)	$78.22_{\pm 0.54}$ (0.52)	$69.48_{\pm 0.56}$ (7.18)	11.80	$29.75_{\pm 0.58}$ (17.64)	2.09
UAM	$99.86_{\pm 0.02}$ (0.12)	$78.16_{\pm 0.30}$ (0.46)	$74.70_{\pm 0.07}$ (1.96)	<u>2.54</u>	$38.55_{\pm 0.95}$ (8.84)	2.41
ROSU	$99.78_{\pm 0.00}$ (0.20)	$77.98_{\pm 0.08}$ (0.28)	$76.85_{\pm 0.18}$ (0.19)	**0.67**	$41.37_{\pm 0.75}$ (6.02)	2.98

relaxations. Additional results and extended ablations are provided in Appendix C. Unless noted otherwise, all results use the full ROSU update in Algorithm 1, and WMDP uses the representation-level variant in Algorithm 2.

5.1 Image classification unlearning

The vision configuration and the class-wise and random forgetting protocols follow UAM [8], with ResNet-18 [32] for CIFAR-10/100 [30] and VGG [33] for Tiny-ImageNet [31]. We compare against

FT [5, 29], NG [5], Fisher forgetting (FF) [5], influence unlearning (IU) [3, 34], ℓ_1-sparse [9], a PCGrad-style projection [23], OrthoGrad [25], GU [26], and UAM [8]. The tables also report runtime in minutes, measured as the wall-clock time of the unlearning loop on a single GPU. Baseline formulations, protocol details, and selection rules are provided in Appendix B.

We report retain accuracy (RA), forget accuracy (FA), test accuracy (TA), MIA-Eff. [9], $\Delta\mathrm{Acc}$, and runtime. MIA-Eff. (membership-inference attack efficacy) is the forget-set non-member rate. The aggregate $\Delta\mathrm{Acc}$ is computed from the mean RA, FA, and TA by summing their absolute differences from Retrain.

Tables 1, 2, and 6 compare ROSU with the approximate-unlearning baselines. ROSU has the smallest $\Delta\mathrm{Acc}$ in each of the six dataset–protocol settings. Its reduction in $\Delta\mathrm{Acc}$ relative to UAM is largest in random forgetting, which Figure 1a shows to be highly coupled. PCGrad, OrthoGrad, and GU enforce orthogonality only at the final update, and each has a larger $\Delta\mathrm{Acc}$ than ROSU in all six settings. Section 5.3 examines the update components and derivative variants separately.

5.2 LLM unlearning

For TOFU, we use `Llama-3.2-1B-Instruct` [35] with the OpenUnlearning pipeline [11], following its checkpoint-selection and evaluation protocol. For WMDP, we follow the setup of UAM [8], which uses the `Zephyr-7B-`β [36] backbone of the WMDP benchmark [20]. Appendices B.2 and B.6 give the setup and the formal definitions.

For TOFU, we compare ROSU with GradDiff, IdkDPO, IdkNLL [10], NPO [19], SimNPO [37], AltPO [38], UNDIAL [39], PCGrad [23], RMU [20], and UAM [8]. For WMDP, we compare against SSD [17], SCRUB [15], LLMU [18], RMU [20], and UAM [8].

For TOFU, Mem, Priv, and Util summarize memorization removal, membership-attack agreement with the retain model, and retained utility. Agg is their harmonic mean (HM), while $\mathrm{Agg}_{\mathrm{MU}}$ uses Mem and Util for model selection. For WMDP, MMLU [40] measures general knowledge [20], while Bio and Cyber measure hazardous-domain accuracy. The score $\Delta\mathrm{Score}$ sums the absolute MMLU change from the base model and the Bio and Cyber distances from chance. Appendix B.3 gives the full metric definitions.

Table 3 shows that ROSU attains the best aggregate score on TOFU among the unlearning methods. It matches UAM on memorization and exceeds it in privacy and utility. Figure 2 shows two forget-set prompts and the responses before and after unlearning. In both examples, the post-unlearning response is fluent but gives a name that differs from the ground truth.

On WMDP, ROSU attains the smallest $\Delta\mathrm{Score}$ among the unlearning methods (Table 3). It also has the highest MMLU apart from the base model and the lowest Bio accuracy. Its Cyber accuracy is near the chance level.

Table 3: **Unlearning on TOFU and WMDP.** *Init. ft.* and Base are the TOFU and WMDP models before unlearning, and *Retain* is the model fine-tuned on the retain data only. Bold and underline mark the best and second-best Agg. on TOFU and ΔScore on WMDP among the unlearning methods we compare.

TOFU				
Method	**Agg.↑**	**Mem.↑**	**Priv.↑**	**Util.↑**
Init. ft.	0.00	0.00	0.01	1.00
Retain	0.61	0.35	1.00	0.99
GradDiff	0.00	0.99	0.00	0.77
NPO	0.02	0.52	0.01	0.97
IdkDPO	0.03	0.56	0.01	0.93
AltPO	0.05	0.67	0.02	0.94
PCGrad	0.06	0.98	0.02	0.61
IdkNLL	0.13	0.10	0.08	0.91
UNDIAL	0.44	0.30	0.46	0.79
SimNPO	0.49	0.36	0.44	0.99
UAM	0.50	0.38	0.62	0.57
RMU	<u>0.54</u>	0.48	0.53	0.62
ROSU	**0.58**	0.38	0.74	0.82

WMDP				
Method	**ΔScore↓**	**MMLU↑**	**Bio**	**Cyber**
Base	0.577	0.581	0.637	0.440
SSD	0.526	0.407	0.502	0.350
SCRUB	0.400	0.512	0.438	0.393
LLMU	0.624	0.447	0.595	0.395
RMU	0.101	0.566	0.310	0.276
UAM	<u>0.077</u>	0.564	0.293	0.233
ROSU	**0.054**	0.570	0.289	0.246

5.3 Ablations and sensitivity

Appendix C.2 separates retain correction from retain-neutral amplification. Amplification alone forgets but degrades retention in every setting. Correction alone keeps retention but forgets too little, with forget accuracy above Retrain in all six vision settings and no memorization removal on TOFU. On WMDP, it instead reduces general utility

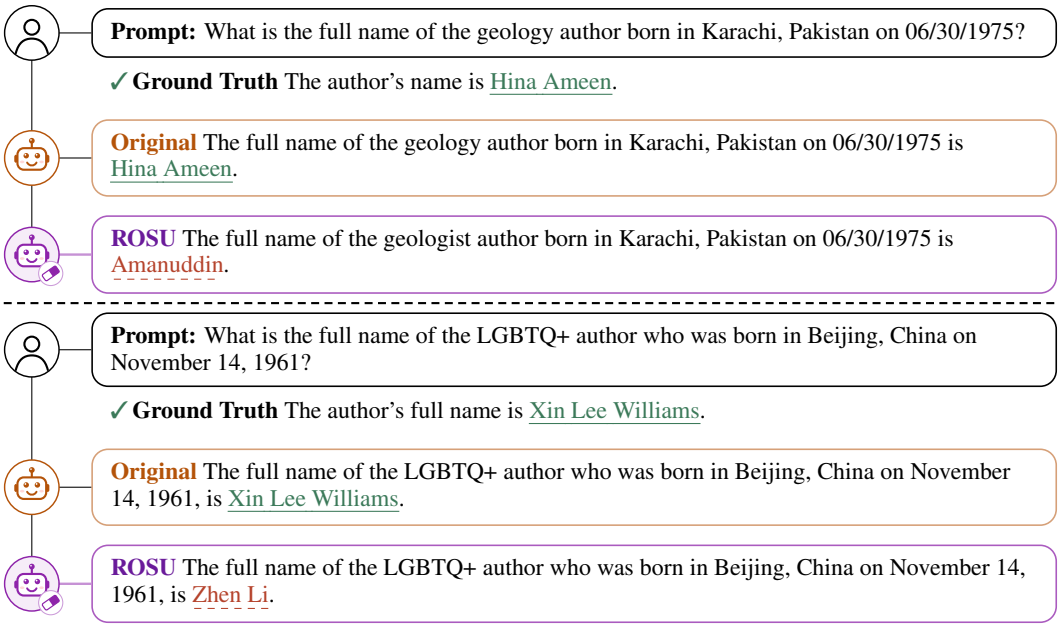


Figure 2: **Qualitative TOFU forgetting examples.** Responses to two forget-set prompts before and after unlearning with ROSU. Green underline marks content consistent with the ground truth, and red dashed underline marks incorrect content.

toward chance. Without the Jacobian correction, the ΔAcc gap to full ROSU is larger in the more coupled settings of each protocol (Figure 1a), and Agg on TOFU drops. Appendices B.4 and D.1 discuss amplification tuning and its retain-loss bound.

Figure 3 compares four derivative variants. The retain-neutral constraint is kept in every variant, so A1 and A2 change only how the correction $\mathbf{v}$ is computed. A1+A2 drops the retain-projector differential and replaces the forget Hessian by the identity. No A1 keeps the projector differential, No A2 keeps the forget Hessian through a Hessian–vector product, and Exact (no A1, no A2) keeps both. No A1 closely tracks A1+A2, whereas No A2 and Exact use the true forget Hessian and diverge during training in the two random-forgetting settings with the highest coupling. Both collapse early in training, and the identity-Hessian relaxation keeps training stable.

Appendix C.4 bounds full-retain orthogonality violation by gradient-estimation error and a regularization residual. Figure 5 shows lower measured error with larger retain batches. Appendix C.3 reports a radius sweep and safeguard audits. For every radius in Table 4, ROSU keeps ΔAcc close to the maintable values on CIFAR-10 and CIFAR-100 random forgetting. On TOFU, Agg stays close to its maintable value for every radius in Table 5. The recorded projected-gradient norms in Figure 4 stay above the degeneracy threshold ε_q in all audited records. Appendix C.5 extends ROSU to sampled rank-k protected subspaces on CIFAR-100 random forgetting.

Table 4: **Sensitivity to ρ on CIFAR-100 and CIFAR-10.** Results are for random forgetting, with standard deviations over three seeds.

ρ	RA	FA	TA	ΔAcc.↓
CIFAR-100				
0.001	$99.95_{\pm 0.00}$	$77.31_{\pm 0.62}$	$76.35_{\pm 0.17}$	0.73
0.005	$99.97_{\pm 0.00}$	$77.33_{\pm 0.56}$	$76.37_{\pm 0.08}$	0.67
0.01	$99.97_{\pm 0.01}$	$77.47_{\pm 0.20}$	$76.22_{\pm 0.04}$	0.68
CIFAR-10				
0.001	$99.99_{\pm 0.00}$	$95.05_{\pm 0.02}$	$94.18_{\pm 0.02}$	0.26
0.003	$99.99_{\pm 0.01}$	$95.02_{\pm 0.22}$	$94.20_{\pm 0.12}$	0.27
0.005	$100.00_{\pm 0.00}$	$95.23_{\pm 0.18}$	$94.27_{\pm 0.11}$	0.24

Table 5: **Sensitivity to ρ on TOFU.** Following [11], the selected ρ maximizes Agg_{MU}, and bold marks this ρ and its Agg_{MU}.

ρ	Agg.↑	Agg_{MU}↑	Mem.↑	Priv.↑	Util.↑
TOFU					
0.20	0.59	0.49	0.35	1.00	0.81
0.25	0.59	0.49	0.36	1.00	0.78
0.30	0.60	0.50	0.38	0.98	0.73
0.35	0.58	0.49	0.35	0.93	0.80
0.40	0.58	**0.52**	0.38	0.74	0.82
0.45	0.58	0.49	0.35	0.92	0.81
0.50	0.56	0.49	0.36	0.82	0.76

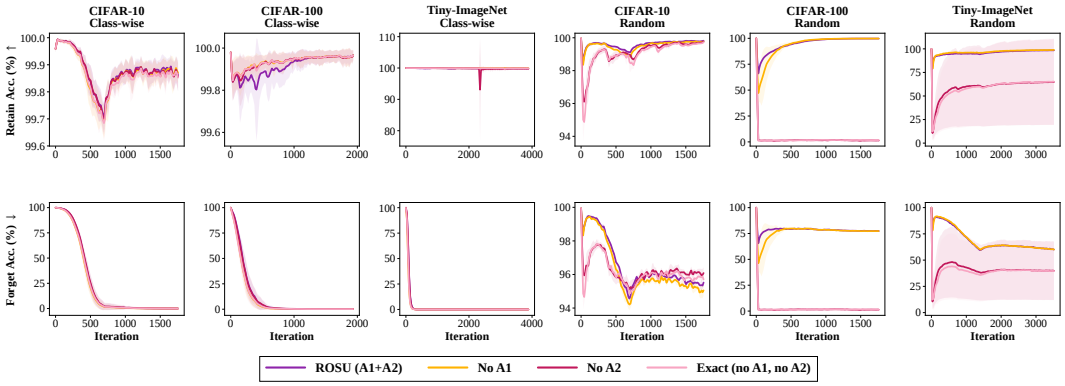


Figure 3: **A1/A2 derivative variants.** Retain and forget accuracy of four ways of computing the retain correction. Curves and bands are the mean and one standard deviation over ten forgotten classes or three seeds.

6 Limitations and broader impacts

Deletion requests, memorized sensitive content, and hazardous knowledge all call for making a trained model forget designated data without retraining it from scratch. Unlearning could also be misused, for example to remove the safety behaviors of a model or to hide which data it was trained on. The difficulty is to forget the designated data without hurting the model's performance on the data to keep. It is greater when the fastest way to raise the model's error on the data to forget also raises its error on the data to keep. We find this effect strong when random training examples or facts about fictitious authors are forgotten, and weak when a whole class is forgotten. ROSU looks for the change that forgets fastest among changes that leave the error on the data to keep unchanged to a first approximation.

ROSU uses the data to keep at every step, and each step requires more computation than a step of fine-tuning on these data. We see three directions for future research. One is unlearning when the data to keep are no longer available, which would require an unlearning algorithm to protect performance on data it cannot access. Another is unlearning without iterative training, which would require a single change to the model to forget without hurting performance on the data to keep. A third is unlearning when deletion requests arrive one after another, which would require each change to leave the data of earlier requests forgotten.

7 Conclusion

In min–max unlearning, an inner adversary perturbs the model to maximize the forget loss, and the retain loss is minimized at the resulting surrogate point. When the forget and retain gradients are positively coupled, this perturbation also moves toward higher retain loss. We propose ROSU, whose inner adversary avoids the retain cost that coupling induces by admitting only perturbations orthogonal to the retain gradient. The one that raises the forget loss fastest has a closed form, sets the surrogate point, and is amplified into the forgetting step. ROSU has the smallest accuracy gap to retraining in every vision setting and is competitive with the baselines on the language benchmarks.

Acknowledgments and Disclosure of Funding

This work is supported in part by the NRF grants funded by the Korea government(MSIT) (RS-2026-25488671, 33% and RS-2022-NR070834, 33%), and in part by ICT Creative Consilience Program through the IITP grant funded by the Korea government(MSIT)(IITP-2026-RS-2020-II201819, 33%).

References

[1] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

[3] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pages 2008–2016. PMLR, 2021.

[4] Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

[5] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312, 2020.

[6] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 792–801, 2021.

[7] Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[8] Hoki Kim, Keonwoo Kim, Sungwon Chae, and Sangwon Yoon. Unlearning-aware minimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

[10] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024.

[11] Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, Zachary Chase Lipton, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[12] Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. ARCANE: An efficient architecture for exact machine unlearning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4006–4013. International Joint Conferences on Artificial Intelligence Organization, 2022.

[13] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3832–3842. PMLR, 2020.

[14] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

[15] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

[16] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12043–12051, 2024.

[18] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024.

[20] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 28525–28550. PMLR, 2024.

[21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[22] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pages 5905–5914. PMLR, 2021.

[23] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[24] Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[25] Aviv Shamsian, Eitan Shaar, Aviv Navon, Gal Chechik, and Ethan Fetaya. Go beyond your means: Unlearning with per-sample gradient orthogonalization. *arXiv preprint arXiv:2503.02312*, 2025.

[26] Duo Zhou, Yuji Zhang, Tianxin Wei, Ruizhong Qiu, Ke Yang, Xiao Lin, Cheng Qian, Jingrui He, Hanghang Tong, Chengxiang Zhai, et al. Geometric-disentanglement unlearning. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 57231–57255. Association for Computational Linguistics, 2026.

[27] Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[28] Jinwei Hu, Zhenglin Huang, Xiangyu Yin, Wenjie Ruan, Guangliang Cheng, Yi Dong, and Xiaowei Huang. FALCON: Fine-grained activation manipulation by contrastive orthogonal unalignment for large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[29] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Proc. of the 30th Network and Distributed System Security (NDSS)*, 2023.

[30] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[31] Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[35] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[36] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.

[37] Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[38] Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752, 2025.

[39] Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, 2025.

[40] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[41] Alessandro Mantelero. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

[42] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The European Union general data protection regulation: what it is and what it means. *Information & communications technology law*, 28(1):65–98, 2019.

[43] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[44] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11186–11194, 2024.

[45] Jacopo Bonato, Marco Cotogni, and Luigi Sabetta. Is retain set all you need in machine unlearning? Restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[46] Junhao Cai, Dohun Kim, Sung Il Choi, Juhyun Park, Chengjun Jin, Dowon Kim, and Changhee Joo. SCOPE: Entanglement frontier escape for source-free class unlearning. *arXiv preprint arXiv:2608.02058*, 2026.

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[48] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[49] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, 2023.

[50] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

[51] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

[52] Yaxuan Wang, Jiaheng Wei, Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations*, pages 43076–43104, 2025.

[53] Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. ReLearn: Unlearning via learning for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5967–5987, 2025.

[54] Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[55] Andrei Ioan Muresanu, Anvith Thudi, Michael R. Zhang, and Nicolas Papernot. Fast exact unlearning for in-context learning data for LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[56] Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, 2024.

[57] Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. Opt-Out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28280–28297, 2025.

[58] Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4046–4056, 2025.

[59] Chenlu Ding, Jiancan Wu, Yancheng Yuan, Jinda Lu, Kai Zhang, Alex Su, Xiang Wang, and Xiangnan He. Unified parameter-efficient unlearning for LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[60] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[61] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[62] Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

[63] Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei S Wu. Reconstruction attacks on machine unlearning: Simple models are vulnerable. *Advances in Neural Information Processing Systems*, 37:104995–105016, 2024.

[64] Jacob L. Block, Aryan Mokhtari, and Sanjay Shakkottai. Machine unlearning under overparameterization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[65] Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Damian Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[66] Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *Forty-second International Conference on Machine Learning*, 2025.

[67] Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4211–4221, 2025.

[68] Mark He Huang, Lin Geng Foo, and Jun Liu. Learning to unlearn for robust machine unlearning. In *European conference on computer vision*, pages 202–219. Springer, 2024.

[69] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 241–257, 2019.

[70] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[71] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

[72] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[73] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

[74] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[75] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

A Extended related work

Machine unlearning is motivated by data-deletion requirements and the "right to be forgotten" in privacy regulation [41, 42]. Unlearning methods trade the removal of the forget data against performance on the retained data. For deep classifiers, early baselines take the two extremes of this trade-off [5]. One ascends the loss on the forget data, and the other fine-tunes on the retained data and relies on catastrophic forgetting [43]. Later methods refine the trade-off by changing what is removed, which data and targets the update uses, or which parameters change. Some unlearn features and labels or individual instances [29, 44]. Others replace the retain set with out-of-distribution images [45], use the forget data alone [46], or take their targets from the original model [15]. Parameter-selection methods decide which parameters to dampen, mask, or sparsify [9, 16, 17].

For large language models [35, 47], the same trade-off arises for new forget targets, such as memorized text, copyrighted content, unsafe behavior, and hazardous knowledge [20, 48–50]. The goal is then to remove specific facts or capabilities while preserving general knowledge and utility [18, 51–54]. Exact unlearning becomes efficient when in-context learning replaces fine-tuning [55]. Approximate methods instead change the forgetting objective [19, 37–39], the optimizer [56], the regularizer [57], or the trained parameters [58, 59]. Benchmarks measure the trade-off on fictitious, hazardous, memorized, and real-world knowledge [10, 20, 60, 61], and OpenUnlearning unifies methods, metrics, and stress tests such as relearning and quantization [11]. Privacy attacks reveal the limits of unlearning, since unlearned texts remain detectable [62] and the unlearning update can expose the deleted data to reconstruction [63].

Across both settings, projection methods restrict the final update, most often to resolve a conflict between the forget and retain gradients. As in gradient surgery for multi-task learning [23], some methods project the forget gradient away from the retain gradient only when the two conflict [27, 28]. Others protect retention more strictly and keep the forget gradient orthogonal to every per-sample retain gradient [25] or to the layer features of the retained data [24]. When the model interpolates the retained data, the retain gradients vanish, so one method keeps the update orthogonal to the gradients of the model outputs and minimizes a complexity objective [64]. The constraint itself can also change, with orthogonality measured in an optimizer-induced metric [26] or the update restricted to low-rank subspaces of the forget gradients [65]. In every case the constraint acts on the final update.

Min–max methods act earlier, at the point where an objective is evaluated. UAM maximizes the forget loss over norm-bounded perturbations and minimizes the retain loss at the resulting surrogate point [8]. A sharpness-aware formulation for LLMs applies this structure to the forgetting objective, with an inner perturbation that simulates a relearning attack [66]. Two related methods replace the inner maximization with a temporary step on one objective and evaluate the other objective after it. One steps on the retain loss before evaluating the forget loss [67], and the other steps on the forget loss before testing retention and projects away gradient conflicts [68]. Each of these methods chooses its evaluation point by one objective alone, and the projection methods constrain only the final update. ROSU uses both gradients to choose its surrogate point. Its inner perturbation raises the forget loss only in directions orthogonal to the retain gradient, which places the constraint on the surrogate point as well as on the forgetting step.

B Experimental setup and metrics

Appendix B.1 describes vision experiments, B.2 covers TOFU and WMDP, B.3–B.4 define metrics and model selection, B.5 provides implementation details, and B.6 documents algorithm variants.

B.1 Image classification protocol

The vision configuration and the class-wise and random forgetting protocols follow the image-classification setup of UAM [8]. CIFAR-10 and CIFAR-100 [30] use ResNet-18 [32], while Tiny-ImageNet [31] uses VGG [33].

The unlearning budgets are 5 epochs for UAM, ROSU, PCGrad, OrthoGrad, and GU, and 10 epochs for FT [5, 29], NG [5], and ℓ_1-sparse [9]. FF and IU apply a single parameter update [9].

Supervised pretraining uses SGD with initial learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and a 200-epoch schedule with learning-rate drops at epochs 100 and 150 [8].

The update-orthogonalization baselines are defined as follows:

- The PCGrad-style [23] baseline uses a trade-off coefficient λ_{pc} and the projected forget gradient $\mathbf{q}_{\mathrm{pc}} = \mathbf{g}_f - \frac{\mathbf{g}_f^\top \mathbf{g}_r}{\|\mathbf{g}_r\|_2^2}\,\mathbf{g}_r$ for $\mathbf{g}_r \neq \mathbf{0}$, with $\mathbf{q}_{\mathrm{pc}} = \mathbf{g}_f$ for $\mathbf{g}_r = \mathbf{0}$. The update is
$$\mathbf{w} \leftarrow \mathbf{w} - \eta\big(\mathbf{g}_r - \lambda_{\mathrm{pc}}\,\mathbf{q}_{\mathrm{pc}}\big).$$
- The OrthoGrad [25] baseline uses a mixing coefficient α and the mini-batch average retain gradient $\mathbf{g}_r$. The forget gradient is projected as $\mathbf{g}_f^\perp = \mathbf{g}_f - \mathbf{Q}_r\mathbf{Q}_r^\top\mathbf{g}_f$, where a QR decomposition gives the orthonormal columns of $\mathbf{Q}_r$, which span the per-sample retain gradients. The update is
$$\mathbf{w} \leftarrow \mathbf{w} - \eta\big(\alpha\,\mathbf{g}_r - (1-\alpha)\,\mathbf{g}_f^\perp\big).$$
- The GU [26] baseline in its Euclidean case uses a learning rate η_{GU} and the same basis $\mathbf{Q}_r$ as above. The update is
$$\mathbf{w} \leftarrow \mathbf{w} + \eta_{\mathrm{GU}}\big((\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^\top)\,\mathbf{g}_f - \mathbf{Q}_r\mathbf{Q}_r^\top\mathbf{g}_r\big).$$

These baselines modify gradient directions at the final update. ROSU imposes its ideal first-order retain-neutral constraint on the inner surrogate perturbation. Hyperparameters are selected under the protocols of Appendix B.4. Some results are taken from UAM [8].

B.2 TOFU and WMDP protocols

For TOFU, we follow the standardized OpenUnlearning pipeline for implementation and evaluation [11], using `Llama-3.2-1B-Instruct` as the target model. This framework standardizes tokenization, collation, generation-time evaluation, and the aggregation of memorization, privacy, and utility metrics. The compared methods are listed in Section 5.2.

For WMDP, we use the representation-level instantiation given in Algorithm 2. The target layer is $\ell = 7$, and the trainable coordinate set $\mathcal{T}$ indexes scalar parameters in the MLP output (`down_proj`) matrices in transformer layers 5–7, with all other parameters frozen. The forget and retain losses are the representation-level objectives of Algorithm 2. The forget corpus is WMDP, and the retain corpus is WikiText. Evaluation uses a four-choice question-answering benchmark, where random selection corresponds to an expected accuracy of 0.25 on both WMDP-Bio and WMDP-Cyber. The restricted coordinate set and representation-level configuration follow UAM [8]. Some baseline results are taken directly from UAM [8].

B.3 Evaluation metrics

For vision, let $\mathbf{x}$ be an input, y its class label, and $\mathbf{f}_{\mathbf{w}}(\mathbf{x})$ the vector of class scores. Define $\hat{y}_{\mathbf{w}}(\mathbf{x}) := \arg\max_j \mathbf{f}_{\mathbf{w}}(\mathbf{x})_j$, with class index j and ties resolved by a fixed rule. The nonempty sets $\mathcal{D}_r^{\mathrm{eval}}$, $\mathcal{D}_f^{\mathrm{eval}}$, and $\mathcal{D}_{\mathrm{test}}$ contain retain, forget, and test evaluation examples, respectively. In class-wise forgetting, $\mathcal{D}_{\mathrm{test}}$ excludes the test examples of the forgotten class. With $\mathbb{I}\{\cdot\}$ equal to 1 when its condition holds and 0 otherwise, we evaluate retain accuracy (RA), forget accuracy (FA), and test accuracy (TA) as
$$\mathrm{RA}(\mathbf{w}) = \frac{1}{|\mathcal{D}_r^{\mathrm{eval}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_r^{\mathrm{eval}}} \mathbb{I}\{\hat{y}_{\mathbf{w}}(\mathbf{x}) = y\},$$
$$\mathrm{FA}(\mathbf{w}) = \frac{1}{|\mathcal{D}_f^{\mathrm{eval}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_f^{\mathrm{eval}}} \mathbb{I}\{\hat{y}_{\mathbf{w}}(\mathbf{x}) = y\},$$
$$\mathrm{TA}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\mathrm{test}}} \mathbb{I}\{\hat{y}_{\mathbf{w}}(\mathbf{x}) = y\}.$$
Let A_{Retrain} denote metric A evaluated on the retrained reference. We summarize the aggregate accuracy discrepancy by
$$\Delta\mathrm{Acc}(\mathbf{w}) := \sum_{A\in\{\mathrm{RA},\mathrm{FA},\mathrm{TA}\}} \big|A_{\mathrm{Retrain}} - A(\mathbf{w})\big|,$$

where lower values are better [8]. In the vision tables, accuracies and MIA-Eff. are reported as percentages, and $\Delta\mathrm{Acc}$ is reported in percentage points (i.e., the ratio-valued definitions are multiplied by 100). Table values of $\Delta\mathrm{Acc}$ are computed from the mean RA, FA, and TA, taking the absolute differences after averaging over targets.

The membership-inference efficacy metric reports the fraction of forget examples classified as non-members [8, 9]. Let $\mathcal{A}_{\mathrm{mia}}(\mathbf{x}; \mathbf{w}) \in \{0, 1\}$ denote the membership predictor, where 1 indicates "member". It is the confidence-based attack [69, 70] used by UAM, implemented as a support-vector classifier on output probabilities. We define

$$\text{MIA-Eff.}(\mathbf{w}) := \frac{1}{|\mathcal{D}_f^{\mathrm{eval}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_f^{\mathrm{eval}}} \mathbb{I}\big\{\mathcal{A}_{\mathrm{mia}}(\mathbf{x}; \mathbf{w}) = 0\big\}.$$

Higher values indicate more forget examples classified as non-members. For random forgetting, agreement with the Retrain value is the evaluation target.

For TOFU, we adopt OpenUnlearning's metric categories and aggregation [11]. HM denotes the harmonic mean of nonnegative arguments, with value zero whenever any argument is zero. Memorization measures forgetting through four knowledge metrics from OpenUnlearning [11]. They are extraction strength (ES), exact memorization (EM) [71], paraphrased probability (Para. Prob., denoted $\mathrm{ParaProb}$), and truth ratio (TR), each inverted so that a higher value means less forget-set knowledge. Their harmonic mean is

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

Privacy measures agreement with the retain model under the four membership inference attacks implemented in OpenUnlearning [11], LOSS [70], ZLib [72], MinK [73], and MinK++ [74]. For each attack a, OpenUnlearning computes a score $s_a \in [0, 1]$ that quantifies how closely the unlearned model's behavior under that attack aligns with the retain model's, higher meaning closer. Their harmonic mean is

$$\mathrm{Priv} = \mathrm{HM}\big(s_{\mathrm{LOSS}},\, s_{\mathrm{ZLib}},\, s_{\mathrm{MinK}},\, s_{\mathrm{MinK++}}\big).$$

Utility measures retention through TOFU's model utility (MU) [10] and through fluency. MU is the harmonic mean of probability, ROUGE [75], and truth ratio on the retain, real-author, and world-fact splits. Fluency is a classifier score that penalizes gibberish in answers to forget-related prompts. The harmonic mean of MU and fluency is

$$\mathrm{Util} = \mathrm{HM}(\mathrm{MU},\, \mathrm{Fluency}).$$

Metric preprocessing and normalization follow the OpenUnlearning evaluation pipeline used for these experiments. The headline score and the model-selection score are

$$\mathrm{Agg} = \mathrm{HM}(\mathrm{Mem},\, \mathrm{Priv},\, \mathrm{Util}), \qquad \mathrm{Agg}_{\mathrm{MU}} = \mathrm{HM}(\mathrm{Mem},\, \mathrm{Util}),$$

so a high Agg requires forgetting, agreement with the retain model, and retained utility at the same time. The detailed computation of each metric follows OpenUnlearning.

For WMDP, Bio and Cyber are four-choice question-answering accuracies on the biosecurity and cybersecurity splits of WMDP [20], and MMLU is four-choice accuracy on general academic knowledge. Higher MMLU indicates better general utility, while Bio and Cyber closer to the chance level 0.25 indicate more complete forgetting without oversuppression. We summarize performance using

$$\Delta\mathrm{Score}(\mathbf{w}) := \big|\mathrm{MMLU}(\mathbf{w}) - \mathrm{MMLU}(\mathbf{w}_0)\big| + \big|\mathrm{Bio}(\mathbf{w}) - 0.25\big| + \big|\mathrm{Cyber}(\mathbf{w}) - 0.25\big|,$$

where $\mathbf{w}_0$ denotes the pre-unlearning weights. The first term measures the absolute change in general-knowledge accuracy, and the remaining terms measure deviation from chance-level forgetting on hazardous domains. Lower $\Delta\mathrm{Score}$ is better. We additionally report raw Bio and Cyber accuracies. Under this metric, the ideal corresponds to values close to 0.25 rather than arbitrarily low accuracy. The OpenUnlearning evaluation [11] penalizes both under- and over-unlearning, and $\Delta\mathrm{Score}$ likewise penalizes deviation in both directions.

B.4 Model selection and amplification tuning

For vision, we use the epoch budgets of the UAM setup [8]. Hyperparameters are tuned per method following UAM [8], and for each method and setting we report the best configuration and checkpoint found under this protocol.

Algorithm 2 Representation-Level ROSU

Input: frozen pre-unlearning weights $\mathbf{w}_0$, trainable coordinate set $\mathcal{T}$, current trainable vector $\boldsymbol{\theta} = \mathbf{w}_{\mathcal{T}}$, nonempty forget mini-batch $\mathcal{B}_f$ and retain mini-batch $\mathcal{B}_r$

Parameters: target layer ℓ, learning rate $\eta > 0$, perturbation radius $\rho > 0$, amplification coefficient $\beta \geq 0$, bootstrap standard deviation $\sigma_b \geq 0$, stabilizer $\tau > 0$, threshold $\varepsilon_q > 0$

Output: updated trainable vector $\boldsymbol{\theta}$ and the full model $\mathbf{w}(\boldsymbol{\theta})$

Let $\mathbf{w}(\boldsymbol{\theta})$ be the full model obtained by inserting $\boldsymbol{\theta}$ into coordinates $\mathcal{T}$ and keeping all other coordinates fixed at $\mathbf{w}_0$, and let $\mathbf{h}^{(\ell)}_{\mathbf{w}}(\mathbf{x})$ be the layer-ℓ representation of $\mathbf{x}$

Sample $\boldsymbol{\sigma} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$ with the shape of $\mathbf{h}^{(\ell)}$ ▷ Mini-batch shared bootstrap noise

$L_f(\boldsymbol{\theta}) \leftarrow \frac{1}{2|\mathcal{B}_f|} \sum_{\mathbf{x} \in \mathcal{B}_f} \left\| \mathbf{h}^{(\ell)}_{\mathbf{w}(\boldsymbol{\theta})}(\mathbf{x}) + \boldsymbol{\sigma} - \mathbf{h}^{(\ell)}_{\mathbf{w}_0}(\mathbf{x}) \right\|_2^2$

$L_r(\boldsymbol{\theta}) \leftarrow \frac{1}{2|\mathcal{B}_r|} \sum_{\mathbf{x} \in \mathcal{B}_r} \left\| \mathbf{h}^{(\ell)}_{\mathbf{w}(\boldsymbol{\theta})}(\mathbf{x}) - \mathbf{h}^{(\ell)}_{\mathbf{w}_0}(\mathbf{x}) \right\|_2^2$

$\mathbf{g}_f \leftarrow \nabla_{\boldsymbol{\theta}} L_f(\boldsymbol{\theta}), \quad \mathbf{g}_r \leftarrow \nabla_{\boldsymbol{\theta}} L_r(\boldsymbol{\theta})$

$\mathbf{q}_\tau \leftarrow \mathbf{g}_f - \frac{\mathbf{g}_f^\top \mathbf{g}_r}{\|\mathbf{g}_r\|_2^2 + \tau} \mathbf{g}_r$

if $\|\mathbf{q}_\tau\|_2 \leq \varepsilon_q$ **then**

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \, \mathbf{g}_r$ ▷ Degenerate fallback, Section 4.3

else

  $\mathbf{u}_\tau \leftarrow \frac{\mathbf{q}_\tau}{\|\mathbf{q}_\tau\|_2}, \quad \boldsymbol{\delta}_\tau \leftarrow \rho \, \mathbf{u}_\tau, \quad \alpha_\tau \leftarrow \frac{\rho}{\|\mathbf{q}_\tau\|_2}$

  $\widetilde{\mathbf{g}}_{r,\tau} \leftarrow \nabla_{\mathbf{z}} L_r(\mathbf{z})|_{\mathbf{z} = \boldsymbol{\theta} + \boldsymbol{\delta}_\tau}$ ▷ Surrogate retain gradient in the trainable subspace

  $\mathbf{v} \leftarrow \widetilde{\mathbf{g}}_{r,\tau} + \alpha_\tau \left(\widetilde{\mathbf{g}}_{r,\tau} - \frac{\mathbf{g}_r^\top \widetilde{\mathbf{g}}_{r,\tau}}{\|\mathbf{g}_r\|_2^2 + \tau} \mathbf{g}_r - (\mathbf{u}_\tau^\top \widetilde{\mathbf{g}}_{r,\tau}) \, \mathbf{u}_\tau \right)$

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \, \boldsymbol{\delta}_\tau - \eta \, \mathbf{v}$

end if

For TOFU, we follow OpenUnlearning's benchmarking protocol [11]. Each method is tuned over its own search space, and model selection uses $\mathrm{Agg}_{\mathrm{MU}} = \mathrm{HM}(\mathrm{Mem}, \mathrm{Util})$.

For WMDP, we follow the training protocol of UAM and report the full (MMLU, Bio, Cyber) triple together with the composite $\Delta\mathrm{Score}$ defined in Appendix B.3.

The amplification schedule of Section 4 and its level are selected together with ρ and η under the selection rules above.

Raising the amplification level scales the first-order forget gain along the retain-neutral direction linearly (Section 4) and the second-order retain bound quadratically (Appendix D.1). Two ablations isolate the two update components. Setting $\beta = 0$ gives the $-\eta\mathbf{v}$-only update, while the $\beta\boldsymbol{\delta}_\tau$-only variant keeps only retain-neutral amplification.

B.5 Implementation details and hyperparameter search

The vision configuration uses SGD with momentum 0.9 and weight decay 5×10^{-4}, with default forget and retain batch sizes of 128 unless stated otherwise. ROSU passes $\mathbf{v}$ and the fallback gradient to this optimizer and adds $\beta\boldsymbol{\delta}_\tau$ to the weights directly. OrthoGrad and GU apply the updates of Appendix B.1 without momentum or weight decay.

For TOFU, we use `Llama-3.2-1B-Instruct` within the released OpenUnlearning runners [11]. Each method retains its benchmark implementation and search space, while checkpoint selection follows Appendix B.4.

For WMDP, we use `Zephyr-7B-`β with the representation-level setup and training protocol of UAM [8]. The bootstrap configuration [8] uses noise in Algorithm 2 sampled once per mini-batch from $\mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$ with $\sigma_b = 0.01$, which is fixed throughout.

Across all benchmarks, ROSU uses each pipeline's standard training setup. We search ρ in $[0.001, 2.5]$ and the learning rate in $[10^{-5}, 10^{-1}]$, and choose c_t so that the largest step length $c_t \eta_t$ lies in $[10^{-4}, 1]$. These hyperparameters are selected under the protocols of Appendix B.4. Appendix C.3 reports a separate sweep over ρ.

In Algorithms 1–2, β denotes the current-step amplification coefficient β_t of Section 4. In the radius sweep of Appendix C.3 it is switched off after the early epochs.

The stabilizer $\tau = 10^{-8}$ and degeneracy threshold $\varepsilon_q = 10^{-6}$ are fixed in all runs. All experiments are run on NVIDIA RTX A6000 GPUs with 48 GB of memory.

B.6 Algorithm variants

Here we describe the No-Jacobian variant and the representation-level update for WMDP.

The No-Jacobian variant uses the regularized surrogate construction, amplification, and fallback of Algorithm 1. On nonfallback steps, it uses $\mathbf{v} = \widetilde{\mathbf{g}}_{r,\tau}$, without the Jacobian correction.

Algorithm 2 extends ROSU to the representation-level WMDP setting of UAM [8]. In Algorithm 2, both L_f and L_r are squared-error losses between current and frozen pre-unlearning representations at a target layer ℓ. Here $\mathbf{h}_{\mathbf{w}}^{(\ell)}(\mathbf{x})$ denotes the layer-ℓ representation of input $\mathbf{x}$ under weights $\mathbf{w}$. Let $\mathbf{w}_0$ denote the frozen pre-unlearning weights, $\mathcal{T}$ the trainable coordinate set, and $\boldsymbol{\theta} = \mathbf{w}_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$ the corresponding trainable vector. We write $\mathbf{w}(\boldsymbol{\theta})$ for the full model obtained by inserting $\boldsymbol{\theta}$ into $\mathcal{T}$ while keeping all other parameters fixed at $\mathbf{w}_0$. The retain-orthogonal construction and update are then performed in the $\boldsymbol{\theta}$-space. In Algorithm 2, $\mathbf{z} \in \mathbb{R}^{|\mathcal{T}|}$ is the loss argument used to evaluate the surrogate retain gradient.

The losses in Algorithm 2 average over examples and sum squared representation coordinates, with a factor of $\frac{1}{2}$. Here $\boldsymbol{\sigma}$ has the same representation shape as $\mathbf{h}^{(\ell)}$, and the identity in its covariance acts on representation coordinates. The identity in the projector $\mathbf{I} - \mathbf{P}_r^{(\tau)}$ of Section 4.3 instead acts on the $|\mathcal{T}|$ trainable coordinates. The bootstrap sample is held fixed during the update. At $\mathbf{w}_0$, $\sigma_b = 0$ makes both loss gradients in Algorithm 2 zero, so its fallback step leaves the parameters unchanged.

C Additional experimental results

Appendix C.1 gives the Tiny-ImageNet results, and Appendix C.2 isolates the update components. Amplification tuning is described in Appendix B.4. Appendices C.3, C.4, and C.5 examine the perturbation radius and numerical safeguards, retain mini-batch size, and sampled retain subspaces.

C.1 Additional vision benchmark results

Table 6: **Unlearning on Tiny-ImageNet.** Same notation as Table 1. FF is not reported.

Method	RA	FA	TA	ΔAcc.↓	MIA-Eff.	Time (min)
Class-wise forgetting						
Retrain	$99.98_{\pm 0.00}$	$0.00_{\pm 0.00}$	$61.89_{\pm 0.18}$	0.00	$100.00_{\pm 0.00}$	189.8
FT	$98.54_{\pm 2.03}$ (1.44)	$26.88_{\pm 13.27}$ (26.88)	$57.52_{\pm 1.95}$ (4.37)	32.69	$97.37_{\pm 2.36}$ (2.63)	9.41
NG	$92.84_{\pm 3.34}$ (7.14)	$0.23_{\pm 0.32}$ (0.23)	$52.28_{\pm 2.23}$ (9.61)	16.98	$99.92_{\pm 0.17}$ (0.08)	9.43
IU	$95.83_{\pm 3.51}$ (4.15)	$1.20_{\pm 2.05}$ (1.20)	$55.03_{\pm 2.64}$ (6.86)	12.21	$99.84_{\pm 0.31}$ (0.16)	1.27
ℓ_1-sparse	$98.19_{\pm 0.16}$ (1.79)	$0.00_{\pm 0.00}$ (0.00)	$59.66_{\pm 0.19}$ (2.23)	4.02	$100.00_{\pm 0.00}$ (0.00)	10.47
GU	$99.54_{\pm 0.22}$ (0.44)	$0.00_{\pm 0.00}$ (0.00)	$58.59_{\pm 0.62}$ (3.30)	3.74	$100.00_{\pm 0.00}$ (0.00)	17.12
OrthoGrad	$99.67_{\pm 0.18}$ (0.31)	$0.13_{\pm 0.31}$ (0.13)	$58.96_{\pm 0.64}$ (2.93)	3.37	$100.00_{\pm 0.00}$ (0.00)	17.64
PCGrad	$99.84_{\pm 0.09}$ (0.14)	$0.08_{\pm 0.23}$ (0.08)	$59.59_{\pm 0.71}$ (2.30)	2.52	$100.00_{\pm 0.00}$ (0.00)	9.14
UAM	$99.97_{\pm 0.02}$ (0.01)	$0.23_{\pm 0.26}$ (0.23)	$60.86_{\pm 0.64}$ (1.03)	<u>1.27</u>	$99.97_{\pm 0.08}$ (0.03)	12.85
ROSU	$99.97_{\pm 0.00}$ (0.01)	$0.00_{\pm 0.00}$ (0.00)	$61.89_{\pm 0.22}$ (0.00)	**0.01**	$100.00_{\pm 0.00}$ (0.00)	14.29
Random forgetting						
Retrain	$99.98_{\pm 0.00}$	$60.25_{\pm 0.19}$	$60.72_{\pm 0.22}$	0.00	$64.50_{\pm 0.60}$	169.5
FT	$99.98_{\pm 0.00}$ (0.00)	$99.96_{\pm 0.01}$ (39.71)	$61.43_{\pm 0.07}$ (0.71)	40.42	$3.89_{\pm 1.21}$ (60.61)	8.48
NG	$0.60_{\pm 0.08}$ (99.38)	$0.57_{\pm 0.06}$ (59.68)	$0.61_{\pm 0.09}$ (60.11)	219.17	$66.55_{\pm 46.56}$ (2.05)	8.47
IU	$89.83_{\pm 8.53}$ (10.15)	$89.25_{\pm 8.56}$ (29.00)	$50.84_{\pm 4.81}$ (9.88)	49.03	$14.42_{\pm 6.60}$ (50.08)	1.25
ℓ_1-sparse	$98.86_{\pm 0.05}$ (1.12)	$59.66_{\pm 0.55}$ (0.59)	$58.16_{\pm 0.23}$ (2.56)	<u>4.27</u>	$54.54_{\pm 0.62}$ (9.96)	9.11
GU	$99.98_{\pm 0.00}$ (0.00)	$99.98_{\pm 0.01}$ (39.73)	$61.72_{\pm 0.03}$ (1.00)	40.73	$1.42_{\pm 0.13}$ (63.08)	15.29
OrthoGrad	$98.63_{\pm 0.09}$ (1.35)	$65.39_{\pm 1.73}$ (5.14)	$52.54_{\pm 0.35}$ (8.18)	14.67	$41.19_{\pm 1.65}$ (23.31)	15.60
PCGrad	$88.39_{\pm 0.69}$ (11.59)	$65.44_{\pm 0.80}$ (5.19)	$50.93_{\pm 0.37}$ (9.79)	26.57	$37.20_{\pm 0.80}$ (27.30)	7.96
UAM	$97.61_{\pm 0.89}$ (2.37)	$69.88_{\pm 1.06}$ (9.63)	$56.37_{\pm 0.38}$ (4.35)	16.35	$46.63_{\pm 1.06}$ (17.87)	11.69
ROSU	$98.70_{\pm 0.03}$ (1.28)	$60.72_{\pm 0.25}$ (0.47)	$58.73_{\pm 0.27}$ (1.99)	**3.74**	$47.31_{\pm 0.48}$ (17.19)	13.16

C.2 Component ablations

The component ablations in Tables 7, 8, and 9 remove one part of the update at a time. The No-Jacobian variant uses $\mathbf{v} = \widetilde{\mathbf{g}}_{r,\tau}$ without the Jacobian correction in (4). The $\beta\boldsymbol{\delta}_\tau$-only variant removes the retain step $-\eta\mathbf{v}$, and the $-\eta\mathbf{v}$-only variant sets $\beta = 0$. Tables 10 and 11 repeat the ablation on TOFU and WMDP. In Tables 7–11, $\boldsymbol{\delta}$ abbreviates the implemented $\boldsymbol{\delta}_\tau$. Figure 3 in Section 5.3 reports the derivative variants.

Table 7: **Component ablation on CIFAR-10.** Check marks show which update terms each variant keeps, and bold marks the lowest ΔAcc in each protocol.

Method	$\beta\boldsymbol{\delta}$	$\eta\mathbf{v}$	ΔAcc.↓	RA	FA	TA	MIA-Eff.
Class-wise forgetting							
ROSU	✓	✓	0.17	$99.96_{\pm 0.01}$	$0.00_{\pm 0.00}$	$95.01_{\pm 0.53}$	$100.00_{\pm 0.00}$
No-Jacobian	✓	✓*	**0.15**	$99.97_{\pm 0.01}$	$0.00_{\pm 0.00}$	$95.02_{\pm 0.51}$	$100.00_{\pm 0.00}$
$\beta\boldsymbol{\delta}$-only	✓	✗	26.22	$86.96_{\pm 1.71}$	$0.00_{\pm 0.00}$	$81.96_{\pm 1.67}$	$100.00_{\pm 0.01}$
$-\eta\mathbf{v}$-only	✗	✓	99.96	$100.00_{\pm 0.00}$	$99.91_{\pm 0.07}$	$95.11_{\pm 0.39}$	$3.34_{\pm 2.26}$
Random forgetting							
ROSU	✓	✓	**0.26**	$99.84_{\pm 0.03}$	$95.13_{\pm 0.23}$	$94.29_{\pm 0.09}$	$7.88_{\pm 0.16}$
No-Jacobian	✓	✓*	1.03	$99.88_{\pm 0.02}$	$96.01_{\pm 0.31}$	$94.40_{\pm 0.16}$	$6.78_{\pm 0.18}$
$\beta\boldsymbol{\delta}$-only	✓	✗	285.64	$1.41_{\pm 0.37}$	$0.32_{\pm 0.02}$	$2.11_{\pm 0.65}$	$0.31_{\pm 0.05}$
$-\eta\mathbf{v}$-only	✗	✓	4.93	$99.91_{\pm 0.01}$	$99.66_{\pm 0.03}$	$94.68_{\pm 0.19}$	$3.83_{\pm 0.20}$

* No-Jacobian: $\eta\mathbf{v}$ present but without Jacobian correction ($\mathbf{v} = \widetilde{\mathbf{g}}_{r,\tau}$).

Table 8: **Component ablation on CIFAR-100.** Same notation as Table 7.

Method	$\beta\boldsymbol{\delta}$	$\eta\mathbf{v}$	ΔAcc.↓	RA	FA	TA	MIA-Eff.
Class-wise forgetting							
ROSU	✓	✓	**0.14**	$99.98_{\pm 0.00}$	$0.12_{\pm 0.26}$	$77.78_{\pm 0.13}$	$100.00_{\pm 0.00}$
No-Jacobian	✓	✓*	1.03	$99.98_{\pm 0.00}$	$0.14_{\pm 0.31}$	$76.91_{\pm 0.22}$	$99.88_{\pm 0.26}$
$\beta\boldsymbol{\delta}$-only	✓	✗	16.46	$93.49_{\pm 0.91}$	$1.16_{\pm 2.37}$	$68.99_{\pm 0.69}$	$99.34_{\pm 1.11}$
$-\eta\mathbf{v}$-only	✗	✓	99.96	$99.98_{\pm 0.00}$	$99.94_{\pm 0.13}$	$77.82_{\pm 0.17}$	$14.50_{\pm 4.05}$
Random forgetting							
ROSU	✓	✓	**0.67**	$99.78_{\pm 0.00}$	$77.98_{\pm 0.08}$	$76.85_{\pm 0.18}$	$41.37_{\pm 0.75}$
No-Jacobian	✓	✓*	12.86	$99.96_{\pm 0.01}$	$90.01_{\pm 0.28}$	$77.19_{\pm 0.13}$	$33.09_{\pm 0.69}$
$\beta\boldsymbol{\delta}$-only	✓	✗	251.97	$0.84_{\pm 0.13}$	$0.63_{\pm 0.05}$	$0.90_{\pm 0.11}$	$33.53_{\pm 46.57}$
$-\eta\mathbf{v}$-only	✗	✓	8.75	$99.74_{\pm 0.01}$	$86.17_{\pm 0.19}$	$76.70_{\pm 0.16}$	$36.27_{\pm 0.48}$

Table 9: **Component ablation on Tiny-ImageNet.** Same notation as Table 7.

Method	$\beta\boldsymbol{\delta}$	$\eta\mathbf{v}$	ΔAcc.↓	RA	FA	TA	MIA-Eff.
Class-wise forgetting							
ROSU	✓	✓	**0.01**	$99.97_{\pm 0.00}$	$0.00_{\pm 0.00}$	$61.89_{\pm 0.22}$	$100.00_{\pm 0.00}$
No-Jacobian	✓	✓*	**0.01**	$99.98_{\pm 0.00}$	$0.00_{\pm 0.00}$	$61.88_{\pm 0.20}$	$100.00_{\pm 0.00}$
$\beta\boldsymbol{\delta}$-only	✓	✗	148.38	$8.82_{\pm 23.98}$	$0.00_{\pm 0.00}$	$4.67_{\pm 11.63}$	$50.00_{\pm 50.00}$
$-\eta\mathbf{v}$-only	✗	✓	100.29	$99.98_{\pm 0.00}$	$99.97_{\pm 0.08}$	$61.57_{\pm 0.14}$	$0.83_{\pm 0.61}$
Random forgetting							
ROSU	✓	✓	**3.74**	$98.70_{\pm 0.03}$	$60.72_{\pm 0.25}$	$58.73_{\pm 0.27}$	$47.31_{\pm 0.48}$
No-Jacobian	✓	✓*	8.16	$99.19_{\pm 0.07}$	$65.88_{\pm 0.03}$	$58.98_{\pm 0.45}$	$44.49_{\pm 0.63}$
$\beta\boldsymbol{\delta}$-only	✓	✗	219.45	$0.50_{\pm 0.00}$	$0.50_{\pm 0.00}$	$0.50_{\pm 0.00}$	$66.51_{\pm 46.68}$
$-\eta\mathbf{v}$-only	✗	✓	31.69	$99.08_{\pm 0.00}$	$90.24_{\pm 0.10}$	$59.92_{\pm 0.34}$	$23.14_{\pm 0.36}$

Table 10: **Component ablation on TOFU.** Same variants as Table 7, and bold marks the best Agg.

TOFU						
Method	$\beta\boldsymbol{\delta}$	$\eta\mathbf{v}$	**Agg.↑**	**Mem.↑**	**Priv.↑**	**Util.↑**
ROSU	✓	✓	**0.58**	0.38	0.74	0.82
No-Jacobian	✓	✓*	0.53	0.53	0.54	0.53
$\beta\boldsymbol{\delta}$*-only*	✓	✗	0.00	0.98	0.48	0.00
$-\eta\mathbf{v}$*-only*	✗	✓	0.00	0.00	0.43	0.88

Table 11: **Component ablation on WMDP.** Same variants as Table 7. As in ΔScore, Bio and Cyber are judged by their distance from the chance level of 0.25.

WMDP					
Method	$\beta\boldsymbol{\delta}$	$\eta\mathbf{v}$	**MMLU↑**	**Bio**	**Cyber**
ROSU	✓	✓	0.570	0.289	0.246
No-Jacobian	✓	✓*	0.577	0.420	0.261
$\beta\boldsymbol{\delta}$*-only*	✓	✗	0.255	0.266	0.246
$-\eta\mathbf{v}$*-only*	✗	✓	0.252	0.251	0.266

C.3 Perturbation radius and numerical safeguards

Table 5 reports ROSU's sensitivity to the perturbation radius ρ on TOFU. Table 4 reports the corresponding sweep on CIFAR-100 and CIFAR-10 random forgetting, with checkpoints selected for each radius. In this vision sweep, amplification is active only during the early epochs and switched off afterwards. Within each dataset, all runs use the same learning rate and c_t. Means and standard deviations are over three seeds. ΔAcc remains close to the main-table values for every radius. Figure 4 audits the numerical scales relevant to the stabilizer τ and degeneracy threshold ε_q. We denote the recorded projected-gradient norm by n_q. In the audited records, $n_q > \varepsilon_q$ throughout. The retain-gradient squared norm exceeds τ on the vision benchmarks and TOFU. On WMDP, some records fall below τ (Figure 4). This is consistent with a representation-level retain loss initially near zero. The stabilizer keeps the retain projector defined on these records. On the vision benchmarks, random forgetting yields smaller recorded projected-gradient norms than class-wise forgetting. This is consistent with higher forget–retain coupling.

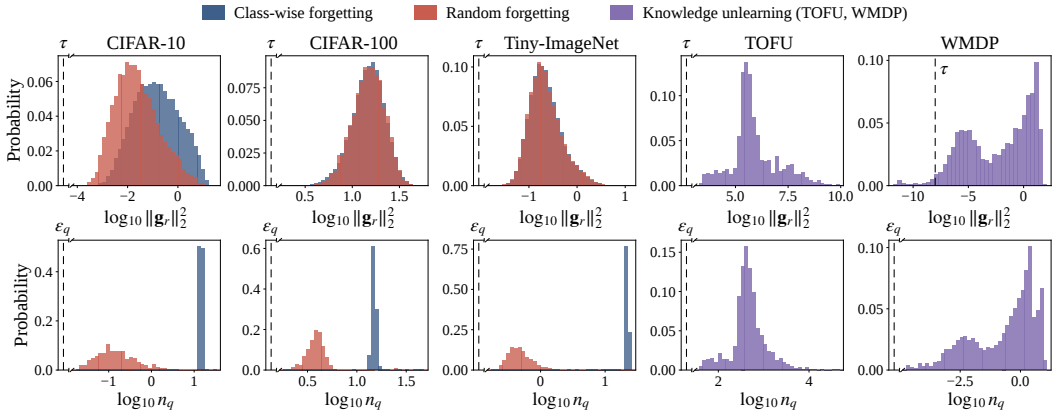


Figure 4: **Gradient-norm audit.** Distributions of $\|\mathbf{g}_r\|_2^2$ in the upper row and of the projected-gradient norm n_q in the lower row. The vision panels overlay class-wise and random forgetting. Dashed lines mark $\tau = 10^{-8}$ in the upper row and $\varepsilon_q = 10^{-6}$ in the lower row.

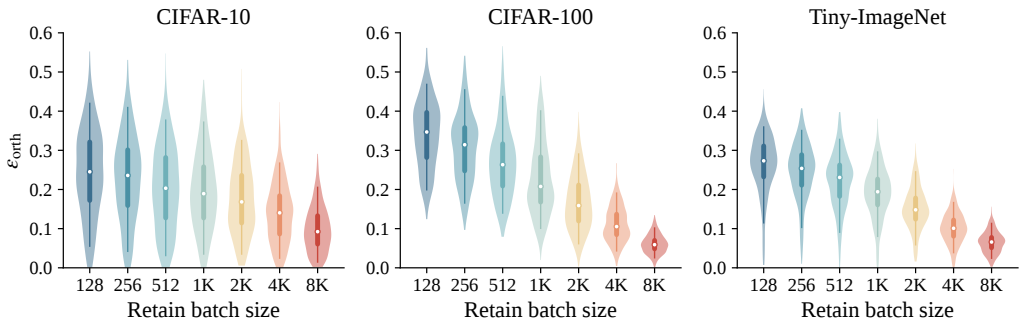


Figure 5: **Stochastic orthogonality audit.** Full-retain orthogonality error $\varepsilon_{\text{orth}}$ for each retain batch size at the pre-unlearning checkpoint under random forgetting. White dots mark medians, and thick and thin sticks mark the interquartile and 5–95% ranges.

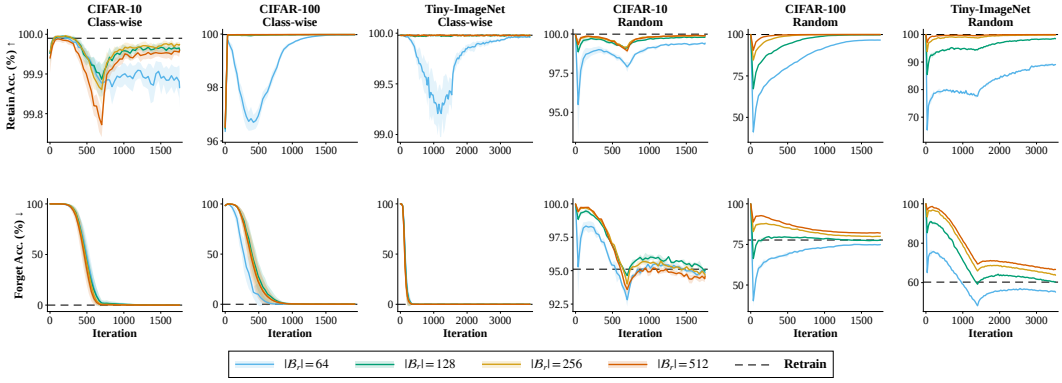


Figure 6: **Sensitivity to the retain mini-batch size.** Retain and forget accuracy during unlearning for $|\mathcal{B}_r| \in \{64, 128, 256, 512\}$. Curves and bands are the mean and one standard deviation over ten forgotten classes or three seeds. Dashed lines mark the Retrain values of Tables 1, 2, and 6.

C.4 Retain mini-batch size and orthogonality

ROSU estimates the retain direction from a mini-batch and regularizes the resulting projection. We examine the full-retain residual of this estimate and its effect on training. Figure 5 measures the residual at the pre-unlearning checkpoint under random forgetting, with the model held fixed so that only the sampling varies. Figure 6 tracks retain–forget trajectories across retain batch sizes under both class-wise and random forgetting.

For the pre-unlearning audit, we evaluate all gradients at the pre-unlearning checkpoint, with $|\mathcal{B}_r| \in \{128, 256, 512, 1024, 2048, 4096, 8192\}$. Let $\mathbf{g}_r^{\text{full}} := \nabla_{\mathbf{w}} L_r(\mathbf{w}; \mathcal{D}_r)$, $\mathbf{g}_f^{\text{mini}} := \nabla_{\mathbf{w}} L_f(\mathbf{w}; \mathcal{B}_f)$, and $\mathbf{g}_r^{\text{mini}} := \nabla_{\mathbf{w}} L_r(\mathbf{w}; \mathcal{B}_r)$. We sample 500 forget/retain mini-batch pairs per cell, with forget batch size 128. The diagnostic implementation uses the same stabilizer $\tau = 10^{-8}$ as training and computes

$$\mathbf{q}_\tau^{\text{mini}} := \mathbf{g}_f^{\text{mini}} - \frac{(\mathbf{g}_f^{\text{mini}})^\top \mathbf{g}_r^{\text{mini}}}{\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau} \mathbf{g}_r^{\text{mini}}.$$

The direction diagnostic uses samples with $\|\mathbf{q}_\tau^{\text{mini}}\|_2 > \varepsilon_q = 10^{-6}$. For these samples, define $\mathbf{u}_\tau^{\text{mini}} := \frac{\mathbf{q}_\tau^{\text{mini}}}{\|\mathbf{q}_\tau^{\text{mini}}\|_2}$.

The full-retain residual is bounded by a gradient-estimation term and a regularization term (Appendix D.3),

$$\left|(\mathbf{g}_r^{\text{full}})^\top \mathbf{u}_\tau^{\text{mini}}\right| \le \|\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}}\|_2 + \frac{\tau \left|(\mathbf{g}_f^{\text{mini}})^\top \mathbf{g}_r^{\text{mini}}\right|}{(\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau)\|\mathbf{q}_\tau^{\text{mini}}\|_2}. \tag{7}$$

When $\mathbf{g}_r^{\text{full}} \neq \mathbf{0}$, Figure 5 reports

$$\begin{aligned}\varepsilon_{\text{orth}} &:= \frac{\left|(\mathbf{g}_r^{\text{full}})^\top \mathbf{u}_\tau^{\text{mini}}\right|}{\|\mathbf{g}_r^{\text{full}}\|_2} \\ &= \left|\cos\big(\angle(\mathbf{g}_r^{\text{full}}, \mathbf{u}_\tau^{\text{mini}})\big)\right| \\ &\leq \frac{\|\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}}\|_2 + \left|(\mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}}\right|}{\|\mathbf{g}_r^{\text{full}}\|_2}.\end{aligned}$$

Here 0 means orthogonality to the full retain gradient and 1 means collinearity, either parallel or antiparallel. This is a relative error, scaled by the full retain-gradient norm.

The distributions in Figure 5 concentrate toward zero as $|\mathcal{B}_r|$ increases across all three datasets, and CIFAR-100 random forgetting shows the largest visible shift.

Figure 6 shows how the same sampling noise affects training. In class-wise forgetting, forgetting succeeds with every retain batch size, and forget accuracy drops to near zero on every dataset. The main batch-size effect is on retain stability. Very small retain batches can introduce transient retain-accuracy drops on CIFAR-100 and Tiny-ImageNet, and can leave a small residual retain gap on CIFAR-10. In random forgetting, the effect is stronger and more systematic. Larger retain batches give a more stable retain-gradient estimate, so retain accuracy is higher and optimization is smoother. With smaller retain batches, forget accuracy is often lower and retention is weaker. This is consistent with the coupling measured in Figure 1a. When the coupling between the forget and retain gradients is higher, the quality of the retain-gradient estimate matters more.

C.5 Extension with sampled rank-k retain protection

The rank-one construction of Section 3.3 protects a single retain direction. It extends to a protected subspace as follows. Let $\mathbf{U} \in \mathbb{R}^{p\times k'}$ have $k' \leq k$ orthonormal columns and set $\mathbf{P}_\mathbf{U} := \mathbf{U}\mathbf{U}^\top$ and $\mathbf{q}_\mathbf{U} := (\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{g}_f$, where k is the number of sampled directions and k' is their retained rank. The protected-subspace inner problem is

$$\max_{\boldsymbol{\delta}} \ \mathbf{g}_f^\top \boldsymbol{\delta} \quad \text{s.t.} \quad \|\boldsymbol{\delta}\|_2 \leq \rho, \quad \mathbf{U}^\top \boldsymbol{\delta} = \mathbf{0}.$$

On its feasible set $\mathbf{g}_f^\top \boldsymbol{\delta} = \mathbf{q}_\mathbf{U}^\top \boldsymbol{\delta}$, so the Cauchy–Schwarz argument of Appendix D.1 applies without change. If $\mathbf{q}_\mathbf{U} \neq \mathbf{0}$, the unique maximizer is $\boldsymbol{\delta}_\mathbf{U} = \rho \frac{\mathbf{q}_\mathbf{U}}{\|\mathbf{q}_\mathbf{U}\|_2}$ with value $\rho\|\mathbf{q}_\mathbf{U}\|_2$, and $\mathbf{h}^\top \boldsymbol{\delta}_\mathbf{U} = 0$ for every $\mathbf{h} \in \operatorname{span}(\mathbf{U})$. If $\mathbf{q}_\mathbf{U} = \mathbf{0}$, every feasible perturbation has zero forget gain. At a fixed point and fixed gradients, enlarging a nested protected subspace shrinks the feasible set, so the optimal forget gain cannot increase.

The component of $\mathbf{g}_r$ outside $\operatorname{span}(\mathbf{U})$ determines the first-order retain effect. Since $\mathbf{P}_\mathbf{U}\mathbf{g}_r$ lies in $\operatorname{span}(\mathbf{U})$ and is orthogonal to $\boldsymbol{\delta}_\mathbf{U}$,

$$\begin{aligned}\mathbf{g}_r^\top \boldsymbol{\delta}_\mathbf{U} &= \big((\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{g}_r\big)^\top \boldsymbol{\delta}_\mathbf{U} \\ &\leq \rho\, \|(\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{g}_r\|_2,\end{aligned}$$

and this bound vanishes exactly when $\mathbf{g}_r \in \operatorname{span}(\mathbf{U})$. If L_r is C^2 on an open neighborhood of the segment $\{\mathbf{w} + \zeta\boldsymbol{\delta}_\mathbf{U} : \zeta \in [0,1]\}$ with Hessian operator norm at most M_r on it, then

$$L_r(\mathbf{w} + \boldsymbol{\delta}_\mathbf{U}) \leq L_r(\mathbf{w}) + \rho\, \|(\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{g}_r\|_2 + \frac{M_r}{2}\rho^2.$$

To test whether protecting multiple retain-gradient directions lowers ΔAcc in a high-coupling regime, we run sampled rank-k ROSU on CIFAR-100 random forgetting. We hold the pretrained checkpoint, data split, perturbation radius, learning-rate schedule, training budget, and seed protocol fixed. The $k = 1$ row in Table 12 is copied from the main-table rank-1 ROSU run. The rank-k diagnostic reruns $k \in \{2, 4, 8\}$ with retain subspaces spanned by sampled per-sample retain gradients. These runs use the same amplification schedule as the main-table run.

For sampled retain examples $(\mathbf{x}_i, y_i)$, define $\mathbf{g}_r^{(i)} := \nabla_\mathbf{w} L_r(\mathbf{w}; \{(\mathbf{x}_i, y_i)\})$ at the current weights $\mathbf{w}$, using the supervised singleton loss. For $k \geq 2$, let $\mathbf{U}$ have orthonormal columns spanning the

retained sampled directions from $[\mathbf{g}_r^{(1)}, \ldots, \mathbf{g}_r^{(k)}]$. On steps with $\mathbf{q}_\mathbf{U} \neq \mathbf{0}$, define

$$\mathbf{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^\top, \qquad \mathbf{q}_\mathbf{U} = (\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{g}_f, \qquad \mathbf{P}_{\perp,\mathbf{U}} = \frac{\mathbf{q}_\mathbf{U}\mathbf{q}_\mathbf{U}^\top}{\|\mathbf{q}_\mathbf{U}\|_2^2}.$$

Under the threshold rule of Algorithm 1, nonfallback steps satisfy $\|\mathbf{q}_\mathbf{U}\|_2 > \varepsilon_q$. On these steps, define $\widetilde{\mathbf{g}}_{r,\mathbf{U}} := \nabla_\mathbf{z} L_r(\mathbf{z}; \mathcal{B}_r)|_{\mathbf{z}=\mathbf{w}+\boldsymbol{\delta}_\mathbf{U}}$. The rank-$k$ relaxation drops the differential of the protected-subspace projector and replaces the forget Hessian by $\mathbf{I}$. The transported update uses

$$\mathbf{v} = \left[\mathbf{I} + \frac{\rho}{\|\mathbf{q}_\mathbf{U}\|_2}\big(\mathbf{I} - \mathbf{P}_\mathbf{U} - \mathbf{P}_{\perp,\mathbf{U}}\big)\right]\widetilde{\mathbf{g}}_{r,\mathbf{U}},$$

and degenerate steps use the same retain-gradient fallback as Algorithm 1.

The ideal surrogate perturbation protects every direction in the retained span of $\mathbf{U}$ exactly. The mini-batch mean gradient $\mathbf{g}_r^{\text{mini}}$ contributes the residual term $\rho\|(\mathbf{I}-\mathbf{P}_\mathbf{U})\mathbf{g}_r^{\text{mini}}\|_2$ in the bound above. This term is zero when the sampled span contains that mean gradient. Table 12 reports the results.

Table 12: **Rank-k retain subspace on CIFAR-100.** The $k = 1$ row is the main-table run, and bold marks the lowest ΔAcc and its k.

CIFAR-100 random forgetting					
k	**RA**	**FA**	**TA**	**ΔAcc.$\downarrow$**	**MIA-Eff.**
1	$99.78_{\pm 0.00}$	$77.98_{\pm 0.08}$	$76.85_{\pm 0.18}$	0.67	$41.37_{\pm 0.75}$
2	$99.94_{\pm 0.01}$	$77.47_{\pm 0.33}$	$76.62_{\pm 0.14}$	0.31	$42.54_{\pm 0.56}$
4	$99.94_{\pm 0.01}$	$77.88_{\pm 0.40}$	$76.92_{\pm 0.22}$	0.48	$42.81_{\pm 0.74}$
8	$99.94_{\pm 0.00}$	$77.68_{\pm 0.23}$	$76.54_{\pm 0.07}$	**0.18**	$42.32_{\pm 0.85}$

Each rank-k rerun has a smaller ΔAcc than the main-table run. The ΔAcc values vary nonmonotonically with k, and $k = 8$ gives the lowest, which is the closest aggregate match to Retrain. MIA-Eff. varies little across k.

D Theory and derivations

This appendix proves Proposition 3.2, Theorem 3.3, and Proposition 4.1, and derives the quantities used in Sections 3.3, 3.4, and 4. Appendix D.1 treats the inner problem and the surrogate retain bounds, Appendix D.2 the exact and relaxed retain corrections, and Appendix D.3 the regularized implementation, the fallback, and mini-batch residuals. We hold ρ fixed within each step and suppress $\mathbf{w}$ when all quantities share the same evaluation point.

D.1 Inner geometry and surrogate retain bounds

Here we prove Proposition 3.2 and Theorem 3.3 and derive the gain ratio, the forgetting step, and the near-orthogonality gap used in Sections 3.3, 3.4, and 4.

Proof of Proposition 3.2. A feasible $\boldsymbol{\delta}$ satisfies $\mathbf{g}_r^\top \boldsymbol{\delta} = 0$, hence $\mathbf{P}_r\boldsymbol{\delta} = \mathbf{0}$. For every such $\boldsymbol{\delta}$,

$$\begin{aligned}\mathbf{g}_f^\top \boldsymbol{\delta} &= \mathbf{g}_f^\top (\mathbf{I} - \mathbf{P}_r)\boldsymbol{\delta} \\ &= \mathbf{q}^\top \boldsymbol{\delta} \\ &\leq \|\mathbf{q}\|_2 \, \|\boldsymbol{\delta}\|_2 \\ &\leq \rho \, \|\mathbf{q}\|_2,\end{aligned}$$

where the first equality uses $\mathbf{P}_r\boldsymbol{\delta} = \mathbf{0}$, the second uses the symmetry of $\mathbf{I} - \mathbf{P}_r$, the first inequality is Cauchy–Schwarz, and the last uses $\|\boldsymbol{\delta}\|_2 \leq \rho$.

Next, we show that the candidate $\rho\frac{\mathbf{q}}{\|\mathbf{q}\|_2}$ attains this bound. It is feasible, since $\mathbf{g}_r^\top\mathbf{q} = 0$ and its norm is ρ. Its forget gain is

$$\begin{aligned}\mathbf{g}_f^\top\Big(\rho\,\frac{\mathbf{q}}{\|\mathbf{q}\|_2}\Big) &= \frac{\rho}{\|\mathbf{q}\|_2}\,\mathbf{g}_f^\top(\mathbf{I}-\mathbf{P}_r)^\top(\mathbf{I}-\mathbf{P}_r)\,\mathbf{g}_f\\ &= \frac{\rho}{\|\mathbf{q}\|_2}\,\|\mathbf{q}\|_2^2\\ &= \rho\,\|\mathbf{q}\|_2,\end{aligned}$$

where the first equality uses that $\mathbf{I}-\mathbf{P}_r$ is symmetric and idempotent, and the second uses the definition of $\mathbf{q}$. Equality in the Cauchy–Schwarz step requires a nonnegative multiple of $\mathbf{q}$, and equality in the last step requires norm ρ. Since $\mathbf{q}\neq\mathbf{0}$, the candidate is the only point that satisfies both. □

For the gain ratio, decompose $\mathbf{g}_f = \mathbf{P}_r\mathbf{g}_f + \mathbf{q}$. The two components are orthogonal, and $\|\mathbf{P}_r\mathbf{g}_f\|_2 = |\mathbf{u}_r^\top\mathbf{g}_f| = \|\mathbf{g}_f\|_2|\cos\theta|$, so

$$\begin{aligned}\|\mathbf{q}\|_2^2 &= \|\mathbf{g}_f\|_2^2 - \|\mathbf{P}_r\mathbf{g}_f\|_2^2\\ &= \|\mathbf{g}_f\|_2^2\big(1-\cos^2\theta\big)\\ &= \|\mathbf{g}_f\|_2^2\sin^2\theta,\end{aligned}$$

where the first equality is Pythagoras. Since $\theta\in[0,\pi]$, $\sin\theta\geq 0$ and $\|\mathbf{q}\|_2 = \|\mathbf{g}_f\|_2\sin\theta$. The constrained and unconstrained gains are $\rho\|\mathbf{q}\|_2$ and $\rho\|\mathbf{g}_f\|_2$, so their ratio is $\sin\theta$ at every common positive radius.

For retain-neutral amplification with $\beta>0$, the argument above at radius $\beta\rho$ shows that the unique maximizer is $\beta\boldsymbol{\delta}_{\mathrm{rosu}}$, with gain $\beta\rho\|\mathbf{q}\|_2$. For $\beta=0$, the feasible set is $\{\mathbf{0}\}$.

Proof of Theorem 3.3. Fix a perturbation $\boldsymbol{\delta}$ whose segment $\{\mathbf{w}+\zeta\boldsymbol{\delta} : \zeta\in[0,1]\}$ satisfies the Hessian bound. Taylor's theorem with integral remainder gives

$$\begin{aligned}L_r(\mathbf{w}+\boldsymbol{\delta}) &= L_r(\mathbf{w}) + \mathbf{g}_r^\top\boldsymbol{\delta} + R_{\boldsymbol{\delta}},\\ R_{\boldsymbol{\delta}} &= \int_0^1 (1-\zeta)\,\boldsymbol{\delta}^\top\nabla^2 L_r(\mathbf{w}+\zeta\boldsymbol{\delta})\,\boldsymbol{\delta}\,\mathrm{d}\zeta.\end{aligned}$$

The Hessian bound controls the remainder,

$$\begin{aligned}|R_{\boldsymbol{\delta}}| &\leq \int_0^1 (1-\zeta)\,M_r\,\|\boldsymbol{\delta}\|_2^2\,\mathrm{d}\zeta\\ &= \frac{M_r}{2}\,\|\boldsymbol{\delta}\|_2^2.\end{aligned}$$

For (i), take $\boldsymbol{\delta}=\boldsymbol{\delta}_{\mathrm{rosu}}$. Then

$$\begin{aligned}\big|F_{\mathrm{rosu}}(\mathbf{w}) - L_r(\mathbf{w})\big| &= \big|\mathbf{g}_r^\top\boldsymbol{\delta}_{\mathrm{rosu}} + R_{\boldsymbol{\delta}_{\mathrm{rosu}}}\big|\\ &= \big|R_{\boldsymbol{\delta}_{\mathrm{rosu}}}\big|\\ &\leq \frac{M_r}{2}\,\rho^2,\end{aligned}$$

where the second equality uses $\mathbf{g}_r^\top\boldsymbol{\delta}_{\mathrm{rosu}} = \frac{\rho}{\|\mathbf{q}\|_2}\mathbf{g}_r^\top\mathbf{q} = 0$ (the first claim of (i)), and the inequality applies the remainder bound with $\|\boldsymbol{\delta}_{\mathrm{rosu}}\|_2=\rho$.

For (ii), take $\boldsymbol{\delta}=\boldsymbol{\delta}_{\mathrm{std}}$ as well. Subtracting the two expansions gives

$$\begin{aligned}F_{\mathrm{std}}(\mathbf{w}) - F_{\mathrm{rosu}}(\mathbf{w}) &= \mathbf{g}_r^\top\boldsymbol{\delta}_{\mathrm{std}} + R_{\boldsymbol{\delta}_{\mathrm{std}}} - R_{\boldsymbol{\delta}_{\mathrm{rosu}}}\\ &= \rho\,\|\mathbf{g}_r\|_2\cos\theta + R_{\boldsymbol{\delta}_{\mathrm{std}}} - R_{\boldsymbol{\delta}_{\mathrm{rosu}}}\\ &\geq \rho\,\|\mathbf{g}_r\|_2\cos\theta - |R_{\boldsymbol{\delta}_{\mathrm{std}}}| - |R_{\boldsymbol{\delta}_{\mathrm{rosu}}}|\\ &\geq \rho\,\|\mathbf{g}_r\|_2\cos\theta - M_r\rho^2,\end{aligned}$$

where the first equality uses $\mathbf{g}_r^\top\boldsymbol{\delta}_{\mathrm{rosu}}=0$, the second uses (1), and the last inequality applies the remainder bound to both segments with $\|\boldsymbol{\delta}_{\mathrm{std}}\|_2 = \|\boldsymbol{\delta}_{\mathrm{rosu}}\|_2 = \rho$. □

Whenever $\rho\|\mathbf{g}_r\|_2 \cos\theta > M_r\rho^2$, the bound in (ii) is positive and therefore $F_{\mathrm{std}}(\mathbf{w}) > F_{\mathrm{rosu}}(\mathbf{w})$. For $M_r > 0$, this condition reads $\cos\theta > \frac{M_r\rho}{\|\mathbf{g}_r\|_2}$, which requires positive coupling and $\rho < \frac{\|\mathbf{g}_r\|_2 \cos\theta}{M_r}$. If $\cos\theta \le 0$, the standard perturbation has a nonpositive first-order retain effect.

The curvature bound also covers the forgetting step. If $0 \le \beta \le 1$, the amplified segment lies within the original segment. For $\beta > 1$, require that L_r is C^2 near $\{\mathbf{w} + \zeta\beta\boldsymbol{\delta}_{\mathrm{rosu}} : \zeta \in [0,1]\}$ and that the Hessian bound holds on it. With $\boldsymbol{\delta} = \beta\boldsymbol{\delta}_{\mathrm{rosu}}$, the linear term of the expansion above vanishes, and the expansion then gives $|L_r(\mathbf{w} + \beta\boldsymbol{\delta}_{\mathrm{rosu}}) - L_r(\mathbf{w})| \le \frac{M_r}{2}\beta^2\rho^2$.

Near orthogonality the two perturbations are close. Both perturbations are ρ times a unit vector, and $\mathbf{q}^\top \mathbf{g}_f = \|\mathbf{q}\|_2^2$. Expanding the squared distance gives

$$\begin{aligned}
\|\boldsymbol{\delta}_{\mathrm{rosu}} - \boldsymbol{\delta}_{\mathrm{std}}\|_2^2 &= \rho^2 \Big\| \frac{\mathbf{q}}{\|\mathbf{q}\|_2} - \frac{\mathbf{g}_f}{\|\mathbf{g}_f\|_2} \Big\|_2^2 \\
&= 2\rho^2 \Big(1 - \frac{\mathbf{q}^\top \mathbf{g}_f}{\|\mathbf{q}\|_2 \, \|\mathbf{g}_f\|_2}\Big) \\
&= 2\rho^2 \big(1 - \sin\theta\big) \\
&= \frac{2\rho^2 \cos^2\theta}{1 + \sin\theta} \\
&\le 2\rho^2 \cos^2\theta,
\end{aligned}$$

where the second equality expands the square of a difference of unit vectors and the third uses $\mathbf{q}^\top \mathbf{g}_f = \|\mathbf{q}\|_2^2$ and the gain ratio. The fourth equality multiplies and divides by $1 + \sin\theta$, and the inequality uses $\sin\theta \ge 0$. Taking square roots gives $\|\boldsymbol{\delta}_{\mathrm{rosu}} - \boldsymbol{\delta}_{\mathrm{std}}\|_2 \le \sqrt{2}\rho|\cos\theta|$. If L_r is G_r-Lipschitz on the segment joining the two surrogate points, then $|F_{\mathrm{rosu}} - F_{\mathrm{std}}| \le G_r\|\boldsymbol{\delta}_{\mathrm{rosu}} - \boldsymbol{\delta}_{\mathrm{std}}\|_2 \le \sqrt{2}G_r\rho|\cos\theta|$.

D.2 Exact and relaxed retain corrections

Here we derive the exact chain-rule gradient of F_{rosu}, obtain the relaxed Jacobian of Section 4, and prove Proposition 4.1.

Let $L_f, L_r \in C^2$, $\mathbf{g}_r \neq \mathbf{0}$, and $\mathbf{q} \neq \mathbf{0}$. Write $\widetilde{\mathbf{g}}_r(\mathbf{w}) := \nabla_{\mathbf{z}} L_r(\mathbf{z})|_{\mathbf{z}=\mathbf{w}+\boldsymbol{\delta}_{\mathrm{rosu}}(\mathbf{w})}$. Differentiating the surrogate retain objective with ρ fixed gives

$$\nabla F_{\mathrm{rosu}}(\mathbf{w}) = \big(\mathbf{I} + \mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w})^\top\big)\, \widetilde{\mathbf{g}}_r(\mathbf{w}).$$

Let $\boldsymbol{\xi} \in \mathbb{R}^p$ be a parameter direction and $\mathbf{J}_{\mathbf{q}}$ the Jacobian of $\mathbf{q}$. All quantities below are evaluated at $\mathbf{w}$ unless another argument is shown. At a nonzero vector $\mathbf{z} \in \mathbb{R}^p$, the norm has derivative $\mathrm{D}\|\mathbf{z}\|_2[\boldsymbol{\xi}] = \frac{\mathbf{z}^\top \boldsymbol{\xi}}{\|\mathbf{z}\|_2}$, and the quotient rule gives

$$\mathrm{D}\left(\frac{\mathbf{z}}{\|\mathbf{z}\|_2}\right)[\boldsymbol{\xi}] = \frac{1}{\|\mathbf{z}\|_2}\left(\mathbf{I} - \frac{\mathbf{z}\mathbf{z}^\top}{\|\mathbf{z}\|_2^2}\right)\boldsymbol{\xi}.$$

We next apply this derivative to the normalized retain gradient. Since $\mathbf{u}_r = \frac{\mathbf{g}_r}{\|\mathbf{g}_r\|_2}$,

$$\begin{aligned}
\mathrm{D}\mathbf{u}_r[\boldsymbol{\xi}] &= \mathrm{D}\left(\frac{\mathbf{g}_r}{\|\mathbf{g}_r\|_2}\right)[\boldsymbol{\xi}] \\
&= \frac{1}{\|\mathbf{g}_r\|_2}\left(\mathbf{I} - \frac{\mathbf{g}_r\mathbf{g}_r^\top}{\|\mathbf{g}_r\|_2^2}\right)\mathrm{D}\mathbf{g}_r[\boldsymbol{\xi}] \\
&= \frac{1}{\|\mathbf{g}_r\|_2}(\mathbf{I} - \mathbf{P}_r)\nabla^2 L_r(\mathbf{w})\boldsymbol{\xi},
\end{aligned}$$

where the second equality applies the normalization derivative through the chain rule, and the third uses the definition of $\mathbf{P}_r$ and $\mathrm{D}\mathbf{g}_r[\boldsymbol{\xi}] = \nabla^2 L_r(\mathbf{w})\boldsymbol{\xi}$. Differentiating $\mathbf{P}_r = \mathbf{u}_r\mathbf{u}_r^\top$ now yields

$$\begin{aligned}
\mathrm{D}\mathbf{P}_r[\boldsymbol{\xi}] &= \mathrm{D}(\mathbf{u}_r\mathbf{u}_r^\top)[\boldsymbol{\xi}] \\
&= (\mathrm{D}\mathbf{u}_r[\boldsymbol{\xi}])\mathbf{u}_r^\top + \mathbf{u}_r(\mathrm{D}\mathbf{u}_r[\boldsymbol{\xi}])^\top \\
&= \frac{(\mathbf{I} - \mathbf{P}_r)\nabla^2 L_r(\mathbf{w})\boldsymbol{\xi}\mathbf{u}_r^\top + \mathbf{u}_r\boldsymbol{\xi}^\top\nabla^2 L_r(\mathbf{w})(\mathbf{I} - \mathbf{P}_r)}{\|\mathbf{g}_r\|_2},
\end{aligned}$$

where the second equality uses the product rule, and the third substitutes the preceding derivative and uses the symmetry of $\mathbf{P}_r$ and $\nabla^2 L_r$. The Hessian is symmetric because L_r is C^2.

Finally, apply the normalization derivative at $\mathbf{z} = \mathbf{q}$ to $\boldsymbol{\delta}_{\mathrm{rosu}}$ in (3). Together with the product rule for $\mathbf{q} = (\mathbf{I} - \mathbf{P}_r)\mathbf{g}_f$, this gives

$$\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w}) = \frac{\rho}{\|\mathbf{q}\|_2}\big(\mathbf{I} - \mathbf{P}_\perp\big)\mathbf{J}_{\mathbf{q}}(\mathbf{w}), \tag{8a}$$

$$\mathbf{J}_{\mathbf{q}}(\mathbf{w})[\boldsymbol{\xi}] = (\mathbf{I} - \mathbf{P}_r)\nabla^2 L_f(\mathbf{w})\,\boldsymbol{\xi} - \mathrm{D}\mathbf{P}_r(\mathbf{w})[\boldsymbol{\xi}]\,\mathbf{g}_f, \tag{8b}$$

where (8a) uses $\frac{\mathbf{q}\mathbf{q}^\top}{\|\mathbf{q}\|_2^2} = \mathbf{P}_\perp$, and (8b) uses $\mathrm{D}\mathbf{g}_f[\boldsymbol{\xi}] = \nabla^2 L_f \boldsymbol{\xi}$. These are the two derivative terms that A1 and A2 replace.

A1 sets $\mathrm{D}\mathbf{P}_r[\boldsymbol{\xi}]$ to zero, and A2 replaces $\nabla^2 L_f$ by $\mathbf{I}$ in (8b). These replacements leave the forward constraint and the projector unchanged, and the relaxed derivative of $\mathbf{q}$ becomes $\mathbf{I} - \mathbf{P}_r$. Substituting it into (8a) gives

$$\begin{aligned}
\widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}(\mathbf{w}) &= \frac{\rho}{\|\mathbf{q}\|_2}(\mathbf{I} - \mathbf{P}_\perp)(\mathbf{I} - \mathbf{P}_r) \\
&= \frac{\rho}{\|\mathbf{q}\|_2}\big(\mathbf{I} - \mathbf{P}_r - \mathbf{P}_\perp + \mathbf{P}_\perp\mathbf{P}_r\big) \\
&= \frac{\rho}{\|\mathbf{q}\|_2}\big(\mathbf{I} - \mathbf{P}_r - \mathbf{P}_\perp\big),
\end{aligned} \tag{9}$$

where the last equality uses $\mathbf{P}_\perp\mathbf{P}_r = \mathbf{0}$, since $\mathbf{u}_\perp$ is orthogonal to $\mathbf{u}_r$. The ideal relaxed Jacobian is therefore symmetric, and substituting it into the chain-rule expression yields $\mathbf{v}$ in (4). Appendix D.3 evaluates the cross term for the regularized operator.

Proof of Proposition 4.1. Subtracting (9) from (8a) gives

$$\begin{aligned}
\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}} &= \frac{\rho}{\|\mathbf{q}\|_2}(\mathbf{I} - \mathbf{P}_\perp)\big(\mathbf{J}_{\mathbf{q}} - (\mathbf{I} - \mathbf{P}_r)\big), \\
\big(\mathbf{J}_{\mathbf{q}} - (\mathbf{I} - \mathbf{P}_r)\big)\boldsymbol{\xi} &= (\mathbf{I} - \mathbf{P}_r)\big(\nabla^2 L_f - \mathbf{I}\big)\boldsymbol{\xi} - \mathrm{D}\mathbf{P}_r[\boldsymbol{\xi}]\,\mathbf{g}_f.
\end{aligned}$$

The two terms on the right of the second line are the A2 and A1 errors. For $\|\boldsymbol{\xi}\|_2 = 1$,

$$\begin{aligned}
\big\|\big(\mathbf{J}_{\mathbf{q}} - (\mathbf{I} - \mathbf{P}_r)\big)\boldsymbol{\xi}\big\|_2 &\le \|\mathbf{I} - \mathbf{P}_r\|_{\mathrm{op}}\,\|\nabla^2 L_f - \mathbf{I}\|_{\mathrm{op}}\,\|\boldsymbol{\xi}\|_2 + \|\mathrm{D}\mathbf{P}_r[\boldsymbol{\xi}]\|_{\mathrm{op}}\,\|\mathbf{g}_f\|_2 \\
&\le \varepsilon_H + \varepsilon_P\|\mathbf{g}_f\|_2,
\end{aligned}$$

where the first inequality is the triangle inequality with submultiplicativity, and the second uses $\|\mathbf{I} - \mathbf{P}_r\|_{\mathrm{op}} \le 1$, $\|\boldsymbol{\xi}\|_2 = 1$, and the two assumptions. Since $\|\mathbf{I} - \mathbf{P}_\perp\|_{\mathrm{op}} \le 1$, the first line gives (5a). For (5b), the chain rule, the definition of $\mathbf{v}$, and the symmetry of $\widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}$ give

$$\begin{aligned}
\big\|\nabla F_{\mathrm{rosu}} - \mathbf{v}\big\|_2 &= \big\|\big(\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}\big)^\top\widetilde{\mathbf{g}}_r\big\|_2 \\
&\le \big\|\mathbf{J}_{\boldsymbol{\delta}_{\mathrm{rosu}}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}\big\|_{\mathrm{op}}\,\|\widetilde{\mathbf{g}}_r\|_2 \\
&\le \frac{\rho\,\|\widetilde{\mathbf{g}}_r\|_2}{\|\mathbf{q}\|_2}\big(\varepsilon_H + \varepsilon_P\|\mathbf{g}_f\|_2\big),
\end{aligned}$$

where the first inequality uses that a matrix and its transpose have the same operator norm, and the second is the first bound. □

The bound also yields a descent condition. Suppose $\|\nabla F_{\mathrm{rosu}} - \mathbf{v}\|_2 < \|\mathbf{v}\|_2$, which also forces $\mathbf{v} \neq \mathbf{0}$. Then

$$\begin{aligned}
\mathbf{v}^\top\nabla F_{\mathrm{rosu}} &= \|\mathbf{v}\|_2^2 + \mathbf{v}^\top\big(\nabla F_{\mathrm{rosu}} - \mathbf{v}\big) \\
&\ge \|\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2\,\|\nabla F_{\mathrm{rosu}} - \mathbf{v}\|_2 \\
&> 0,
\end{aligned}$$

where the first inequality is Cauchy–Schwarz and the second uses the assumption. This makes $-\mathbf{v}$ a descent direction, and small enough steps along it decrease F_{rosu}.

Under the Hessian bound on the ideal surrogate segment,

$$\begin{aligned}\|\widetilde{\mathbf{g}}_r\|_2 &= \Big\|\mathbf{g}_r + \int_0^1 \nabla^2 L_r(\mathbf{w} + \zeta\boldsymbol{\delta}_{\mathrm{rosu}})\,\boldsymbol{\delta}_{\mathrm{rosu}}\,\mathrm{d}\zeta\Big\|_2 \\ &\le \|\mathbf{g}_r\|_2 + M_r\rho,\end{aligned}$$

where the equality is the fundamental theorem of calculus along the segment, and the inequality uses the Hessian bound with $\|\boldsymbol{\delta}_{\mathrm{rosu}}\|_2 = \rho$. Substituting it into (5b) expresses the deviation through the retain gradient at the current weights.

D.3 Regularization and mini-batch residuals

Here we quantify how the stabilizer τ, the degenerate fallback, and mini-batch estimation change the ideal construction.

For $\tau \ge 0$ and $\mathbf{g}_r \neq \mathbf{0}$, the operator and direction satisfy

$$\begin{aligned}\mathbf{P}_r^{(\tau)} &:= \frac{\mathbf{g}_r\mathbf{g}_r^\top}{\|\mathbf{g}_r\|_2^2 + \tau} \\ &= \frac{\|\mathbf{g}_r\|_2^2}{\|\mathbf{g}_r\|_2^2 + \tau}\,\mathbf{P}_r, \qquad (10a)\\ \mathbf{q}_\tau &:= (\mathbf{I} - \mathbf{P}_r^{(\tau)})\mathbf{g}_f \\ &= \mathbf{q} + \frac{\tau}{\|\mathbf{g}_r\|_2^2 + \tau}\,\mathbf{P}_r\mathbf{g}_f, \qquad (10b)\end{aligned}$$

where (10b) uses (10a) and $1 - \frac{\|\mathbf{g}_r\|_2^2}{\|\mathbf{g}_r\|_2^2+\tau} = \frac{\tau}{\|\mathbf{g}_r\|_2^2+\tau}$. For $\tau > 0$, $\mathbf{P}_r^{(\tau)}$ is not idempotent. The retain inner product of $\mathbf{q}_\tau$ is

$$\begin{aligned}\mathbf{g}_r^\top\mathbf{q}_\tau &= \mathbf{g}_r^\top\mathbf{q} + \frac{\tau}{\|\mathbf{g}_r\|_2^2 + \tau}\,\mathbf{g}_r^\top\mathbf{P}_r\mathbf{g}_f \\ &= \frac{\tau\,\mathbf{g}_f^\top\mathbf{g}_r}{\|\mathbf{g}_r\|_2^2 + \tau},\end{aligned}$$

where the second equality uses $\mathbf{g}_r^\top\mathbf{q} = 0$ and $\mathbf{g}_r^\top\mathbf{P}_r = \mathbf{g}_r^\top$.

The operator difference has norm $\|\mathbf{P}_r^{(\tau)} - \mathbf{P}_r\|_{\mathrm{op}} = \frac{\tau}{\|\mathbf{g}_r\|_2^2+\tau}$, because $\|\mathbf{P}_r\|_{\mathrm{op}} = 1$. Applying that difference to $\mathbf{g}_f$ yields

$$\|\mathbf{q}_\tau - \mathbf{q}\|_2 \le \frac{\tau}{\|\mathbf{g}_r\|_2^2 + \tau}\,\|\mathbf{g}_f\|_2. \qquad (11)$$

The two components in (10b) are orthogonal, so

$$\begin{aligned}\|\mathbf{q}_\tau\|_2^2 &= \|\mathbf{q}\|_2^2 + \Big(\frac{\tau}{\|\mathbf{g}_r\|_2^2 + \tau}\Big)^2\|\mathbf{P}_r\mathbf{g}_f\|_2^2 \\ &\ge \|\mathbf{q}\|_2^2,\end{aligned}$$

where the equality is Pythagoras. If $\mathbf{g}_r = \mathbf{0}$ and $\tau > 0$, the implemented formula gives $\mathbf{q}_\tau = \mathbf{g}_f$, and the ideal projector is undefined.

Let $\mathbf{q}_\tau \neq \mathbf{0}$, $\mathbf{u}_\tau := \frac{\mathbf{q}_\tau}{\|\mathbf{q}_\tau\|_2}$, and $\mathbf{P}_\tau := \mathbf{u}_\tau\mathbf{u}_\tau^\top$. Algorithm 1 uses the symmetric matrix $\widehat{\mathbf{J}}_\tau^{\mathrm{impl}} := \frac{\rho}{\|\mathbf{q}_\tau\|_2}(\mathbf{I} - \mathbf{P}_r^{(\tau)} - \mathbf{P}_\tau)$. Repeating A1/A2 with $\mathbf{P}_r^{(\tau)}$ instead gives the product-form Jacobian

$$\begin{aligned}\widehat{\mathbf{J}}_\tau^{\mathrm{prod}} &:= \frac{\rho}{\|\mathbf{q}_\tau\|_2}(\mathbf{I} - \mathbf{P}_\tau)(\mathbf{I} - \mathbf{P}_r^{(\tau)}) \\ &= \frac{\rho}{\|\mathbf{q}_\tau\|_2}\big(\mathbf{I} - \mathbf{P}_r^{(\tau)} - \mathbf{P}_\tau\big) + \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\mathbf{P}_\tau\mathbf{P}_r^{(\tau)} \\ &= \widehat{\mathbf{J}}_\tau^{\mathrm{impl}} + \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\mathbf{P}_\tau\mathbf{P}_r^{(\tau)}.\end{aligned}$$

Rearranging and transposing gives

$$\widehat{\mathbf{J}}_\tau^{\mathrm{prod}} - \widehat{\mathbf{J}}_\tau^{\mathrm{impl}} = \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\mathbf{P}_\tau \mathbf{P}_r^{(\tau)},$$
$$\big(\widehat{\mathbf{J}}_\tau^{\mathrm{prod}}\big)^\top - \big(\widehat{\mathbf{J}}_\tau^{\mathrm{impl}}\big)^\top = \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\mathbf{P}_r^{(\tau)} \mathbf{P}_\tau,$$

where the second line uses the symmetry of $\mathbf{P}_\tau$ and $\mathbf{P}_r^{(\tau)}$. The cross term vanishes at $\tau = 0$. For $\tau > 0$, it can be nonzero, and it still vanishes when $\mathbf{g}_f^\top \mathbf{g}_r = 0$. The two cross products are transposes of each other with equal operator norms, and the transported correction uses the transpose.

Writing $d_\tau := \|\mathbf{g}_r\|_2^2 + \tau$, the cross term has the rank-one form $\mathbf{P}_\tau \mathbf{P}_r^{(\tau)} = \frac{\mathbf{u}_\tau^\top \mathbf{g}_r}{d_\tau}\,\mathbf{u}_\tau \mathbf{g}_r^\top$, and the retain inner product above gives $|\mathbf{u}_\tau^\top \mathbf{g}_r| = \frac{\tau |\mathbf{g}_f^\top \mathbf{g}_r|}{d_\tau \|\mathbf{q}_\tau\|_2}$. Taking operator norms gives

$$\begin{aligned}
\big\|\widehat{\mathbf{J}}_\tau^{\mathrm{prod}} - \widehat{\mathbf{J}}_\tau^{\mathrm{impl}}\big\|_{\mathrm{op}} &= \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\frac{|\mathbf{u}_\tau^\top \mathbf{g}_r|\,\|\mathbf{g}_r\|_2}{d_\tau} \\
&= \frac{\rho\,\tau\,|\mathbf{g}_f^\top \mathbf{g}_r|\,\|\mathbf{g}_r\|_2}{\|\mathbf{q}_\tau\|_2^2\, d_\tau^2} \\
&\le \frac{\rho\,\tau\,\|\mathbf{g}_f\|_2\,\|\mathbf{g}_r\|_2^2}{\|\mathbf{q}_\tau\|_2^2\, d_\tau^2},
\end{aligned} \tag{12}$$

where the first equality uses $\|\mathbf{x}\mathbf{y}^\top\|_{\mathrm{op}} = \|\mathbf{x}\|_2\|\mathbf{y}\|_2$ with $\|\mathbf{u}_\tau\|_2 = 1$, the second substitutes $|\mathbf{u}_\tau^\top \mathbf{g}_r|$, and the inequality is Cauchy–Schwarz. The transported-vector difference is $\frac{\rho}{\|\mathbf{q}_\tau\|_2}\mathbf{P}_r^{(\tau)}\mathbf{P}_\tau \widetilde{\mathbf{g}}_{r,\tau}$, whose norm is at most the bound in (12) times $\|\widetilde{\mathbf{g}}_{r,\tau}\|_2$. Here $\widetilde{\mathbf{g}}_{r,\tau}$ is evaluated at $\mathbf{w} + \rho\mathbf{u}_\tau$. The bound in (12) is uniformly $O(\tau)$ when $\|\mathbf{g}_r\|_2$ and $\|\mathbf{q}_\tau\|_2$ are bounded away from zero and ρ and $\|\mathbf{g}_f\|_2$ are bounded above.

Substituting $\boldsymbol{\delta}_\tau = \rho\mathbf{u}_\tau$ into the retain inner product above gives

$$\begin{aligned}
\mathbf{g}_r^\top \boldsymbol{\delta}_\tau &= \frac{\rho}{\|\mathbf{q}_\tau\|_2}\,\mathbf{g}_r^\top \mathbf{q}_\tau \\
&= \frac{\rho\,\tau\,\mathbf{g}_f^\top \mathbf{g}_r}{(\|\mathbf{g}_r\|_2^2 + \tau)\,\|\mathbf{q}_\tau\|_2}.
\end{aligned} \tag{13}$$

Assume L_r is C^2 near the segment from $\mathbf{w}$ to $\mathbf{w} + \boldsymbol{\delta}_\tau$, with Hessian operator norm at most M_r on that segment. The expansion of Appendix D.1 then gives

$$\begin{aligned}
L_r(\mathbf{w} + \boldsymbol{\delta}_\tau) - L_r(\mathbf{w}) &= \mathbf{g}_r^\top \boldsymbol{\delta}_\tau + R_{\boldsymbol{\delta}_\tau} \\
&\le \mathbf{g}_r^\top \boldsymbol{\delta}_\tau + \frac{M_r}{2}\,\rho^2,
\end{aligned}$$

where the inequality applies the remainder bound with $\|\boldsymbol{\delta}_\tau\|_2 = \rho$. Scaling the regularized displacement by β multiplies the linear residual by β. The ideal construction with $\tau = 0$ is exactly retain-neutral to first order.

When $\|\mathbf{q}_\tau\|_2 \le \varepsilon_q$, Algorithm 1 takes a plain retain-gradient step. For $\mathbf{g}_r \neq \mathbf{0}$ and $\tau \ge 0$, the forgone forget gain satisfies

$$\begin{aligned}
\max_{\boldsymbol{\delta}} \big\{ \mathbf{g}_f^\top \boldsymbol{\delta} \text{ s.t. } \|\boldsymbol{\delta}\|_2 \le \rho,\ \mathbf{g}_r^\top \boldsymbol{\delta} = 0 \big\} &= \rho\,\|\mathbf{q}\|_2 \\
&\le \rho\,\|\mathbf{q}_\tau\|_2 \\
&\le \rho\,\varepsilon_q,
\end{aligned}$$

where the equality is Proposition 3.2, the first inequality uses $\|\mathbf{q}_\tau\|_2 \ge \|\mathbf{q}\|_2$ from the decomposition above, and the second is the fallback test. The equality also holds when $\mathbf{q} = \mathbf{0}$, because every feasible gain is then zero.

If $\mathbf{g}_r = \mathbf{0}$ and $\tau > 0$, the retain constraint is vacuous and $\mathbf{q}_\tau = \mathbf{g}_f$. The same test then bounds the ball optimum by $\rho\|\mathbf{g}_f\|_2 = \rho\|\mathbf{q}_\tau\|_2 \le \rho\varepsilon_q$. Applying the argument at radius $\beta\rho$ bounds the forgone gain from retain-neutral amplification by $\beta\rho\varepsilon_q$.

For mini-batch residuals, use the gradients and unit direction defined in Appendix C.4. Dividing the mini-batch form of (13) by ρ gives

$$(\mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}} = \frac{\tau (\mathbf{g}_f^{\text{mini}})^\top \mathbf{g}_r^{\text{mini}}}{(\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau)\|\mathbf{q}_\tau^{\text{mini}}\|_2}.$$

Adding and subtracting $\mathbf{g}_r^{\text{mini}}$ in the full-retain inner product yields

$$\begin{aligned}
\left|(\mathbf{g}_r^{\text{full}})^\top \mathbf{u}_\tau^{\text{mini}}\right| &= \left|(\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}} + (\mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}}\right| \\
&\le \left|(\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}}\right| + \left|(\mathbf{g}_r^{\text{mini}})^\top \mathbf{u}_\tau^{\text{mini}}\right| \\
&\le \|\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}}\|_2 + \frac{\tau \left|(\mathbf{g}_f^{\text{mini}})^\top \mathbf{g}_r^{\text{mini}}\right|}{(\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau)\|\mathbf{q}_\tau^{\text{mini}}\|_2},
\end{aligned}$$

where the first inequality is the triangle inequality, and the second uses Cauchy–Schwarz with $\|\mathbf{u}_\tau^{\text{mini}}\|_2 = 1$ and the mini-batch residual above. This is (7). For $\mathbf{g}_r^{\text{full}} \neq \mathbf{0}$, dividing by its norm gives the normalized diagnostic in Appendix C.4.

If $\mathbf{g}_r^{\text{mini}} \neq \mathbf{0}$, define $\mathbf{P}_r^{\text{mini}} := \frac{\mathbf{g}_r^{\text{mini}} (\mathbf{g}_r^{\text{mini}})^\top}{\|\mathbf{g}_r^{\text{mini}}\|_2^2}$ and $\mathbf{q}^{\text{mini}} := (\mathbf{I} - \mathbf{P}_r^{\text{mini}})\mathbf{g}_f^{\text{mini}}$. For $\mathbf{q}^{\text{mini}} \neq \mathbf{0}$, the normalized direction $\mathbf{u}_\perp^{\text{mini}} := \frac{\mathbf{q}^{\text{mini}}}{\|\mathbf{q}^{\text{mini}}\|_2}$ has exactly zero mini-batch retain inner product. Its full-retain violation is therefore bounded by the estimation error alone, $|(\mathbf{g}_r^{\text{full}})^\top \mathbf{u}_\perp^{\text{mini}}| \le \|\mathbf{g}_r^{\text{full}} - \mathbf{g}_r^{\text{mini}}\|_2$. By (11), the ideal and implemented components satisfy

$$\|\mathbf{q}_\tau^{\text{mini}} - \mathbf{q}^{\text{mini}}\|_2 \le \frac{\tau \|\mathbf{g}_f^{\text{mini}}\|_2}{\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau}.$$

To compare the normalized directions, take nonzero vectors $\mathbf{z}_1, \mathbf{z}_2$. Adding and subtracting $\frac{\mathbf{z}_2}{\|\mathbf{z}_1\|_2}$ gives

$$\begin{aligned}
\left\| \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|_2} - \frac{\mathbf{z}_2}{\|\mathbf{z}_2\|_2} \right\|_2 &\le \frac{\|\mathbf{z}_1 - \mathbf{z}_2\|_2}{\|\mathbf{z}_1\|_2} + \|\mathbf{z}_2\|_2 \left| \frac{1}{\|\mathbf{z}_1\|_2} - \frac{1}{\|\mathbf{z}_2\|_2} \right| \\
&= \frac{\|\mathbf{z}_1 - \mathbf{z}_2\|_2 + \big|\|\mathbf{z}_2\|_2 - \|\mathbf{z}_1\|_2\big|}{\|\mathbf{z}_1\|_2} \\
&\le \frac{2\|\mathbf{z}_1 - \mathbf{z}_2\|_2}{\|\mathbf{z}_1\|_2},
\end{aligned}$$

where the last inequality is the reverse triangle inequality. Exchanging the roles of $\mathbf{z}_1$ and $\mathbf{z}_2$ gives the same bound with $\|\mathbf{z}_2\|_2$ in the denominator, so the bound holds with $\min\{\|\mathbf{z}_1\|_2, \|\mathbf{z}_2\|_2\}$. With $\mathbf{z}_1 = \mathbf{q}^{\text{mini}}$ and $\mathbf{z}_2 = \mathbf{q}_\tau^{\text{mini}}$, the normalized ideal and regularized directions are therefore close when the relative quantity

$$\frac{\tau \|\mathbf{g}_f^{\text{mini}}\|_2}{(\|\mathbf{g}_r^{\text{mini}}\|_2^2 + \tau) \min\{\|\mathbf{q}^{\text{mini}}\|_2, \|\mathbf{q}_\tau^{\text{mini}}\|_2\}}$$

is small. This criterion compares the direction gap with the projected-gradient norms. Figure 4 reports the individual gradient-norm and fallback diagnostics.

We now compare the implemented correction with its ideal counterpart at a fixed $\mathbf{w}$, with $\mathbf{g}_r \neq \mathbf{0}$ and $\mathbf{q} \neq \mathbf{0}$. Keep ρ fixed. Both corrections are computed from the same retain loss, so τ is the only difference between them. The reverse triangle inequality gives

$$\begin{aligned}
\left| \frac{\rho}{\|\mathbf{q}_\tau\|_2} - \frac{\rho}{\|\mathbf{q}\|_2} \right| &= \frac{\rho \big|\|\mathbf{q}_\tau\|_2 - \|\mathbf{q}\|_2\big|}{\|\mathbf{q}\|_2 \|\mathbf{q}_\tau\|_2} \\
&\le \frac{\rho \|\mathbf{q}_\tau - \mathbf{q}\|_2}{\|\mathbf{q}\|_2 \|\mathbf{q}_\tau\|_2}.
\end{aligned}$$

Next, the two direction projectors satisfy

$$\mathbf{P}_\tau - \mathbf{P}_\perp = (\mathbf{u}_\tau - \mathbf{u}_\perp)\mathbf{u}_\tau^\top + \mathbf{u}_\perp(\mathbf{u}_\tau - \mathbf{u}_\perp)^\top,$$

and therefore

$$\begin{aligned}\|\mathbf{P}_\tau - \mathbf{P}_\perp\|_{\mathrm{op}} &\le 2\|\mathbf{u}_\tau - \mathbf{u}_\perp\|_2 \\ &\le \frac{4\|\mathbf{q}_\tau - \mathbf{q}\|_2}{\|\mathbf{q}\|_2},\end{aligned}$$

where the first inequality uses the rank-one matrix norm and unit-vector norms, and the second uses the normalization bound above and $\|\mathbf{q}_\tau\|_2 \ge \|\mathbf{q}\|_2$.

Adding and subtracting the ideal matrix at the regularized scale gives

$$\begin{aligned}\widehat{\mathbf{J}}_\tau^{\mathrm{impl}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}} &= \left(\frac{\rho}{\|\mathbf{q}_\tau\|_2} - \frac{\rho}{\|\mathbf{q}\|_2}\right)(\mathbf{I} - \mathbf{P}_r - \mathbf{P}_\perp) \\ &\quad + \frac{\rho}{\|\mathbf{q}_\tau\|_2}(\mathbf{P}_r - \mathbf{P}_r^{(\tau)} + \mathbf{P}_\perp - \mathbf{P}_\tau).\end{aligned}$$

The ideal matrix $\mathbf{I} - \mathbf{P}_r - \mathbf{P}_\perp$ is an orthogonal projector, so its operator norm is at most one. Applying the preceding bounds yields

$$\begin{aligned}\left\|\widehat{\mathbf{J}}_\tau^{\mathrm{impl}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}\right\|_{\mathrm{op}} &\le \left|\frac{\rho}{\|\mathbf{q}_\tau\|_2} - \frac{\rho}{\|\mathbf{q}\|_2}\right| \\ &\quad + \frac{\rho}{\|\mathbf{q}_\tau\|_2}\left(\|\mathbf{P}_r - \mathbf{P}_r^{(\tau)}\|_{\mathrm{op}} + \|\mathbf{P}_\perp - \mathbf{P}_\tau\|_{\mathrm{op}}\right) \\ &\le \frac{\rho}{\|\mathbf{q}_\tau\|_2}\left(\frac{\tau}{d_\tau} + \frac{5\|\mathbf{q}_\tau - \mathbf{q}\|_2}{\|\mathbf{q}\|_2}\right) \\ &\le \frac{\rho\tau}{d_\tau\|\mathbf{q}_\tau\|_2}\left(1 + \frac{5\|\mathbf{g}_f\|_2}{\|\mathbf{q}\|_2}\right),\end{aligned}$$

where the first inequality uses the triangle inequality and submultiplicativity, the second combines the scale and projector bounds, and the third uses (11). Since $d_\tau \ge \|\mathbf{g}_r\|_2^2 > 0$ and $\|\mathbf{q}_\tau\|_2 \ge \|\mathbf{q}\|_2 > 0$, this matrix difference is $O(\tau)$.

It remains to account for the change in the surrogate retain gradient. The C^2 assumption in Section 4 gives a neighborhood of the ideal surrogate on which the Hessian is continuous. Choose a closed ball around that surrogate within this neighborhood. The direction bound implies $\boldsymbol{\delta}_\tau \to \boldsymbol{\delta}_{\mathrm{rosu}}$. For sufficiently small τ, the connecting segment lies in this ball, where the Hessian norm has a common bound, which we again denote by M_r. Integrating the Hessian along this segment gives

$$\begin{aligned}\|\widetilde{\mathbf{g}}_{r,\tau} - \widetilde{\mathbf{g}}_r\|_2 &\le M_r\|\boldsymbol{\delta}_\tau - \boldsymbol{\delta}_{\mathrm{rosu}}\|_2 \\ &= M_r\rho\|\mathbf{u}_\tau - \mathbf{u}_\perp\|_2 \\ &\le \frac{2M_r\rho\|\mathbf{q}_\tau - \mathbf{q}\|_2}{\|\mathbf{q}\|_2},\end{aligned}$$

where the last inequality uses the normalization bound. This bound is also $O(\tau)$ by (11).

Write $\mathbf{v}_\tau := (\mathbf{I} + \widehat{\mathbf{J}}_\tau^{\mathrm{impl}})\widetilde{\mathbf{g}}_{r,\tau}$ for the correction formula in the nonfallback branch. Both $\mathbf{I} - \mathbf{P}_r^{(\tau)}$ and $\mathbf{P}_\tau$ are positive semidefinite contractions, so

$$\begin{aligned}-\mathbf{I} \preceq \mathbf{I} - \mathbf{P}_r^{(\tau)} - \mathbf{P}_\tau \preceq \mathbf{I}, \\ \|\widehat{\mathbf{J}}_\tau^{\mathrm{impl}}\|_{\mathrm{op}} \le \frac{\rho}{\|\mathbf{q}_\tau\|_2}.\end{aligned}$$

Adding and subtracting the implemented matrix applied to the ideal surrogate gradient gives

$$\begin{aligned}\mathbf{v}_\tau - \mathbf{v} &= (\widehat{\mathbf{J}}_\tau^{\mathrm{impl}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}})\widetilde{\mathbf{g}}_r \\ &\quad + (\mathbf{I} + \widehat{\mathbf{J}}_\tau^{\mathrm{impl}})(\widetilde{\mathbf{g}}_{r,\tau} - \widetilde{\mathbf{g}}_r),\end{aligned}$$

and therefore

$$\begin{aligned}\|\mathbf{v}_\tau - \mathbf{v}\|_2 &\le \left\|\widehat{\mathbf{J}}_\tau^{\mathrm{impl}} - \widehat{\mathbf{J}}_{\boldsymbol{\delta}_{\mathrm{rosu}}}\right\|_{\mathrm{op}}\|\widetilde{\mathbf{g}}_r\|_2 \\ &\quad + \left(1 + \frac{\rho}{\|\mathbf{q}_\tau\|_2}\right)\frac{2M_r\rho\|\mathbf{q}_\tau - \mathbf{q}\|_2}{\|\mathbf{q}\|_2} \\ &= O(\tau),\end{aligned}$$

where the inequality uses the triangle inequality, submultiplicativity, and the gradient bound. The final order follows from the two preceding estimates at this fixed nondegenerate point.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction reflect the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the main text.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We state the assumptions of each proposition and theorem in the main text and give the proofs in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide our experimental setups in the main text and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use public datasets and models.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide our experimental setups in the main text and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The vision tables and the accuracy curves show one standard deviation over ten forgotten classes for class-wise forgetting and over three seeds for random forgetting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the runtime of each method in the vision tables and the hardware in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss positive and negative impacts in the main text.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, architectures, pretrained models, and pipelines used in the main text and Appendix are cited to their original works and used in accordance with their public licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments in the main text and Appendix use existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so IRB approval is not applicable. All experiments in the main text and Appendix use existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs (Llama-3.2-1B-Instruct, Zephyr-7B-β) are used as standard base models for the unlearning experiments in the main text and Appendix, not as an original or non-standard component of the core methodology. We also used LLMs to assist with writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.