A DYNAMIC EXAMPLE SELECTION FRAMEWORK FOR COMPUTE-EFFICIENT MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove designated training data's influence while preserving utility at a lower cost than retraining. In sample-level unlearning, forget and retain examples can share classes and features, making their selection important for balancing forgetting and preservation. Repeated full-data updates incur substantial cost. We ask which examples should receive computation to achieve an existing unlearning method's performance at a lower cost. We observe that the initial low-confidence forget subset is enriched for examples with low true-label confidence after retraining on the retain set. Motivated by this observation, we introduce Dynamic Source Selection (DSS), which refreshes retain and forget subsets using the current model's true-label confidence while preserving the base loss and update rule. DSS+ reduces rescoring by keeping selected examples in training. A Gaussian logistic analysis identifies conditions under which updates on these sources improve prediction agreement with the retrained model while preserving retain performance. Empirically, dynamic selection allocates more training to these sources. Experiments on CIFAR-10 and CIFAR-100 demonstrate that DSS can attain accuracy comparable to full-data updates at lower computational cost across multiple methods, with TOFU experiments extending applicability to language models. These results support dynamic source selection as an effective strategy for efficient unlearning.

1 INTRODUCTION

Training-data deletion requests can arise after a machine learning model has been deployed (Bourtoule et al., 2021). Machine unlearning (MU) seeks to remove the influence of the designated forget set while preserving utility on the remaining retain set (Kurmanji et al., 2023). Retraining from scratch on the retain set provides a natural reference, but its computational cost motivates approximate updates to the existing model (Chourasia & Shah, 2023; Georgiev et al., 2025).

In sample-level unlearning, forget and retain examples can share classes and learned features, as in random and class-concentrated deletion. Updates to either set can therefore affect both forgetting and preservation (Cheng et al., 2026). Repeated full-data updates are costly, while restricting training to a fixed subset may miss examples that become useful later. Refreshing the subset can address this limitation, but repeatedly scoring examples also incurs computational overhead. We study how to allocate computation as the model changes: given a fixed deletion request, which retain and forget examples should receive updates to attain an existing method's performance at lower cost?

Our starting observation is that forget examples with low true-label confidence before unlearning are enriched for examples whose confidence is low after retraining on the retain set. A Gaussian logistic analysis of single-source updates identifies conditions under which such sources yield greater progress toward the retrained model's predictions while preserving retain performance, for both gradient ascent and relabeling. Confidence under the retrained model guides the source analysis, while the original model's confidence provides an observable proxy for initial selection. Our empirical analyses further show how dynamic selection concentrates training on these sources through method dependent selection transitions.

We propose Dynamic Source Selection (DSS), a method-agnostic framework that refreshes low-confidence retain and forget subsets each epoch. On the retain side, the score prioritizes predictions that need reinforcement; on the forget side, it is motivated by this source analysis. The base method

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

retains its loss formulation, target-generation rules, and optimizer update rule. Examples can enter or leave as their confidence changes. DSS+ reduces rescoring by keeping previously selected examples in training and scoring only those not yet admitted.

Experiments on CIFAR-10 and CIFAR-100 demonstrate that DSS can attain accuracy comparable to full-data updates with lower computation across multiple methods. On CIFAR-10, we compare equal-sized random and class-concentrated deletion requests. We also evaluate prediction agreement with the retrained model beyond accuracy and extend the framework to language models through TOFU experiments.

Our contributions are:

- **Framework.** We introduce DSS for dynamic retain and forget selection and DSS+ for reduced rescoring. DSS+ keeps selected examples in training and scores only those not yet admitted. Both integrate with existing unlearning methods while preserving their loss formulations and update rules, using the original labels for selection independently of the base method's training targets.
- **Selection mechanism.** We analyze source utility relative to a retrained model and connect it to initial confidence and selection transitions.
- **Empirical evaluation.** We demonstrate computational savings across multiple methods, evaluate prediction agreement beyond accuracy, and extend the framework to language models.

2 Related Work

Sample-level unlearning. Sample-level unlearning aims to remove the influence of specified training examples, with retraining on the remaining data as a reference. Bourtoule et al. (2021) shard training data to restrict retraining to the affected models; Kurmanji et al. (2023) train a student to match the original model on retain examples and diverge from it on forget examples; to target prediction agreement with a retrained model, Georgiev et al. (2025) estimate its predictions and fine-tune the model to match them. These approaches develop procedures for updating models after deletion; DSS determines which retain and forget examples contribute to an existing unlearning update as the model's confidence changes.

Data selection for unlearning. Distributional unlearning selects a removal subset within already identified unwanted-domain data to suppress the domain's statistical influence while preserving a retained distribution (Allouah et al., 2026). RASLIK retrieves forget and retain data from a corpus before unlearning, using gradient-based influence similarities (Liu et al., 2026). Within an unlearning procedure, ConMU combines EL2N-based selection from both pools with forget-input perturbations and matching a retain-trained proxy's predictions (Liu et al., 2024); SGA refreshes forget-sequence subsets between ascent epochs using current memorization scores (Bărbulescu & Triantafillou, 2024). For a fixed deletion request, DSS refreshes low-confidence subsets of both retain and forget pools each epoch while preserving the base method's loss formulation, target-generation rules, and update rule.

Coresets and dynamic data selection. For standard model training, CRAIG uses submodular optimization to select weighted subsets that approximate the training gradient (Mirzasoleiman et al., 2020); Grad-Match uses matching pursuit to approximate training or validation gradients (Killamsetty et al., 2021); InfoBatch prunes examples dynamically and rescales gradients to preserve their expectation (Qin et al., 2024). GraNd and EL2N rank examples by early-training gradient norms and prediction errors, respectively (Paul et al., 2021). In LLM unlearning, Pal et al. (2025) show that small forget subsets can match full-set results on WMDP and MUSE with sufficient repeated updates. These results motivate selecting which examples receive unlearning updates. DSS adapts this allocation over time using current confidence in both retain and forget pools; it does not aim to preserve the full-data gradient. In our selection-rule ablation, confidence-based selection reaches a given Avg. Gap with less cumulative computation than GraNd and InfoBatch adapted to unlearning, and with computation similar to EL2N, for both NegGrad+ and NPO (Figure 5).

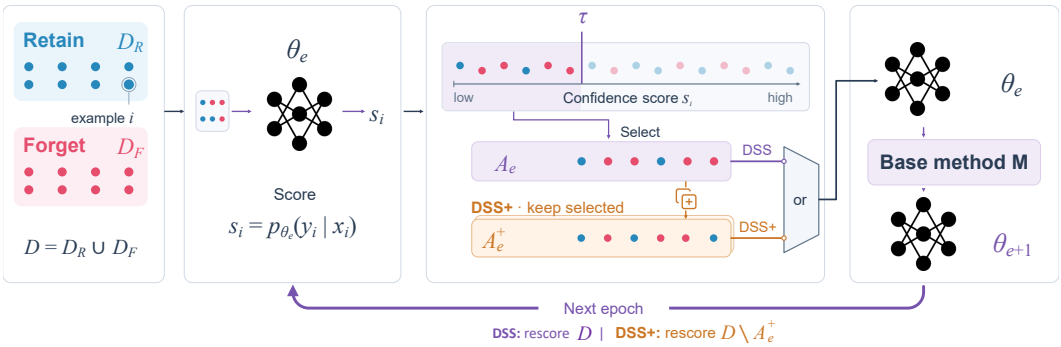


Figure 1: Overview of the DSS framework.

3 Dynamic Source Selection

3.1 Problem Setting

Let D_R and D_F index the retain and forget examples, respectively, and let $D = D_R \cup D_F$. Each index i corresponds to an original training example (x_i, y_i). FULL denotes the base unlearning method without source selection.

Let θ_0 denote the initial model trained on D. We compute its true-label confidence scores over D and set the threshold τ to their empirical ρ-quantile. This threshold remains fixed throughout unlearning. Thus, ρ determines the initial calibration level; subsequent selection fractions can differ between the retain and forget sets and change as the model evolves.

3.2 DSS: Epoch-wise Source Selection

Let θ_e denote the model at the start of epoch e in the policy being run. DSS scores all examples in D using θ_e in evaluation mode and selects those whose true-label confidence is at most τ:

$$A_e = \{i \in D : p_{\theta_e}(y_i \mid x_i) \leq \tau\}. \tag{1}$$

The initial calibration scores are reused for the first epoch, so DSS requires one scoring pass over D per training epoch. The selected set remains fixed during that epoch.

Scoring uses the original examples and labels. The base unlearning method trains on the selected examples according to their retain or forget membership, using its own input transformations, training targets, loss formulation, and update rules. Examples can enter or leave the selected set at the next refresh. Figure 1 illustrates this procedure, and Appendix A details the base-method integration.

3.3 DSS+: Reduced Rescoring

DSS+ reduces scoring overhead by continuing to train on previously selected examples without rescoring them. It starts from the same initial selection $A_0^+ = A_0$ and uses the same fixed threshold τ.

At each subsequent epoch, DSS+ scores only examples not yet selected and adds those meeting the threshold:

$$A_{e+1}^+ = A_e^+ \cup \{i \in D \setminus A_e^+ : p_{\theta_{e+1}}(y_i \mid x_i) \leq \tau\}.$$

Training follows the same base-method procedure using A_e^+. Previously selected examples remain in training even if their confidence later exceeds τ.

Algorithm 1 summarizes both policies and skips rescoring after the final epoch.

Algorithm 1: Dynamic Source Selection

Input: initial model θ_0, retain set D_R, forget set D_F, initial calibration level ρ, base method M, selection policy DSS or DSS+, maximum epochs $E \geq 1$, and a stopping rule.

Set $D \leftarrow D_R \cup D_F$ and initialize the base method's training state once.
Score $s_i \leftarrow p_{\theta_0}(y_i \mid x_i)$ for every $i \in D$ in evaluation mode.
Set τ to the empirical ρ-quantile of $\{s_i : i \in D\}$.
Select $A_0 \leftarrow \{i \in D : s_i \leq \tau\}$ *(DSS+: initialize $A_0^+ \leftarrow A_0$).*
for $e = 0, \ldots, E-1$ **do**
    Apply base method: run one epoch of M from θ_e on the currently selected examples, preserving their retain or forget membership, to obtain θ_{e+1} and the updated training state.
    if $e = E-1$ or the stopping condition is met, **break**.
    Refresh: rescore D using θ_{e+1} and form A_{e+1} from examples at or below the fixed threshold τ *(DSS+: rescore only $D \setminus A_e^+$ and add examples meeting the threshold to A_e^+ to form A_{e+1}^+).*
end for
return the updated model θ_{e+1}.

4 Selection Mechanism

We select low-confidence retain examples to prioritize predictions that need reinforcement. For forget examples, we ask which examples provide updates that move the model toward a model retrained on the retain set while preserving retained predictions. We first compare one-step updates from two forget examples in a logistic model, then examine how initial confidence and dynamic selection allocate training to the sources identified by this analysis.

4.1 Source utility relative to retraining

A forget example provides an update to the shared model parameters, so its usefulness depends on how that update changes predictions across inputs. Let θ_R denote the model retrained from scratch on D_R. We initialize two copies of θ_0, apply one update using forget examples L and H, respectively, and compare the resulting models.

Consider binary logistic prediction $p_\theta(1 \mid x) = \sigma(x^\top \theta)$, where x is an input vector, θ is the model's parameter vector, $\sigma(z) = (1 + e^{-z})^{-1}$, and the label y belongs to $\{-1, +1\}$. For orthogonal unit vectors u, v, assume

$$\theta_R = \beta v, \qquad \theta_0 = \beta v + \gamma u, \qquad 0 < \gamma \leq \beta. \tag{2}$$

Here v is the direction of θ_R and u the direction of $\theta_0 - \theta_R$; β is the norm of θ_R and γ the norm of that difference. Thus γu is the difference to be corrected. Retain inputs are centered Gaussian, with variance $\varepsilon \in (0, 1]$ along u and independent unit variance in orthogonal directions; their labels follow p_{θ_R}. The parameter ε controls how much retained inputs vary along the difference direction, so correcting that difference can also affect retained predictions.

We measure prediction agreement by the average KL divergence

$$Q(\theta) = \mathbb{E}_{X \sim \mathcal{N}(0,I)} \, \mathrm{KL}(p_{\theta_R}(\cdot \mid X) \,\|\, p_\theta(\cdot \mid X)) \,. \tag{3}$$

Here $\mathcal{N}(0, I)$ is a zero-mean Gaussian with identity covariance, and the expectation averages over these reference inputs. Smaller Q means closer predicted label distributions.

For each forget example (x_i, y_i), indexed by $i \in \{L, H\}$, write $c_i^0 = p_{\theta_0}(y_i \mid x_i)$ and $c_i^R = p_{\theta_R}(y_i \mid x_i)$ for its true-label confidence under θ_0 and θ_R. Assume $\|x_i\| = 1$ and $c_i^0 \geq c_i^R$. With $a_i = y_i x_i$ and a common learning rate $\eta > 0$, one true-label cross-entropy ascent update gives

$$\theta_{i,1} = \theta_0 - \eta(1 - c_i^0) a_i,$$

where $\theta_{i,1}$ denotes the parameters after this update.

To isolate the role of confidence under θ_R, we match initial confidence and compare

$$c_L^0 = c_H^0 =: c^0, \qquad c_L^R < c_H^R.$$

The two updates then have the same magnitude, $\eta(1 - c^0)$, and differ only in their source directions a_L, a_H. To relate these directions to the confidence comparison, define $\mathrm{logit}(c) = \log[c/(1-c)]$ and

$$r_i = \mathrm{logit}(c_i^0) - \mathrm{logit}(c_i^R) = a_i^\top (\theta_0 - \theta_R) = \gamma u^\top a_i.$$

Thus, r_i measures both the decrease in true-label log odds after retraining and the source component along the initial model difference. Because logit is increasing and the initial confidence is shared, $r_L - r_H = \text{logit}(c_H^R) - \text{logit}(c_L^R) > 0$. The L update therefore has a larger component that reduces this difference.

Proposition 1. *Under the preceding model and source conditions, for all sufficiently small common learning rates $\eta > 0$, one CE-ascent update satisfies*

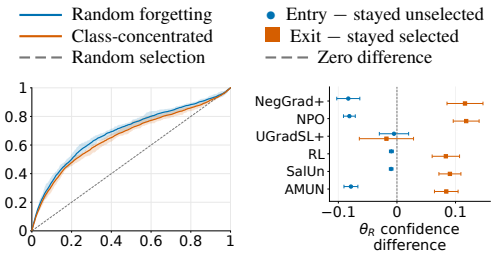
$$Q(\theta_{L,1}) < Q(\theta_{H,1}). \tag{4}$$

The L update also yields lower retain cross-entropy and misclassification probability than the H update.

Improvement over θ_0 requires the additional conditions in Appendix C.3. Appendix C provides the proof and extensions to repeated updates and unequal initial confidence.

4.2 From Initial Confidence to Dynamic Allocation

DSS uses the current model's scores; θ_R is an offline reference. The initial low-confidence forget subset has lower mean confidence under θ_R than the whole forget set (Table 16). True-label confidences under θ_0 and θ_R are positively associated. Figure 2(a) plots the forget fraction selected by increasing true-label confidence under θ_0 (x) against its share of the total positive confidence decrease from θ_0 to θ_R (y). Under both deletion settings, low-confidence examples account for a disproportionate share of this decrease. Shading shows the range across deletion requests.



(a) Initial confidence (b) Selection transitions

Figure 2: Initial confidence and selection transitions on CIFAR-10. (a) Selected forget fraction (x) and its share of the positive confidence decrease after retraining (y). (b) Mean confidence difference under θ_R (x), by base method (y).

For random forgetting, we examine whether dynamic refreshes favor these sources beyond their initial scores. After stratifying forget examples by confidence under θ_0, entrants generally have lower mean true-label confidence under θ_R than those remaining unselected, while exits have higher confidence than those remaining selected. This pattern is less consistent for UGradSL+ (Figure 2(b)). Panel (b) plots these mean differences horizontally for each base method, with bars showing the sample SD across deletion requests. The unstratified comparisons show a similar method-dependent pattern (Table 18). Appendix D details both analyses.

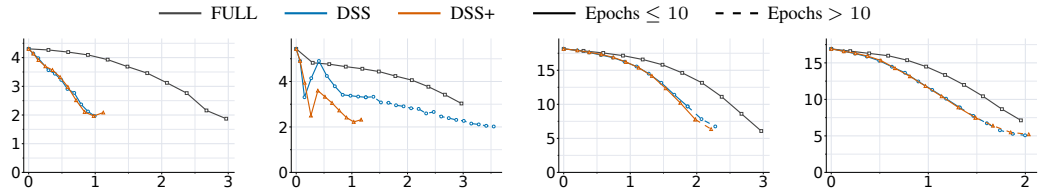
5 Experiments

5.1 Experiment Setup

Datasets and Models. We evaluate DSS on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). We use ResNet-18 (He et al., 2016) on both datasets and additionally evaluate VGG16 (Simonyan & Zisserman, 2015; Ioffe & Szegedy, 2015) on CIFAR-100. Training and implementation details are provided in Appendix A.

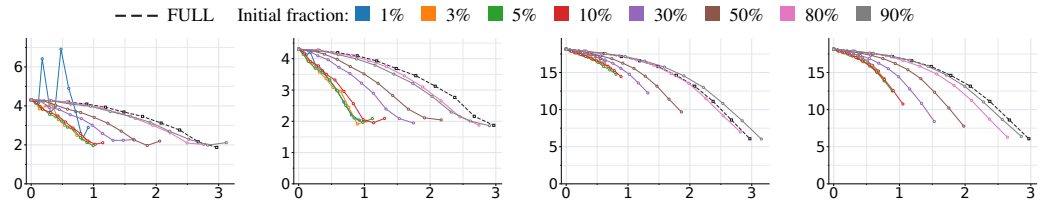
Baseline Methods. We use a model retrained from scratch on the retain set as a reference and apply DSS and DSS+ to six unlearning methods: NegGrad+ (Kurmanji et al., 2023), NPO (Zhang et al., 2024), UGradSL+ (Di et al., 2026), random labeling (RL) (Golatkar et al., 2020), SalUn (Fan et al., 2024), and AMUN (Ebrahimpour-Boroojeny et al., 2025). FULL uses the complete available data pools, while DSS and DSS+ apply sample selection while preserving the base loss formulation. Tables 1 and 2 compare FULL, DSS, and DSS+ after ten unlearning epochs.

Evaluation Metrics. Following Jia et al. (2023); Fan et al. (2024), we report retain accuracy (RA), unlearning accuracy (UA), test accuracy (TA), membership inference (MIA), and runtime efficiency (RTE). We summarize agreement with the retrained model using Avg. Gap and measure predictive-distribution differences using Jensen–Shannon (JS) divergence. We also report cumulative estimated



(a) Table 1 (left): NegGrad+ (b) Table 1 (right): NegGrad+ (c) Table 2 (left): NegGrad+ (d) Table 2 (right): NPO

Figure 3: Unlearning trajectories for the indicated table settings. Axes show cumulative computation (PFLOPs; x) and Avg. Gap (percentage points; y).



(a) Table 1 (left): DSS (b) Table 1 (left): DSS+ (c) Table 2 (left): DSS (d) Table 2 (left): DSS+

Figure 4: Effect of the initial selection fraction for NegGrad+ with ResNet-18 under random forgetting. Panels (a,b): CIFAR-10; (c,d): CIFAR-100. Axes show cumulative computation (PFLOPs; x) and Avg. Gap (percentage points; y).

computation in PFLOPs (10^{15} floating-point operations). Both RTE and PFLOPs include selection-scoring costs (Appendix B.1). Metric definitions are provided in Appendix A.4.

Unlearning Scenarios. We consider random and class-concentrated forgetting. Random forgetting removes 10% of the training examples sampled uniformly across all classes. Class-concentrated forgetting removes 20% of the training examples from each of five fixed CIFAR-10 classes, resulting in the same overall 10% deletion budget while retaining examples from every class.

LLM Setup. We evaluate LLM unlearning on TOFU (Maini et al., 2024) using Llama-3.2-1B-Instruct (Meta, 2024). We compare FULL and DSS for GradDiff (Maini et al., 2024), NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), and RMU (Li et al., 2024). Using an evaluation protocol adapted from OpenUnlearning (Dorna et al., 2025), we report normalized forgetting (Mem.), privacy (Priv.), and utility scores, and their harmonic mean (Agg.). We also report algorithm runtime in minutes (RTE).

5.2 Results

Performance. On CIFAR-10, DSS and DSS+ generally reduce runtime while attaining Avg. Gap values comparable to FULL under random forgetting (Table 1, left). DSS and DSS+ also generally reduce runtime while maintaining Avg. Gap comparable to FULL for class-concentrated deletion requests (Table 1, right).

On CIFAR-100, both DSS and DSS+ reduce runtime with ResNet-18 (Table 2, left). With VGG16, DSS+ reduces runtime across all six methods, whereas DSS is slower than FULL for some methods (Table 2, right). For RL on VGG16, DSS+ shows a larger mean Avg. Gap and greater variation across runs than FULL (Appendix B.2). Some methods retain larger Avg. Gap values, particularly NegGrad+ and NPO with ResNet-18.

The initial selection fraction should also be considered together with training duration. A more selective configuration may require more epochs to reach a given performance level, while each epoch uses less computation. Comparisons at a common epoch horizon therefore capture only part of its potential benefit. The extended trajectories in Figure 3 show that additional training can bring selected variants closer to the FULL reference while retaining substantial computational savings. This motivates using some of the saved computation for further updates, with the training endpoint chosen alongside the initial fraction.

Table 1: Random and class-concentrated forgetting on CIFAR-10 with ResNet-18. The lowest Avg. Gap and RTE within each method and setting are bold.

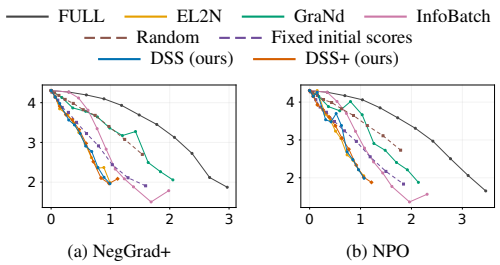
	Random						Class-concentrated					
Method	RA	UA	TA	MIA	Avg. Gap(↓)	RTE(↓, min)	RA	UA	TA	MIA	Avg. Gap(↓)	RTE(↓, min)
Retrain	$100.00_{\pm 0.00}$	$5.04_{\pm 0.33}$	$94.73_{\pm 0.14}$	$12.02_{\pm 0.42}$	$0.00_{\pm 0.00}$	$14.72_{\pm 0.55}$	$100.00_{\pm 0.00}$	$6.50_{\pm 0.32}$	$94.66_{\pm 0.22}$	$15.07_{\pm 0.84}$	$0.00_{\pm 0.00}$	$14.78_{\pm 0.50}$
NegGrad+	$99.01_{\pm 0.40}$	$5.44_{\pm 0.77}$	$92.61_{\pm 0.56}$	$8.18_{\pm 1.06}$	$\mathbf{1.87}_{\pm 0.16}$	$2.68_{\pm 0.07}$	$98.14_{\pm 0.15}$	$4.75_{\pm 0.52}$	$90.93_{\pm 0.17}$	$10.29_{\pm 0.81}$	$3.03_{\pm 0.24}$	$2.70_{\pm 0.09}$
+DSS	$98.85_{\pm 0.44}$	$5.42_{\pm 0.60}$	$92.30_{\pm 0.34}$	$8.30_{\pm 0.98}$	$1.96_{\pm 0.17}$	$\mathbf{0.91}_{\pm 0.03}$	$96.72_{\pm 0.18}$	$7.41_{\pm 0.39}$	$89.60_{\pm 0.22}$	$19.05_{\pm 1.82}$	$3.30_{\pm 0.56}$	$1.14_{\pm 0.01}$
+DSS+	$98.76_{\pm 0.46}$	$7.20_{\pm 0.73}$	$92.00_{\pm 0.77}$	$9.82_{\pm 0.86}$	$2.09_{\pm 0.23}$	$1.03_{\pm 0.05}$	$97.42_{\pm 0.28}$	$5.86_{\pm 0.71}$	$90.09_{\pm 0.32}$	$14.14_{\pm 1.12}$	$\mathbf{2.32}_{\pm 0.27}$	$\mathbf{1.06}_{\pm 0.01}$
NPO	$99.44_{\pm 0.16}$	$5.23_{\pm 0.68}$	$93.15_{\pm 0.35}$	$7.87_{\pm 0.74}$	$\mathbf{1.67}_{\pm 0.09}$	$2.86_{\pm 0.07}$	$98.90_{\pm 0.12}$	$2.98_{\pm 0.48}$	$91.84_{\pm 0.18}$	$7.80_{\pm 0.75}$	$3.68_{\pm 0.26}$	$2.87_{\pm 0.07}$
+DSS	$99.31_{\pm 0.30}$	$4.48_{\pm 0.64}$	$92.91_{\pm 0.53}$	$7.33_{\pm 0.69}$	$1.99_{\pm 0.15}$	$\mathbf{0.94}_{\pm 0.04}$	$96.91_{\pm 0.16}$	$6.92_{\pm 0.33}$	$89.77_{\pm 0.22}$	$19.22_{\pm 2.30}$	$3.22_{\pm 0.29}$	$1.17_{\pm 0.01}$
+DSS+	$99.20_{\pm 0.28}$	$5.82_{\pm 0.48}$	$92.56_{\pm 0.45}$	$8.26_{\pm 0.53}$	$1.88_{\pm 0.22}$	$1.05_{\pm 0.05}$	$97.73_{\pm 0.20}$	$5.15_{\pm 0.50}$	$90.45_{\pm 0.21}$	$13.41_{\pm 0.89}$	$\mathbf{2.40}_{\pm 0.27}$	$\mathbf{1.10}_{\pm 0.01}$
UGradSL+	$98.32_{\pm 1.51}$	$5.77_{\pm 1.69}$	$91.73_{\pm 1.64}$	$15.86_{\pm 3.52}$	$2.45_{\pm 1.57}$	$2.73_{\pm 0.09}$	$98.08_{\pm 0.17}$	$4.84_{\pm 0.56}$	$90.86_{\pm 0.17}$	$10.55_{\pm 0.79}$	$2.98_{\pm 0.28}$	$2.73_{\pm 0.09}$
+DSS	$99.95_{\pm 0.06}$	$3.05_{\pm 0.28}$	$94.12_{\pm 0.27}$	$10.40_{\pm 0.77}$	$1.08_{\pm 0.22}$	$2.15_{\pm 0.06}$	$96.40_{\pm 0.34}$	$8.08_{\pm 0.78}$	$89.20_{\pm 0.26}$	$18.80_{\pm 1.26}$	$3.60_{\pm 0.71}$	$1.17_{\pm 0.01}$
+DSS+	$99.89_{\pm 0.13}$	$3.08_{\pm 0.19}$	$93.97_{\pm 0.40}$	$11.23_{\pm 1.79}$	$\mathbf{1.06}_{\pm 0.33}$	$\mathbf{1.95}_{\pm 0.04}$	$97.03_{\pm 0.19}$	$6.69_{\pm 0.50}$	$89.67_{\pm 0.18}$	$15.25_{\pm 0.46}$	$\mathbf{2.31}_{\pm 0.25}$	$\mathbf{1.09}_{\pm 0.01}$
RL	$99.07_{\pm 0.29}$	$3.67_{\pm 0.46}$	$93.41_{\pm 0.40}$	$13.63_{\pm 0.63}$	$\mathbf{1.31}_{\pm 0.20}$	$1.47_{\pm 0.04}$	$98.40_{\pm 0.20}$	$4.18_{\pm 0.75}$	$92.61_{\pm 0.27}$	$12.59_{\pm 1.96}$	$2.19_{\pm 0.72}$	$1.48_{\pm 0.05}$
+DSS	$99.88_{\pm 0.07}$	$3.03_{\pm 0.38}$	$94.21_{\pm 0.31}$	$16.65_{\pm 1.18}$	$1.82_{\pm 0.30}$	$\mathbf{1.30}_{\pm 0.03}$	$98.01_{\pm 1.17}$	$4.43_{\pm 1.80}$	$92.21_{\pm 0.97}$	$15.10_{\pm 3.34}$	$2.32_{\pm 0.66}$	$\mathbf{1.29}_{\pm 0.08}$
+DSS+	$99.41_{\pm 0.19}$	$2.85_{\pm 0.43}$	$93.74_{\pm 0.30}$	$13.48_{\pm 0.74}$	$\mathbf{1.31}_{\pm 0.17}$	$1.31_{\pm 0.01}$	$97.55_{\pm 0.67}$	$5.09_{\pm 1.42}$	$92.01_{\pm 0.74}$	$14.31_{\pm 2.26}$	$\mathbf{2.14}_{\pm 0.39}$	$1.33_{\pm 0.07}$
SalUn	$98.13_{\pm 0.51}$	$4.63_{\pm 0.64}$	$92.51_{\pm 0.50}$	$13.83_{\pm 0.62}$	$1.62_{\pm 0.29}$	$1.67_{\pm 0.05}$	$96.80_{\pm 1.15}$	$5.71_{\pm 0.71}$	$91.44_{\pm 1.01}$	$15.05_{\pm 1.46}$	$\mathbf{2.02}_{\pm 0.55}$	$1.66_{\pm 0.05}$
+DSS	$99.82_{\pm 0.12}$	$2.90_{\pm 0.45}$	$94.11_{\pm 0.31}$	$15.42_{\pm 1.19}$	$1.58_{\pm 0.18}$	$\mathbf{1.39}_{\pm 0.04}$	$97.54_{\pm 0.76}$	$5.28_{\pm 1.77}$	$91.89_{\pm 0.68}$	$16.81_{\pm 3.04}$	$2.38_{\pm 0.80}$	$\mathbf{1.45}_{\pm 0.05}$
+DSS+	$98.63_{\pm 0.30}$	$3.83_{\pm 0.38}$	$92.90_{\pm 0.24}$	$13.15_{\pm 0.80}$	$\mathbf{1.38}_{\pm 0.19}$	$1.47_{\pm 0.03}$	$96.36_{\pm 1.10}$	$6.65_{\pm 2.15}$	$91.19_{\pm 0.97}$	$17.23_{\pm 3.60}$	$3.13_{\pm 1.36}$	$1.52_{\pm 0.05}$
AMUN	$99.55_{\pm 0.12}$	$5.54_{\pm 0.56}$	$93.45_{\pm 0.34}$	$11.04_{\pm 0.64}$	$0.83_{\pm 0.19}$	$2.08_{\pm 0.11}$	$99.40_{\pm 0.28}$	$6.99_{\pm 0.79}$	$93.10_{\pm 0.38}$	$13.72_{\pm 1.39}$	$\mathbf{1.16}_{\pm 0.27}$	$2.10_{\pm 0.05}$
+DSS	$99.55_{\pm 0.20}$	$5.47_{\pm 0.30}$	$93.47_{\pm 0.53}$	$11.26_{\pm 0.51}$	$\mathbf{0.73}_{\pm 0.27}$	$\mathbf{1.24}_{\pm 0.03}$	$99.28_{\pm 0.40}$	$6.52_{\pm 1.69}$	$93.08_{\pm 0.67}$	$13.51_{\pm 2.48}$	$1.54_{\pm 0.54}$	$\mathbf{1.29}_{\pm 0.05}$
+DSS+	$99.49_{\pm 0.26}$	$6.02_{\pm 0.79}$	$93.46_{\pm 0.40}$	$12.24_{\pm 1.08}$	$0.83_{\pm 0.46}$	$1.31_{\pm 0.06}$	$99.49_{\pm 0.21}$	$6.97_{\pm 1.36}$	$93.29_{\pm 0.50}$	$14.00_{\pm 2.01}$	$1.22_{\pm 0.46}$	$1.35_{\pm 0.05}$

Table 2: Random forgetting on CIFAR-100 with ResNet-18 and VGG16. The lowest Avg. Gap and RTE within each method and architecture are bold.

	ResNet-18						VGG16					
Method	RA	UA	TA	MIA	Avg. Gap(↓)	RTE(↓, min)	RA	UA	TA	MIA	Avg. Gap(↓)	RTE(↓, min)
Retrain	$99.98_{\pm 0.00}$	$24.03_{\pm 0.53}$	$75.90_{\pm 0.31}$	$49.78_{\pm 0.74}$	$0.00_{\pm 0.00}$	$16.68_{\pm 0.75}$	$99.97_{\pm 0.01}$	$27.32_{\pm 0.58}$	$72.53_{\pm 0.23}$	$40.71_{\pm 0.53}$	$0.00_{\pm 0.00}$	$19.38_{\pm 0.14}$
NegGrad+	$98.59_{\pm 0.27}$	$22.22_{\pm 1.70}$	$70.92_{\pm 0.52}$	$33.89_{\pm 1.47}$	$\mathbf{6.08}_{\pm 0.57}$	$2.65_{\pm 0.05}$	$99.80_{\pm 0.08}$	$3.74_{\pm 0.84}$	$71.99_{\pm 0.38}$	$7.10_{\pm 0.82}$	$14.47_{\pm 0.43}$	$1.91_{\pm 0.02}$
+DSS	$98.76_{\pm 0.71}$	$14.95_{\pm 5.09}$	$71.73_{\pm 1.12}$	$25.88_{\pm 4.20}$	$9.69_{\pm 1.70}$	$\mathbf{1.70}_{\pm 0.04}$	$98.55_{\pm 1.87}$	$10.10_{\pm 8.32}$	$70.02_{\pm 2.09}$	$13.93_{\pm 8.83}$	$\mathbf{12.11}_{\pm 2.92}$	$\mathbf{1.36}_{\pm 0.12}$
+DSS+	$98.70_{\pm 0.32}$	$18.75_{\pm 2.70}$	$71.34_{\pm 0.91}$	$29.80_{\pm 2.22}$	$7.77_{\pm 1.00}$	$1.79_{\pm 0.03}$	$99.24_{\pm 0.53}$	$9.10_{\pm 3.76}$	$70.46_{\pm 1.14}$	$12.83_{\pm 4.16}$	$12.22_{\pm 1.62}$	$1.40_{\pm 0.07}$
NPO	$99.40_{\pm 0.11}$	$22.46_{\pm 1.40}$	$72.29_{\pm 0.39}$	$34.23_{\pm 1.20}$	$\mathbf{5.37}_{\pm 0.43}$	$2.84_{\pm 0.05}$	$97.13_{\pm 0.30}$	$20.77_{\pm 0.83}$	$68.36_{\pm 0.36}$	$25.76_{\pm 0.90}$	$\mathbf{7.13}_{\pm 0.35}$	$1.99_{\pm 0.01}$
+DSS	$99.59_{\pm 0.10}$	$12.68_{\pm 2.21}$	$73.25_{\pm 0.43}$	$22.29_{\pm 2.41}$	$10.47_{\pm 1.10}$	$\mathbf{1.73}_{\pm 0.01}$	$97.63_{\pm 0.63}$	$17.16_{\pm 2.79}$	$68.44_{\pm 0.74}$	$21.75_{\pm 2.78}$	$8.89_{\pm 1.10}$	$\mathbf{1.44}_{\pm 0.03}$
+DSS+	$99.49_{\pm 0.05}$	$16.59_{\pm 1.84}$	$72.91_{\pm 0.34}$	$26.85_{\pm 1.65}$	$8.46_{\pm 0.85}$	$1.87_{\pm 0.02}$	$97.10_{\pm 0.66}$	$20.42_{\pm 4.74}$	$68.29_{\pm 0.74}$	$25.51_{\pm 4.84}$	$7.49_{\pm 1.82}$	$1.57_{\pm 0.08}$
UGradSL+	$99.58_{\pm 0.07}$	$27.87_{\pm 1.83}$	$68.86_{\pm 0.53}$	$54.13_{\pm 1.27}$	$3.91_{\pm 0.96}$	$2.70_{\pm 0.06}$	$99.41_{\pm 0.44}$	$6.65_{\pm 1.62}$	$69.98_{\pm 1.28}$	$20.98_{\pm 2.26}$	$\mathbf{10.88}_{\pm 0.71}$	$2.28_{\pm 0.12}$
+DSS	$99.55_{\pm 0.10}$	$23.83_{\pm 1.31}$	$69.11_{\pm 0.56}$	$45.54_{\pm 1.21}$	$3.17_{\pm 0.43}$	$2.40_{\pm 0.04}$	$99.58_{\pm 0.20}$	$5.75_{\pm 0.85}$	$70.47_{\pm 0.68}$	$19.87_{\pm 1.84}$	$11.22_{\pm 0.58}$	$2.43_{\pm 0.23}$
+DSS+	$99.54_{\pm 0.08}$	$25.63_{\pm 1.53}$	$68.88_{\pm 0.51}$	$47.45_{\pm 1.24}$	$\mathbf{2.91}_{\pm 0.24}$	$\mathbf{2.19}_{\pm 0.01}$	$99.26_{\pm 0.74}$	$6.60_{\pm 2.31}$	$69.83_{\pm 1.58}$	$20.61_{\pm 2.39}$	$11.05_{\pm 0.75}$	$\mathbf{2.15}_{\pm 0.20}$
RL	$96.52_{\pm 0.20}$	$16.45_{\pm 0.96}$	$72.85_{\pm 0.25}$	$34.51_{\pm 1.36}$	$7.34_{\pm 0.49}$	$1.45_{\pm 0.03}$	$98.22_{\pm 0.29}$	$15.75_{\pm 0.33}$	$70.24_{\pm 0.41}$	$29.66_{\pm 0.63}$	$6.67_{\pm 0.25}$	$1.07_{\pm 0.01}$
+DSS	$99.33_{\pm 0.16}$	$19.08_{\pm 1.86}$	$73.23_{\pm 0.48}$	$47.54_{\pm 2.45}$	$\mathbf{2.70}_{\pm 0.88}$	$1.25_{\pm 0.05}$	$96.06_{\pm 3.75}$	$18.95_{\pm 4.39}$	$68.72_{\pm 2.76}$	$31.84_{\pm 3.44}$	$\mathbf{6.32}_{\pm 1.60}$	$1.14_{\pm 0.03}$
+DSS+	$98.57_{\pm 1.08}$	$16.32_{\pm 1.84}$	$73.71_{\pm 0.95}$	$41.01_{\pm 5.38}$	$5.02_{\pm 1.87}$	$\mathbf{1.24}_{\pm 0.05}$	$87.48_{\pm 26.41}$	$23.76_{\pm 22.13}$	$63.78_{\pm 17.66}$	$34.16_{\pm 17.36}$	$12.92_{\pm 17.22}$	$\mathbf{1.05}_{\pm 0.02}$
SalUn	$91.78_{\pm 2.26}$	$17.49_{\pm 2.16}$	$70.54_{\pm 1.17}$	$27.79_{\pm 1.53}$	$10.52_{\pm 0.66}$	$1.64_{\pm 0.03}$	$99.48_{\pm 0.13}$	$12.22_{\pm 0.56}$	$71.80_{\pm 0.45}$	$30.88_{\pm 0.80}$	$6.57_{\pm 0.34}$	$1.20_{\pm 0.01}$
+DSS	$98.27_{\pm 0.67}$	$12.46_{\pm 0.91}$	$72.93_{\pm 0.41}$	$33.89_{\pm 3.41}$	$\mathbf{8.04}_{\pm 1.25}$	$1.50_{\pm 0.07}$	$99.52_{\pm 0.18}$	$12.51_{\pm 1.04}$	$71.86_{\pm 0.41}$	$31.92_{\pm 1.82}$	$\mathbf{6.18}_{\pm 0.84}$	$1.30_{\pm 0.02}$
+DSS+	$93.84_{\pm 1.45}$	$16.02_{\pm 1.86}$	$71.05_{\pm 1.09}$	$28.65_{\pm 1.50}$	$10.04_{\pm 0.50}$	$\mathbf{1.48}_{\pm 0.04}$	$99.12_{\pm 0.92}$	$11.60_{\pm 0.99}$	$71.54_{\pm 0.50}$	$28.82_{\pm 4.06}$	$7.36_{\pm 1.64}$	$\mathbf{1.17}_{\pm 0.01}$
AMUN	$96.69_{\pm 0.84}$	$23.07_{\pm 0.90}$	$69.80_{\pm 0.87}$	$33.54_{\pm 1.48}$	$\mathbf{6.71}_{\pm 0.58}$	$2.03_{\pm 0.06}$	$95.30_{\pm 0.80}$	$29.03_{\pm 0.86}$	$67.53_{\pm 0.68}$	$38.03_{\pm 1.41}$	$3.59_{\pm 0.53}$	$1.64_{\pm 0.03}$
+DSS	$97.39_{\pm 0.60}$	$18.54_{\pm 1.22}$	$70.60_{\pm 0.90}$	$28.02_{\pm 1.25}$	$8.79_{\pm 0.61}$	$\mathbf{1.48}_{\pm 0.04}$	$93.71_{\pm 3.20}$	$28.65_{\pm 2.16}$	$66.20_{\pm 2.03}$	$34.79_{\pm 1.68}$	$5.10_{\pm 1.64}$	$1.51_{\pm 0.10}$
+DSS+	$97.61_{\pm 0.58}$	$20.75_{\pm 1.02}$	$71.01_{\pm 0.51}$	$32.26_{\pm 1.65}$	$7.02_{\pm 0.70}$	$1.63_{\pm 0.04}$	$95.58_{\pm 0.90}$	$28.36_{\pm 1.15}$	$67.97_{\pm 0.77}$	$37.58_{\pm 1.64}$	$\mathbf{3.40}_{\pm 0.77}$	$\mathbf{1.48}_{\pm 0.04}$

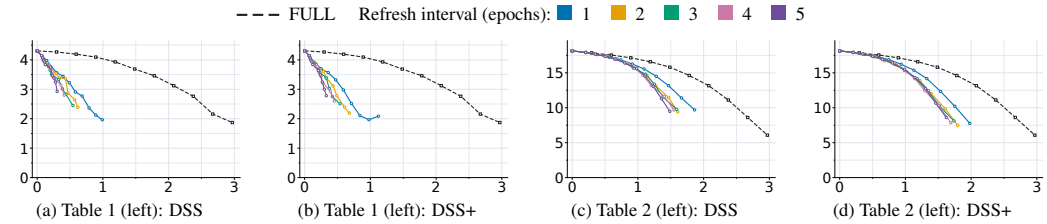
Initial selection fraction. With rescoring every epoch, smaller initial fractions generally reduce cumulative computation, while their effect on Avg. Gap depends on the dataset and policy (Figure 4). Very small initial fractions can produce pronounced fluctuations in Avg. Gap, as seen for NegGrad+ with DSS at 1% on CIFAR-10. The fraction sets the initial threshold. Subsequent subset sizes evolve with confidence.

Effect of the selection rule. On CIFAR-10 with ResNet-18 under random forgetting, Figure 5 compares DSS and DSS+ with EL2N, GraNd (Paul et al., 2021), InfoBatch (Qin et al., 2024), Random, and Fixed initial scores for NegGrad+ and NPO. Random and Fixed initial scores expand their subsets linearly from 5% to 100%, retaining earlier selections and adding examples in random order or ascending order of initial confidence, respectively. For EL2N and GraNd, we replace DSS's confidence score while retaining its selection and refresh procedure. To accommodate the retain and forget objectives, we adapt InfoBatch's scoring and loss weighting. Implementation details are given in Appendix A.3.



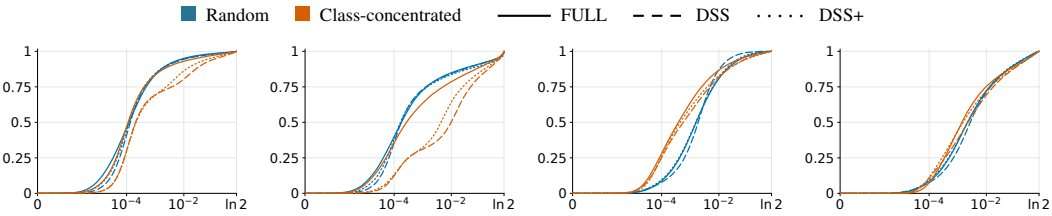
(a) NegGrad+ (b) NPO

Figure 5: Selection-rule comparison: cumulative PFLOPs (x) and Avg. Gap (y).



(a) Table 1 (left): DSS (b) Table 1 (left): DSS+ (c) Table 2 (left): DSS (d) Table 2 (left): DSS+

Figure 6: Effect of the refresh interval for NegGrad+ with ResNet-18 under random forgetting. Panels (a,b): CIFAR-10; (c,d): CIFAR-100. Axes show cumulative computation (PFLOPs; x) and Avg. Gap (percentage points; y).



(a) NegGrad+ — Retain **(b) NegGrad+ — Forget** **(c) RL — Retain** **(d) RL — Forget**

Figure 7: Diagnostic JS distributions relative to the retrained model. Panels (a,c) show retain examples and (b,d) show forget examples. Axes show per-example JS divergence (nats; x) and its CDF (y).

For both NegGrad+ and NPO, confidence-based selection generally reaches comparable Avg. Gap values with less cumulative computation. Like confidence-based selection, EL2N prioritizes examples with larger prediction errors relative to the true label and exhibits broadly similar computation–gap trajectories, with transient increases in Avg. Gap.

Refresh interval. We vary the refresh interval from one to five epochs, using initial selection fractions of 5% for CIFAR-10 and 50% for CIFAR-100 (Figure 6). Selected subsets remain fixed between refreshes. Longer intervals reduce scoring passes and cumulative computation. Reusing selected sources over multiple epochs may also benefit unlearning. On CIFAR-10, several longer refresh intervals reach comparable gaps with less computation than rescoring every epoch, although their gaps are higher at the endpoint. On CIFAR-100, less frequent rescoring can reduce both Avg. Gap and computation. Additional trajectories and sensitivity analyses across all six base methods are provided in Appendix B.2.

DSS versus DSS+. DSS and DSS+ start from the same fixed threshold τ and initial selected set, but differ in how their selected sets evolve. Under DSS, examples can enter or leave the selected set at each refresh. Under DSS+, previously selected examples remain in training even if their confidence later exceeds the threshold, while only the remaining unselected examples are rescored. The transition statistics in Appendix D.2 reflect this difference: DSS continues to add and remove examples across epochs, whereas DSS+ retains substantially more examples across consecutive epochs.

This persistence also changes the training updates. For NegGrad+ and NPO, Appendix D.3 shows that DSS+ has smaller average total gradient norms than DSS, consistent with updates being distributed over a larger persistent selected set. At the same time, keeping more examples in training can increase the amount of training computation. The runtime breakdown in Appendix B.1 shows that DSS+ reduces scoring time but can spend more time on training than DSS, so lower rescoring cost does not necessarily translate into lower total runtime. These differences help explain why DSS and DSS+ can follow different optimization trajectories and why their runtime ordering varies across methods and settings.

Prediction differences beyond accuracy. Training on selected examples may affect predicted probabilities in ways that RA and UA do not capture. Figure 7 therefore compares per-example JS divergence from the retrained model for FULL, DSS, and DSS+, separately on retain and forget examples. We evaluate NegGrad+ and RL under random and class-concentrated forgetting.

The three policies show broadly similar JS distributions for RL and for NegGrad+ under random forgetting. Under class-concentrated forgetting, NegGrad+ with DSS and DSS+ shows a clearer shift toward larger JS divergences, particularly on forget examples. JS distributions for all six base methods are provided in Appendix D.5.

5.3 LLM UNLEARNING

For LLMs, we apply DSS to answer tokens, refreshing their selection within each visited minibatch after the initial epoch. Appendix A.5 details the selection procedure and its integration with the base methods (Algorithm 2).

Method	Variant	Mem. ↑	Priv. ↑	Utility ↑	Agg. ↑	RTE (min) ↓
Retrain	–	0.3131	0.9964	0.9921	0.5763	7.664
GradDiff	FULL	$0.9681_{\pm 0.0033}$	$0.0106_{\pm 0.0173}$	$0.3727_{\pm 0.2285}$	$\mathbf{0.0270}_{\pm 0.0420}$	$\mathbf{2.336}_{\pm 0.344}$
	+DSS	$0.9667_{\pm 0.0044}$	$0.0094_{\pm 0.0199}$	$0.1934_{\pm 0.1100}$	$0.0237_{\pm 0.0500}$	$3.194_{\pm 0.615}$
NPO	FULL	$0.2732_{\pm 0.0077}$	$0.9059_{\pm 0.0391}$	$0.9883_{\pm 0.0060}$	$0.5192_{\pm 0.0100}$	$3.805_{\pm 0.083}$
	+DSS	$0.2837_{\pm 0.0057}$	$0.8457_{\pm 0.0350}$	$0.9908_{\pm 0.0131}$	$\mathbf{0.5246}_{\pm 0.0068}$	$\mathbf{2.284}_{\pm 0.051}$
SimNPO	FULL	$0.2626_{\pm 0.0045}$	$0.8204_{\pm 0.0121}$	$0.5428_{\pm 0.0793}$	$0.4348_{\pm 0.0190}$	$2.762_{\pm 0.070}$
	+DSS	$0.2947_{\pm 0.0059}$	$0.8392_{\pm 0.0223}$	$0.5837_{\pm 0.0823}$	$\mathbf{0.4742}_{\pm 0.0167}$	$\mathbf{2.093}_{\pm 0.033}$
RMU	FULL	$0.2745_{\pm 0.0090}$	$0.4095_{\pm 0.0388}$	$0.6309_{\pm 0.0607}$	$0.3899_{\pm 0.0178}$	$\mathbf{1.546}_{\pm 0.127}$
	+DSS	$0.2196_{\pm 0.0068}$	$0.6986_{\pm 0.0403}$	$0.7874_{\pm 0.0165}$	$\mathbf{0.4133}_{\pm 0.0112}$	$2.601_{\pm 0.473}$

Table 3: LLM unlearning results. Bold marks the better Agg. and RTE within each FULL/DSS pair.

Unlike the vision setting, selection operates at the token level rather than by removing individual training examples. The selected positions determine which answer-token positions contribute to the unlearning objective, while the full causal context is retained for sequences containing selected tokens. Consequently, reducing the number of selected tokens does not translate directly into a proportional reduction in model computation, and the resulting runtime benefit depends on the base method.

DSS improves the aggregate score across most methods (Table 3), with trade-offs among forgetting, privacy, and utility. NPO and SimNPO reduce runtime, whereas GradDiff and RMU take longer than FULL. Initial and online scoring together take approximately 0.53–0.83 minutes on average, depending on the method. The total runtimes for NPO and SimNPO remain below FULL after including this overhead.

5.4 DISCUSSION

The initial selection fraction should reflect both the dataset and the base unlearning method. For datasets with many classes, such as CIFAR-100, we consider larger initial fractions to encourage broader class participation and reduce concentration on a narrow subset of features.

The base unlearning objective also matters. NegGrad+ combines retain training with cross-entropy ascent on forget examples, whereas RL trains forget examples against sampled labels. We believe that broader selection can help distribute ascent updates and limit interference with retained predictions, while RL may accommodate more concentrated reuse of smaller subsets. This motivates adapting the initial selection fraction to the base method.

When selection reduces per-epoch computation but leaves a larger gap after ten unlearning epochs, part of the saved computation can be used for additional training, as shown in Figure 3. Our refresh-interval study further suggests that less frequent rescoring may provide an additional way to reduce selection overhead, although we use rescoring every epoch in the main experiments.

6 CONCLUSION

We introduced DSS to dynamically select low-confidence retain and forget examples within existing unlearning methods, and DSS+ to reduce rescoring by reusing previously selected examples. Both preserve the base method's loss formulation and update rules. Our source-utility analysis motivates selecting forget examples with low confidence under the retrained model, and our empirical analyses link this criterion to initial confidence and selection transitions. Experiments on CIFAR-10 and CIFAR-100 show that DSS can reduce computational cost across multiple unlearning methods. Experiments on TOFU further extend its application to language models. These findings support dynamic source selection as a practical approach to improving the computational efficiency of existing unlearning methods.

AI USE STATEMENT

We used generative AI tools to assist with translation, mathematical derivations, and proofs. We also used these tools to find relevant information, obtain icons, and edit the manuscript for readability.

REPRODUCIBILITY STATEMENT

Algorithms 1 and 2 describe the selection procedures for vision and language models, respectively. We will share a link to an anonymized code repository with reviewers and area chairs through an OpenReview comment once the discussion forum opens.

REFERENCES

Youssef Allouah, Rachid Guerraoui, and Sanmi Koyejo. Distributional machine unlearning via selective data removal. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c1b3d1e2cf53bb28cabd801bd58b3521-Abstract-Conference.html.

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)*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

George-Octavian Bărbulescu and Peter Triantafillou. To each (Textual sequence) its own: Improving memorized-data unlearning in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3003–3023. PMLR, 2024. URL https://proceedings.mlr.press/v235/barbulescu24a.html.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain–forget entanglement. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/bcc2566624d8a63ea6b305e544ceff72-Abstract-Conference.html.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6028–6073. PMLR, 2023. URL https://proceedings.mlr.press/v202/chourasia23a.html.

Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu. Label smoothing improves machine unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2d19debb2ea6a4620f7fe67372cca800-Abstract-Conference.html.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/file/3e4a38f228427ab819ba7899003a44b1-Paper-Datasets_and_Benchmarks_Track.pdf.

Ali Ebrahimpour-Boroojeny, Hari Sundaram, and Varun Chandrasekaran. Not all wrong is bad: Using adversarial examples for unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14950–14971. PMLR, 2025. URL https://proceedings.mlr.press/v267/ebrahimpour-boroojeny25a.html.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ec4d2e436794d1bf55ca83f5ebb31887-Abstract-Conference.html.

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 *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/02443e4e008231e0af0855f3cc70ed17-Abstract-Conference.html.

Kristian Georgiev, Roy Rinberg, Sung Min Park, Shivam Garg, Andrew Ilyas, Aleksander Madry, and Seth Neel. Attribute-to-delete: Machine unlearning via datamodel matching. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3vXpZpOn29.

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*, pp. 9304–9312, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html.

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*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51584–51605, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a204aa68ab4e970e1ceccfb5b5cdc5e4-Abstract-Conference.html.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract.html.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *Proceedings of the ACM Web Conference 2024*, pp. 1260–1271. Association for Computing Machinery, 2024. doi: 10.1145/3589334.3645669. URL https://doi.org/10.1145/3589334.3645669.

Ziwen Liu, Huawei Lin, Yide Ran, Denghui Zhang, Jianwen Xie, Chuan Li, Weijie Zhao, and Zhaozhuo Xu. Randomized antipodal search done right for data Pareto improvement of LLM unlearning. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008939.

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. URL https://openreview.net/forum?id=B41hNBoWLo.

Meta. Llama 3.2 1B Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct. Released September 25, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960. PMLR, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. LLM unlearning reveals a stronger-than-expected coreset effect in current benchmarks. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/pdf/3ebe54a4517ba5c5146eb4601360e351ec40ffee.pdf.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper/2021/hash/ac56f8fe9eea3e4a365f29f0f1957c55-Abstract.html.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. InfoBatch: Lossless training speed up by unbiased dynamic data pruning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ad1d7a4df30a9c0c46b387815a774a84-Abstract-Conference.html.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/.

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. URL https://openreview.net/forum?id=MXLBXjQkmb.

APPENDIX OVERVIEW

- **Appendix A: Experimental details.** Datasets, deletion scenarios, base methods, optimization, selection-rule implementations, metrics, and online token selection.
- **Appendix B: Additional results.** Computational cost, runtime breakdown, results across all six base methods, additional deletion settings, and class unlearning.
- **Appendix C: Theory.** Source comparisons at matched and unequal initial confidence, repeated-update bounds for CE ascent and relabeling, and realizable learning objectives.
- **Appendix D: Retrain response and prediction distributions.** Initial confidence, confidence decreases, selection transitions, low/high selection and gradient norms, and prediction distributions after ten epochs.

A Experimental Details

A.1 Base unlearning methods and why they were chosen

We choose six methods that use both retain and forget data but produce unlearning updates in different ways. NegGrad+ and RL provide the simplest examples of two mechanisms: increasing the loss on forget labels and learning alternative labels. We use these two methods for the detailed gradient and selection analyses. NPO, UGradSL+, SalUn, and AMUN test whether DSS also works with changes to the loss, parameter updates, or training data.

NegGrad+. NegGrad+ minimizes cross-entropy on retain data and maximizes it on forget data (Kurmanji et al., 2023). Its two gradients make it a direct starting point for studying how selecting examples changes the balance between retaining and forgetting.

NPO. Negative Preference Optimization reduces the likelihood of forget labels relative to a frozen copy of the original model. We use its NPO+RT variant, which adds retain cross-entropy (Zhang et al., 2024). Its multiplier on a forget example's gradient decreases as that likelihood ratio falls. We include NPO to test DSS with a loss that reduces forgetting pressure on already suppressed examples, in contrast to ordinary cross-entropy ascent.

UGradSL+. UGradSL+ combines retain cross-entropy with ascent on a label-smoothed forget loss (Di et al., 2026). It tests whether DSS remains useful when smoothing changes the forget gradient. We use the fixed smoothing parameter $\alpha = -0.4$, not the adaptive version.

Random labeling (RL). RL trains forget examples on random labels and retain examples on their true labels (Golatkar et al., 2020). In our implementation, each exposure draws a new label uniformly from all classes, including the true class. Forget updates precede retain updates in each epoch. RL is the basic comparison for studying selection without gradient ascent.

SalUn. SalUn restricts relabeling updates to parameters selected by forget-gradient saliency (Fan et al., 2024). It provides an established baseline for testing data selection alongside parameter selection. We use the same fixed mask for FULL and DSS, retaining the half of coordinates with the largest saliency.

AMUN. AMUN fine-tunes with adversarial examples generated from forget data (Ebrahimpour-Boroojeny et al., 2025). It tests DSS with a method that changes the training inputs, rather than only the loss or updated parameters. Adversarial examples are generated once before unlearning. Selecting a forget example also selects its corresponding adversarial example.

A.2 Experimental setup

The main comparisons use ResNet-18 on CIFAR-10 and both ResNet-18 and VGG16 on CIFAR-100. Additional experiments also evaluate VGG16 on CIFAR-10. Throughout the paper, VGG16 refers to the batch-normalized VGG16 architecture. Random forgetting samples 10% of the training examples without replacement. Class-concentrated forgetting samples 20% from each of CIFAR-10 classes 0–4. Training uses random 32×32 crops with four-pixel padding and random horizontal flips. Selection scoring and evaluation use inputs without augmentation.

Reference-model training. The original model is trained on the full training set, and the retrained model is trained from scratch on the retain set using the same recipe. Both use SGD with initial learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and batch size 256. Training lasts 182 epochs for ResNet-18 on CIFAR-10 and 200 epochs for VGG16 on CIFAR-10 and for both models on CIFAR-100, with tenfold learning-rate reductions at epochs 91 and 136, and 100 and 150, respectively.

Common unlearning settings. The main comparisons use SGD for ten epochs with batch size 256, momentum 0.9, and weight decay 5×10^{-4}. Learning rates remain constant during unlearning and are shared by FULL, DSS, and DSS+ within each method and setting.

For CIFAR-10 with ResNet-18 and VGG16, we use learning rates in $[10^{-3}, 5\times10^{-2}]$ under random forgetting. Under class-concentrated forgetting, learning rates are in $[10^{-5}, 10^{-2}]$ for NegGrad+, NPO, and UGradSL+, and $[10^{-3}, 5 \times 10^{-2}]$ for RL, SalUn, and AMUN. For CIFAR-100 under random forgetting, learning rates are in $[10^{-3}, 5 \times 10^{-2}]$ for both ResNet-18 and VGG16.

For the vision experiments, NegGrad+ and NPO pair retain and forget minibatches, cycling the shorter loader until the longer loader completes one pass. UGradSL+ traverses the retain loader once per epoch and cycles the forget loader as needed.

Method-specific hyperparameters.

- **NegGrad+:** retain and forget loss weights of 0.9 and 0.1, respectively.
- **NPO:** $\beta = 0.1$, retain and forget loss weights of 0.9 and 0.1, respectively, and a frozen original model as the reference.
- **UGradSL+:** retain weight $p = 0.9$ and fixed smoothing parameter $\alpha = -0.4$.
- **RL:** a fresh label sampled uniformly from all classes, including the true class, at each exposure.
- **SalUn:** a fixed mask selecting the 50% of parameter coordinates with the largest absolute gradients accumulated over the complete forget set.
- **AMUN:** adaptive untargeted ℓ_2 PGD with up to 50 iterations per radius and step size 0.1ϵ. Starting at 0.01, we calibrate the initial radius using the median successful radius from the first 50 forget examples. Unresolved examples use doubled radii, with a safety limit of 32 doublings.

Selection and repetitions. On CIFAR-10, DSS and DSS+ use an initial selection fraction of 5% for all six methods. On CIFAR-100, the fraction is 50% for NegGrad+, NPO, and UGradSL+, and 10% for RL, SalUn, and AMUN. Selection is refreshed each epoch, and the initial confidence threshold remains fixed. For each setting, we use ten deletion seeds to generate different forget sets. For each seed, the policies share the initial checkpoint and deletion request. We report means and sample standard deviations across these ten seeds. Experiments are conducted on NVIDIA RTX 6000 Ada.

A.3 Selection-rule comparison

Figure 5 compares selection rules for NegGrad+ and NPO on CIFAR-10 with ResNet-18 under random forgetting. Each policy starts from the same model checkpoint and uses the same deletion request and retrained reference.

Score replacements. Confidence-based selection ranks examples by negative true-label probability. EL2N uses the norm of the difference between the predicted probability vector and the one-hot true label, while GraNd uses the norm of the per-example true-label cross-entropy gradient over all trainable parameters (Paul et al., 2021). These scores are recomputed under the current model at each epoch refresh on clean inputs in evaluation mode. Each rule initially selects the highest-scoring 5% of the pooled retain and forget examples, fixes the resulting numerical threshold, and subsequently allows examples to enter or leave the selected set. Confidence and EL2N require a forward pass without backpropagation; GraNd additionally computes per-example gradients. EL2N depends on the probability mass among incorrect classes as well as the true-label probability, so its rankings and subset sizes can differ from confidence-based selection.

InfoBatch adaptation. InfoBatch (Qin et al., 2024) uses cached loss scores for pruning and inverse inclusion-probability weighting. The first epoch trains on all examples and caches their true-label cross-entropy from the training forward passes. Scores are unweighted and computed before the parameter update. Only visited examples update their cached scores. At each subsequent pruning step, we retain all examples at or above the pooled retain/forget mean score and uniformly sample approximately half of those below it without replacement. Selected examples receive inverse inclusion-probability weights in the base method's retain and forget objectives before minibatch averaging, with its loader cycling unchanged. The final two epochs use all examples with unit weights. Scoring requires no additional model forward passes.

A.4 Evaluation metrics and compute

Retain accuracy (RA) is the accuracy on the retain set, unlearning accuracy (UA) is the error rate on the forget set, and test accuracy (TA) is the accuracy on the complete test set. UA is interpreted by closeness to the corresponding retrained model rather than maximization. MIA is the percentage of forget examples classified as nonmembers by an SVC attack using true-label confidence.

Avg. Gap is the mean absolute difference in RA, UA, TA, and MIA between the unlearned and corresponding retrained models; lower is better.

RTE reports algorithm runtime in minutes, and PFLOPs report estimated computation in units of 10^{15} floating-point operations. For vision experiments, each multiply-add counts as two operations, and the combined student forward and backward computation is estimated as three profiled forward passes. Both include the costs of the base unlearning method, selection scoring, and method-specific preprocessing through the reported checkpoint. Diagnostic evaluation, prediction storage, and checkpoint I/O are excluded.

A.5 Online token selection for LLM unlearning

For the experiments in Table 3, we apply DSS to answer tokens on TOFU forget10/retain90 with Llama-3.2-1B-Instruct. Selection uses the probability of the observed answer token given the question and preceding answer tokens. Prompt, padding, and special-token positions are excluded. Before unlearning, we score all eligible answer tokens in both pools and fix τ at their pooled 30% quantile, with one vote per token. A token is selected when its probability is below τ.

The first epoch uses these stored token masks and the sequences containing at least one selected token. From the second epoch onward, we iterate over all forget sequences once per epoch and sample one retain sequence uniformly with replacement from the complete retain pool for each forget sequence. Within each minibatch, we rescore answer tokens using the current model and update their selection using the fixed threshold. Thus, unlike the epoch-wise image procedure, this variant does not rescore the complete retain pool at every epoch.

Algorithm 2: Online token-level DSS

Input: model θ_0, forget and retain sequence pools, initial quantile ρ, base unlearning method, and epoch count E.

Score eligible answer tokens in both pools under θ_0 and fix τ at the pooled ρ-quantile. Store masks selecting tokens with probability below τ.
For epochs $e = 1, \ldots, E$:
  If $e = 1$, use sequences with nonempty initial masks as the two sampling pools; otherwise, use both complete sequence pools.
  Shuffle the forget pool and pair each forget sequence with one retain sequence sampled uniformly with replacement.
  For each paired minibatch:
    If $e = 1$, retrieve the initial masks; otherwise, make a no-gradient forward pass and select tokens with current probability below τ.
    Omit sequences with empty masks from the gradient-enabled forward; keep the full causal context of the remaining sequences.
    Compute the base method's forget and retain objectives on selected positions, accumulate gradients, and update the model according to the gradient-accumulation schedule.

All four methods use the same true-label-confidence selection rule. The resulting token masks restrict each base method's objective to the selected positions. For GradDiff and retain cross-entropy, selected token losses are summed and divided by the number of eligible tokens in the original minibatch. NPO uses selected token log-likelihoods for both the current and reference models. SimNPO normalizes each selected sequence loss by that sequence's eligible-token count. RMU applies the token mask to the activation loss. For NPO and SimNPO, sequence-level reductions retain the original batch denominator and the constant contribution from an omitted sequence.

We follow the common training configuration and default method-specific hyperparameters of OpenUnlearning (Dorna et al., 2025): ten epochs, batch size eight, and four gradient-accumulation steps. We search candidates within $[10^{-5}, 5 \times 10^{-5}]$ and select the learning rate using the harmonic mean

of Mem. and Utility on FULL runs. FULL and DSS share the selected learning rate within each method. Algorithm runtime includes initial calibration, online scoring, required reference-model computation, and optimization; diagnostic evaluation and checkpoint writing are excluded. Masking token losses alone does not remove the forward computation for their causal context: sequences with selected tokens still use dense model passes.

B Additional Unlearning Results

B.1 Computational cost and runtime breakdown

Tables 4 and 5 report computation and runtime. Training time includes minibatch processing and required reference-model forward passes. Diagnostic evaluation and checkpoint I/O are excluded. Runtime measurements retain their original execution environments, including historical FULL runs.

Table 4: Cumulative estimated computation (PFLOPs) for the settings in Tables 1 and 2.

		CIFAR-10 / ResNet-18		CIFAR-100 / random forgetting	
Method	Variant	Random	Class-concentrated	ResNet-18	VGG16
NegGrad+	FULL	$2.969_{\pm 0.000}$	$2.969_{\pm 0.000}$	$2.969_{\pm 0.000}$	$1.678_{\pm 0.000}$
	DSS	$0.992_{\pm 0.030}$	$1.251_{\pm 0.011}$	$1.862_{\pm 0.042}$	$1.143_{\pm 0.102}$
	DSS+	$1.121_{\pm 0.048}$	$1.166_{\pm 0.010}$	$1.978_{\pm 0.035}$	$1.201_{\pm 0.067}$
NPO	FULL	$3.459_{\pm 0.000}$	$3.459_{\pm 0.000}$	$3.459_{\pm 0.000}$	$1.954_{\pm 0.000}$
	DSS	$1.070_{\pm 0.040}$	$1.353_{\pm 0.012}$	$2.023_{\pm 0.014}$	$1.324_{\pm 0.027}$
	DSS+	$1.215_{\pm 0.059}$	$1.280_{\pm 0.012}$	$2.237_{\pm 0.030}$	$1.487_{\pm 0.083}$
UGradSL+	FULL	$2.999_{\pm 0.000}$	$2.999_{\pm 0.000}$	$3.000_{\pm 0.000}$	$1.695_{\pm 0.000}$
	DSS	$2.369_{\pm 0.069}$	$1.276_{\pm 0.008}$	$2.654_{\pm 0.038}$	$1.883_{\pm 0.005}$
	DSS+	$2.156_{\pm 0.039}$	$1.189_{\pm 0.012}$	$2.432_{\pm 0.008}$	$1.628_{\pm 0.001}$
RL	FULL	$1.666_{\pm 0.000}$	$1.666_{\pm 0.000}$	$1.666_{\pm 0.000}$	$0.941_{\pm 0.000}$
	DSS	$1.453_{\pm 0.034}$	$1.434_{\pm 0.100}$	$1.384_{\pm 0.046}$	$0.956_{\pm 0.021}$
	DSS+	$1.477_{\pm 0.018}$	$1.503_{\pm 0.075}$	$1.384_{\pm 0.064}$	$0.900_{\pm 0.013}$
SalUn	FULL	$1.683_{\pm 0.000}$	$1.683_{\pm 0.000}$	$1.683_{\pm 0.000}$	$0.951_{\pm 0.000}$
	DSS	$1.451_{\pm 0.037}$	$1.510_{\pm 0.054}$	$1.553_{\pm 0.075}$	$1.018_{\pm 0.010}$
	DSS+	$1.512_{\pm 0.021}$	$1.558_{\pm 0.054}$	$1.505_{\pm 0.038}$	$0.906_{\pm 0.009}$
AMUN	FULL	$2.307_{\pm 0.007}$	$2.284_{\pm 0.004}$	$2.172_{\pm 0.005}$	$1.304_{\pm 0.002}$
	DSS	$1.353_{\pm 0.028}$	$1.346_{\pm 0.047}$	$1.512_{\pm 0.018}$	$1.142_{\pm 0.064}$
	DSS+	$1.444_{\pm 0.053}$	$1.427_{\pm 0.056}$	$1.691_{\pm 0.009}$	$1.137_{\pm 0.007}$

Table 5: Algorithm runtime breakdown in minutes. Scoring includes the initial scoring pass and subsequent rescoring. Other includes cold preprocessing, loader construction, subset indexing, and reference-model setup.

CIFAR-10 / ResNet-18

		Random forgetting				Class-concentrated forgetting			
Method	Variant	Scoring	Training	Other	Total	Scoring	Training	Other	Total
NegGrad+	FULL	$0.000_{\pm 0.000}$	$2.683_{\pm 0.068}$	$0.000_{\pm 0.000}$	$2.683_{\pm 0.068}$	$0.000_{\pm 0.000}$	$2.699_{\pm 0.087}$	$0.000_{\pm 0.000}$	$2.699_{\pm 0.087}$
	DSS	$0.516_{\pm 0.003}$	$0.397_{\pm 0.027}$	$0.001_{\pm 0.000}$	$0.914_{\pm 0.028}$	$0.515_{\pm 0.006}$	$0.624_{\pm 0.010}$	$0.001_{\pm 0.000}$	$1.140_{\pm 0.013}$
	DSS+	$0.416_{\pm 0.008}$	$0.608_{\pm 0.053}$	$0.001_{\pm 0.000}$	$1.025_{\pm 0.047}$	$0.392_{\pm 0.003}$	$0.671_{\pm 0.012}$	$0.001_{\pm 0.000}$	$1.065_{\pm 0.013}$
NPO	FULL	$0.000_{\pm 0.000}$	$2.861_{\pm 0.066}$	$0.000_{\pm 0.000}$	$2.861_{\pm 0.066}$	$0.000_{\pm 0.000}$	$2.869_{\pm 0.074}$	$0.000_{\pm 0.000}$	$2.870_{\pm 0.074}$
	DSS	$0.514_{\pm 0.005}$	$0.428_{\pm 0.034}$	$0.001_{\pm 0.000}$	$0.943_{\pm 0.035}$	$0.513_{\pm 0.005}$	$0.658_{\pm 0.009}$	$0.001_{\pm 0.000}$	$1.172_{\pm 0.009}$
	DSS+	$0.419_{\pm 0.008}$	$0.631_{\pm 0.056}$	$0.001_{\pm 0.000}$	$1.052_{\pm 0.049}$	$0.396_{\pm 0.004}$	$0.707_{\pm 0.012}$	$0.002_{\pm 0.000}$	$1.105_{\pm 0.013}$
UGradSL+	FULL	$0.000_{\pm 0.000}$	$2.733_{\pm 0.093}$	$0.000_{\pm 0.000}$	$2.733_{\pm 0.093}$	$0.000_{\pm 0.000}$	$2.732_{\pm 0.090}$	$0.000_{\pm 0.000}$	$2.732_{\pm 0.090}$
	DSS	$0.517_{\pm 0.006}$	$1.632_{\pm 0.058}$	$0.003_{\pm 0.000}$	$2.152_{\pm 0.058}$	$0.516_{\pm 0.005}$	$0.653_{\pm 0.008}$	$0.001_{\pm 0.000}$	$1.171_{\pm 0.011}$
	DSS+	$0.245_{\pm 0.009}$	$1.702_{\pm 0.046}$	$0.003_{\pm 0.000}$	$1.950_{\pm 0.039}$	$0.389_{\pm 0.004}$	$0.695_{\pm 0.015}$	$0.001_{\pm 0.000}$	$1.086_{\pm 0.013}$
RL	FULL	$0.000_{\pm 0.000}$	$1.473_{\pm 0.039}$	$0.000_{\pm 0.000}$	$1.473_{\pm 0.039}$	$0.000_{\pm 0.000}$	$1.479_{\pm 0.046}$	$0.000_{\pm 0.000}$	$1.479_{\pm 0.046}$
	DSS	$0.515_{\pm 0.006}$	$0.787_{\pm 0.030}$	$0.002_{\pm 0.000}$	$1.304_{\pm 0.032}$	$0.517_{\pm 0.005}$	$0.767_{\pm 0.084}$	$0.002_{\pm 0.000}$	$1.286_{\pm 0.082}$
	DSS+	$0.163_{\pm 0.009}$	$1.140_{\pm 0.019}$	$0.004_{\pm 0.000}$	$1.307_{\pm 0.012}$	$0.153_{\pm 0.030}$	$1.174_{\pm 0.100}$	$0.004_{\pm 0.000}$	$1.330_{\pm 0.070}$
SalUn	FULL	$0.000_{\pm 0.000}$	$1.647_{\pm 0.047}$	$0.019_{\pm 0.007}$	$1.666_{\pm 0.052}$	$0.000_{\pm 0.000}$	$1.644_{\pm 0.043}$	$0.019_{\pm 0.007}$	$1.663_{\pm 0.049}$
	DSS	$0.513_{\pm 0.003}$	$0.860_{\pm 0.036}$	$0.017_{\pm 0.000}$	$1.389_{\pm 0.036}$	$0.515_{\pm 0.004}$	$0.916_{\pm 0.054}$	$0.017_{\pm 0.000}$	$1.448_{\pm 0.054}$
	DSS+	$0.155_{\pm 0.008}$	$1.299_{\pm 0.033}$	$0.018_{\pm 0.000}$	$1.472_{\pm 0.026}$	$0.137_{\pm 0.023}$	$1.364_{\pm 0.073}$	$0.019_{\pm 0.000}$	$1.519_{\pm 0.050}$
AMUN	FULL	$0.000_{\pm 0.000}$	$1.622_{\pm 0.050}$	$0.459_{\pm 0.077}$	$2.081_{\pm 0.109}$	$0.000_{\pm 0.000}$	$1.615_{\pm 0.035}$	$0.489_{\pm 0.029}$	$2.103_{\pm 0.055}$
	DSS	$0.518_{\pm 0.005}$	$0.285_{\pm 0.028}$	$0.437_{\pm 0.018}$	$1.240_{\pm 0.029}$	$0.519_{\pm 0.007}$	$0.298_{\pm 0.040}$	$0.470_{\pm 0.034}$	$1.287_{\pm 0.052}$
	DSS+	$0.384_{\pm 0.015}$	$0.491_{\pm 0.061}$	$0.438_{\pm 0.018}$	$1.314_{\pm 0.057}$	$0.387_{\pm 0.019}$	$0.490_{\pm 0.065}$	$0.470_{\pm 0.034}$	$1.347_{\pm 0.048}$

CIFAR-100 / random forgetting

		ResNet-18				VGG16			
Method	Variant	Scoring	Training	Other	Total	Scoring	Training	Other	Total
NegGrad+	FULL	$0.000_{\pm 0.000}$	$2.653_{\pm 0.046}$	$0.000_{\pm 0.000}$	$2.653_{\pm 0.046}$	$0.000_{\pm 0.000}$	$1.913_{\pm 0.018}$	$0.000_{\pm 0.000}$	$1.913_{\pm 0.018}$
	DSS	$0.523_{\pm 0.007}$	$1.173_{\pm 0.039}$	$0.002_{\pm 0.000}$	$1.698_{\pm 0.044}$	$0.403_{\pm 0.006}$	$0.951_{\pm 0.118}$	$0.002_{\pm 0.000}$	$1.356_{\pm 0.117}$
	DSS+	$0.260_{\pm 0.007}$	$1.529_{\pm 0.038}$	$0.003_{\pm 0.000}$	$1.791_{\pm 0.032}$	$0.181_{\pm 0.017}$	$1.220_{\pm 0.091}$	$0.003_{\pm 0.000}$	$1.404_{\pm 0.075}$
NPO	FULL	$0.000_{\pm 0.000}$	$2.835_{\pm 0.047}$	$0.000_{\pm 0.000}$	$2.835_{\pm 0.047}$	$0.000_{\pm 0.000}$	$1.992_{\pm 0.014}$	$0.000_{\pm 0.000}$	$1.992_{\pm 0.014}$
	DSS	$0.521_{\pm 0.005}$	$1.203_{\pm 0.011}$	$0.002_{\pm 0.000}$	$1.726_{\pm 0.010}$	$0.403_{\pm 0.002}$	$1.039_{\pm 0.029}$	$0.002_{\pm 0.000}$	$1.444_{\pm 0.030}$
	DSS+	$0.261_{\pm 0.004}$	$1.607_{\pm 0.027}$	$0.003_{\pm 0.000}$	$1.871_{\pm 0.024}$	$0.159_{\pm 0.019}$	$1.406_{\pm 0.095}$	$0.003_{\pm 0.000}$	$1.568_{\pm 0.077}$
UGradSL+	FULL	$0.000_{\pm 0.000}$	$2.702_{\pm 0.055}$	$0.000_{\pm 0.000}$	$2.702_{\pm 0.055}$	$0.000_{\pm 0.000}$	$1.934_{\pm 0.013}$	$0.000_{\pm 0.000}$	$1.934_{\pm 0.013}$
	DSS	$0.519_{\pm 0.006}$	$1.881_{\pm 0.039}$	$0.003_{\pm 0.000}$	$2.403_{\pm 0.039}$	$0.401_{\pm 0.006}$	$1.798_{\pm 0.021}$	$0.004_{\pm 0.000}$	$2.203_{\pm 0.026}$
	DSS+	$0.181_{\pm 0.002}$	$2.006_{\pm 0.010}$	$0.003_{\pm 0.000}$	$2.190_{\pm 0.010}$	$0.071_{\pm 0.001}$	$1.799_{\pm 0.009}$	$0.004_{\pm 0.000}$	$1.874_{\pm 0.010}$
RL	FULL	$0.000_{\pm 0.000}$	$1.453_{\pm 0.028}$	$0.000_{\pm 0.000}$	$1.453_{\pm 0.028}$	$0.000_{\pm 0.000}$	$1.068_{\pm 0.007}$	$0.000_{\pm 0.000}$	$1.068_{\pm 0.007}$
	DSS	$0.521_{\pm 0.005}$	$0.725_{\pm 0.043}$	$0.002_{\pm 0.000}$	$1.248_{\pm 0.047}$	$0.403_{\pm 0.006}$	$0.736_{\pm 0.028}$	$0.003_{\pm 0.000}$	$1.142_{\pm 0.032}$
	DSS+	$0.205_{\pm 0.027}$	$1.030_{\pm 0.079}$	$0.003_{\pm 0.000}$	$1.239_{\pm 0.053}$	$0.091_{\pm 0.008}$	$0.954_{\pm 0.024}$	$0.004_{\pm 0.000}$	$1.048_{\pm 0.017}$
SalUn	FULL	$0.000_{\pm 0.000}$	$1.621_{\pm 0.033}$	$0.015_{\pm 0.001}$	$1.636_{\pm 0.033}$	$0.000_{\pm 0.000}$	$1.188_{\pm 0.011}$	$0.013_{\pm 0.001}$	$1.202_{\pm 0.011}$
	DSS	$0.520_{\pm 0.003}$	$0.961_{\pm 0.074}$	$0.017_{\pm 0.001}$	$1.499_{\pm 0.075}$	$0.400_{\pm 0.005}$	$0.881_{\pm 0.015}$	$0.016_{\pm 0.000}$	$1.298_{\pm 0.016}$
	DSS+	$0.162_{\pm 0.017}$	$1.301_{\pm 0.055}$	$0.018_{\pm 0.000}$	$1.482_{\pm 0.039}$	$0.092_{\pm 0.005}$	$1.056_{\pm 0.018}$	$0.017_{\pm 0.000}$	$1.165_{\pm 0.013}$
AMUN	FULL	$0.000_{\pm 0.000}$	$1.600_{\pm 0.026}$	$0.427_{\pm 0.050}$	$2.027_{\pm 0.056}$	$0.000_{\pm 0.000}$	$1.149_{\pm 0.012}$	$0.490_{\pm 0.026}$	$1.640_{\pm 0.026}$
	DSS	$0.520_{\pm 0.007}$	$0.539_{\pm 0.014}$	$0.419_{\pm 0.038}$	$1.478_{\pm 0.040}$	$0.401_{\pm 0.004}$	$0.622_{\pm 0.075}$	$0.486_{\pm 0.032}$	$1.509_{\pm 0.100}$
	DSS+	$0.260_{\pm 0.006}$	$0.950_{\pm 0.009}$	$0.420_{\pm 0.038}$	$1.630_{\pm 0.036}$	$0.151_{\pm 0.005}$	$0.843_{\pm 0.014}$	$0.487_{\pm 0.031}$	$1.481_{\pm 0.038}$

B.2 Additional results across base methods

The following figures report unlearning trajectories and the effects of the initial selection fraction and refresh interval across all six base methods. Unless otherwise stated, these trajectories report epoch-wise means over the same ten deletion seeds used in the corresponding main experiments.

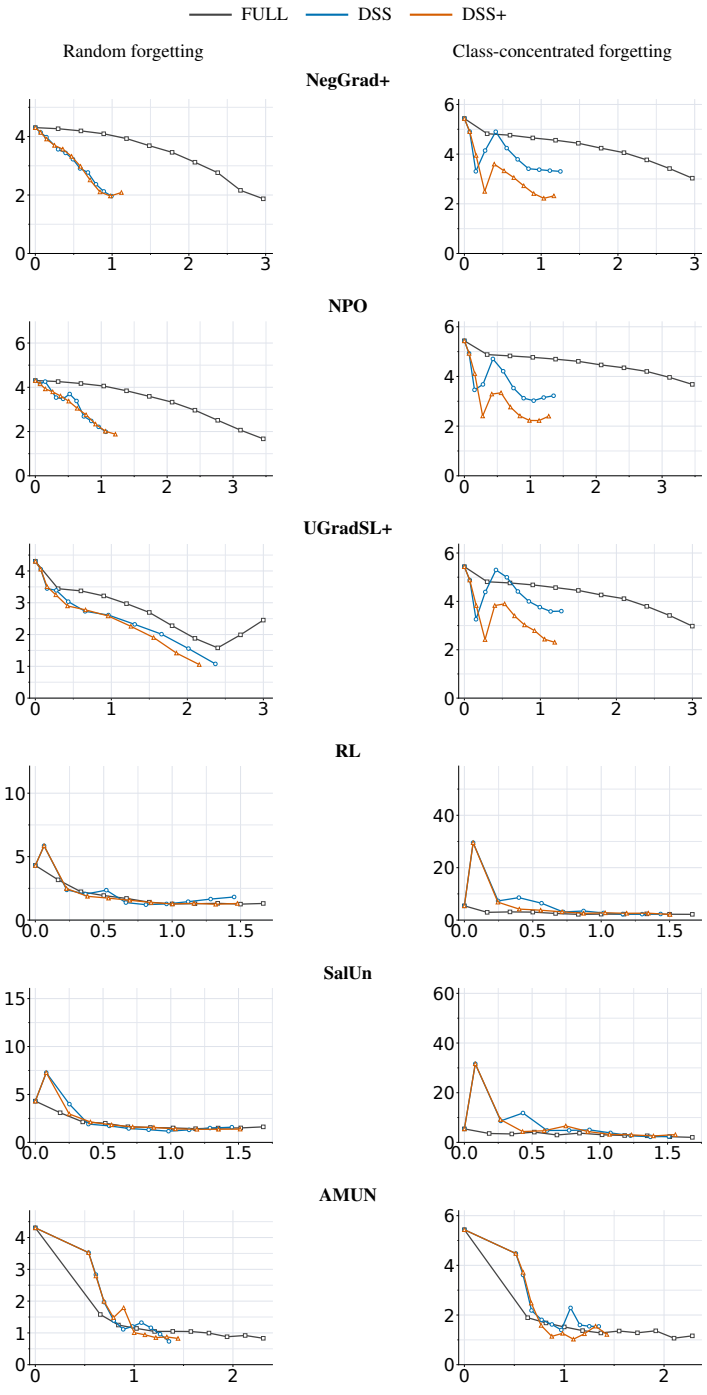


Figure 8: Unlearning trajectories on CIFAR-10 with ResNet-18. Columns show random and class-concentrated forgetting. Rows show the base methods. Axes show cumulative computation (PFLOPs; x) and Avg. Gap (percentage points; y). Markers denote epoch checkpoints.

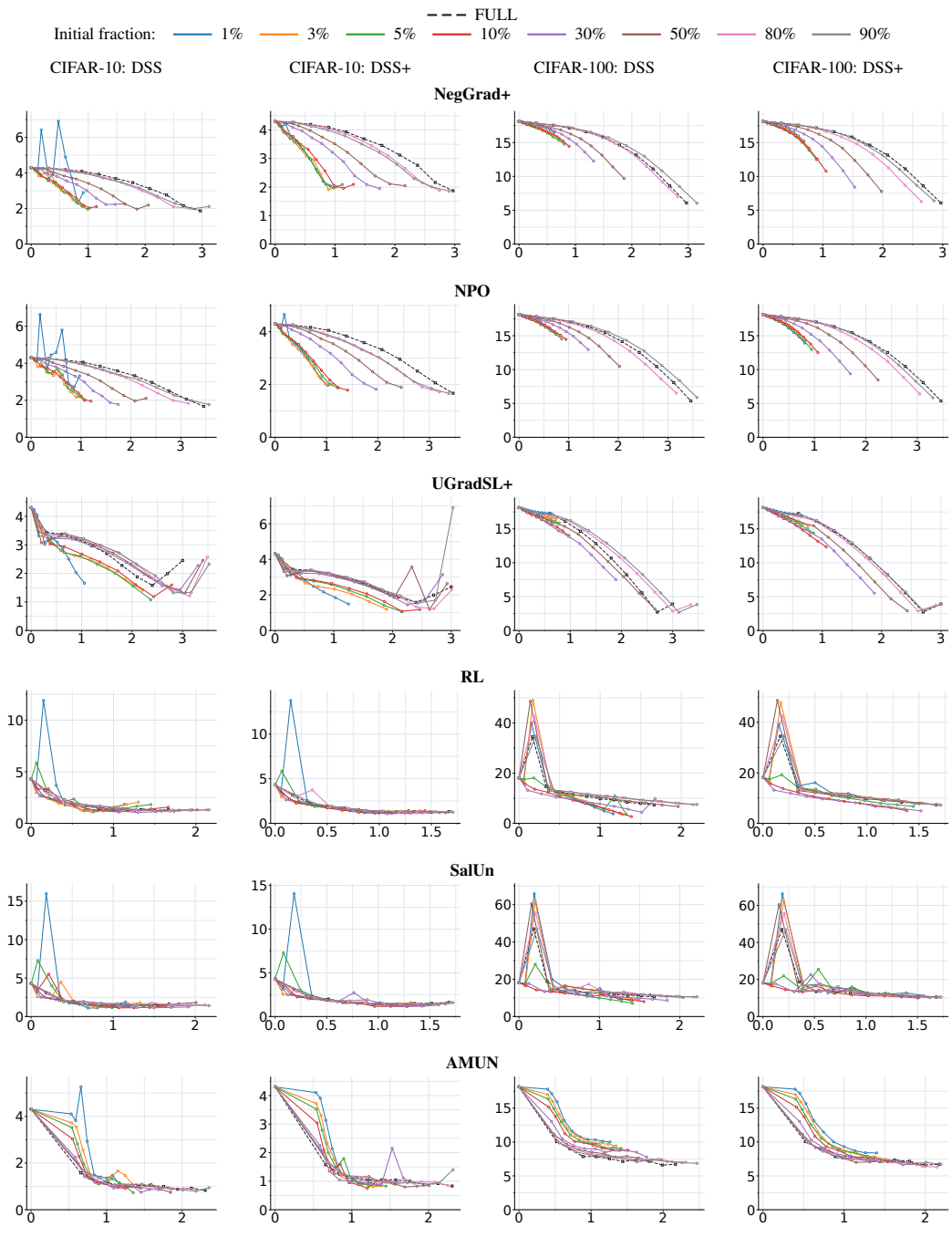


Figure 9: Effect of the initial selection fraction with ResNet-18 under random forgetting and selection refreshed every epoch. Columns identify the dataset and selection policy. Rows show the base methods. Axes show cumulative PFLOPs (x) and Avg. Gap in percentage points (y).

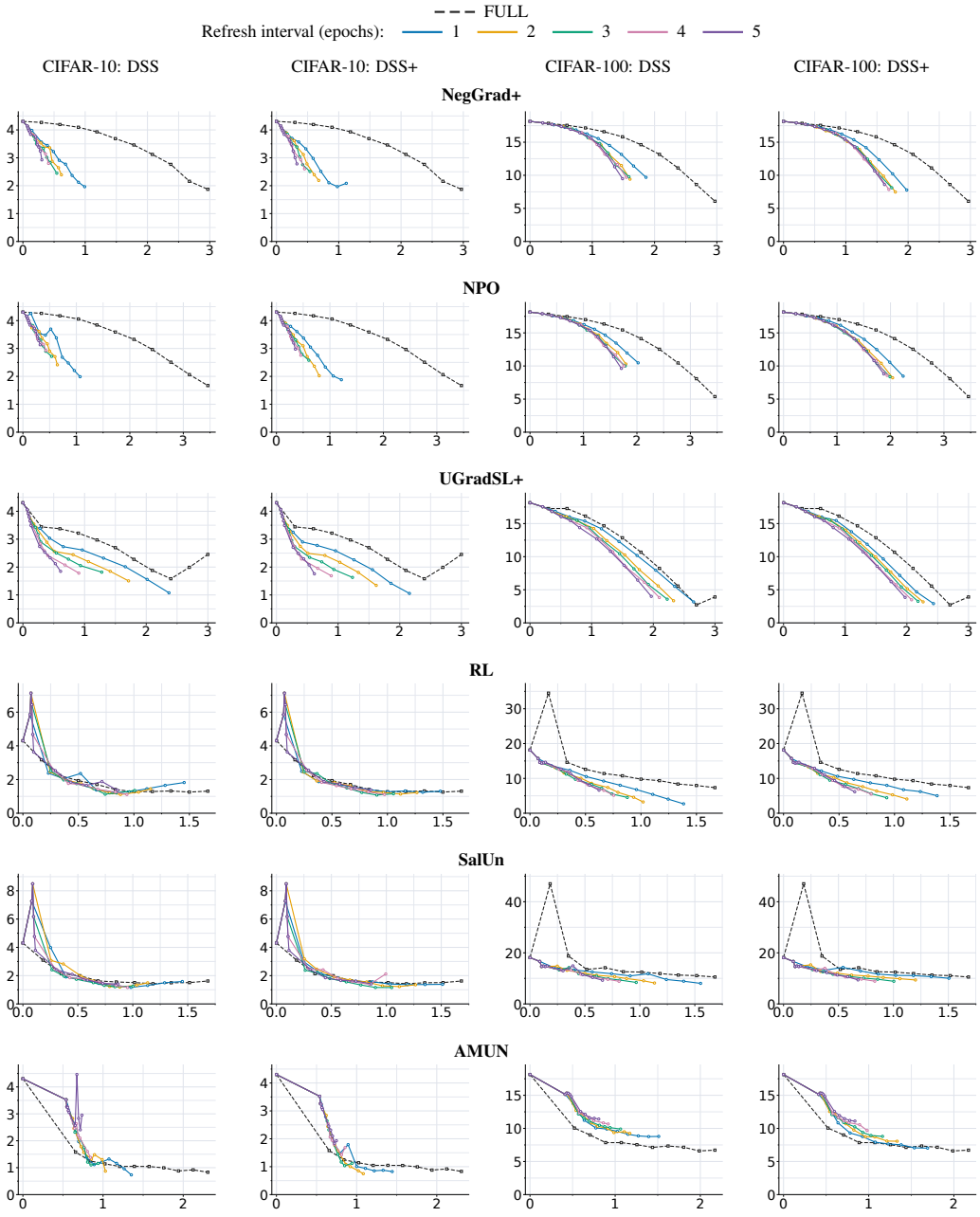


Figure 10: Effect of the refresh interval with ResNet-18 under random forgetting. Initial fractions are 5% on CIFAR-10. On CIFAR-100, they are 50% for NegGrad+, NPO, and UGradSL+, and 10% for RL, SalUn, and AMUN. Axes show cumulative PFLOPs (x) and Avg. Gap in percentage points (y).

RL on CIFAR-100 with VGG16. The larger mean Avg. Gap of DSS+ is strongly influenced by one run with a sharp accuracy drop during early training. This run achieves its lowest Avg. Gap of 14.44 after the first epoch, with test accuracy of 63.57%. By the endpoint, its Avg. Gap increases to 61.64 and test accuracy falls to 13.87%, compared with 70.23% for FULL on the same deletion request. FULL also experiences an early accuracy drop but subsequently recovers. DSS+ uses the full dataset from the fourth epoch onward, indicating that expanding the selected set does not necessarily restore performance after an early collapse.

B.3 Additional deletion settings

We additionally evaluate VGG16 on CIFAR-10 under random and class-concentrated forgetting. We also increase the randomly selected forget fraction to 30% and 50% for both ResNet-18 and VGG16.

Table 6: CIFAR-10 with VGG16 under random forgetting of 10% of the training data.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$6.23_{\pm 0.35}$	$93.50_{\pm 0.15}$	$10.58_{\pm 0.28}$	$0.00_{\pm 0.00}$	16.945	$19.36_{\pm 0.17}$
NegGrad+	FULL	$98.00_{\pm 0.47}$	$9.48_{\pm 0.66}$	$90.41_{\pm 0.42}$	$12.00_{\pm 1.01}$	$2.45_{\pm 0.59}$	$1.677_{\pm 0.000}$	$1.91_{\pm 0.01}$
	DSS	$97.63_{\pm 2.90}$	$7.66_{\pm 2.71}$	$90.11_{\pm 2.62}$	$10.42_{\pm 3.01}$	$2.29_{\pm 2.56}$	$0.611_{\pm 0.035}$	$0.75_{\pm 0.04}$
	DSS+	$97.57_{\pm 0.84}$	$9.40_{\pm 1.60}$	$89.76_{\pm 0.97}$	$12.07_{\pm 1.93}$	$2.76_{\pm 1.16}$	$0.652_{\pm 0.045}$	$0.79_{\pm 0.06}$
NPO	FULL	$98.71_{\pm 0.22}$	$7.93_{\pm 0.59}$	$91.24_{\pm 0.31}$	$10.31_{\pm 0.83}$	$1.48_{\pm 0.16}$	$1.954_{\pm 0.000}$	$1.98_{\pm 0.02}$
	DSS	$97.18_{\pm 5.35}$	$7.41_{\pm 4.85}$	$89.87_{\pm 4.24}$	$10.25_{\pm 5.25}$	$2.86_{\pm 4.47}$	$0.671_{\pm 0.049}$	$0.78_{\pm 0.05}$
	DSS+	$98.97_{\pm 0.20}$	$7.13_{\pm 0.92}$	$91.47_{\pm 0.28}$	$9.46_{\pm 1.16}$	$1.33_{\pm 0.19}$	$0.754_{\pm 0.081}$	$0.84_{\pm 0.07}$
UGradSL+	FULL	$99.97_{\pm 0.01}$	$1.19_{\pm 0.23}$	$93.56_{\pm 0.11}$	$4.32_{\pm 0.34}$	$2.87_{\pm 0.19}$	$1.695_{\pm 0.000}$	$1.93_{\pm 0.02}$
	DSS	$99.93_{\pm 0.02}$	$2.94_{\pm 0.37}$	$93.19_{\pm 0.19}$	$6.54_{\pm 0.42}$	$1.93_{\pm 0.24}$	$1.410_{\pm 0.048}$	$1.66_{\pm 0.05}$
	DSS+	$99.93_{\pm 0.03}$	$2.86_{\pm 0.53}$	$93.15_{\pm 0.18}$	$6.45_{\pm 0.70}$	$1.98_{\pm 0.27}$	$1.225_{\pm 0.021}$	$1.42_{\pm 0.03}$
RL	FULL	$100.00_{\pm 0.00}$	$0.44_{\pm 0.09}$	$93.90_{\pm 0.06}$	$4.86_{\pm 0.32}$	$2.98_{\pm 0.17}$	$0.941_{\pm 0.000}$	$1.07_{\pm 0.01}$
	DSS	$99.99_{\pm 0.02}$	$0.45_{\pm 0.16}$	$93.95_{\pm 0.13}$	$4.61_{\pm 0.40}$	$3.05_{\pm 0.18}$	$1.084_{\pm 0.023}$	$1.28_{\pm 0.02}$
	DSS+	$99.99_{\pm 0.00}$	$0.43_{\pm 0.07}$	$93.95_{\pm 0.10}$	$4.58_{\pm 0.26}$	$3.06_{\pm 0.14}$	$0.860_{\pm 0.014}$	$1.00_{\pm 0.01}$
SalUn	FULL	$99.99_{\pm 0.01}$	$0.59_{\pm 0.10}$	$93.91_{\pm 0.08}$	$5.15_{\pm 0.27}$	$2.87_{\pm 0.14}$	$0.951_{\pm 0.000}$	$1.20_{\pm 0.01}$
	DSS	$99.99_{\pm 0.00}$	$0.59_{\pm 0.07}$	$93.98_{\pm 0.12}$	$4.83_{\pm 0.45}$	$2.97_{\pm 0.13}$	$1.101_{\pm 0.021}$	$1.41_{\pm 0.03}$
	DSS+	$99.99_{\pm 0.00}$	$0.55_{\pm 0.14}$	$93.93_{\pm 0.12}$	$4.78_{\pm 0.45}$	$2.98_{\pm 0.14}$	$0.876_{\pm 0.013}$	$1.12_{\pm 0.02}$
AMUN	FULL	$99.94_{\pm 0.01}$	$1.98_{\pm 0.26}$	$93.46_{\pm 0.17}$	$4.30_{\pm 0.58}$	$2.69_{\pm 0.23}$	$1.391_{\pm 0.095}$	$1.54_{\pm 0.07}$
	DSS	$99.90_{\pm 0.03}$	$2.44_{\pm 0.27}$	$93.26_{\pm 0.17}$	$4.50_{\pm 0.72}$	$2.55_{\pm 0.26}$	$0.787_{\pm 0.095}$	$0.93_{\pm 0.07}$
	DSS+	$99.95_{\pm 0.04}$	$2.17_{\pm 0.41}$	$93.59_{\pm 0.12}$	$4.75_{\pm 1.21}$	$2.52_{\pm 0.38}$	$0.791_{\pm 0.093}$	$0.93_{\pm 0.06}$

Table 7: CIFAR-10 with VGG16 under class-concentrated forgetting of 20% from each of five classes.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$7.70_{\pm 0.46}$	$93.45_{\pm 0.06}$	$13.14_{\pm 0.78}$	$0.00_{\pm 0.00}$	16.945	$19.36_{\pm 0.13}$
NegGrad+	FULL	$98.24_{\pm 0.09}$	$4.17_{\pm 0.44}$	$90.29_{\pm 0.16}$	$7.40_{\pm 0.46}$	$3.54_{\pm 0.39}$	$1.677_{\pm 0.000}$	$2.05_{\pm 0.33}$
	DSS	$97.03_{\pm 0.57}$	$6.66_{\pm 1.28}$	$89.34_{\pm 0.37}$	$12.03_{\pm 1.75}$	$2.38_{\pm 0.51}$	$0.592_{\pm 0.026}$	$0.75_{\pm 0.07}$
	DSS+	$97.73_{\pm 0.16}$	$5.10_{\pm 0.33}$	$89.84_{\pm 0.13}$	$9.15_{\pm 0.59}$	$3.12_{\pm 0.29}$	$0.584_{\pm 0.019}$	$0.72_{\pm 0.05}$
NPO	FULL	$97.73_{\pm 0.86}$	$7.13_{\pm 2.23}$	$89.86_{\pm 1.10}$	$13.12_{\pm 4.77}$	$3.07_{\pm 1.14}$	$1.954_{\pm 0.000}$	$1.98_{\pm 0.02}$
	DSS	$48.44_{\pm 19.22}$	$68.12_{\pm 20.69}$	$45.56_{\pm 17.45}$	$70.46_{\pm 20.66}$	$54.29_{\pm 16.94}$	$1.551_{\pm 0.092}$	$1.68_{\pm 0.10}$
	DSS+	$79.49_{\pm 17.14}$	$39.81_{\pm 25.39}$	$72.96_{\pm 15.59}$	$53.02_{\pm 26.08}$	$28.25_{\pm 20.84}$	$1.277_{\pm 0.082}$	$1.35_{\pm 0.08}$
UGradSL+	FULL	$97.45_{\pm 0.12}$	$6.06_{\pm 0.34}$	$89.34_{\pm 0.15}$	$9.55_{\pm 0.47}$	$2.97_{\pm 0.36}$	$1.694_{\pm 0.000}$	$2.02_{\pm 0.18}$
	DSS	$96.61_{\pm 0.38}$	$7.54_{\pm 0.99}$	$89.00_{\pm 0.28}$	$13.43_{\pm 0.96}$	$2.30_{\pm 0.42}$	$0.591_{\pm 0.001}$	$0.75_{\pm 0.04}$
	DSS+	$97.63_{\pm 0.16}$	$5.35_{\pm 0.42}$	$89.75_{\pm 0.14}$	$9.59_{\pm 0.62}$	$2.99_{\pm 0.29}$	$0.586_{\pm 0.003}$	$0.72_{\pm 0.02}$
RL	FULL	$99.14_{\pm 0.19}$	$1.67_{\pm 0.39}$	$92.16_{\pm 0.26}$	$4.36_{\pm 0.66}$	$4.24_{\pm 0.36}$	$0.941_{\pm 0.000}$	$1.07_{\pm 0.01}$
	DSS	$95.16_{\pm 8.90}$	$7.46_{\pm 11.92}$	$88.43_{\pm 8.12}$	$11.76_{\pm 11.45}$	$6.84_{\pm 6.52}$	$0.993_{\pm 0.088}$	$1.17_{\pm 0.10}$
	DSS+	$98.58_{\pm 0.67}$	$2.88_{\pm 1.50}$	$91.58_{\pm 0.73}$	$6.99_{\pm 2.88}$	$3.59_{\pm 0.75}$	$0.884_{\pm 0.018}$	$1.02_{\pm 0.02}$
SalUn	FULL	$98.21_{\pm 0.64}$	$3.61_{\pm 1.03}$	$91.03_{\pm 0.72}$	$7.55_{\pm 1.32}$	$3.47_{\pm 0.50}$	$0.951_{\pm 0.000}$	$1.20_{\pm 0.01}$
	DSS	$97.13_{\pm 3.39}$	$4.39_{\pm 2.36}$	$90.26_{\pm 2.60}$	$9.20_{\pm 2.62}$	$3.49_{\pm 1.14}$	$0.997_{\pm 0.073}$	$1.27_{\pm 0.09}$
	DSS+	$97.20_{\pm 2.21}$	$5.37_{\pm 3.72}$	$90.29_{\pm 1.85}$	$10.65_{\pm 4.85}$	$3.68_{\pm 1.72}$	$0.902_{\pm 0.015}$	$1.15_{\pm 0.02}$
AMUN	FULL	$99.92_{\pm 0.03}$	$2.46_{\pm 0.49}$	$93.27_{\pm 0.16}$	$5.25_{\pm 0.75}$	$3.36_{\pm 0.45}$	$1.366_{\pm 0.096}$	$1.55_{\pm 0.06}$
	DSS	$99.88_{\pm 0.06}$	$2.92_{\pm 0.59}$	$93.11_{\pm 0.21}$	$5.74_{\pm 0.97}$	$3.16_{\pm 0.47}$	$0.779_{\pm 0.101}$	$0.95_{\pm 0.07}$
	DSS+	$99.88_{\pm 0.09}$	$2.91_{\pm 0.61}$	$93.19_{\pm 0.34}$	$6.04_{\pm 1.28}$	$3.08_{\pm 0.38}$	$0.777_{\pm 0.105}$	$0.95_{\pm 0.07}$

Table 8: CIFAR-10 with ResNet-18 under random forgetting of 30% of the training data.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$5.89_{\pm 0.26}$	$93.68_{\pm 0.19}$	$14.62_{\pm 0.70}$	$0.00_{\pm 0.00}$	21.228	$13.99_{\pm 2.61}$
NegGrad+	FULL	$97.18_{\pm 0.65}$	$7.57_{\pm 0.70}$	$90.35_{\pm 0.66}$	$13.24_{\pm 0.66}$	$2.33_{\pm 0.38}$	$2.328_{\pm 0.000}$	$2.21_{\pm 0.40}$
	DSS	$97.02_{\pm 0.94}$	$6.68_{\pm 0.87}$	$90.29_{\pm 0.90}$	$12.69_{\pm 1.36}$	$2.33_{\pm 0.41}$	$1.174_{\pm 0.056}$	$1.16_{\pm 0.17}$
	DSS+	$97.79_{\pm 0.49}$	$6.19_{\pm 0.69}$	$91.07_{\pm 0.51}$	$11.59_{\pm 0.87}$	$2.10_{\pm 0.25}$	$1.325_{\pm 0.065}$	$1.29_{\pm 0.23}$
NPO	FULL	$97.09_{\pm 0.51}$	$7.67_{\pm 0.64}$	$90.30_{\pm 0.52}$	$13.96_{\pm 0.85}$	$2.24_{\pm 0.29}$	$2.716_{\pm 0.000}$	$2.39_{\pm 0.36}$
	DSS	$96.45_{\pm 1.87}$	$7.31_{\pm 1.78}$	$89.71_{\pm 1.86}$	$14.00_{\pm 2.85}$	$2.76_{\pm 1.79}$	$1.347_{\pm 0.079}$	$1.23_{\pm 0.19}$
	DSS+	$97.82_{\pm 0.48}$	$6.03_{\pm 0.46}$	$91.23_{\pm 0.58}$	$11.71_{\pm 0.75}$	$2.00_{\pm 0.22}$	$1.524_{\pm 0.055}$	$1.39_{\pm 0.22}$
UGradSL+	FULL	$99.09_{\pm 0.47}$	$3.22_{\pm 0.82}$	$92.30_{\pm 0.78}$	$10.20_{\pm 1.40}$	$2.35_{\pm 0.44}$	$2.333_{\pm 0.000}$	$2.25_{\pm 0.30}$
	DSS	$99.58_{\pm 0.25}$	$2.39_{\pm 0.36}$	$93.09_{\pm 0.46}$	$9.56_{\pm 0.96}$	$2.41_{\pm 0.22}$	$2.284_{\pm 0.035}$	$2.22_{\pm 0.25}$
	DSS+	$99.50_{\pm 0.34}$	$2.36_{\pm 0.61}$	$92.97_{\pm 0.57}$	$9.49_{\pm 0.97}$	$2.47_{\pm 0.32}$	$1.909_{\pm 0.021}$	$1.89_{\pm 0.29}$
RL	FULL	$97.31_{\pm 0.83}$	$4.88_{\pm 0.89}$	$91.73_{\pm 0.87}$	$12.85_{\pm 1.03}$	$1.91_{\pm 0.44}$	$1.666_{\pm 0.000}$	$1.43_{\pm 0.18}$
	DSS	$98.07_{\pm 0.80}$	$3.74_{\pm 1.11}$	$92.44_{\pm 0.62}$	$11.83_{\pm 1.32}$	$2.03_{\pm 0.42}$	$1.761_{\pm 0.034}$	$1.65_{\pm 0.25}$
	DSS+	$97.80_{\pm 0.70}$	$4.14_{\pm 0.85}$	$92.19_{\pm 0.63}$	$12.37_{\pm 1.47}$	$1.93_{\pm 0.38}$	$1.575_{\pm 0.022}$	$1.39_{\pm 0.19}$
SalUn	FULL	$97.00_{\pm 0.42}$	$5.10_{\pm 0.44}$	$91.52_{\pm 0.26}$	$12.34_{\pm 1.29}$	$2.10_{\pm 0.34}$	$1.716_{\pm 0.000}$	$1.86_{\pm 0.45}$
	DSS	$96.97_{\pm 2.21}$	$4.75_{\pm 2.34}$	$91.50_{\pm 1.86}$	$12.17_{\pm 2.93}$	$2.75_{\pm 1.41}$	$1.819_{\pm 0.032}$	$1.97_{\pm 0.49}$
	DSS+	$97.62_{\pm 1.11}$	$4.21_{\pm 1.15}$	$92.00_{\pm 0.88}$	$12.15_{\pm 0.99}$	$2.11_{\pm 0.41}$	$1.620_{\pm 0.012}$	$1.88_{\pm 0.54}$
AMUN	FULL	$99.06_{\pm 0.28}$	$4.56_{\pm 0.54}$	$92.53_{\pm 0.43}$	$8.95_{\pm 0.49}$	$2.27_{\pm 0.25}$	$3.535_{\pm 0.006}$	$3.79_{\pm 0.42}$
	DSS	$99.15_{\pm 0.40}$	$4.25_{\pm 0.68}$	$92.81_{\pm 0.50}$	$8.75_{\pm 0.94}$	$2.31_{\pm 0.33}$	$2.346_{\pm 0.033}$	$2.76_{\pm 0.25}$
	DSS+	$99.21_{\pm 0.58}$	$4.41_{\pm 0.77}$	$92.79_{\pm 0.80}$	$8.97_{\pm 0.88}$	$2.23_{\pm 0.17}$	$2.509_{\pm 0.032}$	$2.88_{\pm 0.32}$

Table 9: CIFAR-10 with ResNet-18 under random forgetting of 50% of the training data.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$7.22_{\pm 0.23}$	$92.36_{\pm 0.26}$	$17.86_{\pm 0.55}$	$0.00_{\pm 0.00}$	15.163	$8.98_{\pm 1.32}$
NegGrad+	FULL	$97.87_{\pm 0.37}$	$5.62_{\pm 0.43}$	$90.79_{\pm 0.52}$	$11.28_{\pm 0.69}$	$2.97_{\pm 0.15}$	$1.666_{\pm 0.000}$	$1.69_{\pm 0.29}$
	DSS	$97.05_{\pm 1.92}$	$5.98_{\pm 2.03}$	$89.85_{\pm 1.63}$	$12.22_{\pm 2.56}$	$3.27_{\pm 0.66}$	$1.050_{\pm 0.035}$	$1.06_{\pm 0.12}$
	DSS+	$98.17_{\pm 0.73}$	$4.84_{\pm 1.04}$	$91.16_{\pm 0.72}$	$10.12_{\pm 1.27}$	$3.29_{\pm 0.24}$	$1.061_{\pm 0.048}$	$1.08_{\pm 0.15}$
NPO	FULL	$97.67_{\pm 0.93}$	$6.12_{\pm 1.03}$	$90.44_{\pm 1.06}$	$12.11_{\pm 1.27}$	$2.83_{\pm 0.19}$	$1.944_{\pm 0.000}$	$1.77_{\pm 0.31}$
	DSS	$96.67_{\pm 2.25}$	$6.69_{\pm 2.26}$	$89.41_{\pm 2.11}$	$13.14_{\pm 2.43}$	$3.25_{\pm 1.06}$	$1.180_{\pm 0.053}$	$1.14_{\pm 0.17}$
	DSS+	$97.79_{\pm 0.63}$	$5.52_{\pm 0.74}$	$90.81_{\pm 0.82}$	$11.36_{\pm 0.80}$	$2.99_{\pm 0.22}$	$1.242_{\pm 0.122}$	$1.25_{\pm 0.27}$
UGradSL+	FULL	$98.56_{\pm 0.91}$	$4.18_{\pm 1.25}$	$91.43_{\pm 1.12}$	$12.26_{\pm 0.76}$	$2.79_{\pm 0.32}$	$1.666_{\pm 0.000}$	$1.64_{\pm 0.20}$
	DSS	$99.11_{\pm 0.52}$	$3.46_{\pm 0.63}$	$91.96_{\pm 0.73}$	$12.09_{\pm 1.03}$	$2.82_{\pm 0.29}$	$1.826_{\pm 0.019}$	$1.73_{\pm 0.19}$
	DSS+	$98.81_{\pm 0.73}$	$3.83_{\pm 0.99}$	$91.57_{\pm 0.86}$	$11.44_{\pm 1.17}$	$2.97_{\pm 0.26}$	$1.452_{\pm 0.005}$	$1.48_{\pm 0.35}$
RL	FULL	$96.73_{\pm 0.50}$	$5.10_{\pm 0.49}$	$91.26_{\pm 0.54}$	$13.61_{\pm 0.63}$	$2.68_{\pm 0.18}$	$1.666_{\pm 0.000}$	$1.51_{\pm 0.39}$
	DSS	$97.73_{\pm 0.80}$	$3.82_{\pm 0.87}$	$92.10_{\pm 0.71}$	$11.52_{\pm 1.30}$	$3.16_{\pm 0.32}$	$1.939_{\pm 0.019}$	$1.75_{\pm 0.26}$
	DSS+	$97.36_{\pm 0.84}$	$4.24_{\pm 0.84}$	$91.85_{\pm 0.68}$	$12.40_{\pm 1.24}$	$2.95_{\pm 0.28}$	$1.601_{\pm 0.011}$	$1.42_{\pm 0.23}$
SalUn	FULL	$95.73_{\pm 0.72}$	$5.99_{\pm 0.80}$	$90.44_{\pm 0.58}$	$13.98_{\pm 1.08}$	$2.83_{\pm 0.16}$	$1.750_{\pm 0.000}$	$2.00_{\pm 0.60}$
	DSS	$96.53_{\pm 1.05}$	$4.90_{\pm 1.20}$	$91.15_{\pm 0.87}$	$13.04_{\pm 1.68}$	$2.95_{\pm 0.23}$	$2.030_{\pm 0.016}$	$2.26_{\pm 0.48}$
	DSS+	$96.22_{\pm 1.42}$	$5.25_{\pm 1.31}$	$90.84_{\pm 1.07}$	$12.83_{\pm 1.21}$	$3.13_{\pm 0.22}$	$1.688_{\pm 0.005}$	$1.92_{\pm 0.46}$
AMUN	FULL	$98.73_{\pm 0.55}$	$4.22_{\pm 0.70}$	$92.20_{\pm 0.63}$	$8.29_{\pm 0.86}$	$3.61_{\pm 0.25}$	$4.763_{\pm 0.006}$	$4.77_{\pm 0.43}$
	DSS	$99.32_{\pm 0.19}$	$3.22_{\pm 0.34}$	$92.94_{\pm 0.35}$	$7.08_{\pm 0.43}$	$4.01_{\pm 0.33}$	$3.320_{\pm 0.022}$	$3.74_{\pm 0.40}$
	DSS+	$99.07_{\pm 0.22}$	$3.84_{\pm 0.38}$	$92.59_{\pm 0.35}$	$7.87_{\pm 0.57}$	$3.67_{\pm 0.27}$	$3.590_{\pm 0.082}$	$3.93_{\pm 0.42}$

Table 10: CIFAR-10 with VGG16 under random forgetting of 30% of the training data.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$7.22_{\pm 0.22}$	$92.51_{\pm 0.18}$	$12.25_{\pm 0.32}$	$0.00_{\pm 0.00}$	13.179	$16.71_{\pm 1.78}$
NegGrad+	FULL	$96.55_{\pm 0.63}$	$8.72_{\pm 0.53}$	$88.99_{\pm 0.79}$	$14.85_{\pm 0.62}$	$2.77_{\pm 0.61}$	$1.315_{\pm 0.000}$	$1.64_{\pm 0.18}$
	DSS	$96.00_{\pm 1.00}$	$8.73_{\pm 0.92}$	$88.32_{\pm 1.03}$	$14.68_{\pm 0.91}$	$3.03_{\pm 0.97}$	$0.830_{\pm 0.045}$	$1.07_{\pm 0.18}$
	DSS+	$89.36_{\pm 22.95}$	$15.02_{\pm 21.68}$	$82.52_{\pm 20.59}$	$19.79_{\pm 19.02}$	$8.99_{\pm 21.02}$	$0.928_{\pm 0.057}$	$1.19_{\pm 0.17}$
NPO	FULL	$97.99_{\pm 0.37}$	$5.89_{\pm 0.48}$	$90.28_{\pm 0.53}$	$10.47_{\pm 0.55}$	$1.84_{\pm 0.07}$	$1.534_{\pm 0.000}$	$1.73_{\pm 0.18}$
	DSS	$97.08_{\pm 3.79}$	$5.95_{\pm 3.90}$	$89.46_{\pm 3.42}$	$10.90_{\pm 4.35}$	$3.14_{\pm 2.93}$	$0.784_{\pm 0.061}$	$0.89_{\pm 0.06}$
	DSS+	$98.04_{\pm 0.98}$	$5.34_{\pm 0.84}$	$90.60_{\pm 1.02}$	$9.88_{\pm 0.92}$	$2.06_{\pm 0.22}$	$0.911_{\pm 0.069}$	$1.08_{\pm 0.25}$
UGradSL+	FULL	$98.18_{\pm 0.37}$	$6.16_{\pm 0.48}$	$90.30_{\pm 0.44}$	$10.94_{\pm 0.62}$	$1.61_{\pm 0.15}$	$1.318_{\pm 0.000}$	$1.62_{\pm 0.17}$
	DSS	$98.22_{\pm 0.35}$	$6.26_{\pm 0.62}$	$90.41_{\pm 0.45}$	$10.96_{\pm 0.87}$	$1.57_{\pm 0.08}$	$1.358_{\pm 0.013}$	$1.66_{\pm 0.14}$
	DSS+	$98.15_{\pm 0.37}$	$6.25_{\pm 0.38}$	$90.25_{\pm 0.57}$	$10.87_{\pm 0.67}$	$1.63_{\pm 0.21}$	$1.145_{\pm 0.009}$	$1.39_{\pm 0.12}$
RL	FULL	$97.22_{\pm 2.24}$	$4.20_{\pm 2.52}$	$89.34_{\pm 2.45}$	$8.54_{\pm 4.27}$	$3.70_{\pm 1.03}$	$0.941_{\pm 0.000}$	$1.07_{\pm 0.10}$
	DSS	$95.18_{\pm 5.17}$	$5.96_{\pm 5.14}$	$87.89_{\pm 4.65}$	$9.58_{\pm 6.20}$	$5.11_{\pm 2.69}$	$1.163_{\pm 0.002}$	$1.35_{\pm 0.08}$
	DSS+	$96.19_{\pm 7.63}$	$5.04_{\pm 7.49}$	$88.63_{\pm 7.28}$	$9.73_{\pm 5.97}$	$4.72_{\pm 4.58}$	$0.913_{\pm 0.000}$	$1.10_{\pm 0.10}$
SalUn	FULL	$97.55_{\pm 2.27}$	$3.89_{\pm 2.60}$	$90.07_{\pm 2.29}$	$10.45_{\pm 6.23}$	$3.58_{\pm 1.01}$	$0.970_{\pm 0.000}$	$1.36_{\pm 0.26}$
	DSS	$95.98_{\pm 6.87}$	$5.47_{\pm 6.85}$	$88.47_{\pm 6.31}$	$8.78_{\pm 3.65}$	$4.58_{\pm 4.16}$	$1.182_{\pm 0.007}$	$1.60_{\pm 0.22}$
	DSS+	$97.06_{\pm 2.99}$	$4.35_{\pm 3.10}$	$89.49_{\pm 2.74}$	$8.75_{\pm 5.05}$	$3.97_{\pm 1.60}$	$0.941_{\pm 0.000}$	$1.37_{\pm 0.27}$
AMUN	FULL	$98.68_{\pm 0.31}$	$6.97_{\pm 0.46}$	$90.94_{\pm 0.30}$	$12.23_{\pm 0.56}$	$0.95_{\pm 0.27}$	$2.051_{\pm 0.316}$	$2.65_{\pm 0.36}$
	DSS	$98.89_{\pm 0.19}$	$6.60_{\pm 0.31}$	$91.23_{\pm 0.22}$	$11.70_{\pm 0.61}$	$0.92_{\pm 0.25}$	$1.503_{\pm 0.331}$	$2.20_{\pm 0.44}$
	DSS+	$98.65_{\pm 0.49}$	$6.95_{\pm 0.35}$	$90.98_{\pm 0.69}$	$12.43_{\pm 0.85}$	$1.01_{\pm 0.41}$	$1.638_{\pm 0.332}$	$2.32_{\pm 0.42}$

Table 11: CIFAR-10 with VGG16 under random forgetting of 50% of the training data.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	$100.00_{\pm 0.00}$	$8.57_{\pm 0.23}$	$91.15_{\pm 0.19}$	$14.89_{\pm 0.34}$	$0.00_{\pm 0.00}$	9.414	$11.42_{\pm 0.94}$
NegGrad+	FULL	$97.05_{\pm 0.65}$	$7.38_{\pm 0.75}$	$88.90_{\pm 0.50}$	$13.16_{\pm 0.99}$	$2.03_{\pm 0.12}$	$0.941_{\pm 0.000}$	$1.09_{\pm 0.06}$
	DSS	$93.02_{\pm 7.53}$	$10.97_{\pm 7.17}$	$85.35_{\pm 6.72}$	$17.01_{\pm 7.25}$	$5.15_{\pm 6.72}$	$0.743_{\pm 0.041}$	$0.95_{\pm 0.11}$
	DSS+	$96.79_{\pm 0.67}$	$7.16_{\pm 0.85}$	$88.80_{\pm 0.65}$	$12.97_{\pm 1.09}$	$2.22_{\pm 0.16}$	$0.707_{\pm 0.043}$	$0.88_{\pm 0.09}$
NPO	FULL	$97.10_{\pm 0.51}$	$7.36_{\pm 0.66}$	$88.90_{\pm 0.67}$	$13.40_{\pm 0.69}$	$1.99_{\pm 0.13}$	$1.098_{\pm 0.000}$	$1.15_{\pm 0.08}$
	DSS	$95.52_{\pm 1.45}$	$8.63_{\pm 1.46}$	$87.49_{\pm 1.49}$	$14.91_{\pm 1.95}$	$2.61_{\pm 1.00}$	$0.838_{\pm 0.041}$	$0.99_{\pm 0.08}$
	DSS+	$96.32_{\pm 1.22}$	$7.76_{\pm 1.27}$	$88.30_{\pm 1.16}$	$14.18_{\pm 1.53}$	$2.26_{\pm 0.43}$	$0.837_{\pm 0.049}$	$0.91_{\pm 0.05}$
UGradSL+	FULL	$98.88_{\pm 0.33}$	$3.22_{\pm 0.48}$	$91.15_{\pm 0.54}$	$7.35_{\pm 0.56}$	$3.61_{\pm 0.15}$	$0.941_{\pm 0.000}$	$1.12_{\pm 0.09}$
	DSS	$98.85_{\pm 0.51}$	$3.62_{\pm 0.62}$	$91.01_{\pm 0.58}$	$8.11_{\pm 0.61}$	$3.34_{\pm 0.19}$	$1.038_{\pm 0.012}$	$1.31_{\pm 0.11}$
	DSS+	$99.02_{\pm 0.22}$	$3.34_{\pm 0.38}$	$91.20_{\pm 0.31}$	$7.71_{\pm 0.72}$	$3.42_{\pm 0.21}$	$0.825_{\pm 0.010}$	$0.98_{\pm 0.05}$
RL	FULL	$83.94_{\pm 11.35}$	$17.27_{\pm 11.25}$	$77.33_{\pm 10.41}$	$15.60_{\pm 14.07}$	$13.06_{\pm 7.79}$	$0.941_{\pm 0.000}$	$0.98_{\pm 0.02}$
	DSS	$88.08_{\pm 8.03}$	$13.35_{\pm 7.86}$	$80.57_{\pm 7.27}$	$20.58_{\pm 9.27}$	$9.57_{\pm 6.20}$	$1.164_{\pm 0.001}$	$1.34_{\pm 0.06}$
	DSS+	$86.94_{\pm 13.18}$	$14.43_{\pm 13.09}$	$80.03_{\pm 12.10}$	$20.31_{\pm 12.59}$	$10.58_{\pm 11.22}$	$0.913_{\pm 0.000}$	$0.97_{\pm 0.01}$
SalUn	FULL	$99.97_{\pm 0.02}$	$0.25_{\pm 0.08}$	$93.57_{\pm 0.13}$	$4.76_{\pm 0.68}$	$5.23_{\pm 0.25}$	$0.988_{\pm 0.000}$	$1.47_{\pm 0.60}$
	DSS	$99.95_{\pm 0.05}$	$0.38_{\pm 0.16}$	$93.42_{\pm 0.25}$	$5.99_{\pm 3.74}$	$4.85_{\pm 0.88}$	$1.212_{\pm 0.000}$	$1.59_{\pm 0.18}$
	DSS+	$99.94_{\pm 0.07}$	$0.40_{\pm 0.18}$	$93.36_{\pm 0.40}$	$4.25_{\pm 0.60}$	$5.27_{\pm 0.22}$	$0.960_{\pm 0.000}$	$1.29_{\pm 0.19}$
AMUN	FULL	$98.28_{\pm 0.31}$	$6.32_{\pm 0.56}$	$90.34_{\pm 0.36}$	$11.25_{\pm 0.83}$	$2.11_{\pm 0.32}$	$2.817_{\pm 0.580}$	$3.00_{\pm 0.41}$
	DSS	$98.66_{\pm 0.52}$	$5.74_{\pm 0.79}$	$91.15_{\pm 0.54}$	$10.42_{\pm 1.11}$	$2.24_{\pm 0.35}$	$2.163_{\pm 0.603}$	$2.48_{\pm 0.43}$
	DSS+	$98.42_{\pm 0.70}$	$6.10_{\pm 0.84}$	$90.60_{\pm 0.77}$	$11.37_{\pm 1.71}$	$2.08_{\pm 0.38}$	$2.328_{\pm 0.595}$	$2.61_{\pm 0.41}$

B.4 CLASS UNLEARNING

We evaluate class unlearning by removing all training examples of class 0 from CIFAR-10 and CIFAR-100, using ResNet-18 and VGG16.

We evaluate class unlearning by removing all training examples of class 0 from CIFAR-10 and CIFAR-100, using ResNet-18 and VGG16. Each run stops at the first epoch at which UA reaches 100%, or after 50 epochs otherwise; this per-epoch check is excluded from the reported cost. Learning rates are in $[10^{-5}, 7 \times 10^{-2}]$, and the initial selection fractions are the same as for random forgetting. Because the deletion request is fixed, all seeds share one retrained model, and standard deviations reflect only unlearning randomness.

Table 12: Class unlearning on CIFAR-10 with ResNet-18, removing all training examples of class 0.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	100.00	100.00	85.77	100.00	0.00	27.293	14.85
NegGrad+	FULL	$92.00_{\pm 1.85}$	$100.00_{\pm 0.00}$	$78.23_{\pm 1.48}$	$100.00_{\pm 0.00}$	$3.88_{\pm 0.83}$	$0.297_{\pm 0.000}$	$0.33_{\pm 0.06}$
	DSS	$80.22_{\pm 7.91}$	$100.00_{\pm 0.00}$	$68.76_{\pm 6.39}$	$100.00_{\pm 0.00}$	$9.20_{\pm 3.57}$	$0.372_{\pm 0.029}$	$0.42_{\pm 0.07}$
	DSS+	$97.71_{\pm 1.22}$	$100.00_{\pm 0.00}$	$82.74_{\pm 1.01}$	$100.00_{\pm 0.00}$	$1.33_{\pm 0.56}$	$0.347_{\pm 0.025}$	$0.38_{\pm 0.08}$
NPO	FULL	$98.18_{\pm 0.50}$	$100.00_{\pm 0.00}$	$83.22_{\pm 0.45}$	$100.00_{\pm 0.00}$	$1.09_{\pm 0.23}$	$0.346_{\pm 0.000}$	$0.35_{\pm 0.05}$
	DSS	$90.45_{\pm 4.84}$	$100.00_{\pm 0.00}$	$76.95_{\pm 3.78}$	$100.00_{\pm 0.00}$	$4.59_{\pm 2.15}$	$0.371_{\pm 0.017}$	$0.41_{\pm 0.06}$
	DSS+	$99.19_{\pm 0.34}$	$100.00_{\pm 0.00}$	$84.16_{\pm 0.52}$	$100.00_{\pm 0.00}$	$0.60_{\pm 0.21}$	$0.341_{\pm 0.042}$	$0.37_{\pm 0.08}$
UGradSL+	FULL	$93.65_{\pm 0.37}$	$100.00_{\pm 0.00}$	$79.41_{\pm 0.36}$	$100.00_{\pm 0.00}$	$3.18_{\pm 0.18}$	$1.800_{\pm 0.000}$	$1.85_{\pm 0.24}$
	DSS	$99.82_{\pm 0.09}$	$99.97_{\pm 0.03}$	$84.93_{\pm 0.21}$	$100.00_{\pm 0.00}$	$0.26_{\pm 0.07}$	$3.533_{\pm 0.107}$	$4.01_{\pm 0.56}$
	DSS+	$98.64_{\pm 0.17}$	$100.00_{\pm 0.00}$	$83.71_{\pm 0.23}$	$100.00_{\pm 0.00}$	$0.85_{\pm 0.10}$	$1.734_{\pm 0.052}$	$1.95_{\pm 0.28}$
RL	FULL	$99.47_{\pm 0.13}$	$100.00_{\pm 0.00}$	$84.73_{\pm 0.27}$	$100.00_{\pm 0.00}$	$0.39_{\pm 0.10}$	$0.167_{\pm 0.000}$	$0.18_{\pm 0.03}$
	DSS	$99.56_{\pm 0.05}$	$100.00_{\pm 0.00}$	$84.74_{\pm 0.11}$	$100.00_{\pm 0.00}$	$0.37_{\pm 0.04}$	$0.272_{\pm 0.007}$	$0.30_{\pm 0.04}$
	DSS+	$99.57_{\pm 0.09}$	$100.00_{\pm 0.00}$	$84.79_{\pm 0.20}$	$100.00_{\pm 0.00}$	$0.35_{\pm 0.07}$	$0.270_{\pm 0.007}$	$0.29_{\pm 0.04}$
SalUn	FULL	$99.02_{\pm 0.32}$	$100.00_{\pm 0.00}$	$84.32_{\pm 0.40}$	$100.00_{\pm 0.00}$	$0.61_{\pm 0.18}$	$0.183_{\pm 0.000}$	$0.26_{\pm 0.07}$
	DSS	$96.59_{\pm 8.51}$	$100.00_{\pm 0.00}$	$82.25_{\pm 7.18}$	$99.98_{\pm 0.05}$	$1.74_{\pm 3.93}$	$0.266_{\pm 0.066}$	$0.33_{\pm 0.11}$
	DSS+	$96.39_{\pm 8.45}$	$100.00_{\pm 0.00}$	$82.04_{\pm 7.11}$	$99.98_{\pm 0.05}$	$1.84_{\pm 3.90}$	$0.264_{\pm 0.065}$	$0.32_{\pm 0.10}$
AMUN	FULL	$97.78_{\pm 0.41}$	$34.16_{\pm 12.76}$	$89.06_{\pm 1.35}$	$57.78_{\pm 15.80}$	$28.39_{\pm 7.36}$	$9.573_{\pm 0.000}$	$10.53_{\pm 1.27}$
	DSS	$98.49_{\pm 0.75}$	$26.45_{\pm 8.69}$	$90.16_{\pm 0.73}$	$51.26_{\pm 13.24}$	$32.05_{\pm 5.64}$	$6.342_{\pm 0.075}$	$6.99_{\pm 0.84}$
	DSS+	$97.81_{\pm 0.34}$	$39.85_{\pm 12.88}$	$88.67_{\pm 1.21}$	$63.78_{\pm 14.95}$	$25.37_{\pm 7.12}$	$8.353_{\pm 0.112}$	$9.22_{\pm 1.04}$

Table 13: Class unlearning on CIFAR-10 with VGG16, removing all training examples of class 0.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	100.00	100.00	84.55	100.00	0.00	16.945	19.80
NegGrad+	FULL	$99.41_{\pm 0.02}$	$100.00_{\pm 0.00}$	$83.28_{\pm 0.07}$	$100.00_{\pm 0.00}$	$0.47_{\pm 0.02}$	$1.677_{\pm 0.000}$	$1.93_{\pm 0.13}$
	DSS	$99.68_{\pm 0.05}$	$100.00_{\pm 0.00}$	$83.47_{\pm 0.11}$	$100.00_{\pm 0.00}$	$0.35_{\pm 0.04}$	$2.215_{\pm 0.073}$	$2.78_{\pm 0.15}$
	DSS+	$99.28_{\pm 0.07}$	$100.00_{\pm 0.00}$	$83.08_{\pm 0.12}$	$100.00_{\pm 0.00}$	$0.55_{\pm 0.05}$	$2.013_{\pm 0.026}$	$2.57_{\pm 0.18}$
NPO	FULL	$99.62_{\pm 0.08}$	$100.00_{\pm 0.00}$	$83.70_{\pm 0.10}$	$100.00_{\pm 0.00}$	$0.31_{\pm 0.04}$	$0.195_{\pm 0.000}$	$0.20_{\pm 0.00}$
	DSS	$87.43_{\pm 11.69}$	$100.00_{\pm 0.00}$	$73.64_{\pm 9.26}$	$100.00_{\pm 0.00}$	$5.87_{\pm 5.24}$	$0.194_{\pm 0.006}$	$0.24_{\pm 0.01}$
	DSS+	$99.78_{\pm 0.03}$	$100.00_{\pm 0.00}$	$83.91_{\pm 0.13}$	$100.00_{\pm 0.00}$	$0.21_{\pm 0.04}$	$0.247_{\pm 0.033}$	$0.30_{\pm 0.04}$
UGradSL+	FULL	$98.63_{\pm 0.12}$	$100.00_{\pm 0.00}$	$82.47_{\pm 0.12}$	$100.00_{\pm 0.00}$	$0.86_{\pm 0.06}$	$0.169_{\pm 0.000}$	$0.20_{\pm 0.00}$
	DSS	$99.74_{\pm 0.14}$	$100.00_{\pm 0.00}$	$83.90_{\pm 0.22}$	$100.00_{\pm 0.00}$	$0.22_{\pm 0.09}$	$0.392_{\pm 0.086}$	$0.50_{\pm 0.11}$
	DSS+	$99.67_{\pm 0.05}$	$100.00_{\pm 0.00}$	$83.60_{\pm 0.08}$	$100.00_{\pm 0.00}$	$0.32_{\pm 0.03}$	$0.279_{\pm 0.019}$	$0.35_{\pm 0.02}$
RL	FULL	$97.93_{\pm 0.83}$	$100.00_{\pm 0.00}$	$82.46_{\pm 0.64}$	$100.00_{\pm 0.00}$	$1.04_{\pm 0.36}$	$0.094_{\pm 0.000}$	$0.11_{\pm 0.00}$
	DSS	$95.88_{\pm 5.16}$	$100.00_{\pm 0.00}$	$80.47_{\pm 4.56}$	$100.00_{\pm 0.01}$	$2.05_{\pm 2.43}$	$0.115_{\pm 0.054}$	$0.14_{\pm 0.06}$
	DSS+	$95.95_{\pm 5.19}$	$100.00_{\pm 0.00}$	$80.47_{\pm 4.56}$	$100.00_{\pm 0.01}$	$2.03_{\pm 2.44}$	$0.114_{\pm 0.054}$	$0.14_{\pm 0.06}$
SalUn	FULL	$93.45_{\pm 3.49}$	$100.00_{\pm 0.00}$	$79.07_{\pm 2.61}$	$100.00_{\pm 0.00}$	$3.01_{\pm 1.52}$	$0.104_{\pm 0.000}$	$0.14_{\pm 0.00}$
	DSS	$83.17_{\pm 7.10}$	$100.00_{\pm 0.00}$	$70.01_{\pm 5.59}$	$100.00_{\pm 0.01}$	$7.84_{\pm 3.17}$	$0.057_{\pm 0.036}$	$0.08_{\pm 0.05}$
	DSS+	$83.21_{\pm 7.19}$	$100.00_{\pm 0.00}$	$70.06_{\pm 5.72}$	$100.00_{\pm 0.01}$	$7.82_{\pm 3.23}$	$0.057_{\pm 0.035}$	$0.08_{\pm 0.04}$
AMUN	FULL	$99.17_{\pm 0.26}$	$38.13_{\pm 7.10}$	$88.97_{\pm 0.80}$	$71.07_{\pm 10.40}$	$24.01_{\pm 4.43}$	$5.373_{\pm 0.000}$	$6.25_{\pm 0.45}$
	DSS	$99.54_{\pm 0.31}$	$20.87_{\pm 7.18}$	$90.70_{\pm 0.86}$	$49.98_{\pm 16.09}$	$33.94_{\pm 5.76}$	$3.185_{\pm 0.035}$	$3.91_{\pm 0.22}$
	DSS+	$99.46_{\pm 0.20}$	$34.88_{\pm 9.07}$	$89.62_{\pm 0.99}$	$63.08_{\pm 12.25}$	$26.91_{\pm 5.38}$	$4.060_{\pm 0.336}$	$4.88_{\pm 0.44}$

Table 14: Class unlearning on CIFAR-100 with ResNet-18, removing all training examples of class 0.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	99.98	100.00	76.43	100.00	0.00	32.995	26.43
NegGrad+	FULL	$94.66_{\pm 0.25}$	$100.00_{\pm 0.00}$	$68.29_{\pm 0.29}$	$100.00_{\pm 0.00}$	$3.36_{\pm 0.13}$	$0.327_{\pm 0.000}$	$0.29_{\pm 0.00}$
	DSS	$96.71_{\pm 0.14}$	$100.00_{\pm 0.00}$	$69.50_{\pm 0.10}$	$100.00_{\pm 0.00}$	$2.55_{\pm 0.05}$	$0.206_{\pm 0.000}$	$0.19_{\pm 0.00}$
	DSS+	$96.71_{\pm 0.14}$	$100.00_{\pm 0.00}$	$69.50_{\pm 0.10}$	$100.00_{\pm 0.00}$	$2.55_{\pm 0.05}$	$0.206_{\pm 0.000}$	$0.19_{\pm 0.00}$
NPO	FULL	$97.80_{\pm 0.13}$	$100.00_{\pm 0.00}$	$70.82_{\pm 0.22}$	$100.00_{\pm 0.00}$	$1.95_{\pm 0.08}$	$0.380_{\pm 0.000}$	$0.31_{\pm 0.00}$
	DSS	$98.37_{\pm 0.12}$	$100.00_{\pm 0.00}$	$71.24_{\pm 0.20}$	$100.00_{\pm 0.00}$	$1.70_{\pm 0.07}$	$0.229_{\pm 0.000}$	$0.19_{\pm 0.00}$
	DSS+	$98.37_{\pm 0.12}$	$100.00_{\pm 0.00}$	$71.24_{\pm 0.20}$	$100.00_{\pm 0.00}$	$1.70_{\pm 0.07}$	$0.229_{\pm 0.000}$	$0.19_{\pm 0.00}$
UGradSL+	FULL	$95.86_{\pm 0.20}$	$100.00_{\pm 0.00}$	$69.05_{\pm 0.18}$	$100.00_{\pm 0.00}$	$2.87_{\pm 0.09}$	$0.660_{\pm 0.000}$	$0.60_{\pm 0.00}$
	DSS	$96.45_{\pm 0.12}$	$100.00_{\pm 0.00}$	$69.53_{\pm 0.14}$	$100.00_{\pm 0.00}$	$2.61_{\pm 0.06}$	$0.849_{\pm 0.071}$	$0.77_{\pm 0.06}$
	DSS+	$96.32_{\pm 0.09}$	$100.00_{\pm 0.00}$	$69.38_{\pm 0.16}$	$100.00_{\pm 0.00}$	$2.68_{\pm 0.06}$	$0.680_{\pm 0.001}$	$0.62_{\pm 0.00}$
RL	FULL	$69.08_{\pm 5.26}$	$100.00_{\pm 0.00}$	$55.12_{\pm 3.63}$	$99.98_{\pm 0.06}$	$13.06_{\pm 2.22}$	$0.200_{\pm 0.070}$	$0.18_{\pm 0.06}$
	DSS	$43.43_{\pm 19.26}$	$100.00_{\pm 0.00}$	$36.06_{\pm 14.52}$	$99.84_{\pm 0.26}$	$24.27_{\pm 8.47}$	$0.134_{\pm 0.100}$	$0.13_{\pm 0.09}$
	DSS+	$44.56_{\pm 21.13}$	$100.00_{\pm 0.00}$	$36.72_{\pm 15.60}$	$99.84_{\pm 0.26}$	$23.82_{\pm 9.21}$	$0.166_{\pm 0.176}$	$0.16_{\pm 0.16}$
SalUn	FULL	$99.16_{\pm 0.31}$	$100.00_{\pm 0.00}$	$72.46_{\pm 0.64}$	$100.00_{\pm 0.00}$	$1.20_{\pm 0.23}$	$2.251_{\pm 0.251}$	$2.18_{\pm 0.24}$
	DSS	$98.37_{\pm 1.89}$	$99.70_{\pm 0.95}$	$71.16_{\pm 2.16}$	$100.00_{\pm 0.00}$	$1.80_{\pm 0.97}$	$2.926_{\pm 0.821}$	$2.80_{\pm 0.79}$
	DSS+	$98.72_{\pm 1.80}$	$100.00_{\pm 0.00}$	$72.07_{\pm 2.36}$	$100.00_{\pm 0.00}$	$1.40_{\pm 1.04}$	$2.506_{\pm 0.272}$	$2.44_{\pm 0.27}$
AMUN	FULL	$85.73_{\pm 3.41}$	$25.18_{\pm 10.98}$	$63.01_{\pm 2.11}$	$27.94_{\pm 12.99}$	$43.64_{\pm 6.51}$	$8.471_{\pm 0.000}$	$7.60_{\pm 0.05}$
	DSS	$89.92_{\pm 1.20}$	$33.68_{\pm 8.12}$	$64.51_{\pm 0.84}$	$40.24_{\pm 9.19}$	$37.01_{\pm 4.37}$	$6.946_{\pm 0.035}$	$6.38_{\pm 0.05}$
	DSS+	$86.59_{\pm 1.19}$	$36.18_{\pm 18.37}$	$63.71_{\pm 0.86}$	$39.32_{\pm 18.18}$	$37.65_{\pm 9.01}$	$8.244_{\pm 0.027}$	$7.47_{\pm 0.05}$

Table 15: Class unlearning on CIFAR-100 with VGG16, removing all training examples of class 0.

Method	Variant	RA	UA	TA	MIA	Avg. Gap ↓	PFLOPs ↓	RTE (min) ↓
Retrain	–	99.97	100.00	72.28	100.00	0.00	18.640	21.38
NegGrad+	FULL	$94.75_{\pm 0.20}$	$100.00_{\pm 0.00}$	$65.86_{\pm 0.14}$	$100.00_{\pm 0.00}$	$2.91_{\pm 0.08}$	$1.642_{\pm 0.058}$	$1.91_{\pm 0.19}$
	DSS	$94.49_{\pm 0.21}$	$100.00_{\pm 0.00}$	$65.70_{\pm 0.16}$	$100.00_{\pm 0.00}$	$3.01_{\pm 0.09}$	$1.903_{\pm 0.003}$	$2.28_{\pm 0.16}$
	DSS+	$94.74_{\pm 0.17}$	$100.00_{\pm 0.00}$	$65.91_{\pm 0.16}$	$100.00_{\pm 0.00}$	$2.90_{\pm 0.08}$	$1.622_{\pm 0.003}$	$2.02_{\pm 0.15}$
NPO	FULL	$97.38_{\pm 0.10}$	$100.00_{\pm 0.00}$	$68.17_{\pm 0.12}$	$100.00_{\pm 0.00}$	$1.67_{\pm 0.04}$	$0.215_{\pm 0.000}$	$0.22_{\pm 0.00}$
	DSS	$96.94_{\pm 0.21}$	$100.00_{\pm 0.00}$	$67.87_{\pm 0.15}$	$100.00_{\pm 0.00}$	$1.86_{\pm 0.08}$	$0.124_{\pm 0.000}$	$0.14_{\pm 0.00}$
	DSS+	$96.94_{\pm 0.21}$	$100.00_{\pm 0.00}$	$67.87_{\pm 0.15}$	$100.00_{\pm 0.00}$	$1.86_{\pm 0.08}$	$0.124_{\pm 0.000}$	$0.14_{\pm 0.00}$
UGradSL+	FULL	$93.85_{\pm 0.12}$	$100.00_{\pm 0.00}$	$65.33_{\pm 0.07}$	$100.00_{\pm 0.00}$	$3.27_{\pm 0.05}$	$0.559_{\pm 0.000}$	$0.64_{\pm 0.00}$
	DSS	$93.99_{\pm 0.07}$	$100.00_{\pm 0.00}$	$65.36_{\pm 0.08}$	$100.00_{\pm 0.00}$	$3.22_{\pm 0.04}$	$0.617_{\pm 0.004}$	$0.73_{\pm 0.01}$
	DSS+	$93.78_{\pm 0.17}$	$100.00_{\pm 0.00}$	$65.24_{\pm 0.13}$	$100.00_{\pm 0.00}$	$3.31_{\pm 0.07}$	$0.586_{\pm 0.002}$	$0.68_{\pm 0.01}$
RL	FULL	$97.17_{\pm 1.29}$	$100.00_{\pm 0.00}$	$68.52_{\pm 0.97}$	$100.00_{\pm 0.00}$	$1.64_{\pm 0.56}$	$0.424_{\pm 0.584}$	$0.49_{\pm 0.67}$
	DSS	$93.04_{\pm 5.99}$	$100.00_{\pm 0.00}$	$65.39_{\pm 3.45}$	$100.00_{\pm 0.00}$	$3.45_{\pm 2.35}$	$1.508_{\pm 1.049}$	$1.84_{\pm 1.28}$
	DSS+	$88.41_{\pm 14.88}$	$100.00_{\pm 0.00}$	$63.60_{\pm 7.50}$	$100.00_{\pm 0.00}$	$5.06_{\pm 5.59}$	$1.392_{\pm 0.869}$	$1.66_{\pm 1.04}$
SalUn	FULL	$98.78_{\pm 0.71}$	$100.00_{\pm 0.00}$	$70.00_{\pm 0.91}$	$100.00_{\pm 0.00}$	$0.87_{\pm 0.39}$	$0.490_{\pm 0.674}$	$0.62_{\pm 0.84}$
	DSS	$95.20_{\pm 5.19}$	$100.00_{\pm 0.00}$	$66.96_{\pm 3.34}$	$100.00_{\pm 0.00}$	$2.52_{\pm 2.13}$	$1.407_{\pm 1.072}$	$1.87_{\pm 1.39}$
	DSS+	$94.47_{\pm 5.07}$	$100.00_{\pm 0.00}$	$66.82_{\pm 3.29}$	$100.00_{\pm 0.00}$	$2.74_{\pm 2.08}$	$1.021_{\pm 0.686}$	$1.35_{\pm 0.86}$
AMUN	FULL	$86.84_{\pm 1.61}$	$41.30_{\pm 20.41}$	$63.17_{\pm 1.02}$	$44.86_{\pm 21.88}$	$34.02_{\pm 10.80}$	$4.806_{\pm 0.000}$	$5.68_{\pm 0.39}$
	DSS	$84.87_{\pm 7.98}$	$45.98_{\pm 18.82}$	$61.21_{\pm 4.78}$	$48.08_{\pm 18.35}$	$33.03_{\pm 8.21}$	$5.072_{\pm 0.198}$	$6.25_{\pm 0.44}$
	DSS+	$79.70_{\pm 22.15}$	$54.62_{\pm 25.39}$	$58.02_{\pm 15.38}$	$56.82_{\pm 24.56}$	$30.77_{\pm 10.05}$	$4.312_{\pm 1.318}$	$5.08_{\pm 1.49}$

C SOURCE UTILITY UNDER A RETRAIN REFERENCE

We prove Proposition 1, then consider unequal initial confidence, repeated updates, and relabeling. The final subsection constructs training objectives whose optima are the model pair used in the analysis.

C.1 MODEL, SOURCES, AND COMPARISON CRITERIA

We use the binary logistic model from Section 4. For an input x, label $y \in \{-1, +1\}$, and parameter vector θ, the true-label probability is $p_\theta(y \mid x) = \sigma(yx^\top\theta)$, where $\sigma(z) = (1 + e^{-z})^{-1}$. Its cross-entropy loss is $\ell(\theta; x, y) = \log(1 + e^{-yx^\top\theta})$. The parameters before unlearning, θ_0, and after retraining on the retain set, θ_R, satisfy

$$\theta_R = \beta v, \qquad \theta_0 = \beta v + \gamma u, \qquad 0 < \gamma \le \beta, \tag{C.1}$$

where u, v are orthogonal unit vectors. The component γu is the difference between the two models.

Retain inputs have variance $\varepsilon \in (0, 1]$ along u and unit variance in orthogonal directions. With I the identity matrix, their covariance is $\Sigma_\varepsilon = I - (1 - \varepsilon)uu^\top$:

$$X_R \sim \mathcal{N}(0, \Sigma_\varepsilon), \qquad \Pr(Y_R = 1 \mid X_R) = \sigma(X_R^\top\theta_R).$$

We compare three quantities, all smaller when better: predictive KL Q on isotropic Gaussian reference inputs, retain cross-entropy L_R, and retain misclassification probability $\mathcal{R}_{01}$:

$$Q(\theta) = \mathbb{E}_{X\sim\mathcal{N}(0,I)}\,\mathrm{KL}(p_{\theta_R}(\cdot \mid X)\|p_\theta(\cdot \mid X))\,, \tag{C.2}$$

$$L_R(\theta) = \mathbb{E}\ell(\theta; X_R, Y_R), \qquad \mathcal{R}_{01}(\theta) = \Pr\big(\mathrm{sign}(X_R^\top\theta) \ne Y_R\big)\,. \tag{C.3}$$

The expectations in the second line use the retain distribution above.

Source confidence and direction. For each source $i \in \{L, H\}$, assume $\|x_i\| = 1$ and write $a_i = y_i x_i$ for its signed feature vector. The initial and retrained true-label log odds are $m_i = a_i^\top\theta_0$ and $t_i = a_i^\top\theta_R$, respectively. Thus their confidences and difference satisfy

$$c_i^0 = \sigma(m_i), \qquad c_i^R = \sigma(t_i), \qquad r_i = m_i - t_i = \gamma u^\top a_i. \tag{C.4}$$

Here r_i measures the source component along $\theta_0 - \theta_R$; we assume $r_i \ge 0$ throughout. The squared source component orthogonal to both u and v is

$$d_i = 1 - r_i^2/\gamma^2 - t_i^2/\beta^2 \ge 0. \tag{C.5}$$

In particular, $r_i \le \gamma$ and $|t_i| \le \beta$. The initial CE-ascent multiplier is $h_i = 1 - c_i^0$. One update with learning rate $\eta > 0$ gives

$$\theta_{i,1} = \theta_0 - \eta h_i a_i. \tag{C.6}$$

Gaussian loss identities. For a standard normal variable Z, define

$$F(q) = \mathbb{E}\log(1 + e^{qZ}), \qquad A(q) = \mathbb{E}\sigma'(qZ), \qquad q \ge 0.$$

We use $A_\beta = A(\beta)$, $A_{\mathrm{init}} = A(\sqrt{\beta^2 + \gamma^2})$, and $A_\varepsilon = A(\sqrt{\beta^2 + \varepsilon\gamma^2})$. The function A is positive and strictly decreasing for $q > 0$, so $A_\beta > A_\varepsilon \ge A_{\mathrm{init}} > 0$. Gaussian integration by parts gives $F'(q) = qA(q)$ and

$$Q(\theta) = F(\|\theta\|) - A_\beta\theta_R^\top\theta + \text{constant}, \tag{C.7}$$

$$L_R(\theta) = F\Big(\sqrt{\theta^\top\Sigma_\varepsilon\theta}\Big) - A_\beta\theta_R^\top\theta + \text{constant}. \tag{C.8}$$

Each constant is independent of θ. For example, integration by parts yields $\mathbb{E}[X\sigma(X^\top\theta_R)] = A_\beta\theta_R$ for isotropic inputs; the same identity holds for retain inputs because $\Sigma_\varepsilon\theta_R = \theta_R$.

Retain classification error. A scalar displacement s along a source gives parameters $\theta_0 - sa_i$. Its retain logit can be written as $b_i(s)Z + \sqrt{N_i(s)}Z'$, where Z, Z' are independent standard normals. The coefficient $b_i(s)$ multiplies the retrained signal, while $N_i(s)$ is the variance of the independent component:

$$b_i(s) = \beta - \frac{t_i s}{\beta}, \qquad N_i(s) = \varepsilon \left(\gamma - \frac{r_i s}{\gamma}\right)^2 + d_i s^2. \tag{C.9}$$

When $b_i(s) > 0$, define the noise-to-signal ratio $w_i(s) = \sqrt{N_i(s)}/b_i(s)$. Let ϕ and Φ denote the standard normal density and CDF. Conditioning on Z gives

$$\begin{aligned} \mathcal{R}_{01}(\theta_0 - sa_i) &= \mathcal{R}_{01}(\theta_R) + \mathcal{E}(w_i(s)), \\ \mathcal{E}(w) &= 2\int_0^\infty (2\sigma(\beta z) - 1)\Phi(-z/w)\phi(z)\, dz, \qquad w > 0. \end{aligned} \tag{C.10}$$

The excess error $\mathcal{E}$ is smooth and strictly increasing for $w > 0$. At the initial model, the ratio is $w_0 = \sqrt{\varepsilon}\gamma/\beta > 0$. The analysis uses $\varepsilon > 0$; at $\varepsilon = 0$, the two models already make identical retain predictions, so strict retain improvement over θ_0 is unavailable.

Initial change along a source. To compare updates, let f stand for one of $Q, L_R, \mathcal{R}_{01}$. Write $\varphi_i^f(s) = f(\theta_0 - sa_i)$ for that criterion along source i, and $g_i^f = -(\varphi_i^f)'(0)$ for its initial rate of decrease. A positive g_i^f means a small positive displacement improves f. Differentiating the preceding identities gives

$$\begin{aligned} g_i^Q &= A_{\text{init}} r_i - (A_\beta - A_{\text{init}})t_i, \\ g_i^{L_R} &= \varepsilon A_\varepsilon r_i - (A_\beta - A_\varepsilon)t_i, \\ g_i^{\mathcal{R}_{01}} &= \kappa\left(r_i - \frac{\gamma^2}{\beta^2}t_i\right), \qquad \kappa = \mathcal{E}'(w_0)\frac{\sqrt{\varepsilon}}{\gamma\beta} > 0. \end{aligned} \tag{C.11}$$

For the last identity, $w_i'(0) = -\sqrt{\varepsilon}(r_i - \gamma^2 t_i/\beta^2)/(\gamma\beta)$. The positive factor κ is shared by both sources. These rates separate the source direction from the multiplier used by a particular update rule.

C.2 Matched initial confidence

We first prove Proposition 1. Suppose the sources share their initial confidence but differ under θ_R:

$$m_L = m_H = m, \qquad t_L < t_H, \qquad r_L, r_H \geq 0.$$

Define the retrained log-odds gap $\Delta = t_H - t_L > 0$. Since $r_i = m - t_i$, we also have $r_L - r_H = \Delta$. The two updates have the same displacement $s = \eta(1 - \sigma(m))$.

Exact predictive KL comparison. For either source, $\|\theta_0 - sa_i\|^2 = \beta^2 + \gamma^2 - 2ms + s^2$. The norm terms in equation C.7 therefore cancel:

$$Q(\theta_0 - sa_L) - Q(\theta_0 - sa_H) = -A_\beta \Delta s < 0 \qquad (s > 0). \tag{C.12}$$

Thus the source with lower retrained confidence gives lower predictive KL at every common positive displacement.

Retain comparisons. Equation C.11 gives

$$\begin{aligned} g_L^{L_R} - g_H^{L_R} &= [A_\beta - (1-\varepsilon)A_\varepsilon]\Delta > 0, \\ g_L^{\mathcal{R}_{01}} - g_H^{\mathcal{R}_{01}} &= \kappa(1 + \gamma^2/\beta^2)\Delta > 0. \end{aligned}$$

The coefficient in the first line is positive because $A_\beta > A_\varepsilon$. For either retain criterion, the difference between the two paths starts at zero and has a strictly negative derivative at $s = 0$. It is therefore negative for all sufficiently small positive s. Together with equation C.12, this proves the three comparisons in Proposition 1. These comparisons do not require $m > 0$ or $t_L \leq 0$.

Repeated updates at matched confidence. For repeated CE ascent, let $s_{i,k}$ be the displacement after k updates, starting from $s_{i,0} = 0$. Unit source norm gives the current confidence $c_{i,k} = \sigma(m - s_{i,k})$ and the recurrence

$$s_{i,k+1} = s_{i,k} + \eta_k \sigma(s_{i,k} - m), \qquad \theta_{i,k} = \theta_0 - s_{i,k} a_i. \tag{C.13}$$

Here $\eta_k > 0$ is the learning rate shared by the two paths. Their recurrences are identical, so $s_{L,k} = s_{H,k} > 0$ for every $k \geq 1$. The predictive KL ordering consequently holds at every such update. The retain comparisons hold while the common displacement remains in the neighborhood just established. Improvement over θ_0 requires additional conditions, given next.

C.3 Unequal initial confidence and improvement over the initial model

For unequal initial confidence, we give sufficient conditions for the L update to improve each criterion relative to both the H update and θ_0. Assume

$$0 < m_L \leq m_H, \qquad t_L < t_H, \qquad t_L \leq 0, \qquad r_L, r_H \geq 0. \tag{C.14}$$

Thus both sources are initially classified correctly, and L has retrained confidence at most one half. Define the initial and retrained log-odds gaps by $\delta = m_H - m_L \geq 0$ and $\Delta = t_H - t_L > 0$. We require $\delta \leq \Delta$, which is equivalent to $r_L \geq r_H$.

Proposition C.1 (Unequal initial confidence)**.** *Under equation C.14 and $\delta \leq \Delta$, one CE-ascent update with a sufficiently small common $\eta > 0$ satisfies*

$$f(\theta_{L,1}) < \min\{f(\theta_0), f(\theta_{H,1})\}, \qquad f \in \{Q, L_R, \mathcal{R}_{01}\}. \tag{C.15}$$

Proof. For each criterion, the initial improvement per unit learning rate is $G_i^f = h_i g_i^f$, because the displacement is ηh_i. We show $G_L^f > \max\{0, G_H^f\}$. Since $m_L \leq m_H$, we have $h_L \geq h_H > 0$. Moreover, $h_L t_L < h_H t_H$: when $t_H \geq 0$ this follows from $t_L \leq 0$; when $t_H < 0$, it follows from $t_L < t_H$ and $h_L \geq h_H$. Using $r_i = m_i - t_i$ then gives

$$h_L r_L - h_H r_H = (h_L - h_H) r_L + h_H(\Delta - \delta) \geq 0.$$

Every expression in equation C.11 has the form $a r_i - b t_i$ with $a, b > 0$. The preceding inequalities therefore imply $G_L^f > G_H^f$. Also $r_L = m_L - t_L > 0$ and $t_L \leq 0$, so $G_L^f > 0$. Finally, $f(\theta_{i,1}) = f(\theta_0) - \eta G_i^f + O(\eta^2)$, which proves both strict comparisons for sufficiently small η. □

C.4 Repeated cross-entropy ascent

We now repeat updates using each fixed source, recomputing its confidence at every step. Use a common, predetermined schedule $\eta_j > 0$ and define $T_k = \sum_{j=0}^{k-1} \eta_j$, the cumulative learning rate after k updates. Writing $\theta_{i,k} = \theta_0 - s_{i,k} a_i$, the displacement and current confidence satisfy

$$s_{i,k+1} = s_{i,k} + \eta_k \sigma(s_{i,k} - m_i), \qquad c_{i,k} = \sigma(m_i - s_{i,k}), \qquad s_{i,0} = 0. \tag{C.16}$$

We retain the conditions of the preceding proposition.

A common bound for the three criteria. We first choose a local displacement radius $R = \sqrt{\varepsilon}\gamma/2$. For $|s| \leq R$, the signal $b_i(s)$ is at least $\beta - R > 0$, and the independent-noise standard deviation is at least $\sqrt{\varepsilon}\gamma - R > 0$. These bounds follow from unit source norm and $0 < \gamma \leq \beta$. Consequently each scalar function $\varphi_i^f(s) = f(\theta_0 - s a_i)$ is twice continuously differentiable throughout this interval. Let M_i^f bound its absolute second derivative there, and define the remainder coefficient

$$C_i^f = \frac{|g_i^f|}{8} + \frac{M_i^f}{2}, \qquad M_i^f \geq \max_{|s| \leq R} |(\varphi_i^f)''(s)|. \tag{C.17}$$

For predictive KL, $M_i^Q = 1/4$ is valid because $\sigma' \leq 1/4$. For retain CE, one can use $M_i^{L_R} = [1 - (1 - \varepsilon) r_i^2/\gamma^2]/4$, one quarter of the retain variance along a_i. For retain error, a finite bound is obtained from the smooth expression in equation C.10 on this same interval.

Each displacement increment lies between zero and η_j, so $0 \le s_{i,k} \le T_k$. Since $|\sigma'(z)| \le 1/4$, its difference from the initial-rate approximation obeys

$$|s_{i,k} - h_i T_k| \le \tfrac{1}{4} \sum_{j<k} \eta_j T_j \le T_k^2/8.$$

Taylor's theorem therefore gives, whenever $T_k \le R$,

$$\left| f(\theta_{i,k}) - f(\theta_0) + G_i^f T_k \right| \le C_i^f T_k^2, \qquad G_i^f = h_i g_i^f. \tag{C.18}$$

The linear term records the initial useful direction; the quadratic remainder bounds how much repeated updates can depart from it.

A sufficient interval. For initial improvement rates satisfying $G_L^f > \max\{0, G_H^f\}$, define a common cumulative-learning-rate bound

$$H(G) = \min\left\{ R, \min_{f \in \{Q, L_R, \mathcal{R}_{01}\}} \frac{G_L^f}{C_L^f}, \min_{f \in \{Q, L_R, \mathcal{R}_{01}\}} \frac{G_L^f - G_H^f}{C_L^f + C_H^f} \right\} > 0. \tag{C.19}$$

Here G denotes the collection of initial improvement rates. The first ratio ensures improvement over θ_0; the second ensures superiority to the H path. Subtracting the bounds in equation C.18 proves

$$f(\theta_{L,k}) < \min\{f(\theta_0), f(\theta_{H,k})\} \quad \text{whenever } 0 < T_k < H(G), \qquad f \in \{Q, L_R, \mathcal{R}_{01}\}. \tag{C.20}$$

For a constant learning rate, $T_k = k\eta$; choosing $\eta < H(G)/2$ includes at least two updates.

Expressing the interval in confidence. Along CE ascent, $s_{L,k} \ge h_L T_k$ because its multiplier is at least its initial value h_L. A sufficient lower confidence threshold is therefore

$$c_{\text{end}} = \max\{c_L^R, \sigma(m_L - h_L H(G))\} < c_L^0. \tag{C.21}$$

If $c_{\text{end}} < c_{L,k} < c_L^0$, then $s_{L,k} < h_L H(G)$ and hence $T_k < H(G)$, proving the preceding comparisons. This is a sufficient interval of superiority to the initial model and the H path. It does not assert improvement over the preceding iterate, or require the interval to reach c_L^R.

C.5 Learning alternative labels

In the binary model, let $\zeta \in [0, 1/2]$ be the training target's probability of the observed label; the other label has probability $1 - \zeta$. The resulting cross-entropy is

$$\ell_\zeta(\theta; a_i) = \zeta \log(1 + e^{-a_i^\top \theta}) + (1 - \zeta) \log(1 + e^{a_i^\top \theta}). \tag{C.22}$$

Thus $\zeta = 0$ trains on the opposite label and $\zeta = 1/2$ uses a uniform target. We use the source conditions in equation C.14, so $c_i^0 > 1/2$. The initial gradient multiplier is $k_i = c_i^0 - \zeta > 0$. One gradient-descent step is

$$\theta_{i,1} = \theta_0 - \eta k_i a_i. \tag{C.23}$$

Alternatively, sample a label and let $B_i = 1$ indicate the observed label, with $B_i \sim \text{Bernoulli}(\zeta)$. The actual sampled update is $\theta_{i,1} = \theta_0 - \eta(c_i^0 - B_i)a_i$, whose multiplier has mean k_i.

Let $\xi = (c_H^0 - c_L^0)/(c_L^0 - \zeta)$ denote the relative increase in H's initial multiplier, so $k_H/k_L = 1 + \xi$.

Proposition C.2 (Source comparison under relabeling)**.** *Assume equation C.14 and $0 < \varepsilon \le 1$. Recall that $\delta = m_H - m_L$ and $\Delta = t_H - t_L$ compare the two sources' initial and retrained margins. If*

$$\Delta > \delta + \xi r_H, \tag{C.24}$$

then, for sufficiently small $\eta > 0$, both fixed-target and sampled-label updates satisfy

$$\mathbb{E} f(\theta_{L,1}) < \min\{f(\theta_0), \mathbb{E} f(\theta_{H,1})\}, \qquad f \in \{Q, L_R, \mathcal{R}_{01}\}.$$

Expectations concern the sampled label and are omitted for fixed targets.

Proof. Since $r_L = r_H + \Delta - \delta$, the gap condition gives

$$k_L r_L - k_H r_H = k_L(\Delta - \delta - \xi r_H) > 0, \qquad k_H t_H - k_L t_L = k_L(\Delta + \xi t_H) > 0.$$

For the second inequality, $t_H \geq 0$ is immediate; otherwise use $r_H = m_H - t_H > -t_H$ in the gap condition. The initial improvement per unit learning rate is $G_i^f = k_i g_i^f$. Each directional rate in equation C.11 is a positive multiple of r_i minus a positive multiple of t_i. The two strict inequalities above therefore give $G_L^f > G_H^f$ for all three criteria. Also $r_L > 0$ and $t_L \leq 0$ imply $G_L^f > 0$. Taylor expansion along each source gives $\mathbb{E}f(\theta_{i,1}) = f(\theta_0) - \eta G_i^f + O(\eta^2)$ for both update rules. For sampled labels, this expansion averages the two actual updated models. Both desired differences have a strictly negative linear term, proving the result. □

The gap condition accounts for relabeling's larger multiplier on the higher-confidence source. At matched initial confidence $c_L^0 = c_H^0 =: c^0$, we have $k_L = k_H$ and $\delta = \xi = 0$, so $\Delta > 0$ suffices. In this case, radial cancellation also gives the exact first-update identity

$$\mathbb{E}[Q(\theta_{L,1}) - Q(\theta_{H,1})] = -\eta(c^0 - \zeta)A_\beta \Delta < 0$$

for any $\eta > 0$, for either fixed or sampled targets. The retain comparisons above require a sufficiently small step.

C.6 Repeated binary relabeling

We now repeatedly update each copy using its fixed source. At step j, draw $B_{i,j} \sim \mathrm{Bernoulli}(\zeta)$ independently of past updates. With a common, predetermined positive learning-rate schedule η_j, the actual displacement and model are

$$\begin{aligned} s_{i,j+1} &= s_{i,j} + \eta_j[\sigma(m_i - s_{i,j}) - B_{i,j}], \qquad s_{i,0} = 0, \\ \theta_{i,j} &= \theta_0 - s_{i,j} a_i. \end{aligned} \tag{C.25}$$

For fixed-target training, replace every $B_{i,j}$ by ζ. As before, $T_k = \sum_{j<k} \eta_j$ is the cumulative learning rate. Use the initial gains and curvature bounds from Appendix C.4, now setting $G_i^f = k_i g_i^f$, and evaluate $H(G)$ in equation C.19 with these gains.

Proposition C.3 (A common relabeling horizon)**.** *Under the assumptions and gap condition of Proposition C.2, every $k \geq 1$ with $0 < T_k < H(G)$ satisfies*

$$\mathbb{E}f(\theta_{L,k}) < \min\{f(\theta_0), \mathbb{E}f(\theta_{H,k})\}, \qquad f \in \{Q, L_R, \mathcal{R}_{01}\}.$$

For fixed targets, the same comparison holds without expectations.

Proof. Each displacement increment has absolute value at most η_j, hence $|s_{i,j}| \leq T_j$. The sigmoid has derivative at most $1/4$, so

$$|\mathbb{E}s_{i,k} - k_i T_k| \leq \frac{1}{4} \sum_{j<k} \eta_j \mathbb{E}|s_{i,j}| \leq T_k^2/8.$$

For $T_k \leq R$, the ray curvature bound and $\mathbb{E}s_{i,k}^2 \leq T_k^2$ yield

$$\left|\mathbb{E}f(\theta_{i,k}) - f(\theta_0) + G_i^f T_k\right| \leq C_i^f T_k^2, \qquad C_i^f = |g_i^f|/8 + M_i^f/2.$$

The same bounds hold for the deterministic recurrence. Equation C.19 makes the remainder smaller than each strict gain advantage, proving both comparisons. □

The sampled result compares expected criteria of the updated models at fixed update counts satisfying the bound. It does not order every realization of the sampled labels.

Uniform targets with multiple classes. For a logit vector z, let $p = \mathrm{softmax}(z)$ be the class probabilities. If the target T is sampled uniformly among K classes, write $\pi_{\mathrm{unif}} = K^{-1}\mathbf{1}$ for that target distribution. Then $\mathbb{E}_T[-\log p_T] = \log K + \mathrm{KL}(\pi_{\mathrm{unif}} \| p)$ and its logit gradient is $p - \pi_{\mathrm{unif}}$, so the expected loss favors uniform predictions. Sampled multiclass gradients need not share a single direction, as the binary source updates above do.

C.7 Realizability of the model pair

We finish by constructing a training problem whose initial and retain-only objectives have optima θ_0 and θ_R. Assume the feature space has at least three dimensions, and choose a unit vector w orthogonal to u, v. Take feasible sources with $t_L < 0 < t_H$ and $r_L, r_H > 0$, and define

$$a_i^{\pm} = (r_i/\gamma)u + (t_i/\beta)v \pm \sqrt{d_i}\, w.$$

For example, $\beta = \gamma$, $m_L = m_H = \gamma/2$, $t_L = -\gamma/4$, and $t_H = \gamma/4$ give $d_L = 3/8$ and $d_H = 7/8$, so such unit sources exist. A forget distribution chooses L with probability π and H otherwise, then chooses either sign with equal probability. Independently draw Y uniformly from $\{-1, +1\}$ and set $X = Y a_i^{\pm}$. Let L_F be cross-entropy averaged over this forget distribution.

Let $\omega \in (0, 1)$ be the forget mixture weight in the full-data objective. To balance the gradient along v, set

$$b = \frac{1-\omega}{\omega}(A_\beta - A_\varepsilon)\beta^2, \qquad \pi = \frac{h_H t_H + b}{h_H t_H - h_L t_L}.$$

To balance the gradient along u, use the regularization coefficient

$$\lambda = \frac{\omega[\pi h_L r_L + (1-\pi) h_H r_H]}{\gamma^2} - (1-\omega)\varepsilon A_\varepsilon.$$

Choosing ω sufficiently close to one gives $0 < b < -h_L t_L$, $0 < \pi < 1$, and $\lambda > 0$. The full-data and retain-only objectives, with the same regularizer, are

$$\begin{aligned} \mathcal{J}_0(\theta) &= (1-\omega)L_R(\theta) + \omega L_F(\theta) + \tfrac{\lambda}{2}(u^\top \theta)^2, \\ \mathcal{J}_R(\theta) &= L_R(\theta) + \tfrac{\lambda}{2}(u^\top \theta)^2. \end{aligned} \tag{C.26}$$

The retain distribution is generated by θ_R and $u^\top \theta_R = 0$, so $\nabla \mathcal{J}_R(\theta_R) = 0$. For the full-data objective, symmetry cancels the w gradient. The definitions of π and λ cancel the v and u gradients, respectively, giving $\nabla \mathcal{J}_0(\theta_0) = 0$. Since $\varepsilon > 0$, retain CE is strictly convex, so the two stated parameters are the unique optima of their respective objectives. This construction establishes that the model pair is realizable; it does not prescribe the deletion fraction used in the experiments.

D Retrain Response and Prediction Distributions

We analyze the CIFAR-10/ResNet-18 runs corresponding to the two deletion settings in Table 1. The runs share the initial model θ_0 and use a model θ_R retrained for each deletion request. Unless otherwise stated, statistics are computed within each request before averaging across requests.

D.1 Initial confidence and retraining responses

For each forget example, let $c_i^0 = p_{\theta_0}(y_i \mid x_i)$ and $c_i^R = p_{\theta_R}(y_i \mid x_i)$ denote its true-label confidence before unlearning and under the retrained model, respectively. The retrained predictions are used only for analysis.

Initial selection. Table 16 compares the complete forget set with $A_0 \cap D_F$, the forget examples actually selected by the pooled threshold in Section 3. The initial selected forget examples have lower mean confidence under θ_R, and the two models' true-label confidences are positively correlated in both deletion settings. These measurements support initial confidence as an empirical proxy for identifying low-c_i^R sources.

Concentration of confidence decrease. We measure how much of the positive confidence decrease after retraining is concentrated among forget examples with low initial confidence:

$$w_i = \max(c_i^0 - c_i^R, 0), \qquad C(q) = \frac{\sum_{i \in S_q} w_i}{\sum_{i \in D_F} w_i},$$

where S_q contains the lowest-confidence fraction q of the forget set, ranked by c_i^0. We sort the saved unrounded probabilities, breaking exact ties by ascending sample ID. Here q is a within-forget fraction, distinct from the pooled calibration level ρ. The curve is normalized within each

Table 16: Initial true-label confidence and confidence under θ_R on CIFAR-10.

Setting	Selected forget n	Whole c^R	Selected c^R	Spearman
Random	$255.2_{\pm 17.7}$	$0.9431_{\pm 0.0033}$	$0.7473_{\pm 0.0252}$	$0.4652_{\pm 0.0142}$
Class-concentrated	$256.6_{\pm 13.4}$	$0.9274_{\pm 0.0034}$	$0.7091_{\pm 0.0245}$	$0.4282_{\pm 0.0150}$

Whole and selected columns report mean true-label confidence under θ_R. Spearman is the correlation between c_i^0 and c_i^R over the complete forget set.

Table 17: Concentration of positive true-label confidence decreases within the forget set.

Setting	$C(0.05)$	$C(0.10)$	$C(0.20)$	Mean w_i
Random	$0.2198_{\pm 0.0213}$	$0.3403_{\pm 0.0213}$	$0.5004_{\pm 0.0186}$	$0.0561_{\pm 0.0032}$
Class-concentrated	$0.1996_{\pm 0.0136}$	$0.3146_{\pm 0.0154}$	$0.4771_{\pm 0.0108}$	$0.0718_{\pm 0.0034}$

Mean w_i includes examples with zero decrease.

request before averaging; $C(q) = q$ is the expected share under uniformly random selection of the same number of examples. Table 17 reports the captured shares at fixed fractions and the mean positive decrease over the entire forget set. In both settings, the lowest-confidence fifth accounts for approximately half of the total positive decrease.

D.2 Selection transitions

We examine random forgetting using the recorded selection sets at the starts of epochs $e = 0, \ldots, 9$. The analysis contains nine adjacent transitions, from $0 \to 1$ through $8 \to 9$. No refresh is performed after the final training epoch. Let C denote unselected examples and T selected examples. The four groups are continued exclusion ($C \to C$), entry ($C \to T$), exit ($T \to C$), and continued selection ($T \to T$).

Confidence of entrants and exits. For DSS, we compare entrants with examples remaining unselected and exits with examples remaining selected. Each group's response is measured by its fixed confidence c_i^R under the request's retrained model.

Conditioning on initial confidence. To assess whether these differences persist among examples with similar initial scores, we divide each request's forget set into 20 quantile bins of c_i^0, keeping tied scores together. At each refresh, only bins containing both comparison groups contribute. The two groups receive the same bin weights: the shares of the starting unselected population for entry, or the starting selected population for exit, normalized over the included bins. We average the within-bin confidence differences using these weights, then average over refreshes with at least one eligible bin within each request.

Table 19 reports the resulting differences. NegGrad+, NPO, RL, SalUn, and AMUN have lower-confidence entrants and higher-confidence exits in all ten requests, with smaller entry differences for RL and SalUn. UGradSL+ shows a less consistent pattern: the entry difference is negative in six requests, whereas the exit difference is positive in four and negative on average. The unstratified comparisons also give a negative mean exit difference for UGradSL+. Thus, the separation depends on the base method.

Selection-transition counts. Table 20 reports the sizes of all four groups for retain and forget examples under DSS and DSS+. DSS allows both entry and exit. DSS+ retains every previously selected example, so its $T \to C$ count is zero by construction. Selection membership describes eligibility for training. The base method determines subsequent minibatch use.

D.3 Gradient magnitudes under source selection

We compare NegGrad+ and NPO on CIFAR-10 with ResNet-18 under random forgetting. Before each training epoch, we measure gradients under each policy's current model on four draws of 128

Table 18: True-label confidence under θ_R by forget-set transition for DSS under random forgetting.

Method	$C \to C$	$C \to T$	$T \to C$	$T \to T$
NegGrad+	$0.9765_{\pm 0.0031}$	$0.8793_{\pm 0.0179}$	$0.8744_{\pm 0.0144}$	$0.7456_{\pm 0.0294}$
NPO	$0.9765_{\pm 0.0033}$	$0.8789_{\pm 0.0127}$	$0.8805_{\pm 0.0092}$	$0.7463_{\pm 0.0266}$
UGradSL+	$0.9562_{\pm 0.0305}$	$0.9492_{\pm 0.0113}$	$0.8703_{\pm 0.0398}$	$0.8809_{\pm 0.0146}$
RL	$0.9944_{\pm 0.0018}$	$0.9791_{\pm 0.0044}$	$0.9689_{\pm 0.0175}$	$0.8754_{\pm 0.0094}$
SalUn	$0.9943_{\pm 0.0020}$	$0.9793_{\pm 0.0039}$	$0.9707_{\pm 0.0168}$	$0.8728_{\pm 0.0079}$
AMUN	$0.9821_{\pm 0.0019}$	$0.8947_{\pm 0.0108}$	$0.8988_{\pm 0.0146}$	$0.7964_{\pm 0.0102}$

For each entry or exit comparison, both group means are averaged over the same refreshes in which both groups are nonempty.

Table 19: Forget-set transition differences after stratifying by initial confidence for DSS under random forgetting.

Method	Entry Δ	Exit Δ
NegGrad+	$-0.0831_{\pm 0.0197}$	$0.1162_{\pm 0.0307}$
NPO	$-0.0812_{\pm 0.0101}$	$0.1182_{\pm 0.0221}$
UGradSL+	$-0.0051_{\pm 0.0250}$	$-0.0179_{\pm 0.0463}$
RL	$-0.0095_{\pm 0.0033}$	$0.0835_{\pm 0.0234}$
SalUn	$-0.0099_{\pm 0.0030}$	$0.0904_{\pm 0.0188}$
AMUN	$-0.0786_{\pm 0.0119}$	$0.0842_{\pm 0.0202}$

Entry Δ compares $(C \to T) - (C \to C)$. Exit Δ compares $(T \to C) - (T \to T)$.

retain and 128 forget examples from its selected pools. We use clean inputs in evaluation mode and average the four measurements without updating the model. Let g_R and g_F denote the gradients of the weighted retain and forget objectives, including the sign of the forget objective and any selection weights. We report their norms, the total norm $\|g_R + g_F\|_2$, and their cosine similarity. Negative cosine indicates opposing objective gradients.

Table 21 averages the measurements before each epoch within each request. Random and Fixed initial scores use the growing subsets described in the selection-rule comparison (Figure 5).

For both methods, DSS has a larger average total gradient norm than FULL, with DSS+ between them. The retain component accounts for most of the increase, and the more negative cosine under DSS indicates stronger opposition between the weighted retain and forget gradients. EL2N and GraNd also produce larger average gradient norms than FULL, while Random and Fixed initial scores have smaller norms than DSS.

D.4 Low/high selection: gradient norms and training outcomes

To examine the role of low-confidence examples in source selection, we independently select low- or high-confidence examples from the retain and forget sets, yielding Low/Low, Low/High, High/Low, and High/High. This diagnostic uses the same Random 10% deletion request within each method. At each epoch, the four variants use the same numbers of retain and forget examples as Low/Low; High selects the corresponding number of examples from the high-confidence end of the current ranking. Thus, the four variants have the same optimizer-step count within each method and use the same learning rate. FULL is included as a reference and is not step-matched to the four selection variants.

Initial norms use the gradient of the mean loss over each complete selected set before training, evaluated on clean inputs in evaluation mode with the base objective coefficients. Mean training norms average the actual total-gradient norms over optimizer steps.

D.5 Prediction distributions

We compare the predicted label distributions of FULL, DSS, and DSS+ with those of θ_R, separately on retain and forget examples. We use the endpoint checkpoints from Table 1. For two probability

Table 20: Mean numbers of examples in each selection transition under random forgetting, averaged over refreshes within each request.

Method	Policy	$C \to C$	$C \to T$	$T \to C$	$T \to T$
Retain examples					
NegGrad+	DSS	34,005.3	4,115.5	3,586.7	3,292.5
NegGrad+	DSS+	33,588.8	1,569.0	0.0	9,842.2
NPO	DSS	33,792.9	4,213.9	3,659.8	3,333.3
NPO	DSS+	33,834.4	1,461.8	0.0	9,703.9
UGradSL+	DSS	14,692.7	5,053.9	323.8	24,929.7
UGradSL+	DSS+	13,626.4	4,747.1	0.0	26,626.5
RL	DSS	14,660.9	6,952.5	3,733.4	19,653.2
RL	DSS+	6,103.7	4,382.4	0.0	34,513.8
SalUn	DSS	14,678.6	7,211.9	4,295.2	18,814.4
SalUn	DSS+	5,291.4	4,428.2	0.0	35,280.5
AMUN	DSS	33,609.2	3,900.2	3,295.7	4,194.9
AMUN	DSS+	30,016.0	1,973.8	0.0	13,010.3
Forget examples					
NegGrad+	DSS	3,715.7	443.4	364.2	476.8
NegGrad+	DSS+	3,716.1	175.6	0.0	1,108.2
NPO	DSS	3,701.3	449.1	368.9	480.6
NPO	DSS+	3,745.4	164.7	0.0	1,089.9
UGradSL+	DSS	1,623.0	556.1	31.2	2,789.8
UGradSL+	DSS+	1,506.1	526.9	0.0	2,967.0
RL	DSS	1,585.9	740.1	372.6	2,301.4
RL	DSS+	669.2	486.4	0.0	3,844.4
SalUn	DSS	1,582.9	764.9	425.3	2,226.9
SalUn	DSS+	587.3	491.2	0.0	3,921.5
AMUN	DSS	3,553.0	385.8	266.2	795.0
AMUN	DSS+	3,324.5	220.9	0.0	1,454.6

All transitions contribute, including those with empty groups. These are counts per refresh, not numbers of distinct examples across training.

Table 21: Gradient diagnostics for NegGrad+ and NPO under random forgetting.

Method	Selection	$\|g_R + g_F\|_2$	$\|g_R\|_2$	$\|g_F\|_2$	$\cos(g_R, g_F)$
NegGrad+	FULL	$1.062_{\pm 0.304}$	$0.967_{\pm 0.314}$	$0.305_{\pm 0.059}$	$-0.125_{\pm 0.027}$
	DSS	$8.487_{\pm 1.645}$	$8.900_{\pm 1.838}$	$1.330_{\pm 0.176}$	$-0.336_{\pm 0.056}$
	DSS+	$4.243_{\pm 0.354}$	$4.208_{\pm 0.359}$	$0.959_{\pm 0.100}$	$-0.178_{\pm 0.033}$
	EL2N	$8.385_{\pm 1.561}$	$8.811_{\pm 1.740}$	$1.288_{\pm 0.183}$	$-0.351_{\pm 0.060}$
	GraNd	$7.544_{\pm 1.017}$	$7.933_{\pm 1.149}$	$1.196_{\pm 0.164}$	$-0.370_{\pm 0.052}$
	InfoBatch	$3.360_{\pm 0.561}$	$3.318_{\pm 0.557}$	$0.587_{\pm 0.065}$	$-0.127_{\pm 0.023}$
	Random	$0.548_{\pm 0.250}$	$0.460_{\pm 0.226}$	$0.201_{\pm 0.080}$	$-0.148_{\pm 0.032}$
	Fixed initial scores	$1.611_{\pm 0.499}$	$1.478_{\pm 0.485}$	$0.475_{\pm 0.118}$	$-0.098_{\pm 0.022}$
NPO	FULL	$0.907_{\pm 0.163}$	$0.838_{\pm 0.168}$	$0.235_{\pm 0.024}$	$-0.116_{\pm 0.022}$
	DSS	$8.581_{\pm 1.253}$	$9.034_{\pm 1.383}$	$1.123_{\pm 0.122}$	$-0.374_{\pm 0.040}$
	DSS+	$3.369_{\pm 0.530}$	$3.352_{\pm 0.587}$	$0.664_{\pm 0.045}$	$-0.173_{\pm 0.034}$
	EL2N	$7.895_{\pm 1.317}$	$8.266_{\pm 1.451}$	$1.043_{\pm 0.099}$	$-0.361_{\pm 0.040}$
	GraNd	$6.978_{\pm 0.607}$	$7.292_{\pm 0.707}$	$0.984_{\pm 0.091}$	$-0.353_{\pm 0.050}$
	InfoBatch	$3.403_{\pm 0.536}$	$3.394_{\pm 0.552}$	$0.513_{\pm 0.056}$	$-0.143_{\pm 0.026}$
	Random	$0.471_{\pm 0.123}$	$0.417_{\pm 0.114}$	$0.147_{\pm 0.039}$	$-0.159_{\pm 0.030}$
	Fixed initial scores	$1.322_{\pm 0.421}$	$1.241_{\pm 0.420}$	$0.334_{\pm 0.086}$	$-0.106_{\pm 0.024}$

vectors p and r, we compute

$$\mathrm{JS}(p, r) = \tfrac{1}{2}\,\mathrm{KL}(p\|m) + \tfrac{1}{2}\,\mathrm{KL}(r\|m), \qquad m = \tfrac{1}{2}(p + r),$$

using natural logarithms. For each request and split, we compute the mean, median, and 95th percentile of the per-example divergences. Tables 23 and 24 report averages of these request-specific statistics.

Under random forgetting, UGradSL+ and SalUn with DSS have lower average retain and forget JS than FULL. RL with DSS has lower retain JS but higher forget JS. NegGrad+ and NPO with DSS+ have higher average forget JS than FULL. Under class-concentrated forgetting, FULL has lower average JS on both splits for NegGrad+, NPO, UGradSL+, and RL. For the first three methods, DSS+ reduces both divergences relative to DSS (Table 24). The mean and upper-tail summaries need not rank policies in the same order.

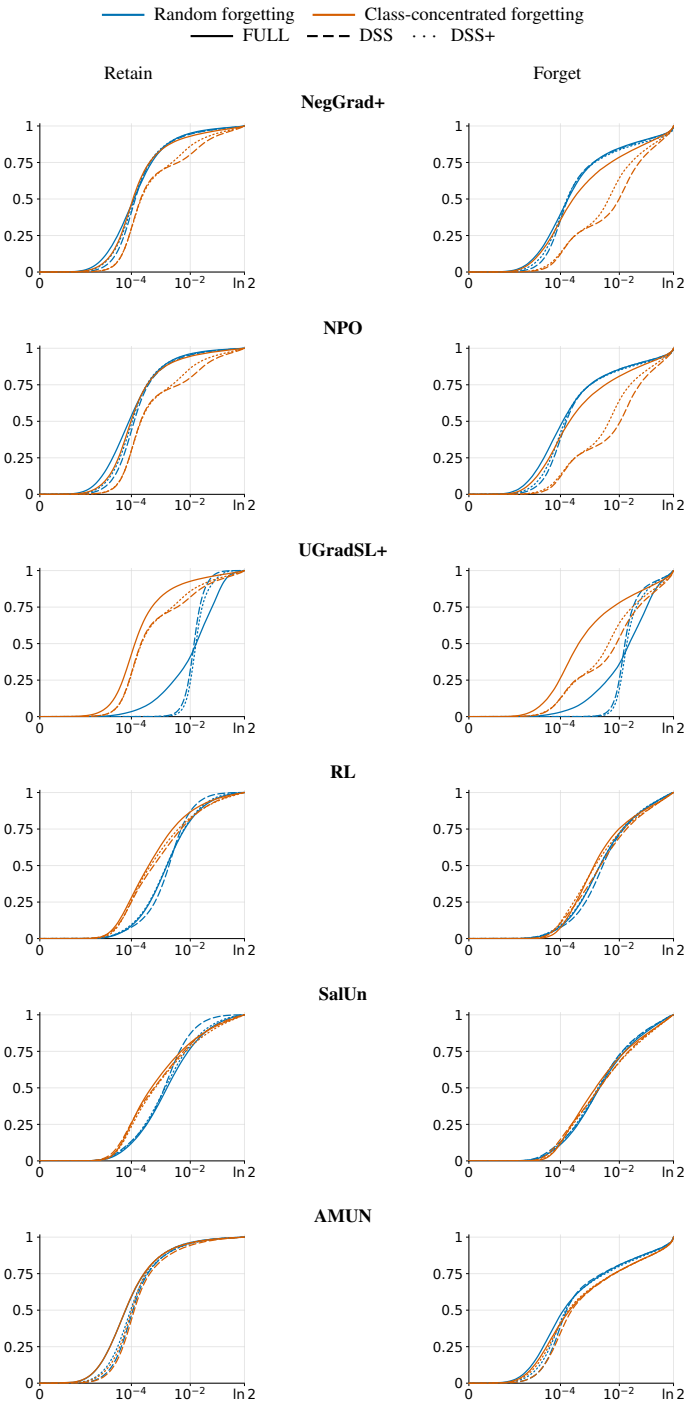


Figure 11: Prediction distributions relative to the retrained model on CIFAR-10 with ResNet-18. Columns show retain and forget examples. Rows show the base methods. The x-axis shows per-example JS divergence in nats, and the y-axis shows its CDF. Curves average the per-request CDFs.

Table 22: Low/high confidence selection under random forgetting.

Method	Retain selection	Forget selection	Initial gradient norm		Mean total gradient norm	Accuracy (%)			Steps
			Retain	Forget		R	F	Test	
NegGrad+	All	All	0.0387	0.00420	0.832	99.29	95.00	92.69	1760
	Low	Low	0.401	0.0442	2.47	98.31	94.88	92.06	294
	Low	High	0.401	0.000284	2.11	11.70	11.62	11.79	294
	High	Low	0.00249	0.0442	6.38	10.37	9.96	10.48	294
	High	High	0.00249	0.000284	0.934	48.66	48.64	47.03	294
NPO	All	All	0.0387	0.00420	0.788	99.39	94.90	93.19	1760
	Low	Low	0.401	0.0442	2.35	99.62	95.88	93.61	273
	Low	High	0.401	0.000284	2.52	43.85	44.42	42.89	273
	High	Low	0.00249	0.0442	1.25	28.66	29.02	28.67	273
	High	High	0.00249	0.000284	1.05	68.73	69.68	65.93	273
UGradSL+	All	All	0.0387	1.65	1.67	99.42	95.92	93.00	1760
	Low	Low	0.401	1.73	0.791	99.96	97.20	93.98	1052
	Low	High	0.401	2.12	1.17	99.99	99.76	94.73	1052
	High	Low	0.00249	1.73	2.40	99.15	96.20	92.48	1052
	High	High	0.00249	2.12	0.658	99.99	99.58	94.79	1052
RL	All	All	0.0430	41.1	1.58	99.21	95.94	93.29	1960
	Low	Low	0.445	42.0	1.91	99.74	97.40	93.80	1033
	Low	High	0.445	53.0	1.82	99.81	99.28	94.13	1033
	High	Low	0.00277	42.0	1.09	85.13	82.82	82.07	1033
	High	High	0.00277	53.0	1.06	78.76	78.22	75.03	1033
SalUn	All	All	0.0423	40.4	1.52	97.97	94.98	92.22	1960
	Low	Low	0.433	41.2	1.56	99.88	97.30	94.50	1014
	Low	High	0.433	51.2	1.60	99.70	98.90	94.24	1014
	High	Low	0.00268	41.2	0.963	84.30	81.96	80.38	1014
	High	High	0.00268	51.2	0.895	83.72	82.84	79.88	1014
AMUN	All	All	0.0352	1.06	1.70	99.63	94.90	93.74	2150
	Low	Low	0.364	1.21	2.93	99.53	94.48	93.31	382
	Low	High	0.364	1.47	2.65	99.59	97.76	93.21	382
	High	Low	0.00226	1.21	1.93	97.98	94.76	92.37	382
	High	High	0.00226	1.47	1.31	96.69	95.42	90.43	382

All/All is FULL. Low/Low is DSS.

Table 23: Random forgetting: JS divergence from the retrained model. All values are scaled by 10^3.

Method	Policy	Split	Mean JS	Median JS	95th percentile JS
NegGrad+	FULL	Retain	$7.084_{\pm 2.387}$	$0.122_{\pm 0.017}$	$10.843_{\pm 5.177}$
		Forget	$40.701_{\pm 4.474}$	$0.182_{\pm 0.028}$	$390.365_{\pm 60.882}$
	DSS	Retain	$8.133_{\pm 2.541}$	$0.136_{\pm 0.013}$	$14.656_{\pm 8.568}$
		Forget	$41.332_{\pm 2.942}$	$0.178_{\pm 0.021}$	$394.159_{\pm 31.342}$
	DSS+	Retain	$8.631_{\pm 3.009}$	$0.123_{\pm 0.010}$	$13.532_{\pm 10.515}$
		Forget	$49.763_{\pm 3.512}$	$0.167_{\pm 0.017}$	$516.050_{\pm 34.408}$
NPO	FULL	Retain	$4.449_{\pm 0.966}$	$0.084_{\pm 0.014}$	$6.114_{\pm 1.660}$
		Forget	$37.850_{\pm 3.187}$	$0.122_{\pm 0.022}$	$349.823_{\pm 44.091}$
	DSS	Retain	$5.454_{\pm 1.722}$	$0.124_{\pm 0.007}$	$9.024_{\pm 2.884}$
		Forget	$37.201_{\pm 2.857}$	$0.158_{\pm 0.011}$	$336.228_{\pm 46.607}$
	DSS+	Retain	$5.855_{\pm 1.724}$	$0.107_{\pm 0.009}$	$7.390_{\pm 3.261}$
		Forget	$43.136_{\pm 3.247}$	$0.140_{\pm 0.014}$	$428.449_{\pm 47.547}$
UGradSL+	FULL	Retain	$44.365_{\pm 30.240}$	$25.522_{\pm 30.061}$	$147.162_{\pm 83.152}$
		Forget	$75.056_{\pm 29.130}$	$30.599_{\pm 34.419}$	$379.298_{\pm 59.663}$
	DSS	Retain	$18.730_{\pm 9.518}$	$14.942_{\pm 7.787}$	$43.154_{\pm 22.788}$
		Forget	$51.598_{\pm 8.468}$	$16.998_{\pm 8.630}$	$305.246_{\pm 12.233}$
	DSS+	Retain	$25.114_{\pm 20.144}$	$19.974_{\pm 16.869}$	$58.855_{\pm 47.445}$
		Forget	$56.943_{\pm 17.929}$	$22.745_{\pm 18.945}$	$306.648_{\pm 21.890}$
RL	FULL	Retain	$13.843_{\pm 2.201}$	$1.606_{\pm 0.342}$	$63.262_{\pm 12.342}$
		Forget	$34.034_{\pm 2.150}$	$2.190_{\pm 0.435}$	$224.974_{\pm 15.105}$
	DSS	Retain	$6.486_{\pm 1.064}$	$1.874_{\pm 0.219}$	$24.884_{\pm 5.152}$
		Forget	$35.698_{\pm 1.262}$	$2.721_{\pm 0.233}$	$233.154_{\pm 12.265}$
	DSS+	Retain	$11.573_{\pm 1.618}$	$1.543_{\pm 0.233}$	$51.144_{\pm 9.042}$
		Forget	$33.025_{\pm 1.927}$	$2.163_{\pm 0.305}$	$213.306_{\pm 19.709}$
SalUn	FULL	Retain	$20.350_{\pm 3.500}$	$1.636_{\pm 0.303}$	$106.187_{\pm 24.229}$
		Forget	$37.049_{\pm 2.373}$	$2.157_{\pm 0.379}$	$246.563_{\pm 15.852}$
	DSS	Retain	$6.792_{\pm 1.226}$	$1.307_{\pm 0.257}$	$27.772_{\pm 6.149}$
		Forget	$34.530_{\pm 3.197}$	$2.013_{\pm 0.322}$	$229.257_{\pm 33.686}$
	DSS+	Retain	$16.693_{\pm 2.249}$	$1.406_{\pm 0.323}$	$81.693_{\pm 14.417}$
		Forget	$34.912_{\pm 2.543}$	$1.898_{\pm 0.426}$	$232.694_{\pm 18.143}$
AMUN	FULL	Retain	$3.821_{\pm 0.820}$	$0.063_{\pm 0.005}$	$5.911_{\pm 1.962}$
		Forget	$47.746_{\pm 2.710}$	$0.127_{\pm 0.011}$	$446.525_{\pm 29.402}$
	DSS	Retain	$4.049_{\pm 1.280}$	$0.112_{\pm 0.014}$	$7.752_{\pm 2.914}$
		Forget	$46.549_{\pm 1.892}$	$0.163_{\pm 0.021}$	$427.586_{\pm 25.179}$
	DSS+	Retain	$4.486_{\pm 2.098}$	$0.096_{\pm 0.019}$	$9.690_{\pm 8.163}$
		Forget	$49.064_{\pm 4.695}$	$0.155_{\pm 0.037}$	$434.930_{\pm 37.926}$

Table 24: Class-concentrated forgetting: JS divergence from the retrained model. All values are scaled by 10^3.

Method	Policy	Split	Mean JS	Median JS	95th percentile JS
NegGrad+	FULL	Retain	$11.839_{\pm 0.909}$	$0.113_{\pm 0.010}$	$26.674_{\pm 4.205}$
		Forget	$46.607_{\pm 2.597}$	$0.274_{\pm 0.039}$	$398.847_{\pm 28.120}$
	DSS	Retain	$24.631_{\pm 1.089}$	$0.205_{\pm 0.016}$	$135.584_{\pm 9.699}$
		Forget	$73.712_{\pm 1.704}$	$9.918_{\pm 1.121}$	$486.871_{\pm 10.651}$
	DSS+	Retain	$18.593_{\pm 1.798}$	$0.215_{\pm 0.017}$	$82.751_{\pm 13.287}$
		Forget	$60.996_{\pm 3.680}$	$4.401_{\pm 0.670}$	$452.397_{\pm 22.268}$
NPO	FULL	Retain	$7.695_{\pm 0.718}$	$0.097_{\pm 0.009}$	$12.828_{\pm 1.899}$
		Forget	$41.727_{\pm 2.205}$	$0.204_{\pm 0.030}$	$358.192_{\pm 22.520}$
	DSS	Retain	$23.973_{\pm 0.716}$	$0.205_{\pm 0.017}$	$128.000_{\pm 6.202}$
		Forget	$73.150_{\pm 1.937}$	$10.904_{\pm 1.512}$	$480.512_{\pm 16.571}$
	DSS+	Retain	$16.961_{\pm 1.240}$	$0.212_{\pm 0.017}$	$72.117_{\pm 9.764}$
		Forget	$58.627_{\pm 2.253}$	$4.535_{\pm 0.658}$	$436.779_{\pm 17.889}$
UGradSL+	FULL	Retain	$12.229_{\pm 0.974}$	$0.131_{\pm 0.012}$	$29.806_{\pm 4.363}$
		Forget	$47.101_{\pm 2.491}$	$0.415_{\pm 0.044}$	$401.606_{\pm 23.368}$
	DSS	Retain	$25.683_{\pm 2.253}$	$0.211_{\pm 0.018}$	$149.394_{\pm 22.615}$
		Forget	$74.511_{\pm 4.211}$	$8.025_{\pm 1.021}$	$501.593_{\pm 16.543}$
	DSS+	Retain	$20.856_{\pm 1.203}$	$0.221_{\pm 0.020}$	$101.179_{\pm 10.928}$
		Forget	$64.070_{\pm 2.417}$	$4.297_{\pm 0.306}$	$470.626_{\pm 19.846}$
RL	FULL	Retain	$14.213_{\pm 1.513}$	$0.385_{\pm 0.038}$	$60.924_{\pm 11.211}$
		Forget	$39.534_{\pm 2.220}$	$1.415_{\pm 0.251}$	$294.401_{\pm 21.569}$
	DSS	Retain	$18.651_{\pm 8.088}$	$0.657_{\pm 0.377}$	$100.606_{\pm 60.542}$
		Forget	$42.876_{\pm 5.072}$	$2.496_{\pm 1.166}$	$294.184_{\pm 21.975}$
	DSS+	Retain	$20.163_{\pm 4.481}$	$0.459_{\pm 0.105}$	$109.646_{\pm 36.054}$
		Forget	$41.769_{\pm 3.448}$	$1.550_{\pm 0.222}$	$299.377_{\pm 20.932}$
SalUn	FULL	Retain	$24.906_{\pm 6.915}$	$0.490_{\pm 0.122}$	$153.075_{\pm 64.870}$
		Forget	$44.530_{\pm 4.163}$	$1.581_{\pm 0.383}$	$314.468_{\pm 33.874}$
	DSS	Retain	$21.999_{\pm 5.042}$	$0.607_{\pm 0.136}$	$127.295_{\pm 42.794}$
		Forget	$45.202_{\pm 6.426}$	$2.232_{\pm 0.827}$	$303.172_{\pm 33.291}$
	DSS+	Retain	$28.374_{\pm 7.650}$	$0.717_{\pm 0.413}$	$180.415_{\pm 64.609}$
		Forget	$47.791_{\pm 8.919}$	$2.460_{\pm 1.819}$	$318.900_{\pm 36.344}$
AMUN	FULL	Retain	$4.684_{\pm 1.771}$	$0.062_{\pm 0.005}$	$7.204_{\pm 4.481}$
		Forget	$57.670_{\pm 3.667}$	$0.211_{\pm 0.036}$	$504.127_{\pm 24.444}$
	DSS	Retain	$5.987_{\pm 2.988}$	$0.128_{\pm 0.022}$	$15.102_{\pm 12.812}$
		Forget	$55.725_{\pm 5.899}$	$0.274_{\pm 0.118}$	$475.346_{\pm 34.590}$
	DSS+	Retain	$4.167_{\pm 1.435}$	$0.103_{\pm 0.007}$	$6.361_{\pm 2.622}$
		Forget	$57.304_{\pm 6.054}$	$0.198_{\pm 0.050}$	$493.396_{\pm 38.885}$