LOSS AND MARGIN POLICIES FOR CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Models that are updated as new tasks arrive must learn new capabilities while preserving those already in use. The loss applied to new-task data is a rarely examined control over this trade-off: cross-entropy (CE) keeps pushing examples that are already classified correctly, whereas a hinge loss stops once a prescribed logit gap is reached. We compare CE, a multiclass hinge, and the high error margin (HEM) loss for task-incremental learning under matched update budgets and replay opportunities, with validation-based selection and separate measurement of new-task acquisition and old-task retention. On Split-CIFAR-100 the selected hinge–HEM ranking reverses with the training budget: HEM leads at 10 epochs per task and hinge at 40, because quadrupling the budget costs the selected CE and HEM policies 3.6–3.8 points of old-task accuracy while hinge gains 1.0; the size of the longer-budget advantage depends on the acquisition requirement used for selection. A 2×2 ablation with margin and learning rate fixed reproduces the reversal within one loss family: HEM-style active-sample averaging helps at the short budget ($+0.55$ points) and hurts at the long one (-3.02). On Split-TinyImageNet, whose test split was held out until the confirmation run, both margin losses beat CE with replay by about five points through retention rather than acquisition, but hinge and HEM are not distinguished ($+0.13$ [-0.34, 0.59]). In an all-hinge study, geometrically discounting the margin across tasks improves final accuracy over a fixed margin by $+1.80$ points (5% discount, 128 seeds) and $+4.80$ (20%, 32 seeds), because forgetting falls by more than acquisition does. An elementary gradient analysis explains why margin, normalization, and learning rate are distinct controls. The results give budget-dependent operating points for the new-task loss and quantify the acquisition cost of the retention they buy.

1 INTRODUCTION

An image recognition service may gain a new task without being allowed to discard its earlier capabilities. A personalized model may encounter one user's task after another while sharing a common representation. In both settings, repeatedly retraining on the entire history can be costly or impossible, and an update that succeeds on the newest task can silently degrade functions that remain useful. Continual learning studies this problem of sequential acquisition and retention (De Lange et al., 2022; van de Ven et al., 2022). We focus on the task-incremental setting, where a task identifier selects an output head at inference. Private heads allow us to study interference through a shared representation while holding the inference task identity fixed. Extending the conclusions to recognition without task identity requires separate evidence.

A central choice is what to ask the network to do on its new data. Cross-entropy (CE) continues to reward greater confidence even after a sample is correctly classified. A hinge loss ceases its direct contribution once a prescribed logit gap is reached. This distinction makes the loss a plausible control over continued adaptation of a shared representation. Yet the question is more subtle than whether a margin loss "forgets less." A loss can retain previous tasks by learning the new task less well; additional training may improve acquisition and increase forgetting simultaneously; replay and normalization can change the comparison. Useful guidance must therefore account for the entire training policy and the route by which final accuracy improves.

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.

We ask: *under a specified update and memory budget, which loss policy gives a useful acquisition–retention trade-off?* We use two complementary designs. First, a shared CE-trained prefix fixes the learner's past before comparing CE, a standard multiclass hinge on the gap between the true-class logit and the largest other logit, and HEM (Spratling & Schütt, 2026), with matched later-task budgets and common replay rules. Second, an all-hinge experiment changes only the margin schedule after a shared first task, testing whether gradually reducing the demanded separation improves the final model. These designs answer different questions and their absolute accuracies are not directly comparable.

Our contributions are three empirical findings, supported by paired measurements and a scoped analytical interpretation:

1. **Loss rankings depend on the available budget, and normalization is one reason.** With the same replay opportunities on Split-CIFAR-100, validation-selected HEM leads at the shorter budget while hinge leads at the longer one. The reversal has a measurable retention component: the selected longer-budget CE and HEM policies lose old-task accuracy whereas hinge retains it, and hinge loses the least at every learning rate of the no-replay validation grid. A 2×2 ablation with margin and learning rate fixed isolates one cause: HEM-style active-sample averaging helps at the shorter budget and hurts at the longer one, so the reversal appears even within a single loss family. Matching budgets removes the confound of unequal training duration but does not produce a universal ranking.

2. **The transferable result is a margin-loss advantage over CE, not hinge dominance.** On Split-TinyImageNet, a second benchmark whose test split was held out until the confirmation run, both margin losses outperform CE under the tested longer-budget policy, while hinge and HEM with replay are not statistically distinguished. Acquisition and retention measurements show what the final scores conceal: the margin losses win by forgetting less, not by learning more.

3. **Margin discounting offers a measurable retention–acquisition trade-off.** Fixed-sample replications establish gains over a constant margin of 0.25 for mild and stronger discounts. Stronger discounting preserves earlier tasks better while reducing immediate new-task accuracy. This establishes a useful policy comparison; superiority to a tuned constant margin remains open.

The losses, replay principle, and checkpoint branching are established components. The contribution is the controlled evidence about their operating regimes, including where an apparent advantage disappears, and a quantitative account of the cost of retaining old skills. The central studies include raw result rows and analysis code that regenerates their tables and figures; historical pilot summaries are identified separately.

2 Related Work

Continual learning and the optimization policy. Continual-learning methods address forgetting through mechanisms including regularization, memory, and architectural changes; De Lange et al. (2022) review this landscape. The task-, domain-, and class-incremental distinction matters because inference information changes the problem (van de Ven et al., 2022). EWC protects parameters important to previous tasks (Kirkpatrick et al., 2017); OGD modifies gradient directions to preserve earlier outputs (Farajtabar et al., 2020); function-space regularization directly constrains predictive change (Benjamin et al., 2019). Our interventions instead change the objective used for new-task examples; combinations with these approaches remain untested.

The importance of a training policy is established: Mirzadeh et al. (2020) study learning-rate decay, batch size, and regularization, while Hess et al. (2024) distinguish the objective from its optimization trajectory. Compute-aware comparisons can also change conclusions about continual-learning methods (Prabhu et al., 2023). We build on these observations by comparing loss families under explicit update budgets and a common replay design. Our matched quantity is optimizer updates and replay exposure, rather than equal elapsed time or exhaustive tuning. Work on loss of plasticity also emphasizes the ability to acquire future tasks (Dohare et al., 2024); we measure immediate acquisition only.

Margins and selective updates. Our hinge is a standard multiclass margin objective (Crammer & Singer, 2001). HEM is the closest prior work: it changes how violating class logits and positive-loss samples are aggregated, and already evaluates margin losses in continual learning (Spratling & Schütt, 2026). Consequently, neither margin-based continual learning nor the idea of stopping contributions from sufficiently separated examples is new here. Our comparison adds budget matching, independent validation selection, a dataset transfer check, and explicit acquisition–retention accounting. Focal loss reduces easy-example influence in detection (Lin et al., 2017); LDAM uses class-dependent margins for imbalance (Cao et al., 2019); LUCIR includes inter-class separation in a class-incremental system (Hou et al., 2019). Our task-wise margin discounts address a different axis: the amount of separation demanded as a sequence progresses. Margin trade-offs also arise in few-shot class-incremental generalization (Zou et al., 2022), and Margin Dampening combines constrained predictions with distillation and a gated classifier (Pomponi et al., 2025). These works establish that controlling margins for retention is not itself new; what we study is the simplest task-wise hinge discount, in isolation, with its acquisition cost measured.

Replay and theory. Small episodic memories can be effective (Chaudhry et al., 2019), and exemplar selection is an established part of incremental learning (Rebuffi et al., 2017). DER combines replay with stored logits (Buzzega et al., 2020). We use ordinary label replay as a common resource for every loss rather than as a contribution. For theory, Soudry et al. (2018) show that continued optimization of logistic-type losses can have useful max-margin directional behavior in separable linear models. Evron et al. (2023) analyze sequential max-margin problems arising from weakly regularized linear classification and bound forgetting with recurring tasks. Schliserman et al. (2026) extend the projection perspective to homogeneous deep models, obtaining local convergence under additional regularity while showing that global convergence can fail. These results concern weak regularization and repeated-task projection dynamics; they do not imply that our decreasing-margin, finite-budget policies improve retention. Our analysis below is a diagnostic of update contributions, not an extension of those convergence guarantees.

3 Problem, Policies, and Measurements

The learning problem. Tasks $t = 0, \ldots, K-1$ arrive sequentially. A shared feature extractor has parameters θ and a private linear head for each task. During task t, the current head and the extractor train on that task's data; earlier heads remain fixed. Task identity is supplied at evaluation. Let $R_{t,j}$ be task t accuracy after task j, and define

$$P_t = R_{t,t}, \qquad F_t = P_t - R_{t,K-1}, \qquad \mathrm{ACC} = \frac{1}{K}\sum_{t=0}^{K-1} R_{t,K-1}. \tag{1}$$

P_t measures immediate acquisition; F_t measures loss since acquisition, rather than loss from the best historical score, and may be negative. All accuracy differences are percentage points.

Comparisons with a common past. In the loss comparisons (studies A and B), tasks 0–6 are trained once per seed with CE. Each arm restores the resulting parameters, BatchNorm buffers, and random-number state before learning tasks 7–9. We report final accuracy on tasks 0–6 (*old-task retention*), immediate acquisition on tasks 7–9, and their mean forgetting. For the shared-prefix tasks,

$$\Delta R_{t,K-1} = -\Delta F_t \quad (t \le 6). \tag{2}$$

Their acquisition cannot explain a subsequent difference. For any task population S, the descriptive identity $\Delta\overline{R}_S = \Delta\overline{P}_S - \Delta\overline{F}_S$ separates how much was learned from how much was later lost.

Losses and replay. For sample i, let $d_i = z_{i,y_i} - \max_{j \neq y_i} z_{i,j}$ be the true-class gap. The full-batch hinge is

$$L_{\mathrm{H}}(m) = \frac{1}{B}\sum_{i=1}^{B}[m - d_i]_+. \tag{3}$$

We compare it with ordinary CE and an implementation checked against the authors' HEM source. HEM averages selected high class errors within a sample and then positive-loss samples within a

batch. Its nominal margin follows $\sqrt{M/N_{\text{train}}}$, with class weights from realized training counts (Appendix A). A/B hold the hinge margin at 0.25 and HEM's M at 70.3125, yielding nominal margins 0.125 on CIFAR and approximately 0.0884 on TinyImageNet; neither margin is searched in A/B.

Replay stores one real image per class, nearest to its class feature mean after that task. At 50 evenly spaced updates per later task, a temperature-2 label-CE gradient from all stored old-task examples is added to the new-data trunk gradient with weight α. Replay uses frozen old heads and evaluation-mode BatchNorm. This is a small-memory comparison with a fixed replay event budget; all losses receive the same selection rule, memory capacity, and event count, although the images selected after branching may differ. The weight is independently selected for each loss, so neither identical gradients nor identical α are implied by matched replay.

Data, model, and selection. Split-CIFAR-100 (Krizhevsky, 2009) has ten ten-class tasks; Split-TinyImageNet (Le & Yang, 2015) has ten twenty-class tasks. The latter is resized to 32×32 by an exact 2×2 integer mean. We use a CIFAR-stem ResNet-18 (He et al., 2016), SGD with momentum 0.9, batch size 128, augmentation, no weight decay, fresh momentum each task, and a per-task cosine learning-rate schedule. A deterministic 10% within-task split supplies validation; it is not class-stratified. Training/validation counts per task are 4,500/500 and 9,000/1,000, respectively. The task partition is fixed across seeds.

Study A compares the three losses at two later-task budgets, 10 and 40 epochs per task (E10 and E40, i.e., 1,080 and 4,320 updates over tasks 7–9). Each loss receives the same size learning-rate/replay grid at both budgets (three rates around its incumbent value, four replay weights, four validation seeds). For each loss, budget, and replay status, selection maximizes final validation accuracy while keeping old-task accuracy and acquisition within two points of that loss's same-budget no-replay reference. Because the reference is loss-specific, the acquisition floor differs across losses; Section 5 reports where it was binding, and Appendix B lists every configuration. In study B, A's validation-selected E40 policies receive the same three-point learning-rate adaptation on TinyImageNet. The frozen A and B configurations are then evaluated on eight fresh paired seeds each.

Inference and evaluation history. Each study preregistered its own two-contrast primary family and fixed seed count before collecting its rows. We report paired two-sided Student t 95% intervals; primary p values use Holm correction within the relevant family. The intervals themselves are pointwise, not simultaneous. Other contrasts are secondary and uncorrected. They quantify training-seed variability for a fixed task partition and test set. CIFAR's official test images had been used in earlier development of this work; fresh seeds do not restore a pristine test set. TinyImageNet's labeled official validation split was held out as test and first evaluated in the B confirmation, with seven exact train/test image duplicates disclosed by the data audit. These qualifications apply to every result below.

4 Why Margin and Normalization Are Distinct Controls

The loss policies have a simple update-level interpretation. At a fixed model and batch, let $A(m) = \{i : d_i < m\}$ and $a(m) = |A(m)|/B$. Away from hinge ties,

$$\nabla_\theta L_{\text{H}}(m) = \frac{1}{B} \sum_{i \in A(m)} -\nabla_\theta d_i = a(m)\, \overline{g}_{A(m)}. \tag{4}$$

If the active set is empty, the gradient is zero. Decreasing m can remove samples from this sum: it changes both the active fraction and the composition of the mean gradient, whereas multiplying the learning rate rescales the existing sum. Except when these sums happen to be collinear, one cannot reproduce a margin change by a scalar learning-rate adjustment. Within an interval that crosses no sample gap, however, changing the hinge margin leaves its gradient unchanged.

Observation 1 (An update contribution bound)**.** *At an update with $\|\nabla_\theta d_i\| \le G$ for each sample, the full-batch hinge satisfies $\|\nabla_\theta L_{\text{H}}\| \le a(m)G$. If its positive contributions instead use active-sample averaging, the corresponding bound is G. For a fixed model and batch, reducing the margin makes the first bound no larger when a common G is used.*

This follows from the triangle inequality; the complete argument and a momentum extension are in Appendix E. The statement bounds a contribution, not the actual norm ordering across two margins: removing a gradient can remove cancellation. It also does not compare the active fractions of different training trajectories. HEM's positive-sample normalization removes the analogous active-fraction denominator effect, while its within-sample aggregation changes directions as well. This makes normalization a plausible reason that the useful training duration differs across losses. The A/B endpoints cannot isolate it, because HEM also changes within-sample aggregation and its margin; Section 5 therefore tests the two factors separately in a 2×2 ablation.

Smaller direct new-task contributions need not freeze previous predictions. Momentum can continue moving parameters; replay also contributes updates; and BatchNorm statistics (Ioffe & Szegedy, 2015) can change without a parameter gradient. Under an additional Lipschitz assumption, old-task prediction stability depends on cumulative displacement and buffer changes relative to existing decision gaps, not just on a zero current loss (Appendix E). We therefore test a trade-off rather than assume a forgetting guarantee: progressively smaller margins may reduce further adaptation, helping retention while weakening new-task acquisition. This prediction motivates the separate schedule experiment in Section 6.

5 Matched Budgets Change the Loss Ranking

Table 1: Final test accuracy (%, mean±SD over eight paired seeds per study). Every entry is a separately validation-selected configuration. Budgets and replay opportunities match across losses within a column. CIFAR's test set has historical exposure; TinyImageNet is held out as described in Section 3.

	Split-CIFAR-100 (A)		Split-TinyImageNet (B)
Loss	10 epochs/task	40 epochs/task	40 epochs/task
CE	51.27±1.90	49.00±2.46	26.76±0.93
CE + replay	56.72±1.69	54.70±2.05	31.32±1.04
Hinge	54.23±1.70	54.54±1.34	34.41±0.88
Hinge + replay	57.05±1.46	58.52±1.07	36.66±0.63
HEM	54.63±1.52	53.58±1.53	34.92±0.94
HEM + replay	57.93±1.06	55.14±0.89	36.54±0.30

At the longer budget, hinge leads on CIFAR. The E40 primary comparison gives hinge with replay $+3.38$ points over HEM with replay (95% CI [2.56, 4.20]) and $+3.82$ over CE with replay [2.92, 4.72]; both are positive on all eight seeds and have Holm $p < 10^{-4}$. The hinge advantage is therefore not an artifact of unequal training duration: it remains at this matched budget with replay available to every loss. Against HEM, hinge retains tasks 0–6 better by $+3.20$ [2.31, 4.09], while acquiring tasks 7–9 less accurately by -3.06 [-3.66, -2.46]. Its later-task forgetting is also lower, by 6.86 points. The higher final accuracy comes from retention, not from acquisition.

At the shorter budget, HEM leads. At E10 with replay, hinge minus HEM is -0.87 [-1.39, -0.35] (secondary; HEM wins all eight seeds), and the two margin losses are close to CE: hinge minus CE is $+0.33$ [-0.17, 0.83] and HEM minus CE $+1.20$ [0.40, 2.01]. Increasing the selected policy's budget from E10 to E40 improves hinge by $+1.47$ [1.08, 1.85], while the selected HEM and CE policies decline by 2.79 and 2.02 points. Learning rate and replay weight were independently selected at each budget, so this is a reversal of *selected policy rankings*, not an isolated causal effect of extra epochs. A practitioner who can choose either budget sees a much smaller difference between hinge E40 and HEM E10: $+0.60$ [0.27, 0.92], at four times the new-data updates. This last comparison is post hoc and its interval is not selection-adjusted.

More updates cost the selected CE and HEM policies old-task accuracy; hinge keeps it. Figure 1b shows the retention component of the reversal. Between the selected E10 and E40 replay policies, old-task accuracy changes by -3.78 [-4.49, -3.08] for CE and -3.61 [-4.34, -2.87] for HEM, but by $+1.02$ [0.41, 1.62] for hinge, while acquisition increases for CE ($+4.63$), HEM ($+7.28$), and hinge ($+3.40$). The same ordering holds without replay (-4.71, -2.05, and -0.57

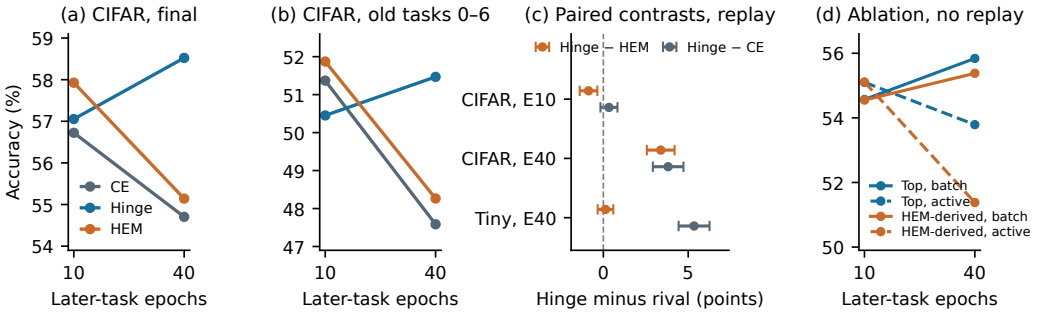

Figure 1: **A budget- and dataset-dependent ranking.** (a) Final and (b) old-task (0–6) test accuracy of the separately validation-selected replay policies on CIFAR; connecting lines do not isolate duration from learning rate or replay weight. For these selected policies, the longer budget raises acquisition for every loss, but only hinge avoids a decline in old-task accuracy. (c) Paired hinge-minus-rival contrasts of final accuracy with pointwise 95% t intervals. CIFAR E10 is secondary; E40 and Tiny E40 are the primary comparisons of separate studies. (d) Final accuracy of the four 2×2 ablation cells (fixed margin 0.25, rate 0.005, no replay, eight seeds): batch-averaged cells gain from the longer budget, active-averaged cells lose.

$[-1.35, 0.21]$ for CE, HEM, and hinge) and, more stringently, inside the validation grid at every fixed learning rate without replay: quadrupling the budget changes old-task validation accuracy by -0.7 to -11.4 points for CE and -2.2 to -6.4 for HEM, but by $+1.0$, -0.7, and -4.2 for hinge at its three rates, and hinge loses the least at each absolute rate shared by two losses (Table 6). The recorded trunk displacement is consistent with this: at E40 without replay, hinge moves the trunk parameters less than CE by 2.07 [2.04, 2.10] and less than HEM by 1.09 [1.03, 1.15] in ℓ_2 norm on CIFAR (Appendix C). Displacement alone does not explain retention; replay policies can have both greater displacement and better old-task accuracy.

The acquisition guardrail was binding at E40. Selection required acquisition within two points of each loss's own no-replay reference. At E40 this excluded each loss's unconstrained best replay configuration, and the exclusion mattered most for HEM: its best unconstrained validation policy (initial rate 0.005, $\alpha = 0.6$) scored 57.5% against 54.8% for the selected one, whereas for hinge the corresponding values are 58.8% and 58.4% (Table 3). On validation, the hinge–HEM gap at E40 is therefore 3.7 points under the preregistered rule but 1.2 points without the acquisition constraint; the reversal survives in the unconstrained grid (HEM ahead by 0.7 at E10, hinge by 1.2 at E40) with a smaller E40 magnitude. The hinge–CE gap is less sensitive (5.0 versus 3.8 on validation). Test accuracy of the unconstrained optima was not collected, so we report the preregistered contrasts as primary and read their magnitude as conditional on the acquisition requirement: a user willing to give up two to three points of new-task accuracy would see hinge and HEM much closer at E40 on validation, and their test gap under that choice is unknown.

A normalization ablation reproduces the reversal within one family. Section 4 identified HEM's positive-sample normalization as a candidate cause of the budget dependence, but HEM also changes class aggregation and its margin. A 2×2 ablation therefore holds the margin (0.25), initial rate (0.005), and no-replay protocol fixed while crossing top-wrong versus HEM-derived class aggregation with full-batch versus active-sample averaging (eight fresh paired seeds, both budgets; Appendix D). Figure 1d shows the same crossing as panel (a): the two batch-averaged cells improve from E10 to E40 ($+1.27$ and $+0.83$ points), while the two active-averaged cells decline (-1.31 and -3.73). At E40, active rather than batch averaging reduces final accuracy by 3.02 points [2.20, 3.85], averaged over aggregation choices, and HEM-derived rather than top-wrong aggregation reduces it by 1.43 [1.02, 1.85] (both Holm $p < .001$); active averaging improves acquisition by 2.54 points but loses 3.38 points of old-task accuracy. At E10 active averaging instead improves final accuracy by 0.55 [0.33, 0.76] (secondary). The top-wrong/batch cell is the A hinge no-replay configuration on new seeds and lands within seed variation of that arm (55.8 versus 54.5 ± 1.3). The ablation establishes a budget-dependent normalization effect at a shared margin; it does not isolate the cause of A's ranking,

because A's HEM baseline also uses a class-adjusted margin, and at a fixed learning rate the two reductions also change the effective step size (Appendix D).

Replay helps all three losses. Within each CIFAR budget, every selected replay policy improves over its family's selected no-replay policy (Table 1). At E40 the gains are $+5.71$ for CE, $+3.98$ for hinge, and $+1.56$ for HEM; their paired intervals exclude zero (Appendix C). Replay is useful here, but its benefit is not specific to hinge. The gains compare selected configurations, whose learning rates may differ.

What transfers to a new dataset? On Split-TinyImageNet at E40, hinge with replay improves over CE with replay by $+5.35$ points [4.43, 6.26] (Holm $p < 10^{-5}$, 8/8 positive). HEM with replay also improves over CE by $+5.22$ [4.28, 6.15] (secondary, 8/8 positive). But hinge minus HEM is only $+0.13$ [-0.34, 0.59] (Holm $p = .538$, 4/8 positive): the CIFAR hinge advantage is not reproduced, and equivalence is not established. Without replay, HEM is slightly ahead of hinge, by 0.51 [0.13, 0.89] in a secondary comparison. The acquisition guardrail was binding for CE in B as well; without it the validation gap to hinge shrinks from 5.5 to 3.5 points (Appendix B), showing validation sensitivity to selection. The reported test gain remains conditional on the frozen acquisition requirement.

TinyImageNet further clarifies the source of the CE gap. Hinge with replay has $+6.50$ [5.68, 7.32] higher old-task accuracy than CE with replay, despite -3.24 [-3.66, -2.81] lower immediate acquisition. Hinge and HEM acquire the later tasks almost identically with replay ($+0.01$ [-0.54, 0.56]), while HEM retains the prefix slightly better (-0.51 [-0.89, -0.13] for hinge). Thus a final average alone can obscure both a real gain over CE and a meaningful difference between the two margin policies. An earlier rotated-CIFAR-10 condition also found no clear hinge–HEM separation (Appendix F).

What does the gain cost? E10 and E40 correspond to 1,080 and 4,320 later-task updates on CIFAR; TinyImageNet E40 uses 8,520. The replay event count stays at 150 over the three later tasks. On the recorded RTX 4090 runs, mean CIFAR suffix times with replay are 26.1/73.5 seconds for hinge and 25.7/72.3 for HEM at E10/E40, and 162.1/161.7 seconds on TinyImageNet E40; the measured times are similar at a fixed budget. They include evaluation and exclude the shared prefix. Maximum stored image-and-label payloads are approximately 1.23 MB and 2.46 MB, with capacity growing by one image per class. Full timings and compute accounting appear in Appendix G.

6 Discounting the Margin Improves Retention at an Acquisition Cost

The preceding comparisons change loss families after a CE prefix. We now isolate a simpler policy within one family. Study C trains all ten Split-CIFAR-100 tasks with hinge, no replay, 40 epochs per task, and initial learning rate 0.005. Each seed first learns task 0 at margin 0.25, then branches with identical state and randomness into

$$m_t = 0.25(1 - r)^t, \qquad r \in \{0, 0.05, 0.10, 0.20\}. \tag{5}$$

The constant policy has $r = 0$. The schedule changes neither the number of updates nor stored memory. Its hypothesis is operational: gradually ask for less separation as tasks accumulate, and measure whether improved retention outweighs weaker acquisition.

Fixed-sample confirmations. The preregistered primary comparisons are 5% versus constant over 128 seeds, and 20% versus constant over the first 32 of those seeds. The 10% arm uses the same 32 seeds as a secondary comparison. All 320 rows are complete; no seed failed the shared first-task fit floor, and sample sizes were not increased after examining these results. Table 2 reports the paired effects. The 5% discount raises final accuracy from 46.04% to 47.85%, a $+1.80$-point gain [1.49, 2.12], positive on 108/128 seeds. The 20% discount raises it from 46.08% to 50.88% on its 32-seed block: $+4.80$ [3.93, 5.68], positive on all 32 seeds. Both primary contrasts remain significant after Holm correction ($p < 10^{-11}$).

The gain has an interpretable price. For 5%, earlier-task final accuracy improves by $+2.47$ [2.06, 2.88], while immediate acquisition falls by -0.65 [-0.72, -0.58]. At 20%, the corresponding

Table 2: Margin discounts versus fixed $m = 0.25$ (C). Differences in accuracy points with paired, pointwise 95% intervals. Earlier: tasks 0–6 final. Acquisition: immediate accuracy on tasks 1–9. Forgetting: mean loss since acquisition over tasks 0–8. These task populations differ from A/B. *Secondary comparison on the 32 common seeds.

Discount	n	ΔACC	ΔEarlier	ΔAcq.	ΔForget
5%	128	+1.80 [1.49, 2.12]	+2.47 [2.06, 2.88]	−0.65 [−0.72, −0.58]	−2.65 [−3.00, −2.31]
20%	32	+4.80 [3.93, 5.68]	+7.65 [6.57, 8.73]	−3.13 [−3.31, −2.95]	−8.47 [−9.41, −7.53]
10%*	32	+3.83 [3.09, 4.57]	+5.25 [4.35, 6.16]	−1.28 [−1.41, −1.15]	−5.53 [−6.35, −4.72]

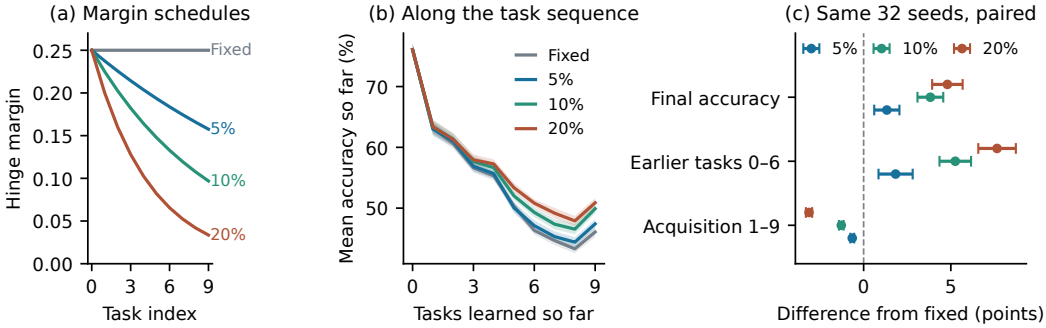

Figure 2: **Margin discounting trades acquisition for retention.** (a) The four schedules. (b) Mean test accuracy over the tasks seen so far, after each task, on the common 32 seeds (bands: pointwise 95% t intervals of the mean); the policies separate progressively along the sequence. (c) Paired differences from the fixed margin on the same 32 seeds. The primary 5% estimate uses 128 seeds and is in Table 2; it is not mixed with the 32-seed means to rank rates.

changes are $+7.65$ [6.57, 8.73] and -3.13 [-3.31, -2.95]. Unlike A/B, C's tasks 1–6 are learned after the fork, so "old 0–6" is descriptive rather than a pure common-prefix retention effect. Direct forgetting measurements resolve this: mean forgetting falls by 2.65 and 8.47 points for the two primary discounts. Because task-0 acquisition is shared and the last task has zero subsequent forgetting, the exact decomposition here is

$$\Delta\mathrm{ACC} = \frac{9}{10}\left(\Delta\overline{P}_{1:9} - \Delta\overline{F}_{0:8}\right). \tag{6}$$

For 20%, $0.9(-3.13 + 8.47) = 4.80$ after rounding. Final performance improves because less is lost later, even though less is acquired immediately. The benefit accrues along the sequence (Figure 2b): after two tasks the four policies are within 0.8 points of one another, and the gap widens as tasks accumulate. The discount does not prevent the large first drop of task 0 (from 76.1% to 51–52% after task 1 under every policy), but it slows the later erosion, leaving task 0 at 20.9, 22.4, 25.8, and 28.3% for the fixed, 5%, 10%, and 20% policies (Appendix C).

Which discount is preferable? On the common 32 seeds, 10% gives $+3.83$ [3.09, 4.57] over constant. Increasing the discount from 10% to 20% adds $+0.98$ [0.11, 1.84] to final accuracy but costs another 1.85 points of acquisition (both secondary). The highest final mean among these tested rates is therefore not an acquisition-preserving recommendation. A mild discount is attractive when a smaller acquisition penalty matters; a stronger one is attractive when final average accuracy has priority. No utility function or universal optimum was established.

What does this establish about a schedule? The practical result is a reproducible benefit over the stated constant baseline, not a requirement to match average margins. A mean-matched flat policy addresses a narrower mechanism question: does the temporal allocation add value beyond lowering the overall margin? Earlier eight-seed replications did not distinguish a geometric policy from that control (Appendix F), without proving equivalence, and C does not compare against a validation-tuned constant margin. The open comparator is therefore a smaller constant margin chosen with the same validation opportunity; until it is tested, the discount should be read as a reliable way

to lower the demanded separation over a sequence, not as evidence that the geometric ordering itself matters. No completed experiment combines the discount with the A/B replay policies, so the C and A/B gains cannot be added.

7 Discussion and Limitations

Practical guidance. The main lesson is to select a loss *policy* for a budget and an acquisition requirement, and to report acquisition, retention, and compute alongside final accuracy. Within the tested settings: at the short budget, HEM with replay has the highest final accuracy and CE with replay the highest immediate acquisition; at the long budget, hinge with replay retains old tasks best and gives the highest final accuracy on CIFAR, and on TinyImageNet the frozen margin-loss policies beat CE by about five points; if a hinge-family loss must be trained long, batch averaging should be preferred to active-sample averaging; within an all-hinge learner, discounting the margin by 5–10% per task buys 2.5–5.3 points of earlier-task accuracy for 0.6–1.3 points of immediate acquisition. These operating points make the new-task objective a useful experimental control alongside memory and optimization.

Limitations. The evidence is conditional on one backbone, private heads, fixed task partitions, a CE prefix in A/B, and a finite validation search. Three design choices deserve emphasis. The selection guardrail was loss-specific, and its acquisition floor was binding at E40; the reported hinge–HEM margin therefore depends on how much acquisition one requires, as Section 5 quantifies. The hinge margin (0.25) and HEM's nominal margin (0.125 on CIFAR, 0.088 on TinyImageNet) were held at their incumbent values in A/B; given C's sensitivity to the margin, a joint margin search could change the hinge–HEM ranking, and the present comparisons do not optimize margins jointly with the other settings. The replay memory is one image per class with 50 replay events per task, which specifies a deliberately small replay resource; whether the loss effects survive a large buffer replayed at every step, or a combination with regularization methods such as EWC, is untested. Some selected learning rates and replay weights lie at search boundaries, and equal search size is not a guarantee that every family is globally optimized. Historical CIFAR test exposure limits confirmatory interpretation beyond the fixed seed blocks; the held-out TinyImageNet test split provides an additional transfer check but only at one budget and a reduced image resolution. The analytical observations do not prove that normalization causes the ranking reversal or that decreasing margins reduce forgetting along arbitrary trajectories. Broader architecture, class-incremental, tuned-constant, and combined-policy evaluations would answer distinct questions; the present results do not substitute for them.

8 Conclusion

The loss used to learn a new task is also a control over how the shared model changes. Matched-budget experiments show that the relative merits of hinge and HEM depend on the budget, the dataset, and the acquisition one is willing to forgo, while the margin losses' advantage over CE at a long budget is reproducible and comes from retention. A normalization ablation reproduces the budget dependence within one loss family, and a separate margin-discount study establishes final-accuracy gains through reduced forgetting with a graded acquisition cost. These results provide concrete operating points and a clearer basis for choosing loss policies when both existing and future capabilities matter.

REPRODUCIBILITY STATEMENT

The anonymous supplementary package provides 464 A/B/C confirmation rows, 360 validation rows, and the 64 normalization-ablation rows, with checksums and CPU analysis scripts (NumPy, SciPy, Matplotlib; tested with Python 3.12, and `recompute.py` with 3.8 or later). These regenerate every table, figure, paired interval, and frozen validation selection in this paper. Appendix B specifies the grids and selection rules; Appendices A and D specify the implementations; Appendix E gives the analytical details. The package reproduces analysis from recorded runs, not full model training. Historical pilot summaries in Appendix F are not independently regenerated by this package.

AI USE STATEMENT

Generative AI tools (ChatGPT and Claude, including Claude Code) were used in this work for developing the conceptual framework and hypotheses, designing the experimental protocols, implementing the training and analysis code, formulating and proving the elementary mathematical claims, Korean–English translation, cleaning and reformatting data, and interpreting results, as well as for figures, literature search, reference formatting, and drafting and editing the manuscript. They were not used to generate synthetic data or any experimental observation. The authors reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Ari S. Benjamin, David Rolnick, and Konrad P. Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.08289.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b704ea2c39778f07c617f6b7ce480e9e-Abstract.html.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.nips.cc/paper_files/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. URL https://arxiv.org/abs/1902.10486.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. URL https://www.jmlr.org/papers/v2/crammer01a.html.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632: 768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9440–9484. PMLR, 2023. URL https://proceedings.mlr.press/v202/evron23a.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773. PMLR, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.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://arxiv.org/abs/1512.03385.

Timm Hess, Tinne Tuytelaars, and Gido M. van de Ven. Two complementary perspectives to continual learning: Ask not only what to optimize, but also how. In *Proceedings of the 1st ContinualAI Unconference, 2023*, volume 249 of *Proceedings of Machine Learning Research*, pp. 37–61. PMLR, 2024. URL https://proceedings.mlr.press/v249/hess24a.html.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hou_Learning_a_Unified_Classifier_Incrementally_via_Rebalancing_CVPR_2019_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.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

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.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Cs231n course project report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/518a38cc9a0173d0b2dc088166981cf8-Abstract.html.

Jary Pomponi, Alessio Devoto, and Simone Scardapane. Class incremental learning with probability dampening and cascaded gated classifier. *Neurocomputing*, 640:130295, 2025. URL https://arxiv.org/abs/2402.01262.

Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K. Dokania, Philip H. S. Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual learning: What does matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3698–3707, 2023. URL https://arxiv.org/abs/2303.11165.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1611.07725.

Matan Schliserman, Gon Buzaglo, Itay Evron, and Daniel Soudry. Convergence of continual learning in homogeneous deep networks. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5743–5784. PMLR, 2026. URL https://proceedings.mlr.press/v336/schliserman26a.html.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Michael W. Spratling and Heiko H. Schütt. HEM: a margin-based loss for visual categorisation tasks. In *Computer Vision – ECCV 2026*, volume 17029 of *Lecture Notes in Computer Science*, pp. 209–227. Springer, 2026. doi: 10.1007/978-3-032-37627-5_12.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022. doi: 10.1038/s42256-022-00568-3.

Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae817e85f71ef86d5c9566598e185b89-Abstract-Conference.html.

A Experimental Protocol and Implementation

Architecture and optimization. All central experiments use the same ResNet-18 variant: width 64, a 3×3 stride-1 input convolution, no initial max-pooling, BatchNorm, and a separate linear head per task. Only the current head and shared trunk are optimized. Momentum is 0.9 and is reset at task boundaries; weight decay is zero. The base learning rate is 0.05 multiplied by the reported loss-specific multiplier. A cosine schedule spans each task's prescribed updates. Consequently, E40 is a different duration policy, including its learning-rate horizon, rather than continuing an E10 run after its learning rate reaches zero. Batches have size 128, including the final partial batch. Augmentation is enabled. In A/B the shared prefix has seven tasks, ten epochs each, and CE with initial learning rate 0.05. It contributes 2,520 CIFAR or 4,970 TinyImageNet updates. Each C branch contributes 12,960 updates after its shared 1,440-update first task.

Dataset preparation and split correction. Task partitions are fixed rather than resampled across seeds. The training/validation split uses a within-task random permutation with seed 987654, withholding 10% of each task. It is not stratified by class. Thus the TinyImageNet preregistration's shorthand "450/50 per class" should be read as expected counts; the implementation produces exactly 9,000/1,000 per task, with realized per-class counts varying. HEM uses those realized counts. The TinyImageNet classes are sorted by WordNet ID and divided into ten consecutive groups of twenty. The prepared 32×32 array uses integer averaging of each 2×2 block from the original 64×64 image and train-set normalization. Its official labeled validation split supplies 1,000 test images per task. The preparation audit found seven exact train/test image duplicates; these were not removed after evaluation, and they can move a task's test accuracy by at most 0.7 points and the ten-task average by at most 0.07. The source archive MD5 is `90528d7ca1a48142e341f4ef8d21d0de`, and the prepared-array MD5 is `c1605b41c7702a088ce71c968e171c96`.

HEM implementation. The implementation is checked against the authors' vendored source at commit `8bfc7606`, obtained before these confirmations. It uses the target-class weight, as in that code. If a task contains C classes with training counts n_c, let $N = \sum_c n_c$, $\bar{n} = N/C$, $w_c = \sqrt{\bar{n}/(n_c + 10^{-9})}$, and $m = \sqrt{M/N}$. For sample i, the errors are

$$e_{i,j} = \mathbf{1}[j \neq y_i]\,[z_{i,j} - z_{i,y_i} + m w_{y_i}]_+. \tag{7}$$

The sample loss averages errors at least as large as the detached mean over all C logits, including the zero true-class entry. The batch loss sums the sample losses and divides by the count of strictly positive sample losses plus 10^{-6}. Both the detached threshold and the non-strict comparison are preserved. This differs from our hinge in both within-sample aggregation and batch reduction. HEM's nominal margin and class-count weights are fixed in A/B; its learning rate and replay weight are selected as below.

Replay details. For each class, the stored example is the real training image whose feature is nearest in squared Euclidean distance to the class feature mean. Prefix exemplars are shared because their capturing model is shared; later exemplars can differ across arms. At every replay event the loss is the mean over stored tasks of their label cross-entropy evaluated on logits divided by 2. There is no extra temperature-squared factor. The replay gradient updates the trunk, while old heads remain fixed. Its raw gradient is multiplied by α with no adaptive norm cap in A/B. Current-task batches run in training mode; replay runs in evaluation mode and does not update BatchNorm buffers. With S updates in a task, replay event indices are $\lfloor (k + 1/2)S/50 \rfloor$, $k = 0, \ldots, 49$. The CIFAR suffix processes $50(70 + 80 + 90) = 12,000$ stored-image presentations; TinyImageNet processes 24,000. These counts match across losses and across the two CIFAR epoch budgets.

B Selection, Replicate Units, and Statistical Families

A: matched budgets. For each loss and E$\in \{10, 40\}$, the learning-rate multiplier is $q \in \{0.5, 1, 2\}$ times its incumbent multiplier: CE 0.05, hinge 0.10, HEM 0.20 (initial rates $\eta_0 = 0.05\times$multiplier). Replay weight is $\alpha \in \{0, 0.1, 0.3, 0.6\}$. This gives $3 \times 2 \times 3 \times 4 = 72$ configurations over seeds 2230–2233 (288 rows), evaluated on validation only. Selection is separate for each loss, budget, and replay status. The reference is that loss at the same budget with $q = 1, \alpha = 0$. Eligible candidates

must have finite final accuracy and have mean acquisition and old-task accuracy at least reference minus 2.0 points. The highest validation accuracy wins; candidates within 0.05 points of the highest are tied, with ties resolved by smaller α, then q closest to 1 in absolute log distance, then smaller learning rate. A fallback to the best finite candidate is specified if none passes; the selected policies did not require it. The frozen twelve policies are evaluated on seeds 2240–2247 (96 rows). Table 3 lists every configuration's validation accuracy, its eligibility, and the selected policies.

B: validation-only transfer. A's E40 validation selections, rather than its test results, define the starting policies. For each loss and replay status, B varies the transferred learning-rate multiplier by $\{0.5, 1, 2\}$ and holds the transferred replay weight fixed. Eighteen configurations over seeds 2250–2253 give 72 validation-only rows. The guardrail reference is the same loss's transferred no-replay policy at factor 1; the selection rule is otherwise unchanged. Six frozen policies are evaluated on seeds 2260–2267 (48 rows). Table 4 lists the B grid; Table 5 reports the selected multipliers and initial rates. Boundary choices emphasize that these are finite-grid selections, not globally optimized losses.

Fixed-learning-rate budget effect. Because the selected E10 and E40 policies use different learning rates, Table 6 reads the budget effect directly off the validation grid at each fixed initial rate without replay. For every loss, quadrupling the budget raises immediate acquisition by 5.6–10.8 points. Old-task accuracy falls at every rate for CE and HEM; for hinge it rises at the lowest rate, is nearly unchanged at the middle rate, and falls least at the highest rate. At matched absolute rates (0.0025 and 0.01 appear for two losses each, 0.005 for all three) hinge always loses the least. These are four-seed validation means, reported as a descriptive sensitivity analysis; they do not establish a universal ordering or remove all optimization-policy effects.

C: fixed samples and task populations. Seeds 3000–3127 compare the constant and 5% policies. The first 32, seeds 3000–3031, additionally compare 10% and 20%. Task 0 is trained once at margin 0.25, E40, initial rate 0.005; the validation fit floor is 0.50. A failing seed would be reported and excluded without replacement. There were no failures. All arms share task-0 parameters, BatchNorm buffers, random state, and subsequent image order. The sample sizes and analysis were fixed before this new block; they were not a continuation of a stopping rule based on significance. The historical 5% estimate informed the prospective sample-size calculation, not the final effect estimate.

Primary families. Each of A and B has exactly two primary final-accuracy contrasts: hinge+replay minus HEM+replay, and hinge+replay minus CE+replay, both at E40. C has exactly two: 5% minus constant on 128 seeds and 20% minus constant on 32. Holm adjustment is applied separately within each study, not across all project history. The C contrasts overlap in their first 32 seeds; Holm control does not require independence between tests. For paired differences d_s over n seeds, we compute

$$\widehat{\Delta} = \overline{d}, \qquad \mathrm{CI}_{.95} = \overline{d} \pm t_{n-1,.975} \frac{s_d}{\sqrt{n}}. \tag{8}$$

These intervals rely on the usual paired-t assumptions and are pointwise; they do not account for adaptive reuse of a benchmark, between-dataset variation, or test-based selection among budgets. A non-rejection is not evidence of equivalence. Secondary acquisition and forgetting contrasts are descriptive, uncorrected analyses rather than additional primary discoveries.

C Complete Central Results

Primary paired effects and corrected probabilities. A: hinge+R minus HEM+R is $+3.381$ [2.564, 4.199], and hinge+R minus CE+R is $+3.819$ [2.916, 4.721], both with Holm $p = 4.27 \times 10^{-5}$. B: the respective effects are $+0.126$ [-0.335, 0.588], Holm $p = .538$, and $+5.345$ [4.434, 6.256], Holm $p = 4.77 \times 10^{-6}$. C: 5% minus constant is $+1.803$ [1.490, 2.117], Holm $p = 8.01 \times 10^{-21}$; 20% minus constant is $+4.802$ [3.928, 5.676], Holm $p = 1.97 \times 10^{-12}$. Full precision, paired standard deviations, and signs appear in `results.json`.

Budget changes of the selected policies. E40 minus E10 for the same loss and replay status, paired over the eight seeds. With replay, old-task accuracy: CE -3.784 [-4.493, -3.075], hinge $+1.016$

Table 3: The complete A validation grid: final validation accuracy (%, mean over seeds 2230–2233) for every loss, later-task budget, initial learning rate η_0, and replay weight α. Bold: the policy selected by the preregistered rule (separately for $\alpha = 0$ and $\alpha > 0$). †: ineligible because its mean acquisition or old-task accuracy fell more than two points below the same loss's reference ($q = 1$, $\alpha = 0$). At E40, the acquisition floor excluded each loss's unconstrained best replay configuration, including the validation optima of hinge (58.76), HEM (57.53), and CE (54.98); these optima were never evaluated on test.

			Replay weight α			
Loss	E	η_0	0	0.1	0.3	0.6
CE	10	0.00125	50.43†	51.30†	52.43†	53.69†
CE	10	0.0025	**52.19**	53.30	54.78	56.36
CE	10	0.005	51.90	54.11	56.05	**57.06**
CE	40	0.00125	51.98†	52.66†	53.77†	54.98†
CE	40	0.0025	**48.45**	50.05	51.76	**53.42**
CE	40	0.005	43.44†	46.71	49.84	51.74
Hinge	10	0.0025	53.33†	54.69†	56.09†	55.88†
Hinge	10	0.005	54.47	56.32	57.55	51.59†
Hinge	10	0.01	**55.06**	**57.51**	55.34	36.27†
Hinge	40	0.0025	55.35†	56.32†	57.78†	58.76†
Hinge	40	0.005	**54.92**	56.52	**58.44**	56.75
Hinge	40	0.01	52.64†	55.21	56.72	50.09†
HEM	10	0.005	54.98	56.83	**58.22**	53.65†
HEM	10	0.01	**55.09**	57.19	57.22	41.25†
HEM	10	0.02	54.28	56.97	51.80	35.13†
HEM	40	0.005	**53.95**	55.40†	57.21†	57.53†
HEM	40	0.01	50.39	52.32	54.40	**54.79**
HEM	40	0.02	49.73	51.72	53.52	54.53

Table 4: The complete B validation grid on TinyImageNet: final validation accuracy (%, mean over seeds 2250–2253) at each transferred initial rate η_0 (given before the colon); the transferred replay weight α is fixed per row. Bold and † as in Table 3. The acquisition floor was binding for CE (with and without replay) and for hinge without replay; for hinge with replay and both HEM policies the selected policy is the unconstrained optimum.

Policy	α	Rate $\times 0.5$	Rate $\times 1$	Rate $\times 2$
CE	0.0	0.00125: 29.88†	0.0025: **25.74**	0.005: 19.64†
CE + replay	0.6	0.00125: 32.49†	0.0025: **30.44**	0.005: 25.64
Hinge	0.0	0.0025: 34.00†	0.005: **33.93**	0.01: 31.69†
Hinge + replay	0.3	0.0025: 35.80†	0.005: **35.96**	0.01: 34.70
HEM	0.0	0.0025: **33.85**	0.005: 33.62	0.01: 31.33†
HEM + replay	0.6	0.005: **35.69**	0.01: 33.42	0.02: 31.64

[0.409, 1.623], HEM -3.607 [-4.342, -2.872]; acquisition: CE $+4.629$, hinge $+3.400$, HEM $+7.279$; later-task forgetting: CE $+2.533$, hinge $+0.875$, HEM $+8.150$. Without replay, old-task accuracy: CE -4.707 [-5.863, -3.551], hinge -0.570 [-1.350, 0.211], HEM -2.050 [-3.030, -1.070]. The E10 and E40 policies of a loss can differ in learning rate and replay weight (Table 5); Table 6 gives the fixed-rate version.

Trunk displacement. Paired displacement contrasts (hinge minus rival, ℓ_2 norm) on CIFAR: at E40 without replay, -2.067 [-2.097, -2.038] versus CE and -1.091 [-1.148, -1.035] versus HEM; at E40 with replay, -1.992 [-2.130, -1.854] and -6.866 [-7.043, -6.689]; at E10 without replay, -0.185 [-0.219, -0.151] and -0.542 [-0.584, -0.500]. On TinyImageNet without replay, hinge moves less than CE by 3.819 [3.775, 3.863] but more than HEM by 1.041 [1.033, 1.050], where HEM's selected rate is half of hinge's. Across the no-replay arms, larger displacement tends to go with larger later-task forgetting within each dataset (rank correlation 0.77 over the six CIFAR arms; the three TinyImageNet arms are ordered identically; Table 7), but the relation is not monotone once

Table 5: Selected policies, read from the confirmation records. Initial rate $\eta_0 = 0.05\times$multiplier. The last column is the mean over seeds and later tasks of each task's median realized replay-to-new trunk-gradient ratio; zero denotes no replay. The same α need not produce the same ratio.

Study	Arm	LR multiplier	η_0	α	Gradient ratio
A	ce_E10_R	0.1	0.005	0.6	1.020
A	ce_E10_noR	0.05	0.0025	0.0	0.000
A	ce_E40_R	0.05	0.0025	0.6	1.223
A	ce_E40_noR	0.05	0.0025	0.0	0.000
A	h_E10_R	0.2	0.01	0.1	0.418
A	h_E10_noR	0.2	0.01	0.0	0.000
A	h_E40_R	0.1	0.005	0.3	1.368
A	h_E40_noR	0.1	0.005	0.0	0.000
A	hem_E10_R	0.1	0.005	0.3	0.796
A	hem_E10_noR	0.2	0.01	0.0	0.000
A	hem_E40_R	0.2	0.01	0.6	2.063
A	hem_E40_noR	0.1	0.005	0.0	0.000
B	ce_R	0.05	0.0025	0.6	1.040
B	ce_noR	0.05	0.0025	0.0	0.000
B	h_R	0.1	0.005	0.3	1.322
B	h_noR	0.1	0.005	0.0	0.000
B	hem_R	0.1	0.005	0.6	2.695
B	hem_noR	0.05	0.0025	0.0	0.000

Table 6: Budget effect at a fixed learning rate without replay, from the A validation grid: E40 minus E10 in old-task (0–6) validation accuracy and in immediate acquisition (tasks 7–9), points, for each loss at its three initial rates η_0.

	Rate $\times 0.5$			Rate $\times 1$			Rate $\times 2$		
Loss	η_0	ΔOld	ΔAcq.	η_0	ΔOld	ΔAcq.	η_0	ΔOld	ΔAcq.
CE	0.00125	-0.67	+10.75	0.0025	-6.89	+8.35	0.005	-11.43	+6.75
Hinge	0.0025	+1.03	+6.30	0.005	-0.65	+5.88	0.01	-4.23	+5.57
HEM	0.005	-2.24	+5.97	0.01	-6.35	+6.18	0.02	-5.66	+5.83

replay is added, and replay can coincide with greater displacement and better old-task accuracy. Displacement is a parameter-space quantity and is reported as a diagnostic, not as a predictor of function change.

Replay contrasts. All are secondary comparisons of separately selected configurations. CIFAR E10 gains for CE, hinge, and HEM are 5.458, 2.820, and 3.299 points; at E40 they are 5.706, 3.981, and 1.560. TinyImageNet gains are $+4.555$ [4.219, 4.891] for CE, $+2.255$ [1.861, 2.649] for hinge, and $+1.618$ [0.978, 2.257] for HEM. Full paired intervals for all six CIFAR gains are recomputed in the accompanying analysis. Each of the nine intervals excludes zero. Because the learning rate can differ between replay and no-replay policies, their difference-in-differences does not identify a same-rate, same-weight loss–replay interaction.

C: compare rates on the same seeds. On the common 32 seeds, final means are 46.078 (constant), 47.408 (5%), 49.904 (10%), and 50.880 (20%). The common-block 5% effect is $+1.330$ [0.607, 2.053]; its primary 128-seed effect is $+1.803$, and the two should not be mixed in rate comparisons. The 10% policy changes old-task accuracy by $+5.254$ [4.349, 6.160], acquisition by -1.282 [-1.414, -1.149], and forgetting by -5.533 [-6.351, -4.715]. The 20% versus 10% comparison gives $+0.976$ [0.110, 1.842] final accuracy, $+2.396$ [1.350, 3.443] on tasks 0–6, and -1.850 [-2.009, -1.690] acquisition. The earlier two-point acquisition reference is met by 5% and 10% but not 20%; it was not a veto on C's final-accuracy hypothesis. Along the sequence, task-0 test accuracy after tasks 1, 3, 6, and 9 is 51.4/37.6/24.0/20.9 (constant), 51.1/38.2/25.4/22.4 (5%), 51.8/39.8/27.9/25.8 (10%), and 52.4/40.5/29.8/28.3 (20%); the full curves and the mean accuracy over tasks seen so far are stored in `results.json` and drawn in Figure 2b.

Table 7: A/B arm means over eight paired seeds. Accuracy columns are percent; Final is mean±SD across seeds, Old is tasks 0–6 final, Acq. is tasks 7–9 immediate, and Forget is their mean loss since acquisition. Displacement is the recorded ℓ_2 distance between the trunk parameters at the end of task 9 and at the shared checkpoint (mean±SD). Seconds is the suffix duration including evaluation.

Study	Arm	Final	Old	Acq.	Forget	Displacement	Seconds
A	ce_E10_R	56.72±1.69	51.37	77.15	7.94	4.22±0.13	26.4
A	ce_E10_noR	51.27±1.90	44.83	73.29	7.01	2.18±0.02	18.2
A	ce_E40_R	54.70±2.05	47.59	81.78	10.47	5.15±0.15	75.3
A	ce_E40_noR	49.00±2.46	40.13	81.73	12.04	4.72±0.05	67.3
A	h_E10_R	57.05±1.46	50.45	76.32	3.86	2.22±0.06	26.1
A	h_E10_noR	54.23±1.70	46.56	76.42	4.28	1.99±0.04	17.8
A	h_E40_R	58.52±1.07	51.47	79.72	4.74	3.16±0.08	73.5
A	h_E40_noR	54.54±1.34	45.99	79.91	5.42	2.65±0.05	65.7
A	hem_E10_R	57.93±1.06	51.87	75.50	3.45	2.25±0.09	25.7
A	hem_E10_noR	54.63±1.52	47.11	77.73	5.56	2.54±0.09	17.4
A	hem_E40_R	55.14±0.89	48.26	82.78	11.60	10.02±0.27	72.3
A	hem_E40_noR	53.58±1.53	45.06	81.47	8.01	3.74±0.11	64.3
B	ce_R	31.32±1.04	22.83	64.35	13.23	7.78±0.05	166.5
B	ce_noR	26.76±0.93	17.57	64.57	16.36	7.48±0.03	153.6
B	h_R	36.66±0.63	29.33	61.12	7.33	4.08±0.06	162.1
B	h_noR	34.41±0.88	26.38	60.99	7.84	3.66±0.04	153.3
B	hem_R	36.54±0.30	29.84	61.11	8.94	5.86±0.26	161.7
B	hem_noR	34.92±0.94	27.75	59.15	7.51	2.62±0.04	155.0

Table 8: C margins and per-task final accuracy on the common 32 seeds. The policy changes every task after task 0; it does not apply a one-time discount to all tasks. Final accuracies are percent.

	Training margin				Final accuracy			
Task	Fixed	5%	10%	20%	Fixed	5%	10%	20%
0	0.2500	0.2500	0.2500	0.2500	20.94	22.40	25.77	28.26
1	0.2500	0.2375	0.2250	0.2000	27.09	27.76	31.88	34.60
2	0.2500	0.2256	0.2025	0.1600	28.44	30.57	34.40	37.70
3	0.2500	0.2143	0.1823	0.1280	33.07	34.21	38.90	41.51
4	0.2500	0.2036	0.1640	0.1024	36.41	39.87	44.27	48.31
5	0.2500	0.1934	0.1476	0.0819	47.23	50.44	52.68	54.27
6	0.2500	0.1838	0.1329	0.0655	54.07	54.81	56.12	56.15
7	0.2500	0.1746	0.1196	0.0524	57.21	58.30	60.20	58.27
8	0.2500	0.1659	0.1076	0.0419	66.85	67.29	67.57	65.69
9	0.2500	0.1576	0.0969	0.0336	89.47	88.42	87.25	84.05

D Normalization and Aggregation Ablation

Design and implementation. Eight fresh paired seeds (5200–5207) each share the same CE prefix on tasks 0–6 and branch into eight cells on tasks 7–9: two aggregation rules, two sample reductions, and E10/E40. This gives 64 rows, with 1,080/4,320 suffix updates. All cells use margin 0.25, initial learning rate 0.005, no replay, and the A architecture, split, momentum, and per-task cosine schedule. Top-wrong aggregation uses $\ell_i = [0.25 - d_i]_+$. HEM-derived aggregation forms wrong-class errors at the same unweighted margin and averages those at least as large as their detached mean (including the zero true-class entry), as in Appendix A. These shared-margin cells do not use A/B's class-adjusted HEM margin and are not interchangeable with that baseline.

Let $n_+ = \sum_i \mathbf{1}[\ell_i > 0]$. Full-batch reduction is $L_B = B^{-1} \sum_i \ell_i$, while active-sample reduction is $L_+ = (n_+ + \epsilon)^{-1} \sum_i \ell_i$ with $\epsilon = 10^{-6}$. Away from boundaries and with the active count treated as constant, $\nabla L_B = ((n_+ + \epsilon)/B) \nabla L_+$. Ignoring the numerical stabilizer gives the familiar active-fraction factor n_+/B. At fixed learning rate these reductions therefore also change update scale; they are not a control that holds effective step size fixed, especially with momentum. The recorded `active_frac` column contains unmeasured 0.0 placeholders for these loss keys. No active-fraction trajectory is inferred from it.

Table 9: C per-task immediate acquisition (P) and forgetting (F), common 32 seeds, percent or points. Task-0 acquisition is identical by construction, and task-9 forgetting is zero. These are descriptive task means, not independent statistical replicates.

	Fixed		5%		10%		20%	
Task	P	F	P	F	P	F	P	F
0	76.07	55.13	76.07	53.67	76.07	50.29	76.07	47.81
1	74.97	47.88	74.91	47.14	74.92	43.04	74.35	39.74
2	80.70	52.26	80.45	49.88	79.98	45.58	79.33	41.63
3	79.95	46.88	79.64	45.43	78.95	40.04	77.73	36.22
4	83.24	46.83	82.67	42.80	82.32	38.05	80.95	32.64
5	83.63	36.40	83.01	32.57	82.53	29.86	81.30	27.02
6	83.63	29.56	82.63	27.82	81.94	25.82	79.32	23.17
7	82.60	25.39	81.75	23.45	80.95	20.75	78.27	20.00
8	86.10	19.24	84.95	17.66	83.93	16.36	80.84	15.15
9	89.47	0.00	88.42	0.00	87.25	0.00	84.05	0.00

Prespecified comparisons and complete cells. The fixed-sample primary family contains two E40 final-accuracy effects: active minus batch reduction, averaged over the two aggregation choices within each seed, and HEM-derived minus top-wrong aggregation, averaged over the two reductions. We compute paired t intervals over the resulting eight seed differences and apply Holm correction to these two tests. Table 10 reports the absolute cells. E10, task-specific metrics, interactions, and within-cell budget changes are secondary and uncorrected. The study uses historically exposed CIFAR test images; fresh seeds do not undo that exposure.

Table 10: Normalization ablation, means over eight paired seeds. Final and Old (tasks 0–6) are final test accuracy; Acq. and Forget are immediate acquisition and subsequent loss on tasks 7–9. All entries are percent or points. Both factors have the same margin and initial learning rate.

E	Aggregation / reduction	Final	Old	Acq.	Forget
10	Top / batch	54.57	47.41	73.72	2.44
10	Top / active	55.10	47.80	75.09	2.97
10	HEM-derived / batch	54.55	47.58	74.23	3.42
10	HEM-derived / active	55.11	47.82	76.45	4.33
40	Top / batch	55.84	47.62	79.73	4.71
40	Top / active	53.79	45.20	81.83	8.00
40	HEM-derived / batch	55.38	47.14	80.09	5.48
40	HEM-derived / active	51.38	42.79	83.07	11.63

The E40 reduction effect is -3.023 [-3.850,-2.195], and the aggregation effect is -1.433 [-1.848,-1.017]; both Holm $p = .000111$. The reduction effect changes old-task accuracy by -3.382 [-4.378,-2.386], acquisition by $+2.540$ [2.316,2.763], and later-task forgetting by $+4.723$ [3.772,5.674]. The secondary interaction, the active-minus-batch difference for HEM-derived minus that for top-wrong, is -1.953 [-2.859,-1.046]. Thus the two factors are not simply additive. At E10 the reduction effect is $+0.546$ [0.334,0.758], while the aggregation effect is -0.005 [-0.259,0.249].

E40 minus E10 improves the top/batch and HEM-derived/batch cells by $+1.271$ [1.014,1.528] and $+0.834$ [0.517,1.151], but lowers top/active and HEM-derived/active by -1.308 [-2.054,-0.561] and -3.725 [-4.879,-2.571]. This is direct evidence for a budget-dependent effect of the reduction choice under this protocol. It does not establish a universal normalization disadvantage, an observed active-fraction mechanism, or the explanation for A/B's class-adjusted HEM comparison. Learning-rate-compensated, replay, other-margin, and other-dataset controls are not supplied by this ablation. The accompanying `recompute_normx.py` verifies the eight CSVs, common prefix signatures, vector-to-scalar metric identities, and all reported contrasts.

E Analytical Details and Scope

Proof of Observation 1. For a fixed batch and model, the gradient of each active hinge term is $-\nabla_\theta d_i$, while inactive terms contribute zero. Thus

$$\|\nabla_\theta L_{\mathrm{H}}\| \le \frac{1}{B}\sum_{i\in A(m)} \|\nabla_\theta d_i\| \le \frac{|A(m)|}{B} G = a(m)G. \tag{9}$$

Dividing by $|A(m)|$ instead yields the bound G when $A(m)$ is nonempty. If $m' \le m$, then $A(m') \subseteq A(m)$ at the same model, so $a(m') \le a(m)$. This last fact compares bounds with a common G, not actual norms or trained endpoints. At ties, a consistent subgradient convention is required; the empirical implementation uses the usual zero derivative at the ReLU boundary.

Cancellation prevents norm monotonicity. For example, suppose an active batch contains two scalar gap-gradient contributions $g_1 = 1$ and $g_2 = -1$. Its full-batch gradient is zero. If reducing the margin deactivates only the second sample, the full-batch gradient becomes $1/2$, even though the active fraction falls. Moreover, if two active gradients are non-collinear, removing one generally changes the direction of their sum. This establishes why neither a smaller gradient norm nor equivalence to learning-rate rescaling follows from a smaller margin alone.

Momentum, replay, and old-task stability. Consider an interval with no momentum reset, indexed by update $s = 0, \dots, T-1$. Let the trunk update be

$$v_s = \mu v_{s-1} + g_s + \alpha_s r_s, \qquad \theta_{s+1} = \theta_s - \eta_s v_s, \tag{10}$$

with $0 \le \mu < 1$, nonnegative η_s, α_s, and $\alpha_s = 0$ off replay events. Suppose $\|g_s\| \le G_s a_s$ and $\|r_s\| \le H_s$. Expanding the recursion and using triangle inequalities gives

$$\|\theta_T - \theta_0\| \le D_T := \sum_{s=0}^{T-1} \eta_s \left[\mu^{s+1}\|v_{-1}\| + \sum_{k=0}^{s} \mu^{s-k}(G_k a_k + \alpha_k H_k)\right]. \tag{11}$$

At a task boundary, the experiments reset momentum; one can apply this bound on each task interval with $v_{-1} = 0$ and add the displacement bounds. This construction makes no assumption that a_s follows the same trajectory under two margin policies.

For an old example x, define its true-class logit gap with the frozen old head as $q_x(\theta, b)$, where b denotes non-parameter state such as BatchNorm buffers. Suppose $q_x(\cdot, b_0)$ is L_x-Lipschitz on a region containing the relevant endpoints. Define the buffer term $\beta_x = |q_x(\theta_T, b_T) - q_x(\theta_T, b_0)|$. Then

$$q_x(\theta_T, b_T) \ge q_x(\theta_0, b_0) - L_x D_T - \beta_x. \tag{12}$$

Hence an initially correct prediction is guaranteed to remain correct if its initial positive gap exceeds $L_x D_T + \beta_x$. This sufficient condition is conditional on quantities we have not bounded empirically. It is not a claimed performance guarantee for the networks. In particular, $a_s = 0$ does not eliminate stored momentum, replay gradients, or buffer drift. A small pre-momentum gradient norm is likewise insufficient to establish a small realized update or a small function change. The measured displacements in Table 7 are realized values of $\|\theta_T - \theta_0\|$, not of the bound D_T.

Acquisition–forgetting identities. Equation 2 follows by subtracting $R_{t,K-1} = P_t - F_t$ between arms with identical P_t. For C, sum that identity over ten tasks. The task-0 acquisition difference is zero and $F_9 = 0$, so the sums of acquisition and forgetting differences can be written as nine times the corresponding means. Dividing by ten yields Equation 6. These exact identities do not identify a causal intermediate variable.

F Earlier Evidence and Alternative Variants

The central A/B/C studies were designed after exploratory and confirmatory work on the same project. We retain the relevant history here to avoid presenting the CIFAR hypotheses as arising without prior test exposure. The effects in this section are historical summaries; they are not pooled with the central studies or included in their multiplicity correction.

Earlier loss comparisons. An eight-seed CIFAR comparison without replay found hinge minus CE $+3.334$ [1.959, 4.709] and hinge minus HEM $+0.414$ [-0.006, 0.834]. A separate sixteen-seed replication found a small hinge–HEM difference, $+0.335$ [0.018, 0.652], with greater acquisition but lower old-task retention for hinge. Those selected policies used hinge E40 versus HEM E10. An earlier combination campaign found hinge+replay minus HEM+replay $+1.501$ [0.903, 2.100] at the same replay weight but again different epoch budgets. These observations motivated A's matched-budget design; they should not replace A/B as the main comparison.

An additional eight-seed rotated-CIFAR-10 condition used a fixed 5,000-image base pool, ten rotations spaced by 18°, and private task heads. It found hinge minus CE $+0.963$ [0.614, 1.312] and hinge minus HEM $+0.026$ [-0.076, 0.127], all at E10. The same base images recur across tasks, so this is a task-incremental domain-shift condition, not an additional disjoint-class benchmark. Earlier project work had already used its official test images.

Duration and the mean-matched flat control. A fixed hinge-margin/fixed-initial-rate duration comparison found E40 minus E10 $+1.030$ [0.585, 1.475] final accuracy, $+6.146$ [5.748, 6.543] acquisition, and $+2.404$ [1.733, 3.076] later-task forgetting (eight seeds). Longer training can therefore improve the final score by learning more while also forgetting more. Its changing cosine horizon means this is a duration-policy comparison.

For a geometric margin schedule, a mean-matched flat control retains $m_0 = 0.25$ on task 0 and uses

$$m_t^{\mathrm{flat}} = \overline{m}_r := \frac{1}{9}\sum_{j=1}^{9} 0.25(1-r)^j, \quad t = 1, \ldots, 9. \tag{13}$$

It matches the arithmetic mean over later tasks, not the active fraction, gradient direction, cumulative displacement, or learned function. This is a mechanism control for temporal allocation, not a necessary criterion for a useful policy versus a specified fixed baseline. In earlier eight-seed replications, 10% discount versus constant was $+3.130$ [1.153, 5.107], while versus its mean-matched flat control it was $+1.540$ [-0.653, 3.733]. For 5%, the corresponding effects were $+0.771$ [-1.554, 3.097] and $+0.435$ [-1.478, 2.348]. The earlier preregistered joint gates therefore failed. Those failures remain historical results; the wide 5% interval is compatible with C's later fixed-sample estimate. C establishes a different, narrower practical comparison and does not retrospectively validate a geometric-order mechanism.

Gradient-size and correct/wrong balancing variants. Small pilots also tested rescaling hinge's parameter gradient toward a CE reference and balancing aggregate contributions from correctly and incorrectly classified examples. The tested hinge variants did not improve their validation reference, so they were not included in the central policies. These outcomes apply to those implementations and small pilots, not to all possible balancing methods. Matching the norm of a raw parameter gradient is different from matching a post-momentum parameter step or the realized changes in correct-versus-wrong example margins. The latter causal quantities were not matched in the completed evidence here. Separate tiny CE/CUT pilots showed suggestive benefits only with the cutoff and were not confirmed. They do not support a general positive or negative claim about balancing.

What worked across variants. The completed evidence most strongly supports ordinary label replay as a broadly useful resource, the longer-budget hinge policy on CIFAR under A's comparison, and margin discounts over the specified constant baseline in C. Mean-image replay, layer-freezing variants, special gradient caps, and the small hinge balancing pilots do not establish additional contributions. A replay pilot comparing label replay with stored-output KL did not distinguish them; this is not an equivalence result or evidence that distillation is generally unnecessary. The central paper therefore does not combine promising means from separate protocols into an untested composite method.

G Compute and Evidence Integrity

The recorded A/B rental total is 14.78 GPU-hours, including setup and spares. C used 20.95 rented GPU-hours. Together the A/B/C campaigns account for 35.73 rented GPU-hours; this excludes the

separate normalization ablation and earlier project exploration. C's measured branch training-plus-evaluation time, adding the shared first task only once per seed, is 18.763 GPU-hours. Counting the same shared prefix once per CSV arm would overcount it. The mean C branch duration after task 0 is 202.3 seconds and the shared first task takes 22.0 seconds. All these timings are observed on the rented RTX 4090 configuration and include evaluation; they are not portable hardware-independent FLOP measurements. A finalized monetary invoice was not available in the audited records, so no dollar cost is asserted.

The A/B prefix costs average approximately 39.8 and 92.0 seconds per seed. At fixed budget the losses have similar measured suffix times; across CIFAR budgets the E40 hinge replay policy costs about 2.82 times its E10 suffix time, despite four times the new-data updates, because replay and other work also contribute. Comparing hinge E40 with HEM E10 is therefore an accuracy–compute trade-off, even when the final accuracy difference is positive.

For central confirmations, we checked 96 A rows, 48 B rows, and 320 C rows for expected seed/arm keys, uniqueness, finite results, and prescribed update counts. A seed's arms agree in their recorded pre-fork signature, and the branching code restores the full model state and RNG state. The scalar signature is a consistency check, not a cryptographic proof of checkpoint identity. The 32 confirmation CSVs were compared with their recorded MD5 checksums; transferred bytes were also verified and SHA-256 hashes are supplied for the analysis bundle. All 288 A and 72 B development rows have empty official-test columns. The accompanying `recompute.py` verifies the raw-file hashes, replays the finite-grid selection rule on the validation rows, and recomputes the central summaries from the same rows. Checksums establish file consistency; they do not remove test exposure, benchmark limitations, or potential implementation assumptions.

The normalization ablation adds 64 unique rows on eight seeds. All eight raw CSVs matched their collected MD5 checksums, and SHA-256 checksums are included in the anonymous package. Its measured training-plus-evaluation time is 0.784 GPU-hours, counting the shared prefix once per seed. This is not a rental total; setup, spares, and idle time are excluded, and no finalized rental invoice for this ablation is asserted.