Catastrophic forgetting in class-incremental learning is a verifier-exploitability problem

W. Schulz

Jun 27, 2026

Abstract

Freeze one near-perfect expert per task and all of class-incremental difficulty becomes selection — which expert to trust. Accuracy factorizes exactly as selection accuracy × within-task accuracy, and the selection term is pinned by the selector's exploitability ε, the same quantity that bounds test-time compute.

⚠ ε CORRECTION (2026-07-26)

Claims in this document about the ε law being a novel cross-domain law or a universality class are superseded. $\varepsilon(\sigma)=\Phi(-\text{gap}/(\sigma\sqrt{2}))$ is Thurstone's Law of Comparative Judgment, Case V (1927) — an identity, not a discovery — and it is not a universality class (it breaks 723× under heavy tails). The class-IL row is a proved theorem (Kim et al., NeurIPS 2022). The measured constants, the shipped tools, and the formula $\varepsilon^*=p/(1-p)$ all stand. Canonical correction: **CORRECTION-epsilon-thurstone.md**.

Preprint draft, 2026-06-27. Reproducible: `continual-learning/` (numpy, seed=0). Consolidates the run-once, pre-registered results in EXPERIMENTS-breaking-ground.md (H4, H8, H9) and the follow-on campaign EXPERIMENTS-selector-epsilon.md (SE1–SE7). Sister paper: When More Sampling Hurts, whose exploitability ε is imported here verbatim.

Abstract

Class-incremental learning (class-IL) — learn disjoint groups of classes in sequence, then classify across all of them with no task label at test time — is the regime where catastrophic forgetting is worst, and where regularization (EWC), gradient-orthogonality (GPM), and capacity-recruitment all collapse toward the finetune floor. We show *why*, with a single decomposition. Train one small expert per task on its own classes and freeze it; each expert is individually near-perfect (oracle-task accuracy 0.987 on Split-MNIST). All of class-IL difficulty is then which expert to trust — a cross-task *selection* problem. We prove empirically that class-IL accuracy factorizes exactly as `ACC = A_sel · p_within` (selection accuracy × within-task accuracy, ±0.01 on MNIST, ±0.05 across synthetic and hardened data), and that the selection term is pinned by the exploitability ε of the selector — the rate at which a wrong task's expert out-scores the correct one — exactly the quantity that bounds Best-of-N test-time compute. The consequences transfer wholesale: (i) you cannot out-vote exploitability — consensus over K experts saturates at $N^*\approx 2$ and never reaches the 0.987 ceiling; (ii) expert diversity does not help — decorrelating architectures leaves ε unmoved (0.388→0.397), because every expert reports the same exploitable

signal (its own confidence); (iii) the only lever is the feature space — moving the selector into drift-free streaming-PCA features drops ε 0.40→0.28 and explains *why* simple streaming features beat the strong GDumb baseline; and (iv) ε scales with task count, climbing 0.21→0.57 as tasks grow 2→10, which bounds selection-based CL to small task counts. The framework is substrate-robust (the factorization holds on hard data) but the strong claim "forgetting *is* selection" is bounded to the regime where experts are individually strong; when data is hard, representation quality re-binds (p_within 0.95→0.62). The practical rule mirrors the verifier paper: don't add experts or replay buffer — lower selector ε by improving the shared feature space.

1. Setup and the one modeling move

Split-MNIST class-IL: 5 tasks × 2 digits, single shared head, no task id at test (`cl/`, seed=0). The finetune baseline scores 0.169 (near the 5-task chance/recency floor); GDumb (balanced buffer + retrain) is the standing strong baseline at 0.859.

The modeling move: treat class-IL as selection. Train K experts per task, each a 2-class MLP on its own digits, then freeze. At test, an input is classified by (a) a *selector* that picks a task from the experts' outputs, then (b) the chosen task's experts voting on the class. Because experts are frozen, they never forget (BWT = +0.000) and stay near-perfect within task (plasticity 0.987). All difficulty is pushed into the selector.

Exploitability ε (imported from the verifier paper): the fraction of test inputs on which *some* wrong-task expert out-scores *every* correct-task expert under the selection signal. ε is to a selector what a Goodhart blind-spot is to a reward model.

2. The decomposition (H4, SE3, SE7)

On Split-MNIST the experts are near-perfect (oracle-task ceiling 0.987), so class-IL accuracy is governed entirely by selection. We measure, sweeping the number of experts K and across data hardness:

> `ACC ≈ A_sel · p_within`, holding to ±0.008 on MNIST, ±0.05 across synthetic separability sweeps and sample-starvation (SE7). The factorization is not an MNIST artifact.

with the selector's `A_sel` pinned near `1 − ε` (a conservative lower bound once consensus lifts it slightly above). The naive shared-softmax finetune is the maximally exploitable selector: 100% recency bias = ε→1. Reframing as explicit selection lifts accuracy 0.169 → 0.635 with no buffer at equal parameters, but ε ≈ 0.39 caps it well below GDumb. The 0.35 gap from 0.635 to the 0.987 ceiling is *entirely* selector exploitability.

3. What does and does not lower ε

Intervention	Result	ε	Verdict
Shared softmax (finetune)	recency-biased selector	→1	floor 0.169
Explicit per-task experts + consensus (H4)	selection reframing	0.388	0.635
More experts (K=1→15, SE3)	consensus voting	0.396	saturates at N*≈2, no gain

Intervention	Result	ε	Verdict
Diverse experts (varied arch + feature subspaces, SE1)	decorrelation	0.397	no change (+0.013 ACC)
Energy / OOD readout on same experts (SE2)	better verifier signal	0.385 / 0.963	no better (energy ≈ confidence)
Streaming-PCA shared features (SE5)	richer feature space	0.279	ε halves; ACC 0.635→0.788

Three negative results sharpen the mechanism. (SE3) You cannot out-vote ε. Consensus over experts saturates almost immediately (N*≈2) and plateaus at 1−ε; the verifier paper's "buy a better verifier, not more samples" holds verbatim for forgetting. (SE1) Diversity within a model family is not the lever — architecturally decorrelated experts leave ε unmoved, because they all report the same exploitable signal: their own max-softmax confidence, which a wrong-task expert emits just as confidently on out-of-distribution input. (SE2) No readout of the frozen experts' own outputs beats confidence — free-energy −`logsumexp` is monotone in the max logit (same information), and naive feature-distance OOD is far worse (ε 0.96). ε ≈ 0.39 is a property of overconfident-OOD experts, robust to the readout.

The one lever that works is the feature space (SE5). Moving the experts into a drift-free shared representation (incremental PCA over the full unlabeled stream) drops ε from 0.40 to 0.28 and lifts the selector to 0.788. This unifies with the independent H8 result: a buffer-free streaming-PCA + nearest- class-mean classifier beats GDumb (0.871 vs 0.859). H8 wins *because* good shared features make task identity easy — i.e. they drive the implicit selection ε down. (Notably, once features are good, a single shared metric space — global NCM — solves selection *better* than explicit per-task experts; the selector framing is the durable *diagnosis*, while the simplest method wins as *practice*.)

4. Recruitment is free; selection is the tax (SE4, H9)

Capacity-recruitment — give each task its own hidden block, freeze the rest — is widely proposed as a forgetting cure. It fails (H9: plasticity collapses to 0.33 on MNIST) for a reason the decomposition predicts: with a shared output head, every frozen block injects interference into the shared logits. Replace the shared head with the selector and recruitment is rescued exactly:

method (Split-MNIST class-IL)	ACC	BWT	plasticity
GPM(e=0.9)	0.235	—	—
Recruit (naive, shared head)	0.128	—	—
Recruit + selector	0.635	+0.000	0.987

Per-task capacity is cheap (zero forgetting, near-perfect plasticity); the entire class-IL gap is the selector. This unifies H4/H5/H9: catastrophic forgetting in shared-output class-IL is fundamentally a cross-task selection / exploitability problem, not a capacity or regularization problem.

5. Bounds (SE6, SE7)

The story has two honest limits. (SE6) ε scales with task count — at fixed per-expert difficulty, ε climbs monotonically 0.214→0.569 as the stream is split into 2→10 tasks, because each task adds another confuser expert that can out-confidence the correct one. Selection-based CL degrades ~linearly in T: it is a small-task-count tool. (SE7) The strong claim is regime-bounded — the factorization `ACC = A_sel·p_within` is substrate-robust (holds on synthetic and MNIST, easy and hard), but "all forgetting is selection" rests on experts being individually strong. As class separability falls, p_within drops 0.95→0.62 and representation quality re-binds as a co-bottleneck (ε rises in parallel, 0.43→0.78). On genuinely hard data, selection and representation degrade together.

6. Related work and the transferable claim

Class-IL's hardness has been attributed to feature drift and classifier (recency) bias; bias-correction methods address the head, replay addresses the features. We reframe both as one quantity: the exploitability ε of the cross-task selector, the same ε that bounds test-time-compute scaling in Best-of-N. This connects two literatures — continual learning and verifier-bounded inference — through a shared selection-ceiling law, and imports the verifier track's tooling (ε, consensus, N*) directly: a CL system's safe "expert budget" is N*(ε), and the way to spend surplus is a less-exploitable selector (better shared features), never more experts or buffer.

7. Reproducibility and honesty ledger

Every number is from a pre-registered, run-once experiment with a kill criterion; failures are reported in full. Of the seven follow-on hypotheses: SE3, SE4, SE7 confirmed; SE5, SE6 confirmed in mechanism with a stated threshold/loose-form miss; SE1, SE2 fired their kills (diversity and alternative readouts do not lower ε). The campaign did not produce a method beating GDumb via the explicit selector; its contribution is the diagnosis (forgetting = feature-controlled selection ε) and the unification of three prior results (H4/H8/H9) plus a sister inference-time literature under one law. Scripts: `continual-learning/experiments/selector_eps/se0{1..7}_*.py`, `experiments/breaking_ground/bg0{4,8,9}_*.py`.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/forgetting-as-selection