DATA SELECTION FOR CONTINUAL LEARNING: A THEORETICAL ANALYSIS OF SAMPLE SCORING UNDER TASK SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In continual learning, a training example that helps a model learn a new task may also interfere with knowledge acquired earlier. Yet most theories of data selection assume a stationary learning objective, leaving unclear how sample prioritization affects learning and forgetting across sequential tasks. We develop a theory of score-based data selection in a structured linear teacher–student model, showing how feature and readout similarity between tasks determine a scoring rule's performance relative to uniform sampling. We analyze two representative scoring rules based on feature novelty and residual error. Novelty consistently improves the trade-off between learning the new task and retaining the old one, whereas residual-based selection depends on the relationship between the tasks: high-error samples help when the tasks are sufficiently aligned but increase forgetting when shared features support conflicting outputs. We further show that this qualitative ordering extends from score-proportional reweighting to hard budgeted selection, while tighter selection amplifies the effect. Experiments on CIFAR-100, CUB-200, and AwA2 qualitatively support the predicted dependence on task similarity and show that hard selection can produce larger retention effects than reweighting, while nonlinear CIFAR-100 experiments show a similar score ordering.

1 INTRODUCTION

Continual learning requires a model to acquire new tasks without erasing what it has already learned. Most approaches address this stability-plasticity problem by controlling *how* the model is updated, for example by penalizing movement of important parameters (Kirkpatrick et al., 2017; Zenke et al., 2017) or by restricting updates to directions that interfere less with previous tasks (Farajtabar et al., 2020; Saha et al., 2021). A complementary question is which *data* the model should prioritize while learning a new task.

Data selection is well studied in single-task learning, where examples can differ substantially in their contribution to training (Paul et al., 2021; Sorscher et al., 2022). In continual learning, however, sample value is inherently sequential. An example from the current task may be useful for learning that task while also modifying directions that support previously learned ones. Scoring rules developed for single-task learning do not explicitly account for this trade-off, which depends on both the current task and what the model has already learned.

We study this interaction in a two-task teacher-student model, making task relationships explicit and analytically tractable (Lee et al., 2021; Evron et al., 2022; Hiratani, 2024). Building on the overparameterized linear framework of Hiratani (2024), we introduce a structured latent representation that separates each sample's features into shared directions, whose overlap determines input similarity across tasks, and task-specific directions that are uncorrelated across tasks. Feature overlap and readout similarity describe the relationship between tasks, allowing us to analyze how prioritizing task-2 examples affects task-1 retention at matched task-2 performance.

We examine two scoring rules that capture different notions of sample value. *Feature novelty* favors examples whose inputs have larger components orthogonal to the previous task's feature subspace, while *residual error* favors examples that the current model predicts poorly, closely related to hardness-based data selection (Paul et al., 2021). We first analyze score-proportional reweighting to characterize how each score changes the extent of learning along shared and task-specific directions under a constraint on weight changes. This yields a criterion for improved retention: at matched

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.

task-2 performance, a scoring rule helps when it leads the model to learn proportionally more along task-specific directions than along shared directions.

The two scores lead to different outcomes. Within the block approximation, novelty improves the learning-retention trade-off, whereas residual scoring depends on task similarity. When tasks share features but disagree in their outputs, prioritizing high-error samples increases forgetting; when tasks are sufficiently aligned, it improves retention relative to uniform sampling. Feature overlap and readout similarity define the boundary between these regimes.

We further show that the same qualitative conclusions extend to hard budgeted selection, with tighter budgets amplifying each scoring rule's effect on retention. This connects the reweighting analysis to the practical setting of retaining a fixed fraction of the highest-scoring training examples (Sorscher et al., 2022). Experiments with pretrained features on CIFAR-100, CUB-200, and AwA2 qualitatively support the predicted dependence on task similarity, including changes in the sign of the residual retention gap. A nonlinear CIFAR-100 classifier also recovers the conflict-regime score ordering.

2 Related Work

Continual learning theory. Analytically tractable models have been widely used to study transfer and forgetting in continual learning, including nonlinear teacher-student and statistical-mechanical analyses (Saad & Solla, 1995; Asanuma et al., 2021; Lee et al., 2021; Mori et al., 2025; Shan et al., 2026), overparameterized linear teacher-student models (Hiratani, 2024), and continual linear regression (Evron et al., 2022; Lin et al., 2023; Goldfarb et al., 2024). Prior work has clarified how task similarity, including similarity in feature representations and readouts, shapes transfer and interference between sequentially learned tasks (Goldfarb & Hand, 2023; Hiratani, 2024). We build on this framework but shift the focus from overall learning dynamics to sample-level reweighting.

Data selection in single-task learning. Early work showed that actively querying informative examples can improve sample efficiency (Seung et al., 1992). Sample efficiency also motivates subsequent work on active learning and coreset formation (Sener & Savarese, 2017; Mirzasoleiman et al., 2020). Modern data-selection methods similarly ask which examples from an available training set should be emphasized or retained. Some selection rules prioritize difficult or informative examples using loss, gradient norm, or forgetting events (Katharopoulos & Fleuret, 2018; Toneva et al., 2018; Paul et al., 2021), while others emphasize coverage in feature space or gradient space (Mirzasoleiman et al., 2020; Killamsetty et al., 2021). Theory has shown that difficulty-based pruning can substantially improve data efficiency (Sorscher et al., 2022). These studies largely consider sample value within a single task objective.

Data selection in continual learning. In continual learning, sample selection has been studied most extensively in experience replay. Past examples are selected using criteria such as gradient interference (Aljundi et al., 2019), approximation of class feature statistics (Rebuffi et al., 2017), or their contribution to stability and plasticity (Sun et al., 2022). Theory has further examined how replay should depend on task similarity (Mori et al., 2025). Active continual learning instead asks which incoming examples should be labeled under a limited annotation budget (Vu et al., 2024; Park et al., 2025). Selection of current-task training data has also been evaluated empirically across continual-learning methods (Yildirim et al., 2026). Online coreset selection (Yoon et al., 2021) combines selection with replay, training on and storing only current-task samples whose gradients are aligned with those of the replay buffer. Another related work on fine-tuning shows that upweighting examples with low pretrained-model loss can reduce forgetting of the pretraining task (Sanyal et al., 2025). These studies show that selection and reweighting can improve retention, but how their effects depend on the relationship between tasks remains less understood.

3 Set-up

Structured teacher-student model. We extend the linear teacher-student continual-learning framework of Hiratani (2024) by introducing a structured latent representation for two sequential tasks, $\mu \in \{1, 2\}$. A latent vector $\boldsymbol{s} \sim \mathcal{N}(0, I_{N_s})$ is mapped through task-specific Gaussian feature and readout matrices $A_\mu \in \mathbb{R}^{N_x \times N_s}$ and $B_\mu \in \mathbb{R}^{N_y \times N_s}$ to produce input-target pairs $\boldsymbol{x} = A_\mu \boldsymbol{s}$, $\boldsymbol{y}^* = B_\mu \boldsymbol{s}$ (e.g., $\boldsymbol{s}$ a bird's underlying attributes, $\boldsymbol{x}$ its image, $\boldsymbol{y}^*$ its species label or a queried trait).

A linear student $\boldsymbol{y} = W\boldsymbol{x}$, with $W \in \mathbb{R}^{N_y \times N_x}$, learns to match the teacher target $\boldsymbol{y}^*$. To study how sample selection interacts with task similarity, we distinguish broadly shared features (e.g., low-level or domain-general) from task-specific features encoded independently across tasks, allowing their relative contributions to vary across samples. A minimal setup capturing this partitions the latent into a shared and a task-specific block of equal dimension $d = N_s/2$, writing $\boldsymbol{s} = (\boldsymbol{s}_s, \boldsymbol{s}_p)$ [1] (Fig. 1a,b). The tasks are coupled on this partition through two similarity parameters: shared feature columns are correlated across tasks with coefficient ρ_a (task-specific columns are independent), and corresponding readout columns are correlated across tasks with coefficient ρ_b in both blocks (Appendix A.1).

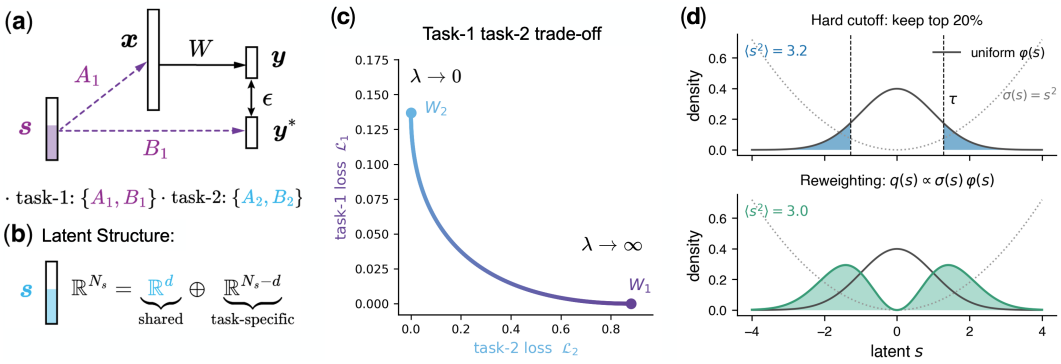

Figure 1: Model and sampling schemes. **(a)** Teacher-student setup: a latent $\boldsymbol{s}$ generates the input $\boldsymbol{x} = A_\mu \boldsymbol{s}$ and target $\boldsymbol{y}^* = B_\mu \boldsymbol{s}$, and a linear student predicts $\boldsymbol{y} = W\boldsymbol{x}$; task-1 is defined by the maps $\{A_1, B_1\}$ and task-2 by $\{A_2, B_2\}$. **(b)** Latent structure: $\mathbb{R}^{N_s}$ splits into a shared subspace $\mathbb{R}^d$ and a task-specific subspace $\mathbb{R}^{N_s - d}$. **(c)** Sweeping the penalty λ of Eq. (1) traces the task-1/task-2 trade-off between the endpoints W_2 ($\lambda \to 0$) and W_1 ($\lambda \to \infty$). **(d)** The two sampling schemes, illustrated on one latent direction with score $\sigma(s) = s^2$: a hard cutoff keeping the top 20% of scores, $|s| > \tau$ (top), and score-proportional reweighting, $q(s) \propto \sigma(s)\,\varphi(s)$ (bottom). Both concentrate mass at large $|s|$ and inflate the latent variance relative to uniform sampling ($\langle s^2 \rangle = 3.2$ and 3.0 vs. 1).

Sequential learning with constrained plasticity. We work in the over-parameterized regime with aspect ratio $\gamma := N_s/N_x \ll 1$. We take the infinite-sample limit to isolate the effect of sample selection from finite-sample fluctuations and characterize how selection changes the training distribution (Appendix A.7). The student is evaluated on each task by the population loss $\mathcal{L}_\mu[W] = N_y^{-1}\langle \|WA_\mu \boldsymbol{s} - B_\mu \boldsymbol{s}\|^2 \rangle$, where $\langle \cdot \rangle$ denotes the average over the latent $\boldsymbol{s} \sim \mathcal{N}(0, I_{N_s})$. For task-1, we take the minimum-norm solution $W_1 := B_1 A_1^\dagger \approx \gamma B_1 A_1^\top$, where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of A and the approximation uses $A_1^\top A_1 \approx \gamma^{-1} I_{N_s}$. Task-2 training updates the weights from W_1 to $W = W_1 + \Delta W$. Motivated by regularization-based continual-learning methods, we penalize the squared Euclidean distance from the task-1 weights (Kirkpatrick et al., 2017; Zenke et al., 2017; Li et al., 2018):

$$\Delta W_\lambda \;=\; \arg\min_{\Delta W} \; \frac{1}{N_y} \langle \|(W_1 + \Delta W)\, A_2 \boldsymbol{s} - B_2 \boldsymbol{s}\|^2 \rangle \;+\; \frac{\lambda}{N_y} \, \|\Delta W\|_F^2 \,. \tag{1}$$

We choose this simple isotropic weight penalty to focus on how data selection affects sequential learning. Sweeping $\lambda \in (0, \infty)$ traces a learning-retention trade-off between two endpoints: $\lambda \to \infty$ returns W_1 untouched, while $\lambda \to 0$ recovers the fixed point of full task-2 training, $W_2 = W_1(I_{N_x} - U_2 U_2^\top) + B_2 A_2^\dagger$, which replaces the components of W_1 along $\mathrm{col}(A_2)$ with the task-2 fit while preserving the orthogonal complement (Fig. 1c). Here, the columns of U_2 form an orthonormal basis for $\mathrm{col}(A_2)$. While these endpoints have been characterized (Hiratani, 2024), we study how selecting task-2 samples changes the learning-retention trade-off at intermediate penalties.

Selection and reweighting. We assign each task-2 sample a nonnegative score $\sigma(\boldsymbol{x}, \boldsymbol{y}^*) \geq 0$, computed before task-2 training and held fixed throughout. In practice, data selection often keeps the top-scoring samples under a fixed budget (Fig. 1d, top). The cutoff makes the selected distribution

[1] Equal blocks are used for clarity of presentation, as any shared-to-specific dimension ratio with both blocks nonempty leaves the main conclusions unchanged. The appendix keeps the shared fraction general.

harder to characterize analytically. We therefore begin with score-proportional reweighting (Fig. 1d, bottom), replacing the objective in Eq. (1) with

$$\Delta W_\lambda^\sigma = \arg\min_{\Delta W} \frac{\langle \sigma(\boldsymbol{x}, \boldsymbol{y}^*) \, \|(W_1 + \Delta W)\boldsymbol{x} - \boldsymbol{y}^*\|^2 \rangle}{N_y \, \langle \sigma(\boldsymbol{x}, \boldsymbol{y}^*) \rangle} + \frac{\lambda}{N_y} \|\Delta W\|_F^2 \,. \tag{2}$$

Normalizing scores by their population mean gives unit mean weight, and constant positive scores recover uniform sampling. This formulation lets us derive how a scoring rule changes the learned solution in closed form. We later show that the covariances induced by hard selection and reweighting can be approximated by a common family of exponential tilts with different strengths, extending the qualitative conclusions to hard selection.

Scoring for data value. We study two scoring rules that capture different notions of sample value. For retention, we ask which samples can be learned without disturbing the predictions of W_1 on task-1 inputs. A gradient step on a single sample changes the weights only along its input direction $\boldsymbol{x}$, so a sample with $\boldsymbol{x} \in \mathrm{col}(A_1)^\perp$ leaves task-1 predictions unchanged. The *feature novelty* score measures how much of the sample lies in this subspace. Let $P_1 := A_1(A_1^\top A_1)^{-1}A_1^\top$ denote the orthogonal projector onto $\mathrm{col}(A_1) \subset \mathbb{R}^{N_x}$. We define

$$\sigma_{\mathrm{nov}}(\boldsymbol{x}) \;:=\; \|(I_{N_x} - P_1)\,\boldsymbol{x}\|^2 \,. \tag{3}$$

Selecting high-novelty samples shares the motivation of gradient projection methods that restrict updates along directions important to previous tasks (Farajtabar et al., 2020; Chaudhry et al., 2020; Saha et al., 2021).

For task-2 learning, we consider prioritizing samples that the task-1 student predicts poorly. The *residual error* score is the squared error of W_1 on the sample,

$$\sigma_{\mathrm{res}}(\boldsymbol{x}, \boldsymbol{y}^*) \;:=\; \|W_1\boldsymbol{x} - \boldsymbol{y}^*\|^2 \,. \tag{4}$$

High-residual samples contribute most to the task-2 loss before training, motivating their prioritization as in hard-example selection in single-task data pruning (Paul et al., 2021; Sorscher et al., 2022).

4 Theory: score-proportional reweighting

Task performance trade-off. We first characterize the trade-off under uniform sampling, then examine how score-based reweighting of task-2 samples changes it. Solving Eq. (1) at leading order gives $\Delta W_\lambda \approx p\,\Delta W_{\lambda=0}$, where $\Delta W_{\lambda=0} = W_2 - W_1$ is the full task-2 correction and the gain $p := \gamma^{-1}/(\gamma^{-1}+\lambda) \in (0,1)$ (Appendix A.2). On task-2 inputs, this gives $W\boldsymbol{x} = (1-p)W_1\boldsymbol{x} + p\,\boldsymbol{y}^*$, so p directly measures the fraction of the task-2 error corrected by learning. As $\lambda \to \infty$, $p \to 0$ and the student remains at the task-1 solution, whereas as $\lambda \to 0$, $p \to 1$ and it approaches the task-2 solution W_2 (Fig. 2a).

To connect task-2 learning to retention, let p_s and p_p denote the gains on the shared and task-specific latent blocks. Under uniform sampling, $p_s = p_p = p$. Under the leading-order block approximation, the population test losses are

$$\mathcal{L}_2 \;=\; \tfrac{1}{2}\,\eta_s\,(1-p_s)^2 + \tfrac{1}{2}\,(1-p_p)^2, \qquad \mathcal{L}_1 \;=\; \tfrac{1}{2}\,\rho_a^2\,\eta_s\,p_s^2, \tag{5}$$

where $\eta_s := 1 - 2\rho_a\rho_b + \rho_a^2$ is the error coefficient for shared directions, while task-specific directions have coefficient one (Appendix A.3). Learning shared directions reduces task-2 error but changes predictions on task-1. Learning task-specific directions reduces task-2 error without contributing to forgetting at this order. Thus, the allocation of learning between the two blocks determines the retention-learning tradeoff.

Reweighting changes the block gains. Since inputs and targets are linear in $\boldsymbol{s}$, the reweighted objective in Eq. (2) depends on the latent distribution only through its covariance $C = \frac{\langle \sigma(A_2\boldsymbol{s}, B_2\boldsymbol{s})\,\boldsymbol{s}\boldsymbol{s}^\top \rangle}{\langle \sigma(A_2\boldsymbol{s}, B_2\boldsymbol{s}) \rangle}$. For $\gamma \ll 1$, and additionally $N_y \gg N_s$ for residual scoring, the block approximation gives $C \approx \mathrm{diag}(c_s I_d,\, c_p I_{N_s - d})$ with $c_s, c_p \geq 1$ (Appendix A.4). Reweighting can inflate the two blocks unequally. The population losses retain the form of Eq. (5), but the gains become block-dependent (Appendix A.5):

$$p_s \;=\; \frac{c_s\,\gamma^{-1}}{c_s\,\gamma^{-1} + \lambda}; \qquad\qquad p_p \;=\; \frac{c_p\,\gamma^{-1}}{c_p\,\gamma^{-1} + \lambda}. \tag{6}$$

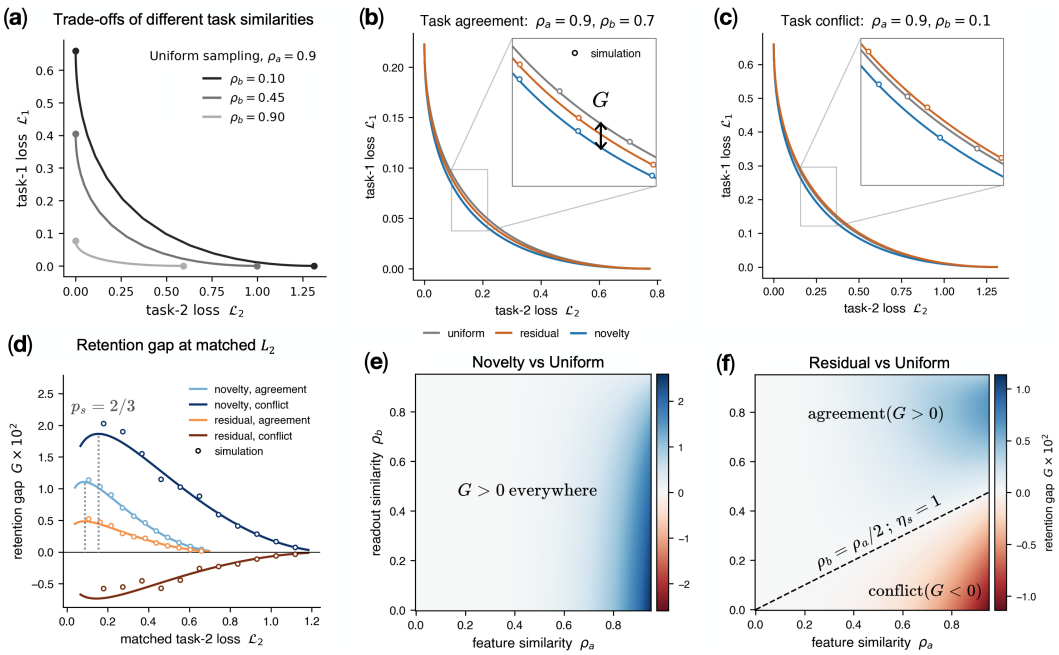

Figure 2: Score-proportional reweighting changes the retention-learning trade-off according to task similarity. **(a)** Trade-off curves under uniform sampling at feature similarity $\rho_a = 0.9$ and readout similarity $\rho_b \in \{0.10, 0.45, 0.90\}$. **(b)** Under task agreement ($\rho_a = 0.9$, $\rho_b = 0.7$), novelty and residual reweighting both reduce forgetting at matched task-2 loss. **(c)** Under task conflict ($\rho_a = 0.9$, $\rho_b = 0.1$), novelty improves retention while residual reweighting increases forgetting. Lines: theory; circles: simulation. **(d)** Retention gap G at matched task-2 loss for novelty and residual reweighting under the agreement and conflict task pairs of (b, c): the gap vanishes at both endpoints and is largest at intermediate gains, at $p_s = 2/3$ (gray dotted line). **(e, f)** Retention gaps across the task plane (ρ_a, ρ_b). Novelty improves retention in all nondegenerate task pairs, with larger gains at strong feature overlap and low readout similarity. Residual changes sign at the boundary $\rho_b = \rho_a/2$ (dashed line), improving retention under agreement and increasing forgetting under conflict. See Appendix A.9 for simulation details.

For score-proportional reweighting, the variance inflation in each latent direction is $c - 1 = O(1/N_s)$. To compare a scoring rule with uniform sampling, define the retention gap at matched task-2 error, $G(\mathcal{L}_2) := \mathcal{L}_1^{\text{unif}}(\mathcal{L}_2) - \mathcal{L}_1^{\text{score}}(\mathcal{L}_2)$. The gap is positive when the score forgets less than uniform. We find that for every task pair with $\rho_a^2 \eta_s > 0$ and every $\lambda > 0$ (Appendix A.6):

$$\operatorname{sign} G = \operatorname{sign}(c_p - c_s). \tag{7}$$

When $c_p > c_s$, task-specific directions are learned more strongly than shared directions. The student can therefore match uniform sampling's task-2 performance with less learning along shared directions, and hence less forgetting. When $c_p < c_s$, the reverse holds. This ordering holds for every $\lambda > 0$, which changes the gap's magnitude but not its sign.

Task similarity determines the effect of each scoring rule. To determine the sign of G for each score, we compare the extra variance it assigns to shared and task-specific directions. Since $c - 1 > 0$ for both scores, the condition $c_p > c_s$ is equivalent to an excess-variance ratio $(\frac{c_s - 1}{c_p - 1})$ below one. These ratios are (Appendix A.4)

$$\frac{c_s^{\text{nov}} - 1}{c_p^{\text{nov}} - 1} = 1 - \rho_a^2; \qquad \frac{c_s^{\text{res}} - 1}{c_p^{\text{res}} - 1} = \eta_s = 1 - 2\rho_a\rho_b + \rho_a^2. \tag{8}$$

For novelty reweighting, $1 - \rho_a^2 \leq 1$, with strict inequality whenever $\rho_a > 0$, independent of ρ_b; hence $G_{\text{nov}} \geq 0$ throughout the task plane. Residual reweighting changes its effect at $\rho_b = \rho_a/2$. Below this boundary, shared directions have larger residual error and receive greater weight, increasing forgetting relative to uniform sampling at matched task-2 performance (Fig. 2c). Above it, task-

specific directions receive greater weight, improving retention (Fig. 2b). At the boundary, $\eta_s = 1$ and $c_s = c_p$, so the trade-off curve coincides with that of uniform sampling (Fig. 2f).

To characterize the magnitude of the gap, we factorize G at leading order into score, gain, and task-pair factors (Appendix A.6). The score factor determines its sign, while the gain factor $(1 - p_s)p_s^2$, with p_s the uniform-sampling gain at matched task-2 loss, vanishes at both endpoints and peaks at $p_s = 2/3$ (Fig. 2d). The task-pair factor grows with feature overlap ρ_a^2 and shared-direction error η_s: overlap makes updates more disruptive to task-1, while readout disagreement increases the required correction. Together, these factors predict the largest positive novelty gaps and most negative residual gaps where features strongly overlap and readouts disagree (Fig. 2e,f). In this conflict regime, downweighting high-residual examples reduces emphasis on interfering directions, suggesting a possible mechanism for the improved retention from favoring low-loss examples during fine-tuning (Sanyal et al., 2025).

5 Extension to hard selection

Reweighting and hard selection are approximated by a family of tilts. Hard selection retains the fraction $B \in (0, 1)$ of samples with the highest scores, corresponding to latent vectors satisfying $\sigma(\boldsymbol{s}) > \tau_B$, where τ_B is the score's upper-B quantile. The original latent distribution is $\boldsymbol{s} \sim \mathcal{N}(0, I_{N_s})$, with density $\varphi(\boldsymbol{s})$. To approximate the covariances induced by reweighting and hard selection, we introduce a one-parameter family of latent probability densities (Appendix B):

$$q_\theta(\boldsymbol{s}) \;\propto\; e^{\theta\,\sigma(\boldsymbol{s})}\,\varphi(\boldsymbol{s}), \tag{9}$$

where φ is the standard Gaussian density and $\theta \geq 0$ controls how strongly high-scoring samples are favored. This exponential tilt increases the variance along score-weighted directions. Under the same block approximation, its covariance determines the shared and task-specific gains through Eq. (6). At the level of this covariance approximation, reweighting and hard selection are represented by the same exponential-tilt family with different strengths. At leading order,

$$\theta_{\mathrm{rw}} \;=\; \frac{1}{\langle\sigma\rangle}\,, \qquad\qquad \theta(B) \;=\; \frac{\varphi(z_B)/B}{\sqrt{\mathrm{Var}(\sigma)}}\,, \tag{10}$$

where $\langle\sigma\rangle$ and $\mathrm{Var}(\sigma)$ are the mean and variance of the score over the full population, and z_B is the standard normal upper-B quantile. For hard selection, we approximate the survival probability conditional on one latent coordinate using a Gaussian tail. Linearizing its logarithm yields an exponential weighting of that coordinate's score contribution, leading to the covariance approximation associated with Eq. (9). The tilt strength is proportional to the selection intensity $\varphi(z_B)/B$. Error estimates and validity conditions are discussed in Appendices B.5-B.6.

Hard selection and reweighting produce retention gaps of the same sign because they favor the same block (Fig. 3a,b). Within the tilt approximation's validity regime, hard selection is typically stronger: at a fixed non-extreme budget, $\theta(B)/\theta_{\mathrm{rw}} = O(\sqrt{N_s})$, and tighter budgets further increase the tilt strength (Appendix B.8). Selection can therefore amplify the retention effects predicted from task similarity (Fig. 3c).

6 CIFAR-100 in the task-conflict regime

This section and the next test the theory's qualitative predictions with empirical features and finite data. The main experiments use small output spaces, outside the regime assumed by the residual-score block approximation. Within the theoretical model, we verify numerically that smaller output dimensions still recover the broad retention ordering (Appendix A.8). We also provide larger-output experiments as complementary results closer to the assumed regime (Appendices C and D).

CIFAR-100 task pairs lie in the conflict regime. We first test the predictions on CIFAR-100, where task pairs lie in the conflict regime, $\rho_b < \rho_a/2$: novelty is predicted to improve the retention-learning trade-off and residual scoring to worsen it. We test this ordering under reweighting and hard selection, and whether tighter selection budgets amplify the effects. To obtain fixed feature representations for the linear model, we embed CIFAR-100 images (Krizhevsky et al., 2009) using a frozen pretrained ViT-B/16 backbone (Dosovitskiy et al., 2020; Steiner et al., 2021), yielding $\phi(\boldsymbol{x}) \in \mathbb{R}^{768}$. For task-1,

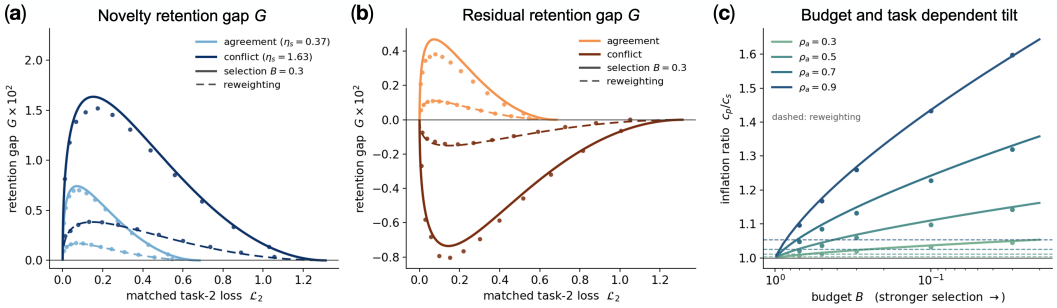

Figure 3: Hard selection preserves the predicted score ordering and amplifies the retention gaps at the displayed budget. **(a, b)** Retention gaps at matched task-2 loss for novelty (a) and residual scoring (b), under hard selection at $B = 0.3$ (solid) and score-proportional reweighting (dashed). Both preserve the same sign for the agreement ($\eta_s = 0.37$) and conflict ($\eta_s = 1.63$) pairs. Dots: simulation; solid curves use the exponential-tilt covariance with moment-matched strength (Appendix B.7). **(c)** Block-inflation ratio c_p/c_s of the novelty score as a function of the budget B (log axis; smaller budget = stronger selection), for $\rho_a \in \{0.3, 0.5, 0.7, 0.9\}$: tightening the budget pushes the tilt monotonically further from uniform sampling, and the effect grows with feature similarity. Solid: selection theory; dashed: reweighting; dots: simulation. See Appendix A.9 for simulation details.

we take $V_1 \in \mathbb{R}^{768\times k}$ ($k = 20$) to be the top-k PC basis of its features and define $P_\perp := I - V_1 V_1^\top$ to project onto directions orthogonal to the estimated task-1 feature subspace, with k serving as a proxy for the latent dimension (Appendix C.1).

Each task classifies five classes, and the two tasks share an output head $W \in \mathbb{R}^{5\times 768}$, with corresponding class positions mapped to the same output rows. Appendix C.4 reports experiments with 50-class outputs. We draw 50 pairs of tasks with disjoint classes. The estimated similarities (Appendix C.1) occupy a narrow conflict regime: $\hat{\rho}_a$ concentrates around 0.7-0.8, whereas $\hat{\rho}_b$ remains near zero (random remapping of output classes), so $\hat{\rho}_b < \hat{\rho}_a/2$ for all pairs.

Empirical score reweighting and retention gap. We fit the task-1 head W_1 by ridge-regularized least squares and train task-2 using the finite-sample weighted analogue of Eq. (1), sweeping λ to trace the empirical retention-learning trade-off (Appendix C.1, C.2). The two scores are adapted to empirical feature space:

$$\sigma_{\mathrm{nov}}(\boldsymbol{x}) = \frac{\|P_\perp\,\phi(\boldsymbol{x})\|^2}{\|\phi(\boldsymbol{x})\|^2}, \qquad \sigma_{\mathrm{res}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\|W_1\,\phi(\boldsymbol{x}) - \boldsymbol{y}\|^2}{\|\phi(\boldsymbol{x})\|^2}. \tag{11}$$

Both scores are normalized by squared feature norm to reduce their dependence on input magnitude. Score-proportional reweighting assigns each task-2 sample a score-proportional weight, normalized to mean 1 as under uniform sampling. Hard selection instead keeps the top BN task-2 samples by score, with $B \in \{0.7, 0.5, 0.3\}$ (total $N = 2500$), and is compared with a uniformly sampled subset of the same size.

Let ϵ_1 and ϵ_2 denote task-1 and task-2 test MSE. At matched ϵ_2, we measure the relative retention gap $G_{\mathrm{rel}} = \frac{\epsilon_1^{\mathrm{unif}} - \epsilon_1^{\mathrm{score}}}{\epsilon_1^{\mathrm{unif}}}$, where positive values indicate better retention than uniform sampling.

The predicted ordering holds, and hard selection partly amplifies it. Under reweighting, novelty improves retention and residual worsens it at matched task-2 error, both for individual task pairs (Figs. 4a, 9a) and on average across 50 pairs throughout the reliable mid-range (Fig. 4b). Hard selection at $B = 70\%$ amplifies residual's disadvantage and novelty's advantage at lower task-2 errors, although novelty's mean gap approaches zero and becomes slightly negative at higher errors. We also evaluate selection without retuning regularization. We fix λ across budgets and compare both tasks' errors with random subsets of the same size. This complements the matched-task-2-error comparison by allowing both retention and new-task performance to vary. Tightening the budget from $B = 70\%$ to 30% increases novelty's retention advantage (Fig. 4c). Across these budgets, novelty's mean relative reduction in task-1 error also exceeds its mean relative increase in task-2 error compared to uniform selection (Fig. 9b).

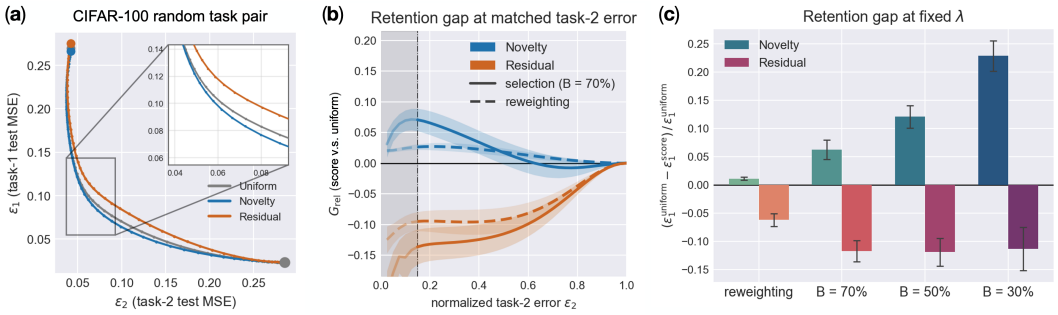

Figure 4: The predicted score ordering holds under reweighting on CIFAR-100 and is partly amplified by hard selection. **(a)** Trade-off curves ϵ_1 vs. ϵ_2 for one task pair under uniform, novelty, and residual score-proportional reweighting. Novelty lies below uniform and residual above, as predicted for the conflict regime; markers denote the $\lambda \to \infty$ (W_1) and $\lambda \to 0$ endpoints. **(b)** Matched-error retention gap G_{rel}, averaged over 50 task pairs with 95% CI bands, for score-proportional reweighting (dashed) and hard selection at $B = 70\%$ (solid). Hard selection makes the residual retention gap more negative: novelty's advantage increases in the lower-error part of the matched range but approaches zero or becomes slightly negative at higher normalized task-2 error. Shaded: regions where task-2 error matching is unreliable (Appendix C.3). **(c)** Task-1 error advantage over uniform at fixed λ (Appendix C.4), from reweighting through hard-selection budgets $B \in \{70\%, 50\%, 30\%\}$. Tightening the budget increases novelty's advantage, consistent with the budget dependence of Eq. (10), whereas residual's disadvantage changes little across the three hard-selection budgets. Error bars: 95% CI over task pairs.

Finally, we replace the linear ridge solve with sequential gradient training of a two-layer MLP to test whether the score ordering persists under nonlinear training. Across budgets $B \in \{70\%, 50\%, 30\%\}$, novelty achieves the highest task-1 accuracy and residual performs worse than random selection at the same budget, while all three rules reach comparable task-2 accuracy. The separation is largest at $B = 30\%$ (Appendix C.5).

7 Experiments spanning the task-similarity plane

CIFAR-100 classification places all task pairs in a narrow conflict regime, so it does not test the theory across the broader task-similarity plane. To examine a broader range of feature and readout similarities, we turn to attribute regression. Here, readout similarity reflects how attribute mappings transfer across tasks and can be positive, rather than being centered near zero by random class remapping. We use CUB-200 (Wah et al., 2011) and AwA2 (Xian et al., 2018), with task constructions that cover broader parts of the (ρ_a, ρ_b) plane (Fig. 5a). Both use the frozen-feature pipeline as CIFAR-100 classification, with a ViT-S/16 backbone and PCA to 100 dimensions (Appendix D.1).

CUB-200 varies readout similarity at high feature overlap. CUB-200 contains 200 bird species with approximately 30 training images per class, providing a sparse finite-data test of the population-level prediction. Each task contains 10 species and regresses onto a different group of annotated attributes, such as back color for task-1 and crown color for task-2 (Appendix D.2). Because all classes are birds, the estimated feature similarity $\hat{\rho}_a$ remains high. Changing the pair of attribute groups instead varies $\hat{\rho}_b$, producing task pairs predominantly in the conflict regime (Fig. 5a).

AwA2 extends into the agreement regime. AwA2 contains 50 more visually diverse animal classes. Tasks use disjoint class subsets but regress onto the same eight attributes, fixed across all pairs and chosen for readout transferability. These attributes allow high readout similarity across tasks. To broaden the range of readout similarities without changing feature similarity, we permute attribute assignments across varying subsets of task-2 output dimensions (Appendix D.2). The resulting task pairs span $\hat{\rho}_a \approx 0.6$-0.8 and cover readout similarities ranging from low values to $\hat{\rho}_b \approx 1$, including pairs in the predicted agreement regime.

The retention gap broadly follows task similarity. For both datasets, we compute G_{rel} at matched task-2 test error under score-proportional reweighting. We average the gap over the reliably matched

mid-range of each trade-off curve, then average across task pairs within bins on the estimated $(\hat{\rho}_a, \hat{\rho}_b)$ plane (Appendix D.2).

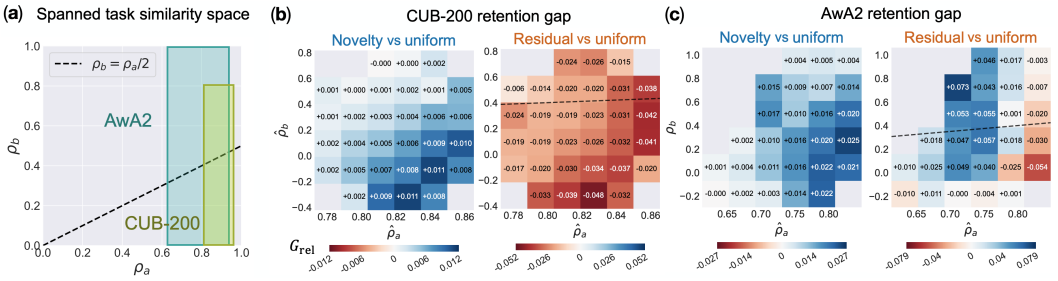

Figure 5: Attribute-regression tasks cover a broader range of task similarities and qualitatively support the predicted retention gap structure. **(a)** Approximate regions of the estimated task-similarity plane $(\hat{\rho}_a, \hat{\rho}_b)$ covered by CUB-200 and AwA2 task pairs; dashed: the predicted boundary $\rho_b = \rho_a/2$. AwA2 spans a broader range. **(b)** CUB-200 retention gap at matched task-2 error under score-proportional reweighting, for novelty (left) and residual (right), binned on the estimated $(\hat{\rho}_a, \hat{\rho}_b)$ plane. Novelty's retention gains and residual's retention losses are largest in the high-$\hat{\rho}_a$, low-$\hat{\rho}_b$ region. **(c)** AwA2 covers a wider region of the task plane. The residual gap changes sign, with positive gaps in the agreement regime.

The observed structure broadly follows the theory (Fig. 5b,c). Novelty gains are concentrated where feature overlap is strong and readout similarity is low, where residual reweighting also shows substantial forgetting. On AwA2, the residual gap transitions between negative and positive values across the task plane, qualitatively consistent with the predicted dependence on task similarity, although the transition does not coincide with $\hat{\rho}_b = \hat{\rho}_a/2$. CUB-200 residual gaps remain negative above this estimated boundary. Differences between the empirical representations and the model assumptions, together with the approximate mapping from measured similarities to model parameters, may contribute to these deviations. Similar qualitative patterns persist under hard selection across budgets from $B = 70\%$ to 30% (Fig. 12) and in larger-output experiments closer to the dimensional regime of the residual analysis (Figs. 13 and 14).

8 Discussion

Using two scoring rules, our framework explains how prioritizing training examples affects learning and retention depending on the relationship between successive tasks. In the model, each rule's effect is determined by how it allocates learning between shared and task-specific directions at matched new-task performance. Feature novelty consistently favors directions that can be learned with less interference and therefore improves the retention-learning trade-off. The effect of residual scoring depends on task similarity: prioritizing high-error samples helps when the two tasks are sufficiently aligned, but increases forgetting when shared features support conflicting outputs (Fig. 2). Within the block approximation, hard selection preserves this qualitative ordering, and the exponential-tilt approximation further predicts stronger effects as the selection budget tightens (Fig. 3). Our experiments qualitatively support the predicted dependence on scoring rule and task similarity, suggesting that the theory helps explain selection effects beyond the linear model (Sections 6-7). These results suggest that selecting training samples with task relationships in mind is a factor worth considering in sequential learning.

Limitations. Our theory assumes fixed feature representations to isolate how data selection affects learning and retention. Extending the analysis to jointly evolving representations remains an open direction for end-to-end continual learning. In our model, both shared and task-specific features contribute to the target outputs. A natural extension is to study domain shifts in which the prediction task remains the same and domain-specific features affect the inputs but not the targets. We also leave the effects of input and label noise, and interactions with other continual-learning mechanisms, for future study.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability and to identify and summarize related literature. We also used them to assist with preparing datasets and with changing experimental settings. For the theoretical results, the authors formulated the main conjectures; generative AI tools assisted with algebraic derivations, and proposed the exponential-family approximation used in Section 5. We did not use generative AI tools to originate the research question or the paper's central claims, or to generate the manuscript text. Synthetic data generation and translation are not applicable to this work. AI-assisted derivations were independently verified by the authors, and AI-assisted code was reviewed before use. We take responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Code to reproduce the theoretical results and the experiments in Figs. 2-4 is available at an anonymized repository: `https://anonymous.4open.science/r/DS4CL-1EB6`. Full proofs of the theoretical results are provided in Appendices A and B, and experimental details are given in Appendices C and D. Code for the Section 7 experiments and the nonlinear training on CIFAR-100 (Section 6) will be added to the same repository.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021.

Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33:9900–9911, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pp. 4028–4079. PMLR, 2022.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3762–3773. PMLR, 2020.

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pp. 2975–2993. PMLR, 2023.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting–an analytical model. *arXiv preprint arXiv:2401.12617*, 2024.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. *Advances in Neural Information Processing Systems*, 37:3243–3274, 2024.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pp. 2525–2534. PMLR, 2018.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*. PMLR, 2018.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, pp. 21078–21100. PMLR, 2023.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084004, 2025.

Jaehyun Park, Dongmin Park, and Jae-Gil Lee. Active learning for continual learning: Keeping the past alive in the present. *arXiv preprint arXiv:2501.14278*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

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*, pp. 2001–2010, 2017.

David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. Upweighting easy samples in fine-tuning mitigates forgetting. *arXiv preprint arXiv:2502.02797*, 2025.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Haozhe Shan, Qianyi Li, and Haim Sompolinsky. Order parameters and phase transitions of continual learning in deep neural networks. *Proceedings of the National Academy of Sciences*, 123(6): e2501899123, 2026.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Qing Sun, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Exploring example influence in continual learning. *Advances in Neural Information Processing Systems*, 35:27075–27086, 2022.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Thuy-Trang Vu, Shahram Khadivi, Mahsa Ghorbanali, Dinh Phung, and Gholamreza Haffari. Active continual learning: On balancing knowledge retention and learnability. In *Australasian Joint Conference on Artificial Intelligence*, pp. 137–150. Springer, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Elif Ceren Gok Yildirim, Murat Onur Yildirim, and Joaquin Vanschoren. Coresets are more than replay: a data-centric view of continual learning. *Neural Computing and Applications*, 38(13):559, 2026.

Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085*, 2021.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

A Derivation for reweighting analysis

Throughout the appendix, "at leading order" refers to the strongly over-parameterized regime, $N_x \to \infty$ at fixed and small aspect ratio $\gamma = N_s/N_x$, and it bundles two error scales that we keep distinct. Entrywise statements, such as Gram entries, cross-overlaps, score-matrix entries, and column norms, carry multiplicative $1 + O(N_x^{-1/2})$ corrections from concentration of random inner products; these vanish as $N_x \to \infty$ at fixed γ. In our calculation, we also replace the full Gram matrix by $\gamma^{-1} I_{N_s}$, particularly inside an inverse such as $(A_2^\top A_2 + \lambda I)^{-1}$, neglecting the variation across its eigenvalues. Their relative spread is of order $\sqrt{\gamma}$, so this approximation requires small γ; its error does not vanish as $N_x \to \infty$ at fixed γ. The block approximation for the residual score requires additional dimensional assumptions, which we discuss in Appendix A.4. Throughout this appendix, we retain a general shared fraction $\alpha := d/N_s \in (0, 1)$; the main text uses equal blocks, $\alpha = 1/2$.

A.1 Model conventions and concentration facts

Coupling construction. Writing $\boldsymbol{a}_{\mu,i}$ and $\boldsymbol{b}_{\mu,i}$ for the i-th columns of A_μ and B_μ (the feature and readout signatures of latent direction i), the task-1 matrices have i.i.d. $\mathcal{N}(0, N_s^{-1})$ entries, and task-2 is coupled to task-1 by

$$\boldsymbol{a}_{2,i} = \rho_a \, \boldsymbol{a}_{1,i} + \sqrt{1-\rho_a^2}\, \tilde{\boldsymbol{a}}_{2,i} \quad (i \le d), \qquad \boldsymbol{b}_{2,i} = \rho_b \, \boldsymbol{b}_{1,i} + \sqrt{1-\rho_b^2}\, \tilde{\boldsymbol{b}}_{2,i} \quad (\text{all } i), \tag{12}$$

with task-specific feature columns ($i > d$) drawn independently and all fresh entries i.i.d. $\mathcal{N}(0, N_s^{-1})$.

Preliminary identities. Two facts are used repeatedly below. First, latent averages of squared linear maps are Frobenius norms: for any matrix X, $\langle \|X\boldsymbol{s}\|^2 \rangle = \operatorname{tr}\left(X^\top X \langle \boldsymbol{s}\boldsymbol{s}^\top \rangle\right) = \|X\|_F^2$, since $\|X\boldsymbol{s}\|^2$ equals its own trace and $\langle \boldsymbol{s}\boldsymbol{s}^\top \rangle = I_{N_s}$. This converts every population loss into a Frobenius norm. Second, the random feature maps concentrate: a column has $\langle \|\boldsymbol{a}_{\mu,i}\|^2 \rangle_A = N_x/N_s = \gamma^{-1}$ with relative fluctuation $O(N_x^{-1/2})$, while inner products of independent columns are zero-mean sums of N_x terms, of size $O(\sqrt{N_x}/N_s)$, smaller than the diagonal entries by the same relative factor, so the Gram matrix concentrates,

$$A_\mu^\top A_\mu \;\approx\; \gamma^{-1} I_{N_s}, \tag{13}$$

and, applying the same argument to Eq. (12) entrywise ($\boldsymbol{a}_{2,i}^\top \boldsymbol{a}_{1,i} = \rho_a \|\boldsymbol{a}_{1,i}\|^2 + \sqrt{1-\rho_a^2}\, \tilde{\boldsymbol{a}}_{2,i}^\top \boldsymbol{a}_{1,i}$, the second term fresh and hence subleading), so does the cross-overlap,

$$A_2^\top A_1 \;\approx\; \gamma^{-1} \operatorname{diag}(\rho_i), \qquad \rho_i = \rho_a \text{ (shared)}, \quad \rho_i = 0 \text{ (specific)}. \tag{14}$$

Task-1 solution and the error matrix. Training on task-1 minimizes the population loss $\mathcal{L}_1[W] = N_y^{-1} \langle \|WA_1\boldsymbol{s} - B_1\boldsymbol{s}\|^2 \rangle = N_y^{-1} \|WA_1 - B_1\|_F^2$. Since $N_x > N_s$, A_1 has full column rank almost surely and the loss attains zero on the solution set of $WA_1 = B_1$; the minimum-norm solution of this system is given by the pseudoinverse, $W_1 = B_1 A_1^\dagger$ with $A_1^\dagger = (A_1^\top A_1)^{-1} A_1^\top$, so that $W_1 A_1 = B_1$ exactly and $\mathcal{L}_1[W_1] = 0$. By Eq. (13), $(A_1^\top A_1)^{-1} \approx \gamma I_{N_s}$, giving the main-text expression $W_1 \approx \gamma B_1 A_1^\top$. The student then enters task-2 carrying the error matrix

$$\mathcal{E} \;:=\; B_2 - W_1 A_2 \;\approx\; B_2 - \gamma B_1 A_1^\top A_2. \tag{15}$$

This matrix describes the task-2 prediction error before learning and determines the residual score derived in Appendix A.4.

A.2 The regularized update: solution, gain form, and endpoints

Substituting $(W_1 + \Delta W)A_2\boldsymbol{s} - B_2\boldsymbol{s} = \Delta W A_2 \boldsymbol{s} - \mathcal{E}\boldsymbol{s}$ into Eq. (1) and averaging over the latent, the objective becomes $N_y^{-1} \|\Delta W A_2 - \mathcal{E}\|_F^2 + \lambda N_y^{-1} \|\Delta W\|_F^2$, so fitting task-2 from W_1 is equivalent to fitting the error matrix $\mathcal{E}$. The objective is strictly convex for $\lambda > 0$, and its gradient vanishes at

$$\left(\Delta W A_2 - \mathcal{E}\right) A_2^\top + \lambda \, \Delta W = 0 \qquad \Longleftrightarrow \qquad \Delta W \left(A_2 A_2^\top + \lambda I_{N_x}\right) = \mathcal{E}\, A_2^\top.$$

Using the push-through identity $A^\top(AA^\top + \lambda I_{N_x})^{-1} = (A^\top A + \lambda I_{N_s})^{-1}A^\top$, which follows by multiplying the rearrangement $(A^\top A + \lambda I)A^\top = A^\top(AA^\top + \lambda I)$ by the two inverses, the unique solution is

$$\Delta W_\lambda \;=\; \mathcal{E}\left(A_2^\top A_2 + \lambda I_{N_s}\right)^{-1} A_2^\top. \tag{16}$$

Its row space lies inside $\mathrm{col}(A_2)$ at every λ, so directions orthogonal to all task-2 features are not updated.

Gain form. By the concentration (13), $\left(A_2^\top A_2 + \lambda I\right)^{-1} \approx (\gamma^{-1} + \lambda)^{-1} I_{N_s}$, so

$$\Delta W_\lambda \;\approx\; \frac{1}{\gamma^{-1} + \lambda}\, \mathcal{E} A_2^\top \;=\; p\, \underbrace{\gamma\, \mathcal{E} A_2^\top}_{\approx \Delta W_{\lambda=0}}, \qquad p = \frac{\gamma^{-1}}{\gamma^{-1} + \lambda},$$

where $\Delta W_{\lambda=0} = \mathcal{E}(A_2^\top A_2)^{-1} A_2^\top \approx \gamma \mathcal{E} A_2^\top$ is the $\lambda = 0$ solution. This is the main-text statement $\Delta W_\lambda \approx p\, \Delta W_{\lambda=0}$. Applying the update to a sample $\boldsymbol{x} = A_2 \boldsymbol{s}$ and using Eq. (13) once more, $\Delta W_\lambda\, \boldsymbol{x} \approx p\, \mathcal{E} \boldsymbol{s} = p\,(\boldsymbol{y}^* - W_1 \boldsymbol{x})$, so the prediction is the convex combination $W\boldsymbol{x} = (1-p)\, W_1 \boldsymbol{x} + p\, \boldsymbol{y}^*$, and the gain p is the fraction of the old student's error removed on every sample.

Endpoints. As $\lambda \to \infty$, $\Delta W_\lambda \to 0$ and the student remains at W_1. As $\lambda \to 0$, Eq. (16) gives, exactly,

$$W \;=\; W_1 + \mathcal{E}\left(A_2^\top A_2\right)^{-1} A_2^\top \;=\; W_1 \Big(I_{N_x} - A_2 \left(A_2^\top A_2\right)^{-1} A_2^\top \Big) + B_2 A_2^\dagger;$$

the matrix $A_2(A_2^\top A_2)^{-1} A_2^\top$ is symmetric, idempotent, and has range $\mathrm{col}(A_2)$, hence equals the orthogonal projector $U_2 U_2^\top$, and the limit is the fixed point $W_2 = W_1(I_{N_x} - U_2 U_2^\top) + B_2 A_2^\dagger$ of full task-2 training (Hiratani, 2024), as stated in the main text.

A.3 Losses in blocked form

Error contraction per task. Right-multiplying Eq. (16) by A_2 and using Eq. (13), $\Delta W_\lambda A_2 \approx p\, \mathcal{E}$, so the task-2 error matrix contracts uniformly, $WA_2 - B_2 = -\mathcal{E} + \Delta W_\lambda A_2 \approx -(1-p)\, \mathcal{E}$. For task-1, $W_1 A_1 = B_1$ gives $WA_1 - B_1 = \Delta W_\lambda A_1$; inserting Eq. (16) and the cross-overlap (14),

$$\Delta W_\lambda A_1 \;\approx\; \frac{1}{\gamma^{-1} + \lambda}\, \mathcal{E}\left(A_2^\top A_1\right) \;\approx\; \mathcal{E}\ \mathrm{diag}\left(\rho_i\, p\right),$$

so the task-1 error along direction i is the same error column that drives task-2 learning, scaled by the overlap ρ_i and the gain. When the gain differs by block (Appendix A.5), the same computations hold blockwise with (p_s, p_p).

Column norms of $\mathcal{E}$ and the weight η_s. The i-th column of $\mathcal{E} \approx B_2 - \gamma B_1 A_1^\top A_2$ is $\boldsymbol{b}_{2,i} - \gamma B_1 A_1^\top \boldsymbol{a}_{2,i}$. The vector $A_1^\top \boldsymbol{a}_{2,i}$ is the i-th column of $A_1^\top A_2$, which by Eq. (14) has a single dominant entry, $\rho_i \gamma^{-1}$ in position i; the subleading entries contribute only $O(\gamma)$ to squared norms, so at leading order $\gamma B_1 A_1^\top \boldsymbol{a}_{2,i} \approx \rho_i\, \boldsymbol{b}_{1,i}$ and the column is $\boldsymbol{b}_{2,i} - \rho_i\, \boldsymbol{b}_{1,i}$. On a specific direction ($\rho_i = 0$) this is $\boldsymbol{b}_{2,i}$, with mean squared norm N_y/N_s. On a shared direction, inserting the readout coupling (12),

$$\boldsymbol{b}_{2,i} - \rho_a \boldsymbol{b}_{1,i} \;=\; (\rho_b - \rho_a)\, \boldsymbol{b}_{1,i} + \sqrt{1 - \rho_b^2}\, \tilde{\boldsymbol{b}}_{2,i},$$

a sum of two independent zero-mean vectors, so the squared norms add:

$$\langle \|\mathcal{E}_{:,i}\|^2 \rangle_{A,B} \;=\; \Big[(\rho_b - \rho_a)^2 + 1 - \rho_b^2\Big] \frac{N_y}{N_s} \;=\; \big(1 - 2\rho_a \rho_b + \rho_a^2\big) \frac{N_y}{N_s} \;=\; \eta_s\, \frac{N_y}{N_s}.$$

The scalar η_s of the main text is the mean squared size of a shared error column: the old student carries a ρ_a-scaled copy of the task-1 readout onto each shared direction, which cancels against the target where the readouts agree (the $-2\rho_a\rho_b$ term) and adds to the error where they do not (the $+\rho_a^2$ term). Distinct columns involve independent fresh readout columns at leading order and are uncorrelated.

Assembling the losses. The squared Frobenius norm is the sum of squared column norms, so the contractions above give

$$\mathcal{L}_2 \;=\; \frac{1}{N_y}\sum_i (1-p_i)^2\,\langle \|\mathcal{E}_{:,i}\|^2\rangle_{A,B} \;=\; \frac{1}{N_s}\sum_i \eta_i\,(1-p_i)^2, \qquad \mathcal{L}_1 \;=\; \frac{1}{N_s}\sum_i \rho_i^2\,\eta_i\,p_i^2,$$

with $\eta_i = \eta_s$ on shared and 1 on specific directions; summing αN_s shared terms at gain p_s and $(1-\alpha)N_s$ specific terms at gain p_p yields the blocked form

$$\mathcal{L}_2 \;=\; \alpha\,\eta_s\,(1-p_s)^2 + (1-\alpha)\,(1-p_p)^2, \qquad \mathcal{L}_1 \;=\; \alpha\,\rho_a^2\,\eta_s\,p_s^2, \tag{17}$$

which at equal blocks, $\alpha = 1/2$, is Eq. (5) of the main text. Evaluating at the endpoints, $\mathcal{L}_2[W_1] = \alpha\eta_s + (1-\alpha)$ at $p \equiv 0$ and $\mathcal{L}_1[W_2] = \alpha\rho_a^2\eta_s$ at $p \equiv 1$; with $\mathcal{L}_1[0] = N_y^{-1}\,\|B_1\|_F^2 = 1$ for the untrained student, the $\alpha = 1$ retention $1 - \rho_a^2(\rho_a^2 - 2\rho_a\rho_b + 1)$ reproduces the closed form of Hiratani (2024).

A.4 Scores in latent space and covariance rescaling

Here we derive the reweighted covariance C and the per-block inflations (c_s, c_p) used in the main text, starting from the scores as defined on samples.

Latent form of the scores. Because the sample is a linear function of the latent, each score is a quadratic form in $\boldsymbol{s}$. For novelty, idempotency of the projector gives

$$\sigma_{\mathrm{nov}}(\boldsymbol{x}) \;=\; \|(I_{N_x} - P_1)\,A_2\boldsymbol{s}\|^2 \;=\; \boldsymbol{s}^\top \Sigma_\perp\,\boldsymbol{s}, \qquad \Sigma_\perp := A_2^\top (I_{N_x} - P_1)\,A_2.$$

For residual, the per-sample error of the task-1 student is $W_1\boldsymbol{x} - \boldsymbol{y}^* = (W_1A_2 - B_2)\boldsymbol{s} = -\mathcal{E}\boldsymbol{s}$, so

$$\sigma_{\mathrm{res}}(\boldsymbol{x}, \boldsymbol{y}^*) \;=\; \boldsymbol{s}^\top R\,\boldsymbol{s}, \qquad R := \mathcal{E}^\top\mathcal{E},$$

the quadratic form of the error matrix (15). Both matrices are positive semidefinite, each being the Gram matrix of a set of columns, so the scores are nonnegative. Below we write M for either score matrix.

Block approximation of the score matrices. For novelty, projecting onto $\mathrm{col}(A_1)^\perp$ removes the component inherited from task-1. Under the small-γ approximation, the remaining squared column norms concentrate around $(1-\rho_a^2)\gamma^{-1}$ on shared directions and γ^{-1} on task-specific directions. For residual scoring, Appendix A.3 gives the corresponding mean squared column norms, $\eta_s N_y/N_s$ and N_y/N_s. These quantities define the block approximations

$$\Sigma_\perp \;\approx\; \gamma^{-1}\begin{pmatrix}(1-\rho_a^2)I_d & 0\\ 0 & I_{N_s-d}\end{pmatrix}, \qquad R \;\approx\; \frac{N_y}{N_s}\begin{pmatrix}\eta_s I_d & 0\\ 0 & I_{N_s-d}\end{pmatrix}. \tag{18}$$

For the residual score, concentration of the full matrix around its block mean additionally assumes $N_y \gg N_s$. Using the feature Gram and cross-overlap approximations of Eqs. (13)-(14), off-diagonal entries of R have zero mean and typical magnitude of order $N_y^{-1/2}$ relative to the corresponding diagonal scale. Individual off-diagonal entries are therefore smaller, but their collective contribution need not be negligible when N_y is comparable to or smaller than N_s.

At smaller output dimensions, the block energies still provide a qualitative interpretation of residual scoring by describing its average emphasis on shared and task-specific directions. We therefore examine how finite output dimension affects the retention gap G (Appendix A.8) and additionally report experiments with output dimensions larger than the assumed latent dimensions (Appendices C and D), bringing the evaluation closer to the regime of the block approximation.

Below, we write $M \approx \mathrm{diag}(m_s I_d, m_p I_{N_s-d})$ for either score matrix. Under these approximations, the score matrices and the feature Gram matrix are scalar on the same shared and task-specific blocks, allowing the blockwise solution in Appendix A.5.

Reweighting as a change of distribution. The normalized data term in Eq. (2) can be written as

$$\frac{\langle \sigma(\boldsymbol{s})\ \|(\Delta W A_2 - \mathcal{E})\boldsymbol{s}\|^2 \rangle}{N_y \langle \sigma(\boldsymbol{s}) \rangle} = \frac{1}{N_y} \mathbb{E}_q \Big[\|(\Delta W A_2 - \mathcal{E})\boldsymbol{s}\|^2 \Big], \qquad q(\boldsymbol{s}) := \frac{\sigma(\boldsymbol{s})\, \varphi(\boldsymbol{s})}{\langle \sigma(\boldsymbol{s}) \rangle},$$

where φ is the standard Gaussian density and $\sigma(\boldsymbol{s}) = \boldsymbol{s}^\top M \boldsymbol{s}$. Thus score-proportional reweighting is equivalent to changing the latent training distribution from φ to q, while leaving the regularizer unchanged. Because the squared residual is quadratic in $\boldsymbol{s}$, this change enters the objective only through the second moment $C := \mathbb{E}_q[\boldsymbol{s}\boldsymbol{s}^\top]$, giving the covariance-rescaling view used in the main text.

Lemma 1 (Reweighted covariance)**.** *For any symmetric positive semidefinite score matrix M, exactly,*

$$C \ := \ \langle \boldsymbol{s}\boldsymbol{s}^\top \rangle_q \ = \ I_{N_s} + \frac{2M}{\operatorname{tr}(M)}\,; \tag{19}$$

in particular $\operatorname{tr} C = N_s + 2$ for every quadratic score, so reweighting only redistributes a fixed total of extra variance across directions.

Proof. A reweighted average is a ratio of two ordinary Gaussian averages: writing $\langle \cdot \rangle_q$ for the average under q, for any observable g,

$$\langle g(\boldsymbol{s}) \rangle_q \ = \ \int g(\boldsymbol{s})\, q(\boldsymbol{s})\, d\boldsymbol{s} \ = \ \frac{\langle g(\boldsymbol{s})\, \sigma(\boldsymbol{s}) \rangle}{\langle \sigma(\boldsymbol{s}) \rangle}.$$

The normalizer is the average score of an unselected sample: with $\sigma(\boldsymbol{s}) = \sum_{kl} M_{kl}\, s_k s_l$ and $\langle s_k s_l \rangle = \delta_{kl}$,

$$\langle \boldsymbol{s}^\top M \boldsymbol{s} \rangle \ = \ \sum_{kl} M_{kl}\, \delta_{kl} \ = \ \operatorname{tr}(M),$$

so the overall scale of M drops out of every reweighted average. For the second moment, the numerator is $\langle s_i s_j\, \boldsymbol{s}^\top M \boldsymbol{s} \rangle = \sum_{kl} M_{kl} \langle s_i s_j s_k s_l \rangle$, a Gaussian fourth moment given by Wick's rule, the sum over the three ways of pairing the four indices:

$$\langle s_i s_j s_k s_l \rangle \ = \ \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}.$$

Contracting with the symmetric M_{kl}, the first pairing yields $\delta_{ij} \sum_k M_{kk} = \delta_{ij} \operatorname{tr}(M)$, and the second and third each yield M_{ij}, so $\langle s_i s_j\, \boldsymbol{s}^\top M \boldsymbol{s} \rangle = \delta_{ij} \operatorname{tr}(M) + 2M_{ij}$. Dividing numerator by normalizer gives Eq. (19); taking the trace gives $\operatorname{tr} C = N_s + 2$. No concentration or large-N_s limit is used. □

Per-block inflation. Substituting the block approximation (18) into Eq. (19) gives, with the same relative corrections,

$$C \ \approx \ \operatorname{diag}\big(c_s I_d,\ c_p I_{N_s - d}\big), \qquad c = 1 + \frac{2m}{\operatorname{tr}(M)} \ \geq \ 1. \tag{20}$$

Each latent direction keeps its unit variance and gains the extra variance

$$c - 1 \ = \ \frac{2m}{\operatorname{tr}(M)} \ = \ O(1/N_s), \tag{21}$$

proportional to the weight m of its block. This is the per-block inflation used in the main text, larger on the block the score weights more. The normalizer $\operatorname{tr}(M)$ cancels from the ratio between blocks,

$$\frac{c_s - 1}{c_p - 1} \ = \ \frac{m_s}{m_p}\,,$$

and reading the block weights off the concentrated score matrices (18), $(m_s, m_p) \propto (1 - \rho_a^2,\ 1)$ for novelty and $(m_s, m_p) \propto (\eta_s,\ 1)$ for residual, which gives Eq. (8) of the main text.

A.5 THE REWEIGHTED SOLUTION

Under the reweighted latent distribution with covariance C, the objective becomes

$$\frac{1}{N_y}\operatorname{tr}\big[(\Delta W A_2 - \mathcal{E})^\top (\Delta W A_2 - \mathcal{E})\, C\big] + \frac{\lambda}{N_y}\,\|\Delta W\|_F^2 .$$

For $\lambda > 0$, this objective is strictly convex. Setting its gradient to zero gives

$$\Delta W\big(A_2 C A_2^\top + \lambda I_{N_x}\big) = \mathcal{E} C A_2^\top .$$

Solving and applying the weighted push-through identity yields

$$\begin{aligned}\Delta W_\lambda &= \mathcal{E} C A_2^\top \big(A_2 C A_2^\top + \lambda I_{N_x}\big)^{-1} \\ &= \mathcal{E}\big(C A_2^\top A_2 + \lambda I_{N_s}\big)^{-1} C A_2^\top .\end{aligned} \tag{22}$$

This solution is exact for the realized covariance C and reduces to Eq. (16) when $C = I_{N_s}$. The second expression places the inverse in latent space, where we next apply the block approximation. Writing c_i for the inflation of the block containing direction i,

$$\Delta W_\lambda \;\approx\; \mathcal{E}\;\operatorname{diag}\!\left(\frac{c_i}{c_i\,\gamma^{-1} + \lambda}\right) A_2^\top ,$$

the same correction-at-reduced-gain form with the gain set per block: right-multiplying by A_2 and using Eq. (13) once more, $\Delta W_\lambda A_2 \approx \mathcal{E}\ \operatorname{diag}(p_i)$ with the per-block gains of Eq. (6) in the main text. The loss formulas hold with the tilted gains: the task-2 contraction is blockwise as in Appendix A.3, and for retention the cross-overlap computation reads

$$\Delta W_\lambda A_1 \;\approx\; \mathcal{E}\;\operatorname{diag}\!\left(\frac{c_i}{c_i\,\gamma^{-1} + \lambda}\right)\big(A_2^\top A_1\big) \;\approx\; \mathcal{E}\;\operatorname{diag}\big(\rho_i\, p_i\big),$$

Here we identify the reweighted gain on direction i as

$$p_i := \frac{c_i \gamma^{-1}}{c_i \gamma^{-1} + \lambda}.$$

Setting $c_i = 1$ recovers the uniform-sampling gain $p = \gamma^{-1}/(\gamma^{-1} + \lambda)$. Reweighting therefore preserves the form of the task-1 residual, $\mathcal{E}\,\operatorname{diag}(\rho_i p_i)$, while changing the gain according to the covariance inflation c_i of each direction. A block-uniform inflation $c_s = c_p = c$ gives $p = \gamma^{-1}/(\gamma^{-1} + \lambda/c)$ on both blocks, recovering the uniform-sampling solution at penalty λ/c. Thus a common inflation only reparameterizes the trade-off curve; changes to the curve depend on the ratio c_p/c_s, which determines the comparison in Appendix A.6.

A.6 THE SIGN AND MAGNITUDE OF THE RETENTION GAP

Here we compare the trade-off curve of a score with that of uniform sampling: we determine the sign of the gap G defined in the main text Section 4 and then compute its magnitude at leading order.

Proposition 1 (Ordering of the trade-off curves)**.** *Fix a task pair and a sampling scheme with block inflations $c_s, c_p > 0$. If $\alpha\rho_a^2\eta_s > 0$, then at every interior point of the trade-off curves, at matched task-2 loss,*

$$\operatorname{sign} G \;=\; \operatorname{sign}(c_p - c_s),$$

with no expansion in $c_p - c_s$. If $\alpha\rho_a^2\eta_s = 0$, both retention losses vanish identically and $G \equiv 0$.

Proof. The degenerate case is immediate from Eq. (17), so assume $\alpha\rho_a^2\eta_s > 0$. Both curves are parameterized by λ, and the same penalty lands at different points of each curve, so we first eliminate λ. By Eq. (17), $\mathcal{L}_1$ depends on the gains only through the shared gain p_s, so $p_s \in (0, 1)$ determines $\mathcal{L}_1$ exactly, at any tilt. We fix the value of p_s, solve for the penalty each curve requires to realize it, and compare the remaining quantity, the specific residual $1 - p_p$, which carries the remaining part of $\mathcal{L}_2$.

From Eq. (6), the penalty realizing shared gain p_s is $\lambda = c_s\gamma^{-1}(1-p_s)/p_s$; substituting into $1-p_p = \lambda/(c_p\gamma^{-1}+\lambda)$ and dividing numerator and denominator by γ^{-1},

$$1 - p_p \;=\; \frac{c_s\,(1-p_s)}{c_p\,p_s + c_s\,(1-p_s)}. \tag{23}$$

Uniform sampling is the case $c_s = c_p$, for which $1-p_p = 1-p_s$: both blocks are learned at the same rate. Whether Eq. (23) lies above or below $1-p_s$ is decided by comparing its denominator with c_s, using the identity

$$c_p\,p_s + c_s\,(1-p_s) - c_s \;=\; (c_p - c_s)\,p_s. \tag{24}$$

For $p_s \in (0,1)$ the factor p_s is positive, so the denominator exceeds c_s exactly when $c_p > c_s$: the score's specific residual is smaller than uniform sampling's precisely when $c_p > c_s$, at every interior p_s and at any tilt strength, since no expansion in $c_p - c_s$ is made. At $p_s \in \{0,1\}$ the two residuals coincide, so the curves share their endpoints.

Next we compare the losses at matched retention. At equal p_s, the two values of $\mathcal{L}_1$ are equal and the shared contributions to $\mathcal{L}_2$ are identical, so the difference comes from the specific block alone:

$$\mathcal{L}_2^{\text{score}} - \mathcal{L}_2^{\text{unif}}\Big|_{\text{equal }\mathcal{L}_1} \;=\; (1-\alpha)\left[\left(\frac{c_s\,(1-p_s)}{c_p\,p_s + c_s\,(1-p_s)}\right)^2 - \;(1-p_s)^2\right],$$

which is negative exactly when $c_p > c_s$ by Eq. (24). The score's curve therefore lies strictly on one side of uniform sampling's in the interior and meets it only at the shared endpoints.

Along either curve, increasing p_s decreases $\mathcal{L}_2$ and increases $\mathcal{L}_1$. Thus, when $c_p > c_s$, the score achieves any interior task-2 loss with a smaller shared gain than uniform sampling, and hence less forgetting. The reverse holds when $c_p < c_s$, while equal inflations give identical curves. Therefore, at matched $\mathcal{L}_2$,

$$\operatorname{sign} G = \operatorname{sign}\left(\mathcal{L}_1^{\text{unif}} - \mathcal{L}_1^{\text{score}}\right) = \operatorname{sign}(c_p - c_s).$$

□

Gap magnitude at leading order. The proposition determines the sign of G for arbitrary block inflations. We now compute its leading-order magnitude under score-proportional reweighting, working within the same block approximation. Let $p_s \in (0,1)$ denote the uniform-sampling gain at the matched task-2 loss, and write the score's shared gain as $p_s - \delta$. Then

$$G = \alpha\rho_a^2\eta_s\left[p_s^2 - (p_s-\delta)^2\right].$$

To determine δ, we equate the task-2 losses using Eqs. (17) and (23):

$$\alpha\eta_s(1-p_s+\delta)^2 + (1-\alpha)\left(\frac{c_s(1-p_s+\delta)}{c_p(p_s-\delta)+c_s(1-p_s+\delta)}\right)^2 = (\alpha\eta_s + 1 - \alpha)(1-p_s)^2. \tag{25}$$

Define the relative difference between block inflations as

$$\varepsilon := \frac{c_p - c_s}{c_s} = \frac{2(m_p - m_s)}{N_s\bar{m} + 2m_s}, \qquad \bar{m} := \alpha m_s + (1-\alpha)m_p.$$

Under reweighting, $\varepsilon = O(N_s^{-1})$. When $\varepsilon = 0$, the curves coincide and $\delta = 0$, so we expand with $\delta = O(\varepsilon)$. Dividing the specific-block residual in Eq. (25) by c_s gives

$$\frac{1-p_s+\delta}{1+\varepsilon(p_s-\delta)} = (1-p_s+\delta)(1-\varepsilon p_s) + O(\varepsilon^2).$$

Substituting this expression into Eq. (25), expanding the squares, and cancelling the common zeroth-order term yields

$$0 = 2(\alpha\eta_s + 1 - \alpha)(1-p_s)\delta - 2(1-\alpha)\varepsilon p_s(1-p_s)^2 + O(\varepsilon^2).$$

Thus, at a fixed interior gain p_s,

$$\delta = \frac{(1-\alpha)p_s(1-p_s)}{\alpha\eta_s + 1 - \alpha}\,\varepsilon + O(\varepsilon^2).$$

For $\varepsilon > 0$, stronger task-specific learning therefore allows a smaller shared gain at the same task-2 loss.

Substituting into the retention gap gives

$$\begin{aligned} G &= \alpha\rho_a^2\eta_s(2p_s\delta - \delta^2) \\ &= \frac{2\alpha(1-\alpha)\rho_a^2\eta_s}{\alpha\eta_s + 1 - \alpha}(1-p_s)p_s^2\,\varepsilon + O(\varepsilon^2). \end{aligned}$$

Finally, using $\varepsilon = 2(m_p - m_s)/(N_s\bar{m}) + O(N_s^{-2})$,

$$G = \underbrace{\frac{4\alpha(1-\alpha)\rho_a^2\eta_s}{\alpha\eta_s + 1 - \alpha}}_{\text{task pair}}\ \underbrace{(1-p_s)p_s^2}_{\text{gain}}\ \underbrace{\frac{m_p - m_s}{N_s\bar{m}}}_{\text{score}} + O(N_s^{-2}). \tag{26}$$

The remainder describes the reweighting expansion within the block approximation.

Classification of the two scores. By Eq. (21) the inflations share the positive normalizer $\mathrm{tr}(M)$, so $\mathrm{sign}(c_p - c_s) = \mathrm{sign}(m_p - m_s)$, and the Proposition classifies the scores through their block weights alone: novelty has $m_s < m_p$ whenever $\rho_a > 0$, so $G_{\mathrm{nov}} > 0$ for every task pair, while residual has $\mathrm{sign}(m_p - m_s) = \mathrm{sign}(1 - \eta_s) = \mathrm{sign}(\rho_b - \rho_a/2)$, the boundary stated in the main text. This is the sign of the score factor in Eq. (26), here established exactly rather than at leading order.

A.7 Finite sample size

The theory is stated in the infinite-sample limit, where the student solves against the second moment of its sampling distribution. For finite sample size with n task-2 samples, the population second moment is replaced by its empirical counterpart. Under score-proportional reweighting, $\hat{C} = \frac{\sum_i \sigma_i \boldsymbol{s}_i \boldsymbol{s}_i^\top}{\sum_i \sigma_i}$, while under hard selection the covariance is computed over the retained samples. Finite sampling therefore introduces fluctuations around the population covariance considered in the theory.

These fluctuations are largely averaged when performance is aggregated across the shared and task-specific blocks, whereas the effect of selection acts coherently across directions within each block. As a result, the finite-sample curves approach the infinite-sample prediction as the number of samples per latent dimension, n/N_s, increases. Fig. 6 shows that once the sample size is comparable to or larger than the latent dimension, the retention gap approaches its population value and preserves the predicted score ordering. We therefore focus on the population effect of sample selection rather than finite-sample estimation noise.

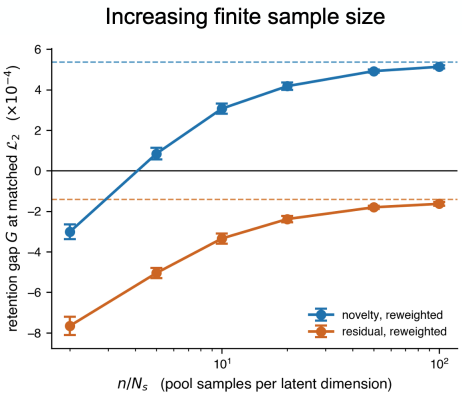

Figure 6: Retention gap at matched task-2 loss (evaluated at $p_s = 2/3$) between score-based reweighting and uniform sampling on the same pool of n samples, against samples per latent dimension n/N_s (mean $\pm 95\%$ CI) over pools; $\alpha = 0.5$, $\rho_a = 0.8$, $\rho_b = 0.2$, $N_s = 200$, $\gamma = 0.05$. Dashed lines: infinite-sample values. The gap carries the predicted sign from moderate n/N_s and converges to the theory.

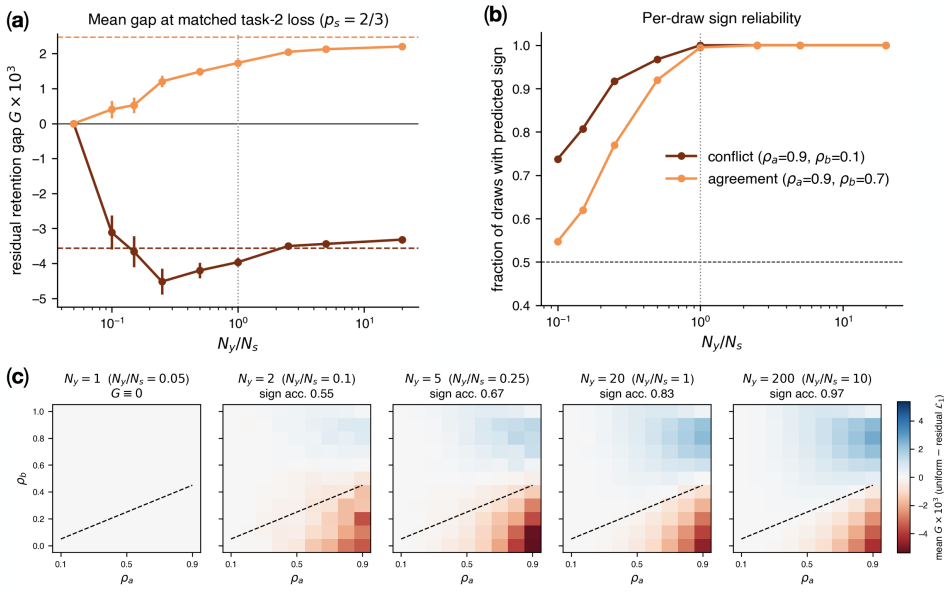

Figure 7: Residual scoring at small output dimension, using the realized score matrix R with $N_s = 20$ and $\alpha = 1/2$. **(a)** Mean retention gap G at matched task-2 loss, referenced to uniform gain $p_s = p_p = 2/3$, for a conflict pair $(\rho_a, \rho_b) = (0.9, 0.1)$ and an agreement pair $(0.9, 0.7)$. Error bars show 95% CIs over 400 realizations. Horizontal colored dashed lines show the leading-order reweighting prediction of Eq. (26). **(b)** Fraction of realizations whose gap has the sign predicted by the block approximation. The horizontal dashed line marks chance level; $N_y = 1$ is excluded because its population gap is zero. Vertical dotted lines in (a,b) mark $N_y = N_s$. **(c)** Mean reweighting gap across the task-similarity plane for $N_y \in \{1, 2, 5, 20, 200\}$, using 100 realizations per cell. Dashed lines mark $\rho_b = \rho_a/2$; panel titles report per-realization sign accuracy averaged over off-boundary cells. Positive gaps indicate better retention than uniform sampling.

A.8 Residual scoring beyond the large-output regime

The block approximation for residual scoring assumes $N_y \gg N_s$. To examine sensitivity to this assumption, we retain the realized residual score matrix $R = \mathcal{E}^\top \mathcal{E}$ rather than replacing it by its block mean. We keep the isotropic feature-Gram approximation used in the main analysis and compute the reweighted solution using the exact covariance

$$C = I_{N_s} + \frac{2R}{\mathrm{tr}(R)}.$$

We fix $N_s = 20$ and $\alpha = 1/2$, vary N_y, and evaluate the retention gap G at matched task-2 loss $\mathcal{L}_2$, using the uniform-sampling gain $p_s = p_p = 2/3$ as the reference.

For the representative conflict and agreement pairs, the mean gap retains the predicted sign at all tested output dimensions $N_y \geq 2$, although its magnitude varies with N_y (Fig. 7a). The fraction of individual realizations with the predicted sign increases with output dimension (Fig. 7b). Across the task-similarity plane, the mean gaps recover the broad conflict–agreement pattern even at small N_y, while showing deviations from the block prediction near the theoretical boundary (Fig. 7c). These results support the qualitative interpretation of residual scoring beyond the large-output regime.

A.9 Simulation details

We use pools of 2-3$\times 10^5$ latent samples, implementing score-proportional reweighting by weighted resampling or direct weighting, and hard selection example is by retaining the top 30% of scores. Sweeping λ traces the learning-retention trade-off, with losses evaluated analytically. Fig. 2(b-d) shows the mean over 32 independent seeds; SEM error bars are smaller than the markers. Fig. 3 uses a single realization.

	Fig. 2 (b–d)	Fig. 3
Latent dim. N_s	8	50
Shared latent dims. $d = \alpha N_s$ ($\alpha = 0.5$)	4	25
Feature dim. N_x	4000	2000
Target dim. N_y	256	120
$\gamma = N_s/N_x$	0.002	0.025

Table 1: Model dimensions used in the simulations.

B HARD SELECTION AND REWEIGHTING APPROXIMATED BY THE SAME TILT FAMILY

This appendix derives the claim of the main text: for the same scoring function, score-proportional reweighting and hard top-B selection produce, at leading order, latent covariances in the exponential-tilt family of Eq. (9), at the strengths of Eq. (10). All statements are conditional on a fixed realization of the task maps (A_μ, B_μ) and are made for the score matrix this realization defines; the block concentration of Eq. (18) enters only in the final step (Appendix B.8). The sign of the retention comparison rests on an exact ordering result (Lemma 3) and holds at every budget; the approximations enter only the quantitative statements.

B.1 SETUP, NOTATION, AND OVERVIEW

As in Appendix A.4, we study selection through the latent covariance it induces: the quadratic training objective depends on the sampling distribution only through its second moment. We therefore compare the covariances produced by score-proportional reweighting and hard selection.

Fix a realization of the task maps and write the score as $\sigma(\boldsymbol{s}) = \boldsymbol{s}^\top M \boldsymbol{s}$, where $M \in \{\Sigma_\perp, R\}$ is symmetric and positive semidefinite. To separate the contributions of individual directions, we express $\boldsymbol{s}$ in an orthonormal eigenbasis of M and denote the resulting coordinates by $\boldsymbol{t}$. Because the latent is standard Gaussian, this rotation preserves its distribution: the coordinates $t_1, \dots, t_{N_s}$ remain independent standard normals. The score then becomes

$$\sigma(\boldsymbol{s}) = \sum_{i=1}^{N_s} m_i t_i^2, \qquad m_i \geq 0$$

where m_i are the eigenvalues of M. This representation lets us analyze how selection changes the variance along each score eigendirection. In particular,

$$\langle \sigma \rangle = \operatorname{tr}(M), \qquad \operatorname{Var}(\sigma) = 2 \operatorname{tr}(M^2).$$

Hard selection retains a fraction $B \in (0, 1)$ of the population. Let τ_B satisfy $\Pr(\sigma > \tau_B) = B$, and define the retained set $\mathcal{S}_B := \{\boldsymbol{s} : \sigma(\boldsymbol{s}) > \tau_B\}$. The covariances we compare are

$$C^{\mathrm{rw}} := \langle \boldsymbol{s}\boldsymbol{s}^\top \rangle_q, \qquad C^{\mathrm{sel}} := \langle \boldsymbol{s}\boldsymbol{s}^\top \mid \mathcal{S}_B \rangle,$$

where q is the score-proportional density defined in Appendix A.4. Both sampling laws are symmetric under $\boldsymbol{s} \mapsto -\boldsymbol{s}$, so their means vanish and these second moments are covariances.

For the Gaussian approximation below, write φ for the standard normal density, Φ for its distribution function, and $\bar{\Phi} := 1 - \Phi$ for its survival function. Let z_B satisfy $\bar{\Phi}(z_B) = B$, and define

$$H(z) := \frac{\varphi(z)}{\bar{\Phi}(z)} = -\frac{d}{dz} \log \bar{\Phi}(z).$$

This Gaussian hazard function measures the sensitivity of the log-survival probability to a threshold shift; at z_B, $H(z_B) = \varphi(z_B)/B$ is the selection intensity appearing in Eq. (10). For other distribution functions F, we similarly write $\bar{F} := 1 - F$.

B.2 THE COVARIANCE FAMILY

Definition. For the quadratic score $\sigma(\boldsymbol{s}) = \boldsymbol{s}^\top M \boldsymbol{s}$, define the exponentially tilted density

$$q_\theta(\boldsymbol{s}) := \frac{1}{Z_\theta} e^{\theta\sigma(\boldsymbol{s})}\varphi(\boldsymbol{s}), \qquad 0 \le \theta < \frac{1}{2m_{\max}}, \qquad m_{\max} := \max_i m_i,$$

where φ is the standard Gaussian density on $\mathbb{R}^{N_s}$ and Z_θ is the normalizing constant. The parameter θ controls how strongly the distribution favors high-scoring samples; $\theta = 0$ recovers the original latent distribution.

Lemma 2 (Covariance of the exponential tilt)**.** *The density q_θ is a centered Gaussian with covariance*

$$C(\theta) := \mathbb{E}_{q_\theta}[\boldsymbol{s}\boldsymbol{s}^\top] = \big(I_{N_s} - 2\theta M\big)^{-1}. \tag{27}$$

Proof. In the eigencoordinates introduced in Appendix B.1, the score is $\sigma = \sum_i m_i t_i^2$ and the original Gaussian density factorizes. Exponential weighting therefore gives

$$q_\theta(\boldsymbol{t}) \propto \prod_{i=1}^{N_s} \exp\left[-\frac{1}{2}(1 - 2\theta m_i)t_i^2\right].$$

This density is normalizable precisely when $1 - 2\theta m_i > 0$ for every i, which gives the stated domain. Within this domain, the coordinates remain independent centered Gaussians, with

$$\mathbb{E}_{q_\theta}[t_i^2] = \mathrm{Var}_{q_\theta}(t_i) = \frac{1}{1 - 2\theta m_i}, \qquad \mathbb{E}_{q_\theta}[t_i t_j] = 0 \quad (i \neq j).$$

Rotating back to the original coordinates gives $C(\theta) = (I_{N_s} - 2\theta M)^{-1}$. □

Properties. Write $c_i(\theta) := (1 - 2\theta m_i)^{-1}$ for the variance along eigendirection i. The family starts at $C(0) = I_{N_s}$. As θ increases, every direction with $m_i > 0$ gains variance, while a zero-weight direction retains unit variance. At any $\theta > 0$, larger score weights give larger variances.

The relative emphasis on differently weighted directions also increases with θ. For $m_i > m_j$,

$$\frac{c_i(\theta)}{c_j(\theta)} = \frac{1 - 2\theta m_j}{1 - 2\theta m_i}, \qquad \frac{d}{d\theta}\log\frac{c_i(\theta)}{c_j(\theta)} = \frac{2m_i}{1 - 2\theta m_i} - \frac{2m_j}{1 - 2\theta m_j} > 0.$$

Thus a stronger tilt increases both the variance in score-weighted directions and the variance ratio favoring the more heavily weighted direction. This ratio determines the retention comparison after the block reduction in Appendix B.8.

B.3 SCORE-PROPORTIONAL SAMPLING IS THE FIRST-ORDER MEMBER

The reweighted covariance is exact for the realized score matrix (Lemma 1):

$$C^{\mathrm{rw}} = I_{N_s} + \frac{2M}{\mathrm{tr}(M)}.$$

In the eigenframe of M, its entries are $1 + 2m_i/\,\mathrm{tr}(M)$. To compare these with the exponential-tilt family, expand its directional variances:

$$c_i(\theta) = \frac{1}{1 - 2\theta m_i} = 1 + 2\theta m_i + O\big((\theta m_i)^2\big).$$

Choosing

$$\theta_{\mathrm{rw}} := \frac{1}{\mathrm{tr}(M)} = \frac{1}{\langle\sigma\rangle}$$

therefore matches the reweighted covariance to first order, with a per-direction error $O\big((m_i/\,\mathrm{tr}(M))^2\big)$. This approximation applies when $m_{\max}/\,\mathrm{tr}(M) \ll 1$, so no single direction contributes a substantial fraction of the mean score.

B.4 HARD SELECTION: EXACT REDUCTIONS

To approximate C^{sel}, we first reduce the calculation to individual eigendirections. Hard selection preserves sign symmetry in each t_i, because both the Gaussian density and the event $\mathcal{S}_B$ depend on these coordinates only through their squares. Thus the conditional means and off-diagonal second moments vanish, and it remains to determine $\langle t_i^2 \mid \mathcal{S}_B \rangle$.

For a fixed direction i, write the contribution from the remaining directions as

$$\sigma_{-i} := \sum_{k \neq i} m_k t_k^2.$$

This quantity is independent of t_i, with survival function $\bar{F}_{-i}$, mean $\mathrm{tr}(M) - m_i$, and variance $v_{-i} := 2\big(\mathrm{tr}(M^2) - m_i^2\big)$. Conditional on $t_i = t$, a sample is retained when $\sigma_{-i} > \tau_B - m_i t^2$, giving the acceptance probability

$$a_i(t) := \Pr(\mathcal{S}_B \mid t_i = t) = \bar{F}_{-i}(\tau_B - m_i t^2). \tag{28}$$

The conditional density of t_i is therefore $\varphi(t) a_i(t)/B$, and its variance has the exact one-dimensional representation

$$\langle t_i^2 \mid \mathcal{S}_B \rangle = \frac{\mathbb{E}_t[t^2 a_i(t)]}{\mathbb{E}_t[a_i(t)]} = \frac{1}{B}\mathbb{E}_t[t^2 a_i(t)], \qquad t \sim \mathcal{N}(0,1). \tag{29}$$

Before approximating these variances, we show that directions with larger score weights have larger conditional variances under hard selection. Under the block approximation, this comparison determines which block is favored and hence the sign of the retention gap.

Lemma 3 (Variance ordering under hard selection)**.** *For every budget $B \in (0,1)$, $m_i > m_j$ implies*

$$\langle t_i^2 \mid \mathcal{S}_B \rangle > \langle t_j^2 \mid \mathcal{S}_B \rangle,$$

while $m_i = m_j$ implies equality.

Proof. Write $\sigma = m_i t_i^2 + m_j t_j^2 + \sigma_{-ij}$, where $\sigma_{-ij} := \sum_{k \neq i,j} m_k t_k^2$ is independent of (t_i, t_j). Since t_i and t_j are exchangeable, averaging an expression with its coordinate-swapped version gives

$$\begin{aligned}\mathbb{E}[(t_i^2 - t_j^2)\mathbf{1}_{\mathcal{S}_B}] = \frac{1}{2}\mathbb{E}\Big[(t_i^2 - t_j^2)\Big(&\mathbf{1}\{m_i t_i^2 + m_j t_j^2 + \sigma_{-ij} > \tau_B\} \\ &- \mathbf{1}\{m_i t_j^2 + m_j t_i^2 + \sigma_{-ij} > \tau_B\}\Big)\Big].\end{aligned}$$

The two scores inside the indicators differ by $(m_i - m_j)(t_i^2 - t_j^2)$. For $m_i > m_j$, the indicator difference therefore has the same sign as $t_i^2 - t_j^2$ whenever it is nonzero, making the integrand nonnegative. It is strictly positive with positive probability: since $\tau_B > 0$ and $\Pr(\sigma_{-ij} < \tau_B) > 0$, there is a positive-probability set of coordinate values for which swapping t_i^2 and t_j^2 moves the score across the threshold. Dividing by B proves the strict inequality. Equal weights give equality by exchangeability. □

The following sections approximate the variances in the remaining directions through Eq. (29).

B.5 GAUSSIAN APPROXIMATION OF THE ACCEPTANCE PROBABILITY

The acceptance probability in Eq. (28) depends on the distribution of $\sigma_{-i} = \sum_{k \neq i} m_k t_k^2$, a sum of independent, unequally weighted squared Gaussian variables. We approximate this sum by a Gaussian when no single direction dominates its variance. Under the block approximation of Eq. (18), the score receives contributions from a number of directions proportional to N_s, with comparable nonzero weights. The Berry–Esseen bound then gives

$$\bar{F}_{-i}(x) = \bar{\Phi}\left(\frac{x - \big(\mathrm{tr}(M) - m_i\big)}{\sqrt{v_{-i}}}\right) + O(N_s^{-1/2}), \tag{30}$$

uniformly in x, where $v_{-i} = 2\big(\operatorname{tr}(M^2) - m_i^2\big)$. This is an absolute error bound on the survival probability.

Applying the same approximation to the full score, whose mean and variance are $\operatorname{tr}(M)$ and $2\operatorname{tr}(M^2)$, gives

$$B = \Pr(\sigma > \tau_B) = \bar{\Phi}\left(\frac{\tau_B - \operatorname{tr}(M)}{\sqrt{2\operatorname{tr}(M^2)}}\right) + O(N_s^{-1/2}).$$

For a fixed budget $B \in (0,1)$, inverting the Gaussian survival function therefore yields

$$\tau_B = \operatorname{tr}(M) + \sqrt{2\operatorname{tr}(M^2)}\left(z_B + O(N_s^{-1/2})\right), \tag{31}$$

where $\bar{\Phi}(z_B) = B$.

B.6 Linearizing the log-acceptance: the exponential tilt

We now approximate $\langle t_i^2 \mid \mathcal{S}_B\rangle$ through the acceptance probability $a_i(t)$. Conditional on $t_i = t$, the remaining score must exceed $\tau_B - m_i t^2$. Its mean is $\operatorname{tr}(M) - m_i$, so the threshold measured relative to that mean is

$$(\tau_B - m_i t^2) - (\operatorname{tr}(M) - m_i) = \tau_B - \operatorname{tr}(M) - m_i(t^2 - 1).$$

Thus $m_i(t^2-1)$ is direction i's contribution above its average contribution m_i: a value $t^2 > 1$ lowers the threshold faced by the remaining directions.

The Gaussian approximation of Appendix B.5 therefore gives

$$a_i(t) \approx \bar{\Phi}\left(\frac{\tau_B - \operatorname{tr}(M) - m_i(t^2-1)}{\sqrt{v_{-i}}}\right).$$

Removing direction i reduces the full score variance by $2m_i^2$:

$$v_{-i} = 2\operatorname{tr}(M^2) - 2m_i^2 = 2\operatorname{tr}(M^2)\left(1 - \frac{m_i^2}{\operatorname{tr}(M^2)}\right).$$

Since no single direction contributes a substantial fraction of the score variance, $m_i^2/\operatorname{tr}(M^2) \ll 1$, and we use $\sqrt{v_{-i}} \approx \sqrt{2\operatorname{tr}(M^2)}$. Together with the leading-order threshold in Eq. (31), this yields

$$a_i(t) \approx \bar{\Phi}\left(z_B - \frac{m_i(t^2-1)}{\sqrt{2\operatorname{tr}(M^2)}}\right).$$

To connect this acceptance probability to the exponential tilt, we approximate $\log a_i(t)$ to first order in the contribution from direction i. Recall that we introduce the Gaussian hazard function

$$H(z) = \frac{\varphi(z)}{\bar{\Phi}(z)}, \qquad \frac{d}{dz}\log\bar{\Phi}(z) = \frac{-\varphi(z)}{\bar{\Phi}(z)} = -H(z).$$

Expanding around z_B therefore gives

$$\begin{aligned}\log a_i(t) &\approx \log\bar{\Phi}(z_B) - \frac{m_i(t^2-1)}{\sqrt{2\operatorname{tr}(M^2)}}\,\frac{d}{dz}\log\bar{\Phi}(z)\Big|_{z=z_B} \\ &= \log\bar{\Phi}(z_B) + \frac{H(z_B)}{\sqrt{2\operatorname{tr}(M^2)}} m_i(t^2-1).\end{aligned}$$

Define

$$\theta(B) := \frac{H(z_B)}{\sqrt{2\operatorname{tr}(M^2)}}. \tag{32}$$

After exponentiating, terms independent of t cancel upon normalization, leaving an acceptance weight proportional to $e^{\theta(B)m_i t^2}$. Multiplying this approximate weight by the original Gaussian density gives

$$\varphi(t)e^{\theta(B)m_i t^2} \propto \exp\left(-\frac{1}{2}\big(1 - 2\theta(B)m_i\big)t^2\right).$$

This is a centered Gaussian with variance $(1 - 2\theta(B)m_i)^{-1}$, so the conditional second moment is approximated by

$$\langle t_i^2 \mid \mathcal{S}_B \rangle \approx \frac{1}{1 - 2\theta(B)m_i}.$$

The hard cutoff therefore induces approximately the same directional variances as the exponential tilt.

At fixed B, the small-contribution condition for each direction also makes $\theta(B)m_i$ small, since $H(z_B)$ is fixed. Requiring this for every direction gives $\theta(B)m_{\max} \ll 1$. This keeps the predicted variances close to one, consistent with treating each direction's contribution as a perturbation.

Proposition 2 (Exponential-tilt approximation of hard selection)**.** *Under the Gaussian approximation of Appendix B.5, at a fixed budget $B \in (0, 1)$ with $\theta(B)m_{\max} \ll 1$, linearizing the log-acceptance gives the covariance approximation*

$$C^{\mathrm{sel}} \approx \big(I_{N_s} - 2\theta(B)M\big)^{-1},$$

where

$$\theta(B) = \frac{H(z_B)}{\sqrt{2\,\mathrm{tr}(M^2)}} = \frac{\varphi(z_B)/B}{\sqrt{\mathrm{Var}(\sigma)}}. \tag{33}$$

Validity. The approximation requires both a Gaussian description of the remaining score and a small contribution from each direction. Under the block approximation, $\theta(B)m_{\max} = O(H(z_B)/\sqrt{N_s})$, so the small-perturbation condition holds for large N_s at fixed B. Tighter selection increases $H(z_B)$ and can make the linearization inaccurate. The weaker condition $2\theta(B)m_{\max} < 1$ only ensures that the approximating Gaussian is normalizable. Outside the approximation's range, the exact variance ordering of Lemma 3 still holds.

B.7 Moment matching for the tilt strength

For the theory curves in Fig. 3, we use the same covariance family but calibrate its strength using the mean score of the retained samples. Under q_θ, the mean score is $\sum_i m_i/(1 - 2\theta m_i)$. We therefore choose $\hat{\theta}$ to satisfy

$$\sum_i \frac{m_i}{1 - 2\hat{\theta}m_i} = \langle \sigma \mid \sigma > \tau_B \rangle, \tag{34}$$

using the measured mean score of the retained samples on the right. The left side increases continuously from $\mathrm{tr}(M)$ to infinity over $0 \le \theta < 1/(2m_{\max})$, so a unique solution exists whenever the measured mean exceeds $\mathrm{tr}(M)$.

We then evaluate $C(\hat{\theta}) = (I_{N_s} - 2\hat{\theta}M)^{-1}$ and compute the shared and task-specific inflations by averaging its diagonal entries within each block in the original latent coordinates. The measured survivor covariance is summarized in the same way. This calibration matches the retained mean score; the individual directional variances remain approximate.

B.8 Back to blocks: consequences for the loss

Under the block approximation of Eq. (18), the selected covariance has the form $C^{\mathrm{sel}} = \mathrm{diag}(c_s I_d, c_p I_{N_s - d})$, so the block gains and losses derived in Appendices A.5 and A.6 apply directly.

By Lemma 3, the block with the larger score weight has the larger conditional variance at every budget $B \in (0, 1)$. Combining this ordering with Proposition 1 gives, for nondegenerate task pairs,

$$\mathrm{sign}\, G = \mathrm{sign}(c_p - c_s) = \mathrm{sign}(m_p - m_s).$$

Thus hard selection preserves the classification obtained for reweighting.

To compare the strength of the two sampling schemes, we divide their tilt parameters:

$$\frac{\theta(B)}{\theta_{\mathrm{rw}}} = H(z_B)\frac{\mathrm{tr}(M)}{\sqrt{2\,\mathrm{tr}(M^2)}} = \Theta\big(H(z_B)\sqrt{N_s}\big).$$

The spectral ratio $(\mathrm{tr}\, M)^2/\,\mathrm{tr}(M^2)$ measures the effective number of latent directions contributing to the score. Under our block approximation, this number is of order N_s for both scoring rules.

Consequently, at a fixed budget and within the tilt approximation's validity regime, hard selection induces a tilt of order $\sqrt{N_s}$ stronger than reweighting. Tightening the budget increases $H(z_B)$ and hence the tilt strength. By the variance-ratio monotonicity established in Appendix B.2, this increases the magnitude of the retention gap at matched task-2 loss while preserving its sign.

C Experimental details: the task-conflict regime (CIFAR-100)

C.1 Theory-to-experiment mapping

Each image is mapped once to $\phi(\boldsymbol{x}) \in \mathbb{R}^{768}$ by a frozen ViT-B/16 pretrained on ImageNet (Dosovitskiy et al., 2020; Steiner et al., 2021). The frozen backbone plays the role of the fixed feature map A_μ. Both tasks share a single 5-output head $W \in \mathbb{R}^{5\times 768}$, and each task pair consists of two disjoint sets of five CIFAR-100 fine classes drawn uniformly at random. We use the official CIFAR-100 split, giving 2500 training and 500 test images per task. All model fitting, score computation, and sample selection use the training split only; the reported ϵ_1 and ϵ_2 are test MSE.

We approximate $\mathrm{col}(A_\mu)$ by the top-k PCA subspace of each task's features. For task-1, let $V_1 \in \mathbb{R}^{768\times k}$ denote this basis and $P_\perp = I - V_1 V_1^\top$. We use a task-subspace rank of (k=20), which retains approximately 96.1% held-out accuracy on five-class CIFAR-100 classification, compared with 96.8% using the full feature representation, averaged over 100 task draws.

Estimation of ρ_a and ρ_b. To estimate feature similarity, we use the canonical correlations between the two tasks' rank-k PCA subspaces. In the structured model, these concentrate near ρ_a along shared directions and near zero otherwise (Eq. (14)). We therefore estimate $\hat{\rho}_a$ by averaging canonical correlations above 0.3, excluding the lower tail of the spectrum associated approximately with task-specific directions.

Readout similarity is estimated from independently fitted task heads as

$$\hat{\rho}_b = \frac{1}{\sqrt{\alpha}} \frac{\mathrm{tr}\big(V_1^\top W_1^\top W_2 V_1\big)}{\|W_1 V_1\|_F \, \|W_2 V_1\|_F}. \tag{35}$$

The ratio is the cosine similarity of the two readouts restricted to the task-1 subspace. Under the leading-order structured-model approximation, this cosine equals $\sqrt{\alpha}\rho_b$, so the prefactor corrects for the shared fraction. We use the equal-block convention $\alpha = 1/2$, consistent with the main-text model. Both similarity estimates are computed before sample selection.

We evaluate the estimators on simulated task pairs from our theoretical model with $\alpha = 1/2$, additionally adding isotropic Gaussian noise to the features to test sensitivity to deviations from the idealized setting. The noise standard deviation ranges over $\{0, 0.1, 0.3, 0.5, 1\}$ relative to the unit signal scale, with zero recovering the original setup.

We choose dimensions matching CIFAR-100: $N_x = D = 768$, $N = 2500$ training samples per task, and $N_y = 5$ outputs. We vary $N_s \in \{10, 20, 50\}$ around the fixed PCA rank $k = 20$ used for CIFAR-100, and sweep $\rho_a \in \{0.5, 0.7, 0.9\}$ and $\rho_b \in \{0, 0.2, 0.4, 0.6, 0.8\}$ with three independent draws per combination. Figure 8a shows that both estimators correlate strongly with the true similarities, with Pearson $r \geq 0.93$ across all tested latent dimensions at noise levels up to 0.5. At noise level 1, correlations remain at least 0.93 except for $\hat{\rho}_b$ at $N_s = 10$, where $r = 0.86$. Figure 8b plots individual estimates against the true parameters at noise level 0.3. Feature-similarity estimates lie close to the identity line for $N_s = 10$ and 20, but underestimate ρ_a at $N_s = 50$, where the retained PCA rank is smaller than the true latent dimension. Overall, the estimators track changes in task similarity well, although retaining fewer PCs than the true latent dimension can bias the estimated feature overlap.

C.2 Task fitting, reweighting, and selection

Let $X_\mu \in \mathbb{R}^{N_\mu \times D}$ contain the task-μ features as rows, and $Y_\mu \in \mathbb{R}^{N_\mu \times 5}$ the corresponding one-hot targets. The task-1 head is fit once per pair by ridge regression,

$$W_1^\top = \big(X_1^\top X_1 + \lambda_1 I\big)^{-1} X_1^\top Y_1, \qquad \lambda_1 = 10^{-3} \, \frac{\mathrm{tr}(X_1^\top X_1)}{D}. \tag{36}$$

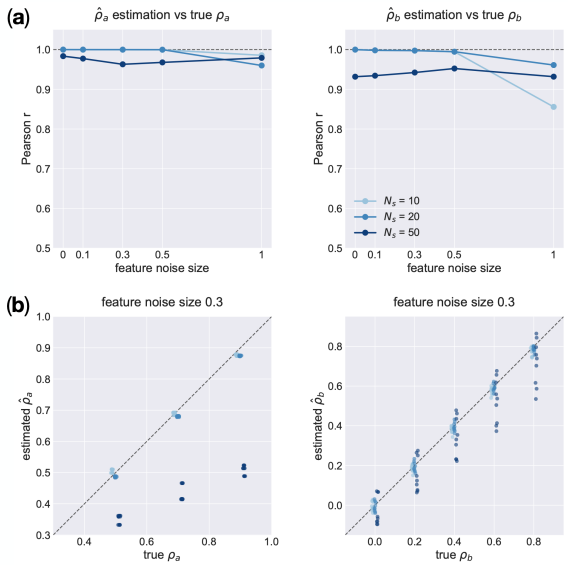

Figure 8: Task similarity estimation on simulated task pairs with CIFAR-100 dimensions, using fixed PCA rank $k = 20$ and latent dimensions $N_s \in \{10, 20, 50\}$. (**a**) Pearson correlations between estimated and true feature similarity (left) and readout similarity (right) across feature-noise levels. Each point summarizes 45 task pairs; dashed lines mark $r = 1$. (**b**) Estimated versus true similarities at feature-noise level 0.3. Each point represents one task pair, with colors indicating N_s as in (a); dashed lines indicate exact agreement. Despite high correlations, feature similarity is systematically underestimated when the latent dimension exceeds the PCA rank ($N_s = 50$, $k = 20$).

This is the empirical counterpart of the minimum-norm task-1 solution of Section 3, with a small ridge added only to condition the inverse.

The novelty and residual scores of Eq. (11) are then computed from W_1, and the task-1 PCA subspace on the full task-2 training set. For score-proportional reweighting, we normalize the scores to have mean 1,

$$\tilde{\sigma}_i = \frac{\sigma_i}{N^{-1} \sum_j \sigma_j},$$

and solve the finite-sample version of Eq. (2),

$$\left(X_2^\top \operatorname{diag}(\tilde{\sigma}) X_2 + \lambda I\right) W^\top = X_2^\top \operatorname{diag}(\tilde{\sigma}) Y_2 + \lambda W_1^\top. \tag{37}$$

Uniform reweighting corresponds to $\tilde{\sigma}_i = 1$ for all samples.

For hard selection at budget $B \in \{70\%, 50\%, 30\%\}$, we retain the top BN task-2 samples according to the same score and solve Eq. (37) on the retained subset with uniform weights. The budget-matched reference randomly selects the same number BN of samples and applies the identical solve.

Score normalization. Empirical feature norms vary across samples and can influence both novelty and residual scores. We divide both scores by $\|\phi(x)\|^2$ to reduce their sensitivity to input magnitude. For novelty, this gives the fraction of the squared feature norm lying outside the task-1 subspace; for residual scoring, it scales prediction error relative to the squared feature norm. We apply the same normalization before score-proportional reweighting and hard selection.

C.3 Matched-error protocol

Because scoring changes both task-1 and task-2 performance, we compare methods at matched task-2 test error, following the definition of G in Section 4. For each task pair and sampling rule, we first retain the monotone branch of the λ sweep: scanning from large to small λ, we keep points that set a new minimum in ϵ_2, excluding the small-λ region where test error can increase again due to overfitting. We then restrict to the overlapping ϵ_2 range of the scored and reference branches,

interpolate the reference task-1 error at each scored value of ϵ_2, and compute

$$G_{\text{rel}} = \frac{\epsilon_1^{\text{ref}} - \epsilon_1^{\text{score}}}{\epsilon_1^{\text{ref}}}.$$

The reference is uniform weighting for reweighting, and a random subset of the same size for hard selection. We report the relative retention gap G_{rel} in empirical experiments to account for differences in baseline task-1 error across task pairs, facilitating comparison of selection effects across tasks and datasets. To average across task pairs, each matched range is rescaled to its relative position along the reference trade-off curve, from the untouched W_1 to the best task-2 fit.

For each pair, G_{rel} is averaged over the middle 40-85% of the trade-off range. Near the task-2-optimal endpoint, the curves become nearly vertical and the matched overlap shrinks. We therefore truncate the upper edge to the last position at which all 50 task pairs contribute. This excluded region is shaded in the gap figures.

C.4 Supplementary results

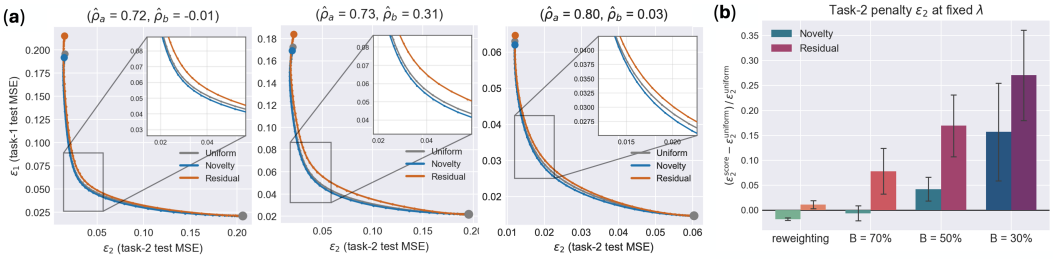

Figure 9: **CIFAR-100 main section supplementary results. (a)** Trade-off curves ϵ_1 vs. ϵ_2 for three more task pairs under uniform, novelty, and residual score-proportional reweighting, with each pair's estimated similarities above the panel. Insets magnify the bend of the trade-off, where novelty lies below uniform and residual above it, as predicted in the conflict regime. Circles mark the sweep endpoints, W_1 at large λ (lower right) and the full task-2 fit at $\lambda \to 0$ (upper left). **(b)** Relative task-2 test-error penalty against uniform at the fixed reference scale (Appendix C.4), for reweighting and hard selection at budgets $B \in \{70\%, 50\%, 30\%\}$. Error bars are 95% CIs over the 50 task pairs.

Fig 9a shows trade-off curves for three more example task pairs in the same style as Fig. 4a.

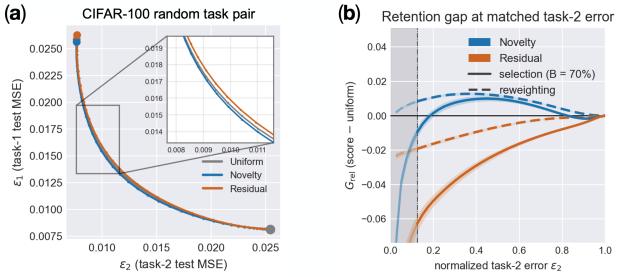

Figure 10: **Larger-output CIFAR-100 experiments.** Each task contains 50 classes with 100 images per class. **(a)** Task-1 versus task-2 test-error curves for one random task pair. The smaller separation under reweighting is consistent with residual scores becoming more uniform as output dimension increases. **(b)** Relative retention gaps at matched task-2 error under reweighting (dashed) and hard selection at $B = 70\%$ (solid). Mean and 95%CI obtained over 50 random task pairs. The qualitative pattern resembles the five-class setting: novelty improves retention over most of the matched range, while residual scoring increases forgetting. Hard selection strengthens the residual disadvantage.

Fig. 9b reports the task-2 error penalty for the fixed-λ comparison shown in Fig. 4c. At fixed $\lambda = 10^{-2}\,\frac{\text{tr}(X_1^\top X_1)}{D}$ across budgets, novelty's mean relative reduction in task-1 error exceeds its mean relative increase in task-2 error.

We also repeat the CIFAR-100 experiment with 50 classes per task and 100 images per class (Fig. 10). The larger-output setting shows qualitatively similar behavior to the five-class experiment, that novelty generally improves retention at matched task-2 error, whereas residual scoring worsens it, although the gaps are smaller.

C.5 Nonlinear CIFAR-100 training

Here, we test whether the predicted score ordering extends to nonlinear gradient-based training under fixed data budgets, a setting closer to practical data pruning.

The nonlinear experiment keeps the continual structure of the linear set-up in Section 6 (two sequential five-class tasks with a single shared head) and replaces the anchored ridge solve by gradient training of a nonlinear classifier. Each image is embedded as the same ViT-B/16 feature, and the classifier consists of two dense MLP layers: a 256-dimensional hidden layer with ReLU followed by a 5-dimensional output passed through a softmax. The model is trained on the MSE loss with vanilla SGD without momentum, learning rate 0.1 and batch size 16. Each run consists of 30 training epochs on task-1 followed by 30 epochs on task-2. We evaluate this setup on 30 random task pairs. During task-2 training only a fixed budget fraction $B \in \{70\%, 50\%, 30\%\}$ of the task-2 training samples is used, selected by novelty score, residual score, or uniformly at random.

Figure 11 reports test accuracy for both tasks as a function of the global epoch, task-1 training for epochs 1 to 30 and task-2 training for epochs 31 to 60. At every budget, novelty selection retains task-1 best, random selection is intermediate, and residual selection forgets most, the same ordering the theory predicts for the conflict regime, while all three rules reach comparable task-2 accuracy. The separation grows as the budget tightens and is largest at $B = 0.3$.

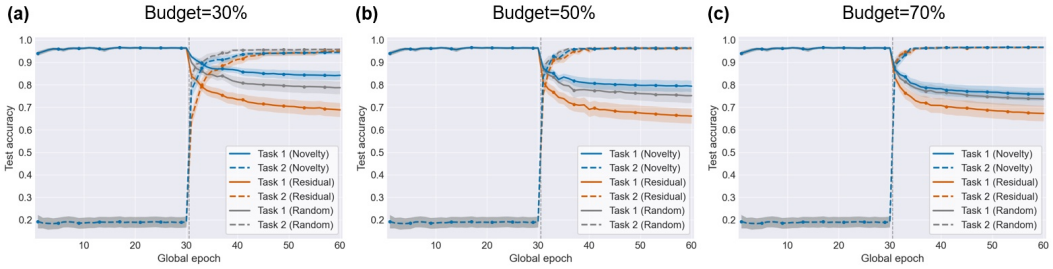

Figure 11: **Nonlinear CIFAR-100 results under task-2 data budgets $B \in \{70\%, 50\%, 30\%\}$.** Test accuracy against global training epoch for novelty, residual, and random task-2 selection. Shaded regions are the SEM over 30 test tasks.

D Experimental details: spanning the task-similarity plane (CUB-200 and AwA2)

CUB-200 and AwA2 reuse the CIFAR-100 machinery, the anchored solve of Appendix C.2 and the matched-error protocol of Appendix C.3. This section states the datasets, task constructions, and estimators, and the points where the protocol differs.

D.1 Datasets, features, and projection

CUB-200 contains 200 bird species, each annotated with class-level and per-image visual attributes organized into semantic groups (back color, crown color, and so on). We use class-level continuous attribute vectors as regression targets, assigning the same target vector to all images of a class. AwA2 contains 50 animal classes, each annotated with a class-level vector of semantic attributes. Both annotations are continuous, rescaled to $[0, 1]$, and used without binarization. CUB-200 uses its official train/test split. AwA2 only has a class-level zero-shot split, so we use a deterministic per-class 80/20 split. As on CIFAR-100, every head is fit on the training side and every reported error is held-out test MSE.

The backbone is a frozen ViT-S/16 ($D = 384$), from the same supervised ImageNet pretraining family as the ViT-B/16 used on CIFAR-100 (Dosovitskiy et al., 2020; Steiner et al., 2021). The smaller embedding is a practical necessity for CUB-200, where a task carries only around 300 training images, and is kept for AwA2 for consistency. For the same reason, a global PCA to 100 dimensions is applied before pair-level computation.

Both datasets use the similarity estimators presented in Appendix C.1. We use a task-subspace rank of $k = 10$. Held-out attribute-regression performance plateaus at this rank: on AwA2 (eight attribute targets; task-1s of 100 experiment pairs, in each pair's 100-dimensional PCA space), R^2 is 0.896 at $k = 10$, compared with 0.915 using the full 100-dimensional representation; on CUB-200 (ten-species tasks, eight attribute targets averaged over 15 attribute groups, in the 100-dimensional global PCA space), R^2 is 0.882 at $k = 10$, compared with 0.883 using the full representation. Values are averaged over 100 task draws.

D.2 Task construction

For each task pair, we draw disjoint classes from the same dataset, and both tasks regress onto 8 continuous attribute targets through a shared linear head $W \in \mathbb{R}^{8\times 100}$ on the projected features, the attribute-regression counterpart of the shared classification head on CIFAR-100. On CUB-200, tasks are random groups of 10 species. Each task reads out one semantic attribute group, keeping the top eight attributes from that group, and task-2's group differs from task-1's. Sampling from 100 species groups, 15 attribute groups, and 2 partner groups per condition gives 3000 conditions. The two attribute groups are different but often near-synonymous; for example, back and crown color share most of their selected attributes. Thus, CUB-200 readouts differ by body part while largely sharing the feature content on which the outputs rely.

On AwA2, tasks are disjoint class subsets. Because AwA2 classes have very unequal image counts, each task contains 3 to 10 classes so that every task has a comparable amount of training data, capped at 1500 images per task. Both tasks regress onto the same 8 attributes, fixed across all pairs and pre-selected for readout transferability, which allows natural task pairs to reach the agreement regime. We use two ways to broaden coverage across the similarity plane. First, taxonomic similarity between the two tasks' class subsets can raise $\hat{\rho}_a$ and $\hat{\rho}_b$ together. Second, we vary readout similarity by permuting task-2 target columns. Equal proportions of pairs receive no permutation or a permutation of 2, 4, 6, or 8 output dimensions (0%, 25%, 50%, 75%, or 100% of the outputs). The permuted dimensions are chosen uniformly at random. This leaves the features, and hence $\hat{\rho}_a$, unchanged while lowering $\hat{\rho}_b$, extending coverage away from the diagonal. In total, we generate and report 3000 pairs.

We fit W_1 and an independent task-2 head once on the training split, using ridge parameter $\lambda = 10^{-2}\ \mathrm{tr}(X_1^\top X_1)/D$ and the procedure in Appendix C.2. The independent task-2 head is used only to estimate $\hat{\rho}_b$. We compare score-proportional reweighting with uniform weighting on both datasets. Hard selection is evaluated on AwA2 at budgets $B \in \{30\%, 50\%, 70\%\}$ against random subsets of the same size. We omit hard-selection experiments on CUB-200 because of its small number of training samples per class. The task-2-loss-matched retention gap is computed over the middle 40-60% of the trade-off, from W_1 toward the best task-2 fit.

D.3 Supplementary results

Fig. 12 reports AwA2 retention gap under hard selection at budgets $B \in \{30\%, 50\%, 70\%\}$. Novelty generally improves retention, whereas residual selection increases task-1 forgetting most strongly where feature overlap is high and readout similarity is low.

We additionally evaluate larger output spaces using the full 15 attributes within an attribute group on CUB-200 and all 85 attributes on AwA2, with 3000 and 5000 task-pair conditions, respectively (Fig.13, 14). Novelty generally improves retention, while residual reweighting increases forgetting across the sampled CUB-200 conditions and produces both benefits and harms on AwA2. Breaking down the AwA2 results by output-permutation level reveals that residual scoring increases forgetting more under more extensive remapping and strong feature overlap (Fig.14). These observations support the qualitative dependence on task relationships.

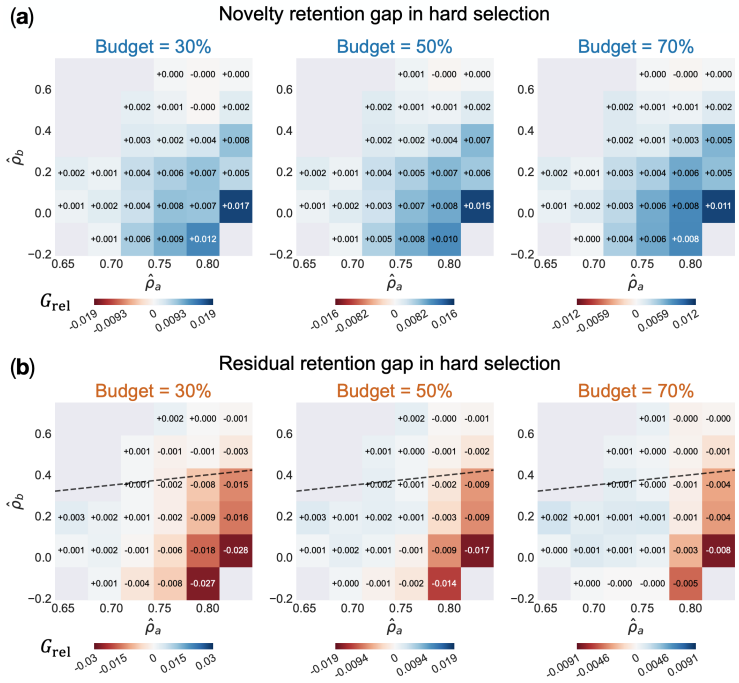

Figure 12: **Hard selection on AwA2 across selection budgets.** Matched-error retention gap G_{rel} relative to random subsets of the same size, binned on the estimated $(\hat{\rho}_a, \hat{\rho}_b)$ plane. **(a)** Novelty selection. **(b)** Residual selection. Columns show budgets $B = 30\%, 50\%, 70\%$ from left to right. Positive values indicate better task-1 retention than the random reference. Cell values are averaged over task pairs within each bin; dashed lines mark the estimated theoretical boundary $\rho_b = \rho_a/2$.

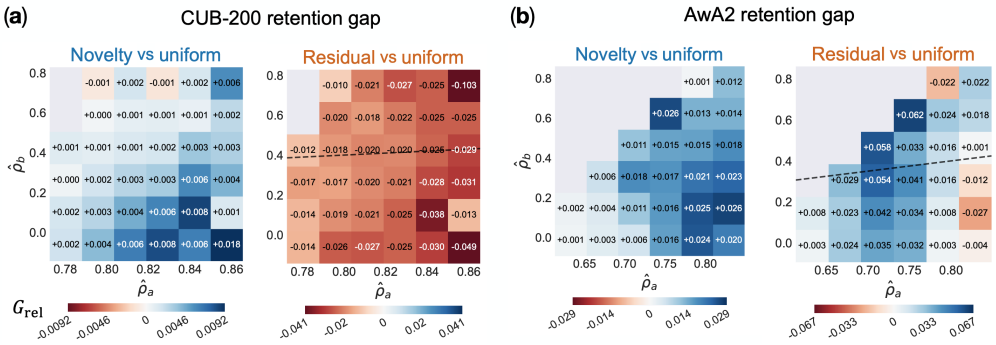

Figure 13: **Retention gaps with larger attribute output spaces.** Relative retention gap G_{rel} at matched task-2 test error under score-proportional reweighting, compared with uniform weighting. **(a)** CUB-200 with 15 outputs and 3000 task-pair conditions. **(b)** AwA2 with 85 outputs and 5000 task-pair conditions. Cell values are averages within bins of estimated task similarity. Dashed lines mark the estimated theoretical boundary $\hat{\rho}_b = \hat{\rho}_a/2$ for reference.

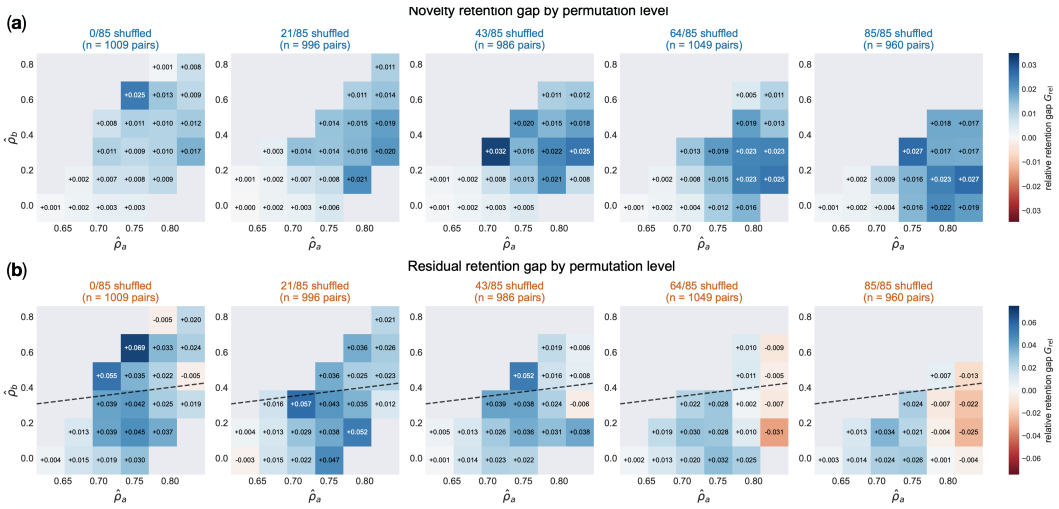

Figure 14: **AwA2 retention gaps by output-permutation level.** The 85-output experiment is partitioned by the number of task-2 output coordinates selected for permutation: 0, 21, 43, 64, and 85 (columns, left to right). Panel titles report the number of task pairs in each group. **(a)** Novelty reweighting yields overall positive mean retention gaps across the displayed bins at every permutation level. **(b)** Residual reweighting exhibits a stronger dependence on task relationships: negative gaps become more prominent under extensive output permutation, particularly at high feature similarity. Color scales are shared across columns within each row. Dashed lines mark $\hat{\rho}_b = \hat{\rho}_a/2$.