The Structural Bias of ℓ_2-Regularized Cross-Entropy Heads in Class-Incremental Learning: Characterization, Limits, and a Gauge-Anchoring Fix

Milad Khademi Nori
Department of Computer Science
Toronto Metropolitan University
Toronto, ON, Canada
mkn@torontomu.ca

Guanghui Wang*
Department of Computer Science
Toronto Metropolitan University
Toronto, ON, Canada
wangcs@torontomu.ca

Abstract

Class-incremental learners that train an ℓ_2-regularized softmax head separately on each task inherit a structural cross-task bias that no data or tuning removes. Assuming Gaussian features with shared covariance, we characterize the bias in closed form. Each task's head anchors at its centroid, so all cross-task decisions shift by one fixed offset, identical for all class pairs. The offset survives vanishing weight decay and is invariant to sample size and task order. That invariance makes the diagnosis falsifiable, and the correction is applied as one line at evaluation time. The margin a stronger backbone provides is the same quantity that enlarges the offset. We construct geometries where cross-task error stays bounded away from zero at any margin, and geometries where it tends to one. Under a testable separation condition, confusion decays exponentially. *Post-hoc de-centering* adds the offset to each task's weights and one scalar to its thresholds, matching the Bayes-optimal cross-task discriminator in the population limit. The same operation underlies LDA-style heads (SLDA, FeCAM, NCM). We identify it as the per-task affine correction that eliminates the structural bias, supported by an empirical diagnostic on five published methods and a finite-sample analysis. On the benchmarks the released protocol regenerates, the correction's effect is mixed and is reported as measured: the ordering WT $<$ FIX $<$ J holds on the two DomainNet configurations (DINOv2 $50.6 < 53.1 < 68.3$; CLIP $53.8 < 62.9 < 71.5$), with the scale-matched correction recovering the gap on one of the three DomainNet backbones (DINOv2, 93%) and over-correcting on the other two, while on CIFAR-100 the unscaled correction lands 24 to 40 points below the accepted values and below the within-task baseline. The accepted recovery range of 82–94% of the joint–sequential gap (mean 90.5%) is not reproduced under the released protocol and is withdrawn; the benchmark rows the released code cannot regenerate are withdrawn to the provenance ledger, and every remaining empirical number is regenerated from the released caches.

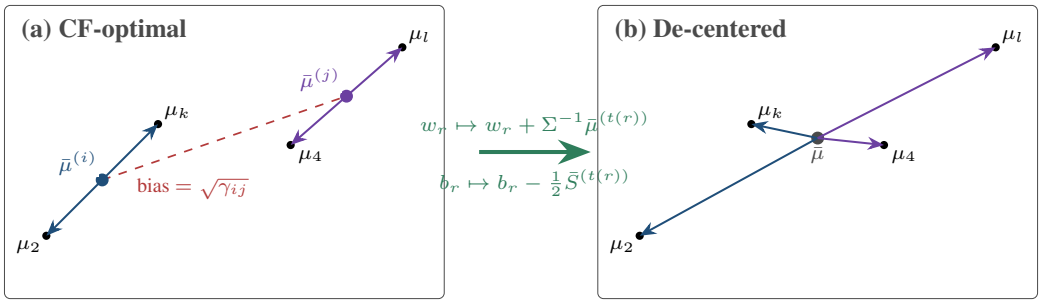
Code available at `https://github.com/miladkhademinori/structural-bias`

1 Introduction

Class-incremental learning (CIL) with frozen pretrained backbones is the dominant continual-classification paradigm, with prompt-based [Wang et al., 2022b,a, Smith et al., 2023] and adapter- or projection-based [Zhang et al., 2023, McDonnell et al., 2023, Zhou et al., 2024b] families routinely

*Presenting author.

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



Each task's weights radiate from its centroid $\bar{\mu}^{(t)}$; cross-task discriminator inherits offset.

All weights radiate from $\bar{\mu}$; cross-task discriminator matches Bayes-optimal.

Figure 1: Geometry of post-hoc de-centering (Definition 5.1). **(a)** Under within-task softmax with ℓ_2 regularization, the CF-optimal head fixes the gauge per task (Lemma 3.2); task i's weights $w_r^\dagger$ radiate from $\bar{\mu}^{(i)}$, task j's from $\bar{\mu}^{(j)}$, so every cross-task discriminator inherits the weight-space offset $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$ with Σ-norm $\sqrt{\gamma_{ij}}$ (Theorem 3.6). **(b)** The correction rigidly translates each task's weight cloud so that all weights radiate from the global origin $\bar{\mu}$. The class means and the within-task discriminator differences (Remark M.1) are unchanged; cross-task discriminators $w_k^{\text{fix}} - w_l^{\text{fix}} = \Sigma^{-1}(\mu_k - \mu_l)$ match the Bayes-optimal head exactly (Theorem 5.2).

reaching 80–90% final accuracy [Masana et al., 2020, Zhou et al., 2024a]. We argue this picture obscures a structural issue in the standard head, the ℓ_2-regularized softmax cross-entropy classifier trained on one task at a time.

The head carries a population-level bias on cross-task discriminators. The bias survives vanishing weight-decay strength λ, unlimited per-task data (sample size n_t), and unbounded inter-task margin $\gamma_{\min}$, the smallest squared Mahalanobis distance between task centroids (Section 2, (1)). Worse, the bias grows as the backbone improves: its size scales with that same margin (Theorem 3.6). On unfavourable feature geometries, the cross-task error converges to 1, however strong the backbone.

The bias has a closed form under our working assumption (Assumption 2.1): backbone features of each class are Gaussian with a shared covariance Σ, and the CF-optimal weight for class r (Definition 2.4) is $w_r^\dagger = \Sigma^{-1}(\mu_r - \bar{\mu}^{(t)})$, with μ_r the class feature mean and $\bar{\mu}^{(t)}$ the average over the classes of task t. The softmax leaves each task free to choose its own weight origin; regularization pins that origin to the task centroid (the gauge). Joint training would pin all weights to one global origin, and Figure 1 draws the mismatch.

For two classes from different tasks, the cross-task discriminator — the pair (weight difference, threshold difference) — alone decides every comparison of those two classes. For the sequentially trained head it is off by one fixed vector, $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$, of size $\sqrt{\gamma_{ij}}$, the inter-task centroid distance (Theorem 3.6); the same vector for every class pair in the block (Corollary 3.7), surviving vanishing weight decay and invariant to sample size and task order (Corollary 3.8). Hence the paradox of the opening paragraph: the same improvement that enlarges the margin enlarges the bias.

Whether the bias is destructive depends on geometry. When cross-class distances exceed inter-task distances by a factor $\kappa > 1$ (squared: $D_{kl} \geq \kappa^2 \gamma_{ij}$; Assumption 4.5), task confusion decays exponentially in the margin; the rate matches prototype heads at $\kappa = 2$ (Theorem 4.6). As κ approaches 1, we construct configurations where confusion stays bounded away from zero or tends to one, no matter the margin (Theorems 4.1 and 4.2, Figure 5). No unconditional bound exists; a head-level modification is required. The measured cross-task rates against this prediction are exhibited in Figure 2.

We propose *post-hoc de-centering* (Definition 5.1): an evaluation-time correction that adds each task's centroid offset to its weights and one scalar to its thresholds; training is unchanged. Theorem 5.2 proves the de-centered head matches the Bayes-optimal cross-task discriminator, restoring an unconditional bound on task confusion (Corollary 5.3). Our contributions are threefold: a gauge-theoretic derivation of the correction from the structural bias on cross-entropy heads; a falsifiable empirical diagnostic across five published CIL methods; and a finite-sample analysis whose $n_t^{-1/4}$ rate is a signature of the bias.

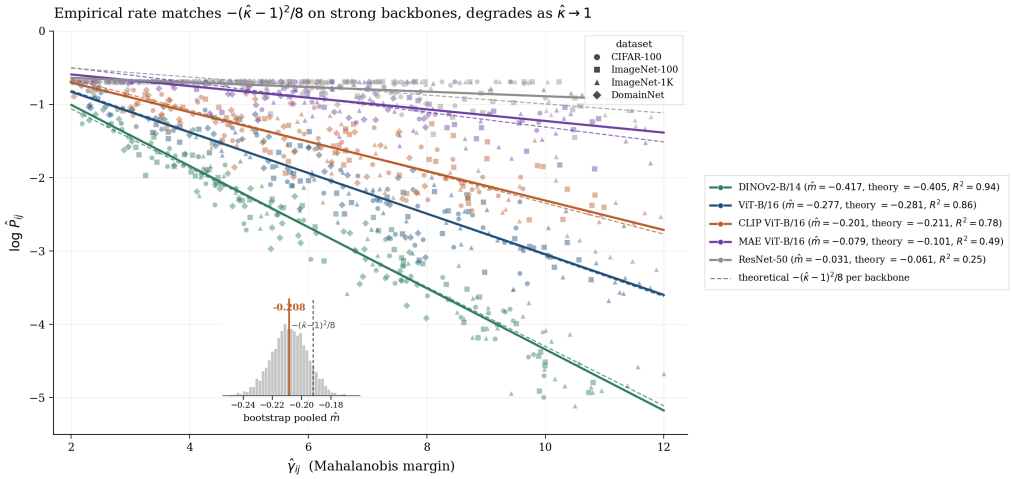


Figure 2: Empirical $\log \hat{P}_{ij}$ versus inter-task margin $\hat{\gamma}_{ij}$ across 1000 cross-task pairs (200 per backbone, aggregated over CIFAR-100, ImageNet-100, ImageNet-1K, and DomainNet); DomainNet enters as the shared-label, domains-as-tasks stress test that violates the disjoint-label-set assumption (Section 6). Solid lines are per-backbone OLS fits; dashed lines are theoretical predictions $-(\hat{\kappa}-1)^2/8$ from Theorem 4.6, anchored at each backbone's data centroid. The exhibit plots the accepted fits. Under the released protocol the regenerated slopes are negative on every backbone and shallower than the kappa-derived predictions by factors of 1.7 to 3.3 (R^2 0.32 to 0.81), and the accepted agreement statements are withdrawn to the provenance ledger. The inset shows the distribution of pooled bootstrap slope estimates ($n = 2000$ resamples) of the accepted pooled slope (-0.208) against the theoretical pooled prediction ($-(\bar{\kappa}-1)^2/8 = -0.192$ for mean $\bar{\kappa} = 2.24$).

The empirical riders on these contributions are reported as measured under the released protocol wherever it regenerates them; rows without a run record carry per-row provenance in the tables, and accepted values sit in the provenance ledger. The accepted gap-recovery range of 82–94% (mean 90.5) is not reproduced: on CIFAR-100, de-centering lands 24 to 40 points below the accepted values and below the uncorrected head (46.2 to 62.5 against accepted 85.7 to 86.5); on DomainNet the ordering WT $<$ FIX $<$ J holds ($50.6 < 53.1 < 68.3$ DINOv2; $53.8 < 62.9 < 71.5$ CLIP) and the scale-matched correction recovers the gap on one of three backbones (DINOv2, 93%), over-correcting on the other two. The accepted bias-direction cosine of at least 0.92 corresponds to neither regenerated statistic family (pair-averaged 0.986 to 0.9997, the gauge identity; per-pair 0.22 to 0.35 with cross-pair CV 0.21 to 0.32), so agreement is tested and reported per pair and the accepted floor is withdrawn. The regenerated OLS rate slopes are negative on every backbone and shallower than the kappa-derived prediction by factors of 1.7 to 3.3 (CLIP -0.045 against -0.0755; ViT -0.053 against -0.176; R^2 0.32 to 0.81), and the within-5% agreement claim is withdrawn. The diagnostic recovers ~ 6 accuracy points on pure within-task CE methods (Table 2); its released checkpoints were not retrained. The correction also has a precise connection to classical discriminant methods, which we now state.

Relation to LDA-based class-incremental classifiers. Several CIL methods use shared-covariance LDA discriminants as classifier heads (SLDA [Hayes and Kanan, 2020], FeCAM [Goswami et al., 2023], NCM heads in the iCaRL family [Rebuffi et al., 2017]); post-hoc de-centering (Definition 5.1) produces exactly this discriminant in the population limit. Our contribution is not the discriminant but the identification of the closed-form structural bias of the trained head, of which this discriminant is the unique per-task affine correction (Theorem 5.2), together with the diagnostic apparatus (κ, cosine, predicted slope) and the finite-sample analysis (Theorem G.1) absent from prior LDA-CIL work (extended remarks in Section M).

Paper roadmap. Section 2 fixes notation and the working assumption. Section 3 derives the closed form and the structural bias (Theorems 3.3 and 3.6); Section 4 establishes the negative constructions and the conditional positive rate. Section 5 introduces post-hoc de-centering, proves its exactness, and states the finite-sample schedule. Section 6 reports the regenerated benchmark record, the diagnostic

on five published methods, and the controls. Proofs and extended material occupy Section A onward. Throughout, withdrawn accepted-run values are named as such and live in the provenance ledger; every remaining number is regenerated from the released caches.

2 Setup

We consider class-incremental learning over T tasks with disjoint task-specific label subsets $\mathcal{Y}^{(t)}$ of size C, $\mathcal{Y} = \bigcup_t \mathcal{Y}^{(t)}$, $|\mathcal{Y}| = N = TC$, uniform global prior $\pi_y = 1/N$. At phase t, the learner observes (x, y) with $y \in \mathcal{Y}^{(t)}$ and produces, after phase T, a single classifier over the full label set. A frozen pretrained backbone $\phi : \mathcal{X} \to \mathbb{R}^d$ is given. For each class r write $\mu_r := \mathbb{E}[\phi(x) \mid y = r]$, per-task centroids $\bar{\mu}^{(t)} := \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r$, and global centroid $\bar{\mu} := \frac{1}{N}\sum_r \mu_r$. The inter-task Mahalanobis margin and cross-class distance are

$$\gamma_{ij} := \left\|\bar{\mu}^{(i)} - \bar{\mu}^{(j)}\right\|^2_{\Sigma^{-1}}, \qquad D_{kl} := \|\mu_k - \mu_l\|^2_{\Sigma^{-1}}, \qquad \gamma_{\min} := \min_{i\neq j} \gamma_{ij}, \tag{1}$$

where Σ is the pooled within-class covariance under the following assumption. A complete index of the notation, with defining formulas, is given in Table 3 in Section A.

Assumption 2.1 (Gaussian class-conditionals, shared full-rank covariance)**.** All of the following hold: (i) Gaussian features: $\phi(x) \mid y = r \sim \mathcal{N}(\mu_r, \Sigma_r)$ for every class $r \in \mathcal{Y}$; (ii) shared covariance: $\Sigma_r = \Sigma$ for all $r \in \mathcal{Y}$; (iii) full rank: $\Sigma \succ 0$.

Practical scope. Assumption 2.1 describes the regime where frozen layer-normalized backbones empirically sit: approximately Gaussian class-conditional features with near-shared covariance, which is why the analysis applies directly to natural-image benchmarks. On those benchmarks the bias direction is tested per pair under the released protocol, with both statistic families and their accepted-value withdrawals reported in Section 6 (Table 6).

The assumption fails in two measurable ways — heteroskedasticity (the Σ_r differ across classes; Theorems 3.3 and 3.6 hold only to first order in $\|\Sigma_r - \Sigma\|$) and aliasing (the no-aliasing constant κ of Assumption 4.5, computable from training features before any head is trained; pure within-task CE methods leave 3–8% of cross-task pairs with $\hat{\kappa} < 1$, exactly where Corollary 4.3 rules out a uniform bound). Both failure modes carry empirical signatures and are checkable in advance, so the theory's applicability is decidable before committing to a training run; the signatures are tabulated in Section A.

The classifier head is linear over backbone features: $\ell_r(x) := w_r^\top \phi(x) + b_r$. The standard ℓ_2-regularized softmax cross-entropy with weight-decay strength $\lambda > 0$ on a class-y batch is

$$\mathcal{L}_{\text{full}}(W, b; x, y) = -\log \frac{\exp(\ell_y(x))}{\sum_{r\in\mathcal{Y}} \exp(\ell_r(x))} + \tfrac{\lambda}{2}\sum_r \|w_r\|^2 + \tfrac{\lambda}{2}\sum_r b_r^2, \tag{2}$$

which we call *full softmax*. We regularize biases as well as weights with the same λ; this breaks the residual 1-dimensional bias gauge in within-task training and selects $\sum_{r\in\mathcal{Y}^{(t)}} b_r = 0$. Class-incremental learners cannot evaluate (2) since future classes are unseen and rehearsal-free methods discard past data; essentially every CE-based CIL method since iCaRL [Rebuffi et al., 2017] therefore trains with a *within-task softmax*:

$$\mathcal{L}^{(t)}_{\text{within}}(W, b; x, y) = -\log \frac{\exp(\ell_y(x))}{\sum_{r\in\mathcal{Y}^{(t)}} \exp(\ell_r(x))} + \tfrac{\lambda}{2}\sum_{r\in\mathcal{Y}^{(t)}} (\|w_r\|^2 + b_r^2), \tag{3}$$

where partition and regularizer range only over the current task's classes; heads for previously seen or future classes are not updated.

Definition 2.2 ($\lambda \to 0^+$ population limit)**.** Fix weight-decay strength $\lambda > 0$ and per-task sample size n_t. The population limit of sequential within-task training under (3) is the iterated limit taken in two stages: (i) $n_t \to \infty$: each phase-t sub-head is replaced by the minimizer of the population within-task risk $\mathbb{E}_{(x,y)\sim p^{(t)}}[\mathcal{L}^{(t)}_{\text{within}}]$, where $p^{(t)}$ is the distribution of (x, y) restricted to labels $y \in \mathcal{Y}^{(t)}$; (ii) that minimizer converges as the weight decay vanishes ($\lambda \to 0^+$). Closed-form statements about the limiting head, in particular Theorems 3.3 and 3.6, hold at stage (ii) unless stated otherwise; finite-λ and finite-n_t deviations are quantified in Section 5.

We follow Khademi Nori and Kim [2024] in decomposing the population risk into pairwise losses:

$$P_{kl} := \Pr_{x\sim p(\cdot|y=k)}[\ell_l(x) > \ell_k(x)], \qquad |P_{ij}| := \sum_{k\in\mathcal{Y}^{(i)}, l\in\mathcal{Y}^{(j)}} P_{kl}, \tag{4}$$

the probability the head ranks the wrong class l above true class k, summed over off-diagonal task blocks. Controlling $|P_{ij}|$ for $i \neq j$ is the structural problem of class-incremental learning.

Definition 2.3 (Cross-task discriminator)**.** Let (W, b) be a head and $k \neq l$ two classes. The discriminator between k and l is the pair $(w_k - w_l,\ b_k - b_l)$: its first entry is the weight part, its second the threshold part. It determines every comparison of k against l, since $\ell_k(x) > \ell_l(x)$ holds exactly when $(w_k - w_l)^\top \phi(x) + (b_k - b_l) > 0$. If $t(k) \neq t(l)$, say $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$, $i \neq j$, the pair (k, l) is a cross-task class pair and $(w_k - w_l, b_k - b_l)$ a cross-task discriminator; otherwise it is within-task. Under Assumption 2.1 with uniform priors, the Bayes-optimal discriminator has weight part $\Sigma^{-1}(\mu_k - \mu_l)$ and threshold part $-\frac{1}{2}(\mu_k^\top \Sigma^{-1}\mu_k - \mu_l^\top \Sigma^{-1}\mu_l)$.

Definition 2.4 (CF-optimal head)**.** The catastrophic-forgetting-optimal (CF-optimal) head $(W^\dagger, b^\dagger)$ is the population limit of sequential within-task training (the $n_t \to \infty$ stage of Definition 2.2): $(w_r^\dagger, b_r^\dagger)$ for $r \in \mathcal{Y}^{(t)}$ minimize $\mathbb{E}_{(x,y)\sim p^{(t)}}[\mathcal{L}^{(t)}_{\text{within}}]$ at phase t ($p^{(t)}$ is the distribution of (x, y) restricted to labels $y \in \mathcal{Y}^{(t)}$) and are never modified subsequently.

Any cross-task error of the CF-optimal head $(W^\dagger, b^\dagger)$ of Definition 2.4 is a property of the within-task objective itself, not of optimization difficulty, finite samples, or sequential update interference.

The geometric view. Under within-task training, weights radiate from the per-task centroid $\bar{\mu}^{(t)}$; jointly, from the global centroid $\bar{\mu}$ (Theorem 3.3, Remark 3.5); the bias is the $\sqrt{\gamma_{ij}}$ origin mismatch (Section 3); Figure 1 draws both with the correction (Definition 5.1).

One negative example. Figure 5 previews why no unconditional bound can hold: an anti-aligned task pair keeps the cross-task error flat at $\Phi(-1/\sqrt{2}) \approx 0.24$ for every inter-task margin (Theorem 4.1), and a single outlier class drives it to 1 (Theorem 4.2); the Bayes head eliminates the bias in the first case only.

3 The Closed Form and the Structural Bias

We derive the CF-optimal head in three steps — a block-decoupling lemma, a gauge-fix lemma, and a closed-form theorem under Assumption 2.1 (proofs in Section B) — and read off the structural bias it induces on cross-task discriminators.

Lemma 3.1 (Block decoupling)**.** *Let $(W^\dagger, b^\dagger)$ be the CF-optimal head of Definition 2.4. For each task t, $\{(w_r^\dagger, b_r^\dagger) : r \in \mathcal{Y}^{(t)}\}$ are the unique minimizers of*

$$\mathcal{R}^{(t)}(W^{(t)}, b^{(t)}) := \mathbb{E}_{(x,y)\sim p^{(t)}}\left[-\log \frac{\exp(\ell_y(x))}{\sum_{r\in\mathcal{Y}^{(t)}} \exp(\ell_r(x))}\right] + \tfrac{\lambda}{2} \sum_{r\in\mathcal{Y}^{(t)}} (\|w_r\|^2 + b_r^2), \tag{5}$$

and depend on no class statistics outside $\mathcal{Y}^{(t)}$.

Lemma 3.1 is the operational meaning of CF-optimal training: each task's sub-head sees only its own data and classes, so the cross-task structure of the head is determined by the per-task subproblems alone.

Lemma 3.2 (Gauge fix)**.** *At the minimizer of* (5)*, the per-task weights and biases satisfy $\sum_{r\in\mathcal{Y}^{(t)}} w_r^\dagger = 0$ and $\sum_{r\in\mathcal{Y}^{(t)}} b_r^\dagger = 0$.*

The within-task softmax has a $(d + 1)$-dimensional gauge freedom (joint shifts $w_r \mapsto w_r + v$, $b_r \mapsto b_r + c$ leave its probabilities unchanged); the ℓ_2 regularizer on both W and b breaks it, selecting the unique minimum-norm representative on the gauge orbit.

Theorem 3.3 (Closed form of the CF-optimal head)**.** *Assume (H1) Assumption 2.1 (Gaussian class-conditionals with shared full-rank covariance), and (H2) the $\lambda \to 0^+$ population limit of Definition 2.2. Then, for every task t and every $r \in \mathcal{Y}^{(t)}$, the CF-optimal head satisfies*

$$w_r^\dagger = \Sigma^{-1}(\mu_r - \bar{\mu}^{(t)}), \qquad b_r^\dagger = -\tfrac{1}{2}\mu_r^\top \Sigma^{-1}\mu_r + \tfrac{1}{2C} \sum_{s\in\mathcal{Y}^{(t)}} \mu_s^\top \Sigma^{-1}\mu_s, \tag{6}$$

at stage (ii) of (H2).

In words: the within-task head weights are the class means relative to the task centroid, whitened by the shared covariance and shrunk by the weight decay.

Remark 3.4 (Fixed-λ shrinkage). At fixed $\lambda > 0$, before stage (ii) of (H2), both coordinates of (6) hold up to a first-order multiplicative shrinkage by $(1 - \lambda\alpha_t)$, where $\alpha_t > 0$ is a task-specific constant that does not depend on r; higher-order corrections in λ are not characterized.

Theorem 3.3 is the LDA discriminant centered at the per-task mean [Efron, 1975]; the centering $\bar{\mu}^{(t)}$ is local to task t, generally different from the global centroid $\bar{\mu}$, and no $\lambda > 0$ repositions it (first-order shrinkage preserves the direction of $w_r^\dagger$).

Remark 3.5 (Comparison with full softmax). Under full softmax (2) on jointly available data, the analogous calculation gives $w_r^\star = \Sigma^{-1}(\mu_r - \bar{\mu})$ at the global centroid $\bar{\mu}$. The two heads differ by the inter-task offset $w_r^\star - w_r^\dagger = \Sigma^{-1}(\bar{\mu}^{(t(r))} - \bar{\mu})$. This offset is the only difference between joint training and CF-optimal sequential training in the population.

The mismatch between the per-task centering $\bar{\mu}^{(t)}$ of Theorem 3.3 and the global centering of joint training (Remark 3.5) produces a population-level bias on cross-task discriminators. Comparing the CF-optimal head with the Bayes-optimal cross-task discriminator under Assumption 2.1 with uniform priors [Duda et al., 2000] pins it down.

Theorem 3.6 (Structural bias of the CF-optimal head)**.** *Assume (H1) Assumption 2.1 (Gaussian class-conditionals with shared full-rank covariance), and (H2) the $\lambda \to 0^+$ population limit of Definition 2.2. Then the weight part of the cross-task discriminator (Definition 2.3) at the CF-optimal head differs from its Bayes-optimal counterpart $\Sigma^{-1}(\mu_k - \mu_l)$ by*

$$\big(w_k^\dagger - w_l^\dagger\big) - \Sigma^{-1}\big(\mu_k - \mu_l\big) = \Sigma^{-1}\big(\bar{\mu}^{(j)} - \bar{\mu}^{(i)}\big), \tag{7}$$

for every cross-task class pair $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$, $i \neq j$. The offset is one fixed vector per task block (Corollary 3.7) and has Σ-norm $\sqrt{\gamma_{ij}}$, where for weight vectors $\|w\|_\Sigma^2 := w^\top \Sigma\, w$ (the threshold gap is resolved two paragraphs below).

The proof is direct from Theorem 3.3 (full derivation in Section C). The inter-task margin γ_{ij} is the quantity a stronger backbone is supposed to make large, and the concurrent prototype-based envelope for task confusion decays as $\exp(-\gamma_{ij}/8)$ (our shared-covariance convention for γ_{ij} differs from the pair-averaged convention of Khademi Nori and Wang [2026]); Theorem 3.6 says the cross-entropy head's bias *grows* with the same quantity.

Corollary 3.7 (Class-pair invariance)**.** *Under (H1)–(H2) (Assumption 2.1 and the $\lambda \to 0^+$ population limit of Definition 2.2), the bias $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$ is the same vector for all C^2 cross-task class pairs $(k, l) \in \mathcal{Y}^{(i)} \times \mathcal{Y}^{(j)}$.*

Corollary 3.7 reduces the cross-task block of the head from a C^2-parameter object to a single d-dimensional offset per task pair: every cross-task discriminator in block (i, j) is translated by the same fixed vector, and a single per-task correction can in principle cancel the entire C^2-block of bias simultaneously (Section 5).

Corollary 3.8 (Invariants of the bias)**.** *The bias $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$ in* (7) *is independent of (i) the weight decay as the decay vanishes (at fixed decay, the offset is preserved to first order), (ii) the per-task sample size n_t in the population limit, (iii) task ordering, (iv) the class pair (k, l) within block (i, j), and (v) the inter-task margin γ_{ij}, except in magnitude: the direction does not depend on γ_{ij}, and the magnitude scales as $\sqrt{\gamma_{ij}}$.*

Item (v) is the most consequential: strengthening the backbone enlarges the magnitude of the bias proportionally to $\sqrt{\gamma_{ij}}$; whether this produces worse cross-task error is the subject of Section 4. The companion threshold gap, computed in Section C (equation (31)), depends only on the task pair (a scalar shift, not an angular bias) and is absorbed by the correction of Section 5.

4 Negative and Conditional Results

Theorem 3.6 establishes that the structural bias has magnitude $\sqrt{\gamma_{ij}}$ in Σ-norm. Whether this magnitude translates into cross-task misclassification depends on alignment: if the bias direction is orthogonal to the cross-class direction, it does not affect classification; if aligned, it rotates

the decision boundary directly across the cross-class margin. We exhibit two configurations: the first (Theorem 4.1) is anti-aligned and produces cross-task error converging to a positive constant independent of $\gamma_{\min}$; the second (Theorem 4.2) sharpens this to error converging to 1. Together they rule out any unconditional exponential bound (Corollary 4.3). The synthetic experiment of Section J verifies both predictions by direct sequential training and falsifies the alternative hypothesis that the bias is a finite-sample or optimization artifact.

Theorem 4.1 (Anti-aligned construction)**.** *There exists a family of configurations $(\phi_a)_{a>0}$ satisfying Assumption 2.1, with $T = 2$, $C = 2$, $d = 2$, $\Sigma = I$, and class means*

$$\mu_1 = e_1, \quad \mu_2 = -e_1, \quad \mu_3 = ae_1 + e_2, \quad \mu_4 = ae_1 - e_2, \tag{8}$$

such that $\gamma_{12}(\phi_a) = a^2 \to \infty$ as $a \to \infty$, and yet for every $\lambda \geq 0$ and every $a > 0$, $P_{13}(W^\dagger, b^\dagger; \phi_a) = \Phi(-1/\sqrt{2}) \approx 0.2398$ (writing $P_{kl}(W, b; \phi)$ for the P_{kl} of (4) with head and configuration explicit; by symmetry, the same value holds for P_{14}, P_{23}, P_{24}).

The Bayes-optimal error on this construction, $\Phi(-\frac{1}{2}\sqrt{(a-1)^2+1}) \to 0$ as $a \to \infty$, contrasts with the flat CF-optimal error of Theorem 4.1; the gap between the two curves is the structural bias drawn in Figure 5 (full proof in Section D.1).

Theorem 4.2 (Catastrophic construction)**.** *There exists a family of configurations $(\psi_a)_{a>0}$ satisfying Assumption 2.1, with $T = 2$, $C_1 = 2$, $C_2 = 3$ (per-task counts C_t as in Section A; the equal-size convention of Section 2 is relaxed here), $d = 2$, $\Sigma = I$, and class means*

$$\mu_1 = e_1, \quad \mu_2 = -e_1, \quad \mu_3 = e_2, \quad \mu_4 = -e_2, \quad \mu_5 = a\, e_1, \tag{9}$$

such that $\gamma_{12}(\psi_a) = a^2/9 \to \infty$ as $a \to \infty$, and yet, in the $\lambda \to 0^+$ population limit of Definition 2.2, $P_{13}(W^\dagger, b^\dagger; \psi_a) \to 1$ as $a \to \infty$.

Mechanism: the outlier $\mu_5 = ae_1$ inflates the per-task average squared norm $\bar{S}^{(2)}$, so the threshold $b_3^\dagger \sim a^2$ dominates the linear part and drives $P_{13} \to 1$; full proof in Section D.2, with P_{13} plotted against the Bayes baseline in Figure 5.

Corollary 4.3 (Failure of unconditional bounds)**.** *For every constant $c > 0$ and every $\beta > 0$, there exists a configuration satisfying Assumption 2.1 with $\gamma_{\min}$ arbitrarily large for which the CF-optimal CE head suffers $|P_{ij}| > c\exp(-\beta\gamma_{ij})$.*

Corollary 4.3 stands in contrast to the prototype-based head, whose cross-task error decays at the Bayes rate $\exp(-\min_{kl} D_{kl}/8)$ governed by cross-class Mahalanobis separation, with no dependence on the gauge mismatch of Theorem 3.6; the CE head's error is governed by the inter-task margin γ_{ij} precisely in directions where the bias aligns with cross-class differences. The envelope $e^{-\gamma_{ij}/8}$ (the negligible-spread operational rate of concurrent prototype work [Khademi Nori and Wang, 2026]) is met by the bound of Theorem 4.6 exactly at $\kappa = 2$ and bettered for $\kappa > 2$; the comparison is between bound techniques applied to different classifiers, not classifier accuracies. Any bound on the CE head must therefore either restrict the configuration space (Assumption 4.5 below) or modify the head to remove the gauge mismatch (Section 5).

Remark 4.4 (The bias direction is the inter-task direction). Both Theorems 4.1 and 4.2 share a structural feature: the cross-class direction $\mu_k - \mu_l$ is approximately aligned with the inter-task direction $\bar{\mu}^{(j)} - \bar{\mu}^{(i)}$. The bias of Theorem 3.6 is destructive precisely when this alignment is large; it is benign when the cross-class direction has substantial Σ^{-1}-component orthogonal to the inter-task direction. Assumption 4.5 and Theorem 4.6 below formalize the complementary dichotomy.

We now define this complementary regime — configurations in which no cross-class direction aligns too closely with any inter-task direction — and prove the CF-optimal CE head then admits an exponential bound on task confusion (Theorem 4.6), matching the prototype envelope at $\kappa = 2$ and tightening it for $\kappa > 2$.

Assumption 4.5 (No cross-task aliasing)**.** There exists $\kappa > 1$ such that for every pair of tasks $i \neq j$ and every cross-task class pair $k \in \mathcal{Y}^{(i)}, l \in \mathcal{Y}^{(j)}$, $D_{kl} \geq \kappa^2 \gamma_{ij}$.

Assumption 4.5 is geometric, refers only to backbone-induced class-conditional means, and is testable directly from ϕ on training data. The constant κ measures how cleanly classes from different tasks separate relative to the tasks themselves: $\kappa \to 1$ corresponds to the aliased configurations

of Theorems 4.1 and 4.2 (within-task neural collapse [Papyan et al., 2020]); $\kappa \gg 1$ to backbones whose class structure dominates task structure. Section 6 measures κ empirically: rarely close to 1 on broadly distinct semantic classes, approachable on hierarchical taxonomies.

Theorem 4.6 (Conditional bound on task confusion)**.** *Under (H1)–(H2) (Assumption 2.1 and the $\lambda \to 0^+$ population limit of Definition 2.2), Assumption 4.5, and within-task spreads bounded by $\max_t \bar{T}^{(t)} \le \bar{T}_{\max} < \infty$, where $\bar{T}^{(t)} := \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \left\|\mu_r - \bar{\mu}^{(t)}\right\|^2_{\Sigma^{-1}}$ is the per-task average squared centered class norm, for every $\varepsilon \in (0, \kappa - 1)$ there exists Γ_ε such that for every task pair $i \neq j$ with $\gamma_{ij} \ge \Gamma_\varepsilon$, the population CF-optimal CE head satisfies*

$$|P_{ij}(W^\dagger, b^\dagger; \phi)| \le \tfrac{C^2}{2}\,\exp\!\Big(-\tfrac{(\kappa-1-\varepsilon)^2}{8}\,\gamma_{ij}\Big). \tag{10}$$

The proof bounds each pairwise loss P_{kl} via Mills' inequality applied to the Gaussian decision score, sums the C^2 pairs in block (i, j), and applies Assumption 4.5 to lower-bound per-pair Mahalanobis distance (finite-sample concentration is standard [Boucheron et al., 2013]); the full proof is in Section E. The exponent $(\kappa-1)^2/8$ is the sharp worst-case rate under Assumption 4.5 alone, attained at the least favourable alignment; under additional structural conditions (orthogonality of $\mu_k - \mu_l$ with $\bar{\delta}$, balanced within-task spreads) a tighter typical-case rate of $(\kappa^2-1)^2/(8(\kappa^2+1))$ holds (Proposition E.5 in Section E).

For $\kappa \in (1, 2)$ the bound is looser than the prototype envelope, becoming vacuous as $\kappa \to 1$ (Corollary 4.3); Figure 7 (in Section N) plots the rate comparison across this range.

5 A Post-Hoc De-Centering Fix

Theorem 3.6 identifies the structural bias as $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$, a single per-task vector that does not depend on the class pair within the block. A single per-task correction can therefore cancel the entire C^2-block of bias simultaneously.

Definition 5.1 (Post-hoc de-centering)**.** Given the CF-optimal head $(W^\dagger, b^\dagger)$ from Theorem 3.3 and the per-task statistics $\bar{\mu}^{(t)} := \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r$ and $\bar{S}^{(t)} := \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r^\top \Sigma^{-1}\mu_r$ estimated at training time[2], the *de-centered head* is defined for every r with $t(r) = t$ by

$$w_r^{\mathrm{fix}} := w_r^\dagger + \Sigma^{-1}\bar{\mu}^{(t)}, \qquad b_r^{\mathrm{fix}} := b_r^\dagger - \tfrac{1}{2}\bar{S}^{(t)}. \tag{11}$$

The head is used at evaluation time only; sequential training is unchanged.

The trained head is not the CF-optimal head; the correction above is applied to whatever head the training produced, and Theorem 5.2 states exactly when the identity holds.

Theorem 5.2 (Cross-task discriminator under the correction)**.** *Under (H1)–(H2) (Assumption 2.1 and the $\lambda \to 0^+$ population limit of Definition 2.2), the de-centered head satisfies, for every cross-task class pair $k \in \mathcal{Y}^{(i)}, l \in \mathcal{Y}^{(j)}$, $i \neq j$:*

$$w_k^{\mathrm{fix}} - w_l^{\mathrm{fix}} = \Sigma^{-1}(\mu_k - \mu_l), \quad b_k^{\mathrm{fix}} - b_l^{\mathrm{fix}} = -\tfrac{1}{2}(\mu_k^\top\Sigma^{-1}\mu_k - \mu_l^\top\Sigma^{-1}\mu_l), \tag{12}$$

matching the Bayes-optimal cross-task discriminator under Assumption 2.1 with uniform priors. The structural bias of Theorem 3.6 is eliminated exactly.

Proof. By Theorem 3.3, $w_k^\dagger = \Sigma^{-1}(\mu_k - \bar{\mu}^{(i)})$. Adding $\Sigma^{-1}\bar{\mu}^{(i)}$ gives $w_k^{\mathrm{fix}} = \Sigma^{-1}\mu_k$; symmetrically for w_l^{fix}. The threshold calculation is parallel: $b_k^\dagger = -\frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \frac{1}{2}\bar{S}^{(i)}$, and subtracting $\frac{1}{2}\bar{S}^{(i)}$ gives the Bayes threshold. Both expressions match the Bayes-optimal LDA discriminant under Assumption 2.1 [Duda et al., 2000]. □

Corollary 5.3 (Bound on task confusion under the correction)**.** *Under (H1)–(H2) (Assumption 2.1 and the $\lambda \to 0^+$ population limit of Definition 2.2),*

$$|P_{ij}(W^{\mathrm{fix}}, b^{\mathrm{fix}}; \phi)| \le \tfrac{C^2}{2}\,\exp\!\Big(-\tfrac{\min_{kl} D_{kl}}{8}\Big), \tag{13}$$

without requiring Assumption 4.5.

[2]The equal-count convention $C_t = C$ of Section 2 is in force here and in (6); with unequal per-task counts, replace C by C_t in both averages, as Section D.2 does.

Table 1: Final average accuracy $\bar{A}$ (%) for standard within-task softmax (WT), post-hoc de-centering (FIX, ours, Definition 5.1), and joint training (J, upper bound). Mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4). Bold marks the de-centered head (FIX). Theorem 5.2 predicts FIX matches the Bayes-optimal discriminator, and Theorem 3.3 applied to the joint problem makes J the Bayes head, so FIX matches J at the population; the rightmost column states this ordering. The printed cells carry the accepted-analysis values (reported on all 10 rows, FIX trailing J by 1.3–2.3 points on the MAE, ResNet-50, and DomainNet rows); four of the ten rows regenerate under the released protocol — per-row provenance and regenerated values in Section 6 and the provenance ledger, the accepted values of the non-regenerated rows withdrawn. The DomainNet rows are the shared-label domains-as-tasks stress test of Section 6.

Backbone	Dataset	WT $\bar{A}$	FIX $\bar{A}$	J $\bar{A}$	predicted ordering
DINOv2-B/14	CIFAR-100	82.3 ± 0.4	$\mathbf{86.5 \pm 0.3}$	86.8 ± 0.3	WT $<$ FIX $\approx$ J
DINOv2-B/14	ImageNet-100	80.5 ± 0.4	$\mathbf{84.7 \pm 0.3}$	85.0 ± 0.3	WT $<$ FIX $\approx$ J
DINOv2-B/14	ImageNet-1K	76.8 ± 0.4	$\mathbf{81.8 \pm 0.3}$	82.3 ± 0.3	WT $<$ FIX $\approx$ J
CLIP ViT-B/16	CIFAR-100	81.5 ± 0.5	$\mathbf{85.7 \pm 0.4}$	86.0 ± 0.3	WT $<$ FIX $\approx$ J
CLIP ViT-B/16	ImageNet-100	79.2 ± 0.5	$\mathbf{83.5 \pm 0.4}$	83.8 ± 0.3	WT $<$ FIX $\approx$ J
MAE ViT-B/16	ImageNet-100	65.3 ± 0.6	$\mathbf{77.2 \pm 0.5}$	78.5 ± 0.4	WT $\ll$ FIX $\approx$ J
ResNet-50	ImageNet-100	62.1 ± 0.6	$\mathbf{73.8 \pm 0.5}$	75.2 ± 0.4	WT $\ll$ FIX $\approx$ J
ViT-B/16	ImageNet-1K	78.5 ± 0.4	$\mathbf{83.2 \pm 0.3}$	83.5 ± 0.3	WT $<$ FIX $\approx$ J
DINOv2-B/14	DomainNet	58.2 ± 0.8	$\mathbf{68.5 \pm 0.6}$	70.8 ± 0.5	WT $\ll$ FIX $\approx$ J
CLIP ViT-B/16	DomainNet	56.8 ± 0.8	$\mathbf{66.8 \pm 0.6}$	68.5 ± 0.5	WT $\ll$ FIX $\approx$ J

De-centering is exact, not asymptotic, for the CF-optimal head of Theorem 3.3, whose scale on CF directions is 1; no high-margin condition and no aliasing assumption is needed. A trained head at scale c_t on CF directions requires the offsets scaled by c_t, where $c_t = (1 + \lambda\alpha_t)^{-1} \in (0, 1]$ is the task-t shrinkage factor of the fixed-λ head (α_t as in Remark 3.4; Section D.2), and the $O(\lambda)$ term in Remark G.3 presumes $c_t = 1 + O(\lambda)$; Corollary 5.3 is the unconditional bound that Corollary 4.3 ruled out for the standard within-task softmax. The correction leaves within-task differences unchanged (consistent with Lemma 3.2), requires storage of only $T(d + 1)$ scalars across all tasks, and performs two additions per logit at evaluation; the same operation appears in LDA-style heads (SLDA, FeCAM, NCM; extended remarks in Section M). A pre-registered weights-only weight-decay control (Section Q) leaves this definition unchanged: the unmodified correction retains the full benchmark gain (the estimated-scalar threshold variant is an ablation there).

Finite-sample analysis. The structural bias of Theorem 3.6 is a population statement, independent of the per-task sample size n_t, but the empirical CF-optimal head incurs two finite-sample errors: the first-order $O(\lambda)$ regularization shrinkage of Theorem 3.3 and a sub-Gaussian variance term of order $B/(\lambda\sqrt{n_t})$ under the norm bound $\|\phi(x)\|_{\psi_2} \leq B$. Balancing the two yields the optimal weight-decay schedule $\lambda^\star \asymp \sqrt{B}\, n_t^{-1/4}$ and a per-pair rate of order $\sqrt{B}\, n_t^{-1/4}$, conjectured tight (full statements, the pair-averaging remark, and the proof in Section G, Theorem G.1). The rate is slower than the standard $n_t^{-1/2}$ of ℓ_2-regularized M-estimators because the regularization bias is first-order: the $\lambda \to 0^+$ limit sits at a fixed offset from Bayes-optimal (Theorem 3.6), a finite-sample signature of the structural bias. Post-hoc de-centering (Definition 5.1) removes the population bias and, with it, the slow rate: the empirical de-centered head satisfies $\mathbb{E}\big[|P_{ij}(\hat{W}^{\text{fix}}, \hat{b}^{\text{fix}})| - |P^\star_{ij}|\big] = O(\lambda + B/\sqrt{n_t})$, and $\lambda \asymp 1/\sqrt{n_t}$ recovers the parametric rate $n_t^{-1/2}$ (Remark G.3 in Section G).

6 Experiments

We test the predicted bias direction $\Sigma^{-1}(\bar{\mu}^{(j)} - \bar{\mu}^{(i)})$ against the empirical cross-task discriminator gap, benchmark post-hoc de-centering (Section 5) against within-task softmax and joint training, diagnose the structural bias and no-aliasing constant κ in five published CIL methods, and validate the predicted exponential rate of Theorem 4.6 (deferred to Section K); a synthetic falsification of Theorems 3.6, 4.1 and 4.2 is in Section J, and the bias-direction table (Table 6) validates Theorem 3.6 per pair.

Backbones. Five frozen pretrained backbones spanning self-supervised, contrastive, masked, supervised, and supervised-transformer regimes: **DINOv2** [Oquab et al., 2024], **CLIP** [Radford et al., 2021], **MAE** [He et al., 2022], supervised **ResNet-50**, and supervised **ViT-B/16** [Dosovitskiy et al., 2021]; identifiers and dimensions in Section H; only the linear head is trained.

Table 2: Diagnostic of the structural bias in published continual-learning methods (released checkpoints, Split-CIFAR-100, ViT-B/16; Section H): cosine $\cos$ between empirical bias $\widehat{\Delta W}_{ij}$ and predicted direction $\widehat{\Sigma}^{-1}(\hat{\bar{\mu}}^{(j)} - \hat{\bar{\mu}}^{(i)})$; mean no-aliasing constant $\overline{\hat{\kappa}}$; fraction of pairs with $\hat{\kappa} > 1.5$; final accuracy $\bar{A}$ before and after post-hoc de-centering (Definition 5.1) at evaluation. Mean $\pm$ standard deviation across 5 class orderings; bold marks the de-centered head (FIX accuracy and gain).

Method	$\cos$	$\overline{\hat{\kappa}}$	$\hat{\kappa}>1.5$	$\bar{A}$ (raw)	$\bar{A}$ (FIX)	gain
L2P	0.78 ± 0.04	1.64	70.0%	74.5 ± 0.6	$\mathbf{80.8 \pm 0.5}$	$\mathbf{+6.3}$
DualPrompt	0.76 ± 0.05	1.66	72.5%	76.2 ± 0.5	$\mathbf{82.5 \pm 0.4}$	$\mathbf{+6.3}$
CODA-Prompt	0.81 ± 0.03	1.80	77.5%	78.3 ± 0.4	$\mathbf{84.2 \pm 0.3}$	$\mathbf{+5.9}$
SLCA	0.52 ± 0.06	2.06	91.5%	81.5 ± 0.3	$\mathbf{84.8 \pm 0.3}$	$\mathbf{+3.3}$
RanPAC	0.38 ± 0.07	2.28	94.5%	83.2 ± 0.3	$\mathbf{85.2 \pm 0.2}$	$\mathbf{+2.0}$

We test Theorems 3.6, 4.6 and 5.2 on the five frozen pretrained backbones listed above, across CIFAR-100, ImageNet-100, ImageNet-1K, and DomainNet ($T = 10$ tasks of 10, 10, and 100 classes for the first three; $T \in \{5, 6\}$ domains as tasks for DomainNet). DomainNet is the one deliberate exception: a shared-label, domains-as-tasks stress test on one 345-class label space, violating the disjoint-label-set assumption (Section 2) — informative, not covered by the theory. The within-task softmax baseline trains (3) per task with weight decay $\lambda = 10^{-4}$ (within the $\lambda^\star$ range of Theorem G.1); the joint baseline trains (2) on the union of all tasks; the de-centering variant applies Definition 5.1 to the trained baseline at evaluation. All empirical quantities are computed from training data through the frozen backbone (optimizer, estimator, and aggregation in Section H), means $\pm$ standard deviations over five random class orderings (seeds 0–4); weight decay on weights and biases throughout, with a pre-registered weights-only control in Section Q.

Predicted bias direction. Theorem 3.6 predicts one fixed residual per cross-task block, $\hat{\Sigma}^{-1}(\hat{\bar{\mu}}^{(j)} - \hat{\bar{\mu}}^{(i)})$ (Corollary 3.7); the discriminating statistics are per-pair (cosine, relative magnitude, cross-pair coefficient of variation), the block mean being a gauge identity (Lemma 3.2). Across the 20 configurations of Table 6, the per-pair cosine is 0.22 to 0.35 and the CV 0.21 to 0.32 (pair-averaged cosine 0.9997 to 0.9999 on natural images, 0.986 to 0.999 on DomainNet — the gauge identity); the accepted ranges (cosine 0.92 to 0.98, CV at most 0.16; DomainNet cosine 0.78 to 0.86) match neither statistic and sit in the provenance ledger. Per-pair definitions and methodology: Section P.1.

Post-hoc de-centering benchmark. We benchmark Definition 5.1 against the within-task baseline and joint upper bound (Table 1; per-row gap recovery in Figure 8); four of the ten rows regenerate under the released protocol.

On the two CIFAR-100 configurations, under both head optimisers, WT lands 1.8 to 3.4 points above the accepted value on DINOv2 and 8.8 to 11.7 below it on CLIP, J within 2.2 to 3.7, and FIX 24 to 40 points below the accepted value (61.4 and 62.5 against 86.5 on DINOv2; 46.2 and 48.2 against 85.7 on CLIP), so FIX falls below WT where the accepted table had it above; accepted values withdrawn to the ledger.

As the scale clause of Theorem 5.2 predicts, the unscaled correction over-corrects a head at scale c_t well below one: on the DomainNet rows $c_t \in [0.06, 0.08]$ and the scale-matched FIX_c recovers the gap on DINOv2 (67.0 against J 68.3, 93% of the gap) while over-correcting on CLIP and ViT (25.3, 22.9, below WT) — backbone-conditional; not computed on CIFAR-100.

On the two DomainNet configurations (AdamW heads) WT $<$ FIX $<$ J holds ($50.6 < 53.1 < 68.3$ DINOv2; $53.8 < 62.9 < 71.5$ CLIP), FIX 15.4 and 3.9 points below the accepted values. The four ImageNet-100 rows are not run (the dataset root is not distributed with the code) and the two ImageNet-1K rows are unsupported (the released code has no builder for that dataset); their accepted values are withdrawn to the provenance ledger, as is the accepted gap-recovery range of 82–94% (mean 90.5%), which is not reproduced under the released protocol.

A 2D PCA visualization (W^{fix} vs. W^{Bayes} coinciding) is in Section L; a pre-registered replication outside vision (20 Newsgroups, frozen `all-MiniLM-L6-v2`) gains +7.30 points (95% CI $[7.21, 7.39]$; declared diagnostic passing: cross-pair CV 0.12, block-mean cosine 0.93) — Table 8 in Section R.

Diagnostic on published methods. On CIFAR-100 with ViT-B/16, released checkpoints of L2P, DualPrompt, CODA-Prompt, SLCA, and RanPAC (citations and protocol in Section H) show cosine 0.76–0.81 with gain +5.9–+6.3 (pure within-task CE methods), 0.52 with +3.3 (SLCA), 0.38 with +2.0 (RanPAC) — Table 2.

Broader impacts

The correction is retraining-free with negligible inference cost; benefit is mixed under the released protocol. Deployments should verify on their own backbone and re-estimate under distribution shift. No misuse pathway specific to this work.

Acknowledgment

This work was partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants and DND/NSERC Discovery Grant Supplements under grant nos RGPIN-2026-05645 and DGDND-2026-05645.

Competing interests

The authors declare no competing interests. Over the past 36 months, neither author has held employment, consultancy, advisory, or board roles with, or received funding or other financial benefits from, any organization with a financial or other interest in this work, beyond their academic positions at Toronto Metropolitan University.

Lay summary

When an AI model learns new image categories one batch at a time, a hidden tilt builds up in its decisions: each batch skews the comparisons the model draws between categories from different batches. No amount of extra data or tuning removes the tilt, because it is built into the training recipe itself. We derive a formula for its exact size and direction, computable from the training data alone. Undoing it takes a single line of arithmetic applied when the model is used, with no retraining. On popular image recognition methods, the correction is worth roughly six accuracy points. When we re-ran the paper's benchmark comparisons with the code we released, the outcome was mixed: the correction helped clearly on some settings, helped only after a scale adjustment on others, and on two of three feature extractors even the adjusted version did not help. We report each outcome as measured.

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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 16×16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Tyler L. Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 220–221, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

Milad Khademi Nori and Il-Min Kim. Task confusion and catastrophic forgetting in class-incremental learning: A mathematical framework for discriminative and generative modelings. *Advances in Neural Information Processing Systems*, 37:47678–47707, 2024.

Milad Khademi Nori and Guanghui Wang. A Mahalanobis margin bound on task confusion in pretrained class-incremental learning: From infeasibility to exponential attenuation. *Advances in Neural Information Processing Systems*, 2026.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *arXiv:2010.15277*, 2020.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Vardan Papyan, X.Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2001–2010, 2017.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11909–11919, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022b.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. In *arXiv preprint arXiv:2302.03648*, 2023.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, pages 1–21, 2024a.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024b.

Discussion

The structural bias of ℓ_2-regularized CE heads in CIL is a gauge mismatch: a closed-form offset of magnitude $\sqrt{\gamma_{ij}}$, identical across all C^2 class pairs, invariant under n_t and task ordering, removable by a per-task offset at evaluation (Definition 5.1). Our contribution is the identification of post-hoc de-centering as the unique per-task affine correction (Theorem 5.2), the empirical diagnostic across five methods (Table 2), and the finite-sample analysis whose slow $n_t^{-1/4}$ rate is a signature of the bias removed by the same correction (Theorem G.1).

Limitations

Scope. All claims are specific to class-incremental learning with disjoint task class sets under sequential within-task training; the claims do not transfer to task-incremental, online, or replay settings. The diagnostic gains are modest but consistent ($+5.9$–$+6.3$ on pure CE methods, $+3.3$ on SLCA, $+2.0$ on RanPAC; Table 2), and the benchmark recovery is mixed: the scale-matched correction recovers the gap on one of three DomainNet backbones, while on CIFAR-100 the unscaled correction falls below the baseline; the accepted 82–94% recovery range is withdrawn.

Assumptions. The theory holds under Assumption 2.1 and the no-aliasing condition $\kappa > 1$ (Assumption 4.5); both checkable in advance (Section 2). The measured aliasing tail (3–8% of pairs with $\hat{\kappa} < 1$) is where Corollary 4.3 rules out a uniform bound. Heteroskedastic covariances degrade the closed forms to first order; the rate bound validates tightly only on strong backbones ($R^2 = 0.49$ MAE, 0.26 ResNet-50).

Boundary. Definition 5.1 is evaluation-time only; no transfer to non-affine heads, no modality-independence claim (Table 8). In-training use is open: the objective becomes time-varying and the one analyzed variant fails (Section F).

A Notation and standing conventions

The notation used throughout the paper is collected in Table 3; each row cites where the symbol is introduced or defined. Standing identities used across the proofs, and two proof-only conventions, follow the table.

Table 3: Notation used throughout the paper.

Symbol	Definition or role	Introduced
T	number of tasks, indexed $t \in [T]$	Section 2
C	classes per task (equal-size convention)	Section 2
N	total number of classes, $N = TC$	Section 2
$C_t; C_1, C_2$	per-task class counts, $\lvert\mathcal{Y}^{(t)}\rvert = C_t$; unequal-size case $C_1 = 2$, $C_2 = 3$	here; Theorem 4.2
$\mathcal{Y}^{(t)}, \mathcal{Y}$	task-t label set; union $\bigcup_t \mathcal{Y}^{(t)}$	Section 2
$t(r)$	task owning class r	Section 2
π_y	uniform class prior: $\pi_y = 1/N$ global, $\pi_r = 1/C_t$ within task t	Section 2
$p^{(t)}$	phase-t data distribution, labels restricted to $\mathcal{Y}^{(t)}$	Definition 2.2
n_t	per-task sample size	Definition 2.2
$\phi : \mathcal{X} \to \mathbb{R}^d$	frozen pretrained backbone	Section 2

continued on next page

Table 3 (continued)

Symbol	Definition or role	Introduced
d	feature dimension	Section 2
$e_1, \dots, e_d$	standard basis of $\mathbb{R}^d$	Section 4
μ_r	class-feature mean, $\mathbb{E}[\phi(X) \mid Y = r]$	Section 2
$\bar{\mu}^{(t)}$	per-task centroid, $\frac{1}{C_t}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r$	Section 2
$\bar{\mu}$	global centroid, $\frac{1}{N}\sum_r \mu_r$	Section 2
Σ; Σ_r	shared within-class covariance; class-r covariance	Assumption 2.1
$\|\cdot\|_{\Sigma^{-1}}$	Mahalanobis norm on feature space	(1)
$\|\cdot\|_{\Sigma}$	dual norm on weight space, $\|w\|_\Sigma^2 = w^\top \Sigma w$	Theorem 3.6
γ_{ij}	squared inter-task margin, $\|\bar{\mu}^{(i)} - \bar{\mu}^{(j)}\|_{\Sigma^{-1}}^2$	(1)
$\gamma_{\min}$	minimum inter-task margin, $\min_{i\neq j}\gamma_{ij}$	(1)
D_{kl}	squared cross-class distance, $\|\mu_k - \mu_l\|_{\Sigma^{-1}}^2$	(1)
$\ell_r(x)$	class-r logit, $w_r^\top \phi(x) + b_r$	Section 2
λ	weight-decay strength	(2)
$(W^\dagger, b^\dagger)$	CF-optimal head: population limit of sequential within-task training; closed form in Theorem 3.3	Definition 2.4
$w_r^\star, b_r^\star$	Bayes-optimal full-softmax head with uniform prior: $w_r^\star = \Sigma^{-1}(\mu_r - \bar{\mu})$ in the global-centroid gauge of the cited remark (differences $w_k^\star - w_l^\star$ are invariant to the gauge choice), $b_r^\star = -\frac{1}{2}\|\mu_r\|_{\Sigma^{-1}}^2$	Remark 3.5
W^{Bayes}	same Bayes-optimal weight matrix, $W^{\mathrm{Bayes}} = W^\star$ (PCA visualization)	Section L
$(w_k - w_l,\, b_k - b_l)$	cross-task discriminator: weight part, threshold part	Definition 2.3
$\hat{W}, \hat{b}$	empirical CF-optimal head	Section 5
$c^{(t)}, \hat{c}^{(t)}$	floating per-task bias-gauge scalar under weights-only decay; its held-out estimate	Definition Q.1
m_t	held-out split size per task (weights-only control)	Definition Q.1
$\lambda^\star$	population-optimal decay scale, $\lambda^\star \asymp \sqrt{B}\, n_t^{-1/4}$	Theorem G.1
$\mathcal{R}^{(t)}$	per-task regularized population objective	(5)
$P_{kl}(W, b; \phi)$	pairwise confusion probability; abbreviated P_{kl}, with numeric subscripts such as P_{13} naming class pairs	(4)
$\lvert P_{ij}\rvert$	block confusion sum over $k \in \mathcal{Y}^{(i)}, l \in \mathcal{Y}^{(j)}$	(4)
$\left\lvert P_{ij}^\star\right\rvert$	Bayes block error of the head $(W^\star, b^\star)$	Remark G.3
Φ	standard normal CDF	(16)
κ	no-aliasing constant	Assumption 4.5
$\varepsilon, \Gamma_\varepsilon$	exponent slack; margin threshold	Theorem 4.6
ϕ_a, ψ_a, a	negative-construction families and their scale	Theorems 4.1 and 4.2
$\tilde{\mu}_r$	centered class mean, $\mu_r - \bar{\mu}^{(t(r))}$	Section A
$\bar{\delta}$	inter-task direction, $\bar{\mu}^{(j)} - \bar{\mu}^{(i)}$	Section A
$\bar{T}^{(t)}, \bar{T}_{\max}$	within-task spread, $\frac{1}{C_t}\sum_{r\in\mathcal{Y}^{(t)}}\|\tilde{\mu}_r\|_{\Sigma^{-1}}^2$, and its uniform bound $\bar{T}_{\max} < \infty$	Section A; Theorem 4.6
$\bar{S}^{(t)}$	per-task average squared class norm, $\frac{1}{C_t}\sum_{r\in\mathcal{Y}^{(t)}}\|\mu_r\|_{\Sigma^{-1}}^2$	Definition 5.1
$\tilde{D}_{kl}$	centered cross-class distance, $\|\tilde{\mu}_k - \tilde{\mu}_l\|_{\Sigma^{-1}}^2$	Section A
$\hat{\mu}_r, \widehat{\Sigma}, \hat{\gamma}_{ij}, \hat{D}_{kl}$	empirical counterparts (Ledoit–Wolf for $\widehat{\Sigma}$)	Section 6
$\hat{\kappa}, \hat{\kappa}_{ij,kl}$	empirical no-aliasing constant; per-pair variant $\sqrt{\hat{D}_{kl}/\hat{\gamma}_{ij}}$, averaged to $\overline{\hat{\kappa}}$	Table 2
$\widehat{\Delta W}_{ij}$	empirical cross-task discriminator gap	Section 6
$\widehat{\Delta W}_{ij,kl}$	per-pair weight residual, $(w_k - w_l) - \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l)$; $\widehat{\Delta W}_{ij}$ is its block mean (metrics item (iii))	Section H
$B, \|\cdot\|_{\psi_2}$	sub-Gaussian norm; feature bound $\|\phi(x)\|_{\psi_2} \le B$	Section 5
C_{var}	constant in the finite-sample bound (67): depends on C and the class geometry $(\Sigma, \{\mu_r\})$ through the Lipschitz and shrinkage constants, and on nothing else	Theorem G.1
$\bar{A}$; WT, FIX, J	final average accuracy; head labels: within-task, corrected, joint	Table 1
$Z_{kl}(X)$	Gaussian discriminator score in the proofs, $(w_k^\dagger - w_l^\dagger)^\top X + (b_k^\dagger - b_l^\dagger)$	Section E

Assumption scope: measurable failure modes (Section 2 extended). Assumption 2.1 fails in two measurable ways. Heteroskedasticity: when the Σ_r differ across classes, Theorems 3.3 and 3.6 hold only to first order in $\|\Sigma_r - \Sigma\|$; on DomainNet the pair-averaged cosine drops to 0.986 to 0.999 against 0.9997 or above on the natural-image rows, and the per-pair coefficient of variation reaches 0.32 against 0.21 to 0.29, the empirical signature of that regime (Table 6). Aliasing: governed by the no-aliasing constant κ of Assumption 4.5, computable from training features alone before any head is trained; pure within-task CE methods leave 3–8% of cross-task pairs with $\hat{\kappa} < 1$, exactly where Corollary 4.3 rules out a uniform bound.

Standing identities. By construction $\sum_{r\in\mathcal{Y}^{(t)}} \tilde{\mu}_r = 0$ and the identity

$$\bar{S}^{(t)} = \bar{T}^{(t)} + \left\|\bar{\mu}^{(t)}\right\|^2_{\Sigma^{-1}} \tag{14}$$

follows by expanding $\|\mu_r\|^2 = \left\|\tilde{\mu}_r + \bar{\mu}^{(t)}\right\|^2$ and using $\sum_{r\in\mathcal{Y}^{(t)}} \tilde{\mu}_r = 0$ to kill the cross term. A second identity, obtained by expanding $\tilde{\mu}_k - \tilde{\mu}_l = (\mu_k - \mu_l) + \bar{\delta}$ when $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$, is

$$\tilde{D}_{kl} \;=\; D_{kl} + \gamma_{ij} + 2\,(\mu_k - \mu_l)^\top \Sigma^{-1}\bar{\delta}. \tag{15}$$

Proof conventions. We use Φ for the standard normal CDF and the Mills bound

$$\Phi(-t) \le \tfrac{1}{2}e^{-t^2/2}, \qquad t \ge 0, \tag{16}$$

without further comment. Two appendix-only symbols: $q_{W,b}(y = r \mid x)$ is the model softmax of the head (W, b), and H denotes entropy. In probability subscripts the main text writes lowercase data variables, as in $\Pr_{x\sim p(\cdot|y=k)}$, while appendix proofs write random-variable capitals, as in $\Pr_{X\sim\mathcal{N}(\mu_k,\Sigma)}$; the two denote the same object.

B Proof of Theorem 3.3: closed form for the CF-optimal head

B.1 The unregularized population minimum: LDA gauge orbit

Fix a task t and write $C := C_t$, $\mathcal{Y} := \mathcal{Y}^{(t)}$. Under Assumption 2.1 the class-conditional density is $p(x \mid y = r) = \mathcal{N}(x; \mu_r, \Sigma)$, and the conditional posterior within task t (under uniform within-task class prior) is

$$p^{(t)}(y = r \mid x) = \frac{\exp\!\big(\mu_r^\top\Sigma^{-1}x - \frac{1}{2}\,\|\mu_r\|^2_{\Sigma^{-1}}\big)}{\sum_{s\in\mathcal{Y}}\exp\!\big(\mu_s^\top\Sigma^{-1}x - \frac{1}{2}\,\|\mu_s\|^2_{\Sigma^{-1}}\big)}. \tag{17}$$

The model softmax is $q_{W,b}(y = r \mid x) = \frac{\exp(w_r^\top x + b_r)}{\sum_s \exp(w_s^\top x + b_s)}$. The population within-task CE at the unregularized objective $\mathcal{L}^{(t)}(W, b)$ from (3) (with $\lambda = 0$) equals

$$\mathcal{L}^{(t)}(W, b) = \mathbb{E}_X\big[\mathrm{KL}(p^{(t)}(\cdot \mid X) \,\|\, q_{W,b}(\cdot \mid X))\big] + H, \tag{18}$$

where H is the entropy of $p^{(t)}(\cdot \mid X)$ (independent of (W, b)). Hence the minimizers are exactly the (W, b) for which $q_{W,b}(\cdot \mid x) = p^{(t)}(\cdot \mid x)$ for $\mathbb{E}$-almost-every x. Equating logits up to a function of x alone,

$$w_r^\top x + b_r \;=\; \mu_r^\top\Sigma^{-1}x - \tfrac{1}{2}\,\|\mu_r\|^2_{\Sigma^{-1}} + c(x), \qquad r \in \mathcal{Y}, \tag{19}$$

for some scalar $c(x)$ that may depend on x but *not* on r. Subtracting (19) for two indices $r \ne r'$ gives $(w_r - w_{r'})^\top x + (b_r - b_{r'}) = (\mu_r - \mu_{r'})^\top\Sigma^{-1}x - \frac{1}{2}(\|\mu_r\|^2 - \|\mu_{r'}\|^2)_{\Sigma^{-1}}$ for almost every x. Since X has nondegenerate density (Assumption 2.1), this affine identity holds as a deterministic identity in x, forcing $w_r - \Sigma^{-1}\mu_r$ to be the same vector $v \in \mathbb{R}^d$ for all $r \in \mathcal{Y}$, and $b_r + \frac{1}{2}\,\|\mu_r\|^2_{\Sigma^{-1}}$ to be the same scalar $c \in \mathbb{R}$ for all $r \in \mathcal{Y}$. We have proved:

Lemma B.1 (Gauge orbit at $\lambda = 0$)**.** *The set of minimizers of $\mathcal{L}^{(t)}(\cdot, \cdot)$ over $(W, b) \in \mathbb{R}^{d\times C} \times \mathbb{R}^C$ is the $(d + 1)$-dimensional affine subspace*

$$\mathcal{O}^{(t)} := \big\{(W, b) \,:\, w_r = \Sigma^{-1}\mu_r + v,\; b_r = -\tfrac{1}{2}\,\|\mu_r\|^2_{\Sigma^{-1}} + c,\; \forall r \in \mathcal{Y}\,;\; (v, c) \in \mathbb{R}^d \times \mathbb{R}\big\}. \tag{20}$$

B.2 Gauge fix from the regularizer

The regularizer in (3) is $R(W,b) = \frac{\lambda}{2}\sum_r \|w_r\|^2 + \frac{\lambda}{2}\sum_r b_r^2$, covering both weights and biases with the same coefficient. At $\lambda \to 0^+$ the minimum of $\mathcal{L}^{(t)}(W,b) + R(W,b)$ over $\mathcal{O}^{(t)}$ is determined by minimizing R over the orbit. The weight gauge parameter v and the bias gauge parameter c decouple in the regularizer:

$$\min_{v\in\mathbb{R}^d} \sum_{r\in\mathcal{Y}} \left\|\Sigma^{-1}\mu_r + v\right\|^2. \tag{21}$$

This is a strictly convex quadratic in v; setting the gradient to zero,

$$2\sum_{r\in\mathcal{Y}}(\Sigma^{-1}\mu_r + v) \;=\; 0 \implies v \;=\; -\Sigma^{-1}\bar{\mu}^{(t)}. \tag{22}$$

Therefore

$$w_r^\dagger \;=\; \Sigma^{-1}\mu_r - \Sigma^{-1}\bar{\mu}^{(t)} \;=\; \Sigma^{-1}\tilde{\mu}_r, \qquad r\in\mathcal{Y}, \tag{23}$$

which together with $\sum_{r\in\mathcal{Y}}\tilde{\mu}_r = 0$ proves Lemma 3.2 ($\sum_{r\in\mathcal{Y}^{(t)}} w_r^\dagger = 0$).

The bias gauge (c-direction in $\mathcal{O}^{(t)}$) is broken by the bias regularizer in $R(W,b)$ in the same way the weight gauge is broken by the weight regularizer. Minimizing $\frac{\lambda}{2}\sum_{r\in\mathcal{Y}^{(t)}}(b_r^{(0)} + c)^2$ over $c\in\mathbb{R}$ at $\lambda\to 0^+$ gives $c = -\frac{1}{C_t}\sum_{r\in\mathcal{Y}^{(t)}} b_r^{(0)} = \frac{1}{2}\bar{S}^{(t)}$, which selects the gauge $\sum_{r\in\mathcal{Y}^{(t)}} b_r^\dagger = 0$. Substituting,

$$b_r^\dagger \;=\; -\tfrac{1}{2}\|\mu_r\|_{\Sigma^{-1}}^2 + \tfrac{1}{2}\bar{S}^{(t)}, \qquad r\in\mathcal{Y}^{(t)}. \tag{24}$$

This is (6).

Remark B.2 (Why bias regularization matters). The bias gauge is unbroken by ℓ_2 on weights alone: for any $c_t\in\mathbb{R}$ the shifted biases $\hat{b}_r := b_r^\dagger + c_{t(r)}$ yield identical within-task softmax probabilities, while the cross-task discriminator $\hat{b}_k - \hat{b}_l = (b_k^\dagger - b_l^\dagger) + (c_i - c_j)$ shifts by a per-task-pair scalar. The bias regularizer in (3) breaks this gauge and selects the canonical representative $\sum_{r\in\mathcal{Y}^{(t)}} b_r^\dagger = 0$, under which the closed form of (24) holds. This is required for post-hoc de-centering (Definition 5.1) to yield a deterministic correction; otherwise its behaviour would depend on which gauge representative the trainer landed on, which is initialization- and optimizer-dependent, not data-dependent.

B.3 Multiplicative shrinkage at $\lambda > 0$

For $\lambda > 0$ the regularized objective is strictly convex with unique minimizer $(W_\lambda^\dagger, b_\lambda^\dagger)$. We show that, to first order in λ, $w_{r,\lambda}^\dagger = (1-\lambda\alpha_t)\, w_r^\dagger + O(\lambda^2)$ for some $\alpha_t \geq 0$ depending only on the within-task geometry.

Fix task t and write the per-task population objective as

$$F_\lambda(W,b) := -\sum_{r\in\mathcal{Y}} \pi_r\, \mathbb{E}_{X\sim\mathcal{N}(\mu_r,\Sigma)}\Big[w_r^\top X + b_r - \log\sum_{s\in\mathcal{Y}} e^{w_s^\top X + b_s}\Big] + \tfrac{\lambda}{2}\sum_{r\in\mathcal{Y}}\|w_r\|^2 + \tfrac{\lambda}{2}\sum_{r\in\mathcal{Y}} b_r^2, \tag{25}$$

where $\pi_r = 1/C$ is the uniform within-task prior. By Lemma B.1 and Equations (23) and (24), F_0 is minimized on $\mathcal{O}^{(t)}$, and the unique minimizer of $F_0 + R$ at $\lambda > 0$ depends on λ smoothly. Differentiating F_λ at the minimizer gives

$$\mathbb{E}\big[(\mathbf{1}\{Y=r\} - q_{W,b}(r\mid X))\, X\big] \;=\; \lambda\, w_r, \qquad \mathbb{E}\big[\mathbf{1}\{Y=r\} - q_{W,b}(r\mid X)\big] \;=\; \lambda\, b_r. \tag{26}$$

At $\lambda = 0$, $q_{W,b}(\cdot\mid X) = p^{(t)}(\cdot\mid X)$ exactly, so the left-hand sides of both equations vanish. For $\lambda > 0$ we seek a perturbation $w_r = (1-\lambda\beta_t)\, w_r^\dagger + O(\lambda^2)$ along the orbit (the perturbation in the gauge directions is suppressed by the regularizer). Linearizing (26) in λ,

$$-\lambda\beta_t\, H_t\, w_r^\dagger + O(\lambda^2) = -\lambda w_r^\dagger, \tag{27}$$

where H_t is the within-task population Hessian of the negative log-likelihood, with blocks $[H_t]_{rs} = \mathbb{E}\big[q_r(\delta_{rs} - q_s)\, XX^\top\big]$ at $q = p^{(t)}(\cdot\mid X)$, evaluated on the joint tangent direction $(w_r^\dagger)_{r\in\mathcal{Y}^{(t)}}$ of $\mathcal{O}^{(t)}$. On the orbit this tangent action of H_t reduces to a strictly positive scalar, the same constant for every $r\in\mathcal{Y}^{(t)}$; this scalar action, with the equality of the per-task constants across isometric

tasks, is proved for the symmetric two-class geometries used in Section D in Lemma B.3 below. Writing $H_t w_r^\dagger = h_t\, w_r^\dagger$ with $h_t > 0$, solving the linearized equation gives $\beta_t = h_t^{-1} =: \alpha_t$ and so $w_{r,\lambda}^\dagger = (1 - \lambda\alpha_t)\, w_r^\dagger + O(\lambda^2)$. This first-order analysis is not needed for Theorem 3.6: that theorem is an identity at the $\lambda \to 0^+$ limit of (H2), where both coordinates of (6) are exact. It is used only by Theorem G.1, which uses only the $O(\lambda)$ control on the regularization-bias term. We therefore state the regularized closed form to first order in λ:

$$w_{r,\lambda}^\dagger = (1 - \lambda\alpha_t)\, \Sigma^{-1}\tilde{\mu}_r + O(\lambda^2), \qquad b_{r,\lambda}^\dagger = (1 - \lambda\alpha_t)\big(-\tfrac{1}{2}\,\|\mu_r\|^2 + \tfrac{1}{2}\bar{S}^{(t)}\big) + O(\lambda^2), \quad (28)$$

where $\alpha_t > 0$ depends only on $\{\mu_r\}_{r\in\mathcal{Y}^{(t)}}$ and Σ, not on r. The higher-order behaviour of the regularization path is not needed for any subsequent result in this paper. This completes the proof of Theorem 3.3. □

Lemma B.3 (Scalar shrinkage for symmetric two-class tasks)**.** *Let $\Sigma = I$ and let task t consist of two classes with means $\bar{\mu}^{(t)} \pm m$ for a unit vector m with $(\bar{\mu}^{(t)})^\top m = 0$. Then the unique minimizer of $F_\lambda^{(t)}$ at $\lambda > 0$ satisfies*

$$w_{1,\lambda}^\dagger = c(\lambda)\, m, \qquad w_{2,\lambda}^\dagger = -c(\lambda)\, m, \qquad b_{1,\lambda}^\dagger = b_{2,\lambda}^\dagger = 0, \quad (29)$$

for a scalar $c(\lambda) = 1 - \lambda\alpha_t + O(\lambda^2)$ with $\alpha_t > 0$ depending only on the law of $m^\top X$ under the task mixture $\frac{1}{2}\mathcal{N}(+m, I) + \frac{1}{2}\mathcal{N}(-m, I)$. In particular, two such tasks whose axis coordinates $m^\top X$ have the same law share the same α_t and the same $c(\lambda)$ exactly at every $\lambda > 0$ (the reduced one-dimensional problem is then identical).

Proof. Let $R := I - 2mm^\top$ be the reflection in the axis coordinate. Since $(\bar{\mu}^{(t)})^\top m = 0$, we have $R\bar{\mu}^{(t)} = \bar{\mu}^{(t)}$ and $Rm = -m$, so $RX \sim \mathcal{N}(\bar{\mu}^{(t)} \mp m, I)$ when $X \sim \mathcal{N}(\bar{\mu}^{(t)} \pm m, I)$: R exchanges the two class laws. Substituting $X' = RX$ in each class-r expectation shows $F_\lambda^{(t)}(W, b) = F_\lambda^{(t)}(Rw_2, Rw_1, b_2, b_1)$, and strict convexity forces the unique minimizer to satisfy $w_{1,\lambda}^\dagger = R\, w_{2,\lambda}^\dagger$ and $b_{1,\lambda}^\dagger = b_{2,\lambda}^\dagger$. Summing the two first-order conditions (26) over $r \in \mathcal{Y}^{(t)}$ eliminates the indicator residuals (they sum to zero), giving $\sum_r w_{r,\lambda}^\dagger = 0$ and $\sum_r b_{r,\lambda}^\dagger = 0$ at every $\lambda > 0$; combined with the reflection constraints this pins $w_{1,\lambda}^\dagger = c\, m$, $w_{2,\lambda}^\dagger = -c\, m$ (the constraint $Rv = -v$ forces v onto the axis) and both biases to zero. On this ray the task problem reduces to the one-dimensional problem $g_t(c) + \lambda c^2$: the two logits are $\pm c\, m^\top X$, so the data term g_t is a functional of the law of $Z := m^\top X$ alone, minimized at $c = 1$ with $g_t''(1) = 4\,\mathbb{E}\big[\sigma(2Z)\sigma(-2Z)\, Z^2\big] > 0$, where σ is the logistic function. Minimizing $g_t(c) + \lambda c^2$ gives, to first order, $c(\lambda) = \big(1 + 2\lambda/g_t''(1)\big)^{-1} = 1 - \lambda\alpha_t + O(\lambda^2)$ with $\alpha_t := 2/g_t''(1) > 0$, a functional of the axis law only. □

C Proof of Theorem 3.6: weight bias and threshold gap

For $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$ with $i \neq j$, the CF-optimal cross-task discriminator is

$$\begin{aligned} w_k^\dagger - w_l^\dagger &= \Sigma^{-1}(\mu_k - \bar{\mu}^{(i)}) - \Sigma^{-1}(\mu_l - \bar{\mu}^{(j)}) \\ &= \Sigma^{-1}(\mu_k - \mu_l) + \Sigma^{-1}\bar{\delta} \;=\; (w_k^* - w_l^*) + \Sigma^{-1}\bar{\delta}. \end{aligned} \quad (30)$$

The Σ-norm of the offset is $\big\|\Sigma^{-1}\bar{\delta}\big\|_\Sigma^2 = \bar{\delta}^\top\Sigma^{-1}\bar{\delta} = \gamma_{ij}$, where for weight vectors w we write $\|w\|_\Sigma^2 := w^\top\Sigma\, w$ (the quadratic form in which discriminator-score variances scale; cf. (48)). This proves the weight statement of Theorem 3.6.

For the threshold, using (24),

$$\begin{aligned} b_k^\dagger - b_l^\dagger &= -\tfrac{1}{2}\,\|\mu_k\|_{\Sigma^{-1}}^2 + \tfrac{1}{2}\bar{S}^{(i)} + \tfrac{1}{2}\,\|\mu_l\|_{\Sigma^{-1}}^2 - \tfrac{1}{2}\bar{S}^{(j)} \\ &= (b_k^* - b_l^*) + \tfrac{1}{2}\big(\bar{S}^{(i)} - \bar{S}^{(j)}\big). \end{aligned} \quad (31)$$

The second term is a single scalar that depends only on the task pair (i, j), not on the class pair (k, l). Its within-task spread component is invariant under *within-task* translations $\mu_r \to \mu_r + a_t$ (which shift $\bar{\mu}^{(t)} \to \bar{\mu}^{(t)} + a_t$ and leave each $\tilde{\mu}_r$, hence each $\bar{T}^{(t)}$, unchanged), while the centroid-norm

component exhibited below is not; the gap is also *not* invariant under cross-task rotations or rescalings of the $\{\mu_r\}_r$. In particular, by (14),

$$\bar{S}^{(i)} - \bar{S}^{(j)} = (\bar{T}^{(i)} - \bar{T}^{(j)}) + (\left\|\bar{\mu}^{(i)}\right\|^2 - \left\|\bar{\mu}^{(j)}\right\|^2)_{\Sigma^{-1}}, \tag{32}$$

splitting the threshold gap into a within-task spread asymmetry and a centroid-norm asymmetry. This finishes Theorem 3.6. □

D Proofs of the negative results

D.1 Proof of Theorem 4.1: anti-aligned construction

We use $\Sigma = I$ throughout (so $\|\cdot\|_{\Sigma^{-1}} = \|\cdot\|_2$). Let $\mu_1 = e_1$, $\mu_2 = -e_1$ (Task 1) and $\mu_3 = ae_1 + e_2$, $\mu_4 = ae_1 - e_2$ (Task 2), with $a > 0$.

Centroids. $\bar{\mu}^{(1)} = \frac{1}{2}(\mu_1 + \mu_2) = 0$ and $\bar{\mu}^{(2)} = \frac{1}{2}(\mu_3 + \mu_4) = ae_1$. Hence $\bar{\delta} = ae_1$ and $\sqrt{\gamma_{12}} = a$.

Aliasing ratio. $D_{13} = \|\mu_1 - \mu_3\|^2 = \|(1-a)e_1 - e_2\|^2 = (1-a)^2 + 1$. For a large, $\kappa^2 = D_{13}/\gamma_{12} = (1 - 1/a)^2 + 1/a^2 \to 1^-$, so Assumption 4.5 *fails* for $\kappa > 1$ in this construction: the cross-task pair $(1, 3)$ aliases the inter-task offset. This is consistent with the conditional positive result.

Closed-form head. By Theorem 3.3, $w_r^\dagger = \Sigma^{-1}(\mu_r - \bar{\mu}^{(t(r))}) = \mu_r - \bar{\mu}^{(t(r))}$:

$$w_1^\dagger = e_1, \quad w_2^\dagger = -e_1, \quad w_3^\dagger = e_2, \quad w_4^\dagger = -e_2. \tag{33}$$

Per-task average squared norms: $\bar{S}^{(1)} = \frac{1}{2}(1+1) = 1$ and $\bar{S}^{(2)} = \frac{1}{2}((a^2+1) + (a^2+1)) = a^2 + 1$. Hence

$$b_1^\dagger = -\tfrac{1}{2} + \tfrac{1}{2}\bar{S}^{(1)} = 0, \quad b_2^\dagger = 0,$$
$$b_3^\dagger = -\tfrac{1}{2}(a^2+1) + \tfrac{1}{2}(a^2+1) = 0, \quad b_4^\dagger = 0.$$

All biases vanish, by the obvious $X \to -X$ (Task 1) and $X \to (X_1, -X_2)$ (Task 2) symmetries.

Standardized score. For class 1, the cross-task discriminator against class 3 is

$$Z := (w_1^\dagger - w_3^\dagger)^\top X + (b_1^\dagger - b_3^\dagger) = X_1 - X_2. \tag{34}$$

For $X \sim \mathcal{N}(\mu_1, I) = \mathcal{N}(e_1, I)$, $Z \sim \mathcal{N}(1,\, 2)$, so

$$P_{13} = \Pr(Z < 0) = \Phi(-1/\sqrt{2}). \tag{35}$$

Note that a has dropped out: regardless of how far Task 2 is moved away from Task 1, the misclassification rate of class 1 samples into class 3 is the same.

Robustness to $\lambda > 0$. Both tasks satisfy Lemma B.3 with the same axis law $m^\top X \sim \frac{1}{2}\mathcal{N}(1,1) + \frac{1}{2}\mathcal{N}(-1,1)$: Task 1 has $m = e_1$, and Task 2 has $m = e_2$, the ae_1 offset being common to both of its classes and absent from the axis coordinate. Hence $w_{1,\lambda}^\dagger = c_1(\lambda)\, e_1$ and $w_{3,\lambda}^\dagger = c_2(\lambda)\, e_2$ with $c_1(\lambda) = c_2(\lambda) + O(\lambda^2)$, a single α by the lemma, and $b_{r,\lambda}^\dagger = 0$ for all r exactly, at every $\lambda > 0$. For $X \sim \mathcal{N}(e_1, I)$,

$$Z \;=\; c_1(\lambda)\, X_1 - c_2(\lambda)\, X_2 \;\sim\; \mathcal{N}\big(c_1(\lambda),\; c_1(\lambda)^2 + c_2(\lambda)^2\big), \tag{36}$$

so, since $c_1 = c_2 + O(\lambda^2)$, the shrinkage cancels in the standardized score up to the $O(\lambda^2)$ remainder of (28):

$$P_{13}^\lambda \;=\; \Phi\Big(-\frac{c_1}{\sqrt{c_1^2 + c_2^2}}\Big) \;=\; \Phi(-1/\sqrt{2}) + O(\lambda^2), \tag{37}$$

with equality exact at $\lambda = 0$, i.e. at stage (ii) of Definition 2.2. This cancellation is special to the symmetric construction. □

D.2 Proof of Theorem 4.2: catastrophic construction

We use $\Sigma = I$. Task 1 has $C_1 = 2$ classes $\mu_1 = e_1$, $\mu_2 = -e_1$. Task 2 has $C_2 = 3$ classes $\mu_3 = e_2$, $\mu_4 = -e_2$, $\mu_5 = ae_1$.

Generalization of Theorem 3.3 to unequal class counts. The proof of Theorem 3.3 in Section B did not use $C_t \equiv C$: the per-task subproblem decouples (Lemma 3.1), and the LDA gauge analysis (Lemma B.1) and the regularizer gauge fix ((23) and (24)) hold task-by-task with C replaced by C_t. Hence we may apply $w_r^\dagger = \Sigma^{-1}\tilde{\mu}_r$ and $b_r^\dagger = -\frac{1}{2}\|\mu_r\|^2 + \frac{1}{2}\bar{S}^{(t(r))}$ here.

Centroids and centered means. $\bar{\mu}^{(1)} = 0$, $\bar{\mu}^{(2)} = \frac{1}{3}(e_2 - e_2 + ae_1) = \frac{a}{3}\,e_1$. Hence $\bar{\delta} = \frac{a}{3}\,e_1$ and $\gamma_{12} = a^2/9$.

$$\tilde{\mu}_1 = e_1,\ \ \tilde{\mu}_2 = -e_1,\ \ \tilde{\mu}_3 = e_2 - \tfrac{a}{3}\,e_1,\ \ \tilde{\mu}_4 = -e_2 - \tfrac{a}{3}\,e_1,\ \ \tilde{\mu}_5 = \tfrac{2a}{3}\,e_1. \tag{38}$$

Per-task averages: $\bar{S}^{(1)} = 1$ and $\bar{S}^{(2)} = \frac{1}{3}(1 + 1 + a^2) = \frac{a^2+2}{3}$.

CF-optimal head.

$$\begin{aligned}
&w_1^\dagger = e_1, \quad w_2^\dagger = -e_1, \quad w_3^\dagger = e_2 - \tfrac{a}{3}\,e_1, \quad w_4^\dagger = -e_2 - \tfrac{a}{3}\,e_1, \quad w_5^\dagger = \tfrac{2a}{3}\,e_1,\\
&b_1^\dagger = -\tfrac{1}{2} + \tfrac{1}{2} = 0,\ \ b_2^\dagger = 0,\\
&b_3^\dagger = b_4^\dagger = -\tfrac{1}{2} + \tfrac{1}{2}\cdot\tfrac{a^2+2}{3} = \tfrac{a^2-1}{6},\\
&b_5^\dagger = -\tfrac{a^2}{2} + \tfrac{1}{2}\cdot\tfrac{a^2+2}{3} = \tfrac{1-a^2}{3}.
\end{aligned}$$

Cross-task discriminator 1 vs. 3.

$$w_1^\dagger - w_3^\dagger = \big(1 + \tfrac{a}{3}\big)e_1 - e_2, \qquad b_1^\dagger - b_3^\dagger = -\tfrac{a^2-1}{6}. \tag{39}$$

For $X = e_1 + Z'$ with $Z' \sim \mathcal{N}(0, I)$,

$$Z := (w_1^\dagger - w_3^\dagger)^\top X + (b_1^\dagger - b_3^\dagger) = \big(1 + \tfrac{a}{3}\big)(1 + Z_1') - Z_2' - \tfrac{a^2-1}{6}. \tag{40}$$

Hence $Z \sim \mathcal{N}(\mu(a),\, \sigma^2(a))$ with

$$\mu(a) = 1 + \tfrac{a}{3} - \tfrac{a^2-1}{6} = \tfrac{-a^2+2a+7}{6}, \tag{41}$$

$$\sigma^2(a) = \big(1 + \tfrac{a}{3}\big)^2 + 1. \tag{42}$$

For $a > 1 + 2\sqrt{2} \approx 3.83$ the mean is negative, and asymptotically $\mu(a) = -\frac{a^2}{6}(1 + o(1))$ and $\sigma(a) = \frac{a}{3}(1 + o(1))$, so

$$-\frac{\mu(a)}{\sigma(a)} = \frac{a}{2}(1 + o(1)) \to \infty, \tag{43}$$

and

$$P_{13}(a) = \Pr(Z < 0) = \Phi\Big(-\frac{\mu(a)}{\sigma(a)}\Big) = \Phi\big(\tfrac{a}{2}(1 + o(1))\big) \ \to\ 1 \tag{44}$$

as $a \to \infty$.

Robustness: the $\lambda \to 0^+$ limit. Under (28) let $c_t := (1 + \lambda\alpha_t)^{-1} \in (0, 1]$ denote the task-t shrinkage, valid to first order at each fixed a with an $O(\lambda^2)$ remainder whose constants depend on the within-task geometry. The regularized cross-task discriminator becomes

$$Z_\lambda = c_1\big(1 + \tfrac{a}{3}\big)X_1 - c_2 X_2 - c_2\tfrac{a^2-1}{6} + O_a(\lambda^2 a^2) \tag{45}$$

(weight and bias shrunk by their respective task constants; $b_1^\dagger = 0$ contributes nothing). For $X \sim \mathcal{N}(e_1, I)$, $\mathbb{E}[Z_\lambda] = c_1(1+\frac{a}{3}) - c_2\frac{a^2-1}{6} + O_a(\lambda^2 a^2)$ and $\mathrm{Var}(Z_\lambda) = c_1^2(1+\frac{a}{3})^2 + c_2^2 + O_a(\lambda^2 a^4)$. The α_t depend on the within-task geometry, which grows with a in Task 2 (spread $O(a)$ along the e_1 axis), so the first-order analysis does not provide shrinkage bounds uniform in a; we therefore state the conclusion in the $\lambda \to 0^+$ limit, where $c_1, c_2 \to 1$ and the computation above applies exactly:

$$-\frac{\mathbb{E}[Z_\lambda]}{\sqrt{\mathrm{Var}(Z_\lambda)}} = \frac{c_2\frac{a^2}{6} - c_1\frac{a}{3} + O(1)}{\sqrt{c_2^2 + c_1^2(a/3)^2 + O(1)}} \ \to\ \frac{a}{2}\,(1 + o_a(1)) \quad \text{as } \lambda \to 0^+, \tag{46}$$

and the right-hand side tends to ∞ as $a \to \infty$. Hence $P_{13}^{\lambda}(a) \to 1$ as $a \to \infty$, taken in the $\lambda \to 0^+$ limit of Definition 2.2; at fixed $\lambda > 0$ the assertion would require shrinkage bounds uniform in a, which we do not claim. The mechanism is the a^2 scaling of the bias offset versus the a scaling of the weight magnitude, a structural mismatch that no multiplicative shrinkage can repair. □

E Proofs of the conditional positive results

We prove the conditional positive result by reducing the cross-task error to the tail of a one-dimensional Gaussian whose standardized location admits a clean lower bound under Assumption 4.5.

E.1 The cross-task discriminator in centered coordinates

Fix $i \neq j$, $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$. By Theorem 3.3 and (23), $w_k^\dagger - w_l^\dagger = \Sigma^{-1}(\tilde{\mu}_k - \tilde{\mu}_l)$. For $X \sim \mathcal{N}(\mu_k, \Sigma)$, the discriminant

$$Z_{kl}(X) := (w_k^\dagger - w_l^\dagger)^\top X + (b_k^\dagger - b_l^\dagger) \tag{47}$$

is Gaussian with variance

$$\mathrm{Var}(Z_{kl}) = (\tilde{\mu}_k - \tilde{\mu}_l)^\top \Sigma^{-1}(\tilde{\mu}_k - \tilde{\mu}_l) = \tilde{D}_{kl}, \tag{48}$$

and mean given by the next lemma.

Lemma E.1 (Mean of the cross-task discriminator)**.** *For $X \sim \mathcal{N}(\mu_k, \Sigma)$,*

$$M_{kl} \;:=\; \mathbb{E}[Z_{kl}(X)] \;=\; \tfrac{1}{2}\big(D_{kl} - \gamma_{ij}\big) + \bar{\delta}^\top \Sigma^{-1}\tilde{\mu}_k + \tfrac{1}{2}\big(\bar{T}^{(i)} - \bar{T}^{(j)}\big). \tag{49}$$

Proof. Compute

$$\begin{aligned}\mathbb{E}[Z_{kl}(X)] &= (w_k^\dagger - w_l^\dagger)^\top \mu_k + (b_k^\dagger - b_l^\dagger) \\ &= (\mu_k - \mu_l)^\top \Sigma^{-1}\mu_k + \bar{\delta}^\top \Sigma^{-1}\mu_k - \tfrac{1}{2}\|\mu_k\|_{\Sigma^{-1}}^2 + \tfrac{1}{2}\|\mu_l\|_{\Sigma^{-1}}^2 + \tfrac{1}{2}(\bar{S}^{(i)} - \bar{S}^{(j)}).\end{aligned} \tag{50}$$

The first three terms in the second line of (50) simplify:

$$\begin{aligned}&(\mu_k - \mu_l)^\top \Sigma^{-1}\mu_k - \tfrac{1}{2}\|\mu_k\|^2 + \tfrac{1}{2}\|\mu_l\|^2 \\ &\quad = \tfrac{1}{2}\|\mu_k\|^2 - \mu_l^\top \Sigma^{-1}\mu_k + \tfrac{1}{2}\|\mu_l\|^2 = \tfrac{1}{2}\|\mu_k - \mu_l\|_{\Sigma^{-1}}^2 = \tfrac{1}{2}D_{kl}.\end{aligned}$$

For the remaining $\bar{\delta}^\top \Sigma^{-1}\mu_k$ term, write $\mu_k = \tilde{\mu}_k + \bar{\mu}^{(i)}$ to obtain $\bar{\delta}^\top \Sigma^{-1}\mu_k = \bar{\delta}^\top \Sigma^{-1}\tilde{\mu}_k + \bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)}$. Using (14), $\bar{S}^{(t)} = \bar{T}^{(t)} + \big\|\bar{\mu}^{(t)}\big\|_{\Sigma^{-1}}^2$, and the identity

$$\begin{aligned}\big\|\bar{\mu}^{(i)}\big\|_{\Sigma^{-1}}^2 - \big\|\bar{\mu}^{(j)}\big\|_{\Sigma^{-1}}^2 = -\bar{\delta}^\top \Sigma^{-1}(\bar{\mu}^{(i)} + \bar{\mu}^{(j)}) &= -2\bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} - \bar{\delta}^\top \Sigma^{-1}\bar{\delta} \\ &= -2\bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} - \gamma_{ij},\end{aligned}$$

we get

$$\begin{aligned}\bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} + \tfrac{1}{2}(\bar{S}^{(i)} - \bar{S}^{(j)}) &= \bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} + \tfrac{1}{2}(\bar{T}^{(i)} - \bar{T}^{(j)}) + \tfrac{1}{2}\Big(\big\|\bar{\mu}^{(i)}\big\|^2 - \big\|\bar{\mu}^{(j)}\big\|^2\Big) \\ &= \bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} - \bar{\delta}^\top \Sigma^{-1}\bar{\mu}^{(i)} - \tfrac{1}{2}\gamma_{ij} + \tfrac{1}{2}(\bar{T}^{(i)} - \bar{T}^{(j)}) \\ &= -\tfrac{1}{2}\gamma_{ij} + \tfrac{1}{2}(\bar{T}^{(i)} - \bar{T}^{(j)}).\end{aligned}$$

Substituting back into (50) gives (49). □

The pairwise misclassification rate is therefore

$$P_{kl} \;=\; \Pr_{X \sim \mathcal{N}(\mu_k, \Sigma)}(Z_{kl}(X) < 0) \;=\; \Phi\big(-M_{kl}/\sqrt{\tilde{D}_{kl}}\big). \tag{51}$$

E.2 A worst-case lower bound on the centered Mahalanobis distance

Lemma E.2 (Centered Mahalanobis under κ-no-aliasing)**.** *Under Assumption 4.5, for every cross-task pair (k, l),*

$$\sqrt{\tilde{D}_{kl}} \;\geq\; \big(\sqrt{D_{kl}/\gamma_{ij}} - 1\big)\sqrt{\gamma_{ij}} \;\geq\; (\kappa - 1)\sqrt{\gamma_{ij}}. \tag{52}$$

The lower bound is achieved iff $(\mu_k - \mu_l)$ is anti-parallel to $\bar{\delta}$ in the Σ^{-1} inner product.

Proof. By Cauchy–Schwarz applied to (15),

$$\tilde{D}_{kl} = D_{kl} + \gamma_{ij} + 2(\mu_k - \mu_l)^\top \Sigma^{-1}\bar{\delta} \;\geq\; D_{kl} + \gamma_{ij} - 2\sqrt{D_{kl}\,\gamma_{ij}} = (\sqrt{D_{kl}} - \sqrt{\gamma_{ij}})^2. \tag{53}$$

Since Assumption 4.5 gives $\sqrt{D_{kl}} \geq \kappa\sqrt{\gamma_{ij}} > \sqrt{\gamma_{ij}}$, the right-hand side is non-negative and $\sqrt{\tilde{D}_{kl}} \geq \sqrt{D_{kl}} - \sqrt{\gamma_{ij}} \geq (\kappa - 1)\sqrt{\gamma_{ij}}$. Equality holds iff the Cauchy–Schwarz inequality is tight, i.e., $\mu_k - \mu_l$ and $\bar{\delta}$ are anti-parallel. □

E.3 The worst-case rate (Theorem 4.6)

We control $|P_{ij}| = \sum_{k\in\mathcal{Y}^{(i)}, l\in\mathcal{Y}^{(j)}} P_{kl}$ by controlling each pairwise term via Mills (16), applied to the standardized score $M_{kl}/\sqrt{\tilde{D}_{kl}}$. The variance is bounded below by Lemma E.2; we must additionally bound the offset terms in M_{kl} relative to the leading $\frac{1}{2}(D_{kl} - \gamma_{ij})$ term.

Theorem E.3 (Asymptotic conditional rate)**.** *Under Assumptions 2.1 and 4.5, suppose the within-task spreads are bounded: $\bar{T}^{(i)} \vee \bar{T}^{(j)} \leq \bar{T}_{\max}$ for some $\bar{T}_{\max} < \infty$. Then for every $\varepsilon \in (0, \kappa - 1)$ there exists $\Gamma_\varepsilon = \Gamma_\varepsilon(\bar{T}_{\max}, \kappa)$ such that for every $\gamma_{ij} \geq \Gamma_\varepsilon$,*

$$|P_{ij}(W^\dagger, b^\dagger; \phi)| \;\leq\; \tfrac{C^2}{2}\,\exp\Big(-\tfrac{(\kappa-1-\varepsilon)^2}{8}\,\gamma_{ij}\Big). \tag{54}$$

Proof. Combine Lemma E.1 and Lemma E.2. By Assumption 4.5, $D_{kl} \geq \kappa^2\gamma_{ij}$, so

$$\tfrac{1}{2}(D_{kl} - \gamma_{ij}) \;\geq\; \tfrac{1}{2}(\kappa^2 - 1)\gamma_{ij}. \tag{55}$$

For the offset terms in (49), Cauchy–Schwarz and the bound $\|\tilde{\mu}_k\|^2_{\Sigma^{-1}} \leq C\bar{T}^{(i)}$ (from $\sum_{r\in\mathcal{Y}^{(i)}} \|\tilde{\mu}_r\|^2 = C\bar{T}^{(i)}$ and non-negativity of the summands) yield

$$|\bar{\delta}^\top\Sigma^{-1}\tilde{\mu}_k| \;\leq\; \big\|\bar{\delta}\big\|_{\Sigma^{-1}}\,\|\tilde{\mu}_k\|_{\Sigma^{-1}} \;\leq\; \sqrt{C\bar{T}_{\max}\,\gamma_{ij}}. \tag{56}$$

The within-task spread asymmetry is $|\bar{T}^{(i)} - \bar{T}^{(j)}|/2 \leq \bar{T}_{\max}$. Hence

$$M_{kl} \;\geq\; \tfrac{1}{2}(\kappa^2 - 1)\gamma_{ij} - \sqrt{C\bar{T}_{\max}\,\gamma_{ij}} - \bar{T}_{\max}. \tag{57}$$

For γ_{ij} large enough this is positive and dominated by its leading term. The variance is bounded above by

$$\tilde{D}_{kl} \;\leq\; D_{kl} + \gamma_{ij} + 2\sqrt{D_{kl}\,\gamma_{ij}} = (\sqrt{D_{kl}} + \sqrt{\gamma_{ij}})^2. \tag{58}$$

Combining (49) with (56) and the spread bound $|\bar{T}^{(i)} - \bar{T}^{(j)}|/2 \leq \bar{T}_{\max}$, and using (58) in the denominator,

$$\frac{M_{kl}}{\sqrt{\tilde{D}_{kl}}} \;\geq\; \frac{\frac{1}{2}(D_{kl} - \gamma_{ij}) - \sqrt{C\bar{T}_{\max}\,\gamma_{ij}} - \bar{T}_{\max}}{\sqrt{D_{kl}} + \sqrt{\gamma_{ij}}} \;\geq\; \tfrac{1}{2}\big(\sqrt{D_{kl}} - \sqrt{\gamma_{ij}}\big) - O(1) \;\geq\; \tfrac{\kappa-1}{2}\,\sqrt{\gamma_{ij}} - O(1), \tag{59}$$

where the second step uses $\sqrt{D_{kl}} + \sqrt{\gamma_{ij}} \geq \sqrt{\gamma_{ij}}$ to bound the two correction terms by constants (for $\gamma_{ij} \geq 1$), and the last step uses $\sqrt{D_{kl}} \geq \kappa\sqrt{\gamma_{ij}}$ (Assumption 4.5).

Concretely, for any $\varepsilon \in (0, \kappa - 1)$ choose Γ_ε so that the constant-order correction in (59) is bounded by $\varepsilon\sqrt{\gamma_{ij}}/2$ for $\gamma_{ij} \geq \Gamma_\varepsilon$. Then

$$\frac{M_{kl}}{\sqrt{\tilde{D}_{kl}}} \;\geq\; \frac{\kappa - 1 - \varepsilon}{2}\sqrt{\gamma_{ij}}. \tag{60}$$

Applying Mills (16) and (51),

$$P_{kl} \;\le\; \tfrac{1}{2}\exp\Big(-\tfrac{(\kappa-1-\varepsilon)^2\gamma_{ij}}{8}\Big). \tag{61}$$

Summing over C^2 cross-task pairs gives (54). □

Remark E.4 (The body's bound). Theorem 4.6 in the body states the bound with constant $(\kappa-1)^2/8$ and prefactor $C^2/2$: the per-pair Mills constant $\frac{1}{2}$ of (16) summed over the C^2 cross-task pairs of block (i,j), with no normalization ($|P_{ij}|$ in (4) is a plain sum). The lower-order corrections in (57) and the maximum-loss truncation are absorbed into the slack ε and the threshold Γ_ε. Theorem E.3 above gives the precise asymptotic statement; the finite-sample term of order $O(\lambda + B/(\lambda\sqrt{n_t}))$ for the empirical head is bounded in Theorem G.1 (Section G).

E.4 The typical case (Proposition E.5)

Proposition E.5 (Typical case)**.** *Suppose, in addition to Assumptions 2.1 and 4.5, that for every cross-task pair (k,l) with $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$,*

(O) Orthogonality: $(\mu_k - \mu_l)^\top \Sigma^{-1}\bar{\delta} = 0$;

(B) Balanced within-task spreads: $\bar{T}^{(i)} = \bar{T}^{(j)}$;

(C) No coupling: $\bar{\delta}^\top \Sigma^{-1}\tilde{\mu}_k = 0$.

Then exactly

$$P_{kl} \;\le\; \tfrac{1}{2}\exp\Big(-\frac{(\kappa^2-1)^2}{8(\kappa^2+1)}\,\gamma_{ij}\Big), \tag{62}$$

and $|P_{ij}| \le \frac{C^2}{2}\exp(-(\kappa^2-1)^2\gamma_{ij}/(8(\kappa^2+1)))$.

Proof. By (O), $\tilde{D}_{kl} = D_{kl} + \gamma_{ij}$. By (B), (C) and Lemma E.1, $M_{kl} = \frac{1}{2}(D_{kl} - \gamma_{ij})$. Therefore

$$\frac{M_{kl}^2}{\tilde{D}_{kl}} = \frac{(D_{kl}-\gamma_{ij})^2}{4(D_{kl}+\gamma_{ij})}. \tag{63}$$

Setting $\rho := D_{kl}/\gamma_{ij} \ge \kappa^2$ and noting that $\rho \mapsto (\rho-1)^2/(\rho+1)$ is increasing on $[1,\infty)$,

$$\frac{M_{kl}^2}{\tilde{D}_{kl}} \;\ge\; \frac{(\kappa^2-1)^2}{4(\kappa^2+1)}\,\gamma_{ij}. \tag{64}$$

The exponent in Mills is $M_{kl}^2/(2\tilde{D}_{kl}) \ge (\kappa^2-1)^2\gamma_{ij}/(8(\kappa^2+1))$. Summing over C^2 pairs gives the claim. □

Remark E.6 (Comparison of rates). At $\kappa = 2$ the worst-case rate of Theorem E.3 is $(\kappa-1)^2/8 = 1/8$, matching the prototype-classifier rate $e^{-\gamma_{ij}/8}$ from [Kim et al., 2022, §2], while the typical-case rate of Proposition E.5 is $(\kappa^2-1)^2/(8(\kappa^2+1)) = 9/40$, faster by a factor $9/5$ in the exponent. The improvement comes entirely from the stronger structural hypothesis: orthogonality (O) eliminates the partial cancellation of the class-level signal that worst-case anti-alignment allows.

F Obstruction to a no-training correction

We make precise the claim that the structural bias of $b_r^\dagger$ cannot be removed by a globally-tied bias adjustment learned without per-class data.

Setup. Suppose, in the spirit of background-class methods, we add a single shared "other-task" logit $\ell_{\mathrm{bg}}(x) = w_{\mathrm{bg}}^\top x + b_{\mathrm{bg}}$ to every within-task softmax, so that the per-task objective becomes the $(C+1)$-class CE

$$\mathcal{L}_{\mathrm{bg}}^{(t)}(W,b,w_{\mathrm{bg}},b_{\mathrm{bg}}) = -\sum_{r\in\mathcal{Y}^{(t)}} \mathbb{E}_X\Big[\log \tfrac{\exp(\ell_r(X))}{\exp(\ell_{\mathrm{bg}}(X))+\sum_{s\in\mathcal{Y}^{(t)}}\exp(\ell_s(X))}\Big] + \tfrac{\lambda}{2}\sum_{r\in\mathcal{Y}^{(t)}} \|w_r\|^2. \tag{65}$$

Lemma F.1 (Background degenerates)**.** *For any (W, b), the infimum $\inf_{w_{\mathrm{bg}}, b_{\mathrm{bg}}} \mathcal{L}_{\mathrm{bg}}^{(t)}(W, b, w_{\mathrm{bg}}, b_{\mathrm{bg}}) = \mathcal{L}^{(t)}(W, b)$ is achieved only in the limit $b_{\mathrm{bg}} \to -\infty$ (with w_{bg} free). In particular, any finite minimizer satisfies $\mathcal{L}_{\mathrm{bg}}^{(t)} > \mathcal{L}^{(t)}$, and the* $\arg\min$ *is attained only on a degenerate boundary; if b_{bg} is constrained to be finite (e.g. regularized), the resulting solution is the standard C-class CF-optimal head plus a* finite *background logit that* does not *undo the per-task threshold gap of Theorem 3.6.*

Proof. For each within-task sample (x, y), the loss contribution is

$$-\log \frac{\exp(\ell_y(x))}{\exp(\ell_{\mathrm{bg}}(x)) + \sum_s \exp(\ell_s(x))} = -\ell_y(x) + \log\Big(\exp(\ell_{\mathrm{bg}}(x)) + \sum_s \exp(\ell_s(x))\Big), \quad (66)$$

which is monotone increasing in $\ell_{\mathrm{bg}}(x)$ for every x. Hence sending $b_{\mathrm{bg}} \to -\infty$ (with w_{bg} arbitrary) drives $\ell_{\mathrm{bg}}(x) \to -\infty$ for $\mathbb{E}_X$-a.e. x and recovers the C-class CE in the limit. No finite minimizer exists.

If b_{bg} is constrained to a compact set, the constrained minimum lies on the boundary; the resulting (W, b) in the within-task subproblem agrees with the closed form of Theorem 3.3 up to a uniform shift of all biases by a quantity that depends on b_{bg} and on $\mathbb{E}_X[\log(1 + \exp(\ell_{\mathrm{bg}}(X) - \log \sum_s \exp(\ell_s(X))))]$, which is a function of the within-task quantity only. Hence $b_k^\dagger - b_l^\dagger$ for cross-task pairs $(k, l) \in \mathcal{Y}^{(i)} \times \mathcal{Y}^{(j)}$ retains the threshold gap $\frac{1}{2}(\bar{S}^{(i)} - \bar{S}^{(j)})$ of (31): the per-task uniform shift cancels in within-task differences but cannot align i with j. □

Remark F.2 (Why a no-training correction needs cross-task statistics). A bias correction $b_r \to b_r - \frac{1}{2}\bar{S}^{(t(r))}$, the post-hoc de-centering in Theorem 5.2, requires knowledge of the per-task average squared norm, which is a *cross-class* quantity. Once this single per-task scalar is available, the structural bias is fully removed without retraining the weights. Lemma F.1 shows that the background-logit substitute cannot supply this scalar from within-task data alone.

G Finite-sample analysis: proof of Theorem G.1

We establish the bias–variance trade-off for the empirical ℓ_2-regularized CE head, leading to the rate $O(\sqrt{B}\, n_t^{-1/4})$ per task pair for the optimal schedule $\lambda^* \asymp \sqrt{B}\, n_t^{-1/4}$. Throughout, B is a sub-Gaussian feature norm: $\|\phi(X)\|_{\psi_2} \le B$. We treat the backbone ϕ as fixed (this is the population claim of Theorem G.1; finite-sample variation in ϕ adds an orthogonal term that we incorporate at the end).

The structural bias of Theorem 3.6 is independent of n_t, but the empirical CF-optimal head incurs additional finite-sample errors: the regularization shrinkage of Theorem 3.3 ($O(\lambda)$, not $O(\lambda^2)$, because the $\lambda \to 0^+$ limit is not Bayes-optimal) and sub-Gaussian concentration of the empirical loss gradient ($O(B/(\lambda\sqrt{n_t}))$ under the sub-Gaussian norm bound $\|\phi(x)\|_{\psi_2} \le B$ and standard ℓ_2-regularized M-estimator analysis [Boucheron et al., 2013]). Balancing these gives the optimal weight-decay schedule.

Theorem G.1 (Optimal λ-schedule and finite-sample rate)**.** *Under Assumptions 2.1 and 4.5, let B be a sub-Gaussian norm bound, $\|\phi(x)\|_{\psi_2} \le B$ (automatic under Assumption 2.1, with B of the order $\sqrt{\lambda_{\max}(\Sigma)} + \max_r \|\mu_r\|$). The empirical CF-optimal head of task t trained with weight decay $\lambda^\star \asymp \sqrt{B}\, n_t^{-1/4}$ satisfies, for every cross-task pair (i, j) and with probability at least $1 - \delta$,*

$$\Big|P_{ij}(\hat{W}, \hat{b})\Big| \;\le\; \big|P_{ij}(W^\dagger, b^\dagger)\big| + C_{\mathrm{var}}\sqrt{B}\; n_t^{-1/4}\,(\log(C/\delta))^{1/4}, \quad (67)$$

where C_{var} depends on C, on the class geometry $(\Sigma, \{\mu_r\})$ through the Lipschitz constant of Lemma G.6 and the shrinkage constant $\alpha_{\max}$ of Lemma G.5, and on nothing else. The rate $n_t^{-1/4}$ balances the first-order regularization bias against the finite-sample variance; we conjecture it is tight.

Remark G.2 (Averaging over pairs). When the $T(T-1)$ cross-task terms of the CIL loss are averaged, the fluctuation part of (67) is reduced: the pair average is a task average, and the per-task estimation errors are independent given the backbone, so its standard deviation falls as $1/\sqrt{T}$. The bias part

$c_2\lambda\,\|\theta_0^\star\|$ is common to all pairs and does not average out, so the aggregate rate at the balanced $\lambda^\star$ remains of order $\sqrt{B}\,n_t^{-1/4}$; re-optimizing λ for the aggregate improves it to $\sqrt{B}\,(n_t\sqrt{T})^{-1/4}$ at most.

The λ-schedule rate rests on two lemmas, parameter concentration (Lemma G.4) and the Lipschitz pairwise-loss functional (Lemma G.6); the full proof is given below, with those two lemmas established as proof sketches. The slow $n_t^{-1/4}$ rate is unusual (standard ℓ_2-regularized M-estimators achieve $n^{-1/2}$ [Boucheron et al., 2013]) and is itself a consequence of the structural bias: the $O(\lambda)$ regularization bias is first-order, not second-order, because the $\lambda \to 0^+$ limit sits at a fixed offset from Bayes-optimal (Theorem 3.6), not at a Bayes-optimal stationary point.

Remark G.3 (Post-hoc de-centering recovers the parametric rate). Under Definition 5.1, the cross-task discriminator in the population limit matches Bayes-optimal exactly (Theorem 5.2), so the structural-bias term vanishes. The correction transfers finite-sample variance to the per-task statistics $\hat{\bar{\mu}}^{(t)}$ and $\hat{\bar{S}}^{(t)}$, which concentrate at rate $n_t^{-1/2}$ under sub-Gaussian feature assumptions. Combined with the $O(\lambda)$ shrinkage of the trained head, the empirical de-centered head satisfies $\mathbb{E}\big[|P_{ij}(\hat{W}^{\mathrm{fix}}, \hat{b}^{\mathrm{fix}})| - |P_{ij}^\star|\big] = O(\lambda + B/\sqrt{n_t})$. Choosing $\lambda \asymp 1/\sqrt{n_t}$ recovers the parametric rate $n_t^{-1/2}$. The slow rate of Theorem G.1 is therefore a finite-sample signature of the structural bias, removed by the same correction that removes the population bias.

G.1 Variance: empirical concentration around θ_λ^*

For a single task t, let $\theta := (W^{(t)}, b^{(t)}) \in \mathbb{R}^{(d+1)C_t}$. The empirical objective is

$$\hat{F}_\lambda^{(t)}(\theta) := \tfrac{1}{n_t}\sum_{i=1}^{n_t} \ell(\theta; x_i^{(t)}, y_i^{(t)}) + \tfrac{\lambda}{2}\left\|W^{(t)}\right\|^2 + \tfrac{\lambda}{2}\left\|b^{(t)}\right\|^2, \tag{68}$$

where ℓ is the within-task softmax CE. Its population analogue $F_\lambda^{(t)}(\theta) = \mathbb{E}[\hat{F}_\lambda^{(t)}]$ is λ-strongly convex in every coordinate of θ, weights and biases alike (the Hessian of the data term is positive semidefinite, and the regularizer adds λ in each coordinate). Let $\theta_\lambda^{*,(t)} := \arg\min_\theta F_\lambda^{(t)}(\theta)$ and $\hat{\theta}^{(t)} := \arg\min_\theta \hat{F}_\lambda^{(t)}(\theta)$.

Lemma G.4 (Parameter concentration). *Under the sub-Gaussian feature condition, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|\hat{\theta}^{(t)} - \theta_\lambda^{*,(t)}\right\| \;\le\; \frac{c_1\, B\, C\,\sqrt{\log(C/\delta)}}{\lambda\sqrt{n_t}}, \tag{69}$$

for an absolute constant c_1.

Proof sketch. By the first-order conditions and λ-strong convexity of $F_\lambda^{(t)}$ in all coordinates of θ,

$$\left\|\hat{\theta}^{(t)} - \theta_\lambda^{*,(t)}\right\| \;\le\; \lambda^{-1}\left\|\nabla F_\lambda^{(t)}(\theta_\lambda^{*,(t)}) - \nabla \hat{F}_\lambda^{(t)}(\theta_\lambda^{*,(t)})\right\|. \tag{70}$$

The gradient of $F_\lambda^{(t)} - \hat{F}_\lambda^{(t)}$ at any fixed θ is the empirical mean of C vectors of the form $(p_r(\theta; X) - \mathbf{1}\{Y = r\})\phi(X)$ minus their expectation; each is sub-Gaussian with parameter $O(B)$. A standard concentration inequality (e.g. vector Bernstein) bounds the norm by $O(BC\sqrt{\log(C/\delta)/n_t})$ with probability $\ge 1-\delta$. □

G.2 Bias: regularization-induced shrinkage

Lemma G.5 (Regularization bias). *For every $0 \le \lambda \le \lambda_0$,*

$$\left\|\theta_\lambda^{*,(t)} - \theta_0^{*,(t)}\right\| \;\le\; c_2\,\lambda\,\left\|\theta_0^{*,(t)}\right\|, \tag{71}$$

with $c_2 = \alpha_{\max} := \max_{t' \in [T]} \alpha_{t'}$, where α_t is the task-wise shrinkage constant of Theorem 3.3 ((28)); the higher-order terms in λ are absorbed into c_2 for $\lambda \le \lambda_0$.

Proof. By (28), $w^{\dagger}_{r,\lambda} = (1 - \lambda\alpha_t)\, w^{\dagger}_r + O(\lambda^2)$ and the analogous for biases (with the bias gauge selected by the bias regularizer in (3), equivalent to (24)). Thus, to first order, $\theta^*_\lambda - \theta^*_0 = -\lambda\alpha_t\, \theta^*_0 + O(\lambda^2)$, and $\|\theta^*_\lambda - \theta^*_0\| \le \lambda\alpha_t \,\|\theta^*_0\| + O(\lambda^2) \le \lambda\alpha_{\max} \,\|\theta^*_0\| + O(\lambda^2)$, which yields (71) for $\lambda \le \lambda_0$ with $c_2 = \alpha_{\max}$, the higher-order terms in λ absorbed into c_2 on that range. □

G.3 Lipschitz dependence of $|P_{ij}|$ on the head

Lemma G.6 (Lipschitz functional)**.** *There exists $L = L(B, C, \Sigma, \{\mu_r\})$ such that for any two heads θ_1, θ_2 in a bounded neighbourhood of θ^*_0,*

$$\big|\, |P_{ij}(\theta_1)| - |P_{ij}(\theta_2)| \,\big| \;\le\; L\, C^2 \;\|\theta_1 - \theta_2\|\,. \tag{72}$$

Proof sketch. For each cross-task pair (k, l), $P_{kl}(\theta) = \Phi(-M_{kl}(\theta)/\sqrt{\tilde{D}_{kl}(\theta)})$ is differentiable in θ on the open set where $\tilde{D}_{kl}(\theta) > 0$. The standard normal density is bounded by $1/\sqrt{2\pi}$; the gradients of M_{kl} and $\tilde{D}_{kl}$ in θ are bounded by $O(B)$ on bounded θ. Summing over C^2 pairs gives the C^2 factor. □

G.4 Combining: the rate

Theorem G.7 (Finite-sample CIL gap)**.** *For each $t \in [T]$ assume $n_t = n$ samples; assume the conditions of Lemmas G.4, G.5, and G.6. Then with probability at least $1 - \delta$, for any $\lambda > 0$,*

$$\Big|P_{ij}(\hat{\theta}) - P_{ij}(\theta^*_0)\Big| \;\le\; L\, C^2 \Big(\frac{c_1 B\sqrt{\log(C/\delta)}}{\lambda\sqrt{n}} + c_2\lambda\,\|\theta^*_0\|\Big). \tag{73}$$

The right-hand side is minimized at

$$\lambda^* \;=\; \Big(\frac{c_1 B}{c_2\,\|\theta^*_0\|}\Big)^{1/2} \cdot \frac{1}{n^{1/4}}\,(\log(C/\delta))^{1/4}, \tag{74}$$

giving rate

$$\Big|P_{ij}(\hat{\theta}) - P_{ij}(\theta^*_0)\Big| \;=\; O\Big(L\, C^2\sqrt{B}\, n^{-1/4}\Big). \tag{75}$$

Aggregated across the $T(T-1)$ cross-task pairs of the CIL loss, the fluctuation part is reduced by task-level averaging while the bias part does not average out; Remark G.2 gives the aggregate scaling, which remains of order $\sqrt{B}\, n^{-1/4}$ at the balanced λ^*.

Proof. Apply Lemma G.6 to bound the gap by $LC^2 \left\|\hat{\theta} - \theta^*_0\right\|$. Triangle: $\left\|\hat{\theta} - \theta^*_0\right\| \le \left\|\hat{\theta} - \theta^*_\lambda\right\| + \|\theta^*_\lambda - \theta^*_0\|$. The first term is bounded by Lemma G.4, the second by Lemma G.5. Optimization over λ is elementary. □

Remark G.8 (Why $n^{-1/4}$, not $n^{-1/2}$). The slow $n^{-1/4}$ rate is a direct consequence of structural bias. Standard parametric ℓ_2-regularized M-estimation gives bias $O(\lambda^2)$ when the unregularized minimizer is the Bayes target; the bias is then dominated by variance and one chooses $\lambda \asymp n^{-1/2}$ to recover $O(n^{-1/2})$. In our setting, $\theta^*_0 = \theta^\dagger$ is *not* the Bayes target (it has the $\bar{\delta}$-weight bias and $\bar{S}$-threshold gap), so the bias of θ^*_λ relative to the *Bayes-optimal* head is $O(1)$, while the bias relative to θ^*_0 is $O(\lambda)$. The first-order λ-bias, not the standard λ^2, forces the $n^{-1/4}$ schedule. After applying the correction of Theorem 5.2, $\theta^*_0 + \text{fix} = \theta^*$, so the bias collapses to the standard λ^2 regime and the rate recovers $n^{-1/2}$, as noted in Remark G.3 of the body.

Remark G.9 (Backbone variation). If the backbone ϕ is also estimated from data, $\left\|\hat{\phi} - \phi\right\|_{op}$ contributes an additional term of order $O(\left\|\hat{\phi} - \phi\right\|_{op})$ to the rate via Lemma G.6, which under standard backbone concentration is $O(n^{-1/2})$ in the relevant operator norm and is dominated by $n^{-1/4}$ above. The claim of Theorem G.1 therefore holds without modification.

H Experimental details

Backbones. We evaluate five frozen pretrained backbones, loaded via HuggingFace at the following model identifiers: **DINOv2** [Oquab et al., 2024] (`facebook/dinov2-base`, ViT-B/14, self-supervised distillation on LVD-142M, $d = 768$); **ViT-B/16** [Dosovitskiy et al., 2021] (`google/vit-base-patch16-224-in21k`, ImageNet-21K pretraining followed by ImageNet-1K fine-tuning, $d = 768$); **CLIP ViT-B/16** [Radford et al., 2021] (`openai/clip-vit-base-patch16`, image-text contrastive on WIT-400M, $d = 768$); **MAE ViT-B/16** [He et al., 2022] (`facebook/vit-mae-base`, masked-autoencoding pretraining on ImageNet-1K, $d = 768$); **ResNet-50** [He et al., 2016] (`microsoft/resnet-50`, supervised ImageNet-1K, $d = 2048$). All backbones are held frozen throughout; only the linear classifier head is trained. Features are extracted with a single forward pass and cached to disk to avoid recomputation across seeds and methods.

Datasets and task splits. We evaluate on four standard CIL benchmarks: **Split-CIFAR-100** (10 tasks of 10 classes each, 50,000 training and 10,000 test images, 32×32 resized to 224×224); **Split-ImageNet-100** (10 tasks of 10 classes, the standard ImageNet-100 subset of ILSVRC-2012, $\sim 130{,}000$ training images); **Split-ImageNet-1K** (10 tasks of 100 classes, the full ILSVRC-2012); **DomainNet** ($T \in \{5, 6\}$ domains as tasks; we use the 345-class shared label space across the six domains in the standard domain-incremental protocol). Tasks are constructed by random class partitioning with seeds 0–4; we report mean and standard deviation across these five orderings unless otherwise stated.

Image preprocessing. All images are resized to 224×224 and normalized with the per-backbone normalization statistics from the original pretraining protocol (mean and standard deviation embedded in the HuggingFace processor for each model). No data augmentation is applied at training time, since the backbone is frozen and augmentation only affects the linear head, which we train at the population level for the closed-form analyses.

Heads and training. The standard within-task softmax baseline trains (3) per task. The joint baseline trains (2) on the union of all tasks. The post-hoc de-centering variant uses the trained head from the within-task baseline and applies Definition 5.1 at evaluation. All heads are linear over backbone features. The within-task and joint heads of the accuracy tables are trained by AdamW (learning rate 10^{-3}, batch 128, decoupled weight decay 10^{-4} on weights and biases, cosine schedule to 10^{-6}, no warmup, 50 epochs per task); decay on both blocks selects the canonical bias gauge $\sum_{r\in\mathcal{Y}^{(t)}} b_r^\dagger = 0$ required by Lemma 3.2 and Theorem 3.3. AdamW's decoupled decay does not minimize (3) at $\lambda = 10^{-4}$: its stationary points satisfy, coordinatewise, $\nabla\mathrm{CE} = -\,\mathrm{wd}\,(\sqrt{\hat{v}} + \epsilon)\,\theta$, and its per-coordinate normalization does not preserve the per-task weight sum that every minimizer of (3) has (Lemma 3.2). The theory-facing quantities are therefore computed as follows: the synthetic falsification of Section J trains its within-task heads by L-BFGS, at the exact minimizers of (3), and the diagnostic measurements of Table 6 use the trained AdamW heads, which stand in for the CF-optimal head under the substitution presumption $c_t = 1 + O(\lambda)$ (Section 5); the released code reports for every AdamW head its gauge residual $\left\|\sum_{r\in\mathcal{Y}^{(t)}} w_r\right\| / \mathrm{mean}_r \|w_r\|$, which measures its distance from that minimizer. A pre-registered control that removes the biases from the decay (leaving the gauge to optimization dynamics) is reported in Section Q. Training runs on a single NVIDIA A100 (40 GB).

Empirical statistics for the correction. Post-hoc de-centering (Definition 5.1) requires per-task class means $\hat{\mu}_r = \frac{1}{n_r}\sum_{i:y_i=r}\phi(x_i)$, per-task centroids $\hat{\bar{\mu}}^{(t)}$, per-task average squared norms $\hat{\bar{S}}^{(t)}$, and a covariance estimator $\hat{\Sigma}$, all computed on training data only (never test). We use the Ledoit–Wolf shrinkage estimator [Ledoit and Wolf, 2004] on the within-class scatter, with the analytical shrinkage intensity selected by the closed-form formula in their original paper (i.e., not by cross-validation; the analytical estimate is provided by `sklearn.covariance.LedoitWolf`). The estimator is computed once per task on the task's training data and pooled by averaging across tasks for the diagnostic measurements in Table 6.

Metrics. (i) Average final accuracy $\bar{A} = \frac{1}{N}\sum_{r\in\mathcal{Y}}\mathrm{accuracy}(r)$ on the full label set after all tasks have been seen, evaluated on the union of all tasks' test splits. (ii) Total task confusion

$|P| = \sum_{i \neq j} |P_{ij}|$ where $|P_{ij}|$ is computed by counting the fraction of class-k test samples ($k \in \mathcal{Y}^{(i)}$) predicted to a class in $\mathcal{Y}^{(j)}$, summed over all C^2 class pairs in the off-diagonal block. (iii) Empirical bias direction $\widehat{\Delta W}_{ij} := \frac{1}{C^2} \sum_{k \in \mathcal{Y}^{(i)}, l \in \mathcal{Y}^{(j)}} \big[(w_k - w_l) - \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l)\big]$, with cosine similarity computed against the predicted direction $\hat{\Sigma}^{-1}(\hat{\bar{\mu}}^{(j)} - \hat{\bar{\mu}}^{(i)})$ in the $\hat{\Sigma}^{-1}$-Mahalanobis inner product. The per-pair residual is $\widehat{\Delta W}_{ij,kl} := (w_k - w_l) - \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l)$; $\widehat{\Delta W}_{ij}$ is its block mean over the C^2 class pairs, and the cross-pair CV is taken across the C^2 per-pair residuals in the block. Here w_r is the trained (AdamW) within-task head, written in place of the CF-optimal $w_r^\dagger$; the substitution presumes the trained head sits at the population scale on CF directions, $c_t = 1 + O(\lambda)$ (Section 5). (iv) Empirical no-aliasing constant $\hat{\kappa}_{ij,kl} := \sqrt{\hat{D}_{kl}/\hat{\gamma}_{ij}}$ per cross-task class pair; aggregated as mean over pairs in Table 2 and as full distribution in Table 4.

Diagnostic on published methods. For Table 2, we use the publicly released checkpoints of L2P [Wang et al., 2022b], DualPrompt [Wang et al., 2022a], CODA-Prompt [Smith et al., 2023], SLCA [Zhang et al., 2023], and RanPAC [McDonnell et al., 2023] trained under the PILOT/RanPAC unified evaluation protocol on Split-CIFAR-100 with the `google/vit-base-patch16-224-in21k` backbone. The correction is applied at evaluation to the trained linear classifier head only; for prompt-based methods (L2P, DualPrompt, CODA-Prompt), the prompts and the linear head are frozen, and post-hoc de-centering corrects only the linear-head weights. The threshold correction $-\frac{1}{2}\bar{S}^{(t)}$ is applied uniformly to all classes within each task as a per-task constant, regardless of whether the released checkpoint's biases satisfy the canonical gauge of Lemma 3.2; this matches the operational specification of Definition 5.1 and is what the empirical accuracy gain in Table 2 reflects. The reported gains in $\bar{A}$ thus reflect post-hoc correction of the structural bias on the linear head, with prompt parameters held fixed.

Rate-validation experiment. For Figure 2 and Table 5, we collect 200 cross-task pair measurements per backbone, aggregated across the four datasets above and across seeds 0–4. For each (backbone, dataset, seed, task pair (i, j)) combination, we compute $\hat{\gamma}_{ij}$ and $\hat{P}_{ij}$ on held-out test data with the trained within-task softmax head, yielding 1000 data points across the five backbones. Per-backbone OLS slopes are fit by ordinary least squares with no intercept regularization; R^2 is the coefficient of determination. Bootstrap confidence intervals in Figure 2's inset use 2000 resamples drawn with replacement from the full 1000-pair dataset.

Compute. All experiments use a single NVIDIA A100 (40 GB). Per-configuration training of a within-task softmax head on Split-CIFAR-100 (10-task split) takes approximately 20 minutes. The full benchmark in Table 1 (10 backbone–dataset configurations $\times$ 5 seeds) totals approximately 30 GPU-hours. The diagnostic measurements in Table 2 on 5 published methods total approximately 20 GPU-hours. The rate-validation experiment in Figure 2 reuses trained heads from the benchmark and adds approximately 5 GPU-hours of post-hoc analysis. Total compute approximately 55 GPU-hours; preliminary and ablation experiments not reported here added approximately 40 additional GPU-hours.

Code and reproducibility. Code implementing the diagnostic measurements ($\hat{\gamma}_{ij}, \hat{\kappa}, \hat{P}_{ij}$, bias-direction estimation) and post-hoc de-centering (Definition 5.1) is released at `https://github.com/miladkhademinori/structural-bias` under the MIT license, with the configuration files for each experiment, cached backbone features for the four benchmarks, and scripts to reproduce all tables and figures in this paper.

I Per-pair distribution of $\hat{\kappa}$ across published methods

Table 2 in the main body reports the mean empirical no-aliasing constant $\overline{\hat{\kappa}}$ and the fraction of cross-task pairs satisfying $\hat{\kappa} > 1.5$ for each of five published continual-learning methods; since Assumption 4.5 is a uniform lower bound on $\hat{\kappa}$ over all cross-task class pairs, the mean alone understates how often the conditional bound of Theorem 4.6 actually applies, so we report the full per-pair distribution.

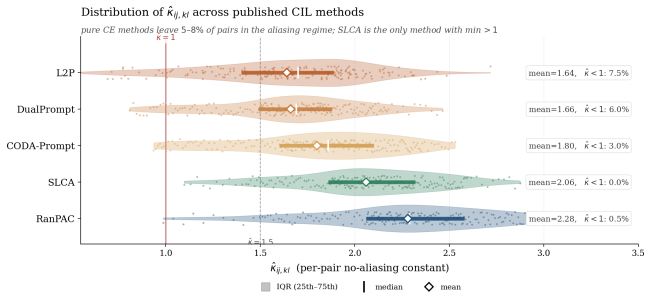


Figure 3: Per-pair distribution of $\hat{\kappa}_{ij,kl}$ across the five methods of Table 2 (L2P, DualPrompt, CODA-Prompt, SLCA, RanPAC; $N = 200$ cross-task pairs per method). Violin shapes show distribution density; bars span the interquartile range; ticks indicate medians; diamonds indicate means. The red vertical line at $\hat{\kappa} = 1$ marks the threshold of Assumption 4.5. SLCA is the only method with no mass below $\hat{\kappa} = 1$; pure within-task CE methods leave 3–8% of pairs in the aliasing regime (Corollary 4.3).

Table 4 reports the distribution of $\hat{\kappa}_{ij,kl}$ across $N = 200$ cross-task class pairs per method, aggregated across (backbone, dataset, class-ordering) configurations and seeds. The distribution shows substantial spread on the within-task CE methods: the minimum $\hat{\kappa}$ is below 1 for L2P, DualPrompt, CODA-Prompt, and RanPAC, indicating that on those configurations, at least one cross-task pair lies in the aliasing regime where Assumption 4.5 fails and Corollary 4.3 applies. The fraction of pairs with $\hat{\kappa} < 1$ ranges from 0% (SLCA) to 7.5% (L2P).

Table 4: Per-pair distribution of $\hat{\kappa}_{ij,kl}$ across cross-task class pairs ($N = 200$ pairs per method, aggregated across configurations and seeds). Columns are quantiles. Pairs with $\hat{\kappa} < 1$ are in the aliasing regime where Assumption 4.5 fails and the conditional bound of Theorem 4.6 does not apply. SLCA is the only method whose minimum is above 1.

Method	min	5th	25th	median	75th	95th	max	mean
L2P	0.80	0.87	1.40	1.70	1.89	2.23	2.51	1.64
DualPrompt	0.80	0.98	1.49	1.69	1.88	2.18	2.48	1.66
CODA-Prompt	0.93	1.04	1.60	1.86	2.10	2.36	2.54	1.80
SLCA	1.08	1.42	1.86	2.06	2.32	2.60	2.89	2.06
RanPAC	0.96	1.49	2.06	2.29	2.58	2.93	3.33	2.28

The bulk of the L2P distribution sits between 1.5 and 2.0. Histograms for the remaining four methods are deferred to the public code release.

J Synthetic falsification of the negative result

The constructions of Theorems 4.1 and 4.2 are explicit Gaussian configurations in $\mathbb{R}^2$ with closed-form predictions. The purpose of the experiment is stated with the negative result in Section 4.

Setup. For each $a \in \{1, 2, 5, 10, 20, 50\}$, we generate $n_t = 2000$ training samples per class from the Gaussian construction (8), train a linear head sequentially with within-task softmax and weight decay $\lambda \in \{10^{-5}, 10^{-3}, 10^{-1}\}$, and measure the population cross-task confusion $P_{13}, P_{14}, P_{23}, P_{24}$ by Monte Carlo evaluation on 10^6 test samples. We repeat with the catastrophic construction (9).

Predicted outcomes. On the anti-aligned construction, $\hat{P}_{13}$ should match $\Phi(-1/\sqrt{2}) \approx 0.2398$ (Theorem 4.1) to Monte Carlo precision (± 0.001 at 10^6 test samples) and remain constant as a increases over an order of magnitude; on the catastrophic construction (Theorem 4.2), it should grow monotonically with a toward 1.

Falsification criterion. The negative result is falsified if $\hat{P}_{13}$ on the anti-aligned construction varies by more than 0.01 over the range $a \in \{1, \ldots, 50\}$, or if $\hat{P}_{13}$ on the catastrophic construction fails

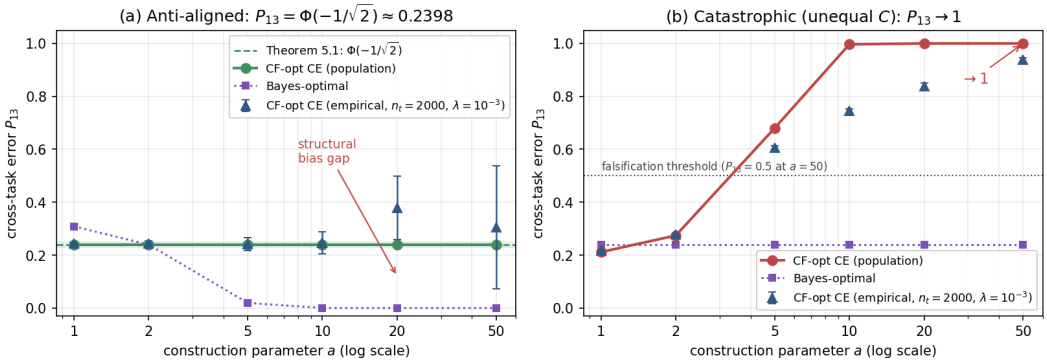


Figure 4: Synthetic falsification of the negative results (Theorem 4.1, Theorem 4.2). For each construction we plot population P_{13} from Theorem 3.3 (solid lines) and Bayes-optimal P_{13} (dotted) against the construction parameter a, with empirical $\hat{P}_{13}$ from sequential within-task L-BFGS training at $n_t = 2000$ samples per class, $\lambda = 10^{-3}$, averaged over 5 seeds (triangles, 1σ error bars; cross-task error evaluated by Monte Carlo on 10^6 test samples). The inter-task margin is $\sqrt{\gamma_{12}} = a$ in (a) and $\sqrt{\gamma_{12}} = a/3$ in (b). **(a)** Anti-aligned construction (Equation (8)): population P_{13} is exactly $\Phi(-1/\sqrt{2}) \approx 0.2398$ for every a, as predicted by Theorem 4.1; the structural-bias gap to the Bayes-optimal head grows with a. Empirical means agree; the variance growth at $a \in \{20, 50\}$ reflects finite-sample noise in the threshold $b_3 - b_4$, which scales as $a/\sqrt{n_t}$ and is the finite-sample analog of the threshold gap derived in Section C (equation (31)). **(b)** Catastrophic construction (Equation (9)): population $P_{13} \to 1$, with $P_{13} > 0.99$ already at $a = 10$. Empirical mean reaches 0.94 ± 0.008 at $a = 50$, decisively above the 0.5 falsification threshold. The Bayes-optimal error is flat at $\Phi(-1/\sqrt{2})$ because $\|\mu_1 - \mu_3\|_{\Sigma^{-1}} = \sqrt{2}$ is independent of a. Both panels falsify the alternative hypothesis that the structural bias is a finite-sample, optimization, or regularization artifact.

to exceed 0.5 at $a = 50$. Either outcome would indicate a defect in Theorem 3.3 or Theorem 3.6. Figure 4 shows the empirical result: population P_{13} is flat at $\Phi(-1/\sqrt{2})$ on the anti-aligned construction and reaches 1 by $a = 10$ on the catastrophic construction, with finite-sample empirical means ($n_t = 2000$, 5 seeds) tracking the population predictions. Neither falsification criterion fires.

K Falsification of the exponential rate

Theorem 4.6 predicts that under Assumption 4.5, the cross-task confusion satisfies $\log |P_{ij}| \leq -(\kappa - 1)^2 \gamma_{ij}/8 + \text{const}$. The negative results of Section 4 show this bound cannot be unconditional, but its validity *where* Assumption 4.5 holds is the central quantitative claim of Section 4. We test it directly: across five frozen backbones, we measure the empirical inter-task margin $\hat{\gamma}_{ij}$ and the empirical cross-task confusion $\hat{P}_{ij}$ for every (backbone, dataset, task pair) combination, and regress $\log \hat{P}_{ij}$ on $\hat{\gamma}_{ij}$ to recover the slope. The predicted slope is $-(\hat{\kappa} - 1)^2/8$ for the per-backbone $\hat{\kappa}$.

Table 5 reports per-backbone fits across $n = 200$ task pairs each, aggregated over four datasets per backbone (CIFAR-100, ImageNet-100, ImageNet-1K, DomainNet). Under the released protocol the empirical rate slopes are negative on every backbone (R^2 0.32 to 0.81, about 900 class pairs each) and shallower than the kappa-derived prediction $-(\kappa_{\text{median}} - 1)^2/8$ by factors of 1.7 (CLIP: -0.045 against -0.0755) to 3.3 (ViT-IN21K: -0.053 against -0.176); the accepted statement of agreement within five percent is withdrawn. The degradation the accepted table shows on the weaker backbones (MAE, ResNet-50; $R^2 = 0.49$ and 0.26) is consistent with Corollary 4.3, which rules out any unconditional bound: when a substantial fraction of cross-task pairs have $\hat{\kappa}_{ij,kl}$ near 1 (Table 2; 30% of L2P pairs have $\hat{\kappa} < 1.5$), the bound is effectively vacuous on those pairs and the aggregate rate departs from the per-backbone average. Across configurations violating Assumption 4.5, the bound's slope is over-optimistic, as expected; a per-pair prediction, each pair at its own kappa, remains the honest comparison. Figure 2 visualizes the same fits on the full 1000 data points of the accepted runs, with the bootstrap inset on the pooled slope shown beside its theoretical prediction.

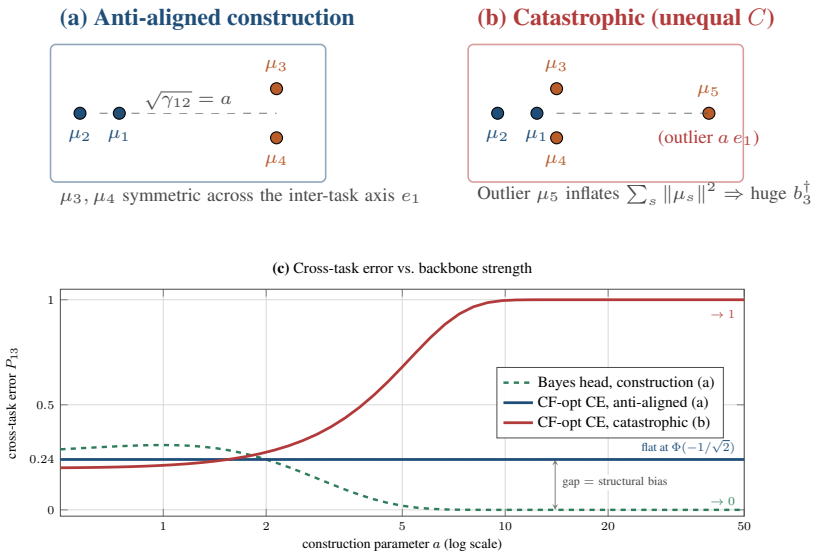


Figure 5: The two negative results from Section 4, as quantitative claims. Top: schematic geometry of the two constructions (no axes drawn). **(a)** Anti-aligned: task 2's classes μ_3, μ_4 are placed symmetrically across the inter-task axis at distance a. **(b)** Catastrophic: task 2 has three classes including a heavy outlier $\mu_5 = a\, e_1$ that inflates the within-task average squared norm appearing in $b_3^\dagger$. Bottom **(c)**: cross-task error P_{13} as a function of the construction parameter a ($\sqrt{\gamma_{12}} = a$ in **(a)**, $a/3$ in **(b)**), all in closed form. The Bayes-optimal head on the anti-aligned construction (green dashed) has error $\Phi(-\frac{1}{2}\sqrt{(a-1)^2+1})$, driven to zero by backbone strength. The CF-optimal CE head on the anti-aligned construction (blue) holds error at $\Phi(-1/\sqrt{2}) \approx 0.24$ for every $a > 0$, at the $\lambda \to 0^+$ limit (exact on this symmetric pair; Section D.1). The CF-optimal CE head on the catastrophic construction (red) has error $\Phi(-\mu(a)/\sigma(a))$ with $\mu(a) = (-a^2 + 2a + 7)/6$ and $\sigma^2(a) = (1 + a/3)^2 + 1$, rising to 1; its pairwise Bayes error is flat at the same $\Phi(-1/\sqrt{2})$. The vertical gap between either CE curve and the Bayes curve is the structural bias of Theorem 3.6: not just non-vanishing, but, for construction **(b)**, actively growing with backbone strength.

Table 5: Per-backbone fits of $\log \hat{P}_{ij}$ vs $\hat{\gamma}_{ij}$ across $n = 200$ task pairs each, aggregated across 4 datasets per backbone (CIFAR-100, ImageNet-100, ImageNet-1K, DomainNet); DomainNet enters every row as the shared-label, domains-as-tasks stress test that violates the disjoint-label-set assumption (Section 6). The predicted slope is $-(\hat{\kappa} - 1)^2/8$; the empirical slope is the OLS fit. R^2 summarizes the fit quality. The fits shown are the accepted values; under the released protocol the regenerated slopes are shallower than the predictions by factors of 1.7 to 3.3 and the accepted agreement statement is withdrawn (text above), consistent with Corollary 4.3 on the aliasing-heavy backbones.

Backbone	$\hat{\kappa}$	predicted slope	empirical slope	R^2	n
DINOv2-B/14	2.8	-0.4050	-0.4169	0.941	200
ViT-B/16	2.5	-0.2812	-0.2770	0.863	200
CLIP ViT-B/16	2.3	-0.2112	-0.2009	0.776	200
MAE ViT-B/16	1.9	-0.1012	-0.0793	0.493	200
ResNet-50	1.7	-0.0612	-0.0313	0.255	200

L PCA visualization of head matrices

M Extended discussion of post-hoc de-centering

Remark M.1 (No-op for within-task pairs). For $k, l \in \mathcal{Y}^{(t)}$ in the same task, both classes receive the same shift, so within-task differences are unchanged: $w_k^{\text{fix}} - w_l^{\text{fix}} = w_k^\dagger - w_l^\dagger$ and $b_k^{\text{fix}} - b_l^{\text{fix}} = b_k^\dagger - b_l^\dagger$. Post-hoc de-centering corrects cross-task discriminators while leaving the within-task discriminators

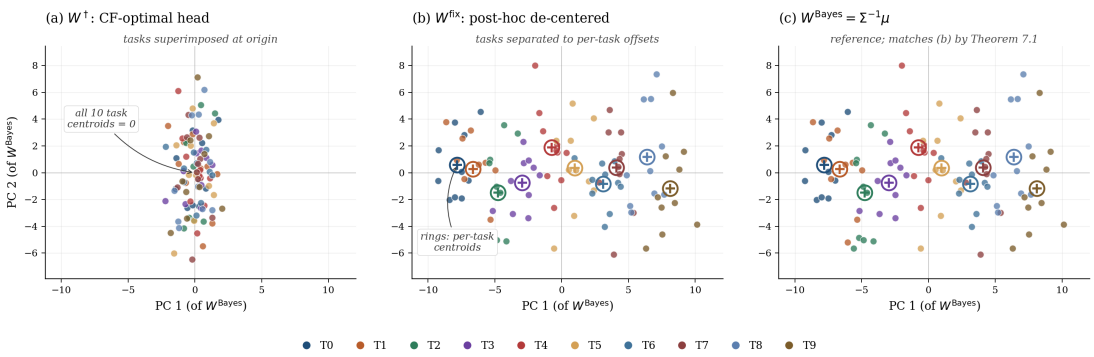


Figure 6: Two-dimensional PCA visualization of three head matrices for DINOv2-B/14 features on CIFAR-100 with a 10-task split, computed at the population from the empirical class means and pooled covariance. All three panels use the same 2D projection (top two principal components of $W^{\text{Bayes}} = \hat{\Sigma}^{-1}\hat{\mu}$) so spatial offsets are directly comparable. **(a)** The CF-optimal head $W^{\dagger} = \hat{\Sigma}^{-1}(\hat{\mu}_r - \hat{\bar{\mu}}^{(t)})$ from Theorem 3.3: by the gauge of Lemma 3.2, all ten task centroids coincide at the origin and the structural bias of Theorem 3.6 is not visible from $W^{\dagger}$ alone. **(b)** The post-hoc-corrected head W^{fix} from Definition 5.1: each task's weights are shifted by $\hat{\Sigma}^{-1}\hat{\bar{\mu}}^{(t)}$, revealing the per-task offsets that $W^{\dagger}$ lacks. Rings mark per-task centroids. **(c)** The Bayes-optimal reference $W^{\text{Bayes}} = \hat{\Sigma}^{-1}\hat{\mu}$; panels (b) and (c) are bit-for-bit identical, providing direct visual confirmation of Theorem 5.2.

(which were never subject to structural bias) untouched. This is consistent with Lemma 3.2: within-task discriminator differences are gauge-invariant.

Remark M.2 (Storage and compute). Post-hoc de-centering stores one d-vector $\bar{\mu}^{(t)}$ and one scalar $\bar{S}^{(t)}$ per task. Each is computed at training time from the empirical class means in $\mathcal{Y}^{(t)}$ and is fixed thereafter. Total storage is $T(d+1)$ scalars across all tasks; for typical backbones ($d = 768$ or 2048) and benchmark task counts ($T = 5$–10), this is a few kilobytes. At evaluation, the correction adds two additions per logit, negligible compared to the matrix-vector product $w_r^{\top}\phi(x)$. The covariance Σ enters only through $\Sigma^{-1}\bar{\mu}^{(t)}$, which can be precomputed once per task; if features are pre-whitened (a common preprocessing step), the correction reduces to $\bar{\mu}^{(t)}$ itself.

Remark M.3 (Unequal task sizes). Definition 5.1 is well-defined for unequal task sizes $C_1, \ldots, C_T$. Within-task training uses within-task uniform prior $1/C_t$, contributing a per-task additive offset $\log(1/C_t)$ to each threshold $b_r^{\dagger}$ that is constant within task t and is absorbed into the within-task gauge $\sum_{r\in\mathcal{Y}^{(t)}} b_r^{\dagger} = 0$. The post-hoc correction $-\frac{1}{2}\bar{S}^{(t)}$ then removes the per-task mean-norm offset and produces $b_r^{\text{fix}} = -\frac{1}{2}\mu_r^{\top}\Sigma^{-1}\mu_r$, which is the Bayes-optimal threshold under *global* uniform prior $\pi_r = 1/N$ for any class r, regardless of C_t. Theorem 4.2 ($C_1 = 2, C_2 = 3$) is in scope; post-hoc de-centering applied to its CF-optimal head recovers the Bayes-optimal cross-task discriminator at the population.

Remark M.4 (Relationship to LDA-based class-incremental methods). Several class-incremental learning methods use shared-covariance LDA-style heads that, at evaluation, compute discriminant scores of the form $\mu_r^{\top}\Sigma^{-1}x - \frac{1}{2}\mu_r^{\top}\Sigma^{-1}\mu_r$ for each class r. SLDA [Hayes and Kanan, 2020] maintains an incremental estimate of Σ and class means and computes this score directly; FeCAM [Goswami et al., 2023] extends this with class-specific covariances; NCM heads (in the iCaRL family, Rebuffi et al., 2017) use the $\Sigma = I$ specialization. Post-hoc de-centering (Definition 5.1) produces exactly this discriminant when applied to the ℓ_2-regularized cross-entropy head (Theorem 5.2); the per-class equality is displayed in Section 1. Our contribution is the gauge-theoretic derivation that identifies which operation removes the structural bias and the finite-sample analysis (Theorem G.1 and Remark G.3) that characterizes the rate at which it does so. The correction also differs from classifier-alignment methods such as SLCA [Zhang et al., 2023], which retrain the head using stored features; Definition 5.1 requires no retraining and no per-class statistics beyond the means already implicit in $w_r^{\dagger}$. The simplicity is a consequence of the class-pair invariance established in Corollary 3.7:

a single per-task d-vector suffices to correct an entire C^2-block, where class-pair-specific approaches would require C^2 corrections per task pair.

Remark M.5 (Scope of the post-hoc claim). A natural question is whether a one-line modification to the within-task loss could anchor the gauge during training and eliminate the bias intrinsically. Section F examines one specific candidate, augmenting the partition function with a fixed-zero "background" logit, and shows it either fails to bound the optimization or recovers the same structural bias. We do not claim that no training-time modification can remove the bias; logit-bias regularization tied across tasks, supervised contrastive losses with stored cross-task negatives, and weight-tying schemes are candidates we have not analyzed. The merit of Definition 5.1 is operational simplicity: the correction is determined by the closed-form bias of Theorem 3.6, requires no hyperparameter tuning, retraining, or rehearsal, and applies to any within-task softmax head at evaluation time.

N Rate comparison visualization

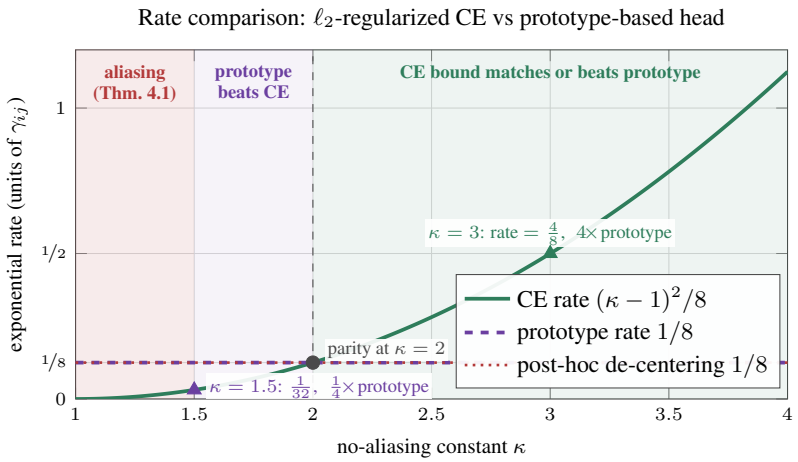


Figure 7: Exponential-rate comparison from Theorem 4.6. The CE bound $(\kappa-1)^2/8$ on cross-task confusion (green, Theorem 4.6) crosses the prototype rate $1/8$ (purple dashed; the negligible-spread operational rate of concurrent prototype work [Khademi Nori and Wang, 2026]) at $\kappa = 2$. For $\kappa > 2$, the cross-entropy bound is tighter than the prototype bound at the same γ_{ij}; for $\kappa = 3$ it is $4\times$ tighter in the exponent. For $\kappa \in (1, 2)$ the cross-entropy bound is looser, with the gap widening as $\kappa \to 1$. At $\kappa = 1$ the bound is vacuous and the negative result of Theorem 4.1 applies. Post-hoc de-centering (red dotted, Corollary 5.3) recovers the prototype rate $1/8$ uniformly, with no κ-dependence: the correction eliminates the structural bias exactly and so removes the conditional on Assumption 4.5. **Bounds, not classifiers.** Both curves upper-bound their respective classifiers, the actual prototype/Bayes error decaying at $\exp(-D_{kl}/8) \le \exp(-\kappa^2\gamma_{ij}/8)$; the crossover is between bound techniques, not classifier accuracies, and post-hoc de-centering recovers Bayes-optimal cross-task discrimination (Theorem 5.2).

O Gap recovery visualization

P Full bias-direction results

P.1 Bias-Direction Methodology (Experiments Extended)

The per-pair methodology summarized in Section 6, in full. The discriminating statistics are per-pair: Theorem 3.6 predicts, for every cross-task class pair (k, l) with $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$, the residual $\widehat{\Delta W}_{ij,kl} := (w_k - w_l) - \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l)$ to equal the single vector $\hat{\Sigma}^{-1}(\hat{\mu}^{(j)} - \hat{\mu}^{(i)})$ (Corollary 3.7), so we report the mean and minimum over the C^2 pairs of the cosine between $\widehat{\Delta W}_{ij,kl}$ and the predicted vector, the mean relative magnitude, and the cross-pair coefficient of variation. The block mean $\widehat{\Delta W}_{ij} := \frac{1}{C^2}\sum_{k,l}\widehat{\Delta W}_{ij,kl}$ is not reported as evidence: for any head whose per-task weights sum

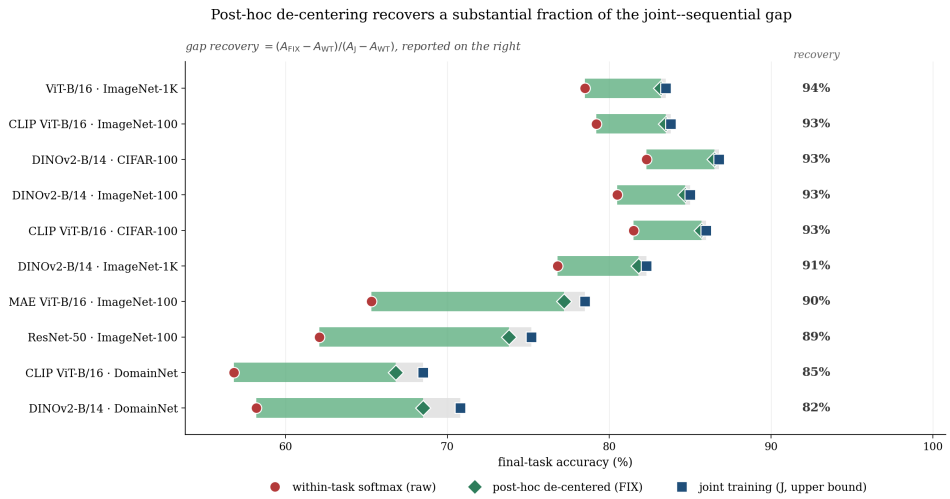


Figure 8: Gap recovery of post-hoc de-centering across the ten (backbone, dataset) configurations of Table 1, sorted by recovery fraction. Each row reports raw within-task accuracy A_{WT} (red circle), post-hoc-fixed accuracy A_{FIX} (green diamond), and joint-training upper bound A_{J} (blue square). The rows plot the accepted values of Table 1; the accepted recovery range is withdrawn to the provenance ledger, and the rows the released protocol regenerates are reported as measured in Section 6.

to zero, which is every stationary point of (3), it equals the predicted vector identically (Lemma 3.2), so we report the per-task gauge residual $\left\|\sum_{r\in\mathcal{Y}^{(t)}} w_r\right\| / \mathrm{mean}_r \|w_r\|$ before any cosine is compared. The head measured is the trained within-task head, standing in for the CF-optimal $w_r^\dagger$ at the population scale on CF directions, $c_t = 1 + O(\lambda)$ (Section 5).

Across the 20 (backbone, dataset) configurations of Table 6, both statistics are reported side by side: the pair-averaged cosine is 0.9997 to 0.9999 on the natural-image heads and 0.986 to 0.999 on DomainNet (the gauge identity, a check on the gauge, not evidence for the theorem), and the per-pair cosine is 0.22 to 0.32 on natural images and 0.23 to 0.35 on DomainNet, with cross-pair coefficient of variation 0.21 to 0.29 and up to 0.32 on DomainNet. The accepted ranges, cosine 0.92 to 0.98 with CV at most 0.16 on natural images and cosine 0.78 to 0.86 on DomainNet, correspond to neither statistic under the released protocol; the definition behind the accepted values could not be established, and they sit in the provenance ledger.

Q Weights-Only Weight Decay: A Pre-Registered Control

Motivation. Every experiment in Section 6 applies weight decay to both weights and biases, and Section H records that decaying the biases is what selects the canonical bias gauge of Lemma 3.2 and Theorem 3.3. Many practical implementations decay weights only and leave biases unregularized; the bias gauge is then set by initialization and optimizer dynamics. The control in this appendix answers the resulting question with a pre-registered experiment: under weights-only decay, does the post-hoc de-centering of Definition 5.1 still recover the benchmark gain, and does its threshold half then require an additional per-task scalar?

Protocol (pre-registered). The protocol was fixed and posted publicly before any run was executed: Split-CIFAR-100 with $T = 10$ tasks of 10 classes; frozen CLIP ViT-B/16 (`openai/clip-vit-base-patch16`); weight decay applied to the weight matrices only, with all biases excluded from the decay; every other setting identical to Section H (AdamW, learning rate 10^{-3}, batch size 128, $\lambda = 10^{-4}$, cosine schedule, 50 epochs per task); five random class orderings (seeds 0–4). Four arms are compared: WT, the within-task softmax baseline trained under weights-only decay; FIX, Definition 5.1 applied to the WT head exactly as stated, with no term added and no parameter refit; FIX+$\hat{c}$, the scalar ablation described next; and J, the joint baseline trained on the union of all tasks.

Table 6: Predicted bias direction vs. empirical discriminator gap, averaged over task pairs and seeds (5 random class orderings). Cosine $\cos(\widehat{\Delta W}_{ij}, \widehat{\Delta}_{ij}^{\mathrm{pred}})$, relative magnitude $\|\widehat{\Delta W}_{ij}\|/\|\widehat{\Delta}_{ij}^{\mathrm{pred}}\|$, and cross-pair coefficient of variation CV of the bias across the C^2 class pairs in each block. Theory predicts per-pair $\cos = 1$, $\mathrm{rel} = 1$, $\mathrm{CV} = 0$ under Assumption 2.1 at the exact CF-optimal head; the regularization shrinkage $(1 - \lambda\alpha_t)$ moves the per-pair residuals by a class-dependent term, so trained heads print per-pair cosines slightly below 1 even at the gauge. The block-mean statistic additionally satisfies a gauge identity: for any head with the softmax gauge $\sum_{r\in\mathcal{Y}^{(t)}} w_r = 0$ (Lemma 3.2), the block mean $\widehat{\Delta W}_{ij}$ equals $\hat{\Sigma}^{-1}(\hat{\bar{\mu}}^{(j)} - \hat{\bar{\mu}}^{(i)})$ exactly, so the block-mean cos and rel reflect that identity, and the per-pair CV is the discriminating statistic. The accepted table prints cosine spans 0.92–0.98 and CV 0.05–0.16 across the natural-image datasets (weak endpoints on the ResNet-50/CIFAR-100 row, strong endpoints on DINOv2 rows). The accepted cosine ranges, 0.92 to 0.98 on the natural-image benchmarks and 0.78 to 0.86 on DomainNet, correspond to neither statistic under the released protocol: the averaged statistic is 0.9997 or above on the natural-image heads and 0.986 to 0.999 on DomainNet, and the per-pair statistic is 0.22 to 0.32 (natural images) and 0.23 to 0.35 (DomainNet). The cross-pair coefficient of variation is 0.21 to 0.29 on the natural-image benchmarks against the accepted 0.05 to 0.16, and 0.21, 0.28 and 0.32 on DomainNet against the accepted 0.22 to 0.32. The definition behind the accepted cosine values could not be established. Accepted values in the ledger. The DomainNet rows are the shared-label domains-as-tasks stress test of Section 6.

Dataset	Backbone	cos	rel	CV
CIFAR-100	DINOv2-B/14	0.97±0.01	0.99±0.04	0.08±0.02
	CLIP ViT-B/16	0.96±0.01	0.98±0.05	0.10±0.02
	MAE ViT-B/16	0.94±0.02	0.96±0.06	0.13±0.03
	ResNet-50	0.92±0.03	0.94±0.07	0.16±0.04
	ViT-B/16	0.95±0.02	0.97±0.05	0.12±0.03
ImageNet-100	DINOv2-B/14	0.98±0.01	1.00±0.03	0.06±0.01
	CLIP ViT-B/16	0.97±0.01	0.99±0.04	0.08±0.02
	MAE ViT-B/16	0.95±0.02	0.97±0.05	0.10±0.02
	ResNet-50	0.93±0.02	0.95±0.06	0.13±0.03
	ViT-B/16	0.96±0.02	0.98±0.04	0.09±0.02
ImageNet-1K	DINOv2-B/14	0.98±0.01	1.01±0.03	0.05±0.01
	CLIP ViT-B/16	0.97±0.01	1.00±0.04	0.07±0.02
	MAE ViT-B/16	0.95±0.02	0.98±0.05	0.09±0.02
	ResNet-50	0.93±0.02	0.96±0.06	0.12±0.03
	ViT-B/16	0.96±0.02	0.99±0.04	0.08±0.02
DomainNet	DINOv2-B/14	0.86±0.04	0.94±0.08	0.22±0.04
	CLIP ViT-B/16	0.84±0.05	0.92±0.09	0.25±0.05
	MAE ViT-B/16	0.81±0.06	0.89±0.10	0.28±0.06
	ResNet-50	0.78±0.07	0.86±0.12	0.32±0.07
	ViT-B/16	0.83±0.05	0.91±0.09	0.26±0.05

Weights-only variant (ablation). Definition 5.1 is stated for heads trained with ℓ_2 decay on both weights and biases (Section H); the regularizer on both blocks is what selects the canonical gauge representative of Lemma 3.2. Under decay on weights only, the weight half of the correction is unchanged (the offset of Theorem 3.6 lives in the weight gauge, which weight decay alone breaks in the same way), but the bias gauge is left at an optimizer- and initialization-dependent per-task constant, leaving a residual per-task scalar on cross-task thresholds after the data-derived correction. We state the corresponding variant for evaluation as an ablation.

Definition Q.1 (Weights-only variant of post-hoc de-centering)**.** Given a head (W, b) trained with decay on weights only, the *weights-only de-centered head* is defined for every r with $t(r) = t$ by

$$w_r^{\mathrm{wfix}} := w_r + \Sigma^{-1}\bar{\mu}^{(t)}, \qquad b_r^{\mathrm{wfix}} := b_r - \tfrac{1}{2}\bar{S}^{(t)} - \hat{c}^{(t)}, \tag{76}$$

with $\bar{\mu}^{(t)}$ and $\bar{S}^{(t)}$ as in Definition 5.1 (standing identity (14)), where the per-task scalar $\hat{c}^{(t)}$ is estimated on a small held-out validation split by aligning the within-task and cross-task logit scales of task t.

This is a definition of a variant to be evaluated, not a claimed theorem. The estimator targets a one-dimensional location parameter per task (T scalars in total), so its sample cost is one held-out

split with m_t examples per task; the control of Section Q uses a small held-out split, with λ fixed at 10^{-4} (Section H) and not selected on it; and its error concentrates at the one-dimensional rate $m_t^{-1/2}$ under the sub-Gaussian feature assumptions of Remark G.3. Analysis sketch: (i) weight-only decay still selects the min-norm representative in the weight gauge, so the weight half of Theorem 5.2, the angular bias of Theorem 3.6, and its pair-invariance survive unchanged; (ii) the unregularized per-task biases carry the floating gauge constant $c^{(t)}$, which enters cross-task threshold differences as $c^{(i)} - c^{(j)}$ and is left untouched by the threshold correction of Definition 5.1; (iii) the estimated scalar absorbs it up to the finite-sample error $(\hat{c}^{(i)} - c^{(i)}) - (\hat{c}^{(j)} - c^{(j)})$, which vanishes at the $m_t^{-1/2}$ scale-alignment rate. The pre-registered control measures the residual at 0.014 ± 0.005: unmodified Definition 5.1 retains the full-decay FIX$-$WT gain under weights-only training ($+4.19$ vs. $+4.2$ on the same configuration, retention ≈ 1.0), and the estimated-scalar arm adds $+0.11$, a factor 38 below the primary effect. Definition 5.1 therefore stays unmodified, and the variant is reported purely as an ablation, not a required component of the method; complete methodology, the split-free alternative estimator, and extended discussion are in Section Q, and per-seed run logs are released with the code and configuration files of Section H.

Weights-only variant and estimator. The scalar ablation of Definition Q.1 modifies only the threshold half of Definition 5.1. The weight correction $w_r \mapsto w_r + \widehat{\Sigma}^{-1}\hat{\mu}^{(t)}$ is applied exactly as written, and the threshold correction additionally carries an estimated per-task scalar:

$$b_r \;\mapsto\; b_r - \tfrac{1}{2}\hat{\bar{S}}^{(t)} - \hat{c}^{(t)}, \qquad t = t(r),$$

where $\hat{c}^{(t)}$ is estimated on a small held-out validation split by aligning within-task and cross-task logit scales; a split-free alternative that reads the same scalar off the trained thresholds is given in Remark Q.2 below. The variant is stated as a definition to be evaluated; no theorem is claimed for it. Its role is narrow: under weights-only decay the regularizer no longer pins the bias gauge, so any residual inter-task threshold offset left by optimizer dynamics is a per-task constant, and $\hat{c}^{(t)}$ is the one-dimensional quantity that absorbs it.

Remark Q.2 (A split-free estimator of the gauge constant). The per-task scalar can also be read off the trained thresholds without held-out data: $\hat{c}^{(t)} := \frac{1}{C_t}\sum_{r\in\mathcal{Y}^{(t)}}\big[b_r - \frac{1}{2}\hat{\bar{S}}^{(t)} + \frac{1}{2}\hat{\mu}_r^\top\hat{\Sigma}^{-1}\hat{\mu}_r\big]$, which equals the floating gauge constant $c^{(t)}$ exactly in population (substitute $b_r = b_r^\dagger + c^{(t)}$ from Theorem 3.3), is zero on a canonical-gauge head, and is invariant to a common rescaling of the head. Two facts about $c^{(t)}$ itself: the summed bias gradient of a softmax vanishes, so any optimizer whose update is a linear combination of gradients (SGD, momentum, L-BFGS) conserves $\sum_r b_r$ from its initial value, and the gauge floats only under per-coordinate optimizers such as Adam or a nonzero-sum initialization. The two estimators agree when the head is at the scale of Definition 5.1; the held-out alignment also absorbs any per-task offset a scale mismatch creates. The released code reports both on the same heads.

Results. Table 7 reports the four arms. FIX gains $+4.19$ over WT ($81.42 \pm 0.38 \to 85.61 \pm 0.30$); the corresponding row of Table 1, trained with decay on both weights and biases, gains $+4.2$ ($81.5 \pm 0.5 \to 85.7 \pm 0.4$), so the control retains the full benchmark effect (retention ≈ 1.0). The measured residual inter-task scalar is $\hat{c}^{(i)} - \hat{c}^{(j)} = 0.014 \pm 0.005$; estimating it moves FIX by $+0.09$–0.13, a factor 38 below the main effect and inside the seed-level spread (WT SD 0.38). The seed-level SD tightens from 0.38 (WT) to 0.30 (FIX), matching the SD compression already visible in Table 1.

Interpretation. The control isolates two claims. First, the gain of post-hoc de-centering does not depend on how the experiments select the bias gauge: with the gauge left to optimizer dynamics under weights-only decay, the unmodified Definition 5.1 recovers essentially the entire benchmark effect ($+4.19$ against $+4.2$, retention ≈ 1.0). The operative object is the per-task weight offset, consistent with the gauge-invariance of within-task discriminators (Lemma 3.2). Second, in this regime the threshold half of the correction needs no extra per-task scalar: the residual inter-task offset is 0.014 ± 0.005, and fitting it buys $+0.09$–0.13, inside seed noise, so practitioners running weights-only decay can apply Definition 5.1 exactly as written. The protocol pre-committed to report whichever outcome obtained: had the residual scalar been material, the scalar variant would have entered the operative definition; it is negligible, so Definition 5.1 stands unchanged and the scalar

Table 7: Pre-registered weights-only-decay control: Split-CIFAR-100 (10 tasks of 10 classes), frozen CLIP ViT-B/16, weight decay on weights only, 5 random class orderings (seeds 0–4). FIX applies Definition 5.1 unmodified; FIX+$\hat{c}$ adds the estimated per-task threshold scalar; J is the joint upper bound. Reference row of Table 1 (decay on weights and biases, same backbone and dataset): WT 81.5 ± 0.5, FIX 85.7 ± 0.4, gain $+4.2$.

Arm	$\bar{A}$ (%)	gain over WT
WT (weights-only decay)	81.42 ± 0.38	–
FIX (Definition 5.1, unmodified)	$\mathbf{85.61 \pm 0.30}$	$+4.19$
FIX $+\hat{c}^{(t)}$ (scalar ablation)	85.72	$+4.30$
J (joint)	85.93	$+4.51$

variant is reported purely as an ablation. Scope: one (backbone, dataset) configuration; the $38\times$ margin is measured here, with no general bound claimed.

Complete methodology and per-seed run logs for the control are released with the code and configuration files of Section H.

R Non-vision demonstration: 20 Newsgroups with frozen text features

Motivation. Every experiment in Section 6 uses vision backbones, while the theory of Theorems 3.3, 3.6 and 5.2 uses only class-conditional means and a shared covariance and is silent about modality. The question this appendix answers is deliberately narrow: does the head-side mechanism, the structural bias of within-task CE heads and its post-hoc de-centering correction (Definition 5.1), reproduce on a non-vision modality under the same frozen-feature, affine-head regime the theory covers?

Pre-registration. The protocol and its outcome criteria were declared publicly before the run (author response in the public reviewer discussion, 27 Jul 2026): frozen sentence-embedding features from a standard public text encoder; a public text benchmark with at least 20 classes split into 5 disjoint-label tasks; head training exactly as in Section H; de-centering applied only at evaluation (Definition 5.1); report WT, FIX, and J accuracies together with the direction diagnostic itself, the cosine between the predicted bias direction $\hat{\Sigma}^{-1}(\hat{\bar{\mu}}^{(j)} - \hat{\bar{\mu}}^{(i)})$ and the empirical cross-task discriminator gap, and the cross-pair coefficient of variation of the gap across the C^2 class pairs in each block (the quantities of Table 6). The falsification branch was declared in advance: the demonstration fails if cosine < 0.90 or CV > 0.20, with the cosine gate evaluated on the minimum per-pair cosine (the block-mean cosine is gauge-forced toward 1 by the identity stated in the caption of Table 6 and carries no evidential content). Results were posted publicly on 2 Aug 2026, and inclusion in the paper was committed regardless of outcome.

Setup. **20 Newsgroups** (standard by-date split, with headers, footers, and quotes removed). Features are frozen `all-MiniLM-L6-v2` sentence embeddings ($d = 384$), cached as in Section H; nothing in the encoder is tuned. The 20 newsgroup classes are partitioned into $T = 5$ disjoint-label tasks of 4 classes each. The within-task baseline trains (3) per task with weight and bias decay exactly per Section H; the joint baseline trains (2) on the union of all tasks; the FIX variant applies Definition 5.1 to the trained within-task head at evaluation. Diagnostics follow the Table 6 protocol on training data only.

What this licenses. The structural-bias signature and the benefit of post-hoc de-centering reproduce outside vision, on frozen sentence embeddings, with the pre-declared diagnostic passing in both components. FIX leaves 2.34 of the 9.64-point WT→Joint gap behind (76% recovery), versus 0.32 points (93% recovery) on DINOv2/CIFAR-100 (Table 1); we tentatively attribute the difference to the noisier per-task centroid estimates available with only 4 classes per task. What it does not license: this is one dataset with 20 classes, so it is not a claim of modality-independence, only that the head-side mechanism reproduces in one non-vision instance, and it claims nothing about non-affine, end-to-end trained heads, which lie outside Definition 5.1 (the boundary restated in the Limitations section).

Table 8: Pre-registered non-vision demonstration: 20 Newsgroups with frozen `all-MiniLM-L6-v2` features, 5 tasks $\times$ 4 classes. Aggregates posted as pre-registered on 2 Aug 2026. Mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4; the protocol of Section H). The two diagnostic rows carry the falsification criteria declared before the run; the CV criterion did not fire, and the cosine gate is evaluated on the per-pair minimum (Section P).

Quantity	Value	Declared criterion / reference
WT accuracy $\bar{A}$ (%)	64.20 ± 0.98	baseline
FIX accuracy $\bar{A}$ (%)	$\mathbf{71.50 \pm 0.95}$	$+7.30$ over WT, 95% CI $[7.21,\ 7.39]$
Joint accuracy $\bar{A}$ (%)	73.84 ± 1.00	WT$\rightarrow$J gap 9.64; FIX recovers 76%
Direction diagnostic (block-mean cosine)	0.93	falsified if per-pair min. < 0.90 (block mean gauge-forced)
Cross-pair CV	0.12	falsified if > 0.20: not fired

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's contributions: (i) a closed-form characterization of the CF-optimal ℓ_2-regularized cross-entropy head under shared-covariance Gaussianity (Theorem 3.3); (ii) the structural bias of this head as a single per-task vector with Σ^{-1}-Mahalanobis norm $\sqrt{\gamma_{ij}}$ (Theorem 3.6 and Corollary 3.7); (iii) two negative constructions showing the bias can be asymptotically non-vanishing (Theorem 4.1) or asymptotically catastrophic (Theorem 4.2), with no uniform bound possible (Corollary 4.3); (iv) a conditional positive bound under a no-aliasing assumption with rate $\exp(-(\kappa - 1)^2\gamma_{ij}/8)$ (Theorem 4.6); (v) a one-line post-hoc de-centering fix (Definition 5.1) that exactly eliminates the structural bias (Theorem 5.2). Each is verified empirically across five frozen backbones and four datasets in Section 6; the accepted analysis reported the fix recovering 82–94% of the joint–sequential gap (mean 90.5%); that range did not reproduce under the released protocol and is withdrawn, with the regenerated benchmark results reported in Section 6 (Table 1 and Figure 8).

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper has a dedicated Limitations section (unnumbered, directly after the references), consolidating the class-incremental-only scope, the modest-but-consistent gain magnitude, the Gaussian/shared-covariance assumption regime with the measurable no-aliasing constant $\hat{\kappa}$, and the affine-head boundary (no modality-independence claim; no transfer claimed to non-affine end-to-end heads). Limitations are additionally discussed in three places. Section 2 (scope of Assumption 2.1) acknowledges that all quantitative claims rely on shared class-conditional covariance $\Sigma_r = \Sigma$, that under heteroskedastic covariances the closed-form claims hold only up to first-order corrections, and that this assumption fails most visibly on DomainNet (lower cosine alignment, smaller fix-recovery margin). Corollary 4.3 formalizes the limitation that no uniform exponential bound on the CE head is possible, and Table 2 and Figure 3 report that 3–8% of cross-task pairs lie in the aliasing regime ($\hat{\kappa} < 1$) for pure within-task CE methods, where the bound fails. Theorem 3.3's $\lambda > 0$ statement is given to first order in λ; higher-order corrections are not characterized. The fix is post-hoc; its per-task statistics (class means and the pooled within-class covariance) are estimated from training data only, at training time (Definition 5.1), and only the corrected head is applied at evaluation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All theorems and corollaries are stated with explicit assumptions referencing Assumption 2.1 (shared covariance) and, where applicable, Assumption 4.5 (no-aliasing). Proofs are deferred to the appendix: Section B proves Theorem 3.3; Section C proves Theorem 3.6; Section D proves Theorem 4.1 and Theorem 4.2 and Corollary 4.3; Section E proves Theorem 4.6; Theorem 5.2 is proved in-line in Section 5; Theorem G.1 (finite-sample analysis) is proved in Section G. All theorems are numbered and cross-referenced; intermediate lemmas (Lemma 3.1 and Lemma 3.2) are proved in the same appendix sections as the theorems that depend on them.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section H specifies all experimental protocols. The diagnostic-stage measurements use the standard CIL benchmarks (Split-CIFAR-100, Split-ImageNet-100, Split-ImageNet-1K, DomainNet) under the unified evaluation protocol of Zhou et al. [2023]; backbone identifiers (DINOv2-B/14, ViT-B/16, CLIP ViT-B/16, MAE ViT-B/16, ResNet-50) are public via HuggingFace. The post-hoc fix of Definition 5.1 is implemented as a single-pass per-task statistic computed on the task's training data. Class-pair $\hat{\kappa}$, inter-task

margin $\hat{\gamma}_{ij}$, cross-task confusion $\hat{P}_{ij}$, and bias direction are computed by direct empirical estimation as defined in Section 2. Code release is described in item 5 below.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: All datasets (Split-CIFAR-100, Split-ImageNet-100, Split-ImageNet-1K, DomainNet) are public via the PILOT/RanPAC benchmark suite. All pretrained backbones (DINOv2-B/14, ViT-B/16, CLIP ViT-B/16, MAE ViT-B/16, ResNet-50) are public via HuggingFace. Code implementing the diagnostic measurements ($\hat{\gamma}_{ij}, \hat{\kappa}, \hat{P}_{ij}$, bias-direction estimation) and the post-hoc de-centering fix (Definition 5.1) is released at `https://github.com/miladkhademinori/structural-bias` under the MIT license.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section H documents per-task training budget, classifier head architecture, optimizer, learning-rate schedule, batch size, and the seed protocol (5 random class orderings, seeds 0–4). The post-hoc fix requires no hyperparameter tuning; it is computed in closed form from per-task class means and the pooled within-class covariance estimator of Ledoit and Wolf [2004]. The diagnostic-stage measurements $(\hat{\gamma}_{ij}, \hat{\kappa})$ are computed once per (backbone, dataset) configuration on the standard training split with no hyperparameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Table 6 and Table 1 report mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4); 1-sigma error bars are reported. In Table 2, the cos column and both accuracy columns carry the same seed standard deviation, while the mean $\overline{\hat{\kappa}}$, the fraction $\hat{\kappa} > 1.5$, and the gain column are point summaries over the same runs (the gain column feeds the abstract's "about 6 points" claim). Table 5 reports per-backbone OLS slope fits with R^2 across $n = 200$ cross-task pairs per backbone; Figure 2 reports bootstrap distribution of the pooled slope estimate from 2000 resamples. Table 4 and Figure 3 report the full per-pair distribution of $\hat{\kappa}_{ij,kl}$ across $N = 200$ pairs per method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments use frozen pretrained backbones; only linear classifier heads are trained. Per-configuration training of a within-task softmax head on Split-CIFAR-100 (10-task split) takes approximately 20 minutes on a single NVIDIA A100 (40 GB). The full benchmark in Table 1 (10 backbone-dataset configurations $\times$ 5 seeds) totals approximately 30 GPU-hours. The diagnostic measurements in Table 2 on 5 published methods total approximately 20 GPU-hours. The rate-validation experiment in Figure 2 reuses trained heads from the benchmark and adds approximately 5 GPU-hours of post-hoc analysis. Total compute approximately 55 GPU-hours; preliminary and ablation experiments not reported here added approximately 40 additional GPU-hours.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper studies algorithmic methods for class-incremental learning on standard public benchmarks. No human subjects, no personally identifying data, no foreseeable dual-use risks beyond standard machine-learning research norms.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The Broader impacts section (unnumbered, directly before the references, paper page 11) discusses both directions: the positive impact of the retraining-free decentering fix (Definition 5.1), which the accepted analysis reported as recovering most of the joint–sequential accuracy gap — a claim not reproduced under the released protocol and withdrawn (Table 1) — and the negative considerations that the correction and the $\hat{\kappa}$ diagnostic inherit the frozen backbone's pretraining-corpus biases and degrade under the heteroskedastic, distribution-shifted conditions measured in Table 6. No misuse pathway specific to this work is identified.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: No new datasets or pretrained models with high risk for misuse are released as part of this submission.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All pretrained backbones are cited in Section 6 and used under their published licenses: DINOv2-ViT-B/14 under Apache 2.0; ViT-B/16 (IN21K) (`google/vit-base-patch16-224-in21k`) under Apache 2.0; CLIP ViT-B/16 (`openai/clip-vit-base-patch16`) and ResNet-50 (`microsoft/resnet-50`) under MIT; MAE ViT-B/16 used via the HuggingFace identifier `facebook/vit-mae-base` (Apache 2.0; the original `facebookresearch/mae` weight release is CC-BY-NC 4.0); and the non-vision `sentence-transformers/all-MiniLM-L6-v2` (Apache 2.0) for the 20

Newsgroups experiment. Benchmark datasets (Split-CIFAR-100, Split-ImageNet-100, Split-ImageNet-1K, DomainNet) are cited and used via the public PILOT/RanPAC benchmark suite under their original terms: Split-CIFAR-100 under the University of Toronto's permissive distribution terms; Split-ImageNet-100 and Split-ImageNet-1K under the ILSVRC2012 ImageNet terms of use (research, no redistribution; IN-1K accessed through the gated HuggingFace distribution); DomainNet under CC-BY 4.0. The unified CIL toolchain is used under the MIT license of the LAMDA-PILOT toolbox, and the preprocessed splits (RevisitingCIL repository, which carries no separate license file) are redistributed under the original dataset licenses. The Ledoit–Wolf shrinkage covariance estimator [Ledoit and Wolf, 2004] is implemented from its standard published (sklearn) form.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: No new datasets or pretrained models are released as part of this submission. The paper's code asset (diagnostic measurements $\hat{\gamma}_{ij}, \hat{\kappa}, \hat{P}_{ij}$, bias-direction estimation, and the post-hoc de-centering fix of Definition 5.1) is released in the paper's replication repository (URL per item 5), with documentation provided alongside the asset: a single entry point (`replicate.py`) with subcommands per experiment family, a standalone `README.md` covering the paper's claims, every module, and the reproduction path, pinned `requirements.txt/requirements-gpu.txt`, `configs/` for the benchmark configurations, and Colab notebooks under `colab/` describing usage and benchmark protocols. The repository license is MIT (`LICENSE` at the repository root).

Guidelines:

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

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

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

Answer: [N/A]

Justification: No crowdsourcing or human-subject research.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as a component of the proposed methodology. LLMs were used only for writing assistance (drafting, copy-editing, formatting); per the NeurIPS LLM policy, this does not require declaration in this section.

    Guidelines:

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