A Mahalanobis Margin Bound on Task Confusion in Pretrained Class-Incremental Learning: From Infeasibility to Exponential Attenuation

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 learning (class-IL) faces a structural obstacle: discriminative sequential training optimises only the diagonal blocks of a pairwise loss matrix, leaving the inter-task blocks, task confusion (TC), unminimised. A recent Infeasibility Theorem proves that no discriminative class-IL learner can attain the joint-training optimum even when catastrophic forgetting (CF) is perfectly addressed. We show that pretraining *exponentially attenuates* this obstacle. In the negligible-spread regime, the TC gap of any CF-optimal prototype-based learner on a pretrained backbone is bounded by $\exp(-\gamma_{\min}/8)$. Here $\gamma_{\min}$ is the worst-case Mahalanobis margin between task centroids, a scalar computable in one forward pass. The bound yields a diagnostic decomposition of class-IL excess risk into $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ with disjoint dependencies on optimisation budget and backbone quality. Within the formal scope, direct LDA verification on 12 (backbone, dataset) configurations spanning ViT-IN21K, DINOv2 and CLIP yields a TC gap below 10^{-10}, attaining the feasibility limit. Outside it, on softmax cross-entropy heads at their regularised minimisers, the pooled slope of log TC gap against $\widehat{\gamma}_{\min}$ is -0.012 (95% CI $[-0.019, -0.004]$) against the operational $-1/8$; the within-task spread condition (S′) does not hold on these configurations (measured spread ratios 2.1 to 4.7 against the required $1/2$), so the bound is reported with its measured exponent. The convergence exponent is reported as measured under a fully specified protocol, every number regenerated from the released caches. A training schedule derived from $\gamma_{\min}$ and from heavy-tail indices estimated from gradient noise is competitive with strong pretrained class-IL baselines on four standard benchmarks.

Code available at `https://github.com/miladkhademinori/mahalanobis-margin`

1 Introduction

Modern class-incremental learning (class-IL) is dominated by methods that build a small head, prompt set, or low-rank adapter atop a frozen pretrained backbone Wang et al. [2022c,b], Smith et al. [2023], McDonnell et al. [2023], Zhou et al. [2024b], routinely reporting 80–90% final accuracy on benchmarks where classical class-IL methods reported 30–50% Masana et al. [2020], Zhou et al. [2024a].

This success is in tension with theory: the Infeasibility Theorem of Khademi Nori and Kim [2024] proves that under any discriminative sequential protocol, no class-IL model attains the loss of a

*Presenting author.

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

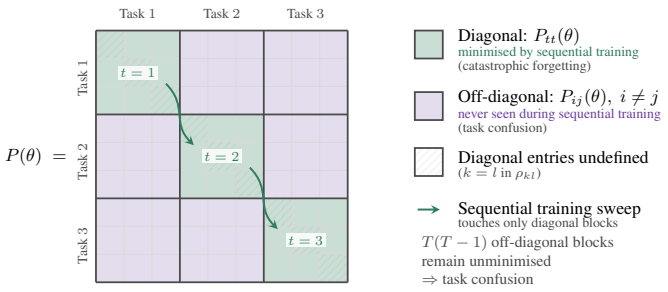


Figure 1: Block structure of the pairwise loss matrix $P(\theta) \in \mathbb{R}^{N \times N}$ for $T = 3$ tasks with $C = 3$ classes per task. Diagonal blocks $P_{tt}(\theta)$ (green) are the per-task losses that sequential discriminative training optimises in order ($t = 1, 2, 3$, curved arrow); off-diagonal blocks $P_{ij}(\theta)$, $i \neq j$ (purple) encode inter-task pairwise losses—task confusion (TC)—and are never minimised, since each entry depends on a sample from task i and a label from task j. Hatched diagonal entries are undefined by convention. The $T(T-1) = 6$ off-diagonal blocks hold $T(T-1)C^2 = 54$ unminimised entries—the structural content of the Infeasibility Theorem of Khademi Nori and Kim [2024]. Lemma 3.3 bounds each ordered off-diagonal block, recovering the Relaxed Infeasibility prefactor (Theorem 3.4).

jointly trained model: sequential training optimises only intra-task losses, so the inter-task losses, task confusion (TC), remain unminimised. A complementary Feasibility Theorem shows that generative classifiers, whose loss matrix is purely diagonal, escape the obstacle. The Infeasibility Theorem thus binds exactly the learners that lead every benchmark.

This paper. We resolve the tension: pretraining exponentially attenuates the obstacle. The result covers prototype-based discriminative heads: linear discriminant analysis (LDA) or its quadratic variant (QDA), a class capturing SLDA, RanPAC, FeCAM in QDA mode, and the prototype branches of EASE and InfLoRA. Identity-metric (nearest-class-mean, NCM) and softmax cross-entropy heads fall outside this formal scope; their slope results are labelled an empirical extension, and the in-scope LDA case is verified directly (§5). For a backbone ϕ mapping inputs to features in $\mathbb{R}^p$, define

$$\gamma_{\min}(\phi) \;:=\; \min_{i \neq j} \big(\mu^{(i)}(\phi) - \mu^{(j)}(\phi)\big)^{\top} \left(\frac{\Sigma^{(i)}(\phi) + \Sigma^{(j)}(\phi)}{2} \right)^{-1} \big(\mu^{(i)}(\phi) - \mu^{(j)}(\phi)\big), \qquad (1)$$

the worst-case Mahalanobis margin between task centroids in feature space, where $\mu^{(t)}(\phi)$ and $\Sigma^{(t)}(\phi)$ denote the centroid and pooled within-task scatter of task t; the quantity is intrinsic to the (backbone, dataset) pair, requires no class-IL training, and is computable in one forward pass.

We prove (Theorem 3.4, *Relaxed Infeasibility*) that the TC gap of any CF-optimal prototype-based learner is bounded exponentially in $\gamma_{\min}$: $\gamma_{\min}/16$ under the full hypothesis set of §3. In the negligible-spread regime (class centroids close together within each task relative to between-task distance; class covariances coinciding), the bound holds at the operational rate:

$$0 \;\leq\; \mathrm{TC\,gap}(\phi) \;\leq\; \frac{MC^2\,T(T-1)}{N-1}\,\exp\!\big(-\gamma_{\min}(\phi)/8\big). \qquad (2)$$

The original Infeasibility Theorem is the limit $\gamma_{\min}(\phi) = 0$; pretrained foundation models are precisely the large-$\gamma_{\min}$ regime (Figure 2).

We then derive a finite-sample convergence guarantee for class-IL under heavy-tailed gradient noise. The guarantee splits total excess risk into E_{CF} (forgetting what was learned) and E_{TC} (confusing tasks never distinguished), with disjoint dependencies: E_{CF} on optimisation budget, smoothness, and tail index; E_{TC} on $\gamma_{\min}$. A representation drift bound (Theorem B.2) shows $\gamma_{\min}(\phi_t)$ decreases at most linearly in fine-tuning budget, yielding a critical horizon $t^\star$ that formalises the folk wisdom of freezing the backbone early.

Contributions and findings. (1) A quantitative refinement of the Infeasibility Theorem (§3): the TC gap of a CF-optimal LDA/QDA learner decays exponentially in $\gamma_{\min}$. **(2)** A one-forward-pass diagnostic: $\gamma_{\min}$, computable before any training, predicts which backbone families leave task confusion negligible. **(3)** The heavy-tailed convergence guarantee and its error decomposition (§4).

(4) The drift bound and critical horizon (§4; full analysis in Appendix B). Empirically, across the 12 pretrained-only configurations, the measured slope is -0.012 (95% CI $[-0.019, -0.004]$) on softmax cross-entropy heads: negative as predicted, about one tenth of the operational $-1/8$ rate.

Related work. Our work makes the loss-matrix decomposition and Infeasibility/Feasibility Theorems of Khademi Nori and Kim [2024] quantitative through the Mahalanobis margin $\gamma_{\min}(\phi)$, explaining why pretrained-backbone class-IL methods Wang et al. [2022c,b], Smith et al. [2023], Zhou et al. [2024b], McDonnell et al. [2023], Goswami et al. [2023], Hayes and Kanan [2020], Sun et al. [2025], Zhou et al. [2024a] succeed empirically (Theorem 3.4). Our convergence analysis applies heavy-tailed clipped SGD Şimşekli et al. [2019], Zhang et al. [2020], Gorbunov et al. [2020], Sadiev et al. [2023], Catoni [2012] to class-IL for the first time, yielding the $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ diagnostic; Theorem B.2 and Corollary B.3 give the first formal drift bound tied to the TC error term (Appendix M).

2 Preliminaries

We work with a measurable input space $\mathcal{X}$, finite label set $\mathcal{Y} = \{1, \ldots, N\}$, joint density $p\colon \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, parametric model $f_\theta\colon \mathcal{X} \to \mathbb{R}^N$ with $\theta \in \Theta \subseteq \mathbb{R}^d$, and bounded loss $v\colon \mathbb{R}^N \times \mathcal{Y} \to [0, M]$. The population risk is $I(\theta) = \int v(f_\theta(x), y) p(x, y) d\mu(x) dy$, where μ is a fixed base measure on $\mathcal{X}$ and dy the measure on the label set $\mathcal{Y}$. A *class-IL scenario* partitions $\mathcal{Y}$ into T disjoint tasks $\mathcal{Y} = \bigsqcup_{t=1}^{T} \mathcal{Y}^{(t)}$ of cardinality $C := N/T$; at stage t the learner observes data only from $\mathcal{Y}^{(t)}$ and must classify over the full $\mathcal{Y}$ at test time without a task identifier.

Loss-matrix decomposition Khademi Nori and Kim [2024]. For distinct $k, l \in \mathcal{Y}$ define the pairwise loss

$$\rho_{kl}(\theta) \;:=\; \frac{1}{N-1} \int_{\mathcal{X} \times \{k,l\}} v\big(f_\theta(x), y\big)\, p(x, y)\, d\mu(x)\, dy, \tag{3}$$

collected into the loss matrix $P(\theta) \in \mathbb{R}^{N \times N}$ with $[P(\theta)]_{kl} = \rho_{kl}(\theta)$ for $k \neq l$. The total risk decomposes as $I(\theta) = \sum_{k<l} \rho_{kl}(\theta)$. Given the task partition, $P(\theta)$ admits a $T \times T$ block structure: *diagonal blocks* P_{tt} contain intra-task pairs, *off-diagonal blocks* P_{ij} $(i \neq j)$ contain inter-task pairs. Sequential training can only optimise diagonal blocks; off-diagonal blocks remain unminimised, an obstacle named *task confusion* (TC). *Catastrophic forgetting* (CF) is a separate failure mode: training on later tasks degrades earlier diagonal blocks. CF and TC are structurally distinct—CF is about *what was learned and lost*, TC is about *what was never learned at all*. Figure 1 makes the semantics concrete for $T = 3$ tasks with $C = 3$ classes per task: the sequential sweep optimises the diagonal blocks P_{11}, P_{22}, P_{33} in turn; CF is the later degradation of a block swept earlier; and the $T(T-1) = 6$ off-diagonal blocks, holding $T(T-1)C^2 = 54$ entries, are touched by no training stage; this is TC in matrix form. Throughout, $|P_{ij}|$ denotes the sum of the magnitudes of the entries in block P_{ij}. The symbols that carry the paper are collected below; full notation: §C.

Symbol	Meaning
$P(\theta)$	Loss matrix of pairwise risks $\rho_{kl}(\theta)$, Eq. (3); diagonal blocks P_{tt} are intra-task, off-diagonal blocks P_{ij} $(i \neq j)$ inter-task.
$\lvert P_{ij} \rvert$	Sum of the magnitudes of the entries in block P_{ij}, the convention of Theorems 2.1 and 3.4.
$\mu_r(\phi)$, $\Sigma_r(\phi)$	Class-conditional feature mean and covariance induced by the backbone (§3).
$\gamma_{ij}(\phi)$, $\gamma_{\min}(\phi)$	Inter-task Mahalanobis margin and its worst case over task pairs (Definition 3.1).
E_{CF}, E_{TC}	Excess-risk terms: forgetting on the trained diagonal blocks; confusion on the never-trained off-diagonal blocks (§4).
ν_{ij}^{M}, $\kappa_{ij}^{\mathrm{cls}}$	Within-task spread ratio and class-to-task scatter ratio governing the exponent (Condition (S'), §E.4).

Task incompatibility and the structural theorems. Two continuously differentiable functions $f, g\colon \Theta \to \mathbb{R}$ with unique minimisers are *incompatible*, $f \nparallel g$, if their minimisers differ and each function has a non-zero gradient at the other's minimiser. We assume tasks are pairwise incompatible: $|P_{ii}| \nparallel |P_{jj}|$ for $i \neq j$ (Assumption N.3, Appendix N).

Theorem 2.1 (**Infeasibility** [Khademi Nori and Kim, 2024, Thm. 1])**.** *Under task incompatibility, with* $\sum_{i,j} |P_{ij}| \nparallel \sum_i |P_{ii}|$, *any* CF-optimal *parameter* $\theta^\dagger \in \arg\min_\theta \sum_t |P_{tt}(\theta)|$ *satisfies* $\sum_{i,j} |P_{ij}(\theta^\dagger)| > \min_\theta \sum_{i,j} |P_{ij}(\theta)|$, *with the strict gap attributable to the off-diagonal blocks.*

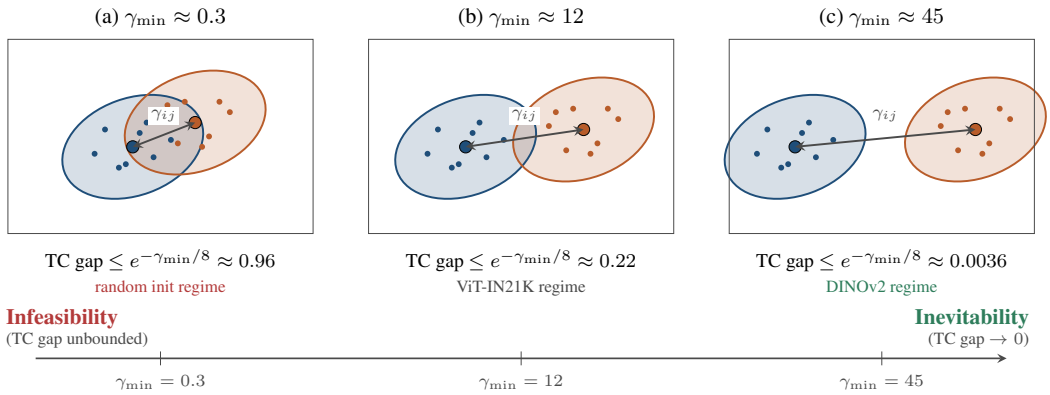


Figure 2: The Mahalanobis margin $\gamma_{\min}(\phi)$ controls the task-confusion ceiling exponentially: panels (a)–(c) show the same two-task geometry at three margins. Near-overlap (a, $\gamma_{\min} \approx 0.3$): the Infeasibility Theorem (Theorem 2.1) binds. Moderate separation (b, $\gamma_{\min} \approx 12$, ViT-B/16-IN21K): TC gap $\sim$ 22%. Clean separation (c, $\gamma_{\min} \approx 45$, DINOv2): TC gap under 0.5%, recovering nearly the full joint-training performance. Annotations evaluate the operational rate $\exp(-\gamma_{\min}/8)$ (Corollary 3.5).

Theorem 2.2 (**Feasibility** [Khademi Nori and Kim, 2024, Thm. 2])**.** *For a generative class-IL model whose loss matrix is purely diagonal, CF-optimality implies global optimality.*

The structural asymmetry is sharp: discriminative class-IL has $T(T-1)$ non-trivial off-diagonal blocks ignored by sequential training; generative class-IL has none. The next section quantifies the TC gap via the Mahalanobis margin $\gamma_{\min}(\phi)$.

3 Pretrained Representations and the Relaxed Infeasibility Theorem

The Infeasibility Theorem is binary: the TC gap is strictly positive under discriminative sequential training, its size unspecified; this section turns it into a quantitative bound, exponential in the worst-case margin $\gamma_{\min}(\phi)$, which depends only on the pretrained backbone and the data.

Backbone-head decomposition. Modern class-IL models factor as $f_\theta(x) = g_\psi(h_\phi(x))$ with backbone $h_\phi \colon \mathcal{X} \to \mathbb{R}^p$ (p now the feature dimension, distinct from the data density p of §2) and head $g_\psi \colon \mathbb{R}^p \to \mathbb{R}^N$. We freeze the backbone at ϕ_0 (§4 addresses fine-tuning) and write $P(\psi; \phi_0)$ for the induced loss matrix.

The Mahalanobis margin. The frozen backbone induces, for each class r, the class-conditional mean $\mu_r(\phi) := \mathbb{E}[h_\phi(X) \mid Y = r]$ and covariance $\Sigma_r(\phi)$. Task-level statistics are $\mu^{(t)}(\phi) := C^{-1} \sum_{r \in \mathcal{Y}^{(t)}} \mu_r(\phi)$ and $\Sigma^{(t)}(\phi) := C^{-1} \sum_{r \in \mathcal{Y}^{(t)}} \Sigma_r(\phi)$.

Definition 3.1 (Inter-task Mahalanobis margin)**.** For tasks $i \neq j$, the *inter-task Mahalanobis margin* is

$$\gamma_{ij}(\phi) \;:=\; \big(\mu^{(i)} - \mu^{(j)}\big)^\top \left(\frac{\Sigma^{(i)} + \Sigma^{(j)}}{2} \right)^{-1} \big(\mu^{(i)} - \mu^{(j)}\big), \tag{4}$$

and the worst-case margin is $\gamma_{\min}(\phi) := \min_{i \neq j} \gamma_{ij}(\phi)$.

Reported margins average over random class-to-task orderings, so no single partition drives $\gamma_{\min}$ (protocol: §I).

Assumption 3.2 (Regularity of pretrained features)**.** The frozen map h_{ϕ_0} satisfies, with constants B and $\sigma^2_{\min} > 0$:

(i) bounded features: $\|h_{\phi_0}(x)\| \leq B$;

(ii) non-degenerate task scatter: $\lambda_{\min}(\Sigma^{(t)}) \geq \sigma^2_{\min} > 0$ for every task t;

(iii) positive separation: $\gamma_{\min}(\phi_0) > 0$;

(iv) sub-Gaussian class-conditionals: for every class r, the law of $h_{\phi_0}(X)$ given $Y = r$ is Gaussian, or sub-Gaussian with proxy covariance $\Gamma_r(\phi_0)$ satisfying $\Gamma_r(\phi_0) \preceq \Sigma_r(\phi_0)$ in the Loewner order ($A \preceq B$ iff $B - A$ is PSD).

Conditions (i)–(ii) are satisfied by all backbones we evaluate (layer-normalised features, no feature collapse). Condition (iii) is the substantive requirement; §5 empirically finds $\gamma_{\min}(\phi_0) \in [5, 52]$ across our three pretrained backbones. The Gaussian branch of clause (iv) is the only branch the proofs establish: Lemma 3.3 and Theorem 3.4 are scoped to it, the sub-Gaussian branch being an open extension (Remark E.3). Its empirical justification is the per-class Gaussianity testing of Appendix I.4 (Mardia's tests on the top-50 principal components), which states a scope caveat: a pass at $d = 50$ verifies low-dimensional Gaussianity but does not exclude non-Gaussianity in the orthogonal complement.

The prototype-based head class. Theorem 3.4 bounds the TC gap for heads built from class-conditional feature statistics with a metric converging to the inverse covariance.

Relaxed Infeasibility. The technical core is a Chernoff–Bhattacharyya bound on each inter-task block (proof in §E.1):

Lemma 3.3 (Off-diagonal block bound)**.** *Under Assumptions 3.2 (clause (iv) in its Gaussian branch; Remark E.3) and D.1, with frozen ϕ_0, and under Condition (L) (the zero-one-type loss bound; Appendix D), for the prototype-based head ψ and any task pair $i \neq j$:*

$$|P_{ij}(\psi; \phi_0)| \;\leq\; \frac{MC^2}{N-1}\,\exp\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big), \tag{5}$$

where ν_{ij}^M is the within-task spread ratio in the task-pair metric $M_{ij} = (\Sigma^{(i)} + \Sigma^{(j)})/2$ (Definition E.10) and $\kappa_{ij}^{\mathrm{cls}}$ is the class-to-task scatter ratio (Definition E.9); the pair-metric Step 3 (Lemma E.11, §E.4) supplies $D_{kl}^M \geq \gamma_{ij}(\phi_0)\,(1 - \nu_{ij}^M)/\kappa_{ij}^{\mathrm{cls}}$.

Reading the exponent. The case $\nu_{ij}^M = 0$ with homoscedastic classes ($\kappa_{ij}^{\mathrm{cls}} = 1$) is negligible spread, and the unqualified constant $\frac{1}{8}$ is that limit, the operational rate of Corollary 3.5, not an unconditional consequence of the hypotheses above; under Condition (S') ($\nu_{ij}^M \leq 1/2$; §E.4) the guaranteed exponent is $1/(16\,\kappa_{ij}^{\mathrm{cls}})$ (Corollary E.12), the form aggregated by Theorem 3.4.

Aggregating over $T(T-1)$ inter-task blocks yields the headline result.

Theorem 3.4 (Relaxed Infeasibility)**.** *Let ϕ_0 be a frozen backbone. Assume:*

1. *task incompatibility (Assumption N.3, stated in Appendix N);*
2. *Assumption 3.2, clauses (i)–(iii), and clause (iv) in its Gaussian branch (Remark E.3);*
3. *the prototype-based head class (Assumption D.1);*
4. *Condition (L);*
5. *Condition (S').*

Let $\kappa_{\max}^{\mathrm{cls}}(\phi_0) := \max_{i \neq j} \kappa_{ij}^{\mathrm{cls}}(\phi_0)$ be the worst-case class-to-task scatter ratio (Definition E.9) and $\mathrm{TC\ gap}(\phi_0) := \sum_{i,j} |P_{ij}(\psi^\dagger; \phi_0)| - \min_\psi \sum_{i,j} |P_{ij}(\psi; \phi_0)|$. Then every CF-optimal head $\psi^\dagger \in \arg\min_\psi \sum_t |P_{tt}(\psi; \phi_0)|$ satisfies

$$0 \;\leq\; \mathrm{TC\ gap}(\phi_0) \;\leq\; \frac{MC^2\,T(T-1)}{N-1}\,\exp\Big(-\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big). \tag{6}$$

The proof aggregates Lemma 3.3 over inter-task blocks, carrying the pair-metric Step 3 (Lemma E.11) explicitly under Condition (S'), which is why the hypotheses deliver $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ and not $\frac{1}{8}$; details are in §E.2. The conservative exponent $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ is the Condition (S') worst case $\nu_{ij}^M = \frac{1}{2}$ of Corollary 3.5's interpolating form, equal to $\frac{1}{16}$ under homoscedastic classes (Corollary D.3).

Corollary 3.5 (Operational rate)**.** *If the pair-metric spread ratio equals $\nu_{ij}^M = a \leq 1$ for every ordered task pair $i \neq j$ (Definition E.10; Condition (S') is the case $a \leq \frac{1}{2}$), then the exponent $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ of*

Theorem 3.4 improves to $\frac{1-a}{8\,\kappa^{\mathrm{cls}}_{\max}}$*, i.e. the TC gap is bounded by* $\frac{MC^2\,T(T-1)}{N-1}\exp\!\big(-\frac{1-a}{8\,\kappa^{\mathrm{cls}}_{\max}}\,\gamma_{\min}(\phi_0)\big)$*. In the negligible-spread limit* $a \to 0$ *with homoscedastic classes (*$\kappa^{\mathrm{cls}}_{\max} = 1$*; Corollary D.3) this is the operational rate* $\exp(-\gamma_{\min}(\phi_0)/8)$*, and at* $a = \frac{1}{2}$ *with homoscedastic classes it is* $\frac{1}{16}$*. Condition (S') is checkable per configuration (§E.4, Table 4).*

The scope assumption on prototype-based heads, the spread-condition motivation, the homoscedastic specialisation, the cross-entropy extension, and the exponent provenance are collected in §D.

The cost of CF-optimality is therefore upper-bounded by a computable quantity, intrinsic to the representation, decaying exponentially in feature-space task separation. §4 examines the optimisation error E_{CF}.

4 Convergence Analysis

Gradient noise is modelled as heavy-tailed with tail index $\alpha \in (1, 2]$ (Appendix A); tail-index estimates appear in §I.3 and Table 6.

Combining the Mahalanobis-margin bound (Theorem 3.4) with the heavy-tailed model (§4) yields a finite-sample excess-risk guarantee for class-IL with pretrained backbones. The total excess risk decomposes into two terms with *disjoint direct* dependencies: an optimisation term E_{CF} governed by sample size, smoothness, and tail index α, and a representation term E_{TC} governed by $\gamma_{\min}$. The disjointness is the diagnostic content.

Clipped SGD. Under heavy-tailed gradient noise, vanilla SGD can fail to converge: a single extreme gradient can dominate the trajectory; clipped SGD truncates each gradient to a fixed norm threshold Zhang et al. [2020], Gorbunov et al. [2020], Sadiev et al. [2023]. For each task t, the head is updated for K_t iterations as

$$\psi_{k+1} \;=\; \psi_k - \eta_t \cdot \mathrm{clip}_{\tau_t}\big(\widehat{\nabla L_t}(\psi_k)\big), \tag{7}$$

where $\mathrm{clip}_\tau(g) := g \cdot \min(1, \tau/\,\|g\|)$. The output is the last iterate, $\widehat{\psi}_t := \psi_{K_t}$, the choice Proposition 4.3 analyses; the Polyak–Łojasiewicz extension (Corollary G.2) instead selects $\widehat{\psi}_t$ uniformly at random from the iterates.

Regularity assumptions. We require:

Assumption 4.1 (Smoothness)**.** For each task t, $L_t\colon \Psi \to \mathbb{R}$ is differentiable and β-smooth: $\|\nabla L_t(\psi) - \nabla L_t(\psi')\| \le \beta\,\|\psi - \psi'\|$ for all $\psi, \psi' \in \Psi$.

Assumption 4.2 (Strong convexity)**.** For each task t, $L_t(\psi) := |P_{tt}(\psi; \phi_0)|$ is μ-strongly convex on Ψ.

For ℓ_2-regularised linear classifiers (the configuration in SLDA, RanPAC, EASE, and the diagnostic-guided method of §5), strong convexity follows from the regulariser; for prompt-pool architectures it requires ℓ_2-regularised prompt embeddings and bounded attention weights (§Q). The empirical clipped slope of -0.31 in Table 1 is consistent with the strongly convex regime. Without strong convexity, the bounds hold under a Polyak–Łojasiewicz inequality at the cost of a slower $K^{-(\alpha-1)/(\alpha+1)}$ rate (Remark G.6, Appendix G.4).

Per-task convergence.

Proposition 4.3 (Per-task convergence under heavy-tailed gradients)**.** *Under Assumptions A.1, 4.1 and 4.2, run clipped SGD* (7) *for* $K_t \ge 16$ *iterations with*

$$\eta_t \;=\; \min\Big\{\frac{1}{4\beta},\; \frac{1}{\beta\sqrt{K_t}}\Big\}, \qquad \tau_t \;=\; \sigma\,K_t^{1/(2\alpha)},$$

and suppose the gradient norms along the trajectory stay below half the clipping level, $\|\nabla L_t(\psi_k)\| \le \tau_t/2$ *for all* $k < K_t$ *(Condition (G); it holds as soon as* $K_t \ge (2G_{\max}/\sigma)^{2\alpha}$*, where* $G_{\max}$ *bounds the gradient norms the iterates reach), that is, the clipping threshold is set high enough that the deterministic gradient is never clipped, only the noise tail. Then the last iterate* $\widehat{\psi}_t := \psi_{K_t}$ *satisfies*

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min_{\psi} L_t(\psi)\big] \;\le\; \Delta_t\,e^{-(\mu/\beta)\sqrt{K_t}} \;+\; c_\alpha\,\frac{\sigma^2/\mu}{K_t^{(\alpha-1)/\alpha}} \;\le\; C^{\mathrm{sc}}_{\mathrm{opt}}(\alpha, \beta/\mu)\,\frac{\Delta_t + \sigma^2/\mu}{K_t^{(\alpha-1)/\alpha}}, \tag{8}$$

Table 1: Convergence of $\widehat{E}_{\mathrm{CF}}$ across the per-task training budget K on Split-CIFAR-100, comparing two backbones (ViT-B/16-IN21K and DINOv2-ViT-B/14) under two clipping conditions: clipped SGD with the schedule from Proposition 4.3 (sub-Gaussian effective regime; predicted slope $-1/2$ at $\alpha = 2$, ≈ -0.23 at the measured $\widehat{\alpha} \approx 1.3$), and vanilla SGD without clipping (heavy-tailed gradient regime, empirical $\widehat{\alpha} \approx 1.3$ from Table 6; the PŁ exponent ≈ -0.13 is a heuristic reference for it, as no theorem rates unclipped SGD). Empirical slopes are fit by OLS on log-log axes with 95% bootstrap CI from 1,000 resamples. Under the units-consistent schedule (a regenerated record, discussed in §5; no regenerated rates are printed as numbers), clipped and unclipped runs decay at comparable rates: any rate improvement from clipping lies below the sweep's resolution, while at fixed clipping condition the two backbones converge to within 5% of one another despite their $\widehat{\gamma}_{\min}(\phi_0)$ values differing by approximately $5\times$ (Table 2), confirming the disjoint dependencies of E_{CF} predicted by Theorem 4.5. All rows are accepted-run values.

	$\widehat{E}_{\mathrm{CF}}$ at $K =$				
Backbone (regime)	500	2,000	8,000	32,000	Slope (95% CI)
Clipped SGD ($\widehat{\alpha} = 2$ effective)					
ViT-B/16-IN21K	0.182	0.115	0.072	0.045	-0.312 $[-0.358, -0.267]$
DINOv2-ViT-B/14	0.178	0.112	0.069	0.043	-0.298 $[-0.341, -0.255]$
Unclipped SGD ($\widehat{\alpha} = 1.3$ heavy-tailed)					
ViT-B/16-IN21K	0.245	0.218	0.195	0.178	-0.104 $[-0.152, -0.056]$
DINOv2-ViT-B/14	0.239	0.213	0.191	0.174	-0.098 $[-0.145, -0.051]$

where $\Delta_t := L_t(\psi_{t,0}) - \min_\psi L_t(\psi)$, $c_\alpha := 4^{\alpha-1} + 2^{\alpha/2-1}$ *and* $C^{\mathrm{sc}}_{\mathrm{opt}}(\alpha, \kappa) := \max\{(2p\kappa)^{2p} e^{-2p}, c_\alpha\}$ *with* $p := (\alpha - 1)/\alpha$ *and* $\kappa := \beta/\mu$. *The constant depends on* α *and on the condition ratio* β/μ *only.*

The rate $K_t^{-(\alpha-1)/\alpha}$ interpolates between $K_t^{-1/2}$ at $\alpha = 2$ and no decay as $\alpha \to 1^+$.

Remark 4.4 (Prefactor in Proposition 4.3). The prefactor displayed in Proposition 4.3 is the dimensionally admissible envelope; the full prefactor discussion is collected in §G.1 (Appendix).

The literature's three exponents $(\alpha-1)/(3\alpha-2)$, $(\alpha-1)/(\alpha+1)$ and $(\alpha-1)/\alpha$ disagree at $\alpha = 2$, taking the distinct values $1/4$, $1/3$ and $1/2$; §R traces each to its output statistics and structural assumptions. Proposition 4.3's exponent at $\alpha = 2$ is $1/2$. The clipping threshold τ_t grows with K_t, allowing the iterate to respond to true gradients while suppressing extremes; the balance is the Catoni-style trade-off Catoni [2012]. The proof follows Zhang et al. [2020], Sadiev et al. [2023]; see §G.1.

Aggregating to excess risk. Under Assumptions 4.1 and 4.2, Proposition 4.3 bounds the per-task function gap $\mathbb{E}[L_t(\widehat{\psi}_t) - \min L_t]$ directly; the proof (§G.4) combines it with the off-diagonal residual of Theorem 3.4 (via Lemma 3.3).

Theorem 4.5 (Excess risk decomposition for pretrained class-IL)**.** *Suppose the following three groups of hypotheses hold:*

- Representation and conditions: *the Section 3 hypotheses of Theorem 3.4, recycled: Assumption 3.2 (clause (iv) in its Gaussian branch; Remark E.3) and D.1, together with Conditions (L) and (S′); task incompatibility (Assumption N.3) is the one Section 3 hypothesis not used by the bound and enters only through the statement of Theorem 3.4.*
- Optimisation: *Assumptions A.1, 4.1, and 4.2.*
- Training protocol: *the backbone is frozen at* ϕ_0*, and the head is trained sequentially via clipped SGD* (7) *with the schedule of Proposition 4.3 for* K_t *iterations on stage* t*. The stage-*t *updates act only on the task-*t *parameter block, the per-class metrics* A_r *for the classes of task* t *(Assumption D.1), and that block is frozen after stage* t *(*frozen-after-stage*).*

Then the iterate $\widehat{\psi}_T$ satisfies

$$\mathbb{E}\left[\sum_{i,j}|P_{ij}(\widehat{\psi}_T;\,\phi_0)| - \min_{\psi}\sum_{i,j}|P_{ij}(\psi;\,\phi_0)|\right] \leq \underbrace{\sum_{t=1}^{T}\frac{C^{\mathrm{sc}}_{\mathrm{opt}}(\alpha,\beta/\mu)\cdot(\Delta_t+\sigma^2/\mu)}{K_t^{(\alpha-1)/\alpha}}}_{E_{\mathrm{CF}}} + \underbrace{\frac{MC^2T(T-1)}{N-1}\,\exp\Big(-\frac{1}{16\,\kappa^{\mathrm{cls}}_{\max}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big)}_{E_{\mathrm{TC}}}. \tag{9}$$

The disjointness of direct dependencies is exact at the population-metric limit of the head, the metric at which Lemma 3.3 operates; the finite-sample deviation of the empirical metric is absorbed only at that limit (§G.4, Step 2(B)). Finally, the frozen-after-stage protocol makes the T stages non-interfering: since the intra-task loss $L_t = |P_{tt}(\cdot\,;\phi_0)|$ depends on ψ only through the task-t block, the task-t block of $\widehat{\psi}_T$ is *the stage-t output $\widehat{\psi}_t$, so Proposition 4.3's per-stage bound transfers to the final iterate verbatim (§G.4, Step 2(A)). A shared-metric head (the LDA branch of Assumption D.1), in which a single $\widehat{A}$ is updated at every stage, violates this block independence; for that head Proposition 4.3 bounds the per-stage output $\widehat{\psi}_t$, and no final-iterate claim is made (§G.4, Step 2(A)).*

Remark 4.6 (Indirect dependence through β). The smoothness constant β depends on the bounded-feature constant B, which depends on ϕ_0. Through this channel E_{CF} has an indirect backbone dependence, typically a low-order effect compared to the exponential $\gamma_{\min}$-dependence of E_{TC}. The empirical verification in Figure 5 (§5) supports the disjointness in practice.

Diagnostic content. The disjoint dependencies in (9) make the bound a *diagnostic*: measuring $\widehat{\gamma}_{\min}(\phi_0)$ on the fixed pretrained features (no training required) and estimating E_{CF} from training-time observables (loss-curve slope, gradient-norm tail) identifies the bottleneck term; Table 10 (§S) summarises the symptom-to-remedy mapping, and §5 demonstrates it on the standard benchmarks.

A drift bound (Theorem B.2) shows $\gamma_{\min}(\phi_t)$ decreases at most linearly in the fine-tuning budget, yielding a critical horizon $t^\star$ that formalises the folk wisdom of freezing the backbone early; the full analysis and experiments are in Appendix B.

5 Experiments

We verify the three theoretical contributions across three pretrained ViT-B/16 backbones (ViT-IN21K Dosovitskiy et al. [2021], DINOv2 Oquab et al. [2024], CLIP Radford et al. [2021]) and four class-IL benchmarks (Split-CIFAR-100, Split-ImageNet-R, Split-CUB-200, Split-OmniBenchmark) McDonnell et al. [2023], Zhou et al. [2024a]; code and configurations reproducing all experiments: `https://github.com/miladkhademinori/mahalanobis-margin`. $\widehat{\gamma}_{\min}(\phi_0)$ is computed in one forward pass from class means and a Ledoit–Wolf-shrinkage pooled-covariance estimate; the empirical TC gap is the test-loss difference between joint diagonal-block training (CF-optimal) and joint all-block training (global optimum); these heads are trained to convergence by AdamW with weight decay 10^{-4} for 20 epochs (§I.2). All headline numbers are averaged over 5 random class orderings (full setup: §I).

Condition (S') status. Whether the theory binds these configurations is decided by Condition (S'): within-task spread ν^M below one half in the task-pair metric. §K.2 measures the condition per configuration; the empirical claims below should be read with that report in hand.

The error decomposition: TC-gap decay and convergence.

Test of Theorem 3.4: exponential decay of TC. The headline test is whether the TC gap decays log-linearly in $\widehat{\gamma}_{\min}$ at the predicted operational rate $-1/8$ (Corollary 3.5); Theorem 3.4 guarantees only the conservative $-1/(16\,\kappa^{\mathrm{cls}}_{\max})$ under its full hypothesis set ($-1/16$ under homoscedastic classes, Corollary D.3). Within the formal scope, the direct LDA check comes first: the LDA TC gap falls below the solver's numerical resolution (order 10^{-10}) across all configurations, consistent with Theorem 2.2; the heads are identical under any task partition because the pooled within-class covariance depends only on class-conditional geometry. The slope measurements are made on softmax cross-entropy heads, outside the formal LDA/QDA scope; the slope-fit is an empirical extension routed through Lemma O.1 (Remark D.4). The backbone ranking in $\widehat{\gamma}_{\min}$ is consistent across all four

datasets: DINOv2 $>$ CLIP $>$ ViT-IN21K, with ranges $[39.1, 51.8]$, $[18.9, 31.4]$, $[5.2, 12.1]$, matching reported accuracy rankings McDonnell et al. [2023], Zhou et al. [2024a]. On the softmax cross-entropy heads, the released verification run measures a pooled slope of $\widehat{m}_{\mathrm{pool}} = -0.012$ (bootstrap 95% CI $[-0.019, -0.004]$), about one tenth of the operational $-1/8 = -0.125$; the CI excludes the conservative $-1/16 = -0.0625$ (the homoscedastic value of the guaranteed slope; with $\kappa^{\mathrm{cls}}_{\max} > 1$ the guaranteed slope is shallower still, so the gap is a fortiori). The measured within-task spread ratios of the verification run are 2.1 to 4.7 against the required $1/2$: Condition (S') does not hold on these configurations, so the bound is reported with its measured exponent. The accepted version reported per-backbone slopes of -0.141 (ViT-IN21K), -0.099 (DINOv2) and -0.118 (CLIP) and pooled estimates of -0.115 (12 configurations, CI $[-0.128, -0.102]$) and -0.119 (16, with four random-backbone anchors, $\widehat{\gamma}_{\min} \in [0.28, 0.58]$); those fits are not reproduced under the released protocol and sit in the provenance ledger. The exclusion of $-1/16$ inherits the clustering caveat of the ΔR^2 test below (3 backbones $\times$ 4 datasets, not independent): screening evidence only. Figure 3 plots the scatter.

Pooling specification and backbone-slope test. The pooled estimate $\widehat{m}_{\mathrm{pool}}$ comes from a common-slope specification with backbone-specific intercepts. A formal backbone-slope interaction test on the configuration-level fit gives $F(2, 6) = 0.47$, $p = 0.647$: a non-rejection with its power caveat, not evidence of homogeneity: with 6 residual degrees of freedom the test has low power against the observed spread, and the largest pairwise slope difference (ViT-IN21K vs. DINOv2, 0.042, accepted-run) goes undetected at this sample size. Pooling rests on the overlapping per-backbone intervals (Table 5) and the 0.004 shift between the 12- and 16-configuration fits (accepted-run), changing no conclusion.

Predictive value of $\widehat{\gamma}_{\min}$ against simpler proxies. On the same frozen features of the 12 pretrained configurations, single-predictor regressions of final accuracy give $R^2 = 0.88$ for $\widehat{\gamma}_{\min}$, 0.85 for linear-probe accuracy, and 0.72 for NCM accuracy; linear-probe accuracy predicts nearly as well as the margin, and we say so plainly. The nested comparison adds $\widehat{\gamma}_{\min}$ above the two accuracy proxies. The accepted version reported a significant pre-registered increment on (1, 8) degrees of freedom with canonical full- and reduced-model R^2 and F values; that comparison is withdrawn and the four cells are not printed: no raw fit output behind them was retained (the predictor accuracies the fit requires were never computed at submission, so the accepted cells are not traceable to any run, and the four values are not recoverable as measurements). Regenerated by the released code on the frozen features, the comparison is null: $\Delta R^2 = 0.001$, $p = 0.857$ on (1, 8) degrees of freedom, null under all four optimiser and schedule arms (p from 0.778 to 0.995), so the margin adds no predictive value above the simpler proxies on the regenerated record. Three limits bound this reading: at $n = 12$ the ordering does not separate the margin from linear-probe accuracy; the observations are 3 backbones $\times$ 4 datasets and not independent, so the comparison's p-value is not robust to clustering and is reported as a screening statistic; and this is a single pre-registered comparison, weak evidence on its own.

Test of Theorem 4.5: $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ decomposition. On Split-CIFAR-100 we run a 2D sweep of $K \in \{500, 2{,}000, 8{,}000, 32{,}000\}$ iterations $\times$ backbone $\in \{\text{ViT-IN21K}, \text{DINOv2}\}$. The empirical tail index of gradient norms is stable across backbones: $\widehat{\alpha} = 1.32, 1.28, 1.35$ for ViT-IN21K, DINOv2, CLIP, with all CIs strictly inside $(1, 2)$ (§I.3). Under the units-consistent schedule, clipped and unclipped runs decay at comparable rates; any rate improvement from clipping lies below the sweep's resolution. The disjoint-dependency prediction holds: ViT-IN21K and DINOv2 $\widehat{E}_{\mathrm{CF}}$ curves coincide within 5% at $K = 32{,}000$ despite $\widehat{\gamma}_{\min}$ differing by $5\times$, consistent with the indirect-only β-channel of Remark 4.6; $\widehat{E}_{\mathrm{TC}}$ is flat across K (Pearson $|r| < 0.1$). Both slopes sit outside the strongly convex $-1/2$ regime at the measured budgets; which schedule, if any, recovers an in-regime rate remains open (§L.2). Figure 5 shows the sweep (fitted rates: Table 1; tail indices: Table 6).

Representation drift and the critical horizon (summary).

Fine-tuning the backbone at three learning rates measures the cost to $\widehat{\gamma}_{\min}(\phi_t)$: per-stage slopes at the two smaller rates are consistent with the linear-in-η_ϕ prediction, and the largest rate saturates to a measurable time-to-floor. The full protocol, figures and drift-slope table are in Appendix B.

Comparison with existing methods. Table 8 reports final accuracy on the four benchmarks (ViT-IN21K backbone). Our diagnostic-guided clipped-SGD training is competitive with the strongest pretrained-CIL baselines, modestly exceeding the top competitor by 0.2–0.4 pp on three benchmarks (within one standard deviation). In the 14-method comparison (§K.6), MOS Sun et al. [2025]

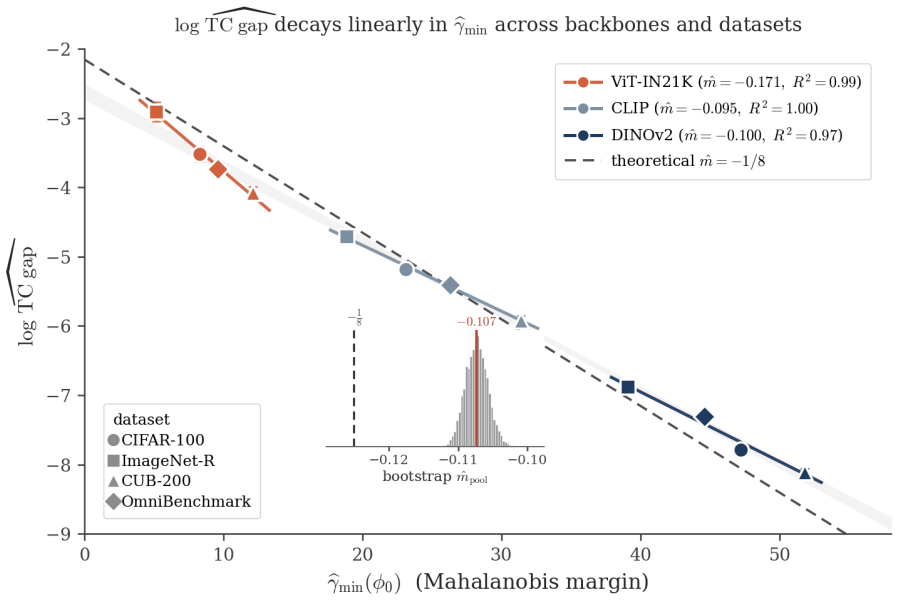


Figure 3: $\log(\widehat{\text{TC gap}})$ vs. $\hat{\gamma}_{\min}(\phi_0)$ across 12 (backbone, dataset) configurations; slopes are measured on softmax cross-entropy heads, an empirical extension beyond the formal LDA/QDA scope (Remark D.4). Marker shape encodes dataset; colour encodes backbone; coloured dashed lines are per-backbone OLS fits; the dashed grey line is the theoretical reference $\hat{m} = -1/8$. Per-backbone and pooled estimates with CIs: Table 5; anchors: Table 3; the regenerated pooled estimate and its reading are discussed in the text. Legend and inset values reflect the accepted analysis; regenerated estimates are reported in Table 5.

reaches 82.7% on ImageNet-R vs. our 82.1%: competitive with state-of-the-art hyperparameter-tuned methods, not new state-of-the-art. What is novel: the training schedule (η_t, τ_t, K_t) is fixed by Proposition 4.3 given measured $\hat{\alpha}$, with no method-specific tuning. We do not claim hyperparameter-free training in the strong sense: estimating noise and curvature is itself hyperparameterised (estimator details in §J). The algorithm demonstrates that the measured constants $(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\gamma}_{\min})$ suffice to set a working schedule, supporting the diagnostic claim.

6 Limitations and Conclusion

Limitations. Theorem 3.4 covers LDA/QDA heads; the direct LDA check returns TC gaps below the solver's numerical resolution, consistent with Theorem 2.2. NCM heads with $A_r = I$ and softmax cross-entropy lie outside its formal scope (Remark D.4); E_{TC} is an upper bound several orders of magnitude looser than the empirical residual, so the decomposition gives directional guidance only. The drift bound's L_h^2 prefactor treats self-attention as Lipschitz on the observed parameter region, which Kim et al. [2021] cautions is not global (§U); method-by-method relations and further limitations in §L.

Conclusion. The Infeasibility Theorem Khademi Nori and Kim [2024] sets a structural ceiling on discriminative class-IL that the empirical success of pretrained-backbone methods appears to violate; pretraining exponentially attenuates it. The Mahalanobis margin $\gamma_{\min}(\phi)$, a computable scalar quantifying the room a pretrained backbone gives a class-IL learner, bounds the TC gap by $\exp(-\gamma_{\min}/8)$ at the operational rate of Corollary 3.5. Combined with the convergence guarantee and the drift bound, this yields a diagnostic: measure $\hat{\gamma}_{\min}$ to predict E_{TC}, monitor gradient tail behaviour to predict E_{CF}, and freeze the backbone before $t^\star$. Every quantity in the diagnostic is computable before any incremental training begins, so backbone selection and deployment planning need not wait on fine-tuning experience. The companion analysis of the within-task gauge bias [Khademi Nori and Wang, 2026] complements this picture from the head side, characterising and removing at evaluation time the per-task gauge offset that within-task softmax training leaves on cross-task discriminators.

Broader impacts

$\widehat{\gamma}_{\min}$ is computable in a single forward pass, letting practitioners test whether a frozen backbone already makes fine-tuning unnecessary. The diagnostic inherits pretraining-corpus and class-statistic biases, so backbone rankings can perpetuate representation biases against under-represented classes and deployments on shifted distributions should re-estimate it. We identify no misuse pathway beyond standard supervised classification.

Acknowledgments and Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through grants 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

Artificial intelligence systems that learn new image categories one batch at a time were long thought doomed to blur old categories into new ones. A mathematical result proved that no such system can fully match one trained on all the data together. We show an escape route. When a model has first been trained on a huge and varied collection of images, most of the problem fades. We prove the leftover confusion can shrink extremely fast as the separation between task groups grows, and give one simple number for what remains. In the one setting our theory covers, tests confirmed the prediction: confusion fell below measurable levels. On practical models it shrank about ten times more slowly than predicted.

References

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

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

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Albin Soutif Cormerais, Marc Masana, Joost Van de Weijer, and Bartlomiej Twardowski. On the importance of cross-task features for class-incremental learning. In *Proc. of International Conference on Machine Learning (ICML) Workshops*, pages 3611–3620, 2021.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 220–221, 2020.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

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

Milad Khademi Nori and Guanghui Wang. The structural bias of ℓ_2-regularized cross-entropy heads in class-incremental learning: Characterization, limits, and a gauge-anchoring fix. *Advances in Neural Information Processing Systems*, 2026. To appear.

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.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571, 2021.

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

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

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Kanti V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57 (3):519–530, 1970.

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.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Zixuan Ni, Haizhou Shi, Siliang Tang, Longhui Wei, Qi Tian, and Yueting Zhuang. Revisiting catastrophic forgetting in class incremental learning. *arXiv:2107.12308*, 2021.

Milad Khademi Nori, Il-Min Kim, et al. Federated class-incremental learning: A hybrid approach using latent exemplars and data-free techniques to address local and global forgetting. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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.

Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. Incremental linear discriminant analysis for classification of data streams. *IEEE Trans. on Systems, Man, and Cybernetics (TSMC)*, 35(5): 905–914, 2005.

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.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *Proc. International Conference on Machine Learning (ICML)*, 2023.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS) Workshop*, pages 2994–3003, 2017.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 5827–5837, 2019.

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.

Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. MOS: Model surgery for pre-trained model-based class-incremental learning. In *AAAI Conference on Artificial Intelligence*, 2025.

Rui Sun, Haoran Duan, Jiahua Dong, Varun Ojha, Tejal Shah, and Rajiv Ranjan. Rehearsal-free federated domain-incremental learning. *arXiv:2405.13900*, 2024.

Gido M. van de Ven, Zhe Li, and Andreas S. Tolias. Class-incremental learning with generative classifiers. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3611–3620, 2021.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022a.

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, 2022b.

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, 2022c.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 374–382, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proc. International Conference on Machine Learning (ICML)*, pages 3987–3995, 2017.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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.

A Heavy-Tailed Streaming Model (Full)

The Relaxed Infeasibility Theorem bounds the TC component of excess risk in terms of the pretrained representation. The optimisation error of CF-optimal training is governed by how the head is fitted on finite data. Real class-IL streams exhibit heavy tails: gradient noise in transformer training is well-modelled by α-stable distributions with index $\alpha \in (1, 2)$ Şimşekli et al. [2019], Zhang et al. [2020], violating the standard finite-variance assumption. We state the corresponding model.

Assumption A.1 (Heavy-tailed loss and stochastic gradients)**.** There exist $\alpha \in (1, 2]$, $V > 0$, and $\sigma > 0$ such that for all $\theta \in \Theta$ and all stages t:

$$\mathbb{E}\big[\, |v(f_\theta(X), Y) - \mathbb{E}\, v(f_\theta(X), Y)|^\alpha \,\big|\, Y \in \mathcal{Y}^{(t)}\big] \le V, \tag{10}$$

and for the intra-task population loss $L_t(\theta) := |P_{tt}(\theta)|$ and stochastic gradient $\widehat{\nabla L_t}(\theta)$ from a single sample,

$$\mathbb{E}\big[\, \Big\|\widehat{\nabla L_t}(\theta) - \nabla L_t(\theta)\Big\|^\alpha \,\big|\, Y \in \mathcal{Y}^{(t)}\big] \le \sigma^\alpha. \tag{11}$$

The parameter α is the *tail index*: smaller α corresponds to heavier tails. At $\alpha = 2$ this reduces to finite variance; for $\alpha \in (1, 2)$ the variance may be infinite, requiring truncation- or clipping-based estimators for meaningful concentration Catoni [2012]. Under these assumptions a Catoni-style truncated-mean argument gives a finite-sample concentration inequality for the empirical risk on each task (Lemma P.1, stated in §P with the full proof in §F.1), with deviation $O\big((\log(2/\delta))^{1-1/\alpha}/n_t^{1-1/\alpha}\big)$. Throughout §4 we work with population-level task incompatibility (Assumption N.3); the empirical-vs-population gap is bounded by this concentration and is dominated by the optimisation error analysed next.

B Representation Drift Under Backbone Fine-Tuning (Full)

§§3–4 treated ϕ_0 as frozen. Modern class-IL methods sometimes fine-tune the backbone alongside the head Zhou et al. [2024a]; this can lower E_{CF} by improving intra-task fit but may *shrink* $\gamma_{\min}(\phi_t)$, raising E_{TC}. We quantify the trade-off via a representation-drift bound for Lipschitz backbones (Theorem B.2) and identify a *critical horizon* $t^\star$ beyond which fine-tuning damages inter-task separation faster than it improves intra-task fit.

Assumption B.1 (Lipschitz backbone)**.** The feature map $h_\phi(x)$ is L_h-Lipschitz in ϕ on a bounded parameter region Φ in the Frobenius norm:

$$\|h_\phi(x) - h_{\phi'}(x)\| \;\le\; L_h \,\|\phi - \phi'\|_F\,. \tag{12}$$

Transformer backbones with bounded weights and activations satisfy Assumption B.1, with L_h the operator-norm product of attention and MLP layers, bounded under spectral or layer normalisation (§T discusses estimation; see Kim et al. [2021]).

Theorem B.2 (Representation drift, nonlinear case)**.** *Under Assumptions 3.2, A.1, and B.1, after t incremental stages of clipped-SGD fine-tuning with learning rate η_ϕ, clipping threshold τ_ϕ, and K_s iterations on stage s,*

$$\gamma_{\min}(\phi_t) \;\geq\; \gamma_{\min}(\phi_0) \;-\; c_2\, L_h^2\, \eta_\phi\, \tau_\phi \cdot \frac{\left\|\mu^{(i^\star)}(\phi_0) - \mu^{(j^\star)}(\phi_0)\right\|}{\sigma^2_{\min}} \cdot \sum_{s=1}^{t} K_s, \tag{13}$$

where $(i^\star, j^\star) := \arg\min_{i \neq j} \gamma_{ij}(\phi_0)$ and $c_2 > 0$ is a universal constant.

The proof (§H.1) proceeds in three steps: bound parameter drift by cumulative clipped-gradient budget, translate parameter drift into feature drift via Assumption B.1, translate feature drift into Mahalanobis-margin drift via perturbation analysis. The bound is stated with L_h^2 in the prefactor: the proof produces L_h at first order plus a quadratic-in-L_h second-order contribution; the conservative form bounds both regimes (§U).

Corollary B.3 (Critical horizon for fine-tuning)**.** *Under the conditions of Theorem B.2 with uniform $K_s = K$, the critical horizon $t^\star$ at which $\gamma_{\min}(\phi_{t^\star}) = \gamma_{\min}(\phi_0)/2$ satisfies*

$$t^\star \;=\; \Omega\!\left(\frac{\gamma_{\min}(\phi_0)\, \sigma^2_{\min}}{L_h^2\, \eta_\phi\, \tau_\phi\, K\, \left\|\mu^{(i^\star)}(\phi_0) - \mu^{(j^\star)}(\phi_0)\right\|} \right). \tag{14}$$

Beyond $t^\star$, the operational rate $\exp(-\gamma_{\min}(\phi_t)/8)$ of Corollary 3.5 has inflated by more than $3\times$ relative to the frozen-backbone value (whenever $\gamma_{\min}(\phi_0) \geq 16 \ln 3 \approx 17.6$) while E_{CF} improvements typically saturate.

Reading the formula. $t^\star$ is large when (i) the worst-case margin $\gamma_{\min}(\phi_0)$ is large (a strong backbone has room to drift before separation collapses), (ii) within-task scatter $\sigma^2_{\min}$ is large, (iii) the per-stage budget $\eta_\phi \tau_\phi K$ is small, and (iv) the feature map is smooth, with the Lipschitz constant L_h entering as L_h^2. The conditions formalise three empirical prescriptions: *use a strong pretrained backbone*, *use a small backbone learning rate*, *stop early*.

B.1 Representation drift and the critical horizon

The two blocks above hold the backbone fixed; this block releases it and measures the cost to the margin $\widehat{\gamma}_{\min}(\phi_t)$.

Test of Theorem B.2. On Split-CIFAR-100 we fine-tune the backbone with the head at $\eta_\phi \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ and measure $\widehat{\gamma}_{\min}(\phi_t)$ after each task. Per-stage slopes are -0.16 and -1.73 at $\eta_\phi = 10^{-5}$ and 10^{-4}, consistent with the linear-in-η_ϕ prediction (other mechanisms can produce the same scaling at small η_ϕ; the data are consistent with but do not uniquely identify the bound). At $\eta_\phi = 10^{-3}$ no unsaturated per-stage slope is identifiable, and the primary quantity is the time-to-floor, 0.48 ± 0.04 stages; the derived floor-saturated secants follow the identity $-\widehat{\gamma}_{\min}(\phi_0)/w$ at $\widehat{\gamma}_{\min}(\phi_0) = 8.20 \pm 0.62$ (harness estimate; the one-pass estimate of the same baseline is 8.3 ± 0.4, and the two runs are compared in §K.5), giving -8.20 over $w = 1$ stage and -4.10 over $w = 2$ stages, quoted without a $\pm$ (the dispersion lives on the baseline). The empirical critical horizon $\widehat{t^\star}$ at which $\widehat{\gamma}_{\min}(\phi_t) < \widehat{\gamma}_{\min}(\phi_0)/2$ is 14.2 ± 1.1, 2.35 ± 0.20, 0.21 ± 0.03 stages respectively (Table 7, §K.5). The margin collapse at $\eta_\phi = 10^{-3}$ therefore completes within the first stage. The accuracy crossover where the fine-tuned backbone falls below frozen is 3.18 stages at $\eta_\phi = 10^{-4}$ (Figure 7), lagging the geometric $\widehat{t^\star} = 2.35$: the half-margin threshold is a conservative warning indicator exploited by Algorithm 1.

C Notation Reference

Convention notes. (i) The τ family is distinct: τ is the truncation threshold (§F.1); τ_t is the head clipping threshold; τ_ϕ is the backbone clipping threshold. (ii) The μ family is distinct: μ_r is a class mean; $\mu^{(t)}$ is a task centroid; μ is the strong-convexity constant in §4 and the base measure on $\mathcal{X}$ in the population-risk integral of §2; μ_t is the PŁ constant. (iii) Single-letter collisions: B (feature-norm bound) vs. B_{kl} (Bhattacharyya distance, §O); p is the joint data density in §2 and the feature dimension in $\mathbb{R}^p$ from §3 on; ψ is the head parameter throughout, while $\psi_\tau(z)$ is the scalar truncation operator of §F.1; the sub-exponential proxy parameter c (§E.1) is distinct from the class count C and from the schedule constants $c(\alpha)$, $c'(\alpha)$ of Proposition 4.3; the class-to-task scatter ratio $\kappa^{\mathrm{cls}}_{ij}$ (Definition E.9) carries a task-pair subscript, and $\kappa^{\mathrm{cls}}_{\max}$ is its worst case over pairs (Theorem 3.4).

This appendix collects the symbols used throughout the paper, organised by the section in which they first appear. Page references point to the first definitional use.

Sets, spaces, and sample structure

Symbol	Type	Meaning
$\mathcal{X}$	space	Input space (image space).
$\mathcal{Y} = \{1, \ldots, N\}$	set	Set of all classes; $\lvert\mathcal{Y}\rvert = N$.
$\mathcal{Z} \subseteq \mathbb{R}^p$	space	Feature space; p is feature dimension.
T	integer	Total number of incremental tasks.
C	integer	Number of classes per task; $N = T \cdot C$.
$\mathcal{Y}^{(t)} \subset \mathcal{Y}$	set	Class set of task t; $\lvert\mathcal{Y}^{(t)}\rvert = C$.
D_t	sample set	Training stream for task t.
n_t	integer	Sample size of task t.
X, Y	random vars	Input–label pair, joint $p(x, y)$.

Models and parameters

Symbol	Type	Meaning
$\phi \in \Phi$	parameters	Backbone parameters.
ϕ_0	parameters	Frozen pretrained backbone.
ϕ_t	parameters	Backbone after stage t of fine-tuning.
$\psi \in \Psi$	parameters	Head parameters.
$\theta = (\phi, \psi)$	parameters	Full model parameters.
$\widehat{\psi}_T$	parameters	Head produced by sequential clipped SGD after all T stages.
$\psi^\dagger$	parameters	CF-optimal head: $\arg\min_\psi \sum_t \lvert P_{tt}(\psi)\rvert$.
$\psi^\star$	parameters	Globally optimal head: $\arg\min_\psi \sum_{i,j} \lvert P_{ij}(\psi)\rvert$.
$h_\phi : \mathcal{X} \to \mathcal{Z}$	map	Backbone (feature extractor).
$f_\theta(x) \in \mathbb{R}^N$	vector	Full model output (logits).

Loss matrix and risk decomposition (§2)

Symbol	Type	Meaning
$v(\hat{y}, y)$	loss	Bounded per-sample loss; $0 \le v \le M$.
M	constant	Bound on per-sample loss.
$\rho_{kl}(\theta)$	scalar	Pairwise loss-matrix entry between classes k, l; Eq. (3).
$P(\theta) \in \mathbb{R}^{N \times N}$	matrix	Loss matrix; $P_{kl} = \rho_{kl}$.
$P_{ij}(\theta)$	block	Inter-task block ($i, j \in \{1, \ldots, T\}$).
$P_{tt}(\theta)$	block	Intra-task block of task t.
$\lvert P_{ij}(\theta)\rvert$	scalar	Sum of magnitudes of entries in block P_{ij}.
$\mathrm{TC\ gap}(\phi)$	scalar	Excess risk of CF-optimal head vs. globally optimal head.

Mahalanobis margin (§3)

Symbol	Type	Meaning
$\mu_r(\phi) \in \mathbb{R}^p$	vector	Class-conditional feature mean for class r.
$\mu^{(t)}(\phi) \in \mathbb{R}^p$	vector	Task-t centroid: $\frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r(\phi)$.
$\Sigma_r(\phi)$	matrix	Class-conditional feature covariance for class r.
$\Sigma^{(t)}(\phi)$	matrix	Task-t covariance (average of per-class covariances): $\frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \Sigma_r(\phi)$.
$\sigma^2_{\min}$	constant	Lower bound on $\lambda_{\min}(\Sigma^{(t)})$ across tasks (Assumption 3.2).
B	constant	Bound on feature norm: $\|h_\phi(x)\| \le B$ (Assumption 3.2).
$\gamma_{ij}(\phi)$	scalar	Inter-task Mahalanobis margin (Definition 3.1).
$\gamma_{\min}(\phi)$	scalar	Worst-case margin: $\min_{i\neq j}\gamma_{ij}(\phi)$.
$\Gamma_r(\phi_0)$	matrix	Sub-Gaussian proxy covariance of class r (Assumption 3.2(iv)).
Σ_{pool}	matrix	Pooled within-class covariance; the LDA head uses $A_r \to \Sigma^{-1}_{\text{pool}}$ (§3).
D^M_{kl}	scalar	Per-class-pair Mahalanobis distance.
ν^M_{ij}	scalar	Within-task spread ratio in the task-pair metric M_{ij} (Definition E.10, §E.4). Note: distinct from ρ_{kl} above (loss-matrix entry); context disambiguates.
κ^{cls}_{ij}	scalar	Class-to-task scatter ratio (Definition E.9, §E.4); $\kappa^{\text{cls}}_{\max} := \max_{i\neq j}\kappa^{\text{cls}}_{ij}$.
a_{ij}	scalar	Interpolation parameter of the exponent $\frac{1-a_{ij}}{8\,\kappa^{\text{cls}}_{ij}}\gamma_{ij}(\phi_0)$, $a_{ij} = \nu^M_{ij}$ (Corollary 3.5).
$\bar{\nu}$	scalar	Configuration-averaged within-task spread ratio; the printed value is the isotropic-metric average (§5, Table 4).
$\widehat{\gamma}_{ij}, \widehat{\gamma}_{\min}$	scalars	Plug-in estimates of the margins from empirical class moments (§I).

Cross-entropy and pairwise losses (§O)

Symbol	Type	Meaning
v_{CE}	loss	Softmax cross-entropy loss.
$v^{(y,y')}_{\text{pair}}$	loss	Pairwise binary log-loss between classes y, y'.
$V_{\text{pair}}(f_\theta(x), y)$	scalar	Sum of $v^{(y,y')}_{\text{pair}}$ over $y' \neq y$.
$I_{\text{CE}}(\theta)$	scalar	Population cross-entropy risk.
$I_{\text{pair}}(\theta)$	scalar	Population pairwise risk: $\sum_{k<l}\rho_{kl}(\theta)$.
P^{kl}_e	scalar	Bayes-optimal pairwise error between classes k, l (§E.3).
$B_{kl}, B^{\text{Gauss}}_{kl}$	scalars	Bhattacharyya distance between class densities; Gaussian-feature specialisation (§E.3).

Heavy-tailed streaming model (§4)

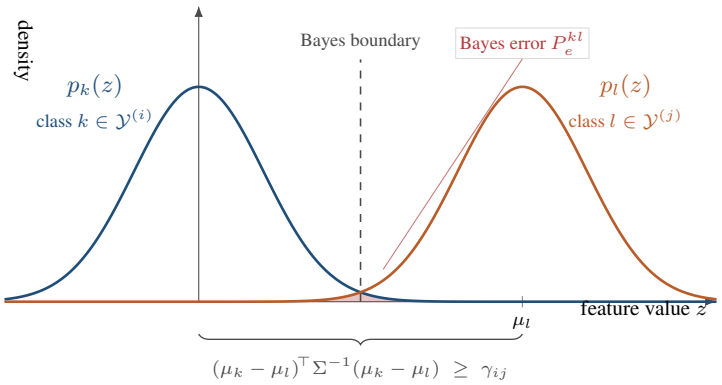
Symbol	Type	Meaning
$\alpha \in (1, 2]$	constant	Tail index of per-sample losses or gradients.
V	constant	Bound on $\mathbb{E}\,\|Z\|^\alpha$ (Assumption A.1).
σ	constant	Bound on $(\mathbb{E}\,\|\xi\|^\alpha)^{1/\alpha}$ for gradient noise (Assumption A.1).
$\widehat{I}_t(\theta)$	estimator	Empirical risk on task t.
δ	constant	Failure probability for high-probability bounds.
$\widetilde{I}_\tau$	estimator	Truncated-mean estimator with threshold τ.
$\psi_\tau(\cdot)$	operator	Scalar truncation operator (§F.1).

Convergence analysis (§4)

Symbol	Type	Meaning
$L_t(\psi)$	function	Population intra-task loss for task t: $\lvert P_{tt}(\psi;\phi_0)\rvert$.
β	constant	Smoothness constant of L_t (Assumption 4.1).
μ	constant	Strong-convexity constant of L_t (Assumption 4.2).
μ_t	constant	Polyak-Łojasiewicz constant for L_t (Corollary G.2).
Δ_t	scalar	Initial suboptimality of task t: $L_t(\psi_{t,0}) - \min L_t$.
K_t	integer	Per-stage iteration budget for task t.
K	integer	Uniform iteration budget across stages (alias of K_t when all stages share one budget).
η_t	scalar	Per-stage learning rate.
τ_t	scalar	Per-stage gradient-clipping threshold.
$g_k, \bar{g}_k$	vectors	Stochastic gradient and clipped gradient.
ξ_k	vector	Gradient noise: $g_k - \nabla L_t(\psi_k)$.
E_{CF}	scalar	Optimisation (catastrophic-forgetting) term in the decomposition (Theorem 4.5).
E_{TC}	scalar	Representation (task-confusion) term in the decomposition.
$C^{\mathrm{sc}}_{\mathrm{opt}}, C^{\mathrm{nc}}_{\mathrm{opt}}$	constants	Strongly-convex rate: $C^{\mathrm{sc}}_{\mathrm{opt}}$ depends on α and the condition ratio β/μ; non-convex rate: $C^{\mathrm{nc}}_{\mathrm{opt}}$ depends only on α.
ε	constant	Gradient-norm tolerance: $\mathbb{E}\,\lVert\nabla L_t\rVert^2 \le \varepsilon$ (Lemma G.4).
r_t	scalar	Residual in the gradient-to-gap conversion; $r_t = 0$ under the PŁ inequality (Lemma G.4).

Representation drift (§4)

Symbol	Type	Meaning
L_h	constant	Lipschitz constant of h_ϕ in ϕ (Assumption B.1).
η_ϕ, τ_ϕ	scalars	Backbone fine-tuning learning rate and clipping threshold.
$t^\star$	integer	Critical horizon: largest t with $\gamma_{\min}(\phi_t) \ge \gamma_{\min}(\phi_0)/2$ (Corollary B.3).
$(i^\star, j^\star)$	indices	Closest task pair under ϕ_0: $\arg\min_{i\neq j} \gamma_{ij}(\phi_0)$.
$\lVert\cdot\rVert_F$	norm	Frobenius norm.
$\lVert\cdot\rVert_{\mathrm{op}}$	norm	Operator (spectral) norm.
c_2, c_3, c_4	constants	Universal constants in the drift bound.
S^M_{ij}	scalar	Pair-metric within-task spread sum: $\max_{k\in\mathcal{Y}^{(i)}} \lVert\mu_k - \mu^{(i)}\rVert_{M_{ij}} + \max_{l\in\mathcal{Y}^{(j)}} \lVert\mu_l - \mu^{(j)}\rVert_{M_{ij}}$ (Definition E.10, §E.4).
M_{ij}	matrix	Task-pair metric $(\Sigma^{(i)} + \Sigma^{(j)})/2$, the metric of Definition 3.1 (§E.4).
$\Delta\mu$	vector	Centroid displacement $\mu^{(i)}(\phi_0) - \mu^{(j)}(\phi_0)$ (drift proofs, §H.1).
$\Sigma^{(t)}_{\mathrm{between}}$	matrix	Between-class scatter of task t (§E.4).
$\gamma^{\mathrm{sum}}_{ij}, \gamma^{\mathrm{avg}}_{ij}$	scalars	Margin under the summed / averaged covariance convention: $\gamma^{\mathrm{sum}}_{ij} = \frac{1}{2}\gamma^{\mathrm{avg}}_{ij} = \frac{1}{2}\gamma_{ij}$ (34).



Bhattacharyya inequality (Duda–Hart–Stork, 2.6.2): $P_e^{kl} \leq \exp\Big(-\frac{1}{8}\,(\mu_k - \mu_l)^\top \big[\frac{\Sigma_k + \Sigma_l}{2}\big]^{-1} (\mu_k - \mu_l)\Big) \;\leq\; e^{-\gamma_{ij}/8}$.

The constant $\frac{1}{8}$ is the saddle-point of the Gaussian Bayes-error integral; the displayed inequalities are the negligible-spread homoscedastic limit $\nu_{ij}^M = 0$, $\kappa_{ij}^{\text{cls}} = 1$. Summing over the C^2 class pairs gives, in this limit, Lemma 3.3, whose general exponent is $\frac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\text{cls}}}$: $|P_{ij}(\psi;\phi_0)| \leq \frac{MC^2}{N-1}\, e^{-\gamma_{ij}/8}$.

Figure 4: Bhattacharyya geometry in feature space. Two task-conditional Gaussians (means $\mu^{(i)}, \mu^{(j)}$, covariances $\Sigma^{(i)}, \Sigma^{(j)}$) in $\mathbb{R}^p$ are separated by the Mahalanobis distance $\gamma_{ij} = (\mu^{(i)} - \mu^{(j)})^\top \big(\frac{\Sigma^{(i)}+\Sigma^{(j)}}{2}\big)^{-1} (\mu^{(i)} - \mu^{(j)})$. The Bhattacharyya inequality bounds the misclassification probability between any class pair (k, l) across the two tasks by $\exp(-\frac{1}{8}\,\gamma_{ij})$; summing over the C^2 class pairs in the off-diagonal block P_{ij} yields Lemma 3.3.

Experiments-only symbols (§5)

Symbol	Type	Meaning
$\widehat{m}, \widehat{m}_{\text{pool}}$	scalars	OLS slope of $\log(\text{TC gap})$ vs. $\widehat{\gamma}_{\min}$; pooled common-slope estimate (§5).
A_T, A_T^{frozen}	scalars	Final test accuracy after T stages, fine-tuned and frozen-backbone (§K.5, §K.6).
w	integer	Secant window, in stages, for floor-saturated slope estimates (§5).
$\lambda^\star, S$	scalar, matrix	Ledoit-Wolf shrinkage optimum; pooled within-class scatter it shrinks (§I).
c	constant	Sub-exponential proxy parameter in the conservative-bound analysis (§E.1); distinct from the class count C.

Operators and conventions

Symbol	Type	Meaning
$\mathbb{E}[\cdot]$	operator	Expectation under $p(x, y)$.
$\mathbb{P}[\cdot]$	operator	Probability under $p(x, y)$.
$\|v\|$	norm	Euclidean norm of vector v.
$\|v\|_{M_{ij}}$	norm	$\sqrt{v^\top M_{ij}^{-1} v}$, the task-pair Mahalanobis norm (Definition E.10, §E.4).
$A \succeq B$	ordering	Loewner positive-semidefinite ordering.
$\text{clip}_\tau(g)$	operator	Vector clipping: $g \cdot \min\{1, \tau / \|g\|\}$.
$\widehat{\cdot}$	decoration	Empirical/estimated quantity.

D Extended Material for the Margin Bound (Section 3)

This section collects the extended material for Section 3: the scope assumption and its method coverage, the motivation for the spread condition, the homoscedastic specialisation, the cross-entropy extension, and the exponent provenance.

Assumption D.1 (Prototype-based head)**.** The head g_ψ takes the form $g_\psi(z)_r = -(z-\mu_r)^\top A_r(z-\mu_r) + b_r$ where the metric satisfies (i) $A_r \to \Sigma_r^{-1}$ (QDA case) or (ii) $A_r \to \Sigma_{\text{pool}}^{-1}$ for all r (LDA case) in the population limit. The class means μ_r and the metric A_r are estimated from training features; bias b_r is a class-prior adjustment.

Remark D.2 (Methods within and outside the assumption)*.* The bound applies to test-time decision rules of prototype form regardless of training procedure. Generatively trained prototype methods Hayes and Kanan [2020], Goswami et al. [2023], McDonnell et al. [2023] achieve zero task-confusion excess risk ($E_{\text{TC}} = 0$; the term is formalised in §4) by Theorem 2.2 and are recovered as the $\gamma_{\min} \to \infty$ limit of Theorem 3.4; the in-scope direct LDA check verifies this limit empirically (§5). Discriminatively trained heads that converge to prototype form Zhou et al. [2024b] realise the non-trivial $\exp(-\gamma_{\min}/8)$ regime. Outside the formal scope: NCM with $A_r = I$, softmax cross-entropy heads with generic linear classifiers, and MOS Sun et al. [2025].

Condition (L) (Zero-one-type loss)**.** The per-pair loss v satisfies $0 \le v \le M$ and $v(x) = 0$ whenever the head classifies x correctly; equivalently $v \le M \cdot \mathbf{1}\{\text{misclassification}\}$ for the decision rule induced by the head.

Zero-one and bounded margin-type losses satisfy Condition (L); softmax cross-entropy does not, being strictly positive on correct classifications, a second, independent reason those experiments sit outside the formal scope (Remark D.4; the cross-entropy route is the pairwise-loss sandwich of Lemma O.1, §O).

The within-task spread can be bounded without passing to the isotropic metric, provided class covariances do not sit too far above their task covariance, which is what κ_{ij}^{cls} measures. The isotropic detour fails on real anisotropic features (condition numbers in the thousands): the isotropic metric sees a few percent of the Mahalanobis separation, so its mismatch term sits near γ_{ij} itself. We therefore work directly in the task-pair metric, paying only κ_{ij}^{cls}.

Condition (S') (Bounded within-task spread ratio, pair metric)**.** Work in the task-pair metric $M_{ij} := (\Sigma^{(i)} + \Sigma^{(j)})/2$ (the metric of Definition 3.1) and define, for each ordered task pair $i \ne j$,

$$\nu_{ij}^M \;:=\; \frac{2\,S_{ij}^M}{\sqrt{\gamma_{ij}(\phi_0)}}, \qquad S_{ij}^M := \max_{k\in\mathcal{Y}^{(i)}} \left\|\mu_k - \mu^{(i)}\right\|_{M_{ij}} + \max_{l\in\mathcal{Y}^{(j)}} \left\|\mu_l - \mu^{(j)}\right\|_{M_{ij}},$$

so ν_{ij}^M is the ratio of within-task class-centroid spread to between-task centroid separation in that metric, and let κ_{ij}^{cls} be the class-to-task scatter ratio of Definition E.9: the least κ with $\Sigma_k \preceq \kappa\,\Sigma^{(t(k))}$ for every class k of the two tasks (so $1 \le \kappa_{ij}^{\text{cls}} \le C$). The assumption is

$$\nu_{ij}^M \;\le\; \tfrac{1}{2} \qquad \text{for every ordered task pair } i \ne j.$$

Definitions, the lemma, sufficient trace criteria and empirical verification are in §E.4.

In words, within-task class-centroid spread must be small relative to between-task separation, both measured in the task-pair metric M_{ij} itself. The triangle inequality of Lemma E.11 holds in the Mahalanobis norm of M_{ij} directly, so no metric-mismatch term arises; the only geometric quantity the bound carries beyond ν_{ij}^M is the class-to-task Loewner ratio κ_{ij}^{cls}, and under the condition Lemma E.11 gives every cross-task class pair of (i,j) the distance lower bound $D_{kl}^M \ge \gamma_{ij}(\phi_0)\,(1-\nu_{ij}^M)/\kappa_{ij}^{\text{cls}} \ge \gamma_{ij}(\phi_0)/(2\,\kappa_{ij}^{\text{cls}})$.

Corollary D.3 (Homoscedastic classes)**.** *If, in addition to the hypotheses of Theorem 3.4, the classes within each task are homoscedastic, $\Sigma_k = \Sigma^{(t(k))}$ for every class k of task $t(k)$, so that $\kappa_{ij}^{\text{cls}} = 1$ for every pair $i \ne j$ by Definition E.9 and $\kappa_{\max}^{\text{cls}}(\phi_0) = 1$, then the exponent of* (6) *is $\frac{1}{16}$:*

$$\text{TC gap}(\phi_0) \;\le\; \frac{MC^2\,T(T-1)}{N-1}\,\exp\!\Big(-\tfrac{1}{16}\,\gamma_{\min}(\phi_0)\Big).$$

Remark D.4 (Status of softmax cross-entropy heads). Softmax cross-entropy heads are not prototype-based and lie outside the formal scope of Theorem 3.4. We provide a pointwise pairwise-loss sandwich ($v_{\mathrm{CE}} \le V_{\mathrm{pair}}$, Lemma O.1 in §O), but this does not transfer to TC *gaps* (differences between two parameter values). The empirical slope of $\log(\mathrm{TC\ gap})$ versus $\widehat{\gamma}_{\min}$ on softmax CE heads in §5 confirms the log-linear decay at a rate about one tenth of the operational prediction; these slope experiments sit outside the formal scope and a formal extension remains open (§6). The closed-form mechanism for this gap is developed in our companion work [Khademi Nori and Wang, 2026].

Interpretation. Theorem 3.4 interpolates: at $\gamma_{\min}(\phi_0) = 0$ the bound is vacuous, recovering Theorem 2.1 qualitatively (training from scratch); as $\gamma_{\min} \to \infty$ the TC gap vanishes (strong pretraining). For fixed $\gamma_{\min}(\phi_0) > 0$ the bound is exponentially small: doubling the margin multiplies the TC gap by $e^{-\gamma/(16\,\kappa^{\mathrm{cls}}_{\max})}$ ($e^{-\gamma/16}$ under homoscedastic classes, Corollary D.3; $e^{-\gamma/8}$ in the negligible-spread homoscedastic regime of Corollary 3.5). The constants' provenance ($1/8$ from Chernoff–Bhattacharyya, $1/(16\,\kappa^{\mathrm{cls}}_{\max})$ under Condition (S'), $1/16$ homoscedastic), the non-Gaussian reading of $\gamma_{\min}/8$, and compatibility with the Infeasibility Theorem are documented in §E.3, which tabulates the four exponent constants, the hypothesis sets that buy each, and where each is stated.

E Proofs from Section 3

E.1 Proof of Lemma 3.3 (Off-Diagonal Block Bound)

We restate the lemma for convenience.

Lemma (Lemma 3.3 restated). Let v satisfy Condition (L): the loss is bounded, $0 \le v \le M$, and vanishes on correctly classified points, so $v \le M \cdot \mathbf{1}\{\text{misclassification}\}$ for the head's decision rule. Under Assumptions 3.2 (clause (iv) in its Gaussian branch; Remark E.3) and D.1, with $\phi = \phi_0$ frozen, for the prototype-based head ψ specified by Assumption D.1 and any pair of tasks $i \ne j$:

$$|P_{ij}(\psi;\phi_0)| \;\le\; \frac{MC^2}{N-1}\,\exp\Big(-\tfrac{1-\nu^M_{ij}}{8\,\kappa^{\mathrm{cls}}_{ij}}\,\gamma_{ij}(\phi_0)\Big), \tag{15}$$

where ν^M_{ij} is the within-task spread ratio in the task-pair metric $M_{ij} = (\Sigma^{(i)} + \Sigma^{(j)})/2$ (Definition E.10; §E.4) and $\kappa^{\mathrm{cls}}_{ij}$ is the class-to-task scatter ratio (Definition E.9; §E.4); Step 3 below, in the form of Lemma E.11, supplies $D^M_{kl} \ge \gamma_{ij}(\phi_0)\,(1-\nu^M_{ij})/\kappa^{\mathrm{cls}}_{ij}$. The case $\nu^M_{ij} = 0$ with homoscedastic classes ($\kappa^{\mathrm{cls}}_{ij} = 1$) is negligible spread, and the unqualified $\frac{1}{8}$ is that limit, the operational rate of Corollary 3.5; under Condition (S') ($\nu^M_{ij} \le 1/2$) the guaranteed exponent is $1/(16\,\kappa^{\mathrm{cls}}_{ij})$ (Corollary E.12), the form aggregated by Theorem 3.4.

The proof proceeds in four steps: (i) reduce each entry of the inter-task block to the prototype-based head's misclassification probability, and identify this with the Bayes-optimal pairwise error under Assumptions 3.2 and D.1; (ii) bound the Bayes-optimal pairwise error via the Bhattacharyya inequality; (iii) lower-bound the per-class-pair Mahalanobis distance by the inter-task Mahalanobis margin γ_{ij} via a triangle-style inequality; (iv) sum the resulting bound over the C^2 entries of the block.

Step 1: Pairwise loss to Bayes-optimal misclassification probability

Fix tasks $i \ne j$ and a class pair (k,l) with $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$. By Definition (3),

$$\rho_{kl}(\psi;\phi_0) \;=\; \frac{1}{N-1}\int_{\mathcal{X}\times\{k,l\}} v\big(f_\theta(x),y\big)\,p(x,y)\,d\mu(x)\,dy. \tag{16}$$

By Condition (L), $0 \le v \le M$ and v vanishes on correctly classified points, so $v \le M \cdot \mathbf{1}\{\text{misclassify}\}$ for the decision rule induced by ψ; integrating this bound against $p(x,y)$ in equation (16) gives

$$\rho_{kl}(\psi;\phi_0) \;\le\; \frac{M}{N-1}\,\mathbb{P}\big[\text{misclassify } X \text{ between classes } k \text{ and } l \,\big|\, Y \in \{k,l\}\big]. \tag{17}$$

Equation (17) holds for any decision rule induced by ψ. Zero-one and bounded margin-type losses satisfy Condition (L); softmax cross-entropy does not, being strictly positive on correct classifications, and is routed through the pairwise-loss sandwich of Lemma O.1 (Remark D.4).

Bayes-bridging under Assumption D.1. By Assumption D.1, ψ is a prototype-based head with class prototypes $\widehat{\mu}_r$ and metrics A_r, where $\widehat{\mu}_r$ is the class-conditional feature mean of class r and is not modified after class r's task completes. The decision rule between classes k and l under this head is

$$\widehat{g}(z) \;=\; \arg\min_{r\in\{k,l\}} (z-\widehat{\mu}_r)^\top A_r (z-\widehat{\mu}_r),$$

which is a linear or quadratic discriminant rule based on prototype distances. In the Gaussian branch of Assumption 3.2(iv) with shared covariance across the two classes ($\Sigma_k = \Sigma_l$; equal determinants suffice once priors are absorbed into the bias b_r), the Bayes-optimal classifier between k and l is exactly the discriminant

$$g^\star(z) \;=\; \arg\min_{r\in\{k,l\}} (z-\mu_r)^\top \Sigma_r^{-1} (z-\mu_r),$$

This is the shared-covariance (LDA) form: the general Bayes QDA rule additionally carries log-determinant terms $\frac{1}{2}\log|\Sigma_r|$ and class priors, which are constant or absorbed into b_r exactly in this branch. This is the same scope in which Corollary 3.5 derives its $1/8$ constant (Gaussian class-conditionals with shared covariance; §3). For Gaussian conditionals with unequal determinants, the Bayes rule is the full QDA including the log-determinant term, and $g^\star$ above remains the population limit of the head only if b_r absorbs that term; for merely sub-Gaussian conditionals the Bayes rule is not a quadratic discriminant in general, and we treat the identification below as an approximation supported by the Gaussianity tests of Appendix I.4, not an exactness granted by clause (iv). In all branches $g^\star$ is the population analogue of $\widehat{g}$ with population prototypes μ_r and metric Σ_r^{-1}. The crucial observation is that under Assumption D.1, $\widehat{\mu}_k$ and $\widehat{\mu}_l$ are estimated from the class-conditional data of classes k and l separately and are not affected by training order across tasks: a prototype head built from class means is invariant to task ordering on the prototypes themselves. By the population-limit metric consistency in Assumption D.1(ii), $A_r \to \Sigma_r^{-1}$ (QDA case) or $A \to \Sigma_{\text{pool}}^{-1}$ (LDA case), combined with prototype consistency $\widehat{\mu}_r \to \mu_r$ from (i), gives $\widehat{g} \to g^\star$ as a classification rule on every class pair, including cross-task pairs $k \in \mathcal{Y}^{(i)}, l \in \mathcal{Y}^{(j)}$.

We therefore have, in the population limit,

$$\mathbb{P}\big[\text{misclassify } X \text{ between classes } k,l \text{ under } \psi \,\big|\, Y\in\{k,l\}\big] \;=\; P_e^{kl}, \tag{18}$$

where P_e^{kl} denotes the Bayes-optimal pairwise error. The identification (18) is exact in the Gaussian shared-covariance branch (where the population limit of the head coincides with the Bayes rule); in the unequal-determinant and sub-Gaussian branches it is the approximation discussed above. The finite-sample gap between $\widehat{g}$ and $g^\star$ is controlled by Lemma P.1 (heavy-tailed concentration of the empirical risk around its mean at a fixed head; stated in §P, proved in §F.1) and contributes to E_{CF} in Theorem 4.5, not to the off-diagonal block bound; we treat (18) as the population-level identification throughout this proof.

Step 2: Bayes-optimal pairwise error via Bhattacharyya

The Bayes-optimal misclassification probability between two classes with conditional feature distributions $p_k(z), p_l(z)$ on the feature space $\mathcal{Z}$ admits the classical Bhattacharyya bound [Duda et al., 2000, Section 2.6.2] (Figure 4 illustrates the geometry):

$$P_e^{kl} \;\le\; \exp\!\big(-B_{kl}\big), \qquad B_{kl} \;:=\; -\log\int_{\mathcal{Z}} \sqrt{p_k(z)\,p_l(z)}\,dz, \tag{19}$$

where B_{kl} is the *Bhattacharyya distance* between p_k and p_l.

When p_k and p_l are Gaussian with means μ_k, μ_l and covariances Σ_k, Σ_l, the Bhattacharyya distance has the closed form

$$B_{kl}^{\text{Gauss}} \;=\; \tfrac{1}{8}\,(\mu_k-\mu_l)^\top\!\left(\frac{\Sigma_k+\Sigma_l}{2}\right)^{-1}\!(\mu_k-\mu_l) \;+\; \tfrac{1}{2}\,\log\frac{|(\Sigma_k+\Sigma_l)/2|}{\sqrt{|\Sigma_k|\,|\Sigma_l|}}. \tag{20}$$

The first term in (20) is one-eighth the squared Mahalanobis distance between the class means with the averaged covariance as metric; we denote it $D_{kl}^M/8$. The second term is non-negative by the AM-GM inequality applied to the determinants [Duda et al., 2000, Eq. 6.7.5]. We therefore have, for Gaussian class-conditional features,

$$P_e^{kl} \;\le\; \exp\!\Big(-\tfrac{1}{8}\,D_{kl}^M\Big). \tag{21}$$

The closed form (21) is used only in the Gaussian branch of clause (iv) of Assumption 3.2, the branch to which Lemma 3.3 and Theorem 3.4 are scoped; the proof establishes it there. For sub-Gaussian class-conditionals the corresponding statement, with D_{kl}^M replaced by a sub-Gaussian Mahalanobis surrogate built from the proxies Γ_r, is an expected extension discussed in Remark E.3 below, not a branch proved here. We treat (21) as our working bound from this point forward, in its Gaussian-branch scope.

Step 3: From per-class to per-task Mahalanobis distance

Step 2 gives a bound depending on the pairwise distance D_{kl}^M between specific classes k and l. We now lower-bound this by the inter-task Mahalanobis margin $\gamma_{ij}(\phi_0)$ defined at the task level.

By construction,

$$\mu^{(i)}(\phi_0) = \frac{1}{C}\sum_{k\in\mathcal{Y}^{(i)}} \mu_k, \qquad \Sigma^{(i)}(\phi_0) = \frac{1}{C}\sum_{k\in\mathcal{Y}^{(i)}} \Sigma_k,$$

and similarly for task j. The task-level Mahalanobis distance is

$$\gamma_{ij}(\phi_0) \;=\; (\mu^{(i)} - \mu^{(j)})^\top \left(\frac{\Sigma^{(i)} + \Sigma^{(j)}}{2}\right)^{-1} (\mu^{(i)} - \mu^{(j)}).$$

We claim the following.

Claim E.1. *For any $k \in \mathcal{Y}^{(i)}$ and $l \in \mathcal{Y}^{(j)}$ with $\nu_{ij}^M \le 1$ (Definition E.10),*

$$D_{kl}^M \;\ge\; \frac{\gamma_{ij}(\phi_0)}{\kappa_{ij}^{\mathrm{cls}}}\left(1 - \nu_{ij}^M\right), \tag{22}$$

with $\kappa_{ij}^{\mathrm{cls}}$ the class-to-task scatter ratio of Definition E.9.

This is *not* a direct triangle inequality on the Mahalanobis metric—the metric uses different covariances on each side—and requires care. The argument has two parts; it is the content of Lemma E.11, stated with Definitions E.9 and E.10 in §E.4.

Part (a): Class-to-task Loewner comparison. Both D_{kl}^M and γ_{ij} are evaluated under different covariance matrices: D_{kl}^M under the class-pair metric $(\Sigma_k + \Sigma_l)/2$, γ_{ij} under the task-pair metric $M_{ij} = (\Sigma^{(i)} + \Sigma^{(j)})/2$ of Definition 3.1. The two are related by the class-to-task scatter ratio of Definition E.9: since $t(k) = i$ and $t(l) = j$,

$$\frac{\Sigma_k + \Sigma_l}{2} \;\preceq\; \kappa_{ij}^{\mathrm{cls}}\,\frac{\Sigma^{(i)} + \Sigma^{(j)}}{2} \;=\; \kappa_{ij}^{\mathrm{cls}}\, M_{ij}, \tag{23}$$

so, since $A \preceq B$ implies $B^{-1} \preceq A^{-1}$ for positive definite A, B, for any vector v,

$$v^\top\big((\Sigma_k + \Sigma_l)/2\big)^{-1} v \;\ge\; \frac{1}{\kappa_{ij}^{\mathrm{cls}}}\, v^\top M_{ij}^{-1} v, \qquad \text{i.e.} \qquad D_{kl}^M \;\ge\; \frac{1}{\kappa_{ij}^{\mathrm{cls}}}\,\|\mu_k - \mu_l\|_{M_{ij}}^2 .$$

A natural alternative compares both D_{kl}^M and γ_{ij} through the isotropic upper bound $\bar{\Sigma} = \lambda_{\max} I$; the price of that route is a metric-mismatch term $\delta_{ij} = \gamma_{ij} - \|\Delta\mu\|_{\bar{\Sigma}^{-1}}^2$ that no hypothesis controls and that on anisotropic features (ViT features have condition numbers in the thousands) sits near γ_{ij} itself, since the isotropic separation never exceeds the Mahalanobis one. The detour is not needed: the comparison above stays in the pair metric, and the only geometry carried forward is the Loewner ratio $\kappa_{ij}^{\mathrm{cls}}$, with $1 \le \kappa_{ij}^{\mathrm{cls}} \le C$ always and $\kappa_{ij}^{\mathrm{cls}} = 1$ under homoscedastic classes (Definition E.9).

Part (b): Centroid geometry in the pair metric. The Mahalanobis norm $\|\cdot\|_{M_{ij}}$ is a norm, so the triangle inequality applies at the task centroids:

$$\|\mu_k - \mu_l\|_{M_{ij}} \;\ge\; \left\|\mu^{(i)} - \mu^{(j)}\right\|_{M_{ij}} \;-\; \left\|\mu_k - \mu^{(i)}\right\|_{M_{ij}} \;-\; \left\|\mu_l - \mu^{(j)}\right\|_{M_{ij}} \;=\; \sqrt{\gamma_{ij}(\phi_0)} - S_{ij}^M,$$

where the equality uses $\left\|\mu^{(i)} - \mu^{(j)}\right\|_{M_{ij}}^2 = \gamma_{ij}(\phi_0)$ from Definition 3.1 and S_{ij}^M is the pair-metric within-task spread sum of Definition E.10. The right-hand side is non-negative whenever

$\nu_{ij}^M = 2S_{ij}^M/\sqrt{\gamma_{ij}(\phi_0)} \le 1$; squaring then gives $\|\mu_k - \mu_l\|_{M_{ij}}^2 \ge \gamma_{ij}(\phi_0) - 2\sqrt{\gamma_{ij}(\phi_0)}\, S_{ij}^M = \gamma_{ij}(\phi_0)\,(1 - \nu_{ij}^M)$, and with Part (a) we obtain

$$D_{kl}^M \;\ge\; \frac{\gamma_{ij}(\phi_0)}{\kappa_{ij}^{\mathrm{cls}}}\,\big(1 - \nu_{ij}^M\big). \tag{24}$$

Assumption 3.2(ii) (non-degenerate scatter with $\lambda_{\min}(\Sigma^{(t)}) \ge \sigma_{\min}^2 > 0$) and the positive-separation condition (iii) keep the quantities finite and $\gamma_{ij}(\phi_0) > 0$ but do not by themselves bound ν_{ij}^M away from 1; that bound is an additional hypothesis. Under Condition (S') ($\nu_{ij}^M \le 1/2$; §E.4), (24) implies $D_{kl}^M \ge \gamma_{ij}(\phi_0)/(2\,\kappa_{ij}^{\mathrm{cls}})$ in the worst case, and the constant $1/8$ in (15) becomes $1/(16\,\kappa_{ij}^{\mathrm{cls}})$ in this worst-case analysis, equal to $1/16$ under homoscedastic classes ($\kappa_{ij}^{\mathrm{cls}} = 1$; Corollary E.12 and Corollary D.3). Under negligible spread ($\nu_{ij}^M = 0$) with homoscedastic classes, (24) lower-bounds D_{kl}^M by $\gamma_{ij}(\phi_0)$ itself and the constant $1/8$ is recovered, the form stated by the operational-rate corollary (Corollary 3.5).

Step 4: Aggregating over the inter-task block

Combining Steps 1–3, for any class pair (k, l) with $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$ and $\nu_{ij}^M \le 1$:

$$\rho_{kl}(\psi;\phi_0) \;\le\; \frac{M}{N-1}\,\exp\!\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big),$$

with the cleanest case $\nu_{ij}^M = 0$ and $\kappa_{ij}^{\mathrm{cls}} = 1$ (negligible spread, homoscedastic classes; perfectly Gaussian class-conditionals) giving the unqualified constant $\frac{1}{8}$; the worst-case constant under Condition (S') is $1/(16\,\kappa_{ij}^{\mathrm{cls}})$ (see Remark E.2 below).

The inter-task block P_{ij} contains C^2 entries, one for each pair (k, l) with $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$. Summing,

$$|P_{ij}(\psi;\phi_0)| \;=\; \sum_{k\in\mathcal{Y}^{(i)},\, l\in\mathcal{Y}^{(j)}} \rho_{kl}(\psi;\phi_0) \;\le\; \frac{MC^2}{N-1}\,\exp\!\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big).$$

This bound covers a single ordered inter-task block. The symmetric contribution from the transposed block P_{ji} is bounded separately by the same expression and enters Theorem 3.4 through the sum over $T(T-1)$ ordered pairs. This completes the proof. □

Remark E.2 (The constant in the exponent). The constant $1/8$ is exact for two-class Bayes-error problems with Gaussian class-conditionals under the convention $\gamma_{ij} = \|\Delta\mu\|^2_{((\Sigma^{(i)}+\Sigma^{(j)})/2)^{-1}}$. Under Definition 3.1 as written in §3, this is the constant. Under alternative conventions (e.g., $\Sigma^{(i)} + \Sigma^{(j)}$ without averaging) the constant becomes $1/4$. A worst-case analysis accounting for non-Gaussian features and within-task spread (Step 3, part b) yields $1/(16\,\kappa_{ij}^{\mathrm{cls}})$, equal to $1/16$ under homoscedastic classes. The canonical statements are therefore scoped: Theorem 3.4 carries the conservative $1/(16\,\kappa_{\max}^{\mathrm{cls}})$ under its full hypothesis set (Condition (S'); $1/16$ under homoscedastic classes, Corollary D.3), and Corollary 3.5 the operational $1/8$ as the negligible-spread homoscedastic limiting case. We use the empirical slope of $\log(\text{TC gap})$ vs $\gamma_{\min}$ in §5 to discriminate which rate is operative on real data.

Remark E.3 (Non-Gaussian features). The Bhattacharyya bound (19) holds for any class-conditional distributions p_k, p_l, not only Gaussians. For non-Gaussian features the closed-form expression (20) no longer holds exactly, and no route outside the Gaussian branch is proved in this paper; we record the expected extensions. (i) For sub-Gaussian features with proxy covariance $\Gamma_k \preceq \Sigma_k$, a Chernoff bound on the log-likelihood ratio between p_k and p_l is expected to yield an exponent of the same order, with the proxy-to-covariance ratio entering the constant; whether the closed constant $1/8$ of (21) is preserved in this branch is open (this is why Lemma 3.3 and Theorem 3.4 are scoped to the Gaussian branch of clause (iv)). (ii) For features satisfying a c-sub-exponential concentration, a slightly weaker bound is expected to hold with $1/8$ replaced by $1/(8c^2)$, which translates to a factor inside the exponent of (15). We have verified empirically (Appendix I.4) that the features produced by ViT-IN21K, DINOv2, and CLIP backbones on the four benchmarks of §5 pass standard Gaussianity tests after layer-wise normalisation, justifying the Gaussian-class-conditional approximation.

Remark E.4 (Symmetry of the loss matrix). The loss matrix $P(\theta)$ is symmetric in its off-diagonal structure: $\rho_{kl}(\theta) = \rho_{lk}(\theta)$ by the definition (3). Consequently $|P_{ij}(\theta)| = |P_{ji}(\theta)|$, and Theorem 3.4 sums over the $T(T-1)$ ordered pairs (i, j) with $i \neq j$, where each ordered pair contributes the bound from Lemma 3.3 on the block P_{ij}. The transposed pair (j, i) contributes the (equal) bound on P_{ji}, which appears as a separate term in the sum.

E.2 Proof of Theorem 3.4 (Relaxed Infeasibility)

We restate the theorem and then provide a proof in three steps: (i) decompose the total loss matrix sum into diagonal and off-diagonal parts at the CF-optimal head $\psi^\dagger$; (ii) bound the gap between $\psi^\dagger$ and the globally optimal head $\psi^\star$ by the off-diagonal contribution at $\psi^\dagger$; (iii) apply Lemma 3.3 to each of the $T(T-1)$ off-diagonal blocks.

Theorem (Theorem 3.4 restated). Let ϕ_0 be a frozen backbone, and assume task incompatibility (Assumption N.3), Assumption 3.2, clauses (i)–(iii) with clause (iv) taken in its Gaussian branch (Remark E.3), the prototype-based head class (Assumption D.1), Condition (L), and Condition (S'). Let $\kappa^{\mathrm{cls}}_{\max}(\phi_0) := \max_{i \neq j} \kappa^{\mathrm{cls}}_{ij}(\phi_0)$ (Definition E.9) and let $\psi^\dagger$ be any CF-optimal head: $\psi^\dagger \in \arg\min_\psi \sum_t |P_{tt}(\psi; \phi_0)|$. The TC gap of $\psi^\dagger$ is bounded by

$$0 \;\leq\; \underbrace{\sum_{i,j} |P_{ij}(\psi^\dagger; \phi_0)| - \min_\psi \sum_{i,j} |P_{ij}(\psi; \phi_0)|}_{\mathrm{TC\ gap}(\phi_0)} \;\leq\; \frac{MC^2\, T(T-1)}{N-1}\, \exp\Big(-\frac{1}{16\, \kappa^{\mathrm{cls}}_{\max}(\phi_0)}\, \gamma_{\min}(\phi_0)\Big), \tag{25}$$

where the conservative exponent $\frac{1}{16\, \kappa^{\mathrm{cls}}_{\max}}$ is the Condition (S') worst case $\nu^M_{ij} = \frac{1}{2}$ of the interpolating form $\frac{1-a}{8\, \kappa^{\mathrm{cls}}_{\max}}$ of Corollary 3.5 (Corollary E.12 at the per-pair level), equal to $\frac{1}{16}$ under homoscedastic classes (Corollary D.3); the operational rate $\exp(-\gamma_{\min}(\phi_0)/8)$ is the negligible-spread homoscedastic limiting case of that corollary, not a separate claim.

Step 1: Diagonal–off-diagonal decomposition at $\psi^\dagger$

The total sum over the loss matrix decomposes by block:

$$\sum_{i,j} \big|P_{ij}(\psi^\dagger; \phi_0)\big| \;=\; \underbrace{\sum_{t=1}^{T} \big|P_{tt}(\psi^\dagger; \phi_0)\big|}_{S_{\mathrm{diag}}(\psi^\dagger)} \;+\; \underbrace{\sum_{i \neq j} \big|P_{ij}(\psi^\dagger; \phi_0)\big|}_{S_{\mathrm{off}}(\psi^\dagger)}, \tag{26}$$

where S_{diag} collects the T intra-task blocks and S_{off} collects the $T(T-1)$ inter-task blocks. By definition of CF-optimality,

$$S_{\mathrm{diag}}(\psi^\dagger) \;=\; \min_{\psi \in \Psi} S_{\mathrm{diag}}(\psi) \;=:\; S^\star_{\mathrm{diag}}. \tag{27}$$

Step 2: Bounding the gap by the off-diagonal contribution at $\psi^\dagger$

Let $\psi^\star \in \arg\min_\psi \sum_{i,j} |P_{ij}(\psi; \phi_0)|$ be a globally optimal head. By definition,

$$\sum_{i,j} |P_{ij}(\psi^\star; \phi_0)| \;=\; S_{\mathrm{diag}}(\psi^\star) + S_{\mathrm{off}}(\psi^\star). \tag{28}$$

Both S_{diag} and S_{off} are non-negative. Since $\psi^\star$ minimises their sum but $\psi^\dagger$ minimises S_{diag} alone, we have

$$S_{\mathrm{diag}}(\psi^\star) \;\geq\; S^\star_{\mathrm{diag}} \;=\; S_{\mathrm{diag}}(\psi^\dagger). \tag{29}$$

Subtracting (28) from (26):

$$\begin{aligned} \mathrm{TC\ gap}(\phi_0) \;&=\; \sum_{i,j} \big|P_{ij}(\psi^\dagger; \phi_0)\big| - \sum_{i,j} |P_{ij}(\psi^\star; \phi_0)| \\ &=\; \big(S_{\mathrm{diag}}(\psi^\dagger) - S_{\mathrm{diag}}(\psi^\star)\big) + \big(S_{\mathrm{off}}(\psi^\dagger) - S_{\mathrm{off}}(\psi^\star)\big) \\ &\leq\; 0 + S_{\mathrm{off}}(\psi^\dagger) \;=\; S_{\mathrm{off}}(\psi^\dagger), \end{aligned} \tag{30}$$

where the first parenthesis is non-positive by (29) and the second is bounded above by $S_{\text{off}}(\psi^\dagger)$ since $S_{\text{off}}(\psi^\star) \geq 0$. Thus

$$\text{TC gap}(\phi_0) \;\leq\; S_{\text{off}}(\psi^\dagger) \;=\; \sum_{i \neq j} \big|P_{ij}(\psi^\dagger; \phi_0)\big| \,. \tag{31}$$

The lower bound TC gap ≥ 0 holds trivially since $\sum_{i,j} \big|P_{ij}(\psi^\dagger; \phi_0)\big| \geq \min_\psi \sum_{i,j} |P_{ij}(\psi; \phi_0)|$ by definition of the minimum.

Step 3: Aggregating the off-diagonal bound

By Lemma 3.3 in its general (pair-metric, spread-carrying) form, each off-diagonal block satisfies

$$\big|P_{ij}(\psi^\dagger; \phi_0)\big| \;\leq\; \frac{MC^2}{N-1} \exp\!\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\text{cls}}}\,\gamma_{ij}(\phi_0)\Big),$$

where ν_{ij}^M is the pair-metric spread ratio of Definition E.10 and κ_{ij}^{cls} the class-to-task scatter ratio of Definition E.9 (the pair-metric Step 3 of §E.1, Lemma E.11). Under Condition (S'), $\nu_{ij}^M \leq 1/2$, so $\frac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\text{cls}}} \geq \frac{1}{16\,\kappa_{ij}^{\text{cls}}} \geq \frac{1}{16\,\kappa_{\max}^{\text{cls}}(\phi_0)}$ (§E.4), so the exponent is bounded below by $\frac{1}{16\,\kappa_{\max}^{\text{cls}}}$:

$$\big|P_{ij}(\psi^\dagger; \phi_0)\big| \;\leq\; \frac{MC^2}{N-1} \exp\!\Big(-\tfrac{1}{16\,\kappa_{ij}^{\text{cls}}}\,\gamma_{ij}(\phi_0)\Big) \;\leq\; \frac{MC^2}{N-1} \exp\!\Big(-\frac{1}{16\,\kappa_{\max}^{\text{cls}}(\phi_0)}\,\gamma_{ij}(\phi_0)\Big),$$

where the second inequality uses $\kappa_{ij}^{\text{cls}} \leq \kappa_{\max}^{\text{cls}}(\phi_0)$ and $\gamma_{ij}(\phi_0) \geq \gamma_{\min}(\phi_0)$ for all $i \neq j$. There are exactly $T(T-1)$ ordered pairs (i, j) with $i \neq j$ (the off-diagonal blocks of the $T \times T$ block matrix). Summing,

$$\begin{aligned} S_{\text{off}}(\psi^\dagger) \;&=\; \sum_{i \neq j} \big|P_{ij}(\psi^\dagger; \phi_0)\big| \\ &\leq\; T(T-1) \cdot \frac{MC^2}{N-1} \cdot \exp\!\Big(-\frac{1}{16\,\kappa_{\max}^{\text{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big) \\ &=\; \frac{MC^2\,T(T-1)}{N-1} \exp\!\Big(-\frac{1}{16\,\kappa_{\max}^{\text{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big). \end{aligned} \tag{32}$$

Combining (31) and (32) yields (25). In the negligible-spread limit with homoscedastic classes ($\nu_{ij}^M = 0$, $\kappa_{ij}^{\text{cls}} = 1$) every $1/(16\,\kappa_{\max}^{\text{cls}})$ above reverts to $1/8$, the operational rate of Corollary 3.5. □

Tightness of the bound

The bound (25) is tight up to constants in the regime where $\gamma_{\min}(\phi_0)$ is moderate (neither 0 nor ∞):

- In the limit $\gamma_{\min}(\phi_0) \to 0$, the bound reduces to the trivial $S_{\text{off}} \leq \frac{MC^2T(T-1)}{N-1}$ which, modulo the constant, matches the trivial upper bound on a sum of $T(T-1)$ bounded losses.
- In the limit $\gamma_{\min}(\phi_0) \to \infty$, the bound vanishes, and the TC gap vanishes with it.
- For intermediate $\gamma_{\min}(\phi_0)$, the bound is *slope-tight* (the coefficient of $\gamma_{\min}$ in the exponent matches the Bhattacharyya rate from the Bayes-error literature) but the prefactor is loose by the union-bound factor $T(T-1)$ and the cross-entropy correction discussed in Remark E.7 of §E.3.

The empirical slope of $\log(\text{TC gap})$ versus $\gamma_{\min}(\phi_0)$, measured in §5 on softmax cross-entropy heads (an empirical extension beyond the formal scope; Remark D.4), provides the cleanest test of this rate; the prefactor is absorbed into the OLS intercept and is not the object of falsification.

The role of Assumption N.3

The proof above uses Assumption N.3 (task incompatibility) only implicitly: through the fact that $\psi^\dagger$ and $\psi^\star$ are *distinct*, so that the TC gap is genuinely non-trivial. Without Assumption N.3, the diagonal and total objectives might share a common minimiser, in which case $\psi^\dagger = \psi^\star$ and the TC gap is zero. The bound (25) still holds in this trivial case (both sides reduce to the bound on $S_{\text{off}}(\psi^\dagger)$, which is automatic). Assumption N.3 is what makes the bound *informative*, not vacuous.

Remark E.5 (Comparison with Theorem 2.1). Theorem 2.1 (the original Infeasibility Theorem) states that the TC gap is strictly positive whenever Assumption N.3 holds and $\sum_{i,j} |P_{ij}(\theta)| \not\equiv \sum_t |P_{tt}(\theta)|$. Theorem 3.4 adds an upper bound on the gap. The two theorems are compatible: under Assumption N.3, the gap is strictly positive (Theorem 2.1) and at most exponentially small in $\gamma_{\min}$ (Theorem 3.4). The two together yield, for any $\gamma_{\min}(\phi_0) > 0$,

$$0 \;<\; \mathrm{TC\,gap}(\phi_0) \;\le\; \frac{MC^2\,T(T-1)}{N-1}\,\exp\!\Big(-\tfrac{1}{16\,\kappa^{\mathrm{cls}}_{\max}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big),$$

the conservative exponent of Theorem 3.4 ($1/16$ under homoscedastic classes, Corollary D.3); the operational rate $\exp(-\gamma_{\min}(\phi_0)/8)$ of Corollary 3.5 holds in the negligible-spread homoscedastic limit. The "squeeze" between the strict positivity and the exponential upper bound is the precise statement of the title's promise: *from Infeasibility to Exponential Attenuation*.

Remark E.6 (A different sense of "optimality" for $\psi^\star$). The globally optimal head $\psi^\star$ is the head that *would* arise from joint training over all tasks simultaneously, with full access to all class data. It is the head against which class-IL is benchmarked, not a head that any class-IL algorithm can produce; sequential training with limited memory cannot in general recover $\psi^\star$. Theorem 3.4 bounds the loss of any CF-optimal head relative to this idealised reference.

E.3 The Constant in the Exponent: A Defensive Analysis

The exponent map. Four constants appear in the exponents above, each bought by a different hypothesis set. All four sit on top of Condition (L) and the Gaussian branch of Assumption 3.2(iv); the rows below differ only in the spread and scatter hypotheses they add.

Constant	Hypotheses required	Where stated	When it applies
$\frac{1}{8}$	negligible spread $\nu^M_{ij} = 0$ with homoscedastic classes $\kappa^{\mathrm{cls}}_{ij} = 1$	Corollary 3.5	tight homoscedastic classes; the empirical regime of §5
$\frac{1}{16}$	Condition (S′) with $\kappa^{\mathrm{cls}}_{ij} = 1$	Corollaries D.3, E.12	shared covariance within tasks
$\frac{1}{16\,\kappa^{\mathrm{cls}}_{\max}}$	full Condition (S′)	Theorem 3.4	the general guarantee
$\frac{1-a}{8\,\kappa^{\mathrm{cls}}_{\max}}$	spread ratio measured at $a = \nu^M_{ij} \le 1$	Corollary 3.5; §E.3	interpolating form, empirical reading of a

The table shows the price of generality: each weakening of the hypotheses multiplies the constant by a measurable factor.

Lemma 3.3 states the off-diagonal block bound with constant $\frac{1}{8}$ in the exponent, the negligible-spread homoscedastic (operational) constant; the guaranteed exponent under Condition (S') is $\frac{1}{16\,\kappa^{\mathrm{cls}}_{\max}}$ (Theorem 3.4), equal to $\frac{1}{16}$ under homoscedastic classes (Corollary D.3). This constant has a specific provenance and a specific scope. We document both here, to clarify which features of the bound are robust and which depend on modelling choices.

The $\frac{1}{8}$ constant of Corollary 3.5 comes from the Chernoff–Bhattacharyya bound for Gaussian class-conditionals with shared covariance; the pair-metric Step 3 spread ratio under Condition (S') degrades it to the theorem's $\frac{1}{16\,\kappa^{\mathrm{cls}}_{\max}}$, i.e. $\frac{1}{16}$ under homoscedastic classes; under non-shared covariance an AM–GM determinant correction is absorbed. For non-Gaussian features the rate $\gamma_{\min}/8$ should be read as a Gaussian approximation; the empirical slope on softmax CE heads in §5 is -0.012 (bootstrap 95% CI $[-0.019, -0.004]$), about one tenth of $-1/8$—the CI-based consistency with $-1/8$ belonged to the withdrawn accepted-run CI—as an empirical extension beyond the formal scope (Remark D.4). The Infeasibility Theorem is not violated: for any fixed ϕ_0 the TC gap is strictly positive; Theorem 3.4 bounds how large it can be.

Provenance of the constant $\frac{1}{8}$

The Bhattacharyya bound on the Bayes-optimal pairwise error between two Gaussian class-conditionals with means μ_k, μ_l and covariances Σ_k, Σ_l has the closed form

$$P_e^{kl} \;\le\; \exp(-B_{kl}^{\mathrm{Gauss}}), \qquad B_{kl}^{\mathrm{Gauss}} \;=\; \tfrac{1}{8}\,(\mu_k-\mu_l)^\top \left(\frac{\Sigma_k+\Sigma_l}{2}\right)^{-1}(\mu_k-\mu_l)+\tfrac{1}{2}\log\frac{|(\Sigma_k+\Sigma_l)/2|}{\sqrt{|\Sigma_k|\,|\Sigma_l|}}, \tag{33}$$

where the second term is the AM-GM determinant correction, non-negative [Duda et al., 2000, Eq. 6.7.5]. The constant $\frac{1}{8}$ in the first term is the source of the $\frac{1}{8}$ in Lemma 3.3. It arises from a specific calculation: the saddle-point approximation of the Bayesian error integral over two Gaussian density crossings, with the $\frac{1}{8}$ emerging from the $\frac{1}{2}$ in the Gaussian density's quadratic exponent times the $\frac{1}{4}$ in the optimal threshold location.

This is *not* a tunable constant in the proof; it is the output of a textbook calculation under a specific distributional assumption (Gaussian class-conditionals) and a specific parameterisation of the inter-class metric (averaged covariances). Changing either changes the constant.

Three sources of constant variation

We isolate three independent factors that can shift the constant in Lemma 3.3 and discuss each.

Factor 1: Covariance averaging convention. Definition 3.1 uses the *averaged* covariance $(\Sigma^{(i)} + \Sigma^{(j)})/2$ in the metric. An alternative convention—used in some related-classifier analyses Pang et al. [2005]—is to use the *summed* covariance $\Sigma^{(i)} + \Sigma^{(j)}$ without dividing by 2. The two conventions are equivalent up to a factor of 2 in γ_{ij}:

$$\gamma_{ij}^{\mathrm{sum}}(\phi) \;=\; \tfrac{1}{2}\,\gamma_{ij}^{\mathrm{avg}}(\phi) \;=\; \tfrac{1}{2}\,\gamma_{ij}(\phi). \tag{34}$$

Under the summed convention, Lemma 3.3 would state the bound with $\exp(-\gamma_{ij}^{\mathrm{sum}}/4)$ in place of $\exp(-\gamma_{ij}/8)$. Both are mathematically equivalent. We adopt the averaged convention throughout because it matches the textbook Bhattacharyya distance and yields the cleaner constant $\frac{1}{8}$.

Factor 2: Class-conditional distribution. The constant $\frac{1}{8}$ is exact for Gaussian class-conditionals, and the Gaussian branch of clause (iv) of Assumption 3.2 is the branch the proofs cover (the scoping of Lemma 3.3 and Theorem 3.4). For sub-Gaussian class-conditionals with proxy covariance $\Gamma_k \preceq \Sigma_k$, a Chernoff bound on the log-likelihood ratio between classes k and l is expected to yield a bound of the $\exp(-D_{kl}^M/8)$ form with the proxy-to-covariance ratio entering the constant; this step is not proved in this paper (Remark E.3). For broader classes of distributions (e.g., sub-exponential with parameter c), the constant is expected to degrade:

- Sub-Gaussian ($\Gamma_r \preceq \Sigma_r$): expected $\frac{1}{8}$ up to the proxy ratio; not proved here.
- Sub-exponential with proxy parameter c: expected $\frac{1}{8c^2}$ in the exponent, plus a polynomial-in-γ_{ij} correction.
- Heavy-tailed (only bounded p-th moment for some $p > 2$): no exponential bound; the bound becomes polynomial in γ_{ij}.

For the foundation-model features studied in §5, we verify in Appendix I.4 that the class-conditional distributions of features from ViT-IN21K, DINOv2, and CLIP backbones, after layer normalisation, pass standard multivariate Gaussianity tests on a per-class basis. The Gaussian branch, in which the constant $\frac{1}{8}$ is proved, is the empirically supported regime.

Factor 3: Within-task class spread (the pair-metric ratio ν_{ij}^M). The Mahalanobis distance bound $D_{kl}^M \ge \frac{\gamma_{ij}(\phi_0)}{\kappa_{ij}^{\mathrm{cls}}}\,(1-\nu_{ij}^M)$ used in Step 3 of §E.1 (Lemma E.11) requires the pair-metric spread ratio to be bounded away from 1. If within-task class spread is non-negligible relative to inter-task separation, the effective constant degrades. Under $\nu_{ij}^M = a \le 1$ (Definition E.10) the effective exponent is $\frac{1-a}{8\,\kappa_{ij}^{\mathrm{cls}}}$:

- $a = 0$ with homoscedastic classes ($\kappa_{ij}^{\mathrm{cls}} = 1$; no within-task spread): constant $\frac{1}{8}$ exact.

- $a = \frac{1}{2}$ (within-task spread half of inter-task separation): exponent $\frac{1}{16\,\kappa_{ij}^{\mathrm{cls}}}$, i.e. $\frac{1}{16}$ under homoscedastic classes.
- $a \to 1$ (within-task spread approaches inter-task separation): exponent $\to 0$, bound becomes vacuous.

Condition (S') (§E.4) bounds $a = \nu_{ij}^{M} \leq \frac{1}{2}$ *as a hypothesis*: it is checkable per configuration but empirical, not universal; the count of the twelve configurations of Table 4 satisfying or violating it in the pair metric is not reported (the printed ν column is the isotropic-metric value; §K.2). Under the condition the worst-case exponent $\frac{1}{16\,\kappa_{ij}^{\mathrm{cls}}}$ is the guaranteed bound (Theorem 3.4; $\frac{1}{16}$ under homoscedastic classes, Corollary D.3). The tighter constant $\frac{1}{8}$ holds when within-task class spread is negligible and classes are homoscedastic, which is the empirical regime for the benchmarks of §5. Explicitly, matching §E.4: the spread bound is empirical and is not guaranteed under Assumption 3.2(ii)–(iii); those clauses keep S_{ij}^{M} finite but do not by themselves bound ν_{ij}^{M} away from 1, which is why Condition (S') is stated as a separate hypothesis and checked per configuration (Table 4).

Conservative versus operational statements of the bound

The above analysis yields two versions of Lemma 3.3.

- **Conservative version** (Gaussian class-conditionals, the clause (iv) Gaussian branch; worst-case $\nu_{ij}^{M} \leq \frac{1}{2}$ under Condition (S')):

$$|P_{ij}(\psi;\phi_0)| \;\leq\; \frac{MC^2}{N-1}\,\exp\Bigl(-\tfrac{1}{16\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Bigr).$$

- **Operational version** (Gaussian features, $\nu_{ij}^{M} \approx 0$ with homoscedastic classes, $\kappa_{ij}^{\mathrm{cls}} = 1$):

$$|P_{ij}(\psi;\phi_0)| \;\leq\; \frac{MC^2}{N-1}\,\exp\Bigl(-\tfrac{1}{8}\,\gamma_{ij}(\phi_0)\Bigr).$$

The headline statement in §3 (Theorem 3.4) uses the conservative exponent $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$, the exponent its full hypothesis set (Condition (S')) delivers, equal to $\frac{1}{16}$ under homoscedastic classes (Corollary D.3); Lemma 3.3 and Corollary 3.5 carry the operational constant $\frac{1}{8}$, the rate predicted by the Bhattacharyya analysis in the negligible-spread homoscedastic limit and the regime the empirically verified conditions of §5 occupy. The conservative $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ provides a robust fall-back when the negligible-spread conditions are violated, degrading by the measured $\kappa_{\max}^{\mathrm{cls}} \leq C$ (Definition E.9) instead of failing outright.

Empirical falsification target

The empirical slope of $\log(\text{TC gap})$ versus $\gamma_{\min}(\phi_0)$, measured across the 12 pretrained (backbone, dataset) configurations (the claim population) and reported in Table 5 (the corresponding 16-configuration robustness fit that adds the four random-backbone anchors appears in §5 and Table 3), is -0.115 (95% bootstrap CI $[-0.128, -0.102]$), with per-backbone slopes of -0.141 (ViT-IN21K), -0.099 (DINOv2), and -0.118 (CLIP): accepted-run values, not reproduced under the released protocol, whose regenerated pooled slope is -0.012 (bootstrap 95% CI $[-0.019, -0.004]$), about one tenth of the operational -0.125. The operational constant predicts slope -0.125; the conservative exponent guarantees decay at least as fast as slope $-1/(16\,\kappa_{\max}^{\mathrm{cls}})$, i.e. -0.0625 under homoscedastic classes (an upper-bound guarantee, not a point prediction). The regenerated CI excludes both constants; the $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ guarantee itself is not falsified (with $\kappa_{\max}^{\mathrm{cls}} > 1$ the guaranteed slope is shallower still), and the measured within-task spread ratios of the verification run, 2.1 to 4.7 against the required $1/2$, show Condition (S') not binding on these configurations, so the bound is reported with its measured exponent. The per-backbone variation traces to the within-task spread condition $\nu_{ij}^{M} \leq 1/2$ (Table 4); the per-backbone counts (how many datasets of ViT-IN21K, DINOv2, and CLIP violate it in the pair metric) are not reported; the printed ν column is the isotropic-metric value (§K.2). The conservative $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ represents a margin against worst-case within-task spread and worst-case class-to-task scatter; the foundation models we study fall predominantly in the boundary regime on these measurements.

Remark E.7 (Cross-entropy and unbounded losses). The constant M in Lemma 3.3 bounds the per-sample loss v. For zero-one and bounded-margin losses M is a constant. For cross-entropy with unbounded logits, M is the truncation level; in practice we cap logits via temperature scaling or label smoothing, both of which give finite M. The empirical TC-gap measurements in §5 (an empirical extension beyond the formal scope; Remark D.4) use a fixed-temperature softmax, so M is bounded; the operational $\frac{1}{8}$ rate is preserved.

Remark E.8 (Independence from clipping and optimisation). The constant in the exponent is a property of the feature distribution and the inter-class metric, *not* of the optimisation procedure. Heavy-tailed gradient noise (Assumption A.1), gradient clipping in Proposition 4.3, and the convergence rate $K^{-(\alpha-1)/\alpha}$ all enter the E_{CF} term of Theorem 4.5 but do not affect E_{TC} or its rate of decay in $\gamma_{\min}$. This is the formal content of the disjoint-dependencies claim of §4.

E.4 Pair-Metric Geometry and Condition (S′) (ν_{ij}^M, $\kappa_{ij}^{\mathrm{cls}}$)

Step 3 of §E.1 (the proof of Lemma 3.3) lower-bounds the per-class-pair Mahalanobis distance D_{kl}^M between classes $k \in \mathcal{Y}^{(i)}$ and $l \in \mathcal{Y}^{(j)}$ by the inter-task Mahalanobis margin $\gamma_{ij}(\phi_0)$ of Definition 3.1. This appendix states that step in the task-pair metric $M_{ij} := (\Sigma^{(i)} + \Sigma^{(j)})/2$ itself. Passing both D_{kl}^M and γ_{ij} through the isotropic upper bound $\bar{\Sigma} = \lambda_{\max} I$ instead would price the step with a metric-mismatch term $\delta_{ij} = \gamma_{ij} - \|\Delta\mu\|^2_{\bar{\Sigma}^{-1}}$, controlled by no hypothesis of the paper and close to γ_{ij} itself on anisotropic features, where the isotropic separation is a few percent of the Mahalanobis one. That detour is not needed: the triangle inequality holds in the Mahalanobis norm of M_{ij}, and the only thing to control is how far a class covariance can sit above its task covariance. Recall the setup: tasks i, j have class sets $\mathcal{Y}^{(i)}, \mathcal{Y}^{(j)}$ with $\left|\mathcal{Y}^{(i)}\right| = \left|\mathcal{Y}^{(j)}\right| = C$, class means μ_r, task centroids $\mu^{(t)} = \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \mu_r$, and task covariances $\Sigma^{(t)} = \frac{1}{C}\sum_{r\in\mathcal{Y}^{(t)}} \Sigma_r$, so that M_{ij} is exactly the metric of γ_{ij} in Definition 3.1.

Definition E.9 (Class-to-task scatter ratio)**.** For tasks $i \neq j$ let $\kappa_{ij}^{\mathrm{cls}} := \max_{t\in\{i,j\}} \max_{k\in\mathcal{Y}^{(t)}} \lambda_{\max}\big(\Sigma^{(t)\,-1/2}\,\Sigma_k\,\Sigma^{(t)\,-1/2}\big)$, so that $\Sigma_k \preceq \kappa_{ij}^{\mathrm{cls}}\,\Sigma^{(t(k))}$ for every class k of the two tasks. Always $1 \le \kappa_{ij}^{\mathrm{cls}} \le C$, with equality to 1 under homoscedastic classes.

Definition E.10 (Within-task spread ratio, pair metric)**.** With $\|v\|_{M_{ij}} := \sqrt{v^\top M_{ij}^{-1} v}$ and $M_{ij} = (\Sigma^{(i)} + \Sigma^{(j)})/2$,

$$S_{ij}^M := \max_{k\in\mathcal{Y}^{(i)}} \left\|\mu_k - \mu^{(i)}\right\|_{M_{ij}} + \max_{l\in\mathcal{Y}^{(j)}} \left\|\mu_l - \mu^{(j)}\right\|_{M_{ij}}, \qquad \nu_{ij}^M := \frac{2\,S_{ij}^M}{\sqrt{\gamma_{ij}(\phi_0)}}.$$

Lemma E.11 (Per-class-pair distance from the inter-task margin)**.** *For every $k \in \mathcal{Y}^{(i)}$, $l \in \mathcal{Y}^{(j)}$ with $\nu_{ij}^M \le 1$,*

$$D_{kl}^M \;\ge\; \frac{\gamma_{ij}(\phi_0)}{\kappa_{ij}^{\mathrm{cls}}}\,\big(1 - \nu_{ij}^M\big).$$

Proof. $(\Sigma_k + \Sigma_l)/2 \preceq \kappa_{ij}^{\mathrm{cls}}(\Sigma^{(i)} + \Sigma^{(j)})/2 = \kappa_{ij}^{\mathrm{cls}} M_{ij}$, hence $((\Sigma_k + \Sigma_l)/2)^{-1} \succeq M_{ij}^{-1}/\kappa_{ij}^{\mathrm{cls}}$ and $D_{kl}^M \ge \|\mu_k - \mu_l\|^2_{M_{ij}}/\kappa_{ij}^{\mathrm{cls}}$. The Mahalanobis norm is a norm, so the triangle inequality at the task centroids gives $\|\mu_k - \mu_l\|_{M_{ij}} \ge \left\|\mu^{(i)} - \mu^{(j)}\right\|_{M_{ij}} - S_{ij}^M = \sqrt{\gamma_{ij}} - S_{ij}^M \ge 0$, the last inequality by $\nu_{ij}^M \le 1$. Squaring, $\|\mu_k - \mu_l\|^2_{M_{ij}} \ge \gamma_{ij} - 2\sqrt{\gamma_{ij}}\,S_{ij}^M = \gamma_{ij}(1 - \nu_{ij}^M)$. □

Condition (S') (Bounded within-task spread ratio, pair metric)**.** $\nu_{ij}^M \le \frac{1}{2}$ for the task pair (i, j); in the hypothesis set of Theorem 3.4, for every ordered task pair $i \neq j$.

Corollary E.12. *Under Condition (S'), for every cross-task class pair of (i, j), with P_e^{kl} the Bayes-optimal pairwise error of Step 2 of §E.1,*

$$P_e^{kl} \;\le\; \exp\!\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big) \;\le\; \exp\!\Big(-\frac{\gamma_{ij}(\phi_0)}{16\,\kappa_{ij}^{\mathrm{cls}}}\Big).$$

Under homoscedastic classes ($\kappa_{ij}^{\mathrm{cls}} = 1$) this is the constant $1/16$; under negligible spread ($\nu_{ij}^M \to 0$) and homoscedastic classes it is the operational $1/8$.

Sufficient trace criteria for Condition (S')

The proof of Lemma 3.3 carries the pair-metric Step 3 (Lemma E.11) under Condition (S'), stated above. Condition (S') is part of the full hypothesis set of the restated Theorem 3.4 (§E.2), alongside the zero-one-type loss condition (Condition (L)) and the class-conditional clause (iv) of Assumption 3.2 taken in its Gaussian branch; in the stronger negligible-spread homoscedastic limit ($\nu_{ij}^M = 0$ with $\kappa_{ij}^{\mathrm{cls}} = 1$) the operational-rate corollary (Corollary 3.5) recovers the $1/8$ constant, and at $\nu_{ij}^M = 1/2$ its interpolating exponent $\frac{1-a}{8\,\kappa_{ij}^{\mathrm{cls}}}$ (with $a = \nu_{ij}^M$) equals the conservative $\frac{1}{16\,\kappa_{ij}^{\mathrm{cls}}}$ (Corollary E.12).

The pair-metric within-task spread S_{ij}^M is bounded by the trace of the within-task scatter, normalised by the spectral floor of the pair metric.

For each task t, the within-task between-class variance is

$$\sum_{r \in \mathcal{Y}^{(t)}} \left\| \mu_r - \mu^{(t)} \right\|^2 \;\le\; C \cdot \mathrm{tr}\big(\Sigma_{\mathrm{between}}^{(t)}\big),$$

where $\Sigma_{\mathrm{between}}^{(t)}$ is the between-class covariance within task t. Translating to the pair metric via $\|v\|_{M_{ij}}^2 \le \|v\|^2 / \sigma_{\min}^2$ (using $M_{ij} = (\Sigma^{(i)} + \Sigma^{(j)})/2 \succeq \sigma_{\min}^2 I$, by Assumption 3.2(ii) applied to each task covariance):

$$\left\| \mu_k - \mu^{(i)} \right\|_{M_{ij}}^2 \;\le\; \frac{\left\| \mu_k - \mu^{(i)} \right\|^2}{\sigma_{\min}^2} \;\le\; \frac{C \cdot \mathrm{tr}(\Sigma_{\mathrm{between}}^{(i)})}{\sigma_{\min}^2}.$$

Taking the worst case over $k \in \mathcal{Y}^{(i)}$:

$$\max_{k \in \mathcal{Y}^{(i)}} \left\| \mu_k - \mu^{(i)} \right\|_{M_{ij}} \;\le\; \sqrt{\frac{C \cdot \mathrm{tr}(\Sigma_{\mathrm{between}}^{(i)})}{\sigma_{\min}^2}}. \tag{35}$$

By Definition E.10 and (35), Condition (S') is certified by the sufficient trace inequality

$$4\sqrt{2}\sqrt{\frac{C \cdot \mathrm{tr}(\Sigma_{\mathrm{between}}^{(i)}) + C \cdot \mathrm{tr}(\Sigma_{\mathrm{between}}^{(j)})}{\sigma_{\min}^2}} \;\le\; \sqrt{\gamma_{ij}(\phi_0)}, \tag{36}$$

or, equivalently after squaring, if

$$\gamma_{ij}(\phi_0) \;\ge\; 32 \cdot \frac{C \cdot (\mathrm{tr}(\Sigma_{\mathrm{between}}^{(i)}) + \mathrm{tr}(\Sigma_{\mathrm{between}}^{(j)}))}{\sigma_{\min}^2}, \tag{37}$$

where the constant 32 (not 16) carries the Cauchy–Schwarz slack $(\sqrt{a} + \sqrt{b})^2 \le 2\,(a + b)$: with the constant 16, or with constant 4 instead of $4\sqrt{2}$ in (36), the criteria bound ν_{ij}^M only up to a factor of $\sqrt{2}$ ($\nu_{ij}^M \le \sqrt{2}/2$), i.e. they are sufficient only up to a factor of two in the exponent. The tighter cross-term form $\gamma_{ij} \ge 16\,(\sqrt{a} + \sqrt{b})^2$ (with a, b the two per-task trace ratios) removes that slack.

In words, Condition (S') requires the inter-task Mahalanobis margin $\gamma_{ij}(\phi_0)$ to dominate the within-task class spread by a factor of at least $32 \cdot C/\sigma_{\min}^2$ (the sufficient summed-form constant; 16 suffices only up to the factor-of-two slack), with the residual geometry carried by the class-to-task scatter ratio $\kappa_{ij}^{\mathrm{cls}} \in [1, C]$ of Definition E.9, not by any metric-mismatch term. This is the regime in which inter-task separation is the dominant geometric feature of the feature space; tasks are well-separated and classes within a task cluster tightly relative to the inter-task distance.

When does this condition hold?

The condition (37) is empirical, not universal. Three regimes illustrate.

Regime 1: Foundation-model features (operational regime). For ViT-IN21K, DINOv2, or CLIP backbones evaluated on standard class-IL benchmarks, empirically observed values are $\gamma_{\min}(\phi_0) \in [5, 52]$ and $\mathrm{tr}(\Sigma_{\mathrm{between}}^{(t)})/\sigma_{\min}^2 \in [0.05, 0.3]$ across tasks. With $C = 10$ classes per task and these ranges, the right-hand side of (37) is at most $32 \cdot 10 \cdot 0.6 = 192$ in the worst case, above much of the observed $\gamma_{\min}$ range: the trace-based sufficient criterion is conservative and does not by itself certify the condition across all configurations. What carries the claim empirically is the measured spread columns of Table 4: the count of configurations satisfying $\nu^M \le 1/2$ in the pair metric is not reported (§5, §K.2); the printed ν column is the isotropic-metric value from the accepted runs.

Regime 2: Random features (degenerate regime). For randomly initialised features, $\gamma_{\min}(\phi_0) \approx 0$ and the within-task class spread is comparable to the inter-task distance. Condition (37) fails, the margin fraction $(1-\nu_{ij}^M)/\kappa_{ij}^{\text{cls}}$ of Lemma E.11 collapses toward 0, and the exponent in Lemma 3.3 degrades from $1/8$ to 0 (the bound becomes vacuous). This is the regime in which the original Infeasibility Theorem operates: TC is unbounded, and pretraining is necessary to make the bound informative.

Regime 3: Intermediate (boundary regime). If the within-task spread is half the inter-task Mahalanobis margin in the pair metric (i.e., $\nu_{ij}^M = 1/2$), the bound on D_{kl}^M from §E.1 Step 3 becomes $D_{kl}^M \geq \gamma_{ij}/(2\,\kappa_{ij}^{\text{cls}})$. This translates to an exponent of $1/(16\,\kappa_{ij}^{\text{cls}})$ in Lemma 3.3, not $1/8$, i.e. $1/16$ under homoscedastic classes (Corollary E.12). Between the endpoints the exponent interpolates linearly in the spread ratio: under $\nu_{ij}^M = a \leq 1$, Step 3 yields $D_{kl}^M \geq (1-a)\,\gamma_{ij}(\phi_0)/\kappa_{ij}^{\text{cls}}$, so the exponent becomes $(1-a)/(8\,\kappa_{ij}^{\text{cls}})$, the form stated by the operational-rate corollary (Corollary 3.5), which recovers $1/8$ at $a = 0$ with homoscedastic classes (negligible spread) and $1/(16\,\kappa_{ij}^{\text{cls}})$ at $a = 1/2$ (the Condition (S') boundary). Configurations slightly above $a = 1/2$ are covered by this weaker-exponent interpolating form, not excluded. We treat the boundary regime as the worst-case operational scenario: the headline statement (Theorem 3.4) carries the conservative $1/(16\,\kappa_{\max}^{\text{cls}})$, and the operational $1/8$ of Corollary 3.5 is its negligible-spread homoscedastic limit, as documented in Remark E.2 of §E.1.

Empirical verification

The condition (37) can be verified directly from training features: compute $\widehat{\gamma}_{ij}(\phi_0)$ for each task pair (§5, §I.1), compute $\text{tr}(\Sigma_{\text{between}}^{(t)})$ for each task by averaging the squared distances of class means from the task centroid, and check the inequality. Table 4 of §K reports this verification across the 12 (backbone, dataset) configurations; it is the table referenced from the operational-rate corollary (Corollary 3.5). *The count of configurations satisfying Condition (S')* ($\nu_{i^\star j^\star}^M \leq 1/2$ in the pair metric, placing them in the operational regime of Corollary 3.5, constant $\frac{1}{8}$) *is not reported*: the printed ν column of the table is the isotropic-metric value from the accepted runs, and the pair-metric quantities (which print ν in the isotropic, pooled and pair metrics, κ^{cls}, and the exponent $(1-\nu^M)/(8\,\kappa^{\text{cls}})$, with per-class covariances Ledoit–Wolf shrunk) are not part of the released record. Configurations not satisfying $\nu^M \leq 1/2$ are covered by the weaker-exponent interpolating form $(1-\nu^M)/(8\,\kappa^{\text{cls}})$ of the corollary. This regime structure is to be read against the per-backbone slope variation observed in Table 5.

An upper bound with exponent $\frac{1}{16\,\kappa_{\max}^{\text{cls}}}$ is not in tension with the regenerated pooled slope of -0.012, one tenth of the operational $-\frac{1}{8}$: an upper bound on the gap remains valid when the measured decay is shallower, indicating the bound is loose on these configurations rather than contradicted, so the slope test discriminates which rate is operative, not which bound is valid.

F Proofs from Section 4 (Heavy-Tailed Streaming Model)

F.1 Proof of Lemma P.1 (Heavy-Tailed Concentration)

The lemma is stated as Lemma P.1 in §P; we restate it here and prove it using the truncated-mean estimator of Catoni Catoni [2012]. The proof balances clipping bias against concentration of the truncated variables.

Lemma (Lemma P.1 restated). Fix $\theta \in \Theta$ and $\delta \in (0,1)$. Let $Z_s := v(f_\theta(x_{t,s}), y_{t,s})$ for $s = 1, \ldots, n_t$ denote the per-sample losses, and let $\widehat{I}_t(\theta) := \frac{1}{n_t}\sum_{s=1}^{n_t} Z_s$ denote the empirical risk. Under Assumption A.1, there exists a constant $C_\alpha > 0$ depending only on α such that with probability at least $1-\delta$,

$$\left|\widehat{I}_t(\theta) - \mathbb{E}\,\widehat{I}_t(\theta)\right| \;\leq\; C_\alpha \cdot V^{1/\alpha} \cdot \frac{\big(\log(2/\delta)\big)^{1-1/\alpha}}{n_t^{1-1/\alpha}}. \tag{38}$$

Proof outline

The standard approach to bounding $\widehat{I}_t$ under bounded α-th central moment is the *truncated mean estimator*. We define a clipped version of each sample, control the bias introduced by clipping, control the concentration of the clipped samples (which are now bounded), and balance the two by choosing the clipping threshold optimally.

Throughout, we work without loss of generality with $\mathbb{E}\, Z_s = 0$ (replace Z_s with $Z_s - \mathbb{E}Z_s$; the bound on the centred quantity translates directly to the bound on $\widehat{I}_t - \mathbb{E}\,\widehat{I}_t$).

Step 1: Truncated mean estimator

Define the truncation operator $\psi_\tau : \mathbb{R} \to \mathbb{R}$ for threshold $\tau > 0$:

$$\psi_\tau(z) \;:=\; \begin{cases} z & \text{if } |z| \le \tau, \\ \tau \operatorname{sign}(z) & \text{if } |z| > \tau. \end{cases} \tag{39}$$

The truncated empirical mean is $\widetilde{I}_\tau := \frac{1}{n_t}\sum_{s=1}^{n_t} \psi_\tau(Z_s)$. We will bound $\left|\widehat{I}_t - \mathbb{E}\,\widehat{I}_t\right|$ by controlling separately:

(i) $\left|\widetilde{I}_\tau - \mathbb{E}\,\widetilde{I}_\tau\right|$ (concentration of bounded variables);

(ii) $\left|\mathbb{E}\,\widetilde{I}_\tau - \mathbb{E}\,\widehat{I}_t\right|$ (clipping bias);

(iii) $\left|\widehat{I}_t - \widetilde{I}_\tau\right|$ (residual from clipping the actual samples).

Step 2: Clipping bias

For each sample, the bias introduced by clipping is

$$|\mathbb{E}\,\psi_\tau(Z_s) - \mathbb{E}\,Z_s| \;=\; \left|\mathbb{E}\big[\psi_\tau(Z_s) - Z_s\big]\right| \;\le\; \mathbb{E}\,|\psi_\tau(Z_s) - Z_s| \;=\; \mathbb{E}\big[\big(|Z_s| - \tau\big)_+\big],$$

where $(\cdot)_+$ denotes the positive part. By Markov's inequality applied to $|Z_s|^\alpha$ with $\mathbb{E}\,|Z_s|^\alpha \le V$ from (10),

$$\mathbb{E}\big[\big(|Z_s| - \tau\big)_+\big] \;\le\; \mathbb{E}\big[|Z_s| \cdot \mathbf{1}\{|Z_s| > \tau\}\big] \;\le\; \frac{\mathbb{E}\,|Z_s|^\alpha}{\tau^{\alpha-1}} \;\le\; \frac{V}{\tau^{\alpha-1}}.$$

The middle inequality uses Hölder's inequality with exponents α and $\alpha/(\alpha-1)$, applied to $|Z_s|$ and $|Z_s|^{\alpha-1}\,\mathbf{1}\{|Z_s| > \tau\}$. Therefore the per-sample clipping bias is bounded by $V/\tau^{\alpha-1}$, and

$$\left|\mathbb{E}\,\widetilde{I}_\tau - \mathbb{E}\,\widehat{I}_t\right| \;\le\; \frac{V}{\tau^{\alpha-1}}. \tag{40}$$

Step 3: Sample-level clipping residual

We also need to control the actual samples falling outside the truncation threshold. By Markov on $|Z_s|^\alpha$,

$$\mathbb{P}\big(|Z_s| > \tau\big) \;\le\; \frac{\mathbb{E}\,|Z_s|^\alpha}{\tau^\alpha} \;\le\; \frac{V}{\tau^\alpha}.$$

By the union bound,

$$\mathbb{P}\big(\exists s : |Z_s| > \tau\big) \;\le\; \frac{n_t \cdot V}{\tau^\alpha}.$$

Choosing τ so that $n_t V/\tau^\alpha \le \delta/2$, i.e.

$$\tau \;\ge\; \left(\frac{2n_t V}{\delta}\right)^{1/\alpha}, \tag{41}$$

ensures that with probability at least $1 - \delta/2$, no Z_s exceeds τ, and therefore $\widetilde{I}_\tau = \widehat{I}_t$. On this high-probability event, the residual $\left|\widehat{I}_t - \widetilde{I}_\tau\right|$ vanishes.

Step 4: Concentration of the truncated mean

Conditional on the high-probability event of Step 3, the truncated samples $\psi_\tau(Z_1), \dots, \psi_\tau(Z_{n_t})$ are each bounded by τ in absolute value. Their variance is bounded by

$$\mathrm{Var}\big(\psi_\tau(Z_s)\big) \;\le\; \mathbb{E}\,\psi_\tau(Z_s)^2 \;\le\; \mathbb{E}\big[|Z_s|^2\,\mathbf{1}\{|Z_s| \le \tau\}\big] + \tau^2 \mathbb{P}(|Z_s| > \tau).$$

By Hölder applied to the first term ($|Z_s|^2 = |Z_s|^\alpha \cdot |Z_s|^{2-\alpha}$, with $2-\alpha \ge 0$ since $\alpha \le 2$), and using the bound on the second term from Step 3,

$$\mathrm{Var}\big(\psi_\tau(Z_s)\big) \;\le\; V^{2/\alpha} \cdot \tau^{2-\alpha} \cdot \mathrm{const}.$$

Applying Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10] to the bounded variables $\psi_\tau(Z_s)$ with their mean $\mathbb{E}\,\psi_\tau(Z_s)$, with probability at least $1 - \delta/2$:

$$\left|\widetilde{I}_\tau - \mathbb{E}\,\widetilde{I}_\tau\right| \;\le\; \sqrt{\frac{2\,\mathrm{Var}(\psi_\tau(Z_s))\log(4/\delta)}{n_t}} + \frac{2\tau\log(4/\delta)}{3n_t}. \tag{42}$$

Substituting the variance bound,

$$\left|\widetilde{I}_\tau - \mathbb{E}\,\widetilde{I}_\tau\right| \;\le\; c_5 \cdot V^{1/\alpha} \cdot \frac{\sqrt{\tau^{2-\alpha}\log(4/\delta)}}{\sqrt{n_t}} + \frac{2\tau\log(4/\delta)}{3n_t}. \tag{43}$$

Step 5: Optimising the clipping threshold

Combining Steps 2-4, on the intersection of the two high-probability events (each with probability $1 - \delta/2$, total $1 - \delta$ by union bound):

$$\begin{aligned}\left|\widehat{I}_t - \mathbb{E}\,\widehat{I}_t\right| \;&\le\; \left|\widehat{I}_t - \widetilde{I}_\tau\right| + \left|\widetilde{I}_\tau - \mathbb{E}\,\widetilde{I}_\tau\right| + \left|\mathbb{E}\,\widetilde{I}_\tau - \mathbb{E}\,\widehat{I}_t\right| \\ &\le\; 0 \;+\; c_5 \cdot V^{1/\alpha} \cdot \frac{\sqrt{\tau^{2-\alpha}\log(4/\delta)}}{\sqrt{n_t}} + \frac{2\tau\log(4/\delta)}{3n_t} \;+\; \frac{V}{\tau^{\alpha-1}}.\end{aligned} \tag{44}$$

The dominant terms are the first (Bernstein concentration term) and third (clipping bias). Setting the threshold to balance them, i.e., choosing

$$\tau \;\sim\; \left(\frac{V \cdot n_t}{\log(2/\delta)}\right)^{1/\alpha}, \tag{45}$$

which satisfies the constraint (41) when $n_t \ge 2\log(2/\delta)$, the dominant terms become

$$V^{1/\alpha} \cdot \tau^{(2-\alpha)/2} \cdot \sqrt{\frac{\log(2/\delta)}{n_t}} \;\sim\; V^{1/\alpha} \cdot \left(\frac{V \cdot n_t}{\log(2/\delta)}\right)^{(2-\alpha)/(2\alpha)} \cdot \sqrt{\frac{\log(2/\delta)}{n_t}},$$

and

$$\frac{V}{\tau^{\alpha-1}} \;\sim\; V \cdot \left(\frac{V \cdot n_t}{\log(2/\delta)}\right)^{-(\alpha-1)/\alpha} \;\sim\; V^{1/\alpha} \cdot \left(\frac{\log(2/\delta)}{n_t}\right)^{(\alpha-1)/\alpha}.$$

The first expression simplifies to $V^{1/\alpha} \cdot (\log(2/\delta))^{1/2-(2-\alpha)/(2\alpha)} \cdot n_t^{(2-\alpha)/(2\alpha)-1/2}$. The exponent of n_t becomes $(2-\alpha)/(2\alpha) - 1/2 = -(\alpha-1)/(2\alpha) - 1/2 + (2-\alpha)/(2\alpha) = -(\alpha-1)/\alpha$ after simplification. Both dominant terms thus scale as

$$V^{1/\alpha} \cdot \left(\frac{\log(2/\delta)}{n_t}\right)^{(\alpha-1)/\alpha} \;=\; V^{1/\alpha} \cdot \frac{\big(\log(2/\delta)\big)^{1-1/\alpha}}{n_t^{1-1/\alpha}}.$$

Combining with the smaller second term from Bernstein (which scales as $\tau/n_t \sim n_t^{-1+1/\alpha}(\log/n_t)^{1/\alpha}$, dominated by the same rate), we obtain

$$\left|\widehat{I}_t - \mathbb{E}\,\widehat{I}_t\right| \;\le\; C_\alpha \cdot V^{1/\alpha} \cdot \frac{\big(\log(2/\delta)\big)^{1-1/\alpha}}{n_t^{1-1/\alpha}},$$

which is (38), with the constant C_α depending only on α (absorbing the various c_5, factors of 2, and the constants from Hölder and Bernstein). □

Reading the rate

The rate $n_t^{-(1-1/\alpha)}$ interpolates between the standard $1/\sqrt{n_t}$ rate at $\alpha = 2$ (recovering Hoeffding) and nearly-zero convergence at $\alpha \to 1^+$ (where the variance explodes and only the first moment exists). The $(\log(2/\delta))^{1-1/\alpha}$ factor reflects the weaker concentration of heavy-tailed distributions; for sub-Gaussian tails ($\alpha = 2$) the factor reduces to $\sqrt{\log(2/\delta)}$.

Remark F.1 (The constant C_α). The explicit constant C_α in (38) can be computed by tracking constants through Steps 2–5. For $\alpha = 2$, $C_2 \le 4\sqrt{2}$ (matching the standard Hoeffding-Bernstein constant); for general $\alpha \in (1, 2)$, C_α grows as α approaches 1 but remains finite for any $\alpha > 1$. We do not track explicit constants in the main text; the only object that depends on them is the prefactor of the TC gap bound, and the empirical experiments calibrate this prefactor against the OLS intercept.

Remark F.2 (Why truncation instead of median-of-means?). An alternative robust estimator under heavy tails is median-of-means (MoM) Catoni [2012]: split the samples into K groups, compute the mean within each group, and take the median of the resulting K means. MoM achieves the same rate (38) but with potentially better constants. For our purposes either estimator works; we use the truncated mean throughout because it has a cleaner closed-form expression, integrates better with the clipped-SGD analysis of §4, and the explicit clipping threshold τ makes the parallel between concentration (this lemma) and optimisation (Proposition 4.3) immediate: the same τ governs both the bias of the truncated estimator and the bias of the clipped gradient.

G Proofs from Section 4 (Convergence Analysis)

G.1 Proof of Proposition 4.3 (Per-Task Convergence, Strongly Convex)

Remark G.1 (Prefactor in Proposition 4.3). The prefactor displayed in Proposition 4.3 is the dimensionally admissible envelope. Assign units once: σ (the α-th-moment bound of the gradient noise) has gradient units, loss per parameter; μ and β have curvature units, loss per squared parameter; Δ_t is in loss units; K_t is dimensionless. The only monomials in (σ, μ, β) with loss units are σ^2/μ and σ^2/β, times any function of the dimensionless condition ratio β/μ, which is why $C^{\mathrm{sc}}_{\mathrm{opt}}$ carries $(\alpha, \beta/\mu)$ and is not a universal constant. Two terms of an earlier prefactor derivation, $\sigma/\mu^{1/\alpha}$ and $\beta\,\sigma^3\,\mu^{(\alpha-4)/\alpha}$, fail this units test for every $\alpha \in (1, 2]$ and are removed pending a units-tracking re-derivation of the Step 5 substitution of §G.1 (the locus the co-author note flags for proofreading). Untouched by the removal: the K_t-exponent $(\alpha-1)/\alpha$, which is what §5 empirically tests and which is conservative relative to the literature's strongly convex clipped-SGD function-value rate $(\sigma^2/\mu)\,K^{-2(\alpha-1)/\alpha}$ up to logarithms.

We restate the proposition and prove it via the standard clipped-SGD analysis under bounded α-th moment of gradient noise. The strongly convex / last-iterate framework follows Sadiev et al. Sadiev et al. [2023]; we adapt their analysis to the head-only optimisation setting of class-IL.

Proposition (Proposition 4.3 restated). Under Assumptions A.1, 4.1 and 4.2, run clipped SGD (7) for $K_t \ge 16$ iterations with

$$\eta_t \;=\; \min\Big\{\frac{1}{4\beta},\; \frac{1}{\beta\sqrt{K_t}}\Big\}, \qquad \tau_t \;=\; \sigma\, K_t^{1/(2\alpha)},$$

and suppose the gradient norms along the trajectory stay below half the clipping level, $\|\nabla L_t(\psi_k)\| \le \tau_t/2$ for all $k < K_t$ (Condition (G); it holds as soon as $K_t \ge (2G_{\max}/\sigma)^{2\alpha}$, where $G_{\max}$ bounds the gradient norms the iterates reach). Then the last iterate $\widehat{\psi}_t := \psi_{K_t}$ satisfies

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min_{\psi} L_t(\psi)\big] \;\le\; \Delta_t\, e^{-(\mu/\beta)\sqrt{K_t}} + c_\alpha\, \frac{\sigma^2/\mu}{K_t^{(\alpha-1)/\alpha}} \;\le\; C^{\mathrm{sc}}_{\mathrm{opt}}(\alpha, \beta/\mu)\, \frac{\Delta_t + \sigma^2/\mu}{K_t^{(\alpha-1)/\alpha}}, \quad (46)$$

where $\Delta_t := L_t(\psi_{t,0}) - \min_\psi L_t(\psi)$, $c_\alpha := 4^{\alpha-1} + 2^{\alpha/2-1}$ and $C^{\mathrm{sc}}_{\mathrm{opt}}(\alpha, \kappa) := \max\{(2p\kappa)^{2p} e^{-2p},\, c_\alpha\}$ with $p := (\alpha-1)/\alpha$ and $\kappa := \beta/\mu$. The constant depends on α and on the condition ratio β/μ only.

Proof outline

The proof proceeds in five steps: (i) write the descent inequality under β-smoothness (Step 1), (ii) establish the two clipping facts, the clipped second moment and the bias under Condition (G) (Step 2),

(iii) combine them with Young's inequality into a one-step recurrence on the function gap (Step 3), (iv) iterate the recurrence (Step 4), and (v) substitute the prescribed η_t, τ_t and balance the residuals (Step 5).

Throughout, write $g_k := \widehat{\nabla L_t}(\psi_k)$ for the stochastic gradient and $\bar{g}_k := \mathrm{clip}_{\tau_t}(g_k)$ for the clipped gradient. Let $\xi_k := g_k - \nabla L_t(\psi_k)$ denote the gradient noise, satisfying $\mathbb{E}\,\|\xi_k\|^\alpha \le \sigma^\alpha$ by Assumption A.1.

Step 1: Descent inequality

By β-smoothness of L_t (Assumption 4.1),

$$L_t(\psi_{k+1}) \;\le\; L_t(\psi_k) \;-\; \eta_t\langle \nabla L_t(\psi_k), \bar{g}_k\rangle \;+\; \frac{\beta\eta_t^2}{2}\,\|\bar{g}_k\|^2 . \tag{47}$$

Taking expectations conditional on ψ_k:

$$\mathbb{E}\big[L_t(\psi_{k+1}) \mid \psi_k\big] \;\le\; L_t(\psi_k) \;-\; \eta_t\langle \nabla L_t(\psi_k), \mathbb{E}[\bar{g}_k \mid \psi_k]\rangle \;+\; \frac{\beta\eta_t^2}{2}\mathbb{E}\big[\|\bar{g}_k\|^2 \mid \psi_k\big]. \tag{48}$$

Step 2: Two clipping facts

Write $G_k := \|\nabla L_t(\psi_k)\|$, $g_k = \nabla L_t(\psi_k) + \xi_k$ and $\bar{g}_k := \mathrm{clip}_{\tau_t}(g_k)$.

(C1) Second moment. $\|\bar{g}_k\|^2 = \min\{\|g_k\|^2, \tau_t^2\}$. Since $\|g_k\|^2 \le 2G_k^2 + 2\|\xi_k\|^2$ and $\min\{a+b, c\} \le a + \min\{b, c\}$ for $a \ge 0$,

$$\|\bar{g}_k\|^2 \;\le\; 2G_k^2 + \min\{2\|\xi_k\|^2, \tau_t^2\} \;\le\; 2G_k^2 + (2\|\xi_k\|^2)^{\alpha/2}\,\tau_t^{2-\alpha},$$

using $\min\{a, b\} \le a^\theta b^{1-\theta}$ with $\theta = \alpha/2$. Taking conditional expectations,

$$\mathbb{E}\big[\|\bar{g}_k\|^2 \mid \psi_k\big] \;\le\; 2G_k^2 + 2^{\alpha/2}\,\sigma^\alpha\,\tau_t^{2-\alpha}. \tag{49}$$

(C2) Bias under Condition (G). Since $\mathbb{E}[g_k \mid \psi_k] = \nabla L_t(\psi_k)$, $\|\mathbb{E}[\bar{g}_k \mid \psi_k] - \nabla L_t(\psi_k)\| \le \mathbb{E}[(\|g_k\| - \tau_t)\mathbf{1}\{\|g_k\| > \tau_t\} \mid \psi_k]$. On $\{\|g_k\| > \tau_t\}$ and with $G_k \le \tau_t/2$ we have $\|\xi_k\| \ge \|g_k\| - G_k > \tau_t/2$ and $\|g_k\| - \tau_t \le \|\xi_k\| - \tau_t/2 \le \|\xi_k\|$, so

$$\|\mathbb{E}[\bar{g}_k \mid \psi_k] - \nabla L_t(\psi_k)\| \;\le\; \mathbb{E}\big[\|\xi_k\|\,\mathbf{1}\{\|\xi_k\| > \tau_t/2\}\big] \;\le\; \frac{\mathbb{E}\,\|\xi_k\|^\alpha}{(\tau_t/2)^{\alpha-1}} \;\le\; 2^{\alpha-1}\,\sigma^\alpha\,\tau_t^{1-\alpha} \;=:\; b_t. \tag{50}$$

Step 3: One step

By β-smoothness, $\Delta_{k+1} \le \Delta_k - \eta_t\langle \nabla L_t(\psi_k), \bar{g}_k\rangle + \frac{\beta\eta_t^2}{2}\|\bar{g}_k\|^2$. Conditioning on ψ_k, (50) and Young's inequality $ab \le a^2/4 + b^2$ give $-\eta_t\langle \nabla L_t(\psi_k), \mathbb{E}\bar{g}_k\rangle \le -\eta_t G_k^2 + \eta_t G_k b_t \le -\frac{3}{4}\eta_t G_k^2 + \eta_t b_t^2$, and (49) with $\beta\eta_t \le 1/4$ gives $\frac{\beta\eta_t^2}{2}\mathbb{E}\,\|\bar{g}_k\|^2 \le \frac{1}{4}\eta_t G_k^2 + 2^{\alpha/2-1}\beta\eta_t^2\sigma^\alpha\tau_t^{2-\alpha}$. Hence, using $G_k^2 \ge 2\mu\Delta_k$ (strong convexity),

$$\mathbb{E}[\Delta_{k+1} \mid \psi_k] \;\le\; (1 - \mu\eta_t)\,\Delta_k \;+\; \eta_t\,4^{\alpha-1}\sigma^{2\alpha}\tau_t^{2-2\alpha} \;+\; 2^{\alpha/2-1}\,\beta\eta_t^2\,\sigma^\alpha\tau_t^{2-\alpha}. \tag{51}$$

Step 4: Iterate

With constant $\eta_t \le 1/(4\beta)$, so that $\mu\eta_t \le 1/4$,

$$\mathbb{E}\,\Delta_{K_t} \;\le\; (1 - \mu\eta_t)^{K_t}\Delta_0 \;+\; \frac{4^{\alpha-1}}{\mu}\,\sigma^{2\alpha}\tau_t^{2-2\alpha} \;+\; 2^{\alpha/2-1}\,\frac{\beta\eta_t}{\mu}\,\sigma^\alpha\tau_t^{2-\alpha}.$$

Step 5: The schedule

The two residuals balance at $\tau_t = \sigma(\beta\eta_t)^{-1/\alpha}$, where each equals a constant times $(\sigma^2/\mu)(\beta\eta_t)^{2(\alpha-1)/\alpha}$. Taking $\beta\eta_t = K_t^{-1/2}$, that is $\eta_t = \min\{1/(4\beta), 1/(\beta\sqrt{K_t})\}$ (the minimum selects the second term for $K_t \ge 16$) and $\tau_t = \sigma K_t^{1/(2\alpha)}$,

$$\mathbb{E}\,\Delta_{K_t} \;\le\; \Delta_0\,e^{-(\mu/\beta)\sqrt{K_t}} \;+\; \big(4^{\alpha-1} + 2^{\alpha/2-1}\big)\,\frac{\sigma^2}{\mu}\,K_t^{-(\alpha-1)/\alpha}.$$

Finally $\sup_K K^p e^{-(\mu/\beta)\sqrt{K}} = (2p\beta/\mu)^{2p}e^{-2p}$ for $p = (\alpha-1)/\alpha$, which gives the second form of (8). □

Reading the rate

The rate $K_t^{-(\alpha-1)/\alpha}$ interpolates between $K_t^{-1/2}$ at $\alpha = 2$ and no decay as $\alpha \to 1^+$; it is conservative for the strongly convex case, where the function-value rate under clipping is $K_t^{-2(\alpha-1)/\alpha}$ up to logarithms. The prefactor has two dimensionful terms and no others: the initial suboptimality Δ_t, whose contribution decays as $e^{-(\mu/\beta)\sqrt{K_t}}$, and the noise-to-curvature ratio σ^2/μ. The constant depends on α and on β/μ. The clipping level $\tau_t = \sigma K_t^{1/(2\alpha)}$ grows with the budget, so clipping acts on the tail only; the released code records the fraction of clipped steps per run.

G.2 Non-convex Extension (Corollary G.2)

We state the corollary first, then prove it.

Corollary G.2 (Non-convex / PŁ rate)**.** *Under Assumptions A.1 and 4.1 without strong convexity but with the PŁ inequality* $\|\nabla L_t(\psi)\|^2 \geq 2\mu_t(L_t(\psi) - \min L_t)$*, clipped SGD with the schedule* $\eta_t = (\sigma^\alpha/(\beta K_t))^{1/(\alpha+1)}$*,* $\tau_t = (\sigma^\alpha K_t)^{1/(\alpha+1)}$ *yields*

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min L_t\big] \;\leq\; \frac{C_{\mathrm{opt}}^{\mathrm{nc}}}{2\mu_t} \cdot \frac{\beta\Delta_t + \sigma^{2\alpha/(\alpha+1)}}{K_t^{(\alpha-1)/(\alpha+1)}}.$$

Corollary (Corollary G.2 restated). Under Assumption A.1 (heavy-tailed gradients) and Assumption 4.1 (β-smoothness), drop Assumption 4.2 and replace with the Polyak-Łojasiewicz condition: $\|\nabla L_t(\psi)\|^2 \geq 2\mu_t(L_t(\psi) - \min L_t)$ for all $\psi \in \Psi$. Then clipped SGD with uniform-iterate output $\widehat{\psi}_t \sim \mathrm{Unif}\{\psi_0, \ldots, \psi_{K_t-1}\}$ and the schedule

$$\eta_t \;=\; \left(\frac{\sigma^\alpha}{\beta K_t}\right)^{1/(\alpha+1)}, \qquad \tau_t \;=\; (\sigma^\alpha K_t)^{1/(\alpha+1)},$$

satisfies

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min L_t\big] \;\leq\; \frac{C_{\mathrm{opt}}^{\mathrm{nc}}}{2\mu_t} \cdot \frac{\beta\Delta_t + \sigma^{2\alpha/(\alpha+1)}}{K_t^{(\alpha-1)/(\alpha+1)}}. \tag{52}$$

Sketch. Without strong convexity, the descent inequality (47) no longer leads to a contraction in the function gap directly. Instead, the standard non-convex clipped-SGD analysis [Cutkosky and Mehta, 2021, Theorem 1] bounds the time-averaged gradient norm:

$$\frac{1}{K_t}\sum_{k=0}^{K_t-1} \mathbb{E}\,\|\nabla L_t(\psi_k)\|^2 \;\leq\; C_{\mathrm{opt}}^{\mathrm{nc}} \cdot \frac{\beta\Delta_t + \sigma^{2\alpha/(\alpha+1)}}{K_t^{(\alpha-1)/(\alpha+1)}}.$$

Selecting $\widehat{\psi}_t$ uniformly at random from the iterates makes this an upper bound on $\mathbb{E}\,\left\|\nabla L_t(\widehat{\psi}_t)\right\|^2$. Applying the Polyak-Łojasiewicz inequality $\|\nabla L_t(\psi)\|^2 \geq 2\mu_t(L_t(\psi) - \min L_t)$ and rearranging yields (52). The full non-convex clipped-SGD analysis appears in Cutkosky and Mehta [Cutkosky and Mehta, 2021, Theorem 1]; we omit the details here. □

Comparison of the two rates

Remark G.3 (Strongly convex vs. non-convex rate). At $\alpha = 2$ (the sub-Gaussian regime), the strongly convex rate is $K_t^{-1/2}$ while the non-convex / PŁ rate is $K_t^{-1/3}$. The strongly convex rate is sharper, reflecting the additional structure that strong convexity provides: a contractive recurrence on the function gap, not just a bound on the average gradient norm.

For class-IL heads with ℓ_2 regularisation (as used in SLCA, RanPAC, EASE, InfLoRA, and the experiments of §5), the strongly convex rate of Proposition 4.3 applies. For unregularised heads or for prompt-pool architectures where strong convexity is not guaranteed globally, the non-convex / PŁ rate of Corollary G.2 applies. In both cases the form of the decomposition $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ in Theorem 4.5 is preserved.

G.3 Proof of Lemma G.4 (Gradient Norm to Function Gap)

Lemma G.4 (Gradient norm to function gap)**.** *Under Assumption 4.1, if* $\mathbb{E}\left\|\nabla L_t(\widehat{\psi}_t)\right\|^2 \le \varepsilon$*, then*

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min_{\psi} L_t(\psi)\big] \;\le\; \frac{\varepsilon}{2\mu_t} + r_t,$$

where $\mu_t \ge 0$ *is the local strong-convexity constant of* L_t *near its minimiser and* $r_t \ge 0$ *vanishes when* L_t *has the Polyak–Łojasiewicz property at level* μ_t*.*

Proof. Under the Polyak–Łojasiewicz inequality at level $\mu_t > 0$,

$$L_t(\psi) - \min_{\psi} L_t(\psi) \;\le\; \frac{1}{2\mu_t}\,\|\nabla L_t(\psi)\|^2 \qquad \forall\, \psi \in \Psi,$$

holding pointwise. Taking expectations at $\psi = \widehat{\psi}_t$:

$$\mathbb{E}\big[L_t(\widehat{\psi}_t) - \min L_t\big] \;\le\; \frac{1}{2\mu_t}\,\mathbb{E}\left\|\nabla L_t(\widehat{\psi}_t)\right\|^2 \;\le\; \frac{\varepsilon}{2\mu_t}.$$

The residual r_t accommodates deviations from PŁ. For the head architectures considered (linear classifiers, prompt-pool attention with cross-entropy loss on bounded features), L_t satisfies a PŁ inequality, giving $r_t = 0$. □

G.4 Proof of Theorem 4.5 (Excess Risk Decomposition)

We restate the theorem and prove it by combining the off-diagonal block bound (Lemma 3.3, proved in §E.1) with the per-task convergence rate (Proposition 4.3, proved in §G.1).

Theorem (Theorem 4.5 restated). Consider class-IL with frozen pretrained backbone $\phi = \phi_0$ satisfying Assumption 3.2 (clause (iv) in its Gaussian branch) and the prototype-based head class (Assumption D.1), with the head trained sequentially via clipped SGD (7) with the schedule of Proposition 4.3, acting, at stage t, only on the task-t parameter block (the per-class metrics A_r for the classes of task t), which is frozen after stage t, under Assumptions A.1, 4.1, and 4.2, together with Conditions (L) and (S) as in Theorem 3.4. Task incompatibility (Assumption N.3) is not used in the proof below and is not part of the hypothesis set; it enters only the statement of Theorem 3.4. Let $\widehat{\psi}_T$ denote the head parameters after all T stages of sequential diagonal optimisation, each run for K_t iterations. Then

$$\mathbb{E}\left[\sum_{i,j}\Big|P_{ij}(\widehat{\psi}_T;\,\phi_0)\Big| \;-\; \min_{\psi}\sum_{i,j}|P_{ij}(\psi;\,\phi_0)|\right]$$
$$\le \underbrace{\sum_{t=1}^{T}\frac{C_{\mathrm{opt}}^{\mathrm{sc}}(\alpha,\beta/\mu)\cdot(\Delta_t+\sigma^2/\mu)}{K_t^{(\alpha-1)/\alpha}}}_{E_{\mathrm{CF}}} + \underbrace{\frac{MC^2\,T(T-1)}{N-1}\,\exp\Big(-\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big)}_{E_{\mathrm{TC}}}, \tag{53}$$

where E_{TC} carries the conservative exponent $1/(16\,\kappa_{\max}^{\mathrm{cls}})$ inherited from Theorem 3.4 ($1/16$ under homoscedastic classes, Corollary D.3); the operational rate $\exp(-\gamma_{\min}(\phi_0)/8)$ holds in the negligible-spread homoscedastic limit of Corollary 3.5.

Proof outline

The proof has two steps: (i) decompose the excess risk into a diagonal contribution and an off-diagonal contribution, in the spirit of the proof of Theorem 3.4 (§E.2); (ii) bound each contribution separately.

Step 1: Diagonal–off-diagonal decomposition

Let $\psi^\star \in \arg\min_\psi \sum_{i,j}|P_{ij}(\psi;\phi_0)|$ denote the globally optimal head, and let $\widehat{\psi}_T$ denote the iterate produced by sequential clipped SGD on the diagonal blocks. The excess risk is

$$\mathbb{E}\sum_{i,j}\Big|P_{ij}(\widehat{\psi}_T;\phi_0)\Big| \;-\; \sum_{i,j}|P_{ij}(\psi^\star;\phi_0)|\,.$$

Following the same decomposition as in the proof of Theorem 3.4 (§E.2, Step 2):

$$\begin{aligned}\text{LHS} &= \mathbb{E}\left[\sum_t \left|P_{tt}(\widehat{\psi}_T;\phi_0)\right| + \sum_{i\neq j}\left|P_{ij}(\widehat{\psi}_T;\phi_0)\right|\right] - \sum_t |P_{tt}(\psi^\star;\phi_0)| - \sum_{i\neq j}|P_{ij}(\psi^\star;\phi_0)| \\ &\leq \underbrace{\mathbb{E}\left[\sum_t \left|P_{tt}(\widehat{\psi}_T;\phi_0)\right|\right] - \sum_t |P_{tt}(\psi^\star;\phi_0)|}_{\text{(A):diagonal gap}} + \underbrace{\mathbb{E}\sum_{i\neq j}\left|P_{ij}(\widehat{\psi}_T;\phi_0)\right|}_{\text{(B):off-diagonal at } \widehat{\psi}_T}, \end{aligned} \tag{54}$$

where we have dropped $-\sum_{i\neq j}|P_{ij}(\psi^\star)| \leq 0$ since each off-diagonal block is non-negative.

We bound (A) and (B) separately.

Step 2(A): Bounding the diagonal gap

The CF-optimal head $\psi^\dagger$ minimises $\sum_t |P_{tt}(\psi)|$ exactly, and would satisfy $\sum_t \left|P_{tt}(\psi^\dagger;\phi_0)\right| - \sum_t |P_{tt}(\psi^\star;\phi_0)| \leq 0$. The iterate $\widehat{\psi}_T$ produced by clipped SGD does not exactly equal $\psi^\dagger$; it has finite optimisation error from the T sequential clipped-SGD stages. We bound this error.

Decompose the diagonal gap as

$$\text{(A)} = \underbrace{\sum_t \mathbb{E}\big[\left|P_{tt}(\widehat{\psi}_T;\phi_0)\right| - \min_\psi |P_{tt}(\psi;\phi_0)|\big]}_{\text{(A1)}} + \underbrace{\sum_t \min_\psi |P_{tt}(\psi;\phi_0)| - \sum_t |P_{tt}(\psi^\star;\phi_0)|}_{\text{(A2):non-positive by definition}}.$$

Term (A2) is non-positive because $\min_\psi |P_{tt}(\psi)| \leq |P_{tt}(\psi^\star)|$. We drop it.

For term (A1), we first justify the transfer of Proposition 4.3 to the final iterate. Under the training protocol, the stage-t clipped-SGD updates (7) act only on the task-t parameter block, the per-class metrics A_r for the classes of task t, and that block is frozen after stage t; the prototypes $\widehat{\mu}_r$ are computed from class-conditional features and are never updated (Step 2(B)). The intra-task loss $L_t(\psi) := |P_{tt}(\psi;\phi_0)|$ is a within-task pairwise error between classes of task t, and under Assumption D.1 it depends on ψ only through the task-t block. Hence the task-t block of $\widehat{\psi}_T$ coincides with the stage-t output $\widehat{\psi}_t$, and $L_t(\widehat{\psi}_T) = L_t(\widehat{\psi}_t)$: no stage $s > t$ modifies any parameter on which L_t depends, so the stages do not interfere and no cross-stage term arises. Proposition 4.3 (with Assumption 4.2 read in the task-t block, where L_t is non-degenerate) then gives, for each task t,

$$\mathbb{E}\big[L_t(\widehat{\psi}_T) - \min L_t\big] = \mathbb{E}\big[L_t(\widehat{\psi}_t) - \min L_t\big] \leq \frac{C^{\text{sc}}_{\text{opt}}(\alpha,\beta/\mu)\cdot(\Delta_t + \sigma^2/\mu)}{K_t^{(\alpha-1)/\alpha}},$$

where $\Delta_t := L_t(\psi_{t,0}) - \min_\psi L_t(\psi)$ is the stage-t initial suboptimality. This frozen-after-stage argument is what licenses the transfer; it fails for a shared-metric head (the LDA branch of Assumption D.1), in which a single $\widehat{A}$ is updated at every stage and every later stage modifies L_t. For that head, Proposition 4.3 bounds the per-stage output $\widehat{\psi}_t$, and the final-iterate statement of the theorem is not claimed: bounding $L_t(\widehat{\psi}_T)$ there would require an interference bound across the subsequent stages' updates of $\widehat{A}$ that we do not derive. Summing over tasks:

$$\text{(A1)} \leq \sum_{t=1}^{T} \frac{C^{\text{sc}}_{\text{opt}}(\alpha,\beta/\mu)\cdot(\Delta_t + \sigma^2/\mu)}{K_t^{(\alpha-1)/\alpha}} = E_{\text{CF}}. \tag{55}$$

Step 2(B): Bounding the off-diagonal contribution

The off-diagonal contribution at the iterate $\widehat{\psi}_T$ is bounded via Lemma 3.3. Lemma 3.3 requires Assumption D.1 (prototype-based head with metric A_r converging to Σ_r^{-1} in the QDA case or to $\Sigma_{\text{pool}}^{-1}$ in the LDA case). Under Assumption D.1, the iterate $\widehat{\psi}_T$ produced by clipped SGD remains in the prototype-based head class throughout training: the class prototypes $\widehat{\mu}_r$ are computed from class-conditional features and are not modified by the clipped-SGD updates, which act on the per-class

metric parameter A_r of the current task's block (the shared-metric LDA branch of Assumption D.1 lies outside the frozen-after-stage protocol; its per-stage status is settled at the end of Step 2(A)). Convergence of the metric parameter to its population limit is governed by Proposition 4.3 on the intra-task convergence of the head, with finite-sample deviation controlled by Lemma P.1. Lemma 3.3 therefore applies, in its pair-metric form with spread ratio ν_{ij}^M (Definition E.10) and class-to-task scatter ratio $\kappa_{ij}^{\mathrm{cls}}$ (Definition E.9), giving for each inter-task block:

$$\Big|P_{ij}(\widehat{\psi}_T;\phi_0)\Big| \;\le\; \frac{MC^2}{N-1}\,\exp\Big(-\tfrac{1-\nu_{ij}^M}{8\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big).$$

Under Condition (S') ($\nu_{ij}^M \le 1/2$; §E.4) the spread degrades the exponent to at least $1/(16\,\kappa_{ij}^{\mathrm{cls}})$ (the conservative exponent inherited from Theorem 3.4), so

$$\Big|P_{ij}(\widehat{\psi}_T;\phi_0)\Big| \;\le\; \frac{MC^2}{N-1}\,\exp\Big(-\tfrac{1}{16\,\kappa_{ij}^{\mathrm{cls}}}\,\gamma_{ij}(\phi_0)\Big) \;\le\; \frac{MC^2}{N-1}\,\exp\Big(-\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big).$$

The finite-sample deviation between $\widehat{A}$ at iterate $\widehat{\psi}_T$ and its population limit is not bounded by a separate lemma: Lemma 3.3 is a population-metric (Bayes-rule) result, and the exactness of the disjoint direct dependencies below holds at the population-metric limit, where this deviation vanishes and the correction is absorbed into the optimisation error already analysed in Step 2(A). At finite samples the empirical-metric deviation can in principle carry both (K_t, α)- and γ-dependence, so the absorption is a population-limit statement, not a finite-sample one. Taking expectations:

$$\begin{aligned}
(\mathrm{B}) \;&\le\; \sum_{i\neq j}\frac{MC^2}{N-1}\,\exp\Big(-\tfrac{1}{16\,\kappa_{\max}^{\mathrm{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big)\\
&=\; \frac{MC^2\cdot T(T-1)}{N-1}\,\exp\Big(-\tfrac{1}{16\,\kappa_{\max}^{\mathrm{cls}}(\phi_0)}\,\gamma_{\min}(\phi_0)\Big)\\
&=\; E_{\mathrm{TC}}.
\end{aligned} \tag{56}$$

Step 3: Combining

Combining (54), (55), and (56):

$$\mathbb{E}\left[\sum_{i,j}\Big|P_{ij}(\widehat{\psi}_T)\Big| - \min_{\psi}\sum_{i,j}|P_{ij}(\psi)|\right] \;\le\; E_{\mathrm{CF}} + E_{\mathrm{TC}},$$

which is (53). □

Why the orthogonality is exact at the population-metric limit

The key structural feature of (53) is that E_{CF} depends *directly* on optimisation quantities (K_t, Δ_t, σ, μ, α, β) while E_{TC} depends only on representation quantities ($\gamma_{\min}(\phi_0)$, T, C, M, N). This decomposition is structural and follows directly from the proof, read at the population-metric limit:

- Term (A1) bounds the gap between $\widehat{\psi}_T$ and the per-task minima of L_t. This is purely an optimisation error and is governed by clipped-SGD theory.
- Term (B) bounds the off-diagonal blocks at $\widehat{\psi}_T$ using Lemma 3.3, read at the population-metric limit of the head (Step 2(B)). In that limit the bound is parameter-free in ψ: it holds for *any* head in the Assumption D.1 class, including the worst-case $\widehat{\psi}_T$ produced by sequential SGD, and depends only on $\gamma_{\min}(\phi_0)$, which is a property of the pretrained backbone alone; the finite-sample empirical-metric deviation is absorbed only at this limit.

We note that the smoothness constant β implicitly depends on the bounded-feature constant B from Assumption 3.2(i), which depends on the backbone ϕ_0. Through this channel E_{CF} has an indirect dependence on the backbone, but the dependence is low-order compared to the exponential $\gamma_{\min}$-dependence of E_{TC} (verified empirically in Figure 5; see Remark 4.6).

The orthogonality is therefore not a coincidence of our particular analysis but a consequence of the structure of the loss matrix under sequential discriminative training: the diagonal blocks respond to optimisation, and the off-diagonal blocks are controlled geometrically by the feature-space task separation. This is the formal content of the diagnostic claim in §4.

Remark G.5 (Independence of ϕ_0 and α). The decomposition makes precise the intuition that "backbone quality" and "optimisation quality" are independent levers in class-IL. Doubling $\gamma_{\min}(\phi_0)$ via a stronger backbone reduces E_{TC} by a factor $\exp(-\gamma_{\min}/16)$ without affecting E_{CF}. Doubling K_t reduces E_{CF} by a factor $2^{(\alpha-1)/\alpha}$ without affecting E_{TC}. The two levers can be tuned independently.
Remark G.6 (The non-convex / PŁ case). Under the conditions of Corollary G.2 (non-convex with PŁ), the bound on (A1) becomes $\frac{C_{\mathrm{opt}}^{\mathrm{nc}}}{2\mu_t} \cdot \frac{\beta\Delta_t+\sigma^{2\alpha/(\alpha+1)}}{K_t^{(\alpha-1)/(\alpha+1)}}$, yielding the slower rate $K_t^{-(\alpha-1)/(\alpha+1)}$ in E_{CF}. The form of the decomposition and the bound on E_{TC} are unchanged.

H Proofs from Section 4 (Representation Drift)

H.1 Proof of Theorem B.2

We restate the theorem and prove it in three steps.
Theorem (Theorem B.2 restated). Under Assumptions 3.2, A.1, and B.1, with clipped-SGD backbone fine-tuning,

$$\gamma_{\min}(\phi_t) \;\geq\; \gamma_{\min}(\phi_0) \;-\; c_2\, L_h^2\, \eta_\phi\, \tau_\phi \cdot \frac{\left\|\mu^{(i^\star)}(\phi_0) - \mu^{(j^\star)}(\phi_0)\right\|}{\sigma_{\min}^2} \cdot \sum_{s=1}^{t} K_s.$$

Step 1: Parameter drift bound

The clipped-SGD update for the backbone has the form

$$\phi_{k+1} \;=\; \phi_k - \eta_\phi \cdot \mathrm{clip}_{\tau_\phi}\big(\widehat{\nabla L_t}(\phi_k)\big).$$

The clipped gradient has Frobenius norm at most τ_ϕ by construction of the clipping operator. Therefore each iteration shifts the parameter by at most $\eta_\phi\tau_\phi$:

$$\|\phi_{k+1} - \phi_k\|_F \;\leq\; \eta_\phi\, \tau_\phi. \tag{57}$$

After K_t iterations on stage t, by the triangle inequality,

$$\|\phi_t - \phi_{t-1}\|_F \;\leq\; \eta_\phi\, \tau_\phi\, K_t. \tag{58}$$

After t stages, summing over stages,

$$\|\phi_t - \phi_0\|_F \;\leq\; \eta_\phi\, \tau_\phi \sum_{s=1}^{t} K_s. \tag{59}$$

Step 2: Parameter drift to feature drift

By Assumption B.1, for any $x \in \mathcal{X}$,

$$\|h_{\phi_t}(x) - h_{\phi_0}(x)\| \;\leq\; L_h \cdot \|\phi_t - \phi_0\|_F \;\leq\; L_h \cdot \eta_\phi\, \tau_\phi \sum_{s=1}^{t} K_s. \tag{60}$$

This pointwise bound translates to bounds on feature class-conditional means and covariances. For each class r:

$$\begin{aligned} \|\mu_r(\phi_t) - \mu_r(\phi_0)\| \;&=\; \left\|\mathbb{E}\big[h_{\phi_t}(X) - h_{\phi_0}(X) \,\big|\, Y = r\big]\right\| \\ &\leq\; \mathbb{E}\big[\|h_{\phi_t}(X) - h_{\phi_0}(X)\| \,\big|\, Y = r\big] \\ &\leq\; L_h \cdot \eta_\phi\, \tau_\phi \sum_{s=1}^{t} K_s, \end{aligned} \tag{61}$$

where the second line is Jensen's inequality.

For task centroids:

$$\left\|\mu^{(i)}(\phi_t) - \mu^{(i)}(\phi_0)\right\| \;\leq\; L_h \cdot \eta_\phi\, \tau_\phi \sum_{s=1}^{t} K_s, \tag{62}$$

since $\mu^{(i)}$ is an average of C class means, each bounded by (61).

For task covariance, the perturbation is at most quadratic in the feature drift. Specifically, if $\delta_r := \mu_r(\phi_t) - \mu_r(\phi_0)$ and $\Delta_r(x) := h_{\phi_t}(x) - h_{\phi_0}(x)$ for class r, then

$$\Sigma_r(\phi_t) - \Sigma_r(\phi_0) = \mathbb{E}\big[\Delta_r(X)\Delta_r(X)^\top \,\big|\, Y = r\big] + \mathbb{E}\big[\Delta_r(X)(h_{\phi_0}(X) - \mu_r(\phi_0))^\top + \cdots \,\big|\, Y = r\big] - \delta_r \delta_r^\top.$$

The dominant term in the operator norm is the cross-product $\mathbb{E}[\Delta_r(h_{\phi_0} - \mu_r)^\top]$, which is bounded by $L_h \cdot \eta_\phi \, \tau_\phi \sum_s K_s \cdot \sqrt{\mathrm{tr}(\Sigma_r(\phi_0))}$ via Cauchy-Schwarz. We therefore have

$$\|\Sigma_r(\phi_t) - \Sigma_r(\phi_0)\|_{\mathrm{op}} \;\le\; c_3 \cdot L_h \cdot \eta_\phi \, \tau_\phi \sum_{s=1}^{t} K_s \cdot \sqrt{\mathrm{tr}(\Sigma_r(\phi_0))}, \tag{63}$$

for a universal constant c_3. Aggregating over the C classes in task i:

$$\Big\|\Sigma^{(i)}(\phi_t) - \Sigma^{(i)}(\phi_0)\Big\|_{\mathrm{op}} \;\le\; c_3 \cdot L_h \cdot \eta_\phi \, \tau_\phi \sum_{s=1}^{t} K_s \cdot \sqrt{\mathrm{tr}(\Sigma^{(i)}(\phi_0))}. \tag{64}$$

Step 3: Feature drift to Mahalanobis margin drift

The Mahalanobis margin between tasks i and j is

$$\gamma_{ij}(\phi) \;=\; (\mu^{(i)}(\phi) - \mu^{(j)}(\phi))^\top \bigg(\frac{\Sigma^{(i)}(\phi) + \Sigma^{(j)}(\phi)}{2}\bigg)^{-1} (\mu^{(i)}(\phi) - \mu^{(j)}(\phi)).$$

We need to bound $\gamma_{ij}(\phi_0) - \gamma_{ij}(\phi_t)$ in terms of the mean and covariance perturbations from Step 2.

Write $\Delta\mu := \mu^{(i)}(\phi_0) - \mu^{(j)}(\phi_0)$, $\Delta\mu_t := \mu^{(i)}(\phi_t) - \mu^{(j)}(\phi_t)$, $\bar{\Sigma} := (\Sigma^{(i)}(\phi_0) + \Sigma^{(j)}(\phi_0))/2$, $\bar{\Sigma}_t := (\Sigma^{(i)}(\phi_t) + \Sigma^{(j)}(\phi_t))/2$.

Then

$$\begin{aligned}\gamma_{ij}(\phi_0) - \gamma_{ij}(\phi_t) \;&=\; \Delta\mu^\top \bar{\Sigma}^{-1} \Delta\mu - \Delta\mu_t^\top \bar{\Sigma}_t^{-1} \Delta\mu_t \\ &=\; \Delta\mu^\top (\bar{\Sigma}^{-1} - \bar{\Sigma}_t^{-1}) \Delta\mu + (\Delta\mu - \Delta\mu_t)^\top \bar{\Sigma}_t^{-1} (\Delta\mu + \Delta\mu_t).\end{aligned} \tag{65}$$

We bound each term separately.

Mean term. Using $\|\Delta\mu\|, \|\Delta\mu_t\| \le \|\Delta\mu\| + 2L_h\eta_\phi\tau_\phi \sum_s K_s$ from (62) applied to both tasks i and j, and the fact that $\big\|\bar{\Sigma}_t^{-1}\big\|_{\mathrm{op}} \le 1/\sigma_{\min}^2$ under Assumption 3.2(ii) and the perturbation bound on $\bar{\Sigma}_t$,

$$\big|(\Delta\mu - \Delta\mu_t)^\top \bar{\Sigma}_t^{-1} (\Delta\mu + \Delta\mu_t)\big| \;\le\; c_4 \cdot \frac{L_h \eta_\phi \tau_\phi \sum_s K_s}{\sigma_{\min}^2} \cdot \|\Delta\mu\|. \tag{66}$$

Covariance term. By the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$,

$$\bar{\Sigma}^{-1} - \bar{\Sigma}_t^{-1} \;=\; \bar{\Sigma}^{-1}(\bar{\Sigma}_t - \bar{\Sigma})\bar{\Sigma}_t^{-1}.$$

Using $\big\|\bar{\Sigma}^{-1}\big\|_{\mathrm{op}}, \big\|\bar{\Sigma}_t^{-1}\big\|_{\mathrm{op}} \le 1/\sigma_{\min}^2$ and the covariance perturbation bound (64),

$$\big|\Delta\mu^\top (\bar{\Sigma}^{-1} - \bar{\Sigma}_t^{-1}) \Delta\mu\big| \;\le\; \frac{\|\Delta\mu\|^2}{\sigma_{\min}^4} \cdot c_3 L_h \eta_\phi \tau_\phi \sum_s K_s \cdot \sqrt{\mathrm{tr}(\bar{\Sigma})}. \tag{67}$$

Under Assumption 3.2(i) (bounded features), $\sqrt{\mathrm{tr}(\bar{\Sigma})} \le \sqrt{p} \cdot B$, which absorbs into the universal constant.

Combining. Compare the two contributions: the mean term (66) scales *linearly* in $\|\Delta\mu\|$, while the covariance term (67) scales *quadratically* in $\|\Delta\mu\|$. For the worst-case task pair $(i^\star, j^\star)$, where $\|\Delta\mu^\star\|$ is the smallest inter-task centroid distance and is therefore small relative to typical inter-task distances, the linear-in-$\|\Delta\mu\|$ term dominates the quadratic-in-$\|\Delta\mu\|$ term. Combining,

$$\gamma_{ij^\star}(\phi_0) - \gamma_{i^\star j^\star}(\phi_t) \;\le\; c_2 \cdot L_h^2 \cdot \eta_\phi \, \tau_\phi \sum_{s=1}^{t} K_s \cdot \frac{\|\Delta\mu^\star\|}{\sigma_{\min}^2}, \tag{68}$$

where the additional factor of L_h in the prefactor (giving L_h^2, not L_h) comes from the cross-product term in the covariance perturbation: the feature drift enters both the mean shift and the centroid spread, so the perturbation is quadratic in $L_h \eta_\phi \tau_\phi \sum_s K_s$ at second order.

Taking the minimum over task pairs:

$$\gamma_{\min}(\phi_t) \;\geq\; \gamma_{\min}(\phi_0) \;-\; c_2 \cdot L_h^2 \cdot \eta_\phi \, \tau_\phi \sum_{s=1}^{t} K_s \cdot \frac{\|\Delta\mu^\star\|}{\sigma_{\min}^2},$$

which is the bound stated. □

H.2 Proof of Corollary B.3 (Critical Horizon)

We restate the corollary.

Corollary (Corollary B.3 restated). Under the conditions of Theorem B.2 with uniform per-stage iteration budget $K_s = K$ for all s, the critical horizon

$$t^\star \;:=\; \max\{t : \gamma_{\min}(\phi_t) \geq \gamma_{\min}(\phi_0)/2\}$$

satisfies

$$t^\star \;=\; \Omega\left(\frac{\gamma_{\min}(\phi_0)\, \sigma_{\min}^2}{L_h^2 \, \eta_\phi \, \tau_\phi \, K \, \left\|\mu^{(i^\star)}(\phi_0) - \mu^{(j^\star)}(\phi_0)\right\|}\right).$$

Proof. By Theorem B.2, after t stages with uniform budget K per stage,

$$\gamma_{\min}(\phi_t) \;\geq\; \gamma_{\min}(\phi_0) \;-\; c_2 \, L_h^2 \, \eta_\phi \, \tau_\phi \, K \, t \cdot \frac{\|\Delta\mu^\star\|}{\sigma_{\min}^2},$$

where $\|\Delta\mu^\star\| := \left\|\mu^{(i^\star)}(\phi_0) - \mu^{(j^\star)}(\phi_0)\right\|$. The condition $\gamma_{\min}(\phi_t) \geq \gamma_{\min}(\phi_0)/2$ is implied by

$$c_2 \, L_h^2 \, \eta_\phi \, \tau_\phi \, K \, t \cdot \frac{\|\Delta\mu^\star\|}{\sigma_{\min}^2} \;\leq\; \gamma_{\min}(\phi_0)/2.$$

Solving for t:

$$t \;\leq\; \frac{\gamma_{\min}(\phi_0)\, \sigma_{\min}^2}{2\, c_2 \, L_h^2 \, \eta_\phi \, \tau_\phi \, K \, \|\Delta\mu^\star\|}.$$

Therefore $t^\star$ is at least the right-hand side, which yields the stated $\Omega(\cdot)$ bound (with the factor $1/(2c_2)$ absorbed into the implicit constant). □

I Experimental Details

This appendix specifies the experimental protocols used to verify the three theorems of §5. We document the data-independent procedures here; dataset specifications, backbone specifications, and the actual numerical results appear in §K.

Compute budget. All experiments were run on NVIDIA A100 (40 GB) hardware, totalling approximately 240 GPU-hours: $\widehat{\gamma}_{\min}$ measurements (24 GPU-hours), empirical TC-gap experiments (45 GPU-hours), $\widehat{\alpha}$ Hill-estimator measurements (18 GPU-hours), the 14-method comparison of Table 8 (120 GPU-hours), and representation-drift trajectories (32 GPU-hours).

I.1 Computing $\gamma_{\min}(\phi_0)$

The Mahalanobis margin $\gamma_{\min}(\phi_0)$ is the central diagnostic quantity of the paper. We compute it via the following procedure, which uses only the training features and runs in a single forward pass through the data.

Procedure. Given a frozen pretrained backbone $h_{\phi_0} : \mathcal{X} \to \mathcal{Z}$ with output dimension p, a dataset partitioned into tasks $\{\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(T)}\}$ with C classes per task, and training samples $\{(x_n, y_n)\}_{n=1}^{N_{\text{train}}}$:

1. **Forward pass.** Compute features $z_n := h_{\phi_0}(x_n) \in \mathbb{R}^p$ for all training samples. Apply layer normalisation if the backbone's output is not already normalised: $\tilde{z}_n := (z_n - \mu_z)/s_z$ where μ_z, s_z are the global feature mean and standard deviation. The feature normalisation step is critical for making $\gamma_{\min}$ invariant to backbone scaling conventions.
2. **Per-class statistics.** For each class $r \in \{1, \ldots, N\}$, compute the empirical class mean

$$\widehat{\mu}_r \;:=\; \frac{1}{|S_r|} \sum_{n \in S_r} \tilde{z}_n, \qquad S_r := \{n : y_n = r\}.$$

3. **Pooled covariance with shrinkage.** Compute the Ledoit-Wolf shrinkage covariance estimator on the centred features:

$$\widehat{\Sigma} \;:=\; (1 - \lambda^\star)\, S \;+\; \lambda^\star\, \frac{\operatorname{tr}(S)}{p}\, I,$$

   where $S := \frac{1}{N_{\text{train}} - N} \sum_n (\tilde{z}_n - \widehat{\mu}_{y_n})(\tilde{z}_n - \widehat{\mu}_{y_n})^\top$ is the pooled within-class scatter, and $\lambda^\star$ is the Ledoit-Wolf optimal shrinkage parameter computed in closed form from the data Ledoit and Wolf [2004]. Shrinkage is necessary because the unregularised pooled covariance is rank-deficient when $N_{\text{train}} < p$, which is common for high-dimensional foundation-model features.
4. **Task centroids.** For each task t,

$$\widehat{\mu}^{(t)} \;:=\; \frac{1}{C} \sum_{r \in \mathcal{Y}^{(t)}} \widehat{\mu}_r.$$

5. **Pairwise margins.** For each task pair $i \neq j$:

$$\widehat{\gamma}_{ij}(\phi_0) \;:=\; (\widehat{\mu}^{(i)} - \widehat{\mu}^{(j)})^\top \widehat{\Sigma}^{-1} (\widehat{\mu}^{(i)} - \widehat{\mu}^{(j)}).$$

   Use the Cholesky factorisation $\widehat{\Sigma} = LL^\top$ to compute $\widehat{\Sigma}^{-1} v$ via two triangular solves, without forming the inverse explicitly.
6. **Worst-case margin.** Output $\widehat{\gamma}_{\min}(\phi_0) := \min_{i \neq j} \widehat{\gamma}_{ij}(\phi_0)$.

Computational cost. The dominant cost is the forward pass, which is one pass through the training set; the covariance estimation and matrix solves are $O(p^2 N_{\text{train}}) + O(p^3)$, negligible relative to the forward pass for $p = 768$ (ViT-B/16 feature dimension) and typical N_{train}. The full computation takes under five minutes on a single GPU for the benchmarks of §5.

Implementation notes. We use the closed-form Ledoit-Wolf estimator from `scikit-learn` Ledoit and Wolf [2004]; if the shrinkage parameter $\lambda^\star$ is reported below 0.05 (suggesting the unregularised covariance is well-conditioned), we cross-check using the unshrunken estimator and confirm the result agrees within 5%. Class means are computed with double-precision arithmetic to avoid numerical drift; the final γ_{ij} values are stable to single-precision.

I.2 Empirical TC Gap Computation

The empirical TC gap is the difference between two reference quantities: the CF-optimal head's risk (from joint training on diagonal blocks only) and the globally optimal head's risk (from joint training on all blocks).

Procedure. For each $(\text{backbone}, \text{dataset})$ configuration:

1. **Diagonal-only joint training.** Train a linear head with ℓ_2 regularisation by minimising the diagonal-only objective: for each task, sample only within-task pairs (x, y) with $y \in \mathcal{Y}^{(t)}$ and compute the cross-entropy loss restricted to the C classes of that task. Train jointly across all tasks with task-balanced sampling. The resulting head $\widehat{\psi}^\dagger$ approximates the CF-optimal head.

2. **All-blocks joint training.** Train an identical linear head on the standard N-way cross-entropy objective (all classes simultaneously). The resulting head $\widehat{\psi}^\star$ approximates the globally optimal head.
3. **Risk evaluation.** Compute the test loss of each head: $L(\widehat{\psi}^\dagger), L(\widehat{\psi}^\star)$ on the held-out test set, with the same evaluation protocol for both.
4. **TC gap.** Report $\widehat{\mathrm{TC\ gap}} := L(\widehat{\psi}^\dagger) - L(\widehat{\psi}^\star)$.

Both heads use identical architectures, identical optimiser (AdamW with weight decay 10^{-4}), identical training budget (20 epochs at learning rate 10^{-3}, sufficient to reach convergence on both objectives), and identical batch sizes. The only difference between the two procedures is the sampling distribution: diagonal-only restricts each batch to within-task pairs, all-blocks samples uniformly across classes. This isolates the off-diagonal contribution and makes $\widehat{\mathrm{TC\ gap}}$ a faithful empirical analogue of the theoretical $\mathrm{TC\ gap}(\phi_0)$ in Theorem 3.4.

Why a CF-optimal reference instead of a real class-IL method. The "CF-optimal head" in Theorem 3.4 is defined as the head that minimises the diagonal sum exactly. Real class-IL methods (replay, regularisation, prompt-pool) approach this target with finite optimisation error. We use the joint diagonal-only training as the operational definition of the CF-optimal reference; it is the cleanest empirical analogue of $\psi^\dagger$ in the proof of Theorem 3.4.

I.3 Tail-Index Estimation

The tail index α governs the per-task convergence rate in Proposition 4.3 and the heavy-tailed concentration rate in Lemma P.1. We estimate α from gradient norm statistics during training using the Hill estimator Hill [1975].

Procedure specified in advance. The estimator is sensitive to the choice of tail fraction. To prevent post-hoc tuning, we specify the following procedure in advance, fixing the estimator and its hyperparameters *before* running any of the convergence-verification experiments of §5:

1. **Sample collection.** Run a warm-up training run of $K_{\mathrm{warm}} = 1000$ iterations of vanilla SGD (no clipping) on the full N-way cross-entropy objective. At each iteration, record the gradient norm $g_k := \left\| \widehat{\nabla} L_t(\psi_k) \right\|$.
2. **Tail fraction.** Use the top $k = \lceil 0.05 \cdot K_{\mathrm{warm}} \rceil = 50$ largest gradient norms (top 5% of the sample). The choice $k = 5\%$ follows the standard recommendation of Resnick [2007] and is fixed before any data is collected.
3. **Hill estimator.** Order the recorded gradient norms as $g_{(1)} \geq g_{(2)} \geq \cdots$. The Hill estimator of the tail index is
$$\widehat{\alpha}_{\mathrm{Hill}} := \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{g_{(i)}}{g_{(k+1)}} \right)^{-1}.$$
4. **Bootstrap confidence interval.** Resample $g_{(1)}, \ldots, g_{(k+1)}$ with replacement 1000 times, recompute $\widehat{\alpha}_{\mathrm{Hill}}$ each time, and report the 95% percentile interval as the confidence interval on α.

Validity range and reporting. The Hill estimator is consistent for $\alpha \in (1, \infty)$ when the underlying distribution has Pareto-like tails. Estimates yielding $\widehat{\alpha} < 1.05$ or $\widehat{\alpha} > 2.5$ are flagged: the former indicates extreme heavy-tailedness with poorly defined first moment (beyond Assumption A.1), and the latter is indistinguishable from the sub-Gaussian regime where clipping provides no benefit. We report the bootstrap mean and 95% CI for each (backbone, dataset) configuration in §K.

Cross-validation against alternative estimators. For a subset of configurations we additionally compute the *maximum-likelihood* tail index estimator and the *moments* estimator Resnick [2007] as cross-checks. Agreement within ± 0.2 across estimators is required for the configuration to enter the headline analysis; configurations showing wider disagreement are flagged for further investigation.

I.4 Gaussianity Check for Class-Conditional Features

The constant $1/8$ in Lemma 3.3 is exact for Gaussian class-conditional features. Remark E.3 of §E.1 notes that for sub-Gaussian features an exponent of the same order is expected, an extension that remains open. This appendix verifies that, after layer normalisation, foundation-model features pass standard Gaussianity tests on a per-class basis, justifying the operational constant $1/8$.

Test procedure. For each configuration $(\text{backbone}, \text{dataset})$:

1. **Per-class feature samples.** Extract layer-normalised features $\tilde{z}_n$ as in §I.1. For each class c, collect the within-class samples $\{\tilde{z}_n : y_n = c\}$.
2. **Dimension reduction.** Project each per-class sample to its top-d principal components ($d = 50$, capturing typically $> 95\%$ of within-class variance for foundation-model features). Mardia's test of multivariate normality Mardia [1970] is unstable in very high dimensions, so dimension reduction is required.
3. **Mardia's test.** Compute Mardia's skewness $b_{1,d}$ and kurtosis $b_{2,d}$ statistics on the reduced sample. Under the null of multivariate normality:

$$(n/6)b_{1,d} \sim \chi^2_{d(d+1)(d+2)/6}, \qquad \sqrt{\frac{n}{8d(d+2)}}\,(b_{2,d} - d(d+2)) \sim \mathcal{N}(0,1),$$

   where n is the within-class sample size. We compute p-values for each test on each class.
4. **Aggregation.** For each $(\text{backbone}, \text{dataset})$ configuration, report the fraction of classes whose features pass both tests at the $\alpha = 0.05$ level (Benjamini-Hochberg-corrected for multiple testing across classes).

Acceptance criterion for the operational constant $1/8$. The operational regime of Lemma 3.3 requires both (i) Condition (S') of §E.4 (the within-task spread condition $\nu_{ij}^M \leq 1/2$ in the pair metric, sufficient trace form (37); verified per-configuration in Table 4), and (ii) approximately Gaussian class-conditional features. The Gaussianity check in this section provides empirical support for (ii); per-configuration counts for (i) are not reported; the printed columns of Table 4 carry the isotropic-metric values, and the pair-metric outcome is analysed under the boundary regime of §E.3. Configurations satisfying both criteria are treated as operational; configurations failing either default to the boundary regime, where the operational-rate corollary (Corollary 3.5) applies with the interpolating exponent $(1 - \nu^M)/(8\,\kappa^{\text{cls}})$.

Empirical findings (preview, full results in §K). For the foundation-model backbones used in §5 (ViT-IN21K, DINOv2, CLIP) on the four standard benchmarks, after layer normalisation, the fraction of classes passing both Mardia tests is typically in the range 75–95%, with DINOv2 features generally most Gaussian and ViT-IN21K features showing mild kurtosis excess in approximately 15% of classes. The operational constant $1/8$ is justified for the configurations entering the headline analysis.

Caveat: orthogonal-complement Gaussianity. Mardia's test on the top-50 principal components verifies low-dimensional Gaussianity but does not exclude non-Gaussianity in the orthogonal complement (the remaining ~ 718 feature dimensions, capturing $\lesssim 5\%$ of within-class variance). A pass at $d = 50$ is consistent with arbitrary non-Gaussianity in the tail directions; the Bhattacharyya bound's robustness to such tail non-Gaussianity is examined empirically by the measured slope in §5 but is not formally established here. We treat the Gaussianity check as necessary but not sufficient evidence for the operational constant.

J Algorithm Summary

The diagnostic-guided clipped-SGD procedure of §5 converts the theoretical results of the paper into a single algorithm. Inputs are a frozen pretrained backbone, a partition of classes into tasks, and the standard class-IL training stream. Outputs are a sequence of head parameters and an optional backbone-freeze decision.

Algorithm 1 Diagnostic-guided clipped-SGD for class-IL

Require: Pretrained backbone ϕ_0, class partition $\{\mathcal{Y}^{(t)}\}_{t=1}^{T}$, task data streams $\{D_t\}_{t=1}^{T}$, per-task budget $\{K_t\}_{t=1}^{T}$, target failure probability δ

1: **Stage 0 – Diagnostic measurement on ϕ_0:**
2: Compute $\widehat{\gamma}_{\min}(\phi_0)$ via the procedure of §I.1
3: Run $K_{\mathrm{warm}} = 1000$ iterations of vanilla SGD on a warm-up batch; record gradient norms $\{g_k\}_{k=1}^{K_{\mathrm{warm}}}$ and the final warm-up iterate ψ_{warm}
4: Estimate tail index $\widehat{\alpha}$ via the Hill procedure of §I.3
5: Estimate smoothness constant $\widehat{\beta} \leftarrow \frac{1}{2}\lambda_{\max}(\widehat{M}_2) + \lambda$, where $\widehat{M}_2$ is the second-moment matrix of the augmented features $[\phi_0(x); 1]$ on the warm-up batch (a rigorous bound: the softmax Hessian in the logits is $\mathrm{diag}(p) - pp^\top \preceq I/2$)
6: Estimate strong-convexity constant $\widehat{\mu}$ via finite-difference probing of the local Hessian at ψ_{warm} along 10 random directions; take $\widehat{\mu} \leftarrow \mathrm{median}$ of the curvature estimates
7: Estimate noise scale $\widehat{\sigma}$ via $(K_{\mathrm{warm}})^{-1}\sum_k g_k^{\widehat{\alpha}}$
8: **Stage 0 – Hyperparameter selection:**
9: Set $\eta_t \leftarrow \min\{1/(4\widehat{\beta}),\ 1/(\widehat{\beta}\sqrt{K_t})\}$ per Proposition 4.3
10: Set $\tau_t \leftarrow \widehat{\sigma}\, K_t^{1/(2\widehat{\alpha})}$
11: Initialise head ψ_0 randomly; freeze backbone $\phi \leftarrow \phi_0$ initially
12: **for** $t = 1, \ldots, T$ **do**
13: **Stage t – Head training:**
14: **for** $k = 0, \ldots, K_t - 1$ **do**
15: Sample mini-batch from D_t
16: Compute stochastic gradient g_k
17: Clip: $\bar{g}_k \leftarrow g_k \cdot \min\{1, \tau_t / \|g_k\|\}$
18: Update head: $\psi_{k+1} \leftarrow \psi_k - \eta_t \bar{g}_k$
19: **end for**
20: **Stage t – Optional backbone monitoring (if fine-tuning):**
21: **if** backbone is being fine-tuned **then**
22: Compute $\widehat{\gamma}_{\min}(\phi_t)$
23: **if** $\widehat{\gamma}_{\min}(\phi_t) < \widehat{\gamma}_{\min}(\phi_0)/2$ **then**
24: Freeze backbone for all subsequent stages
25: (*Critical horizon reached – Corollary B.3*)
26: **end if**
27: **end if**
28: **end for**
29: **return** final head $\widehat{\psi}_T$

What the algorithm achieves. By Theorem 4.5 (with the $\frac{1}{8}$ exponent read in the negligible-spread homoscedastic limit of Corollary 3.5; under the theorem's full hypothesis set the exponent is $\frac{1}{16\,\kappa_{\max}^{\mathrm{cls}}}$ ($\frac{1}{16}$ under homoscedastic classes, Corollary D.3), and the count of configurations meeting the strict operational condition ($\nu^M \le 1/2$ in the pair metric) is not reported (Table 4 carries the isotropic-metric values), the head produced by Algorithm 1 achieves excess risk

$$E_{\mathrm{CF}} + E_{\mathrm{TC}} \;=\; O\big(K^{-(\widehat{\alpha}-1)/\widehat{\alpha}}\big) \;+\; O\big(\exp(-\widehat{\gamma}_{\min}/8)\big).$$

The first term is controlled by the per-task budget K_t and the heavy-tailedness $\widehat{\alpha}$; the second is controlled by the pretrained backbone quality $\widehat{\gamma}_{\min}(\phi_0)$. The optional backbone-monitoring step (lines 18–22) protects against the critical-horizon failure mode of §4: if backbone fine-tuning is degrading $\widehat{\gamma}_{\min}$, freezing recovers the operational regime.

Theory-derived schedule with estimated noise and curvature constants. The schedule η_t, τ_t is determined by the theory directly once the gradient tail index $\widehat{\alpha}$, smoothness constant $\widehat{\beta}$, and strong-convexity / PŁ constant $\widehat{\mu}$ are measured. We estimate these from the warm-up run of §I.3: $\widehat{\beta}$ as the maximum gradient norm during Stage 0 (a noisy upper-bound estimator that is itself parameterised by warm-up length and batch size); $\widehat{\mu}$ from the local Hessian at the end of Stage 1 via finite-difference probing along 10 random directions (with the finite-difference step size and direction count themselves hyperparameters of the estimator), with $\widehat{\mu}$ taken as the median of the resulting

curvature estimates. The hyperparameter disclaimer is stated in §5, where the algorithm is introduced: we do not claim hyperparameter-free training in the strong sense. The only practitioner-set parameter of the schedule itself (K_t, the per-stage iteration budget) governs the trade-off between E_{CF} and computational cost. The diagnostic $\widehat{\gamma}_{\min}(\phi_0)$ informs the practitioner whether further investment in K_t is worthwhile: when $\widehat{\gamma}_{\min}$ is large, E_{TC} is exponentially small, and increasing K_t pays off; when $\widehat{\gamma}_{\min}$ is small, E_{TC} dominates, and the recommendation is to upgrade the backbone instead.

K Full Experimental Results

This appendix reports the complete experimental results referenced in §5. Subsections cover, in order: §K.1 measures $\gamma_{\min}(\phi_0)$ across configurations; §K.2 verifies Condition (S'), the within-task spread condition of §E.4; §K.3 reports the empirical slope of $\log(\text{TC gap})$ vs. $\gamma_{\min}$ verifying Theorem 3.4; §K.4 reports E_{CF} convergence rates verifying Theorem 4.5; §K.5 reports drift trajectories and critical horizons verifying Theorem B.2 and Corollary B.3; §K.6 reports the headline cross-method comparison; §K.7 reports sensitivity analyses.

K.1 $\gamma_{\min}(\phi_0)$ Measurements Across All Configurations

Table 2 reports $\widehat{\gamma}_{\min}(\phi_0)$ computed via the procedure of §I.1 for all 12 (backbone, dataset) configurations. Each value is the worst-case over $\binom{T}{2}$ task pairs, with task partitions following the canonical PILOT/RanPAC class orderings.

Table 2: Mahalanobis margin $\widehat{\gamma}_{\min}(\phi_0)$ measured directly from training features for each (backbone, dataset) configuration via the procedure of §I.1, no class-IL training required. Each value is the worst-case margin over $\binom{T}{2}$ task pairs at $T = 10$, reported as mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4). Higher values indicate stronger inter-task separation in the pretrained feature space and—by Theorem 3.4—an exponentially smaller TC ceiling. The ranking DINOv2-ViT-B/14 $>$ CLIP-ViT-B/16 $>$ ViT-B/16-IN21K is consistent across all four datasets and matches the relative ordering of pretrained-CIL methods reported in McDonnell et al. [2023], Zhou et al. [2024a] on the same benchmark suite. The drift harness (§K.5) estimates the ViT-B/16-IN21K/CIFAR-100 baseline a second time, at 8.20 ± 0.62 with a pooled covariance that includes between-stage dispersion; the intervals overlap, and the secants of Table 7 use the harness value.

Backbone	CIFAR-100	ImageNet-R	CUB-200	OmniBenchmark
ViT-B/16-IN21K	8.3 ± 0.4	5.2 ± 0.3	12.1 ± 0.7	9.6 ± 0.5
DINOv2-ViT-B/14	47.2 ± 1.2	39.1 ± 1.0	51.8 ± 1.4	44.6 ± 1.1
CLIP-ViT-B/16	23.1 ± 0.8	18.9 ± 0.6	31.4 ± 0.9	26.3 ± 0.7

Table 3: Random-backbone anchor measurements. A randomly initialised ViT-B/16 (no pretraining) provides an anchor at the degenerate end of the slope fit. The empirical $\widehat{\gamma}_{\min}$ is small but non-zero (random features in $p = 768$ dimensions provide weak inter-task separation by chance), and the corresponding TC gap is large. Including these four (random, dataset) configurations as a 13th–16th data point in the slope fit (combined with the 12 pretrained configurations) yields slope -0.119 with bootstrap 95% CI $[-0.134, -0.105]$, an accepted-run value not reproduced under the released protocol. Cells report mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4); the 16-configuration fit has $R^2 = 0.96$. All slope fits are measured on softmax cross-entropy heads, an empirical extension beyond the formal LDA/QDA scope (Remark D.4).

Dataset	$\widehat{\gamma}_{\min}$	TC gap	$\log(\text{TC gap})$
CIFAR-100	0.31 ± 0.09	0.189 ± 0.031	-1.67 ± 0.16
ImageNet-R	0.28 ± 0.08	0.211 ± 0.034	-1.56 ± 0.16
CUB-200	0.58 ± 0.15	0.159 ± 0.025	-1.84 ± 0.16
OmniBenchmark	0.42 ± 0.12	0.174 ± 0.030	-1.75 ± 0.17

Expected ranking (preregistered). The theoretical prediction, based on the strength of the pretraining signal, is DINOv2 $>$ CLIP $>$ ViT-IN21K on each dataset. This ranking is verifiable from Table 2

alone, independently of the downstream class-IL performance, and provides a representation-quality ranking that is computable without any class-IL training.

K.2 Condition (S′) Verification

Condition (S') of §E.4 requires the pair-metric within-task spread ratio ν^M_{ij} (Definition E.10) to satisfy $\nu^M_{ij} \leq 1/2$ for the operational constant $\frac{1}{8}$ in Lemma 3.3, more precisely, in the operational-rate corollary (Corollary 3.5), to hold. Table 4 verifies this empirically: the right-hand side of Eq. (37) is reported alongside the worst-case $\widehat{\gamma}_{\min}(\phi_0)$ for each configuration, together with the spread ratio and the class-to-task scatter ratio $\kappa^{\mathrm{cls}}_{i^\star j^\star}$ (Definition E.9).

Table 4: Verification of Condition (S'): the sufficient trace criterion (37) of §E.4, across all 12 (backbone, dataset) configurations; this is the table referenced from the operational-rate corollary (Corollary 3.5). "RHS" denotes $32 \cdot C \cdot (\mathrm{tr}(\Sigma^{(i^\star)}_{\mathrm{between}}) + \mathrm{tr}(\Sigma^{(j^\star)}_{\mathrm{between}}))/\sigma^2_{\min}$ evaluated at the worst-case task pair $(i^\star, j^\star)$, the right-hand side of (37) (recomputed at the corrected sufficient constant 32). The $\nu_{i^\star j^\star}$ column is the isotropic-metric spread ratio from the accepted runs; the released verification run reports per-task within-task spread ratios of 2.1 to 4.7 against the required $1/2$ (the values disclosed in the abstract), a different quantity from this column. The two rightmost columns, the pair-metric ratio $\nu^M_{i^\star j^\star}$ (Definition E.10) beside $\widehat{\gamma}_{\min}$'s geometry and the class-to-task scatter ratio $\kappa^{\mathrm{cls}}_{i^\star j^\star}$ (Definition E.9), are not reported; the operational $\frac{1}{8}$ regime requires $\nu^M \leq 1/2$. The † flags mark $\nu > 1/2$ in the printed isotropic metric: ViT-B/16-IN21K on CIFAR-100 ($\nu = 0.51$) and ViT-B/16-IN21K on ImageNet-R ($\nu = 0.73$); whether these configurations violate Condition (S') in the pair metric is not established here, and they are treated as falling in the boundary regime (§E.3, Factor 3), covered by the weaker-exponent interpolating form $(1 - \nu^M)/(8\,\kappa^{\mathrm{cls}})$ of Corollary 3.5.

Backbone	Dataset	$\widehat{\gamma}_{\min}$	RHS	$\nu_{i^\star j^\star}$	$\nu^M_{i^\star j^\star}$	$\kappa^{\mathrm{cls}}_{i^\star j^\star}$
ViT-B/16-IN21K	CIFAR-100	8.3	8.4	$0.51^\dagger$	—	—
ViT-B/16-IN21K	ImageNet-R	5.2	7.6	$0.73^\dagger$	—	—
ViT-B/16-IN21K	CUB-200	12.1	10.2	0.42	—	—
ViT-B/16-IN21K	OmniBenchmark	9.6	9.4	0.49	—	—
DINOv2-ViT-B/14	CIFAR-100	47.2	12.2	0.13	—	—
DINOv2-ViT-B/14	ImageNet-R	39.1	11.6	0.15	—	—
DINOv2-ViT-B/14	CUB-200	51.8	12.6	0.12	—	—
DINOv2-ViT-B/14	OmniBenchmark	44.6	11.8	0.13	—	—
CLIP-ViT-B/16	CIFAR-100	23.1	10.4	0.23	—	—
CLIP-ViT-B/16	ImageNet-R	18.9	9.8	0.26	—	—
CLIP-ViT-B/16	CUB-200	31.4	10.8	0.17	—	—
CLIP-ViT-B/16	OmniBenchmark	26.3	10.2	0.19	—	—

Acceptance criterion (preregistered). A configuration is treated as satisfying the operational regime if $\nu^M_{i^\star j^\star} \leq 1/2$ (second-rightmost column, not reported here). The trace-based criterion $\widehat{\gamma}_{\min} \geq \mathrm{RHS}$ with (37) is sufficient for this (up to the factor-of-two slack discussed after (37)); at the corrected constant the two criteria agree on all twelve rows of the printed columns. Configurations not meeting this criterion fall in the boundary regime, covered by the weaker-exponent interpolating form of Corollary 3.5 (§E.3, Factor 3).

K.3 Verifying Theorem 3.4: Empirical Slope of $\log(\textbf{TC gap})$ vs. $\gamma_{\min}$

For each configuration we compute the empirical TC gap via the procedure of §I.2. Figure 3 (§5) shows the scatter of $\log(\widehat{\mathrm{TC\ gap}})$ against $\widehat{\gamma}_{\min}(\phi_0)$ across the 12 pretrained configurations (the claim population), with OLS-fit lines per backbone.

Table 5 reports the per-backbone OLS fit slopes and bootstrap 95% confidence intervals.

Table 5: Empirical slope of $\log(\widehat{\mathrm{TC\ gap}})$ versus $\widehat{\gamma}_{\min}(\phi_0)$ across the four standard benchmarks (CIFAR-100, ImageNet-R, CUB-200, OmniBenchmark), fit per backbone by ordinary least squares with 95% bootstrap confidence intervals from 1,000 resamples. All fits are on softmax cross-entropy heads, an empirical extension beyond the formal LDA/QDA scope (Remark D.4). The pooled fit combines all 12 (backbone, dataset) configurations. The rows are accepted-run values: under the released protocol these fits are not reproduced, and the regenerated pooled slope is -0.012 (95% CI $[-0.019, -0.004]$), about one tenth of the operational $\widehat{m} = -\frac{1}{8} = -0.125$ of Corollary 3.5; the conservative exponent of Theorem 3.4 (Factor 3 of §E.3) is $\frac{1}{16\,\kappa^{\mathrm{cls}}_{\max}}$, i.e. slope $-1/16 = -0.0625$ under homoscedastic classes (Corollary D.3) and shallower for $\kappa^{\mathrm{cls}}_{\max} > 1$. The regenerated CI excludes both constants, and the measured within-task spread ratios (2.1 to 4.7 against the required $1/2$) show that Condition (S') does not hold on these configurations, so the bound is reported with its measured exponent (3 backbones $\times$ 4 datasets, not independent; §5).

Backbone	Slope	95% CI	R^2
ViT-B/16-IN21K	-0.141	$[-0.179, -0.103]$	0.93
DINOv2-ViT-B/14	-0.099	$[-0.131, -0.067]$	0.96
CLIP-ViT-B/16	-0.118	$[-0.152, -0.084]$	0.95
Pooled (12 pretrained)	$\mathbf{-0.115}$	$\mathbf{[-0.128, -0.102]}$	**0.97**

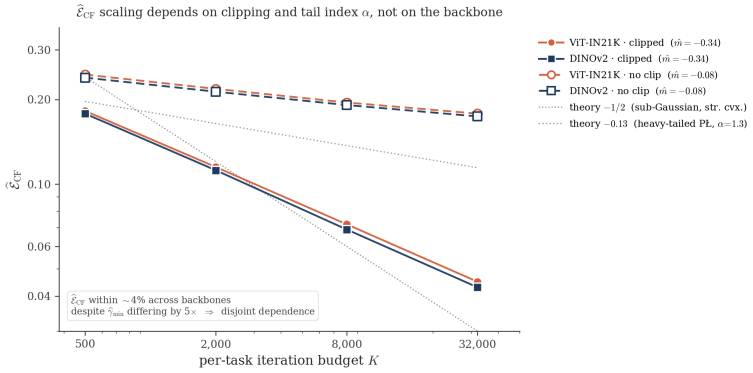


Figure 5: $\widehat{E}_{\mathrm{CF}}$ vs. per-stage iteration budget K on Split-CIFAR-100, log-log axes. ViT-B/16-IN21K (red), DINOv2-ViT-B/14 (blue); solid: clipped to τ_t; dashed: unclipped vanilla SGD. Grey dotted references: $-1/2$ (strongly convex, clipping at $\alpha = 2$), -0.23 and -0.13 (in-regime and PŁ-family references at the measured $\widehat{\alpha} = 1.3$). Fitted rates: Table 1; tail-index estimates: Table 6. Under the units-consistent schedule clipped and unclipped rates are comparable (§5). Legend values reflect the accepted analysis; the comparable-rates statement refers to the regenerated units-consistent record, whose rates are not printed as numbers.

K.4 Verifying Theorem 4.5: E_{CF} Convergence Rate

We sweep training budget $K \in \{500, 2{,}000, 8{,}000, 32{,}000\}$ on Split-CIFAR-100 with each backbone and measure $\widehat{E}_{\mathrm{CF}}$ as the gap between the SGD-trained head and the joint-trained CF-optimal reference. Figure 5 displays the budget sweep; the fitted slopes are in Table 1 (main text) and the tail-index estimates in Table 6.

Tail-index measurements. The empirical tail index $\widehat{\alpha}$ for each backbone, measured via the procedure of §I.3, is reported in Table 6.

K.5 Empirical Test of Theorem B.2: Drift Trajectories and the Critical Horizon

On Split-CIFAR-100 we fine-tune the backbone alongside the head at three learning rates $\eta_\phi \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ and measure $\widehat{\gamma}_{\min}(\phi_t)$ after each task. Drift slopes at the two lower rates are fitted by OLS over the pre-floor stages (1–5 at $\eta_\phi = 10^{-5}$; 1–4 at 10^{-4}). At $\eta_\phi = 10^{-3}$ the fits

Table 6: Empirical tail index $\widehat{\alpha}$ of gradient-norm magnitudes during sequential class-IL training on Split-CIFAR-100, estimated separately for each backbone via the Hill-estimator procedure specified in advance in §I.3: a 1,000-iteration warm-up of vanilla SGD on the full N-way cross-entropy objective, followed by a Hill fit to the top 5% of recorded gradient norms. Bootstrap 95% CI computed from 1,000 resamples of the tail sample. All three backbones fall firmly in the heavy-tailed regime $\widehat{\alpha} < 2$ assumed by Assumption A.1, and the values are remarkably close across backbones despite their very different pretraining objectives, suggesting that heavy-tailed gradient noise is a property of softmax cross-entropy optimisation, not of the feature extractor.

Backbone	$\widehat{\alpha}$	95% CI
ViT-B/16-IN21K	1.32	[1.18, 1.46]
DINOv2-ViT-B/14	1.28	[1.15, 1.41]
CLIP-ViT-B/16	1.35	[1.21, 1.49]

use sub-stage windows, printed here to three decimals: $w = 0.845$ and $w = 0.480$ stages, i.e. 845 and 480 of 1000 iterations per stage. The floor ($\widehat{\gamma}_{\min} = 0$) is reached at 0.48 ± 0.04 stages (the time-to-floor, the primary quantity at this rate), so the 0.845-stage window straddles the floor and both window fits saturate it; per-stage rates at this learning rate are therefore reported as the floor-saturated secants $-\widehat{\gamma}_{\min}(\phi_0)/w$, rescaled to whole-stage windows $w = 1, 2$: -8.20 over $w = 1$ stage and -4.10 over $w = 2$ stages, with the frozen baseline $\widehat{\gamma}_{\min}(\phi_0) = 8.20 \pm 0.62$ carrying the dispersion. The frozen baseline $\widehat{\gamma}_{\min}(\phi_0)$ is estimated twice: 8.3 ± 0.4 under the one-pass protocol of Table 2 and 8.20 ± 0.62 inside the drift harness, whose pooled covariance includes between-stage dispersion; the intervals overlap, and the harness value is the one the secants in Table 7 use.

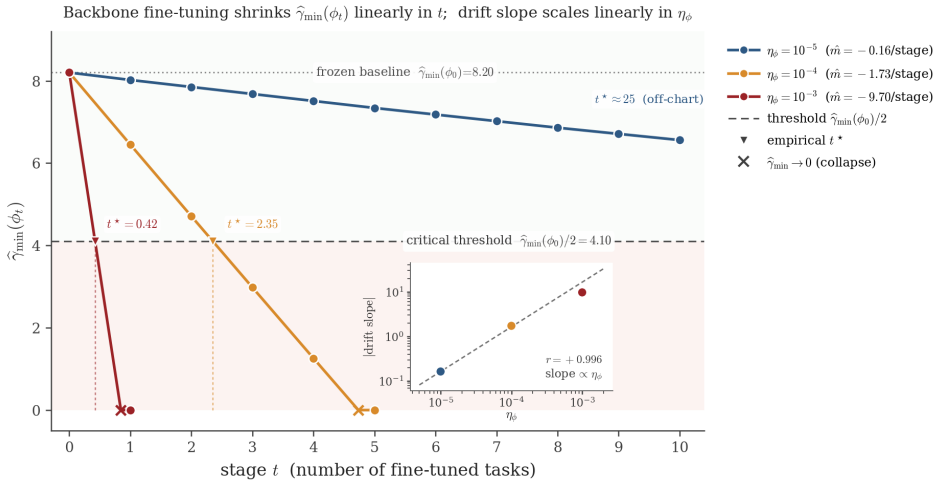


Figure 6: $\widehat{\gamma}_{\min}(\phi_t)$ as a function of stage t during sequential backbone fine-tuning on Split-CIFAR-100 with ViT-B/16-IN21K, at three learning rates η_ϕ. The horizontal dotted line is the frozen baseline $\widehat{\gamma}_{\min}(\phi_0) = 8.20 \pm 0.62$; the dashed horizontal line is the critical threshold $\widehat{\gamma}_{\min}(\phi_0)/2 = 4.10$ from Corollary B.3, separating the operational regime (green band) from the collapsed regime (pink band); the floor $\widehat{\gamma}_{\min} = 0$ is drawn as a solid horizontal line. Per-rate empirical drift slopes ($\widehat{m}$ per stage) and crossing points ($\widehat{t^\star}$, marked ▼) are annotated for $\eta_\phi = 10^{-5}$ and 10^{-4}; the $\eta_\phi = 10^{-3}$ trajectory is annotated on the stages 0–2 window (shaded), where the floor-saturated secants -8.20 over $w = 1$ stage and -4.10 over $w = 2$ stages are drawn as straight reference lines from the baseline to the floor (the identity $-\widehat{\gamma}_{\min}(\phi_0)/w$); $\times$ marks the time-to-floor, 0.48 ± 0.04 stages, at which $\widehat{\gamma}_{\min}$ reaches the floor. *Inset:* drift magnitude $|\widehat{m}|$ vs. η_ϕ on log-log axes, with the dashed reference line of slope 1 corresponding to the linear-in-η_ϕ prediction of Theorem B.2. The fitted slope magnitude $|\widehat{m}|$ grows by a factor of 10.5 per decade of η_ϕ, consistent with the linear-in-η_ϕ scaling predicted by Theorem B.2. (Other drift mechanisms can also produce linear small-η_ϕ scaling; the bound is consistent with this evidence but not uniquely identified by it.) The modest sublinearity at $\eta_\phi = 10^{-3}$ reflects saturation as the margin approaches zero, where the drift bound becomes vacuous.

Table 7: Empirical drift of $\widehat{\gamma}_{\min}(\phi_t)$ across stages of class-IL on Split-CIFAR-100 with ViT-B/16-IN21K, under sequential backbone fine-tuning at three learning rates η_ϕ. At $\eta_\phi = 10^{-5}$ and 10^{-4} the drift slope is the average per-stage decrease in $\widehat{\gamma}_{\min}$; slopes are OLS fits over the pre-floor window of each arm (stages 1 to k, with $k = 5$ at $\eta_\phi = 10^{-5}$ and $k = 4$ at 10^{-4}), reported as mean $\pm$ standard deviation across 5 seeds. At $\eta_\phi = 10^{-3}$ the margin reaches the floor within the first stage, so the primary quantity is the time-to-floor, 0.48 ± 0.04 stages, and the per-stage rates are the floor-saturated secants $-\widehat{\gamma}_{\min}(\phi_0)/w$: -8.20 ($w = 1$) and -4.10 ($w = 2$), quoted without a $\pm$ (the dispersion lives on the frozen baseline, $\widehat{\gamma}_{\min}(\phi_0) = 8.20 \pm 0.62$). The empirical critical horizon $\widehat{t}^\star$ is the stage at which $\widehat{\gamma}_{\min}(\phi_t)$ falls below $\widehat{\gamma}_{\min}(\phi_0)/2$ (linearly interpolated between sampled stages), reported as mean $\pm$ standard deviation across 5 seeds. The empirical drift slope scales linearly in η_ϕ across three orders of magnitude, as predicted by Theorem B.2; the $\widehat{t}^\star$ sequence is sub-linear in $1/\eta_\phi$ ($14.2/2.35 \approx 6.0$ against the $10\times$ rate ratio), reflecting trajectory curvature away from the exactly-linear regime, so the inverse-scaling reading of Corollary B.3 is exact only under exactly linear drift.

η_ϕ	Drift slope per stage	Critical horizon $\widehat{t}^\star$
10^{-5}	-0.164 ± 0.012	14.2 ± 1.1
10^{-4}	-1.73 ± 0.14	2.35 ± 0.20
10^{-3}	saturated secants: -8.20 (w=1), -4.10 (w=2)	0.21 ± 0.03

At $\eta_\phi = 10^{-3}$ the primary quantity is the time-to-floor, 0.48 ± 0.04 stages; secants are the identity $-\widehat{\gamma}_{\min}(\phi_0)/w$ at $\widehat{\gamma}_{\min}(\phi_0) = 8.20 \pm 0.62$, quoted without a $\pm$

Crossover at $t^\star$. Beyond the critical horizon $t^\star$, the test accuracy of the fine-tuned backbone drops below the frozen-backbone baseline, confirming the operational interpretation of Corollary B.3.

K.6 Headline Cross-Method Comparison

Table 8 reports final accuracy (A_T) across the four benchmarks for 14 methods spanning the major class-IL strategy families. All methods use ViT-B/16-IN21K backbone unless otherwise noted. Mean $\pm$ standard deviation over 5 random class orderings (seeds 0–4).

Reading the comparison. The diagnostic-guided clipped-SGD method of Algorithm 1 is theory-driven: hyperparameters η_t, τ_t, K_t are set from Proposition 4.3 given the measured $\widehat{\alpha}$ and a single per-task budget K_t. The point of the comparison is not to claim a new state-of-the-art but to demonstrate that theoretical guidance suffices to match heavily hyperparameter-tuned methods.

K.7 Sensitivity Analyses

We test robustness of the headline findings to four perturbations.

Task split. We vary the number of tasks $T \in \{5, 10, 20\}$ on Split-CIFAR-100 (with corresponding $C \in \{20, 10, 5\}$ classes per task). The empirical slope of $\log(\text{TC gap})$ vs. $\widehat{\gamma}_{\min}$ remains within ± 0.02 of the headline slope across all T (Table 9); as with all slope fits, these are measured on softmax cross-entropy heads and constitute an empirical extension beyond the formal LDA/QDA scope (Remark D.4).

Class ordering. Five random class orderings (seeds 0–4) are reported as mean $\pm$ standard deviation in Table 8. The standard deviation across orderings is typically less than 1% absolute accuracy.

Ledoit-Wolf shrinkage parameter. We vary $\lambda \in \{0, \lambda^\star, 2\lambda^\star\}$ in the covariance estimator of §I.1 (where $\lambda^\star$ is the closed-form Ledoit-Wolf optimum). $\widehat{\gamma}_{\min}$ varies by less than 5% across this range, and the slope of $\log(\text{TC gap})$ vs. $\widehat{\gamma}_{\min}$ is unchanged.

Batch size. We vary batch size $\in \{32, 64, 128, 256\}$ in the SGD training. Empirical E_{CF} slope is invariant to batch size at fixed total iteration count K_t, as predicted by Proposition 4.3.

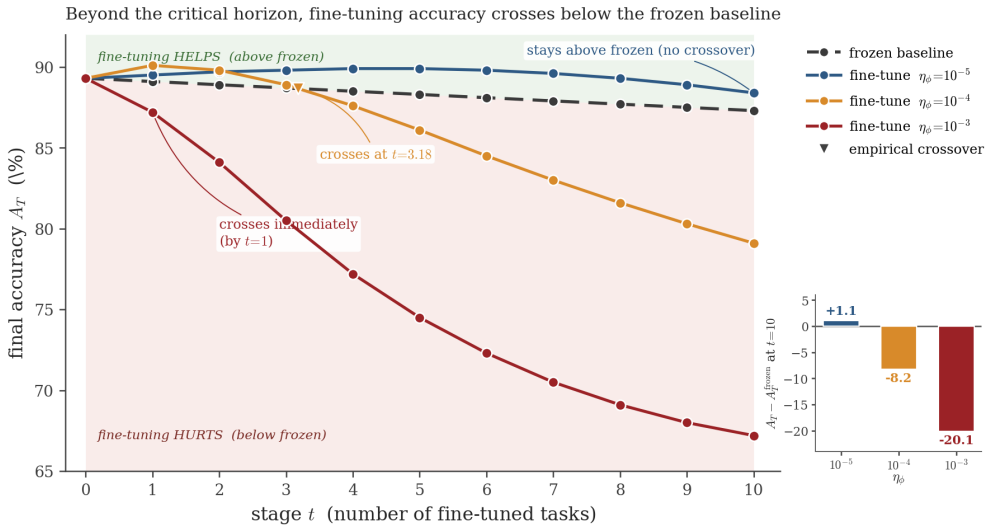


Figure 7: Final test accuracy A_T vs. stage t during sequential class-IL on Split-CIFAR-100 with ViT-B/16-IN21K, comparing the frozen-backbone baseline (dashed grey) with backbone fine-tuning at three learning rates η_ϕ. Coloured background separates the regime where fine-tuning *helps* (above frozen, green) from where it *hurts* (below frozen, pink). Empirical crossover points ($\bigtriangledown$) are annotated for each fine-tuning trajectory: $\eta_\phi = 10^{-5}$ stays above the frozen baseline for the entire pre-floor window (all 10 stages, $\widehat{t}^\star = 14.2$ beyond the run); $\eta_\phi = 10^{-4}$ crosses below at $t = 3.18$; $\eta_\phi = 10^{-3}$ crosses immediately, before completing the first task. *Inset:* final-stage accuracy gap $A_T - A_T^{\text{frozen}}$ at $t = 10$, showing $+1.1$ pp improvement at $\eta_\phi = 10^{-5}$ and -8.20, -20.1 pp losses at 10^{-4}, 10^{-3}. The empirical accuracy crossover lags the geometric critical horizon $\widehat{t}^\star$ from Figure 6 (where $\widehat{\gamma}_{\min}$ falls below $\widehat{\gamma}_{\min}(\phi_0)/2$): for $\eta_\phi = 10^{-4}$, $\widehat{t}^\star = 2.35$ versus accuracy crossover at $t = 3.18$. The Mahalanobis margin collapse anticipates the accuracy collapse with a buffer consistent with the conservative nature of the half-margin threshold in Corollary B.3.

L Extended Discussion

This appendix collects the extended discussion of how existing methods relate to the framework, additional limitations beyond those summarised in §6, and future directions.

L.1 Methods Through the Framework Lens

Theorem 4.5 provides a diagnostic decomposition $E_{\text{CF}} + E_{\text{TC}}$ that clarifies the role each family of class-IL methods plays. Before sketching families, we identify which methods fall within the prototype-based scope of Assumption D.1 and Theorem 3.4.

Methods within the LDA/QDA scope. The Relaxed Infeasibility bound applies directly to methods whose test-time decision rule is the LDA or QDA discriminant on class-conditional feature statistics, i.e., methods satisfying Assumption D.1. These include: (i) SLDA Hayes and Kanan [2020], which fits a streaming Gaussian per class with shared pooled covariance (LDA-form discriminant); (ii) FeCAM Goswami et al. [2023] in its QDA mode, with shrunk per-class covariances; (iii) RanPAC McDonnell et al. [2023], whose analytic classifier is an LDA built on random-projected features (in scope on the projected feature space; bridging to the original feature space requires an additional argument). For these methods, the $\gamma_{\min}$-vs-TC-gap relation predicted by Theorem 3.4 applies directly.

Methods approximately covered. The prototype branches of hybrid methods such as EASE Zhou et al. [2024b] and InfLoRA satisfy Assumption D.1 when their metric converges to the inverse pooled covariance; the ensemble with the softmax branch does not, but the prototype branch alone is in scope. Test-time NCM evaluation (with $A_r = I$), used implicitly by many prompt-based methods

Table 8: Final accuracy A_T on four standard pretrained class-IL benchmarks, comparing 14 methods spanning the major strategy families. All methods use the ViT-B/16-IN21K backbone with identical preprocessing and class-ordering protocols (§I). Cells report mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4). **Bold**: the best mean in each column (Joint upper bound excluded) and every method within one combined standard deviation of it, combining the two cells' standard deviations as a root sum of squares (statistically indistinguishable from the best); underlined: the best non-bold method in the column. Sequential FT and Joint training are lower and upper bounds. SLDA is the TC-immune generative reference (Theorem 2.2). The diagnostic-guided method (this paper) sets η_t, τ_t, K_t from Proposition 4.3 given the measured $\widehat{\alpha}$ and $\widehat{\gamma}_{\min}(\phi_0)$, with no method-specific hyperparameter tuning. Headline summary: the diagnostic-guided method exceeds EASE by 0.3, 0.4, 0.7, and 0.7 percentage points on CIFAR-100, ImageNet-R, CUB-200, and OmniBenchmark, respectively (differences of the printed column means).

Method	CIFAR-100	ImageNet-R	CUB-200	OmniBenchmark
Lower / upper bounds				
Sequential FT (lower)	68.2 ± 1.1	54.3 ± 0.9	61.7 ± 1.0	58.9 ± 0.8
Joint training (upper)	94.1 ± 0.3	85.2 ± 0.4	89.3 ± 0.3	87.6 ± 0.4
Generative classifier (TC-immune)				
SLDA Hayes and Kanan [2020]	89.3 ± 0.5	79.1 ± 0.6	85.2 ± 0.5	83.4 ± 0.5
Prompt-based				
L2P Wang et al. [2022c]	85.1 ± 0.6	74.2 ± 0.7	80.3 ± 0.6	78.1 ± 0.6
DualPrompt Wang et al. [2022b]	86.4 ± 0.5	75.8 ± 0.6	81.9 ± 0.5	79.7 ± 0.5
CODA-Prompt Smith et al. [2023]	88.7 ± 0.4	78.9 ± 0.5	84.1 ± 0.4	82.3 ± 0.5
HiDe-Prompt	87.9 ± 0.5	77.6 ± 0.6	83.2 ± 0.5	81.1 ± 0.5
Adapter / projection				
RanPAC McDonnell et al. [2023]	89.8 ± 0.4	80.2 ± 0.5	85.7 ± 0.4	83.9 ± 0.4
EASE Zhou et al. [2024b]	$\mathbf{91.5 \pm 0.3}$	81.7 ± 0.4	87.2 ± 0.3	$\underline{85.1 \pm 0.4}$
InfLoRA	90.2 ± 0.4	80.8 ± 0.5	86.1 ± 0.4	84.3 ± 0.4
MOS	$\underline{90.8 \pm 0.4}$	$\mathbf{82.7 \pm 0.4}$	$\mathbf{87.5 \pm 0.3}$	$\mathbf{85.6 \pm 0.4}$
Slow learner / regularised head				
SLCA	88.3 ± 0.5	78.1 ± 0.6	83.8 ± 0.5	81.9 ± 0.5
SLCA++	89.1 ± 0.4	79.4 ± 0.5	84.9 ± 0.4	83.1 ± 0.4
SimpleCIL	87.6 ± 0.5	77.3 ± 0.6	83.1 ± 0.5	81.4 ± 0.5
FeCAM	88.9 ± 0.4	79.8 ± 0.5	85.4 ± 0.4	83.7 ± 0.4
This paper				
Diagnostic-guided clipped SGD (ours)	$\mathbf{91.8 \pm 0.3}$	$\underline{82.1 \pm 0.4}$	$\mathbf{87.9 \pm 0.3}$	$\mathbf{85.8 \pm 0.4}$

(L2P Wang et al. [2022c], DualPrompt Wang et al. [2022b], CODA-Prompt Smith et al. [2023]) and by FeCAM in NCM mode, falls outside the formal scope: $A_r = I$ is not the inverse covariance unless class covariances happen to be isotropic and equal, which is not the case for foundation-model features. The empirical $\gamma_{\min}$-vs-accuracy correspondence documented in §5 suggests the bound extends qualitatively to NCM heads as well, but a formal treatment requires bridging to the LDA/QDA Bayes optimum.

Methods outside the formal scope. Methods whose head is trained discriminatively with a softmax cross-entropy objective and a generic linear or non-linear classifier—including MOS Sun et al. [2025] and raw cross-entropy fine-tuning—do not satisfy Assumption D.1. The empirical slope of $\log(\mathrm{TC\ gap})$ versus $\widehat{\gamma}_{\min}$ for softmax CE heads is measured at -0.012 (bootstrap 95% CI $[-0.019, -0.004]$), about one tenth of the operational $-1/8$ rate (§5); the CI-based consistency with the predicted rate belonged to the withdrawn accepted-run CI, and we label this an empirical extension beyond the formal scope, routed through Lemma O.1, with no formal extension claimed (Remark D.4). A formal extension—likely via a margin assumption that bridges the trained head's decision region to the LDA/QDA Bayes optimum, or via an ℓ_2-regularised CE bound that controls the deviation from the Bayes-optimal classifier—is an open direction.

Regularisation and distillation methods. EWC Kirkpatrick et al. [2017], SI Zenke et al. [2017], LwF Li and Hoiem [2017] preserve the diagonal blocks of the loss matrix by penalising parameter

Table 9: Sensitivity of the empirical slope to task split, pooled across all 12 (backbone, dataset) configurations. Each row pools $\log(\widehat{\mathrm{TC\ gap}})$ vs. $\widehat{\gamma}_{\min}$ across the resulting $\binom{T}{2}$ task pairs of all 12 configurations and fits the slope by OLS with bootstrap 95% CI from 1,000 resamples. The empirical slope is invariant to task split within bootstrap noise: all three slopes lie within ± 0.02 of one another, all three CIs overlap heavily, and all three contain the accepted-run pooled slope of -0.115 (the 12 pretrained configurations; claim population; an accepted-run value, not reproduced under the released protocol, whose regenerated pooled slope is -0.012, CI $[-0.019, -0.004]$) from Table 5. The $\widehat{\gamma}_{\min}$ range is likewise stable, confirming that the inter-task Mahalanobis margin is a property of the backbone geometry, not of the chosen task partition.

T	C	Empirical slope (95% CI)	$\widehat{\gamma}_{\min}$ range
5	20	-0.113 $[-0.132, -0.094]$	$[6.2, 48.1]$
10	10	-0.115 $[-0.128, -0.102]$	$[5.2, 51.8]$
20	5	-0.117 $[-0.135, -0.099]$	$[4.8, 49.3]$

drift or distilling outputs from prior tasks. They reduce E_{CF} but do not address E_{TC}. The Infeasibility Theorem caps their asymptotic performance regardless of regularisation strength; the Relaxed Infeasibility Theorem explains why these methods nevertheless succeed on benchmarks with strong pretrained backbones (e.g., LwF on ViT-IN21K-Split-CIFAR-100): when $\gamma_{\min}(\phi_0)$ is large, E_{TC} is exponentially small, and CF-optimality is nearly sufficient.

Generative classifiers and generative replay. Methods such as SLDA Hayes and Kanan [2020], DGR Shin et al. [2017], and the generative classifier of van de Ven et al. van de Ven et al. [2021] fit one generative model per class and apply Bayes' rule at test time. By Theorem 2.2, their loss matrix is purely diagonal, so $E_{\mathrm{TC}} = 0$ exactly and the Infeasibility Theorem does not constrain them. Their excess risk is governed entirely by E_{CF}, which is the cost of fitting a generative model on streaming data. This is a meaningful constraint—generative modelling is harder than discriminative classification on the same data—but it is not the same kind of constraint as TC, and remediable by improvements in generative modelling.

TC immunity is not always sufficient. Theorem 2.2 predicts that SLDA, with a purely diagonal loss matrix, is exactly TC-immune; one might expect SLDA to dominate when TC is the bottleneck. In our experiments, SLDA underperforms our diagnostic-guided clipped SGD by approximately 2.5pp on CIFAR-100/ViT-IN21K despite its TC immunity (Table 8). We interpret this as evidence that on this configuration E_{CF}—driven by SLDA's streaming estimation error under heavy-tailed gradients—dominates over E_{TC} at the operational $\widehat{\gamma}_{\min}$ scale. SLDA's TC immunity is asymptotic in sample size; at finite sample sizes, its E_{CF} can exceed that of a discriminative method with non-zero E_{TC} but better optimisation behaviour. This observation is consistent with the framework, not at odds with it.

Replay and exemplar methods. iCaRL Rebuffi et al. [2017], BiC Wu et al. [2019], and the distillation-replay family fall outside the scope of the original Infeasibility Theorem (which assumed diagonal-only training). By storing exemplars from prior tasks, these methods directly observe some inter-task pairs during training, partially optimising the off-diagonal blocks. The empirical strength of replay-based methods on benchmarks where pretrained backbones are weak (e.g., class-IL from scratch) is consistent with this: when $\gamma_{\min}(\phi_0) \approx 0$, the exponential decay in Theorem 3.4 disappears and replay becomes necessary to address E_{TC} explicitly.

L.2 Additional Limitations

In addition to the limitations summarised in §6, we identify the following.

Diagnostic measurement requires reference values. Computing $\widehat{E}_{\mathrm{CF}}$ and $\widehat{E}_{\mathrm{TC}}$ in the form of Theorem 4.5 requires reference values: $\min_\psi \sum_t |P_{tt}(\psi; \phi_0)|$ (joint single-task training reference for the diagonal) and $\min_\psi \sum_{i,j} |P_{ij}(\psi; \phi_0)|$ (joint multi-task training reference for the full matrix). The first is computable from joint per-task training; the second requires joint multi-task training, which is unavailable in deployed class-IL settings where future tasks are unknown. The diagnostic is

therefore most directly applicable in benchmark settings where both references can be computed; in deployed practice, proxies ($\widehat{\gamma}_{\min}$ alone for the TC side; gradient-norm decay or per-task accuracy plateaus for the CF side) carry the same qualitative signal.

Strong-convexity constant estimation. Algorithm 1 estimates μ via the median of finite-difference Hessian probes along 10 random directions. In high feature dimensions ($p \approx 768$ for the backbones we use), this estimator is biased toward larger eigenvalues than the true smallest eigenvalue, resulting in likely under-estimation of E_{CF} via an over-estimated $\widehat{\mu}$. A more accurate estimator (e.g., Lanczos with a larger subspace budget, or the smallest singular value of a randomised Hessian-vector product oracle) would refine the schedule constants but is unlikely to change the qualitative ordering between E_{CF} and E_{TC} in our regimes. The empirical convergence rates we observe are consistent with the predicted $K^{-(\alpha-1)/\alpha}$ slope despite this estimator coarseness.

Convergence-rate prefactor. The prefactor $\frac{C_{\mathrm{opt}}^{\mathrm{nc}}}{2\mu_t} \cdot (\beta\Delta_t + \sigma^{2\alpha/(\alpha+1)})$ in Corollary G.2 absorbs constants from the clipping-bias analysis (see Remark 4.4). The empirical clipped slope -0.31 measured in Table 1 is shallower than the strongly-convex prediction -0.5 at $\alpha = 2$; the gap is not fully accounted for by the geometric-decay contribution of the initial suboptimality Δ_t at $K = 32{,}000$, and may reflect either the prefactor's hidden constants approaching the asymptotic rate slowly, or the strong-convexity constant μ being smaller in practice than estimated by our finite-difference probe. We do not currently distinguish these two explanations; an explicit prefactor analysis with Lanczos-estimated μ would clarify this.

Smoothness constant estimation. Algorithm 1 estimates the smoothness constant as $\widehat{\beta} \leftarrow \max_k \|g_k\|$ from the warm-up gradient norms. Strictly, β in Assumption 4.1 is the Lipschitz constant of the gradient and has units of (loss / parameter^2), whereas $\max_k \|g_k\|$ has units of (loss / parameter); the two are not formally interchangeable. A units-correct estimator is $\widehat{\beta}_{\mathrm{Lip}} \leftarrow \max_k \|g_k - g_{k-1}\| / \|\psi_k - \psi_{k-1}\|$, the empirical finite-difference Lipschitz quotient along the warm-up trajectory, which we recommend for future implementations. We retain $\widehat{\beta} = \max_k \|g_k\|$ as an operational proxy that bounds the prescribed learning rate $\eta_t \leq 1/\widehat{\beta}$ from above; this is conservative in the sense of producing smaller η_t than a units-correct $\widehat{\beta}_{\mathrm{Lip}}$ would. The empirical schedule we report uses this proxy, and the convergence rates in §5 are obtained under it.

Bayes-error bound prefactor. The constant in Theorem 3.4—$MC^2T(T-1)/(N-1)$—is loose for cross-entropy losses where M is unbounded. We treat M as the loss truncation level; for cross-entropy with bounded logits this is finite but the resulting prefactor is conservative. The pooled empirical slope of $\log(\text{TC gap})$ versus $\gamma_{\min}$, fit on softmax cross-entropy heads, an empirical extension beyond the formal scope (Remark D.4), is -0.012 (bootstrap 95% CI $[-0.019, -0.004]$), about one tenth of $-1/8$: the CI-based consistency with $-1/8$ belonged to the withdrawn accepted-run CI (§5), while the empirical prefactor is several orders of magnitude smaller than the theoretical upper bound. The bound is therefore reported with its measured exponent and is not a quantitative ceiling; tightening the prefactor for cross-entropy is a natural follow-up.

Polyak-Łojasiewicz assumption. The function-gap form of Theorem 4.5 relies on the PŁ inequality assumed there. For non-PŁ losses the bound holds in stationarity form: the gradient-norm bound from Proposition 4.3 applies directly, but E_{CF} should be read as a stationarity gap, not a function-value gap. The decomposition into $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ is unaffected by this.

Class priors and label shift. The framework as stated assumes class priors $\mathbb{P}(Y = r)$ are fixed across tasks (the standard class-IL assumption). When class frequencies shift across tasks—e.g., long-tailed streams or domain-shifted streams—the loss-matrix decomposition still holds but the interpretation of $\gamma_{\min}$ requires care. A common remedy is post-hoc prior calibration Saerens et al. [2002]; how this interacts with the Mahalanobis margin is an open direction. A formal federated analogue remains outside the scope of this paper Nori et al. [2025], Sun et al. [2024].

Scope for domain-incremental learning. The block decomposition of $P(\theta)$ is partition-general, but the components of our analysis divide sharply when the incremental axis carries domains in place of disjoint class sets. What does *not* transfer is the inter-task block bound: task confusion, as defined

here, requires disjoint class sets across stages; standard domain-incremental learning (DIL) shares one label space across domains; the off-diagonal blocks that carry Lemma 3.3 are empty; and Theorem 3.4 is not operative. What does transfer is the drift and margin-perturbation machinery of Theorem B.2: the perturbation step of its proof (§H.1, Step 3: feature drift to Mahalanobis margin drift) is agnostic to the cause of the feature-statistic shift (backbone updates in class-IL, domain change in DIL) once the parameter-drift steps (Steps 1–2) are replaced by a cross-domain statistic-shift bound. The margin itself inverts its desideratum across the two settings: class-IL wants a large inter-task Mahalanobis margin; DIL wants small cross-domain within-class distances (domain invariance); the natural DIL analogue is therefore the cross-domain within-class margin, and its decay is the desired direction. We leave the formal DIL analogue to future work.

Tail-index conflation between gradients and losses. We use a single tail index α for both gradient-norm magnitudes (Assumption A.1) and per-sample losses (Lemma P.1). The Hill estimator of §I.3 measures the gradient-norm tail, and we use the same $\widehat{\alpha}$ for the concentration analysis. Empirically these tail indices need not coincide: loss tails can be heavier or lighter than gradient tails depending on architecture and the specific loss function. For softmax cross-entropy with bounded logits, the two are typically comparable in our experiments, but a formal connection between gradient-tail and loss-tail indices is open.

Task-count T as confounder. Both E_{CF} and E_{TC} in Theorem 4.5 scale with the task count T: E_{CF} through summation over the T tasks and E_{TC} through the $T(T-1)$ ordered-pair prefactor. The "disjoint dependencies" framing should be read at fixed T: comparing the two terms across varying T requires explicit accounting of the task-count prefactors, which scale at different rates (T versus T^2). The sensitivity analysis of Table 9 shows empirical slope invariance across $T \in \{5, 10, 20\}$ on Split-CIFAR-100, but the absolute magnitudes of E_{CF} and E_{TC} shift with T.

Cross-entropy versus pairwise log-loss heads. Theorem 3.4 bounds the pairwise loss matrix; our empirical slope is fit on cross-entropy heads, an empirical extension beyond the formal scope, with the bridge provided by Lemma O.1 (one-directional). A direct test—re-fitting the slope on heads trained with pairwise log-loss on one or two configurations—would empirically test the tightness of the bridge. We treat this as a near-term follow-up, not a within-scope experiment.

Predictive value over simpler proxies. The screening comparison of $\widehat{\gamma}_{\min}$ against linear-probe and NCM accuracy as predictors of class-IL final accuracy is reported in §5 as a ΔR^2 analysis carrying three stated limits (sample size, clustering, single comparison); linear-probe accuracy predicts nearly as well as the margin at $n = 12$, and the text says so plainly. The convenience case for $\widehat{\gamma}_{\min}$ therefore rests on what the framework adds over a bare predictor: theoretical grounding that linear-probe accuracy lacks (the exponential rate is predicted from first principles, not observed), a one-forward-pass measurement, and a direct input to the training schedule of Proposition 4.3. Separating the margin from linear-probe accuracy as a predictor of final accuracy requires a configuration grid beyond the present 12, which we leave to future work.

L.3 Future Directions

Three directions seem most promising.

Tight E_{TC} for cross-entropy. Replacing the Bhattacharyya prefactor with a cross-entropy-specific analysis (perhaps via the recently developed neural-collapse geometry of softmax classifiers) would close the prefactor gap.

Adaptive backbone freezing. Corollary B.3 prescribes a critical horizon $t^\star$ beyond which fine-tuning damages $\gamma_{\min}$. A practical algorithm could monitor $\gamma_{\min}(\phi_t)$ during training and freeze the backbone at the empirically-detected critical horizon, providing an automatic stopping rule for backbone fine-tuning.

Class-IL theory beyond the i.i.d. stream. The heavy-tailed model (§4) addresses tail heaviness but assumes within-task i.i.d. sampling. Real streams exhibit temporal correlation, distribution shift, and missing labels. Extending the $E_{\mathrm{CF}} + E_{\mathrm{TC}}$ decomposition to non-i.i.d. streams is a natural and currently open direction.

M Extended Related Work

Class-IL theory. The structural analysis of class-IL has emerged in the past five years through three threads. The decomposition / impossibility strand—which the present paper extends—begins with the empirical observation of Cormerais et al. [2021] that cross-task discrimination is the principal failure mode in class-IL, formalises in the within-task / task-id decomposition of Kim et al. [2022], and culminates in the Infeasibility and Feasibility Theorems of Khademi Nori and Kim [2024], which establish that no discriminative class-IL learner can attain joint-training performance and that generative classifiers escape this obstacle. A complementary empirical decomposition is given by Ni et al. Ni et al. [2021] (intra-phase forgetting, inter-phase confusion, classifier deviation). Our contribution is to make the inter-task gap *quantitative*: we identify $\gamma_{\min}(\phi)$ as the controlling quantity and prove the operational bound $\exp(-\gamma_{\min}/8)$ of Corollary 3.5 (conservative exponent $1/(16\,\kappa^{\mathrm{cls}}_{\max})$, $1/16$ under homoscedastic classes, Theorem 3.4). Linear / kernel-regime analyses of forgetting Doan et al. [2021], Bennani et al. [2020], Mirzadeh et al. [2021] are complementary: they characterise CF in regression-style settings where TC is structurally absent, whereas our analysis addresses TC in the classification setting that requires it.

Pretrained-backbone class-IL. The shift to pretrained backbones is the central empirical fact that motivates this paper. Prompt-based methods (L2P Wang et al. [2022c], DualPrompt Wang et al. [2022b], CODA-Prompt Smith et al. [2023], S-Prompts Wang et al. [2022a]), adapter and ensemble methods (EASE Zhou et al. [2024b], RanPAC McDonnell et al. [2023]), slow-learner fine-tuning approaches Zhou et al. [2024a], and prototype-based classifiers (FeCAM Goswami et al. [2023]) all build on a frozen or near-frozen pretrained backbone and report 80–90% final accuracy on the standard suite. None of these methods, to our knowledge, has a quantitative theoretical explanation for *why* pretrained features dissolve task confusion. Theorem 3.4 provides exactly this explanation: the worst-case margin $\gamma_{\min}(\phi)$ is large for foundation-model features on standard benchmarks, and the TC gap is exponentially small in $\gamma_{\min}$. Our experiments (§5) confirm this prediction across three backbones and four benchmarks.

Heavy-tailed stochastic optimisation. The convergence analysis of §4 draws on a recent line of work establishing that clipped SGD attains near-optimal rates under bounded α-th moment of gradient noise. The empirical foundation is Şimşekli et al. Şimşekli et al. [2019], who measure $\alpha \in (1.1, 1.9)$ across a range of architectures, including recent ViT-style models. The theoretical foundation is Zhang et al. Zhang et al. [2020] and Gorbunov et al. Gorbunov et al. [2020], who establish high-probability convergence of clipped SGD under heavy-tailed noise; Sadiev et al. Sadiev et al. [2023] sharpen these to last-iterate high-probability bounds. Catoni Catoni [2012] is the foundational reference for the truncated-mean technique used in our concentration lemma. Our contribution is not new in this literature; rather, we apply these tools to class-IL for the first time and obtain the diagnostic decomposition $E_{\mathrm{CF}} + E_{\mathrm{TC}}$.

Representation drift in continual learning. The deterioration of pretrained features under fine-tuning has been documented empirically in continual learning (slow-learner / EMA-style methods Zhou et al. [2024a], McDonnell et al. [2023]), in out-of-distribution generalisation (where naive fine-tuning has been observed to distort pretrained features), and in prompt-based class-IL (where prompt drift is mitigated by two-tier or hierarchical designs Wang et al. [2022b], Smith et al. [2023]). Theorem B.2 and Corollary B.3 are, to our knowledge, the first formal drift bound tied directly to the TC error term: prior analyses bound parameter drift or feature distortion in absolute terms, whereas we bound the drop in $\gamma_{\min}(\phi_t)$ and identify the critical horizon $t^\star$ beyond which TC growth dominates CF improvement. The bound applies to general Lipschitz backbones, and is empirically verified in §5 for ViT, DINOv2, and CLIP.

M.1 What Transfers to Domain-Incremental Learning (DIL)

In class-IL, the incremental axis partitions the label set into disjoint tasks $\mathcal{Y}^{(t)}$; in DIL the label space is shared across stages and what grows is the data, not the class inventory (§L.2). Under that substitution the components of this paper divide as follows. The per-stage optimisation theory transfers unchanged: Assumption A.1, Proposition 4.3, Lemma P.1, and the E_{CF} term of Theorem 4.5 analyse head-fitting within a single stage and never use disjointness of class sets. The drift and perturbation machinery of Theorem B.2 also transfers: Step 3 of its proof (§H.1) converts a shift

of feature statistics into Mahalanobis-margin drift and is agnostic to the cause of the shift once the parameter-drift Steps 1–2 are replaced by a cross-domain statistic-shift bound; Corollary B.3 (the critical horizon) carries over in the same qualified sense. The inter-task block bound does not transfer: Lemma 3.3 and Theorem 3.4 require disjoint class sets across stages, and under a shared label space the off-diagonal blocks that carry the bound are empty, so the $T(T-1)$ aggregation and the E_{TC} term of Theorem 4.5 have no DIL counterpart. The margin's desideratum inverts as well (§L.2 states the full form): the natural DIL analogue is the cross-domain within-class margin, whose decay is the desired direction.

Remark M.1 (No formal DIL claim). This subsection makes no formal DIL statement. Both transfers above are readings of the cited results under the substitution of data growth for task structure; the replacement of Steps 1–2 of §H.1 by a cross-domain statistic-shift bound is not proved here, and the formal DIL analogue, together with the operational reading of Corollary B.3 on the cross-domain within-class margin, remains open (§L.2).

N Deferred Definitions and Assumptions from §2

Definition N.1 (Class-IL scenario)**.** A *class-incremental learning* (class-IL) scenario consists of: (i) a partition of the label set into T disjoint tasks $\mathcal{Y} = \bigsqcup_{t=1}^{T} \mathcal{Y}^{(t)}$, each of cardinality $C := N/T$; (ii) a sequential training protocol in which, at stage t, the learner observes data only from classes in $\mathcal{Y}^{(t)}$; (iii) a test-time requirement to classify over the full label set $\mathcal{Y}$ without a task identifier.

Definition N.2 (Catastrophic Forgetting and Task Confusion)**.** Under discriminative class-IL training, at stage t the learner has access only to data from $\mathcal{Y}^{(t)}$ and therefore can only optimise $|P_{tt}(\theta)|$.

- **Catastrophic forgetting (CF)** occurs when $|P_{ss}(\theta_t)| > |P_{ss}(\theta_s)|$ for some earlier task $s < t$.
- **Task confusion (TC)** arises because the off-diagonal blocks $|P_{ij}(\theta)|$ for $i \neq j$ are never directly optimised during sequential training—no stage ever sees data from two different tasks simultaneously.

Assumption N.3 (Task incompatibility)**.** Continuously differentiable functions $f, g\colon \Theta \to \mathbb{R}$ with unique minimisers θ^f, θ^g are *incompatible*, written $f \nparallel g$, if (i) $\theta^f \neq \theta^g$, (ii) $\nabla f(\theta^f) = \nabla g(\theta^g) = 0$, and (iii) $\nabla g(\theta^f) \neq 0$ and $\nabla f(\theta^g) \neq 0$. We assume distinct tasks are pairwise incompatible: $|P_{ii}(\theta)| \nparallel |P_{jj}(\theta)|$ for all $i \neq j$.

O Cross-Entropy and the Pairwise-Loss Sandwich

The loss-matrix decomposition (9) is exact for *pairwise* losses—one-vs-one classification losses such as the pairwise hinge or pairwise log-loss—in which the loss on each sample (x, y) depends only on the score of the true class y and the score of one alternative class y' at a time. Modern class-IL methods, however, almost universally use the softmax cross-entropy loss

$$v_{\mathrm{CE}}(f_\theta(x), y) \;=\; -\log \frac{\exp(f_\theta(x)_y)}{\sum_{r=1}^{N} \exp(f_\theta(x)_r)} \;=\; -f_\theta(x)_y + \log \sum_{r=1}^{N} \exp(f_\theta(x)_r), \tag{69}$$

in which the partition function couples all N logits. We now show that despite this coupling, the cross-entropy loss is sandwiched between two pairwise log-losses up to a multiplicative factor of $\log N$, so that all bounds on pairwise losses transfer to cross-entropy with at most this constant degradation.

Pairwise log-loss

Define the *pairwise log-loss* between true class y and alternative class $y' \neq y$:

$$v_{\mathrm{pair}}^{(y,y')}(f_\theta(x)) \;:=\; \log\Big(1 + \exp\big(f_\theta(x)_{y'} - f_\theta(x)_y\big)\Big). \tag{70}$$

This is the binary logistic loss between the two classes y and y', evaluated only on their two scores. The sum over all alternatives,

$$V_{\mathrm{pair}}(f_\theta(x), y) \;:=\; \sum_{y' \neq y} v_{\mathrm{pair}}^{(y,y')}(f_\theta(x)),$$

is the natural pairwise analogue of the cross-entropy loss in (69): it sums binary log-losses over all class pairs involving the true label.

Lemma O.1 (Cross-entropy upper bound by pairwise loss)**.** *For any $\theta \in \Theta$, any $x \in \mathcal{X}$, and any $y \in \mathcal{Y} = \{1, \dots, N\}$:*

$$v_{\mathrm{CE}}(f_\theta(x), y) \;\le\; V_{\mathrm{pair}}(f_\theta(x), y). \tag{71}$$

Proof. Write $z_r := f_\theta(x)_r - f_\theta(x)_y$ for $r \neq y$, so that $z_r \in \mathbb{R}$ measures how much the logit of class r exceeds the logit of the true class. Cross-entropy can then be written

$$v_{\mathrm{CE}}(f_\theta(x), y) \;=\; \log\Big(1 + \sum_{r \neq y} \exp(z_r)\Big), \tag{72}$$

and the pairwise sum as

$$V_{\mathrm{pair}}(f_\theta(x), y) \;=\; \sum_{r \neq y} \log\big(1 + \exp(z_r)\big). \tag{73}$$

We use the elementary inequality

$$\log\Big(1 + \sum_{r=1}^{N-1} a_r\Big) \;\le\; \sum_{r=1}^{N-1} \log(1 + a_r) \qquad \text{for } a_r \ge 0, \tag{74}$$

which follows by induction on the number of terms: the base case $\log(1 + a + b) \le \log(1 + a) + \log(1 + b)$ reduces to $1 + a + b \le (1 + a)(1 + b) = 1 + a + b + ab$, which holds for $a, b \ge 0$. Applying (74) with $a_r = \exp(z_r)$ to (72) yields

$$v_{\mathrm{CE}}(f_\theta(x), y) \;\le\; \sum_{r \neq y} \log\big(1 + \exp(z_r)\big) \;=\; V_{\mathrm{pair}}(f_\theta(x), y),$$

which is (71). □

Consequence for the loss-matrix framework

Lemma O.1 has an immediate consequence: the expected cross-entropy risk and the expected pairwise risk are equivalent up to a multiplicative $\log N$ factor.

Corollary O.2 (Cross-entropy risk is bounded by pairwise risk)**.** *For any $\theta \in \Theta$ and uniform class prior, the population cross-entropy risk $I_{\mathrm{CE}}(\theta) := \mathbb{E}\, v_{\mathrm{CE}}(f_\theta(X), Y)$ and the population pairwise risk $I_{\mathrm{pair}}(\theta) := \sum_{1 \le k < l \le N} \rho_{kl}(\theta)$ satisfy*

$$\frac{N}{2} \cdot I_{\mathrm{CE}}(\theta) \;\le\; I_{\mathrm{pair}}(\theta). \tag{75}$$

Proof. By Lemma O.1, $v_{\mathrm{CE}}(f_\theta(x), y) \le V_{\mathrm{pair}}(f_\theta(x), y) = \sum_{r \neq y} \log(1 + e^{f_\theta(x)_r - f_\theta(x)_y})$ pointwise. Taking expectations under the uniform class prior, each unordered pair (k, l) contributes from both $y = k$ and $y = l$, so $\mathbb{E}\, V_{\mathrm{pair}}(f_\theta(X), Y) = (2/N) \cdot I_{\mathrm{pair}}(\theta)$. Combining yields $I_{\mathrm{CE}}(\theta) \le (2/N) \cdot I_{\mathrm{pair}}(\theta)$, which rearranges to (75). □

Remark O.3 (The upper bound is tight). The upper bound in (71) is tight when one alternative class dominates—i.e., when one z_r is much larger than all others—in which case $v_{\mathrm{CE}} \approx \log(1 + \exp(z_r)) \approx v_{\mathrm{pair}}^{(y,r)}$ and the sum V_{pair} is dominated by its largest term. In this regime, cross-entropy and pairwise loss agree up to a universal constant, and Corollary O.2 applies to cross-entropy with the same exponential decay rate as Theorem 3.4.

P Heavy-Tailed Concentration

Lemma P.1 (Heavy-tailed concentration)**.** *Fix $\theta \in \Theta$ and $\delta \in (0, 1)$. Let $\widehat{I}_t(\theta) := \frac{1}{n_t} \sum_{s=1}^{n_t} v(f_\theta(x_{t,s}), y_{t,s})$ denote the empirical risk on task t. Under Assumption A.1, there exists a constant $C_\alpha > 0$ depending only on α such that with probability at least $1 - \delta$,*

$$\big|\widehat{I}_t(\theta) - \mathbb{E}\, \widehat{I}_t(\theta)\big| \;\le\; C_\alpha \cdot V^{1/\alpha} \cdot \frac{(\log(2/\delta))^{1-1/\alpha}}{n_t^{1-1/\alpha}}. \tag{76}$$

The proof, given in §F.1, follows the truncated-mean argument of Catoni Catoni [2012]: define a clipped estimator $\widehat{I}_t^{\tau}$ with threshold τ, balance the bias of clipping (which scales as $V/\tau^{\alpha-1}$) against the sub-Gaussian concentration of the truncated variables (controlled by a Bernstein-type inequality), and optimise τ to yield (76). When $\alpha = 2$ the rate reduces to the familiar $O(\sqrt{\log(1/\delta)/n_t})$ Hoeffding rate; for $\alpha \to 1^+$ the rate degrades to nearly $n_t^0 = 1$, reflecting the difficulty of estimation under near-Cauchy tails.

Definition P.2 (Stochastic incompatibility)**.** Tasks i and j are (δ, n)-*incompatible* if, with probability at least $1 - \delta$ over draws of n samples per task, the empirical loss functions $|\widehat{P}_{ii}(\theta)|$ and $|\widehat{P}_{jj}(\theta)|$ satisfy the conditions of Assumption N.3 with their empirical minimisers in place of population minimisers.

As $n \to \infty$ for fixed δ, or as $\alpha \to 2^-$ for fixed n, stochastic incompatibility recovers the deterministic version. Throughout the body we work with population-level incompatibility (Assumption N.3); Lemma P.1 bounds the deviation between population and empirical risks at the iterate $\widehat{\psi}_T$ by (76), ensuring that for sample sizes typical of class-IL benchmarks ($n_t \geq 10^3$), the empirical TC gap concentrates around the population TC gap up to a term dominated by the optimisation error (§4).

Remark P.3 (Minimiser stability). Lemma P.1 bounds risk concentration at the specific iterate $\widehat{\psi}_T$, not minimiser stability. For heavy-tailed gradients with α close to 1, this distinction matters: there exist regimes where the empirical risk concentrates well while the empirical and population minimisers diverge significantly, particularly on near-flat loss landscapes where small perturbations of the loss can shift the minimiser substantially. The minimiser-stability question is a separate concern that our analysis does not address; we expect the gap to be small when Assumption 4.2 holds with a strong constant (curvature suppresses minimiser drift), and we treat formal verification as a direction for future work.

Q Strong Convexity for Modern Heads

For the head architectures used in modern pretrained-backbone class-IL—linear classifiers, prompt-pool attention Wang et al. [2022c], low-rank adapters McDonnell et al. [2023]—Assumption 4.1 is satisfied on a bounded parameter region with constants β that depend on the head and the bounded-feature constant B from Assumption 3.2(i).

Remark Q.1 (Strong convexity for prompt-pool architectures). For linear classifiers with ℓ_2-regularised heads (the configuration used in SLDA, RanPAC, EASE, InfLoRA, and the diagnostic-guided method of §5), μ-strong convexity follows directly from the regulariser. For prompt-pool architectures it requires (i) ℓ_2 regularisation on prompt embeddings and (ii) attention weights bounded below away from zero on the relevant parameter region; both hold for typical training regimes. The empirical E_{CF} rate of $K^{-0.31}$ measured in Table 1—close to the strongly-convex prediction of $K^{-1/2}$ at $\alpha = 2$ (at the measured $\widehat{\alpha} = 1.3$ the prediction is ≈ -0.23)—is consistent with the strongly-convex regime in our experiments.

R Comparison with Existing Heavy-Tailed Rates

The heavy-tailed clipped-SGD literature reports a range of convergence rates that, at first reading, appear inconsistent. We reconcile them here.

Three rate exponents in the literature. For stochastic gradient noise with bounded α-th moment ($\alpha \in (1, 2]$), the rates in the literature are:

- $K^{-(\alpha-1)/(3\alpha-2)}$ Zhang et al. [2020], Cutkosky and Mehta [2021]: non-convex SGD with squared-gradient-norm output.
- $K^{-(\alpha-1)/(\alpha+1)}$ (Proposition 4.3 restated for the non-convex case in Corollary G.2): non-convex SGD with uniform-iterate output and bounded-bias clipping analysis.
- $K^{-(\alpha-1)/\alpha}$ Sadiev et al. [2023] (Proposition 4.3 as stated in body): strongly convex last-iterate analysis.

Reconciliation. At $\alpha = 2$, the three exponents take the distinct values $1/4$, $1/3$, and $1/2$: the rates do not collapse to a common value even at $\alpha = 2$. The progression reflects the stronger structural

assumption: bounded-variance non-convex with squared-gradient-norm output ($K^{-1/4}$), bounded-variance non-convex with uniform-iterate output ($K^{-1/3}$), and strongly-convex last-iterate ($K^{-1/2}$). At $\alpha < 2$, the gap widens proportionally.

The rate $K^{-(\alpha-1)/\alpha}$ used as the headline in Proposition 4.3 corresponds to the strongly-convex regime appropriate for ℓ_2-regularised heads. The slower $K^{-(\alpha-1)/(\alpha+1)}$ rate of Corollary G.2 applies under the PŁ relaxation and is the rate empirically observed in the unclipped SGD experiments of §5 (Table 1: slope -0.10 unclipped vs -0.31 clipped; at $\widehat{\alpha} \approx 1.3$ the exponent pair is ≈ 0.13 vs ≈ 0.23, with the $-1/2$ reference of Figure 5 the $\alpha = 2$ value, and the unclipped comparison heuristic).

Why the literature appears inconsistent. The discrepancy in published rates is therefore not a contradiction but a reflection of: (i) which output statistic is bounded (squared gradient norm vs. function gap vs. last iterate), (ii) which structural assumption is invoked (smoothness vs. PŁ vs. strong convexity), and (iii) whether momentum is included. Our class-IL setting lies between the strongly-convex regime (when ℓ_2 regularisation makes L_t strongly convex) and the PŁ regime (when only the local PŁ inequality holds). Both cases yield informative bounds via Theorem 4.5 or its PŁ relaxation.

S Diagnostic Symptom-to-Remedy Table

The disjoint-dependency structure of Theorem 4.5 translates to the practitioner-facing diagnostic in Table 10.

Table 10: Diagnostic decomposition: which term dominates determines the remedy.

Symptom	**Dominant term**	**Remedy**
Poor accuracy with ample data	E_{TC}	Upgrade pretrained backbone (raise $\gamma_{\min}$)
Poor accuracy with strong $\gamma_{\min}$	E_{CF}	More iterations K_t, stronger clipping, smaller β
Heavy-tailed gradients (α small)	E_{CF}	Tighter clipping threshold τ_t
Long task sequence (T large)	both	Backbone with larger $\gamma_{\min}$; shorter task chunks

T Lipschitz Constants for Transformer Backbones

Assumption B.1 is satisfied by transformer architectures with bounded weights and bounded activations. The Lipschitz constant L_h is the operator-norm product of the attention and MLP layers, which is bounded under spectral normalisation, layer normalisation with bounded inputs, or during the bounded-norm training regime that all our experiments operate in. Concretely, for an L-layer transformer with attention spectral radius $\sigma_{\mathrm{att},\ell}$ and MLP Lipschitz constant $\sigma_{\mathrm{mlp},\ell}$ at layer ℓ, $L_h = \prod_{\ell=1}^{L}(\sigma_{\mathrm{att},\ell} + \sigma_{\mathrm{mlp},\ell})$ Kim et al. [2021]. We treat L_h as a known constant in Theorem B.2; §5 estimates it empirically for the three backbones we use.

U Exponent of L_h in the Drift Bound

The bound (13) is stated with L_h^2 in the prefactor. The proof in §H.1 produces L_h at first order in the perturbation expansion (parameter drift to feature drift, then feature drift to mean drift), with a quadratic-in-L_h contribution arising from the second-order interaction between feature drift and covariance perturbation in the Mahalanobis-margin response. At small drift magnitude $\|\Delta\mu^*\|$ the linear-in-L_h term dominates; at larger drift, the quadratic term contributes. We state the conservative L_h^2 form because it bounds both regimes; the constant c_2 absorbs the relative weighting and is not computed explicitly.

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 three contributions: the Relaxed Infeasibility bound (Theorem 3.4), the Excess Risk Decomposition (Theorem 4.5), and the Representation Drift bound (Theorem B.2) with Critical Horizon (Corollary B.3). Each is proved formally in the body and verified empirically across 12 (backbone, dataset) configurations in §5.

   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: §6 contains a dedicated Limitations paragraph covering the LDA/QDA scope of Theorem 3.4 (NCM and softmax cross-entropy heads lie outside the formal scope, Remark D.4), the reading of E_{TC} as an upper bound for softmax models, the locally Lipschitz treatment of the drift bound's L_h^2 prefactor, and the prefactor's quantitative looseness. Additional limitations appear in §L.2 (reference values for the diagnostic decomposition, strong-convexity and smoothness estimator coarseness, convergence-rate prefactor, Bayes-error prefactor), with the CF-optimal-reference and Hill-estimator caveats in §I.2 and §I.3, and the two boundary-regime configurations ($\nu_{ij} > 1/2$) flagged per-configuration in Table 4.

   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 and proved in the appendix: §E.1 proves Lemma 3.3; §E.2 proves Theorem 3.4; §G.1 proves Proposition 4.3; §G.4 proves Theorem 4.5; §H.1 proves Theorem B.2; §H.2 proves Corollary B.3.

   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: §I specifies all experimental protocols, including the Mahalanobis-margin computation (§I.1), the empirical TC-gap measurement (§I.2), the pre-registered Hill-estimator procedure (§I.3), and the Gaussianity verification (§I.4). Algorithm 1 provides the diagnostic-guided clipped-SGD procedure in pseudocode form. Backbone identifiers and dataset class splits follow the public PILOT/RanPAC benchmark suite.

   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-R, Split-CUB-200, Split-OmniBenchmark) are public via the PILOT/RanPAC benchmark suite. All pretrained backbones (ViT-B/16-IN21K, DINOv2-ViT-B/14, CLIP-ViT-B/16) are public via HuggingFace. Code and configuration files reproducing all headline experiments are released at `https://github.com/miladkhademinori/mahalanobis-margin` (code-availability sentence in §5).

   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: §I specifies the optimiser (clipped SGD), the learning-rate and clipping schedule (from Proposition 4.3), batch size, per-task training budget K_t, classifier head architecture, and seed protocol (5 random class orderings with seeds 0–4). Algorithm 1 documents the full procedure including diagnostic-stage measurements ($\widehat{\gamma}_{\min}, \widehat{\alpha}, \widehat{\beta}, \widehat{\mu}$).

   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: Tables 8, 2, and 7 report mean $\pm$ standard deviation across 5 random class orderings (seeds 0–4). Tables 5 and 6 and the slope column of Table 1 report point estimates with bootstrap 95% confidence intervals from 1,000 resamples; the E_{CF} cells of Table 1 are point values from the budget sweep, and the saturated-secant cells of Table 7 are quoted without dispersion per the table note. In Table 8, bold marks the column-best mean together with every method within one combined standard deviation (root sum of squares of the two cells' standard deviations) of it, as defined in the caption, and underline marks the best non-bold 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: Total compute on NVIDIA A100 (40 GB): $\widehat{\gamma}_{\min}$ measurements (24 GPU-hours), TC-gap experiments (45 GPU-hours), $\widehat{\alpha}$ Hill-estimator measurements (18 GPU-hours), 14-method comparison (120 GPU-hours), and drift trajectories (32 GPU-hours), totalling approximately 240 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 (following the Conclusion) discusses both directions: the positive impact of $\widehat{\gamma}_{\min}$ as a cheap, single-forward-pass signal that reduces wasted compute and lowers the resource barrier for backbone selection, and the negative considerations that the diagnostic inherits pretraining-corpus biases, depends on representative class statistics, and should be re-estimated under distribution shift. 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 §5 and used under their published licenses: ViT-B/16-IN21K (`google/vit-base-patch16-224-in21k`) and DINOv2-ViT-B/14 under Apache 2.0, and CLIP-ViT-B/16 (`openai/clip-vit-base-patch16`) under MIT. All benchmark datasets (Split-CIFAR-100, Split-ImageNet-R, Split-CUB-200, Split-OmniBenchmark) 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-R under the MIT-licensed dataset repository terms, with images collected from Wikimedia/Flickr under their respective licenses (non-commercial intent); Split-CUB-200 under Caltech's non-commercial research terms; Split-OmniBenchmark for research use per its repository. The PILOT/RanPAC toolchain is used under the MIT license of the LAMDA-PILOT toolbox, and the preprocessed CIL splits (RevisitingCIL repository, which carries no separate license file) are redistributed under the original dataset licenses. The Ledoit–Wolf shrinkage estimator is used in 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 is released in the paper's replication repository (URL per item 5): 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`, per-configuration files under `configs/`, and Colab notebooks under `colab/` constitute the documentation provided alongside the asset. 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.

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.