PROBABILISTIC INTERPRETATION OF CATASTROPHIC FORGETTING VIA SOLUTION DENSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In continual learning, standard empirical risk minimization (ERM) can severely degrade a neural network's performance on earlier tasks, which is a phenomenon known as catastrophic forgetting. We study a solution-density interpretation of this behavior: what follows when current-task learning no longer favors low-risk solutions for a distal component of prior data beyond a reference density, and how much prior-task mass can be described in this way? Our framework combines a measure-theoretic risk decomposition with a proximal/distal split relative to the current task. We formulate a solution-density selection hypothesis that bounds the ERM learner's probability of selecting low-risk distal solutions by their normalized parameter volume. Under this hypothesis and label-permutation symmetry, we derive a chance-level lower bound on expected distal risk and a corresponding lower bound on final mixture risk. Within this framework, we further derive an inverse-factorial upper bound of $1/r!$ on optimal solution density and selection probability, where r is the number of classes with positive distal mass. CIFAR-100 proof-of-concept experiments illustrate the framework across class-disjoint, blurry, and generalized class-incremental settings through empirical risk-CDF comparisons and non-vacuous plug-in risk bounds.

1 INTRODUCTION

Catastrophic forgetting refers to a severe deterioration in performance on previously learned tasks during subsequent learning. In sequential Empirical Risk Minimization (ERM), current-task data guide optimization, while knowledge of earlier tasks is carried forward through the model's parameters and optimization state. Even when a model can represent solutions that perform well on both past and current tasks, their existence alone does not determine whether sequential training selects them. We study this selection problem through the density of prior-task solutions in parameter space.

PAC-style transfer bounds relate error under distribution shift to source risk and distribution discrepancy, while PAC-Bayesian analyses provide generalization guarantees for lifelong and online learning (Ben-David et al., 2006; Pentina & Lampert, 2014; Haddouche & Guedj, 2022; Friedman & Meir, 2025). Even without distribution shift, multiclass PAC theory shows that ERM rules need not be equally effective: rules over the same hypothesis class can have different sample complexities (Daniely et al., 2011). These results highlight the relevance of both distributional relationships and the learning rule, while upper bounds alone do not determine a positive floor on prior-task error.

Loss-landscape analyses give a geometric view of this dependence on training. They relate forgetting to displacement from prior-task minima and local curvature (Mirzadeh et al., 2020). Relevantly, mode-connectivity studies examine low-loss paths between solutions, including those obtained by continual and multitask training (Garipov et al., 2018; Mirzadeh et al., 2021). Switching tasks changes the objective that guides optimization, so the model may leave low-loss regions of earlier tasks as it approaches solutions for the current one. The intersection of task-specific low-loss regions describes solutions that preserve performance across tasks.

Together, these perspectives motivate a question about the interaction between solution geometry and learning behavior: *how much of parameter space supports good prior-task performance, and how much probability does current-task ERM assign to these regions?* Our solution-density interpretation connects these quantities to derive an error floor for sequential learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Consider a learner starting at a low-risk solution for an earlier task. Under a substantial task shift, current-task optimization can move it away from that region without directly rewarding a return. If the advantage inherited from prior learning is eroded, we use solution density as a reference for the probability of selecting a low-risk region. Negative transfer, including suppression of classes absent from the current task, can further favor higher prior-task risk.

To accommodate overlapping tasks, our framework combines a measure-theoretic decomposition of prior-task risk with an ε-proximal/distal split relative to the current-task support under a task-relevant geometry. On a distal component with positive mass, our solution-density hypothesis (Hypothesis 2) posits that the learner assigns no more probability to any sublevel risk set than the normalized volume reference measure ρ does. This stochastic ordering links the distribution of selected risks to the volume occupied by prior-task solution sets.

Under this hypothesis, label-permutation symmetry yields a chance-level lower bound on expected distal risk. Weighting this floor by distal mass and the prior-task mixture weight gives an irreducible error term in the final expected risk under the selection model. This term scales with the distal mass and vanishes when that mass becomes zero. The framework also yields an inverse-factorial upper bound $1/r!$ on the optimal solution density and selection probability, where r is the number of classes with positive distal mass. This bound follows by counting disjoint, equal-volume label-permutation copies of the optimal solution set.

Our CIFAR-100 proof-of-concept experiments illustrate the quantitative significance of the distal error term in class-disjoint, blurry, and generalized class-incremental learning. We observe non-vacuous distal components whose empirical risk CDFs are consistent with the selection hypothesis. The corresponding plug-in floors amount to 67.7%-91.1% of measured final error across the six protocols (Table 1). These observations show that the distal error term can remain quantitatively substantial even when task supports overlap.

The primary contributions of our work are as follows:

- We formulate a solution-density framework for sequential learning, linking prior-task risk to parameter selection via measure-theoretic risk decomposition and ε-proximal/distal split.
- We derive a lower bound on expected final risk that scales with distal mass, together with an inverse-factorial upper bound on optimal solution density and selection probability.
- We provide a CIFAR-100 proof-of-concept across class-disjoint and overlapping task sequences, observing non-vacuous distal components whose plug-in risk floors represent a substantial fraction of measured final error.

2 Preliminaries

2.1 Basic Notations

Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ denote the data space consisting of input space $\mathcal{X} \subset \mathbb{R}^m$ and label space $\mathcal{Y} = \{1, 2, \ldots, K\}$. P_i denotes the data distribution belonging to task i.[1] Given a model $f(\cdot; \theta)$ with model parameter θ in parameter space $\Theta \subset \mathbb{R}^p$, the true risk on task i is defined as:

$$R(\theta; P_i) := \mathbf{E}_{(x,y)\sim P_i} L(f(x; \theta), y),$$

where L is a nonnegative loss function. In classical statistical learning theory (Vapnik, 1999; Natarajan, 1989; Shalev-Shwartz & Ben-David, 2014), 0–1 loss is commonly adopted, resulting in:

$$R(\theta; P_i) = P_i(\{f(x; \theta) \neq y\}).$$

While practical optimization employs differentiable surrogates (e.g., cross-entropy), the ultimate classification goal is minimizing the misclassification rate (Mao et al., 2023). We thus focus on the 0–1 loss to analyze the solution space, independent of surrogate-specific geometries. For simplicity, we write $L(\theta, x, y)$ or $L(\theta, z)$ for the $L(f(x; \theta), y)$ when the context is clear.

We write $N \in \mathbb{N}$ for the total number of tasks and $[N] = \{1, 2, \ldots, N\}$. Let $\mathcal{T}$ denote a distribution over task sequences, and write $(P_1, \ldots, P_N) \sim \mathcal{T}$. For $n \in [N]$, we define the overall distribution

[1] Formally, $(\mathcal{X} \times \mathcal{Y}, \Sigma, P_i)$ forms a probability space for task i, where Σ is some σ-algebra on $\mathcal{X} \times \mathcal{Y}$.

up to task n as $P_{1:n} := \sum_{i=1}^{n} w_i P_i$, where $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$, and its corresponding risk as $R_{1:n}(\theta) := R(\theta; P_{1:n})$. Unless stated otherwise, we set $w_i = 1/n$, i.e. $P_{1:n} = \frac{1}{n}\sum_{i=1}^{n} P_i$. Similarly, let $\mathcal{Y}_i \subseteq \mathcal{Y}$ represent the label subset associated with task i, and define $\mathcal{Y}_{1:n} := \bigcup_{i=1}^{n} \mathcal{Y}_i$ as the labels covered by tasks 1 through n, with $\mathcal{Y}_{1:N} = \mathcal{Y}$. We use λ for Lebesgue measure in the relevant Euclidean dimension and $\mathcal{L}$ for the associated σ-algebra. All sets and functions used in probability and integration statements are assumed measurable with respect to the stated σ-algebras.[2]

2.2 Principal Theorems

Lebesgue's decomposition theorem. (Folland, 1999) For any measurable space (Ω, Σ) and any pair of σ-finite measures μ and ν, there exist unique σ-finite measures μ_a and μ_s satisfying

$$\mu = \mu_a + \mu_s, \qquad \mu_a \ll \nu, \qquad \mu_s \perp \nu.$$

μ_a is absolutely continuous with respect to ν, meaning μ_a has the shared mass with ν. Meanwhile, μ_s is singular to ν; that is, μ_s accounts for the alien mass relative to ν.

Radon-Nikodym theorem. (Folland, 1999) Let μ and ν be σ-finite measures on the measurable space (Ω, Σ). If μ is absolutely continuous with respect to ν ($\mu \ll \nu$), then there exists a non-negative measurable function $h : \Omega \to [0, \infty)$ such that for any set $E \in \Sigma$, we have

$$\mu(E) = \int_E h \, d\nu.$$

The function h is called the Radon-Nikodym derivative of μ with respect to ν, is denoted by $\frac{d\mu}{d\nu}$, and is unique up to a ν-null set (i.e., any other such function g will have $h = g$, ν-almost everywhere).

For probability measures P and Q on the same space (Ω, Σ), write $P = P_a + P_s$ for the Lebesgue decomposition of P with respect to Q. The Radon-Nikodym theorem then gives, for every $E \in \Sigma$,

$$P(E) = \int_E \frac{dP_a}{dQ} \, dQ + P_s(E).$$

2.3 Probabilistic Modeling of ERM

A probability measure on parameter space captures both the possible outputs of ERM and their selection probabilities. Let π_i^* denote the output distribution of a sequential ERM procedure after task i on $(\Theta, \mathcal{L}(\Theta))$, induced by initialization, data sampling, and optimization randomness. Then, the expected true risk of ERM on task i can be formulated as: $\mathbf{E}_{\theta\sim\pi_i^*} R(\theta; P_i) = \mathbf{E}_{\theta\sim\pi_i^*}\mathbf{E}_{z\sim P_i} L(\theta, z)$.

We study the expected risk after learning task j, evaluated on an earlier task $i < j$: $\mathbf{E}_{\theta\sim\pi_j^*} R(\theta; P_i)$.

By applying the Lebesgue decomposition to P_i with respect to P_j, we can isolate the absolutely continuous part $P_{i\ll j}$ and the singular part $P_{i\perp j}$. We quantify the singular mass as $m_{i\perp j} := P_{i\perp j}(\mathcal{Z})$ and, when $m_{i\perp j} > 0$, define the normalized singular distribution as $\tilde{P}_{i\perp j} := m_{i\perp j}^{-1} \cdot P_{i\perp j}$. (*Throughout this work, $\tilde{P}$ denotes the normalization of a sub-probability measure P with positive mass. A weighted risk contribution from a component of zero mass is understood to be zero.*) Consequently, the expected true risk of ERM for task j over distribution P_i can be rewritten as:

$$\begin{aligned}\mathbf{E}_{\theta\sim\pi_j^*} R(\theta; P_i) &= \mathbf{E}_{\theta\sim\pi_j^*}\mathbf{E}_{(x,y)\sim P_i} L(\theta, x, y) \\ &= \underbrace{\mathbf{E}_{\theta\sim\pi_j^*}\mathbf{E}_{(x,y)\sim P_j} h_{ij}(x,y)\, L(\theta, x, y)}_{\text{absolutely continuous error}} + \underbrace{m_{i\perp j}\, \mathbf{E}_{\theta\sim\pi_j^*}\mathbf{E}_{(x,y)\sim \tilde{P}_{i\perp j}} L(\theta, x, y)}_{\text{singular error}},\end{aligned}$$

where $h_{ij} = dP_{i\ll j}/dP_j$ is the Radon-Nikodym derivative of $P_{i\ll j}$ with respect to P_j.

The decomposition yields a canonical split into absolutely continuous and singular risk contributions. A lower bound on the expected risk of the normalized singular component therefore yields a lower bound on prior-task risk, weighted by the singular mass.

[2] Non-measurable pathological examples such as Vitali set are excluded from analysis. Appendix B details.

2.4 THE ε-PROXIMAL DECOMPOSITION

Mutual singularity can arise from arbitrarily small perturbations that preserve task-relevant semantics, so singularity alone need not indicate a loss of transferability. We therefore use a metric reflecting task-relevant similarity and a scale ε to isolate prior-task mass beyond the local neighborhood of current-task data. This distal component identifies where proximity to current-task data no longer provides a basis for expecting successful transfer.

Proposition 1 (ε-proximal/distal decomposition)**.** *Let (X, d) be a metric space, let Σ be a σ-algebra containing the d-Borel sets, and let μ, ν be σ-finite measures on (X, Σ). For any $\varepsilon > 0$, there exists a unique decomposition of μ into two mutually singular measures $\mu = \mu_{prox} + \mu_{dist}$ satisfying:*

- μ_{prox} ***(The Proximal Part):*** *The measure μ_{prox} is concentrated on the open ε-neighborhood of* $\operatorname{supp}(\nu)$*:*
$$\mu_{prox}(\{x \in X : d(x, \operatorname{supp}(\nu)) \geq \varepsilon\}) = 0.$$
- μ_{dist} ***(The Distal Part):*** *The measure μ_{dist} is concentrated on the region at distance at least ε from* $\operatorname{supp}(\nu)$*:*
$$\mu_{dist}(\{x \in X : d(x, \operatorname{supp}(\nu)) < \varepsilon\}) = 0.$$

Here, $d(x, S) = \inf_{y \in S} d(x, y)$ is the distance between the point $x \in X$ and the set $S \subseteq X$.

Proof. See Appendix C. □

For task distributions, we apply this decomposition with $\mu = P_i$ and $\nu = P_j$ on $\mathcal{Z}$.

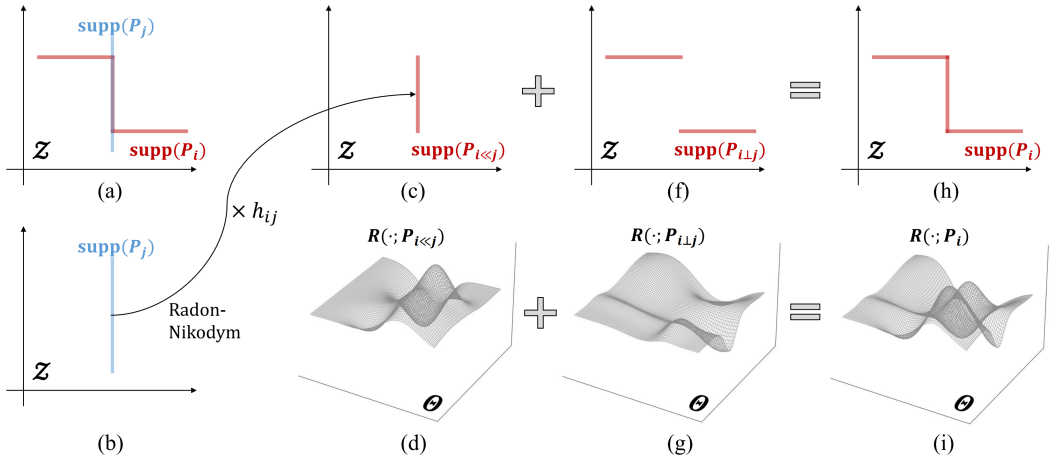


Figure 1: Lebesgue decomposition of P_i with respect to P_j and the induced risk split. P_i (a) uniquely decomposes into an absolutely continuous part (c) and a singular part relative to P_j (f), summing up to the original measure (h). The absolutely continuous part $P_{i \ll j}$ can be represented through the Radon-Nikodym derivative $h_{ij} = dP_{i \ll j}/dP_j$ (b). This yields a canonical split of risk into absolutely continuous and singular contributions (d, g, i).

3 PROBABILISTIC MODELING OF ERM IN CONTINUAL LEARNING

Building on Section 2, we relate the density of distal solutions to their selection probability under standard sequential ERM. We derive bounds on expected risk and optimality probability, extending the latter to the full task mixture when its proximal and distal components share an optimum.

Assumption 1. *The effective parameter space $\Theta \subset \mathbb{R}^p$ is bounded with $0 < \lambda(\Theta) < \infty$.*[3]

[3] In the lazy-training regime, the effective parameter space is de facto bounded. See Appendix B for details.

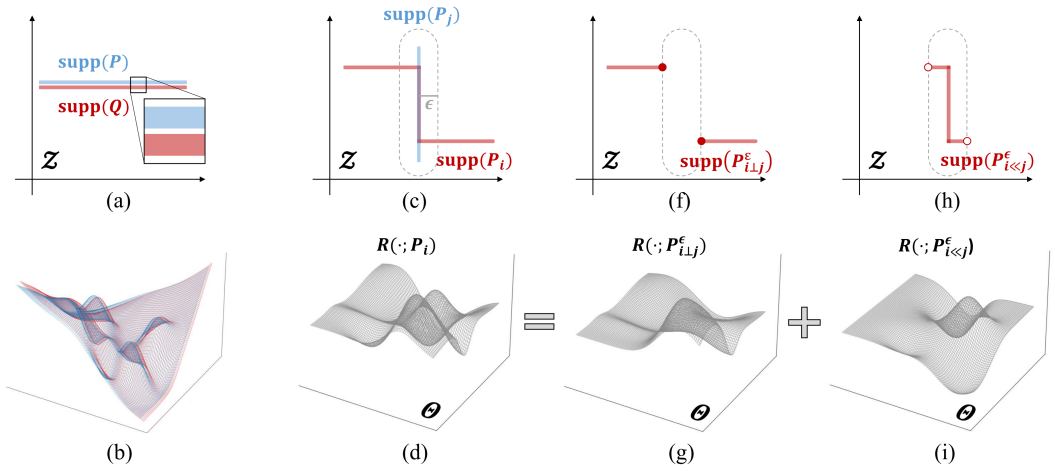


Figure 2: Motivation for the ε-proximal/distal decomposition. Even arbitrarily close distributions may be singular (a), yet nearby distributions may still admit a common low-risk predictor (b). A metric threshold ε separates proximal and distal prior-task mass relative to the current-task support (c–i). This decomposition is metric-dependent and is not intended as a limiting form of the standard Lebesgue decomposition as $\varepsilon \to 0$.

Assumption 2. *The effective parameter space Θ is closed under last-layer permutations, and the predicted labels permute accordingly almost everywhere.*[4]

- Let $\Theta = \Theta_{\text{body}} \times \Theta_{\text{last}}$ with $\Theta_{\text{last}} = \prod_{i=1}^{k} \Theta_{\text{last}}^{(i)}$, where $\Theta_{\text{last}}^{(i)}$ is the parameter block of the last layer associated with class i (e.g., the i-th row of the classifier and its bias) and k is the number of output classes.
- Define an action of the symmetric group S_k on Θ by

$$(\sigma,\ (\theta_{\text{body}}, (\theta_{\text{last}}^{(i)})_{i=1}^{k})) \ \longmapsto\ \big(\theta_{\text{body}},\ (\theta_{\text{last}}^{(\sigma^{-1}(i))})_{i=1}^{k}\big).$$

  We say that Θ is *closed under last-layer permutations* if for every $\sigma \in S_k$ and every $\theta \in \Theta$, the permuted parameter tuple also lies in Θ.

Remark (Symmetry of representation and measure). Assumption 2 implies two crucial properties for the geometry of the solution space:

1. **Representational Equivalence:** For any parameter θ representing a function $f(\cdot;\theta)$ and any permutation σ, there exists a $\theta' = \sigma \cdot \theta$ such that the output logits are exactly permuted. Thus, the functional capacity for class i is the same as for class j.
2. **Volume Preservation:** The permutation action is a volume-preserving isometry. Equivalently, for any $E \in \mathcal{L}(\Theta)$ and $\sigma \in S_k$, $\lambda(\sigma \cdot E) = \lambda(E)$. Every permuted copy of the same parameter region has the identical Lebesgue measure.

Assumption 3. *For any evaluated distal probability measure P on $(\mathcal{Z}, \mathcal{L}(\mathcal{Z}))$ and any optimal parameter $\theta^* \in \Theta$ for P, the true class label is the unique mode of the class-conditional prediction distribution. Formally, for all $y \in \mathcal{Y}$ with $P(Y = y) > 0$ and every output class $c \neq y$:*

$$\Pr_P(f(x;\theta^*) = y \mid Y = y) > \Pr_P(f(x;\theta^*) = c \mid Y = y).$$

Remark (Learnability of the Optimal Parameter). This assumption posits that the best possible model within the hypothesis class (θ^*) is strictly more correlated with the true label than with any specific incorrect label. This acts as a classwise condition for learnability, excluding scenarios where feature collapse or severe model under-capacity renders the true label statistically indistinguishable from noise. The assumption does not require high accuracy and only requires that the optimal parameter breaks the symmetry between the correct class and incorrect classes.

[4]The symmetry holds for standard classification architectures (e.g. CNN, ViT, ResNet). Appendix B details.

3.1 Solution Set and Set Density

Definition 1 (Solution Set and Optimal Solution Set). *For a probability measure P on $(\mathcal{Z}, \mathcal{L}(\mathcal{Z}))$ and tolerance $\tau \in [0, 1]$, the solution set $\mathcal{S}(\tau; P)$ and the optimal solution set $\mathcal{S}^*(P)$ are:*

$$\mathcal{S}(\tau; P) := \{\theta \in \Theta : R(\theta; P) \le \tau\}, \quad \mathcal{S}^*(P) := \mathcal{S}(m; P)$$

where $m := \inf_{\theta \in \Theta} R(\theta; P)$.

Definition 2 (Set Density). *For all $E \in \mathcal{L}(\Theta)$, the density of E is defined as:*

$$\rho(E) := \lambda(E) \,/\, \lambda(\Theta).$$

Remark (Density-probability duality). Under uniform reference sampling, the probability of selecting $\mathcal{S}(\tau; P)$ equals its volume fraction $\rho(\mathcal{S}(\tau; P))$. We compare this density with the learner's probability $\pi_j^*(\mathcal{S}(\tau; P))$; on distal data, our hypothesis posits that the latter is no greater.

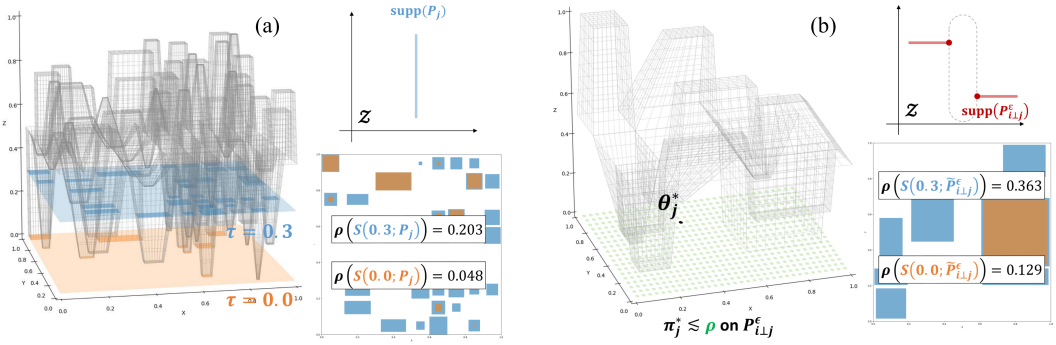


Figure 3: A solution-density view of forgetting. (a) The solution set $\mathcal{S}(\tau; P_j)$ expands with tolerance τ; its volume fraction gives the reference probability of selecting an acceptable solution. (b) On distal data $\tilde{P}_{i\perp j}^{\varepsilon}$, the selection hypothesis bounds ERM's success probability by solution density, modeling the loss of its inherited selection advantage.

3.2 Distal Risk and Optimality Bounds

Current-task optimization can move the learner away from regions of low prior-task risk. We model the resulting loss of inherited advantage on distal data by comparing selection probability with solution density.

At stage j, we consider the distal components $P_{i\perp j}^{\varepsilon}$, $i < j$, and $P_{(1:j-1)\perp j}^{\varepsilon}$ of individual past tasks and their accumulated mixture relative to P_j, evaluated under the same learner law π_j^*.

Hypothesis 2 (Solution-Density Selection). *For the specified sequential ERM procedure and task sequence, for each $1 \le i < j \le N$ there exist a metric d on $\mathcal{Z}$ and a threshold $\varepsilon > 0$ such that, whenever $P_{i\perp j}^{\varepsilon}(\mathcal{Z}) > 0$,*

$$\pi_j^*(\mathcal{S}(\tau; \tilde{P}_{i\perp j}^{\varepsilon})) \le \rho(\mathcal{S}(\tau; \tilde{P}_{i\perp j}^{\varepsilon})) \qquad \text{for all } \tau \in [0, 1].$$

The same statement holds with P_i replaced by $P_{1:j-1}$, with its own d, ε and the same learner law π_j^*.

Remark (Selection on distal data). Equality in Hypothesis 2 represents the loss of inherited selection advantage, matching the reference success rate without making the learner's parameter distribution uniform. Negative transfer can further reduce the chance of finding low-risk distal solutions, motivating strict inequality. For example, in class-incremental training with softmax cross-entropy, new-class samples generate gradients that suppress old-class logits and bias predictions toward recent classes (Ahn et al., 2021). Under the 0–1 evaluation loss, distal risk lies in $[0, 1]$, and these solution-set probabilities are the corresponding risk CDF values. Overlap can leave zero distal mass, yielding a zero contribution to the final-risk bound. Section 3.3 reports nonzero distal mass in both class-disjoint and overlapping CIFAR-100 settings.

We first translate the selection hypothesis into an expected-risk bound.

Theorem 3 (Chance-Level Accuracy). *For $1 \le i < j \le N$, choose d, ε as in Hypothesis 2 for this comparison. If $P^{\varepsilon}_{i\perp j}(\mathcal{Z}) > 0$,*

$$\mathbf{E}_{\theta\sim\pi^*_j} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) \ge 1 - |\mathcal{Y}_{1:j}|^{-1}.$$

The same bound holds for $P_{1:j-1}$ with its own choice and positive distal mass.

Proof Sketch (Appendix D). Last-layer permutations preserve reference volume and give every output class the same reference prediction probability. Thus, the expected risk under ρ is exactly $1 - 1/k$ for any evaluated probability distribution. Hypothesis 2 orders the distal risk CDFs as $F_{\pi^*_j} \le F_\rho$; integrating their survival functions gives an ERM expected risk of at least $1 - 1/k$. Since $k \ge |\mathcal{Y}_{1:j}|$, this yields the stated bound. The argument applies to both individual past tasks and the accumulated past mixture.

Corollary 4 (Lower Bound on Expected Risk). *Choose d, ε as in Hypothesis 2 for $P_{1:N-1}$ relative to P_N. For the overall mixture $P_{1:N}$,*

$$\mathbf{E}_{\theta\sim\pi^*_N} R(\theta; P_{1:N}) \ge \frac{N-1}{N} \cdot m^{\varepsilon}_{(1:N-1)\perp N} \cdot (1 - |\mathcal{Y}|^{-1}),$$

where $m^{\varepsilon}_{(1:N-1)\perp N} := P^{\varepsilon}_{(1:N-1)\perp N}(\mathcal{Z})$ is the distal mass of the past mixture $P_{1:N-1}$ with respect to P_N.

Proof Sketch (Appendix E). Discarding the nonnegative current-task and proximal contributions leaves the past distal risk weighted by $(N-1)/N$ and $m^{\varepsilon}_{(1:N-1)\perp N}$. Apply Theorem 3 to the normalized distal component of $P_{1:N-1}$; zero distal mass gives the bound directly.

Remark (From distal chance level to global lower bound). Distal mass scales the contribution of the chance-level floor to final expected risk. The bound concerns final risk; a forgetting increment would instead compare risk before and after learning a task.

We next bound optimal solution density by counting the distinct output-label permutations of the optimal set. For an empirical distal distribution, r is simply the number of distinct classes observed among the retained distal samples.

Theorem 5 (Prior-Task Optimality Probability). *For $1 \le i < j \le N$, choose d, ε as in Hypothesis 2 for this comparison. With positive distal mass, let r count the classes with positive mass in $\tilde{P}^{\varepsilon}_{i\perp j}$. Then*

$$\pi^*_j(\mathcal{S}^*(\tilde{P}^{\varepsilon}_{i\perp j})) \le \frac{1}{r!}.$$

The same result holds for $P_{1:j-1}$ with its own choice, with r counting its positive-mass distal classes.

Proof Sketch (Appendix F). If the optimal set is empty, the bound is immediate. The group S_k contains $k!$ output-label permutations. Rearranging only the $k - r$ zero-mass labels preserves the optimal set, so each distinct copy is counted $(k-r)!$ times. By Assumption 3, moving a positive-mass label makes an optimal parameter non-optimal for the original labels, so different assignments of these labels yield disjoint copies. There are therefore $k!/(k-r)!$ equal-volume copies in Θ, all but the original lying in its complement. The optimal set thus has density at most $(k-r)!/k! \le 1/r!$. Hypothesis 2, applied at the optimal-risk threshold, transfers this density bound to the learner's selection probability.

To transfer Theorem 5 to the global objective, choose d, ε for $P_{1:N-1}$ relative to P_N as in Hypothesis 2. Using this choice, decompose the uniform mixture $P_{1:N}$ with respect to P_N:

$$P^{\varepsilon}_{\text{dist}} := P^{\varepsilon}_{(1:N)\perp N}, \quad P^{\varepsilon}_{\text{prox}} := P^{\varepsilon}_{(1:N)\ll N}.$$

The current task contributes only to the proximal component, so $P^{\varepsilon}_{\text{dist}} = \frac{N-1}{N} P^{\varepsilon}_{(1:N-1)\perp N}$ comes entirely from the past mixture. Let $m := P^{\varepsilon}_{\text{dist}}(\mathcal{Z}) > 0$. Since $m \le (N-1)/N < 1$, both components can be normalized. In particular, $\tilde{P}^{\varepsilon}_{\text{dist}} = \tilde{P}^{\varepsilon}_{(1:N-1)\perp N}$, placing the full distal component within the scope of Hypothesis 2.

Corollary 6 (Global Optimality Probability). *For the positive-mass decomposition above, let r count the classes with positive mass in $\tilde{P}^{\varepsilon}_{dist}$. If $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{dist}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{prox}) \ne \emptyset$,*

$$\pi^*_N(\mathcal{S}^*(P_{1:N})) \le \frac{1}{r!}.$$

Proof Sketch (Appendix G). Linearity of risk, positive mixture weights, and a common optimum give $\mathcal{S}^*(P_{1:N}) = \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$. Thus global optimality implies distal optimality under the same π^*_N. Theorem 5, applied to the accumulated past mixture, then yields the bound.

Remark (Solution density and sequential selection). The solution-density interpretation separates the existence of a common solution from the probability that sequential ERM finds it. Hypothesis 2 bounds sequential ERM's probability of selecting distal optima by their solution density under ρ. Corollary 6 extends this inverse-factorial selection bound to global optima when the two components share an optimum.

3.3 Empirical Diagnostics

Our proof-of-concept asks whether substantial components of past data exhibit the selection relation in Hypothesis 2 and yield a quantitatively meaningful final-risk floor. We train ResNet-32 by standard sequential ERM on CIFAR-100 under class-disjoint class-incremental learning (CIL), Blurry1 and Blurry2, and the task-varied, uniform, and long-tail gCIL protocols (Rebuffi et al., 2017; Bang et al., 2021; Mi et al., 2020). CIL provides a transparent construction through label separation; Blurry and gCIL examine whether substantial distal components persist when tasks overlap, using DINOv2 features (Oquab et al., 2024) to define the data geometry.

For each protocol, we fix one reference box from coordinatewise minima and maxima over all ERM initializations and task checkpoints across seeds. Only the output-layer bounds are pooled across classes, giving corresponding coordinates identical intervals and ensuring closure under label permutations. This anchors ρ to parameter ranges encountered during learning and keeps the reference fixed across comparisons. We evaluate the ERM models and uniform reference samples on the same distal data, from either an earlier task or the mixture of all earlier tasks. Their risk CDFs estimate the learner's selection probabilities and the reference solution densities in Hypothesis 2. We select the smallest tested ε with positive mass and non-rejection by the one-sided Kolmogorov–Smirnov (KS) diagnostic, retaining the largest eligible mass. Sampling and evaluation details are in Appendix H.

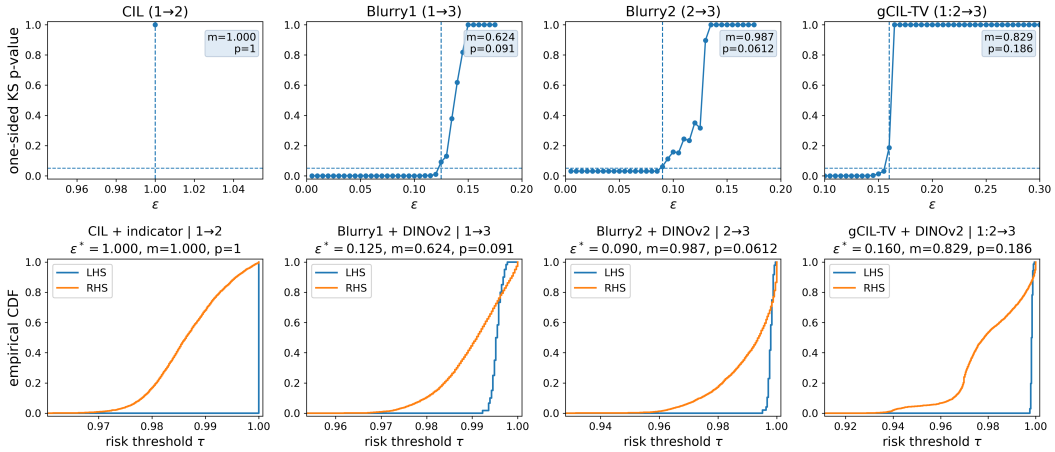


Figure 4: Comparing ERM selection with solution density on past data. In $i \to j$, distal data from task i are evaluated after sequential training through task j. From left to right, the task pairs (i, j) are $(1, 2)$ for CIL, $(1, 3)$ for Blurry1, and $(2, 3)$ for Blurry2; gCIL-TV evaluates the uniform mixture of tasks 1 and 2 after task 3, denoted $1{:}2 \to 3$. Top: KS p-values over ε, with dashed lines at ε^* and $p = 0.05$. Bottom: empirical risk CDFs for ERM (LHS) and uniform reference samples (RHS) on the selected distal component. These diagnostics probe whether ERM selects low-risk solutions on distal data with probability no greater than their reference density (Hypothesis 2).

The CDF comparisons examine the selection relation, while retained mass determines its contribution to final risk. For the past mixture, Table 1 combines this mass with the chance-level factor from Theorem 3 and the mixture weight in Corollary 4. Error coverage is the plug-in floor divided by final risk. Figures 5 and 6 show all selected comparisons and sampled rejected thresholds.

Table 1: CIFAR-100 diagnostics after the final task. Past mass is the fraction of the uniform mixture $P_{1:N-1}$ retained as distal to P_N using one selected threshold. Final risk averages 0–1 risk over all tasks and ERM runs. The plug-in floor combines retained mass with the past-task weight $(N-1)/N$ and the chance-level factor in Corollary 4; error coverage expresses this floor as a percentage of final risk. CIL uses label separation; Blurry and gCIL use DINOv2. See Appendix H and Table 2 for construction and selection details.

Construction	CIL Indicator	Blurry1 DINOv2	Blurry2 DINOv2	gCIL-TV DINOv2	gCIL-U DINOv2	gCIL-L DINOv2
Past mass $\widehat{m}^{\varepsilon}_{(1:N-1)\perp N}$	1.000	0.994	0.731	0.852	0.815	0.805
Final risk $\widehat{R}(\theta; P_{1:N})$	0.725	0.720	0.713	0.857	0.816	0.812
Plug-in floor (Cor. 4)	0.660	0.656	0.483	0.628	0.600	0.593
Error coverage	91.05%	91.08%	67.70%	73.35%	73.56%	73.09%

Mean ERM risks on selected distal components exceed the chance-level factors in Theorem 3 (Table 2). Across Blurry and gCIL, DINOv2 retains 73.1%–99.4% of the past mixture at the final stage. With label separation for CIL, these constructions yield plug-in floors amounting to 67.7%–91.1% of measured final error. The ordering diagnostics identify components with non-negligible mass for which the floor is on the scale of the observed error, including when tasks overlap.

4 Related Work

Learning-theoretic studies analyze generalization and transfer through domain-adaptation bounds and PAC-Bayesian guarantees for lifelong and online learning (Ben-David et al., 2006; Pentina & Lampert, 2014; Haddouche & Guedj, 2022). Analyses of linear regression quantify the roles of task similarity, task order, and overparameterization in forgetting (Lin et al., 2023; Evron et al., 2022; Goldfarb et al., 2024), while Friedman & Meir (2025) derive data-dependent and oracle upper bounds on forgetting. For class-incremental learning, structural analyses characterize the role of out-of-distribution detection in learnability (Kim et al., 2022; 2023). Our focus is the prior-task risk floor obtained by linking solution density to the probabilities assigned by sequential ERM.

Loss-landscape studies examine local curvature and connectivity between learned solutions (Mirzadeh et al., 2020; Garipov et al., 2018; Mirzadeh et al., 2021; Wen et al., 2023; Bian et al., 2024). The existence and connectivity of low-loss solutions leave open how much reference volume they occupy and how likely sequential training is to select them. Methods such as EWC and GEM preserve earlier performance through parameter-importance penalties and gradient constraints, respectively (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017). These interventions also motivate examining how the learning rule influences selection among prior-task solutions.

Hidden-neuron permutations can preserve network function and reveal equivalent minima and symmetry-induced critical points (Brea et al., 2019; Simsek et al., 2021). We instead permute output classes against fixed labels, using disjoint, equal-volume non-optimal copies of the full optimal set to obtain an inverse-factorial bound on its reference density. Appendix I gives further discussion.

5 Conclusion

We interpret forgetting in standard sequential ERM through a hypothesis linking solution density to selection probability. The framework yields a lower bound on final risk weighted by distal mass and an inverse-factorial bound on distal optimality probability. CIFAR-100 diagnostics identify distal components whose plug-in risk floors amount to a substantial fraction of measured final error across class-disjoint and overlapping tasks. The central distinction is between the existence of solutions shared across tasks, their reference density, and the probability that sequential training selects them. This perspective suggests studying how continual-learning methods preserve a selection advantage on past data: replay supplies prior-task supervision, while regularization alters the learning objective. A key direction is to characterize when sequential learning dynamics preserve or erode this advantage.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, support qualitative and thematic data analysis, interpret results, suggest experimental parameters, discover research topics or identify gaps, brainstorming, format references, suggest a structure for a research paper, propose a title or keywords for a research paper. and generate synthetic data sets, clean and reformat dataset, formulate questions for surveys or interviews, and transcribe recordings of research material are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. All the writings in the paper have been checked for hallucinations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work provides a theoretical analysis of forgetting with experiments on the public CIFAR-100 benchmark using pretrained DINOv2 features. It involves no human-subject study or collection of new personal data.

REPRODUCIBILITY STATEMENT

Appendix B explains the assumptions, and Appendices C–G provide the proofs. Appendix H documents the data protocols, training and evaluation settings, reference box construction and sampling, threshold selection, and calculation of the reported statistics.

REFERENCES

Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *International Conference on Computer Vision (ICCV)*, pp. 824–833. IEEE, 2021.

Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8218–8227, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Neural Information Processing Systems (NIPS)*, volume 19, 2006.

Ang Bian, Wei Li, Hangjie Yuan, Chengrong Yu, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, and Tao Feng. Make continual learning stronger via c-flat. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 7608–7630. Curran Associates, Inc., 2024. doi: 10.52202/079017-0244.

Johanni Brea, Berfin Simsek, Bernd Illing, and Wulfram Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *arXiv preprint arXiv:1907.02911*, 2019.

Xi Chen, Christos Papadimitriou, and Binghui Peng. Memory bounds for continual learning. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 519–530, 2022.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In Sham M. Kakade and Ulrike von Luxburg (eds.), *Conference on Learning Theory (COLT)*, volume 19 of *Proceedings of Machine Learning Research*, pp. 207–232, Budapest, Hungary, 09–11 Jun 2011. PMLR.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Po-Ling Loh and Maxim Raginsky (eds.), *Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 02–05 Jul 2022.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Lior Friedman and Ron Meir. Data-dependent and oracle bounds on forgetting in continual learning. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2025.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting–an analytical model. In *International Conference on Learning Representations (ICLR)*, 2024.

Maxime Haddouche and Benjamin Guedj. Online pac-bayes learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 25725–25738. Curran Associates, Inc., 2022.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 5065–5079. Curran Associates, Inc., 2022.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *International Conference on Machine Learning (ICML)*, pp. 16877–16896. PMLR, 2023.

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

Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In *International Conference on Machine Learning*, pp. 5327–5337. PMLR, 2020.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100. PMLR, 23–29 Jul 2023.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23803–23828. PMLR, 23–29 Jul 2023.

Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 240–241, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7308–7320. Curran Associates, Inc., 2020.

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.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning (ICML)*, pp. 991–999. PMLR, 2014.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning (ICML)*, pp. 9722–9732. PMLR, 2021.

Terence Tao. *An Introduction to Measure Theory*, volume 126. American Mathematical Soc., 2011.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Haitao Wen, Haoyang Cheng, Heqian Qiu, Lanxiao Wang, Lili Pan, and Hongliang Li. Optimizing mode connectivity for class incremental learning. In *International Conference on Machine Learning (ICML)*, pp. 36940–36957. PMLR, 2023.

Zhi Zhang, Chris Chow, Yasi Zhang, Yanchao Sun, Haochen Zhang, Eric Hanchen Jiang, Han Liu, Furong Huang, Yuchen Cui, and Oscar Hernan Madrid Padilla. Statistical guarantees for lifelong reinforcement learning using pac-bayes theory. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 5050–5058. PMLR, 2025.

A Additional Discussion, Proof Details, and Extended Explanations

This appendix contains supporting technical material that is omitted from the main body for brevity. Appendix B justifies the standing assumptions and formal setup, Appendices C to G provide full proofs of the main results (Proposition 1, Theorems 3, 5, and Corollaries 4, 6), Appendix H contains detailed experiment results regarding proof-of-concept, Appendix I addresses related work in detail, and Appendix J includes further discussions on theories.

B Justifications on Assumptions and Theoretical Setups

Risks are measurable. We take the predictor $(\theta, x) \mapsto f(x;\theta)$ to be jointly measurable for the parameter and data measures used in the analysis. The bounded 0–1 loss then gives measurable risks and permits the interchange of integrals by Fubini's theorem.

Task-space geometry. For the task-space applications, we take $\Sigma = \mathcal{L}(\mathcal{Z})$. Each metric d used in Hypothesis 2 has its d-Borel sets in Σ, and each task measure is concentrated on its d-support. The latter property holds on separable metric spaces. These conditions make Proposition 1 applicable and ensure that the current task contributes no distal mass relative to its own support.

Bounded reference parameter region. (Assumption 1) Assumption 1 restricts the analysis to a bounded continuous reference region of the theoretical parameter space. This is natural in practice. Lazy-training behavior, where parameters remain near their initialization, provides justification for using such a finite-volume reference region. Floating-point arithmetic also imposes bounded realizable ranges through minimum and maximum representable values, and GD-based training typically explores only a limited part of this range. This should not be confused with the literal finite set of floating-point parameter vectors, which has Lebesgue measure zero.

Label-permutation symmetry (Assumption 2). Standard classification architectures, including ResNet, EfficientNet, and ViT, possess the output-class symmetry in Assumption 2. Their classification heads compute logits $z = W_L\phi(x) + b_L$, where $W_L \in \mathbb{R}^{k\times m}$ and $b_L \in \mathbb{R}^k$. Permuting the rows of W_L and the corresponding entries of b_L permutes the output logits while preserving the shared feature extractor. Regarding the indexing convention, the original class-c block moves to position $\sigma(c)$. Thus, the block occupying the new position i comes from the original position $\sigma^{-1}(i)$, explaining the inverse index in the definition. Standard classifier heads use the same numerical representation for every output class, so corresponding parameter coordinates share the same representable ranges. Our continuous reference region preserves this class symmetry through common coordinate ranges across output blocks, ensuring closure under last-layer permutations.

At the prediction level, the same symmetry holds almost everywhere. For any fixed input and body parameters, equality of two logits defines a Lebesgue-null hyperplane in the last-layer coordinates. Taking the finite union over class pairs and applying Fubini's theorem gives a unique maximizing logit for P-almost every example at ρ-almost every parameter, for each evaluated distribution P. Hence, for every output-label permutation σ, $f(x;\sigma\cdot\theta) = \sigma(f(x;\theta))$ holds with precisely this almost-everywhere meaning under the normalized Lebesgue reference measure ρ.

Learnability of the Optimal Parameter (Assumption 3). For each evaluated distal distribution P, the assumption requires every optimal parameter θ^* to predict the true label more often than any individual incorrect label within each class of positive mass: $\Pr_P(f(x;\theta^*) = y \mid Y = y) > \Pr_P(f(x;\theta^*) = c \mid Y = y)$ for $c \neq y$. These are frequencies of predicted labels across class-conditional inputs, rather than softmax probabilities at a single input.

To illustrate, consider a 3-class problem with labels $\{A, B, C\}$, and the frequencies of these predictions among inputs whose true label is A.

- **Scenario 1: Classwise condition holds.** At an optimum, the prediction frequencies for (A, B, C) are $(0.34, 0.33, 0.33)$. Although the error rate within class A is 66%, its true label is the unique mode. The same condition must hold for the other positive-mass classes and every optimum.

- **Scenario 2: Classwise condition fails.** At an optimum, the prediction frequencies are $(0.33, 0.34, 0.33)$. The incorrect label B is predicted more often than the true label A, violating the assumption.

Optimal-Solution Set and Attainment (Definition 1). If $m = \inf_{\theta\in\Theta} R(\theta; P)$ is not attained, then $\mathcal{S}^*(P) = \mathcal{S}(m; P) = \varnothing$, so its density and selection probability are zero. The risk lower bounds still apply through the risk sublevel sets.

C Proof on Proposition 1

Proposition 1. (ε-proximal/distal decomposition) *Let (X, d) be a metric space, let Σ be a σ-algebra containing the d-Borel sets, and let μ, ν be σ-finite measures on (X, Σ). For any $\varepsilon > 0$, there exists a unique decomposition of μ into two mutually singular measures $\mu = \mu_{prox} + \mu_{dist}$ satisfying:*

- *μ_{prox} **(The Proximal Part):** The measure μ_{prox} is concentrated on the open ε-neighborhood of $\mathrm{supp}(\nu)$:*
$$\mu_{\text{prox}}(\{x \in X : d(x, \mathrm{supp}(\nu)) \geq \varepsilon\}) = 0.$$
- *μ_{dist} **(The Distal Part):** The measure μ_{dist} is concentrated on the region at distance at least ε from $\mathrm{supp}(\nu)$:*
$$\mu_{\text{dist}}(\{x \in X : d(x, \mathrm{supp}(\nu)) < \varepsilon\}) = 0.$$

Here, $d(x, S) = \inf_{y\in S} d(x, y)$ is the distance between the point $x \in X$ and the set $S \subseteq X$.

Proof. Let $S_\nu := \mathrm{supp}(\nu)$ and define $f : X \to [0, +\infty]$ by
$$f(x) := d(x, S_\nu).$$
If $S_\nu \neq \varnothing$, the distance function f is continuous under d, hence Σ-measurable. If $S_\nu = \varnothing$, then $f \equiv +\infty$ is also measurable. Thus the sets
$$A := \{x \in X : f(x) < \varepsilon\}, \qquad B := \{x \in X : f(x) \geq \varepsilon\} = A^c$$
are measurable and partition X.

1. Existence

For each $E \in \Sigma$, define
$$\mu_{\text{prox}}(E) := \mu(E \cap A), \qquad \mu_{\text{dist}}(E) := \mu(E \cap B).$$
These restrictions are σ-finite measures, and the partition gives
$$\mu_{\text{prox}}(E) + \mu_{\text{dist}}(E) = \mu(E).$$
Moreover, $\mu_{\text{prox}}(B) = 0$ and $\mu_{\text{dist}}(A) = 0$, so the measures satisfy the stated concentration conditions and are mutually singular. These conditions do not require their topological supports to be disjoint.

2. Uniqueness

Suppose another pair $(\lambda_{\text{prox}}, \lambda_{\text{dist}})$ satisfies $\mu = \lambda_{\text{prox}} + \lambda_{\text{dist}}$, with $\lambda_{\text{prox}}(B) = 0$ and $\lambda_{\text{dist}}(A) = 0$. For every $E \in \Sigma$,
$$\begin{aligned}\mu_{\text{prox}}(E) &= \mu(E \cap A)\\ &= \lambda_{\text{prox}}(E \cap A) + \lambda_{\text{dist}}(E \cap A)\\ &= \lambda_{\text{prox}}(E \cap A) = \lambda_{\text{prox}}(E),\end{aligned}$$
and similarly,
$$\begin{aligned}\mu_{\text{dist}}(E) &= \mu(E \cap B)\\ &= \lambda_{\text{prox}}(E \cap B) + \lambda_{\text{dist}}(E \cap B)\\ &= \lambda_{\text{dist}}(E \cap B) = \lambda_{\text{dist}}(E).\end{aligned}$$
Therefore both component measures are unique. □

D PROOF ON THEOREM 3

Theorem 3 (Chance-Level Accuracy) *For $1 \le i < j \le N$, choose d, ε as in Hypothesis 2 for this comparison. If $P^{\varepsilon}_{i\perp j}(\mathcal{Z}) > 0$,*

$$\mathbf{E}_{\theta\sim\pi^*_j} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) \ge 1 - |\mathcal{Y}_{1:j}|^{-1}.$$

The same bound holds for $P_{1:j-1}$ with its own choice and positive distal mass.

Proof. We show $\mathbf{E}_{\theta\sim\pi^*_j} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) \ge \mathbf{E}_{\theta\sim\rho} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) = 1 - 1/k$. (As a recap on Definition 2, ρ is the uniform probability measure on Θ defined as $\rho(E) = \lambda(E)/\lambda(\Theta)$.)

First, we prove that for any evaluated probability measure P on $(\mathcal{Z}, \mathcal{L}(\mathcal{Z}))$ whose labels are among the k output classes, $\mathbf{E}_{\theta\sim\rho} R(\theta; P) = 1 - 1/k$.

By Assumption 2 and Fubini's theorem, for P_X-almost every input x, prediction equivariance holds for ρ-almost every parameter and every permutation, where P_X is the input marginal. Fix such an x. We show that $\rho(\{\theta : f(x;\theta) = y\}) = 1/k$ for each output label y.

Assumption 2 states that Θ is closed under the action of the symmetric group S_k (last-layer permutations). Such action has two properties:

- **Equivariance.** For any permutation $\sigma \in S_k$, the predicted label permutes accordingly for ρ-almost every θ: $f(x; \sigma \cdot \theta) = \sigma(f(x;\theta))$.
- **Volume Invariance:** The permutation $\sigma : \Theta \to \Theta$ is a volume-preserving isometry, meaning $\lambda(\sigma \cdot E) = \lambda(E)$ for any measurable set E.

For each output label $y \in \{1, \dots, k\}$, let $A_y := \{\theta \in \Theta : f(x;\theta) = y\}$. Consider any two distinct output labels a, b, and let $\sigma_{a\leftrightarrow b}$ swap them. Applying the equivariance property,

$$f(x; \sigma_{a\leftrightarrow b} \cdot \theta) = \sigma_{a\leftrightarrow b}(f(x;\theta)) = b \qquad (\rho\text{-almost every } \theta \in A_a).$$

The inverse swap gives $\sigma_{a\leftrightarrow b} \cdot A_a = A_b$ up to a Lebesgue-null set.

By the volume invariance property,

$$\lambda(A_b) = \lambda(\sigma_{a\leftrightarrow b} \cdot A_a) = \lambda(A_a).$$

Since the measures are equal for all classes and the sets $\{A_y\}_{y=1}^k$ partition the space Θ:

$$\lambda(\Theta) = \sum_{y=1}^{k} \lambda(A_y) = \sum_{y=1}^{k} \lambda(A_1) = k \cdot \lambda(A_1) = k \cdot \lambda(A_2) = \cdots = k \cdot \lambda(A_k).$$

Therefore, for P_X-almost every x and every output label y,

$$\rho(\{f(x;\theta) = y\}) = \frac{\lambda(A_y)}{\lambda(\Theta)} = \frac{1}{k}$$

Then, by the definition of risk and Fubini's theorem (Tao, 2011), we can swap the integrals:

$$\mathbf{E}_{\theta\sim\rho} R(\theta; P) = \mathbf{E}_{\theta\sim\rho} \int_{\mathcal{Z}} L(\theta, z)\, dP(z) = \int_{\mathcal{Z}} \mathbf{E}_{\theta\sim\rho} L(\theta, z)\, dP(z) = \int_{\mathcal{Z}} \rho(\{f(x;\theta) \ne y\})\, dP(z)$$

Since $\rho(\{f(x;\theta) \ne y\}) = 1 - 1/k$ for P-almost every $z = (x, y)$,

$$\int_{\mathcal{Z}} \rho(\{f(x;\theta) \ne y\})\, dP(z) = \int_{\mathcal{Z}} \left(1 - \frac{1}{k}\right) dP(z) = \left(1 - \frac{1}{k}\right) \int_{\mathcal{Z}} dP(z) = 1 - \frac{1}{k}.$$

Next, we show $\mathbf{E}_{\theta\sim\pi_j^*} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) \geq \mathbf{E}_{\theta\sim\rho} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j})$.

For this comparison's fixed d, ε and positive distal mass, Hypothesis 2 gives

$$\pi_j^*(\mathcal{S}(\tau; \tilde{P}^{\varepsilon}_{i\perp j})) \leq \rho(\mathcal{S}(\tau; \tilde{P}^{\varepsilon}_{i\perp j})) \qquad \text{for all } \tau \in [0, 1].$$

Recall that the solution set $\mathcal{S}(\tau; P)$ is the set of parameters with risk less than or equal to τ. In probabilistic terms, this becomes the cumulative distribution function (CDF) of the risk random variable.

- Let X be the risk of a parameter drawn from ERM: $X := R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}),\ \theta \sim \pi_j^*$.
- Let Y be the risk of a parameter drawn from Uniform: $Y := R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}),\ \theta \sim \rho$.

For a random variable Z taking values in $[0, 1]$, Tonelli's theorem applied to $Z = \int_0^1 \mathbf{1}\{Z > \tau\}\, d\tau$ gives $\mathbf{E}[Z] = \int_0^1 \Pr(Z > \tau)\, d\tau = \int_0^1 (1 - F_Z(\tau))\, d\tau$. The risks X and Y take values in this interval under 0–1 loss.

Using the inequality $F_X(\tau) \leq F_Y(\tau)$ from Hypothesis 2:

$$1 - F_X(\tau) \geq 1 - F_Y(\tau).$$

Integrating both sides over the domain $[0, 1]$:

$$\mathbf{E}[X] = \int_0^1 (1 - F_X(\tau))\, d\tau \geq \int_0^1 (1 - F_Y(\tau))\, d\tau = \mathbf{E}[Y]$$

, i.e. $\mathbf{E}_{\theta\sim\pi_j^*} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j}) \geq \mathbf{E}_{\theta\sim\rho} R(\theta; \tilde{P}^{\varepsilon}_{i\perp j})$.

Since k is the number of output classes, $k \geq |\mathcal{Y}_{1:j}|$ and $1 - 1/k \geq 1 - 1/|\mathcal{Y}_{1:j}|$ trivially. The same argument applies to $\tilde{P}^{\varepsilon}_{(1:j-1)\perp j}$ using the choice of d, ε for $P_{1:j-1}$ relative to P_j in Hypothesis 2, under the same learner distribution π_j^*.

□

Remark. The bound $1 - 1/|\mathcal{Y}_{1:j}|$ has been utilized instead of $1 - 1/k$ to allow the theorem to apply to both "fixed backbone" settings (where k is large and constant) and "growing head" settings (where k grows with tasks).

E Proof on Corollary 4

Proof. Fix the choice of d, ε for $P_{1:N-1}$ relative to P_N in Hypothesis 2 throughout the decomposition below. If the distal mass is zero, the bound follows from non-negativity. Otherwise, decomposing $P_{1:N}$ into the past mixture $P_{1:N-1}$ and current task P_N, and subsequently isolating the ε-distal component of the past mixture, we have:

$$\begin{aligned}
\mathbf{E}_{\theta\sim\pi_N^*} R(\theta; P_{1:N}) &= \frac{N-1}{N} \mathbf{E}_{\theta\sim\pi_N^*} R(\theta; P_{1:N-1}) + \frac{1}{N} \mathbf{E}_{\theta\sim\pi_N^*} R(\theta; P_N) \\
&\geq \frac{N-1}{N} \mathbf{E}_{\theta\sim\pi_N^*} R(\theta; P^{\varepsilon}_{(1:N-1)\perp N}) \qquad \text{(by non-negativity)} \\
&= \frac{N-1}{N} \cdot m^{\varepsilon}_{(1:N-1)\perp N} \cdot \mathbf{E}_{\theta\sim\pi_N^*} R(\theta; \tilde{P}^{\varepsilon}_{(1:N-1)\perp N})
\end{aligned}$$

Apply Hypothesis 2 to the accumulated past mixture $P_{1:N-1}$ with respect to P_N. Theorem 3 then gives $\mathbf{E}_{\theta\sim\pi_N^*} R(\theta; \tilde{P}^{\varepsilon}_{(1:N-1)\perp N}) \geq 1 - |\mathcal{Y}|^{-1}$, completing the proof. □

F Proof on Theorem 5

Theorem 5 (Prior-Task Optimality Probability) *For $1 \le i < j \le N$, choose d, ε as in Hypothesis 2 for this comparison. With positive distal mass, let r count the classes with positive mass in $\tilde{P}^{\varepsilon}_{i\perp j}$. Then*

$$\pi_j^*(\mathcal{S}^*(\tilde{P}^{\varepsilon}_{i\perp j})) \le \frac{1}{r!}.$$

The same result holds for $P_{1:j-1}$ with its own choice, with r counting its positive-mass distal classes.

Proof. For simplicity, write $\mathcal{S}^* := \mathcal{S}^*(P)$ for an evaluated distal distribution P satisfying Assumption 3. Let $A := \{y : P(Y = y) > 0\}$ and $r := |A|$.

We first show $\rho(\mathcal{S}^*) \le (k-r)!/k!$ by counting disjoint permutation copies. If $\rho(\mathcal{S}^*) = 0$, this density bound is immediate; otherwise, the following volume argument applies.

Assumption 2 gives prediction equivariance outside a ρ-null set of parameters. Since S_k is finite and its action preserves volume, we may discard this set and its permuted copies in the following volume calculation. All set identities and disjointness claims in this calculation are understood up to null sets.

The true risk can be decomposed using the confusion matrix. For $y \in A$ and output class c, let $C_{y,c}(\theta) := \Pr_P(f(x;\theta) = c \mid Y = y)$. Then,

$$R(\theta; P) = \sum_{y \in A} P(Y = y)\,(1 - C_{y,y}(\theta))\,.$$

Now, let $\theta^* \in \mathcal{S}^*$, $\sigma \in S_k$, and $\theta' = \sigma \cdot \theta^*$. Due to equivariance, if the original model predicts c, the permuted model predicts $\sigma(c)$. In other words,

$$C_{y,\sigma(c)}(\theta') = C_{y,c}(\theta^*) \quad \text{and} \quad C_{y,y}(\theta') = C_{y,\sigma^{-1}(y)}(\theta^*).$$

Consequently,

$$R(\theta'; P) - R(\theta^*; P) = \sum_{y \in A} P(Y = y)\left[C_{y,y}(\theta^*) - C_{y,\sigma^{-1}(y)}(\theta^*)\right].$$

By Assumption 3, each term is nonnegative and is strictly positive when σ moves the corresponding label. Thus, moving any label in A strictly increases the risk, while fixing every label in A preserves it. Consequently, permutations fixing A preserve $\mathcal{S}^*$; all others map it outside $\mathcal{S}^*$.

If two permutations σ, τ agree on A, then $\tau^{-1}\sigma$ fixes A, so $\sigma \cdot \mathcal{S}^* = \tau \cdot \mathcal{S}^*$. If they differ on A, their copies are disjoint: an intersection would give $\theta_1, \theta_2 \in \mathcal{S}^*$ with $\theta_2 = (\tau^{-1}\sigma) \cdot \theta_1$, contradicting the strict risk increase above.

Among the $k!$ permutations, each assignment of the r labels in A occurs $(k-r)!$ times, corresponding to the permutations of the remaining labels. Hence there are $k!/(k-r)!$ disjoint copies of $\mathcal{S}^*$ in Θ. By volume preservation,

$$\frac{k!}{(k-r)!}\lambda(\mathcal{S}^*) \le \lambda(\Theta), \qquad \rho(\mathcal{S}^*) \le \frac{(k-r)!}{k!} \le \frac{1}{r!}.$$

The last inequality follows from $k!/(k-r)! = r!\binom{k}{r} \ge r!$.

Finally, take $P = \tilde{P}^{\varepsilon}_{i\perp j}$ or $P = \tilde{P}^{\varepsilon}_{(1:j-1)\perp j}$ with positive distal mass, using the respective comparison's choice of d, ε. Applying Hypothesis 2 at $\tau = \inf_{\theta\in\Theta} R(\theta; P)$ gives

$$\pi_j^*(\mathcal{S}^*(P)) \le \rho(\mathcal{S}^*(P)) \le \frac{(k-r)!}{k!} \le \frac{1}{r!}.$$

Here the hypothesis is applied to the full optimal set $\mathcal{S}^*(P)$, so no absolute continuity of π_j^* with respect to ρ is needed. If the infimum is not attained, this set is empty and the selection bound is immediate. □

G PROOF ON COROLLARY 6

Corollary 6 (Global Optimality Probability) *Choose d, ε as in Hypothesis 2 for $P_{1:N-1}$ relative to P_N, and use this choice to decompose the uniform mixture $P_{1:N}$ with respect to P_N as $P^{\varepsilon}_{dist} := P^{\varepsilon}_{(1:N)\perp N}$ and $P^{\varepsilon}_{prox} := P^{\varepsilon}_{(1:N)\ll N}$. Suppose the distal mass is positive, and let r be the number of classes with positive mass in $\tilde{P}^{\varepsilon}_{dist}$. If $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{dist}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{prox}) \neq \emptyset$,*

$$\pi^*_N(\mathcal{S}^*(P_{1:N})) \leq \frac{1}{r!}.$$

Proof. We first show $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}}) \neq \emptyset \implies \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}}) = \mathcal{S}^*(P_{1:N})$.

Throughout, $P^{\varepsilon}_{\text{dist}}$ and $P^{\varepsilon}_{\text{prox}}$ use the choice of d, ε for $P_{1:N-1}$ relative to P_N specified above. Since P_N is concentrated on its d-support and $\varepsilon > 0$, it contributes only to the proximal component. Restricting the uniform mixture to the distal region therefore gives

$$P^{\varepsilon}_{\text{dist}} = \frac{N-1}{N} P^{\varepsilon}_{(1:N-1)\perp N}, \qquad P^{\varepsilon}_{\text{prox}} = \frac{N-1}{N} P^{\varepsilon}_{(1:N-1)\ll N} + \frac{1}{N} P_N.$$

Let $m := P^{\varepsilon}_{\text{dist}}(\mathcal{Z}) > 0$. Then $m \leq (N-1)/N < 1$, so both components can be normalized, and $P_{1:N}$ can be written as:

$$P_{1:N} = P^{\varepsilon}_{\text{dist}} + P^{\varepsilon}_{\text{prox}} = m \cdot \tilde{P}^{\varepsilon}_{\text{dist}} + (1-m) \cdot \tilde{P}^{\varepsilon}_{\text{prox}}$$

Moreover, $\tilde{P}^{\varepsilon}_{\text{dist}} = \tilde{P}^{\varepsilon}_{(1:N-1)\perp N}$, so the normalized distal component is one of the distributions covered by Hypothesis 2 under the learner law π^*_N.

Let $R(\theta, \tilde{P}^{\varepsilon}_{\text{dist}})$ and $R(\theta, \tilde{P}^{\varepsilon}_{\text{prox}})$ be the risk functions for $\tilde{P}^{\varepsilon}_{\text{dist}}$ and $\tilde{P}^{\varepsilon}_{\text{prox}}$. Since $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}})$ and $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$ are the sets of minimizers, for any $\theta^* \in \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$, $R(\theta^*, \tilde{P}^{\varepsilon}_{\text{dist}}) \leq R(\theta, \tilde{P}^{\varepsilon}_{\text{dist}})$ and $R(\theta^*, \tilde{P}^{\varepsilon}_{\text{prox}}) \leq R(\theta, \tilde{P}^{\varepsilon}_{\text{prox}})$ for all $\theta \in \Theta$.

The risk for the overall distribution $P_{1:N} = m \cdot \tilde{P}^{\varepsilon}_{\text{dist}} + (1-m) \cdot \tilde{P}^{\varepsilon}_{\text{prox}}$ is linear:

$$R(\theta, P_{1:N}) = m \cdot R(\theta, \tilde{P}^{\varepsilon}_{\text{dist}}) + (1-m) \cdot R(\theta, \tilde{P}^{\varepsilon}_{\text{prox}}).$$

1. $\mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}}) \subseteq \mathcal{S}^*(P_{1:N})$.

For any $\theta^* \in \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$ and any arbitrary $\theta \in \Theta$,

$$\begin{aligned} R(\theta^*, P_{1:N}) &= m \cdot R(\theta^*, \tilde{P}^{\varepsilon}_{\text{dist}}) + (1-m) \cdot R(\theta^*, \tilde{P}^{\varepsilon}_{\text{prox}}) \\ &\leq m \cdot R(\theta, \tilde{P}^{\varepsilon}_{\text{dist}}) + (1-m) \cdot R(\theta, \tilde{P}^{\varepsilon}_{\text{prox}}) \\ &= R(\theta, P_{1:N}) \end{aligned}$$

2. $\mathcal{S}^*(P_{1:N}) \subseteq \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$.

Let $\hat{\theta}$ be any minimizer of the risk $R(\cdot, P_{1:N})$. Then, $R(\hat{\theta}, P_{1:N}) \leq R(\theta^*, P_{1:N})$. This inequality can be written as:

$$m \cdot [R(\hat{\theta}, \tilde{P}^{\varepsilon}_{\text{dist}}) - R(\theta^*, \tilde{P}^{\varepsilon}_{\text{dist}})] + (1-m) \cdot [R(\hat{\theta}, \tilde{P}^{\varepsilon}_{\text{prox}}) - R(\theta^*, \tilde{P}^{\varepsilon}_{\text{prox}})] \leq 0.$$

Because θ^* minimizes both $\tilde{P}^{\varepsilon}_{\text{dist}}$ and $\tilde{P}^{\varepsilon}_{\text{prox}}$ individually, the terms in the brackets are both nonnegative (≥ 0). Since $m \in (0,1)$, $R(\hat{\theta}, \tilde{P}^{\varepsilon}_{\text{dist}}) - R(\theta^*, \tilde{P}^{\varepsilon}_{\text{dist}}) = 0$ and $R(\hat{\theta}, \tilde{P}^{\varepsilon}_{\text{prox}}) - R(\theta^*, \tilde{P}^{\varepsilon}_{\text{prox}}) = 0$, i.e. $\hat{\theta} \in \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{dist}})$ and $\hat{\theta} \in \mathcal{S}^*(\tilde{P}^{\varepsilon}_{\text{prox}})$. Then,

$$
\begin{aligned}
\pi_N^*(\mathcal{S}^*(P_{1:N})) &= \pi_N^*(\mathcal{S}^*(\tilde{P}^\varepsilon_{\text{dist}}) \cap \mathcal{S}^*(\tilde{P}^\varepsilon_{\text{prox}})) && \text{(equality of the optimal sets)} \\
&\leq \pi_N^*(\mathcal{S}^*(\tilde{P}^\varepsilon_{\text{dist}})) && (\pi_N^*(A \cap B) \leq \pi_N^*(A)) \\
&\leq \frac{1}{r!} && \text{(Theorem 5, applied to the past mixture)}.
\end{aligned}
$$

□

H PROOF-OF-CONCEPT DETAILS

This appendix provides the experimental details for the proof-of-concept (PoC) diagnostics in Tables 1, 2, and 3, and Figures 4, 5, and 6. We examine whether concrete distal constructions identify components whose empirical risk CDFs are consistent with Hypothesis 2, and whether their retained mass yields quantitatively substantial plug-in risk floors through Corollary 4.

Dataset and task protocols. All PoC experiments use CIFAR-100. We consider six continual-learning protocols. First, **CIL** denotes the standard class-disjoint class-incremental protocol with three tasks: a 50-class base task followed by two 25-class increments. This setting is included because the class-disjoint label structure gives an explicit and transparent distal construction: if two tasks contain disjoint labels, then the label-based metric below separates their supports with full distal mass. Second, **Blurry1** and **Blurry2** are blurry CIL protocols with the same nominal 50/25/25 class schedule, but with a small amount of cross-task overlap. Here Blurry$M$ means that $M\%$ of the full CIFAR-100 training set appears outside its nominal task assignment. These protocols are included to test whether the PoC remains meaningful when strict class-disjointness is weakened. Third, we consider three generalized class-incremental learning (gCIL) protocols with four phases: **gCIL-TV** (task-varied), **gCIL-U** (uniform), and **gCIL-L** (long-tail). Each gCIL phase contains 1000 training samples drawn from 20 appearing classes. The class supports and sampling weights are fixed once and reused across all ERM seeds so that differences across ERM runs reflect optimization randomness rather than changes in the task distribution. The three gCIL variants are included because they test the hypothesis beyond clean class-disjoint increments: gCIL-U gives a controlled balanced phase structure, gCIL-TV introduces phase-dependent class composition, and gCIL-L introduces class-frequency imbalance.

Training protocol. For all CIFAR-100 PoC runs, we train ResNet-32 models by standard ERM on the data available at each task or phase. No replay, regularization-based continual-learning method, distillation, exemplar memory, or task-specific correction is used. This choice is intentional: the theory concerns the behavior of naive ERM under task shift, so adding a continual-learning method would confound the diagnostic. Unless otherwise stated, models are trained for 100 epochs using SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, batch size 128, and a multistep schedule with milestones at epochs 60 and 80 and decay factor 0.1. We use the same training hyperparameters across protocols. They are standard CIFAR/ResNet hyperparameters and were fixed before the PoC analysis; they were not tuned to improve the ordered-CDF diagnostic. Evaluation uses the seen-class output space at each stage. The LHS uses at least 30 ERM runs for each CIL and Blurry setting and at least 120 for each gCIL setting. The RHS uses the same $5{,}000$ uniform parameter draws across all stages and comparisons within each protocol. Each LHS sample represents a complete sequential training run; each RHS sample requires evaluation of a sampled parameter vector. The larger reference sample reduces Monte Carlo variability in the density estimate. Per-image 0–1 losses are cached before applying distal masks and reused across thresholds, preserving the task-mixture multiplicities of overlapping samples.

Distal constructions. Let $\phi(x)$ be the cached DINOv2 representation and let $D_{\mathrm{box}} > 0$ be the diagonal length of the coordinate box fitted once to the protocol's full training/evaluation feature union. With raw pixels $x \in [0,1]^{3072}$ and $\eta = 10^{-6}$, we use

$$d((x,y),(x',y')) = \begin{cases} \min\{\|\phi(x) - \phi(x')\|_2 / D_{\mathrm{box}} + \eta \|x - x'\|_2 / \sqrt{3072},\ 1 - \eta\}, & y = y', \\ 1, & y \neq y'. \end{cases}$$

The positive pixel term separates distinct images even when their feature vectors coincide; the cap and label term preserve the metric properties. The feature scale is shared across comparisons within each protocol, and features are not normalized to unit norm. A test sample is distal when its minimum distance to the current task's training support is at least ε. For class-disjoint CIL, $\varepsilon = 1$ reduces this split to the label indicator and retains all prior-task mass. Blurry and gCIL use the same metric with comparison-specific thresholds in Tables 1 and 2; the additional experiment below fixes thresholds from current-task geometry.

Ordered-CDF diagnostic. For each ordered task pair $i \to j$, let $\tilde{P}^{\varepsilon}_{i \perp j}$ denote its normalized empirical distal distribution; for $1{:}j-1 \to j$, use $\tilde{P}^{\varepsilon}_{(1:j-1)\perp j}$ for the accumulated past mixture

relative to P_j. We write $\widehat{R}(\theta; P)$ for the empirical 0–1 risk on evaluation samples representing P. For a mixture, each model's retained losses are combined with uniform task weights and divided by the retained mixture mass before forming the empirical CDF. In Table 2, $\widehat{m}$ abbreviates the indicated component's empirical distal mass, $\widehat{R}$ is the mean of its ERM risk samples, and $k_j = |\mathcal{Y}_{1:j}|$ is the seen-class count. Hypothesis 2 can be read as an ordered-CDF condition over risk sublevel sets: ERM trained on task j should not place more mass on low-risk solutions for $\tilde{P}^{\varepsilon}_{i\perp j}$ than the reference density baseline. For each threshold ε, the LHS contains one risk value per ERM run and the RHS contains 5,000 reference risk values. All displayed curves are right-continuous empirical CDFs of these model-wise 0–1 risks; repeated risks, including risk 1, produce jumps equal to their empirical mass. We report

$$\widehat{\Delta} := \sup_{\tau \in [0,1]} \left(\widehat{F}_{\mathrm{LHS}}(\tau) - \widehat{F}_{\mathrm{RHS}}(\tau) \right),$$

where positive values indicate empirical violations of the ordered-CDF relation. We also report the one-sided Kolmogorov–Smirnov p-value p^{+}_{KS} for the alternative that the LHS CDF exceeds the RHS CDF. A threshold is non-rejected by this diagnostic when $p^{+}_{\mathrm{KS}} \geq 0.05$.

Empirical approximation of the density baseline ρ. We use one axis-aligned reference box Θ per protocol and estimate its normalized volume $\rho(E) = \lambda(E)/\lambda(\Theta)$ using $M = 5{,}000$ uniform parameter draws. The box is constructed from all ERM initializations and task checkpoints across seeds, taking minima and maxima separately for each trainable parameter coordinate. Running statistics and counters are excluded from these coordinates. For classifier weights, bounds are additionally pooled across output classes separately for each input coordinate; all classifier biases share one pooled interval. This makes the box closed under output-label permutations while retaining the original intervals for all other parameters. Using the observed parameter ranges anchors the reference to the scales encountered during learning. The box is fixed across all stages, thresholds, and comparisons. Parameter coordinates are sampled independently and uniformly, and the resulting vectors are reused throughout the protocol. Thus, for a risk threshold τ, the reference density is estimated as

$$\widehat{\rho}\Big(\mathcal{S}(\tau; \tilde{P}^{\varepsilon}_{i\perp j})\Big) = \frac{1}{M} \sum_{\ell=1}^{M} \mathbf{1}\Big\{ \widehat{R}(\theta_\ell; \tilde{P}^{\varepsilon}_{i\perp j}) \leq \tau \Big\}, \qquad \theta_\ell \sim \mathrm{Unif}(\Theta).$$

The RHS empirical CDF estimates this reference density, while the LHS empirical CDF uses the trained ERM models. For an accumulated past mixture, the same expression uses $\tilde{P}^{\varepsilon}_{(1:j-1)\perp j}$.

Threshold selection. CIL fixes $\varepsilon^* = 1$; Blurry and gCIL search $\{0.005, 0.010, \ldots, 0.400, 0.45, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00\}$. For each individual-task or past-mixture comparison in Table 2, ε^* is the smallest tested threshold with positive $\widehat{m}$ and $p^{+}_{\mathrm{KS}} \geq 0.05$. This retains the largest mass among eligible thresholds for that comparison; the threshold is fixed across risk levels and parameter draws. The selected final-mixture thresholds are 0.075, 0.120, 0.140, 0.145, and 0.150 for Blurry1, Blurry2, gCIL-TV, gCIL-U, and gCIL-L, respectively. Table 2 lists 36 individual and mixture comparisons. These are finite-sample diagnostics after threshold selection, rather than a certification of population FOSD.

Final mixture quantities. The quantities in Table 1 use the empirical analogue of Corollary 4 with uniform task weights. For a protocol with N tasks or phases, set $\varepsilon = \varepsilon^*$ for the threshold selected for $P_{1:N-1}$ relative to P_N. The *Past mass* entry is

$$\widehat{m}^{\varepsilon}_{(1:N-1)\perp N} = \frac{1}{N-1} \sum_{i=1}^{N-1} \widehat{m}^{\varepsilon}_{i\perp N}.$$

The *Plug-in floor* entry is

$$\text{Plug-in floor} = \frac{N-1}{N} \cdot \widehat{m}^{\varepsilon}_{(1:N-1)\perp N} \cdot \left(1 - \frac{1}{k_N}\right).$$

Here $k_N = |\mathcal{Y}_{1:N}| = |\mathcal{Y}|$ is the final seen-class count, continuing the k_j notation in Table 2: 100 for CIL and Blurry, and 60, 56, and 57 for gCIL-TV, gCIL-U, and gCIL-L, respectively. Every mass

in this expression uses the same final-mixture threshold; independently selected task-pair floors are not summed. For a model θ after the final task, $\widehat{R}(\theta; P_{1:N})$ is its empirical risk averaged uniformly across tasks. The *Final risk* entry reports the mean of this quantity over all ERM runs. Using this reported mean, error coverage is expressed as a percentage:

$$\text{Error coverage } (\%) = 100 \times \frac{\text{Plug-in floor}}{\text{Final risk}}.$$

It measures the plug-in floor as a fraction of observed final error, not classification accuracy.

Table 2: Individual-task and past-mixture diagnostics. $1{:}i \to j$ denotes the uniform mixture of tasks $1, \ldots, i$ relative to task j; k_j is the seen-class count. The chance-level lower bound $1 - 1/k_j$ from Theorem 3 is compared with mean ERM distal risk $\widehat{R}$. Each comparison reports its selected threshold ε^*, retained mass $\widehat{m}$, maximal CDF gap $\widehat{\Delta}$, one-sided KS p-value, and mean ERM risk $\widehat{R}$. All statistics use the full set of ERM runs, with label separation for CIL and DINOv2 otherwise. Thresholds maximize retained mass among grid candidates with $p \geq 0.05$; CIL fixes $\varepsilon^* = 1$. Across all selected comparisons, mean ERM distal risk exceeds $1 - 1/k_j$. Distal components account for non-negligible fractions of individual past-task distributions and their final mixtures, yielding quantitatively meaningful plug-in floors even when tasks overlap.

Protocol	Comparison	$1 - 1/k_j$	ε^*	$\widehat{m}$	$\widehat{\Delta}$	p^+_{KS}	$\widehat{R}$
CIL	$1 \to 2$	0.9867	1.0000	1.0000	0.0000	1.0000	1.0000
	$1 \to 3$	0.9900	1.0000	1.0000	0.0000	1.0000	1.0000
	$2 \to 3$	0.9900	1.0000	1.0000	0.0000	1.0000	1.0000
	$1{:}2 \to 3$	0.9900	1.0000	1.0000	0.0000	1.0000	1.0000
Blurry1	$1 \to 2$	0.9900	0.1250	0.6206	0.1538	0.0544	0.9948
	$1 \to 3$	0.9900	0.1250	0.6244	0.1393	0.0910	0.9954
	$2 \to 3$	0.9900	0.1200	0.8568	0.1399	0.0892	0.9985
	$1{:}2 \to 3$	0.9900	0.0750	0.9937	0.1485	0.0660	0.9924
Blurry2	$1 \to 2$	0.9900	0.1400	0.1910	0.1564	0.0878	0.9940
	$1 \to 3$	0.9900	0.1450	0.1192	0.1334	0.1684	0.9933
	$2 \to 3$	0.9900	0.0900	0.9868	0.1678	0.0612	0.9979
	$1{:}2 \to 3$	0.9900	0.1200	0.7311	0.1673	0.0622	0.9931
gCIL-TV	$1 \to 2$	0.9722	0.1500	0.8075	0.0977	0.0720	0.9989
	$1 \to 3$	0.9800	0.1600	0.8520	0.0722	0.2350	0.9995
	$2 \to 3$	0.9800	0.1700	0.8000	0.0000	1.0000	1.0000
	$1{:}2 \to 3$	0.9800	0.1600	0.8290	0.0779	0.1858	0.9984
	$1 \to 4$	0.9833	0.1550	0.8050	0.0566	0.4070	0.9991
	$2 \to 4$	0.9833	0.1550	0.8020	0.0000	1.0000	0.9996
	$3 \to 4$	0.9833	0.1400	0.9080	0.0619	0.3428	0.9992
	$1{:}3 \to 4$	0.9833	0.1400	0.8522	0.1036	0.0522	0.9951
gCIL-U	$1 \to 2$	0.9730	0.1450	0.8640	0.0990	0.0538	0.9979
	$1 \to 3$	0.9787	0.1650	0.7020	0.0957	0.0649	0.9993
	$2 \to 3$	0.9787	0.1650	0.7525	0.0000	1.0000	0.9996
	$1{:}2 \to 3$	0.9787	0.1600	0.7302	0.0511	0.4516	0.9983
	$1 \to 4$	0.9821	0.1500	0.8515	0.0000	1.0000	0.9998
	$2 \to 4$	0.9821	0.1500	0.8580	0.0000	1.0000	0.9995
	$3 \to 4$	0.9821	0.1550	0.7075	0.0000	1.0000	0.9992
	$1{:}3 \to 4$	0.9821	0.1450	0.8147	0.0405	0.6035	0.9975
gCIL-L	$1 \to 2$	0.9750	0.0050	1.0000	0.0000	1.0000	1.0000
	$1 \to 3$	0.9796	0.1600	0.7025	0.0000	1.0000	0.9997
	$2 \to 3$	0.9796	0.1500	0.7595	0.1083	0.0595	0.9984
	$1{:}2 \to 3$	0.9796	0.1450	0.7418	0.0988	0.0950	0.9953
	$1 \to 4$	0.9825	0.1600	0.8505	0.0000	1.0000	0.9998
	$2 \to 4$	0.9825	0.1750	0.7025	0.0000	1.0000	0.9995
	$3 \to 4$	0.9825	0.1650	0.8040	0.0000	1.0000	0.9997
	$1{:}3 \to 4$	0.9825	0.1500	0.8053	0.1095	0.0559	0.9941

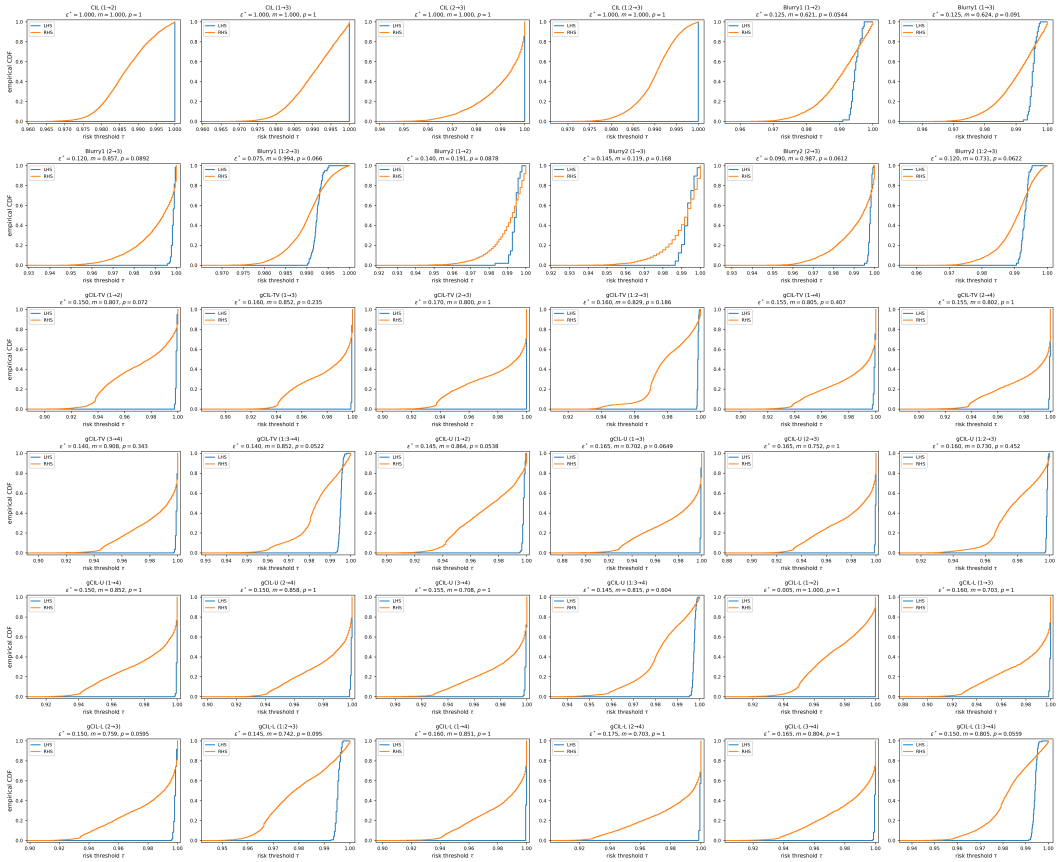


Figure 5: Full empirical CDF diagnostics for CIFAR-100: 27 task pairs and 9 uniform past-task mixtures. Each panel shows the empirical risk CDFs of sequential ERM (LHS) and 5,000 uniform reference draws (RHS) on the same distal component at its selected threshold ε^*. CIL uses label separation; Blurry1, Blurry2, gCIL-TV, gCIL-U, and gCIL-L use DINOv2. The notation $1{:}i \to j$ denotes the uniform mixture of tasks $1, \ldots, i$ relative to task j. Panel titles report ε^*, empirical distal mass, and p^+_{KS}.

Rejected examples. Figure 6 displays 30 rejected threshold comparisons from the same ε grid, including 7 past-mixture panels. We sample 30 eligible comparisons without replacement and select one rejected threshold at random within each comparison. These rows exhibit excess LHS CDF mass at some risk thresholds and are excluded from the selected-threshold summaries in Tables 1 and 2.

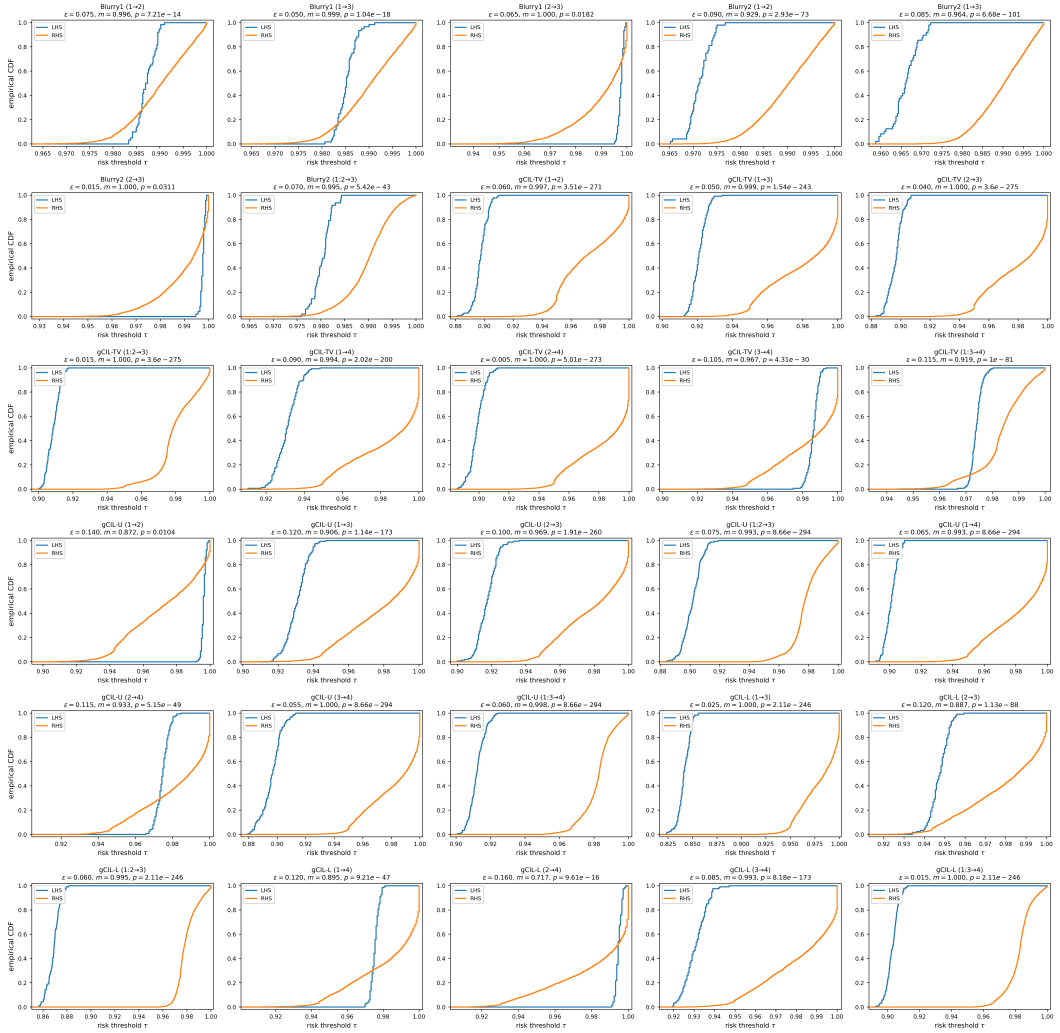


Figure 6: Rejected eCDF diagnostics from the threshold grid. The 30 panels (23 individual-task and 7 past-mixture comparisons) have $p^{+}_{\mathrm{KS}} < 0.05$ and are sampled without replacement (seed 2026). Each panel plots sequential ERM risks (LHS) and 5,000 reference risks (RHS) at that rejected threshold. Titles report ε, empirical distal mass, and p^{+}_{KS}.

H.1 Additional Experiment

We additionally choose ε without using risk CDFs or KS statistics. CIFAR-100 test image IDs are split into 20% calibration and 80% evaluation, consistently across task occurrences. For Blurry and gCIL, ε is determined by the 95th percentile of distances from final-task calibration samples to training support, with ties placed on the proximal side. We retain the metric, reference box, model populations, and equal task weights used above; CIL keeps $\varepsilon = 1$. Table 3 reports the final past mixture on the evaluation subset.

Table 3: Additional diagnostics for the final past mixture, with thresholds selected from current-task geometry. Quantities follow Table 1; CIL retains its label-separation control. All six final-mixture comparisons have one-sided KS $p \geq 0.05$.

	CIL	Blurry1	Blurry2	gCIL-TV	gCIL-U	gCIL-L
Threshold ε	1.0000	0.1299	0.1299	0.1492	0.1512	0.1549
Past mass $\widehat{m}^{\varepsilon}_{(1:N-1)\perp N}$	1.0000	0.6436	0.5810	0.8442	0.8017	0.7973
Final risk	0.7246	0.7201	0.7120	0.8576	0.8158	0.8099
Plug-in floor (Cor. 4)	0.6600	0.4248	0.3834	0.6226	0.5905	0.5875
Error coverage	91.0817%	58.9907%	53.8565%	72.5990%	72.3897%	72.5403%
KS p-value	1.0000	0.9449	0.6216	0.2423	1.0000	0.2870

Across Blurry and gCIL, this choice retains 58.1%–84.4% of the past mixture, with plug-in floors covering 53.9%–72.6% of measured final error.

I Related Work

Learning guarantees, learnability, and forgetting bounds. Domain-adaptation bounds relate target risk to source risk and distribution discrepancy (Ben-David et al., 2006), while PAC-Bayesian analyses provide guarantees for lifelong and online learning (Pentina & Lampert, 2014; Haddouche & Guedj, 2022) and lifelong reinforcement learning (Zhang et al., 2025). Multiclass learnability results further show that ERM rules over the same hypothesis class can have different sample complexities (Daniely et al., 2011). For class-incremental learning, Kim et al. (2022) connect performance to within-task prediction and task identification through out-of-distribution detection, and Kim et al. (2023) establish learnability guarantees. Computational and memory analyses establish hardness results for optimal continual learning (Knoblauch et al., 2020) and PAC-style memory lower bounds (Chen et al., 2022). These results address learnability and resource requirements in their respective settings.

Quantitative analyses of forgetting provide a complementary view. Linear-regression studies characterize the effects of task similarity, task order, and overparameterization (Lin et al., 2023; Evron et al., 2022; Goldfarb et al., 2024), and Friedman & Meir (2025) derive data-dependent and oracle upper bounds on forgetting. Our analysis concerns expected prior-task and final-mixture risk under standard sequential ERM. The selection hypothesis connects the learner's probability of selecting low-risk distal solutions to their reference density; label symmetry then yields a risk floor weighted by distal mass. This selection problem remains meaningful even when the tasks admit a common solution.

Loss geometry and preservation of prior knowledge. Loss-landscape analyses relate forgetting to displacement from prior-task minima and local curvature (Mirzadeh et al., 2020). Mode-connectivity studies examine low-loss paths between solutions, including solutions obtained through continual and multitask training (Garipov et al., 2018; Mirzadeh et al., 2021; Wen et al., 2023), while other work examines flatness in continual learning (Bian et al., 2024). These studies describe geometric relationships between solutions. Our framework examines their reference density and selection probability: the existence or connectivity of good solutions does not determine how much parameter volume they occupy or how much probability subsequent training assigns to them.

Methods that preserve prior knowledge intervene in this selection process. EWC penalizes changes to parameters important for earlier tasks (Kirkpatrick et al., 2017), while GEM constrains gradients using stored examples from previous tasks (Lopez-Paz & Ranzato, 2017). Such mechanisms modify the objective or constraints guiding subsequent learning. The distinction between these procedures and standard sequential ERM is reflected in the learner's selection law. This connects the solution-density interpretation to the question of how learning procedures preserve an advantage in selecting solutions that perform well on past data.

Permutation symmetry and solution density. Brea et al. (2019) and Simsek et al. (2021) investigate the geometry and optimization of neural-network loss landscapes through equivalent minima and symmetry-induced critical points. Our focus is forgetting under sequential ERM, relating the reference volume of prior-task optima to their selection probability through the selection hypothesis. Their hidden-neuron permutations preserve network function; our output-class permutations against fixed labels generate disjoint, equal-volume non-optimal copies of the full optimal set, yielding an inverse-factorial density bound.

J Further Discussions on Hypothesis, Theorem, and Corollaries

Discussion on Hypothesis 2 Hypothesis 2 and the resulting bounds concern the specified sequential ERM procedure and task sequence. The learner law π_j^* includes the actual training history, and its expectations account for initialization, data sampling, and optimization randomness conditional on that task sequence. The proofs use each comparison's choice of d, ε, fixed across risk thresholds and parameter draws, with the same stage-j learner law π_j^*. Corollaries 4 and 6 use the choice for $P_{1:N-1}$ relative to P_N throughout their decompositions.

Discussions on Theorem 5 and Corollary 6. For r positive-mass distal classes, the permutation argument bounds the reference density of distal optima by $(k-r)!/k! \leq 1/r!$. When the proximal and distal components share an optimum, the global optimal set is contained in the distal optimal

set, so the same density bound applies to global optima. Hypothesis 2 connects this density to the selection probability of the specified sequential ERM procedure, even when a common solution exists.