Spectral-Aware Analytic Class-Incremental Learning for Long-Tailed Distributions

Quyen Tran[*1], Hai Nguyen[*2], Quan Dao[1], Zhuowei Li[†1], Nam Le[3], Trung Le[4], and Dimitris Metaxas[1]

[1]Rutgers University [2]Tuft University [3]HUST [4]Monash University

Abstract. Analytic Continual Learning (ACL) offers a computationally efficient alternative to gradient-based approaches. Recent ACL methods are based on Recursive Least Squares (RLS) and have achieved the state-of-the-art results compared to other alternatives. However, they falter significantly in Class-Incremental Learning scenarios characterized by Long-Tailed distributions. While the ill-conditioning of the autocorrelation (Gram) matrix is a known limitation of RLS, we demonstrate that class imbalance exacerbates this issue into a distinct spectral pathology: "tail" classes suffer from severe spectral collapse, rendering their subspaces numerically indistinguishable from noise. Standard Ridge Regression (L_2) fails to address this effectively as it applies *isotropic regularization* - a uniform penalty that is insufficient to stabilize the tail without over-shrinking the head. To address this, we propose ***Geometry-Spectral Rectification (GSR)***, a theoretically grounded framework that treats long-tailed learning as a spectral regularization problem. Unlike standard isotropic regularization (Ridge) which uniformly penalizes all eigenvalues, GSR acts as an *anisotropic spectral filter*, selectively inflating the collapsed eigenvalues of tail classes. We construct a structured, data-dependent spectral perturbation matrix Δ that selectively inflates collapsed tail eigen-directions of the Gram matrix. Theoretical analysis proves that GSR guarantees an improved stable rank for the Gram matrix, ensuring numerical stability. Extensive experiments show that GSR establishes a new state-of-the-art for analytic CIL, offering a superior trade-off between computational efficiency and robust generalization in long-tailed settings.

Keywords: Pretrained-based Continual Learning · Long-tailed distribution

1 Introduction

Analytic Continual Learning (ACL) [17, 19, 33–35] has recently emerged as a compelling "Green AI" alternative to gradient-based methods. By reformulating the learning objective into a Recursive Least Squares (RLS) problem [17,19,33],

[*] Equal contribution.
[†] Work done outside Amazon.

recent ACL methods achieve near-instantaneous updates via closed-form matrix operations, bypassing the computational burden of backpropagation, and even can achieve the upper-bound performance of Class Incremental Learning setting [18,19,26,27]. This efficiency makes them particularly attractive for edge devices and real-time applications where resources are scarce.

However, the theoretical elegance of these methods relies on a critical, often overlooked condition: the numerical stability of the RLS update. In realistic scenarios characterized by Long-Tailed distributions, their performance degrades catastrophically (Fig.1). While the ill-conditioning of the autocorrelation (Gram) matrix is a known limitation of RLS, and existing works often attribute this to simple data scarcity [2, 5, 7, 13, 23], we introduce a different view that the root cause is a fundamental linear algebra failure: ***Spectral Collapse***.

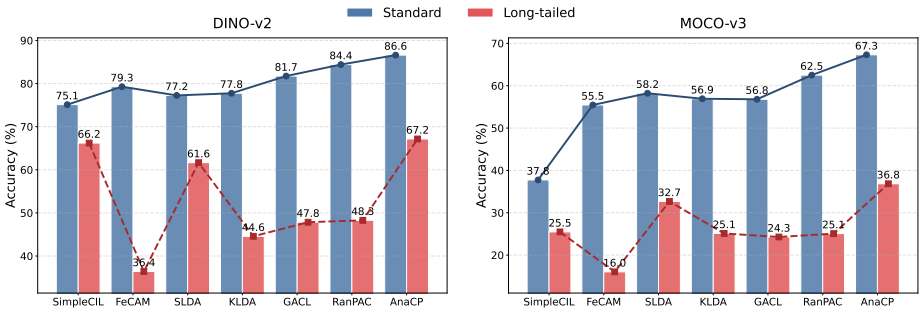

Fig. 1: Performance in *standard and long-tailed settings* of different ACL methods, on Split-Imagenet-R. See Fig.3 in Appendix D for full results on different datasets.

In RLS-based ACL, the optimal weight solution W^* is governed by the inverse of the autocorrelation (Gram) matrix G. We demonstrate that in imbalanced streams, G suffers from severe spectral skewness. "Head" classes dominate the principal components (large eigenvalues), while "Tail" classes correspond to directions with vanishingly small variance (near-zero eigenvalues). This results in *an ill-conditioned matrix* where the inversion process amplifies noise in the tail directions by orders of magnitude. Standard Ridge Regression (λI) fails to address this because it applies a uniform penalty, which is either too weak to stabilize the tail or too strong, suppressing the head.

To bridge this gap, we propose ***Geometry-Spectral Rectification (GSR)***, a non-iterative framework that treats data imbalance as a spectral regularization problem, leveraging the statistical geometry of the feature manifold. Particularly, we introduce a class-adaptive manifold mixing strategy that synthesizes 'virtual' support features to densify the sparse tail manifolds, leading to a better-conditioned empirical covariance estimate. Theoretically, we prove that our mixing strategy acts as a structured spectral filter, selectively inflating the smallest eigenvalues of the Gram matrix without distorting the dominant principal components. This effectively "heals" the broken spectrum, improving the stable rank and ensuring numerical stability during inversion.

Contributions: In summary, our contributions are threefold:

1. **Spectral Diagnosis:** We provide a rigorous analysis showing that the failure of RLS-based ACL in long-tailed settings stems from the spectral collapse phenomenon of the Gram matrix.
2. **Geometry-Spectral Rectification (GSR):** We propose a non-parametric rectification method that respects the hyperspherical geometry of modern PTMs. We prove that this strategy acts as a spectral preconditioner, explicitly improving the stable rank of the inverted matrix.
3. **SOTA Performance:** Extensive experiments on benchmark datasets show that GSR significantly outperforms existing ACL baselines, bridging the gap between analytic efficiency and robust generalization in long-tailed settings.

2 Related Works

2.1 Analytic Continual Learning (ACL)

Unlike gradient-based Continual Learning methods [21, 24, 25, 30] that require iterative optimization and often suffer from catastrophic forgetting or high computational costs, Analytic Continual Learning [8, 9, 17, 19, 33–35] aims to acquire new knowledge via closed-form solutions. Early works [33–35] utilized Recursive Least Squares (RLS) to update linear classifiers incrementally. Recently, the proliferation of large-scale frozen pre-trained models (PTMs) has revitalized ACL [17–19, 27], achieving state-of-the-art results, and even each upper bound performance on balanced benchmarks. While these methods achieve high efficiency, they implicitly assume a relatively balanced data stream. Our analysis reveals that in Long-Tailed (LT) scenarios, the standard RLS update used in SOTA methods like GACL [33], RanPAC [17], and AnaCP [19] becomes numerically unstable. The Gram matrix for tail classes suffers from *spectral collapse*, which standard isotropic regularization (Ridge) cannot fix without harming head classes. Our GSR framework specifically targets this spectral pathology.

2.2 Long-Tailed Learning in Continual Learning Settings

Real-world data streams frequently exhibit long-tailed distributions, where head classes dominate, and tail classes remain severely scarce. In static learning, this imbalance is commonly addressed through *Re-sampling* (e.g., SMOTE [4]), *Re-weighting* (e.g., Class-Balanced Loss [5]), or *Decoupling* strategies [13]. In Continual Learning, these ideas have been extended to mitigate the compounded effects of imbalance and catastrophic forgetting, for example via bias-correction layers [31], cosine normalization [10], and balanced replay buffers [15]. Within the ACL paradigm, Fang et al. [7] introduce inverse-frequency re-weighting into the analytic solution. However, such adjustments remain fundamentally *scalar modulations*: they rescale contributions without enriching the underlying feature geometry, and cannot recover missing directions in the representation space (See results in Tables 1 and 2). We argue that, in RLS-based ACL, tail-class degradation originates from data scarcity but ultimately manifests as *spectral*

collapse—the ill-conditioning of the autocorrelation matrix that governs the analytic solution. While existing approaches operate at the level of data counts or gradient magnitudes, they do not explicitly address this algebraic bottleneck. In contrast, we frame long-tailed ACL as a spectral conditioning problem and target it directly through explicit spectral regularization.

2.3 Implicit Spectral Regularization via Geodesic Interpolation

While standard ACL relies on explicit scalar regularization (Ridge, τI) to invert the Gram matrix, this isotropic penalty is suboptimal for long-tailed distributions. A more efficient alternative is implicit regularization via feature interpolation [29,32], which is asymptotically equivalent to adding a data-dependent regularizer to the Hessian [1,3]. However, existing strategies like Manifold Mixup [28] are ill-suited for our setting. First, they employ linear interpolation to train backbone parameters, whereas our framework targets covariance conditioning for the analytic inverse on frozen PTM embeddings. Second, linear interpolation on normalized embeddings creates a geometric inconsistency: it traverses the chord rather than the arc [6,14], causing norm shrinkage and energy suppression. This prevents effective inflation of collapsed tail-class eigenvalues. To overcome this, we propose Geodesic Interpolation (or Spherical Mixup) [14, 22]. By strictly following the manifold geometry, our approach preserves the unit norm constraint and prevents the variance attenuation characteristic of linear mixing. This ensures the spectral perturbation Δ_{struct} effectively reconditions the tail spectrum (Theorem 1) without the energy decay inherent in linear chordal traversal.

3 Method

3.1 Preliminaries: Analytic Continual Learning

This work consider Class Incremental Learning (CIL) setting, where we have a sequence of distinct tasks $t = 1, \ldots, T$. At each task, the model needs to deal with the dataset $D_t = (X_t, Y_t)$, so that we have (Z_t, Y_t), which are input features extracted from a pretrained model and their corresponding labels. RSL-based ACL methods [17, 19, 33–35] aims to minimize the Ridge Regression objective:

$$\mathcal{L}(W) = \|ZW - Y\|_F^2 + \tau\|W\|_F^2 \tag{1}$$

where Z and Y are all the features and their corresponding labels observed by the model so far. And W is the weight matrix of the model's classification head, whose pretrained backbone is frozen. The optimal solution is given in closed form by:

$$W^* = (G + \tau I)^{-1} Z^T Y \tag{2}$$

where $G = Z^T Z$ is the Gram matrix. In CIL setting, G and $Z^T Y$ can be updated recursively [17, 19]. The numerical stability of the inversion of $(G + \tau I)$ acts as a proxy for the bias-variance tradeoff, thereby influencing the model's generalization capability.

3.2 The Limitation of Isotropic Regularization

In this section, we analyze why standard RLS-based ACL fails in Long-Tailed (LT) scenarios through the lens of spectral geometry. We argue that the standard L_2 regularization creates a bad trade-off due to its *isotropic* nature.

a. The Dilemma of Isotropic Regularization (Ridge) Let $G = U\Sigma U^T$ be the eigendecomposition of the Gram matrix, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. The inversion in Eq. (2) modifies the spectrum as:

$$i^{th} \text{ eigenvalue of } (G + \tau I)^{-1} = \frac{1}{\lambda_i + \tau} \tag{3}$$

Here, the term τI acts as *Isotropic Regularization*. It adds a uniform spherical penalty τ to the variance in *all* directions of the feature space, regardless of the data distribution.

b. The Spectral Pathology of Long-Tailed Data: In imbalanced CIL, "tail" classes span subspaces with vanishingly small variance, leading to spectral collapse ($\lambda_{tail} \approx 0$). This creates a dilemma:

- *Under-regularization (Small τ):* If τ is small, the inverse term $(\lambda_{tail} + \tau)^{-1}$ explodes, causing the model to overfit to noise in the tail directions.
- *Over-regularization (Large τ):* If we increase τ to stabilize the tail, we inadvertently suppress the "head" classes (where $\lambda_{head} \gg \tau$). The term $(\lambda_{head} + \tau)^{-1}$ shrinks significantly, causing bias and underfitting for majority classes.

Thus, isotropic regularization cannot simultaneously stabilize the tail and preserve the head.

3.3 Geometry-Spectral Rectification (GSR)

To resolve this, we propose GSR, which utilizes Class-Adaptive Manifold Mixing to inject *Anisotropic Information* into the collapsed dimensions.

a. GSR as Structured Anisotropic Regularization Let z_1, z_2 be representations of two samples from a tail class with covariance Σ_{class}. We generate a rectified sample $\tilde{z}$ by mixing z_1 with z_2:

$$\tilde{z} = \frac{\gamma z_1 + (1-\gamma) z_2}{\|\gamma z_1 + (1-\gamma) z_2\|_2}, \quad \text{where } \gamma \text{ is mixing coefficient} \tag{4}$$

We utilize Spherical Mixup (detailed in Appendix B.2), which introduces non-linearity via hypersphere projection, which prevents the synthetic sample from lying in the linear span of the supports. This operation prevents the synthetic samples from collapsing onto the linear convex hull, thereby implicitly regularizing the covariance rank. In addition, from the perspective of Vicinal Risk Minimization [3], training on mixed samples is asymptotically equivalent to training

with a data-dependent regularizer. The second moment of the mixed data transforms as:

$$\mathbb{E}[\tilde{x}\tilde{x}^T] \approx G_{original} + \Delta \tag{5}$$

Crucially, unlike τI, the term Δ is proportional to the class covariance Σ_{class}. This implies that GSR acts as an *Anisotropic Spectral Regularizer*:

1. *Directional Alignment:* Δ adds "mass" (variance) specifically along the principal components of the class manifold, not in orthogonal null-spaces.
2. *Tail Rectification:* For a tail class, this selectively inflates the eigenvalues λ_{tail} ("thickening" the manifold) without needing a large global γ.

b. Theoretical Guarantee Define $\tilde{G} = G + \Delta$ is update of G after pertubation. We formally show that this strategy improves numerical stability through the increase of *stable rank*, which supports the success of GSR in recovering the intrinsic manifold structure of the data, while reducing overfitting driven by heavy-tail behavior in the spectrum. Formally, *stable rank* for a matrix G is $\mathrm{sr}(G) = \frac{\sum_{i=1}^{d} \lambda_i^2}{\lambda_1^2}$.

Theorem 1 (Quantitative stable-rank improvement under tail-mixup). *Fix an integer $r \in \{1, \ldots, d-1\}$ and let $U_h \in \mathbb{R}^{d\times r}$ collect the top-r orthonormal eigenvectors of G, and $U_t \in \mathbb{R}^{d\times(d-r)}$ collect the remaining ones. Define the head/tail projectors $P_h = U_h U_h^\top, P_t = U_t U_t^\top$. Assume the following blockwise quantitative bounds:*

$$\|U_h^\top \Delta U_h\|_2 \le \varepsilon, \quad \|U_h^\top \Delta U_t\|_2 \le \eta, \quad \|U_t^\top \Delta U_t\|_2 \le \bar{m}, \quad m \le \lambda_{\min}(U_t^\top \Delta U_t),$$

for some parameters $\varepsilon \ge 0$, $m > 0$, $\bar{m} \ge m$, and $\eta \ge 0$. Define

$$S_\tau \triangleq \|G + \tau I\|_F^2 = \sum_{i=1}^{d} (\lambda_i + \tau)^2,$$

and

$$\Lambda_+ \triangleq \frac{(\lambda_1 + \varepsilon) + (\lambda_{r+1} + \bar{m}) + \sqrt{\big((\lambda_1 + \varepsilon) - (\lambda_{r+1} + \bar{m})\big)^2 + 4\eta^2}}{2}.$$

Then for any ridge parameter $\tau \ge 0$,

$$sr(\widetilde{G} + \tau I) \ge \frac{S_\tau + (d-r)m^2}{(\Lambda_+ + \tau)^2}. \tag{6}$$

Moreover, a sufficient condition for strict *improvement is*

$$(d-r)m^2 > S_\tau \left(\frac{(\Lambda_+ + \tau)^2}{(\lambda_1 + \tau)^2} - 1 \right). \tag{7}$$

Corollary 1. *Assume $\Delta \succ 0$ satisfies $U_h^\top \Delta U_h = 0$ and $U_h^\top \Delta U_t = 0$ (i.e., Δ is supported entirely in the tail eigenspace of G), and the top eigenvalue remains head-dominated: $\lambda_1 \geq \lambda_{r+1} + \|U_t^\top \Delta U_t\|_2$. Then*

$$sr(\tilde{G}) > sr(G)$$

We give the full proof for all these theorems in the Appendix B.4. Theorem 1 shows that GSR *strictly increases* $\mathrm{sr}(\widetilde{G} + \tau I)$ by injecting structured mass into the tail eigenspace (via Δ), while controlling the increase of the top eigenvalue through blockwise bounds. In other words, GSR does not blindly add isotropic noise; it *densifies* the collapsed tail spectrum in a targeted way.

Furthermore, GSR acts as an *Anisotropic Spectral Regularizer*. Unlike standard L_2 regularization (Ridge), which adds a constant damping factor τI uniformly across all directions (penalizing head and tail classes equally), GSR selectively "inflates" only the collapsed eigenvalues of tail subspaces via Δ. This allows the model to retain high precision on head classes (where eigenvalues are already large) while stabilizing the learning of tail classes, effectively resolving the stability-plasticity dilemma at the spectral level.

One might be concerned that injecting Δ into the null-space could amplify noise. However, in the long-tailed regime, the zero eigenvalues in tail classes predominantly stem from sample scarcity (feature collapse) rather than the absence of semantic information. Crucially, unlike standard Tikhonov regularization (Ridge Regression), which adds isotropic noise (white noise), our injected Δ is anisotropic. By leveraging the latent geometric structure within the scarce samples themselves, we selectively inflate the collapsed spectral components of tail classes, ensuring the restored geometry reflects plausible semantic variations rather than arbitrary disturbances.

c. Our technical method

Adaptive Mixing Strategy. To operationalize Theorem 1, we define the mixing intensity α_c for class c inversely proportional to its spectral health:

$$\alpha_c = \alpha_{base} + (1 - \alpha_{base}) \cdot e^{-\xi \cdot N_c} \tag{8}$$

where N_c is the sample count of class c, and ξ is the decay rate. This ensures that tail classes undergo intensive manifold mixing (high variance injection), while head classes preserve their original precise feature statistics.

Rectified Gram Matrix Update via Tail Mixup. We leverage the scarce real samples available for tail classes to induce spectral rectification directly. Let $\mathcal{X}_{real} = \{x_i\}_{i=1}^{N_c}$ be the set of current real samples, and $\mathcal{Z}_{real} = \{z_i\}_{i=1}^{N_c}$ be the set of their corresponding representations for a specific class c.

Due to the scarcity of samples ($N_c \ll D$, D is the feature dimension), the empirical covariance of $\mathcal{Z}_{real}$ is rank-deficient. To counteract this phenomenon, we synthesize a set of augmented samples $\mathcal{Z}_{aug}$ by interpolating along the geodesic

Algorithm 1 Geometry-Spectral Rectification (GSR) for RLS-based ACL

Require: Initial Gram matrix $G_0 = \mathbf{0}$, Cross-correlation $Q_0 = \mathbf{0}$, Regularization τ.
Require: Data coming in a sequence of tasks $\mathcal{T} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$.
Ensure: Class weights W_{final}.
for task $t = 1$ to T **do**
  Receive data $\mathcal{D}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_t}$.
  Extract features $Z_t = f_\theta(X_t) \in \mathbb{R}^{N_t \times D}$, where D is feature dimension.
  Initialize augmented buffer $\mathcal{Z}_{\text{aug}} = \emptyset$.
  for class $c \in \mathcal{C}_t$ **do**
    // **1. Spectral Diagnosis & Adaptive Mixing**
    Calculate sample count N_c.
    Compute mixing intensity α_c ▷ Using Eq.8
      $\alpha_c = \alpha_{\text{base}} + (1 - \alpha_{\text{base}}) \cdot e^{-\xi \cdot N_c}$
    // **2. Geometry-Spectral Rectification**
    Initialize augmented buffer $\mathcal{Z}_{\text{aug},c} = \emptyset$.
    for each sample pair $(\mathbf{z}_i, \mathbf{z}_j) \in (Z_{t,c}, Z_{t,c})$ (in the same class c) **do**
      Sample $\gamma \sim \text{Beta}(\alpha_c, \alpha_c)$.
      Synthesize spherical mixup sample ▷ Using Eq.9
        $\tilde{\mathbf{z}} = \frac{\gamma \mathbf{z}_i + (1-\gamma)\mathbf{z}_j}{\|\gamma \mathbf{z}_i + (1-\gamma)\mathbf{z}_j\|_2}$
      $\mathcal{Z}_{\text{aug},c} \leftarrow \mathcal{Z}_{\text{aug},c} \cup \{\tilde{\mathbf{z}}\}$.
    end for
    $\mathcal{Z}_{\text{aug}} \leftarrow \mathcal{Z}_{\text{aug}} \cup \mathcal{Z}_{\text{aug},c}$.
  end for
  // **3. Block-wise Update (Dual Stream)**
  Update $Q_t \leftarrow Q_{t-1} + Z_t^T Y_t + \beta Z_{\text{aug}}^T Y_{\text{aug}}$.
  Update $G_t \leftarrow G_{t-1} + Z_t^T Z_t + \beta Z_{\text{aug}}^T Z_{\text{aug}}$. ▷ Using Eq.10
  // **4. Analytic Solution**
  Compute weights: $W_t = (G_t + \tau I)^{-1} Q_t$.
end for
return $W_{final} = W_T$

paths between real data pairs. Unlike standard linear mixup, which shrinks the feature norm, our spherical formulation preserves the hyper-spherical geometry:

$$\tilde{z} = \frac{\gamma z_i + (1-\gamma) z_j}{\|\gamma z_i + (1-\gamma) z_j\|_2}, \quad \text{where } z_i, z_j \in \mathcal{Z}_{real}, \text{ and } \gamma \sim \text{Beta}(\alpha_c, \alpha_c) \tag{9}$$

Here, the mixing coefficient γ is drawn from the adaptive Beta distribution defined in Eq.8, the mixing pairs (x_i, x_j) are randomly selected from the available real samples of the same class, and the normalization term ensures that $\tilde{z}$ remains on the unit hyper-sphere, avoiding the norm shrinkage phenomenon after mixing.

Crucially, to ensure both fidelity to the true distribution and structural robustness, we utilize a *Dual-Stream Update* strategy. The Gram matrix is updated using the union of both real and augmented data streams. The update rule be-

comes:

$$\tilde{G}_t = G_{t-1} + \underbrace{\sum_{z \in \mathcal{Z}_{real}} z z^T}_{\text{Fidelity Term}} + \underbrace{\beta \sum_{\tilde{z} \in \mathcal{Z}_{aug}} \tilde{z} \tilde{z}^T}_{\text{Rectification Term}} \tag{10}$$

where β is a balancing coefficient. Particularly,

- *Fidelity Term:* ensures the model learns the precise feature statistics from the real data.
- *Rectification Term* injects variance along the geodesic arcs into the null-space of the tail class manifold, preventing the spectral collapse described in Theorem 1.

By combining both, GSR preserves the original semantic identity while artificially "thickening" the manifold to ensure numerical stability during the inversion of RLS-based ACL methods in long-tailed settings. The complete procedure of our technical method is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Datasets and Protocols. We evaluate GSR on four benchmark datasets widely used in Class-Incremental Learning (CIL), but in Long-Tailed setting: Split-CIFAR-100, Split-ImageNet-R, Split-Tiny-ImageNet, and Split-CUB-200. Further details regarding the long-tailed datasets can be found in Appendix C.
Baselines. We compare GSR against a wide range of representative ACL baselines, categorized into two groups: *(i) RLS-based methods*, including RanPAC [17], AnaCP [19], GACL [33], and AIR [7]; and *(ii) Covariance and Prototype-based methods*, including FECAM [8], SLDA [9], KLDA [18], and SimpleCIL [11].
Evaluation Metrics. We report the performance using: Last Accuracy (A_{last}) and Average Accuracy (A_{avg}).

All experiments use DINO-v2 or MoCo-v3 as pretrained backbones and are averaged over 3 random seeds to ensure statistical significance. See further details in Appendix C.

4.2 Experimental results

Performance analysis on different backbones

a. Performance Reversal: Simple Baselines Outperform RLS-based "SOTA methods" in Long-tailed Settings Observing Table 1 and Figure 1, we identify an intriguing phenomenon: in a long-tailed setting with a robust backbone like DINO-v2, the simple method SimpleCIL achieves remarkably high performance (80.42% on CIFAR-100), far surpassing RLS-based ACL methods such as GACL (48.78%) and RanPAC (49.30%).

Typically, in balanced settings, RLS leverages the Gram matrix to utilize second-order statistics for feature decorrelation and minimizes the global Mean Squared Error (MSE), yielding precise decision boundaries. However, this strength

becomes a liability in long-tailed scenarios. In the high-dimensional feature space of DINO-v2, combined with data scarcity, the covariance estimation becomes ill-conditioned. This leads to severe overfitting, causing performance to degrade significantly compared to the simple baseline.

b. Effectiveness of GSR on DINO-v2 backbone: Our method acts as a powerful regularizer, effectively addressing the critical drawback mentioned above:

- *Revitalizing RLS-based Algorithms:* Integrating our method (+Ours) yields substantial performance gains for GACL and RanPAC (e.g., GACL improves from 48.78% $\rightarrow$ 65.81% on Split-CIFAR-100). This demonstrates that our approach effectively stabilizes parameter estimation in high-dimensional spaces, mitigating the overfitting issue observed in the original baselines.
- *Achieving State-of-the-Art with AnaCP:* While SimpleCIL serves as a highly competitive baseline, the combination of AnaCP and our method (AnaCP + Ours) reaches 82.51%, surpassing SimpleCIL to establish a new State-of-the-Art. This confirms that to fully exploit the power of DinoV2, a strong Analytic algorithm (AnaCP) supported by high-quality augmented data from our method is essential.
- *Our advantage over long-tail-aware RLS methods:* Among the baselines, AIR stands out as the only RLS-based ACL method explicitly designed for long-tailed settings by introducing inverse-frequency re-weighting into the Gram matrix. However, we argue that such adjustments remain fundamentally scalar modulations. They merely rescale the contribution of existing features without enriching the underlying feature geometry, failing to recover missing directions in the representation space. Consequently, as shown in the tables, AIR consistently underperforms compared to our method, which actively rectifies the geometric structure.

c. Consistency on MoCo-v3: The improvement trend remains consistent on MoCo-v3 (Table 2). AnaCP + Ours continues to lead with a significant margin over the baseline (increasing from 57.38% to 66.01% on Split-CIFAR-100). This highlights the generalization capability of the proposed method across different feature extractor architectures.

4.3 Ablation Study

Impact of Geometric Rectification Strategies. To validate the effectiveness of our method, we incrementally evaluate the components of GSR on the Split-CUB-200 dataset. Table 3 reveals a critical insight: applying standard *Linear Mixup* in the high-dimensional feature space of DINO-v2 actually degrades performance (80.73% $\rightarrow$ 77.85%). This confirms that linear interpolation within a hyperspherical embedding space can create "weak" features that drift off the manifold (norm shrinkage), confusing the classifier. In contrast, our GSR respects the geometry, boosting accuracy to 82.01% and 47.45%. Finally, incorporating the Adaptive Mixing strategy (α_c) further refines the spectral density of tail classes, achieving the best performances of 82.46% and 48.33%.

Table 1: Overall performance comparison, using DINO-v2 as PTM. The best result is shown in **bold**, and the second-best is underlined.

Method	Split CIFAR-100		Split ImageNet-R		Split-Tiny-Imagenet		Split CUB-200	
	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)
SLDA	$72.28_{\pm 0.67}$	$76.06_{\pm 1.30}$	$61.65_{\pm 0.32}$	$64.60_{\pm 1.08}$	$71.19_{\pm 0.43}$	$73.90_{\pm 1.49}$	$81.11_{\pm 0.82}$	$86.21_{\pm 0.80}$
FeCAM	$33.35_{\pm 0.48}$	$37.97_{\pm 1.71}$	$36.41_{\pm 1.29}$	$42.89_{\pm 2.77}$	$35.28_{\pm 0.27}$	$39.68_{\pm 1.07}$	$70.55_{\pm 0.54}$	$79.31_{\pm 0.90}$
SimpleCIL	$\underline{80.42}_{\pm 0.33}$	$\underline{87.21}_{\pm 0.40}$	$66.20_{\pm 1.28}$	$\underline{73.73}_{\pm 0.60}$	$\underline{78.21}_{\pm 0.13}$	$\underline{83.60}_{\pm 1.03}$	$82.07_{\pm 0.29}$	$87.84_{\pm 0.83}$
KLDA	$49.30_{\pm 1.37}$	$55.38_{\pm 2.72}$	$44.57_{\pm 1.74}$	$51.26_{\pm 1.97}$	$52.31_{\pm 0.78}$	$54.84_{\pm 1.32}$	$78.16_{\pm 0.98}$	$85.98_{\pm 1.26}$
AIR	$64.87_{\pm 0.93}$	$69.92_{\pm 2.48}$	$54.51_{\pm 1.35}$	$58.87_{\pm 1.29}$	$69.19_{\pm 0.45}$	$69.89_{\pm 1.41}$	$80.74_{\pm 1.04}$	$87.39_{\pm 1.38}$
GACL	$48.78_{\pm 1.32}$	$58.84_{\pm 2.97}$	$47.84_{\pm 2.18}$	$51.53_{\pm 1.34}$	$42.98_{\pm 0.37}$	$53.65_{\pm 1.32}$	$80.73_{\pm 1.07}$	$87.20_{\pm 1.17}$
+Ours	$65.51_{\pm 1.25}$	$71.79_{\pm 2.69}$	$61.58_{\pm 0.58}$	$65.90_{\pm 0.72}$	$68.88_{\pm 0.29}$	$70.70_{\pm 1.27}$	$82.46_{\pm 0.58}$	$87.98_{\pm 1.00}$
RanPAC	$49.30_{\pm 1.27}$	$59.03_{\pm 1.03}$	$48.28_{\pm 1.67}$	$52.94_{\pm 2.97}$	$43.62_{\pm 0.36}$	$54.84_{\pm 1.23}$	$81.11_{\pm 0.89}$	$87.52_{\pm 1.25}$
+Ours	$66.68_{\pm 1.25}$	$71.98_{\pm 2.65}$	$62.08_{\pm 0.62}$	$66.65_{\pm 1.07}$	$69.57_{\pm 0.29}$	$71.45_{\pm 1.25}$	$83.03_{\pm 0.62}$	$88.64_{\pm 1.02}$
AnaCP	$72.90_{\pm 0.01}$	$78.54_{\pm 1.72}$	$\underline{67.15}_{\pm 0.19}$	$71.86_{\pm 0.11}$	$72.97_{\pm 0.26}$	$76.30_{\pm 1.23}$	$\underline{83.56}_{\pm 0.66}$	$\underline{89.07}_{\pm 1.09}$
+Ours	$\mathbf{82.51}_{\pm 0.01}$	$\mathbf{88.29}_{\pm 0.73}$	$\mathbf{72.13}_{\pm 0.84}$	$\mathbf{77.96}_{\pm 0.47}$	$\mathbf{78.81}_{\pm 0.16}$	$\mathbf{83.71}_{\pm 0.96}$	$\mathbf{85.85}_{\pm 0.43}$	$\mathbf{90.53}_{\pm 0.85}$

Table 2: Overall performance comparison, using MoCo-v3 as PTM. The best result is shown in **bold**, and the second-best is underlined.

Method	Split CIFAR-100		Split ImageNet-R		Split-Tiny-Imagenet		Split CUB-200	
	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)	A_{last} (↑)	A_{avg} (↑)
SLDA	$52.67_{\pm 0.59}$	$59.38_{\pm 2.64}$	$32.67_{\pm 1.25}$	$37.22_{\pm 3.05}$	$57.50_{\pm 0.21}$	$62.65_{\pm 1.47}$	$41.75_{\pm 0.53}$	$50.78_{\pm 1.37}$
FeCAM	$27.64_{\pm 0.55}$	$29.28_{\pm 0.03}$	$16.04_{\pm 1.38}$	$20.06_{\pm 2.04}$	$27.00_{\pm 0.22}$	$29.34_{\pm 0.17}$	$24.19_{\pm 1.11}$	$30.43_{\pm 1.68}$
SimpleCIL	$55.32_{\pm 0.89}$	$66.25_{\pm 1.77}$	$25.48_{\pm 0.62}$	$32.32_{\pm 1.89}$	$58.65_{\pm 0.75}$	$66.57_{\pm 1.03}$	$34.48_{\pm 0.81}$	$44.78_{\pm 1.63}$
KLDA	$38.11_{\pm 1.24}$	$44.84_{\pm 2.95}$	$25.11_{\pm 1.93}$	$30.29_{\pm 2.94}$	$41.70_{\pm 0.47}$	$47.12_{\pm 0.90}$	$34.42_{\pm 1.07}$	$44.64_{\pm 0.88}$
AIR	$\underline{65.13}_{\pm 0.21}$	$\underline{74.37}_{\pm 1.12}$	$36.71_{\pm 0.47}$	$\underline{44.80}_{\pm 2.24}$	$\underline{63.75}_{\pm 0.22}$	$\underline{72.59}_{\pm 0.68}$	$49.37_{\pm 0.45}$	$59.57_{\pm 1.54}$
GACL	$30.78_{\pm 0.51}$	$41.48_{\pm 2.35}$	$24.28_{\pm 1.70}$	$30.76_{\pm 2.64}$	$30.01_{\pm 0.52}$	$39.69_{\pm 0.44}$	$34.48_{\pm 1.27}$	$45.77_{\pm 1.05}$
+Ours	$52.40_{\pm 1.54}$	$60.67_{\pm 2.69}$	$34.86_{\pm 1.32}$	$39.49_{\pm 2.75}$	$58.04_{\pm 0.19}$	$63.93_{\pm 1.28}$	$48.33_{\pm 0.80}$	$56.72_{\pm 0.97}$
RanPAC	$31.23_{\pm 0.57}$	$41.84_{\pm 2.63}$	$25.07_{\pm 1.59}$	$32.24_{\pm 2.68}$	$30.62_{\pm 0.59}$	$40.95_{\pm 0.67}$	$35.17_{\pm 1.29}$	$46.44_{\pm 1.23}$
+Ours	$53.28_{\pm 1.55}$	$61.32_{\pm 2.55}$	$35.12_{\pm 1.32}$	$40.94_{\pm 2.77}$	$58.82_{\pm 0.21}$	$64.84_{\pm 1.17}$	$49.02_{\pm 0.80}$	$57.04_{\pm .83}$
AnaCP	$57.38_{\pm 1.14}$	$63.30_{\pm 2.69}$	$\underline{36.84}_{\pm 1.46}$	$41.95_{\pm 2.78}$	$62.02_{\pm 0.34}$	$66.71_{\pm 1.21}$	$\underline{52.11}_{\pm 0.86}$	$\underline{60.27}_{\pm 1.52}$
+Ours	$\mathbf{66.01}_{\pm 0.26}$	$\mathbf{76.34}_{\pm 1.29}$	$\mathbf{37.21}_{\pm 0.65}$	$\mathbf{46.05}_{\pm 1.19}$	$\mathbf{64.37}_{\pm 0.23}$	$\mathbf{73.27}_{\pm 0.52}$	$\mathbf{56.05}_{\pm 0.91}$	$\mathbf{65.52}_{\pm 1.32}$

Sensitivity to Imbalance Ratio (ρ) We investigate the robustness of GSR under varying degrees of data imbalance on Split-CIFAR-100. As shown in Table 4, as the imbalance ratio ρ increases from 10 (moderate) to 150 (extreme), the baseline GACL suffers a catastrophic drop of over 42% (81.23% → 38.60%). However, GSR demonstrates remarkable resilience. The performance gap between our method and the baseline widens significantly as the task becomes harder: from a +3.99% gain at $\rho = 10$ to a massive +16.9% gain at $\rho = 150$. This proves that GSR is not merely a constant regularizer, but an active *spectral rectifier* that becomes increasingly critical as the tail lengthens.

Numerical Stability Analysis. To empirically verify Theorem 1, we monitor the *Stable Rank* ($\mathrm{sr}(G + \tau I)$) of the regularized Gram matrix throughout the

Table 3: Ablation Study on Component Effectiveness. We incrementally add components to the baseline (Standard Ridge) to validate the contribution of each module in GSR. Evaluated on Split-CUB-200, using DINO-v2 as the pretrained backbone.

Method Variant	Components			Last Acc (%)	
	Mixup	Spherical	Adaptive (α_c)	DINO-v2	MoCo-v3
Standard Ridge (GACL)	-	-	-	$80.73_{\pm 1.07}$	$34.48_{\pm 1.27}$
+ Linear Mixup	✓	-	-	$77.85_{\pm 1.63}$	$45.90_{\pm 0.83}$
+ Spherical Mixup	✓	✓	-	$\underline{82.01}_{\pm 0.47}$	$\underline{47.45}_{\pm 0.67}$
+ Our full GSR	✓	✓	✓	$\mathbf{82.46}_{\pm 0.58}$	$\mathbf{48.33}_{\pm 0.80}$

Table 4: Robustness to Data Imbalance. Comparison of Last Accuracy (%) under different imbalance ratios (ρ), on Split-CIFAR-100, using DINO-v2. Higher ρ indicates a more severe imbalance. GACL+GSR demonstrates superior robustness compared to the base method GACL.

Method	Imbalance Ratio (ρ)				
	$\rho = 10$	$\rho = 20$	$\rho = 50$	$\rho = 100$	$\rho = 150$
GACL	81.23	74.19	59.99	48.78	38.60
GACL + GSR	85.22	82.58	75.16	65.51	55.50
Improvement (Δ)	*+3.99*	*+8.39*	*+15.17*	*+16.73*	*+16.9*

incremental learning process. A stable rank close to 1.0 indicates severe spectral collapse, where the matrix is numerically indistinguishable from a rank-1 matrix (dominated by a single direction). Figure 2 shows that the baseline GACL consistently hovers near the collapse point (sr ≈ 1.02). In contrast, GSR effectively "thickens" the spectrum, maintaining a healthy stable rank (sr ≈ 2.80) across all tasks. This confirms that our method successfully injects variance into the null-space of tail classes, ensuring a well-conditioned inversion.

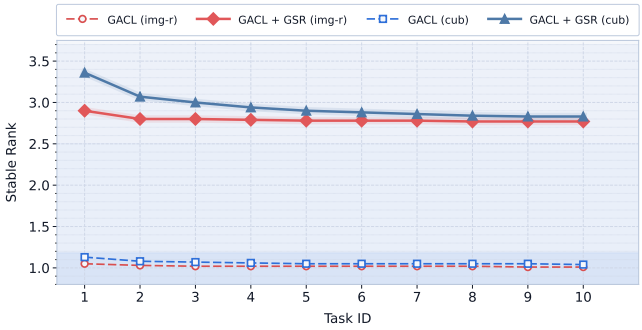

Fig. 2: Numerical Stability Analysis. We report the ***Stable Rank*** of the matrix $(G+\tau I)$ at different incremental stages. While GACL (dashed lines) suffers from severe spectral collapse with a stable rank near 1.0, our proposed GSR (solid lines) consistently helps maintain a higher stable rank across all tasks, indicating a well-conditioned feature space and effective mitigation of collapse. *"img-r"* stands for *Split-Imagenet-R*, and *"cub"* stands for *Split-CUB-200*.

Performance Breakdown by Class Frequency. Analysis of Class Frequency Impact. To investigate the source of the performance gains, Table 5 breaks down the accuracy for the most frequent (Head) and least frequent (Tail) classes.

First, GSR demonstrates a remarkable ability to resurrect tail classes. In standard GACL, tail performance is extremely low (e.g., 12.50% on Split-CIFAR-100 and 13.33% on Split-Tiny-ImageNet), indicating the model mostly ignores these minority classes. However, integrating GSR yields massive improvements, boosting tail accuracy by 3× to 4× (reaching 38.00% and 52.70%, respectively). This confirms that our GSR effectively prevents feature collapse in sparse regions.

Second, regarding the head classes, we observe a negligible trade-off. While there is a slight dip in accuracy for CIFAR-100 and Tiny-ImageNet (approx. 1%), this is a typical phenomenon in long-tailed learning due to the alleviation of majority-class bias. Crucially, the massive gains in the tail far outweigh this minor drop, resulting in a significantly higher overall accuracy.

Interestingly, on ImageNet-R, GSR improves both head (+5.62%) and tail (+26.29%) performance. This indicates that GSR does not merely shift the decision boundary bias. Instead, it rectifies the feature distribution to be more representative, enabling the analytic solver to generalize better across both head and tail classes.

Table 5: Breakdown of accuracy by class frequency groups (Head, Tail). DINO-v2 is used as the pretrained backbone.

Dataset	Method	Group Accuracy (%)		Overall (%)
		Head (5 classes)	Tail (5 classes)	
CIFAR100	GACL	95.38	12.50	48.78
	GACL + GSR	95.25	38.00	65.81
Imagenet-R	GACL	83.38	21.09	47.84
	GACL + GSR	89.00	47.38	61.58
Tiny-Imagenet	GACL	95.00	13.33	42.98
	GACL + GSR	94.00	52.70	68.88

GSR vs. common alternatives. To provide a comprehensive assessment, Table 6 compares GSR against other feature augmentation techniques that enrich the latent space, specifically Gaussian sampling (Covariance Shrinkage/Transfer) and Inter-class Mixup.

Efficiency vs. Accuracy Trade-off. Gaussian-based methods provide strong regularization, particularly for the GACL baseline (e.g., boosting Tiny-ImageNet to 75.21%). However, this comes at a prohibitive computational cost of $O(D^3)$ due to the Cholesky decomposition required for covariance estimation. This complexity makes them ill-suited for real-time continual learning on edge devices with high-dimensional backbones like DINO-v2 ($D = 768$ or 1024).

In contrast, our GSR operates with linear $O(D)$ complexity, similar to Mixup. While GACL+GSR trails behind the computationally heavy Gaussian methods, the combination of AnaCP + GSR yields remarkable results. As shown in the bottom section of Table 6, AnaCP + GSR achieves 78.81% on Tiny-ImageNet

and 72.13% on ImageNet-R, clearly outperforming both Gaussian variants and Mixup. On CUB-200, it remains highly competitive (85.85%) compared to the best Gaussian method (85.96%), while being significantly faster. This confirms that GSR offers the best trade-off between accuracy and efficiency.

Synergy with AnaCP. We want to highlight *why AnaCP integrates naturally with our method.* A key ingredient of AnaCP [19] is the Contrastive Projection Layer, which encourages the learned features to form well-separated clusters—a geometric structure closely related to the Neural Collapse phenomenon in contrastive learning [20]. However, under head–tail (long-tailed) data imbalance, contrastive learning is prone to *spectral (dimensional) collapse* [12]: the representation remains non-degenerate (features are not identical), yet most variation concentrates in only a few directions, making the embedding effectively low-rank. Our method directly addresses this issue. In particular, Theorem 1 shows that we increase the stable rank of the learned representation, mitigating dimensional collapse and thereby strengthening the feature geometry that AnaCP relies on—ultimately improving AnaCP's performance.

Table 6: Comparison of computational complexity and accuracy of different feature augmentation methods, using DINO-v2 as the pretrained backbone. D is the feature dimension. See Appendix A for details on the computational complexity.

Method	**Computational Complexity**	**Last Accuracy (%) ↑**		
		CUB-200	Imagenet-R	Tiny-Imagenet
GACL + Gauss (cov shrinkage)	$O(D^3)$	82.50	62.85	69.38
GACL + Gauss (cov transfer)	$O(D^3)$	**83.04**	**70.85**	**75.21**
GACL + Inter-class mixup	$O(D)$	82.43	61.53	68.76
GACL + GSR (Ours)	$O(D)$	82.46	61.58	68.78
AnaCP + Gauss (cov shrinkage)	$O(D^3)$	**85.96**	71.36	76.73
AnaCP + Gauss (cov transfer)	$O(D^3)$	<u>85.85</u>	71.01	75.78
AnaCP + Inter-class mixup	$O(D)$	85.69	<u>71.76</u>	<u>77.62</u>
AnaCP + GSR (Ours)	$O(D)$	<u>85.85</u>	**72.13**	**78.81**

5 Conclusion

In this work, we identified Spectral Collapse as a critical bottleneck for state-of-the-art Analytic Continual Learning (ACL) methods in long-tailed scenarios. We demonstrated that the standard isotropic regularization in Ridge Regression is insufficient to handle the extreme eigenvalue skewness caused by class imbalance. To address this, we proposed Geometry-Spectral Rectification (GSR), a theoretically grounded framework acting as an anisotropic spectral filter. By leveraging Spherical Geodesic Mixing and Adaptive Spectral Densification, GSR effectively "heals" the collapsed subspaces of tail classes without hallucinating arbitrary noise or inducing semantic bias from head classes. Extensive experiments validate the effectiveness of GSR in the practical yet challenging setting of Long-Tailed Class Incremental Learning.

References

1. Bishop, C.M.: Training with noise is equivalent to tikhonov regularization. Neural Comput. **7**(1), 108–116 (Jan 1995). https://doi.org/10.1162/neco.1995.7.1.108, https://doi.org/10.1162/neco.1995.7.1.108
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (2019)
3. Chapelle, O., Weston, J., Bottou, L., Vapnik, V.: Vicinal risk minimization. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems. vol. 13. MIT Press (2000), https://proceedings.neurips.cc/paper_files/paper/2000/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/JAIR.953, https://doi.org/10.1613/jair.953
5. Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S.J.: Class-balanced loss based on effective number of samples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9268–9277. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00949, http://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
7. Fang, D., Zhu, Y., Lin, Z., Chen, C., Zeng, Z., Zhuang, H.: AIR: Analytic imbalance rectifier for continual learning (2024). https://doi.org/10.48550/arXiv.2408.10349, https://arxiv.org/abs/2408.10349
8. Goswami, D., Liu, Y., Twardowski, B., van de Weijer, J.: FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=Asx5eDqFZl
9. Hayes, T.L., Kanan, C.: Lifelong machine learning with deep streaming linear discriminant analysis. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
10. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
11. Janson, P., Zhang, W., Aljundi, R., Elhoseiny, M.: A simple baseline that questions the use of pretrained-models in continual learning. CoRR **abs/2210.04428** (2022). https://doi.org/10.48550/ARXIV.2210.04428, https://doi.org/10.48550/arXiv.2210.04428
12. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348 (2021)
13. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=r1gRTCVFvB

14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb
15. Kim, C.D., Jeong, J., Kim, G.: Imbalanced continual learning with partitioning reservoir sampling. In: European Conference on Computer Vision. pp. 411–428. Springer (2020)
16. Liu, X., Hu, Y.S., Cao, X.S., Bagdanov, A.D., Li, K., Cheng, M.M.: Long-tailed class incremental learning. In: European Conference on Computer Vision. pp. 495–512. Springer (2022)
17. McDonnell, M., Gong, D., Parvaneh, A., Abbasnejad, E., van den Hengel, A.: RanPAC: Random projections and pre-trained models for continual learning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=aec58UfBzA
18. Momeni, S., Mazumder, S., Liu, B.: Continual learning using a kernel-based method over foundation models. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
19. Momeni, S., Xiao, C., Liu, B.: Anacp: Toward upper-bound continual learning via analytic contrastive projection. CoRR **abs/2511.13880** (2025). https://doi.org/10.48550/ARXIV.2511.13880, https://doi.org/10.48550/arXiv.2511.13880
20. Nguyen, T., Jiang, R., Aeron, S., Ishwar, P., Brown, D.R.: On neural collapse in contrastive learning with imbalanced datasets. In: 2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2024)
21. Saha, G., Garg, I., Roy, K.: Gradient projection memory for continual learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=3AOj0RCNC2
22. Shoemake, K.: Animating rotation with quaternion curves. In: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques. p. 245–254. SIGGRAPH '85, Association for Computing Machinery, New York, NY, USA (1985). https://doi.org/10.1145/325334.325242, https://doi.org/10.1145/325334.325242
23. Thanh, N.X., Le, A.D., Tran, Q., Le, T., Van, L.N., Nguyen, T.H.: Few-shot, no problem: Descriptive continual relation extraction. In: Walsh, T., Shah, J., Kolter, Z. (eds.) Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025. pp. 25282–25290. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I24.34715, https://doi.org/10.1609/aaai.v39i24.34715
24. Tran, Q., Nguyen, H., Phan, H., Dao, Q., Ngo, L., Than, K., Phung, D., Metaxas, D., Le, T.: An optimal transport-driven approach for cultivating latent space in online incremental learning (2026), https://arxiv.org/abs/2211.16780
25. Tran, Q., Phan, H., Le, M., Truong, T., Phung, D., Ngo, L., Nguyen, T., Ho, N., Le, T.: Leveraging hierarchical taxonomies in prompt-based continual learning (2025), https://arxiv.org/abs/2410.04327
26. Tran, Q., Thanh, N.X., Anh, N.H., Hai, N.L., Le, T., Ngo, L.V., Nguyen, T.H.: Preserving generalization of language models in few-shot continual relation extraction. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the

2024 Conference on Empirical Methods in Natural Language Processing. pp. 13771–13784. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.763, https://aclanthology.org/2024.emnlp-main.763/

27. Tran, Q., Tran, T.L., Doan, K., Tran, T., Phung, D., Than, K., Le, T.: Boosting multiple views for pretrained-based continual learning. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=AZR4R3lw7y
28. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6438–6447. PMLR (2019), http://proceedings.mlr.press/v97/verma19a.html
29. Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., Wu, C.: Implicit semantic data augmentation for deep networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 12635–12644 (2019)
30. Wang, Z., Zhang, Z., Lee, C., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J.G., Pfister, T.: Learning to prompt for continual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 139–149. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00024, https://doi.org/10.1109/CVPR52688.2022.00024
31. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 374–382 (2019)
32. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1Ddp1-Rb
33. Zhuang, H., Chen, Y., Fang, D., He, R., Tong, K., Wei, H., Zeng, Z., Chen, C.: GACL: Exemplar-free generalized analytic continual learning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (Dec 2024)
34. Zhuang, H., Weng, Z., He, R., Lin, Z., Zeng, Z.: GKEAL: Gaussian kernel embedded analytic learning for few-shot class incremental task. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7746–7755 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00748
35. Zhuang, H., Weng, Z., Wei, H., Xie, R., Toh, K.A., Lin, Z.: ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 11602–11614. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/4b74a42fc81fc7ee252f6bcb6e26c8be-Paper-Conference.pdf