SCULPT-LoRA: Spectral Sculpting for Continual Low-Rank Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Low-Rank Adaptation (LoRA) enables parameter-efficient continual learning, yet a fixed rank budget neither safeguards pretrained representations nor guarantees effective use of adaptation capacity. We identify two complementary spectral challenges: interference with dominant pretrained directions and concentration of update energy in a few singular modes. Addressing both requires controlling *where* updates act and *how* their energy is distributed. To alleviate these problems, we propose `SCULPT-LoRA`, a framework that couples *Elastic Principal Subspace Protection (EPSP)* with *Polynomial Spectral Balancing (PSB)*. EPSP constructs a fixed orthonormal adaptation basis by attenuating its overlap with the pretrained principal subspace, allowing controlled modification rather than strict exclusion. PSB reshapes the trainable factor through a compact Gram-matrix polynomial during the forward pass, reducing spectral concentration. The two mechanisms are linked by row orthonormality, which transfers the nonzero singular spectrum of the conditioned factor exactly to the effective LoRA update. We derive a sharp worst-case bound on principal-subspace perturbation that decreases with protection strength under a fixed update-energy budget, and establish conditions for condition-number improvement under polynomial spectral shaping. Experiments across five continual learning benchmarks show improved final and average accuracy, with ablations supporting the complementary benefits of elastic protection and spectral balancing.

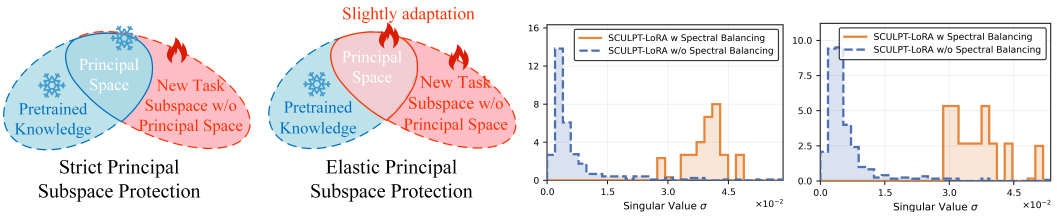
1 Introduction

Continual learning (CL) aims to acquire knowledge from sequential tasks while retaining previously learned capabilities Kirkpatrick et al. (2017); Qiu et al. (2025b); Wang et al. (2021). With the emergence of foundation models, this objective increasingly involves adapting pretrained representations under limited computation and memory budgets Radford et al. (2021); Cheng et al. (2024). Low-Rank Adaptation (LoRA) Hu et al. (2022) offers an efficient approach by parameterizing adaptation as $\Delta W_t = B_t A_t$ while freezing the backbone within each task. However, parameter efficiency does not by itself resolve the stability–plasticity dilemma. Freezing backbone parameters does not prevent the effective weights from changing, and fixing the LoRA rank does not ensure that all available directions contribute meaningfully to adaptation.

Existing LoRA-based CL methods primarily address forgetting through subspace constraints. O-LoRA Wang et al. (2023) and InfLoRA Liang & Li (2024) restrict updates relative to historical task information. Pretrained knowledge protection has also received increasing attention. MiLoRA Wang et al. (2025a) initializes low-rank factors from minor pretrained singular components, whereas KeepLoRA Luo et al. (2026) projects gradients away from principal pretrained and historical feature subspaces in Figure 2. These approaches highlight the importance of choosing adaptation directions. Yet, selecting a protected space leaves two questions unresolved: *how restrictive should protection be, and how effectively is the permitted space utilized?*

Protection need not imply exclusion. Dominant pretrained directions can support both transferable representations and features that require adjustment for a new task. Unconstrained updates along these directions can disrupt pretrained responses, but excluding them entirely can also remove useful adaptation capacity. The relevant distinction is therefore between *controlling modification* and *prohibiting modification*. In continual learning, where successive updates may accumulate along

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.



(a) Strict (old) vs. Elastic (ours) Subspace Protection **(b)** Task 1 singular-value **(c)** Task 20 singular-value

Figure 1: (a) Strict protection excludes the pretrained principal subspace, whereas our SCULPT-LoRA permits controlled overlap to balance preservation and adaptation. (b,c) LoRA update ΔW_t spectra at Tasks 1 and 20. Without spectral balancing, energy concentrates in a few dominant modes; SCULPT-LoRA maintains a more balanced spectrum, mitigating spectral plasticity collapse and improving low-rank capacity utilization.

the same pretrained directions, this distinction becomes particularly important. A useful protection mechanism should limit interference while retaining flexibility to adjust shared features in Figure 1.

Available rank is not effective capacity. Even when an adaptation space is appropriately chosen, its nominal dimension does not determine how effectively it is used. The singular spectrum of ΔW_t may concentrate into a few dominant modes, leaving other directions with negligible contributions. We refer to this failure mode as *spectral plasticity collapse*: the update behaves as a much lower-rank transformation despite an unchanged rank budget. Subspace protection alone does not prevent this concentration in Figure 1 (b,c) and Figure 6. Conversely, a balanced update spectrum does not ensure that the update avoids sensitive pretrained directions. Spectral interaction with the backbone and spectral concentration within the update are therefore complementary aspects of continual adaptation.

This perspective motivates a joint design principle: *preserve flexibility in where adaptation is allowed, while controlling how strongly its available modes contribute*. It also distinguishes the spectrum of an optimization step from that of the accumulated adaptation. Balancing a matrix gradient need not balance the resulting LoRA factor, because successive steps can combine unevenly. To regulate the effective update directly, spectral control should be incorporated into the adaptation parameterization.

How can a fixed-rank update limit interference with pretrained knowledge while making effective use of its adaptation capacity?

We answer this question with Spectrally Controlled Updates for Low-Rank Parameter Transfer (SCULPT-LoRA), a framework that jointly controls where adaptation occurs and how its spectral capacity is utilized through *Elastic Principal Subspace Protection* and *Spectral Balancing via Polynomial Conditioning*. (i) The former attenuates components of a random initialization along the principal pretrained directions, then orthonormalizes the result to construct the LoRA input basis A_t. Its protection strength interpolates between unrestricted initialization and strict principal-subspace exclusion. Row orthonormality ensures that this mechanism controls the orientation of the adaptation space without distorting the nonzero singular spectrum of the trainable factor B_t. We freeze A_t within each task to preserve this geometry throughout optimization. (ii) The second mechanism applies a differentiable polynomial transformation to the trainable factor B_t. The transformation acts through a low-dimensional Gram matrix and relatively strengthens weak nonzero singular modes, counteracting concentration without decomposing the full LoRA update. Crucially, these mechanisms are mathematically connected. Since $A_t A_t^\top = I_r$, the effective update $\widehat{B}_t A_t$ inherits exactly the nonzero singular spectrum of $\widehat{B}_t$. Thus, the protected basis determines *where* adaptation occurs while allowing compact factor conditioning to regulate *how* update energy is distributed.

Our analysis formalizes both aspects of this construction. For a fixed seed satisfying the stated rank condition, we derive a sharp worst-case bound on perturbations along the original principal input subspace. At an equal update-energy budget, this bound decreases monotonically with protection strength and reaches zero under strict exclusion. Separately, we establish conditions under which polynomial shaping contracts spectral disparities and derive condition-number. Exact spectrum preservation transfers these guarantees from the conditioned factor to the effective update. Experiments on ImageNet-R, ImageNet-A, CIFAR-100, CUB, and DomainNet evaluate the framework across different task sequences.

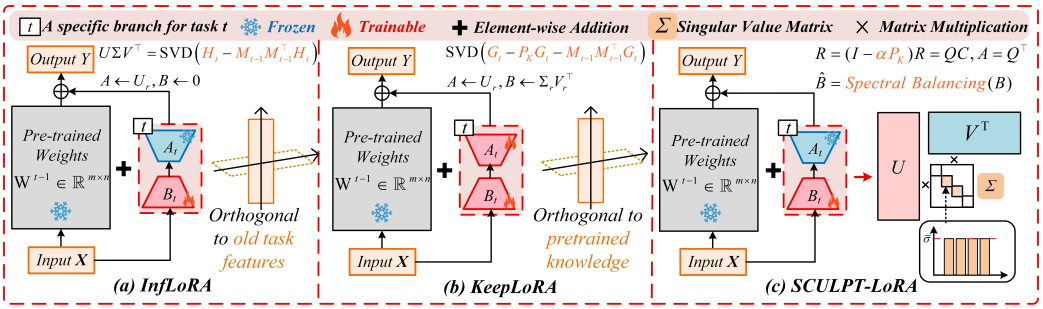


Figure 2: Comparison of different LoRA-based CL paradigms. (a) InfLoRA preserves previous knowledge by restricting updates to isolated directions, (b) KeepLoRA preserves pretrained knowledge by restricting updates to isolated directions, (c) SCULPT-LoRA preserves the principal spectral subspace of the pre-trained weights while balancing the singular-value spectrum of LoRA updates, thereby mitigating interference with pre-trained knowledge and enhancing learning plasticity. P_K denotes the pretrained principal-subspace projector, G_t accumulated gradient statistics, H_t current-task feature statistics, and M_{t-1} historical dominant directions, respectively.

Our main contributions can be summarized as follows:
• We identify two complementary spectral challenges in CL LoRA: interference with dominant pretrained directions and concentration of update energy in a few singular modes. This perspective motivates jointly controlling *where updates act* and *how their spectral capacity is utilized*.
• We propose SCULPT-LoRA, which couples Elastic Principal Subspace Protection (EPSP) with Polynomial Spectral Balancing (PSB). EPSP constructs a fixed row-orthonormal adaptation basis with controlled overlap with the pretrained principal subspace, while PSB reshapes the trainable factor through a compact Gram-matrix polynomial. Row orthonormality links the two mechanisms by transferring the conditioned factor's nonzero singular spectrum exactly to the effective LoRA update.
• We derive a sharp worst-case bound on principal-subspace perturbation that decreases with protection strength under a fixed update-energy budget, and establish conditions for condition-number improvement under polynomial spectral shaping. Experiments across five continual learning benchmarks and multiple incremental settings demonstrate strong performance, while ablations and spectral analyses support the complementary benefits of elastic protection and spectral balancing.

2 Related Works

PEFT-based Continual Learning. PEFT has emerged as a dominant paradigm for continual learning, where a frozen backbone is adapted via lightweight task-specific modules. Prompt-based methods Smith et al. (2023); Gao et al. (2024) maintain a prompt pool and dynamically select or update prompts for task encoding, while PGP Qiao et al. (2023) and VPT-NSP Lu et al. (2024) provide theoretical guarantees by enforcing gradient projection to mitigate forgetting. Adapter-based methods Li et al. (2025); Tan et al. (2024); He et al. (2025); Wang et al. (2026a) insert bottleneck modules into transformer layers to alleviate forgetting via capacity expansion or feature drift compensation, though they typically lack explicit modeling of task-wise representation geometry. Recently, LoRA-based methods have gained increasing attention due to their simplicity and transferability. SD-LoRA Wu et al. (2025b) decouples direction and magnitude learning, while O-LoRA Wang et al. (2023) and InfLoRA Liang & Li (2024) constrain updates to the estimated null space of previous tasks to preserve prior outputs. However, such designs neglect shared subspaces and may hinder transfer. LoDA He et al. (2026) introduces two LoRA branches, aiming to improve the stability–plasticity trade-off. SplitLoRA Qiu et al. (2025a) decomposes gradient spaces into two components to balance the stability–plasticity. BiLoRA Zhu et al. (2025) and PLAN Wang et al. (2025b) further impose orthogonality constraints on LoRA projections, improving stability but relying on rigid subspace partitioning. Beyond preserving task-specific knowledge, recent work also considers the preserving pretrained knowledge. MiLoRA Wang et al. (2025a), developed for LoRA, initializes low-rank factors from minor singular components while keeping the principal pretrained weight component frozen. KeepLoRA Luo et al. (2026) explicitly addresses CL by projecting gradients onto a subspace orthogonal to both the principal pretrained subspace and dominant historical feature directions.

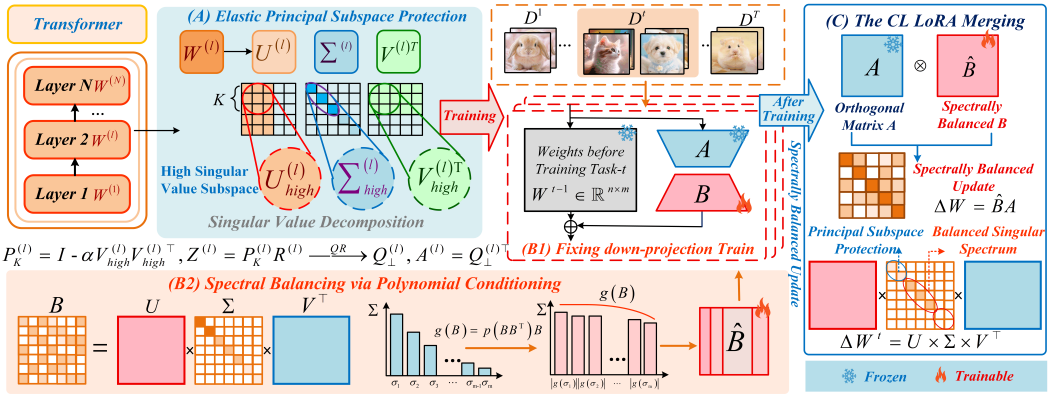


Figure 3: Overview of SCULPT-LoRA. At task t, we freeze the backbone weight W^{t-1} and attach a LoRA module. (A) Elastic principal subspace protection constructs a row-orthonormal down-projection A_t with controlled overlap with dominant pretrained directions. (B1-B2) We freeze A_t and train B_t on $\mathcal{D}^t$, applying polynomial conditioning during the forward pass to balance the effective update spectrum. (C) The conditioned update is merged into the backbone: $W^t = W^{t-1} + \widehat{B}_t A_t$.

3 METHODOLOGY

Problem Definition. In CL, there are T sequential tasks with non-overlapping classes. The datasets are defined as $\{\mathcal{D}^1, \cdots, \mathcal{D}^t, \cdots, \mathcal{D}^T\}$. $\mathcal{D}^t = \{x_i^t, y_i^t\}_{i=1}^{N^t}$ denotes the t-th task's dataset, where $x_i^t \in \mathbb{R}^{H\times W\times C}$ is the i-th input and $y_i^t \in \mathcal{C}^t$ is its class label. $\mathcal{C}^t$ denotes the label space of the t-th task, and $\mathcal{C}^t \cap \mathcal{C}^{t'} = \emptyset$ for $t \neq t'$. The objective of CL is to train a model $f(w, \cdot)$ sequentially on T tasks and performs well on all seen classes $\mathcal{C}^1 \cup \cdots \cup \mathcal{C}^T$. We consider the task-agnostic class-incremental learning setting, where task identity is unavailable during inference.

Overall framework. Figure 3 illustrates SCULPT-LoRA. Following prior PEFT-based continual learning methods Wang et al. (2022b); Tan et al. (2024), we adopt a pretrained Vision Transformer (ViT) Dosovitskiy et al. (2021) as the backbone. At task t, each adapted layer freezes the accumulated weight W^{t-1} and learns a low-rank update $\Delta W_t = B_t A_t \in \mathbb{R}^{n\times m}$. SCULPT-LoRA controls two complementary properties of this update: *where* adaptation occurs and *how* its limited rank capacity is utilized. Specifically, Elastic Principal Subspace Protection (EPSP) constructs a fixed, row-orthonormal basis A_t with controlled overlap with the dominant pretrained subspace, limiting interference while retaining adaptation flexibility. With A_t frozen, Polynomial Spectral Balancing (PSB) conditions the trainable factor B_t during the forward pass to counter spectral concentration and promote effective utilization of its nonzero singular modes. Because $A_t A_t^\top = I_r$, this factor-level spectral conditioning transfers exactly to the effective update $\Delta\widehat{W}_t = \widehat{B}_t A_t$. After task t, the conditioned update is merged into the backbone as $W^t = W^{t-1} + \Delta\widehat{W}_t$.

LoRA Analysis with Fixed Down-Projections. Given the backbone weight $W^{t-1} \in \mathbb{R}^{n\times m}$ before the t-th task, where m and n denote the input and output dimensions, respectively, LoRA parameterizes the low-rank update as $B_t A_t$, with $A_t \in \mathbb{R}^{r\times m}$ and $B_t \in \mathbb{R}^{n\times r}$, while keeping W^{t-1} fixed. For a batch of b input features $X \in \mathbb{R}^{b\times m}$, the layer output is $Y = X\left(W^{t-1} + B_t A_t\right)^\top \in \mathbb{R}^{b\times n}$.

3.1 ELASTIC PRINCIPAL SUBSPACE PROTECTION INITIALIZATION

Freezing W_0 does not preserve pretrained knowledge by itself, since the effective weight $W = W_0 + BA$ can still deviate along dominant pretrained spectral directions. MiLoRA Wang et al. (2025a) initializes adaptation from minor singular components, while KeepLoRA Luo et al. (2026) projects gradients away from principal pretrained and historical feature subspaces. However, principal directions may support both pretrained knowledge and new-task learning, making their complete exclusion unnecessarily restrictive. We therefore propose *Elastic Principal Subspace Protection*, which attenuates adaptation along principal directions without eliminating access to them. This

replaces rigid exclusion with controlled modification, balancing pretrained knowledge retention against the flexibility required for new tasks in Figure 4.

Elastic Principal Subspace Protection (EPSP). The principal directions of pretrained weights encode pretrained knowledge and support new-task learning, motivating elastic protection rather than complete exclusion. Given $W_0 = U\Sigma V^\top \in \mathbb{R}^{n\times m}$, let $V_K \in \mathbb{R}^{m\times K}$ contain its top-K right singular vectors and define $P_K = V_K V_K^\top$, requiring $r \le m - K$. We shape a Gaussian seed $R \in \mathbb{R}^{m\times r}$ before orthonormalization:

$$\widetilde{R} = (I - \alpha P_K)R = QC,\ A = Q^\top,\ \alpha \in [0,1],$$

where QC is a thin QR decomposition. The operator preserves complementary components while attenuating principal components by $1-\alpha$, interpolating between unrestricted initialization and strict exclusion. Crucially, orthonormalization separates *subspace orientation* from *factor scale*: protection is encoded in the geometry of the adaptation space rather than a uniform reduction in update magnitude. Assuming $\widetilde{R}$ has full column rank, we obtain

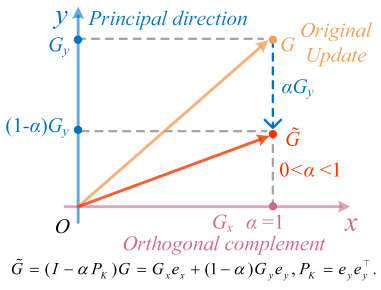


Figure 4: Elastic principal-subspace protection. Increasing α suppresses updates along principal directions while preserving complement.

$$AA^\top = I_r,\ AV_K = (1-\alpha)C^{-\top}R^\top V_K. \tag{1}$$

For $\alpha < 1$, principal directions remain accessible; at $\alpha = 1$, they are excluded exactly, requiring $r \le m - K$. We freeze A during task adaptation and train only B. Since $\|BA\|_F = \|B\|_F$, every nonzero update satisfies $\frac{\|\Delta W V_K\|_F}{\|\Delta W\|_F} \le \|AV_K\|_2, \Delta W = BA$.

Theorem 3.1 (Bound for Elastic Principal Subspace Protection). *Let $W_0 \in \mathbb{R}^{n\times m}$, let $V_K \in \mathbb{R}^{m\times K}$ contain its top-K right singular vectors, and define $P_K = V_K V_K^\top \in \mathbb{R}^{m\times m}$. Fix $R \in \mathbb{R}^{m\times r}$ such that $H = R^\top(I_m - P_K)R \succ 0$, and define $G = R^\top P_K R$ and $\lambda_\star = \lambda_{\max}(H^{-1/2}GH^{-1/2})$. For $\alpha \in [0,1]$, let $(I_m - \alpha P_K)R = Q_\alpha C_\alpha$ be a thin QR decomposition with $Q_\alpha^\top Q_\alpha = I_r$, and set $A_\alpha = Q_\alpha^\top$ and $c_\alpha = \sqrt{(1-\alpha)^2\lambda_\star/[1+(1-\alpha)^2\lambda_\star]}$. Then, for every $B \in \mathbb{R}^{n\times r}$,*

$$\|(W_0 + BA_\alpha)P_K - W_0 P_K\|_F \le c_\alpha \|B\|_F. \tag{2}$$

Interpretation. Theorem 3.1 gives a direct geometric interpretation of the protection strength α. Since $A_\alpha A_\alpha^\top = I_r$, we have $\|BA_\alpha\|_F = \|B\|_F$, and hence, for any nonzero B, $\frac{\|BA_\alpha P_K\|_F}{\|BA_\alpha\|_F} \le c_\alpha$. Thus, c_α bounds the fraction of a fixed update-energy budget that can perturb the pretrained principal input subspace. Importantly, this reduction is achieved by reshaping the orientation of the adaptation space rather than shrinking the overall update magnitude. Since c_α decreases monotonically with α and vanishes at $\alpha = 1$, EPSP provides a continuous interpolation from unrestricted adaptation to strict principal-subspace exclusion. The theorem therefore formalizes the stability side of EPSP: increasing α provably limits interference with dominant pretrained responses. At the same time, $\alpha < 1$ retains access to these directions, which may remain useful for new-task adaptation. This explains why an intermediate protection strength can provide a better stability–plasticity trade-off than either unrestricted adaptation or strict exclusion.

3.2 Spectrally Balanced LoRA Training

Motivation. The subspace construction in the previous section determines *where* LoRA adaptation takes place by fixing the down-projection basis A_t. However, an informative adaptation subspace alone does not guarantee that its low-rank capacity can be effectively utilized during training. Given the LoRA update $\Delta W_t = B_t A_t$, conventional gradient-based optimization may progressively concentrate the update magnitude of B_t along a few dominant singular directions. Consequently, although the nominal LoRA rank is r, only a small fraction of these directions may contribute substantially to the resulting update, leading to an increasingly imbalanced singular spectrum and reduced effective adaptation capacity. Such spectral concentration is particularly undesirable

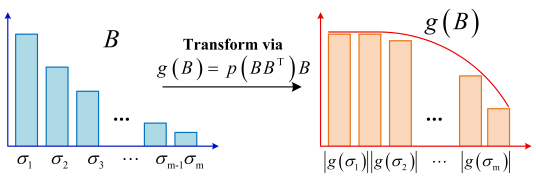
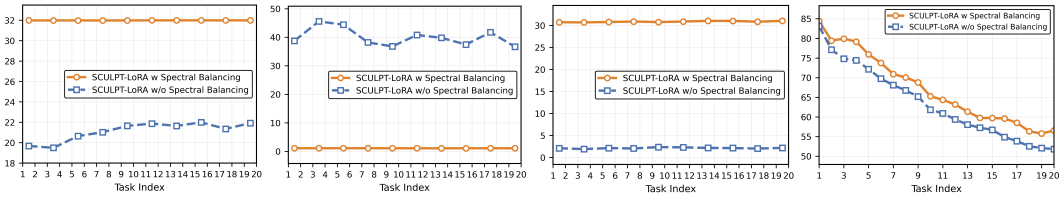


Figure 5: Polynomial spectral balancing.



(a) Effective rank **(b)** Condition number κ_+ **(c)** Stable rank **(d)** Overall Test Acc

Figure 6: **Spectral balancing prevents rank collapse.** We track effective rank, κ_+, stable rank of ΔW_t, and overall test accuracy, comparing SCULPT-LoRA with and without spectral balancing.

in continual learning, where the limited low-rank capacity must be repeatedly reused for sequential tasks. To alleviate this issue, we introduce a *spectrally balanced LoRA training* strategy. Rather than jointly optimizing both LoRA factors, we keep the constructed basis A_t fixed and optimize only the up-projection B_t. During the forward pass, we apply a spectral balancing transformation to B_t, encouraging different low-rank directions to contribute more uniformly to the model update.

Polynomial Spectral Balancing (PSB). To regulate the spectrum of the effective adaptation, we apply a polynomial transformation to the trainable factor $B_t \in \mathbb{R}^{n\times r}$. For $\epsilon > 0$, define

$$s_t = \max\{\|B_t\|_2, \epsilon\},\ X_t = B_t/s_t,\ G_t = X_t^\top X_t. \tag{3}$$

The effective factor is

$$\widehat{B}_t = \gamma_t\, \mathrm{sg}(s_t) X_t p_4(G_t), \tag{4}$$

where sg is the identity in the forward pass and stops gradients in the backward pass in 5. The learnable scalar γ_t is initialized to one, and

$$p_4(z) = 3.625 - 9.261z + 14.097z^2 - 10.351z^3 + 2.890z^4. \tag{5}$$

For matrix arguments, the constant term is interpreted as $3.625I_r$. The polynomial is evaluated through the $r \times r$ Gram matrix, without constructing or decomposing the full update. The transformation is differentiable almost everywhere; spectral-norm normalization can be nondifferentiable at repeated leading singular values and at the scale-floor boundary. Let $f(x) = xp_4(x^2)$. If $X_t = U\,\mathrm{diag}(x_i)V^\top$ is a compact SVD, then

$$\widehat{B}_t = \gamma_t s_t U\,\mathrm{diag}(f(x_i))V^\top. \tag{6}$$

Thus, PSB acts on spectral magnitudes within the same singular subspaces, although their ordering may change. The response satisfies $f(1) = 1$ and $f(x) = 3.625x + O(x^3)$ as $x \to 0$. These identities describe the response of individual modes; they do not imply exact spectral equalization.

Polynomial construction and provenance. We adopt a fixed odd spectral map $f(\sigma) = \sigma p_k(\sigma^2)$, where $\deg(p_k) \le k$. The underlying design approximates the soft spectral-balancing target $h_b(\sigma) = \min\left\{\frac{\sigma}{b}, 1\right\}, \sigma \ge 0$, which amplifies small singular values while saturating larger ones at the target level. Specifically, Wang et al. (2026b) fit this target on $[0, 1.1]$ using the finite-sample weighted least-squares problem $\boldsymbol{a}^\star \in \arg\min_{\boldsymbol{a}\in\mathbb{R}^{k+1}} \sum_{j=1}^{M} \sigma_j^\omega \left(\sum_{\ell=0}^{k} a_\ell \sigma_j^{2\ell+1} - h_b(\sigma_j)\right)^2$, where $\sigma_1, \ldots, \sigma_{M-1}$ are sampled uniformly from $[0, 1.1]$ and $\sigma_M = 1$ is included to encourage agreement at the nominal unit scale. Here, ω denotes the fitting-weight exponent.

We use the published coefficient set associated with $k = 4$ and $b = 0.3$: $p_4(z) = 3.625 - 9.261z + 14.097z^2 - 10.351z^3 + 2.890z^4$. Thus, the polynomial is fully specified by the displayed coefficients and introduces no fitting hyperparameters during continual learning. Under our normalization $s_t = \max\{\|B_t\|_2, \epsilon\}$, the singular values of $X_t = B_t/s_t$ lie in $[0, 1]$. We apply the polynomial through the compact Gram matrix $X_t^\top X_t$, and keep its coefficients fixed across tasks and layers. The resulting response satisfies $f(1) = 1$ and $f(\sigma) = 3.625\sigma + O(\sigma^3)$ near zero, providing soft spectral shaping rather than exact singular-value equalization. Our spectral guarantees are established directly for the displayed coefficients, rather than inferred from the least-squares fitting objective.

Notation. For a nonzero matrix $B \in \mathbb{R}^{n\times r}$ of rank q, write $\sigma_1(B) \ge \cdots \ge \sigma_q(B) > 0$ and define $\kappa_+(B) = \frac{\sigma_1(B)}{\sigma_q(B)}, \mathrm{srank}(B) = \frac{\|B\|_F^2}{\|B\|_2^2}$. We additionally report the entropy-based effective rank $\mathrm{erank}(B) = \exp(-\sum_{i=1}^{q} \pi_i \log \pi_i)\,, \pi_i = \frac{\sigma_i(B)}{\sum_{j=1}^{q}\sigma_j(B)}$. All three quantities are invariant under nonzero scalar multiplication in Figure 6. The following results concern exact forward evaluations.

Table 1: Experimental results for 10 incremental tasks on ImageNetR, ImageNetA, CIFAR100, and CUB. We report the averaged results over 3 random seeds. The highest results are in **bold**.

Method	Venue	10S-ImageNetR		10S-ImageNetA		10S-CIFAR100		10S-CUB	
		$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$
L2P	CVPR'22	$72.34_{\pm 0.17}$	$77.36_{\pm 0.64}$	$44.04_{\pm 0.93}$	$51.24_{\pm 2.26}$	$84.06_{\pm 0.88}$	$88.26_{\pm 1.34}$	$67.02_{\pm 1.90}$	$79.62_{\pm 1.60}$
DualPrompt	ECCV'22	$69.10_{\pm 0.62}$	$74.28_{\pm 0.66}$	$53.19_{\pm 0.74}$	$64.59_{\pm 0.08}$	$86.93_{\pm 0.24}$	$91.13_{\pm 0.32}$	$68.48_{\pm 0.47}$	$80.59_{\pm 1.50}$
CODAPrompt	CVPR'23	$73.31_{\pm 0.50}$	$78.47_{\pm 0.53}$	$52.08_{\pm 0.12}$	$63.92_{\pm 0.12}$	$83.21_{\pm 3.39}$	$87.71_{\pm 3.17}$	$77.23_{\pm 1.12}$	$81.90_{\pm 0.85}$
InfLoRA	CVPR'24	$74.75_{\pm 0.64}$	$80.67_{\pm 0.55}$	$49.20_{\pm 1.12}$	$60.92_{\pm 0.61}$	$86.75_{\pm 0.35}$	$91.72_{\pm 0.15}$	$70.82_{\pm 0.23}$	$81.39_{\pm 0.14}$
SD-LoRA	ICLR'25	$77.34_{\pm 0.35}$	$82.04_{\pm 0.24}$	$55.96_{\pm 0.73}$	$64.95_{\pm 1.63}$	$88.01_{\pm 0.31}$	$92.54_{\pm 0.18}$	$77.48_{\pm 0.20}$	$85.59_{\pm 0.44}$
Bi-LoRA	CVPR'25	$77.95_{\pm 0.14}$	$81.52_{\pm 0.26}$	–	–	$87.46_{\pm 0.76}$	$92.50_{\pm 0.62}$	–	–
PLAN	ICCV'25	$75.25_{\pm 0.42}$	$80.41_{\pm 0.56}$	–	–	$87.54_{\pm 0.31}$	$92.21_{\pm 0.35}$	–	–
LoRA-P&M	NIPS'25	$79.95_{\pm 0.18}$	$85.29_{\pm 0.93}$	$56.57_{\pm 0.78}$	$65.35_{\pm 1.81}$	$88.45_{\pm 0.35}$	$92.89_{\pm 1.13}$	$78.29_{\pm 0.50}$	$83.39_{\pm 0.61}$
CoSO	NIPS'25	$81.10_{\pm 0.39}$	$85.56_{\pm 0.13}$	–	–	$88.77_{\pm 0.16}$	$92.99_{\pm 0.23}$	–	–
MiLoRA	NAACL'25	$77.76_{\pm 0.17}$	$83.84_{\pm 0.43}$	$60.06_{\pm 0.72}$	$67.97_{\pm 1.41}$	$88.47_{\pm 0.06}$	$90.67_{\pm 1.42}$	$78.56_{\pm 0.57}$	$85.47_{\pm 0.78}$
KeepLoRA	ICLR'26	$78.93_{\pm 0.20}$	$84.90_{\pm 0.40}$	$60.59_{\pm 0.64}$	$68.87_{\pm 1.61}$	$87.47_{\pm 0.06}$	$90.46_{\pm 1.42}$	$78.74_{\pm 0.78}$	$86.35_{\pm 0.98}$
SplitLoRA	ICLR'26	$81.00_{\pm 0.17}$	$85.84_{\pm 0.62}$	–	–	$90.33_{\pm 0.73}$	$93.70_{\pm 0.32}$	–	–
SCULPT-LoRA	–	$\mathbf{81.77}_{\pm 0.25}$	$\mathbf{87.28}_{\pm 0.28}$	$\mathbf{62.23}_{\pm 0.52}$	$\mathbf{71.27}_{\pm 1.02}$	$\mathbf{90.40}_{\pm 0.06}$	$\mathbf{93.81}_{\pm 1.42}$	$\mathbf{81.47}_{\pm 0.55}$	$\mathbf{89.47}_{\pm 0.42}$

Table 2: Experimental results for 5 and 20 incremental sessions on ImageNetR, 20 sessions on ImageNetA, and 10 sessions on DomainNet.

Method	Venue	5S-ImageNetR		20S-ImageNetR		20S-ImageNetA		10S-DomainNet	
		$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$
L2P	CVPR'22	$70.83_{\pm 0.58}$	$78.34_{\pm 0.47}$	$69.64_{\pm 0.42}$	$75.28_{\pm 0.57}$	$40.48_{\pm 1.78}$	$49.62_{\pm 1.46}$	$81.17_{\pm 0.83}$	$87.43_{\pm 0.95}$
DualPrompt	ECCV'22	$73.05_{\pm 0.50}$	$79.47_{\pm 0.40}$	$66.61_{\pm 0.58}$	$72.45_{\pm 0.37}$	$42.28_{\pm 1.94}$	$53.39_{\pm 1.64}$	$81.70_{\pm 0.78}$	$87.80_{\pm 0.99}$
CODAPrompt	CVPR'23	$74.91_{\pm 0.33}$	$79.25_{\pm 0.53}$	$69.96_{\pm 0.50}$	$75.34_{\pm 0.85}$	$44.62_{\pm 1.92}$	$54.86_{\pm 0.50}$	$80.04_{\pm 0.79}$	$86.27_{\pm 0.82}$
InfLoRA	CVPR'24	$77.52_{\pm 0.37}$	$82.01_{\pm 0.12}$	$71.01_{\pm 0.45}$	$77.28_{\pm 0.45}$	$46.81_{\pm 2.30}$	$57.13_{\pm 0.95}$	$81.45_{\pm 0.68}$	$87.55_{\pm 0.57}$
SD-LoRA	ICLR'25	$79.15_{\pm 0.20}$	$83.01_{\pm 0.42}$	$75.26_{\pm 0.37}$	$80.22_{\pm 0.72}$	$49.12_{\pm 1.78}$	$59.63_{\pm 1.57}$	$81.01_{\pm 0.42}$	$86.85_{\pm 0.40}$
Bi-LoRA	CVPR'25	$78.32_{\pm 0.37}$	$83.31_{\pm 0.52}$	$72.41_{\pm 0.65}$	$79.28_{\pm 0.76}$	–	–	$76.56_{\pm 0.41}$	$83.20_{\pm 0.35}$
LoRA-P&M	NIPS'25	$81.47_{\pm 0.56}$	$85.96_{\pm 0.52}$	$76.37_{\pm 0.09}$	$82.77_{\pm 0.31}$	$52.27_{\pm 1.84}$	$62.15_{\pm 1.33}$	$84.35_{\pm 0.48}$	$89.43_{\pm 0.27}$
CoSO	NIPS'25	$82.10_{\pm 0.13}$	$86.38_{\pm 0.07}$	$78.19_{\pm 0.28}$	$83.69_{\pm 0.12}$	–	–	–	–
MiLoRA	NAACL'25	$81.45_{\pm 0.23}$	$85.78_{\pm 0.41}$	$76.86_{\pm 0.23}$	$82.94_{\pm 0.17}$	$50.87_{\pm 1.39}$	$61.42_{\pm 1.47}$	$81.56_{\pm 0.21}$	$87.18_{\pm 0.46}$
KeepLoRA	ICLR'26	$81.37_{\pm 0.14}$	$85.46_{\pm 0.41}$	$77.09_{\pm 0.31}$	$82.98_{\pm 0.18}$	$51.74_{\pm 1.89}$	$62.54_{\pm 2.43}$	$82.49_{\pm 0.11}$	$88.15_{\pm 0.42}$
SplitLoRA	ICLR'26	$81.92_{\pm 0.29}$	$85.85_{\pm 0.55}$	$78.82_{\pm 0.28}$	$84.57_{\pm 0.44}$	–	–	$84.31_{\pm 0.23}$	$88.99_{\pm 0.57}$
SCULPT-LoRA	–	$\mathbf{83.04}_{\pm 0.23}$	$\mathbf{87.56}_{\pm 0.32}$	$\mathbf{78.93}_{\pm 0.18}$	$\mathbf{85.02}_{\pm 0.21}$	$\mathbf{55.02}_{\pm 1.71}$	$\mathbf{65.61}_{\pm 0.64}$	$\mathbf{85.61}_{\pm 0.10}$	$\mathbf{90.21}_{\pm 0.32}$

Lemma 3.2 (Exact transfer of the nonzero spectrum). *Let $B \in \mathbb{R}^{n\times r}$ have rank $q > 0$, let $s > 0$, and let $\gamma \neq 0$. Set $X = B/s, \widehat{B} = \gamma s X p(X^\top X), f(x) = xp(x^2)$, where p is a real polynomial. Let $x_i = \sigma_i(B)/s$. The potentially nonzero singular values of $\widehat{B}$ are $|\gamma|s|f(x_i)|, i = 1, \ldots, q$, up to ordering. In particular, if $f(x_i) \neq 0$ for every i, then $\operatorname{rank}(\widehat{B}) = q$.*

If $A \in \mathbb{R}^{r\times m}$ satisfies $AA^\top = I_r$, then $\widehat{B}A$ and $\widehat{B}$ have identical nonzero singular values, including multiplicities. Hence, whenever $\widehat{B} \neq 0$, they have the same κ_+, stable rank, and effective rank.

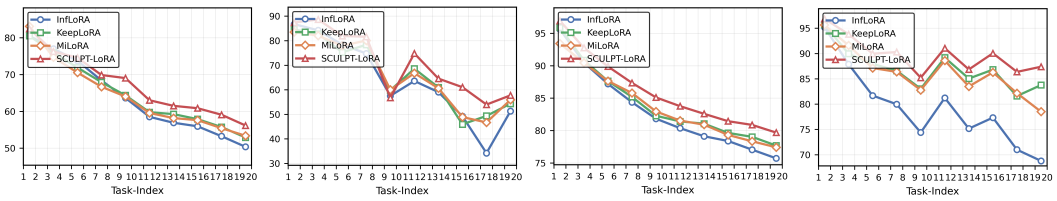
Theorem 3.3 (Condition-number control by log-spectral contraction). *Under the notation of Lemma 3.2, suppose the nonzero singular values of X lie in $[a, b] \subset (0, \infty)$. Assume that f is positive and continuously differentiable on an open interval containing $[a, b]$, and that, for some $\rho \in [0, 1]$, $\left|\frac{xf'(x)}{f(x)}\right| \leq \rho, x \in [a, b]$. Then $\operatorname{rank}(\widehat{B}) = q$ and $\kappa_+(\widehat{B}) \leq \kappa_+(B)^\rho$. If $\rho < 1$ and $\kappa_+(B) > 1$, the condition number strictly decreases. The same conclusions hold for $\widehat{B}A$ relative to BA whenever $AA^\top = I_r$.*

Corollary 3.4 (Condition-number improvement for the PSB polynomial). *Let $p = p_4$ be the polynomial in Eq. equation 5, with the displayed coefficients treated as exact. For any nonzero B, $\epsilon > 0$, $s = \max\{\|B\|_2, \epsilon\}$, and $\gamma \neq 0$, the effective factor preserves rank and satisfies $\kappa_+(\widehat{B}) \leq \kappa_+(B)$. The inequality is strict whenever $\kappa_+(B) > 1$. If $AA^\top = I_r$, the same rank preservation and condition-number improvement hold for the effective update $\widehat{B}A$ relative to BA.*

Together, Lemma 3.2, Theorem 3.3, and Corollary 3.4 establish that PSB preserves rank and strictly improves the nonzero condition number of the effective LoRA update whenever it exceeds one, complementing EPSP's control of the adaptation subspace.

4 Experiments

Datasets and Evaluation Metrics. We conduct experiments on five widely used benchmarks: ImageNet-R Hendrycks et al. (2021a), ImageNet-A Hendrycks et al. (2021b), CIFAR-100 Krizhevsky



(a) 20S-ImageNetA (all) **(b)** 20S-ImageNetA (new) **(c)** 20S-ImageNetR (all) **(d)** 20S-ImageNetR (new)

Figure 7: Stability–plasticity dynamics across incremental tasks. Overall and new-task accuracy on 20S-ImageNetA and 20S-ImageNetR. SCULPT-LoRA better maintains overall performance while preserving strong adaptation to newly introduced classes.

Table 3: Ablation studies on ImageNetR and ImageNetA under 10 and 20 incremental sessions, evaluating the individual and combined contributions of Elastic Principal Subspace Protection (EPSP) and Polynomial Spectral Balancing (PSB). Idx 1 is Random Orthogonal Bases.

Idx	EPSP	PSB	10S-ImageNetR		10S-ImageNetA		20S-ImageNetR		20S-ImageNetA	
			$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$	$A_{Last}\uparrow$	$A_{Avg}\uparrow$
1	–	–	$76.23_{\pm 0.18}$	$82.05_{\pm 0.62}$	$57.48_{\pm 1.47}$	$67.34_{\pm 1.86}$	$76.21_{\pm 0.18}$	$80.26_{\pm 0.51}$	$48.62_{\pm 1.28}$	$60.52_{\pm 1.26}$
2	✓	–	$80.23_{\pm 0.28}$	$85.64_{\pm 0.36}$	$61.27_{\pm 0.56}$	$69.81_{\pm 1.08}$	$77.56_{\pm 0.20}$	$83.67_{\pm 0.32}$	$51.92_{\pm 0.66}$	$62.67_{\pm 0.51}$
3	–	✓	$80.74_{\pm 0.28}$	$86.31_{\pm 0.35}$	$61.86_{\pm 0.55}$	$70.26_{\pm 0.98}$	$78.50_{\pm 0.26}$	$84.25_{\pm 0.36}$	$54.32_{\pm 0.64}$	$64.53_{\pm 0.51}$
4	✓	✓	$\mathbf{81.77}_{\pm 0.25}$	$\mathbf{87.28}_{\pm 0.28}$	$\mathbf{62.23}_{\pm 0.52}$	$\mathbf{71.27}_{\pm 1.02}$	$\mathbf{78.93}_{\pm 0.18}$	$\mathbf{85.02}_{\pm 0.21}$	$\mathbf{55.02}_{\pm 1.71}$	$\mathbf{65.61}_{\pm 0.64}$

et al. (2009), CUB Wah et al. (2011), and DomainNet Peng et al. (2019). ImageNet-R comprises 30K images from 200 categories with diverse visual renditions, whereas ImageNet-A contains 7.5K naturally occurring adversarial images covering 200 categories. CIFAR-100 consists of 60K 32×32 images distributed over 100 classes. CUB is a fine-grained recognition benchmark of bird species. DomainNet contains 345 object categories collected from six distinct domains; following Gao et al. (2024), we retain the 200 classes with the largest number of samples. We adopt two standard CL metrics following prior PEFT-based studies Tan et al. (2024); Wu et al. (2025a). Last Accuracy (A_{Last}) evaluates the model on all observed classes after completing the entire task sequence, while Average Accuracy (A_{Avg}) averages the evaluation accuracy obtained after each incremental session.

Implementation Details. We adopt ImageNet-21K pre-trained ViT-B/16 Dosovitskiy et al. (2021); Russakovsky et al. (2015) as the backbone throughout all experiments. Following Wu et al. (2025a), SCULPT-LoRA is applied to every self-attention block with a fixed rank of $r = 32$, $K = 64$, $\alpha = 0.5$, while a cosine classifier is used for prediction. For each incremental task, the model is trained for 20 epochs with a batch size of 48 using SGD and a cosine annealing learning-rate schedule. We set the initial learning rate to 0.02 for ImageNet-R and 0.005 for CUB, while using 0.01 for the remaining datasets. All experiments are performed on a single NVIDIA RTX 4090 GPU. All reported results are mean $\pm$ standard deviation over three random seeds 1993, 1996, 1997.

4.1 Comparison with Baseline Methods

We compare SCULPT-LoRA with representative PEFT-based continual learning methods, including prompt-based approaches Wang et al. (2022b;a); Jiao et al. (2024), adapter-based approaches Tan et al. (2024); Gao et al. (2023), and recent LoRA-based methods Liang & Li (2024); Wu et al. (2025b); Wang et al. (2025b); Cheng et al. (2025); Qiu et al. (2025b). Results on five benchmarks with diverse incremental settings are reported in Table 1, Table 2, Figure 7. Overall, SCULPT-LoRA achieves state-of-the-art or competitive performance across the evaluated benchmarks, demonstrating the complementary benefits of controlling *where adaptation occurs* and *how the available low-rank capacity is utilized*. Without feature replay or additional memory buffers, SCULPT-LoRA delivers strong performance on both A_{Last} and A_{Avg}. On challenging ImageNet-scale benchmarks, SCULPT-LoRA obtains clear gains by limiting interference with dominant pretrained directions while retaining sufficient adaptation flexibility. It achieves 81.77% and 62.23% A_{Last} on 10S-ImageNetR and 10S-ImageNetA, respectively. Similar improvements are observed in the 5S- and 20S-ImageNetR settings, indicating that the proposed spectral control remains effective under longer continual-learning horizons. The benefits are further evident on the highly shifted 20S-ImageNetA benchmark, where SCULPT-LoRA achieves 55.02% A_{Last} and 65.61% A_{Avg}. These results suggest

Table 4: Sensitivity to protection strength α and protected-subspace dimension K. We report A_{Last}.

	α ($K = 64$)							K ($\alpha = 0.5$)						
Dataset	0.00	0.2	0.40	0.50	0.60	0.80	1.00	0	32	64	128	256	384	512
20S-ImageNetR	78.50	78.62	78.75	**78.93**	78.61	78.12	77.65	78.50	78.65	**78.93**	78.22	77.26	75.03	73.65
20S-ImageNetA	54.32	54.61	54.92	**55.02**	54.62	54.42	54.10	54.32	54.71	**55.02**	54.86	54.62	54.14	53.82

Table 5: Comparison of different principal space protection methods

Method	10S-ImageNetR		10S-ImageNetA		20S-ImageNetR		20S-ImageNetA	
	$A_{Last} \uparrow$	$A_{Avg} \uparrow$	$A_{Last} \uparrow$	$A_{Avg} \uparrow$	$A_{Last} \uparrow$	$A_{Avg} \uparrow$	$A_{Last} \uparrow$	$A_{Avg} \uparrow$
KeepLoRA Luo et al. (2026)	$78.93_{\pm 0.20}$	$84.90_{\pm 0.40}$	$60.59_{\pm 0.64}$	$68.87_{\pm 1.61}$	$77.09_{\pm 0.31}$	$82.98_{\pm 0.18}$	$51.74_{\pm 1.89}$	$62.54_{\pm 2.43}$
MiLoRA Wang et al. (2025a)	$77.76_{\pm 0.17}$	$83.84_{\pm 0.43}$	$60.06_{\pm 0.72}$	$67.97_{\pm 1.41}$	$76.86_{\pm 0.23}$	$82.94_{\pm 0.17}$	$50.87_{\pm 1.39}$	$61.42_{\pm 1.47}$
SCULPT-LoRA (EPSP)	$\mathbf{80.23}_{\pm 0.28}$	$\mathbf{85.64}_{\pm 0.36}$	$\mathbf{61.27}_{\pm 0.56}$	$\mathbf{69.81}_{\pm 1.08}$	$\mathbf{77.56}_{\pm 0.20}$	$\mathbf{83.67}_{\pm 0.32}$	$\mathbf{51.92}_{\pm 0.66}$	$\mathbf{62.67}_{\pm 0.51}$

that elastic principal-subspace protection preserves adaptation flexibility under substantial distribution shift, while polynomial spectral balancing promotes more effective utilization of the available low-rank update space. On CIFAR100, CUB, and DomainNet, SCULPT-LoRA also achieves strong performance, further demonstrating the generality of the proposed spectral adaptation framework across diverse continual learning scenarios.

4.2 Ablation Studies

Both components improve continual adaptation. Table 3 ablates EPSP and PSB on ImageNetR and ImageNetA under 10 and 20 incremental sessions. Relative to the random orthogonal basis baseline without either component, EPSP and PSB individually improve A_{Last} by 1.35–4.00 and 2.29–5.70 percentage points, respectively, while consistently improving A_{Avg}. Combining both yields the highest mean accuracy across all settings: adding EPSP to PSB further improves A_{Last} by 0.37–1.03 points and A_{Avg} by 0.77–1.08 points, resulting in total A_{Last} gains of 2.72–6.40 points over the baseline. These results support the complementary benefits of controlling *where updates act* and *how their spectral capacity is utilized*.

Sensitivity to α and K. Table 4 examines the protection strength α and protected-subspace dimension K with PSB enabled. Both datasets achieve the highest accuracy among the tested settings at $\alpha = 0.5$ and $K = 64$. With $K = 64$, moderate protection outperforms both unrestricted initialization ($\alpha = 0$) and strict exclusion ($\alpha = 1$), improving A_{Last} over strict exclusion by 1.28 and 0.92 percentage points on 20S-ImageNetR and 20S-ImageNetA, respectively. With $\alpha = 0.5$, accuracy improves as K increases from 0 to 64, but declines for larger protected subspaces, particularly on 20S-ImageNetR. These results support selectively attenuating a compact set of principal directions, while suggesting that overly strong or overly broad protection can restrict adaptation.

Comparison with other principal space protection methods. SCULPT-LoRA (EPSP) achieves higher mean final and average accuracy than KeepLoRA and MiLoRA across all evaluated settings (Table 5), although some gains are modest. These results suggest that elastic control over principal pretrained directions offers a useful balance between representation preservation and adaptation flexibility.

5 Conclusion

In this work, we revisit LoRA-based continual learning from a spectral perspective and propose SCULPT-LoRA to jointly control *where* adaptation occurs and *how* its limited rank capacity is utilized. Elastic Principal Subspace Protection (EPSP) provides controlled access to dominant pretrained directions, while Polynomial Spectral Balancing (PSB) counteracts spectral concentration within the trainable update. A row-orthonormal LoRA basis connects these two mechanisms by transferring the conditioned spectrum exactly to the effective update. Together, SCULPT-LoRA provides a principled approach to balancing pretrained knowledge preservation and continual adaptation, with consistent improvements across diverse continual learning benchmarks.

REFERENCES

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23595–23604, 2024.

Quan Cheng, Yuanyu Wan, Lingyu Wu, Chenping Hou, and Lijun Zhang. Continuous subspace optimization for continual learning. In *Conference on Neural Information Processing Systems*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11483–11493, 2023.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28463–28473, 2024.

Lingfeng He, De Cheng, Zhiheng Ma, Huaijie Wang, Dingwen Zhang, Nannan Wang, and Xinbo Gao. Ckaa: Cross-subspace knowledge alignment and aggregation for robust continual learning. *arXiv preprint arXiv:2507.09471*, 2025.

Lingfeng He, De Cheng, Huaijie Wang, Xi Yang, Nannan Wang, and Xinbo Gao. Task-driven subspace decomposition for knowledge sharing and isolation in lora-based continual learning. *arXiv preprint arXiv:2603.00191*, 2026.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Jiao, Qiuxia Lai, Yu Li, and Qiang Xu. Vector quantization prompting for continual learning. *NeurIPS*, 2024.

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

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

Jiashuo Li, Shaokun Wang, Bo Qian, Yuhang He, Xing Wei, Qiang Wang, and Yihong Gong. Dynamic integration of task-specific adapters for class incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30545–30555, 2025.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Nannan Wang, Peng Wang, and Yanning Zhang. Visual Prompt Tuning in Null Space for Continual Learning. In *Conference on Neural Information Processing Systems*, 2024.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. KeepLoRA: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.19659.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Jingyang Qiao, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, Yuan Xie, et al. Prompt gradient projection for continual learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, Weili Guan, Min Zhang, and Liqiang Nie. Splitlora: Balancing stability and plasticity in continual learning through gradient space splitting. *arXiv preprint arXiv:2505.22370*, 2025a.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, and Liqiang Nie. Train with perturbation, infer after merging: A two-stage framework for continual learning. In *Conference on Neural Information Processing Systems*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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

Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vitransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23252–23262, 2024.

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

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025a. URL https://arxiv.org/abs/2406.09044.

Huaijie Wang, De Cheng, Lingfeng He, Yan Li, Jie Li, Nannan Wang, and Xinbo Gao. Ekpc: Elastic knowledge preservation and compensation for class-incremental learning: H. wang et al. *International Journal of Computer Vision*, 134(5):238, 2026a.

Senmiao Wang, Tiantian Fang, Haoran Zhang, Yushun Zhang, Kunxiang Zhao, Alex Schwing, and Ruoyu Sun. Pc layer: Polynomial weight preconditioning for improving llm pre-training. *arXiv preprint arXiv:2606.06470*, 2026b.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Xiequn Wang, Zhan Zhuang, and Yu Zhang. Plan: Proactive low-rank allocation for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2909–2918, 2025b.

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

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

Fangwen Wu, Lechao Cheng, Shengeng Tang, Xiaofeng Zhu, Chaowei Fang, Dingwen Zhang, and Meng Wang. Navigating semantic drift in task-agnostic class-incremental learning. In *Forty-second International Conference on Machine Learning*, 2025a.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-loRA: Scalable decoupled low-rank adaptation for class incremental learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=5U1rlpX68A.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25613–25622, 2025.

ETHICS STATEMENT

This work studies parameter-efficient continual learning using five existing image classification benchmarks: ImageNet-R, ImageNet-A, CIFAR-100, CUB, and DomainNet. Our experiments use benchmark task sequences rather than data collected from real users. Although SCULPT-LoRA does not require storing or replaying past-task examples, this property does not provide a formal privacy guarantee or establish resistance to inference attacks. Adapted models may also inherit biases from the benchmarks and pretrained backbone. Deployment in sensitive applications therefore requires additional privacy safeguards and evaluation of fairness, robustness, and potential misuse.

REPRODUCIBILITY STATEMENT

The paper describes the two components of SCULPT-LoRA: Elastic Principal Subspace Protection (EPSP) and Polynomial Spectral Balancing (PSB). The appendix provides proofs of the principal-subspace perturbation bound and spectral conditioning results, together with the assumptions and limitations of the accuracy-retention analysis. The experimental section specifies the datasets, pre-trained backbone, evaluation metrics, LoRA rank, protection hyperparameters, training epochs, batch size, optimizer, learning-rate schedule, and hardware. Reported results are summarized as mean and standard deviation over three random seeds: 1993, 1996, and 1997. Component ablations and sensitivity analyses examine the effects of EPSP, PSB, protection strength, and protected-subspace dimension. The PSB polynomial uses the fixed coefficients published by Wang et al. (2026b), without refitting. Section B.1 distinguishes their source fitting interval $[0, 1.1]$ from the normalized spectral domain $[0, 1]$ used in our proofs.

AI USE STATEMENT

Large language models were used to assist with manuscript preparation, including language editing, clarity improvements, and feedback on technical exposition. The authors reviewed and revised the AI-assisted content and take full responsibility for the final manuscript, including the correctness of its methods, theoretical arguments, experimental results, and citations.

APPENDIX CONTENTS

A Geometry of Elastic Principal Subspace Protection 15
A.1 Sharp single-update protection bound 15
A.2 Feasibility and choice of orthonormal basis 16
A.3 Accumulated protection across tasks 17
B Proofs of Spectral Conditioning Guarantees 17
B.1 Parameterization and provenance 17
B.2 Exact Transfer of the Nonzero Spectrum 18
B.3 Condition-Number Control by Log-Spectral Contraction 19
B.4 An Exact Polynomial Certificate for PSB 21
B.5 Condition-Number Improvement for the PSB Polynomial 22
B.6 Joint geometric control and separation of shape from scale 23
C From Principal-Subspace Protection to Accuracy Retention 24
C.1 Setup and notation 24
C.2 A general margin-based retention inequality 24
C.3 An EPSP-based logit perturbation bound 25
C.4 Accuracy retention under EPSP 26
C.5 Extension to multiple adapted layers 27
C.6 Expanding label sets in class-incremental learning 28
D Computational Details and Experimental Scope 28
D.1 Evaluation of the spectral parameterization 28
D.2 Reported benchmark configuration and diagnostic definitions 29

A GEOMETRY OF ELASTIC PRINCIPAL SUBSPACE PROTECTION

Throughout the appendix, $W_0 \in \mathbb{R}^{n\times m}$ denotes an original pretrained layer, $A \in \mathbb{R}^{r\times m}$ the fixed input basis, and $B \in \mathbb{R}^{n\times r}$ the trainable factor. The effective factor after PSB is denoted by $\widehat{B}$. The projector $P_K = V_K V_K^\top \in \mathbb{R}^{m\times m}$ is defined from the original pretrained weight and remains fixed across tasks. Task and layer indices are suppressed only for single-update statements. The norm $\|\cdot\|_2$ denotes the spectral norm for matrices and the Euclidean norm for vectors.

A.1 SHARP SINGLE-UPDATE PROTECTION BOUND

Theorem A.1 (Sharp Bound for Elastic Principal Subspace Protection)**.** *Let $W_0 \in \mathbb{R}^{n\times m}$, let $V_K \in \mathbb{R}^{m\times K}$ contain its top-K right singular vectors, and define $P_K = V_K V_K^\top$. Fix $R \in \mathbb{R}^{m\times r}$, with $1 \le r \le m-K$, such that $H = R^\top(I-P_K)R \succ 0$, and let $G = R^\top P_K R$ and $\lambda_\star = \lambda_{\max}(H^{-1/2} G H^{-1/2})$. For $\alpha \in [0,1]$, let $(I-\alpha P_K)R = Q_\alpha C_\alpha$ be a thin QR decomposition with $Q_\alpha^\top Q_\alpha = I_r$, and set $A_\alpha = Q_\alpha^\top$ and $c_\alpha = \sqrt{(1-\alpha)^2\lambda_\star/[1+(1-\alpha)^2\lambda_\star]}$. Then, for every $B \in \mathbb{R}^{n\times r}$,*

$$\|(W_0 + BA_\alpha)P_K - W_0 P_K\|_F \le c_\alpha \|B\|_F. \tag{7}$$

Moreover, this bound is sharp: for every $M \ge 0$,

$$\sup_{\|B\|_F \le M} \|(W_0 + BA_\alpha)P_K - W_0 P_K\|_F = M c_\alpha. \tag{8}$$

The coefficient c_α is nonincreasing in α, with $c_1 = 0$. In particular, $(W_0 + BA_1)P_K = W_0 P_K$ for every B.

Proof. Write $\beta = 1-\alpha$ and

$$F_\alpha = (I-\alpha P_K)R = (I-P_K)R + \beta P_K R.$$

Because P_K is an orthogonal projector,

$$F_\alpha^\top F_\alpha = H + \beta^2 G, \qquad F_\alpha^\top P_K F_\alpha = \beta^2 G. \tag{9}$$

Since $H \succ 0$ and $G \succeq 0$, the matrix $F_\alpha^\top F_\alpha$ is positive definite for every $\alpha \in [0,1]$. Hence F_α has full column rank, and C_α is invertible.

We first characterize the maximal overlap between $\mathrm{col}(Q_\alpha)$ and $\mathrm{col}(V_K)$. Using $A_\alpha = Q_\alpha^\top$ gives

$$\begin{aligned} \|A_\alpha V_K\|_2^2 &= \lambda_{\max}(Q_\alpha^\top P_K Q_\alpha) \\ &= \max_{y\neq 0} \frac{y^\top Q_\alpha^\top P_K Q_\alpha y}{y^\top y} \\ &= \max_{z\neq 0} \frac{\beta^2 z^\top G z}{z^\top H z + \beta^2 z^\top G z}. \end{aligned} \tag{10}$$

The last equality follows from the invertible change of variables $y = C_\alpha z$, together with $Q_\alpha C_\alpha = F_\alpha$ and equation 9.

For $z \neq 0$, define

$$q(z) = \frac{z^\top G z}{z^\top H z}.$$

By the Rayleigh–Ritz characterization,

$$0 \le q(z) \le \lambda_\star, \qquad \max_{z\neq 0} q(z) = \lambda_\star.$$

Since $q \mapsto \beta^2 q/(1+\beta^2 q)$ is nondecreasing for $q \ge 0$, Eq. equation 10 yields

$$\|A_\alpha V_K\|_2^2 = \frac{\beta^2\lambda_\star}{1+\beta^2\lambda_\star} = c_\alpha^2. \tag{11}$$

Next, the change in the protected parameter component is

$$(W_0 + BA_\alpha)P_K - W_0 P_K = BA_\alpha V_K V_K^\top.$$

For every $Z \in \mathbb{R}^{n\times K}$, $V_K^\top V_K = I_K$ implies

$$\|ZV_K^\top\|_F^2 = \operatorname{tr}(ZV_K^\top V_K Z^\top) = \|Z\|_F^2.$$

Consequently,

$$\begin{aligned}\|(W_0 + BA_\alpha)P_K - W_0P_K\|_F &= \|BA_\alpha V_K\|_F \\ &\le \|B\|_F \|A_\alpha V_K\|_2 \\ &= c_\alpha \|B\|_F,\end{aligned}$$

which proves equation 7.

To establish sharpness, let $D = A_\alpha V_K$. If $D = 0$, Eq. equation 8 holds immediately. Otherwise, let $u \in \mathbb{R}^r$ be a unit left singular vector of D associated with $\|D\|_2$. For any unit vector $a \in \mathbb{R}^n$, choose

$$B_\star = Mau^\top.$$

Then $\|B_\star\|_F = M$, and

$$\|B_\star D\|_F = M\|u^\top D\|_2 = M\|D\|_2 = Mc_\alpha.$$

Thus the upper bound is attained.

Finally,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} c_\alpha^2 = -\frac{2(1-\alpha)\lambda_\star}{\left[1 + (1-\alpha)^2\lambda_\star\right]^2} \le 0.$$

Therefore c_α is nonincreasing in α. Since $c_1 = 0$, strict protection preserves W_0P_K exactly. □

Interpretation. Theorem A.1 characterizes protection through the change in the pretrained parameter component W_0P_K. Since $A_\alpha A_\alpha^\top = I_r$, we have $\|BA_\alpha\|_F = \|B\|_F$. Thus, under an equal effective-update Frobenius-norm budget, increasing α does not increase the exact worst-case perturbation of the protected component. At $\alpha = 1$, the layer response is unchanged for every input in $\operatorname{col}(V_K)$. For $\alpha < 1$, the parameterization permits controlled access to these directions. This guarantee concerns the restriction of the effective weight to the original principal input subspace; it does not assert invariance of the leading singular subspace of the adapted weight.

Norm and energy interpretation. For every nonzero effective factor $\widehat{B}$, row orthonormality gives

$$\frac{\|\widehat{B}A_\alpha P_K\|_F}{\|\widehat{B}A_\alpha\|_F} \le c_\alpha, \qquad \frac{\|\widehat{B}A_\alpha P_K\|_F^2}{\|\widehat{B}A_\alpha\|_F^2} \le c_\alpha^2. \tag{12}$$

Thus c_α bounds a norm fraction, whereas c_α^2 bounds the fraction of squared Frobenius energy in the protected component. The result applies directly to the factor produced by PSB. It does not bound the absolute perturbation independently of the learned update magnitude, nor does it prove that any intermediate α maximizes classification accuracy. At $\alpha = 0$, the method uses an unprotected random subspace; this is still a fixed rank-r parameterization, not unrestricted full-matrix adaptation.

A.2 Feasibility and choice of orthonormal basis

Feasibility. Assume $1 \le r \le \min\{n, m-K\}$ and $0 \le K < m$. If the entries of R are independent standard Gaussian variables, then $H = R^\top(I_m - P_K)R \succ 0$ almost surely. Indeed, for an orthonormal complement $V_\perp \in \mathbb{R}^{m\times(m-K)}$, $V_\perp^\top R$ is an $(m-K)\times r$ standard Gaussian matrix, which has full column rank almost surely, and $H = (V_\perp^\top R)^\top(V_\perp^\top R)$. This common assumption makes the construction valid throughout $\alpha \in [0,1]$, including strict exclusion at $\alpha = 1$. For $\alpha < 1$ alone, invertibility of $I_m - \alpha P_K$ means that full column rank of R suffices; $r \le m - K$ is not required merely to compute that particular basis. When $K = 0$, $P_K = 0$ and $c_\alpha = 0$ because the protected component is empty.

QR extracts a basis without changing the chosen subspace. Let $F_\alpha = (I_m - \alpha P_K)R$ have full column rank. Both thin QR and a thin SVD produce an orthonormal basis of $\mathrm{col}(F_\alpha)$. If $\widetilde{Q}$ and Q are two such bases, then $\widetilde{Q} = QO$ for an orthogonal $O \in \mathbb{R}^{r\times r}$. Consequently, $\widetilde{A} = O^\top A$ and $\|\widetilde{A}V_K\|_2 = \|AV_K\|_2$. For $\widetilde{B} = BO$, we also have $\widetilde{B}\widetilde{A} = BA$. Thus the protection guarantee depends on the subspace, not on the particular orthonormal coordinates returned by QR.

The PSB forward map respects the same change of coordinates: with $s = \max\{\|B\|_2, \epsilon\}$,

$$\mathrm{PSB}(BO) = \gamma s \frac{BO}{s} p_4\left(O^\top \frac{B^\top B}{s^2} O\right) = \mathrm{PSB}(B)O.$$

The equality follows from polynomial functional calculus and $\|BO\|_2 = \|B\|_2$. Hence the effective update is unchanged under the paired transformation $(B, A) \mapsto (BO, O^\top A)$. This is a geometric and forward-map equivalence; it does not assert identical training trajectories for arbitrary coordinate-dependent optimizers. Selecting singular vectors of W_0 instead of orthonormalizing F_α generally selects a different subspace and is a different initialization design.

A.3 Accumulated protection across tasks

Corollary A.2 (Cumulative protected-component perturbation)**.** *Fix the original pretrained projector P_K. For each task t, let A_t be an EPSP basis with seed R_t and strength α_t, where $H_t = R_t^\top(I_m - P_K)R_t \succ 0$. Define*

$$G_t^{\mathrm{seed}} = R_t^\top P_K R_t, \quad \lambda_{\star,t} = \lambda_{\max}(H_t^{-1/2} G_t^{\mathrm{seed}} H_t^{-1/2}), \quad c_t = \sqrt{\frac{(1-\alpha_t)^2\lambda_{\star,t}}{1 + (1-\alpha_t)^2\lambda_{\star,t}}}.$$

Let $W^t = W^{t-1} + \widehat{B}_t A_t$, with $W^0 = W_0$. For any $0 \le u < v \le T$,

$$\|(W^v - W^u)P_K\|_F \le \sum_{t=u+1}^{v} c_t \|\widehat{B}_t\|_F \le \left(\sum_{t=u+1}^{v} c_t^2\right)^{1/2} \left(\sum_{t=u+1}^{v} \|\widehat{B}_t\|_F^2\right)^{1/2}. \tag{13}$$

In particular, strict exclusion at every intervening task gives $W^v P_K = W^u P_K$.

Proof. The merged update telescopes: $(W^v - W^u)P_K = \sum_{t=u+1}^{v} \widehat{B}_t A_t P_K$. Apply the triangle inequality and Theorem A.1 to each summand. The second inequality is Cauchy–Schwarz. If every $\alpha_t = 1$, every $c_t = 0$. □

This deterministic bound permits either reused or independently sampled task bases and does not assume independent updates. For fixed seeds and common per-task norm budgets, increasing the protection strengths cannot increase its worst-case upper bound. Elastic protection can nevertheless accumulate drift over tasks; the result is not a task-independent forgetting bound. The projector must remain the same across the compared tasks: protecting a succession of different projectors does not by itself preserve the original pretrained component.

B Proofs of Spectral Conditioning Guarantees

This appendix proves Lemma B.1, Theorem B.2, and Corollary B.3. All statements concern exact forward evaluations. In particular, the stop-gradient operator is the identity in the forward pass and therefore does not affect the spectral identities below.

B.1 Parameterization and provenance

Fixed spectral parameterization. For $\epsilon > 0$, PSB uses

$$\begin{gathered} s = \max\{\|B\|_2, \epsilon\}, \quad X = B/s, \quad G_B = X^\top X, \\ \widehat{B} = \gamma\, \mathrm{sg}(s) X p_4(G_B), \\ p_4(z) = 3.625 - 9.261z + 14.097z^2 - 10.351z^3 + 2.890z^4. \end{gathered} \tag{14}$$

Here G_B is the factor Gram matrix, distinct from the seed Gram matrix G used in the EPSP proof. The operator sg is the identity in the forward pass and stops gradients only through the outer scale-recovery factor; the denominator in $X = B/s$ is not detached. The coefficients are fixed across tasks and layers, and the learnable scalar γ is initialized to one.

Published coefficients and their construction. We adopt the coefficient set published by Wang et al. (2026b) for degree $k = 4$ and target cutoff $b = 0.3$, without refitting. Their construction approximates

$$h_b(\sigma) = \min\{\sigma/b, 1\}, \qquad \sigma \geq 0,$$

by an odd map $f(\sigma) = \sigma p_k(\sigma^2)$ with $\deg(p_k) \leq k$. In particular, their finite-sample fitting procedure solves

$$\boldsymbol{a}^\star \in \arg\min_{\boldsymbol{a}\in\mathbb{R}^{k+1}} \sum_{j=1}^{M} \sigma_j^{\omega} \left(\sum_{\ell=0}^{k} a_\ell \sigma_j^{2\ell+1} - h_b(\sigma_j) \right)^2, \tag{15}$$

where $\sigma_1, \ldots, \sigma_{M-1}$ are sampled uniformly from $[0, 1.1]$ and $\sigma_M = 1$ is included as an additional fitting sample. Here ω is the fitting-weight exponent, not the EPSP protection strength α. The added unit sample encourages agreement at the nominal unit scale; it is not an explicit equality constraint $f(1) = 1$. The expanded fitting interval provides a margin for spectral-norm estimation error in the original PC layer.

Equation equation 15 describes the provenance of the adopted coefficients, not an optimization problem solved by SCULPT-LoRA. Reproducing our forward map requires the five coefficients in Eq. equation 14, not reconstructing the original sampling realization or choosing M and ω. We treat these displayed decimal coefficients as exact rational numbers in the analysis. They satisfy $f(1) = 1$ and $f(\sigma) = 3.625\sigma + O(\sigma^3)$ near zero. The soft target motivates the map; the fitted polynomial need not equal that target or produce equal nonzero singular values.

Fitting domain versus certification domain. The source fitting interval $[0, 1.1]$ is distinct from the spectral domain used in our guarantees. Under the normalization in Eq. equation 14,

$$0 \leq \sigma_i(X) = \frac{\sigma_i(B)}{s} \leq 1, \qquad 0 \preceq G_B \preceq I_r.$$

Thus the scalar response is evaluated at $x \in [0, 1]$, and its polynomial argument $z = x^2$ also lies in $[0, 1]$. The certificates below directly verify the adopted polynomial on this domain; they do not follow merely from least-squares optimality or from the source fitting interval. In particular, using coefficients fitted on $[0, 1.1]$ does not require changing the normalization, the condition-number guarantee, or the norm bounds established below.

Application to the LoRA factor. For any polynomial p, the identity

$$p(XX^\top)X = Xp(X^\top X)$$

follows termwise from $(XX^\top)^j X = X(X^\top X)^j$. We evaluate the right-hand expression through the compact $r \times r$ Gram matrix. This applies the PC spectral map to the trainable LoRA factor. The row-orthonormal EPSP basis then transfers its nonzero spectrum exactly to the effective weight update, as proved in Lemma B.1. The polynomial coefficients themselves are inherited from Wang et al. (2026b).

Notation. For a nonzero matrix $B \in \mathbb{R}^{n\times r}$ of rank q, write $\sigma_1(B) \geq \cdots \geq \sigma_q(B) > 0$ and define $\kappa_+(B) = \frac{\sigma_1(B)}{\sigma_q(B)}, \mathrm{srank}(B) = \frac{\|B\|_F^2}{\|B\|_2^2}$. We additionally report the entropy-based effective rank $\mathrm{erank}(B) = \exp(-\sum_{i=1}^{q} \pi_i \log \pi_i)\,, \pi_i = \frac{\sigma_i(B)}{\sum_{j=1}^{q} \sigma_j(B)}$. All three quantities are invariant under nonzero scalar multiplication. The following results concern exact forward evaluations.

B.2 Exact Transfer of the Nonzero Spectrum

Lemma B.1 (Exact transfer of the nonzero spectrum)**.** *Let $B \in \mathbb{R}^{n\times r}$ have rank $q > 0$, let $s > 0$, and let $\gamma \neq 0$. Set $X = B/s, \widehat{B} = \gamma s X p(X^\top X), f(x) = xp(x^2)$, where p is a real polynomial.*

Let $x_i = \sigma_i(B)/s$. The potentially nonzero singular values of $\widehat{B}$ are $|\gamma|s|f(x_i)|, i = 1, \dots, q$, up to ordering. In particular, if $f(x_i) \neq 0$ for every i, then $\operatorname{rank}(\widehat{B}) = q$.

If $A \in \mathbb{R}^{r\times m}$ satisfies $AA^\top = I_r$, then $\widehat{B}A$ and $\widehat{B}$ have identical nonzero singular values, including multiplicities. Hence, whenever $\widehat{B} \neq 0$, they have the same κ_+, stable rank, and effective rank.

Proof of Lemma B.1. Let

$$B = U_q \Sigma_q V_q^\top$$

be a compact singular value decomposition of B, where $U_q \in \mathbb{R}^{n\times q}$ and $V_q \in \mathbb{R}^{r\times q}$ have orthonormal columns, and

$$\Sigma_q = \operatorname{diag}\big(\sigma_1(B), \dots, \sigma_q(B)\big).$$

Writing

$$D = \Sigma_q/s = \operatorname{diag}(x_1, \dots, x_q),$$

we obtain

$$X = U_q D V_q^\top, \qquad X^\top X = V_q D^2 V_q^\top.$$

For any real polynomial p, polynomial functional calculus gives

$$p(X^\top X) = V_q p(D^2) V_q^\top + p(0)\big(I_r - V_q V_q^\top\big). \tag{16}$$

The second term accounts for the nullspace of X. Since $V_q^\top(I_r - V_q V_q^\top) = 0$, multiplication by X yields

$$\begin{aligned} Xp(X^\top X) &= U_q D p(D^2) V_q^\top \\ &= U_q \operatorname{diag}\big(f(x_1), \dots, f(x_q)\big) V_q^\top, \end{aligned} \tag{17}$$

where $f(x) = xp(x^2)$. Consequently,

$$\widehat{B} = \gamma s U_q \operatorname{diag}\big(f(x_1), \dots, f(x_q)\big) V_q^\top.$$

Because U_q and V_q have orthonormal columns, the potentially nonzero singular values of $\widehat{B}$ are

$$|\gamma|s|f(x_i)|, \qquad i = 1, \dots, q,$$

up to ordering. All remaining singular values are zero. Thus

$$\operatorname{rank}(\widehat{B}) = \#\{i : f(x_i) \neq 0\}.$$

In particular, the rank equals q whenever every $f(x_i)$ is nonzero.

Now suppose that $AA^\top = I_r$. Then

$$(\widehat{B}A)(\widehat{B}A)^\top = \widehat{B}AA^\top\widehat{B}^\top = \widehat{B}\widehat{B}^\top. \tag{18}$$

The squared nonzero singular values of a matrix are precisely the nonzero eigenvalues of its left Gram matrix, including multiplicities. Hence $\widehat{B}A$ and $\widehat{B}$ have identical nonzero singular values.

Whenever $\widehat{B} \neq 0$, this identity preserves both the largest and smallest positive singular values, the sum of their squares, and their normalized proportions. It therefore preserves κ_+, stable rank, and entropy-based effective rank. □

B.3 Condition-Number Control by Log-Spectral Contraction

Theorem B.2 (Condition-number control by log-spectral contraction)**.** *Under the notation of Lemma B.1, suppose the nonzero singular values of X lie in $[a, b] \subset (0, \infty)$. Assume that f is positive and continuously differentiable on an open interval containing $[a, b]$, and that, for some $\rho \in [0, 1]$, $\left|\frac{xf'(x)}{f(x)}\right| \leq \rho, x \in [a, b]$. Then $\operatorname{rank}(\widehat{B}) = q$ and $\kappa_+(\widehat{B}) \leq \kappa_+(B)^\rho$. If $\rho < 1$ and $\kappa_+(B) > 1$, the condition number strictly decreases. The same conclusions hold for $\widehat{B}A$ relative to BA whenever $AA^\top = I_r$.*

Proof of Theorem B.2. Since $x_i \in [a, b]$ and f is positive on this interval, $f(x_i) > 0$ for every i. Lemma B.1 therefore implies

$$\operatorname{rank}(\widehat{B}) = q.$$

Define the logarithmic spectral response

$$h(u) = \log f(e^u), \qquad u \in [\log a, \log b].$$

The assumed positivity and continuous differentiability of f ensure that h is continuously differentiable on an open interval containing this domain. By the chain rule,

$$h'(u) = \frac{e^u f'(e^u)}{f(e^u)}.$$

The assumed log-derivative bound thus gives

$$|h'(u)| \le \rho, \qquad u \in [\log a, \log b].$$

By the mean value theorem, for any u, v in this interval,

$$|h(u) - h(v)| \le \rho |u - v|. \tag{19}$$

Let $u_i = \log x_i$. Applying equation 19 to every pair (u_i, u_j) yields

$$\begin{aligned} \max_i h(u_i) - \min_i h(u_i) &= \max_{i,j} |h(u_i) - h(u_j)| \\ &\le \rho \max_{i,j} |u_i - u_j| \\ &= \rho \log \frac{x_1}{x_q}. \end{aligned} \tag{20}$$

This argument does not require f to be monotone: the transformed singular values may change their ordering.

By Lemma B.1, the common positive factor $|\gamma| s$ cancels in the condition number, giving

$$\begin{aligned} \log \kappa_+(\widehat{B}) &= \log \frac{\max_i f(x_i)}{\min_i f(x_i)} \\ &= \max_i h(u_i) - \min_i h(u_i) \\ &\le \rho \log \frac{x_1}{x_q} \\ &= \rho \log \kappa_+(B). \end{aligned} \tag{21}$$

Exponentiating proves

$$\kappa_+(\widehat{B}) \le \kappa_+(B)^\rho.$$

If $\rho < 1$ and $\kappa_+(B) > 1$, then

$$\kappa_+(B)^\rho < \kappa_+(B),$$

so the condition number strictly decreases.

Finally, if $AA^\top = I_r$, the Gram-matrix identity used in Lemma B.1 applies to both B and $\widehat{B}$. Therefore,

$$\kappa_+(BA) = \kappa_+(B), \qquad \kappa_+(\widehat{B}A) = \kappa_+(\widehat{B}),$$

and

$$\operatorname{rank}(BA) = \operatorname{rank}(B) = q = \operatorname{rank}(\widehat{B}) = \operatorname{rank}(\widehat{B}A).$$

The same conclusions consequently hold for the effective update. □

B.4 AN EXACT POLYNOMIAL CERTIFICATE FOR PSB

Corollary B.3 (Condition-number improvement for the PSB polynomial)**.** *Let $p = p_4$ be the polynomial in Eq. equation 23, with the displayed coefficients treated as exact. For any nonzero B, $\epsilon > 0$, $s = \max\{\|B\|_2, \epsilon\}$, and $\gamma \neq 0$, the effective factor preserves rank and satisfies*

$$\kappa_+(\widehat{B}) \leq \kappa_+(B). \tag{22}$$

The inequality is strict whenever $\kappa_+(B) > 1$. If $AA^\top = I_r$, the same rank preservation and condition-number improvement hold for the effective update $\widehat{B}A$ relative to BA.

We first verify the scalar inequalities required to apply Theorem B.2. The certification domain is $z \in [0, 1]$, as justified in Section B.1. No least-squares optimality assumption is required. Throughout this subsection, the displayed decimal coefficients are treated as exact rational numbers:

$$p_4(z) = \frac{3625 - 9261z + 14097z^2 - 10351z^3 + 2890z^4}{1000}. \tag{23}$$

Bernstein positivity certificate. For a polynomial g of degree at most d on $[\ell, u]$, set $z = \ell + (u-\ell)t$, where $t \in [0, 1]$, and write

$$g\big(\ell + (u-\ell)t\big) = \sum_{k=0}^{d} \beta_k \binom{d}{k} t^k (1-t)^{d-k}. \tag{24}$$

The Bernstein basis functions are nonnegative and sum to one. Consequently,

$$\min_k \beta_k \leq g(z) \leq \max_k \beta_k, \qquad z \in [\ell, u]. \tag{25}$$

In particular, strictly positive Bernstein coefficients certify strict positivity on the entire closed interval.

For completeness, if

$$g\big(\ell + (u-\ell)t\big) = \sum_{j=0}^{d} a_j t^j,$$

then the coefficients in equation 24 are

$$\beta_k = \sum_{j=0}^{k} a_j \frac{\binom{k}{j}}{\binom{d}{j}}, \qquad k = 0, \ldots, d. \tag{26}$$

Thus all certificates below can be checked by exact rational arithmetic.

Define

$$\begin{aligned} R(z) &:= -p_4'(z) \\ &= \frac{9261 - 28194z + 31053z^2 - 11560z^3}{1000}, \end{aligned} \tag{27}$$

$$\begin{aligned} Q(z) &:= p_4(z) + zp_4'(z) \\ &= \frac{3625 - 18522z + 42291z^2 - 41404z^3 + 14450z^4}{1000}. \end{aligned} \tag{28}$$

The degree-three Bernstein coefficients of R on the two half-intervals are

$$\boldsymbol{\beta}^R_{[0,1/2]} = \left(\frac{9261}{1000}, \frac{2281}{500}, \frac{9803}{4000}, \frac{5929}{4000}\right), \tag{29}$$

$$\boldsymbol{\beta}^R_{[1/2,1]} = \left(\frac{5929}{4000}, \frac{411}{800}, \frac{86}{125}, \frac{14}{25}\right). \tag{30}$$

Every coefficient is strictly positive. Hence

$$R(z) > 0, \qquad z \in [0, 1], \tag{31}$$

or equivalently, $p_4'(z) < 0$ throughout $[0, 1]$. Since $p_4(1) = 1$, it follows that

$$p_4(z) \geq 1 > 0, \qquad z \in [0, 1]. \tag{32}$$

Similarly, the degree-four Bernstein coefficients of Q are

$$\boldsymbol{\beta}^Q_{[0,1/2]} = \left(\frac{29}{8}, \frac{5239}{4000}, \frac{6053}{8000}, \frac{2687}{4000}, \frac{1063}{1600}\right), \tag{33}$$

$$\boldsymbol{\beta}^Q_{[1/2,1]} = \left(\frac{1063}{1600}, \frac{657}{1000}, \frac{5817}{8000}, \frac{121}{250}, \frac{11}{25}\right). \tag{34}$$

These coefficients are also strictly positive, proving

$$Q(z) > 0, \qquad z \in [0, 1]. \tag{35}$$

Strict log-spectral contraction. Let $f(x) = xp_4(x^2)$ and $z = x^2$. For $x \in (0, 1]$, equation 32 gives $f(x) > 0$, and differentiation yields

$$\frac{xf'(x)}{f(x)} = \frac{p_4(z) + 2zp_4'(z)}{p_4(z)} = 1 + \frac{2zp_4'(z)}{p_4(z)}. \tag{36}$$

Since $z > 0$, $p_4'(z) < 0$, and $p_4(z) > 0$, we obtain

$$\frac{xf'(x)}{f(x)} < 1.$$

For the lower bound, observe that

$$\begin{aligned} 1 + \frac{xf'(x)}{f(x)} &= \frac{2\big(p_4(z) + zp_4'(z)\big)}{p_4(z)} \\ &= \frac{2Q(z)}{p_4(z)} > 0. \end{aligned} \tag{37}$$

Therefore,

$$-1 < \frac{xf'(x)}{f(x)} < 1, \qquad x \in (0, 1]. \tag{38}$$

This establishes the required inequality over the entire normalized interval by exact polynomial certificates.

B.5 Condition-Number Improvement for the PSB Polynomial

Proof of Corollary B.3. Let $B \neq 0$ have rank q, and define

$$s = \max\{\|B\|_2, \epsilon\}, \qquad a = \frac{\sigma_q(B)}{s}, \qquad b = \frac{\sigma_1(B)}{s}.$$

Since $\epsilon > 0$ and $B \neq 0$,

$$0 < a \leq b \leq 1.$$

By equation 32, $f(x) = xp_4(x^2)$ is positive on $[a, b]$. Since f is a polynomial and is strictly positive on this compact interval, it is positive and continuously differentiable on an open interval containing $[a, b]$.

Define

$$\rho_B := \max_{x \in [a,b]} \left| \frac{xf'(x)}{f(x)} \right|. \tag{39}$$

The function being maximized is continuous on $[a, b]$, so its maximum is attained. By equation 38, its value is strictly less than one at every point of this interval. Consequently,

$$0 \leq \rho_B < 1.$$

Applying Theorem B.2 with $\rho = \rho_B$ gives

$$\operatorname{rank}(\widehat{B}) = q, \qquad \kappa_+(\widehat{B}) \leq \kappa_+(B)^{\rho_B} \leq \kappa_+(B). \tag{40}$$

If $\kappa_+(B) > 1$, the last inequality is strict because $\rho_B < 1$.

If $\kappa_+(B) = 1$, all positive singular values of B are equal, and hence all x_i are equal. Lemma B.1 then shows that the positive singular values of $\widehat{B}$ are also equal, so

$$\kappa_+(\widehat{B}) = 1.$$

This also covers the rank-one case.

Finally, if $AA^\top = I_r$, the nonzero spectra of BA and B coincide, as do those of $\widehat{B}A$ and $\widehat{B}$. Thus the effective update preserves rank and satisfies exactly the same condition-number inequalities. □

Scope of the strict contraction. The contraction constant ρ_B depends on the normalized nonzero spectral interval of B. Indeed,

$$\lim_{x\downarrow 0} \frac{xf'(x)}{f(x)} = \lim_{z\downarrow 0} \left(1 + \frac{2zp_4'(z)}{p_4(z)}\right) = 1.$$

Hence the pointwise strict bound on $(0, 1]$ does not provide a uniform constant $\rho < 1$ over that entire interval. Nevertheless, every fixed nonzero matrix has $a > 0$, which suffices for the strict improvement proved above whenever its nonzero spectrum is nonflat.

The proof does not assume monotonicity of f and allows the transformed singular values to change ordering. The resulting guarantees concern rank and the nonzero condition number at a fixed parameter value; they do not by themselves establish an increase in stable rank or effective rank, or a guarantee of optimization convergence.

B.6 Joint geometric control and separation of shape from scale

Proposition B.4 (Simultaneous geometric and spectral guarantees)**.** *Let A_α satisfy the EPSP assumptions, let $B \neq 0$, and let $\widehat{B}$ be the PSB factor with $\gamma \neq 0$ and exact normalization. Set $\widehat{\Delta W} = \widehat{B}A_\alpha$. Then*

$$\begin{aligned} \operatorname{rank}(\widehat{\Delta W}) &= \operatorname{rank}(B), \\ \kappa_+(\widehat{\Delta W}) &= \kappa_+(\widehat{B}) \leq \kappa_+(BA_\alpha), \\ \frac{\|\widehat{\Delta W} P_K\|_F}{\|\widehat{\Delta W}\|_F} &\leq c_\alpha. \end{aligned} \tag{41}$$

The condition-number inequality is strict if $\kappa_+(B) > 1$. Furthermore,

$$|\gamma|\|B\|_F \leq \|\widehat{B}\|_F \leq 3.625|\gamma|\|B\|_F. \tag{42}$$

Proof. Rank preservation and condition-number control follow from Lemma B.1 and Corollary B.3. Apply Theorem A.1 to $\widehat{B}$ and use $\|\widehat{B}A_\alpha\|_F = \|\widehat{B}\|_F$ for the normalized protection bound. For the scale bounds, positivity and monotonic decrease of p_4 on $[0, 1]$ imply $1 \leq p_4(z) \leq p_4(0) = 3.625$. Each positive singular value of $\widehat{B}$ is $|\gamma|\sigma_i(B)p_4((\sigma_i(B)/s)^2)$. Squaring, summing, and taking square roots proves Eq. equation 42. □

Implication for mechanism attribution. The guarantees control two distinct quantities: the protected fraction of the effective update and its nonzero condition number. PSB can also change the update magnitude, even with $\gamma = 1$. Consequently, an accuracy difference between separately trained models with and without PSB cannot be attributed to spectral shape alone using these guarantees.

A norm-matched diagnostic can isolate this forward-map distinction. For $B \neq 0$ and $\gamma \neq 0$, define

$$\widetilde{B} = \frac{\|B\|_F}{\|\widehat{B}\|_F}\widehat{B}, \qquad \widetilde{\Delta W} = \widetilde{B}A_\alpha.$$

Then $\|\widetilde{\Delta W}\|_F = \|BA_\alpha\|_F$; rank, condition number, stable rank, and effective rank are the same as those of $\widehat{\Delta W}$, and the protected norm fraction remains at most c_α. These identities follow from scalar invariance and row orthonormality. Set $\widetilde{B} = 0$ when $B = 0$. This is an analytical diagnostic, not an additional component of SCULPT-LoRA or a claim of an evaluated ablation. Matched training comparisons would additionally need the same learnable scaling, initialization, optimizer, and tuning budget.

Zero factors and numerical rank. If $B = 0$, then $X = 0$ and $\widehat{B} = 0$; if $\gamma = 0$, the effective update also vanishes. The forward map remains defined, but κ_+ and the normalized rank statistics used above are not defined for a zero matrix. Rank preservation statements thus explicitly require $B \neq 0$ and $\gamma \neq 0$. The polynomial does not create new nonzero singular directions at a fixed parameter value. During optimization, the factor itself can change rank; that is a separate statement from the forward-map theorem. In numerical measurements, the threshold for counting a singular value as nonzero must be specified.

C From Principal-Subspace Protection to Accuracy Retention

This section connects the geometric guarantee of EPSP to classification accuracy through prediction margins. The results are deterministic with respect to the model parameters and hold for any evaluation distribution satisfying the stated assumptions. They do not assume that improved spectral conditioning alone implies improved accuracy.

C.1 Setup and notation

Let $\mathcal{D}_{old}$ be a distribution over examples (x, y) with labels in a finite set $\mathcal{Y}$, where $|\mathcal{Y}| \geq 2$. Initially, both models are evaluated over the same candidate label set $\mathcal{Y}$. Let

$$z, z^+ : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$$

denote the logits before and after adaptation. Predictions use a fixed deterministic tie-breaking rule:

$$h(x) = \arg\max_{j \in \mathcal{Y}} z_j(x), \qquad h^+(x) = \arg\max_{j \in \mathcal{Y}} z_j^+(x).$$

Define

$$\mathrm{Acc}_{old}(z) = \Pr_{\mathcal{D}_{old}}[h(x) = y]$$

and the pre-adaptation true-label margin

$$m_z(x, y) = z_y(x) - \max_{j \in \mathcal{Y} \setminus \{y\}} z_j(x).$$

Correct predictions satisfy $m_z(x, y) \geq 0$; strictly positive margins guarantee correct predictions independently of the tie-breaking rule.

All probability statements below are conditional on the chosen pre-adaptation and adapted models. In particular, the results do not require independence between these models and the training data.

C.2 A general margin-based retention inequality

Lemma C.1 (Margin stability under logit perturbations)**.** *Suppose a measurable function $d : \mathcal{X} \to [0, \infty)$ satisfies*

$$\|z^+(x) - z(x)\|_\infty \leq d(x).$$

Then, for every (x, y),

$$|m_{z^+}(x, y) - m_z(x, y)| \leq 2d(x).$$

Consequently,

$$m_z(x, y) > 2d(x) \quad \Longrightarrow \quad h^+(x) = y.$$

Proof. For any two finite real-valued collections $(a_j)_j$ and $(b_j)_j$,

$$\left| \max_j a_j - \max_j b_j \right| \leq \max_j |a_j - b_j|.$$

Indeed, if $j_\star$ maximizes a_j, then

$$\max_j a_j - \max_j b_j \leq a_{j_\star} - b_{j_\star} \leq \max_j |a_j - b_j|,$$

and the reverse inequality follows by exchanging the two collections.

Applying this fact to the competing-class logits gives

$$\begin{aligned}|m_{z^+}(x,y) - m_z(x,y)| &\le |z_y^+(x) - z_y(x)| \\ &\quad + \left|\max_{j\neq y} z_j^+(x) - \max_{j\neq y} z_j(x)\right| \\ &\le 2d(x).\end{aligned}$$

Therefore,

$$m_{z^+}(x,y) \ge m_z(x,y) - 2d(x).$$

If $m_z(x,y) > 2d(x)$, the adapted true-label margin is strictly positive, which implies $h^+(x) = y$. □

Theorem C.2 (Distribution-dependent accuracy retention)**.** *Under the assumptions of Lemma C.1,*

$$\begin{aligned}\mathrm{Acc}_{old}(z^+) \ge\ & \mathrm{Acc}_{old}(z) \\ & - \Pr_{\mathcal{D}_{old}}[h(x) = y,\ 0 \le m_z(x,y) \le 2d(x)]\,.\end{aligned}$$

In particular, if $d(x) \le d_0$ almost surely and

$$F_z^{\mathrm{corr}}(u) := \Pr_{\mathcal{D}_{old}}[h(x) = y,\ 0 \le m_z(x,y) \le u]\,, \qquad u \ge 0,$$

then

$$\mathrm{Acc}_{old}(z^+) \ge \mathrm{Acc}_{old}(z) - F_z^{\mathrm{corr}}(2d_0).$$

Proof. Let

$$C = \{h(x) = y\}, \qquad C^+ = \{h^+(x) = y\}.$$

The exact identity

$$\Pr(C^+) = \Pr(C) - \Pr(C \setminus C^+) + \Pr(C^+ \setminus C)$$

implies

$$\Pr(C^+) \ge \Pr(C) - \Pr(C \setminus C^+).$$

On C, the original margin is nonnegative. By Lemma C.1, an example with $m_z(x,y) > 2d(x)$ cannot belong to $C \setminus C^+$. Hence

$$C \setminus C^+ \subseteq \{h(x) = y,\ 0 \le m_z(x,y) \le 2d(x)\}\,.$$

Taking probabilities proves the first claim. If $d(x) \le d_0$ almost surely, the event on the right is contained in the event defining $F_z^{\mathrm{corr}}(2d_0)$, which proves the second claim. □

Ties. The event $h(x) = y$ and the non-strict lower endpoint $m_z(x,y) = 0$ explicitly account for correctly classified ties. If $\Pr[m_z(x,y) = 0] = 0$, the probability term simplifies to

$$\Pr_{\mathcal{D}_{old}}[0 < m_z(x,y) \le 2d(x)]\,.$$

C.3 An EPSP-based logit perturbation bound

Consider a layer with pre-adaptation weight $W \in \mathbb{R}^{n\times m}$ and input representation $\phi(x) \in \mathbb{R}^m$. Let

$$z(x) = g(W\phi(x)), \qquad z^+(x) = g((W + \Delta W)\phi(x)),$$

where the input representation and the downstream map $g : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{Y}|}$ are held fixed.

Use the EPSP construction from Section A, and abbreviate $P = P_K$ and $V = V_K$ in this section. The reference pretrained weight defining P need not equal the current pre-adaptation weight W. For the fixed seed and protection strength under consideration, Eq. equation 11 gives

$$A_\alpha A_\alpha^\top = I_r, \qquad \|A_\alpha V\|_2 = c_\alpha.$$

Theorem C.3 (EPSP-induced logit stability)**.** *Let*

$$\Delta W = \widehat{B} A_\alpha, \qquad \widehat{B} \in \mathbb{R}^{n \times r}, \qquad b = \|\widehat{B}\|_2.$$

Assume that, for all relevant layer outputs u, v,

$$\|g(u) - g(v)\|_\infty \le L\|u - v\|_2.$$

Then, pointwise in x,

$$\|z^+(x) - z(x)\|_\infty \le d_\alpha(x),$$

where

$$d_\alpha(x) := Lb \min\{\|\phi(x)\|_2,\ c_\alpha \|P\phi(x)\|_2 + \|(I_m - P)\phi(x)\|_2\}\,.$$

Proof. Since $A_\alpha A_\alpha^\top = I_r$, $\|A_\alpha\|_2 = 1$. Consequently,

$$\|\Delta W \phi(x)\|_2 \le b\|\phi(x)\|_2.$$

For the protected component, write

$$P\phi(x) = VV^\top \phi(x).$$

Using the EPSP overlap identity and $\|V^\top \phi(x)\|_2 = \|P\phi(x)\|_2$ gives

$$\begin{aligned}\|\Delta W P\phi(x)\|_2 &= \|\widehat{B} A_\alpha VV^\top \phi(x)\|_2 \\ &\le bc_\alpha \|P\phi(x)\|_2.\end{aligned}$$

For the complementary component,

$$\|\Delta W (I_m - P)\phi(x)\|_2 \le b\|(I_m - P)\phi(x)\|_2.$$

The triangle inequality therefore yields

$$\|\Delta W \phi(x)\|_2 \le b\,(c_\alpha \|P\phi(x)\|_2 + \|(I_m - P)\phi(x)\|_2)\,.$$

Both bounds hold simultaneously. Taking their minimum and applying the Lipschitz assumption on g proves the claim. □

Applicability to PSB. Theorem C.3 holds for every effective factor $\widehat{B}$, including the output of PSB. It does not require $\widehat{B}$ to have full rank or to be nonzero. A Frobenius-budget version follows immediately from $\|\widehat{B}\|_2 \le \|\widehat{B}\|_F$.

C.4 Accuracy retention under EPSP

Corollary C.4 (EPSP accuracy retention)**.** *Under the assumptions of Theorem C.3,*

$$\begin{aligned}\mathrm{Acc}_{old}(z^+) \ge\ &\mathrm{Acc}_{old}(z) \\ &- \Pr_{\mathcal{D}_{old}}[h(x) = y,\ 0 \le m_z(x,y) \le 2d_\alpha(x)]\,.\end{aligned}$$

Suppose additionally that, almost surely,

$$\|\phi(x)\|_2 \le R_x, \qquad \|(I_m - P)\phi(x)\|_2 \le \eta R_x, \qquad 0 \le \eta \le 1,$$

and that $\|\widehat{B}\|_2 \le M$. Then, with

$$\overline{d}_\alpha := LMR_x \min\{1, c_\alpha + \eta\},$$

we have

$$\mathrm{Acc}_{old}(z^+) \ge \mathrm{Acc}_{old}(z) - F_z^{\mathrm{corr}}(2\overline{d}_\alpha).$$

If, for some $C > 0$, $\nu > 0$, and $u_0 > 0$,

$$F_z^{\mathrm{corr}}(u) \le Cu^\nu, \qquad 0 < u \le u_0,$$

then, whenever $2\overline{d}_\alpha \le u_0$,

$$\boxed{\mathrm{Acc}_{old}(z^+) \ge \mathrm{Acc}_{old}(z) - C\,[2LMR_x \min\{1, c_\alpha + \eta\}]^\nu\,.}$$

Proof. The first claim follows by substituting $d_\alpha(x)$ into Theorem C.2. The feature and update-budget assumptions imply

$$d_\alpha(x) \le LMR_x \min\{1, c_\alpha + \eta\} = \overline{d}_\alpha.$$

Applying the uniform version of Theorem C.2 proves the second claim.

For $\overline{d}_\alpha > 0$, the last claim follows by applying the assumed small-margin bound at $u = 2\overline{d}_\alpha$. If $\overline{d}_\alpha = 0$, the pointwise logit perturbation bound implies $z^+ = z$ almost surely, so accuracy is preserved exactly. □

Interpretation and scope. For a fixed seed and protected subspace, c_α is nonincreasing in α. Under a common update budget and common values of L, R_x, η, C, ν, the upper bound on accuracy loss is therefore nonincreasing in α. This is a statement about a certified bound, not monotonicity of the realized accuracy across separately trained models.

The feature-leakage parameter η is essential: protecting a weight-defined subspace does not by itself establish that old-task features lie in that subspace. At $\alpha = 1$, the protected component is unchanged, but complementary components can still affect predictions. If $\eta = 0$ as well, logits are preserved exactly under the single-layer setting considered here.

C.5 Extension to multiple adapted layers

We next remove the restriction that only one layer changes. Consider a sequential network

$$z = T_d \circ \cdots \circ T_1, \qquad z^+ = T_d^+ \circ \cdots \circ T_1^+,$$

where T_d outputs logits. Suppose

$$T_\ell(u) = \psi_\ell(W_\ell u), \qquad T_\ell^+(u) = \psi_\ell((W_\ell + \Delta W_\ell)u),$$

with unchanged maps ψ_ℓ.

Define the original prefix representation

$$u_{\ell-1}(x) = T_{\ell-1} \circ \cdots \circ T_1(x), \qquad u_0(x) = x,$$

and the adapted suffix map

$$S_\ell^+ = T_d^+ \circ \cdots \circ T_{\ell+1}^+ \circ \psi_\ell.$$

For $\ell = d$, this means $S_d^+ = \psi_d$.

Proposition C.5 (Multilayer perturbation certificate)**.** *Suppose S_ℓ^+ is L_ℓ-Lipschitz from the Euclidean norm to the logit infinity norm on the relevant inputs. For each layer, let*

$$\Delta W_\ell = \widehat{B}_\ell A_{\alpha_\ell}, \qquad b_\ell = \|\widehat{B}_\ell\|_2,$$

where the EPSP construction uses projector P_ℓ and overlap coefficient c_{α_ℓ}. Define

$$d_{\mathrm{net}}(x) := \sum_{\ell=1}^{d} L_\ell b_\ell \min\Big\{ \|u_{\ell-1}(x)\|_2, \; c_{\alpha_\ell}\|P_\ell u_{\ell-1}(x)\|_2 + \|(I - P_\ell)u_{\ell-1}(x)\|_2 \Big\}.$$

Then

$$\|z^+(x) - z(x)\|_\infty \le d_{\mathrm{net}}(x).$$

Consequently,

$$\mathrm{Acc}_{old}(z^+) \ge \mathrm{Acc}_{old}(z) - \Pr_{\mathcal{D}_{old}}\left[h(x) = y,\ 0 \le m_z(x,y) \le 2d_{\mathrm{net}}(x)\right].$$

Proof. For $\ell = 0, \ldots, d$, introduce hybrid networks

$$Z^{(\ell)} = T_d^+ \circ \cdots \circ T_{\ell+1}^+ \circ T_\ell \circ \cdots \circ T_1.$$

Thus $Z^{(0)} = z^+$ and $Z^{(d)} = z$. For each $\ell \ge 1$, the adjacent hybrids share the original prefix through layer $\ell - 1$ and the adapted suffix after layer ℓ. Hence

$$Z^{(\ell-1)}(x) = S_\ell^+((W_\ell + \Delta W_\ell)u_{\ell-1}(x)),$$
$$Z^{(\ell)}(x) = S_\ell^+(W_\ell u_{\ell-1}(x)).$$

The Lipschitz assumption implies

$$\|Z^{(\ell-1)}(x) - Z^{(\ell)}(x)\|_\infty \le L_\ell \|\Delta W_\ell u_{\ell-1}(x)\|_2.$$

The layerwise EPSP argument in Theorem C.3 bounds the right-hand side by the corresponding summand of $d_{\mathrm{net}}(x)$. Finally, telescoping and the triangle inequality give

$$\begin{aligned}\|z^+(x) - z(x)\|_\infty &= \left\| \sum_{\ell=1}^{d} \left(Z^{(\ell-1)}(x) - Z^{(\ell)}(x) \right) \right\|_\infty \\ &\le d_{\mathrm{net}}(x).\end{aligned}$$

The accuracy bound follows from Theorem C.2. □

Architectural scope. The layer factorization above is an explicit assumption. For architectures with residual paths, attention, or additional trainable parameters, the same hybrid argument remains valid once a corresponding perturbation bound is established for each modified block. Changes to biases, normalization parameters, or classifier weights must be included in those bounds. In particular, the sequential factorization stated here is not itself a verified blockwise certificate for the experimental ViT with a trained cosine classifier. Its application to that architecture requires bounds for the actual attention, residual, normalization, and classifier computations. No numerical nonvacuous accuracy certificate for the experimental models is claimed here.

C.6 Expanding label sets in class-incremental learning

The preceding results compare predictions over a fixed candidate label set. To account for newly introduced classes, let $\mathcal{Y}_{old}$ and $\mathcal{Y}_{new}$ be disjoint, and let the adapted classifier predict over their union. For an old-task example, define the new-class intrusion event

$$E_{\mathrm{new}} = \left\{ \max_{j \in \mathcal{Y}_{new}} z_j^+(x) \ge z_y^+(x) \right\}.$$

Proposition C.6 (Retention with new-class competition)**.** *Let z be the original logits over $\mathcal{Y}_{old}$, with $|\mathcal{Y}_{old}| \ge 2$. Suppose the adapted logits restricted to old classes satisfy*

$$\|z^+_{\mathcal{Y}_{old}}(x) - z(x)\|_\infty \le d(x).$$

Then the adapted accuracy over $\mathcal{Y}_{old} \cup \mathcal{Y}_{new}$ satisfies

$$\begin{aligned}\mathrm{Acc}^{all}_{old}(z^+) \ge\ & \mathrm{Acc}_{old}(z) \\ & - \Pr[h(x) = y,\ 0 \le m_z(x,y) \le 2d(x)] \\ & - \Pr[h(x) = y,\ E_{\mathrm{new}}]\,,\end{aligned}$$

where all probabilities are under $\mathcal{D}_{old}$.

Proof. Consider an originally correct example outside both subtracted events. Its original old-class margin is greater than $2d(x)$, so Lemma C.1 implies that its adapted true-label logit strictly exceeds every competing old-class logit. Since E_{new} does not occur, that logit also strictly exceeds every new-class logit. The adapted classifier therefore predicts y over the full candidate set. Thus, every originally correct example that becomes incorrect belongs to the union of the two events. The claim follows from the union bound and the same accuracy decomposition used in Theorem C.2. □

Limitations of the guarantee. The EPSP-based certificate controls the contribution of bounded parameter perturbations to old-class prediction changes. It does not, without further assumptions, control the new-class intrusion probability, the usefulness of adaptation for new tasks, or optimization convergence. Likewise, a reduction in the nonzero condition number of an update does not alone imply improved accuracy. These distinctions separate the proved retention guarantee from empirical claims about overall class-incremental performance.

D Computational Details and Experimental Scope

D.1 Evaluation of the spectral parameterization

For a fixed pretrained layer, its top-K right singular vectors can be computed once and reused across tasks. The filtered seed can be formed as $F_\alpha = R - \alpha V_K (V_K^\top R)$ without explicitly storing the dense

projector P_K. This costs $O(mKr)$ operations; a thin QR of the $m \times r$ filtered seed costs $O(mr^2)$. These are basis-construction costs, not per-minibatch costs.

For a factor $B \in \mathbb{R}^{n\times r}$, form $S_B = B^\top B$. The exact scale in the analysis is $s = \max\{\sqrt{\lambda_{\max}(S_B)}, \epsilon\}$ and $G_B = S_B/s^2$. Horner evaluation gives

$$p_4(G_B) = 3.625 I_r + G_B\big(-9.261 I_r + G_B(14.097 I_r + G_B(-10.351 I_r + 2.890 G_B))\big).$$

The constant term is a multiple of I_r, not an entrywise constant matrix. Forming the Gram matrix and multiplying $Xp_4(G_B)$ cost $O(nr^2)$; a dense eigendecomposition of S_B and fixed-degree polynomial evaluation cost $O(r^3)$. Thus an exact forward evaluation has $O(nr^2 + r^3)$ arithmetic cost, with $O(r^2)$ Gram and polynomial workspace in addition to the factor-sized tensors. Autodifferentiation requires further saved tensors; these operation counts are not a measurement of training time or peak GPU memory.

The spectral results apply whenever the scale satisfies $s \geq \|B\|_2$ and $s > 0$. If an implementation instead uses a norm estimate that underestimates $\|B\|_2$, some normalized singular values can exceed one, and the certificate proved here cannot be invoked without verifying its domain. The source fitting interval $[0, 1.1]$ alone does not supply such a certificate: an extension to normalized singular values up to 1.1 would require checking the scalar inequalities on $x \in (0, 1.1]$, equivalently polynomial arguments $z = x^2 \in (0, 1.21]$. The exact normalization specified in the main text avoids this additional requirement. The denominator in $X = B/s$ and the outer $\mathrm{sg}(s)$ are distinct gradient paths; detaching both is a different backward rule from the written parameterization. Nondifferentiability at repeated leading singular values or the scale-floor boundary does not alter the forward identities, but the proofs do not specify an optimizer's behavior at those points.

At the end of a task, merge the *conditioned* factor: $W^t = W^{t-1} + \widehat{B}_t A_t$. The merged layer has the same forward output as the adapted layer at that parameter value. No separate polynomial evaluation is required during inference after merging.

D.2 Reported benchmark configuration and diagnostic definitions

The main-text experiments use an ImageNet-21K-pretrained ViT-B/16 and a cosine classifier, with rank $r = 32$, protected dimension $K = 64$, and protection strength $\alpha = 0.5$ unless varied in an ablation. Training uses SGD for 20 epochs per task, batch size 48, and cosine learning-rate annealing. Initial learning rates are 0.02 for ImageNet-R, 0.005 for CUB, and 0.01 for the remaining benchmarks. The reported hardware is a single NVIDIA RTX 4090, and the three random seeds are 1993, 1996, and 1997. The evaluated sequences comprise ImageNet-R with 5, 10, and 20 tasks; ImageNet-A with 10 and 20 tasks; and CIFAR-100, CUB, and DomainNet with 10 tasks each. The DomainNet protocol retains the 200 classes with the largest sample counts, as described in the main text. These are eight settings across five datasets.

Let a_t denote the evaluation accuracy over all seen classes after task t, using a classifier that does not receive the test task identity. The reported aggregate metrics are

$$\mathcal{A}_{\mathrm{Last}} = a_T, \qquad \mathcal{A}_{\mathrm{Avg}} = \frac{1}{T}\sum_{t=1}^{T} a_t.$$

A difference in either metric does not by itself distinguish initial acquisition, subsequent forgetting, and competition from new classes. Similarly, the highest sample mean does not establish statistical significance. The component ablations establish the performance of the evaluated variants; they do not isolate all changes in scale and optimization introduced by PSB.

For interpreting spectral measurements, three different objects must be distinguished: the raw factor B_t, the conditioned single-task update $\widehat{B}_t A_t$, and the cumulative update $\sum_{i=1}^{t} \widehat{B}_i A_i$. Lemma B.1 identifies the spectrum of $\widehat{B}_t A_t$ with that of $\widehat{B}_t$, but it does not identify the spectrum of the cumulative sum with any one factor. Likewise, the condition-number theorem compares B_t and its own forward transformation at fixed parameters; it is not a comparison between factors learned in two separate training runs. Dataset, layer, projection matrix, checkpoint, aggregation rule, and numerical-rank threshold are needed to interpret such plots.

Table 6: Ablation of spectral shaping, learnable scaling, and gradient paths on 20S-ImageNetA with EPSP enabled. Results are reported as mean ± standard deviation over three random seeds.

ID	Variant	Effective factor	$A_{\mathrm{Last}} \uparrow$	$A_{\mathrm{Avg}} \uparrow$
1	Without PSB	B	51.92 ± 0.66	62.67 ± 0.51
2	Learnable scale only	γB	51.72 ± 0.34	62.34 ± 0.23
3	Norm-matched PSB	$\gamma \mathcal{N}(B)$	53.86 ± 0.45	64.12 ± 0.56
4	PSB without learnable scale	$\mathrm{sg}(s) X p_4(X^\top X)$	54.45 ± 0.34	65.09 ± 0.45
5	PSB with full scale gradients	$\gamma s X p_4(X^\top X)$	54.06 ± 0.56	65.34 ± 0.25
6	PSB with detached normalization	$\gamma\, \mathrm{sg}(s) X_{\mathrm{det}} p_4(X_{\mathrm{det}}^\top X_{\mathrm{det}})$	54.67 ± 0.27	65.18 ± 0.57
7	Full PSB	$\gamma\, \mathrm{sg}(s) X p_4(X^\top X)$	55.02 ± 1.71	65.61 ± 0.64

Disentangling spectral shaping, scaling, and gradient paths. Table 6 examines the contributions of PSB with EPSP fixed on 20S-ImageNetA. Full PSB improves A_{Last} and A_{Avg} over the variant without PSB by 3.10 and 2.94 percentage points, respectively. Learnable scaling alone provides no improvement in mean accuracy, whereas norm-matched PSB outperforms the scale-only variant by 2.14 and 1.78 points. This comparison supports the value of spectral conditioning beyond learnable scaling and forward norm amplification, although it does not isolate all optimization effects. Removing learnable scaling or modifying the scale-gradient paths yields smaller differences from full PSB: 0.35–0.96 points in A_{Last} and 0.27–0.52 points in A_{Avg}. Full PSB achieves the highest mean accuracy, but its larger final-accuracy variability warrants caution in interpreting these smaller gains. Overall, the results support spectral conditioning as a useful component, without establishing that every scaling or gradient-path choice is individually necessary.