Catastrophic Forgetting is Low-Rank: A Function-Space Theory for Continual Adaptation

Anonymous Authors

Abstract

Catastrophic forgetting is the defining obstacle to continual adaptation: it disrupts instruction-tuning and alignment maintenance in large language models, destabilizes domain-incremental vision systems under distribution shift, and limits the reuse of pretrained backbones across downstream tasks. Despite a decade of mitigation strategies such as parameter regularization, replay, gradient projection or functional distillation, none explains *why* forgetting occurs or identifies the mechanism that produces it. We propose a function-space account in the Neural Tangent Kernel regime that derives the mechanism of forgetting mathematically: new-task training induces interference along a structured, low-rank subspace of output space defined by the cross-task NTK. A closed-form predictor identifies the forgetting vector - direction and magnitude jointly, with cosine similarity indistinguishable from 1 on transformer backbones in the PEFT-CL regime. The theory further shows that forgetting concentrates in only 1-6 NTK eigenmodes and yields a Kronecker scaling rule for the vulnerable rank under frozen heads. These structural results explain why parameter-space methods fail on shared-head benchmarks and motivate a targeted spectral regularizer.

1. Introduction

Continual adaptation fine-tuning, alignment maintenance, domain and distribution shift is bottlenecked by catastrophic forgetting. The problem has gained renewed urgency as adaptation increasingly targets large pretrained foundation models: recent empirical work documents substantial forgetting during continual instruction tuning of LLMs across the 1B-7B scale (Luo et al., 2024; Wang et al., 2024), and an entire subfield of parameter-efficient continual learning has emerged to adapt frozen backbones to new tasks without forgetting (Wang et al., 2022b;a; Smith et al., 2023; Liang & Li, 2024). Despite a decade of mitigation strategies spanning parameter regularization (Kirkpatrick et al., 2017; Zenke et al., 2017), replay (Rolnick et al., 2019; Buzzega et al., 2020), knowledge distillation (Li & Hoiem, 2017), gradient projection (Farajtabar et al., 2020; Saha et al., 2021), and prompt- or adapter-based PEFT methods, the field still lacks a *mechanistic* account of forgetting. The closest theoretical prior work (Doan et al., 2021; Bennani et al., 2020) introduces the cross-task NTK overlap matrix and bounds forgetting magnitude through task alignment, but stops at magnitude bounds: it does not identify the function-space directions along which drift concentrates, nor the structural reasons some directions are vulnerable while others are not.

This paper targets the theoretical foundations of continual adaptation. Before deciding whether adaptation should occur through PEFT, full fine-tuning, replay, RAG, or functional distillation, one needs to identify the output-space directions along which adaptation induces interference. We supply such a diagnostic - by characterizing forgetting as NTK interference in function space, identifying its low-rank structure, and predicting the forgetting vector in closed form before new-task training begins. The theory is meant as a foundation that future methods, including current PEFT-CL methods, can be analyzed against rather than replaced by.

Contributions.

1. **A closed-form forgetting predictor** (Proposition 1, with full proof) that identifies the forgetting vector direction and magnitude jointly from pre-training-time quantities, with cosine similarity indistinguishable from 1 at float32 precision on transformer backbones and 0.994 on ResNet-18 in the PEFT-CL regime (Section 3).

2. **A structural characterization of the vulnerable subspace**: 50-90% of forgetting energy concentrates in 1-6 NTK eigenmodes, a Kronecker factorization $K_{AA} = I_C \otimes G$ under frozen heads yields a derivable scaling rule $k^\star \approx C \cdot k_G$ for the vulnerable rank (Remark 1).

3. **A diagnostic lens for method design**: the theory explains why parameter-space methods fail on shared-head benchmarks and why targeted and broad function-space methods converge under the scaling rule. Spectral regularization is offered as an instrument of the theory, not a performance claim (Sections 4, 5).

2. Related Work

NTK analyses of continual learning. Doan et al. (2021) introduce the NTK *overlap matrix* K_{AB} and show forgetting scales with task alignment, Bennani et al. (2020) derive generalization bounds for SGD and OGD (Farajtabar et al., 2020) under NTK linearization, Rathin Chandra (2025) use NTK-spectrum analysis for path-based plasticity bounds. These works bound forgetting *magnitude*. Proposition 1 delivers the forgetting *vector* in closed form, Section 3 identifies the low-rank eigenstructure of K_{AA}, and Remark 1 derives the factorization $K_{AA} = I_C \otimes G$ - none of which appear in prior NTK-CL theory.

PEFT-based continual learning. Recent empirical progress on pretrained-backbone CL is driven by parameter-efficient adaptation: prompt pools (Wang et al., 2022b;a), decomposed attention prompting (Smith et al., 2023), and low-rank adapters explicitly motivated by interference mitigation (Liang & Li, 2024). These are method proposals engineering around forgetting, they do not characterize the structure of the interference they mitigate. Our eigenmode concentration and Kronecker factorization are properties of K_{AA} under any frozen-head PEFT-CL setup, providing a principled diagnostic applicable to all of them.

Functional regularization. Spectral regularization belongs to the functional-regularization family (Benjamin et al., 2019; Titsias et al., 2020; Li & Hoiem, 2017): LwF and FRCL are *broad*, distilling all output directions. We concentrate the penalty on the NTK-identified vulnerable subspace, yielding 75:1 targeting (Section 5.2) - though the two approaches converge under the scaling rule (Section 5.3).

Parameter-space methods. GPM (Saha et al., 2021), OGD (Farajtabar et al., 2020), EWC (Kirkpatrick et al., 2017), and SI (Zenke et al., 2017) constrain parameter drift. Section 5.1 shows all four fail on shared-head benchmarks, consistent with our claim that vulnerable directions live in output space, not parameter space.

Concurrent mechanistic analyses. Imanov (2026) decompose LLM forgetting into attention-head interference, representational drift, and loss-landscape flattening - a *parameter-space* decomposition over architectural components. Ours is a *function-space* decomposition over NTK eigenmodes. The two are complementary: they identify which components change, we identify which output directions drift.

3. Theory: Forgetting as NTK Interference

3.1. Setup

We consider two-task continual regression. A model trained on Task A reaches parameters $\theta_A \in \mathbb{R}^p$, training then proceeds on Task B from θ_A, producing θ_B. *Forgetting* is the induced shift in Task A predictions:

$$\Delta f_A := f_A(\theta_B) - f_A(\theta_A) \in \mathbb{R}^{n_A d}. \tag{1}$$

We aim to predict Δf_A-direction and magnitude-without running Task B training.

Notation. Let $f(\cdot, \theta) : \mathcal{X} \to \mathbb{R}^d$ with $\theta \in \mathbb{R}^p$. Given probe set $X_A = \{x_i^A\}_{i=1}^{n_A}$ and training set $(X_B, y_B) = \{(x_i^B, y_i^B)\}_{i=1}^{n_B}$, stack outputs as $f_A(\theta) \in \mathbb{R}^{n_A d}$, $f_B(\theta)$, $y_B \in \mathbb{R}^{n_B d}$. All Jacobians and kernels are evaluated at θ_A:

$$J_A := \nabla_\theta f_A(\theta)\big|_{\theta_A}, \quad J_B := \nabla_\theta f_B(\theta)\big|_{\theta_A},$$
$$K_{AA} = J_A J_A^\top, \quad K_{BB} = J_B J_B^\top, \quad K_{AB} = J_A J_B^\top.$$

Crucially, J_B uses Task B inputs but Task A weights-this is why the predictor is computable before Task B training.

Task B objective. MSE with L_2 penalty on drift from θ_A:

$$\mathcal{L}_B(\theta) = \tfrac{1}{2}\|f_B(\theta) - y_B\|^2 + \tfrac{\lambda}{2}\|\theta - \theta_A\|^2, \quad \lambda \geq 0. \tag{2}$$

$\lambda = 0$ recovers vanilla MSE (interpreted as the minimum-$\|\delta\|$ interpolant under gradient flow from θ_A), $\lambda > 0$ ensures a unique minimizer. We take $\theta_B \in \arg\min \mathcal{L}_B$.

3.2. The Predictor

Proposition 1 (Forgetting Predictor)**.** *Under NTK linearization of f around θ_A and convergence of Task B training to a minimizer of $\mathcal{L}_B$,*

$$\Delta f_A = -K_{AB}\,(K_{BB} + \lambda I)^{-1}\, r_B, \tag{3}$$

where $r_B := f_B(\theta_A) - y_B$ is the Task B residual at θ_A, and the inverse is the Moore-Penrose pseudoinverse when $\lambda = 0$.

Proof. Let $\delta := \theta - \theta_A$.

Step 1: Linearize around θ_A.

$$f_B(\theta_A + \delta) \approx f_B(\theta_A) + J_B\delta. \tag{4}$$

Step 2: Reduce to quadratic. Substituting (4) into (2):

$$\mathcal{L}_B(\theta_A + \delta) \approx \tfrac{1}{2}\|J_B\delta + r_B\|^2 + \tfrac{\lambda}{2}\|\delta\|^2. \tag{5}$$

Convex in δ, with a unique minimizer (minimum-norm for $\lambda = 0$).

Step 3: First-order condition. Setting $\nabla_\delta \mathcal{L}_B = 0$ yields the ridge normal equations $(J_B^\top J_B + \lambda I_p)\,\delta = -J_B^\top r_B$, hence

$$\delta^\star = -(J_B^\top J_B + \lambda I_p)^{-1} J_B^\top\, r_B. \tag{6}$$

Step 4: Push-through identity. For any $M \in \mathbb{R}^{m\times p}$,

$$(M^\top M + \lambda I_p)^{-1} M^\top \;=\; M^\top (M M^\top + \lambda I_m)^{-1}, \quad (7)$$

Applied to (6) with $M = J_B$:

$$\delta^\star \;=\; -J_B^\top \left(K_{BB} + \lambda I\right)^{-1} r_B. \quad (8)$$

The inverse is now $n_B d \times n_B d$ rather than $p \times p$ - the tractable dual form.

Conclusion. Applying the linearization to Task A:

$$\begin{aligned}\Delta f_A \;&\approx\; J_A\, \delta^\star \\ &=\; -J_A J_B^\top \left(K_{BB} + \lambda I\right)^{-1} r_B \\ &=\; -K_{AB} \left(K_{BB} + \lambda I\right)^{-1} r_B. \qquad \square\end{aligned}$$

Empirically, (3) achieves $\cos\mathrm{sim}(\Delta f_A^{\text{pred}}, \Delta f_A^{\text{real}}) > 0.99$ on Split-MNIST and Split-CIFAR-10 (Fig. 1, left), and is structurally exact under frozen-head PEFT-CL (Table 1).

3.3. Structural Consequences

Proposition 1 delivers Δf_A as the action of the cross-task kernel K_{AB} on the Task-B residual. Three structural properties of this object follow directly and shape the rest of the paper: forgetting lives in a low-rank subspace, linearization is exact under frozen heads, and a Kronecker factorization fixes the vulnerable rank.

Low-rank structure. In the eigenbasis $K_{AA} = U\Lambda U^\top$, write $\Delta f_A = \sum_i c_i u_i$. Whenever Λ is rapidly decaying - as it is for kernels of standard architectures, by spectral bias (Rahaman et al., 2019), the residual-projection coefficients c_i inherit this decay through Eq. (3), and Δf_A concentrates in the top eigenmodes of K_{AA}. We call $\mathrm{span}(u_1, \ldots, u_k)$ the *vulnerable subspace*: the directions in output space along which Task-B training can move predictions on Task A, ordered by how much it is geometrically allowed to do so. The complement is protected by construction — drift there requires alignment between K_{AB} and a low-eigenvalue mode of K_{BB}, which the predictor suppresses by the $(K_{BB} + \lambda I)^{-1}$ factor. Section 5 measures k empirically and finds 1–6 modes carry 50–90% of forgetting energy.

Exact linearization under a linear probe. When f is linear in θ - as in linear probing on a frozen backbone - the Taylor expansion in Step 1 of the proof is an equality, every $\approx$ becomes $=$, and Proposition 1 holds without approximation. This is the regime where the predictor can be checked at machine precision, we do so in Section 5.

Remark 1 (Kronecker structure and k-scaling)**.** *For a linear head $W \in \mathbb{R}^{C\times F}$ on features $\phi(x) \in \mathbb{R}^F$, each output f_c depends only on row W_c, so the Jacobian is block-diagonal in output index:*

$$K_{AA} \;=\; I_C \otimes G, \qquad G_{ij} = \phi(x_i)^\top \phi(x_j). \quad (9)$$

Every eigenvalue of G has multiplicity C in K_{AA}, under MSE the rows evolve independently, yielding $k^\star \approx C \cdot k_G$ with $k_G \in [1, 5]$ empirically (CE couples rows and breaks this, App. A). Thus C=10 needs $k \in [10, 50]$, C=100 needs $k \approx 100$.

Table 1. Predictor precision across frozen backbones (Split-CIFAR-100, MSE, 3 seeds). Values are $1 - \cos\mathrm{sim}$, lower is sharper. ResNet-18's gap is an SGD-convergence artifact on a less well-conditioned feature Gram.

Backbone	10 tasks	20 tasks
ResNet-18	6.4 ± 1.2 (e-3)	6.1 ± 1.4 (e-3)
ViT-B/16	6 ± 4 (e-6)	35 ± 35 (e-6)
DINOv2	< 1 (e-6)	< 1 (e-6)

4. A Theory-Derived Probe: Spectral Regularization

Since forgetting concentrates in the top-k eigenspace of K_{AA}, we penalize drift specifically there. After Task τ, compute top-k eigenvectors $\{u_j^{(\tau)}\}$ of $K_{\tau\tau}$ on n_{probe} probes and store $f_\tau^{\text{ref}} = f_\tau(\theta_\tau)$, during subsequent training,

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{new}}(\theta) + \sum_{\tau<t} \frac{\mu}{k} \sum_{j=1}^{k} \left(u_j^{(\tau)\top} \left[f_\tau(\theta) - f_\tau^{\text{ref}} \right] \right)^2. \quad (10)$$

Drift in the $(nd{-}k)$-dimensional complement is *unconstrained*, granting full plasticity outside the vulnerable subspace in contrast to EWC (constrains all p parameter directions) and LwF (constrains all nd output directions).

5. Experiments

Setup. Three benchmarks: Split-MNIST (5 tasks, MLP) and Split-CIFAR-10 (5 tasks, ∼200k-parameter CNN), each in shared-head (single output layer over union of classes - the harder regime, since new-task training overwrites old-task logits) and multi-head (per-task heads, task identity given at inference) variants; Split-CIFAR-100 with a frozen ImageNet ResNet-18 (10 tasks, PEFT-CL, where the linearization is exact). Baselines: parameter regularization (EWC (Kirkpatrick et al., 2017), SI (Zenke et al., 2017)), gradient projection (GPM (Saha et al., 2021)), functional distillation (LwF (Li & Hoiem, 2017)), replay (random, DER++ (Buzzega et al., 2020)), plus no-reg control. We report final-task average accuracy and *forgetting* (average drop in per-task accuracy from when learned to end of training);

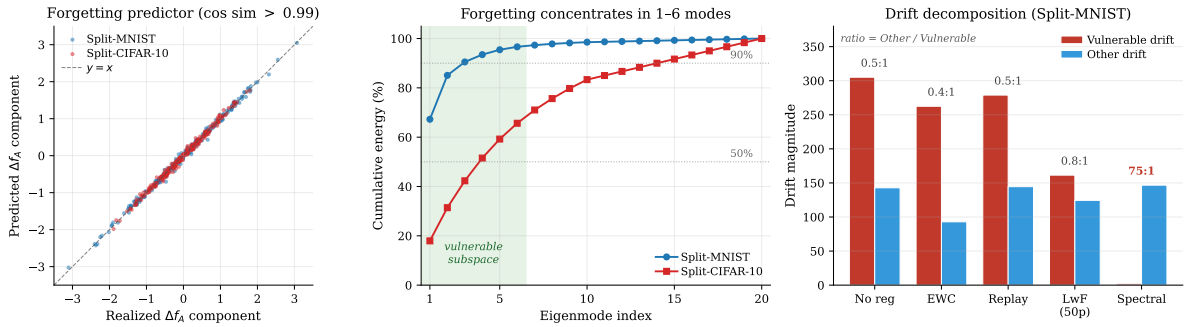

Figure 1. **Left**: Predicted vs. realized Δf_A (cos sim > 0.99) on Split-MNIST/CIFAR-10. **Center**: Cumulative forgetting energy-50-90% in 1-6 eigenmodes. **Right**: Drift decomposition-spectral reg achieves 75:1 targeting vs. $<$ 1:1 for baselines.

Table 2. Shared-head results. Parameter-space methods fail, function-space methods succeed. Std shown only for top two methods (5/10 seeds for MNIST/CIFAR-10). Spectral beats LwF on CIFAR-10 (p=0.002).

	Split-MNIST		Split-CIFAR-10	
Method	Acc ↑	Fgt ↓	Acc ↑	Fgt ↓
No reg	19.7	99.6	17.5	85.8
EWC	19.8	99.5	18.7	88.7
SI	22.2	96.4	17.8	85.5
Replay	56.7	53.3	21.5	82.7
DER++	65.9	42.0	28.4	78.6
LwF	**80.3**±1.4	23.9±1.8	29.0±2.0	77.4±2.4
Spectral	77.9±3.0	**11.3**±0.5	**32.0**±2.1	**51.6**±5.1

mean±std over 5–10 seeds. Apps. D–G cover multi-head, probe scaling, sensitivity.

5.1. Parameter-Space Methods Fail on Shared-Head

Table 2 shows a clean divide: parameter-space methods (EWC, SI) match no-reg on both benchmarks while function-space methods succeed. The mechanism is structural, forgetting concentrates in $k{\approx}10$ output directions, but EWC's diagonal Fisher spreads regularization across $\sim$200k parameters, no λ both protects the vulnerable subspace and preserves plasticity. The failure is not specific to diagonal Fisher: GPM (Saha et al., 2021), a non-diagonal method that projects gradients orthogonal to the top-energy subspace of old-task gradients, achieves 9.7–10.1% on Split-CIFAR-100 across energy thresholds, matching EWC (9.9%) and dominated by spectral reg (39.9%) and LwF (39.8%). Gradient-energy bases simply do not align with output-interference directions - which function-space methods target by construction.

5.2. Drift Decomposition: Targeted Protection

Decomposing Δf_A into drift within the vulnerable subspace (top-k eigenmodes of K_{AA}) and drift outside isolates the targeting mechanism (Table 3). LwF reduces both components proportionally, confirming broad but untargeted protection - exactly the structural distinction the theory predicts.

Table 3. **Direct test of the low-rank claim.** Drift decomposition after 5 tasks on Split-MNIST (shared-head, k=10). Spectral reg suppresses vulnerable-subspace drift 150× while leaving the complement untouched. CIFAR-10 in App. E.

Method	Vuln. ↓	Other	Ratio
No reg	305.1	142.8	0.5:1
EWC (best)	262.5	92.9	0.4:1
Replay (100)	278.9	144.5	0.5:1
LwF (50p)	161.4	124.4	0.8:1
Spectral (μ=10)	**2.0**	146.7	**75:1**

5.3. PEFT-CL: Function-Space Methods Converge Under Scaling Rule

On Split-CIFAR-100 (C=100, 20 probes), spectral reg at k=100 and LwF are statistically indistinguishable (39.9±1.5% vs 39.8±1.2%): matched at $k^\star \approx C \cdot k_G$, targeted and broad function-space methods converge. Outside this matched regime they trade off as the theory predicts: targeted regularization wins at low probe count, broad protection wins at high probe count (App. F).

6. Conclusion

Catastrophic forgetting has long been treated as a pathology of optimization - something that happens, gets measured, and gets suppressed. We have argued it is something more tractable: a geometric phenomenon, concentrated in a handful of NTK eigenmodes, whose direction and magnitude can be written down in closed form before any gradient step on the new task. The Kronecker factorization $K_{AA} = I_C \otimes G$ says plainly what the vulnerable rank must be, the 75:1 drift asymmetry under spectral regularization says the theory points at the right subspace. A full account will need to handle compound drift across long sequences, feature learning beyond the linearized regime, and the cross-entropy case. But treating forgetting as interference with exploitable structure is, we think, the right starting point.

References

Benjamin, A. S., Rolnick, D., and Kording, K. Measuring and regularizing networks in function space. *ICLR*, 2019.

Bennani, M. A., Doan, T., and Sugiyama, M. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 2020.

Doan, T., Abbana Bennani, M., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. *AISTATS*, 2021.

Imanov, O. Y. L. Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2601.18699*, 2026.

Liang, Y.-S. and Li, W.-J. InfLoRA: Interference-free low-rank adaptation for continual learning. *CVPR*, pp. 23638-23647, 2024.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2024.

Path-coordinated continual learning with neural tangent kernel-justified plasticity. *arXiv preprint arXiv:2511.02025*, 2025.

Smith, J. S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., and Kira, Z. CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning. *CVPR*, 2023.

Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.-Y., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. DualPrompt: Complementary prompting for rehearsal-free continual learning. *ECCV*, 2022.

Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. Learning to prompt for continual learning. *CVPR*, 2022.

Wang, H., Lu, H., Yao, L., and Gong, D. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Farajtabar, M., Azizan, N., Mott, A., and Li, A. Orthogonal gradient descent for continual learning. *AISTATS*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*, 2018.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521-3526, 2017.

Li, Z. and Hoiem, D. Learning without forgetting. *TPAMI*, 40(12):2935-2947, 2017.

Rahaman, N., Baratin, A., Arpit, D., et al. On the spectral bias of neural networks. *ICML*, 2019.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *NeurIPS*, 2019.

Saha, G., Garg, I., and Roy, K. Gradient projection memory for continual learning. *ICLR*, 2021.

Titsias, M. K., Schwarz, J., Matthews, A. G., Pascanu, R., and Teh, Y. W. Functional regularisation for continual learning using Gaussian processes. *ICLR*, 2020.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. *ICML*, 2017.

A. Limitations

NTK regime and feature drift. The predictor is exact under frozen-head PEFT-CL and achieves > 0.99 cosine similarity on MLPs and small CNNs trained from scratch. For full-network training on deeper architectures, the last-layer Jacobian approximation cannot see backbone feature drift, and K_{AA} itself becomes time-varying during training with eigenstructure that evolves as representations do. Extension via gradient sketching, block-diagonal approximations, or an online estimate of K_{AA} tracked over training is future work.

Cross-entropy breaks strict block-decoupling. Remark 1's factorization $K_{AA} = I_C \otimes G$ relies on MSE loss, under which output rows are gradient-decoupled. Softmax cross-entropy couples rows through the normalization, so the factorization holds only approximately. The $k^\star \approx C \cdot k_G$ rule remains a useful design heuristic under CE but may require modest empirical adjustment consistent with the soft plateau we observe on MNIST at $k \in [10, 50]$ rather than a sharp optimum (Appendix G).

B. Subspace Stability

We measure whether the vulnerable subspace $\mathrm{span}(u_1, \ldots, u_k)$ remains valid as the model trains on subsequent tasks, via principal angles between U_k at θ_A (after Task 0) and U_k recomputed at each subsequent task boundary on the *same* Task-0 probes. On Split-MNIST, the mean principal angle plateaus at 25.9 after 4 task transitions the bulk geometry is preserved, though individual directions may rotate up to 83. On Split-CIFAR-10, rotation is faster (mean 34.0), quantitatively explaining why spectral regularization gains are smaller on CNNs than MLPs. The gap between mean ($\sim$25-35) and max ($\sim$80-90) angles reveals anisotropic rotation: a few eigendirections rotate substantially while the majority remain stable.

C. Cross-Task Coupling Predicts Forgetting Magnitude

The predictor identifies the *direction* of forgetting, the Frobenius norm of K_{AB} also enables *magnitude forecasting* across task pairs. Across all 10 task pairs on Split-MNIST (3 seeds, 30 measurements), $\|K_{AB}\|_F$ strongly predicts realized forgetting magnitude (Spearman $\rho = 0.88$, $p < 10^{-10}$): before training on any new task, one can cheaply estimate which prior tasks will suffer the most damage. On CIFAR-10 the correlation is weaker ($\rho = 0.36$, $p = 0.053$), consistent with stronger NTK regime violations on CNNs.

D. Multi-Head Evaluation

Task-specific heads eliminate the cross-output interference that makes shared-head catastrophic. EWC recovers from 19.8% (shared) to 88.4% (multi) on CIFAR, confirming that its shared-head failure arises from output-layer interference, not a fundamental flaw in parameter-space regularization. Spectral reg is competitive on MNIST but no longer leads on multi-head CIFAR - exactly what Remark 1 predicts: the C-fold eigenvalue multiplicity does not apply per-head, so the scaling-rule advantage that drives shared-head dominance disappears. The result confirms the theory's scope rather than contradicting it.

Table 4. Multi-head evaluation. Task-specific heads eliminate shared-head output interference: EWC recovers on CIFAR, and LwF leads.

	MNIST Multi-Head		CIFAR Multi-Head	
Method	Acc (%) ↑	Fgt (%) ↓	Acc (%) ↑	Fgt (%) ↓
No reg	92.2±2.3	9.4±2.8	79.2±2.2	16.7±2.7
SI	97.3±0.8	3.0±0.9	83.1±4.0	11.1±4.8
Spectral k=10	99.1±0.2	0.5±0.2	81.1±1.1	10.4±0.9
Spectral k=50	**99.4**±0.1	**0.4**±0.1	86.3±1.5	7.0±1.9
EWC	98.4±0.7	1.7±0.9	88.4±0.6	**1.2**±0.4
LwF (50p)	99.3±0.1	0.5±0.1	**89.2**±1.2	4.6±1.0

E. CIFAR-10 Drift Decomposition

Table 5. Drift decomposition after 5 tasks on Split-CIFAR-10 (shared-head, k=10).

Method	Vuln. ↓	Other	Ratio
No reg	146.6	106.1	0.7:1
EWC (best)	425.4	263.3	0.6:1
Replay (200)	302.1	203.7	0.7:1
LwF (50p)	139.5	76.6	0.5:1
Spectral (μ=10)	**30.8**	52.6	**1.7:1**

Targeting is weaker on CIFAR-10 (ratio 1.7:1) than MNIST (75:1), consistent with the faster subspace rotation on CNNs (Appendix B). Spectral reg remains the only method where other drift exceeds vulnerable drift.

F. Probe Scaling

Table 6. Probe scaling on Split-MNIST. Spectral reg extracts more anti-forgetting signal per probe at $\leq$100 probes, LwF overtakes at 200.

Probes	Spectral k=50	LwF
20	**71.8**±2.4	67.1
50	**84.6**±1.2	80.3±1.4
100	**87.9**±0.7	87.2
200	88.3±0.4	**92.4**

At low probe count, KL divergence is diluted across all output dimensions each direction receives a weak supervisory signal. Eigenmode projection concentrates the penalty on k directions, extracting more anti-forgetting value per stored point. At high probe count, LwF's broad coverage pays off: per-direction signal becomes strong across all dimensions. Targeted regularization wins at low memory, broad regularization wins at high memory.

G. Sensitivity Analysis

Table 7. Sensitivity to k (Split-MNIST, μ=1, C=10). The plateau at $k \in [10, 50]$ matches Remark 1: $k^\star \approx C \cdot k_G$ with $k_G \in [1, 5]$ predicts $k^\star \in [10, 50]$.

k	1	5	10	20	50	100
Acc (%)	43.2±7.8	72.7±5.8	79.6±1.2	82.1±0.7	**84.6**±0.5	83.2±0.2

μ-sensitivity: MNIST optimal μ=10, CIFAR optimal μ=1 (CNNs need more plasticity). The range $\mu \in [1, 10]$ is robust. L2 projection is the natural loss for eigenmode-specific regularization.