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

Ido Nitzan Hidekel [1] **Dan Raviv** [1]

Abstract

Catastrophic forgetting in continual adaptation is usually studied through parameter drift, replay, or distillation, but these views do not identify which output-space directions are vulnerable. We give a function-space account in the NTK regime: new-task training induces old-task prediction drift through the cross-task kernel, yielding a closed-form predictor for the forgetting vector before any new-task gradient step. In frozen-backbone linear-head PEFT-CL, where the model is linear in the trainable parameters, the predictor is exact up to numerical precision; for nonlinear adapters/full fine-tuning it is a local NTK approximation. The same expression reveals that forgetting concentrates in a small number of old-task NTK eigenmodes, and under frozen linear heads gives a Kronecker scaling rule for the vulnerable rank. These results clarify the relation to prior NTK-overlap theory, explain why parameter-space regularizers can miss output-space interference, and motivate a targeted spectral regularizer.

1. Introduction

Continual adaptation—fine-tuning, alignment maintenance, and adaptation under domain or distribution shift—is bottlenecked by catastrophic forgetting. The problem has gained renewed urgency as adaptation increasingly targets large pretrained models: recent work documents substantial forgetting during continual instruction tuning of LLMs (Luo et al., 2024; Wang et al., 2024), and a subfield of parameter-efficient continual learning has emerged to adapt frozen backbones 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 emphasizes scalar magnitude/risk control rather than the eigenspace structure of the realized forgetting vector: it does not identify which output directions drift, nor why some directions are vulnerable while others are not.

This paper targets the theoretical foundations of continual adaptation. We study forgetting as NTK interference in function space. The goal is not to replace PEFT-CL methods, but to identify the output-space directions along which adaptation induces drift. Our main experiments use the analytically exact linear limit: a frozen pretrained backbone with a task-shared trainable linear head. More general frozen-backbone modules such as adapters or LoRA are covered by the same local NTK linearization, but the predictor is then approximate.

Contributions.

1. **A closed-form forgetting predictor** (Proposition 1, with full proof) that identifies the forgetting vector direction and magnitude jointly from quantities computable before any new-task gradient step. The predicted shift achieves cosine similarity indistinguishable from 1 with the realized shift, at float32 precision, on transformer backbones under frozen-backbone linear-head adaptation, and 0.994 on a frozen ResNet-18 (Section 3).
2. **A structural characterization of the vulnerable subspace**: in the eigenbasis of the Task-A NTK K_{AA}, the forgetting vector concentrates in only a handful of eigenmodes (1–6 modes carry 50–90% of its energy on Split-MNIST/CIFAR-10). Under a frozen backbone with a trainable linear head of C outputs, K_{AA} admits an exact Kronecker factorization that fixes the vulnerable rank to $k^\star \approx C \cdot k_G$ where k_G is the effective rank of the feature Gram (Remark 1).
3. **A diagnostic lens for method design**: the theory ex-

[1]Tel Aviv University, Tel Aviv, Israel. Correspondence to: Ido Nitzan Hidekel <idon@mail.tau.ac.il>, Dan Raviv <darav@tauex.tau.ac.il>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

plains 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; its consistent shared-head gains (Table 1) confirm the diagnostic rather than constituting the central claim (Sections 4, 5).

2. Related Work

NTK analyses of continual learning. Doan et al. (2021) introduce the NTK *overlap matrix* K_{AB} and derive a two-task expression for prediction drift that is closely related to Eq. (3); their analysis emphasizes a scalar magnitude bound governed by task alignment. Bennani et al. (2020) derive generalization bounds for SGD and OGD (Farajtabar et al., 2020) under NTK linearization. Building on this line, our contribution is to read the same two-task expression as a structured operator on output space: we identify the eigenstructure of the old-task kernel that controls which output directions are vulnerable, derive its Kronecker factorization under a frozen linear head, and use the resulting low-rank picture to motivate a targeted spectral regularizer (all formalized in Section 3).

PEFT-based continual learning. Recent pretrained-backbone CL is driven by parameter-efficient adaptation: prompt pools (Wang et al., 2022b;a), decomposed attention prompting (Smith et al., 2023), and interference-motivated low-rank adapters (Liang & Li, 2024). These engineer around forgetting without characterizing the interference they mitigate. Our eigenmode concentration and Kronecker factorization are exact under frozen-backbone linear-head adaptation, and provide a diagnostic lens for broader PEFT-CL methods when treated through local NTK linearization.

Functional regularization. Spectral regularization belongs to the functional-regularization family (Benjamin et al., 2019; Titsias et al., 2020; Li & Hoiem, 2017): LwF and FRCL distill drift across all output directions, whereas we concentrate the penalty on the NTK-identified vulnerable subspace. Sections 5.2–5.3 show targeted and broad approaches converge under the scaling rule of Remark 1.

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 over architectural components (a *parameter-space* view); ours is a complementary *function-space* decomposition over NTK eigenmodes – which output directions drift, rather than which components change.

3. Theory: Forgetting as NTK Interference

3.1. Setup

We consider two-task continual regression: train on Task A to $\theta_A \in \mathbb{R}^p$, then on Task B from θ_A to θ_B. *Forgetting* is the induced shift in Task-A predictions, written as a flat vector:

$$\Delta f_A \;:=\; \mathrm{vec}[\, f_A(\theta_B) - f_A(\theta_A)\,] \;\in\; \mathbb{R}^{n_A d}, \qquad (1)$$

where $\mathrm{vec}[\cdot]$ stacks the n_A rows of the $n_A \times d$ output-shift matrix into a single column. 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:

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

$\lambda = 0$ 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 (Jacot et al., 2018) and convergence of Task B training to a minimizer of $\mathcal{L}_B$,*

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

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 sketch. With $\delta := \theta - \theta_A$, linearizing $f_B(\theta_A + \delta) \approx f_B(\theta_A) + J_B\delta$ turns $\mathcal{L}_B$ into the convex quadratic $\frac{1}{2}\|J_B\delta + r_B\|^2 + \frac{\lambda}{2}\|\delta\|^2$, whose minimizer solves the ridge normal equations $(J_B^\top J_B + \lambda I)\delta = -J_B^\top r_B$. The push-through identity $(M^\top M + \lambda I)^{-1}M^\top = M^\top(MM^\top + \lambda I)^{-1}$ with $M = J_B$ rewrites this in the tractable dual form $\delta^\star = -J_B^\top(K_{BB} + \lambda I)^{-1} r_B$. Applying the Task-A linearization, $\Delta f_A \approx J_A\delta^\star = -K_{AB}(K_{BB} + \lambda I)^{-1} r_B$. Full step-by-step derivation in App. A. □

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

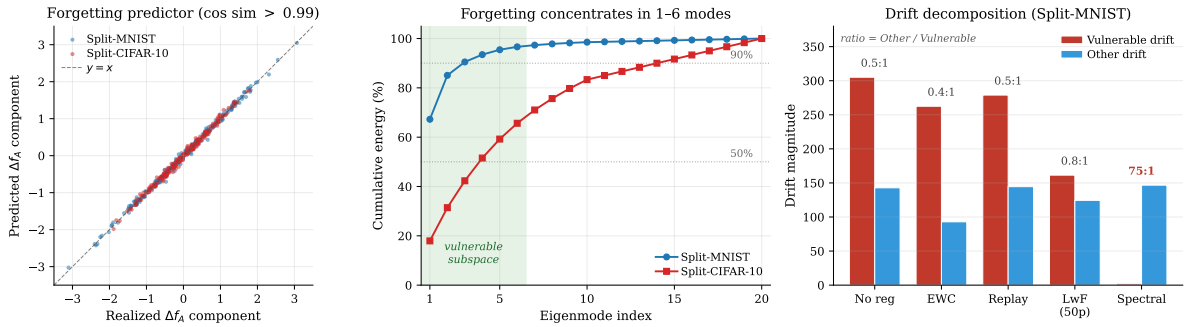

Figure 1. **Left**: Predicted vs. realized Δf_A (cos sim > 0.99) on Split-MNIST/CIFAR-10. **Center**: Cumulative forgetting energy: 50–90% in 1–6 eigenmodes. **Right**: Drift decomposition — spectral reg targets the vulnerable subspace at 75:1 on Split-MNIST (1.7:1 on the CNN-based Split-CIFAR-10, App. L) vs. $<$1:1 for baselines.

structurally exact in the frozen-backbone linear-head PEFT-CL regime, with $1 - \cos\mathrm{sim}$ down to 10^{-6} on ViT-B/16 and DINOv2 (App. C).

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 follow directly and shape the rest of the paper: forgetting lives in a low-rank subspace, linearization is exact under a frozen backbone with a trainable linear head, and a Kronecker factorization fixes the vulnerable rank.

Low-rank structure. The predictor in Eq. (3) factors as

$$\Delta f_A \;=\; -\,J_A J_B^\top (K_{BB} + \lambda I)^{-1}\, r_B, \tag{4}$$

so Δf_A lies in $\mathrm{Im}(J_A) = \mathrm{Im}(K_{AA})$ (for any real J_A, $\mathrm{Im}(J_A J_A^\top) = \mathrm{Im}(J_A)$). Expanding in the eigenbasis $K_{AA} = U\Lambda U^\top$ as $\Delta f_A = \sum_i c_i u_i$ with $c_i = u_i^\top \Delta f_A$, and substituting the SVD $J_A = U\Sigma V_A^\top$ (so $u_i^\top J_A = \sigma_i v_{A,i}^\top$) into Eq. (4), the coefficient on mode i is

$$c_i \;=\; -\,\sigma_i\, v_{A,i}^\top J_B^\top (K_{BB} + \lambda I)^{-1}\, r_B, \tag{5}$$

so $|c_i|$ inherits the decay of σ_i modulated by the alignment of $v_{A,i}$ with the residual-driven Task-B direction $J_B^\top (K_{BB} + \lambda I)^{-1} r_B$. When Λ decays rapidly – as it does for standard architectures by spectral bias (Rahaman et al., 2019) – and the cross-task alignment factor is not adversarially concentrated on small-σ_i directions, Δf_A is expected to concentrate in the top eigenmodes of K_{AA}. We call $\mathrm{span}(u_1, \ldots, u_k)$ the *vulnerable subspace*: the output directions along which Task-B training can move Task-A predictions. 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 – a frozen backbone with a trainable linear head – the Taylor expansion is an equality and Proposition 1 holds exactly, checkable at machine precision (Section 5). For nonlinear adapters or full fine-tuning the model is not linear in the trainable parameters, and the predictor becomes a local NTK approximation around θ_A.

Is it forgetting that is low-rank, or learning? Spectral bias makes the NTK low-rank in any task (Rahaman et al., 2019); the novel content of Eq. (5) is not that Λ decays but that the *forgetting* coefficient c_i is set by a specific cross-task product $v_{A,i}^\top J_B^\top (K_{BB} + \lambda I)^{-1} r_B$. Two factors compound: spectral-bias decay of σ_i (how Task-A learning is itself low-rank), and a cross-task alignment factor that selects which of those modes Task-B can actually move (App. E).

Remark 1 (Kronecker structure and k-scaling)**.** *For a frozen feature map $\phi(x) \in \mathbb{R}^F$ and trainable linear head $W \in \mathbb{R}^{C\times F}$, with $\theta = \mathrm{vec}(W)$ and $f_c(x) = W_c^\top \phi(x)$, the MSE Jacobian is block-diagonal across output classes. Therefore*

$$K_{AA} \;=\; I_C \otimes G, \qquad G_{ij} = \phi(x_i)^\top \phi(x_j), \tag{6}$$

where $G \in \mathbb{R}^{n_A \times n_A}$ is the (single-output) feature Gram and I_C is the $C \times C$ identity. Consequently every eigenvalue of G has multiplicity C in K_{AA}, and the dimension of the vulnerable subspace scales as

$$k^\star \;\approx\; C \cdot k_G, \tag{7}$$

where k_G is the effective rank of G – the number of dominant eigenvalues of the feature Gram. We observe $k_G \in [1, 5]$ empirically on standard pretrained features (Appendix N). Thus for C=10 outputs $k^\star \in [10, 50]$, and for C=100 outputs $k^\star \in [100, 500]$ – collapsing toward $k^\star \approx 100$ on the near-rank-one frozen CIFAR-100 features ($k_G \approx$ 1–2 measured; App. N), the matched point used in Section 5.3. Under softmax cross-entropy, the normalization couples rows of W, so Eq. (6) holds only approximately and the rule survives as a design heuristic with a softer plateau (Section B). This scaling drives the convergence in Section 5.3: with C large, the LwF-style "broad" penalty already lives mostly inside the $C \cdot k_G$-dimensional vulnerable subspace.

Table 1. **Shared-head results:** parameter-space methods fail, function-space methods succeed. Std shown for the top two methods (5/10 seeds). 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**$\pm$1.4	23.9$\pm$1.8	29.0$\pm$2.0	77.4$\pm$2.4
Spectral	77.9$\pm$3.0	**11.3**$\pm$0.5	**32.0**$\pm$2.1	**51.6**$\pm$5.1

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,

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{new}}(\theta) + \sum_{\tau<t} \frac{\mu}{k} \sum_{j=1}^{k} \left(u_j^{(\tau)\top} \left[f_\tau(\theta) - f_\tau^{\text{ref}} \right] \right)^2 . \tag{8}$$

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

5. Experiments

Setup. Three benchmarks: Split-MNIST (5 tasks, MLP) and Split-CIFAR-10 (5 tasks, ∼200k-param CNN), each in shared-head (single output layer over the union of classes – the harder regime) and multi-head variants; Split-CIFAR-100 with a frozen ImageNet ResNet-18 (10 tasks, frozen-backbone linear-head PEFT-CL regime, where linearization is exact). Baselines span 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)), and a no-reg control. We report final-task accuracy and *forgetting* (mean per-task accuracy drop) over 5–10 seeds. Apps. K–N cover multi-head, probe scaling, and sensitivity.

5.1. Parameter-Space Methods Fail on Shared-Head

Table 1 shows a clean divide: parameter-space methods (EWC, SI) match no-reg while function-space methods improve substantially. The mechanism is geometric – forgetting concentrates in a low-dimensional subspace of *output* space, but diagonal Fisher is anisotropic in parameter coordinates and is not constructed to align with the rank-k output-space projector $U_k U_k^\top$, so EWC can slow drift broadly without preferentially protecting the vulnerable K_{AA} eigenmodes. A single-step drift-decomposition probe confirms this: EWC and no-reg leave indistinguishable Other:Vuln ratios (0.033:1 vs 0.028:1), whereas spectral reg flips it to 32.7:1; EWC scales both components down uniformly but cannot translate per-parameter Fisher mass into selective output-mode protection (App. F). Non-diagonal methods (GPM) fail for the same reason (App. H).

Table 2. **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×. CIFAR-10 in App. L.

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**

5.2. Drift Decomposition: Targeted Protection

Decomposing Δf_A inside vs. outside the vulnerable subspace isolates the targeting mechanism (Table 2): LwF reduces both proportionally, while spectral reg suppresses only the vulnerable component – exactly the structural distinction the theory predicts.

5.3. Function-Space Methods Converge Under the Scaling Rule

On Split-CIFAR-100 (C=100, 20 probes), spectral reg at k=100 and LwF are 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 regime they trade off as predicted (App. M), and spectral protection stabilizes the old-class decision boundary against inter-task *confusion* (App. G).

6. Conclusion

This work reframes catastrophic forgetting as low-rank interference in function space. In the NTK regime, a cross-task kernel expression predicts the forgetting vector before new-task training and identifies a small vulnerable eigenspace of K_{AA}. For frozen-backbone linear heads, the analysis is exact and yields $k^\star \approx Ck_G$; for nonlinear adapters or full fine-tuning, it is a local approximation. Empirically, this explains why parameter-space regularizers can fail on shared-head benchmarks and why spectral function-space regularization protects the relevant modes. Extending the mechanism to cross-entropy, evolving representations, and long task sequences remains open.

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.

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. Proof of Proposition 1

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

Step 1: Linearize around θ_A.

$$f_B(\theta_A + \delta) \approx f_B(\theta_A) + J_B\delta. \tag{9}$$

Step 2: Reduce to quadratic. Substituting (9) into (2):

$$\mathcal{L}_B(\theta_A + \delta) \approx \tfrac{1}{2}\|J_B\delta + r_B\|^2 + \tfrac{\lambda}{2}\|\delta\|^2. \tag{10}$$

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

$$\delta^\star = -(J_B^\top J_B + \lambda I_p)^{-1} J_B^\top r_B. \tag{11}$$

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

$$(M^\top M + \lambda I_p)^{-1} M^\top = M^\top (MM^\top + \lambda I_m)^{-1}, \tag{12}$$

applied to (11) with $M = J_B$:

$$\delta^\star = -J_B^\top (K_{BB} + \lambda I)^{-1} r_B. \tag{13}$$

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:

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

B. Limitations

Scope: where the predictor is exact and where it is approximate. Proposition 1 is derived under four assumptions: (i) NTK linearization of f around θ_A, (ii) convergence of Task-B training to a minimizer of $\mathcal{L}_B$, (iii) MSE loss with an optional ridge anchor, and (iv) trainable parameters restricted to a head (or low-rank adapter) on top of frozen pretrained features (the frozen-backbone linear-head PEFT-CL regime). When all four hold simultaneously – as in frozen-backbone linear probing – the predictor is an exact identity at machine precision (Table 3): it is exact precisely when the model is linear in the trainable parameters (a trainable linear head on a frozen backbone), and is a local NTK approximation around θ_A for nonlinear adapters or full fine-tuning. When the head is trainable but the backbone is frozen and shallow MLPs/CNNs are trained from scratch, the predictor remains tight ($\cos\mathrm{sim} > 0.99$) because the operating regime is close to lazy. Outside these regimes the predictor degrades along three orthogonal axes, which we list explicitly because they delimit the claims of this paper:

- *Full fine-tuning of deep backbones.* The Jacobian J_A evaluated at θ_A does not see feature drift in lower layers, and K_{AA} itself becomes time-varying during Task-B training with eigenstructure that evolves as representations do. The predictor then captures only the linearized component of Δf_A.

- *Continual instruction tuning of LLMs.* The frozen-backbone linear-head PEFT-CL regime assumption (small adapter/head) is closer to LoRA-style instruction tuning than to full fine-tuning, but the MSE assumption is violated: token-level cross-entropy with a softmax over a large vocabulary breaks both the Kronecker factorization of Remark 1 and the closed-form residual r_B.

- *Long task sequences with cumulative drift.* Even when each pairwise step is well predicted, error in K_{AA} compounds across tasks; we measure this drift in App. I but do not give a multi-step predictor.

Extension via gradient sketching, block-diagonal Jacobian approximations, an online estimate of K_{AA} tracked through training, and a softmax-aware analogue of Eq. (3) are natural next steps.

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 (App. N).

Relation to prior NTK-CL theory. The two-task expression in Proposition 1 is closely related to the form derived by Doan et al. (2021): both arise from substituting the NTK linearization into a quadratic loss and applying the push-through identity. Our contribution is not to add a new derivation but to read this expression as a structured operator on output space, exposing the eigenstructure of K_{AA}, the Kronecker factorization in the frozen-backbone linear-head regime, and the resulting scaling rule that ties together targeted and broad function-space methods. The closed-form predictor and the structural-rank picture should be taken as a single package.

C. Predictor Precision Across Backbones

In the frozen-backbone linear-head PEFT-CL regime the NTK linearization is exact, so Proposition 1 holds at machine precision; the residual gap on ResNet-18 reflects SGD convergence rather than linearization error.

Table 3. 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)

D. Method Overview Figure

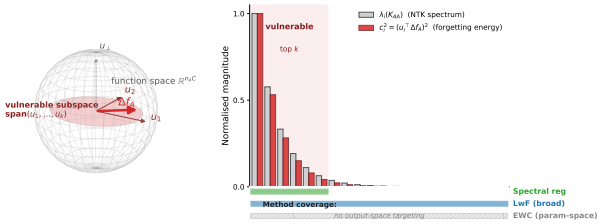

Figure 2. **Method overview.** *Left*: Forgetting Δf_A lies in the column space of K_{AA} (Prop. 1), and its energy concentrates on a low-rank slice – the vulnerable subspace $\mathrm{span}(u_1, \ldots, u_k)$ spanned by the top eigenvectors of K_{AA}. The complementary $u_\perp$ directions are unprotected by construction. *Right*: the NTK spectrum decays rapidly (grey), and the forgetting-energy coefficients $c_i^2 = (u_i^\top \Delta f_A)^2$ inherit this decay (red); 50–90% of Δf_A's energy lives in the first 1–6 modes (Fig. 1, center). *Spectral reg* (§4) penalizes drift exactly on those modes and leaves the complement free; *LwF* penalizes every output direction uniformly; *EWC* acts in parameter space and has no mechanism to selectively reach output modes (§5.1, App. F).

E. Low-Rank Forgetting vs. Low-Rank Learning

The confound. The center panel of Fig. 1 shows the forgetting vector concentrating in the top eigenmodes of K_{AA}. Taken alone this is not conclusive: the NTK spectrum λ_i already decays rapidly by spectral bias (Rahaman et al., 2019), so *any* vector expressed in this eigenbasis tends to look top-heavy. Concentration could therefore be a generic property of the architecture rather than a fact about forgetting. To see which, write the per-mode coefficient of Eq. (5) as a product of two factors:

$$c_i = \underbrace{\sigma_i}_{\text{(i) spectral decay}} \cdot \underbrace{v_{A,i}^\top J_B^\top (K_{BB} + \lambda I)^{-1} r_B}_{\text{(ii) cross-task alignment}}. \quad (14)$$

Factor (i) is the singular value $\sigma_i = \sqrt{\lambda_i}$; it shrinks high modes *regardless of the task* – pure spectral bias, shared by learning and forgetting. Factor (ii) depends on the *actual* new task: it measures how the Task-B residual r_B, passed through the new-task kernel, projects onto the i-th old-task direction. The eigenvalue decay (i) is not the novel content; the task-specific selection (ii) is.

A control that isolates the two factors. To test whether the observed concentration is (i) or (ii), we keep the *same* operators K_{AB}, K_{BB} (so factor (i) is untouched) but replace the real residual r_B in Eq. (3) with a random residual of equal norm, which randomizes factor (ii). If the concentration were just spectral bias, the random residual – fed through the identical decaying kernel – would concentrate just as tightly. On Split-MNIST (3 seeds) we report, for four energy profiles in the relevant NTK eigenbasis, the participation ratio $\mathrm{PR} = (\sum_i e_i)^2 / \sum_i e_i^2$ (lower = more concentrated) and the number of modes carrying 50/90% of the energy (k_{50}/k_{90}), in an $n_A C$=500-dimensional output space:

Energy profile	k_{50}	k_{90}	top-1	PR
K_{AA} spectrum λ_i (capacity)	6.0	47.3	0.15	16.5
Realized forgetting Δf_A	1.0	4.0	0.78	**1.6**
Random-residual control	52.0	217.0	0.01	123.0
Realized learning Δf_B	1.3	3.7	0.65	2.2

The realized forgetting vector is far more concentrated (PR=1.6, 97% of its energy in 6 modes, 78% in a single mode) than the bare K_{AA} spectrum (PR=16.5): spectral bias alone does *not* explain it. The random-residual control, fed through the identical operators, instead scatters across PR=123 modes – *more* spread than even the bare spectrum. So the tight concentration is produced by factor (ii), the cross-task alignment of the *real* residual, not by the kernel's decay. This is precisely why a cross-task predictor is needed and the spectrum is not enough: factor (i) says the low-eigenvalue modes *exist*, but only factor (ii) – the K_{AB}/r_B product – says *which* of them the new task will actually disturb.

This also settles the question of whether it is forgetting or learning that is low-rank. The learning drift Δf_B, decomposed in the K_{BB} eigenbasis, is itself low-rank (PR=2.2) – a direct consequence of spectral bias, factor (i). Both learning and forgetting are low-rank for that shared reason; forgetting is *additionally* sharpened by cross-task alignment (factor (ii)) into a still lower-dimensional, task-pair-specific subspace. The novelty of the low-rank claim is thus not the decay of Λ (generic) but the alignment-driven selection of a particular vulnerable direction.

F. Why EWC Cannot Protect the Vulnerable Subspace

EWC (Kirkpatrick et al., 2017) penalizes parameter drift weighted by the diagonal Fisher information F_{ii}, anchoring each parameter to its Task-A value with strength proportional to F_{ii}. The vulnerable subspace, by contrast, is defined in *output* space: it is the span of the top-k eigenvectors of $K_{AA} = J_A J_A^\top$. There is no general mechanism by which a per-parameter diagonal penalty maps onto a selected set

of output-space eigenmodes. Diagonal Fisher is anisotropic in parameter coordinates, but it is not constructed to align with the rank-k output-space projector $U_k U_k^\top$. Thus EWC can slow drift broadly, but it need not preferentially protect the vulnerable K_{AA} eigenmodes.

We verify this with a single-step drift-decomposition experiment that asks exactly the question of interest: does EWC's regularization preferentially suppress drift inside the top-k eigenspace of K_{AA}? After training Task A, we record U_k on probes, then train Task B under (i) no regularization, (ii) EWC, and (iii) spectral regularization, and decompose the realized Δf_A into its vulnerable and complementary components (Split-MNIST, k=10, 3 seeds):

Method	$\|U_k^\top \Delta f_A\|^2$	$\|(I - U_k U_k^\top)\Delta f_A\|^2$	Other:Vuln
No reg	$177\text{k} \pm 20\text{k}$	$5.0\text{k} \pm 0.1\text{k}$	$0.028:1$
EWC	$85\text{k} \pm 9\text{k}$	$2.8\text{k} \pm 0.5\text{k}$	$0.033:1$
Spectral	25 ± 6	820 ± 255	$32.7:1$

EWC and no-reg have indistinguishable Other:Vuln ratios: EWC scales both components down by roughly $2\times$, but does not preferentially shrink the vulnerable component. Spectral regularization flips the ratio by three orders of magnitude. This is the operational answer to how Fisher-based regularization aligns with the vulnerable subspace: under realized drift, it does not. The five-task drift decomposition in Section 5.2 confirms the same pattern at scale (Table 2).

The failure is not mis-weighting – it is the parameter-space penalty form. A reviewer might expect EWC to fail because diagonal Fisher *mis-points*, putting its mass on the wrong (non-vulnerable) parameters. The opposite is true. Pushing the diagonal Fisher F_{ii} through J_A into output space and comparing it to the per-parameter vulnerable-mode mass $\mathrm{diag}(J_A^\top U_k U_k^\top J_A)$ (Split-MNIST, k=10, 3 seeds), the two are *strongly* aligned: Spearman $\rho = 0.996$, cosine 0.94, and 55% of the total Fisher trace already lies inside the top-k vulnerable subspace. EWC's weights sit on essentially the right parameters. Yet realized drift (above) shows no selective protection, because a diagonal *quadratic* parameter anchor only rescales how fast each coordinate moves; it cannot impose the rank-k output-space constraint $U_k^\top \Delta f_A \approx 0$ that the geometry requires. Selective protection of a low-rank *output* subspace needs a function-space penalty (Eq. (8)), not a better-weighted parameter-space one – which is precisely why spectral regularization succeeds where EWC, with near-identical Fisher targeting, does not.

G. Inter-Task Confusion in Class-Incremental Learning

In class-incremental settings, the shared-head classifier is never jointly trained to discriminate between classes from different tasks; the dominant failure mode is inter-task *confusion* – the relative ordering of old-task logits is corrupted by drift introduced when fitting new-task logits. Our predictor identifies precisely this corruption in closed form: the top eigenmodes of K_{AA} that carry Δf_A are the directions along which the old-class logit field can move. Protecting them with spectral regularization stabilizes the old-class decision boundary against the cross-task force $K_{AB}(K_{BB} + \lambda I)^{-1} r_B$, while leaving the head free to fit new-task logits in the complement. This is a structural reason for the consistent gains seen on shared-head benchmarks (Table 1), and explains why parameter-space methods, which lack a way to selectively constrain the old-class output field, cannot recover inter-task discrimination.

We make this concrete on Split-MNIST: after training the new task on a shared 10-way head, we evaluate on held-out *old-task* test images (true labels 0–4) and measure the inter-task confusion rate – the fraction predicted into a new-task class $\{5, \ldots, 9\}$ – alongside accuracy restricted to the old-class logits (task identity known), 3 seeds:

Method	Old-class acc. ↑	Confusion → new ↓	Acc. | task ↑
No reg	0.0%	100.0%	29.8%
EWC	1.3%	98.7%	84.4%
Spectral	$\mathbf{93.5}\%$	$\mathbf{0.0}\%$	$\mathbf{93.5}\%$

The decomposition is revealing: EWC actually preserves the *within*-old-task ordering (84.4% accuracy once the task is known), yet 98.7% of old-class images are still classified as new-task classes – the new logits overwhelm the old ones because EWC cannot constrain the cross-task output direction. Spectral regularization, by protecting exactly that direction, drives inter-task confusion from 100% to 0% and recovers old-class accuracy without any task oracle.

H. GPM and Other Non-Diagonal Parameter-Space Methods

GPM (Saha et al., 2021) extends parameter-space regularization beyond EWC's diagonal Fisher by projecting new-task gradients orthogonal to the top-energy subspace of old-task gradients. It is therefore not subject to the rank mismatch of Section 5.1: its protection operator is itself low-rank. The mechanism it targets, however, is different from ours. GPM identifies directions in *parameter* space along which old-task gradients had large energy, whereas the vulnerable subspace identified by Proposition 1 consists of directions in *output* space along which K_{AA} has large eigenvalues. On Split-CIFAR-100 (10 tasks, frozen ResNet-18, frozen-backbone linear-head PEFT-CL regime), GPM achieves 9.7–10.1% final-task accuracy across energy thresholds, matching diagonal-Fisher EWC (9.9%) and dominated by spectral regularization (39.9%) and LwF (39.8%). Gradient-energy

bases simply do not align with output-interference directions – which function-space methods target by construction. The conclusion is that the failure of parameter-space methods on shared-head benchmarks is not specific to diagonal Fisher: it is a consequence of solving the wrong geometric problem (input/parameter gradient energy) for the wrong target (output-space drift).

I. 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.

J. 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.

K. 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

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**

L. CIFAR-10 Drift Decomposition

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

M. 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.

N. Sensitivity Analysis

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-

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$\pm$7.8	72.7$\pm$5.8	79.6$\pm$1.2	82.1$\pm$0.7	**84.6**$\pm$0.5	83.2$\pm$0.2

specific regularization.

Measured k_G. We verify the Kronecker rule's input by measuring the effective rank of the (raw) single-output feature Gram $G_{ij} = \phi(x_i)^\top \phi(x_j)$ – the object that $K_{AA} = I_C \otimes G$ depends on – using 10 probes/class. We report both the participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ and k_{50} (modes for 50% energy):

Features	PR	k_{50}	top-1
TinyMLP penultimate (MNIST)	3.1	1	0.50
Frozen ResNet-18 (CIFAR-10)	2.1	1	0.69
Frozen ResNet-18 (CIFAR-100)	2.4	1	0.65

All three sit in $k_G \in [1, 5]$ (a single dominant direction, k_{50}=1, with PR $\approx$ 2–3), confirming the input to $k^\star \approx C \cdot k_G$. For CIFAR-100 ($C$=100) this gives $k^\star \approx$ 100–240, consistent with the k=100 matched point at which targeted and broad function-space methods converge (Section 5.3).