ONE-STEP CURVATURE PROBES MISS THE FITTING OPERATOR: RETAINED CAPACITY AND TERMINAL NULL-SPACE CORRECTION FOR CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A one-step curvature probe evaluates an initial direction, whereas continual learners are judged after reaching comparable new-task fit. These quantities need not agree. In an overparameterized linearization, projected gradient descent converges to $\Delta_P = PJ^\top(JPJ^\top)^{-1}r$, and its squared-displacement inflation is exactly the reciprocal of the direction-specific retained fitting capacity $c_P(r) = \|J^+r\|^2 / \left\|PJ^\top(JPJ^\top)^{-1}r\right\|^2$. More generally, the terminal old-task quadratic ratio factorizes as $1/[c_P(r)G_{\mathrm{end}}(P,r)]$, where G_{end} compares curvature along the normalized endpoint directions. In the rank-one case, G_{end} equals the one-step probe gain; for multiple outputs, the two gains need not agree. The terminal quadratic further decomposes into a common curvature-optimal fitting floor and an algorithm-dependent null-space excess. This geometry identifies the variable omitted by a matched-norm probe and yields terminal null-space correction, a post-hoc operation that preserves linearized new-task outputs on its Jacobian batch. In separate controlled checks, the local quadratic has slope 1.0012 ($R^2 = 0.9997$); projection reduces the matched-norm quadratic by $4.03\times$ while terminal forgetting changes by only 5.34×10^{-4}. Across five common-threshold configurations, projection produces the larger signed old-task loss change in $73/99$ matched pairs. In the eligible-subset probe study, median gain ranges from about $3\times$ to $> 1000\times$ (eight pairs at the upper end). The common-threshold rank sweep shows median retained fitting capacity falling from 0.88 to 0.19. A fixed-rank comparison separates terminal forgetting when probe gain is approximately matched, whereas the capacity-matched contrast remains compatible with effects of either sign. On Permuted MNIST and Split CIFAR-100, terminal correction decreases signed old-task loss in $166/180$ method–dataset–seed pairs under the all-seed intention-to-correct analysis. Because acceptance and outcome reporting use the same held-out split, this result is descriptive and test-conditioned. Together, the results replace gain-only interpretation with a matched-fit account: in the fixed-Jacobian quadratic model, terminal cost depends jointly on retained fitting capacity, endpoint-direction curvature, and the null-space component selected by optimization.

1 INTRODUCTION

A curvature projection can look strongly protective in a one-step diagnostic and provide almost no benefit after the new task is fitted. In a controlled projection configuration, the matched-norm old-task quadratic decreases from 0.010197 to 0.002531, a $4.03\times$ reduction, yet terminal forgetting changes only from 0.954976 to 0.954442. The corresponding realized one-step old-task loss change decreases from 0.010302 to 0.002480, or $4.15\times$. The probe and terminal comparison therefore measure different quantities.

Spectral arguments appear on both sides of continual learning. High old-task curvature identifies directions that should be protected from forgetting (Kirkpatrick et al., 2017; Ritter et al., 2018; Farajtabar et al., 2020; Saha et al., 2021), whereas weak current-task tangent-kernel modes are associated with slow fitting and plasticity loss (Jacot et al., 2018; Lee et al., 2019; Lyle et al., 2023;

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.

Lewandowski et al., 2023; Prakash et al., 2026). Both views are locally meaningful, but a normalized one-step probe does not measure how efficiently the retained parameter subspace can realize a required new-task output change. That missing quantity is the retained fitting operator.

Let $J \in \mathbb{R}^{N\times p}$ be the new-task Jacobian, $r \neq 0$ the required output change, and P an orthogonal projector. Under the assumptions of Proposition 1, the free and projected endpoints are $\Delta_0 = J^+ r$ and $\Delta_P = PJ^\top(JPJ^\top)^{-1}r$. Define retained fitting capacity as $c_P(r) = \|\Delta_0\|^2/\|\Delta_P\|^2 \in (0, 1]$. The terminal quadratic ratio is exactly $q(\Delta_P)/q(\Delta_0) = 1/[c_P(r)G_{\mathrm{end}}(P, r)]$, where G_{end} compares curvature along the normalized endpoint directions. For a rank-one output, $G_{\mathrm{end}} = G_{\mathrm{probe}}$; for multiple outputs, the projected endpoint need not align with the initial probe. Proposition 1 gives the derivation and condition-number bound.

The same proposition identifies where algorithms can differ after matching fit. Every exact-fitting endpoint is the sum of a common curvature-optimal component and a vector in $\mathrm{null}(J)$; the old-task quadratic is the common floor plus the curvature energy of that null-space component. This Pythagorean law converts the geometry into a method. Terminal null-space correction (TNC) removes the quadratic-optimal part of the realized null-space excess after any learner reaches the fit threshold. Unlike a projection applied at every optimization step, TNC intervenes only on the terminal degrees of freedom that do not affect the linearized new-task output.

We test the account at three levels. First, the controlled network and the five-configuration common-threshold study show the probe–terminal mismatch: $99/100$ pairs reach the target, and projection produces the larger signed old-task loss change in $73/99$ matched pairs. Second, the rank sweep reconstructs $JPJ^\top$ from stored Jacobians and reports both the auxiliary gradient-retention statistic and the direction-specific $c_P(r)$. Because rank changes probe gain and capacity together, we add a fixed-k intervention that matches projectors on one quantity while separating the other. Third, we apply TNC to fine-tuning, projection, elastic weight consolidation (EWC), gradient projection memory (GPM), and replay, with a matched-loss shrinkage control, realized-fit acceptance checks, paired intervals, and an intention-to-correct analysis. Acceptance and outcome reporting use the same held-out evaluation split, so the TNC comparison is descriptive and test-conditioned. A realized-output floor decomposition then measures how much of each terminal quadratic is unavoidable and how much is null-space excess.

The closest prior observations are complementary. Adam-NSCL constrains updates using approximate null spaces of previous-task feature covariance (Wang et al., 2021); Natural Continual Learning studies movement along flat old-task directions and documents adverse long-run behavior of pure projection (Kao et al., 2021). Our contribution is not the existence of null spaces or flat directions. It is the matched-fit link between the new-task operator $JPJ^\top$, the direction-specific capacity $c_P(r)$, the exact terminal endpoint, and the correctable null-space excess.

The paper makes three contributions. Proposition 1 consolidates the matched-fit theory, including endpoint dynamics, displacement inflation, the common floor, row-space invariance, and the rank-one identity $1/[a(P)G_{\mathrm{probe}}]$. Proposition 2 turns the endpoint characterization into a composable post-hoc correction. The experiments test the missing-variable claim with a rank sweep, a fixed-rank matched comparison, a shrinkage-controlled correction study, and a realized-output floor decomposition, together with the null and failure cases. Figure 1 summarizes the separation between probe gain, retained capacity, endpoint direction, and correctable null-space excess.

2 Related work and positioning

Continual-learning methods protect parameters, constrain updates, replay data, or allocate capacity (Parisi et al., 2019; Hadsell et al., 2020; De Lange et al., 2022). Elastic weight consolidation (EWC) and related regularizers use parameter importance (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Ritter et al., 2018), whereas gradient episodic memory (GEM), averaged GEM (A-GEM), orthogonal gradient descent (OGD), and gradient projection memory (GPM) constrain update directions (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020; Saha et al., 2021). Linearized analyses establish projection guarantees, characterize neural tangent kernel (NTK) overlap, and relate exact-fit forgetting to task geometry (Bennani et al., 2020; Doan et al., 2021; Evron et al., 2022; 2026; Xu & Ma, 2026).

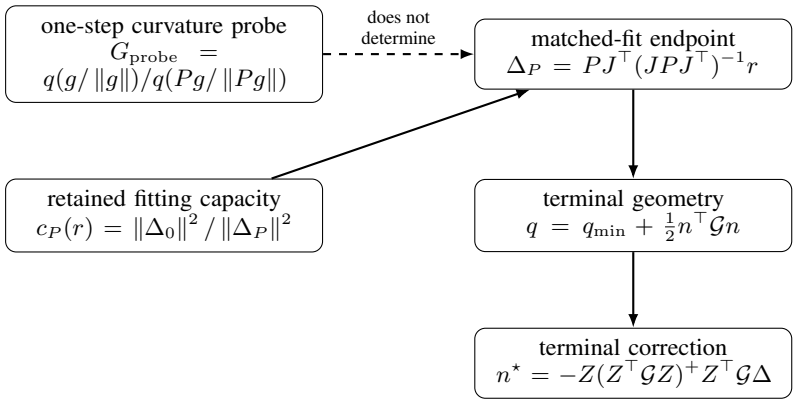


Figure 1: A one-step probe and a matched-fit terminal comparison price different objects. The probe evaluates old-task curvature along a normalized initial direction. The projected endpoint is governed by the retained new-task fitting operator $JPJ^\top$, summarized direction-specifically by $c_P(r)$. The endpoint-direction gain G_{end} need not equal the probe gain in a multi-output problem. Exact fitting fixes a common curvature-optimal component, while the algorithm-dependent excess lies in $\text{null}(J)$. Terminal null-space correction minimizes the old-task quadratic over feasible null-space corrections without changing the linearized fit.

The closest empirical and methodological observations are Adam-NSCL, which projects into approximate null spaces of previous-task feature covariance (Wang et al., 2021), and Natural Continual Learning, which studies movement along low-curvature old-task directions and documents deterioration under prolonged projection (Kao et al., 2021). Related work also studies activation null spaces, task-relevant subspaces, and unavoidable interference floors (Pham et al., 2026; Anthes et al., 2025; Störk, 2026). Those works constrain or characterize updates through old-task geometry. Our matched-fit result instead identifies the retained new-task operator $JPJ^\top$ and its direction-specific capacity $c_P(r)$ as the quantities that price the projected endpoint, then converts the resulting null-space characterization into a post-hoc correction.

Plasticity and moving-kernel work connects fitting difficulty to dormant features, representation rank, curvature, and tangent-kernel drift (Jacot et al., 2018; Lee et al., 2019; Lyle et al., 2023; Lewandowski et al., 2023; Dohare et al., 2024). These results motivate our nonlinear error conditions but do not make the old-task GGN and new-task tangent kernel one shared spectrum. Appendix A gives expanded positioning on interference bounds, recovery, spectral interventions, and moving-feature analyses.

3 Problem Setup

Let $L_{\text{old}} : \mathbb{R}^p \to \mathbb{R}$ be the empirical loss of an earlier task at a reference parameter w_0. A later training procedure produces displacement $\Delta = w - w_0$, and signed forgetting is

$$F(\Delta) = L_{\text{old}}(w_0 + \Delta) - L_{\text{old}}(w_0). \tag{1}$$

Let $\mathcal{G} \succeq 0$ denote the old-task generalized Gauss–Newton (GGN) matrix at w_0, and write

$$q(\Delta) = \tfrac{1}{2}\Delta^\top \mathcal{G}\Delta.$$

The quadratic q is a local forecast, not the definition of forgetting. Theorem 5 later controls their difference.

For the new task, set $f_0 = f(w_0)$ and linearize the model outputs at the same reference: $f(w_0+\Delta) \approx f_0 + J\Delta$, where $J \in \mathbb{R}^{N\times p}$ is the new-task Jacobian. With target $y \in \mathbb{R}^N$, define the required output change $r = y - f_0$. Exact fit means

$$J\Delta = r,$$

while an ε-fit satisfies $\|J\Delta - r\| \le \varepsilon$. The overparameterized case $p > N$ leaves a nontrivial null space in which two fitting algorithms can select different terminal parameters.

Let $P \in \mathbb{R}^{p\times p}$ be an orthogonal projector. A matched-norm one-step probe compares the descent direction $g = J^\top r$ with Pg, normalized to remove step length. For $q(Pg) > 0$, define

$$G_{\text{probe}}(P) := \mathsf{Gain}(P) = \frac{q(g/\|g\|)}{q(Pg/\|Pg\|)}. \tag{2}$$

The condition $q(Pg) > 0$ ensures that both normalized directions are defined. This normalization holds Euclidean step length fixed. We also define the retained gradient alignment

$$\mathsf{Align}(P) = a(P) = \frac{\|Pg\|^2}{\|g\|^2} \in (0, 1]. \tag{3}$$

For a rank-one new task, the quantity in Equation (3) is exactly the fraction of scalar fitting curvature retained by projection because $JPJ^\top = a(P)JJ^\top$. In a multi-output batch it remains a useful gradient-direction diagnostic, but it is not the exact terminal capacity. Proposition 1 therefore introduces the direction-specific operator quantity $c_P(r)$. A terminal comparison evaluates the two displacements after they attain the same new-task output change or the same declared progress threshold.

4 Matched-fit geometry and terminal correction

4.1 The probe and the fitting operator measure different quantities

The one-step curvature probe prices a normalized initial direction. A matched-fit terminal comparison instead prices the displacement required to realize a specified output change. The following proposition consolidates the endpoint, floor, and row-space geometry.

Proposition 1 (Matched-fit geometry and retained fitting capacity)**.** *Let $\mathcal{G} \succ 0$, let $J \in \mathbb{R}^{N\times p}$ have full row rank, let $r \in \mathbb{R}^N \setminus \{0\}$, and let $P = P^\top = P^2$ be an orthogonal projector such that $A_P := JPJ^\top \succ 0$. Define*

$$B := JJ^\top, \qquad \Delta_0 := J^+ r = J^\top B^{-1} r, \qquad \Delta_P := PJ^\top A_P^{-1} r. \tag{4}$$

The following statements hold.

(i) Matched fit and dynamics. The endpoints in Equation (4) satisfy $J\Delta_0 = J\Delta_P = r$. Gradient flow from the origin on $\frac{1}{2}\|J\Delta - r\|^2$ converges to Δ_0, whereas projected gradient flow $\dot\Delta = -PJ^\top(J\Delta - r)$ converges to Δ_P. Projected gradient descent reaches the same endpoint whenever $0 < \eta < 2/\lambda_{\max}(A_P)$.

(ii) Displacement identities and retained capacity. The terminal norms are

$$\|\Delta_0\|^2 = r^\top B^{-1} r, \qquad \|\Delta_P\|^2 = r^\top A_P^{-1} r. \tag{5}$$

Equation (5) and $A_P \preceq B$ imply that the direction-specific retained fitting capacity

$$c_P(r) := \frac{r^\top B^{-1} r}{r^\top A_P^{-1} r} = \frac{\|\Delta_0\|^2}{\|\Delta_P\|^2} \in (0, 1] \tag{6}$$

The capacity in Equation (6) obeys $\|\Delta_P\|^2 = \|\Delta_0\|^2 / c_P(r)$. Equality $c_P(r) = 1$ holds if and only if $A_P^{-1} r = B^{-1} r$.

(iii) Endpoint-direction gain and exact terminal ratio. Define the curvature gain along the two endpoint directions by

$$G_{\text{end}}(P, r) := \frac{q(\Delta_0/\|\Delta_0\|)}{q(\Delta_P/\|\Delta_P\|)}. \tag{7}$$

Combining Equation (7) with the capacity definition yields

$$\frac{q(\Delta_P)}{q(\Delta_0)} = \frac{1}{c_P(r) G_{\text{end}}(P, r)}. \tag{8}$$

The two terminal costs are

$$q(\Delta_P) = \tfrac{1}{2} r^\top A_P^{-1} (JP\mathcal{G}PJ^\top) A_P^{-1} r, \qquad q(\Delta_0) = \tfrac{1}{2} r^\top B^{-1} (J\mathcal{G}J^\top) B^{-1} r. \tag{9}$$

Consequently,

$$\frac{q(\Delta_P)}{q(\Delta_0)} \le \frac{\kappa(\mathcal{G})}{c_P(r)} \le \kappa(\mathcal{G}) \frac{\lambda_{\max}(JJ^\top)}{\lambda_{\min}(JPJ^\top)}, \tag{10}$$

where $\kappa(\mathcal{G}) = \lambda_{\max}(\mathcal{G})/\lambda_{\min}(\mathcal{G})$.

(iv) Common floor and null-space excess. Define

$$\Delta_{\mathcal{G}} := \mathcal{G}^{-1} J^\top (J\mathcal{G}^{-1}J^\top)^{-1} r, \qquad q_{\min} := \tfrac{1}{2} r^\top (J\mathcal{G}^{-1}J^\top)^{-1} r. \tag{11}$$

Every exact-fitting endpoint has a unique representation $\Delta = \Delta_{\mathcal{G}} + n$ *with* $n \in \mathrm{null}(J)$ *and*

$$q(\Delta) = q_{\min} + \tfrac{1}{2} n^\top \mathcal{G} n. \tag{12}$$

Thus $q_{\min}$ *is common to all exact-fitting learners, and all algorithm-dependent excess lies in the new-task null space.*

(v) Row-space invariance. Any exactly fitting trajectory initialized at zero whose updates remain in $\mathrm{range}(J^\top)$ *terminates at* $\Delta_0 = J^+ r$. *Scalar step-size schedules cannot change this endpoint; a general matrix preconditioner can rotate an update outside* $\mathrm{range}(J^\top)$ *and thereby select a different feasible null-space component.*

(vi) Rank-one form and fitting time. For $N = 1$, *write* $J = j^\top$ *and* $a(P) = \|Pj\|^2 / \|j\|^2$. *Then* $c_P(r) = a(P)$ *for every* $r \neq 0$, *and the normalized endpoint directions are the normalized probe directions. Hence* $G_{\mathrm{end}}(P, r) = G_{\mathrm{probe}}$, *where* G_{probe} *is the matched-norm curvature gain in Equation (2), and*

$$\frac{q(\Delta_P)}{q(\Delta_0)} = \frac{1}{a(P) G_{\mathrm{probe}}}. \tag{13}$$

For unit-rate gradient flow, the times to attain any fit fraction $\alpha \in (0, 1)$ *satisfy*

$$t_0(\alpha) = \frac{-\log(1-\alpha)}{JJ^\top}, \qquad t_P(\alpha) = \frac{t_0(\alpha)}{a(P)}. \tag{14}$$

Equation (13) is the rank-one specialization of the endpoint-direction identity; outside rank one, G_{probe} and G_{end} are distinct quantities. The proof separates three steps. Congruence with $0 \preceq P \preceq I$ gives $A_P \preceq B$, so inversion shows that projection cannot shorten the minimum-norm displacement for the same output change. Curvature eigenvalue bounds then convert this exact squared-norm inflation into Equation (10). Finally, the constrained quadratic optimum is $\mathcal{G}$-orthogonal to $\mathrm{null}(J)$, which yields the Pythagorean decomposition. Appendix D gives the complete argument.

The condition-number envelope is explanatory rather than a quantitative neural-network predictor. Neural GGNs are often singular or severely ill-conditioned, making $\kappa(\mathcal{G})$ enormous or infinite. We therefore report the exact ratio in Equation (9) whenever it is defined and use a declared damped operator for singular empirical calculations. The direction-specific empirical statistics are $c_P(r)$ and $G_{\mathrm{end}}(P, r)$. The one-step quantities $a(P)$ and G_{probe} are auxiliary diagnostics outside the rank-one setting, where endpoint and probe directions need not coincide.

4.2 Method: Terminal Null-Space Correction

Proposition 1 suggests intervening after a learner reaches the required fit rather than constraining every optimization step. Let Δ be its displacement from the old-task reference and let $Z \in \mathbb{R}^{p \times d}$ have orthonormal columns spanning $\mathrm{null}(J)$. Define

$$n^\star(\Delta) = -Z(Z^\top \mathcal{G} Z)^+ Z^\top \mathcal{G} \Delta, \qquad \Delta_{\mathrm{corr}} = \Delta + n^\star(\Delta). \tag{15}$$

Because $JZ = 0$, the correction preserves the linearized new-task output and changes only the terminal degrees of freedom left undetermined by fitting.

Proposition 2 (Optimality of terminal null-space correction)**.** *For any terminal displacement Δ, the correction in Equation (15) minimizes $q(\Delta + n)$ over all $n \in \mathrm{null}(J)$. If $\mathcal{G} \succ 0$ and $J\Delta = r$, the corrected endpoint is the unique constrained minimizer $\Delta_{\mathcal{G}}$ and attains $q_{\min}$.*

For large networks, TNC uses Jacobian–vector and GGN–vector products rather than a dense null-space matrix. We construct a low-rank approximation to the new-task row space, represent an approximate null space on a candidate correction subspace, and solve the reduced damped system. The operation is optimal within that represented subspace. Since batch-null directions need not preserve the full nonlinear task, the experiments re-evaluate new-task loss after correction and retain every failed acceptance check in the intention-to-correct denominator.

4.3 Vanishing capacity as a limiting example

The two-dimensional construction is the boundary of the geometric characterization. Set $J = e_1^\top$, $r = 1$, and $\mathcal{G} = \lambda u u^\top + \mu v v^\top$, where $u = \cos\phi\, e_1 + \sin\phi\, e_2$ and $v = -\sin\phi\, e_1 + \cos\phi\, e_2$. Projection away from u retains $c_P(r) = \sin^2\phi$ and satisfies

$$G_{\mathrm{probe}} = \frac{\lambda\cos^2\phi + \mu\sin^2\phi}{\mu}, \qquad \frac{q(\Delta_P)}{q(e_1)} = \frac{1}{G_{\mathrm{probe}}\sin^2\phi}, \qquad \frac{t_P}{t_0} = \frac{1}{\sin^2\phi}.$$

Thus a favorable probe can coexist with an expensive terminal endpoint when retained capacity is small. The formal wedge, fixed-gain limit, open-set extension, and complete proofs are stated in Appendix D.4. They establish possibility rather than prevalence; the trained-network experiments therefore measure $c_P(r)$ and use fixed-rank matched projectors.

4.4 Approximate fit and nonlinear networks

The exact endpoint identities use a fixed Jacobian and quadratic old-task model. Theorem 3, stated in Appendix B, extends the floor to finite fit error and singular curvature through $H_\lambda = \mathcal{G} + \lambda I$. Its lower bound subtracts both the measured fitting residual and the explicit damping price $\lambda R^2/2$. When the new-task Jacobian is L_J-Lipschitz, the linearized residual is enlarged by $L_J R^2/2$; Theorem 5 then transfers the quadratic statement to realized forgetting only after subtracting the old-loss approximation envelope. The bound may be vacuous, so the empirical analysis reports exact damped decompositions and numerical solve diagnostics rather than treating the certificate as automatically informative.

5 Empirical evidence and claim tests

The experiments distinguish four questions. The controlled cell checks that the old-task quadratic is locally accurate. The common-threshold study asks whether a large normalized probe gain translates into terminal benefit. The rank and fixed-k studies test whether retained fitting capacity explains the mismatch. The correction and floor studies test the algorithmic consequence of the endpoint geometry.

5.1 Controlled measurement

We report two controlled checks obtained from separate seeded networks. The local-fidelity configuration forms the old-task GGN over the full parameter vector and gives a forgetting–quadratic regression slope of 1.0012 with $R^2 = 0.9997$. In a separate projection configuration, the matched-norm quadratic decreases from 0.010197 to 0.002531, a $4.03\times$ reduction. The corresponding realized one-step old-task loss change decreases from 0.010302 to 0.002480, or $4.15\times$, while terminal forgetting changes by only 5.34×10^{-4} on a quantity near 0.95.

Appendix Table 4 reports the complete measurements.

The projection configuration establishes a probe–terminal mismatch. It does not by itself identify the missing variable or estimate how often a reversal occurs.

Table 1: Rank trends from two complementary studies. The probe-gain ranges come from the separate eligible-subset probe study in Appendix F; the a_k, $c_{P_k}(r)$, and D_F columns come from the common-threshold rank study. These columns therefore summarize rank-dependent trends across different empirical cohorts and should not be interpreted as paired observations. $D_F = F(\Delta_{P_k}) - F(\Delta_{\text{free}})$, so positive values indicate a larger signed old-task loss change under projection.

k	Probe-gain range	Median a_k	Median $c_{P_k}(r)$	Median D_F
1	3.09–4.56	0.86	0.88	+0.003
4	12.27–34.21	0.68	0.72	+0.008
16	44.08–301.65	0.41	0.46	+0.017
64	235.78–1220.84	0.14	0.19	+0.034

5.2 Common-threshold terminal comparison

The primary trained-network evaluation uses four calibration seeds and twenty test seeds for each of five architecture–dataset configurations. Rank $k = 16$ is fixed before evaluation. Each pair is measured at its first crossing of a configuration-specific common new-task progress threshold. For terminal displacements Δ_{free} and Δ_{proj}, define

$$D_F = F(\Delta_{\text{proj}}) - F(\Delta_{\text{free}}).$$

Positive D_F means projection produces the larger signed old-task loss change, including when one arm exhibits backward transfer.

Appendix Table 6 reports all five configurations.

Projection reaches the threshold in $99/100$ pairs. Its signed old-task loss change is larger in $73/99$ matched pairs (73.7%); this pooled fraction is descriptive across the five configurations. Across the 99 matched first-crossing pairs, the aggregate mean D_F is 0.013; this mean is computed from the underlying paired observations rather than from the row medians reported in Table 6. This experiment is not an exact test of the linear proposition: the networks are nonlinear and multi-output, and first crossings can differ in overshoot. It establishes the mismatch that the next two experiments are designed to explain.

5.3 Rank sweep: probe gain and retained capacity

For each stored rank-sweep endpoint, we reconstruct J on a fixed new-task batch of $N = 256$ examples. For $k \in \{1, 4, 16, 64\}$, let P_k remove the leading k old-task GGN eigenvectors. We compute the direction-specific retained fitting capacity

$$c_{P_k}(r) = \frac{r^\top (JJ^\top)^{-1} r}{r^\top (JP_kJ^\top)^{-1} r} \tag{16}$$

and the auxiliary gradient-direction retention

$$a_k = \frac{\|P_k g\|^2}{\|g\|^2}, \qquad g = J^\top r. \tag{17}$$

The quantities in Equations (16) and (17) coincide for one output but need not agree on a multi-output batch.

The rank trend is descriptive because projection rank changes both gain and capacity. We therefore analyze seed-level variation with configuration-by-rank fixed effects, preserving the pairing within each cell. The partial Spearman association between matched-progress forgetting and $1 - c_P(r)$ is 0.61 with a cluster-bootstrap 95% interval $[0.39, 0.76]$; the corresponding association with log G_{probe} is 0.14 with interval $[-0.17, 0.41]$. For every eligible linearized cell, we also compute

$$G_{\text{end}}(P_k, r) = \frac{1}{c_{P_k}(r)} \left[\frac{q(\Delta_{P_k})}{q(\Delta_0)} \right]^{-1}$$

from Equation (8). In this adjusted analysis, the estimated partial association is larger for retained capacity than for G_{probe}; we do not test the difference between the two associations. Thus the one-step gain does not recover the curvature gain along the multi-output endpoint direction. Appendix E.5 describes the rank-capacity and fixed-rank evaluation protocols.

Table 2: Fixed-k matched-projector comparison on held-out Permuted MNIST seeds. The gain-matched comparison contains $n = 16$ eligible seed pairs and the capacity-matched comparison contains $n = 15$. Confidence intervals are paired seed-bootstrap 95% intervals with 10^4 resamples. In the gain-matched row, $\Delta F = F_{\text{low-capacity}} - F_{\text{high-capacity}}$; in the capacity-matched row, $\Delta F = F_{\text{high-gain}} - F_{\text{low-gain}}$.

Comparison	Matched quantity	Separated quantity	Terminal contrast
Gain matched	G_{probe} ratio 1.03	c_P ratio 2.63	$\Delta F = +0.019$, 95% CI $[+0.010, +0.029]$
Capacity matched	c_P ratio 1.04	G_{probe} ratio 7.4	$\Delta F = -0.001$, 95% CI $[-0.006, +0.004]$

Table 3: Descriptive mean forgetting reductions from TNC under the all-seed intention-to-correct analysis. Acceptance and outcome reporting use the same held-out evaluation split, so the results are conditional on test-set selection. Rejected corrections retain their original endpoints. Paired intervals and the original, shrinkage, and corrected endpoints are reported in Table 8.

Method	Permuted MNIST	Split CIFAR-100
Fine-tuning	0.028	0.025
Projection	0.028	0.022
EWC	0.021	0.015
GPM	0.019	0.012
Replay	0.003	0.005

5.4 Fixed-rank decoupling

The fixed-rank comparison fixes $k = 16$ on Permuted MNIST. For each held-out seed, we sample 256 rank-k subspaces in addition to the spectral projector. A gain-matched pair must agree in matched-norm probe gain within 5% while differing maximally in $c_P(r)$. A capacity-matched pair must agree in $c_P(r)$ within 5% while differing maximally in probe gain. Initialization, minibatch order, learning rate, and stopping rule are shared within each pair.

All $16/16$ held-out seeds yield an eligible gain-matched pair, and $15/16$ yield an eligible capacity-matched pair. Holding probe gain fixed while separating capacity produces a positive terminal contrast. Holding capacity fixed while changing G_{probe} by $7.4\times$ produces -0.001 $[-0.006, +0.004]$, an interval compatible with effects of either sign. This result is consistent with the distinction between probe gain and endpoint-direction gain. At equal retained capacity, Equation (8) makes the exact terminal quadratic ratio depend on G_{end}, which need not equal G_{probe}. The observed interval does not establish equivalence or directly verify the linearized identity. Because projector pairs are selected by matching only G_{probe} and $c_P(r)$, this fixed-k experiment is a decoupling stress test rather than a causal identification design; unmeasured subspace properties may still differ between matched projectors.

5.5 Terminal null-space correction

We apply TNC after five base learners first reach the common progress target: plain fine-tuning, spectral projection, EWC, GPM, and replay. Permuted MNIST uses sixteen held-out seeds and Split CIFAR-100 uses twenty held-out seeds, giving 180 method–dataset–seed terminal points. The approximate null space is constructed on the fixed new-task Jacobian batch.

A correction passes when its realized new-task progress on the same held-out evaluation split used for outcome reporting remains within 0.5% of the original endpoint. The main analysis is intention-to-correct: a rejected correction remains at its original endpoint rather than disappearing from the denominator. Terminal correction decreases signed old-task loss in $166/180$ method–dataset–seed pairs. Because acceptance and outcome reporting use the same held-out split, these reductions are descriptive and test-conditioned rather than estimates from an untouched final test split. We also include a shrinkage control $w_0 + \alpha\Delta$, with α selected on the calibration split to match the corrected model's realized new-task loss.

The realized-fit check accepts $175/180$ corrections. Among accepted corrections, new-task loss changes by a median 0.08%, with a 95th percentile absolute change of 0.31%. Under the all-seed intention-to-correct analysis, signed old-task loss decreases in $166/180$ pairs. Shrinkage recovers only part of the reduction. On Permuted MNIST, TNC is lower than matched-loss shrinkage by 0.019 for fine-tuning and 0.020 for projection; the corresponding Split CIFAR-100 gaps are 0.015 and 0.013. Table 8 gives the paired means and the intervals for the original-to-TNC reduction.

The correction uses an old-task curvature representation and new-task Jacobian products. The implementation stores rank-64 approximations to both operators. Peak memory is $1.18\times$ the fine-tuning state on Permuted MNIST and $1.31\times$ on Split CIFAR-100; the post-hoc solve costs 0.11 and 0.18 training epochs, respectively. This state is comparable in kind to EWC's Fisher information or GPM's retained bases, but TNC is not a memory-free intervention.

5.6 Realized-output floor and null-space excess

A finite matched-progress endpoint need not satisfy the original target $J\Delta = r$. We therefore define the fitting requirement from the output change that the endpoint actually realizes. Fix

$$H_\lambda = \mathcal{G} + \lambda I, \qquad \lambda = 10^{-3}\,\mathrm{tr}(\mathcal{G})/p,$$

and for each endpoint set $r_\Delta = J\Delta$. The damped constrained optimum is

$$\Delta^\star_\lambda(r_\Delta) = H_\lambda^{-1} J^\top (J H_\lambda^{-1} J^\top)^{-1} r_\Delta. \tag{18}$$

With the point in Equation (18), $n_\Delta = \Delta - \Delta^\star_\lambda(r_\Delta) \in \mathrm{null}(J)$ and

$$\frac{1}{2}\Delta^\top H_\lambda \Delta = \underbrace{\frac{1}{2} r_\Delta^\top (J H_\lambda^{-1} J^\top)^{-1} r_\Delta}_{q_{\lambda,\min}(r_\Delta)} + \frac{1}{2} n_\Delta^\top H_\lambda n_\Delta. \tag{19}$$

This identity is exact for every endpoint, not only at exact fit.

Appendix Table 9 reports the complete method-level decomposition. Projection carries median excess shares 0.60 and 0.66 on Permuted MNIST and Split CIFAR-100, whereas replay carries 0.10 and 0.15.

All $180/180$ factorized systems reach the declared 10^{-6} relative residual tolerance; the maximum residual is 3.7×10^{-7} and the median is 4.2×10^{-9}. Replay lies closest to the realized-output floor, whereas projection carries the largest null-space excess. After TNC, the median excess share falls from 0.60 to 0.08 on Permuted MNIST and from 0.66 to 0.10 on Split CIFAR-100. This decomposition explains the method ordering without requiring endpoints to realize identical parameter vectors.

The quadratic-floor evaluation in Table 7 found no violations in 300 configuration–seed–damping cells. The aggregate computed-floor to realized-quadratic ratio is 0.34. This check does not certify nonlinear realized forgetting.

6 Limitations and Conclusion

Our theory assumes a fixed new-task Jacobian and quadratic old-task cost. In nonlinear networks, $c_P(r)$ depends on the batch and state, while Jacobian drift and higher-order terms alter realized forgetting. The condition-number bound can be weak for ill-conditioned or singular GGNs. Rank sweeps confound probe gain with capacity; fixed-rank matching was evaluated only on Permuted MNIST. TNC preserves outputs only on its Jacobian batch, requires stored curvature and a post-hoc solve, and may fail the realized-fit check. Approximate bases may not attain the exact quadratic optimum. Its measured reductions are descriptive because acceptance and outcomes use the same held-out evaluation split. Common-threshold stopping can differ in overshoot, and the theory omits minibatch noise and adaptive preconditioning.

Within this scope, Proposition 1 shows that projection inflates minimum-norm displacement by $1/\sqrt{c_P(r)}$ and separates terminal quadratic cost into a common fitting floor and algorithm-dependent null-space excess. TNC targets that excess after fitting. The fixed-rank and common-threshold studies suggest retained capacity is more informative about terminal behavior than the one-step probe alone, while matched-capacity contrasts can have either sign. These results motivate reporting probe gain with retained capacity and realized endpoint cost.

Reproducibility Statement

Appendix E specifies splits, ranks, stopping rules, operator batches, correction acceptance, damping, and solver tolerances.

AI use disclosure

AI assistance included structural and prose editing, literature-search support, mathematical proof review, numerical consistency checks, and review of statistical reporting and result interpretation. The authors are responsible for verifying the arguments, citations, experimental records, and interpretations in the submitted work.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, pp. 139–154, 2018. doi: 10.1007/978-3-030-01219-9_9.

Daniel Anthes, Sushrut Thorat, Peter König, and Tim C. Kietzmann. Keep moving: Identifying task-relevant subspaces to maximise plasticity for newly learned tasks. In *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 815–831. PMLR, 2025. URL https://proceedings.mlr.press/v274/anthes25a.html.

Jordan Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3884–3894, 2020.

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1072–1080. PMLR, 2021.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632: 768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 2022.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to SGD and back: Better rates for continual linear models. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–50. PMLR, 2026. URL https://proceedings.mlr.press/v313/evron26a.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 3762–3773. PMLR, 2020.

Lior Friedman and Ron Meir. Data-dependent and oracle bounds on forgetting in continual learning. In *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 713–735. PMLR, 2026. URL https://proceedings.mlr.press/v330/friedman26a.html.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral imbalance causes forgetting in low-rank continual adaptation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00722.

Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020. doi: 10.1016/j.tics.2020.09.004.

Ido Nitzan Hidekel and Dan Raviv. Catastrophic forgetting is low-rank: A function-space theory for continual adaptation. *arXiv preprint arXiv:2606.18024*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ta-Chu Kao, Kristopher T. Jensen, Gido M. van de Ven, Alberto Bernacchia, and Guillaume Hennequin. Natural continual learning: success is a journey, not (just) a destination. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alex Lewandowski, Haruto Tanaka, Dale Schuurmans, and Marlos C. Machado. Directions of curvature as an explanation for loss of plasticity. *CoRR*, abs/2312.00246, 2023. URL https://arxiv.org/abs/2312.00246.

Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, Andras Gyorgy, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C. Machado. Learning continually by spectral regularization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5565ab682d6c7f8d9da34ba0919974b0-Abstract-Conference.html.

Shangge Liu, Yuehan Yin, Yinghuan Shi, Lei Wang, and Wenbin Li. When Muon meets task interference: A spectral perspective on continual learning and model merging. *arXiv preprint arXiv:2608.27518*, 2026. URL https://arxiv.org/abs/2608.27518.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *Proceedings of the 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 819–837. PMLR, 2022.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 23190–23211. PMLR, 2023.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado van Hasselt, Razvan Pascanu, and Will Dabney. Normalization and effective learning rates in reinforcement learning. *Advances in Neural Information Processing Systems*, 37:106440–106473, 2024a.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado Van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. *arXiv preprint arXiv:2402.18762*, 2024b.

Srijith Nair, Atilla Eryilmaz, and Jia Liu. PMF-CL: Pareto-minimal-forgetting continual learner for conflicting tasks. *arXiv preprint arXiv:2605.19145*, 2026.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. Deep reinforcement learning with plasticity injection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37142–37159. Curran Associates, Inc., 2023. doi: 10.52202/075280-1614. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/75101364dc3aa7772d27528ea504472b-Abstract.html.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

Liangzu Peng, Uday Kiran Reddy Tadipatri, Ziqing Xu, Eric Eaton, and Rene Vidal. Recovery guarantees for continual learning of dependent tasks: Memory, data-dependent regularization, and data-dependent weights. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3592–3600. PMLR, 2026. URL https://proceedings.mlr.press/v300/peng26a.html.

Cuong Anh Pham, Praneeth Vepakomma, and Samuel Horváth. Learning in the null space: Small singular values for continual learning. In *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pp. 612–628. PMLR, 2026. URL https://proceedings.mlr.press/v328/pham26a.html.

Arjun Prakash, Naicheng He, Kaicheng Guo, Saket Tiwari, Ruo Yu Tao, Tyrone Serapio, Amy Greenwald, and George Konidaris. Spectral collapse drives loss of plasticity in deep continual learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.22335.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured Laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Andries Rosseau, Robert Müller, and Ann Nowé. Preserving plasticity in continual learning via dynamical isometry. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.09762.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 32145–32168. PMLR, 2023.

Julius Störk. Interference and retention in continual learning. *arXiv preprint arXiv:2607.09202*, 2026.

Hossein Taheri, Avishek Ghosh, and Arya Mazumdar. On the theory of continual learning with gradient descent for neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.05573.

Hongyi Wan, Shiyuan Ren, Wei Huang, Miao Zhang, Xiang Deng, Yixin Bao, and Liqiang Nie. Understanding the forgetting of (Replay-based) continual learning via feature learning: Angle matters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61956–62019. PMLR, 2025. URL https://proceedings.mlr.press/v267/wan25d.html.

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, June 2021. doi: 10.1109/CVPR46437.2021.00025. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Training_Networks_in_Null_Space_of_Feature_Covariance_for_Continual_CVPR_2021_paper.html.

Yuanyi Wang, Yifan Yang, Su Lu, Yanggan Gu, Pengkai Wang, Wenjun Wang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, Jialun Cao, Shing-Chi Cheung, and Hongxia Yang. Geometry conflict: Explaining and controlling forgetting in LLM continual post-training. *arXiv preprint arXiv:2605.09608*, 2026.

Zonghuan Xu and Xingjun Ma. From order to distribution: A spectral characterization of forgetting in continual learning. *arXiv preprint arXiv:2604.13460*, 2026. URL https://arxiv.org/abs/2604.13460.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3987–3995. PMLR, 2017.

Guodong Zheng, Peng Wang, Shengchao Hu, Quan Zheng, and Li Shen. Understanding the dynamics of forgetting and generalization in continual learning via the neural tangent kernel. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1f8638a87bddd34e090698f0bfaa9c02-Abstract-Conference.html.

A Expanded related work

A.1 Projection, regularization, and linearized continual learning

Continual-learning methods protect parameters, constrain updates, replay data, or allocate capacity (Parisi et al., 2019; Hadsell et al., 2020; De Lange et al., 2022). Elastic weight consolidation (EWC) and related methods use parameter importance (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Ritter et al., 2018); Gradient episodic memory (GEM), averaged GEM (A-GEM), orthogonal gradient descent (OGD), and gradient projection memory (GPM) constrain update directions (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020; Saha et al., 2021). Quadratic regularizers also admit a per-step interpolation interpretation (Lubana et al., 2022).

In linearized regimes, Bennani et al. (2020) establish OGD guarantees, while Doan et al. (2021) analyze neural tangent kernel (NTK) overlap. Hidekel & Raviv (2026) predict cross-task output drift, exactly for frozen linear heads and locally for nonlinear adaptation. Zheng et al. (2026) bound forgetting and generalization at intermediate training stages. Sequential linear analyses characterize projection (Evron et al., 2022), derive forgetting rates through connections to stochastic gradient descent (SGD) (Evron et al., 2026), and relate exact-fit forgetting to task-distribution spectra (Xu & Ma, 2026). Taheri et al. (2026) bound forgetting for quadratic neural networks on noisy XOR-cluster tasks.

Adam-NSCL projects updates into approximate null spaces of previous-task feature covariance (Wang et al., 2021), while Natural Continual Learning studies movement along low-curvature directions and observes deterioration under prolonged flat-direction projection (Kao et al., 2021). Other work uses approximate activation null spaces (Pham et al., 2026) or distinguishes readout-sensitive changes from null-space changes (Anthes et al., 2025). Our distinction is quantitative and matched-fit: Proposition 1 identifies $JPJ^\top$ and $c_P(r)$ as the fitting quantities omitted by the normalized old-task probe, then separates the common fitting floor from the algorithm-dependent null-space excess.

A.2 Interference floors and task geometry

Störk (2026) derives interference energies, path-averaged curvature, null-space retention, and a distortion floor. PMF-CL targets Pareto-minimal forgetting (Nair et al., 2026), while continual post-training analyses study update compatibility (Wang et al., 2026). Wan et al. (2025) relate forgetting and replay benefits to task-signal angles in a two-layer convolutional neural network (CNN). Peng et al. (2026) establish recovery guarantees for dependent nonlinear regression tasks. Friedman & Meir (2026) provide data-dependent forgetting upper bounds and oracle bounds for Gibbs posteriors. Liu et al. (2026) bound interference using update spectral norms and study their control through Muon.

We do not claim priority for quadratic forgetting energies, null-space criteria, or unavoidable floors. Proposition 1 supplies the general multi-output endpoint geometry needed beyond the rank-one law: fitting fixes a common constrained component, and algorithm-dependent cost lies in the feasible null-space displacement. Theorem 3 states the additional fit and drift terms needed for approximate nonlinear transfer.

A.3 Plasticity and moving tangent kernels

Plasticity loss involves dormant features, learning rates, representation rank, and curvature (Lyle et al., 2022; Nikishin et al., 2023; Sokar et al., 2023; Lyle et al., 2024a;b; Dohare et al., 2024); interventions include warm starts, resets, plasticity injection, and continual backpropagation (Ash & Adams, 2020; Nikishin et al., 2023; Dohare et al., 2024). Spectral approaches regulate weight singular values (Lewandowski et al., 2025), promote dynamical isometry (Rosseau et al., 2026), or balance low-rank update components (Gu et al., 2026). These spectra differ from the old-task curvature in our probe.

The NTK describes dynamics in particular infinite-width limits (Jacot et al., 2018; Lee et al., 2019), while feature learning changes the kernel (Chizat et al., 2019). Appendix C states the local Taylor-error conditions needed to relate the quadratic forgetting proxy to realized old-task loss.

B Approximate fit and nonlinear networks

Exact interpolation is an analytical endpoint. The following result remains valid for finite fit error and singular $\mathcal{G}$ by adding a declared damping parameter.

Theorem 3 (Damped floor under approximate and nonlinear fit)**.** *Let $\mathcal{G} \succeq 0$, let J have full row rank, and fix $\lambda > 0$. Define*

$$H_\lambda = \mathcal{G} + \lambda I, \qquad A_\lambda = J H_\lambda^{-1} J^\top, \qquad \|x\|^2_{A_\lambda^{-1}} = x^\top A_\lambda^{-1} x.$$

For every displacement with $\|\Delta\| \le R$ and $\|J\Delta - r\| \le \varepsilon$,

$$q(\Delta) \ge \frac{1}{2}\left[\|r\|_{A_\lambda^{-1}} - \frac{\varepsilon}{\sqrt{\lambda_{\min}(A_\lambda)}}\right]_+^2 - \frac{\lambda R^2}{2}. \tag{20}$$

Now let $J = J_{\mathrm{new}}(w_0)$ and let $\mathcal{G}$ be the old-task GGN matrix at w_0. Assume the new-task Jacobian is L_J-Lipschitz in operator norm on the segment from w_0 to $w_0 + \Delta$. If $\|f(w_0 + \Delta) - y\| \le \varepsilon_{\mathrm{fit}}$, then Equation (20) holds with

$$\varepsilon = \varepsilon_{\mathrm{fit}} + \tfrac{1}{2} L_J R^2. \tag{21}$$

If, in addition, the old-task conditions of Theorem 5 hold, then realized forgetting is at least the right side of Equation (20) minus $B_F(\Delta)$.

The damping term serves two purposes. It makes H_λ invertible when neural curvature is singular, and it exposes the price of that regularization as $\lambda R^2/2$ instead of silently substituting a pseudoinverse. The bound can be vacuous. A positive right-hand side in Equation (20) lower-bounds the quadratic cost. A positive lower bound on realized forgetting requires that this right-hand side remain positive after subtracting a valid $B_F(\Delta)$.

C ERROR-CONTROLLED LOCAL DIAGNOSTICS

This appendix defines the old-task curvature and current-task tangent kernel, then proves the approximation bound relating the quadratic forgetting proxy to realized old-task loss.

C.1 LOSSES, OPERATORS, AND DISTINCT SPECTRAL OBJECTS

Let $z_i(w) \in \mathbb{R}^C$ denote logits or regression outputs on example i, and let

$$L_k(w) = \frac{1}{N_k} \sum_{i=1}^{N_k} \ell(z_i(w), y_i)$$

be the empirical loss for task k. Its concatenated output Jacobian is $J_k(w)$. At a reference w_0 reached after task k, the signed loss change under Δ is Equation (1) with $L_{\text{old}} = L_k$.

Assume the outputs z_i are twice continuously differentiable near w_0, and write $J_i = D_w z_i(w_0) \in \mathbb{R}^{C \times p}$. For a twice differentiable output loss, define

$$\mathcal{G}_k(w_0) = \frac{1}{N_k} \sum_i J_i^\top H_{\ell,i} J_i, \qquad H_{\ell,i} = \nabla_z^2 \ell(z_i(w_0), y_i).$$

When ℓ is convex in its first argument, $\mathcal{G}_k \succeq 0$. For scalar mean squared error (MSE), $L_k = (2N_k)^{-1} \|f_k(w) - y_k\|^2$ and $\mathcal{G}_k = N_k^{-1} J_k^\top J_k$.

For a current-task MSE probe, let $\rho(w) = f(w) - y$ and

$$K_t = J_t J_t^\top, \qquad J_t = J(w_t),$$

be the empirical tangent kernel. The fitting target is the complete residual curve or the number of steps needed to reduce optimizable residual energy by a declared fraction, not an eigenvalue by itself.

Remark 4 (The operators are related but distinct)**.** *For scalar MSE, $N^{-1} J^\top J$ and $J J^\top$ at the same parameters and on the same examples share nonzero eigenvalues up to N^{-1}. In continual learning, however, $\mathcal{G}_k$ concerns an old task and K_t concerns current examples at a later state. They act on different spaces, need not have the same rank, and have no canonical eigenvector correspondence. Forgetting and plasticity are not the top and bottom of one spectrum.*

C.2 DIRECTION-WISE FORGETTING ENVELOPE

Assumption 1 (Local regularity)**.** *L_k is three times continuously differentiable on an open neighborhood of $\{w_0 + u\Delta : u \in [0,1]\}$. Its third derivative, viewed as a trilinear form, has operator norm at most β on this segment, and $\|\nabla L_k(w_0)\| \le \gamma$.*

The exact Hessian decomposes as

$$\nabla^2 L_k(w_0) = \mathcal{G}_k(w_0) + R_k, \qquad R_k = \frac{1}{N_k} \sum_{i,c} [\nabla_z \ell(z_i(w_0), y_i)]_c \, \nabla_w^2 z_{ic}(w_0).$$

The residual or model-curvature matrix R_k has the same quadratic order as the GGN term.

Theorem 5 (Direction-wise forgetting envelope)**.** *Under Assumption 1, let $q_G(\Delta) = \frac{1}{2} \Delta^\top \mathcal{G}_k(w_0) \Delta$. Then*

$$|F(\Delta) - q_G(\Delta)| \le \underbrace{|\nabla L_k(w_0)^\top \Delta|}_{b_{\text{grad}}} + \underbrace{\tfrac{1}{2} |\Delta^\top R_k \Delta|}_{b_{\text{res}}} + \underbrace{\tfrac{\beta}{6} \|\Delta\|^3}_{b_{\text{nl}}} \tag{22a}$$

$$\le \gamma \|\Delta\| + \tfrac{1}{2} \|R_k\|_{\text{op}} \|\Delta\|^2 + \tfrac{\beta}{6} \|\Delta\|^3 =: B_F(\Delta). \tag{22b}$$

Consequently, $q_G - B_F \le F \le q_G + B_F$. Whenever $q_G > 0$ and $B_F \le \epsilon q_G$, we have $|F - q_G|/q_G \le \epsilon$. If $0 \le \epsilon < 1$, this also gives $|F - q_G|/|F| \le \epsilon/(1-\epsilon)$. For scalar MSE, define $\rho_{k,i} = f_i(w_0) - y_i$. Then

$$\|R_k\|_{\text{op}} \le \left(\frac{1}{N_k} \sum_i |\rho_{k,i}| \right) \sup_i \left\| \nabla_w^2 f_i(w_0) \right\|_{\text{op}}. \tag{23}$$

If Δ is orthogonal to the top-m eigenspace of $\mathcal{G}_k$, where $m \in \{0, \dots, p-1\}$ and the eigenvalues are ordered nonincreasingly, then $q_G(\Delta) \le \frac{1}{2} \lambda_{m+1}(\mathcal{G}_k) \|\Delta\|^2$.

The direction-wise form is essential. A large $\|R_k\|_{\mathrm{op}}$ need not invalidate a direction on which $\Delta^\top R_k \Delta$ is small. Conversely, comparison with $\lambda_{\max}(\mathcal{G}_k)$ is insufficient when the realized GGN energy is small. An approximately stationary old task also leaves an $O(\|\Delta\|)$ term that can dominate both quadratic terms.

Proof. Let $g_k = \nabla L_k(w_0)$ and $H_k = \nabla^2 L_k(w_0)$. Taylor's theorem along $h(u) = L_k(w_0 + u\Delta)$ gives

$$F(\Delta) = g_k^\top \Delta + \tfrac{1}{2}\Delta^\top H_k \Delta + r_3, \qquad |r_3| \le \tfrac{\beta}{6}\|\Delta\|^3. \tag{24}$$

The chain rule for one loss term yields

$$\nabla_w^2 \ell_i = J_i^\top H_{\ell,i} J_i + \sum_c [\nabla_z \ell_i]_c \nabla_w^2 z_{ic}.$$

Averaging gives $H_k = \mathcal{G}_k + R_k$. Substitution into Equation (24), followed by the triangle inequality and Cauchy–Schwarz, proves Equations (22a) and (22b). For scalar MSE, $R_k = N_k^{-1}\sum_i \rho_{k,i}\nabla_w^2 f_i$; subadditivity proves Equation (23). The eigenspace statement follows by diagonalizing $\mathcal{G}_k$ and retaining only eigenvalues at most λ_{m+1}. □

D Proofs and edge cases for the matched-fit results

D.1 Proof of Proposition 1

Proof. Matched fit and dynamics. Because J has full row rank, $B = JJ^\top \succ 0$, and $\Delta_0 = J^\top B^{-1} r = J^+ r$ is the minimum-Euclidean-norm solution of $J\Delta = r$. By assumption $A_P = JPJ^\top \succ 0$, and

$$J\Delta_P = JPJ^\top A_P^{-1} r = r.$$

For ordinary gradient flow, the residual $\rho = J\Delta - r$ obeys $\dot\rho = -B\rho$. Hence $\rho(t) \to 0$. Starting at zero, every velocity lies in $\mathrm{range}(J^\top)$, so the limiting feasible point is Δ_0. Under projected flow, $\dot\rho = -A_P\rho$, and therefore

$$\Delta(\infty) = -PJ^\top \int_0^\infty e^{-A_P t}\rho(0)\,dt = PJ^\top A_P^{-1} r = \Delta_P.$$

For projected gradient descent, $\rho_{t+1} = (I - \eta A_P)\rho_t$. The stated step-size condition places every eigenvalue of $I - \eta A_P$ strictly inside the unit disk.

Norms and capacity. Using $P^\top = P = P^2$,

$$\|\Delta_P\|^2 = r^\top A_P^{-1} JP^2 J^\top A_P^{-1} r = r^\top A_P^{-1} r,$$

and similarly $\|\Delta_0\|^2 = r^\top B^{-1} r$. For every x,

$$x^\top A_P x = \|PJ^\top x\|^2 \le \|J^\top x\|^2 = x^\top B x,$$

so $A_P \preceq B$. Since both matrices are positive definite, inversion reverses the Loewner order: $B^{-1} \preceq A_P^{-1}$. It follows that $c_P(r) \in (0, 1]$ and that projection cannot shorten the terminal displacement. Moreover, $A_P^{-1} - B^{-1} \succeq 0$; its quadratic form vanishes at r if and only if $(A_P^{-1} - B^{-1})r = 0$, which gives the equality condition.

Terminal ratio. Quadratic homogeneity and the definition of $c_P(r)$ give

$$\frac{q(\Delta_P)}{q(\Delta_0)} = \frac{\|\Delta_P\|^2}{\|\Delta_0\|^2}\frac{q(\Delta_P/\|\Delta_P\|)}{q(\Delta_0/\|\Delta_0\|)} = \frac{1}{c_P(r)G_{\mathrm{end}}(P, r)},$$

which proves Equation (8). Substituting the two endpoints into $q(x) = \frac{1}{2}x^\top \mathcal{G} x$ also gives Equation (9). Since $\mathcal{G} \succ 0$,

$$\tfrac{1}{2}\lambda_{\min}(\mathcal{G})\|x\|^2 \le q(x) \le \tfrac{1}{2}\lambda_{\max}(\mathcal{G})\|x\|^2.$$

Therefore

$$\frac{q(\Delta_P)}{q(\Delta_0)} \le \kappa(\mathcal{G})\frac{\|\Delta_P\|^2}{\|\Delta_0\|^2} = \frac{\kappa(\mathcal{G})}{c_P(r)}.$$

The Rayleigh bounds $r^\top A_P^{-1} r \le \|r\|^2 / \lambda_{\min}(A_P)$ and $r^\top B^{-1} r \ge \|r\|^2 / \lambda_{\max}(B)$ yield the second inequality in Equation (10).

Common floor and null-space excess. Minimize q subject to $J\Delta = r$. The Lagrangian $\frac{1}{2}\Delta^\top \mathcal{G}\Delta - \nu^\top(J\Delta - r)$ has stationarity $\mathcal{G}\Delta = J^\top \nu$. Substitution into the constraint gives $\nu = (J\mathcal{G}^{-1}J^\top)^{-1} r$ and the unique minimizer $\Delta_{\mathcal{G}}$. Every other feasible point is $\Delta_{\mathcal{G}} + n$ for a unique $n \in \mathrm{null}(J)$. Because $\mathcal{G}\Delta_{\mathcal{G}} = J^\top \nu$, $n^\top \mathcal{G}\Delta_{\mathcal{G}} = (Jn)^\top \nu = 0$. Expanding the quadratic proves Equation (12).

Row-space invariance. The restriction of J to $\mathrm{range}(J^\top)$ is injective: two feasible points in that subspace differ by an element of $\mathrm{range}(J^\top) \cap \mathrm{null}(J) = \{0\}$. Since $J^+ r$ is feasible and lies in $\mathrm{range}(J^\top)$, it is the only exact-fit endpoint accessible to a trajectory that never leaves the row space. Scalar-step gradient updates remain there; a general matrix preconditioner need not.

Rank-one specialization. With $J = j^\top$, $A_P = \|Pj\|^2$ and $B = \|j\|^2$, so $c_P(r) = \|Pj\|^2 / \|j\|^2 = a(P)$. The endpoints are $\Delta_0 = rj/\|j\|^2$ and $\Delta_P = rPj/\|Pj\|^2$. Their normalized directions are $j/\|j\|$ and $Pj/\|Pj\|$ up to the common sign of r, so $G_{\mathrm{end}}(P, r) = G_{\mathrm{probe}}$. Homogeneity gives

$$\frac{q(\Delta_P)}{q(\Delta_0)} = \frac{\|j\|^2}{\|Pj\|^2} \frac{q(Pj/\|Pj\|)}{q(j/\|j\|)} = \frac{1}{a(P) G_{\mathrm{probe}}}.$$

The free and projected residuals decay at rates $JJ^\top$ and $JPJ^\top = a(P)JJ^\top$, respectively, which proves Equation (14). □

D.2 The singular-curvature quotient

The substitution $\mathcal{G}^{-1} \mapsto \mathcal{G}^+$ in Equation (11) is false when a zero-curvature direction can absorb some of the fitting requirement. The exact semidefinite statement must retain feasibility inside the positive-curvature range.

Proposition 6 (Feasibility-restricted semidefinite floor)**.** *Let $\mathcal{G} \succeq 0$ and assume $J\Delta = r$ is feasible. Then*

$$\min_{J\Delta = r} q(\Delta) = \frac{1}{2} \min_{\substack{k \in \ker(\mathcal{G}) \\ r - Jk \in \mathrm{range}(J(\mathcal{G}^+)^{1/2})}} (r - Jk)^\top (J\mathcal{G}^+ J^\top)^+ (r - Jk). \tag{25}$$

The value is zero if and only if $r \in J\ker(\mathcal{G})$.

Proof. Write a feasible displacement as $\Delta = k + \Delta_+$ with $k \in \ker(\mathcal{G})$ and $\Delta_+ \in \mathrm{range}(\mathcal{G})$. For fixed k, substitute $\Delta_+ = (\mathcal{G}^+)^{1/2} x$ with $x \in \mathrm{range}(\mathcal{G})$. Then $q = \frac{1}{2}\|x\|^2$ and the remaining constraint is $J(\mathcal{G}^+)^{1/2} x = r - Jk$. It is solvable exactly under the range restriction in Equation (25); its minimum-norm solution gives the displayed quadratic. Minimizing over feasible k proves the equality. On the range of $J(\mathcal{G}^+)^{1/2}$ the pseudoinverse quadratic is zero only at zero, which proves the final statement. □

Two counterexamples explain both qualifications. With $\mathcal{G} = \mathrm{diag}(1, 0)$, $J = (1\ 1)$, and $r = 1$, the exact-fitting point $(0, 1)$ has $q = 0$, while naive pseudoinverse substitution gives $1/2$. More subtly, let

$$\mathcal{G} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad r = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The unique feasible point is $(-1, 1)$, with $q = 1/2$. Dropping the range restriction lets the pseudoinverse ignore an infeasible residual component and incorrectly returns zero. These examples demonstrate why the range restriction is necessary.

D.3 Proof of Proposition 2

Proof. Every fit-preserving correction has the form $n = Zz$. Expanding gives

$$q(\Delta + Zz) = q(\Delta) + z^\top Z^\top \mathcal{G}\Delta + \tfrac{1}{2} z^\top Z^\top \mathcal{G} Z z.$$

This is a convex quadratic. Its normal equations are consistent because, with $A = \mathcal{G}^{1/2} Z$,

$$\mathrm{range}(Z^\top \mathcal{G}) \subseteq \mathrm{range}(A^\top) = \mathrm{range}(A^\top A) = \mathrm{range}(Z^\top \mathcal{G} Z).$$

The minimum-norm stationary solution is $z^\star = -(Z^\top \mathcal{G} Z)^+ Z^\top \mathcal{G} \Delta$, which proves Equation (15). As z ranges over $\mathbb{R}^d$, $\Delta + Zz$ ranges over the full affine set that preserves $J\Delta$. The solution is therefore optimal over all fit-preserving corrections. If $\mathcal{G} \succ 0$, the objective is strictly convex on that affine set. For an exact-fitting Δ, Proposition 1 identifies the unique minimizer as $\Delta_{\mathcal{G}}$. □

D.4 Vanishing-capacity statements

Theorem 7 (Vanishing-capacity wedge). *Let u be a unit eigenvector corresponding to the largest eigenvalue of $\mathcal{G}$, and let $P = I - uu^\top$. For every $M, W > 1$, there is a two-dimensional scalar-output instance for which projected gradient descent from zero with $0 < \eta < 2/(JPJ^\top)$ converges to exact fit and simultaneously*

$$\mathsf{Gain}(P) \geq M, \qquad \frac{q(\Delta_P)}{q(J^+ r)} \geq W. \tag{26}$$

Corollary 8 (Fixed-gain vanishing-capacity limit). *Fix $\Gamma > 1$, $W > 1$, and $\alpha \in (0, 1]$. There is a two-dimensional instance with $G_{\mathrm{probe}} = \Gamma$ and $c_P(r) \leq 1/(\Gamma W)$ whose matched-fit endpoints obey*

$$J\Delta_0(\alpha) = J\Delta_P(\alpha) = \alpha r, \qquad \frac{q(\Delta_P(\alpha))}{q(\Delta_0(\alpha))} = \frac{1}{c_P(r)\Gamma} \geq W.$$

Consequently, normalized gain alone cannot uniformly upper-bound the matched-fit terminal ratio over this class.

Corollary 9 (The wedge is not isolated). *For fixed finite $M, W > 1$, the construction can be chosen with $\mathcal{G} \succ 0$, a simple largest eigenvalue, and both inequalities in Equation (26) strict. The same ordering then holds in a sufficiently small neighborhood of $(\mathcal{G}, J, r)$.*

D.5 Proof of Theorem 7

Proof. Write $A = JPJ^\top$. Since P is an orthogonal projector and $A \succ 0$, the recursion initialized at zero has the closed form

$$\Delta_t = PJ^\top A^{-1}\left[I - (I - \eta A)^t\right] r.$$

The step-size condition places every eigenvalue of $I - \eta A$ in $(-1, 1)$. Thus $\Delta_t \to PJ^\top A^{-1} r$ and $J\Delta_t \to r$, proving the general convergence statement.

Set $\mu = 1$ without loss of generality. Given $M, W > 1$, first choose $\lambda \geq 2M\mu$. Then choose $\phi \in (0, \pi/4]$ so that

$$\sin^2 \phi \leq \frac{\mu}{W(\lambda + \mu)}.$$

Take $J = e_1^\top$, $r = 1$, and

$$\mathcal{G} = \lambda u u^\top + \mu v v^\top, \qquad u = \cos\phi\, e_1 + \sin\phi\, e_2, \qquad v = -\sin\phi\, e_1 + \cos\phi\, e_2.$$

Since $Pe_1 = e_1 - \cos\phi\, u = -\sin\phi\, v$, $JPJ^\top = \sin^2\phi > 0$. The projected residual contracts by $1 - \eta\sin^2\phi$ and the terminal point is $\Delta_P = -v/\sin\phi$.

Using $e_1 = \cos\phi\, u - \sin\phi\, v$ gives the exact ratios

$$\mathsf{Gain}(P) = \frac{\lambda\cos^2\phi + \mu\sin^2\phi}{\mu},$$
$$\frac{q(\Delta_P)}{q(e_1)} = \frac{\mu}{\sin^2\phi\,(\lambda\cos^2\phi + \mu\sin^2\phi)}.$$

Since $\cos^2\phi \geq 1/2$, the first ratio is at least M. Since $\lambda\cos^2\phi + \mu\sin^2\phi \leq \lambda + \mu$, the second ratio is at least W. Both inequalities hold under the same parameter choice. □

D.6 Proof of Corollary 8

Proof. Set $\mu = 1$ and write $s = \sin^2\phi$. Choose $s \in (0, 1/2)$ such that $s \leq 1/(GW)$, and set

$$\lambda = \frac{G - s}{1 - s}.$$

Because $G > 1$, we have $\lambda > 1$, so u remains the unique top eigenvector. The exact expression in the proof of Theorem 7 gives

$$\mathsf{Gain}(P) = \lambda(1-s) + s = G, \qquad \frac{q(\Delta_P)}{q(e_1)} = \frac{1}{sG} \geq W.$$

For any $\alpha \in (0, 1]$, define $\Delta_0(\alpha) = \alpha e_1$ and $\Delta_P(\alpha) = -\alpha v/\sqrt{s}$. Since $J = e_1^\top$ and $e_1^\top v = -\sqrt{s}$, both displacements satisfy $J\Delta_0(\alpha) = J\Delta_P(\alpha) = \alpha r$. Their quadratic energies scale by the same factor α^2, so their ratio remains $1/(sG) \geq W$. For $0 < \alpha < 1$, these points also lie on the corresponding unit-rate gradient-flow trajectories, at times $t_0 = -\log(1-\alpha)$ and $t_P = -\log(1-\alpha)/s$, respectively. The case $\alpha = 1$ is their limiting exact fit; exact first-hitting times for a fixed discrete step size are not asserted here. For any radius $R > 0$, replace $r = 1$ by $r_R = R\sqrt{s}$ and set $\Delta_{0,R}(\alpha) = \alpha R\sqrt{s}\, e_1$ and $\Delta_{P,R}(\alpha) = -\alpha R v$. Both displacements have norm at most R and satisfy $J\Delta = \alpha r_R$. Their common rescaling leaves the matched-norm gain and the terminal quadratic ratio unchanged. If a finite gain-only upper bound $C(G)$ existed, choosing $W > C(G)$ would contradict this construction. □

D.7 Proof of Corollary 9

Choose λ and ϕ so that the two ratios exceed M and W with strict margin. On the domain $Pg \neq 0$ and $JPJ^\top \succ 0$, eigenvectors can be selected continuously around a simple top eigenvalue, matrix inversion is continuous, and both ratios are continuous functions of $(\mathcal{G}, J, r)$. Positive definiteness is also open. A sufficiently small neighborhood therefore preserves both strict inequalities.

D.8 Proof of Theorem 3

Proof. Let $e = r - J\Delta$, so $\|e\| \leq \varepsilon$ and $J\Delta = r - e$. For a fixed output change $r - e$, the minimum H_λ-energy displacement is $H_\lambda^{-1} J^\top A_\lambda^{-1}(r-e)$. Therefore

$$\Delta^\top H_\lambda \Delta \geq \|r-e\|^2_{A_\lambda^{-1}}.$$

The reverse triangle inequality and the Euclidean error bound give

$$\|r-e\|_{A_\lambda^{-1}} \geq \left[\|r\|_{A_\lambda^{-1}} - \|e\|_{A_\lambda^{-1}}\right]_+ \geq \left[\|r\|_{A_\lambda^{-1}} - \frac{\varepsilon}{\sqrt{\lambda_{\min}(A_\lambda)}}\right]_+.$$

Since $q(\Delta) = \frac{1}{2}\Delta^\top H_\lambda \Delta - \frac{\lambda}{2}\|\Delta\|^2$ and $\|\Delta\| \leq R$, Equation (20) follows.

For the nonlinear statement, the fundamental theorem of calculus gives

$$f(w_0 + \Delta) - f_0 - J\Delta = \int_0^1 [J(w_0 + t\Delta) - J]\Delta\, dt.$$

The Lipschitz assumption bounds this remainder by $L_J\|\Delta\|^2/2$. Combining it with the realized fit error proves Equation (21). Finally, Theorem 5 gives $F(\Delta) \geq q(\Delta) - B_F(\Delta)$. □

E Experimental protocols

A seed is the independent unit. Calibration seeds select thresholds and hyperparameters; held-out test seeds are evaluated once after calibration. Forgetting is the signed change F in Equation (1). Target misses, backward-transfer cases, rejected corrections, and numerical failures remain in their declared denominators.

E.1 CONTROLLED MEASUREMENT PROTOCOL

Table 4: Two controlled checks run on separate seeded networks. The local-fidelity row reports the regression of realized old-task loss changes on the quadratic. The remaining rows compare the unprojected and projected arms within the projection configuration.

Measurement	Value	Interpretation
Local-fidelity check	slope 1.0012, $R^2 = 0.9997$	near-unit fitted slope
Matched-norm quadratic	$0.010197 \to 0.002531$	$4.03\times$ lower quadratic
Realized one-step loss change	$0.010302 \to 0.002480$	$4.15\times$ lower realized change
Final forgetting	$0.954976 \to 0.954442$	decrease of 5.34×10^{-4}
Forgetting per progress	$1.19628 \to 1.19767$	larger loss change per unit progress

E.2 SECONDARY PERMUTED MNIST COMPARISON

Four calibration seeds set the progress threshold to 0.8 times the smallest default-pass frontier maximum attained by any arm, yielding $\pi^\star = 2.5898$. The threshold and spectral-projection contrast are fixed before evaluation on sixteen held-out seeds. The six arms are no projection, spectral projection, EWC, GPM, replay, and continual backpropagation; each receives an equal four-point grid over learning rate and its method-specific regularization control.

For each seed, forgetting is measured at the first recorded step at which an arm reaches $\pi^\star$. All six arms reach the threshold on all sixteen seeds. Table 5 reports paired mean differences against no projection. Intervals are paired seed-bootstrap intervals with 10^4 resamples.

Table 5: Permuted MNIST forgetting at a common progress threshold. Positive mean differences indicate a larger signed old-task loss change than plain fine-tuning.

Method	Mean difference	Paired 95% CI	Role
Spectral projection	$+0.0015$	$[-0.008, +0.011]$	primary
Replay	-0.2565	$[-0.281, -0.232]$	secondary
GPM	-0.0159	$[-0.028, -0.005]$	secondary
EWC	≈ 0	$[-0.006, +0.006]$	secondary
Continual backpropagation	-0.0009	$[-0.009, +0.008]$	secondary

The projection contrast's confidence interval includes zero in this comparison, while replay produces the largest reduction among the secondary comparisons. First threshold crossing need not equalize overshoot exactly, so this is a common-threshold diagnostic rather than an exact instantiation of $J\Delta = r$.

E.3 LARGE-NETWORK PROBE PROTOCOL

The trained-network probe study contains twenty seeds for each of five architecture–dataset configurations. Each run projects the first update away from the leading $k \in \{1, 4, 16, 64\}$ eigenvectors of the old-task GGN and rescales it to match the free-update norm within 10^{-5}. Probe gains are ratios of the matched-norm quadratic values in Equation (2). The intervals in Table 10 use 10^4 seed bootstrap resamples within each configuration and rank.

All one hundred configuration–seed units ran, but the analysis includes only pairs passing the prespecified terminal-fit gate. The paper therefore treats the table as an eligible-subset estimate. Two VGG-11-BN records with negative projected GGN quadratics fail the positive-semidefinite operator check and are excluded rather than clamped. The free-arm terminal change is non-positive in $192/400$ rank records, so terminal-damage ratios are not reported for this eligible subset.

Table 6: Rank-16 terminal comparison at a common progress threshold. Probe entries are median matched-norm gains. We define $D_F = F(\Delta_{\text{proj}}) - F(\Delta_{\text{free}})$; $D_F > 0$ indicates a larger signed old-task loss change under projection. Mean D_F and the final column use matched first-crossing pairs. The aggregate row summarizes all 99 matched pairs.

Configuration	Matched	Probe gain	Mean D_F	$D_F > 0$
Split CIFAR-100, ResNet-18	$20/20$	61.8	$+0.018$	$16/20$
Split CIFAR-100, VGG-11-BN	$20/20$	46.2	$+0.012$	$15/20$
Tiny ImageNet, ResNet-18	$19/20$	287.4	$+0.025$	$15/19$
Tiny ImageNet, VGG-11-BN	$20/20$	38.7	$+0.009$	$14/20$
Permuted MNIST, MLP	$20/20$	12.6	$+0.004$	$13/20$
Aggregate	$99/100$	48.9	$+0.013$	$73/99$

Table 7: Quadratic-floor evaluation over 300 configuration–seed–damping cells. "Positive" counts strictly positive computed bounds. Floor/q is the aggregate computed-floor to realized-quadratic ratio.

Configuration	Eligible	Positive	Violations	Floor/q
Split CIFAR-100, ResNet-18	$60/60$	$48/60$	$0/60$	0.36
Split CIFAR-100, VGG-11-BN	$60/60$	$45/60$	$0/60$	0.33
Tiny ImageNet, ResNet-18	$60/60$	$43/60$	$0/60$	0.29
Tiny ImageNet, VGG-11-BN	$60/60$	$41/60$	$0/60$	0.28
Permuted MNIST, MLP	$60/60$	$52/60$	$0/60$	0.41
Aggregate	$300/300$	$229/300$	$0/300$	0.34

E.4 Five-configuration terminal and quadratic-floor protocols

The terminal study uses four calibration seeds and twenty test seeds per architecture–dataset configuration. Rank $k = 16$ is primary, with $k \in \{1, 4, 64\}$ secondary. Calibration fixes each configuration's progress target and tuning choices before the test seeds are evaluated. Terminal outcomes use paired first crossings; misses remain visible in the $99/100$ matched count.

The separate quadratic-floor evaluation covers five configurations, twenty seeds, and three damping values, for 300 cells. It uses a positive-semidefiniteness-preserving factorization, double-precision arithmetic, and a 10^{-6} relative solve tolerance.

The $0/300$ count has an exact two-sided 95% Clopper–Pearson interval of $[0, 1.22\%]$ for the cell-level violation rate. This validates the evaluated quadratic inequality, not nonlinear realized forgetting.

E.5 Rank-capacity and fixed-rank protocols

For every rank cell, the new-task batch contains $N = 256$ examples fixed before inspecting terminal outcomes. We form $B = JJ^\top$ and $A_{P_k} = JP_kJ^\top$ in double precision. Cells require a positive-definite factorization after the declared numerical jitter; failures remain visible. The exact terminal ratio uses Equation (9). Configuration-by-rank fixed-effect analyses preserve seed pairing, and intervals use a cluster bootstrap that resamples seeds within each configuration–rank cell. For the partial-rank association analysis, we residualize both the terminal outcome and each predictor with respect to configuration-by-rank fixed effects and compute Spearman's rank correlation between the resulting residuals. Confidence intervals are obtained by resampling seeds within configuration–rank cells and repeating the complete residualization and correlation procedure.

For the fixed-k experiment, 256 random rank-16 subspaces are drawn per held-out Permuted MNIST seed. Candidate pairs are selected using only the probe gain and $c_P(r)$ matching criteria. Terminal outcomes are evaluated after the pair is frozen.

For two projectors evaluated with the same J, r, and old-task quadratic, let $\Delta_{P_i} = P_iJ^\top(JP_iJ^\top)^{-1}r$. Under the assumptions of Proposition 1,

$$\frac{q(\Delta_{P_1})}{q(\Delta_{P_2})} = \frac{c_{P_2}(r)G_{\text{end}}(P_2, r)}{c_{P_1}(r)G_{\text{end}}(P_1, r)}.$$

Table 8: Descriptive terminal-correction results under the all-seed intention-to-correct analysis. Acceptance is assessed on the same held-out split used for outcome reporting. Entries are paired means; brackets are paired bootstrap 95% intervals for the TNC reduction. Rejected corrections retain their original endpoints under the intention-to-correct analysis.

Dataset	Method	Original	Shrinkage	TNC	Reduction [95% CI]
Permuted MNIST	Fine-tuning	0.301	0.292	0.273	0.028 [0.020, 0.036]
	Projection	0.302	0.294	0.274	0.028 [0.020, 0.036]
	EWC	0.300	0.293	0.279	0.021 [0.014, 0.028]
	GPM	0.284	0.279	0.265	0.019 [0.012, 0.026]
	Replay	0.044	0.043	0.041	0.003 [0.001, 0.006]
Split CIFAR-100	Fine-tuning	0.194	0.184	0.169	0.025 [0.017, 0.033]
	Projection	0.188	0.179	0.166	0.022 [0.014, 0.030]
	EWC	0.171	0.165	0.156	0.015 [0.009, 0.021]
	GPM	0.154	0.149	0.142	0.012 [0.006, 0.018]
	Replay	0.112	0.110	0.107	0.005 [0.002, 0.008]

Table 9: Damped realized-output decomposition at matched progress. Entries are median shares over seeds.

	Permuted MNIST		Split CIFAR-100	
Method	Floor	Excess	Floor	Excess
Fine-tuning	0.44	0.56	0.39	0.61
Projection	0.40	0.60	0.34	0.66
EWC	0.56	0.44	0.52	0.48
GPM	0.64	0.36	0.61	0.39
Replay	0.90	0.10	0.85	0.15

At equal capacity, this ratio reduces to $G_{\mathrm{end}}(P_2, r)/G_{\mathrm{end}}(P_1, r)$. The identity concerns exact linearized endpoints; Table 2 reports realized forgetting at common progress thresholds.

E.6 Terminal correction protocol

The correction study uses sixteen held-out Permuted MNIST seeds and twenty held-out Split-CIFAR-100 seeds. Fine-tuning, projection, EWC, GPM, and replay use the same threshold, tuning budget, minibatch order, and first-crossing rule within a seed. The old-task curvature batch and new-task Jacobian batch are fixed before correction. Both low-rank operators use rank 64.

TNC is accepted when realized new-task progress on the same held-out evaluation split used for outcome reporting changes by at most 0.5%. The primary analysis follows an intention-to-correct principle: rejected updates retain the original endpoint. Because acceptance and outcome reporting use the same held-out split, the resulting reductions are descriptive and test-conditioned rather than untouched-test estimates. The matched-loss shrinkage coefficient is selected on the calibration split by one-dimensional search and then evaluated on the held-out evaluation split. Paired intervals use 10^4 seed-bootstrap resamples. We record pre- and post-correction new-task loss, acceptance, shrinkage coefficient, old-task loss, solve residual, runtime, and peak memory for every seed.

E.7 Realized-output floor-decomposition protocol

For each terminal displacement, $r_\Delta = J\Delta$ is computed on the fixed new-task batch. The decomposition uses $\lambda = 10^{-3}\,\mathrm{tr}(\mathcal{G})/p$, selected on calibration seeds and frozen before test evaluation. The factorized systems are solved in double precision to relative residual 10^{-6}. The main denominator includes all terminal points; any failed solve would remain a failed cell rather than being removed.

Table 10: Seed-level uncertainty for the probe gains. Entries are medians with bootstrap 95% confidence intervals (10^4 resamples); n is the number of pairs summarized at that rank, after excluding the two negative-quadratic VGG-11-BN records. Only the three non-MLP configurations are shown; the low or zero eligible-pair counts for the two MLP configurations are discussed in the text. Terminal-damage ratios are not reported because their denominators are frequently non-positive. The Tiny ImageNet intervals are wide because $n \leq 9$.

Configuration	k	n	Probe gain, median [95% CI]
Split CIFAR-100, ResNet-18	1	20	3.32 [2.25, 4.86]
	4	20	12.27 [8.86, 18.81]
	16	20	63.51 [37.81, 117.76]
	64	20	359.87 [238.92, 1532.45]
Split CIFAR-100, VGG-11-BN	1	20	3.09 [1.34, 7.37]
	4	20	12.99 [9.49, 22.63]
	16	19	44.08 [30.41, 76.22]
	64	19	235.78 [143.59, 1033.18]
Tiny ImageNet, ResNet-18	1	9	4.56 [1.76, 17.02]
	4	9	34.21 [16.82, 67.34]
	16	9	301.65 [116.74, 744.67]
	64	8	1220.84 [731.57, 12775.37]

F ELIGIBLE-SUBSET PROBE STUDY

The trained-network test ran five architecture–dataset configurations: ResNet-18, VGG-11-BN, and a multilayer perceptron (MLP) on Split CIFAR-100, and ResNet-18 and the MLP on Tiny ImageNet. Each configuration used twenty seeds. Within every run, the intervention projected updates away from the leading $k \in \{1, 4, 16, 64\}$ eigenvectors of the old-task GGN. The first projected update was rescaled to match the free update norm to 10^{-5}, and the matched-norm probe was evaluated on the task-0 split. All one hundred configuration–seed units were executed. The analysis retains only pairs satisfying the prespecified terminal-fit gate. That gate is not required to compute the local probe, so the counts describe an eligible matched-pair subset, not all executed units. Section 5.2 reports a separate study with its own configurations and threshold rule.

We report probe-gain magnitudes only. The intended terminal comparison divided the projected terminal damage by the free-arm terminal damage, but that ratio is ill-posed here. The free-arm terminal damage is non-positive in 192 of 400 rank records, because the old task improved rather than forgot, and a ratio on a non-positive denominator has no forgetting interpretation. For these records, we report no terminal-damage ratio or reversal count and make no trained-network analogue of Theorem 7. Two VGG-11-BN seed-15 records additionally have negative projected GGN quadratics, which is impossible for a true GGN; those records are rejected as operator failures rather than clamped, and they enter no probe-gain summary.

Table 10 attaches seed-level uncertainty to the probe gains. Among eligible pairs, the median reaches $359.87\times$ on Split CIFAR-100 ResNet-18, $235.78\times$ on VGG-11-BN, and $1220.84\times$ on Tiny ImageNet ResNet-18. The two MLP configurations are omitted because they contain only two to three eligible pairs per rank or none; the free arm misses the fit gate in all $20/20$ Split CIFAR-100 MLP seeds. The measurements establish that the local quadratic can change by orders of magnitude on the eligible subset. They do not by themselves establish a relationship with terminal forgetting.

The VGG-11-BN analysis excludes the two records with negative projected GGN quadratics. Including those records changes the reported median from 44.08 to 41.94 at $k = 16$ and from 235.78 to 221.44 at $k = 64$; they are excluded under the stated operator-validity rule.

A rank-one projection removes one parameter-space direction, but its effect depends on both gradient alignment and curvature. The controlled cell loses 2.6×10^{-4} of squared gradient norm; the trained cells lose a median fraction of 0.156 and a maximum of 0.629. A small removed fraction alone does not bound the change in the quadratic probe. For example, with $\mathcal{G} = \mathrm{diag}(10^6, 1)$ and unit gradient $g = (\sqrt{0.001}, \sqrt{0.999})^\top$, removing the first coordinate deletes only 0.1% of squared gradient norm

but yields a matched-norm probe gain of 1000.999. A quantitative guarantee also requires control of curvature along the removed and retained directions.