# BEYOND ADAPTER RANK: THE SHARED-SUBSPACE COMPLEXITY OF LORA TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Low-rank adaptation constrains each weight update, but its training checkpoints need not share a small input subspace. This distinction matters whenever one fixed representation must describe a training history or preserve its predictions. We study the minimum shared input dimension of a LoRA path and how the weighting of checkpoints affects its approximation. In aligned linear regression, simultaneous gradient descent produces rank-one checkpoints whose shared hidden input space has dimension $d - 1$, even though the initialization space already contains a solution. We also derive the exact minimax error for shared linear measurements when the observer does not know the probe orientation. Spectral approximation, finite-horizon bounds, and a conditional network perturbation bound connect these results to measurable interventions. Across 24 language-model fine-tuning runs, paths move substantially outside the initial input space, yet a few shared directions retain 99% of their energy. Joint projections across adapted matrices approximately preserve aggregate prediction quality in language modeling and arithmetic reasoning, and outperform random directions at the same dimension. Similar aggregate accuracy nevertheless coexists with changes in individual decisions. These results distinguish instantaneous rank, shared path dimension, and predictive fidelity as separate properties of low-rank training.

## 1 INTRODUCTION

Low-rank adaptation (LoRA) reduces the cost of adapting a pretrained model by representing each weight update as a product of two small matrices (Hu et al., 2022). Rank controls the number of trainable parameters and underlies methods for rank allocation, quantization, and initialization (Zhang et al., 2023; Dettmers et al., 2023; Hayou et al., 2024a). Yet the same rank can describe very different learning histories. A rank-one update can retain one input direction throughout training or continually change that direction. These paths have the same instantaneous parameter budget, but a fixed representation of their checkpoints may require very different numbers of input directions.

This distinction matters whenever we use one subspace across training. A direction selected from the final update can represent that checkpoint exactly while missing earlier predictions. A subspace estimated early can capture most later state energy even as the model continues to move in new directions. Measurements and interventions based on a shared subspace therefore need a notion of complexity that includes the training history. Adapter rank alone cannot determine the dimension of that representation, and endpoint accuracy cannot establish that it preserves behavior throughout the path. We ask: *how many shared input directions does a LoRA training path require, and what does retaining those directions preserve?*

Existing accounts of low-dimensional adaptation leave this question open. Intrinsic-dimension studies identify small coordinate spaces in which optimization can reach useful solutions (Li et al., 2018; Aghajanyan et al., 2021). Analyses of LoRA dynamics show that trainable factors can learn features beyond their initial representation (Dayi & Chen, 2025; Hayou et al., 2024b). A low-dimensional solution, however, need not contain the optimizer's path. Nor does substantial departure from initialization imply that many new directions carry substantial energy. We separate three properties: the rank of each checkpoint, the input dimension shared across checkpoints, and the prediction



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.

changes caused by discarding directions. Each property answers a different question about what low-rank training represents.

Our analysis acts on the realized update matrices rather than on a particular choice of factors. We isolate the component outside the initial input space and stack it across checkpoints with explicit weights. The rank of this stack gives the exact shared dimension, while its singular values give the dimension needed at a chosen approximation tolerance. This construction is invariant under factor reparameterization. It also specifies what the representation shares: one input projector across time, rather than coordinates of vectorized weight matrices. The distinction lets us compare complete-path, endpoint, and early-path subspaces under the same direction budget.

The main theoretical challenge is to show that actual optimization can create a large shared dimension. Arbitrarily chosen rotating matrices do not establish such a claim about LoRA. We construct a simultaneous gradient-descent run whose rank-one checkpoints collectively use all $d - 1$ directions outside initialization, although the target is attainable with the initial input factor frozen. The active feature eigenvalues determine the exact dimension. We then derive the optimal worst-case recovery error when all probes share a set of linear measurements and their hidden orientation is unknown. This result gives the path spectrum a precise operational meaning beyond its standard role in low-rank approximation.

The empirical question is whether the directions that distinguish exact rank also matter at practical tolerances. Across 24 fine-tuning runs, the hidden-to-total path norm reaches 81.5–86.4%, yet the group summaries of the dimension needed for 99% energy remain at 1.67–3. These paths substantially leave initialization while learning a small shared input space. Joint interventions across every adapted matrix approximately retain mean loss and aggregate arithmetic accuracy, whereas random directions of the same dimension perform worse. Similar aggregate accuracy still accompanies changes in individual answers. The experiments thus test the learned directions through the complete network instead of treating matrix approximation as a guarantee of prediction fidelity.

We make three contributions:

- We prove that simultaneous gradient descent can produce rank-one checkpoints with shared hidden dimension $d - 1$, and derive the exact recovery error for a specified shared-measurement protocol (Section 3).
- We connect approximation to checkpoint weighting and finite training horizons, and state the activation and sensitivity conditions that convert matrix error into a prediction-loss bound (Section 4).
- We measure substantial subspace departure with concentrated spectra and test the dominant directions through matched-budget, joint network interventions on language modeling and arithmetic reasoning (Section 5).

## 2 Related Work

**Efficient adaptation and update geometry.** Adapters reduce the number of trainable parameters in transfer learning (Houlsby et al., 2019), and LoRA implements this reduction through low-rank weight corrections (Hu et al., 2022). AdaLoRA allocates rank budgets across matrices (Zhang et al., 2023), while QLoRA combines adapter training with quantized base weights (Dettmers et al., 2023). Initialization and factor-specific learning rates alter LoRA dynamics (Hayou et al., 2024a;b). Comparisons with full fine-tuning reveal differences in learning, forgetting, and the singular directions of trained weights (Biderman et al., 2024; Shuttleworth et al., 2025). We study the input space shared across checkpoints, which captures a different property from the cost or spectrum of an individual update.

**Low-dimensional learning and feature dynamics.** Intrinsic dimension measures how few coordinates suffice to reach useful solutions in a restricted optimization space (Li et al., 2018; Aghajanyan et al., 2021). Low-dimensional structure also appears in prediction-distribution trajectories across deep networks (Mao et al., 2024). Kernel analyses describe language-model fine-tuning near an initialization (Malladi et al., 2023), and LoRA has tractable optimization properties in a neural-tangent-kernel regime (Jang et al., 2024). Low-rank fine-tuning can also span lazy and feature-learning regimes

(Dayi & Chen, 2025), and its factor dynamics affect feature stability (Wu et al., 2026). These viewpoints concern reachable solutions, prediction manifolds, or learning regimes. We instead ask for the smallest fixed linear input space containing, or approximating, the particular matrix path actually visited.

**Changing subspaces and trajectory reuse.** ReLoRA accumulates high-rank updates by merging and resetting low-rank factors (Lialin et al., 2024). Flora compresses gradients through resampled low-rank projections (Hao et al., 2024). Gradient subspace stabilization can inform low-rank weight representations (Jaiswal et al., 2025). Recent work models low-rank trajectories to accelerate reinforcement learning through nonlinear prediction or rank-one extrapolation (Chen et al., 2026; Wei et al., 2026). Our exact separation uses one uninterrupted run whose instantaneous and final updates remain low-rank. Our experiments compare subspaces fitted to the complete observed path with subspaces estimated earlier and reused at later checkpoints. Both represent measured states rather than forecast unseen checkpoints.

**Compression and behavioral fidelity.** Data-aware singular-value compression accounts for how weight errors interact with input activations (Wang et al., 2025). Earlier compression research showed that small aggregate accuracy changes can conceal heterogeneous effects on examples (Hooker et al., 2019). A recent public research abstract explicitly studies decision changes under low-rank language-model compression (Yang et al., 2026). Our comparison uses the public abstract hosted by the authors, as a complete manuscript was not available for verification. Our generation analysis is an application boundary for path-based adapter projections. Our main contribution concerns shared path geometry, while the distinction between aggregate accuracy and individual agreement is established.

**Spectral approximation and information.** Classical low-rank approximation gives the optimal shared-projector identity (Eckart & Young, 1936). Eigenvalue extremality (Fan, 1949) and optimal-recovery theory (Hinrichs et al., 2021) provide the mathematical tools for the measurement analysis. We treat these tools as established results. Our results establish the separation through actual factor dynamics and derive the exact population-weighted error bound for the specified LoRA probe-information class. This explicit information model defines the scope of the lower bound.

## 3 Pointwise Rank and Pathwise Complexity

To describe a training history with one input subspace, we must first identify the directions each checkpoint uses. We define this representation for a single adapted linear map, then study how factor training changes it in fixed-feature regression.

Let $x \in \mathbb{R}^d$ be an input and let $o$ denote the output dimension. LoRA adds the correction $\Delta f_t(x) = B_t A_t x$, with $A_t \in \mathbb{R}^{r\times d}$ and $B_t \in \mathbb{R}^{o\times r}$. The factor width $r$ bounds the rank of every update, while the row space of $A_t$ specifies the input directions available at step $t$. We initialize $B_0 = 0$, so the adapter initially leaves the base model unchanged.

For the dynamical results, $X \in \mathbb{R}^{n\times d}$ contains fixed training features and $Y \in \mathbb{R}^{n\times o}$ contains target corrections after subtraction of the frozen base output. We minimize

$$\mathcal{L}(A, B) = \frac{1}{2n}\|XA^\top B^\top - Y\|_F^2 \tag{1}$$

by simultaneous full-batch Euclidean gradient descent without weight decay. Both factor updates use the current pair $(A_t, B_t)$. This model supports exact statements about factor dynamics; the empirical study separately measures paths produced by AdamW in full language models.

Let $\Pi_0 = A_0^\dagger A_0$ and choose $N \in \mathbb{R}^{d\times q}$ with $N^\top N = I_q$ and $NN^\top = I - \Pi_0$, where $q = d - \operatorname{rank}(A_0)$. The hidden input map

$$M_t = B_t A_t N \in \mathbb{R}^{o\times q} \tag{2}$$

records directions outside the initial LoRA input subspace. The choice $N = I_d$ instead measures the total adapter path. The hidden map isolates directions outside initialization and vanishes whenever training keeps $A$ fixed.

### 3.1 Exact and Effective Shared Dimension

The hidden maps describe individual checkpoints; a shared representation must accommodate them together, either exactly or within a chosen error tolerance. For checkpoints $\mathcal{T} = \{t_1, \dots, t_\ell\}$ with positive weights $w_j$ summing to one, define the weighted path operator

$$M_{\mathcal{T},w} = \begin{bmatrix} \sqrt{w_1} M_{t_1} \\ \vdots \\ \sqrt{w_\ell} M_{t_\ell} \end{bmatrix}. \tag{3}$$

**Definition 1** (Endpoint and pathwise complexity)**.** *The endpoint complexity is $C_t = \operatorname{rank}(M_t)$. The pathwise complexity is*

$$C_{\mathrm{path}}(\mathcal{T}) = \operatorname{rank}(M_{\mathcal{T},w}).$$

*For $\varepsilon \in [0,1]$, the effective pathwise complexity is the smallest shared dimension whose relative root-mean-square error is at most $\varepsilon$:*

$$C^{\mathrm{eff}}_{\mathrm{path}}(\varepsilon) = \min \left\{ k : \min_{Q^\top Q = I_k} \sum_{j=1}^{\ell} w_j \|M_{t_j}(I - QQ^\top)\|_F^2 \le \varepsilon^2 \|M_{\mathcal{T},w}\|_F^2 \right\}.$$

*It is zero when the path operator is zero.*

Positive weights do not change exact rank, but they specify which checkpoints matter for approximation.

**Proposition 1** (Standard spectral characterization)**.** *Let $\sigma_1 \ge \sigma_2 \ge \cdots$ be the singular values of $M_{\mathcal{T},w}$. Then $C_{\mathrm{path}}$ is the minimum dimension of a subspace $S \subseteq \mathbb{R}^q$ for which $M_t = M_t \Pi_S$ at every $t \in \mathcal{T}$. Moreover,*

$$\min_{Q^\top Q = I_k} \sum_j w_j \|M_{t_j}(I - QQ^\top)\|_F^2 = \sum_{i>k} \sigma_i^2. \tag{4}$$

*Consequently $C^{\mathrm{eff}}_{\mathrm{path}}(\varepsilon)$ is the smallest $k$ whose discarded spectral energy is at most an $\varepsilon^2$ fraction of the total. Exact path complexity is invariant to the choice of basis $N$, nondecreasing when checkpoints are added, and satisfies*

$$C_t \le r, \qquad C_{\mathrm{path}} \le \min\{q, \ell r\}. \tag{5}$$

*In the gradient-descent model above, the hidden path additionally satisfies $C_{\mathrm{path}} \le \operatorname{rank}(XN)$. For the total path, the corresponding training-span bound is $C_{\mathrm{path}} \le \operatorname{rank}([A_0^\top, X^\top]^\top)$.*

The row-space identity and classical low-rank approximation prove this result (Eckart & Young, 1936). We can also measure the spectrum directly from saved factors: Proposition 3 in Appendix A.3 reduces the decomposition to at most $\ell r$ rows without forming dense updates.

### 3.2 Rank One Can Trace a Near-Full-Dimensional Path

The definitions allow the shared dimension to exceed the rank of each checkpoint. We next show that ordinary gradient descent can realize this separation, even when the initial input space already contains a solution.

**Theorem 1** (Rank-one endpoint–path separation)**.** *For every $q \ge 1$, there is a scalar-output rank-one problem with $d = n = q + 1$, $A_0 = v^\top$, $\|v\|_2 = 1$, and target $Y = \delta X v$ for some $\delta > 0$, such that the target is exactly attainable while freezing $A = A_0$. Nevertheless, for all sufficiently small common positive step sizes, the vanilla gradient-descent input maps at steps $2, \dots, q+1$ satisfy*

$$\operatorname{rank} \begin{bmatrix} M_2 \\ \vdots \\ M_{q+1} \end{bmatrix} = q = d - 1, \qquad \operatorname{rank}(M_t) = 1 \quad (2 \le t \le q+1). \tag{6}$$

*Thus every selected checkpoint is rank one, while their shared hidden path requires all $d-1$ available directions.*

The gradient updates generate the entire construction in one run. Let $\mathcal{H} = X^\top X/n$ be diagonal with distinct positive eigenvalues and let $v$ have nonzero overlap with every eigenvector. With $A_t = a_t^\top$ and scalar $B_t = b_t$, the updates are

$$a_{t+1} = (I - \eta_A b_t^2 \mathcal{H}) a_t + \eta_A b_t \delta \mathcal{H} v,$$
$$b_{t+1} = b_t + \eta_B \{\delta a_t^\top \mathcal{H} v - b_t a_t^\top \mathcal{H} a_t\}. \tag{7}$$

Hence $a_t = p_t(\mathcal{H})v$. Successive polynomials have increasing degree and nonzero leading coefficients. After projection removes $v$, a Vandermonde system establishes that the remaining $q$ directions are independent. The full proof, including a rational three-dimensional instance, is in Appendix B.5.

The active feature spectrum determines how many independent directions this construction can reach.

**Corollary 1** (Active feature second-moment spectrum controls path growth)**.** *For arbitrary $\mathcal{H} \succ 0$, unit $v$, and target $Y = \delta X v$, let $P_\lambda$ be the spectral projectors and*

$$s = |\{\lambda : P_\lambda v \neq 0\}|.$$

*Every finite hidden path has complexity at most $s - 1$. For all sufficiently small common positive step sizes, checkpoints $2, \ldots, s$ attain complexity $s - 1$. If $s = 1$, every hidden input map is zero; in particular this holds for $\mathcal{H} = \lambda I$.*

The active Krylov space $\mathrm{span}\{v, \mathcal{H}v, \ldots\}$ has dimension $s$; removing the initial direction leaves $s-1$. Distinct eigenvalues allow the updates to change the input direction, whereas whitening reduces all eigenvalues to one scale.

A path can attain high shared dimension even when its hidden endpoint response later vanishes. In the full-rank construction,

$$\|M_t\|_F \leq \sqrt{2\mathcal{L}_t / \lambda_{\min}(\mathcal{H})}. \tag{8}$$

Any interpolating limit has zero hidden endpoint response even though an earlier finite path has complexity $d - 1$. The path spectrum measures how much these transient directions contribute to approximation error.

### 3.3 A Sharp Shared-Measurement Frontier

The separation raises a recovery question: how much prediction error remains when we observe only a fixed number of shared input directions? We answer it under the following information model. Let $P \in \mathbb{R}^{m \times d}$ be an unknown probe matrix. The observer knows $P_{\mathrm{v}} = P\Pi_0$, the trained path, and the residual second-moment matrix $\Sigma = P(I - \Pi_0)P^\top$, but not its hidden orientation $H = PN$. Assume $p = \mathrm{rank}(\Sigma) \leq q$. Write $M = M_{\mathcal{T},w}$ and let $V$ concatenate the known visible predictions with the same square-root weights. All compatible hidden orientations satisfy $HH^\top = \Sigma$. A protocol chooses one orthonormal $Q_k \in \mathbb{R}^{q \times k}$ before seeing $H$, observes $HQ_k$ for every probe, and uses an arbitrary decoder $\Psi$.

**Theorem 2** (Sharp frontier for shared linear measurements)**.** *For $0 \leq k \leq q$, the optimal worst-case squared Frobenius error is*

$$\inf_{Q_k^\top Q_k = I_k,\, \Psi} \sup_{HH^\top = \Sigma} \|V + HM^\top - \Psi(HQ_k)\|_F^2 = \sum_{i=1}^{\min(p, q-k)} \lambda_i(\Sigma)\sigma_{i+k}(M)^2. \tag{9}$$

*Eigenvalues and singular values are decreasing and zero-padded. Top right singular directions of $M$ attain the bound. For a fixed query space with orthogonal complement $Q_\perp$ and feasible observation $W = HQ_k$, the conditional squared radius is*

$$\sum_i \lambda_i(\Sigma - WW^\top)\sigma_i(MQ_\perp)^2. \tag{10}$$

*If $\Sigma \neq 0$, exact recovery requires $\mathrm{rank}(M)$ shared directions.*

The lower bound applies to any decoder within this shared linear measurement protocol. Unrestricted nonlinear encodings fall outside the stated information model. Appendix A gives the attainable prediction set, and Appendix B.4 proves the frontier. An exact two-Gram recurrence describes the same bilinear factor dynamics in Appendix B.1.

# 4 From Path Spectra to Predictive Perturbations

The recovery frontier depends on the path spectrum, not only its exact rank. We now examine what shapes this spectrum and when a small matrix approximation error also limits changes in prediction. For a square-integrable matrix path with weighting measure $\mu$, define

$$G_\mu = \int M(t)^\top M(t)\, d\mu(t), \qquad E_k(\mu) = \frac{\sum_{i>k} \lambda_i(G_\mu)}{\operatorname{tr} G_\mu}.$$

When the denominator is positive, $C_{\varepsilon,\mu} = \min\{k : E_k(\mu) \le \varepsilon^2\}$; the zero path has complexity zero. A 99% squared-energy threshold permits a relative root-mean-square error of 10%, not 1%. This is a shared *input*-direction dimension, rather than the dimension obtained by vectorizing each complete matrix.

## 4.1 Weighting the Path Spectrum

The weighting measure determines which parts of training a shared subspace should represent. Checkpoint weighting averages saved states, uniform training-time weighting measures how long states persist, and arc-length weighting measures movement along the matrix path:

$$d\mu_{\mathrm{arc}}(t) = \frac{\|\dot{M}(t)\|_F\, dt}{L}, \qquad L = \int \|\dot{M}(t)\|_F\, dt > 0.$$

For absolutely continuous paths this measure is invariant to admissible order-preserving changes of traversal speed. Repeated traversals contribute separately to this measure. A constant path has no normalized arc measure; we use its single-state spectrum by convention. From the saved factors, we integrate the piecewise-linear *matrix* path exactly. This interpolated path approximates the underlying training trajectory. The arc measure is defined separately for the total operator, hidden operator, and each token-response path.

Because these choices can produce different spectra, it is useful to identify when their approximation errors remain comparable. If $0 < a \le d\nu/d\mu \le b < \infty$, then

$$\frac{a}{b} E_k(\mu) \le E_k(\nu) \le \frac{b}{a} E_k(\mu). \tag{11}$$

A comparison of quadratic-form integrals gives this tail-energy bound, although the integer dimension can change near the tolerance threshold. Time and arc measures need not satisfy a positive density lower bound. Appendix C gives the proofs and the exact segment-integration formula.

## 4.2 Sources of Spectral Concentration

With the weights fixed, we next ask why many independent directions can carry little energy. In the aligned rank-one model of Corollary 1, let $\lambda_i$ be the active eigenvalues and $c_i = \|P_{\lambda_i} v\|$, so $\sum_i c_i^2 = 1$. The response $a_t = p_t(\mathcal{H})v$ gives spectral energies through the weighted Gram matrix of $b_t c_i p_t(\lambda_i)$. The following bound shows how nearby eigenvalues can produce similar responses and hence a small approximation dimension.

**Proposition 2** (A finite-horizon clustering bound)**.** *In the aligned-target rank-one model above, take $\mu = \sum_t w_t \delta_t$ with $w_t \ge 0$ summing to one. Partition the active eigenvalues into $m$ nonempty groups. Choose a center $\zeta_g$ in each group's convex hull, let $d_g = \max_{i\in g} |\lambda_i - \zeta_g|$, $\alpha_g = \sum_{i\in g} c_i^2$, and let $L_{t,g}$ bound $|p_t'(\lambda)|$ on that hull. For a finite weighted hidden path, assume $D > 0$, where $D = \sum_t w_t \|M_t\|_F^2$. Then*

$$E_{m-1}(\mu) \le \frac{\sum_t w_t b_t^2 \sum_g \alpha_g d_g^2 L_{t,g}^2}{D}. \tag{12}$$

The proof approximates $p_t$ by one value per spectral group and removes the initial direction. Thus, many independent directions can admit a compact approximation when responses within each group remain close relative to total hidden energy. The bound depends on the feature spectrum, initialization, and finite horizon. As training continues, effective dimension can decrease when a dominant direction accumulates energy, although exact dimension cannot decrease as checkpoints are added.

Such a decrease leaves open whether the same subspace remains useful later in training. To examine this distinction, we separate the earlier path from the subsequent interval. For a time-integrated path, let $C = \int_0^{T_1} M_t^\top M_t \, dt$ and $D = \int_{T_1}^{T_2} M_t^\top M_t \, dt$; both are unnormalized. Write $R_k(C) = \sum_{i>k} \lambda_i(C)$ and let $P_k$ project onto the top $k$ eigenvectors of $C$. Elementary variational bounds give

$$R_k(C) \le R_k(C + D) \le R_k(C) + \mathrm{tr}((I - P_k)D). \tag{13}$$

The absolute residual must therefore increase or stay constant, even when the relative complexity decreases. For positive traces, put $q = R_k(C)/\,\mathrm{tr}\, C$ and $\beta = \mathrm{tr}((I - P_k)D)/\,\mathrm{tr}\, D$. The new relative tail is at most the energy-weighted average of $q$ and $\beta$; $\beta \le q$ is sufficient for concentration in the previously learned subspace. A decrease in the optimized relative tail alone does not imply this condition, because the best subspace can change. Appendix C.7 gives proofs and distinguishes fixed from evolving data weighting.

The preceding bounds apply at finite times, while convergence gives a complementary explanation for late concentration. If a locally square-integrable matrix path converges to a nonzero matrix $M_\infty$ of rank at most $r$, its uniform-time Gram average converges to $M_\infty^\top M_\infty$, and its relative spectral tail beyond $r$ vanishes (Lemma 2). The nonzero limit is essential: any interpolating limit of our aligned construction has a zero hidden map. Convergence of the measured Transformer paths also remains unestablished.

### 4.3 From Matrix Residuals to Prediction Loss

The spectral bounds concern matrices. To connect them to model behavior, we follow the discarded component through the network. The intervention at site $j$ deletes an operator $E_j$ while retaining the initialization component. Its response on inputs $H_j$ satisfies $\mathbb{E}\|E_j H_j\|_F^2 = \mathrm{tr}(E_j \Sigma_j E_j^\top)$, where $\Sigma_j = \mathbb{E}[H_j H_j^\top]$ is an uncentered second moment. For joint interventions, form hybrid networks that replace one site at a time. Suppose the inputs at each hybrid site obey $\Sigma_j \preceq \beta_j I$ and the conditional downstream logit maps are $L_j$-Lipschitz on the relevant perturbations. Teacher-forced token negative log likelihood then obeys

$$|\widetilde{\mathcal{L}} - \mathcal{L}| \le \sqrt{2} \sum_j L_j \sqrt{\beta_j} \|E_j\|_F. \tag{14}$$

Appendix D proves the hybrid-network bound and its path-averaged version under the same weights. Our experiments do not estimate these constants, so the bound provides conditions for loss control rather than a numerical certificate. Greedy generation also depends on the gap between competing logits. Small average loss changes alone do not guarantee identical generated answers.

## 5 Experiments

The theory separates a compact matrix representation from its effect on predictions. We test this distinction in language modeling and arithmetic reasoning: do trained LoRA paths learn a small space beyond initialization, and what behavior does that space retain?

### 5.1 Experimental Design

We pair measurements of the training path with interventions that keep or remove its learned directions. The language-modeling study uses WikiText-103, a dataset of Wikipedia articles for next-token prediction (Merity et al., 2017). We fine-tune Qwen2.5-1.5B (Qwen et al., 2024) and Pythia-1.4B-deduped (Biderman et al., 2023) with adapter ranks one and four, using three seeds per setting. Each run lasts 1600 steps, and we measure the saved path across all adapted matrices. Alongside ordinary LoRA, which trains both factors, we train a matched control with the input factor $A$ frozen. The frozen factor confines updates to the initial input space, providing a control for departure from initialization.

On GSM8K, a dataset of arithmetic word problems (Cobbe et al., 2021), we fine-tune Qwen with supervision on the answer tokens. We use the same two ranks and three seeds, plus a frozen-$A$ control at each rank. We measure held-out answer loss and generate solutions to all 1319 test questions. Throughout, negative log likelihood (NLL) measures the mean negative log probability of

Table 1: Shared hidden-path geometry after 1600 steps on WikiText-103. The dimensions $k_{99}$ and $k_{99.9}$ retain 99% and 99.9% of path energy. Error is the relative rank-$r$ reconstruction error. Entries average the within-run matrix medians over three seeds, so dimensions can be noninteger.

| Model | $r$ | Endpoint $k_{99}$ | Path $k_{99}$ | Path $k_{99.9}$ | Error | Hidden/total |
|---|---|---|---|---|---|---|
| Qwen | 1 | 1 | 2.00 | 3.00 | 10.74% | 0.864 |
| Qwen | 4 | 2 | 3.00 | 10.17 | 6.46% | 0.850 |
| Pythia | 1 | 1 | 1.67 | 3.00 | 10.42% | 0.850 |
| Pythia | 4 | 2 | 3.00 | 9.33 | 6.54% | 0.815 |

the observed tokens, so lower values are better. Generation accuracy measures the fraction of correct final answers. Appendices E and F give the training settings, evaluation protocol, and aggregation details.

### 5.2 LoRA Learns a Small Shared Space Beyond Initialization

We first ask whether departure from initialization produces a broad input space or a small set of learned directions. At the end of training, the hidden-to-total path norm is 81.5–86.4%, yet only 1.67–3 shared directions retain 99% of hidden path energy in the group summaries (Table 1). Each group needs more directions to represent the path than the endpoint, although both dimensions remain small at this threshold. The frozen-$A$ controls have zero hidden component, as expected. LoRA therefore learns a small shared space that differs substantially from the one supplied at initialization.

A small effective dimension does not mean that the learned subspace has stopped changing. Directions fitted to the first half of training capture 96.7–98.7% of later state energy, but only 70.3–84.9% of the energy in subsequent movement. Most of the learned update remains near the earlier subspace even as training continues to move beyond it. The state spectrum can become more concentrated without the path becoming stationary (Appendix E). Separate fixed-feature controls recover the exact dimension predicted by Corollary 1 (Appendix G).

### 5.3 Shared Directions Retain Average Prediction Quality

A concentrated spectrum describes matrix error; we next test whether its dominant directions also retain predictions. We project the hidden component in every adapted matrix while preserving the component inside the initialization space. We compare three subspaces with the same budget $k = r$, fitted to the full path, final checkpoint, or first half of training. The full-path fit uses all saved checkpoints offline. Evaluations at four training stages test whether a representation of the final update also represents the history that led to it.

Full-path directions give lower time-weighted NLL than endpoint directions in 11 of 12 language-modeling runs. The mean reductions range from 0.000045 to 0.001240 nats per token across model and rank settings. These improvements are small, but Figure 1 shows the distinction they reveal: endpoint directions reconstruct the final update exactly while changing predictions at earlier checkpoints. A subspace that represents the endpoint therefore need not represent the training path equally well.

The math task provides a complementary test of whether a few learned directions retain the effect of adaptation. A joint projection that keeps 99% of hidden path energy changes mean answer NLL by less than 0.0002 at either rank. In contrast, removal of the entire hidden component raises NLL by 0.13–0.15. Random subspaces with the same number of directions also perform worse in all 30 draws. They retain much less path energy, so this comparison tests direction selection at a fixed dimension rather than at equal retained energy. Together, these interventions show that the hidden update contributes to prediction and that a small, selected part of it retains most of its effect on average loss.

### 5.4 Similar Accuracy Can Conceal Different Answers

The loss comparisons use reference answers, whereas generated solutions depend on the model's own earlier tokens. We therefore compare accuracy and answer agreement on complete solutions. Table 2

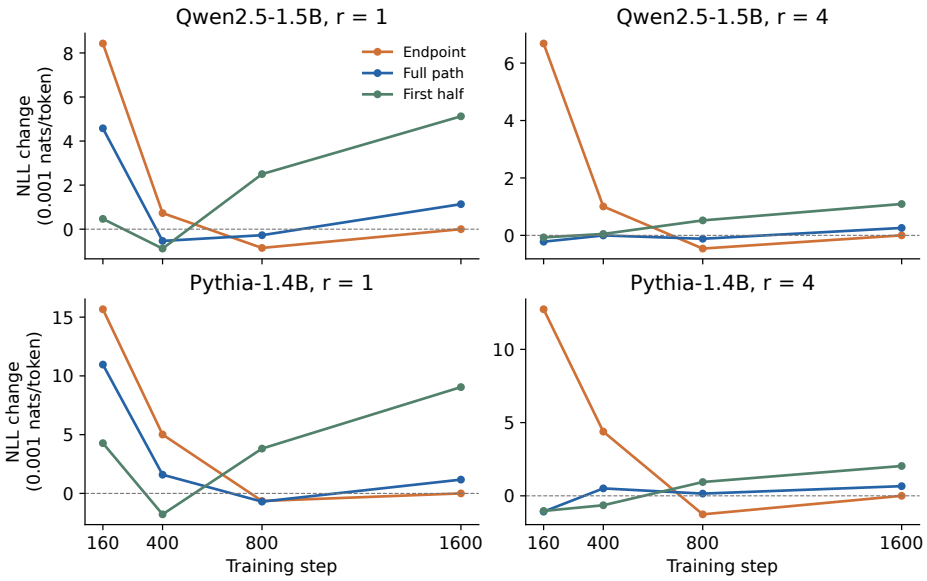



Figure 1: Prediction loss across training under the same direction budget $k = r$. Curves show the NLL change from the original model, averaged over three seeds. Endpoint directions preserve the final checkpoint exactly but can alter earlier predictions.

Table 2: GSM8K results after joint 99%-energy projection, averaged over three seeds on the same 1319 questions. A correctness flip changes whether the answer is right, whereas an answer change can also occur between two wrong answers. Changes in accuracy are in percentage points (pp).

| | Original | 99% path | Change (pp) | Correctness flips | Answers changed |
|---|---|---|---|---|---|
| $r = 1$ | 58.15% | 59.31% | +1.16 | 6.72% | 16.60% |
| $r = 4$ | 58.07% | 57.92% | -0.15 | 7.43% | 17.94% |

shows that 99%-energy projection keeps mean accuracy within 1.2 percentage points of the original at both ranks. These averages describe the measured runs, rather than establish statistical equivalence.

The agreement is weaker at the level of individual questions. For one rank-four run, both versions score 57.24%, yet 53 answers change from correct to wrong and another 53 change in the opposite direction. Equal accuracy can therefore conceal compensating changes. The recorded generations contain one pass per model variant, so repeated original-model and identity-intervention runs are still needed to separate projection effects from execution variability (Appendix F). The comparison illustrates why aggregate accuracy and agreement on individual answers answer different questions about predictive fidelity.

## 6 Discussion and Conclusion

The theory and experiments together explain why adapter rank alone cannot describe a training history. Rank-one gradient descent can visit every hidden input direction, while the measured language-model paths concentrate most of their energy in a few learned directions. These directions retain average prediction quality under joint interventions, but similar aggregate scores can conceal different answers. The dynamical construction concerns fixed-feature regression, and the loss bound requires control of activations and downstream sensitivity. The empirical evidence covers two model families, ranks one and four, and two objectives. Within this scope, shared-subspace analysis separates three questions that require different evidence: where training moves, how many directions describe its path, and how well those directions preserve predictions.

### AI USE STATEMENT

We used generative AI tools to polish the writing, check and improve the theoretical proofs, and assist with the design and implementation of the experimental code. The tools also supported the analysis of the synthetic experiments and literature searches. Numerical checks tested algebraic identities and the assumptions of the measurement protocol. We take responsibility for the final manuscript, including its mathematics, citations, empirical claims, and AI-assisted code.

### ETHICS STATEMENT

We use synthetic features, public pretrained models, and published text and arithmetic datasets. The study involves no human subjects or deployed interventions. Our probe-information model assumes access to noiseless linear observations. Practical applications would require appropriate access and privacy controls. Geometric compression and agreement between predictions alone do not establish semantic benefit or safety.

### REPRODUCIBILITY STATEMENT

The appendices provide complete proofs with explicit optimizer and observation assumptions. Appendices E and F document the training and evaluation configurations, seeds, checkpoints, projection budgets, numerical thresholds, and verification gaps. We preserved source and file hashes, configurations, factors, token identities, per-sequence scores, generations, environment records, and logs. These records support recomputation of the final analysis without new model training. We distinguish post-hoc diagnostics from the geometric outcomes specified before analysis.

## REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, 2021.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023.

Zhipeng Chen, Tao Qian, Wayne Xin Zhao, and Ji-Rong Wen. Low-rank optimization trajectories modeling for LLM RLVR acceleration. arXiv preprint arXiv:2604.11446, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Arif Kerem Dayi and Sitan Chen. Low-rank fine-tuning lies between lazy training and feature learning. In *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1415–1471, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17554–17571, 2024.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117015–117040, 2024a.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806, 2024b.

Aicke Hinrichs, David Krieg, Erich Novak, Joscha Prochno, and Mario Ullrich. Random sections of ellipsoids and the power of random information. *Transactions of the American Mathematical Society*, 374(12):8691–8713, 2021.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26740–26756, 2025.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21306–21328, 2024.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. In *International Conference on Learning Representations*, 2024.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23610–23641, 2023.

Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K. Transtrum, James P. Sethna, and Pratik Chaudhari. The training process of many deep networks explores the same low-dimensional manifold. *Proceedings of the National Academy of Sciences*, 121(12): e2310002121, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, 2025.

Zhepei Wei, Xinyu Zhu, Wei-Lin Chen, Chengsong Huang, Jiaxin Huang, and Yu Meng. You only need minimal RLVR training: Extrapolating LLMs via rank-1 trajectories. arXiv preprint arXiv:2605.21468, 2026.

Yize Wu, Ke Gao, Ling Li, and Yanjun Wu. Stable-LoRA: Stabilizing feature learning of low-rank adaptation. In *International Conference on Learning Representations*, 2026.

Zhen Yang, Wei Tang, Chu Chen, Jeonseong Kim, Sizai Hou, Yinhao Wu, Jack Perry, Liang He, Yixuan Chen, Fan Yang, and Kangning Cui. Accuracy is not fidelity: Decision churn in low-rank compressed language models. Author-hosted research abstract, 2026.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

# A Operational Recovery and Exact Closure

The main text connects the path spectrum to a shared representation of predictions. Here we give the attainable prediction set, an exact recurrence for the trained path, and a way to compute its spectrum from saved factors.

## A.1 A Population-Weighted Approximation Frontier

We now use the path spectrum to characterize prediction error under incomplete probe information. For a probe matrix $P$, write

$$P_{\mathrm{v}} = P\Pi_0, \qquad \Sigma = P(I - \Pi_0)P^\top,$$

and suppose only these two quantities are retained. Every compatible probe is $P = P_{\mathrm{v}} + HN^\top$ with $HH^\top = \Sigma$. Define

$$V_t = P_{\mathrm{v}} A_t^\top B_t^\top, \qquad M_t = B_t A_t N. \tag{15}$$

**Theorem 3** (Attainable predictions and exact conditional risk)**.** *Suppose* $\mathrm{rank}(\Sigma) = p \le q$ *and* $LL^\top = \Sigma$ *with* $L$ *full column rank. The complete prediction set is*

$$\mathcal{D}_t = \{V_t + LQM_t^\top : QQ^\top = I_p\}. \tag{16}$$

*Its unique minimax Frobenius center is* $V_t$*, with squared radius*

$$\mathcal{R}_t^2 = \sum_{i=1}^{p} \lambda_i(\Sigma)\sigma_i(M_t)^2. \tag{17}$$

*The diameter is* $2\mathcal{R}_t$*, and for any* $U$ *the sharp support is*

$$\sup_{D \in \mathcal{D}_t} \langle U, D\rangle_F = \langle U, V_t\rangle_F + \|L^\top U M_t\|_*. \tag{18}$$

Write $A_t = A_0 + J_t^\top X$ with $J_t \in \mathbb{R}^{n\times r}$ and $R_t = J_t B_t^\top$. We can then compute the response spectrum in training coordinates:

$$M_t M_t^\top = R_t^\top G_\perp R_t, \qquad G_\perp = X(I - \Pi_0)X^\top. \tag{19}$$

Theorem 2 gives the corresponding worst-case error when a protocol measures the same $k$ hidden directions for every probe. A stack with one element gives the single-checkpoint result. The information model thus gives path complexity an operational meaning: its spectrum controls recovery error when probe orientation is unknown.

## A.2 Supporting Result: Exact Finite-Step Closure

The recovery analysis takes the trained path as given. To connect this object to the optimizer, we now express the factor updates through training and probe second moments. The following recurrence computes the learned path without storing the ambient feature vectors. Write $Z_0 = XA_0^\top$, $K = Z_0 Z_0^\top$, and $G = XX^\top$. For probes $P$, let $K_{PT}^0 = PA_0^\top Z_0^\top$ and $G_{PT} = PX^\top$.

**Theorem 4** (Exact two-Gram functional closure)**.** *Let* $k = \mathrm{rank}(K) > 0$ *and* $K = ZZ^\top$ *with* $Z$ *full column rank. Set* $Z_P = K_{PT}^0 Z(Z^\top Z)^{-1}$ *and* $\widetilde{C}_0 = \widetilde{B}_0 = 0$*. Using current states on each right-hand side, define*

$$\begin{aligned} \widetilde{Z}_t &= Z + G\widetilde{C}_t, & E_t &= \widetilde{Z}_t\widetilde{B}_t^\top - Y, \\ \widetilde{C}_{t+1} &= \widetilde{C}_t - \frac{\eta_{A,t}}{n} E_t \widetilde{B}_t, & \widetilde{B}_{t+1} &= \widetilde{B}_t - \frac{\eta_{B,t}}{n} E_t^\top \widetilde{Z}_t. \end{aligned} \tag{20}$$

*Then every finite-step probe correction is exactly*

$$\Delta f_t(P) = Z_P \widetilde{B}_t^\top + G_{PT} R_t, \qquad R_t = \widetilde{C}_t \widetilde{B}_t^\top. \tag{21}$$

*Thus* $(K, G, K_{PT}^0, G_{PT}, Y)$ *and the step-size schedule determine the full path. If* $K = 0$*, all corrections remain zero.*

This exact recurrence retains the changing factors throughout training. It also identifies the information absent from the initial kernel: $G$ and $G_{PT}$ retain feature directions that $A_0$ removes. Appendix B.1 gives the constructive proof and examples showing that both missing training geometry and missing train–probe geometry can matter.

### A.3 Computing the path spectrum from factors

The recurrence above describes how training produces a path. For an already saved path, we instead need a compact way to measure its spectrum; the following identity avoids constructing dense weight updates.

**Proposition 3** (Factor-only path spectrum)**.** *Suppose the checkpoint correction is $\Delta W_t = s_t B_t A_t$ and let $S_t = (B_t^\top B_t)^{1/2}$. The nonzero singular values of $\sqrt{w_t}\Delta W_t N$ equal those of $\sqrt{w_t} s_t S_t A_t N$. Therefore the complete path spectrum is the spectrum of a matrix with at most $\ell r$ rows:*

$$\begin{bmatrix} \sqrt{w_1} s_1 S_1 A_1 N \\ \vdots \\ \sqrt{w_\ell} s_\ell S_\ell A_\ell N \end{bmatrix}.$$

Indeed, both blocks have Gram matrix $w_t s_t^2 N^\top A_t^\top B_t^\top B_t A_t N$. This avoids storing or decomposing dense Transformer weight updates.

## B Proofs for Finite-Time Geometry and Probe Information

The results in Section 3 combine factor dynamics with a model of the information available about probes. We first establish the exact dynamics, then derive the recovery bounds and the rank-one path construction.

### B.1 Exact Gradient Closure

We begin with the optimizer because the later construction must arise from a single training run. The recurrence below expresses its factor updates in the span of the fixed training features. All statements concern the adapter correction, with the frozen base output already subtracted from the target $Y$. The feature matrix $X$ is fixed, the adapter scale is one, and

$$E_t = X A_t^\top B_t^\top - Y, \qquad \alpha_t = \eta_{A,t}/n, \quad \beta_t = \eta_{B,t}/n.$$

The simultaneous factor updates are

$$A_{t+1}^\top = A_t^\top - \alpha_t X^\top E_t B_t, \qquad B_{t+1} = B_t - \beta_t E_t^\top X A_t^\top. \tag{22}$$

Both right-hand sides use $(A_t, B_t)$. These identities follow by differentiating $\|XA^\top B^\top - Y\|_F^2/(2n)$.

For an auxiliary coefficient representation, initialize $C_0 = 0 \in \mathbb{R}^{n\times n}$ and $F_0 = 0 \in \mathbb{R}^{o\times n}$, and use

$$\begin{aligned} S_t &= I + G C_t, & E_t &= S_t K F_t^\top - Y, \\ F_{t+1} &= F_t - \beta_t E_t^\top S_t, & C_{t+1} &= C_t - \alpha_t E_t F_t. \end{aligned} \tag{23}$$

*Verification of the full coefficient representation.* Initially $B_0 = 0 = F_0 Z_0$ and $A_0^\top = A_0^\top + X^\top C_0 Z_0$. Suppose the representer identities hold at time $t$. Then

$$X A_t^\top = (I + G C_t) Z_0 = S_t Z_0, \qquad X A_t^\top B_t^\top = S_t K F_t^\top.$$

Substitution into (22) gives

$$\begin{aligned} B_{t+1} &= (F_t - \beta_t E_t^\top S_t) Z_0, \\ A_{t+1}^\top &= A_0^\top + X^\top (C_t - \alpha_t E_t F_t) Z_0. \end{aligned}$$

These expressions match the claimed recurrences and complete the induction. For any probe matrix $P$,

$$\begin{aligned} P A_t^\top B_t^\top &= P A_0^\top Z_0^\top F_t^\top + P X^\top C_t Z_0 Z_0^\top F_t^\top \\ &= K_{PT}^0 F_t^\top + G_{PT} C_t K F_t^\top. \end{aligned}$$

This proves the full coefficient output formula at every finite step. The proof allows singular $K$ and $G$, arbitrary finite step sizes, and trajectories whose loss does not converge. □

**A compact equivalent recurrence.** The $n \times n$ coefficient representation is convenient for the proof but can contain unnecessary coordinates. An equivalent construction uses a minimal factor of the initial kernel.

**Proposition 4** (Minimal-kernel-factor implementation)**.** *Let $k = \operatorname{rank}(K) > 0$ and choose any full-column-rank $Z \in \mathbb{R}^{n\times k}$ satisfying $ZZ^\top = K$. Define*

$$Z_P = K_{PT}^0 Z(Z^\top Z)^{-1}, \qquad \widetilde{C}_0 = 0 \in \mathbb{R}^{n\times k}, \qquad \widetilde{B}_0 = 0 \in \mathbb{R}^{o\times k}.$$

*The recurrence*

$$\begin{aligned} \widetilde{Z}_t &= Z + G\widetilde{C}_t, & E_t &= \widetilde{Z}_t\widetilde{B}_t^\top - Y, \\ \widetilde{C}_{t+1} &= \widetilde{C}_t - \alpha_t E_t \widetilde{B}_t, & \widetilde{B}_{t+1} &= \widetilde{B}_t - \beta_t E_t^\top \widetilde{Z}_t \end{aligned} \tag{24}$$

*produces the exact correction*

$$\Delta f_t(P) = (Z_P + G_{PT}\widetilde{C}_t)\widetilde{B}_t^\top.$$

*If $K = 0$, both original factors stay at initialization and the correction is zero.*

*Proof of Proposition 4 and Theorem 4.* Set $O = Z^\dagger Z_0 \in \mathbb{R}^{k\times r}$. Since $Z_0 Z_0^\top = ZZ^\top$, $OO^\top = I_k$ and $Z_0 = ZO$. We claim that

$$B_t = \widetilde{B}_t O, \qquad A_t^\top = A_0^\top + X^\top \widetilde{C}_t O.$$

These identities hold initially. If they hold at time $t$, then $XA_t^\top = \widetilde{Z}_t O$ and the training correction is $\widetilde{Z}_t OO^\top \widetilde{B}_t^\top = \widetilde{Z}_t \widetilde{B}_t^\top$. The updates in (22) preserve both identities, which completes the induction step. Furthermore

$$K_{PT}^0 = PA_0^\top O^\top Z^\top, \qquad Z_P = PA_0^\top O^\top.$$

Substitution into the probe correction gives the stated formula. If $K = 0$, then $Z_0 = 0$ and both gradients vanish at $B_0 = 0$. The factors therefore remain at initialization by induction. □

**Constructive finite-time predictor from Gram information.** Factor $K = ZZ^\top$ using its nonzero eigenspace, recover $Z_P$ from $K_{PT}^0$, and initialize $\widetilde{C} = \widetilde{B} = 0$. At each step, form $\widetilde{Z} = Z + G\widetilde{C}$ and $E = \widetilde{Z}\widetilde{B}^\top - Y$. Compute both updates in (24) from the current states, then apply them simultaneously. Evaluate probes using $(Z_P + G_{PT}\widetilde{C})\widetilde{B}^\top$. After constructing the Grams and kernel factor, one training step uses $O(n^2k + nok)$ arithmetic operations.

This recurrence implements the original optimizer in compact coordinates. Orthogonal changes of the kernel factor $Z$ rotate both compact states and leave predictions invariant. The minimal state dimension removes redundant coordinates, but floating-point stability requires a separate analysis. Implementations that approximate the positive semidefinite (PSD) factorization should report their numerical rank thresholds and discarded eigenvalues. The dense and compact responses satisfy

$$R_t = C_t K F_t^\top = \widetilde{C}_t \widetilde{B}_t^\top.$$

The equality follows either from the inductions above or by keeping the corresponding coefficient states synchronized.

**What is sufficient and what can be omitted.** For fixed training data, the response $R_t$ is independent of probes. If two probes have the same $K_{PT}^0$, subtracting their predictions gives

$$\Delta f_t(P_1) - \Delta f_t(P_2) = (G_{P_1T} - G_{P_2T})R_t.$$

Thus equality of these products is necessary and sufficient for equality of corrections at that endpoint. Only the action of $G_{PT}$ on $R_t$ matters for this endpoint, so entries that do not affect that product can be omitted. For multiple times, replace $R_t$ by its horizontal concatenation.

Some problems require the raw training Gram in addition to the initial kernel. For a scalar rank-one adapter with $a_0 = e_1$, one training input $x = e_1 + ue_2$, target correction $\delta > 0$, and common step size $0 < \eta < 1$, the initial training kernel is one for every $u$. After two steps,

$$b_2 = \eta\delta(2 - \eta), \qquad a_2 = e_1 + \eta^2\delta^2(1 - \eta)x.$$

The training correction is $b_2[1+\eta^2\delta^2(1-\eta)(1+u^2)]$, which depends on $G=1+u^2$ despite the matched initial kernel.

The raw cross-Gram can also matter when the complete training problem is fixed. Set $x=e_1+e_2$ and compare $p_1=ce_1+he_2$ with $p_2=ce_1+he_3$. The initial cross- and probe kernels, as well as the raw probe norm, match. Their two-step corrections differ by $h\eta^3\delta^3(2-\eta)(1-\eta)$. These examples show that initial kernels alone can omit necessary geometric information. Which individual entries matter still depends on the task and its response matrix.

**Why coordinatewise Adam is outside the closure.** The closure relies on the transformation properties of Euclidean gradient updates. The following example shows why the same Gram information does not determine coordinatewise Adam updates. Take $n=1$, $d=o=3$, $r=2$, $X=e_1^\top$, and $Y=(1,0,0)$. Let $A_0$ have rows $e_1^\top, e_2^\top$. In a second parameterization, apply an orthogonal rotation to these rows so that the training feature changes from $z=(1,0)$ to $z'=(1/\sqrt{2},1/\sqrt{2})$. Both parameterizations have the same raw training Gram, targets, and initial tangent-kernel blocks on any fixed inputs. Initially the $A$ gradient is zero. For bias-corrected Adam with no weight decay and denominator $\sqrt{\widehat{v}}+\epsilon_{\mathrm{Adam}}$, the first active output after one step is

$$\eta\sum_j \frac{z_j^2}{|z_j|+\epsilon_{\mathrm{Adam}}}=\frac{\eta}{1+\epsilon_{\mathrm{Adam}}}$$

in the first parameterization, and $\eta/(1/\sqrt{2}+\epsilon_{\mathrm{Adam}})$ in the second. The two outputs differ for every $\epsilon_{\mathrm{Adam}}>0$. As $\epsilon_{\mathrm{Adam}}\downarrow 0$, they tend to $\eta$ and $\sqrt{2}\,\eta$. Thus Euclidean Gram information does not identify coordinatewise Adam dynamics. This restriction applies to the dynamical closure. The conditional geometric results below hold for any fixed response matrix.

### B.2 The Probe-Information Class

Once the trained path is fixed, recovery depends on which parts of the probe inputs remain unknown. We characterize that uncertainty from the visible projection and the residual second moment. Let $\Pi_0=A_0^\dagger A_0$ and let $N$ be an orthonormal basis of its orthogonal complement. Given $Z_P^0=PA_0^\top$, we recover the visible probe matrix exactly as

$$P_{\mathrm{v}}=Z_P^0(A_0A_0^\top)^\dagger A_0.$$

This remains true when $A_0$ is rank deficient, by the singular-value decomposition of $A_0$. Thus the observations $(PA_0^\top, PP^\top)$ specify

$$\Sigma=PP^\top-P_{\mathrm{v}}P_{\mathrm{v}}^\top.$$

Since visible and hidden subspaces are orthogonal, every consistent probe has a unique representation $P=P_{\mathrm{v}}+HN^\top$ with $HH^\top=\Sigma$. Conversely, every such $H$ produces the same initial projection and the same raw within-probe Gram matrix.

**Lemma 1** (All factors of a fixed Gram matrix)**.** *Let $\Sigma=LL^\top$ have rank $p$, where $L\in\mathbb{R}^{m\times p}$ has full column rank, and suppose $p\le q$. Then*

$$\{H\in\mathbb{R}^{m\times q}: HH^\top=\Sigma\}=\{LQ: Q\in\mathbb{R}^{p\times q},\ QQ^\top=I_p\}.$$

*Proof.* Every matrix $LQ$ on the right has Gram matrix $LL^\top$, which proves the reverse inclusion. Conversely, if $HH^\top=LL^\top$, their column spaces coincide. Hence $H=LQ$ for $Q=L^\dagger H$, and

$$QQ^\top=L^\dagger LL^\top(L^\dagger)^\top=I_p.$$

□

Initial kernel blocks alone need not determine the entire initial projection. If $Z_0$ has full column rank, however, $K_{PT}^0=Z_P^0Z_0^\top$ gives $Z_P^0=K_{PT}^0Z_0(Z_0^\top Z_0)^{-1}$. Full column rank here refers to the actual factor width because the goal is to recover the entire projection. Proposition 4 recovers only the compact projection active on the training data. That projection need not contain every visible coordinate required by the probe-Gram decomposition.

If a linear frozen base $f_0(x) = W_0 x$ contributes additional observed base outputs, one can instead use the row space of $[A_0^\top, W_0^\top]^\top$ as the visible subspace, provided its probe projections are observed. The same proof applies with the smaller unobserved complement. For a nonlinear frozen base, fixed probe outputs can impose nonlinear restrictions on the orientation family. The unrestricted Gram-factor family then contains the feasible set but need not equal it. Exact attainability and the matching lower bound require the stated information model.

### B.3 Attainability, Radius, and Linear Functionals

The probe class gives a set of compatible inputs. We now map that set through the trained adapter to find all attainable predictions and the smallest worst-case recovery error.

*Proof of Theorem 3.* For every consistent probe,

$$PA_t^\top B_t^\top = P_{\mathrm{v}} A_t^\top B_t^\top + HM_t^\top = V_t + HM_t^\top.$$

Lemma 1 proves the exact attainable-set description.

Choose $L = U_\Sigma \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$, where $U_\Sigma^\top U_\Sigma = I_p$. Writing $q_i^\top$ for the rows of $Q$,

$$\|LQM_t^\top\|_F^2 = \sum_{i=1}^{p} \lambda_i q_i^\top M_t^\top M_t q_i.$$

Let $\mu_1 \geq \cdots \geq \mu_q \geq 0$ be the eigenvalues of $M_t^\top M_t$. For any orthonormal $q_1, \ldots, q_j$,

$$\sum_{i=1}^{j} q_i^\top M_t^\top M_t q_i \leq \sum_{i=1}^{j} \mu_i.$$

Indeed, in an eigenbasis the left side is $\sum_\ell \mu_\ell w_\ell$ with $0 \leq w_\ell \leq 1$ and $\sum_\ell w_\ell = j$, whose maximum selects the $j$ largest eigenvalues. Taking $\lambda_{p+1} = 0$ and summing these inequalities with weights $\lambda_j - \lambda_{j+1}$ yields

$$\sum_{i=1}^{p} \lambda_i q_i^\top M_t^\top M_t q_i \leq \sum_{i=1}^{p} \lambda_i \mu_i.$$

The corresponding right singular vectors of $M_t$ attain equality when used as the rows of $Q$, with an orthonormal completion if needed.

Let $T_\star = LQ_\star M_t^\top$ attain this maximum, and write $R^2 = \|T_\star\|_F^2$. The attainable centered set is symmetric: it contains both $T_\star$ and $-T_\star$. For any proposed center $V_t + C$,

$$\begin{aligned}\max\{\|T_\star - C\|_F^2, \| - T_\star - C\|_F^2\} &= R^2 + \|C\|_F^2 + 2|\langle T_\star, C\rangle_F| \\ &\geq R^2 + \|C\|_F^2.\end{aligned}$$

Thus $V_t$ is the unique minimax center. The triangle inequality bounds every pair distance by $2R$, and the pair $V_t \pm T_\star$ attains it.

For a linear functional, set $J = L^\top U M_t \in \mathbb{R}^{p \times q}$. Then

$$\langle U, LQM_t^\top\rangle_F = \langle J, Q\rangle_F.$$

Write a rectangular singular-value decomposition $J = U_J[D \;\; 0]V_J^\top$. For $QQ^\top = I_p$, the matrix $U_J^\top Q V_J$ has operator norm one, so each diagonal entry has absolute value at most one. Hence $\langle J, Q\rangle_F \leq \operatorname{tr} D = \|J\|_*$. Equality holds for $Q = U_J[I_p \;\; 0]V_J^\top$, and its negative attains the opposite deviation. The case $\Sigma = 0$ follows directly from $H = 0$. □

**Interpretation of the set.** The support extrema and diameter are attainable, but the uncertainty set need not fill its enclosing ball. In the elementary case $m = p = q = o = 1$ with $\Sigma > 0$ and $M_t \neq 0$, it contains exactly two points. A support interval therefore specifies attainable extrema, although intermediate values may be absent. The radius measures error around the optimal visible center and can differ from the maximum absolute output norm $\sup_{D \in \mathcal{D}_t} \|D\|_F$ when $V_t \neq 0$.

**Training-space computation of the response spectrum.** The uncertainty radius depends on the hidden response spectrum. The gradient closure lets us compute that spectrum using training-side quantities alone. Because $A_0 N = 0$,

$$\begin{aligned} M_t &= B_t A_t N = F_t K C_t^\top X N = R_t^\top X N \\ &= \widetilde{B}_t \widetilde{C}_t^\top X N. \end{aligned}$$

The product with its transpose gives $M_t M_t^\top = R_t^\top G_\perp R_t$. The singular values, and hence the radius, require no probe cross-Gram. The computation of $G_\perp$ requires the fixed training features and initial visible subspace, which the unweighted initial kernel alone need not determine.

### B.4 Queries and Exact Information–Error Frontiers

Additional shared measurements reduce the uncertainty just characterized. We first compute the remaining error after a particular observation, then optimize the measurement directions against the worst compatible probe.

*Conditional part of Theorem 2.* Complete the orthonormal columns of $Q_k$ to $[Q_k, Q_\perp]$. The observation $W = HQ_k$ leaves the decomposition

$$H = WQ_k^\top + JQ_\perp^\top, \qquad JJ^\top = \Sigma - WW^\top.$$

The cross terms vanish because $Q_k^\top Q_\perp = 0$. Thus an observation is feasible precisely when its residual Gram is PSD and has rank at most $q - k$. The corresponding corrections are

$$V_t + W(M_t Q_k)^\top + J(M_t Q_\perp)^\top.$$

Theorem 3, applied to this residual problem, gives the conditional center and radius. □

*Worst-case protocol part of Theorem 2.* First fix $Q_k$ and put $s = q - k$. For every feasible observation, $0 \preceq \Sigma - WW^\top \preceq \Sigma$ and its rank is at most $s$. The conditional radius is therefore bounded by

$$\sum_{i=1}^{\min(p,s)} \lambda_i(\Sigma)\sigma_i(M_t Q_\perp)^2. \tag{25}$$

To construct a feasible observation that attains this upper bound, factor $\Sigma$ using its ordered eigenvectors. Place its largest $\min(p, s)$ modes in the $s$ unqueried coordinate directions, and place its remaining modes in the $k$ queried directions. The condition $p \le q$ ensures that all modes fit within these coordinates. Then $\Sigma - WW^\top$ has exactly the largest $\min(p, s)$ eigenvalues of $\Sigma$. The conditional minimax radius equals (25). The optimal decoder for this fixed query is its conditional center.

Let $\mu_j = \sigma_j(M_t)^2$, including zeros to length $q$. The eigenvalues $\nu_i$ of $Q_\perp^\top M_t^\top M_t Q_\perp$ satisfy

$$\nu_i \ge \mu_{i+k}, \qquad 1 \le i \le q - k.$$

For completeness, intersect the span of the largest $i + k$ eigenvectors of $M_t^\top M_t$ with $\mathrm{range}(Q_\perp)$. The intersection has dimension at least $i$, and every unit vector in it has Rayleigh quotient at least $\mu_{i+k}$. The variational characterization of the $i$th eigenvalue of the compression gives the inequality. This inequality applies to each term of (25) and gives the lower bound

$$\sum_{i=1}^{\min(p,q-k)} \lambda_i(\Sigma)\sigma_{i+k}(M_t)^2.$$

Choose $Q_k$ to span the largest $k$ right singular directions of $M_t$. Its orthogonal complement then has exactly the remaining singular values, so equality holds.

If $\Sigma \neq 0$ and $k < \mathrm{rank}(M_t)$, the first term $\lambda_1(\Sigma)\sigma_{k+1}(M_t)^2$ is positive. If $k \ge \mathrm{rank}(M_t)$, all terms vanish. Since $M_t = B_t A_t N$ has rank at most $r$ (and at most the compact training-active rank $k_0 = \mathrm{rank}(K)$), at most $r$ measurements per probe suffice. □

**A deterministic adaptive single-probe lower bound.** The main frontier fixes all query directions in advance. For a single probe, an adversarial argument also gives a lower bound when a deterministic procedure chooses later directions from earlier answers. Let one probe's hidden vector satisfy $\|h\|_2 = \rho > 0$. Consider any deterministic procedure taking $k < q$ noiseless linear measurements, whose next direction may depend on previous answers. Follow its all-zero answer branch. The queried directions on that branch span a subspace of dimension at most $k$. Its orthogonal complement contains a unit vector $u$ with $\|M_t u\|_2 \geq \sigma_{k+1}(M_t)$ by the same compression inequality. Both hidden probes $\rho u$ and $-\rho u$ give all-zero answers and are therefore consistent with this branch. Any common prediction has worst-case error at least $\rho\|M_t u\|_2 \geq \rho\sigma_{k+1}(M_t)$. The fixed top-singular-direction protocol attains this value. This adaptive result concerns deterministic linear measurements and worst-case error. Randomized expected-risk criteria and nonlinear real-valued encodings are different information models.

**Why a conditional frontier needs its residual Gram.** The conditional uncertainty depends on $\Sigma - WW^\top$, which generally differs from $\Sigma$ after coordinates have been observed. For a single probe with all its norm in the queried subspace, the residual radius is zero, even if the unqueried response $M_t Q_\perp$ is nonzero. The fixed-query worst-case expression remains correct because its adversary is allowed to allocate norm to unobserved directions. The theorem selects a common, nonadaptive set of directions before the probe scores are observed. Its worst-case optimum need not be optimal for each realized conditional problem.

### B.5 Trajectory Information and Rank-One Dynamics

The recovery bounds depend on a single response matrix or a stacked path. We now justify the shared-path characterization and construct a gradient-descent trajectory whose stack has much higher rank than any checkpoint.

*Proof of Proposition 1.* Let $\mathcal{R}$ be the row space of $M_{\mathcal{T},w}$. A subspace $S$ satisfies $M_t = M_t \Pi_S$ for every checkpoint if and only if it contains every row of every $M_t$, hence if and only if $\mathcal{R} \subseteq S$. The smallest possible dimension is therefore $\dim \mathcal{R} = \operatorname{rank}(M_{\mathcal{T},w})$.

For an orthonormal $Q$, block expansion gives

$$\|M_{\mathcal{T},w}(I - QQ^\top)\|_F^2 = \sum_j w_j \|M_{t_j}(I - QQ^\top)\|_F^2.$$

The Eckart–Young theorem states that the top $k$ right singular vectors minimize this quantity and that its value is $\sum_{i>k} \sigma_i^2$. This proves the approximation formula and the effective-complexity characterization.

Any other orthonormal basis of the same hidden space is $N' = NO$ for an orthogonal $O$, so $M_t' = M_t O$ and the stacked rank and singular values are unchanged. The rank cannot decrease as the stack acquires more rows, which proves monotonicity. Subadditivity gives $\operatorname{rank}(M_{\mathcal{T},w}) \leq \sum_j \operatorname{rank}(M_{t_j}) \leq \ell r$. The ambient hidden dimension also bounds the rank by $q$. Finally, the factor gradient keeps $A_t - A_0$ in the row span of $X$. Since $A_0 N = 0$, every row of $M_t = B_t(A_t - A_0)N$ lies in the row space of $XN$, giving the hidden training-span bound. For the total path, every row of $A_t$ lies in the row span of $[A_0^\top, X^\top]^\top$, which gives the stated total-path bound. □

**Stacking checkpoints.** The preceding recovery argument extends to a path because the same unknown probe acts at every checkpoint. We make this shared dependence explicit by stacking the responses. For times $t_1, \ldots, t_\ell$, concatenate corrections horizontally. With $M_{\mathcal{T}}$ stacked vertically,

$$[\Delta f_{t_1}(P), \ldots, \Delta f_{t_\ell}(P)] = [V_{t_1}, \ldots, V_{t_\ell}] + H M_{\mathcal{T}}^\top.$$

The unknown $H$ is shared across times, so every attainable-set, support-function, and query result applies directly. For positive checkpoint weights, multiply each block by the square root of its weight. These weights preserve algebraic rank but change the singular values and the resulting risk. Since each $M_t = R_t^\top XN$, $\operatorname{row}(M_{\mathcal{T}}) \subseteq \operatorname{row}(XN)$. This proves the training-span upper bound.

*Proof of Theorem 1.* Choose $n = d = q + 1$, $\mathcal{H} = \mathrm{diag}(\lambda_0, \ldots, \lambda_q)$ with distinct positive diagonal entries, and a unit vector $v$ whose coordinates are all nonzero. Take $X = \sqrt{n}\,\mathcal{H}^{1/2}$ and $Y = \delta X v$. Then $X^\top X/n = \mathcal{H}$ and $X^\top Y/n = \delta \mathcal{H} v$. With $A_t = a_t^\top$ and $B_t = b_t$, gradient descent reads

$$\begin{aligned} a_{t+1} &= (I - \eta_A b_t^2 \mathcal{H}) a_t + \eta_A b_t \delta \mathcal{H} v, \\ b_{t+1} &= b_t + \eta_B \{\delta a_t^\top \mathcal{H} v - b_t a_t^\top \mathcal{H} a_t\}. \end{aligned} \tag{26}$$

The choice $a = v, b = \delta$ fits the target exactly within the initial row space.

First assume $b_1 \notin \{0, \delta\}$ and $b_t \neq 0$ for $2 \leq t \leq q+1$. The first step has $a_1 = v$ and $b_1 = \eta_B \delta v^\top \mathcal{H} v$. Inductively $a_t = p_t(\mathcal{H}) v$, where

$$p_0(\lambda) = p_1(\lambda) = 1, \qquad p_{t+1}(\lambda) = (1 - \eta_A b_t^2 \lambda) p_t(\lambda) + \eta_A b_t \delta \lambda.$$

In particular,

$$p_2(\lambda) = 1 + \gamma\lambda, \qquad \gamma = \eta_A b_1 (\delta - b_1) \neq 0.$$

For $t \geq 2$, let $\ell_t$ be the leading coefficient of $p_t - 1$. The recurrence implies

$$\deg(p_t - 1) = t - 1, \qquad \ell_2 = \gamma, \qquad \ell_{t+1} = -\eta_A b_t^2 \ell_t \neq 0.$$

The additive term has degree one and cannot cancel the highest degree for $t \geq 2$.

Let $N$ span $v^\perp$. The $q$ vectors $N^\top \mathcal{H}^j v$, $j = 1, \ldots, q$, are independent. Otherwise, for some nonzero coefficients $c_j$, $\sum_{j=1}^q c_j \mathcal{H}^j v$ would lie in the span of $v$. For a scalar $c_0$, the polynomial $\sum_{j=1}^q c_j \lambda^j - c_0$ would then vanish at all $q+1$ distinct eigenvalues, since every coordinate of $v$ is nonzero. A polynomial of degree at most $q$ with $q+1$ distinct roots is zero, contradicting the choice of coefficients. The vectors $N^\top a_t$, $t = 2, \ldots, q+1$, are an invertible triangular combination of this basis, because their polynomial degrees increase from one to $q$ with nonzero leading coefficients. The nonzero scalars $b_t$ preserve independence, so the rows $M_t = b_t a_t^\top N$ are independent.

It remains to justify the nonzero-iterate condition for sufficiently small common step size $\eta_A = \eta_B = \eta$. For each fixed finite $t$, induction in (26) gives

$$a_t = v + O(\eta^2), \qquad b_t = t\eta\,\delta v^\top \mathcal{H} v + O(\eta^2) \quad (\eta \downarrow 0).$$

The constants may depend on $t, \mathcal{H}, v, \delta$. The first update gives the base case. For the induction step, the increment of $a_t$ is $O(\eta^2)$, while the increment of $b_t$ is $\eta\delta v^\top \mathcal{H} v + O(\eta^2)$. There are only finitely many indices $1, \ldots, q+1$. Since $\delta v^\top \mathcal{H} v > 0$, one can choose a common $\eta_0 > 0$ so that every $b_t > 0$ on $0 < \eta < \eta_0$. If necessary, decrease $\eta_0$ until $\eta_0 v^\top \mathcal{H} v < 1$, which ensures $b_1 < \delta$. All conditions used above now hold. The positive-risk criterion in Theorem 2, applied to the stacked response, proves the exact-measurement requirement. □

*Proof of Corollary 1.* Let $\lambda_1, \ldots, \lambda_s$ be the distinct eigenvalues with $P_{\lambda_j} v \neq 0$. Their projected vectors are nonzero and mutually orthogonal. Therefore

$$\mathcal{K} = \mathrm{span}\{P_{\lambda_1} v, \ldots, P_{\lambda_s} v\} = \mathrm{span}\{v, \mathcal{H}v, \ldots, \mathcal{H}^{s-1} v\}$$

has dimension $s$, where the second equality is the Vandermonde identity on the $s$ distinct active eigenvalues. The polynomial recurrence in (26) keeps $a_t$ in $\mathcal{K}$ at every finite time. Since $v \in \mathcal{K}$, projecting $\mathcal{K}$ onto $v^\perp$ leaves dimension $s - 1$. The scalar factor $b_t$ cannot increase this dimension. This proves the upper bound for arbitrary finite iterates and step sizes in the stated GD model.

For sufficiently small common positive step size, the preceding finite-horizon argument gives $0 < b_1 < \delta$ and $b_t > 0$ through time $s$. The polynomials $p_t - 1$, $t = 2, \ldots, s$, have degrees $1, \ldots, s-1$ with nonzero leading coefficients. The vectors $N^\top \mathcal{H}^j v$, $1 \leq j \leq s-1$, are independent: a dependence would imply that a polynomial of degree at most $s - 1$ vanishes on all $s$ active eigenvalues after subtracting a constant, and hence that all its coefficients vanish. The same triangular-basis argument proves attainability of rank $s - 1$. When $s = 1$, $\mathcal{K} = \mathrm{span}(v)$ and its hidden projection is zero at every step, independently of step size. □

**A rational three-dimensional instance.** The general construction can be checked directly in three dimensions. The following rational parameters give two independent hidden directions from rank-one updates. Take

$$\mathcal{H} = \mathrm{diag}(1,2,3), \quad v = \tfrac{1}{3}(1,2,2)^\top, \quad N = \tfrac{1}{3}\begin{pmatrix} 2 & 2 \\ -2 & 1 \\ 1 & -2 \end{pmatrix}, \quad \delta = 1, \quad \eta_A = \eta_B = \tfrac{1}{10}.$$

Then $N^\top N = I$, $N^\top v = 0$, and

$$b_1 = \tfrac{7}{30}, \qquad b_2 = \tfrac{371}{900}, \qquad \gamma = \tfrac{161}{9000}.$$

The first hidden Krylov vectors are

$$N^\top \mathcal{H} v = (0, -2/3)^\top, \qquad N^\top \mathcal{H}^2 v = (4/9, -26/9)^\top,$$

whose determinant is $8/27$. The third scalar iterate is the positive rational number

$$b_3 = b_2 + \tfrac{1}{10}\left[\tfrac{7}{3} + \gamma\tfrac{53}{9} - b_2\left(\tfrac{7}{3} + 2\gamma\tfrac{53}{9} + \gamma^2\tfrac{47}{3}\right)\right].$$

Consequently the two-row hidden-response matrix satisfies

$$\det\begin{pmatrix} M_2 \\ M_3 \end{pmatrix} = -\frac{8}{27}\,\frac{1}{10}\,b_2^3 b_3 \gamma^2 \neq 0.$$

The second-moment updates in (26) allow exact rational-arithmetic checks of every state. For a four-dimensional instance, take $v = (1,1,1,1)^\top/2$, $\mathcal{H} = \mathrm{diag}(1,2,3,4)$, and the same $\delta$ and step size. The first four scalar iterates are positive, and the same polynomial argument proves that the response rows at times $2, 3, 4$ are independent.

**An embedding with a strictly low-rank matrix update.** The scalar-output construction also embeds in a matrix-valued regression problem. This extension keeps each update rank one while allowing both matrix dimensions to exceed one. For output dimension $o \geq 2$, fix a unit vector $u \in \mathbb{R}^o$, use target $Y = \delta(Xv)u^\top$, and initialize $B = 0$. The gradients keep $B_t = b_t u$, so the scalar proof applies unchanged. Every adapter correction $B_t A_t$ has rank at most one.

**Corollary 2** (Vanishing endpoint uncertainty at an interpolating limit)**.** *In the temporal construction, at every checkpoint*

$$\|M_t\|_F \leq \sqrt{\frac{2\mathcal{L}_t}{\lambda_{\min}(\mathcal{H})}}, \qquad \mathcal{R}_t^2 \leq \frac{2\lambda_{\max}(\Sigma)\mathcal{L}_t}{\lambda_{\min}(\mathcal{H})}.$$

*Thus any trajectory whose training loss tends to zero has vanishing hidden endpoint response and vanishing conditional endpoint radius, even when a fixed collection of earlier checkpoints has temporal rank q.*

*Proof.* Let $w_t = b_t a_t$. Since the target is $\delta X v$,

$$2\mathcal{L}_t = (w_t - \delta v)^\top \mathcal{H}(w_t - \delta v) \geq \lambda_{\min}(\mathcal{H})\|w_t - \delta v\|_2^2.$$

Because $N^\top v = 0$ and $N$ has orthonormal columns, $\|M_t\|_F = \|N^\top(w_t - \delta v)\|_2 \leq \|w_t - \delta v\|_2$. Finally $\mathcal{R}_t^2 \leq \lambda_{\max}(\Sigma)\|M_t\|_F^2$. □

The corollary assumes that the loss converges to zero. Global convergence of discrete bilinear gradient descent requires a separate proof. The construction uses full-column-rank $X$, so the unique interpolating weight map is $\delta v^\top$. Its hidden component is zero. The temporal-rank result therefore describes information needed for the training path even when the endpoint lies in the initial space. At small step size, some higher response singular values can be very small despite nonzero algebraic rank. Quantitative approximate prediction must use the weighted stacked spectrum and the frontier in (9).

## C Path Measures, Spectral Compression, and Finite-Time Responses

Exact path dimension records which directions appear; approximation also depends on their weights and energies. This appendix proves the comparisons in Section 4 and explains when a concentrated spectrum supports reuse of a fixed subspace.

### C.1 Common input directions and the spectral variational identity

We first specify the shared object that the spectrum measures. The same input projector must approximate every checkpoint, which differs from expressing whole matrices in a vectorized basis. For a finite collection, the exact quantity used throughout this paper is

$$\dim \operatorname{span}\left(\bigcup_t \operatorname{row}(M_t)\right),$$

not $\dim \operatorname{span}\{\operatorname{vec}(M_t)\}_t$. A fixed rank-$r$ matrix has the former dimension $r$ and the latter dimension one. Conversely, $M_t = u_t v^\top$ can have an arbitrarily large vectorized span while its shared input dimension is one. For any rank-$k$ orthogonal projector $\Pi$,

$$\int \|M(t)(I-\Pi)\|_F^2 d\mu(t) = \operatorname{tr} G_\mu - \operatorname{tr}(\Pi G_\mu).$$

The spectral variational principle bounds the second term by the sum of the largest $k$ eigenvalues and identifies a projector that attains it. This gives the formula for $E_k(\mu)$. If $G_\mu = 0$, we use $E_k = 0$ and $C_{\varepsilon,\mu} = 0$. For a continuous-time path, the rank of $G_\mu$ describes the row span outside a $\mu$-null set. For a continuous path and a measure with full support, this equals the span over all times: a continuous nonzero quadratic response is positive on an open neighborhood, which a full-support measure cannot ignore.

### C.2 Proof of the measure comparison

The variational identity isolates the approximation error for any fixed projector. We use it to compare two weighting measures whose density ratio stays bounded above and away from zero. Suppose $a \le d\nu/d\mu \le b$ almost everywhere. If $G_\mu = 0$, then $G_\nu = 0$ and the convention above proves the claim. Otherwise both energy denominators are positive. For any fixed projector, both its squared residual integral and the total-energy integral are between $a$ and $b$ times their counterparts under $\mu$. Dividing gives

$$\frac{a}{b}\frac{\int \|M(I-\Pi)\|_F^2 d\mu}{\int \|M\|_F^2 d\mu} \le \frac{\int \|M(I-\Pi)\|_F^2 d\nu}{\int \|M\|_F^2 d\nu} \le \frac{b}{a}\frac{\int \|M(I-\Pi)\|_F^2 d\mu}{\int \|M\|_F^2 d\mu}.$$

The infimum over rank-$k$ projectors then gives (11). Equivalently, $aG_\mu \preceq G_\nu \preceq bG_\mu$. The result requires bounded density ratios in both directions. Intervals with zero speed can violate this condition for time and arc measures.

### C.3 Arc length and exact integration on the measured polygonal path

The comparison above concerns different weights on the same path. For arc-length weighting, we also need to establish invariance to traversal speed and specify the curve represented by saved checkpoints. Let $M : [0,T] \to \mathbb{R}^{o\times q}$ be absolutely continuous with $0 < L < \infty$. For an increasing $C^1$ bijection $\phi : [0,S] \to [0,T]$ with positive derivative, define $\widetilde{M}(u) = M(\phi(u))$. The chain rule and change of variables yield

$$\int_0^S \widetilde{M}(u)^\top \widetilde{M}(u) \|\widetilde{M}'(u)\|_F du = \int_0^T M(t)^\top M(t) \|M'(t)\|_F dt.$$

The same substitution preserves $L$, proving the stated invariance. The formula preserves traversal multiplicity, unlike a measure that depends only on the unordered image of the curve.

For consecutive measured matrices $A, B$, write $\delta = \|B - A\|_F$. On the straight segment $M(u) = (1-u)A + uB$, its unnormalized arc Gram contribution is exactly

$$\delta\left\{\frac{A^\top A + B^\top B}{3} + \frac{A^\top B + B^\top A}{6}\right\}. \tag{27}$$

The integral of this quadratic expression in $u$ gives the formula. Equivalently, evaluate the two matrices at $u = 1/2 \pm 1/\sqrt{12}$, stack them with weight $\delta/(2L)$, and compute the singular values. For factors,

$$(1-u)s_j B_j A_j + u s_{j+1} B_{j+1} A_{j+1} = [(1-u)s_j B_j,\ u s_{j+1} B_{j+1}] \begin{bmatrix} A_j \\ A_{j+1} \end{bmatrix}.$$

This factor representation gives the exact polygonal integral without forming dense updates. The interpolated matrices can have rank up to $2r$ and serve only as integration points, not saved LoRA checkpoints. The integral is unchanged by subdivision of a straight segment, but can differ from the integral along an unobserved curved training path.

Endpoint trapezoidal arc weighting instead adds $\delta(A^\top A + B^\top B)/2$. Its excess is $\delta(B-A)^\top(B-A)/6 \succeq 0$. Across segments, the normalized trace discrepancy is

$$\operatorname{tr}(G_{\mathrm{trap}} - G_{\mathrm{poly}}) = \frac{\sum_j \delta_j^3}{6\sum_j \delta_j} \le \frac{1}{6}\max_j \delta_j^2.$$

We therefore distinguish exact polygonal integration from endpoint trapezoidal approximation. A zero-length constant curve uses its single-state spectrum, including when that state is nonzero.

### C.4 An exact finite-horizon spectral identity

We next turn from the choice of weights to the dynamics that produce the spectrum. In the aligned rank-one model, the polynomial response gives an exact expression for the energy of a finite path. Let $u_i = P_{\lambda_i} v/c_i$, $U = [u_1, \dots, u_s]$, $c = (c_i)$, and $D_c = \operatorname{diag}(c_i)$. The columns of $U$ are orthonormal and $v = Uc$. Write $p_t = (p_t(\lambda_i))_{i=1}^s$. In the aligned rank-one model,

$$a_t = UD_c p_t, \qquad (I - vv^\top) b_t a_t = U(I - cc^\top) b_t D_c p_t.$$

The nonzero hidden path energies are therefore the eigenvalues of

$$(I - cc^\top)\left(\sum_t w_t b_t^2 D_c p_t p_t^\top D_c\right)(I - cc^\top). \tag{28}$$

Equivalently, the checkpoint Gram matrix has entry

$$\sqrt{w_j w_k}\, b_j b_k \left\{\sum_i c_i^2 p_j(\lambda_i) p_k(\lambda_i) - \left(\sum_i c_i^2 p_j(\lambda_i)\right)\left(\sum_i c_i^2 p_k(\lambda_i)\right)\right\}.$$

Both identities follow by taking inner products of the displayed hidden vectors. They express a finite-time spectrum exactly and separate algebraic independence from energy concentration.

### C.5 Proof of Proposition 2

The spectral identity suggests that nearby eigenvalues can have nearly identical responses. We formalize this idea by replacing each group with one representative polynomial value, then bounding the resulting hidden-path error. For group $g$, let $v_g = \sum_{i\in g} P_{\lambda_i} v$ and define $\widetilde{a}_t = \sum_g p_t(\zeta_g) v_g$. The mean-value theorem and orthogonality of eigenspaces give

$$\|a_t - \widetilde{a}_t\|_2^2 = \sum_g \sum_{i\in g} c_i^2 |p_t(\lambda_i) - p_t(\zeta_g)|^2 \le \sum_g \alpha_g d_g^2 L_{t,g}^2.$$

All vectors $(I - vv^\top)\widetilde{a}_t$ lie in $S = \operatorname{span}\{(I - vv^\top)v_g : g = 1, \dots, m\}$. The corresponding subspace in hidden input coordinates is $N^\top S$. Since $\sum_g v_g = v$, the projected generators sum to zero, so $\dim S \le m - 1$. Orthogonal projection cannot increase the error. After multiplication by $b_t$, the weighted sum of squared errors is at most the numerator of (12). The optimum over subspaces satisfies the same bound, and division by $D$ proves the claim. The polynomial derivatives are bounded on compact spectral hulls at any fixed horizon, although these bounds can grow with the horizon.

### C.6 Why effective dimension has no general monotonicity

The clustering bound concerns approximation at a fixed horizon. Before interpreting changes across horizons, we distinguish the monotonicity of exact rank from the behavior of an energy threshold. Always $0 \le C_{\varepsilon,\mu} \le C_{\mathrm{exact}}$, and for positive checkpoint weights $\max_t \operatorname{rank}(M_t) \le C_{\mathrm{exact}}$. The first bound does not compare effective dimension with pointwise algebraic rank: $\operatorname{diag}(1, 0.01)$ has rank two but its 99%-energy dimension is one. The effective dimension can also decrease as the horizon grows. Two equally weighted row vectors $e_1^\top, e_2^\top$ require two directions at 99% energy. An additional checkpoint $10e_1^\top$ changes the directional energies to $101 : 1$, so one direction suffices even though the exact dimension remains two.

### C.7 Cumulative residual and out-of-window transfer

A lower effective dimension at a later horizon does not establish that earlier directions remain useful. The following bounds separate a newly optimized approximation from reuse of the earlier projector. For positive semidefinite $C, D$, the variational identity gives

$$R_k(C+D) = \min_{P=P^\top=P^2,\,\operatorname{rank} P=k} \operatorname{tr}((I-P)(C+D)).$$

The second summand is nonnegative for every $P$, proving the lower bound in (13). The optimizer $P_k$ for $C$ gives a feasible upper bound. Division by the new trace yields

$$\frac{R_k(C+D)}{\operatorname{tr}(C+D)} \le \frac{\operatorname{tr} C}{\operatorname{tr} C + \operatorname{tr} D} q + \frac{\operatorname{tr} D}{\operatorname{tr} C + \operatorname{tr} D}\beta.$$

In particular, $\beta \le q$ implies that the relative tail cannot increase. This sufficient condition is not necessary, because a new optimal subspace can reduce the relative tail even if the earlier projector captures little of the new energy. The same proof applies to additive trapezoidal window Grams when boundaries are saved checkpoints. The argument requires additive, unnormalized Grams. It does not establish monotonicity when normalization or a longer horizon changes the earlier weights arbitrarily.

Our transfer diagnostic is $1-\beta$. It measures late *state* energy captured by an early projector. For the polygonal matrix path we separately compute late squared-velocity capture,

$$\frac{\sum_{j:t_j\ge 800} \|(M_{j+1}-M_j)P_k\|_F^2/(t_{j+1}-t_j)}{\sum_{j:t_j\ge 800} \|M_{j+1}-M_j\|_F^2/(t_{j+1}-t_j)}.$$

This quantity measures movement along the saved path. It differs from the arc-occupancy spectrum and does not estimate unsaved stochastic gradients. Unlike state capture, it is unaffected by adding a constant matrix to all late states.

### C.8 Data weighting does not order effective dimensions

So far the approximation weights have acted across time. Input data introduce another weighting through the response matrix, and this can change relative spectral concentration even when exact rank decreases. For a fixed input bank $X \in \mathbb{R}^{q\times n}$, stacking responses gives $\mathcal{M}X$, where $\mathcal{M}$ is the weighted stack of operators. Hence its exact rank is at most $\operatorname{rank}\mathcal{M}$. Normalized spectral dimensions need not obey this ordering. For the constant operator $M = \operatorname{diag}(10, 1)$ and $X = \operatorname{diag}(0.1, 1)$, the operator's squared energies are $(100, 1)$, so $k_{99}(M) = 1$, whereas $MX = I_2$ and $k_{99}(MX) = 2$. Conversely, $M = I_2$, $X = \operatorname{diag}(1, 0.01)$ gives 2 versus 1.

A conditional comparison is possible. If $aI \preceq XX^\top \preceq bI$ with $0 < a \le b < \infty$, then $a\mathcal{M}\mathcal{M}^\top \preceq \mathcal{M}XX^\top\mathcal{M}^\top \preceq b\mathcal{M}\mathcal{M}^\top$. Eigenvalue monotonicity, followed by bounds on the tail and trace, gives $(a/b)E_k(\mathcal{M}) \le E_k(\mathcal{M}X) \le (b/a)E_k(\mathcal{M})$. Rank-deficient input banks need not satisfy a positive lower bound. With evolving inputs $X_t$, the exact-rank comparison to the operator stack can also fail. For example, the constant row $M_t = e_1^\top$ has shared input dimension one. If $X_1 = I_2$ and $X_2$ swaps the coordinates, the response rows are $e_1^\top, e_2^\top$, with shared token-coordinate dimension two. The lower response dimensions measured on matched input banks therefore describe those observations. They do not establish a universal ordering of dimensions or bound whole-model prediction error.

### C.9 A generic convergence lemma and its zero-limit boundary

The finite-horizon results require no convergence assumption. To complete the comparison, we show how convergence to a nonzero low-rank limit yields concentration under uniform-time weighting, and why a zero limit is different.

**Lemma 2** (Uniform-time concentration at a nonzero limit)**.** *Let $M : [0,\infty) \to \mathbb{R}^{o\times q}$ be locally square-integrable and satisfy $M(t) \to M_\infty \ne 0$. Set $G_T = T^{-1}\int_0^T M(t)^\top M(t)\,dt$. Then $G_T \to M_\infty^\top M_\infty$ in any matrix norm. If $\operatorname{rank}(M_\infty) \le r$, then $\sum_{i>r}\lambda_i(G_T)/\operatorname{tr}(G_T) \to 0$. The analogous statement holds for discrete uniform averages.*

*Proof.* Let $F(t) = M(t)^\top M(t)$ and $F_\infty = M_\infty^\top M_\infty$. Continuity of matrix multiplication gives $F(t) \to F_\infty$. Given $\epsilon > 0$, choose $T_0$ such that $\|F(t) - F_\infty\| \le \epsilon$ for $t \ge T_0$. Local square

integrability makes $C = \int_0^{T_0} \|F(t) - F_\infty\|\, dt$ finite. For $T \geq T_0$, $\|G_T - F_\infty\| \leq C/T + \epsilon$. Continuity of eigenvalues implies that the tail beyond $r$ tends to zero, while $\operatorname{tr} G_T \to \|M_\infty\|_F^2 > 0$. The same argument with sums in place of integrals proves the discrete statement. □

The nonzero condition prevents a vanishing denominator. For a counterexample with rank-one states and zero limit, take $M(t) = [\sqrt{2}, 0]$ on $[0, 1/2)$, $M(t) = [0, \sqrt{2}]$ on $[1/2, 1)$, and $M(t) = 0$ thereafter. For $T \geq 1$, $G_T = I_2/T$, whose relative tail beyond one is $1/2$. Thus convergence to a rank-zero limit alone does not imply a vanishing relative tail. The lemma describes a general averaging effect without specifying a convergence rate or a mechanism specific to LoRA.

# D From Shared Input Directions to Predictive Perturbations

The path spectrum controls a matrix approximation objective. We give sufficient conditions for that approximation to bound teacher-forced prediction loss. We also distinguish the objectives of the endpoint, path, and random-subspace controls. All statements apply to fixed trained networks, regardless of the optimizer used.

## D.1 The actual deleted operator

Before following an intervention through the network, we identify exactly which matrix it removes. The projection keeps the initialization component and acts only on the complementary input directions. For adapter site $\ell$ at checkpoint $t$, let $\Delta W_{\ell,t} = s_{\ell,t} B_{\ell,t} A_{\ell,t}$ and let $\Pi_{\ell,0}$ project onto the row space of its initial $A$. Choose $N_\ell$ with orthonormal columns satisfying $N_\ell N_\ell^\top = I - \Pi_{\ell,0}$, and set $M_{\ell,t} = \Delta W_{\ell,t} N_\ell$. We implement a projector $P_\ell$ in these complementary coordinates through the replacement

$$\widetilde{\Delta W}_{\ell,t} = \Delta W_{\ell,t} \Pi_{\ell,0} + M_{\ell,t} P_\ell N_\ell^\top, \tag{29}$$

$$E_{\ell,t} := \Delta W_{\ell,t} - \widetilde{\Delta W}_{\ell,t} = M_{\ell,t}(I - P_\ell) N_\ell^\top. \tag{30}$$

Thus $\|E_{\ell,t}\|_{\mathrm{F}} = \|M_{\ell,t}(I - P_\ell)\|_{\mathrm{F}}$ exactly. Equivalently, one replaces $A_{\ell,t}$ by $A_{\ell,t}\Pi_{\ell,0} + A_{\ell,t} N_\ell P_\ell N_\ell^\top$ and leaves $B_{\ell,t}$ and its scaling unchanged. The dimension of $P_\ell$ is a shared hidden-input budget, not the rank or storage size of the resulting LoRA factors.

## D.2 Local response requires an uncentered second moment

The deleted operator matters through the activations on which it acts. We first relate its matrix norm to the local response change, before accounting for the rest of the network.

**Proposition 5** (Exact local response and an activation-weighted bound)**.** *Let $E \in \mathbb{R}^{o \times d}$ be a fixed deleted operator and $H \in \mathbb{R}^{d \times n}$ a random input tensor with finite second moment. The number of columns may depend on the input. Put $\Sigma = \mathbb{E}[HH^\top]$. Then*

$$\mathbb{E}\|EH\|_{\mathrm{F}}^2 = \operatorname{tr}(E\Sigma E^\top) \leq \beta \|E\|_{\mathrm{F}}^2 \quad \textit{if } \Sigma \preceq \beta I. \tag{31}$$

*If also $\Sigma \succeq \alpha I$, the left-hand side is at least $\alpha \|E\|_{\mathrm{F}}^2$. For a vector $h$ of mean $m$ and centered covariance $C$,*

$$\Sigma = C + mm^\top, \qquad \mathbb{E}\|Eh\|_2^2 = \operatorname{tr}(ECE^\top) + \|Em\|_2^2. \tag{32}$$

*Proof.* The identity follows by linearity of expectation in $\operatorname{tr}(EHH^\top E^\top)$. The map $\Sigma \mapsto E\Sigma E^\top$ preserves the semidefinite inequalities, and the trace gives the bounds. The expansion $h = (h - m) + m$ gives the last identity because the cross terms have zero mean. □

The moment in (31) sums across the tensor's columns. A moment averaged per position requires the corresponding position-count factor. In causal language modeling, $H$ includes the context needed to predict the scored token, including prompt tokens whose own labels are masked. A per-token activation average must therefore be rescaled before it can be used in this sequence-level perturbation bound.

### D.3 SIMULTANEOUS REPLACEMENT OF MULTIPLE ADAPTERS

A bound at one site does not yet control a joint intervention, because earlier replacements can change later activations. We handle this dependence through hybrid networks that replace one site at a time. Fix a checkpoint and enumerate the $m$ adapter sites in a topological order of the network's computation graph. We assume deterministic evaluation throughout this argument. An example $x$ comprises a fixed teacher-forced token context and a scored next-token label $y(x)$. Let $z(x)$ be its vocabulary logit vector. The evaluation distribution $\mathcal{D}$ is the uniform distribution over all scored tokens when the reported loss is a token-weighted mean. Answer-only scoring changes this distribution, while retaining the question in each context.

Define $F^{(j)}$ to use projected operators at sites $1, \dots, j$ and original operators at sites $j+1, \dots, m$, and denote its logits by $z^{(j)}$. Adjacent hybrids have the same input $H_j(x)$ to site $j$, since all upstream computations are identical. They also have the same other tensors entering the downstream computation. Their local output difference is exactly $E_j H_j(x)$, up to sign. We use a conditional downstream Lipschitz assumption: with these other tensors fixed, changing only this site's output satisfies

$$\|z^{(j)}(x) - z^{(j-1)}(x)\|_2 \le L_j \|E_j H_j(x)\|_{\mathrm{F}} \quad \text{for } \mathcal{D}\text{-almost every } x. \tag{33}$$

A uniform bound on the induced norm of the downstream Jacobian along each relevant line segment implies (33). Its value at one endpoint alone does not provide this guarantee. Here $H_j$ is a *hybrid-network* activation, rather than an activation from an all-original probe run.

**Theorem 5** (A joint-projection bound on teacher-forced loss)**.** *Assume* (33) *with finite constants $L_j$, and let $\Sigma_j = \mathbb{E}_{\mathcal{D}}[H_j H_j^\top] \preceq \beta_j I$. For the token negative log likelihood $\ell(z, y) = \log \sum_v e^{z_v} - z_y$, put $\mathcal{L} = \mathbb{E}\ell(z^{(0)}, y)$ and $\widetilde{\mathcal{L}} = \mathbb{E}\ell(z^{(m)}, y)$, assumed finite. Then*

$$\left(\mathbb{E}\|z^{(m)} - z^{(0)}\|_2^2\right)^{1/2} \le \sum_{j=1}^{m} L_j \sqrt{\mathrm{tr}(E_j \Sigma_j E_j^\top)}, \tag{34}$$

$$|\widetilde{\mathcal{L}} - \mathcal{L}| \le \sqrt{2} \sum_{j=1}^{m} L_j \sqrt{\mathrm{tr}(E_j \Sigma_j E_j^\top)} \le \sqrt{2} \sum_{j=1}^{m} L_j \sqrt{\beta_j} \|E_j\|_{\mathrm{F}}. \tag{35}$$

*Proof.* The telescoping identity is $z^{(m)} - z^{(0)} = \sum_j (z^{(j)} - z^{(j-1)})$. Minkowski's inequality in the space of square-integrable logit vectors, (33), and Proposition 5 prove (34). Writing $p = \mathrm{softmax}(z)$ gives $\nabla_z \ell = p - e_y$ and $\|p - e_y\|_2^2 = \|p\|_2^2 + 1 - 2p_y \le 2$. Integration along the segment between two logit vectors therefore gives $|\ell(z', y) - \ell(z, y)| \le \sqrt{2}\|z' - z\|_2$. The expectation of this inequality, together with Cauchy–Schwarz, gives (35). □

This argument includes attention branches and residual connections through the conditional downstream map. A bound stated in terms of the all-original activation requires an additional comparison with $H_j$. For example, if $\mathbb{E}\|H_j - H_j^{\mathrm{orig}}\|_{\mathrm{F}}^2 \le d_j^2$, then Minkowski gives the valid replacement

$$\sqrt{\mathbb{E}\|E_j H_j\|_{\mathrm{F}}^2} \le \sqrt{\mathbb{E}\|E_j H_j^{\mathrm{orig}}\|_{\mathrm{F}}^2} + \|E_j\|_2 d_j. \tag{36}$$

The second term accounts for the difference between original and hybrid activations. The path spectrum does not estimate the theorem's sensitivity constants, so this bound does not predict the observed $10^{-3}$-scale loss differences.

### D.4 WHICH TIME AVERAGE IS CONTROLLED?

The preceding argument controls loss at one checkpoint. To connect it to a shared path projector, we average across checkpoints under the same weights used to define the approximation objective.

**Corollary 3** (A path-weighted loss bound)**.** *Let $w_t \ge 0$ sum to one over a finite checkpoint set, and use one projector $P_j$ at each site for all these checkpoints. Suppose Theorem 5 holds at every checkpoint with $L_{j,t}\sqrt{\beta_{j,t}} \le a_j$. Define*

$$G_j = \sum_t w_t M_{j,t}^\top M_{j,t}, \qquad R_j(P_j) = \mathrm{tr}((I - P_j)G_j), \qquad D_j = \mathrm{tr}\, G_j.$$

*Then the root-mean-square loss change satisfies*

$$\left(\sum_t w_t|\widetilde{\mathcal{L}}_t - \mathcal{L}_t|^2\right)^{1/2} \le \sqrt{2}\sum_j a_j\sqrt{R_j(P_j)}. \tag{37}$$

*The weighted mean absolute loss change obeys the same upper bound. If $P_j$ is a top-$k_j$ eigenspace of $G_j$, then $R_j(P_j) = \sum_{i>k_j}\lambda_i(G_j)$. In particular, retaining a fraction $1-\varepsilon_j^2$ of the path energy gives the upper bound $\sqrt{2}\sum_j a_j\varepsilon_j\sqrt{D_j}$.*

*Proof.* Apply (35) at each checkpoint, followed by Minkowski's inequality for the weighted Euclidean norm over checkpoints. The squared norm of the $j$th residual sequence is exactly $R_j(P_j)$ by (30). Cauchy–Schwarz bounds the weighted mean absolute change by its root mean square. The spectral assertion follows from the maximum of $\mathrm{tr}(P_jG_j)$ over the admissible projectors. □

The 99% and 99.9% energy thresholds correspond to $\varepsilon = 0.1$ and $\varepsilon = \sqrt{0.001}$, respectively. Both are relative root-mean-square *matrix* tolerances, not percentages of task loss. The checkpoint set and weights in the conclusion must match those in $G_j$. A four-checkpoint functional score is a separate objective from a projector fitted using 57 checkpoints.

For a designated checkpoint $u$ with $w_u > 0$, nonnegativity yields the weaker individual bound

$$\|E_{j,u}\|_{\mathrm{F}} \le \sqrt{R_j(P_j)/w_u}. \tag{38}$$

The factor $w_u^{-1/2}$ cannot in general be removed. A checkpoint with zero weight admits no bound based solely on the average residual. Even a continuous path can concentrate its error near the endpoint: for $M_n(t) = (1, \max\{0, 1-n(1-t)\})$ on $[0,1]$ and the fixed projector $P = \mathrm{diag}(1,0)$, the integrated residual is $1/(3n)$ while the endpoint residual is one. Its relative integrated residual is $1/(3n+1)$. A zero endpoint residual likewise does not control errors at earlier checkpoints.

### D.5 The endpoint, path, and random controls

The path-weighted bound connects one fitting objective to loss. The experiments also use endpoint and random subspaces, so we distinguish the geometric guarantees of these choices.

**Proposition 6** (Matched budgets order geometry)**.** *For one module, write $J_\mu(P) = \sum_t w_t\|M_t(I-P)\|_{\mathrm{F}}^2$ and $J_T(P) = \|M_T(I-P)\|_{\mathrm{F}}^2$. Let $P_{\mathrm{path}}$ minimize $J_\mu$ and $P_{\mathrm{end}}$ minimize $J_T$ among rank-$k$ orthogonal projectors in the same input space. Then*

$$J_\mu(P_{\mathrm{path}}) \le J_\mu(P_{\mathrm{end}}), \qquad J_T(P_{\mathrm{end}}) \le J_T(P_{\mathrm{path}}). \tag{39}$$

*If $k \ge \mathrm{rank}(M_T)$ and $P_{\mathrm{end}}$ contains its row space, then $J_T(P_{\mathrm{end}}) = 0$. If this holds at every adapter, the projected endpoint network equals the original endpoint network exactly.*

*Proof.* Each inequality follows from the defining minimization property. Because the projector contains the row space, $M_TP_{\mathrm{end}} = M_T$, and (30) gives $E_{j,T} = 0$ at each site. Every network operator is therefore unchanged. □

Since $\mathrm{rank}(M_T) \le r$, an exact endpoint row-space projector with budget $r$ is an identity control at the endpoint. If the row space has fewer dimensions than the budget, the control must specify how it chooses the remaining directions. Every completion preserves the endpoint but can affect earlier checkpoints differently. An early-fit projector has the analogous optimum only on its own fitting window. These geometric statements impose no ordering on prediction loss, either at an individual time or averaged over time.

For an explicit counterexample, take the constant operator $M = \mathrm{diag}(2,1)$, budget $k = 1$, input $h = e_2$, and logits $(u_2, 0)$ where $u = MPh$. The first class is the correct label. The matrix-optimal projector onto $e_1$ has squared residual one and loss $\log 2$. The projector onto $e_2$ has squared residual four, yet preserves the original loss $\log(1+e^{-1})$. Activation weighting and downstream sensitivity determine whether geometric directions matter to the task.

**Proposition 7** (A random-subspace reference, including numerical truncation)**.** *Let $G \succeq 0$ be the path Gram matrix, and let $U \in \mathbb{R}^{q\times s}$ have orthonormal columns specifying a fixed candidate space. Draw a uniform random rank-$k$ projector $P = UQQ^\top U^\top$ within it, where $0 \le k \le s$ and $Q$ is uniform on the Stiefel manifold of $k$ orthonormal vectors in $\mathbb{R}^s$. For $s > 0$,*

$$\begin{aligned}\mathbb{E}J_\mu(P) &= \operatorname{tr} G - \frac{k}{s}\operatorname{tr}(U^\top GU) \\ &= \operatorname{tr}((I - UU^\top)G) + \left(1 - \frac{k}{s}\right)\operatorname{tr}(U^\top GU).\end{aligned} \tag{40}$$

*The best rank-$k$ projector within this candidate space has residual $\operatorname{tr} G - \sum_{i=1}^{k}\lambda_i(U^\top GU)$ and cannot have larger residual than any such random draw. If $s = 0$, necessarily $k = 0$ and the residual is $\operatorname{tr} G$.*

*Proof.* Rotational invariance makes $\mathbb{E}[QQ^\top]$ a scalar multiple of $I_s$. Its trace is $k$, so the scalar is $k/s$. Substitution into $J_\mu(P) = \operatorname{tr} G - \operatorname{tr}(PG)$ gives the expected residual. The variational principle on $U^\top GU$ proves the optimal residual formula. □

When $U$ spans the complete path row space, the outside-space term vanishes. For a numerically truncated span, that term remains and must be reported. The rank threshold, randomization law, and actual per-module budget define the control. A random Gaussian matrix in $\mathbb{R}^{s\times k}$ followed by orthonormalization implements the stated distribution. The proposition compares geometric errors. Our experiments separately compare the prediction losses of random and dominant directions.

### D.6 Teacher forcing and generated answers

A small change in teacher-forced loss does not by itself control greedy generation. For two logits and correct label one, take $z = (\delta, 0)$ and $\widetilde{z} = (-\delta, 0)$ with $\delta > 0$. Their greedy tokens differ, while

$$\ell(\widetilde{z}, 1) - \ell(z, 1) = \log(1 + e^{\delta}) - \log(1 + e^{-\delta}) = \delta,$$

which can be arbitrarily small. Subsequent autoregressive contexts can then diverge.

**Proposition 8** (A sufficient margin condition for identical greedy output)**.** *Fix a prompt, deterministic greedy decoding, and identical stopping rules. At generation step $j$, evaluate both models on the original model's generated prefix. Let $v_j$ be the original winning token, let $\gamma_j = z_{j,v_j} - \max_{v\neq v_j} z_{j,v}$, and suppose $\|\widetilde{z}_j - z_j\|_\infty \le \eta_j$. If $\gamma_j > 2\eta_j$ at every step through the original stopping point, including its end-of-sequence token when used, the two generated token sequences and their stopping points are identical.*

*Proof.* For each alternative $v$, $\widetilde{z}_{j,v_j} - \widetilde{z}_{j,v} \ge \gamma_j - 2\eta_j > 0$. The inequality gives the same first token for both models. Their prefixes then remain identical by induction, so the comparison applies at every subsequent step. The shared stopping rule then terminates both sequences at the same point. □

The required perturbation and margin conditions concern every prefix along the generated trajectory, not teacher-forced reference answers or an average token loss. A deterministic answer extractor assigns the same score to identical token sequences, although different sequences can also give the same final answer. We therefore measure generation accuracy separately from token loss.

**Experimental interpretation.** The projection preserving the initial-space component in (29) identifies exactly what is removed. Removal of the entire complement tests whether the trained network uses directions outside initialization. The spectrum-based projection then tests whether a few dominant shared directions preserve the observed prediction behavior. Matched endpoint–path budgets compare two distinct geometric objectives, and a random-subspace control tests whether the spectral selection is functionally informative at fixed budget. The conditional bounds explain how matrix residuals can propagate to prediction loss. Functional measurements, including generated answers where applicable, determine the size and ordering of the effects in the trained language models.

# E LONG-HORIZON PROTOCOL AND ADDITIONAL DIAGNOSTICS

The language-modeling experiments measure departure from initialization and the concentration of the resulting path. This appendix records their configuration and adds diagnostics that distinguish a compact state representation from continued motion outside it.

## E.1 CONFIGURATION, PROVENANCE, AND AGGREGATION

A comparison across ranks and model families requires consistent training and measurement rules. We describe the shared protocol and how module-level spectra become the group summaries in Table 1. We fixed the 24-run configuration after the pilot and kept its primary outcome unchanged: hidden operator $k_{99}$ under training-time weighting. We use Qwen2.5-1.5B, Pythia-1.4B-deduped, and the raw WikiText-103 data. Each run uses 4096 training and 256 validation blocks of 512 tokens. Validation loss averages the first 64 blocks in 32 batches of two. Response probes cover positions across the full 256-block bank. Each optimizer step accumulates two microbatches of two blocks each. The learning rate is $2 \times 10^{-4}$, with 20 warmup steps followed by cosine decay to step 1600. We use zero weight decay and adapter dropout, and clip the gradient norm at one. The adapter scale is one. Qwen adapts the query and value projections in each layer, for a total of 56 matrices. Pythia adapts its combined query/key/value projections, for a total of 24 matrices. The base weights use bfloat16 precision. We store factors and response coordinates in float32 and perform the analysis in float64.

We save 57 checkpoints, sampled densely early in training and every 40 steps from step 160 to 1600. Time weights are normalized trapezoidal weights on these times; arc spectra integrate the polygonal matrix path exactly. A hidden path below $10^{-8}$ of the same-weight total path norm is classified as zero. This threshold removes only roundoff-level responses in the frozen-$A$ controls. Numerical ranks at a relative $10^{-6}$ threshold are retained as diagnostics, not estimates of exact algebraic rank.

All archived outputs and source files match their manifests. The paired runs have identical initializations and token banks, and every frozen-$A$ factor remains unchanged across checkpoints. Recomputation of the saved spectra reproduces the reported summaries, and an independent QR/SVD and analytic-segment implementation agrees on the effective dimensions up to numerical precision.

The experiments ran in a Python/PyTorch environment with Transformers and PEFT on two H100 80GB devices. The recorded training loops required about 2.6 GPU-card hours, or 1.4 hours of wall-clock time, excluding setup and later analysis. We preserved the complete source, actual configurations, token hashes, logs, raw spectra, factors, responses, and environment records.

Unless explicitly noted otherwise, group values average three within-run medians. The seeds, rather than individual layers or text banks, are the replication units. Fractional dimensions can arise both from even-sized matrix medians and from averaging seeds. At step 1600, hidden $k_{99.9}$ seed medians are $(3, 3, 3)$ for Qwen rank one, $(10, 10, 10.5)$ for Qwen rank four, $(3, 3, 3)$ for Pythia rank one, and $(10, 10, 8)$ for Pythia rank four. Hidden $k_{99}$ is $(2, 2, 2)$, $(3, 3, 3)$, $(2, 1.5, 1.5)$, and $(3, 3, 3)$ respectively. We do not assign confidence intervals to pooled matrices, which are not independent training replications.

The hidden-to-total path norm increases from 0.607–0.681 at step 160 to 0.815–0.864 at step 1600 across the four groups. From step 800 to step 1600, the median $k_{99.9}$ and relative rank-$r$ error decrease in all 12 joint-training runs. Under time, checkpoint, and arc weighting, respectively, hidden $k_{99} > r$ in 197, 210, and 309 of 480 repeated module observations. These counts describe the same finite collection rather than independent training trials. Fixed and evolving token-response banks also have concentrated spectra, without a universal ordering relative to operator spectra (Appendix C.7).

## E.2 WINDOW ANALYSIS AND EVOLVING RESPONSES

The full-horizon summaries show concentration, but do not determine whether one early subspace remains useful. We therefore divide training into windows and compare the later states and movements with directions fitted earlier. We specified the window analysis after inspecting the 1600-step results. The windows are $[0, 400]$, $[400, 800]$, and $[800, 1600]$, and we fit early directions on $[0, 800]$. All boundaries are saved, so unnormalized trapezoidal Grams add exactly. Across 960 module observations, the 480 frozen-$A$ paths vanish and all 480 standard paths pass energy-additivity and

residual-bound checks. For each matrix we fit the top $r$, early $k_{99}$, and early $k_{99.9}$ directions, preserving the full spectra and principal angles. Main-text comparisons use the prespecified nominal budget $r$ within this post-hoc analysis, rather than selecting the most favorable budget.

At budget $r$, mean late-state energy capture is 96.95%, 98.44%, 96.73%, and 98.70% in the order of Table 1. The corresponding squared-velocity energy capture is 71.08%, 70.29%, 84.78%, and 84.95%. The largest early-to-late principal-angle medians average $9.95^\circ$, $79.50^\circ$, $10.25^\circ$, and $80.18^\circ$. Unweighted largest angles in the rank-four groups are sensitive to weak trailing directions. We therefore retain all spectral gaps and angles, and use energy capture as the primary transfer diagnostic. Only three Pythia rank-one matrix observations satisfy $\beta \leq q$, and no observations in the other groups satisfy it. The decrease in the optimized relative tail therefore rarely meets this sufficient condition for concentration within the unchanged early subspace.

Response measurements use 16 full 512-token sequences, with 64 evenly spaced positions per sequence. Alternating sequences form two banks of 512 positions each. We probe the query and value matrices at Qwen layers 0, 14, and 27, and the combined projections at Pythia layers 0, 12, and 23. Comparisons to operators use these same matrices. The hidden-to-total response norm is not an orthogonal energy decomposition after data weighting. At step 1600, the relative input-response difference $\|B_t A_t (H_t - H_0)\|_F / \|B_t A_t H_t\|_F$ is near roundoff in the first layer. Its descriptive medians are 19.6–32.7% in Qwen middle/last layers and 17.1–20.5% in Pythia. For this additional diagnostic only, each model/layer median pools both ranks, all three standard-training seeds, the selected matrices, and both text banks. Similar effective dimensions therefore do not imply identical responses or frozen activations.

Mean validation losses for frozen-$A$ versus joint-factor training are 2.344569 versus 2.165907, 2.262455 versus 2.147249, 2.651581 versus 2.427197, and 2.545308 versus 2.396586. The trainable parameter counts are 50,176 versus 136,192; 200,704 versus 544,768; 147,456 versus 196,608; and 589,824 versus 786,432. These losses measure mean next-token negative log probability rather than generated-answer accuracy. Their absolute values are not directly comparable across tokenizers.

# F Joint-Projection Protocol and Functional Diagnostics

The spectral measurements motivate the interventions, but predictions must be evaluated in the complete network. Here we specify how we fit and apply the projections, compare token losses and generated answers, and check the paired evaluation records.

## F.1 Math training and frozen evaluation choices

The arithmetic task lets us compare average answer loss with agreement on generated solutions. We first describe the training data and the evaluation choices fixed before these comparisons. Eight math runs fine-tune Qwen2.5-1.5B on GSM8K. Ranks one and four each use three joint-training seeds (17, 29, 43) and one frozen-$A$ control (17). The first 90% of the training split supplies 1024 training blocks and the last 10% supplies 64 validation blocks. Blocks contain 512 tokens. Prompt labels are masked, while answer and end-of-sequence labels are scored. Packing retains cross-example context within each split. The checkpoint schedule and adapter sites match the Qwen language-modeling study. The test questions do not enter projector fitting or validation-loss measurement.

Projection preserves $\Delta W_t \Pi_0$ and replaces the hidden component by its projection onto a chosen shared input space, as in (29). We apply these replacements to all adapted matrices in one network. The factors keep their original shapes. The intervention therefore tests which directions suffice to represent the learned updates, without reducing parameter storage or measuring inference speed. Full-path and early-path fits use the trapezoidal Gram over 57 checkpoints and its prefix through 800, respectively. The full-path fit uses future states relative to any earlier evaluation. An endpoint fit uses the final hidden map and budget $r$, with padding if its numerical rank is smaller.

The functional collection includes four-checkpoint path evaluations, math-loss evaluations, random controls, and generation evaluations. We evaluate each original model and its interventions with the same restored scoring code. Language-modeling and math losses are computed over the corresponding validation blocks. Absolute NLLs across these tasks or tokenizers do not measure comparable task difficulty.

### F.2 Checkpoint projections and aggregate losses

We next specify where the interventions are evaluated and how losses are combined. This distinction matters because the loss comparison samples fewer checkpoints than the geometric fit. We evaluate the language-modeling interventions at steps 160, 400, 800, and 1600. Their normalized trapezoidal weights are $(1/12, 2/9, 5/12, 5/18)$. This four-checkpoint loss objective differs from the 57-checkpoint geometric objective used to fit the full-path projector. Table 3 gives the absolute losses underlying the main-text comparison, averaged over the three training seeds in each setting.

Table 3: Time-weighted language-modeling NLL at equal direction budget $k = r$. A positive reduction means that the full-path subspace gives lower loss than the endpoint subspace.

| Model | $r$ | Endpoint | Full path | Reduction |
|---|---|---|---|---|
| Qwen | 1 | 2.180743 | 2.180698 | 0.000045 |
| Qwen | 4 | 2.157512 | 2.156927 | 0.000585 |
| Pythia | 1 | 2.459465 | 2.458617 | 0.000848 |
| Pythia | 4 | 2.424275 | 2.423035 | 0.001240 |

On GSM8K, the joint 99%-energy projection changes mean answer NLL from 0.481506 to 0.481553 at rank one and from 0.478678 to 0.478496 at rank four, differences of $+0.000047$ and $-0.000182$. Removal of all hidden directions raises these means to 0.635646 and 0.608605, respectively, increases of 0.154140 and 0.129927. The block-level distributions below show how these changes vary across inputs.

### F.3 Random directions and residual energy

The endpoint and path comparisons test where the directions are fitted. The random controls instead test whether selecting the dominant directions matters at the same per-module dimension. For each module, let $U$ span the numerical hidden input-factor path. Candidate directions retain singular values exceeding $\max(10^{-6}, \max(\text{shape})\epsilon_{64})$ times the largest singular value. Thus $D = \dim U$ is a numerical diagnostic, not an exact algebraic rank. Random controls draw a Haar-uniform $k$-dimensional subspace within $U$, using seeds 101, 211, 307, 401, and 503, where $k$ matches that module's full-path 99%-energy budget. All five draws use the model trained with seed 17, so they test subspace variation rather than independent training replications. For Gram $G$, the expected retained energy is $(k/D)\,\mathrm{tr}(U^\top G U)$, as proved in Appendix D.

Table 4: Random controls for training seed 17. The candidate-space dimension $D$ depends on the numerical rank threshold. We pool energies across modules before computing the retained fraction. Final error is the median relative total-map error across modules. Random entries average each statistic over five draws.

| Task/model | $r$ | $D$ range | Main energy | Random energy | Main error | Random error |
|---|---|---|---|---|---|---|
| Wiki/Pythia | 1 | 48–50 | 99.50% | 2.54% | 2.00% | 85.70% |
| Wiki/Pythia | 4 | 182–201 | 99.37% | 1.31% | 4.99% | 83.58% |
| Math/Qwen | 1 | 46–49 | 99.52% | 2.87% | 1.48% | 84.79% |
| Math/Qwen | 4 | 159–193 | 99.46% | 1.60% | 4.36% | 83.68% |
| Wiki/Qwen | 1 | 47–50 | 99.66% | 3.26% | 2.01% | 86.08% |
| Wiki/Qwen | 4 | 167–195 | 99.35% | 1.55% | 4.04% | 85.92% |

Table 4 pools matrix energies before taking their ratio, whereas the final-map error first takes a median across modules. These aggregations measure different quantities: a median relative error need not equal the energy ratio across all modules. Random controls match the count of directions and preserve the same initialization component, but discard substantially more path energy. They test direction selection at a fixed dimension, but do not compare alternative projectors that retain the same energy.

All 30 random-subspace draws have higher NLL than the principal-direction intervention. Mean gaps across the six model/task/rank settings range from 0.125 to 0.339 nats per token. Random directions retain 1.31–3.26% of pooled path energy, compared with 99.35–99.66% for principal directions.

### F.4 PAIRED LOSS DISTRIBUTIONS

Aggregate loss can hide unequal changes across text blocks. We therefore compare each projected score with the original score on the same block, using the saved token counts to keep each pair aligned. For each block, we divide the saved negative-log-probability sum by its number of scored tokens and subtract the paired original score. Sequence identities and token counts match exactly. The distribution is unweighted over the 64 blocks, while the aggregate NLL weights all scored tokens equally. The quantiles in Table 5 are post-hoc descriptive statistics, not independent-example confidence intervals or a test of equivalence.

Table 5: Paired loss differences after 99%-energy projection on the math validation bank, with 64 blocks per run. Values are nats per supervised token. Quantiles summarize variation across blocks rather than confidence in a population mean.

| $r$ | Seed | 5th percentile | Median | 95th percentile | Maximum absolute |
|---|---|---|---|---|---|
| 1 | 17 | -0.00295 | -0.00004 | +0.00279 | 0.00444 |
| 1 | 29 | -0.00321 | +0.00006 | +0.00295 | 0.00508 |
| 1 | 43 | -0.00249 | +0.00020 | +0.00297 | 0.00428 |
| 4 | 17 | -0.00391 | -0.00048 | +0.00360 | 0.00530 |
| 4 | 29 | -0.00434 | -0.00052 | +0.00325 | 0.00782 |
| 4 | 43 | -0.00306 | -0.00007 | +0.00292 | 0.00598 |

### F.5 GENERATION, PARSING, AND UNCERTAINTY

The paired losses use reference contexts, whereas generated solutions depend on the model's own earlier tokens. We evaluate complete solutions on matched test questions and record both their accuracy and their agreement. Generation uses all 1319 official test questions in their recorded order, batch size 16, left padding, greedy decoding (`do_sample=False`, one beam), and at most 512 new tokens. Each of the six math models has four variants: original, full-path 99%, full-path 99.9%, and removal of all hidden directions. This produces 31,656 answers. The strict parser reads the numerical answer after the last `####` and counts an unparseable answer as incorrect. A prespecified secondary parser takes the final number in the output. Token comparisons retain the first end-of-sequence token and exclude subsequent padding. The tokenizer call explicitly omits token-type IDs, which the model does not accept. The interface repair leaves the prompts and answer parsers unchanged.

Table 6: Generation results for each training seed after 99%-energy projection. All rows use the same 1319 test questions. The counts reveal compensating correctness changes that aggregate accuracy conceals.

| $r$ | Seed | Original (%) | Projected (%) | Correct→wrong | Wrong→correct | Answer changes |
|---|---|---|---|---|---|---|
| 1 | 17 | 57.5436 | 57.9985 | 40 | 46 | 223 |
| 1 | 29 | 58.7566 | 60.0455 | 32 | 49 | 206 |
| 1 | 43 | 58.1501 | 59.8939 | 38 | 61 | 228 |
| 4 | 17 | 57.2403 | 57.2403 | 53 | 53 | 242 |
| 4 | 29 | 58.6050 | 57.2403 | 58 | 40 | 236 |
| 4 | 43 | 58.3776 | 59.2873 | 39 | 51 | 232 |

At 99% retention, 40.38% and 44.43% of complete generated token sequences change at ranks one and four, respectively. At rank four and seed 17, both versions have 57.24% accuracy, despite 53 correct-to-wrong and 53 wrong-to-correct changes. More retained matrix energy does not monotonically improve generated-answer accuracy: the rank-four mean is 57.92% at 99% retention and 57.49% at 99.9% retention.

The three seeds reuse the same questions, so the 3957 model–question records per rank are not 3957 independent questions. An exploratory crossed bootstrap independently resamples seed indices and question indices with replacement, keeping original and projected responses paired. Across 10,000 replicates, percentile 95% intervals for the mean accuracy change are $[-0.076, 2.350]$ percentage points at rank one and $[-1.693, 1.365]$ at rank four. Three seeds provide limited replication, and no equivalence or noninferiority margin was set.

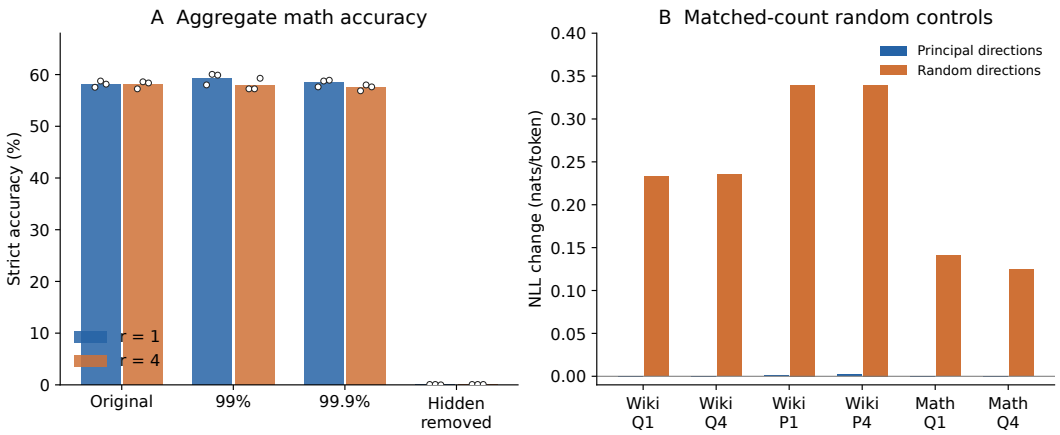



Figure 2: Generation accuracy and random-direction controls. The accuracy panel shows the three training seeds. Random-control bars average five subspace draws for one training seed, with the same dimension but different retained energy from the principal-direction intervention.

After removal of hidden directions, mean accuracy falls to 0.0505% and 0.0758% under strict parsing, or 1.8954% and 2.3755% under the secondary parser, at ranks one and four respectively. The fraction reaching the token limit rises from 0.45%/0.40% to 98.13%/97.70%. This intervention perturbs arithmetic, formatting, and stopping behavior together. The resulting score therefore does not isolate arithmetic competence. The generation study did not include random-subspace controls.

### F.6 Integrity checks and remaining execution baseline

The paired comparisons must start from the same trained checkpoints and use matched inputs and scoring rules. We verify these records and distinguish the completed checks from the repeated-generation baseline that remains necessary. The returned files, source snapshot, and job-completion records pass their integrity checks. Independent reconstruction of the losses and rescoring of the generation batches reproduce the reported summaries. Question order, prompt token IDs, batch signatures, and reference answers match across paired variants. Factor files used for math loss and generation have matching hashes. We found no exact question–answer row overlap with the actual fine-tuning data. This check does not determine whether the base model encountered test questions during pretraining.

Identity restoration and final endpoint reconstruction agree in the matched evaluator. Differences from older stored losses remain within the frozen restoration tolerance. All reported comparisons use the same evaluator version.

The actual-model generation collection has no independently repeated original run or identity-intervention generation run. Repeated tiny-model preflight checks do not establish execution reproducibility for these real models. The reported question-level differences therefore describe the returned paired executions. To measure the contribution of execution variability, a same-model repeat must keep the weights, token IDs, batches, precision, backend, and stopping rules fixed. This missing generation check does not affect the separate geometric and teacher-forced checks.

The functional jobs ran on two H100 80GB devices. They required about 3.5 hours of wall-clock time and 5.9 hours of summed task-slot time, most of it for generation. These are measured evaluation costs, not speedups or precise accelerator utilization.

## G Controlled Fixed-Feature Protocols

The language-model experiments measure realistic training paths, while the rank-one theorem concerns fixed-feature regression. These controlled experiments test the theorem in its stated setting and isolate the role of the active feature spectrum.

### G.1 TRAJECTORY AND MATCHED-SPECTRUM EXPERIMENTS

We first construct trajectories from the aligned-target model, then vary the feature spectrum while keeping the initialization and target relationship fixed. This comparison tests whether the number of active eigenvalues controls the exact shared dimension. For $d \in \{3, 5, 9\}$, set

$$v = d^{-1/2}(1, \ldots, 1)^\top, \quad A_0 = v^\top, \quad X = \sqrt{d}\ \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}), \quad Y = Xv,$$

where the eigenvalues range geometrically from $0.1$ to $4$. These targets are exactly attainable with frozen $A$. Both factors use learning rate $0.08$ for 2,000 steps. A fixed grid combines the earliest steps with approximately 100 logarithmically spaced integer checkpoints, removing duplicates. For $\ell$ saved checkpoints, the stack contains $\ell^{-1/2} M_t$, which weights checkpoints equally rather than by elapsed training time. Its singular values give the single-probe worst-case frontier on the unit hidden sphere.

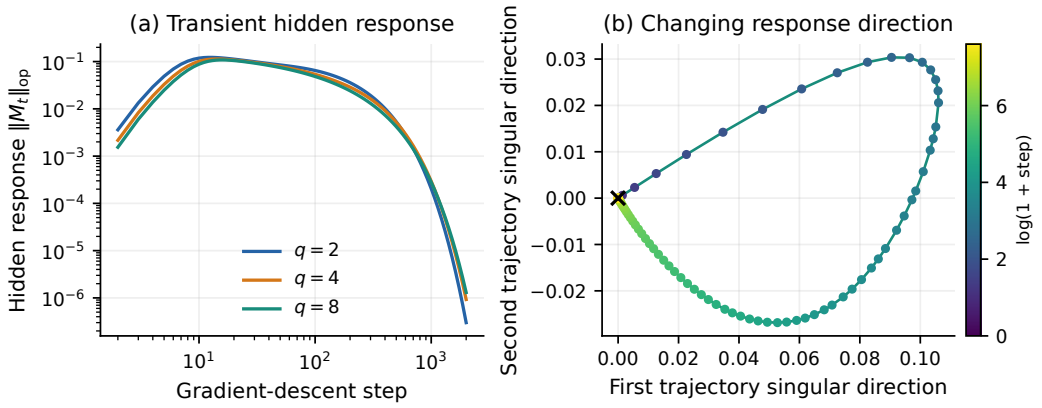



Figure 3: Rank-one training develops transient hidden responses even when the initial space contains a solution. (a) The hidden operator norm changes during training. The log plot omits steps 0 and 1, whose responses are exactly zero. (b) The dimension-nine response appears in its first two trajectory singular directions, with color indicating the training step. This projection serves only to display the trajectory. The response changes direction before returning close to zero at interpolation.

The matched-spectrum control adds $s = 1, 2, 3, d$ active eigenvalues in each dimension, with duplicate choices removed. We assign repeated eigenvalues to consecutive coordinate groups and normalize their sum to the reference second-moment trace. A left orthogonal transformation maps each initial training projection onto the same reference vector. This transformation preserves the eigenvalues of the feature second moment and the adapter initialization. Thus every initial kernel block, target, and probe population matches within each dimension. We run all 11 problems with both joint and frozen factors, for 22 additional fits. All observed trajectory ranks equal $s - 1$ at tolerance $10^{-10}$ times the target scale. The largest initial training-kernel discrepancy is $5.33 \times 10^{-15}$, and isotropic hidden responses remain below $5.09 \times 10^{-16}$.

Rational-arithmetic checks in hidden dimensions two and three verify exact rank independently of floating-point thresholds. Approximate query requirements instead use the actual singular values.

Table 7: Second-moment controls with matched initial kernels. Each row reports a joint-factor run paired with a frozen-factor control. The predicted trajectory rank is $s - 1$, and the observed rank uses absolute tolerance $10^{-10}$ for the unit target scale.

| $d$ | Active eigenvalues $s$ | Predicted rank | Observed rank | Peak $\|M_t\|$ |
|---|---|---|---|---|
| 3 | 1 | 0 | 0 | $4.48 \times 10^{-16}$ |
| 3 | 2 | 1 | 1 | $1.42 \times 10^{-1}$ |
| 3 | 3 | 2 | 2 | $1.23 \times 10^{-1}$ |
| 5 | 1 | 0 | 0 | $3.52 \times 10^{-16}$ |
| 5 | 2 | 1 | 1 | $1.40 \times 10^{-1}$ |
| 5 | 3 | 2 | 2 | $1.22 \times 10^{-1}$ |
| 5 | 5 | 4 | 4 | $1.14 \times 10^{-1}$ |
| 9 | 1 | 0 | 0 | $5.09 \times 10^{-16}$ |
| 9 | 2 | 1 | 1 | $1.37 \times 10^{-1}$ |
| 9 | 3 | 2 | 2 | $1.24 \times 10^{-1}$ |
| 9 | 9 | 8 | 8 | $1.08 \times 10^{-1}$ |