CHARACTERIZING LOCAL LOSS GAPS IN LOW-RANK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) makes fine-tuning efficient, but a suitable rank is not known in advance. Informed rank adjustment requires diagnosing whether the current rank limits attainable improvement and how much additional improvement a specific increase could unlock. Performance gaps alone cannot answer these questions, because optimization and update scale confound the comparison. We propose a local diagnostic framework that compares the best dense and low-rank updates to the same module within a fixed Frobenius ball. A local quadratic approximation makes this comparison tractable. With separable curvature and feasible unconstrained optima, its gap equals the spectral tail of a curvature-whitened gradient. We control the discrepancy from the actual loss and correct for optima outside the radius. Under explicit approximation-error bounds, the resulting conditional intervals distinguish rank limitation, local sufficiency, and inconclusive cases relative to a prescribed tolerance. The reduction in the gap to the same dense comparator quantifies the optimal benefit of increasing rank; newly captured spectral energy predicts this benefit in the interior regime. Controlled objectives evaluate interval coverage and width, while RoBERTa/SST-2 experiments assess fixed-rank prediction and comparison sensitivity. Newly captured energy ranks observed upgrade gains with Spearman correlations of 0.903 on SST-2 and 0.734 on MRPC, compared with 0.333 and 0.414 for the current tail. The framework informs local rank adjustment while distinguishing optimal capacity gains from finite-budget training outcomes.

1 INTRODUCTION

Low-rank adaptation (LoRA) reduces fine-tuning costs by restricting weight updates to a chosen rank (Hu et al., 2022). But *how much rank is enough?* Methods such as AdaLoRA allocate rank across modules during training (Zhang et al., 2023). We complement these methods by asking whether rank limits attainable improvement and how much additional improvement a specific rank increase could unlock. A worse training loss does not by itself identify rank as the bottleneck: optimization, initialization, and update scale also affect the outcome (Hayou et al., 2024; Wang et al., 2024). To isolate rank, we compare the best dense and rank-r updates with the same module, loss, expansion point, and Frobenius-norm budget. Their best attainable losses differ by the *local rank gap*. A small gap means that rank leaves little improvement out of reach within this region (Figure 1).

Directly computing these optimal losses for a neural network is generally intractable, motivating a local quadratic approximation. Even then, general curvature makes rank-constrained approximation difficult. Separable, positive-definite curvature permits a change of coordinates that preserves rank and reduces the objective to ordinary low-rank approximation. When the unconstrained optima satisfy the radius bound, the Eckart–Young–Mirsky theorem gives the gap as half the discarded squared singular-value sum (Eckart & Young, 1936; Mirsky, 1960). We call this quantity the *spectral tail*.

This identity alone does not determine the actual-loss gap. The quadratic proxy can differ from the loss, and its unconstrained optima can exceed the prescribed radius. Even an exact generalized Gauss–Newton (GGN) block may not be separable (Eschenhagen et al., 2023). We account for proxy error through uniform bounds and for boundary effects through a dense convex problem and a feasible scaled low-rank update. Relative to a chosen tolerance, the resulting interval identifies rank limitation or sufficiency, or leaves the question open.

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.

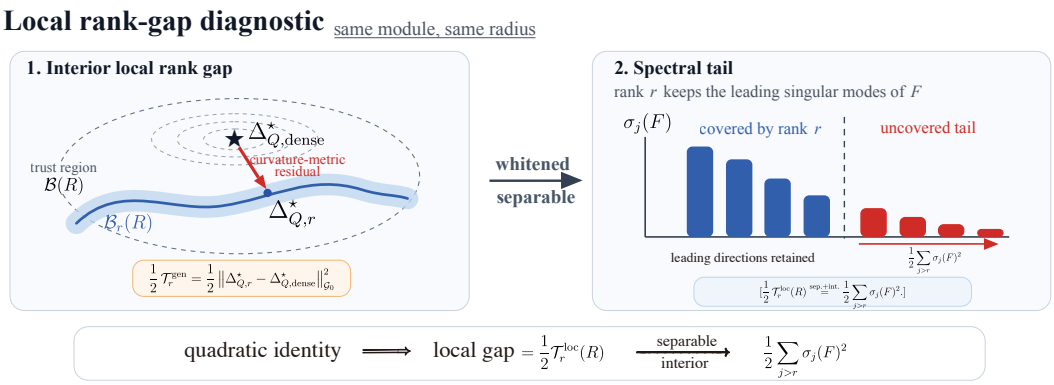

Figure 1: Comparing dense and rank-r updates to the same module within the same radius. When the unconstrained quadratic optima satisfy the radius bound, discarded singular values of the whitened gradient determine the gap. Section 5 accounts for optima on the boundary.

Rank limitation raises a second question: *how much of the gap can a particular rank increase close?* Similar tails beyond rank r can contain different amounts of energy in directions $r+1$ through r'. Only this newly captured energy determines the quadratic gain from r to r' when both rank-constrained optima satisfy the radius bound. Approximation and boundary corrections extend it to an interval for the optimal actual-loss reduction.

Our contributions are threefold: a local rank gap with actual-loss error bounds; computable boundary corrections and rank-upgrade intervals; and separate evaluations of interval validity and predictive usefulness. Controlled objectives test interval coverage against reference values. RoBERTa experiments examine fixed-rank prediction and protocol sensitivity; SST-2 and MRPC upgrades test newly captured energy against current-tail and gradient-only baselines. These training outcomes remain distinct from the best attainable losses addressed by the bounds.

2 Setup and Local Comparison

2.1 Adapting One Module

To isolate rank, we hold the loss and all other model parameters fixed. We study an update $\Delta \in \mathbb{R}^{d_1 \times d_2}$ to one module whose pretrained weights remain frozen. Fix the loss L, its normalization, and the expansion point $\Delta = 0$. The loss may be the population risk $L(\Delta) = \mathbb{E}_{Z\sim\mathcal{D}}\ell(Z;\Delta)$ or an average over a fixed dataset. All comparisons and conclusions refer to this chosen loss. Write $g_0 = \nabla L(0)$ and let $\mathcal{G}_0$ be its GGN operator, using a dataset average for a fixed dataset. For population risk, assume differentiation and expectation can be interchanged at the expansion point, giving $g_0 = \mathbb{E}\nabla\ell(Z;0)$. This holds automatically for finite averages. Using the Frobenius inner product on matrix updates, the local GGN model is

$$Q_0(\Delta) = L(0) + \langle g_0, \Delta\rangle + \tfrac{1}{2}\langle \Delta, \mathcal{G}_0[\Delta]\rangle . \tag{1}$$

The GGN is positive semidefinite and approximates the curvature of the loss; it can differ from the actual Hessian.

2.2 Comparing Dense and Low-Rank Updates

A common radius prevents either competitor from gaining an advantage simply by taking a larger update. For $R > 0$ and $r \in \{0, \dots, \min(d_1, d_2)\}$, define

$$\mathcal{B}(R) = \{\Delta : \|\Delta\|_F \le R\}, \qquad \mathcal{B}_r(R) = \{\Delta : \operatorname{rank}(\Delta) \le r,\ \|\Delta\|_F \le R\}.$$

Let $m_{L,d}(R)$ and $m_{L,r}(R)$ denote the minimum values of L over $\mathcal{B}(R)$ and $\mathcal{B}_r(R)$, respectively. The local rank gap is

$$G_r(R) = m_{L,r}(R) - m_{L,d}(R) \ge 0. \tag{2}$$

Continuity of L ensures that both minima exist on these compact sets. The subscript d denotes dense, and rank zero permits only the zero update. Both updates leave other modules fixed and may use any norm up to the radius. Rank constrains the update matrix, regardless of how LoRA factors are scaled or trained.

2.3 Curvature at the pretrained weights

The spectral reduction requires separable curvature, which the exact GGN need not provide for shared weights. Making its structure explicit separates curvature estimation from the additional approximation used by the proxy. Let a block Z have supervised output positions $t = 1, \dots, T(Z)$ and loss

$$\ell(Z; \Delta) = \sum_t w_t(Z)\ \mathrm{CE}(u_t(\Delta), y_t), \qquad w_t(Z) \geq 0.$$

Expectations average over the target block distribution, or uniformly over the fixed dataset. Weights encode the mask and loss normalization, independently of Δ, and apply to all gradients, curvature estimates, and losses being compared. A classification block may have many input positions but only one supervised output.

At $\Delta = 0$, define $J_t[U] := Du_t(0)[U]$ and $H_t := \mathrm{diag}(p_t) - p_t p_t^\top$, where p_t is the softmax probability vector. The adjoint J_t^* is defined using the Euclidean and Frobenius inner products.

The exact GGN has the second-moment representation

$$\mathcal{G}_0[U] = \mathbb{E}\left[\sum_{t,j} \langle U, \Phi_{t,j} \rangle\, \Phi_{t,j}\right]. \tag{3}$$

Here $\Phi_{t,j}$ are the weighted output-curvature directions pulled back through J_t^*; Proposition B.2.1 in Appendix B.2 gives their construction and the full derivation. For shared modules, each output can aggregate multiple input positions, so these matrices need not have rank one. The exact GGN can also contain cross-position terms and need not reduce to $S[\Delta] = B\Delta A$. Separability is therefore an additional modeling approximation (Eschenhagen et al., 2023), whose discrepancy enters the error budget in Section 4.

Even before introducing a separable proxy, finite-sample estimation creates a distinct source of error. Given i.i.d. blocks, the direct estimators are $\widehat{g}_0 = N^{-1} \sum_i \nabla \ell(Z_i; 0)$ and $\widehat{\mathcal{G}}_0 = N^{-1} \sum_i \mathcal{G}_0(Z_i)$. Appendix B.2 proves these identities and bounds curvature estimation error. Appendix B.3 bounds errors in the empirical quadratic approximation, including ridge terms, for finitely many samples.

3 How Rank Limits Improvement in a Quadratic Model

We first isolate the geometric cost of rank, setting aside discrepancies between the quadratic model and the actual loss. For a self-adjoint operator H, write $q_{g,H}(\Delta) = \langle g, \Delta \rangle + \frac{1}{2} \langle \Delta, H[\Delta] \rangle$ and define

$$\mathcal{T}_r^{\mathrm{loc}}(R; g, H) = 2 \left(\min_{\Delta \in \mathcal{B}_r(R)} q_{g,H}(\Delta) - \min_{\Delta \in \mathcal{B}(R)} q_{g,H}(\Delta) \right). \tag{4}$$

The factor of two converts the gap into a squared approximation error after completing the square. The definition also allows singular H. For the geometric interpretation, assume $H \succ 0$, put $D_H = -H^{-1} g$, and write $\|U\|_H^2 = \langle U, H[U] \rangle$. The choice $(g, H) = (g_0, \mathcal{G}_0)$ gives the GGN model when it is positive definite. Section 5 uses an estimate with regularization for computation.

Theorem 3.1 (The gap as an approximation error). *For $H \succ 0$,*

$$q_{g,H}(\Delta) = q_{g,H}(D_H) + \tfrac{1}{2}\|\Delta - D_H\|_H^2 ,$$

and consequently

$$\mathcal{T}_r^{\mathrm{loc}}(R; g, H) = \min_{\mathcal{B}_r(R)} \|\Delta - D_H\|_H^2 - \min_{\mathcal{B}(R)} \|\Delta - D_H\|_H^2 . \tag{5}$$

If D_H and an unconstrained best rank-r quadratic update both belong to $\mathcal{B}(R)$, the second minimum is zero and the first equals $\mathcal{T}_r^{\mathrm{gen}}(g, H) := \min_{\mathrm{rank}(\Delta)\le r} \|\Delta - D_H\|_H^2$.

Subtracting the dense distance removes the error already imposed by the radius, isolating the additional restriction due to rank. General curvature need not admit a closed form; separable curvature yields a spectral formula.

Corollary 3.2 (The gap from singular values). *Assume $H[\Delta] = B\Delta A$ with $A, B \succ 0$. Set $F = B^{-1/2} g A^{-1/2}$, $M = B^{1/2}\Delta A^{1/2}$, and $\mathcal{E}_R = \{M : \left\|B^{-1/2} M A^{-1/2}\right\|_F \le R\}$. Then* $\mathrm{rank}(M) = \mathrm{rank}(\Delta)$ *and*

$$q_{g,H}(\Delta) = \tfrac{1}{2}\|M + F\|_F^2 - \tfrac{1}{2}\|F\|_F^2 , \tag{6}$$

so that

$$\mathcal{T}_r^{\mathrm{loc}}(R; g, H) = \min_{\substack{M \in \mathcal{E}_R \\ \mathrm{rank}(M)\le r}} \|M + F\|_F^2 - \min_{M \in \mathcal{E}_R} \|M + F\|_F^2 . \tag{7}$$

If $-F$ and some rank-r truncated SVD $-(F)_r$ are in $\mathcal{E}_R$, then

$$\tfrac{1}{2}\mathcal{T}_r^{\mathrm{loc}}(R; g, H) = \tfrac{1}{2}\sum_{j>r} \sigma_j(F)^2 . \tag{8}$$

These coordinates preserve rank and turn quadratic distance into Frobenius distance. The whitened gradient incorporates local curvature geometry into the gradient coordinates, exposing the matrix directions that govern quadratic improvement. Classical low-rank approximation then gives the formula (Eckart & Young, 1936; Mirsky, 1960). We call feasibility of both unconstrained optima the interior regime, including equality at the radius. Otherwise, the ellipsoid may change either optimum. Nonseparable curvature generally requires more than an SVD (Srebro & Jaakkola, 2003). Appendix B.4 gives proofs and a boundary counterexample.

4 Relating the Quadratic Estimate to the Actual Loss

The quadratic identities do not yet control the actual-loss gap. To bound the discrepancy, let $\widehat{g}_0$ and $\widehat{\mathcal{S}}$ be an estimated gradient and self-adjoint curvature operator, and define

$$\widehat{q}(\Delta) := \langle \widehat{g}_0, \Delta\rangle + \tfrac{1}{2}\left\langle \Delta, \widehat{\mathcal{S}}[\Delta]\right\rangle , \quad v_d(R) := \min_{\mathcal{B}(R)} \widehat{q}, \quad v_r(R) := \min_{\mathcal{B}_r(R)} \widehat{q}. \tag{9}$$

Write $\widehat{G}_r(R) := v_r(R) - v_d(R)$. Compactness ensures both minima exist. The loss, module, center, and radius match those in $G_r(R)$.

Theorem 4.1 (Error in the estimated loss gap). *Let L be twice continuously differentiable on a neighborhood of $\mathcal{B}(R)$. For nonnegative budgets $\varepsilon_g, \varepsilon_H, \rho$, suppose that, deterministically or on one common event, the Hessian variation satisfies $\|\nabla^2 L(\Delta) - \nabla^2 L(0)\|_{\mathrm{op}} \le \rho\|\Delta\|_F$ for every $\Delta \in \mathcal{B}(R)$, and the empirical proxy obeys*

$$\|\widehat{g}_0 - g_0\|_F \le \varepsilon_g, \qquad \|\widehat{\mathcal{S}} - \nabla^2 L(0)\|_{\mathrm{op}} \le \varepsilon_H . \tag{10}$$

Then, on the same event,

$$\left|G_r(R) - \widehat{G}_r(R)\right| \le E(R) := 2R\varepsilon_g + R^2\varepsilon_H + \tfrac{\rho}{3}R^3 . \tag{11}$$

Taylor's theorem bounds $|L(\Delta) - L(0) - \widehat{q}(\Delta)|$ uniformly by $E(R)/2$. Each constrained minimum changes by at most this amount; subtracting them proves the result. This argument requires neither convexity of L nor positive definiteness or separability of the proxy. It also covers boundary optima, separating validity of the transfer from efficient computation of the quadratic gap. Appendix B.5 gives the full proof.

Decomposing the curvature error. To connect the transfer bound to the approximations used in practice, we separate Hessian mismatch, sampling error, and proxy error. A sufficient budget is

$$\varepsilon_H = \eta_{\mathrm{GGN}} + \Gamma_G + \widehat{\eta}_{\mathrm{proxy}}, \tag{12}$$

where the terms bound, respectively, $\left\|\nabla^2 L(0) - \mathcal{G}_0\right\|_{\mathrm{op}}$, $\left\|\mathcal{G}_0 - \widehat{\mathcal{G}}_0\right\|_{\mathrm{op}}$, and $\left\|\widehat{\mathcal{G}}_0 - \widehat{\mathcal{S}}\right\|_{\mathrm{op}}$. They account for the GGN approximation, sampling, and the proxy, including ridge regularization in $\widehat{\mathcal{S}}$. Uniform bounds require more than checks along sampled directions. Appendix B.3 gives sufficient bounds on a common event covering all ranks for a fixed proxy and radius. Data-dependent choices of proxies, radii, modules, or centers require bounds across those choices too.

5 Computing the Gap and the Benefit of Higher Rank

The remaining obstacle is the boundary: dense and low-rank optima may incur different penalties when constrained by the radius. To correct the spectral estimate, consider a separable empirical proxy:

$$\widehat{\mathcal{S}}[\Delta] = \widetilde{B}\Delta\widetilde{A}, \qquad \widetilde{A} \succ 0, \quad \widetilde{B} \succ 0, \qquad \widehat{F} := \widetilde{B}^{-1/2}\widehat{g}_0\widetilde{A}^{-1/2}. \tag{13}$$

These curvature matrices may include ridge regularization and are distinct from LoRA factors. Completing the square gives the unconstrained dense minimizer $D_d = -\widetilde{B}^{-1}\widehat{g}_0\widetilde{A}^{-1}$ with minimum value $v_d^\infty = -\frac{1}{2}\|\widehat{F}\|_F^2$. Truncating an SVD of $\widehat{F}$ yields an unconstrained rank-r minimizer

$$D_r = -\widetilde{B}^{-1/2}(\widehat{F})_r\widetilde{A}^{-1/2}. \tag{14}$$

Its minimum value v_r^∞ differs from the dense optimum by the spectral tail

$$s_r := v_r^\infty - v_d^\infty = \tfrac{1}{2}\sum_{j>r}\sigma_j(\widehat{F})^2. \tag{15}$$

When D_d, D_r are both feasible, this recovers Corollary 3.2: $\widehat{G}_r(R) = s_r$. With tied singular values, any feasible optimal truncation suffices.

Outside this regime, the two minima can increase by different amounts. We track their boundary penalties separately by setting

$$d_d(R) := v_d(R) - v_d^\infty, \qquad \alpha_r := \min\{1, R/\|D_r\|_F\}, \qquad d_r(R) := \tfrac{1}{2}(1-\alpha_r)^2\left\|(\widehat{F})_r\right\|_F^2, \tag{16}$$

with $\alpha_r = 1$ when $D_r = 0$. Here d_d is the exact increase in the dense minimum. For the rank-constrained minimum, d_r bounds the increase using the scaled, feasible update $\alpha_r D_r$.

Theorem 5.1 (Bounds when the radius constrains the update)**.** *Under Eq. (13), for every $R > 0$,*

$$\max\{0, s_r - d_d(R)\} \le \widehat{G}_r(R) \le s_r + d_r(R) - d_d(R). \tag{17}$$

If Theorem 4.1 also applies, then

$$\max\{0, s_r - d_d(R) - E(R)\} \le G_r(R) \le s_r + d_r(R) - d_d(R) + E(R). \tag{18}$$

The proof uses $v_r^\infty \le v_r(R) \le \widehat{q}(\alpha_r D_r) = v_r^\infty + d_r(R)$ and subtracts $v_d(R)$ (Appendix B.6). Scaling gives an upper bound from a feasible update, which need not be optimal. When D_d, D_r are feasible, both corrections vanish and Eq. (18) becomes the spectral estimate with error bounds. Its width is controlled by approximation error and the low-rank boundary bound $d_r(R)$; the exact dense correction shifts both endpoints before the lower bound is clipped at zero. For tolerance $\epsilon > 0$, Eq. (18) certifies rank limitation if its lower endpoint exceeds ϵ, local sufficiency if its upper endpoint is at most ϵ, and an inconclusive diagnosis otherwise (Appendix D.1 gives the sufficient-rank rule).

Computing the dense value. Given the approximation-error budget, computing $d_d(R)$ completes the quantities needed for the interval. This correction is tractable through a convex trust-region problem; no globally optimal rank-constrained update is needed. Diagonalize $\widetilde{A} = V\,\mathrm{diag}(a)V^\top$ and $\widetilde{B} = U\,\mathrm{diag}(b)U^\top$, and set $C = U^\top\widehat{g}_0 V$. The dense solution has coordinates $X_{ij} = -C_{ij}/(b_i a_j +$

λ). Use $\lambda = 0$ if feasible; otherwise solve the monotone equation $\sum_{ij} C_{ij}^2/(b_i a_j + \lambda)^2 = R^2$ for $\lambda > 0$. A feasible point and a dual value provide numerical bounds $\ell_d \le v_d(R) \le u_d$. The resulting interval for the actual loss gap is

$$\max\{0, v_r^\infty - u_d - E(R)\} \le G_r(R) \le \widehat{q}(\alpha_r D_r) - \ell_d + E(R). \tag{19}$$

Appendix B.7 gives the dual and its proof. Numerical errors in singular values or objective values must also be bounded.

From rank limitation to a specific upgrade. A gap relative to the dense optimum does not specify how much improvement a particular higher rank can recover. The benefit of increasing rank is $G_r(R) - G_{r'}(R) = m_{L,r}(R) - m_{L,r'}(R)$. The dense optimum cancels; the gain is nonnegative because higher rank retains all lower-rank updates.

Corollary 5.2 (Benefit of increasing rank)**.** *For $0 \le r < r' \le \min(d_1, d_2)$, keep the loss, module, center, radius, and proxy fixed, and define*

$$B_{r\to r'}(R) := m_{L,r}(R) - m_{L,r'}(R), \qquad t_{r,r'} := s_r - s_{r'} = \tfrac{1}{2} \sum_{j=r+1}^{r'} \sigma_j(\widehat{F})^2.$$

Under the assumptions of Theorem 5.1 and Theorem 4.1,

$$\max\{0, t_{r,r'} - d_{r'}(R) - E(R)\} \le B_{r\to r'}(R) \le t_{r,r'} + d_r(R) + E(R). \tag{20}$$

If $D_r, D_{r'} \in \mathcal{B}(R)$, this simplifies to $|B_{r\to r'}(R) - t_{r,r'}| \le E(R)$; feasibility of D_d is not needed.

The quantity $t_{r,r'}$ includes only the spectral energy newly admitted by the proposed upgrade. The current tail s_r also includes directions beyond r', which this upgrade cannot capture. This distinction motivates the upgrade predictor evaluated in Section 6.

Comparing the two rank-constrained minima directly gives error $E(R)$; subtracting separate gap intervals would give $2E(R)$ (Appendix B.8). This result concerns the best achievable loss at each rank. For feasible trained updates with optimization errors $e_k := L(\widetilde{\Delta}_k) - m_{L,k}(R) \ge 0$, the observed difference is $B_{r\to r'}(R) + e_r - e_{r'}$. Training gains therefore depend on how closely each run approaches its optimum.

6 Experiments

We test whether the spectrum predicts the loss gap at a given rank, first on quadratic objectives with analytic references and then on RoBERTa/SST-2. We next predict the benefits of increasing rank on SST-2 and MRPC. These prediction comparisons use continuous scores without a decision tolerance ϵ. Appendices F.4.1 and F.4.2 examine how the comparison protocol and rank affect the observations.

6.1 Controlled fixed-rank prediction

We hold rank fixed at $r \in \{4, 8\}$ to isolate the spectrum's contribution to prediction on controlled quadratic objectives. Each rank has 16 calibration cases and 16 evaluation cases. Each evaluation reference is summarized by its mean and standard deviation over ten repeats. We compare a constant based only on rank, gradient norm, and three versions of the gradient's spectral tail: unwhitened, whitened using diagonal factors, and whitened using complete factors. Whitening accounts for curvature using either the diagonals or the complete moment factors, called diagonal-factor and full-factor whitening in the tables. Both use the factored approximation to the GGN operator.

Table 1 reports Spearman correlation and mean absolute error (MAE) against the evaluation reference means. For the whitened predictors, we report errors before and after calibration. The constant prediction has undefined Spearman correlation within each rank; its calibrated MAE provides a baseline.

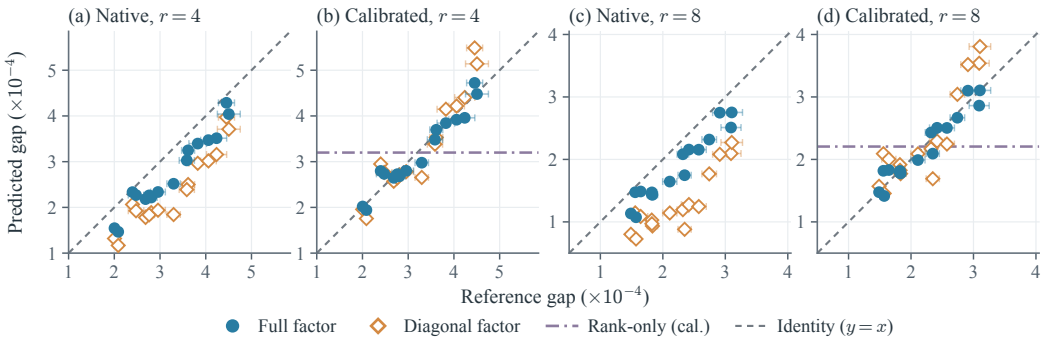

Figure 2: Native and calibrated predictions at fixed rank. Panels (a,b) show $r = 4$ and panels (c,d) show $r = 8$, with 16 evaluation cases per panel. Circles and diamonds denote full-factor and diagonal-factor predictions. Horizontal error bars show the reference mean $\pm$ one standard deviation over ten repeats, not a confidence interval. The gray dashed line is identity; the purple dash-dot line is the calibrated rank-only constant. Axis limits are shared within each rank pair. All gaps are in 10^{-4} loss units.

Table 1: Fixed-rank prediction on 16 evaluation cases per rank, with 16 calibration cases per rank. Spearman ρ measures ordering (↑); native and calibrated MAEs are in 10^{-4} loss units (↓). Best values within each rank are bold. A dash denotes an undefined correlation or a predictor defined only after calibration.

	$r = 4$			$r = 8$		
Predictor	ρ ↑	Native ↓	Calibrated ↓	ρ ↑	Native ↓	Calibrated ↓
Rank-only constant	—	—	0.7126	—	—	0.4785
Gradient norm	0.3559	—	0.6457	0.2647	—	0.4430
Unwhitened gradient tail	0.3118	—	0.6473	0.3500	—	0.4471
Diagonal-factor whitened tail	0.8912	0.8922	0.3102	0.8147	0.9196	0.2868
Full-factor whitened tail	**0.9441**	**0.4770**	**0.1507**	**0.9206**	**0.3580**	**0.1162**

Results. Whitening with complete factors best predicts the ordering at both ranks: Spearman correlations are 0.9441 and 0.9206 for $r = 4$ and $r = 8$, compared with 0.8912 and 0.8147 when only the diagonals are used. Gradient norm and unwhitened tails yield correlations between 0.2647 and 0.3559. For complete factors, calibration reduces MAE from 0.4770 to 0.1507 at $r = 4$ and from 0.3580 to 0.1162 at $r = 8$ (all in 10^{-4} loss units). Its calibrated MAE is respectively 51.4% and 59.5% lower than the predictor using diagonal factors, and 78.9% and 75.7% lower than the rank-only constant. Figure 2 shows the corresponding reduction in underprediction after calibration. The spectrum therefore helps distinguish gaps within each rank.

From quadratic objectives to a neural network. Appendices F.1.2 and F.1.3 verify the analytic and numerical references for these quadratic objectives. We next ask whether the predictions remain useful on a real network, where references must be estimated numerically.

6.2 Fixed-rank prediction on RoBERTa/SST-2

We next compare RoBERTa modules at a fixed rank on SST-2, using $R = 0.05$ and $r \in \{4, 8\}$. The candidates are Q and V modules from 12 layer groups. Odd-numbered groups are used for calibration and even-numbered groups for evaluation, keeping modules from the same layer together. Of 24 candidates, 20 are retained: ten calibration and ten evaluation modules, with six layer groups in each split. Four candidates are excluded because of locality, interior feasibility, restart stability, or a missing score for complete factors. Appendix F.2.1 summarizes this screening, and the exported records list all candidates. We retain the original selection used to evaluate the spectral score. Appendix F.2.2

Table 2: RoBERTa/SST-2 fixed-rank prediction on ten retained evaluation modules from six layer groups at $R = 0.05$, with ten retained calibration modules from six disjoint layer groups. Spearman ρ measures ordering; native and calibrated MAEs are in 10^{-4} loss units. Best point estimates within each rank and column are bold. Dashes denote undefined correlations or native predictions without a specified loss-scale mapping.

	$r = 4$			$r = 8$		
Predictor	$\rho \uparrow$	Native $\downarrow$	Calibrated $\downarrow$	$\rho \uparrow$	Native $\downarrow$	Calibrated $\downarrow$
Rank-only constant	—	—	0.8442	—	—	0.6762
Gradient norm	0.5273	—	0.7992	0.4788	—	0.6333
Unwhitened gradient tail	0.6727	—	0.7606	0.7818	—	0.5895
Diagonal-factor whitened tail	0.7939	0.4988	0.5655	0.7939	0.3954	0.4348
Full-factor whitened tail	**0.9394**	**0.2150**	**0.2212**	**0.9515**	**0.2276**	**0.1769**

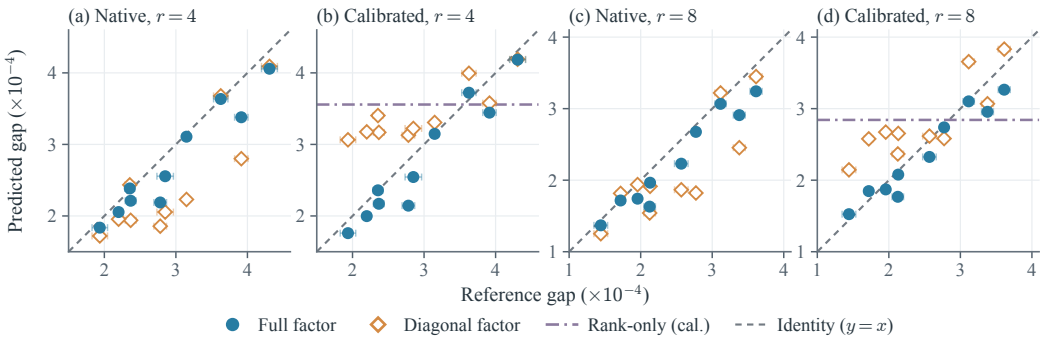

Figure 3: RoBERTa/SST-2 predictions on ten retained evaluation modules per rank. Panels (a,b) show native and calibrated predictions at $r = 4$; (c,d) show $r = 8$. Horizontal bars are reference mean $\pm$ one SD over ten repeats, not confidence intervals. Blue circles and orange diamonds denote full-factor and diagonal-factor predictions. Axis limits are shared within each rank pair. The gray dashed line is identity; the purple dash-dot line is the calibrated rank-only constant. All values are in 10^{-4} loss units.

separately examines the effect of restoring the module that failed the interior check on SST-2 and MRPC.

Each module/rank reference summarizes ten repeats. We compare 400-step results with 800-step references, using the best of three restarts for both dense and rank-restricted updates.

Prediction results. Whitening with complete factors again gives the highest Spearman correlation, 0.9394 and 0.9515, versus 0.7939 for diagonal-factor whitening at each rank (Table 2). Its calibrated MAEs are 0.2212 and 0.1769 in 10^{-4} loss units, respectively 60.9% and 59.3% below the predictor using diagonal factors and 73.8% below the rank-only constant at both ranks. Calibration affects the ranks differently. With complete factors, MAE increases slightly from 0.2150 to 0.2212 at $r = 4$, and decreases from 0.2276 to 0.1769 at $r = 8$. With diagonal factors, MAE increases after calibration at both ranks. Figure 3 shows the native and calibrated predictions separately.

Numerical stability. The trajectories reproduce the reference gaps and their variation with optimization steps and paired restarts. Among retained modules, the largest absolute change from 400 to 800 steps is 0.01203×10^{-4}, and the largest gap spread across paired restarts is 0.04258×10^{-4}. All recorded update norms stay within R. These checks support numerical stability; global optimality remains unverified. Correlations and errors are descriptive estimates across modules; the ten repeats per module are not independent evaluation cases.

6.3 Predicting the benefit of increasing rank

We now ask which modules benefit most when rank increases. For a $4 \to 8$ upgrade, the theory uses the energy in the additional singular directions, $t_{4,8} = s_4 - s_8 = \frac{1}{2}\sum_{j=5}^{8} \sigma_j(\widehat{F})^2$, computed from one proxy without separate calibration for each rank. We compare this score with the current tail s_4 and the first-order score $R(\|\widehat{g}_8\|_F - \|\widehat{g}_4\|_F)$, where $\widehat{g}_r$ is the rank-r truncated SVD of the same center gradient. The target is the observed extra gain $L(\Delta_4^{\text{continue}}) - L(\Delta_8^{\text{upgrade}})$ from paired continuations of a rank-4 checkpoint. This measures the improvement reached by training, rather than the optimal capacity gain $B_{4\to 8}$.

Table 3: Spearman correlation with observed $4 \to 8$ extra gain. Each observation is a module mean over ten paired repeats; SST-2 has ten modules in six layer groups and MRPC has 20 in 12 groups. Larger is better. The largest point estimate in each task is bold.

Predictor	SST-2 ($n = 10$)	MRPC ($n = 20$)
New spectral energy $s_4 - s_8$	**0.903**	**0.734**
Current spectral tail s_4	0.333	0.414
First-order gradient gain	0.285	0.208

The additional-direction score gives the highest rank correlation: 0.903 on SST-2 and 0.734 on MRPC, versus 0.333 and 0.414 for the current tail (ordering, not magnitude). Initial diagnostic timing is 6.6 s versus 90 s for a 400-step continuation (7.3%; Table 13). Appendix F.3 details both evaluations.

6.4 Boundary validation and sensitivity analysis

The boundary interval contains all 10, 750, and 24 global references in three controlled tests; its lower bound is positive in 7, 525, and 13 cases (Table 4(c)). Appendix F.1.1 and Table 5 describe these tests and their error bounds. Appendices F.1 and F.2 document candidates, numerical references, and checks along selected directions. Appendix F.2.2 restores an excluded module while keeping SST-2 calibration fixed; Appendix F.1.4 reports costs on controlled objectives.

7 Related Work

Rank and optimization. AdaLoRA allocates rank across modules (Zhang et al., 2023). Existing theory studies representational capacity (Zeng & Lee, 2024) and NTK-regime optimization (Jang et al., 2024). LoRA+ and LoRA-GA address learning rates and initialization (Hayou et al., 2024; Wang et al., 2024). Our comparison isolates the improvement rank leaves out of reach at a fixed module and radius.

Curvature and spectral approximation. K-FAC and shared-weight extensions use two curvature factors (Martens & Grosse, 2015; Eschenhagen et al., 2023). Low-rank approximation gives the spectral formula for feasible unconstrained optima (Eckart & Young, 1936; Mirsky, 1960); weighted problems may require methods beyond SVD (Srebro & Jaakkola, 2003).

8 Limitations and Conclusion

Limitations. Guarantees are local to one loss, module, center, and radius and require uniform error bounds beyond sampled directions. Fixed-rank SST-2 covers ten modules; sensitivity reintroduces one excluded module per task. Boundary validation uses controlled objectives; network timings are preliminary. Neural upgrades assess gain ordering; optimal-loss accuracy and equal-compute performance at larger scale remain future work.

Conclusion. Residual spectral tails measure dense-update gaps; new spectral energy predicts upgrade gains, with corrected intervals quantifying both.

AI USE STATEMENT

AI-assisted tools were used during the preparation of this work for language polishing and improving the clarity of the manuscript. AI agents were also used to assist with scaffolding parts of the experimental workflow and code infrastructure, as well as to provide suggestions during experimental design, implementation, and manuscript revision. All research questions, methodological development, theoretical analysis, experimental design decisions, result verification, and scientific conclusions were determined and validated by the authors. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The theoretical results state their assumptions explicitly, with proofs and error-budget derivations in Appendix B. Appendix E describes the experimental design, reference construction, calibration, and aggregation. Controlled objective definitions, parameter grids, budgets, and numerical tolerances are given in Appendix F.1.1. Candidate ledgers and reference audits appear in Appendices F.1.2, F.1.3, and F.2.1; Appendices F.2.2 and F.3 define the re-inclusion and rank-upgrade evaluations.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 3194–3201. IOS Press, 2025. doi: 10.3233/FAIA251185. URL https://doi.org/10.3233/FAIA251185.

Elad Ben-Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.1. URL https://aclanthology.org/2022.acl-short.1/.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical Gauss–Newton optimisation for deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 557–565, 2017.

Viktoriia Chekalina, Daniil Moskovskiy, Tatiana Matveeva, Andrey Kuznetsov, and Evgeny Frolov. Generalized fisher-weighted svd: Scalable kronecker-factored fisher approximation for compressing large language models. *arXiv preprint arXiv:2505.17974*, 2025. doi: 10.48550/arXiv.2505.17974. URL https://arxiv.org/abs/2505.17974. Version 2, revised 15 July 2026.

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(3):211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Runa Eschenhagen, Alexander Immer, Richard E. Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33624–33655, 2023. doi: 10.52202/075280-1461. URL https://papers.neurips.cc/paper_files/paper/2023/hash/6a6679e3d5b9f7d5f09cdb79a5fc3fd8-Abstract-Conference.html.

Yihang Gao and Vincent Y. F. Tan. Automatic rank determination for low-rank adaptation via submodular function maximization. *IEEE Transactions on Signal Processing*, 74:1094–1108, 2026. doi: 10.1109/TSP.2026.3671460. URL https://doi.org/10.1109/TSP.2026.3671460.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 573–582, 2016.

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. 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. PMLR, 2024. URL https://proceedings.mlr.press/v235/hayou24a.html.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0RDcd5Axok.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Yen-Chang Hsu, Ting Hua, Sung-En Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uPv9Y3gmAI5.

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.

Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (LoRA) fine-tuning for transformer models. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 6167–6206, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/12143893d9d37c3569dda800b95cabd9-Abstract-Conference.html.

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. PMLR, 2024. URL https://proceedings.mlr.press/v235/jang24d.html.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. MoRA: High-rank updating for parameter-efficient fine-tuning. *arXiv preprint arXiv:2405.12130*, 2024. doi: 10.48550/arXiv.2405.12130. URL https://arxiv.org/abs/2405.12130.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. doi: 10.48550/arXiv.2312.03732. URL https://arxiv.org/abs/2312.03732.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4158–4169, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham K. Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. SVFT: Parameter-efficient fine-tuning with singular vectors. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41425–41446, 2024. doi: 10.52202/079017-1311. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/48c368f105e8145b945227b73255635a-Abstract-Conference.html.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965, 2022. doi: 10.52202/068431-0142. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121, 2024a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. ALoRA: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 622–641. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.naacl-long.35. URL https://aclanthology.org/2024.naacl-long.35/.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072, 2024. doi: 10.52202/079017-3846.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50. URL https://doi.org/10.1093/qmath/11.1.50.

Pritish Saha, Chandrav Rajbangshi, Rudra Goyal, Mohit Goyal, Anurag Deo, Biswajit Roy, Ningthoujam Dhanachandra Singh, Raxit Goswami, and Amitava Das. GRIT – geometry-aware PEFT with K-FAC preconditioning, fisher-guided reprojection, and dynamic rank adaptation. *arXiv preprint arXiv:2601.00231*, 2026.

Steffen Schotthöfer, Emanuele Zangrando, Gianluca Ceruti, Francesco Tudisco, and Jonas Kusch. GeoLoRA: Geometric integration for parameter efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 49852–49881, 2025. URL https://openreview.net/forum?id=bsFWJ0Kget.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Nathan Srebro and Tommi S. Jaakkola. Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 720–727, 2003. URL https://cdn.aaai.org/ICML/2003/ICML03-094.pdf.

Nurbek Tastan, Stefanos Laskaridis, Martin Takáč, Karthik Nandakumar, and Samuel Horváth. LoFT: Low-rank adaptation that behaves like full fine-tuning. In *The Fourteenth International Conference on Learning Representations*, volume 2026, pp. 71294–71324, 2026. URL https://openreview.net/forum?id=86P3sb1dpr.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. doi: 10.1561/2200000048. URL https://doi.org/10.1561/2200000048.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.239. URL https://aclanthology.org/2023.eacl-main.239/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54905–54931, 2024. doi: 10.52202/079017-1741. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/62c4718cc334f6a0a62fb81c4a2095a1-Abstract-Conference.html.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-Pro: Are low-rank adapters properly optimized? In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 93787–93808, 2025. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 5078–5123, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/154926e0b66e2b2a8c1120852f31a12d-Abstract-Conference.html.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59641–59669, 2024a.

Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130819–130847, 2024b. doi: 10.52202/079017-4158.

Hao Zhang, Bo Huang, Zhenjia Li, Xi Xiao, Hui Yi Leong, Zumeng Zhang, Xinwei Long, Tianyang Wang, and Hao Xu. Sensitivity-LoRA : Low-load sensitivity-based fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13185–13199. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-emnlp.709. URL https://aclanthology.org/2025.findings-emnlp.709/.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. AutoLoRA: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5048–5060. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.282. URL https://aclanthology.org/2024.naacl-long.282/.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75513–75574. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhang25ax.html.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143, 2024.

APPENDIX CONTENTS

A Summary of Notation **16**

B Proofs and Error Budgets **16**

B.1 Softmax Curvature 16

B.2 GGN for Shared Modules 16

B.3 Empirical Proxy Construction and Valid Error Budgets 17

B.4 Local Quadratic Geometry and Its Boundary 18

B.5 Empirical Optimal-Value Transfer 19

B.6 Empirical Spectral Values and Boundary Corrections 19

B.7 Dense Trust-Region Value and Numerical Bounds 20

B.8 Rank-Increase Benefits 21

C Extended Related Work **21**

D Additional Theoretical Analysis **22**

D.1 Smallest Certified Sufficient Rank 22

D.2 Numerical Endpoints and Optimization Error 23

D.3 Re-centering and Incremental Rank 23

D.4 First-order Behavior at Small Radius 23

D.5 Scope of Multi-Module Allocation 23

E Experimental Details **24**

F Additional Experiments **25**

F.1 Controlled Validation and Numerical Audits 25

F.2 Fixed-Rank Prediction and Sensitivity 32

F.3 Rank-Upgrade Evaluation 35

F.4 Protocol Sensitivity and Supporting Observations 35

A SUMMARY OF NOTATION

All values use the same specified loss, module, center, and radius unless a new comparison is explicitly declared. The subscript d denotes dense.

Symbol	Meaning
$L, g_0, \mathcal{G}_0$	Target loss, its center gradient, and its exact GGN operator.
$J_t, H_t, \Phi_{t,j}$	Logit Jacobian, softmax Hessian, and general matrix-valued GGN factors.
$\mathcal{B}(R), \mathcal{B}_r(R)$	Frobenius ball and its rank-at-most-r subset.
$m_{L,d}, m_{L,r}, G_r$	Dense and low-rank actual-loss minima and their difference.
$q_{g,H}, D_H$	Quadratic model and unconstrained dense solution $-H^{-1}g$ for $H \succ 0$.
$\mathcal{T}_r^{\mathrm{loc}}, \mathcal{T}_r^{\mathrm{gen}}$	Twice the local quadratic gap and the unconstrained squared approximation error.
$\widehat{g}_0, \widehat{\mathcal{G}}_0$	Direct block-average gradient and GGN estimates.
$\widehat{\mathcal{S}}, \widehat{q}$	Actual empirical curvature proxy, including ridge, and its quadratic objective.
$v_d(R), v_r(R)$	Ball-constrained empirical quadratic optimal values.
$\widetilde{A}, \widetilde{B}, \widehat{F}$	Positive-definite proxy factors and $\widetilde{B}^{-1/2}\widehat{g}_0\widetilde{A}^{-1/2}$.
D_d, D_r	Unconstrained dense and rank-r empirical quadratic minimizers.
s_r	Half of the post-r squared singular-value tail of $\widehat{F}$.
$d_d(R), d_r(R)$	Exact dense boundary cost and a feasible radial upper bound on low-rank boundary cost.
$\varepsilon_g, \varepsilon_H, \rho, E(R)$	Gradient error, center curvature error, Hessian variation, and resulting gap-transfer budget.
Γ_g, Γ_G	Sampling error bounds for direct block-average estimates.
$\widehat{\eta}_{\mathrm{proxy}}$	Valid upper bound on $\left\|\widehat{\mathcal{G}}_0 - \widehat{\mathcal{S}}\right\|_{\mathrm{op}}$.
$B_{r\to r'}, t_{r,r'}$	Actual optimal rank-increase benefit and its unconstrained empirical spectral counterpart.

B PROOFS AND ERROR BUDGETS

B.1 SOFTMAX CURVATURE

Lemma B.1.1 (Softmax Hessian bound)**.** *For every probability vector $p \in \mathbb{R}^V$, $0 \preceq \mathrm{diag}(p) - pp^\top \preceq I_V/2$.*

Proof. For any v, the quadratic form equals $\mathrm{Var}_{i\sim p}(v_i)$, which is nonnegative. For $\|v\|_2 = 1$, the range width satisfies $(\max_i v_i - \min_i v_i)^2 \le 2$: each pair obeys $(v_i - v_j)^2 \le 2(v_i^2 + v_j^2) \le 2$. If a random variable X lies in $[a, b]$, then $\mathrm{Var}(X) \le (\mathbb{E}X - a)(b - \mathbb{E}X) \le (b-a)^2/4$, as follows from $\mathbb{E}[(X-a)(b-X)] \ge 0$. Thus the quadratic form is at most $1/2$ for every unit v. □

B.2 GGN FOR SHARED MODULES

Proposition B.2.1 (GGN second-moment form)**.** *The sample GGN is $\mathcal{G}_0(Z) = \sum_t w_t J_t^* H_t J_t$. If $H_t = \sum_j \lambda_{t,j} b_{t,j} b_{t,j}^\top$ is its nonzero spectral decomposition, set $\Phi_{t,j} := \sqrt{w_t \lambda_{t,j}}\, J_t^*[b_{t,j}]$. Then the target GGN operator satisfies*

$$\mathcal{G}_0[U] = \mathbb{E}\left[\sum_t w_t J_t^*[H_t J_t[U]]\right] = \mathbb{E}\left[\sum_{t,j} \langle U, \Phi_{t,j}\rangle\, \Phi_{t,j}\right], \tag{B.1}$$

$$\langle U, \mathcal{G}_0[U]\rangle = \mathbb{E}\left[\sum_{t,j} \langle U, \Phi_{t,j}\rangle^2\right]. \tag{B.2}$$

The matrices $\Phi_{t,j}$ need not have rank one.

Proof of Proposition B.2.1. For directions U, V, the definition of GGN gives

$$\langle V, \mathcal{G}_0(Z)[U] \rangle = \sum_t w_t \langle J_t[V], H_t J_t[U] \rangle = \sum_{t,j} w_t \lambda_{t,j} \langle b_{t,j}, J_t[V] \rangle \langle b_{t,j}, J_t[U] \rangle .$$

By the adjoint identity, each summand equals $\langle V, \Phi_{t,j} \rangle \langle U, \Phi_{t,j} \rangle$. Since this holds for every V, it identifies the sample operator; taking expectations gives Eq. (B.1). Setting $V = U$ gives Eq. (B.2). No factorization or rank assumption on $J_t^*[b_{t,j}]$ is used. □

Cross-position terms and the Hessian residual. For a linear module whose update is shared across input positions $s = 1, \ldots, S(Z)$, the chain rule gives

$$J_t[U] = \sum_s C_{t,s} U a_s, \qquad J_t^*[v] = \sum_s C_{t,s}^\top v \, a_s^\top ,$$

where inputs a_s and downstream Jacobians $C_{t,s}$ are evaluated at the pretrained weights. The exact sample operator is

$$\mathcal{G}_0(Z)[U] = \sum_t \sum_{s,s'} w_t \, C_{t,s}^\top H_t C_{t,s'} \, U \, a_{s'} a_s^\top .$$

Terms with $s \neq s'$ are present even when the logits are affine in Δ. Conversely, a linear adapted module does not in general make the downstream logits affine. The sample Hessian is

$$\nabla^2 \ell(Z; 0) = \mathcal{G}_0(Z) + \sum_{t,k} w_t \big(p_{t,k} - \mathbf{1}\{y_t = k\}\big) \nabla^2 u_{t,k}(0).$$

The GGN-to-Hessian mismatch must therefore be controlled separately whenever the actual loss is the target.

Lemma B.2.2 (A sufficient sample curvature bound)**.** *Let* $\|J_t\|_{F\to 2} := \sup_{\|U\|_F = 1} \|J_t[U]\|_2$. *If* $\sum_t w_t \|J_t\|_{F\to 2}^2 \le K_J^2$ *almost surely, then*

$$0 \preceq \mathcal{G}_0(Z) \preceq H_{\max} I, \qquad H_{\max} := K_J^2/2.$$

In particular, one may bound $\|J_t\|_{F\to 2}$ *by* $\sum_s \|C_{t,s}\|_{\mathrm{op}} \|a_s\|_2$.

Proof. By Lemma B.1.1, $0 \preceq H_t \preceq I/2$. Consequently, $\langle U, \mathcal{G}_0(Z)[U] \rangle \le \frac{1}{2} \sum_t w_t \|J_t[U]\|_2^2 \le K_J^2 \|U\|_F^2 / 2$. The final bound follows from the triangle inequality and $\|C_{t,s} U a_s\|_2 \le \|C_{t,s}\|_{\mathrm{op}} \|U\|_F \|a_s\|_2$. □

B.3 Empirical Proxy Construction and Valid Error Budgets

The shared-position proxy used in the experiments. The theory permits any specified self-adjoint empirical proxy. Our experiments use the following separable construction, which retains only the diagonal-location terms. For each block set

$$A_i := \frac{1}{S_i} \sum_s a_{i,s} a_{i,s}^\top, \qquad B_i := \sum_{t,s} w_{i,t} C_{i,t,s}^\top H_{i,t} C_{i,t,s}, \quad S_i := S(Z_i) \ge 1.$$

Write $A_N = N^{-1} \sum_i A_i$ and $B_N = N^{-1} \sum_i B_i$. For the declared scale convention and ridges $\tau_A, \tau_B > 0$, we set $\widetilde{A} = A_N + \tau_A I$, $\widetilde{B} = B_N + \tau_B I$, and $\widehat{\mathcal{S}}[U] = \widetilde{B} U \widetilde{A}$. This K-FAC-style proxy discards cross-position terms and replaces coupled moments by a product of aggregated moments; neither step is an identity. Variable position counts and heterogeneous position statistics can add mismatch. The weights $w_{i,t}$ are exactly those of the target loss. Any alternative scaling, estimation rule, or ridge convention defines a different proxy and must be included in its mismatch budget.

Proposition B.3.1 (Joint block-sampling event)**.** *Let* $Z_1, \ldots, Z_N$ *be i.i.d. blocks sampled according to the target objective, with* $g_i := \nabla \ell(Z_i; 0)$, $\mathcal{G}_i := \mathcal{G}_0(Z_i)$, $\|g_i\|_F \le K_g$, *and* $0 \preceq \mathcal{G}_i \preceq H_{\max} I$ *almost surely. Set* $D = d_1 d_2$ *and, for* $\delta \in (0, 1)$,

$$x_g := \log \frac{4(D+1)}{\delta}, \qquad x_G := \log \frac{4D}{\delta},$$

$$\Gamma_g := K_g \left(\sqrt{\frac{2x_g}{N}} + \frac{4x_g}{3N} \right), \qquad \Gamma_G := H_{\max} \left(\sqrt{\frac{2x_G}{N}} + \frac{2x_G}{3N} \right).$$

With probability at least $1 - \delta$*, simultaneously*

$$\|\widehat{g}_0 - g_0\|_F \le \Gamma_g, \qquad \left\|\widehat{\mathcal{G}}_0 - \mathcal{G}_0\right\|_{\mathrm{op}} \le \Gamma_G.$$

For a fixed empirical target, the same statement applies conditionally on that dataset when blocks are drawn uniformly with replacement from it. If the entire fixed target is evaluated exactly, the corresponding sampling errors are zero.

Proof. Vectorize the matrices as vectors in $\mathbb{R}^D$. The centered curvature summands satisfy $\|\mathcal{G}_i - \mathcal{G}_0\|_{\mathrm{op}} \le H_{\max}$ and $\left\|\mathbb{E}(\mathcal{G}_i - \mathcal{G}_0)^2\right\|_{\mathrm{op}} \le H_{\max}^2$. For $z_i := \mathrm{vec}(g_i - g_0)$, the self-adjoint dilation $X_i = \left(\begin{smallmatrix} 0 & z_i^\top \\ z_i & 0 \end{smallmatrix}\right)$ has norm at most $2K_g$ and satisfies $\left\|\mathbb{E}X_i^2\right\|_{\mathrm{op}} \le \mathbb{E}\|z_i\|_2^2 \le K_g^2$. Matrix Bernstein in dimensions D and $D+1$, respectively, gives the two bounds with failure probabilities $\delta/2$; a union bound proves the claim (Tropp, 2012). □

From the event to the curvature budget. Suppose $\left\|\nabla^2 L(0) - \mathcal{G}_0\right\|_{\mathrm{op}} \le \eta_{\mathrm{GGN}}$ and an available bound satisfies

$$\left\|\widehat{\mathcal{G}}_0 - \widehat{\mathcal{S}}\right\|_{\mathrm{op}} \le \widehat{\eta}_{\mathrm{proxy}}.$$

On the joint event of Proposition B.3.1, valid inputs to the empirical-surrogate transfer theorem are

$$\varepsilon_g = \Gamma_g, \qquad \varepsilon_H = \eta_{\mathrm{GGN}} + \Gamma_G + \widehat{\eta}_{\mathrm{proxy}}.$$

Indeed, the curvature bound follows directly by inserting $\mathcal{G}_0$ and $\widehat{\mathcal{G}}_0$ between $\nabla^2 L(0)$ and $\widehat{\mathcal{S}}$ and applying the triangle inequality. The proxy may depend on the same blocks: this argument is pathwise on the event and needs no independence between the proxy and its estimates. If a randomized procedure supplies the proxy bound, its failure probability must also be included in the joint event. Data-dependent selection among candidate proxies is covered only when the selected proxy's bound is valid on that event.

The residual bound must be an actual operator-norm upper bound; a maximum over a few sampled directions is not such a bound. Since $\widehat{\mathcal{S}}$ already contains the chosen ridges, their contribution is included in $\widehat{\eta}_{\mathrm{proxy}}$ and is not charged again. Sampling concentration alone establishes neither a small proxy residual nor uniform local quadratic fit. These remain explicit conditions, not conclusions supplied by the reported real-network diagnostic observations.

B.4 Local Quadratic Geometry and Its Boundary

Proof of Theorem 3.1. Since $H[D_H] = -g$, expansion and self-adjointness give

$$q_{g,H}(\Delta) = -\tfrac{1}{2}\langle g, H^{-1} g\rangle + \tfrac{1}{2}\langle \Delta - D_H, H[\Delta - D_H]\rangle.$$

The first term is $q_{g,H}(D_H)$. Taking minima on the two feasible sets and subtracting proves Eq. (5). Positive definiteness makes $q_{g,H}$ coercive, so a global rank-r minimizer exists on the closed rank-r set. If this minimizer and D_H are feasible on the ball, the ball-constrained and unconstrained values coincide, proving the final claim. □

Proof of Corollary 3.2. The whitening map is invertible and preserves rank. Cyclicity of the trace gives $\langle g, \Delta\rangle = \langle F, M\rangle$ and $\langle \Delta, B\Delta A\rangle = \|M\|_F^2$; completing the Euclidean square yields Eq. (6). The Frobenius constraint on Δ is exactly $M \in \mathcal{E}_R$, proving Eq. (7). The unconstrained optima are $-F$ and any best rank-r approximation $-(F)_r$. When both are feasible, the dense squared residual is zero and the restricted one is $\sum_{j>r} \sigma_j(F)^2$ by Eckart–Young–Mirsky. □

Why the spectrum alone does not describe an active boundary. Fix $F = \mathrm{diag}(f_1, f_2)$ with $f_1 \geq f_2 > 0$, let $H[\Delta] = \kappa\Delta$ and $g = \sqrt{\kappa}F$, and take $r = 1$. Whitening gives the same F for every $\kappa > 0$. If $R < f_1/\sqrt{\kappa}$, both dense and rank-one optima are on the boundary. For a fixed norm t, their best linear terms are $-t\,\|g\|_F$ and $-t\,\|(g)_1\|_F$, respectively, so

$$\tfrac{1}{2}\mathcal{T}_1^{\mathrm{loc}}(R; g, H) = R\sqrt{\kappa}\big(\sqrt{f_1^2 + f_2^2} - f_1\big).$$

This depends on κ despite an unchanged whitened spectrum. It explains why Section 5 retains the geometry through the boundary corrections.

B.5 Empirical Optimal-Value Transfer

This subsection proves Theorem 4.1 for the actual empirical quadratic in Eq. (9). All arguments are deterministic after conditioning on an event where the stated budgets hold. In particular, the empirical proxy may depend on the same data as the gradient estimate; no independence between those estimates is required.

Proof of Theorem 4.1. For any $\Delta \in \mathcal{B}(R)$, Taylor's formula along the segment $t \mapsto t\Delta$ gives

$$\begin{aligned} L(\Delta) - L(0) &= \langle g_0, \Delta\rangle + \int_0^1 (1-t)\left\langle \Delta, \nabla^2 L(t\Delta)[\Delta]\right\rangle dt \\ &= \langle g_0, \Delta\rangle + \tfrac{1}{2}\left\langle \Delta, \nabla^2 L(0)[\Delta]\right\rangle + \mathcal{R}(\Delta). \end{aligned}$$

The locality assumption implies

$$|\mathcal{R}(\Delta)| \leq \int_0^1 (1-t)\rho t\,\|\Delta\|_F^3\; dt = \tfrac{\rho}{6}\,\|\Delta\|_F^3\,.$$

Subtracting $\widehat{q}(\Delta)$ and applying the gradient and operator error bounds therefore yields

$$\begin{aligned} |L(\Delta) - L(0) - \widehat{q}(\Delta)| &\leq \varepsilon_g\,\|\Delta\|_F + \tfrac{1}{2}\varepsilon_H\,\|\Delta\|_F^2 + \tfrac{\rho}{6}\,\|\Delta\|_F^3 \\ &\leq e(R) := R\varepsilon_g + \tfrac{1}{2}R^2\varepsilon_H + \tfrac{\rho}{6}R^3 = \tfrac{1}{2}E(R). \end{aligned} \tag{B.3}$$

For any common nonempty feasible set $\mathcal{C} \subseteq \mathcal{B}(R)$, a uniform bound $|f - h| \leq e(R)$ implies $|\inf_{\mathcal{C}} f - \inf_{\mathcal{C}} h| \leq e(R)$: take the infimum in each of the pointwise inequalities $h - e(R) \leq f \leq h + e(R)$. Here the sets $\mathcal{B}(R)$ and $\mathcal{B}_r(R)$ are compact, since the rank-at-most-r set is closed, and both objectives are continuous. Thus the infima are minima, and

$$|m_{L,d}(R) - L(0) - v_d(R)| \leq e(R), \qquad |m_{L,r}(R) - L(0) - v_r(R)| \leq e(R).$$

Subtract the two value differences and use the triangle inequality to obtain $|G_r(R) - \widehat{G}_r(R)| \leq 2e(R) = E(R)$. □

The same proof applies to any pair of nonempty feasible subsets of the ball; the rank constraint enters only through the chosen sets. The curvature decomposition in Eq. (12) follows from

$$\widehat{\mathcal{S}} - \nabla^2 L(0) = (\widehat{\mathcal{S}} - \widehat{\mathcal{G}}_0) + (\widehat{\mathcal{G}}_0 - \mathcal{G}_0) + (\mathcal{G}_0 - \nabla^2 L(0))$$

and the operator-norm triangle inequality. A probability statement is inherited from a single simultaneous event establishing its gradient, curvature, and locality budgets; selected-direction checks do not supply that event.

B.6 Empirical Spectral Values and Boundary Corrections

We first derive the empirical unconstrained values, including the equality needed for the radial feasible point. Assume Eq. (13) and define $M := \widetilde{B}^{1/2}\Delta\widetilde{A}^{1/2}$. The change of coordinates is invertible and preserves matrix rank, and

$$\widehat{q}(\Delta) = \left\langle \widehat{F}, M\right\rangle + \tfrac{1}{2}\,\|M\|_F^2 = \tfrac{1}{2}\left\|M + \widehat{F}\right\|_F^2 - \tfrac{1}{2}\left\|\widehat{F}\right\|_F^2. \tag{B.4}$$

The dense unconstrained minimizer is $M = -\widehat{F}$; its inverse transform is the dense minimizer D_d defined in Section 5. The Eckart–Young–Mirsky theorem gives an unconstrained rank-at-most-r minimizer $M = -(\widehat{F})_r$, whose inverse transform is D_r. Consequently,

$$v_d^\infty = -\tfrac{1}{2}\left\|\widehat{F}\right\|_F^2, \qquad v_r^\infty = -\tfrac{1}{2}\left\|(\widehat{F})_r\right\|_F^2 = v_d^\infty + s_r.$$

These statements hold for any optimal truncated SVD, including tied singular values. If the selected D_d and D_r lie in $\mathcal{B}(R)$, restricting to the ball cannot change either minimum, proving the empirical interior formula. Checking one selected D_r is sufficient, but need not be necessary when optimal truncations are nonunique.

For every scalar α, the SVD orthogonality identity $\left\langle \widehat{F}, (\widehat{F})_r \right\rangle = \left\|(\widehat{F})_r\right\|_F^2$ implies

$$\begin{aligned}\widehat{q}(\alpha D_r) &= -\alpha\left\|(\widehat{F})_r\right\|_F^2 + \tfrac{1}{2}\alpha^2\left\|(\widehat{F})_r\right\|_F^2 \\ &= v_r^\infty + \tfrac{1}{2}(1-\alpha)^2\left\|(\widehat{F})_r\right\|_F^2.\end{aligned} \tag{B.5}$$

With $\alpha = \alpha_r$ from Eq. (16), $\alpha_r D_r \in \mathcal{B}_r(R)$, including the case $D_r = 0$.

Proof of Theorem 5.1. Relaxing the radius constraint gives $v_r^\infty \le v_r(R)$, and the feasible point in Eq. (B.5) gives

$$v_r^\infty \le v_r(R) \le v_r^\infty + d_r(R). \tag{B.6}$$

Since $v_d(R) = v_d^\infty + d_d(R)$, subtracting this dense value yields

$$s_r - d_d(R) \le \widehat{G}_r(R) \le s_r + d_r(R) - d_d(R).$$

The inclusion $\mathcal{B}_r(R) \subseteq \mathcal{B}(R)$ also gives $\widehat{G}_r(R) \ge 0$, proving Eq. (17). Theorem 4.1 expands these bounds by $E(R)$. The same feasible-set inclusion gives $G_r(R) \ge 0$, so clipping the lower endpoint at zero yields Eq. (18). If both unconstrained minimizers are feasible, $d_d(R) = d_r(R) = 0$, which recovers the interior spectral-tail interval. □

The radial point is used only as a feasible upper bound. No assertion that it minimizes the rank-constrained problem on the ball is needed. A better feasible rank-at-most-r point may replace it to improve the upper bound, provided its objective is evaluated with a valid numerical upper bound.

B.7 Dense Trust-Region Value and Numerical Bounds

Let $\widetilde{A} = V\,\mathrm{diag}(a)V^\top$ and $\widetilde{B} = U\,\mathrm{diag}(b)U^\top$ be orthogonal eigendecompositions, with $a_j, b_i > 0$. Set $X = U^\top \Delta V$, $C = U^\top \widehat{g}_0 V$, and $h_{ij} = b_i a_j > 0$. Orthogonal invariance of the Frobenius norm gives the dense problem

$$v_d(R) = \min_{\sum_{ij} X_{ij}^2 \le R^2} \sum_{ij}\left(C_{ij}X_{ij} + \tfrac{1}{2}h_{ij}X_{ij}^2\right). \tag{B.7}$$

Using the Lagrange term $\lambda(\|X\|_F^2 - R^2)/2$, the unconstrained minimizer of the Lagrangian is

$$X_{ij}(\lambda) = -\frac{C_{ij}}{h_{ij} + \lambda}, \qquad \lambda \ge 0, \tag{B.8}$$

and its minimum is

$$\ell_d(\lambda) := -\tfrac{1}{2}\sum_{ij}\frac{C_{ij}^2}{h_{ij}+\lambda} - \tfrac{\lambda}{2}R^2 \le v_d(R). \tag{B.9}$$

The inequality follows directly from weak duality: the Lagrange term is nonpositive at every feasible X. The objective is strictly convex and the ball has a strictly feasible point because $R > 0$, so its optimality conditions are sufficient and strong duality holds. They can also be verified directly.

If $\|X(0)\|_F \le R$, the unconstrained minimum is feasible and $\lambda = 0$ attains both primal and dual optima. Otherwise, $C \neq 0$ and

$$f(\lambda) := \sum_{ij} \frac{C_{ij}^2}{(h_{ij} + \lambda)^2}$$

is continuous and strictly decreasing for $\lambda \ge 0$, with $f(0) > R^2$ and $f(\lambda) \to 0$ as $\lambda \to \infty$. There is a unique $\lambda_* > 0$ satisfying $f(\lambda_*) = R^2$. The point $X(\lambda_*)$ satisfies stationarity, feasibility, and complementary slackness, and hence solves Eq. (B.7). Substitution gives $v_d(R) = \ell_d(\lambda_*)$.

For numerical evaluation, any $\lambda \ge 0$ provides a dual lower bound via Eq. (B.9), whether or not the radius equation has converged. Any feasible X gives a primal upper bound

$$u_d := \sum_{ij} \left(C_{ij} X_{ij} + \tfrac{1}{2} h_{ij} X_{ij}^2\right) \ge v_d(R).$$

For example, rescale a computed candidate to have norm at most R before using its primal value. To obtain a rigorous floating-point interval, the eigendecomposition, feasibility, and value evaluations must themselves have certified error bounds or conservative rounding; an unverified floating-point dual evaluation is not automatically a rigorous bound.

Suppose verified numerical values satisfy $\ell_d \le v_d(R) \le u_d$. Equation (B.6) gives

$$v_r^\infty - u_d \le \widehat{G}_r(R) \le \widehat{q}(\alpha_r D_r) - \ell_d.$$

Apply Theorem 4.1 and $G_r(R) \ge 0$ to obtain Eq. (19). More generally, if $\ell_r^\infty \le v_r^\infty$ is a verified spectral lower bound and $u_r \ge \widehat{q}(\overline{\Delta}_r)$ is a verified objective upper bound for any certified feasible $\overline{\Delta}_r \in \mathcal{B}_r(R)$, then

$$\max\{0, \ell_r^\infty - u_d - E(R)\} \le G_r(R) \le u_r - \ell_d + E(R). \tag{B.10}$$

This form accounts separately for numerical errors and the statistical or model-approximation error $E(R)$. In particular, the objective from an approximate rank-r SVD generally bounds v_r^∞ from above; it is not the required lower bound unless its approximation error is also bounded.

B.8 Rank-Increase Benefits

Proof of Corollary 5.2. Write the proxy benefit as $\widehat{B}_{r\to r'}(R) := v_r(R) - v_{r'}(R)$ and the boundary excesses as $\beta_k(R) := v_k(R) - v_k^\infty$ for $k \in \{r, r'\}$. Equation (B.6) gives $0 \le \beta_k(R) \le d_k(R)$, so

$$\widehat{B}_{r\to r'}(R) = t_{r,r'} + \beta_r(R) - \beta_{r'}(R), \qquad t_{r,r'} - d_{r'}(R) \le \widehat{B}_{r\to r'}(R) \le t_{r,r'} + d_r(R).$$

The uniform fit in Eq. (B.3) applies to both $\mathcal{B}_r(R)$ and $\mathcal{B}_{r'}(R)$, hence

$$\begin{aligned}\left|B_{r\to r'}(R) - \widehat{B}_{r\to r'}(R)\right| &\le |m_{L,r}(R) - L(0) - v_r(R)| \\ &\quad + |m_{L,r'}(R) - L(0) - v_{r'}(R)| \le 2e(R) = E(R).\end{aligned}$$

Expanding the proxy bounds by $E(R)$ and using $\mathcal{B}_r(R) \subseteq \mathcal{B}_{r'}(R)$, which implies $B_{r\to r'}(R) \ge 0$, proves Eq. (20). If $D_r, D_{r'} \in \mathcal{B}(R)$, their scaling factors equal one and both $d_r(R)$ and $d_{r'}(R)$ vanish. This gives $|B_{r\to r'}(R) - t_{r,r'}| \le E(R)$ without a condition on D_d. □

For any feasible trained updates $\widetilde{\Delta}_r \in \mathcal{B}_r(R)$ and $\widetilde{\Delta}_{r'} \in \mathcal{B}_{r'}(R)$, define their nonnegative optimization errors by $e_k = L(\widetilde{\Delta}_k) - m_{L,k}(R)$. Subtracting these two definitions yields

$$L(\widetilde{\Delta}_r) - L(\widetilde{\Delta}_{r'}) = B_{r\to r'}(R) + e_r - e_{r'}.$$

Thus an oracle benefit bound becomes a bound for particular trained endpoints only after controlling the optimization errors as well. Feasibility under the same radius and rank constraints is essential to this interpretation.

C Extended Related Work

Section 7 discusses the closest connections to rank allocation, LoRA theory, and spectral approximation. This appendix gives broader context on adaptation methods and the tools behind the diagnostic.

PEFT landscape. Parameter-efficient fine-tuning methods include adapters (Houlsby et al., 2019), soft or prefix prompts (Li & Liang, 2021; Lester et al., 2021), bias-only updates and activation rescaling (Ben-Zaken et al., 2022; Liu et al., 2022), and low-rank weight updates (Hu et al., 2022). Unified PEFT views organize these methods by the trainable subspace and insertion mechanism (He et al., 2022). These alternatives provide the wider context for the matrix-update rank constraint studied here.

LoRA variants and rank design. Within the LoRA family, quantized and decomposed variants improve practical fine-tuning efficiency or parameterization (Dettmers et al., 2023; Liu et al., 2024a), while vectorized or extremely small adapters reduce the trainable parameter count further (Kopiczko et al., 2024; Bałazy et al., 2025). Spectral initialization and spectral-coordinate methods exploit singular structure during training (Meng et al., 2024; Zhang & Pilanci, 2024b; Lingam et al., 2024); the diagnostic here uses the spectrum of the curvature-whitened local gradient. High-rank, one-step, and full-fine-tuning-like variants change capacity or the optimization target (Jiang et al., 2024; Zhang et al., 2025b; Tastan et al., 2026). Beyond AdaLoRA, dynamic, automatic, sensitivity-based, and submodular methods choose or allocate ranks during training (Valipour et al., 2023; Zhang et al., 2024; Liu et al., 2024b; Zhang et al., 2025a; Gao & Tan, 2026). Appendix D.5 states the additional equal-cost and feasibility conditions needed for a blockwise allocation rule.

Further optimization perspectives. Gradient-projection and compression views relate low-rank training to memory efficiency (Hao et al., 2024; Zhao et al., 2024), while complexity analyses study its computational limits (Hu et al., 2025). Rank-dependent scaling and gradient correction offer further ways to improve optimization (Kalajdzievski, 2023; Wang et al., 2025). These perspectives complement the distinction between an optimal local capacity gap and the gap reached by finite-budget training.

Additional curvature methods and finite-sample tools. Related approaches include natural-gradient and scalable curvature approximations (Amari, 1998; Grosse & Martens, 2016; Botev et al., 2017; Anil et al., 2020; Martens, 2020), as well as geometry-aware PEFT methods (Zhang & Pilanci, 2024a; Schotthöfer et al., 2025; Saha et al., 2026). Fisher-weighted compression uses related geometry to compress pretrained weights (Hsu et al., 2022; Chekalina et al., 2025). Our diagnostic keeps the distinctions among Fisher, GGN, and Hessian geometry explicit (Kunstner et al., 2019; Martens, 2020). The finite-sample extension uses standard matrix concentration (Tropp, 2012; 2015); optimal-value transfer avoids a separate inverse-matrix and optimizer-perturbation chain.

D Additional Theoretical Analysis

The following consequences use the same loss, module, and Frobenius ball as the main results. Write $p = \min\{d_1, d_2\}$ and $G_r(R) = m_{L,r}(R) - m_{L,d}(R)$.

D.1 Smallest Certified Sufficient Rank

Suppose the empirical-proxy transfer and boundary bounds hold on a common event, with error $E(R)$, spectral score s_r, dense boundary penalty $d_d(R)$, and feasible rank-r penalty $d_r(R)$. Define

$$\begin{aligned} \mathrm{LB}_r(R) &:= \max\{0, s_r - d_d(R) - E(R)\}, \\ \mathrm{UB}_r(R) &:= s_r + d_r(R) - d_d(R) + E(R). \end{aligned} \tag{D.1}$$

These bound $G_r(R)$ simultaneously whenever the underlying uniform error event holds. Valid numerical bounds for the dense optimum can replace its exact value, with the conservative endpoints prescribed by the boundary bound.

Corollary D.1.1 (Smallest rank passing a sufficient test)**.** *For $\epsilon > 0$, let*

$$r_{\mathrm{cert}}(\epsilon) := \min\{r \in \{0, \dots, p\} : \mathrm{UB}_r(R) \le \epsilon\}, \tag{D.2}$$

with $r_{\mathrm{cert}} = +\infty$ if the set is empty. If finite, every $r \ge r_{\mathrm{cert}}$ is sufficient: $G_r(R) \le \epsilon$. It also certifies the true minimum sufficient rank if $r_{\mathrm{cert}} = 0$ or $\mathrm{LB}_{r_{\mathrm{cert}}-1}(R) > \epsilon$.

Proof. The upper bound certifies r_{cert}, and nested rank constraints make $G_r(R)$ nonincreasing. The additional lower bound excludes the preceding rank and hence every smaller rank. Without that exclusion, r_{cert} is only an upper bound on the true minimum; an empty test set means inconclusive, even though $G_p(R) = 0$. □

D.2 Numerical Endpoints and Optimization Error

For feasible numerical endpoints $\widetilde{\Delta}_k$, define their true optimization errors by $\xi_k := L(\widetilde{\Delta}_k) - m_{L,k}(R) \geq 0$, for $k \in \{d, r, r'\}$ and $r' > r$. The observed dense-relative gap and rank-increase gain satisfy the exact identities

$$\begin{aligned} L(\widetilde{\Delta}_r) - L(\widetilde{\Delta}_d) &= G_r(R) + \xi_r - \xi_d, \\ L(\widetilde{\Delta}_r) - L(\widetilde{\Delta}_{r'}) &= B_{r\to r'}(R) + \xi_r - \xi_{r'}, \end{aligned} \tag{D.3}$$

where $B_{r\to r'}(R) := m_{L,r}(R) - m_{L,r'}(R)$. Both identities follow by adding and subtracting the respective oracle values. If independently justified bounds $\xi_k \leq u_k$ are available and the first observed difference is $\widetilde{G}_r$, then $\widetilde{G}_r - u_r \leq G_r(R) \leq \widetilde{G}_r + u_d$; the same rule applies to the rank gain with $u_{r'}$ replacing u_d. Restart and optimization-extension stability alone do not supply these optimization-error bounds. All terms must refer to the same objective; held-out endpoint differences are separate quantities.

D.3 Re-centering and Incremental Rank

Remark D.3.1 (A new low-rank correction at a checkpoint)**.** *At a fixed checkpoint Δ_c, apply the local results to $L_c(U) := L(\Delta_c + U) - L(\Delta_c)$ after verifying their assumptions at the new center. The constraint $\mathrm{rank}(U) \leq r$ controls the additional correction. If $\mathrm{rank}(\Delta_c) = r_c$, it only implies $\mathrm{rank}(\Delta_c + U) \leq r_c + r$. Preserving a total rank-r constraint instead requires $\mathrm{rank}(\Delta_c + U) \leq r$, a different feasible set to which the spectral formula does not apply verbatim. When a checkpoint is selected using the estimation data, a probabilistic guarantee also requires a valid uniform or conditional event, or independent estimation data.*

D.4 First-order Behavior at Small Radius

Proposition D.4.1 (Small-radius rank gap)**.** *Suppose L is twice continuously differentiable on a neighborhood of $\mathcal{B}(R)$ and $\sup_{\Delta\in\mathcal{B}(R)} \left\|\nabla^2 L(\Delta)\right\|_{\mathrm{op}} \leq K$. Let $g_0 = \nabla L(0)$ and let $(g_0)_r$ be an SVD truncation to rank r. Then*

$$\left|G_r(R) - R\big(\|g_0\|_F - \|(g_0)_r\|_F\big)\right| \leq KR^2. \tag{D.4}$$

Proof. Taylor's theorem bounds $|L(\Delta) - L(0) - \langle g_0, \Delta\rangle|$ by $KR^2/2$ throughout the ball. Cauchy–Schwarz gives the dense linear minimum $-R\,\|g_0\|_F$. The singular-value trace inequality followed by Cauchy–Schwarz gives $\langle g_0, \Delta\rangle \geq -R\,\|(g_0)_r\|_F$ on $\mathcal{B}_r(R)$, attained at $-R(g_0)_r/\,\|(g_0)_r\|_F$ when $(g_0)_r \neq 0$; otherwise the linear minimum is zero. Each nonlinear minimum differs from its linear counterpart by at most $KR^2/2$. Subtraction proves the bound. □

Thus the leading term as $R \downarrow 0$ is determined by the gradient. When that term vanishes, including when $g_0 = 0$, the bound only gives an $O(R^2)$ statement; it does not identify the quadratic coefficient.

D.5 Scope of Multi-Module Allocation

Remark D.5.1 (Equal rank costs and feasible block optima)**.** *For an exactly block-additive separable proxy, suppose the unconstrained dense and allocated low-rank block optima are jointly feasible in the shared ball for every allocation under consideration. Under an equal-unit-cost budget $\sum_m r_m \leq R_{\mathrm{tot}}$, minimizing the sum of spectral tails selects the largest R_{tot} block singular-value energies, resolving ties to preserve each block's prefix. This follows by maximizing captured energy. A parameter budget instead charges $(d_{1m} + d_{2m})r_m$ for module m; unequal costs do not admit this general sorting rule. Cross-module curvature and an active shared boundary also require separate treatment.*

E EXPERIMENTAL DETAILS

Experimental design. The evaluation combines controlled objectives, fixed-rank prediction, and rank-upgrade comparisons. Controlled results and their reference audits appear in Appendix F.1. The boundary tests use the parameter grids, analytic error budgets, global references, and numerical tolerances in Appendix F.1.1. Appendices F.2.2 and F.3 describe the additional fixed-rank re-inclusion analysis and SST-2/MRPC rank-upgrade outcomes, respectively.

E.1 TARGET, DATA, AND CANDIDATE SETTINGS

The primary fixed-rank comparisons use dense and low-rank updates to the same module within a common Frobenius ball. The RoBERTa/SST-2 evaluation uses $R = 0.05$ and ranks $4, 8$. Its 24 candidates comprise Q and V modules from 12 layer groups. Odd-numbered groups form the calibration split, and even-numbered groups form the evaluation split. Each split retains ten modules from six layer groups under the original screen (Appendix F.2.1).

The controlled objectives provide analytic references, whereas the network comparisons use numerical references. Prediction metrics are reported separately by task and rank. Earlier train- and evaluation-loss endpoint comparisons are presented separately in Appendix F.4; they do not identify population-risk oracle gaps.

E.2 SCORE CONSTRUCTION AND COMPARISONS

The five fixed-rank predictors are a rank-only constant, gradient norm, unwhitened gradient tail, diagonal-factor whitening, and full-factor whitening (Table 1). Full-factor whitening uses the full matrices of the two separable curvature factors; diagonal-factor whitening uses their diagonal parts. These proxies are distinct from the full output-Hessian GGN operator defined in Section 2.

For rank upgrades, the spectral predictor is $s_4 - s_8$, computed from one proxy without separate rank-wise calibration. The comparisons also include s_4 and the first-order gradient gain (Section 6.3). Directional residuals characterize the sampled directions; their finite maximum does not establish a uniform operator-norm bound.

E.3 NUMERICAL REFERENCES AND OPTIMIZATION UNCERTAINTY

Each RoBERTa/SST-2 module/rank reference summarizes ten repeats. Dense and rank-restricted losses use the best of three restarts at each checkpoint. The primary and extended-reference checkpoints are 400 and 800 steps, respectively. The audit reports the change in their gap and the spread across paired restarts; all recorded update norms satisfy $R = 0.05$ (Appendix F.2.1).

For a common objective, writing the two optimization errors as $e_r, e_d \geq 0$ gives

$$\widetilde{G}_r = G_r + e_r - e_d.$$

If validated upper bounds $e_r \leq u_r$ and $e_d \leq u_d$ are available, then $\widetilde{G}_r - u_r \leq G_r \leq \widetilde{G}_r + u_d$. The reported extension and restart checks characterize numerical stability, without certifying these optimization-error bounds or global optimality.

E.4 CALIBRATION, EVALUATION, AND REPORTING

Affine least-squares mappings are fitted separately at each rank using the ten retained SST-2 calibration modules. The rank-only constant is the corresponding calibration-reference mean. These mappings remain frozen for the re-inclusion analysis, including MRPC. Module/rank reference means aggregate ten repeats before prediction errors are computed.

Spearman correlation measures ordering, and MAE measures loss-scale prediction error. Repeat standard deviations describe within-module variability; repeats do not increase the number of independent modules or layer groups. The network comparisons report descriptive point estimates without statistical significance claims. Controlled scalar-sampling intervals and deterministic boundary enclosures are distinguished in Appendix F.1.1.

E.5 Data summaries and resource accounting

Appendix F contains the candidate ledgers, reference summaries, and directional and optimization diagnostics. Table 9 reports stage timings and process memory for the specified controlled workload. Those measurements characterize that workload and are not estimates of end-to-end network fine-tuning cost. Table 13 separately reports an initial rank-upgrade diagnostic timing, including gradient and factor collection, relative to a 400-step continuation.

F Additional Experiments

This appendix details the controlled validation, candidate coverage, numerical references, and computational costs. It also reports RoBERTa reference audits, fixed-rank re-inclusion sensitivity, and the SST-2/MRPC rank-upgrade evaluation.

F.1 Controlled Validation and Numerical Audits

F.1.1 Controlled Models and Bound Validation

We separate exact quadratic identities, active-boundary corrections, sampling uncertainty, and deterministic approximation error using controlled objectives. Table 4 reports all records; Table 5 specifies their reference constructions and error budgets. These checks use analytic controlled objectives and do not establish real-network certification.

Exactness and the interior boundary. Across 64 separable-quadratic case/rank pairs, the spectral score agrees with the direct objective difference at analytic SVD optima to numerical precision. The radius sweep contains eight interior records with similarly negligible error and ten boundary records. All ten boundary records fail the interior gate; their score MAE is 0.8218×10^{-4} and their maximum error is 2.0178×10^{-4}. These errors describe the uncorrected spectral score. Panel (c) applies Eq. (18) to the same records and two additional controlled grids, retaining every case. The corrected interval contains the reference in all 10, 750, and 24 boundary records, respectively, with mean full widths 0.3821, 0.6738, and 0.8017 in 10^{-4} loss units. Its lower bound is positive in $7/10$, $525/750$, and $13/24$ cases. The corresponding uncorrected intervals $[\max\{0, s_r - E(R)\}, s_r + E(R)]$ contain only $0/10$, $152/750$, and $0/24$ references. These are deterministic inclusion checks, not estimates of a nominal statistical coverage probability. The interior rows retain the remaining records from each grid.

Sampling versus deterministic enclosure. The scalar-sampling check uses 1,000 independent draws per sample size under a known Gaussian noise model. Empirical coverage is 94.9%, 93.9%, and 95.0% at $n = 128, 512, 2048$, respectively; each Wilson 95% interval contains the nominal 95% level. Quadrupling n approximately halves the mean interval full width, from 1.1119 to 0.5557 to 0.2780 in 10^{-4} loss units. This check concerns a scalar amplitude with known variance, rather than the full token-moment estimator. Separately, all 50 proxy-plus-quartic configurations satisfy the deterministic enclosure with gap halfwidth $\eta R^2 + \kappa R^4/2$. This $50/50$ inclusion result is distinct from statistical coverage and has no Wilson interval.

Directional check. Figure 4 shows one controlled instance (S01, repeat 1) along four directions, each with 121 signed steps in $[-0.06, 0.06]$. The exported update norms satisfy $\|\Delta\|_F = |\alpha|$. The GGN and direct objective increments agree to a maximum absolute residual of 1.67×10^{-16}. Every displayed proxy residual lies within the provided uniform envelope; the largest residual-to-envelope ratio over nonzero steps is 5.71%. These directional observations illustrate the controlled construction; they do not establish a uniform bound from finitely many directions or measure statistical coverage.

Validation quantities. The boundary evaluation includes gradient, curvature, and local-remainder budget components, dense primal/dual values, and low-rank reference checks. The constructions below specify each controlled objective and its budget.

Boundary-grid construction. For the retained isotropic family, let $g = \mathrm{diag}(\gamma_1, \ldots, \gamma_{24})$, where $\gamma_j = cj^{-1/2}$ and c makes $\frac{1}{4}\sum_{j>4} \gamma_j^2 = 3.1 \times 10^{-4}$. The curvature is $2I$. Both ranks $4, 8$ use the

Table 4: Controlled validation. Errors and interval full widths are in 10^{-4} loss units. Panel (a) retains the original uncorrected-score errors and interior-gate failures. Panel (b) separates nominal 95% scalar-sampling coverage from the deterministic quartic enclosure. Panel (c) reports inclusion counts for the new boundary interval and the uncorrected interval, widths after lower-bound clipping, and positive lower bounds. Boundary/interior groups use the proxy dense solution's feasibility; these grids carry no nominal coverage probability or Wilson interval.

(a) Score accuracy and interior checks

Controlled case	Records	MAE	Max. error	Gate failures
Separable quadratic, interior	64	2.43×10^{-15}	1.03×10^{-14}	0
Radius sweep, interior	8	2.85×10^{-15}	2.98×10^{-15}	0
Radius sweep, boundary	10	0.8218	2.0178	10
Scalar sampling, $n = 128$	1000	0.2290	1.0995	0
Scalar sampling, $n = 512$	1000	0.1157	0.4841	0
Scalar sampling, $n = 2048$	1000	0.0565	0.2481	0
Proxy + quartic deviation	50	0.0173	0.0604	0

(b) Sampling coverage and deterministic enclosure

Controlled case	Covered	Coverage	Wilson 95% CI	Mean width
Scalar sampling, $n = 128$	949/1000	94.9%	$[93.36, 96.10]\%$	1.1119
Scalar sampling, $n = 512$	939/1000	93.9%	$[92.24, 95.22]\%$	0.5557
Scalar sampling, $n = 2048$	950/1000	95.0%	$[93.47, 96.19]\%$	0.2780
Proxy + quartic deviation	50/50	100.0%	—	0.9688

(c) Boundary intervals: deterministic inclusion and full width

Setting	New	Uncorrected	Mean width	Max. width	LB > 0
Retained sweep, boundary	10/10	0/10	0.3821	1.0822	7/10
Retained sweep, interior	8/8	8/8	0.0000	0.0000	8/8
Nonzero E, boundary	750/750	152/750	0.6738	1.1923	525/750
Nonzero E, interior	600/600	600/600	1.1457	4.0519	598/600
Anisotropic, boundary	24/24	0/24	0.8017	2.2534	13/24
Anisotropic, interior	6/6	6/6	0.0000	0.0000	6/6

Table 5: Controlled-case reference constructions and budgets. Dimension is $d = 24$ where specified; g_r denotes the rank-r SVD truncation of the gradient and t is the update norm. The scalar-sampling row uses a statistical interval; the remaining budgets are deterministic. Panel (b) gives four examples from the retained sweep; all loss values are in 10^{-4} units, while radius R is unscaled.

(a) Reference constructions and budgets

Case	Reference construction	Budget	Configuration
Exact separable quadratic	Direct objective difference at analytic SVD optima.	True factors; no approximation or sampling budget.	32 Table 1 cases $\times$ ranks $4, 8$; perturbation 1; $d = 24$, $R = 0.05$.
Radius / interior boundary	Isotropic Frobenius-ball optimum with $t = \min(R, \|g_r\|_F/\lambda)$.	Exact proxy, $E(R) = 0$; boundary corrections retain all failed interior gates.	9 radii $\times$ 2 ranks; $d = 24$, $\lambda = 2$.
Independent scalar sampling	S01 quadratic, rank 4, perturbation 1; the gap scales as amplitude squared.	Known Gaussian variance; exact 95% mean interval mapped through squaring.	1,000 draws per $n \in \{128, 512, 2048\}$; $\sigma = 0.45$, $R = 0.05$.
Proxy + local deviation	Isotropic quadratic plus radial quartic; $\lambda t + \kappa t^3 = \|g_r\|_F$.	Deterministic uniform bound; gap halfwidth $\eta R^2 + \kappa R^4/2$.	5 values of $\eta \times 5$ values of κ $\times$ ranks $4, 8$; $d = 24$, $R = 0.05$.
Nonzero transfer budget	Isotropic quadratic plus radial quartic with gradient multiplier $1 + \zeta$.	$E(R) = 2R\|\zeta\|\|g\|_F + \eta R^2 + \kappa R^4$.	$5\eta \times 5\kappa \times 3\zeta \times 9R \times 2r$; 1,350 records, 750 at the boundary.
Single-sided anisotropy	Lagrangian spectral solve with matching primal and weak-dual values.	Exact proxy, $E(R) = 0$; radial candidate bounds the rank-side boundary cost.	3 condition numbers $\times$ 5 relative radii $\times$ 2 ranks; 30 records, 24 at the boundary.

(b) Retained-sweep examples: loss values in 10^{-4} units; R unscaled

r	R	Tail s_r	Reference gap	Boundary interval
4	0.008	3.1000	1.0822	[0.0000, 1.0822]
4	0.016	3.1000	2.1645	[2.0396, 2.1645]
8	0.016	1.9379	1.2758	[0.8775, 1.2758]
8	0.020	1.9379	1.5947	[1.5413, 1.5947]

same radius grid

$$\mathcal{R} = \{0.008, 0.012, 0.016, 0.020, 0.024, 0.028, 0.032, 0.040, 0.050\}.$$

The rank-r optimum has norm $\min\{R, \|g_r\|_F/2\}$; the dense optimum uses g instead of g_r. This recovers all 18 earlier reference gaps. The examples in Table 5(b) show that the uncorrected tail can substantially exceed the reference. In this isotropic family, $\alpha_r D_r$ is itself optimal, so the corrected upper endpoint equals the reference when $E = 0$. This equality need not hold under anisotropy.

Nonzero gradient, curvature, and remainder budgets. The second grid uses

$$L(\Delta) = (1+\zeta)\langle g, \Delta\rangle + \|\Delta\|_F^2 + \frac{\kappa}{4}\|\Delta\|_F^4, \qquad \widehat{q}(\Delta) = \langle g, \Delta\rangle + \frac{2+\eta}{2}\|\Delta\|_F^2.$$

We take $\eta \in \{0.0025, 0.005, 0.01, 0.02, 0.04\}$, $\kappa \in \{0.5, 1, 2, 4, 8\}$, and $\zeta \in \{-0.01, 0, 0.01\}$, crossed with $\mathcal{R}$ and both ranks. All 1,350 combinations are retained. The reference norm solves $2t + \kappa t^3 = (1+\zeta)\|g_r\|_F$ and is capped at R. The uniform bounds are $\epsilon_g = |\zeta|\|g\|_F$, $\epsilon_H = \eta$, and $\rho(R) = 3\kappa R$, giving

$$E(R) = 2R|\zeta|\|g\|_F + \eta R^2 + \kappa R^4.$$

This follows the Hessian-based transfer theorem. The earlier 50-case quartic check in Table 4(b) retains its tighter direct function-deviation budget with quartic term $\kappa R^4/2$; its reference values are unchanged. The new grid has 750 boundary and 600 interior records. Nonzero $E(R)$ can leave wide intervals even in the interior, as Table 4(c) shows.

Solvable anisotropic stress cases. The third grid sets $B = I$ and $A = \mathrm{diag}(a_1, \ldots, a_{24})$, with $a_j = 2\chi^{(j-1)/23}$ and $\chi \in \{3, 10, 30\}$. The diagonal gradient retains the $j^{-1/2}$ profile, rescaled within each family to make $s_4 = 3.1 \times 10^{-4}$. For each condition number we use $R/\|D_d\|_F \in \{0.3, 0.5, 0.7, 0.9, 1.05\}$ and ranks $4, 8$, giving 24 boundary and six interior records. The proxy equals the target, so $E = 0$. For every multiplier $\mu \geq 0$, the sequence $\gamma_j^2/(a_j + \mu)$ is decreasing; hence the rank-r Lagrangian minimizer keeps the same first r coordinates. Its weak-dual value is

$$-\frac{1}{2}\sum_{j=1}^{r}\frac{\gamma_j^2}{a_j+\mu} - \frac{\mu R^2}{2}.$$

Choosing $\mu = 0$ when feasible, or solving $\sum_{j=1}^{r}\gamma_j^2/(a_j+\mu)^2 = R^2$, gives a feasible primal value matching this lower bound. This supplies a global reference for the constructed single-sided family; it does not solve arbitrary two-sided low-rank problems.

Numerical interpretation. Inclusion is checked with tolerance 10^{-12} in raw loss units, before rounding table entries. Full width is UB − LB after clipping the lower endpoint at zero. Positive-bound counts use LB $> 10^{-14}$ in raw loss units. The exports also provide the dense primal/dual values and an interval with a 10^{-14} numerical guard. These floating-point consistency checks are not validated-rounding machine certificates. The grids show inclusion and informativeness on the stated controlled objectives, without establishing real-network coverage.

F.1.2 Complete Candidate Ledger and Locality Audit

Table 6 reports all 32 controlled quadratic cases: 16 calibration and 16 evaluation cases, each with ten repeats and ranks $4, 8$. The split labels agree across the candidate ledger, quality audit, and reference records. All cases are retained and pass the reported interior check. The largest exact-solution norm is 0.00653534, leaving a minimum interior margin of 0.04346466 at $R = 0.05$.

The maximum absolute GGN residual across the ledger is 5.00×10^{-16}; the maximum proxy residual is 0.14233 in raw loss units. These are loss-increment residuals, not optimized rank gaps. The ledger covers $|\alpha| \leq 0.05$ with 101 points per direction; Figure 4 separately illustrates one case and repeat on a wider 121-point grid extending to $|\alpha| = 0.06$. These finite-direction checks document the controlled objectives and do not certify a uniform approximation bound or real-network locality.

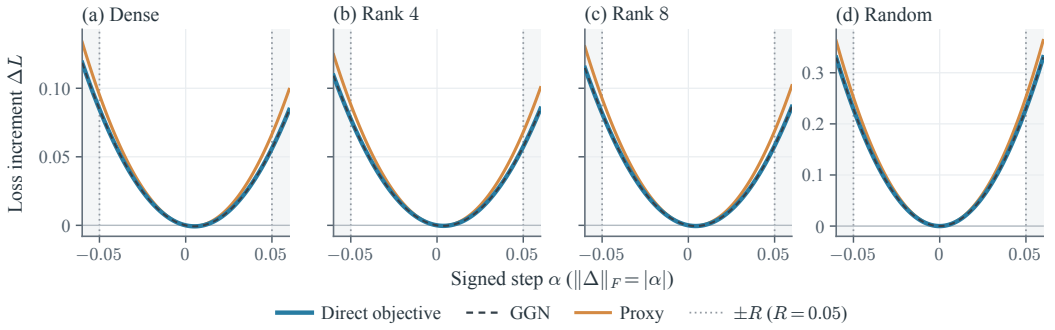

Figure 4: Controlled directional audit (S01, repeat 1; 121 points per direction). GGN (dark dashed) coincides with the direct objective (blue); the proxy is orange. Signed steps preserve the linear-term asymmetry, with $\|\Delta\|_F = |\alpha|$. Dotted lines mark $\pm R$ at $R = 0.05$; shaded regions extend to $|\alpha| = 0.06$. Panels (a–c) share a vertical scale; (d) uses a larger scale. Loss increments are unscaled.

Table 6: Complete candidate ledger for 32 controlled quadratic cases. All use $d = 24$, ranks $4, 8$, ten repeats, and four unit-Frobenius directions on $\alpha \in [-0.05, 0.05]$ with step 0.001 (101 points). All cases pass the interior check and are retained. Residuals are maximum absolute loss-increment errors in raw loss units; the margin is R minus the reported maximum exact solution norm, with $R = 0.05$.

Case	GGN residual	Proxy residual	Max. solution norm	Interior margin
Calibration cases				
S02	2.50×10^{-16}	0.02872	0.004598	0.045402
S03	1.67×10^{-16}	0.03656	0.002806	0.047194
S04	5.00×10^{-16}	0.02857	0.004079	0.045921
S06	3.33×10^{-16}	0.02197	0.004276	0.045724
S07	1.67×10^{-16}	0.02102	0.004075	0.045925
S08	8.33×10^{-17}	0.00836	0.003225	0.046775
S10	1.39×10^{-16}	0.01393	0.003739	0.046261
S11	1.11×10^{-16}	0.05980	0.003714	0.046286
S13	4.16×10^{-16}	0.04618	0.002199	0.047801
S14	1.11×10^{-16}	0.06886	0.004832	0.045168
S20	1.67×10^{-16}	0.03242	0.002991	0.047009
S21	5.55×10^{-17}	0.04797	0.005304	0.044696
S22	1.67×10^{-16}	0.03211	0.004303	0.045697
S23	1.39×10^{-16}	0.00375	0.006352	0.043648
S29	1.67×10^{-16}	0.04703	0.004798	0.045202
S30	2.22×10^{-16}	0.11132	0.004176	0.045824
Evaluation cases				
S01	2.78×10^{-16}	0.02375	0.005297	0.044703
S05	2.78×10^{-16}	0.14233	0.001635	0.048365
S09	2.22×10^{-16}	0.01927	0.005304	0.044696
S12	1.67×10^{-16}	0.02698	0.004550	0.045450
S15	2.78×10^{-16}	0.05940	0.004200	0.045800
S16	1.11×10^{-16}	0.01920	0.006285	0.043715
S17	1.39×10^{-16}	0.02353	0.004744	0.045256
S18	1.11×10^{-16}	0.02889	0.004587	0.045413
S19	8.33×10^{-17}	0.00484	0.005833	0.044167
S24	1.11×10^{-16}	0.05625	0.003505	0.046495
S25	2.78×10^{-16}	0.04280	0.003691	0.046309
S26	8.33×10^{-17}	0.01475	0.005147	0.044853
S27	8.33×10^{-17}	0.00947	0.003166	0.046834
S28	1.39×10^{-16}	0.07230	0.002798	0.047202
S31	6.94×10^{-17}	0.01036	0.004505	0.045495
S32	1.53×10^{-16}	0.03554	0.006535	0.043465

Table 7: Reference-quality audit for all 64 controlled case/rank pairs at $R = 0.05$. Each entry is the reported maximum in 10^{-4} loss units. All pairs record zero feasibility violations in 270 endpoint checks each (17,280 recorded checks in total, not necessarily distinct endpoints). Directional GGN and proxy residuals are identical for both ranks and are listed in Table 6.

	Gap change		Restart spread		Oracle gap error	
Case	$r = 4$	$r = 8$	$r = 4$	$r = 8$	$r = 4$	$r = 8$
Calibration cases						
S02	8.08×10^{-6}	3.21×10^{-4}	5.04×10^{-14}	5.09×10^{-7}	3.25×10^{-15}	1.49×10^{-9}
S03	1.30×10^{-6}	2.41×10^{-5}	1.08×10^{-15}	1.55×10^{-10}	1.08×10^{-15}	9.78×10^{-12}
S04	2.21×10^{-12}	1.30×10^{-7}	3.25×10^{-15}	6.51×10^{-15}	2.17×10^{-15}	2.17×10^{-15}
S06	9.48×10^{-5}	3.91×10^{-4}	5.53×10^{-14}	3.22×10^{-8}	2.71×10^{-15}	1.17×10^{-9}
S07	8.52×10^{-11}	4.55×10^{-8}	3.25×10^{-15}	2.71×10^{-15}	2.17×10^{-15}	2.71×10^{-15}
S08	1.57×10^{-13}	7.60×10^{-8}	1.63×10^{-15}	2.71×10^{-15}	1.36×10^{-15}	2.17×10^{-15}
S10	1.05×10^{-6}	5.43×10^{-5}	1.08×10^{-15}	4.74×10^{-10}	1.63×10^{-15}	2.07×10^{-11}
S11	2.06×10^{-9}	4.33×10^{-7}	1.08×10^{-15}	1.63×10^{-15}	1.63×10^{-15}	3.25×10^{-15}
S13	1.70×10^{-7}	1.57×10^{-4}	1.08×10^{-15}	4.47×10^{-9}	1.36×10^{-15}	1.72×10^{-11}
S14	1.73×10^{-14}	6.58×10^{-9}	2.71×10^{-15}	2.71×10^{-15}	2.71×10^{-15}	3.25×10^{-15}
S20	4.28×10^{-13}	4.59×10^{-8}	1.08×10^{-15}	2.71×10^{-15}	1.08×10^{-15}	2.17×10^{-15}
S21	1.80×10^{-13}	1.60×10^{-8}	4.34×10^{-15}	4.34×10^{-15}	4.34×10^{-15}	4.34×10^{-15}
S22	1.64×10^{-6}	1.46×10^{-4}	2.71×10^{-15}	1.26×10^{-6}	1.63×10^{-15}	1.50×10^{-10}
S23	1.84×10^{-14}	8.52×10^{-9}	5.42×10^{-15}	3.25×10^{-15}	3.25×10^{-15}	3.25×10^{-15}
S29	5.13×10^{-10}	7.82×10^{-7}	2.17×10^{-15}	4.34×10^{-15}	2.71×10^{-15}	3.25×10^{-15}
S30	1.09×10^{-7}	1.40×10^{-5}	1.63×10^{-15}	1.35×10^{-11}	2.17×10^{-15}	2.17×10^{-14}
Evaluation cases						
S01	5.58×10^{-13}	8.32×10^{-7}	2.71×10^{-15}	3.25×10^{-15}	2.17×10^{-15}	2.17×10^{-15}
S05	3.44×10^{-6}	2.69×10^{-5}	8.13×10^{-16}	8.28×10^{-9}	1.36×10^{-15}	1.59×10^{-11}
S09	1.07×10^{-12}	3.47×10^{-8}	3.79×10^{-15}	4.34×10^{-15}	3.25×10^{-15}	5.42×10^{-15}
S12	1.41×10^{-10}	1.52×10^{-7}	1.08×10^{-15}	1.63×10^{-15}	2.17×10^{-15}	2.17×10^{-15}
S15	5.60×10^{-12}	1.79×10^{-7}	2.17×10^{-15}	2.17×10^{-15}	2.71×10^{-15}	3.25×10^{-15}
S16	6.27×10^{-11}	6.00×10^{-8}	3.25×10^{-15}	2.71×10^{-15}	1.63×10^{-15}	1.08×10^{-15}
S17	1.47×10^{-7}	9.51×10^{-5}	2.17×10^{-15}	6.48×10^{-11}	2.17×10^{-15}	3.61×10^{-12}
S18	2.22×10^{-8}	1.88×10^{-4}	1.08×10^{-15}	3.05×10^{-11}	2.71×10^{-15}	9.68×10^{-13}
S19	2.06×10^{-12}	2.56×10^{-7}	4.34×10^{-15}	3.25×10^{-15}	3.25×10^{-15}	3.25×10^{-15}
S24	2.14×10^{-9}	1.17×10^{-5}	2.71×10^{-15}	2.99×10^{-13}	1.63×10^{-15}	1.03×10^{-14}
S25	1.59×10^{-10}	2.28×10^{-6}	2.17×10^{-15}	3.25×10^{-15}	5.96×10^{-15}	4.34×10^{-15}
S26	2.87×10^{-9}	3.97×10^{-6}	2.17×10^{-15}	1.79×10^{-14}	3.25×10^{-15}	3.79×10^{-15}
S27	3.33×10^{-12}	3.92×10^{-8}	1.36×10^{-15}	2.71×10^{-15}	2.71×10^{-15}	3.25×10^{-15}
S28	3.85×10^{-6}	3.58×10^{-4}	3.33×10^{-14}	2.00×10^{-6}	1.08×10^{-15}	2.48×10^{-9}
S31	9.86×10^{-12}	1.11×10^{-6}	1.08×10^{-15}	1.63×10^{-15}	1.08×10^{-15}	2.17×10^{-15}
S32	3.78×10^{-8}	1.76×10^{-5}	1.63×10^{-15}	3.67×10^{-11}	1.63×10^{-15}	1.02×10^{-12}

F.1.3 Reference Optimization and Full Prediction Results

Tables 7 and 8 audit all 64 controlled case/rank pairs. Each reference uses ten repeats at $R = 0.05$. The numerical gap compares rank-restricted and dense objective increments on the same controlled objective:

$$\widetilde{G}_r(R) = \ell(\widetilde{\Delta}_r) - \ell(\widetilde{\Delta}_{\text{dense}}), \qquad \ell(\Delta) := L_{\text{ref}}(\Delta) - L_{\text{ref}}(0).$$

The analytic references permit direct error checks in these controlled problems. Table 7 reports the supplied maximum changes, restart spreads, and errors relative to the analytic oracle gap. Every case/rank pair records zero feasibility violations in 270 endpoint checks; the 17,280 recorded checks need not be distinct endpoints.

Reference agreement and stability. Across the 64 pairs, the largest reported gap change, restart spread, and oracle gap error are 3.91×10^{-4}, 2.00×10^{-6}, and 2.48×10^{-9}, respectively, all in the table's 10^{-4} loss units. Table 8 reports repeat means and standard deviations alongside

mean discrepancies and stability changes. The maximum absolute analytic–numerical mean-gap discrepancy is 6.02×10^{-10} and the maximum mean stability change is 8.53×10^{-5} in the same reporting units. The reference values underlying Figure 2 agree with the analytic means and standard deviations to maximum absolute differences of 8.88×10^{-15} and 1.83×10^{-15}, respectively. This supports consistency of the controlled reference records.

Interpretation. The controlled results quantify numerical agreement with known analytic references. RoBERTa/SST-2 numerical stability is examined separately in Appendix F.2.1. The reported SD measures repeat variability and is not a confidence interval.

Table 8: Controlled reference records for all 64 case/rank pairs, with ten repeats per pair. All values are in 10^{-4} loss units. Dense and rank-r columns are mean objective increments: analytic and numerical values agree to four decimal places and are shown once. The analytic gap includes repeat SD, not a confidence interval. e_{gap} is the absolute difference of analytic and numerical mean gaps; δ_{stab} is the reported mean stability change.

Case	r	Dense	Rank-r	Analytic gap $\pm$ SD	Numerical gap	e_{gap}	δ_{stab}
Calibration cases							
S02	4	-5.3330	-1.4112	3.9219 ± 0.2461	3.9219	0	1.04×10^{-6}
S02	8	-5.3330	-2.3967	2.9363 ± 0.1838	2.9363	1.84×10^{-10}	4.42×10^{-5}
S03	4	-4.0221	-1.1670	2.8551 ± 0.1304	2.8551	4.44×10^{-16}	2.14×10^{-7}
S03	8	-4.0221	-1.9176	2.1046 ± 0.0951	2.1046	1.09×10^{-12}	1.05×10^{-5}
S04	4	-8.7129	-4.2365	4.4763 ± 0.1523	4.4763	0	5.18×10^{-13}
S04	8	-8.7129	-5.7892	2.9237 ± 0.1000	2.9237	4.44×10^{-16}	1.87×10^{-8}
S06	4	-5.8088	-1.5219	4.2869 ± 0.1378	4.2869	8.88×10^{-16}	1.02×10^{-5}
S06	8	-5.8088	-2.5941	3.2147 ± 0.1037	3.2147	1.27×10^{-10}	8.53×10^{-5}
S07	4	-5.4477	-2.5864	2.8612 ± 0.1270	2.8612	0	1.00×10^{-11}
S07	8	-5.4477	-3.5666	1.8811 ± 0.0840	1.8811	4.44×10^{-16}	1.62×10^{-8}
S08	4	-4.3831	-2.3077	2.0753 ± 0.1156	2.0753	4.44×10^{-16}	4.37×10^{-14}
S08	8	-4.3831	-3.0634	1.3197 ± 0.0731	1.3197	0	9.73×10^{-9}
S10	4	-2.9651	-0.7924	2.1727 ± 0.0975	2.1727	0	3.25×10^{-7}
S10	8	-2.9651	-1.3411	1.6240 ± 0.0725	1.6240	3.98×10^{-12}	1.43×10^{-5}
S11	4	-4.7198	-1.8898	2.8300 ± 0.2098	2.8300	4.44×10^{-16}	4.41×10^{-10}
S11	8	-4.7198	-2.7703	1.9495 ± 0.1440	1.9495	6.66×10^{-16}	9.62×10^{-8}
S13	4	-3.4600	-0.9761	2.4838 ± 0.0748	2.4838	0	4.57×10^{-8}
S13	8	-3.4600	-1.6198	1.8401 ± 0.0551	1.8401	3.19×10^{-12}	2.30×10^{-5}
S14	4	-6.2357	-3.5573	2.6784 ± 0.0997	2.6784	1.78×10^{-15}	4.28×10^{-15}
S14	8	-6.2357	-4.5839	1.6519 ± 0.0612	1.6519	0	1.60×10^{-9}
S20	4	-4.3862	-2.2017	2.1845 ± 0.1141	2.1845	4.44×10^{-16}	6.53×10^{-14}
S20	8	-4.3862	-2.9740	1.4122 ± 0.0742	1.4122	2.22×10^{-16}	9.91×10^{-9}
S21	4	-10.2162	-6.0317	4.1844 ± 0.1643	4.1844	1.78×10^{-15}	1.94×10^{-14}
S21	8	-10.2162	-7.6735	2.5427 ± 0.0997	2.5427	8.88×10^{-16}	2.96×10^{-9}
S22	4	-5.3035	-1.4169	3.8865 ± 0.1639	3.8865	0	4.06×10^{-7}
S22	8	-5.3035	-2.3966	2.9069 ± 0.1225	2.9069	2.67×10^{-11}	2.45×10^{-5}
S23	4	-6.2577	-3.5855	2.6723 ± 0.0882	2.6723	4.44×10^{-16}	5.26×10^{-15}
S23	8	-6.2577	-4.6118	1.6459 ± 0.0544	1.6459	0	1.74×10^{-9}
S29	4	-6.4712	-2.7166	3.7545 ± 0.2460	3.7545	0	6.04×10^{-11}
S29	8	-6.4712	-3.9159	2.5553 ± 0.1667	2.5553	0	9.42×10^{-8}
S30	4	-5.7033	-1.8467	3.8567 ± 0.1498	3.8567	8.88×10^{-16}	2.82×10^{-8}
S30	8	-5.7033	-2.9180	2.7854 ± 0.1082	2.7854	4.88×10^{-15}	4.52×10^{-6}
Evaluation cases							
S01	4	-7.4030	-3.7890	3.6141 ± 0.1417	3.6141	0	6.71×10^{-14}
S01	8	-7.4030	-5.0823	2.3207 ± 0.0926	2.3207	0	9.25×10^{-8}
S05	4	-2.7745	-0.7706	2.0039 ± 0.0698	2.0039	4.44×10^{-16}	5.76×10^{-7}
S05	8	-2.7745	-1.2862	1.4884 ± 0.0520	1.4884	3.95×10^{-12}	5.88×10^{-6}
S09	4	-8.2319	-4.4074	3.8245 ± 0.1929	3.8245	8.88×10^{-16}	2.13×10^{-13}
S09	8	-8.2319	-5.8149	2.4170 ± 0.1209	2.4170	8.88×10^{-16}	6.49×10^{-9}

Continued on the next page

Table 8: Controlled reference records (continued). All values are in 10^{-4} loss units.

Case	r	Dense	Rank-r	Analytic gap $\pm$ SD	Numerical gap	e_{gap}	δ_{stab}
S12	4	-4.6594	-1.9761	2.6833 ± 0.0792	2.6833	8.88×10^{-16}	2.98×10^{-11}
S12	8	-4.6594	-2.8373	1.8222 ± 0.0545	1.8222	4.44×10^{-16}	3.47×10^{-8}
S15	4	-5.4596	-2.6522	2.8074 ± 0.1290	2.8074	8.88×10^{-16}	9.38×10^{-13}
S15	8	-5.4596	-3.6262	1.8334 ± 0.0833	1.8334	6.66×10^{-16}	2.64×10^{-8}
S16	4	-5.1868	-2.4248	2.7620 ± 0.1159	2.7620	0	7.29×10^{-12}
S16	8	-5.1868	-3.3620	1.8248 ± 0.0770	1.8248	0	1.47×10^{-8}
S17	4	-6.1070	-1.8706	4.2364 ± 0.2206	4.2364	0	2.01×10^{-8}
S17	8	-6.1070	-3.0158	3.0912 ± 0.1602	3.0912	5.88×10^{-13}	1.81×10^{-5}
S18	4	-5.3688	-1.7834	3.5854 ± 0.1663	3.5854	0	6.06×10^{-9}
S18	8	-5.3688	-2.7938	2.5750 ± 0.1199	2.5750	9.81×10^{-14}	2.03×10^{-5}
S19	4	-8.6318	-4.1789	4.4529 ± 0.1752	4.4529	8.88×10^{-16}	4.47×10^{-13}
S19	8	-8.6318	-5.7207	2.9112 ± 0.1142	2.9112	4.44×10^{-16}	3.08×10^{-8}
S24	4	-4.4950	-1.5392	2.9558 ± 0.1600	2.9558	0	9.46×10^{-10}
S24	8	-4.4950	-2.3845	2.1105 ± 0.1136	2.1105	1.33×10^{-15}	2.35×10^{-6}
S25	4	-7.1492	-3.0903	4.0589 ± 0.1707	4.0589	8.88×10^{-16}	2.02×10^{-11}
S25	8	-7.1492	-4.4070	2.7422 ± 0.1164	2.7422	4.44×10^{-16}	3.09×10^{-7}
S26	4	-7.5106	-3.0063	4.5043 ± 0.2501	4.5043	0	4.24×10^{-10}
S26	8	-7.5106	-4.4078	3.1028 ± 0.1723	3.1028	0	6.27×10^{-7}
S27	4	-4.7404	-2.3424	2.3981 ± 0.0712	2.3981	4.44×10^{-16}	4.65×10^{-13}
S27	8	-4.7404	-3.1826	1.5578 ± 0.0469	1.5578	0	8.19×10^{-9}
S28	4	-2.8021	-0.7165	2.0856 ± 0.0917	2.0856	8.88×10^{-16}	5.58×10^{-7}
S28	8	-2.8021	-1.2313	1.5708 ± 0.0695	1.5708	6.02×10^{-10}	5.86×10^{-5}
S31	4	-4.5799	-2.1050	2.4750 ± 0.1696	2.4750	0	2.81×10^{-12}
S31	8	-4.5799	-2.9351	1.6448 ± 0.1130	1.6448	2.22×10^{-16}	1.42×10^{-7}
S32	4	-5.0441	-1.7476	3.2965 ± 0.1465	3.2965	0	4.90×10^{-9}
S32	8	-5.0441	-2.6957	2.3484 ± 0.1051	2.3484	1.02×10^{-13}	3.06×10^{-6}

F.1.4 Controlled Validation Costs

Table 9 reports the computational cost of the controlled validation pipeline after gradients and moment factors are available. These measurements characterize numerical auditing and reference construction rather than the end-to-end cost of the neural-network diagnostic. The workload comprises 32 cases, ten perturbations per case, ranks $4, 8$, and all three starts. The exported subtotal has median 8.0901 s and interquartile range $[8.0655, 8.1346]$ s. Its maximum process memory high-water mark is 47.625 MiB. Whitening and spectrum computation take a median 0.0364 s, the condition/directional audit 1.2169 s, and reference optimization with the rank sweep 6.6945 s. Shared work and rank-specific solves are identified in panel (b).

Accounting scope. Gradients and moment factors are supplied by construction and their construction is excluded from the subtotal. The subtotal need not equal the sum of stage medians. Process high-water marks are not additive stage allocations or GPU-memory readings. The 13,440 checks counted in the timing workload and the 17,280 endpoint checks in Table 7 use separate accounting and are not combined. Real-network diagnostic overhead, including gradient and factor construction, is reported in Appendix F.3 (Table 13).

F.2 Fixed-Rank Prediction and Sensitivity

F.2.1 RoBERTa/SST-2 Candidate and Reference Records

This section summarizes the real-network evaluation in Section 6.2. The exported records retain all 24 candidates, 48 module/rank summaries, 480 repeat-level references, 15,120 trajectory records, and 240 method-specific predictions. Detailed records include excluded candidates and missing scores; only the ten retained evaluation modules contribute to Table 2.

Table 9: Controlled validation costs for 32 cases, ten perturbations per case, both ranks $4, 8$, and all three starts. Times are medians with 25th–75th percentiles, in seconds. Memory is the reported maximum process high-water mark (HWM), in MiB; it is not GPU memory or an incremental stage allocation. The subtotal excludes gradient and token-moment construction and is reported directly, without summing stage quantiles.

(a) Measured timing and process memory

Stage	Median (s)	$[Q_{25}, Q_{75}]$ (s)	HWM (MiB)
Input loading	0.1456	[0.1446, 0.1459]	45.000
Whitening / spectrum	0.0364	[0.0356, 0.0366]	45.203
Condition / directional audit	1.2169	[1.2104, 1.2180]	47.344
Reference optimization / rank sweep	6.6945	[6.6731, 6.7343]	47.625
Measured subtotal	8.0901	[8.0655, 8.1346]	47.625

(b) Workload and sharing across ranks

Stage	Operations / inputs	Sharing
Input loading	448 model arrays; 2880 initial states; 320 directions; 10 fixed calibration maps	Yes
Whitening / spectrum	64 eigendecompositions; 960 SVDs; both ranks; apply fixed calibration	Yes
Condition / directional audit	64 true-factor eigendecompositions; 320 SVDs; 129280 q evaluations	Shared preparation; rank-specific directions
Reference optimization / rank sweep	61440 dense steps; 491520 block iterations; 13440 checks	Dense reused; rank solves separate
Measured subtotal	32 cases $\times$ 10 perturbations; $r = 4, 8$; all three starts	No double-counting

Table 10: RoBERTa/SST-2 candidate screening at $R = 0.05$. Panel (a) includes all candidates; panel (b) preserves all four exclusions and the corresponding exported diagnostics. All candidates record zero norm violations; full-factor scores are missing only for L10-V. Detailed module-level records accompany the exported records.

(a) Candidate coverage

Split	Candidates	Excluded	Retained	Retained layer groups
Calibration	12	2	10	6
Evaluation	12	2	10	6

(b) Excluded candidates

Case	Reported exclusion	Exported diagnostic
L03-Q	Locality	Relative directional residual 0.0540
L04-Q	Interior	Required radius 0.0590
L07-V	Restart stability	Maximum spread 0.2244×10^{-4}
L10-V	Missing score	Full-factor score unavailable

Candidate coverage. Each of the 12 numbered layer groups contains a Q and a V module. Odd groups are assigned to calibration and even groups to evaluation. Each split retains ten of its 12 candidates, spanning six layer groups. Table 10 reports the coverage and all exclusions. The missing full-factor score for L10-V is preserved as missing; it is not filled with an estimated value.

Reference construction and stability. For module c, repeat j, checkpoint k, and restart u, let $\ell_{d,cju}(k)$ and $\ell_{r,cju}(k)$ be the recorded dense and rank-r objective increments. The numerical gap is reproduced by

$$\widetilde{G}_{r,cj}(k) = \min_{u\in\{1,2,3\}} \ell_{r,cju}(k) - \min_{u\in\{1,2,3\}} \ell_{d,cju}(k).$$

Table 11: RoBERTa/SST-2 reference summary for the 20 retained modules. All values are in 10^{-4} loss units. Each row contains ten modules with ten repeats each. Gap range spans module-level mean gaps; the remaining columns are maxima of module-level SD, absolute 400-to-800-step gap change, and paired-restart gap spread at step 800. Repeat SD is not a confidence interval. Detailed records include all 24 candidates, including exclusions.

Split	r	Mean-gap range	Max. SD	Max. extension	Max. restart spread
Calibration	4	[2.8361, 4.3900]	0.1212	0.0089	0.0426
Calibration	8	[2.1351, 3.3594]	0.1002	0.0092	0.0379
Evaluation	4	[1.9364, 4.3080]	0.1154	0.0092	0.0379
Evaluation	8	[1.4423, 3.6128]	0.0956	0.0120	0.0389

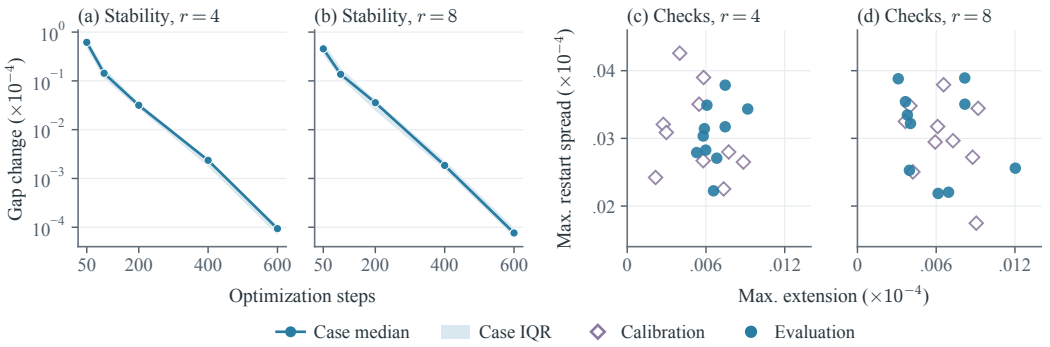

Figure 5: Optimization diagnostics for the 20 retained RoBERTa/SST-2 modules. Panels (a,b) show absolute differences from each repeat's 800-step reference gap at earlier checkpoints, for $r = 4, 8$. Each module contributes its mean over ten repeats; lines and bands show the median and interquartile range across modules. Panels (c,d) compare each module's maximum 400-to-800-step gap change with its maximum paired-restart gap spread at step 800; diamonds and circles denote calibration and evaluation modules. All errors are in 10^{-4} loss units. Bands describe module variation, not confidence intervals; agreement with the terminal reference does not prove optimality.

Primary and reference gaps use $k = 400$ and $k = 800$. Extension change is $|\widetilde{G}_{r,cj}(800) - \widetilde{G}_{r,cj}(400)|$; restart spread is the range of $\ell_{r,cju}(800) - \ell_{d,cju}(800)$ across paired restart indices. These definitions reproduce the exported statistics to rounding precision. Table 11 summarizes the retained modules; Figure 5 shows their checkpoint and restart diagnostics. No exported update norm exceeds R.

Calibration and aggregation. Affine least-squares fits using only the ten retained calibration modules at each rank reproduce all available calibrated predictions; the constant baseline equals the corresponding calibration reference mean. The ten evaluation modules belong to six layer groups, and their repeats do not increase the number of independent evaluation groups. We report descriptive point estimates without significance claims. Error bars in Figure 3 show repeat standard deviations, and bands in Figure 5 show variation across modules.

F.2.2 Fixed-Rank Prediction Sensitivity to Re-inclusion

We reassess the exclusion based only on the original interior-feasibility screen while retaining exclusions for locality, restart instability, and missing scores. The re-included module is L04-Q in each task. Its recorded required interior radius is 0.059, exceeding the evaluation radius $R = 0.05$. Failure of this sufficient screen alone does not establish that the constrained optimum lies on the boundary. We distinguish the re-included interior-screen failure from the original screen-passing cohort. Table 12 reports full-factor predictions under the frozen, rank-specific SST-2 calibration, without refitting on the re-included module or on MRPC.

Table 12: Fixed-rank re-inclusion sensitivity for full-factor whitening. MAE is computed against module-level numerical-reference means using frozen SST-2 calibration and is reported in 10^{-4} loss units. The combined row is the module-count-weighted mean of the two preceding strata. The re-included stratum is L04-Q only. An interior-screen failure does not establish an active constrained boundary.

Task	Evaluation stratum	Modules	Rank-4 MAE	Rank-8 MAE
SST-2	Original interior-screen pass	10	0.221	0.177
	Re-included interior-screen failure	1	0.435	0.112
	Combined	11	0.241	0.171
MRPC	Original interior-screen pass	20	0.204	0.227
	Re-included interior-screen failure	1	0.754	0.837
	Combined	21	0.230	0.256

Re-inclusion changes the SST-2 evaluation cohort from ten to 11 modules and the MRPC cohort from 20 to 21. SST-2 retains ten separate calibration modules; MRPC contributes no calibration modules. Both tasks start with 24 candidates and retain 21 after re-inclusion: L03-Q remains excluded for locality, L07-V for restart instability, and L10-V for a missing score. The original SST-2 main-table and figure cohorts remain unchanged. Ten repeats per module/rank are averaged before computing prediction errors; they do not increase the independent module count. Layer-group counts remain six for SST-2 evaluation and 12 for MRPC.

SST-2 combined MAE changes from 0.221 to 0.241 at rank 4 and from 0.177 to 0.171 at rank 8. MRPC combined MAE increases from 0.204 to 0.230 and from 0.227 to 0.256, respectively (all in 10^{-4} loss units). Thus the effect of re-inclusion differs across tasks and ranks; each re-included stratum contains only one module. This is a sensitivity analysis of prediction errors under the original screening rule; it does not test boundary-interval inclusion or establish active-boundary status.

F.3 Rank-Upgrade Evaluation

Table 3 correlates each predictor with the module mean of the paired difference $L(\Delta_4^{\text{continue}}) - L(\Delta_8^{\text{upgrade}})$. The reported radius is $R = 0.05$; the comparator is defined relative to the original expansion center and total update. The spectral difference uses the same proxy factors and gradient at both ranks. The first-order comparator uses the linearized objective's predicted rank-increase benefit $R(\|\widehat{g}_8\|_F - \|\widehat{g}_4\|_F)$; it requires no additional training. The current tail s_4 is a ranking baseline, rather than a loss-scale prediction of the $4 \to 8$ gain. The upgrade cohorts contain ten SST-2 and 20 MRPC modules; the L04-Q re-inclusion in Appendix F.2.2 does not add a paired-upgrade observation.

These correlations concern finite-budget endpoints, not numerical-reference capacity gains or certified optimal values. They assess predictive ordering, not training efficiency at matched compute. The reported correlations are descriptive point estimates; repeats within a module and modules within a layer are not treated as independent replications for statistical significance.

Diagnostic cost. Table 13 reports an initial timing measurement for the rank-upgrade diagnostic. The measured stages cover gradient and full-factor collection, eigendecomposition and whitening, and SVD with score evaluation. One proxy supplies s_4, s_8, and $t_{4,8} = s_4 - s_8$, without separate calibration at each rank. The reported total is 6.6 s, compared with 90 s for a 400-step continuation. Their ratio, $6.6/90 \approx 7.3\%$, describes diagnostic overhead relative to that continuation; it does not establish a training advantage at matched compute. These timings are preliminary point measurements and do not characterize variation across runs or workloads.

F.4 Protocol Sensitivity and Supporting Observations

F.4.1 Sensitivity to the Comparison Protocol

The `roberta-base`/SST-2 experiments (Liu et al., 2019; Socher et al., 2013; Wang et al., 2018) illustrate that endpoint gaps depend on how update size is handled. In the reported `layer10`

Table 13: Initial timing measurement for the rank-upgrade diagnostic. The total includes gradient and curvature-factor collection. The overhead is relative to the reported 400-step continuation.

Stage	Time
Gradient and full-factor collection	6.3 s
Eigendecomposition and whitening	0.25 s
SVD and evaluation of $s_4, s_8, t_{4,8}$	0.05 s
Diagnostic total	6.6 s
400-step continuation	90 s
Diagnostic overhead relative to continuation	7.3%

Table 14: Sensitivity of `layer10` eval gaps to the comparison protocol. Arrows summarize rank 1 to rank 16; the raw-endpoint row gives the reported range. These are observed loss differences, not certified optimal-value gaps.

Reported protocol	Update-size handling	Eval gap
Raw endpoints	Dense norm ≈ 0.5; LoRA norms $0.1395 \to 0.0449$	0.01294–0.01298
Common-radius projection	Endpoint rescaling/projection to the reported common radius	$(3.98 \to 1.39) \times 10^{-4}$
Ball-constrained optimization	Both updates optimized subject to the same norm cap	$(4.1 \to 1.3) \times 10^{-4}$

comparison, raw dense and LoRA endpoints had different norms and a nearly flat eval-gap profile across ranks. Endpoint projection and direct ball-constrained optimization produced decreasing profiles (Table 14). These observations motivate explicit comparator control; recovering a monotone trend does not itself establish correctness or optimality.

The theoretical comparator uses the same feasible ball, not necessarily equal norms at the two solutions. An endpoint projected to the sphere need not solve the ball-constrained problem. Table 14 measures protocol sensitivity through observed endpoints, rather than oracle values.

F.4.2 Rank-Indexed Observations

At the reported radius $R = 0.05$, `layer6` shows the same ordering of spectral scores and observed gaps across $r \in \{1, 2, 4, 8, 16\}$ (Figure 6). The `layer10` curves also show decreasing observed gaps with increasing rank (Appendix F.4.5, Figure 7). The experiments reported ten seeds; complete mean and standard-deviation values are in Appendix Tables 17 and 18.

Table 15 summarizes descriptive associations using the five rank means. Each listed Spearman coefficient is one; the Pearson coefficients quantify linear association, with a logarithmic score transform only for Llama. These statistics do not establish accurate loss-scale prediction or predictive value beyond rank itself. The `Llama-3.1-8B` (Grattafiori et al., 2024) comparison evaluates ordering; its full table, plot, and directional diagnostics are in Appendix F.4.7.

Table 15: Descriptive score–gap association across the five reported rank means. All settings reported ten seeds. Pearson and Spearman entries are train/eval pairs, computed from the mean tables. Llama uses log tail; RoBERTa uses the untransformed reported score. Correlation does not validate absolute score normalization or finite-sample coverage.

Setting	Radius	Score transform	Pearson (train/eval)	Spearman (train/eval)
RoBERTa `layer6`	0.05	Identity	0.941/0.919	1.000/1.000
RoBERTa `layer10`	0.05	Identity	0.963/0.947	1.000/1.000
Llama `down_proj`	0.02	Logarithm	0.953/0.858	1.000/1.000

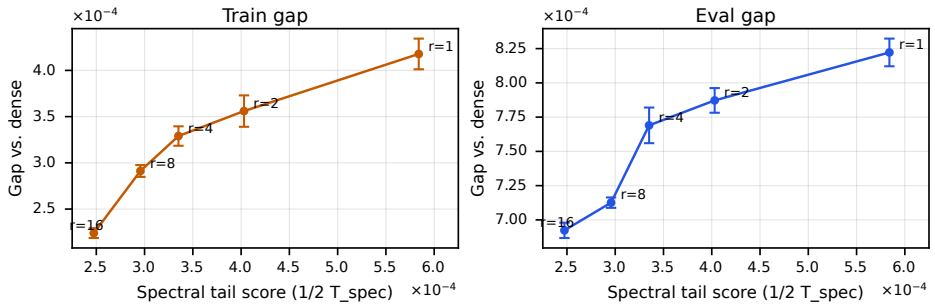

Figure 6: `layer6` observations at $R = 0.05$: the spectral score versus common-radius train and eval gaps across ranks. Error bars show seed standard deviations. These curves illustrate within-module association; they do not establish oracle-gap coverage or predictive value beyond rank.

Table 16: RoBERTa/SST-2 directional screening observations. Errors are reported means over $|\alpha| \leq 0.02$; the symmetric-radius statistic is reported separately. These descriptive statistics are not uniform locality bounds.

Module	Dense r_{sym}	Dense err.	Dense asym.	Rank-4 $(r_{\mathrm{sym}}, \mathrm{err})$	Rank-16 $(r_{\mathrm{sym}}, \mathrm{err})$	Role
`layer6`	0.05	3.38×10^{-4}	2.05×10^{-5}	$(0.05,\ 2.63 \times 10^{-4})$	$(0.05,\ 6.87 \times 10^{-5})$	first module
`layer10`	0.05	5.65×10^{-5}	3.42×10^{-5}	$(0.05,\ 1.07 \times 10^{-3})$	$(0.05,\ 1.32 \times 10^{-3})$	second module

These rank-indexed comparisons assess association on each score's reported scale. The fixed-rank results in Section 6.1 assess predictive information beyond this rank-indexed association on the evaluated cases.

Supporting records. These tables provide the rank-indexed observations and directional diagnostics supporting Appendices F.4.1 and F.4.2. Directional checks describe the evaluated directions and do not supply uniform approximation guarantees.

F.4.3 RoBERTa Directional Screening Records

Table 16 reports directional observations for two RoBERTa/SST-2 modules. The error summaries over $|\alpha| \leq 0.02$ and the reported symmetric-radius statistic $r_{\mathrm{sym}} = 0.05$ are separate descriptive quantities. These finite-direction observations do not establish a uniform guarantee over the Frobenius ball or a candidate-level acceptance rate.

The comparator summaries in Table 14 describe the sensitivity of observed gaps to the comparison protocol.

F.4.4 Layer6 Full Rank Means

Table 17 reports losses and spectral scores from ten seeds at the reported radius $R = 0.05$. The listed gaps describe trained endpoints and are not identified with exact ball-constrained optimal values.

The corresponding score–gap curves are shown in Figure 6 in Appendix F.4.2.

F.4.5 Layer10 Full Rank Means

The tail-energy values for the second RoBERTa module are used only to describe within-table rank ordering. Their numerical differences from the loss gaps are not interpreted as calibrated prediction errors, and the listed tail energy is not identified with the half-tail score s_r. Agreement in ordering does not establish agreement in magnitude or a predictive advantage over rank.

The directional observations include reported mean/max relative errors of $5.65 \times 10^{-5}/1.47 \times 10^{-4}$ for the dense direction, $1.07 \times 10^{-3}/4.00 \times 10^{-3}$ for trained rank 4, $1.32 \times 10^{-3}/5.12 \times 10^{-3}$ for trained rank 16, and $2.59 \times 10^{-2}/5.87 \times 10^{-2}$ for a random rank-4 direction. These observations

Table 17: `layer6` observations at the reported radius $R = 0.05$ over ten seeds. Gaps are means $\pm$ seed standard deviations; these deviations measure variability across seeds, not uncertainty bounds on the oracle gap.

Rank r	Train gap vs dense	Eval gap vs dense	Spectral score s_r
1	$(4.178 \pm 0.166) \times 10^{-4}$	$(8.222 \pm 0.101) \times 10^{-4}$	5.840×10^{-4}
2	$(3.560 \pm 0.170) \times 10^{-4}$	$(7.872 \pm 0.090) \times 10^{-4}$	4.030×10^{-4}
4	$(3.290 \pm 0.105) \times 10^{-4}$	$(7.690 \pm 0.130) \times 10^{-4}$	3.351×10^{-4}
8	$(2.912 \pm 0.064) \times 10^{-4}$	$(7.126 \pm 0.038) \times 10^{-4}$	2.957×10^{-4}
16	$(2.240 \pm 0.053) \times 10^{-4}$	$(6.924 \pm 0.056) \times 10^{-4}$	2.474×10^{-4}

Table 18: `layer10` observations at the reported radius $R = 0.05$ over ten seeds. Tail-energy and gap entries use their reported 10^{-4} scaling; gaps are mean $\pm$ seed standard deviation. Tail energy is used only for within-table ordering.

Rank r	Tail energy	Train gap vs dense	Eval gap vs dense	Observed rank trend
1	15.243	2.867 ± 0.360	3.980 ± 0.430	largest tail and gap
2	8.359	2.673 ± 0.330	3.524 ± 0.390	decreasing from rank 1
4	4.488	2.270 ± 0.290	2.633 ± 0.340	monotone decrease continues
8	2.251	2.155 ± 0.180	1.943 ± 0.210	eval gap sharply reduced
16	0.754	1.933 ± 0.140	1.390 ± 0.160	smallest tail and gap

accompany the reported radius 0.05 and describe only the tested directions; they are not uniform approximation bounds or certificates for the local loss gap.

The corresponding rank-gap curves are shown in Figure 7.

F.4.6 Directional Proxy-Fit Observations

For a tested direction, the reported proxy-fit statistic is

$$e_{\mathrm{sep}}(\Delta) = \frac{|\left\langle \Delta, (\widehat{\mathcal{G}}_0 - \widehat{S}_{\mathrm{mom}})[\Delta] \right\rangle|}{|\left\langle \Delta, \widehat{\mathcal{G}}_0[\Delta] \right\rangle| + \epsilon_{\mathrm{num}}}, \qquad \widehat{S}_{\mathrm{mom}}[\Delta] = \widehat{B}\Delta\widehat{A}.$$

The values below summarize the reported direction sets under this residual definition. Their interpretation is restricted to those directional calculations: no finite-direction maximum is used as an operator-norm bound or a uniform transfer budget for the local gap. These ratios alone do not validate the complete shared-weight Jacobian construction.

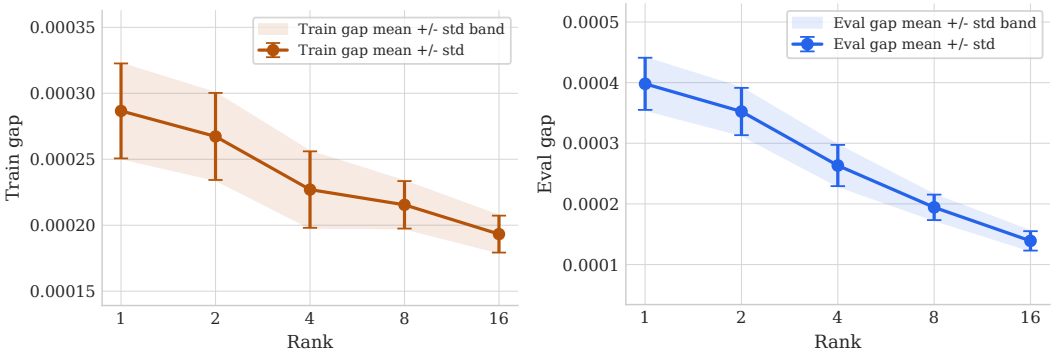

Figure 7: `layer10` train and eval endpoint gaps by rank at the reported radius $R = 0.05$. Error bars and shaded bands show seed standard deviations over ten seeds.

Table 19: Empirical directional quadratic-form residuals on `layer6`. Quantiles summarize the reported direction sets and are descriptive fit statistics, not operator-norm bounds.

Direction family	# dirs	Median	90%	Max	Direction tested
Dense step direction	1	0.051	0.051	0.051	Checks the dense comparator direction used by the local quadratic target.
LoRA rank directions	5	0.042	0.055	0.061	Checks trained rank-restricted directions for $r \in \{1, 2, 4, 8, 16\}$.
Top spectral directions	16	0.034	0.047	0.053	Checks the directions that dominate the ridge-whitened spectral tail.
Random low-rank directions	64	0.030	0.044	0.057	Stress-tests nearby low-rank subspaces not selected by training.

Table 20: Llama directional locality observations at the reported radius $R = 0.02$. These statistics describe the tested directions and do not establish a uniform locality guarantee.

Direction	Mean relative error	Asymmetry	Symmetric radius	Status
Dense direction	0.217323	0.104088	0.02	diagnostic-only
LoRA rank-4 direction	0.080796	0.009318	0.02	diagnostic-only
LoRA rank-16 direction	0.066025	0.079311	0.02	diagnostic-only

F.4.7 Llama Exploratory Observations

A ten-seed `Llama-3.1-8B` probe uses the SST-2 one-token slice `train[512:1024]` using `model.layers.31.mlp.down_proj` at $R = 0.02$. The raw ridge-tail values are used only for within-setting ordering. Their scale is not used to predict absolute loss gaps or instantiate the finite-sample bounds. The logarithmic score plot displays this ordering without asserting agreement on the loss scale.

Table 21: Llama raw ridge-tail and endpoint loss gaps at the reported radius $R = 0.02$ over ten seeds. Gaps are means $\pm$ seed standard deviations. Raw scores support ordering comparisons only; they are not calibrated loss predictions or certified bounds.

Rank r	Raw ridge-tail	Train gap vs dense	Eval gap vs dense
1	1575.147	0.245 ± 0.078	0.154 ± 0.030
2	42.884	0.180 ± 0.048	0.145 ± 0.035
4	21.452	0.128 ± 0.030	0.110 ± 0.032
8	10.196	0.073 ± 0.020	0.066 ± 0.024
16	2.888	0.034 ± 0.014	0.039 ± 0.016

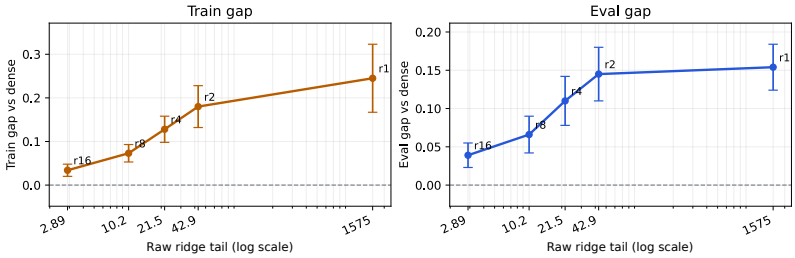

Figure 8: Llama exploratory tail–gap plot. The logarithmic raw-score axis displays within-setting ordering. The raw score is a ranking diagnostic, not a calibrated loss prediction or a certified bound.