WHEN SPECTRAL ACCURACY CONTROLS LOW-RANK UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivalent factor coordinates of a low-rank product lead to different Euclidean factor-gradient updates, and a spectral curvature objective can be precisely optimized over such coordinates while the resulting product update remains undetermined. We characterize when the spectral objective's accuracy controls the product response. A latent coverage matrix is formed by a spectral optimality certificate for convex monotone spectral criteria on a bounded factor-coordinate budget. Compared with the current gradient, this matrix provides an explicit upper bound on the response error that allows for repeated eigenvalues, optima on the budget boundary, and rank-deficient factors. Full-support criteria identify the response under gradient compatibility. For the largest eigenvalue, at interior optimal points with a simple top eigenvalue, the criterion is exact: the response is fixed if and only if the gradient has no leakage into the free subspace. When identification fails, we provide the extremal first-order response pairing for a certified spectral block in closed form. On reduced-rank digit classifiers, coordinates whose relative spectral excess is certified to be less than 7×10^{-6} still change the product response by as much as 4.9% under squared loss. Under softmax cross-entropy, a measured relative objective deviation of at most 7×10^{-6} from a solved reference allows changes of up to 54%. Each evaluated certificate contains the observed change; under squared loss, projecting the gradient onto the predicted zero-leakage subspace eliminates the effect to numerical precision, and the block extremum improves the response pairing over both uniform endpoints exactly when predicted.

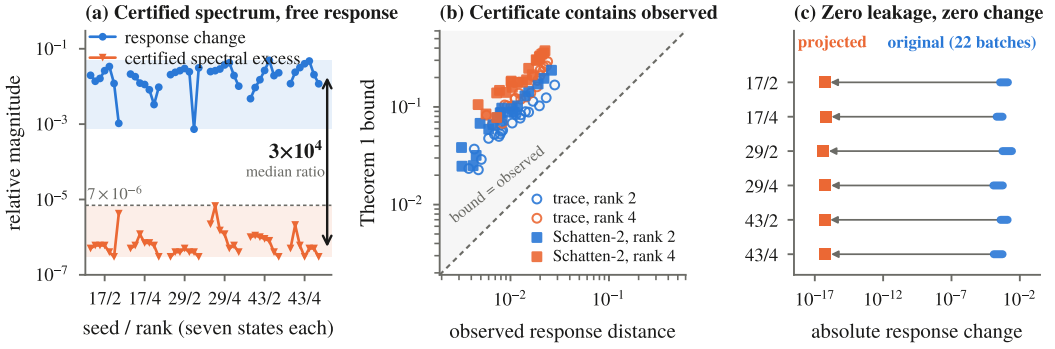


Figure 1: **Spectral accuracy does not identify the update; coverage does. (a)** On 42 controlled states, the rationally certified relative spectral excess of the reference and of both uniform endpoints on the free subspace stays below 7×10^{-6}, while the larger endpoint response change relative to the reference spans 0.07%–4.9%; state by state, this change exceeds that state's upward-rounded certified excess by a median factor of 3×10^4 and by more than 200 in every state. **(b)** Theorem 1 on the full-support criteria, for all 84 trace and Schatten-2 comparisons; every point lies above the diagonal. **(c)** For each of six stationary heads, projecting its 22 fixed batch gradients onto the zero-leakage subspace (Proposition 5) collapses the absolute response change from 1.2×10^{-4}–2.7×10^{-3} to below 10^{-16}. In (a), each seed/rank group has seven correlated states; in (c), each row is one stationary head. Apart from the certified excess, all quantities are floating-point diagnostics.

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.

1 INTRODUCTION

A low-rank product UV has many factorizations: the pair $(UG, G^{-1}V)$ represents the same matrix for every $G \in \mathrm{GL}(k)$, yet its curvature and its gradient step depend on G (Dinh et al., 2017; Arora et al., 2018). In practice one optimizes a criterion: optimal scalings minimize a spectral quantity such as the largest singular value or a condition number over a scaling group (Balakrishnan & Boyd, 1995; Ghadimi et al., 2025), and adapter methods balance, refactorize or remove the factor coordinates (Castin et al., 2026; Zhang et al., 2025; Yen et al., 2025). These guarantees concern a chosen criterion or update rule. They leave a question about the set of coordinates: *if a spectral criterion used to choose coordinates is solved accurately, is the product update determined accurately as well?* It need not be. A consistent refactorization rule can coexist with different product responses among near-minimizers of a curvature criterion: the criterion sees the curvature spectrum, the update sees the gradient, and the condition linking them is gradient-relative. This paper supplies that condition.

A two-coordinate example. A curvature criterion constrains some latent directions (the inner dimension $\mathbb{R}^k$ of UV) more than others; an update moves along those its gradient selects. Take a fixed quadratic loss at $U = V = I_2$ whose product gradient is $D = E_{22}$. The largest-eigenvalue objective and the trace objective are both minimized at I. Moving to $T = \mathrm{diag}(1, 2)$ leaves the largest eigenvalue exactly optimal and changes the product response by $1/2$ in Frobenius norm; the trace objective pays a gap of $1/4$ for the same move, which yields a finite response bound (Appendix A.6).

On the digit classifiers of Section 6, coordinates whose relative spectral excess is certified below 7×10^{-6} still change the product response by up to 4.9% under squared loss. Under softmax cross-entropy, a measured relative objective deviation of at most 7×10^{-6} from a solved reference allows changes of up to 54%. Even where a certificate applies, the two tolerances differ in scale: for our sufficient certificate, a 1% response guarantee needs a relative spectral excess of order 10^{-6} at an exact reference, which solver iterates reach and certify (Section 6.1). A small spectral optimality gap is therefore not by itself a guarantee about the step; the missing quantity is gradient-relative spectral coverage.

Contributions.

- **A gradient-relative certificate.** For convex monotone spectral criteria, Theorem 1 turns spectral optimality information into a $k \times k$ coverage matrix and a coverage coefficient $\kappa_\Phi(D)$ that converts objective error into product-response error on the original bounded budget. Repeated spectra, budget boundaries and rank-deficient factors are allowed. Full-support criteria identify the response under gradient compatibility (Corollary 2).
- **An exact boundary for $\lambda_{\max}$.** At an interior optimum with a simple top eigenvalue, the largest-eigenvalue optimizer set has local dimension $(k - r)(k - r + 1)/2$, with r the active latent Gram rank, and spectral accuracy fixes the response if and only if the gradient has no leakage into that free subspace (Propositions 4 and 5).
- **The remaining freedom, quantified.** When identification fails, Theorem 6 gives the extremal gradient–response pairing over a certified spectral block in closed form, with its exact gain over uniform rescaling; for a loss gradient it is the block's largest first-order decrease.

2 THE COORDINATE-SELECTION PROBLEM

Let $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{k\times n}$, and $M = UV$, with Frobenius inner products on the factor spaces. Equivalent coordinates are $(UG, G^{-1}V)$, $G \in \mathrm{GL}(k)$, a change of basis of the *latent* space $\mathbb{R}^k$; we use *gauge* for this choice of factor coordinates. Throughout, a fixed PSD factor operator H supplies the curvature and a fixed product gradient D the update; they need not come from the same batch.

The quantity to be controlled. Write $S = GG^\top$. One simultaneous ordinary factor-gradient step with a common step size η gives

$$\Delta M_S = -\eta \mathcal{P}_S(D) + \eta^2 D M^\top D, \qquad \mathcal{P}_S(D) = USU^\top D + DV^\top S^{-1} V. \tag{1}$$

The quadratic term is coordinate-independent (Arora et al., 2018; Zhang et al., 2025); hence $\|\Delta M_S - \Delta M_T\|_F = \eta\|\mathcal{P}_S(D) - \mathcal{P}_T(D)\|_F$. Thus controlling the *product response* controls the actual step, not just its linear approximation. Coordinate-wise adaptive updates such as Adam are not gauge-equivariant (Singh, 2026) and lie outside this map. Rank-deficient factors are allowed: we study their specified $\mathrm{GL}(k)$ orbit, not all factorizations of M.

The coordinate budget and spectral objective. The bound $\|G\|_{\mathrm{op}}, \|G^{-1}\|_{\mathrm{op}} \le R, 1 \le R < \infty$, is equivalent to $S \in \mathcal{Q}_R = \{S \in \mathbb{S}^k_{++} : R^{-2}I \preceq S \preceq R^2 I\}$. On factor variations define $\mathcal{T}_G(X,Y) = (XG, G^{-1}Y)$ and $\mathcal{W}(S)(X,Y) = (XS^{-1}, SY)$. The curvature transforms as $H_G = \mathcal{T}_G^{-*} H \mathcal{T}_G^{-1}$. Its spectrum equals that of $K_H(S)$ below, including zero multiplicities:

$$\begin{aligned} K_H(S) &= \mathcal{W}(S)^{1/2} H \mathcal{W}(S)^{1/2}, \qquad & f_H(S) &= \lambda_{\max}(K_H(S)), \\ \mathfrak{F}_R(H) &= \min_{S \in \mathcal{Q}_R} f_H(S), & \mathrm{Opt}_R(H) &= \arg\min_{S \in \mathcal{Q}_R} f_H(S). \end{aligned} \tag{2}$$

For $\mathfrak{F}_R(H) > 0$, the *relative spectral excess* of S is $f_H(S)/\mathfrak{F}_R(H) - 1$. Section 3 treats more general spectral criteria on this same budget. A factor-balanced gauge (Du et al., 2018) equalizes the factor Grams, $(UG)^\top(UG) = (G^{-1}V)(G^{-1}V)^\top$; an interior largest-eigenvalue optimum with a simple top mode instead equalizes the Grams of the two blocks of that curvature mode (Theorem 3).

Core notation. H is the supplied curvature, $F_\Phi(T)$ a spectral coordinate objective with T relative to a reference, P_Φ the latent coverage matrix of a spectral certificate, Q_D the latent directions used by the gradient D, $\kappa_\Phi(D)$ the least κ with $Q_D \preceq \kappa \mathsf{P}_\Phi$, and $\mathcal{P}_T(D)$ the product response (Table 3).

Which curvature is supplied? For a product loss ℓ, put $E = \nabla\ell(M)$, $\mathcal{C} = \nabla^2\ell(M)$, and $J(X,Y) = XV + UY$. By the chain rule, the exact Hessian of $(U,V) \mapsto \ell(UV)$ is $H_{\mathrm{full}} = H_{\mathrm{pb}} + \mathcal{B}_E$, with $\mathcal{B}_E(X,Y) = (EY^\top, X^\top E)$ and the product-Hessian pullback $H_{\mathrm{pb}} = J^*\mathcal{C}J$. For convex ℓ, H_{pb} is PSD at every iterate; H_{full} need not be. Our guarantees apply to whichever PSD H is specified, including a Gauss–Newton proxy, and compare coordinates for that one fixed H; they do not transfer an update guarantee from one curvature approximation to another.

3 From spectral certificates to update accuracy

Let Φ be a finite, convex, orthogonally invariant, Loewner-monotone function on PSD factor operators, such as $\lambda_{\max}$, the trace, or $\mathrm{Tr}(A^p)/p$, $1 \le p < \infty$, but not a condition number (we write Tr on factor operators and tr on $k \times k$ matrices). We study $F_\Phi(T) = \Phi(H^{1/2}\mathcal{W}(T)H^{1/2})$. This operator has the spectrum of $K_H(T)$, so $F_{\lambda_{\max}} = f_H$; the order used here makes the certificate below linear in $\mathcal{W}(T)$. Reanchor a reference optimum to I, transporting U, V, H and the *original* budget to $\overline{\mathcal{Q}} = \{T : \underline{T} \preceq T \preceq \overline{T}\}$, where $0 \prec \underline{T} \preceq I \preceq \overline{T}$ and $\underline{T} \prec \overline{T}$ (Appendix A.1). This does not replace the budget by a new box centered at I. From here on U, V, H denote the transported quantities.

A spectral certificate in latent coordinates. Choose a PSD spectral subgradient $\Xi \in \partial\Phi(H)$ and put $B = H^{1/2}\Xi H^{1/2}$. Its two partial traces are the PSD matrices P_U, P_V defined by

$$\mathrm{Tr}(B\mathcal{W}(T)) = \mathrm{tr}(P_U T^{-1}) + \mathrm{tr}(P_V T), \qquad \mathsf{P}_\Phi = \tfrac{1}{2}(P_U + P_V). \tag{3}$$

For example, if $B = \sum_j b_j q_j q_j^\top$ and $q_j = (X_j, Y_j)$, then $P_U = \sum_j b_j X_j^\top X_j$ and $P_V = \sum_j b_j Y_j Y_j^\top$. We require this subgradient to certify optimality: for some PSD multipliers,

$$\begin{aligned} P_V - P_U + \Lambda_u - \Lambda_l = 0, \qquad \Lambda_u, \Lambda_l \succeq 0, \\ \langle \Lambda_u, \overline{T} - I \rangle = \langle \Lambda_l, I - \underline{T} \rangle = 0. \end{aligned} \tag{4}$$

At an interior optimum $P_U = P_V$; boundary optima need not satisfy this equality. The average P_Φ handles both cases. We fix this certificate; dependence on the chosen Ξ is suppressed in the notation.

The gradient uses latent directions through Q_D, and we call $\kappa_\Phi(D)$ its *coverage coefficient*:

$$Q_D = U^\top D D^\top U + V D^\top D V^\top, \qquad \kappa_\Phi(D) = \inf\{c \ge 0 : Q_D \preceq c\mathsf{P}_\Phi\}. \tag{5}$$

Intuitively, P_Φ marks the latent directions along which the certificate proves that rescaling raises the objective, and Q_D those that the gradient step moves. The coefficient is finite exactly when every used direction is marked, $\operatorname{range} Q_D \subseteq \operatorname{range} \mathsf{P}_\Phi$; it then equals $\lambda_{\max}(\mathsf{P}_\Phi^{\dagger/2} Q_D \mathsf{P}_\Phi^{\dagger/2})$ (zero when $Q_D = 0$).

Theorem 1 (Spectral certificates for product-response accuracy)**.** *Under (4), let $\ell_0 = \lambda_{\min}(\underline{T}) > 0$, $u_0 = \lambda_{\max}(\overline{T})$, and $C_{\mathcal{Q}} = \sqrt{u_0}\|U\|_{\mathrm{op}} + \ell_0^{-1/2}\|V\|_{\mathrm{op}}$ for the transported factors. If $\kappa_\Phi(D) < \infty$, every $T \in \overline{\mathcal{Q}}$ satisfies*

$$\|\mathcal{P}_T(D) - \mathcal{P}_I(D)\|_F \le C_{\mathcal{Q}} \sqrt{\kappa_\Phi(D)[F_\Phi(T) - F_\Phi(I)]}. \tag{6}$$

No simplicity, interiority, or factor-rank assumption is required.

Proof idea and use. The certifying subgradient Ξ and the normal cones of the original and inverse budgets imply $F_\Phi(T) - F_\Phi(I) \ge \operatorname{tr}[\mathsf{P}_\Phi(T + T^{-1} - 2I)]$. Factoring the two gradient blocks through $\mathsf{P}_\Phi^{1/2}$ gives (6) (Appendix A); the square-root order is attained (Appendix A.3). In use, the inputs are a solved reference, its certifying subgradient ($\Xi = I$ for the trace, a top-eigenspace density for $\lambda_{\max}$) and the current gradient. A displacement tolerance τ then holds for *every* coordinate with $\kappa_\Phi(D)[F_\Phi(T) - F_\Phi(I)] \le [\tau/(\eta C_{\mathcal{Q}})]^2$, including any a solver returns; comparing two computed updates gives no such guarantee. A failed certificate is inconclusive in general: another valid subgradient may cover the gradient. Appendix A.5 adds the error term for a feasible inexact reference.

Corollary 2 (Full-support response identification)**.** *Under the same certificate assumptions, suppose the certificate is* full-support*, $\operatorname{range} B = \operatorname{range} H$, and $J^* D = (DV^\top, U^\top D) \in \operatorname{range} H$. Then $\kappa_\Phi(D) < \infty$, and all minimizers of F_Φ on $\overline{\mathcal{Q}}$, whatever their coordinates, give one product response.*

Trace and finite-power criteria have $B = H$ and $B = H^p$, respectively, so they preserve every positive-curvature mode. The compatibility condition holds for actual gradients of linear least-squares and fixed-feature softmax cross-entropy with $H = J^* \mathcal{C} J$ (Appendix A.4).

The example from the introduction. For the quadratic loss specified in Appendix A.6, $\mathsf{P}_{\max} = \operatorname{diag}(1, 0)$, $\mathsf{P}_{\mathrm{tr}} = \operatorname{diag}(5/4, 1/2)$, and $Q_D = 2E_{22}$. The largest-eigenvalue certificate misses the gradient direction, while $\kappa_{\mathrm{tr}}(D) = 4$ (identification, not training performance).

4 The Largest Eigenvalue: An Exact Identification Criterion

For $\Phi = \lambda_{\max}$ the coverage condition becomes necessary as well as sufficient under local spectral assumptions; we first record the optimality criterion that identifies the active directions. For $K_H(S)$ with top eigenvalue $\lambda > 0$, a *top density* $\Omega = \sum_j w_j q_j q_j^\top$ is PSD, trace one and supported within the top eigenspace (not necessarily on all of it). Write $q_j = (X_j, Y_j)$ and define

$$C_U(\Omega) = \sum_j w_j X_j^\top X_j, \qquad C_V(\Omega) = \sum_j w_j Y_j Y_j^\top, \qquad \Gamma_{H,S}(\Omega) = \lambda S^{-1/2}(C_V - C_U)S^{-1/2}.$$

These are the partial traces of (3) applied to Ω, so they do not depend on its decomposition; they concern curvature modes, not the factor Grams.

Theorem 3 (Optimizer membership on a finite gauge box)**.** *For a supplied $H \succeq 0$ and finite $R \ge 1$,*

$$\min_{\substack{G \in \mathrm{GL}(k) \\ \|G\|_{\mathrm{op}},\, \|G^{-1}\|_{\mathrm{op}} \le R}} \|H_G\|_{\mathrm{op}} = \min_{S \in \mathcal{Q}_R} \lambda_{\max}\Big(H^{1/2} \mathcal{W}(S) H^{1/2}\Big) = \mathfrak{F}_R(H). \tag{7}$$

This is an attained convex problem in S. If $R > 1$ and $H \neq 0$, a feasible S belongs to $\mathrm{Opt}_R(H)$ if and only if some top density Ω and PSD multipliers satisfy

$$\begin{aligned} 0 &= \Gamma_{H,S}(\Omega) + \Lambda_u - \Lambda_l, \qquad \Lambda_u, \Lambda_l \succeq 0, \\ 0 &= \big\langle \Lambda_u, R^2 I - S \big\rangle = \big\langle \Lambda_l, S - R^{-2} I \big\rangle . \end{aligned} \tag{8}$$

At an interior point with a simple top mode $q = (X, Y)$, the criterion is $X^\top X = Y Y^\top$. If $H = 0$, every feasible S is optimal; if $R = 1$, the only feasible point is I.

Appendix B proves this, including repeated eigenvalues and budget faces.

Local optimizer geometry. A simple top eigenvalue can leave coordinate directions undetermined. Reanchor an interior optimum to I as in Section 3; Opt is the set of minimizers of f_H on $\overline{\mathcal{Q}}$.

Proposition 4 (Active rank and local optimizer geometry)**.** *Let $H \succeq 0$ and $\lambda = \lambda_{\max}(H) > 0$ be simple with unit mode $q = (X, Y)$. At an interior optimum I, Theorem 3 gives the* active latent Gram *$P = X^\top X = YY^\top$; put $r = \operatorname{rank} P$, and let N span the* free subspace $\ker P$ *orthonormally. In a sufficiently small neighborhood $\mathcal{O}$ of I within the feasible interior, the complete optimizer set and its dimension are*

$$\operatorname{Opt} \cap \mathcal{O} = \{I + NZ_0N^\top : Z_0 = Z_0^\top\} \cap \mathcal{O}, \qquad \dim_{\mathrm{loc}} \operatorname{Opt} = \frac{(k-r)(k-r+1)}{2}. \tag{9}$$

The smooth coordinate Hessian has kernel $\{\Delta = \Delta^\top : \Delta P^{1/2} = 0\}$. If $r = k$, I is the unique global SPD minimizer and $f_H(T) - \lambda \geq \lambda\lambda_{\min}(P)\|\log T\|_F^2$.

The proof rests on the fixed minorant $f_H(T) \geq \lambda \operatorname{tr}[P(T + T^{-1})] \geq \lambda$. Equality forces T to fix $\operatorname{range} P$; conversely, changes confined to $\ker P$ preserve the top mode and remain optimal locally by the gap. Appendix C proves both directions and the Hessian claim, and extends the local-set formula to repeated modes under a balanced density that is positive on the *complete* top eigenspace.

When spectral accuracy controls an update. A small spectral excess need not make a given factor update insensitive to coordinate choice. With the same reanchoring and budget $\overline{\mathcal{Q}}$, measure the response (1) by

$$\omega_D(\varepsilon) = \sup_{\substack{T \in \overline{\mathcal{Q}} \\ f_H(T) \leq \lambda(1+\varepsilon)}} \|\mathcal{P}_T(D) - \mathcal{P}_I(D)\|_F.$$

Proposition 5 (Spectral accuracy and a fixed update)**.** *Under Proposition 4,*

$$\lim_{\varepsilon \downarrow 0} \omega_D(\varepsilon) = 0 \quad \Longleftrightarrow \quad N^\top U^\top D = 0 \ \text{ and } \ DV^\top N = 0. \tag{10}$$

When these conditions hold, $\omega_D(\varepsilon) = O(\sqrt{\varepsilon})$. Otherwise, two explicitly bounded perturbations of I along the free subspace remain spectrally optimal, and at least one has a nonzero response difference.

We call $N^\top U^\top D$ and $DV^\top N$ the *leakage* of D into the free subspace. The tests ask whether the gradient uses directions that the criterion leaves undetermined, for the same operator H and its exact reference optimum; Appendix C.4 gives the constants, a positive lower bound and the repeated-eigenvalue extension under full active support. For a simple top mode, the certificate of Theorem 1 is $\mathsf{P}_\Phi = \lambda P$, and finite $\kappa_\Phi(D)$ is exactly zero leakage. The local geometry therefore supplies necessity in this case, beyond the general sufficient certificate.

Zero leakage thus supplies a uniform spectral stopping tolerance; otherwise the explicit floor (39) shows when even exact spectral optimality cannot meet a requested response tolerance.

Both propositions need a PSD operator with a simple top eigenvalue at an interior optimum, not full-rank factors or stationarity; for convex ℓ the pullback $J^*\mathcal{C}J$ is PSD at every iterate, so they apply there with $D = E$. One-sided quadratic losses give a rank-one active structure (Appendix C.2).

5 Quantifying remaining response freedom

When a spectral criterion does not identify the response, the freedom that remains can still be measured exactly. Let $G_D(T) = \langle D, \mathcal{P}_T(D)\rangle$, the squared factor-gradient norm in coordinates T. For the actual loss gradient this is the coefficient of first-order loss decrease; here it measures the pairing of the response with the fixed gradient inside a certified spectral block. Let $N \in \mathbb{R}^{k\times d}$, $N^\top N = I$, and consider

$$T_Z = I - NN^\top + NZN^\top, \qquad lI \preceq Z \preceq hI, \qquad 0 < l < 1 < h.$$

The whole block box is required to lie in the original budget and the desired spectral sublevel set. Define

$$A_D = N^\top U^\top DD^\top UN, \qquad B_D = N^\top VD^\top DV^\top N, \qquad C_D = A_D - B_D/(lh),$$

so that $g(Z) = G_D(T_Z) - G_D(I) = \operatorname{tr}[A_D(Z - I) + B_D(Z^{-1} - I)]$.

Theorem 6 (Extremal response pairing on a certified spectral block)**.** *Write $\Pi_+(C_D)$ for the orthogonal projector onto the positive eigenspace of C_D. Without assuming that A_D, B_D commute,*

$$Z_* = lI + (h-l)\Pi_+(C_D) \tag{11}$$

is a maximizer of g on the entire block box, with value

$$g_* = (l-1)\operatorname{tr} A_D + (l^{-1}-1)\operatorname{tr} B_D + (h-l)\operatorname{tr}((C_D)_+). \tag{12}$$

Its improvement over uniform endpoint choices is exactly

$$g_* - \max\{g(lI), g(hI)\} = (h-l)\min\{\operatorname{tr}((C_D)_+), \operatorname{tr}((-C_D)_+)\}. \tag{13}$$

Here C_+ denotes the positive part of a symmetric matrix.

The inverse-function chord bounds g by an affine function of Z, and (11) attains the bound at endpoint eigenvalues (Appendix D). Strict improvement over both uniform endpoints occurs exactly when C_D has positive and negative eigenvalues. In one dimension the formula reduces to the two endpoints; a zero eigenspace of C_D can make the maximizer nonunique.

Given U, V, D, an orthonormal N and certified bounds l, h, the extremal coordinate $T_* = I - NN^\top + NZ_*N^\top$ therefore costs one $d \times d$ eigendecomposition.

Certifying the whole box. Under Proposition 4, let $\gamma = \lambda - \lambda_2(H) > 0$ be the gap to the second-largest eigenvalue, $\rho_N = \operatorname{tr}(N^\top P N)$, and $C_\gamma = 2 + 20\lambda/(3\gamma)$. Here N may contain weakly active directions, not only $\ker P$. If the box lies in the original budget and $0 < a \le \min\{1/2, \gamma/(4\lambda+\gamma)\}$, then

$$\sup_{(1-a)I \preceq Z \preceq (1+a)I} \left[\frac{f_H(T_Z)}{\lambda} - 1\right] \le C_\gamma \rho_N a^2. \tag{14}$$

Thus $l = 1-a, h = 1+a$ and $C_\gamma \rho_N a^2 \le \varepsilon$ give a complete ε-near-optimal box. When $\rho_N = 0$ (N inside the free subspace), the box is exactly optimal, and Appendix D.2 gives a less conservative radius. The free dimension is set by $r = \operatorname{rank} P$, not by the gap γ, which only limits how far the top eigenspace can be transported before another mode overtakes it.

Uniform endpoints as a special case. For N spanning the free subspace $\ker P$, choose an admissible radius a; the *uniform endpoints* $T_\pm = I \pm aNN^\top$ are the block coordinates at $Z = hI$ and $Z = lI$, with $l = 1-a$, $h = 1+a$. Then

$$G_D(T_\pm) - G_D(I) = \pm a\Big(\operatorname{tr} A_D - \frac{\operatorname{tr} B_D}{1 \pm a}\Big), \quad \max_{\pm} G_D(T_\pm) - G_D(I) \ge \frac{a^2(\operatorname{tr} A_D + \operatorname{tr} B_D)}{2 - a^2}. \tag{15}$$

Here $\operatorname{tr} A_D$ and $\operatorname{tr} B_D$ are the squared norms of the two leakages, so nonzero leakage gives response freedom at one uniform endpoint (Appendix C.4), while Theorem 6 quantifies the matrix block.

6 Numerical results

The experiments ask whether a covering certificate turns spectral error into update error (Sections 6.1 and 6.5), how far a nearly exact spectral solve without coverage leaves the update undetermined (Sections 6.2, 6.3 and 6.5), and whether the variation is a leakage effect rather than a step-size or numerical effect (Sections 6.2 and 6.4); spectral values carry exact rational witnesses under squared loss, while all other comparisons, including cross-entropy deviations from a solved reference, are floating-point. We use squared-loss reduced-rank digit classifiers (*heads*) of ranks $2, 4$, with three stratified 75/25 splits of the 1,797-example optical-digit subset, 12 unwhitened training principal components, one-hot targets and product ridge 10^{-2} (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). All comparisons keep the product, gradient and original budget $\mathcal{Q}_2$ fixed (Appendix E); at the frozen references the squared-loss largest-eigenvalue results are identical for every $R \ge 1.8$ (Appendix E.9).

Table 1: **Direct check of Theorem 1 on the 42 frozen checkpoints.** For each full-support criterion we re-solve only the small coordinate problem and compare its reference response with that of the $\lambda_{\max}$ coordinate. Each entry is a median over the 21 checkpoints of that rank, taken column by column: observed is the Frobenius response distance, bound is the certificate right-hand side, and ratio is the median of the per-checkpoint ratios bound/observed. Brackets give the range of the three per-seed medians of the ratio. The Schatten-2 rows include the Appendix A.5 KKT-residual correction; all 84 individual inequalities hold.

Criterion	Rank	$\kappa_\Phi(D)$	$C_\mathcal{Q}$	observed	bound	ratio	[seeds]
Trace	2	0.00901	2.68	0.00833	0.0655	6.53	[5.8, 6.9]
Trace	4	0.00883	3.16	0.0144	0.138	10.9	[8.8, 12.0]
$\|A\|_F$ (Schatten-2)	2	0.0662	2.69	0.00815	0.0872	9.23	[8.7, 10.5]
$\|A\|_F$ (Schatten-2)	4	0.106	3.23	0.0132	0.184	16.4	[13.6, 16.7]

6.1 A DIRECT CHECK OF THE GENERAL CERTIFICATE

At the 42 checkpoints of Section 6.2 we re-solve only the small coordinate problem for two full-support criteria, without retraining: trace and the Schatten-2 norm $\Phi(A) = \|A\|_F$. All 84 references are interior (KKT residuals below 1.6×10^{-8}; Appendix E.2). Taking the solved largest-eigenvalue coordinate as a feasible comparator, we evaluate both sides of Theorem 1. The bound is finite in all 84 criterion–checkpoint comparisons and contains the observed response distance in every case (Figure 1b); median bound/observed ratios range from 6.5 to 16.4. Two baselines calibrate this (Table 4): at the comparators' objective excess of 1.6%–10.7% the bound is below the spectrum-free response diameter of the budget in 83 of 84 cases (medians 0.24–0.57) but above the reference response norm in all 84 (medians 1.4–3.4), so gaps of this size give loose guarantees.

Small optimality residuals give tight ones: with the residual term of Appendix A.5, Theorem 1 becomes a stopping rule (Appendix E.10) that bounds, from a feasible solver iterate alone, how far every exact minimizer's response can lie, stops when this bound meets a target, continues otherwise, and reports no guarantee if $\kappa_\Phi(D) = \infty$. Started from the current coordinates, solver iterates for the Schatten-2 criterion certify a 1% response tolerance at all 42 checkpoints within 39 iterations (median 19.5), with actual errors at most 3.1×10^{-5} of the response. For the largest eigenvalue the leakage of Section 6.2 makes κ infinite (Proposition 5), so under squared loss the rule never stops, however accurate the solve; under cross-entropy it certifies 1% at 16 of the 34 simple-top references of Section 6.5 and asks for a more accurate reference at the rest.

6.2 LARGEST-EIGENVALUE ACCURACY WITHOUT UPDATE IDENTIFICATION

From a seeded random initialization, each head takes 200 plain gradient steps with $\eta = 0.1$. At seven checkpoints per head we optimize the pullback on $\mathcal{Q}_2$, obtaining 42 reference coordinates S^*. The inferred active latent Gram has rank one, the relative top gaps exceed 2.9×10^{-3}, and balance residuals are below 10^{-6}. The reference and both uniform endpoints, in original coordinates $(S^*)^{1/2} T_\pm (S^*)^{1/2}$, have rationally certified relative spectral excess below 7×10^{-6} (Figure 1a).

Across these same points, the leakage carries 30%–100% of the norm of each factor-gradient block, the response changes by 0.07%–4.9%, and the measured one-step decrease at the better uniform endpoint by 0.01%–1.7% (Figure 1a; Appendix E.3, Table 5). Nor is this step-size tuning: with any two nonnegative scalar rates on the reference factors, the finite-step matching lower bound (41) of Appendix C.4 leaves a median 1.36% and maximum 4.07% of the reference displacement norm (Figure 2c).

6.3 EXTREMAL RESPONSE PAIRING ON THE SAME CHECKPOINTS

Keeping every reference, radius and gradient unchanged, we evaluate the closed form (11) without retraining. At rank two the free block is one-dimensional, so the matrix extremum equals a uniform endpoint. At rank four it strictly exceeds both endpoints in 20 of 21 cases, exactly those in which C_D has both positive and negative eigenvalues, as predicted by (13). The extra pairing has median 0.281% and maximum 0.981% of $G_D(I)$ (Table 2, Figure 2a).

Table 2: Block extremum versus uniform endpoints on the same 42 checkpoints. $d = k - r$ is the block dimension (the active rank is $r = 1$). Mixed counts the indefinite matrices C_D; gain counts strict pairing extrema above the better uniform endpoint. Extra pairing is normalized by $G_D(I)$. 1-step agrees counts, among the strict gains, the cases in which the larger pairing also gives lower one-step loss at $\eta = 0.1$, a local consistency check only; a dash means there is no strict gain to compare. Counts are out of 21 per rank; spectral-excess bounds are rounded upward.

Rank	d	Mixed / gain	Extra pairing (%) median / max	1-step agrees	Spectral excess (10^{-6})
2	1	0 / 0	0.000 / 0.000	–	4.2
4	3	20 / 20	0.281 / 0.981	20 / 20	6.7

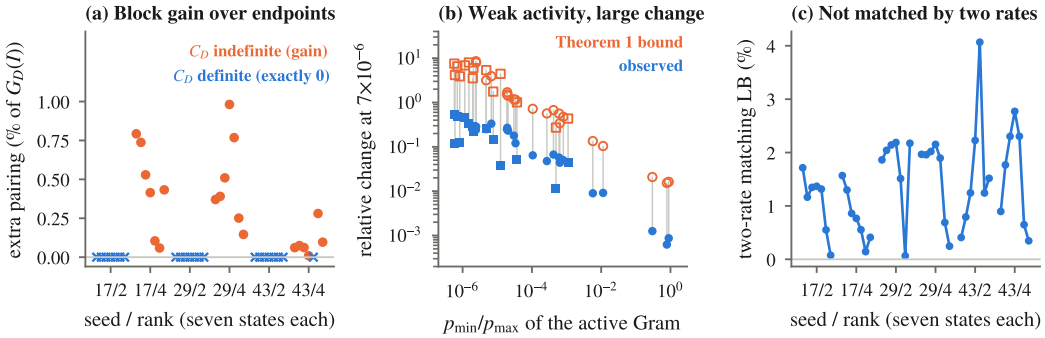


Figure 2: **The remaining freedom is quantified, grows with weak activity, and is not matched by two scalar factor rates. (a)** Extra pairing of the block extremum over the better uniform endpoint, relative to $G_D(I)$: exactly zero whenever C_D is definite (all 21 rank-two states and one rank-four state) and 0.009%–0.98% when it is indefinite, as (13) predicts. **(b)** Cross-entropy, 34 simple-top checkpoints (Section 6.5; circles rank two, squares rank four): response change along the softest direction within relative objective deviation 7×10^{-6} (filled) and the residual-corrected bound of Theorem 1 (open), against $p_{\min}/p_{\max}$ of the active Gram. The change grows as the Gram weakens (rank correlation -0.85); the bound tracks it (0.85) and lies above every point. **(c)** Lower bound on the update component that no pair of nonnegative scalar factor rates reproduces, for the uniform endpoint with the larger pairing, relative to the reference displacement (Appendix E.5); median 1.36%, maximum 4.07% over the 42 states.

At $\eta = 0.1$ the same 20 cases also have the one-step loss ordering predicted by the larger pairing, a consistency check of Appendix D.3, not a long-horizon performance claim. Separate rational witnesses certify the relative pullback spectral excess of all 42 rounded block representatives below 7×10^{-6} on the original budget (Table 2); they certify these representatives, not every point of the numerical block.

6.4 Mechanism controls

Projecting each batch gradient onto the zero-leakage subspace of (10) reduces the largest absolute response change below 6.4×10^{-17}, so the variation is carried by leakage, not by coordinate change as such (Appendix E.6; Figure 1c). The coordinate-Hessian nullities are one and six at ranks two and four, as predicted. At six stationary fitted products, where the full-batch response vanishes, fixed mini-batch gradients still give coordinate-induced changes of 0.178%–2.192% of each batch's own response: the coordinate effect acts on each fixed batch gradient and does not come from resampling.

6.5 Two-sided curvature

Squared loss forces a rank-one active latent Gram (Appendix C.2). Retraining the heads with fixed-feature softmax cross-entropy, with the exact cross-entropy pullback (Appendix E.8, Table 8), gives

42 checkpoints, 34 with a simple top eigenvalue (the eight near-repeated ones are excluded from the statistics). At these 34 the active Gram has full rank, so under exact simple-top geometry the coordinate would be unique, but the Gram can be poorly conditioned: $p_{\min}/p_{\max}$ of P is down to 6×10^{-7}. Within a relative objective deviation of 7×10^{-6} from the solved reference, the response along the softest direction changes by 0.06%–54% of the reference response (median 12%; largest early in training, 0.06%–4.4% at $t = 200$ in the four heads with a simple top there), more for weaker Grams (Figure 2b), and Theorem 1 with the inexact-reference correction of Appendix A.5 contains every one of the 34 observed changes (median bound/observed 13).

7 Related work

Spectral scaling and certificates. Uniqueness of optimal diagonal scalings is already known to depend on the support of the extreme modes (Balakrishnan & Boyd, 1995; Ghadimi et al., 2025). Eigenvalue optimization, spectral subdifferentials and error bounds are standard tools (Overton, 1992; Shapiro & Fan, 1995; Lewis & Overton, 1996; Lewis, 1999; Rebjock & Boumal, 2025), and the count in Proposition 4 has the form of face-dimension counts in semidefinite programming (Pataki, 1998). Li (2026, Theorems 3.6, 5.9) derives log-spectral optimality certificates and Kronecker partial-trace conditions for structured metrics, with residual certificates for point error, objective gap and thresholds (Corollary 5.13, Theorem 5.18); these control the metric, whereas here the certificate controls the product response to a fixed gradient on the original budget (claim table in Appendix G).

Low-rank optimization and coordinate selection. Fixed-rank quotient geometry is the general language of factorization invariance (Mishra et al., 2014; Mishra & Sepulchre, 2016). In low-rank adapters, balancing, factor-specific rates, Riemannian preconditioning and invariant updates address the dependence on factor coordinates (Hu et al., 2022; Castin et al., 2026; Kamber & Parhi, 2025; Hayou et al., 2024; Zhang & Pilanci, 2024; Lion et al., 2025; Yen et al., 2025; Bogachev et al., 2026). RefLoRA minimizes a bound on the next-step loss, whose minimizer need not be unique, and proves that its refactorization rule is consistent (Zhang et al., 2025, Theorems 3 and 6). We ask a different question: whether *all* coordinates near-optimal for a curvature criterion give nearly the same response. Symmetry teleportation moves along loss-invariant orbits to increase the gradient norm (Zhao et al., 2022; 2024), and LoRA-Pro chooses factor gradients whose product update best matches the full gradient (Wang et al., 2025); the block theorem instead computes the extremal pairing inside a certified spectral block. Weight-space symmetry also enters curvature approximations (Artemev et al., 2026); LoRA-Muon, PoLoRA and LoRA-TSD constrain the update itself (Cesista et al., 2026; Ghosh et al., 2026; Andriianov et al., 2026).

8 Limitations

In general the response certificate is only sufficient: failure of one certificate is inconclusive. The block extremum quantifies response freedom inside its certified box, not finite-step loss over the full orbit. The states come from six seed/rank configurations per loss on one small dataset; rational witnesses certify spectral values and budget feasibility, the simple-top geometry is a numerical diagnostic, and near-repeated spectra need the full active-density treatment. Under the simple-top assumptions, the active latent Gram separates three regimes: rank deficiency (exact freedom), full rank with small eigenvalues (sensitivity to finite spectral tolerances), and a well-conditioned Gram (a favorable but not guaranteed tolerance). The squared-loss heads sit in the first (Appendix C.2); the simple-top cross-entropy checkpoints move from the second toward the third. Which regime large adapters occupy is open, and longer runs of the uniform endpoint rules show no stable performance ordering (Appendix E.7).

9 Conclusion

A spectral objective can be solved almost exactly while the low-rank update remains undetermined. The spectral gap alone does not control the step; gradient-relative coverage and the reference's optimality residual do, and together they give a stopping rule that either certifies the update or reports that no guarantee is possible. Next steps are adapter-scale diagnostics and certificates for gauge-equivariant preconditioned updates.

REPRODUCIBILITY STATEMENT

The anonymized supplementary material contains the source code, all fixed data, preprocessing, factors, checkpoint coordinates, numerical results and candidate witnesses needed to reproduce the reported comparisons, together with instructions for rebuilding every table and figure and the software versions used. Appendices E and F specify the protocols and the exact value bounds. The exact witnesses are checked with rational arithmetic alone, in about 30 seconds of single-core CPU time; the floating-point comparisons are recomputed within stated tolerances.

AI USE STATEMENT

Large language models were used for language polishing in the Abstract, Introduction, and selected parts of the appendices, and to assist with code implementation and experiments. The authors take full responsibility for all mathematical statements, proofs, experimental results, interpretations, and conclusions, as well as the final manuscript and code.

REFERENCES

P.-A. Absil, Robert Mahony, and Jochen Trumpf. An extrinsic look at the Riemannian Hessian. In *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pp. 361–368. Springer, 2013.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. Dataset, CC BY 4.0.

Dmitrii Andriianov, Andrey Veprikov, and Aleksandr Beznosikov. LoRA-TSD: Tangent-Space Spectral Descent for LoRA via Muon-Style Updates. *arXiv preprint arXiv:2609.02734*, 2026. Concurrent preprint; first submitted 2 September 2026.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253. PMLR, 2018.

Artem Artemev, Rui Xia, Benjamin M. Boyd, Youjing Yu, Felix Dangel, Guillaume Hennequin, and Alberto Bernacchia. Exploiting weight-space symmetries for approximating curvature. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

V. Balakrishnan and S. Boyd. Existence and uniqueness of optimal matrix scalings. *SIAM Journal on Matrix Analysis and Applications*, 16(1):29–39, 1995.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA Meets Riemannion: Muon Optimizer for Parametrization-Independent Low-Rank Adapters. In *International Conference on Learning Representations*, 2026.

Valérie Castin, Kimia Nadjahi, Pierre Ablin, and Gabriel Peyré. Balanced LoRA: Removing Parameter Invariance to Accelerate Convergence. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. LoRA-Muon: Spectral Steepest Descent on the Low-Rank Manifold. *arXiv preprint arXiv:2606.12921*, 2026.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Saeed Ghadimi, Woosuk L. Jung, Arnesh Sujanani, David Torregrosa-Belén, and Henry Wolkowicz. Optimal diagonal preconditioning beyond worst-case conditioning: Theory and practice of omega scaling. *arXiv preprint arXiv:2509.23439*, 2025. Version 2.

Nikhil Ghosh, Tetiana Parshakova, and Robert M. Gower. PoLoRA: A Preconditioned Orthogonalized LoRA Optimizer. *arXiv preprint arXiv:2607.17620*, 2026.

Frank Hansen and Gert K. Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564, 2003.

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.

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.

Anıl Kamber and Rahul Parhi. Sharpness of minima in deep matrix factorization. *arXiv preprint arXiv:2509.25783*, 2025. Version 6.

Daniel Kunin, Javier Sagastuy-Breña, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Adrian S. Lewis. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3): 576–588, 1996.

Adrian S. Lewis. Nonsmooth analysis of eigenvalues. *Mathematical Programming*, 84(1):1–24, 1999.

Adrian S. Lewis and Michael L. Overton. Eigenvalue optimization. *Acta Numerica*, 5:149–190, 1996.

Zavier Li. Structured preconditioning in affine-invariant geometry: Projection, certificates, and Kronecker separation. *arXiv preprint arXiv:2607.12286*, 2026.

Kai Lion, Liang Zhang, Bingcong Li, and Niao He. PoLAR: Polar-Decomposed Low-Rank Adapter Representation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3–4):591–621, 2014.

Michael L. Overton. Large-scale optimization of eigenvalues. *SIAM Journal on Optimization*, 2(1): 88–120, 1992.

Michael L. Overton and Robert S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62: 321–357, 1993.

Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: Four equivalent conditions for C^2 functions. *Mathematical Programming*, 213(1–2):151–199, 2025.

Alexander Shapiro and Michael K. H. Fan. On eigenvalue optimization. *SIAM Journal on Optimization*, 5:552–569, 1995.

Devender Singh. The Loss Does Not See the Basis, but Adam Does. *arXiv preprint arXiv:2608.05136*, 2026.

Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-Pro: Are low-rank adapters properly optimized? In *International Conference on Learning Representations*, 2025.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Venkata Sai Surya Subramanyam Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA Done RITE: Robust Invariant Transformation Equilibration for LoRA Optimization. In *International Conference on Learning Representations*, 2025.

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. PMLR, 2024.

Yilang Zhang, Bingcong Li, and Georgios B. Giannakis. RefLoRA: Refactored Low-Rank Adaptation for Efficient Fine-Tuning of Large Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. Extended version: arXiv:2505.18877v4.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry Teleportation for Accelerated Optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving Convergence and Generalization Using Parameter Symmetries. In *International Conference on Learning Representations*, 2024.

GUIDE TO THE SUPPLEMENTARY MATERIAL

Appendices A–D prove the response certificate, its largest-eigenvalue specialization and the block-extremum result. Appendix E specifies the fixed-checkpoint study and mechanism controls; Appendix F gives the candidate-value certificates. The final appendix compares the closest mathematical results. Table 3 collects the notation.

A SPECTRAL CERTIFICATES AND PRODUCT-RESPONSE ACCURACY

Throughout this section the reference coordinates are I, and U, V, H have all been transported to that reference. The transported original domain is

$$\overline{\mathcal{Q}} = \{T : \underline{T} \preceq T \preceq \overline{T}\}, \qquad 0 \prec \underline{T} \preceq I \preceq \overline{T}, \quad \underline{T} \prec \overline{T}.$$

It is compact and lies in the SPD cone. If it is obtained from an original spectral box at S_0, then $\underline{T} = R^{-2}S_0^{-1}$ and $\overline{T} = R^2 S_0^{-1}$, as in Appendix A.1. The arguments below also allow noncommuting endpoints.

A.1 REANCHORING PRESERVES THE OPERATOR SPECTRUM AND THE BUDGET

Let S_0 be any feasible point of the original $\mathcal{Q}_R$, including a budget-boundary point, and put $G_0 = S_0^{1/2}$. At $(UG_0, G_0^{-1}V)$, the operator is $\bar{H} = \mathcal{T}_{G_0}^{-*} H \mathcal{T}_{G_0}^{-1}$. For a subsequent gauge G_1 with $T = G_1 G_1^\top$, its coordinate in the original orbit is $S = G_0 T G_0$. Both descriptions represent the same factor pair. Their transformed operators have the same full spectrum, including zero multiplicities. Consequently, for every criterion used here,

$$F_{\Phi,H}(G_0 T G_0) = F_{\Phi,\bar{H}}(T), \qquad \overline{\mathcal{Q}} = \{T : R^{-2}S_0^{-1} \preceq T \preceq R^2 S_0^{-1}\}.$$

Here $F_{\Phi,H}$ indicates which supplied operator is used. Thus a reference optimum maps to an optimum at I, whether or not it is interior. Interiority, which the local geometry requires, is also preserved. The product response is also unchanged: substitution of $\bar{U} = UG_0$, $\bar{V} = G_0^{-1}V$ gives $\mathcal{P}_T^{\bar{U},\bar{V}}(D) = \mathcal{P}_{G_0 T G_0}^{U,V}(D)$. The main text and the remaining appendices drop the bars: after reanchoring, U, V, H always denote $\bar{U}, \bar{V}, \bar{H}$. No step replaces the original budget by a new box centered at I. For $R > 1$ its endpoint difference is positive definite; $R = 1$ has a single coordinate and its update-identification question is trivial.

A.2 SPECTRAL SUPPORT AND THE TWO BUDGET NORMAL CONES

Let Φ be finite, convex, orthogonally invariant and Loewner-monotone on PSD factor operators. We assume a PSD support matrix $\Xi \in \partial\Phi(H)$; explicitly,

$$\Phi(A) - \Phi(H) \geq \operatorname{Tr}(\Xi(A - H)) \quad \text{for every PSD } A.$$

The assumption specifies the subgradient being used, including at a singular H. Since inversion is operator convex, $T \mapsto H^{1/2}\mathcal{W}(T)H^{1/2}$ is matrix convex; monotonicity and convexity of Φ imply convexity of F_Φ. These are standard spectral and convex-analytic constructions (Lewis, 1996; 1999).

Put $B = H^{1/2}\Xi H^{1/2} \succeq 0$, with partial traces P_U, P_V as in (3), and let $\Gamma = P_V - P_U$. The support inequality gives, for every SPD T,

$$F_\Phi(T) - F_\Phi(I) \geq s(T), \qquad s(T) = \operatorname{tr}[P_U(T^{-1} - I) + P_V(T - I)]. \tag{16}$$

We use the explicit certificate (4). It gives $\langle \Gamma, T - I\rangle \geq 0$ for feasible T. Consequently,

$$s(T) = \operatorname{tr}[P_U(T + T^{-1} - 2I)] + \langle \Gamma, T - I\rangle \geq \operatorname{tr}[P_U(T + T^{-1} - 2I)].$$

The inverse variable has the exact feasible set

$$\overline{\mathcal{Q}}^{-1} = \{Q : \overline{T}^{-1} \preceq Q \preceq \underline{T}^{-1}\}.$$

Complementarity of two PSD matrices implies their product is zero. Thus $\Lambda_u \overline{T} = \Lambda_u$ and $\Lambda_l \underline{T} = \Lambda_l$, as well as $\Lambda_u \overline{T}^{-1} = \Lambda_u$ and $\Lambda_l \underline{T}^{-1} = \Lambda_l$. In the inverse budget, the same multipliers exchange their upper and lower roles. It follows that $\langle -\Gamma, T^{-1} - I\rangle \geq 0$, and hence

$$s(T) = \operatorname{tr}[P_V(T + T^{-1} - 2I)] + \langle -\Gamma, T^{-1} - I\rangle \geq \operatorname{tr}[P_V(T + T^{-1} - 2I)].$$

Table 3: Notation. After reanchoring (Appendix A.1), U, V, H denote the transported factors and operator throughout. Q_D is a matrix; $\mathcal{Q}_R$ and $\overline{\mathcal{Q}}$ are sets of coordinates.

Symbol	Meaning
$M = UV, U \in \mathbb{R}^{m\times k}, V \in \mathbb{R}^{k\times n}$	represented product and its factors; $\mathbb{R}^k$ is the latent space
$G \in \mathrm{GL}(k), S = GG^\top$	gauge (a change of latent basis) and its SPD coordinate
$\mathcal{Q}_R$	coordinate budget $\{R^{-2}I \preceq S \preceq R^2 I\}$
$T, \overline{\mathcal{Q}} = \{\underline{T} \preceq T \preceq \overline{T}\}$	coordinate relative to the reference $T = I$; original budget transported to it
$\ell_0 = \lambda_{\min}(\underline{T}), u_0 = \lambda_{\max}(\overline{T})$	scalar bounds of $\overline{\mathcal{Q}}$, used in $C_{\mathcal{Q}}$ (Theorem 1)
$\mathcal{T}_G(X,Y) = (XG, G^{-1}Y)$, $H_G = \mathcal{T}_G^{-*} H \mathcal{T}_G^{-1}$	change of factor coordinates and the transformed curvature
$\mathcal{W}(S)(X,Y) = (XS^{-1}, SY)$	action of a coordinate on factor variations
$\eta, \Delta M_S$	step size; product change of one simultaneous factor-gradient step (1)
$\ell, E = \nabla\ell(M), \mathcal{C} = \nabla^2\ell(M)$	product loss, its gradient and Hessian
$J(X,Y) = XV + UY$, $H_{\mathrm{pb}} = J^*\mathcal{C}J$, $H_{\mathrm{full}} = H_{\mathrm{pb}} + \mathcal{B}_E$	product differential, pullback curvature, exact factor Hessian
Tr, tr	trace of factor operators; trace of $k \times k$ matrices
$H, K_H(S) = \mathcal{W}(S)^{1/2} H \mathcal{W}(S)^{1/2}$	supplied PSD factor curvature and its coordinate form
$f_H(S) = \lambda_{\max}(K_H(S))$, $\mathfrak{F}_R(H)$	largest-eigenvalue objective and its budget minimum
$\mathrm{Opt}_R(H)$, Opt	minimizers of f_H on $\mathcal{Q}_R$; on $\overline{\mathcal{Q}}$ after reanchoring
$\mathrm{Exc}_{H,R}(S)$	relative spectral excess $f_H(S)/\mathfrak{F}_R(H) - 1$ (Appendix F)
Φ, $F_\Phi(T)$	spectral criterion and coordinate objective $\Phi(H^{1/2}\mathcal{W}(T)H^{1/2})$; $F_{\lambda_{\max}} = f_H$
$\Xi, B = H^{1/2}\Xi H^{1/2}$	certifying subgradient; *full-support* means $\operatorname{range} B = \operatorname{range} H$
$P_U, P_V, \mathsf{P}_\Phi = \frac{1}{2}(P_U + P_V)$	partial traces of B; latent coverage matrix
Λ_u, Λ_l	PSD multipliers of the upper and lower budget faces
$D, Q_D = U^\top DD^\top U + VD^\top DV^\top$	fixed product gradient; latent directions it uses
$\kappa_\Phi(D)$	coverage coefficient $\inf\{c \geq 0 : Q_D \preceq c\,\mathsf{P}_\Phi\}$
$\mathcal{P}_T(D) = UTU^\top D + DV^\top T^{-1}V$	product response in coordinate T
$\Omega, C_U(\Omega), C_V(\Omega), \Gamma_{H,S}(\Omega)$	top density of $K_H(S)$, its partial traces, and the subgradient $\lambda S^{-1/2}(C_V - C_U)S^{-1/2}$ (Theorem 3)
$q = (X,Y), P = X^\top X = YY^\top$, $r = \operatorname{rank} P$	simple top mode, active latent Gram, active rank
$\lambda, \lambda_2(H), \gamma = \lambda - \lambda_2(H)$	top eigenvalue, second eigenvalue, top gap
N	orthonormal basis of the free subspace $\ker P$ (Proposition 4) or of a chosen block (Section 5)
$N^\top U^\top D$, $DV^\top N$	leakage of D into the free subspace (Proposition 5)
$\omega_D(\varepsilon)$	largest response change over coordinates with relative spectral excess at most ε (Proposition 5)
$T_Z = I - NN^\top + NZN^\top, lI \preceq Z \preceq hI$	block coordinates, $0 < l < 1 < h$
$T_\pm = I \pm aNN^\top$	uniform endpoints of a block of radius a
$G_D(T) = \langle D, \mathcal{P}_T(D)\rangle$	response pairing (first-order decrease for a loss gradient)
$A_D = N^\top U^\top DD^\top UN$, $B_D = N^\top VD^\top DV^\top N$, $C_D = A_D - B_D/(lh)$	block pairing matrices; C_D indefinite iff the extremum beats both endpoints
$g(Z) = G_D(T_Z) - G_D(I)$, Z_*, g_*, $\Pi_+(C_D)$	block pairing gain, its maximizer and value (11)–(12), projector onto the positive eigenspace
$\rho_N = \mathrm{tr}(N^\top PN)$, $a, C_\gamma = 2 + 20\lambda/(3\gamma)$	weak activity of a block; certified block radius and constant of (14)
S^*	solved reference coordinate of a checkpoint (Section 6)

Averaging the two inequalities proves the global bound

$$F_\Phi(T) - F_\Phi(I) \geq \operatorname{tr}[\mathsf{P}_\Phi(T + T^{-1} - 2I)], \qquad \mathsf{P}_\Phi = \tfrac{1}{2}(P_U + P_V). \tag{17}$$

In particular the reference is optimal. The proof does not impose $P_U = P_V$ at a boundary point and does not require H to be invertible.

A.3 Proof of Theorem 1

Set $C_T = T^{1/2} - T^{-1/2}$. The right-hand side of (17) is $\|C_T \mathsf{P}_\Phi^{1/2}\|_F^2$. Suppose $Q_D \preceq \kappa \mathsf{P}_\Phi$ for finite κ. The two PSD summands of Q_D are each bounded by the same right-hand side. Therefore there exist matrices $A_\dagger, B_\dagger$ with

$$U^\top D = \mathsf{P}_\Phi^{1/2} A_\dagger, \qquad DV^\top = B_\dagger \mathsf{P}_\Phi^{1/2}, \qquad \|A_\dagger\|_{\mathrm{op}}, \|B_\dagger\|_{\mathrm{op}} \leq \sqrt{\kappa}.$$

For instance, take $A_\dagger = \mathsf{P}_\Phi^{\dagger/2} U^\top D$ and $B_\dagger = DV^\top \mathsf{P}_\Phi^{\dagger/2}$; the range inclusion makes these factorizations exact. The identities $T - I = T^{1/2} C_T$ and $T^{-1} - I = -C_T T^{-1/2}$ give

$$\|U(T-I)U^\top D\|_F \leq \sqrt{u_0}\|U\|_{\mathrm{op}}\sqrt{\kappa}\|C_T \mathsf{P}_\Phi^{1/2}\|_F,$$
$$\|DV^\top(T^{-1}-I)V\|_F \leq \ell_0^{-1/2}\|V\|_{\mathrm{op}}\sqrt{\kappa}\|\mathsf{P}_\Phi^{1/2} C_T\|_F.$$

The last two Frobenius norms agree by transposition. The triangle inequality and (17) prove (6). The minimum finite κ is

$$\lambda_{\max}(\mathsf{P}_\Phi^{\dagger/2} Q_D \mathsf{P}_\Phi^{\dagger/2})$$

provided $\operatorname{range} Q_D \subseteq \operatorname{range} \mathsf{P}_\Phi$. Without that inclusion no finite PSD domination is possible; taking a pseudoinverse alone would silently discard uncovered gradient directions. When $Q_D = 0$, the response is zero for every coordinate and the bound is immediate.

If $\mathsf{P}_\Phi \succ 0$, the scalar inequality $t + t^{-1} - 2 \geq (\log t)^2$ also gives

$$F_\Phi(T) - F_\Phi(I) \geq \lambda_{\min}(\mathsf{P}_\Phi)\|\log T\|_F^2.$$

Thus the coordinate itself is unique in this stronger case. The response bound does not require that stronger conclusion. With the same certificate and D, two candidates of objective error at most δ therefore have response distance at most $2C_{\mathcal{Q}}\sqrt{\kappa_\Phi(D)\delta}$. Its square-root error order is consistent with classical growth and error-bound theory (Rebjock & Boumal, 2025); the additional information is the explicit, gradient-dependent coefficient and its spectral certificate.

The square-root order is attained. Let $m = n = k = 2, U = V = I_2$, and $\ell(M) = \frac{1}{2}\|M - (I_2 - E_{12})\|_F^2$, so that at $M = I_2$ the gradient is $D = E_{12}$, $\mathcal{C} = I$ and $H = J^*J$ with $J(X, Y) = X + Y$. For the trace criterion, $F_{\mathrm{tr}}(T) = \operatorname{Tr}(H\mathcal{W}(T)) = 2\operatorname{tr}(T + T^{-1})$, so I is the interior global optimum, $P_U = P_V = 2I$, $\mathsf{P}_{\mathrm{tr}} = 2I$, $Q_D = I$ and $\kappa_{\mathrm{tr}}(D) = 1/2$. For $T_t = \operatorname{diag}(1 + t, 1)$ with $t > 0$ small enough to stay in the budget,

$$F_{\mathrm{tr}}(T_t) - F_{\mathrm{tr}}(I) = \frac{2t^2}{1+t}, \qquad \mathcal{P}_{T_t}(D) - \mathcal{P}_I(D) = tE_{12},$$

since $T_t E_{12} = (1 + t)E_{12}$ and $E_{12} T_t^{-1} = E_{12}$. The response distance therefore equals $\sqrt{(1 + t)/2}$ times the square root of the objective gap: the exponent $1/2$ in (6) cannot be improved for an actual least-squares gradient and pullback, although the constant $C_{\mathcal{Q}}\sqrt{\kappa_\Phi(D)}$ need not be optimal. This identity is also checked numerically in the supplementary material.

A.4 Gradient compatibility and full-support criteria

Proof of Corollary 2. Write $g_D = J^*D = (DV^\top, U^\top D)$ as a vector in the factor space. Under the stated range assumptions, $\chi = \langle g_D, B^\dagger g_D\rangle$ is finite and $g_D g_D^\top \preceq \chi B$. Indeed, apply Cauchy–Schwarz after factoring $g_D = B^{1/2} B^{\dagger/2} g_D$. Partial traces are positive linear maps. Taking the partial trace of each factor block and adding them yields

$$Q_D \preceq \chi(P_U + P_V) = 2\chi \mathsf{P}_\Phi.$$

The theorem applies with $\kappa_\Phi(D) \leq 2\chi$; at every exact minimizer its right-hand side is zero.

For $\Phi(A) = \operatorname{Tr} A$, choose $\Xi = I$ and $B = H$. For $\Phi(A) = \operatorname{Tr}(A^p)/p$, $1 < p < \infty$, choose $\Xi = H^{p-1}$ and $B = H^p$. Both have the same range as H, including when H is singular. By contrast, for $\lambda_{\max}$, $B = \lambda\Omega$ uses only the certified top eigenspace. A top-space partial trace can nevertheless be full rank, and small positive weights can give large constants even for full-support criteria.

Pullback compatibility. If $H = J^*\mathcal{C}J$, $\mathcal{C} \succeq 0$, and $D \in \operatorname{range}\mathcal{C}$, then

$$J^*D \in \operatorname{range}(J^*\mathcal{C}^{1/2}) = \operatorname{range}(J^*\mathcal{C}J).$$

The equality uses $\operatorname{range}(AA^*) = \operatorname{range} A$. For a linear least-squares loss $\ell(M) = \frac{1}{2}\|\mathcal{L}(M) - y\|^2$, its gradient lies in $\operatorname{range}\mathcal{L}^* = \operatorname{range}(\mathcal{L}^*\mathcal{L})$. No feature-rank or factor-rank assumption is needed.

Fixed-feature softmax cross-entropy. Let the logits be $z = \mathcal{L}(M)$, with fixed linear $\mathcal{L}$, and let each target vector sum to one. At finite logits, each probability vector p has positive entries. The logit Hessian $C_z = \operatorname{diag}(p) - pp^\top$ has kernel $\operatorname{span}\{\mathbf{1}\}$, while the residual $r = p - y$ sums to zero. Thus $r \in \operatorname{range} C_z$; the same holds for the block-diagonal Hessian over samples. With $D = \mathcal{L}^*r$ and $\mathcal{C} = \mathcal{L}^*C_z\mathcal{L}$,

$$D \in \operatorname{range}(\mathcal{L}^*C_z^{1/2}) = \operatorname{range}\mathcal{C}.$$

The preceding pullback argument proves compatibility for $J^*\mathcal{C}J$. This is a nonquadratic application of the sufficient condition, not a claim that the exact factor Hessian is PSD.

Why compatibility cannot be omitted. For the scalar linear product loss $\ell(M) = M$, take $U = V = D = 1$. Then $H = H_{\mathrm{pb}} = 0$, so the trace objective is zero everywhere, while $\mathcal{P}_t(D) = t + t^{-1}$ varies with t. Full spectral support alone cannot identify the response to a gradient outside the curvature range.

A.5 Certificate choice and approximate optimality

A failed certificate need not be an obstruction. Let $U = V = I_2$, $D = \operatorname{diag}(0, 1)$, and let $H = I$ on the two factor spaces. The objective is $\max\{\lambda_{\max}(T), \lambda_{\min}(T)^{-1}\}$, whose unique SPD minimizer is I. For either $j = 1$ or $j = 2$, use the density that gives mass $1/2$ to each of the unit factor modes $(e_je_j^\top, 0)$ and $(0, e_je_j^\top)$. Both satisfy the interior KKT condition, but they give $\mathsf{P}_\Phi = \frac{1}{2}e_je_j^\top$. Since $Q_D = 2e_2e_2^\top$, the first certificate has infinite coverage coefficient and the second has coefficient four. Thus failure of one density cannot establish update ambiguity.

At a repeated top eigenvalue and for $Q_D \neq 0$, one may optimize the certificate itself. With the top eigenspace fixed, maximizing $t \geq 0$ subject to $\Omega \succeq 0$, $\operatorname{Tr}\Omega = 1$, top-space support, (4), and $\mathsf{P}_\Omega \succeq tQ_D$, where $\mathsf{P}_\Omega = \frac{\lambda}{2}[C_U(\Omega) + C_V(\Omega)]$ is the coverage matrix of the certificate $\Xi = \Omega$, is a semidefinite program. A feasible $t > 0$ gives $\kappa \leq 1/t$. This optional certificate search is not needed for the main results or the reported experiments.

A feasible reference with an optimality residual. Now let I be feasible rather than exactly optimal. Retain a valid PSD subgradient Ξ, and choose a true normal $\nu \in \mathcal{N}_{\overline{\mathcal{Q}}}(I)$, using the outward-normal convention. Set $\eta_{\mathrm{opt}} = \|\Gamma + \nu\|_F$. For the largest-eigenvalue criterion, minimizing over normals and top densities gives the analogue of Appendix B.2's residual on the transported budget; its Frobenius norm need not equal the residual in the original coordinates. A normal has the multiplier representation used above; complementarity therefore implies $-\nu \in \mathcal{N}_{\overline{\mathcal{Q}}^{-1}}(I)$. It follows that

$$\langle \Gamma, T - I\rangle \geq -\eta_{\mathrm{opt}}\|T - I\|_F, \qquad \langle -\Gamma, T^{-1} - I\rangle \geq -\eta_{\mathrm{opt}}\|T^{-1} - I\|_F.$$

Repeating the argument for (17) gives

$$\operatorname{tr}[\mathsf{P}_\Phi(T + T^{-1} - 2I)] \leq F_\Phi(T) - F_\Phi(I) + \frac{\eta_{\mathrm{opt}}}{2}\big(\|T - I\|_F + \|T^{-1} - I\|_F\big). \tag{18}$$

If $F_\Phi(T) - \min_{\overline{\mathcal{Q}}} F_\Phi \leq \delta$, the right-hand side is at most $\delta + \eta_{\mathrm{opt}}d_{\mathcal{Q}}$, where

$$d_{\mathcal{Q}} = \tfrac{1}{2}\sup_{T \in \overline{\mathcal{Q}}}(\|T - I\|_F + \|T^{-1} - I\|_F).$$

Thus $\delta + \eta_{\mathrm{opt}}d_{\mathcal{Q}}$ replaces the objective gap in (6). The endpoints of the compact budget give a finite explicit bound on $d_{\mathcal{Q}}$. This result requires a valid support matrix and a true normal; numerical eigenvector or curvature errors need their own error bounds.

A local form of the residual term. Let T minimize F_Φ on $\overline{\mathcal{Q}}$, so that $F_\Phi(T) \le F_\Phi(I)$, and suppose $\mathsf{P}_\Phi \succ 0$ with $p_{\min} = \lambda_{\min}(\mathsf{P}_\Phi)$. Put $E = T - I$. Since $T + T^{-1} - 2I = ET^{-1}E$ and $T \preceq u_0 I$, the left-hand side of (18) is at least $p_{\min}\|E\|_F^2/u_0$, while $\|T^{-1} - I\|_F \le \|E\|_F/\ell_0$. Hence $\|E\|_F \le \eta_{\mathrm{opt}}(1 + \ell_0^{-1})u_0/(2p_{\min})$, the residual term is at most $\eta_{\mathrm{opt}}^2(1 + \ell_0^{-1})^2 u_0/(4p_{\min})$, and the response of every minimizer lies within

$$C_{\mathcal{Q}}\sqrt{\kappa_\Phi(D)}\;\eta_{\mathrm{opt}}\,(1 + \ell_0^{-1})\sqrt{u_0/(4p_{\min})}$$

of the reference response, a bound linear in the residual. This form and the uniform form $C_{\mathcal{Q}}\sqrt{\kappa_\Phi(D)\eta_{\mathrm{opt}} d_{\mathcal{Q}}}$ use only quantities at the reference, so either serves as a stopping rule (Appendix E.10).

A.6 An exact example separating the two objectives

Let $m = n = k = 2$, $U = V = I_2$, $M_* = \mathrm{diag}(1, -3)$, and

$$\ell(M) = \tfrac{1}{2}\sum_{i,j=1}^{2} c_{ij}(M_{ij} - (M_*)_{ij})^2, \qquad c_{11} = 1, \quad c_{12} = c_{21} = c_{22} = \tfrac{1}{4}.$$

At $M = I_2$ the actual gradient is $D = E_{22}$, and $J(X, Y) = X + Y$. The pullback has orthonormal positive modes $q_{ij} = (E_{ij}, E_{ij})/\sqrt{2}$ of eigenvalue $2c_{ij}$, and zero modes $(E_{ij}, -E_{ij})/\sqrt{2}$. The top mode is simple. Its two latent Grams are $\frac{1}{2}E_{11}$, yielding $\mathsf{P}_{\max} = E_{11}$. For the trace criterion, $B = H$; column and row sums of (c_{ij}) both yield $\mathsf{P}_{\mathrm{tr}} = \mathrm{diag}(5/4, 1/2)$. Both certificates have zero imbalance, proving that I is a global SPD minimizer of each objective. Since $Q_D = 2E_{22}$, the respective coverage coefficients are infinite and four.

For $T = \mathrm{diag}(1, t)$, its four nonzero transformed eigenvalues are

$$2, \quad \tfrac{1}{4}(1 + t^{-1}), \quad \tfrac{1}{4}(t + 1), \quad \tfrac{1}{4}(t + t^{-1}).$$

When $1/2 \le t \le 2$, the largest remains exactly two, whereas

$$F_{\mathrm{tr}}(T) - F_{\mathrm{tr}}(I) = \tfrac{1}{2}(t + t^{-1} - 2), \qquad \mathcal{P}_T(D) - \mathcal{P}_I(D) = (t + t^{-1} - 2)E_{22}.$$

All these coordinates are feasible in $\mathcal{Q}_{\sqrt{2}}$. For example, $t = 2$ has zero largest-eigenvalue spectral excess, trace gap $1/4$, and response distance $1/2$. The example is an exact substitution into the certificate, not an additional empirical performance result. It also shows a nonzero actual gradient compatible with a singular factor pullback. Different spectral objectives need not share minimizers in general; their common reference here is part of this example.

B Largest-eigenvalue optimization on a finite budget

This appendix proves Theorem 3. We keep the supplied PSD operator and the coordinate budget fixed throughout. The common spectral representation also justifies using any orthogonally invariant criterion in Section 3.

B.1 Finite-budget spectral reduction and convexity

Operator convexity and its Jensen inequalities are classical (Hansen & Pedersen, 2003); the inverse-convexity step used here is proved directly by a Schur complement below.

For the operator-composition calculation, let $\mathcal{U} = \mathbb{R}^{m\times k}$, $\mathcal{V} = \mathbb{R}^{k\times n}$, and write $\mathcal{R}_T(X) = XT$, $\mathcal{L}_T(Y) = TY$ on the respective factor spaces. Fix $G \in \mathrm{GL}(k)$, write $S = GG^\top$, and put $\Psi = H^{1/2}\mathcal{T}_G^{-1}$. Then

$$H_G = \Psi^*\Psi, \qquad \Psi\Psi^* = H^{1/2}\mathcal{T}_G^{-1}\mathcal{T}_G^{-*}H^{1/2}.$$

These PSD operators have the same eigenvalues, including zero multiplicities. Since right multiplication composes in reverse order,

$$\mathcal{T}_G^{-1}\mathcal{T}_G^{-*} = \mathcal{R}^{\mathcal{U}}_{G^{-\top}G^{-1}} \oplus \mathcal{L}^{\mathcal{V}}_{GG^\top} = \mathcal{R}^{\mathcal{U}}_{S^{-1}} \oplus \mathcal{L}^{\mathcal{V}}_{S} = \mathcal{W}(S). \tag{19}$$

Thus H_G, $L_H(S) = H^{1/2}\mathcal{W}(S)H^{1/2}$, and $K_H(S) = \mathcal{W}(S)^{1/2}H\mathcal{W}(S)^{1/2}$ have the same spectrum. Also $\|G\|_{\mathrm{op}}, \|G^{-1}\|_{\mathrm{op}} \le R$ exactly when $S \in \mathcal{Q}_R$, and every such S is realized by $G = S^{1/2}$. This proves the exact equality of gauge and SPD objectives in Theorem 3.

For $S_\theta = \theta S_1 + (1-\theta)S_2$, the Schur complement of the convex combination of $\left[\begin{smallmatrix} S_i^{-1} & I \\ I & S_i \end{smallmatrix}\right] \succeq 0$ gives

$$S_\theta^{-1} \preceq \theta S_1^{-1} + (1-\theta)S_2^{-1}.$$

Left and right multiplication by PSD matrices define positive operators on the corresponding matrix spaces. Hence

$$\mathcal{W}(S_\theta) \preceq \theta\mathcal{W}(S_1) + (1-\theta)\mathcal{W}(S_2).$$

Congruence by $H^{1/2}$, followed by monotonicity and convexity of $\lambda_{\max}$, proves convexity of f_H without a commutativity assumption. For finite R, the spectral box is compact and remains inside the SPD cone; continuity gives attainment. When $R = 1$, it is the singleton $\{I\}$.

B.2 The complete spectral-support balance criterion

The spectral subdifferential and dual-matrix representation used here are standard tools of nonsmooth eigenvalue optimization (Lewis, 1999; Overton & Womersley, 1993). We specialize them to the reciprocal factor action and derive the associated partial traces and box-boundary conditions explicitly.

For fixed $S \succ 0$, abbreviate $W = \mathcal{W}(S)$, $K = W^{1/2}HW^{1/2}$, $L = H^{1/2}WH^{1/2}$, and $\lambda = \lambda_{\max}(K)$. A top density is any

$$\Omega \succeq 0, \qquad \operatorname{Tr}\Omega = 1, \qquad K\Omega = \lambda\Omega.$$

For an operator Ξ on $\mathcal{U} \oplus \mathcal{V}$, let

$$[C_U(\Xi)]_{ab} = \sum_{i=1}^{m} \left\langle E_{ia}^U, \Xi^{UU} E_{ib}^U \right\rangle, \qquad [C_V(\Xi)]_{ab} = \sum_{j=1}^{n} \left\langle E_{aj}^V, \Xi^{VV} E_{bj}^V \right\rangle. \tag{20}$$

These are the partial traces of (3); for $\Omega = qq^\top$, $q = (X, Y)$, they equal $X^\top X$ and $YY^\top$.

Transport without a commutativity assumption. Set $C = W^{1/2}H^{1/2}$. A full singular-value decomposition gives an orthogonal polar factor O, even if H is singular, with

$$C = K^{1/2}O = OL^{1/2}, \qquad K = OLO^*.$$

Consequently $\widetilde{\Omega} = O^*\Omega O$ bijects top densities of K and L. Since $H^{1/2} = W^{-1/2}K^{1/2}O$,

$$H^{1/2}\widetilde{\Omega}H^{1/2} = W^{-1/2}K^{1/2}\Omega K^{1/2}W^{-1/2} = \lambda W^{-1/2}\Omega W^{-1/2}. \tag{21}$$

The last equality uses only the top-space support of Ω; no commutation between S and a latent Gram is involved.

The full subdifferential. The fixed compact density set gives the variational identity

$$f_H(S) = \lambda_{\max}(L) = \max_{\widetilde{\Omega}\succeq 0,\ \operatorname{Tr}\widetilde{\Omega}=1} \operatorname{Tr}(H^{1/2}\widetilde{\Omega}H^{1/2}\mathcal{W}(S)). \tag{22}$$

Each function inside the maximum is differentiable and convex in S. For $B = H^{1/2}\widetilde{\Omega}H^{1/2}$, the partial traces $P_U = C_U(B)$, $P_V = C_V(B)$ are PSD, and direct trace contraction gives, as in (3),

$$\operatorname{Tr}(B\mathcal{W}(S)) = \operatorname{tr}(P_V S) + \operatorname{tr}(P_U S^{-1}). \tag{23}$$

Its gradient is $P_V - S^{-1}P_U S^{-1}$. At a top density, (21) yields

$$P_U = \lambda S^{1/2}C_U(\Omega)S^{1/2}, \qquad P_V = \lambda S^{-1/2}C_V(\Omega)S^{-1/2}.$$

Thus define

$$\Gamma_{H,S}(\Omega) = \lambda S^{-1/2}\big[C_V(\Omega) - C_U(\Omega)\big]S^{-1/2}. \tag{24}$$

Danskin's theorem gives

$$Df_H(S;\Delta) = \max_{\Omega \text{ top density of } K} \langle \Gamma_{H,S}(\Omega), \Delta\rangle.$$

The top-density set and its linear image under Γ are compact and convex. Since f_H is convex, this directional derivative identifies the entire subdifferential:

$$\partial f_H(S) = \{\Gamma_{H,S}(\Omega) : \Omega \text{ top density of } K_H(S)\}. \tag{25}$$

If $H = 0$, then $K = L = 0$, $\lambda = 0$, and this image is $\{0\}$; every feasible point is optimal. Thus singular or zero operators create no missing subgradients.

Box boundaries and exact membership. For $R > 1$ the box has nonempty interior (Slater's condition), which gives the spectral-box normal cone

$$\mathcal{N}_{\mathcal{Q}_R}(S) = \left\{ \Lambda_u - \Lambda_l : \begin{array}{l} \Lambda_u, \Lambda_l \succeq 0, \\ \langle \Lambda_u, R^2 I - S \rangle = 0, \quad \langle \Lambda_l, S - R^{-2} I \rangle = 0 \end{array} \right\}. \tag{26}$$

The PSD multipliers are supported on the corresponding boundary eigenspaces. Convex optimality therefore gives membership in $\mathrm{Opt}_R(H)$ exactly when some top density and complementary multipliers satisfy

$$\Gamma_{H,S}(\Omega) + \Lambda_u - \Lambda_l = 0. \tag{27}$$

For $R = 1$, the normal cone at I is the whole symmetric space and membership is automatic.

The optimality residual. For a feasible S, the outward normal cone is

$$\mathcal{N}_{\mathcal{Q}_R}(S) = \{\nu \in \mathbb{S}^k : \langle \nu, T - S \rangle \le 0 \text{ for all } T \in \mathcal{Q}_R\}.$$

Let $\mathcal{D}_{\max}(K) = \{\Omega \succeq 0 : \operatorname{Tr}\Omega = 1, K\Omega = \lambda_{\max}(K)\Omega\}$. Define

$$\mathrm{Def}_{H,R}(S) = \min_{\substack{\Omega \in \mathcal{D}_{\max}(K_H(S)) \\ \nu \in \mathcal{N}_{\mathcal{Q}_R}(S)}} \|\Gamma_{H,S}(\Omega) + \nu\|_F. \tag{28}$$

Set $\Gamma = 0$ when $H = 0$; for $R = 1$ the normal cone at the sole feasible point I is the whole symmetric space. The subdifferential is compact and the normal cone is closed, so their sum is closed. The distance from zero to this nonempty closed set is attained in finite dimension. Consequently the minimum Frobenius residual vanishes exactly at an optimizer:

$$\mathrm{Def}_{H,R}(S) = 0 \quad \Longleftrightarrow \quad S \in \mathrm{Opt}_R(H). \tag{29}$$

Neither convexity nor a successful membership test implies uniqueness.

Repeated top eigenvalues. For $\lambda > 0$, if Q spans the complete top eigenspace orthonormally, all top densities are $\Omega = QZQ^\top$ with $Z \succeq 0$, $\operatorname{tr} Z = 1$. Off-diagonal entries of Z are allowed; one arbitrary top eigenvector need not supply a balanced density. At an interior point, the criterion is

$$C_U(\Omega) = C_V(\Omega) \quad \text{for some top density } \Omega. \tag{30}$$

For a simple positive top eigenvalue, the only density is $qq^\top$, so this becomes

$$X^\top X = YY^\top, \qquad q = (X, Y), \quad \|q\| = 1. \tag{31}$$

These are Grams of curvature-mode blocks, not of the factors; on a scalar slice $S = xI$ the criterion sees only $\operatorname{tr}[C_V(\Omega) - C_U(\Omega)]$.

C Local optimizer geometry and update identification

This appendix proves Propositions 4 and 5, including the full-active-support extension. It also derives the simultaneous-step identity used in the numerical controls. The reanchoring convention is given in Appendix A.1.

C.1 The full-active-support version and proof of the active-rank formula

Proposition 7 (Local optimizer set under full active support)**.** *Let $H \succeq 0$, $\lambda = \lambda_{\max}(H) > 0$, and $\mathcal{E} = \ker(H - \lambda I)$. Assume I is interior to the transported original budget. Suppose a trace-one PSD density Ω has range exactly $\mathcal{E}$ and is positive definite on that space, with $C_U(\Omega) = C_V(\Omega) = P$. Write $r = \operatorname{rank} P$, and let the columns of N be an orthonormal basis of $\ker P$. Then I is a global SPD minimizer. In a sufficiently small neighborhood $\mathcal{O}$ of I contained in the feasible interior, the* complete *optimizer set is*

$$\mathrm{Opt} \cap \mathcal{O} = \{I + NZ_0N^\top : Z_0 = Z_0^\top\} \cap \mathcal{O}.$$

Its dimension is $(k - r)(k - r + 1)/2$. If $r = k$, I is the unique global SPD minimizer and

$$f_H(T) - \lambda \ge \lambda\lambda_{\min}(P)\|\log T\|_F^2 \quad (T \succ 0). \tag{32}$$

Proof. The two spectral representations of Theorem 3 and $H^{1/2}\Omega H^{1/2} = \lambda\Omega$ give

$$f_H(T) \geq \mathrm{Tr}(\Omega H^{1/2}\mathcal{W}(T)H^{1/2}) = \lambda \,\mathrm{tr}[P(T + T^{-1})] \geq 2\lambda \,\mathrm{tr}\, P = \lambda. \tag{33}$$

Here $2\,\mathrm{tr}\, P = \mathrm{Tr}\,\Omega = 1$, and $T + T^{-1} - 2I \succeq 0$. If T is any global optimizer, the nonnegative trace $\mathrm{tr}[P(T + T^{-1} - 2I)]$ must vanish. The kernel of $T + T^{-1} - 2I$ is exactly $\ker(T - I)$, so $(T - I)P^{1/2} = 0$. Symmetry of T then gives $T = I + NZ_0N^\top$. This proves necessity globally.

For sufficiency locally, decompose $\Omega = \sum_j w_j q_j q_j^\top$ on the complete space $\mathcal{E}$, with $w_j > 0$ and $q_j = (X_j, Y_j)$. Nonnegativity gives

$$\ker P = \bigcap_j \ker X_j = \bigcap_j \ker Y_j^\top.$$

A matrix $T = I + NZ_0N^\top$ therefore satisfies $X_jT^{-1/2} = X_j$ and $T^{1/2}Y_j = Y_j$. The weight $\mathcal{W}(T)^{1/2}$ fixes *every* vector in $\mathcal{E}$, so $K_H(T)q_j = \lambda q_j$. The remaining eigenvalues stay strictly below λ near I by continuity and the gap below the complete top eigenspace. If the complementary space is empty, that step is vacuous. Hence all such nearby T are optimizers. The affine parameterization has the stated dimension. For full-rank P, scalar spectral calculus yields $T + T^{-1} - 2I \succeq (\log T)^2$, proving (32) and uniqueness. □

Simple mode and the coordinate Hessian. Under the hypotheses of Proposition 4, Theorem 3 gives $P = X^\top X = YY^\top$ and the full-active-support density is $qq^\top$. The minorant in (33) has second variation $2\lambda\,\mathrm{tr}(P\Delta^2)$. The difference $f_H(T) - \lambda\,\mathrm{tr}[P(T + T^{-1})]$ has a minimum of zero at I, so the coordinate Hessian of f_H dominates this quadratic form. It is strictly positive when $\Delta P^{1/2} \neq 0$. When $\Delta P^{1/2} = 0$, the entire short segment $I + t\Delta$ lies in the optimizer set, and the second variation is zero. Convexity makes the Hessian PSD, so these are exactly its kernel directions. In particular, for a simple active mode, full latent rank is equivalent to positive-definite coordinate curvature at an interior optimum.

Full active support and boundary conditions. For $m = n = k = 2$, take unit vectors $q_1 = (E_{11}, E_{11})/\sqrt{2}$, $q_2 = (E_{22}, 0)$, and $H = q_1q_1^\top + q_2q_2^\top$. The density $q_1q_1^\top$ is balanced and has latent rank one, but at $T = \mathrm{diag}(1, z)$ the objective is $\max\{1, z^{-1}\}$: only $z \geq 1$ is locally optimal, and no balanced density is positive on both active modes. A density supported on part of a repeated top eigenspace is thus insufficient for the two-sided local formula, and at a feasibility boundary the normal cone changes the dimension. The kernel of P does not depend on the choice among full-active-support densities, although the growth constant may.

C.2 One-sided quadratic curvature forces a rank-one active mode

Consider $\ell(M) = \frac{1}{2}\,\mathrm{tr}(M^\top P_0 M) - \langle C, M\rangle$, $P_0 \succ 0$, with full-column-rank U and full-row-rank V. Let $\mathcal{C}(Z) = P_0Z$. The nonzero spectrum of $H_{\mathrm{pb}} = J^*\mathcal{C}J$ agrees with that of $\mathcal{C}^{1/2}JJ^*\mathcal{C}^{1/2}$, whose action is

$$Z \longmapsto P_0Z(V^\top V) + P_0^{1/2}UU^\top P_0^{1/2}Z. \tag{34}$$

Diagonalizing $V^\top V$ splits this into blocks $\mu_j P_0 + P_0^{1/2}UU^\top P_0^{1/2}$. Their largest eigenvalue is strictly increasing in μ_j, since $P_0 \succ 0$. A simple overall top mode therefore lies in the top μ_j block and has rank-one form $zw^\top$. Pulling it back by $J^*\mathcal{C}^{1/2}$ gives blocks proportional to

$$X = P_0^{1/2}z(Vw)^\top, \qquad Y = (U^\top P_0^{1/2}z)w^\top.$$

Both have rank at most one. At an interior coordinate optimum their partial Grams coincide, and normalization makes their traces $1/2$. Their common rank is therefore exactly one. Applying Proposition 4 proves local dimension $k(k-1)/2$ under the stated simple-top and interior assumptions.

C.3 The exact simultaneous product step

Write $U_G = UG$, $V_G = G^{-1}V$, and $S = GG^\top$. For a fixed product gradient D, expansion of the two simultaneous factor-gradient updates gives

$$\begin{aligned}(U_G - \eta DV_G^\top)(V_G - \eta U_G^\top D) &= M - \eta\mathcal{P}_S(D) + \eta^2 DM^\top D,\\ \mathcal{P}_S(D) &= USU^\top D + DV^\top S^{-1}V.\end{aligned} \tag{35}$$

Indeed, $V_G^\top U_G^\top = M^\top$, so the quadratic term is independent of S. Subtracting this identity at two coordinates proves (1) and its exact displacement relation (Arora et al., 2018; Zhang et al., 2025).

C.4 From spectral accuracy to a fixed update

This subsection proves Proposition 5 and gives its quantitative form. All factors and operators are reanchored at the same exact interior optimum I. The domain is the transported original box $\overline{\mathcal{Q}} = \{T : R^{-2}S_0^{-1} \preceq T \preceq R^2 S_0^{-1}\}$, not a newly centered budget. Let $\ell_0 I \preceq T \preceq u_0 I$ on this domain, where one may take $\ell_0 = R^{-2}/\lambda_{\max}(S_0)$ and $u_0 = R^2/\lambda_{\min}(S_0)$, the constants of Theorem 1. Then $0 < \ell_0 < 1 < u_0$. The result applies both to the simple-mode case and to the balanced density of Proposition 7 that is positive on the entire top eigenspace. Let P be that density's common Gram, Π the orthogonal projector onto its range, $\Pi_0 = I - \Pi = NN^\top$, and $p_+ = \lambda^+_{\min}(P) > 0$. Since $\operatorname{tr} P = 1/2$, this positive part is nonempty even when P is singular. The gradient D is fixed across coordinates; it need not be the full-batch gradient of the loss whose curvature defines H.

A uniform bound on the original domain. Write $e(T) = f_H(T)/\lambda - 1$, $\lambda = f_H(I) > 0$, and $\mathcal{E}_D(T) = \|\mathcal{P}_T(D) - \mathcal{P}_I(D)\|_F$. Specializing (17) to $\mathsf{P}_\Phi = \lambda P$ gives

$$\begin{aligned} e(T) &\geq \operatorname{tr}[P(T + T^{-1} - 2I)] = \|T^{-1/2}(T - I)P^{1/2}\|_F^2, \\ \|(T - I)\Pi\|_F &\leq \sqrt{u_0 e(T)/p_+}. \end{aligned} \tag{36}$$

For the second bound, use $T^{-1} \succeq u_0^{-1} I$ and $P \succeq p_+ \Pi$ in the squared norm; no commutation of P, T is needed. Define

$$\begin{aligned} L_U &= \|\Pi_0 U^\top D\|_F, & L_V &= \|DV^\top \Pi_0\|_F, \\ c_{D,\mathrm{act}} &= \|U\|_{\mathrm{op}}\|\Pi U^\top D\|_{\mathrm{op}} + \ell_0^{-1}\|V\|_{\mathrm{op}}\|DV^\top \Pi\|_{\mathrm{op}}, & & \\ c_T &= \max(u_0 - 1, 1 - \ell_0), & c_{-1} &= \max(\ell_0^{-1} - 1, 1 - u_0^{-1}). \end{aligned}$$

Then for every $T \in \overline{\mathcal{Q}}$,

$$\mathcal{E}_D(T) \leq c_{D,\mathrm{act}}\sqrt{u_0 e(T)/p_+} + c_T\|U\|_{\mathrm{op}} L_U + c_{-1}\|V\|_{\mathrm{op}} L_V. \tag{37}$$

Indeed, decompose both factor-gradient blocks using $\Pi + \Pi_0 = I$. The supported terms are bounded by (36) and $(T^{-1} - I)\Pi = -T^{-1}(T - I)\Pi$, with $\|T^{-1}\|_{\mathrm{op}} \leq \ell_0^{-1}$. Symmetry bounds the corresponding left-supported inverse term by the same Frobenius norm. The remaining terms use $\|T - I\|_{\mathrm{op}} \leq c_T$ and $\|T^{-1} - I\|_{\mathrm{op}} \leq c_{-1}$. Thus zero leakage gives the asserted uniform $O(\sqrt{\varepsilon})$ rate. The constants concern this one operator and its reference optimum; a value certificate for another operator cannot be substituted into (37).

Necessity and an explicit positive floor. When $\Pi_0 \neq 0$, Proposition 4 supplies spectrally optimal $T(t) = I + t\Pi_0$ for both signs of sufficiently small t. Set $A_\Pi = U\Pi_0 U^\top D$ and $B_\Pi = DV^\top \Pi_0 V$. Exactly,

$$\mathcal{P}_{T(t)}(D) - \mathcal{P}_I(D) = tA_\Pi - \frac{t}{1+t} B_\Pi. \tag{38}$$

If $\omega_D(\varepsilon) \to 0$, every one of these exact minimizers has zero response difference, so $(1+t)A_\Pi - B_\Pi = 0$ throughout a small interval. Two different t's force $A_\Pi = B_\Pi = 0$. No factor-rank assumption is needed: the identities

$$\langle D, A_\Pi\rangle = \|\Pi_0 U^\top D\|_F^2, \qquad \langle D, B_\Pi\rangle = \|DV^\top \Pi_0\|_F^2$$

show that $A_\Pi = B_\Pi = 0$ implies $\Pi_0 U^\top D = 0$ and $DV^\top \Pi_0 = 0$. These are equivalent to the two conditions in (10). For $\Pi_0 = 0$, both conditions are vacuous and (37) proves sufficiency directly. This establishes the equivalence, including the repeated-mode extension.

For an explicit radius, let $\gamma > 0$ be the gap below the complete top eigenspace and set

$$\delta_{\mathrm{box}} = \min\{\lambda_{\min}(I - \underline{T}), \lambda_{\min}(\overline{T} - I)\}.$$

Here $\underline{T} = R^{-2}S_0^{-1}$, $\overline{T} = R^2 S_0^{-1}$. If $\Pi_0 \neq 0$, full active support leaves a nonempty complementary spectral space: were the top eigenspace the whole factor space, the modes of the positive-definite

density would span every X, and $X\Pi_0 = 0$ for all of them would force $\Pi_0 = 0$. For $0 < a < \min\{1, \delta_{\mathrm{box}}, \gamma/(\lambda+\gamma)\}$, the whole free block is optimal by the radius proof in Appendix D.2. In particular, $T_\pm = I \pm a\Pi_0$ are feasible exact optimizers. For $F_\pm = \mathcal{P}_{T_\pm}(D) - \mathcal{P}_I(D)$,

$$\frac{F_+ + F_-}{2} = \frac{a^2 B_\Pi}{1-a^2}, \qquad \frac{(1+a)F_+ + (1-a)F_-}{2} = a^2 A_\Pi,$$
$$\frac{F_+ - F_-}{2} = a\Big(A_\Pi - \frac{B_\Pi}{1-a^2}\Big).$$

Triangle inequalities therefore give

$$\omega_D(0) \geq \max\{\mathcal{E}_D(T_+), \mathcal{E}_D(T_-)\} \geq \max\left\{a^2\|A_\Pi\|_F, \frac{a^2\|B_\Pi\|_F}{1-a^2}, a\left\|A_\Pi - \frac{B_\Pi}{1-a^2}\right\|_F\right\}. \quad (39)$$

This is positive whenever either leakage is nonzero. It compares I, T_+, T_-; it is not the same as a bound on the distance between the two endpoint responses alone. In particular, $A_\Pi = B_\Pi \neq 0$ cancels the first derivative, but (38) still equals $t^2 A_\Pi/(1+t)$. A first-order test would miss this case.

First-order choice on the free subspace. Pairing (38) with D at $t = \pm a$, and using $\langle D, A_\Pi\rangle = \|\Pi_0 U^\top D\|_F^2 = \operatorname{tr} A_D$ and $\langle D, B_\Pi\rangle = \|DV^\top \Pi_0\|_F^2 = \operatorname{tr} B_D$, gives the identity in (15). For $0 < a < 1$ the weight $w = \frac{1}{2} + a/[2(2-a^2)]$ lies in $(0,1)$, and

$$w\,[G_D(T_+) - G_D(I)] + (1-w)\,[G_D(T_-) - G_D(I)] = \frac{a^2}{2-a^2}(\operatorname{tr} A_D + \operatorname{tr} B_D),$$

which the larger of the two brackets dominates. Under zero leakage, (37) vanishes at every exact optimizer, so all of them give the same response and the same G_D. For smooth ℓ and $D = E$, the exact one-step identity of Appendix C.3 gives $\ell(M_T^+) = \ell(M) - \eta G_E(T) + O(\eta^2)$ on the three coordinates, hence the first-order statement; for a quadratic loss the coordinate-dependent part of the remainder is $\frac{\eta^2}{2}\langle \mathcal{P}_T(E), \mathcal{C}\,\mathcal{P}_T(E)\rangle + O(\eta^3)$. Factor stationarity gives $U^\top E = 0$ and $EV^\top = 0$, so the gain vanishes there. Away from stationarity the statement needs a PSD operator with the hypotheses above, such as the pullback of a convex loss.

Zero-leakage controls. Let Π_L, Π_R project orthogonally onto $\operatorname{col}(UN)$ and $\operatorname{col}(V^\top N)$. The Frobenius-nearest gradient satisfying both zero-leakage conditions is $D_{\mathrm{safe}} = (I - \Pi_L)D(I - \Pi_R)$, since the two constraints define commuting orthogonal projections. This is a diagnostic control, not a replacement for the training gradient.

A displacement-tolerance rule. Set $b_{D,\mathrm{leak}} = c_T\|U\|_{\mathrm{op}}L_U + c_{-1}\|V\|_{\mathrm{op}}L_V$, the leakage term in the upper bound (37). At a common step size $\eta > 0$, the coordinate-independent quadratic term cancels, so $\|\Delta_T - \Delta_I\|_F = \eta\mathcal{E}_D(T)$ exactly. To ensure a displacement tolerance $\tau_\Delta > 0$ for every candidate with $e(T) \leq \varepsilon$, a sufficient condition, when $c_{D,\mathrm{act}} > 0$ and $b_{D,\mathrm{leak}} < \tau_\Delta/\eta$, is

$$\varepsilon \leq \frac{p_+}{u_0}\left(\frac{\tau_\Delta/\eta - b_{D,\mathrm{leak}}}{c_{D,\mathrm{act}}}\right)^2. \quad (40)$$

If $c_{D,\mathrm{act}} = 0$, use $\eta b_{D,\mathrm{leak}} \leq \tau_\Delta$ directly. The term $b_{D,\mathrm{leak}}$ is not an error floor. Let F_D denote the rightmost maximum in (39). When $\eta F_D > \tau_\Delta$, even zero spectral excess cannot give a uniform tolerance on the optimizer set; this does not assert that every optimizer has a large response difference. When neither bound decides the comparison, or the exact-reference hypotheses are unverified, the test is inconclusive. A relative tolerance additionally requires a specified nonzero reference norm.

Matching with two scalar factor rates. Fix the same product $M = UV$ and product gradient D, and consider one simultaneous ordinary factor-gradient step. Write $A_0 = UU^\top D$, $B_0 = DV^\top V$, and $R_D = DM^\top D$. If the reference factors U, V use rates $s_U, s_V \geq 0$, respectively, their exact product displacement and the displacement at coordinate T with common rate $\eta > 0$ are

$$\Delta_0(s_U, s_V) = -s_V A_0 - s_U B_0 + s_U s_V R_D, \qquad \Delta_T = -\eta\mathcal{P}_T(D) + \eta^2 R_D.$$

This is the simultaneous-step identity, not a first-order approximation (Arora et al., 2018; Zhang et al., 2025). In the Frobenius matrix space put $\mathcal{L}_D = \operatorname{span}\{A_0, B_0, R_D\}$. Then

$$\inf_{s_U, s_V \geq 0}\|\Delta_T - \Delta_0(s_U, s_V)\|_F \geq \eta\big\|\mathrm{Proj}_{\mathcal{L}_D^\perp}\mathcal{P}_T(D)\big\|_F. \quad (41)$$

Indeed, projecting the difference onto $\mathcal{L}_D^{\perp}$ annihilates both the reference displacement and R_D. The bound holds for any SPD T, without an optimality or active-rank assumption. A positive right-hand side rules out reproducing that displacement with two scalar rates; zero does not establish reproducibility because the attainable two-rate set is only a nonlinear subset of $\mathcal{L}_D$. This comparison concerns neither the best achievable loss nor stateful optimizers or multi-step learning-rate schedules. In particular, a different reference update can have a lower loss even when exact displacement matching is impossible.

D Extremal response on a spectrally optimal or near-optimal block

D.1 Proof of Theorem 6

For orthonormal N, $T_Z^{-1} = I - NN^\top + NZ^{-1}N^\top$. Substitution into the product response proves

$$g(Z) = \operatorname{tr}[A_D(Z - I) + B_D(Z^{-1} - I)].$$

For $l \le z \le h$, convexity of the reciprocal function gives $z^{-1} \le (l+h)/(lh) - z/(lh)$. Applying this scalar inequality in the eigenbasis of Z gives

$$Z^{-1} \preceq \frac{l+h}{lh} I - \frac{Z}{lh}. \tag{42}$$

Since $B_D \succeq 0$, the resulting affine upper bound for g has linear term $\operatorname{tr}(C_D Z)$, with $C_D = A_D - B_D/(lh)$. For $lI \preceq Z \preceq hI$, this linear term is maximized by $Z_* = lI + (h-l)\Pi_+(C_D)$. Every eigenvalue of Z_* is either l or h, so equality holds in (42) at that point. The affine upper bound is attained, proving global optimality on the block box and the value in (12). This argument does not require simultaneous diagonalization of A_D and B_D.

Let $c_+ = \operatorname{tr}((C_D)_+)$ and $c_- = \operatorname{tr}((-C_D)_+)$. Subtracting the value at lI from the value at hI gives $(h-l)(c_+ - c_-)$, while the improvement of Z_* over lI is $(h-l)c_+$. Subtracting the better endpoint proves (13). It is strictly positive if and only if $c_+ > 0$ and $c_- > 0$. In particular it is zero for $d = 1$. For $d = 0$, the only coordinate is I, and all gains are zero.

Zero eigenvalues of C_D may be assigned either endpoint without changing this construction. They do not in general allow arbitrary interior values: for scalars $A_D = B_D/(lh) > 0$, the two endpoints have equal value and an interior point has a strictly smaller value. Thus the projector formula supplies an extremal representative, not a complete parameterization of every possible optimizer in degenerate cases.

D.2 A spectral budget for the entire block

Assume the simple interior setting of Proposition 4. Write $\lambda = \lambda_{\max}(H) > 0$, $\gamma = \lambda - \lambda_2(H) > 0$, and let the unit top mode be $q = (X, Y)$, with $X^\top X = YY^\top = P$. For an arbitrary orthonormal N, set $\rho_N = \operatorname{tr}(N^\top P N)$. Let $0 < a \le 1/2$ and $(1-a)I \preceq Z \preceq (1+a)I$. Put

$$E_T = H^{1/2}(\mathcal{W}(T_Z) - I)H^{1/2}.$$

The spectra of T_Z and T_Z^{-1} imply $\|E_T\|_{\mathrm{op}} \le \lambda a/(1-a)$. Under $a \le \gamma/(4\lambda + \gamma)$, this is at most $\gamma/4$. Since $H^{1/2}q = \sqrt{\lambda}q$, the top Rayleigh increment is

$$e = q^\top E_T q = \lambda \operatorname{tr}[P(T_Z + T_Z^{-1} - 2I)] \le 2\lambda\rho_N a^2, \qquad e \ge 0.$$

Here $z + z^{-1} - 2 = (z-1)^2/z \le 2a^2$ on the block and is zero on its complement. Similarly,

$$\|E_T q\|^2 \le \lambda^2 \|(\mathcal{W}(T_Z) - I)q\|^2 \le \lambda^2 \rho_N \left[a^2 + a^2/(1-a)^2\right] \le 5\lambda^2\rho_N a^2.$$

The compression of $H + E_T$ to $q^\perp$ has largest eigenvalue at most $\lambda - 3\gamma/4$. Relative to $\operatorname{span}\{q\} \oplus q^\perp$, write

$$H + E_T = \begin{bmatrix} \lambda + e & b^* \\ b & C \end{bmatrix}, \qquad \|b\| \le \|E_T q\|, \qquad \lambda_{\max}(C) \le \lambda - 3\gamma/4.$$

If $\mu = \lambda_{\max}(H + E_T)$, then $\mu \geq \lambda + e \geq \lambda$, hence $\mu I - C \succeq (3\gamma/4)I$. For a top eigenvector with nonzero q-component, the Schur complement gives $\mu - (\lambda + e) = b^*(\mu I - C)^{-1}b$; a zero q-component would instead give $\mu \leq \lambda - 3\gamma/4$. Therefore

$$\mu - \lambda \leq e + \frac{4}{3\gamma}\|E_T q\|^2 \leq \left(2\lambda + \frac{20\lambda^2}{3\gamma}\right)\rho_N a^2.$$

The equally sized matrices $H^{1/2}\mathcal{W}(T_Z)H^{1/2}$ and $K_H(T_Z)$ have the same spectrum, so division by λ proves (14). The estimate is uniform over all orientations and eigenvalues in the block box, not just a scalar path.

To respect the transported original budget, one sufficient radius is

$$a \leq \delta_{\mathrm{box}} := \min\{\lambda_{\min}(I - \underline{T}), \lambda_{\min}(\overline{T} - I)\}.$$

It is positive at the interior reference. Given $\varepsilon > 0$ and $\rho_N > 0$, the minimum of δ_{box}, $1/2$, $\gamma/(4\lambda + \gamma)$, and $\sqrt{\varepsilon/(C_\gamma \rho_N)}$ thus supplies a valid radius. The same bound applies when P is full rank but weakly covers the chosen block. A small gap can make this sufficient radius conservative.

A larger radius inside the free subspace. If the columns of N lie in $\ker P$, allow $0 < a < 1$. Then both $\mathcal{W}(T_Z)^{1/2}q = q$ and $K_H(T_Z)q = \lambda q$ hold exactly. For $W_Z = \mathcal{W}(T_Z)^{1/2}$,

$$\|W_Z H W_Z - H\|_{\mathrm{op}} \leq \lambda a/(1 - a).$$

Consequently, the less restrictive bound $0 < a < \min\{1, \delta_{\mathrm{box}}, \gamma/(\lambda + \gamma)\}$ keeps every complementary eigenvalue below λ. The entire block box is then exactly optimal. The argument also applies to the complete top eigenspace under the full-active-support conditions of Appendix C, with γ then denoting the gap below the complete top eigenspace. Taking $Z = (1 \pm a)I$ also proves the endpoint radius used in Appendix C.4, independently of the response bound.

D.3 Local loss comparison

By (35), the difference between an extremal and a reference product step is exactly $-\eta[\mathcal{P}_{T_Z}(D) - \mathcal{P}_I(D)]$. For $D = \nabla\ell(M)$ and a twice continuously differentiable loss, Taylor expansion on the compact block box gives, uniformly in Z,

$$\ell(M + \Delta M_{T_Z}) - \ell(M + \Delta M_I) = -\eta g(Z) + O(\eta^2).$$

The extremal representative therefore maximizes the first-order loss decrease. A strict pairing advantage yields a loss advantage for sufficiently small step sizes, not for arbitrary step sizes or trajectories. Candidate-value certification at numerical references is a separate procedure (Appendix F).

E Fixed-checkpoint study and mechanism controls

E.1 Data, loss and coordinate solver

The data are the 1,797-example optical-digit subset distributed with scikit-learn (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). Pixel counts are divided by 16. Seeds $17, 29, 43$ define stratified 75/25 splits with 1,347 training and 450 held-out examples. The held-out rows serve only to report each fitted head's accuracy (37–38% at rank two and 71–72% at rank four, within 1.6 points of training accuracy); no diagnostic uses them. Training-mean centering and the top 12 unwhitened principal components are fitted on training rows only; the model has no intercept. The supplementary material contains the raw data, split indices, preprocessing arrays and the dataset attribution.

For training features $X_{\mathrm{tr}} \in \mathbb{R}^{s\times m}$, $m = 12$, and one-hot labels $Y_{\mathrm{tr}} \in \mathbb{R}^{s\times 10}$ (not to be confused with the mode blocks X, Y of the main text), we use ranks $k = 2, 4$ and

$$\ell(M) = \frac{1}{2s}\|X_{\mathrm{tr}}M - Y_{\mathrm{tr}}\|_F^2 + \frac{1}{200}\|M\|_F^2, \quad P_0 = X_{\mathrm{tr}}^\top X_{\mathrm{tr}}/s + I/100, \quad C_{\mathrm{tr}} = X_{\mathrm{tr}}^\top Y_{\mathrm{tr}}/s. \tag{43}$$

Thus $D = P_0 UV - C_{\mathrm{tr}}$, $\mathcal{C}(Z) = P_0 Z$ and $H = J^*\mathcal{C}J \succeq 0$. All coordinate comparisons use this pullback, the same product and gradient, and the original budget $\mathcal{Q}_2$. The rank-constrained optimum used to scale the initialization is

$$M^* = P_0^{-1/2}[P_0^{-1/2}C_{\mathrm{tr}}]_k,$$

where brackets denote a best rank-k truncated SVD. This follows by completing the square in (43); whitening is used only for this product solution, not to change the factor-space metric.

The coordinate solver minimizes the spectral objective in epigraph form over all symmetric $k \times k$ coordinates, using SLSQP with at most 400 iterations and stopping tolerance 10^{-12}. All eigenvalue inequalities enter the epigraph; candidates are projected onto the original budget before evaluation. Feasible density lower bounds are computed independently of solver status. For a clustered top spectrum, the numerical density is optimized on the cluster and then made PSD and trace one. Membership in the exact top eigenspace affects tightness, not validity, of the global value minorant in Appendix F.

E.2 General-certificate diagnostic on the frozen checkpoints

This diagnostic uses exactly the same 42 factor pairs and product gradients as the training-checkpoint study below, but re-solves only the small coordinate problem for two full-support criteria. No factor training is repeated. For trace, write the two partial traces of the supplied pullback as $A, D_0 \succ 0$, so

$$F_{\mathrm{tr}}(S) = \mathrm{tr}(AS^{-1}) + \mathrm{tr}(D_0 S).$$

All 42 unconstrained minimizers lie strictly inside $\mathcal{Q}_2$ and are therefore the budget optima. We evaluate them in closed form as

$$S_{\mathrm{tr}} = D_0^{-1/2}(D_0^{1/2} A D_0^{1/2})^{1/2} D_0^{-1/2}, \qquad S_{\mathrm{tr}} D_0 S_{\mathrm{tr}} = A.$$

For the Schatten-2 norm $\Phi(A) = \|A\|_F$, the minimizer is the same as for the smooth convex surrogate $\frac{1}{2}\,\mathrm{Tr}(A^2)$. We solve that equivalent full-SPD coordinate problem with an analytic gradient on the same spectral box, then evaluate the Schatten-2 certificate itself. All 42 solutions are interior; every certificate KKT residual is below 1.6×10^{-8}.

For each criterion, the stored $\lambda_{\max}$ coordinate S^* is treated as a natural feasible comparator. We reanchor at the full-support optimum, compute P_Φ, Q_D, $\kappa_\Phi(D)$ and $C_{\mathcal{Q}}$, and compare the observed response distance with the right-hand side of (6). For the numerical Schatten-2 references we include the residual term from Appendix A.5; its median contribution to the objective gap is below 2.5×10^{-9} and its maximum below 8.4×10^{-9}. All 84 individual bounds contain the observed response distance. The complete per-checkpoint results are in the supplementary material; Table 1 reports the medians used in the main text. This is a floating-point diagnostic of the theorem, distinct from the exact rational candidate-value certificates in Appendix F.

Table 4 places the 84 bounds between two baselines computed at the same frozen checkpoints. Since both coordinates lie in $\mathcal{Q}_R$, $\|S - S'\|_{\mathrm{op}}$ and $\|S^{-1} - S'^{-1}\|_{\mathrm{op}}$ are at most $R^2 - R^{-2}$, so the budget alone bounds every response difference by $\Delta_{\mathcal{Q}} = (R^2 - R^{-2})(\|U\|_{\mathrm{op}}\|U^\top D\|_F + \|V\|_{\mathrm{op}}\|DV^\top\|_F)$ in the original gauge. The certificate improves on this spectrum-free diameter in 83 of 84 cases but, at the comparators' objective excess of 1.6%–10.7%, exceeds the norm of the reference response itself in all 84. Because the bound scales as the square root of the objective gap, the relative excess at which the same certificate would guarantee a 1% response error is $\varepsilon_{1\%} = \varepsilon\,(0.01\,\|\mathcal{P}_I(D)\|_F/\beta)^2$, where β is the right-hand side of (6) at excess ε; $\varepsilon_{1\%}$ lies between 1.9×10^{-7} and 3.0×10^{-6} over the 84 comparisons. This is an idealized extrapolation at an exact reference: with an inexact reference the residual term of Appendix A.5 is added to the gap and the attainable tolerance is correspondingly larger; the trace references here are closed-form and the Schatten-2 references carry the residual term above.

Table 4: Two baselines for the 84 certificate comparisons of Table 1. Diameter is the response bound that the budget $\mathcal{Q}_2$ alone implies without spectral information, $\Delta_{\mathcal{Q}} = (R^2 - R^{-2})(\|U\|_{\mathrm{op}}\|U^\top D\|_F + \|V\|_{\mathrm{op}}\|DV^\top\|_F)$; reference is the norm of the full-support optimum's response; the bound is below the diameter in 83 of the 84 comparisons (all but one Schatten-2 rank-four checkpoint). Each entry is the median over the 21 checkpoints of that rank with the range in parentheses. Excess is the comparator's relative objective excess; excess for 1% is the relative objective excess at which the same certificate would guarantee a response error of 1% of the reference, from the square-root law of (6); it is an idealized extrapolation at an exact reference, since an inexact reference adds the optimality-residual term of Appendix A.5 (KKT residuals of these references in Appendix E.2).

Criterion	Rank	bound / diameter	bound / reference	excess	excess for 1%
Trace	2	0.24 (0.15–0.4)	1.4 (1–2.1)	3.8% (1.6–7.2)	1.6×10^{-6} (1.0×10^{-6}–3.0×10^{-6})
Trace	4	0.42 (0.28–0.83)	2.4 (1.7–4.6)	5.5% (2.7–8.8)	8.3×10^{-7} (3.2×10^{-7}–1.5×10^{-6})
$\|A\|_F$ (Schatten-2)	2	0.35 (0.2–0.46)	1.9 (1.1–3)	4.4% (1.6–8.4)	1.4×10^{-6} (6.1×10^{-7}–2.1×10^{-6})
$\|A\|_F$ (Schatten-2)	4	0.57 (0.38–1.1)	3.4 (2.2–6.2)	6.8% (3.4–10.7)	5.0×10^{-7} (1.9×10^{-7}–9.7×10^{-7})

E.3 Training checkpoints and uniform-sign comparisons

For each split and rank, independent standard-normal factor entries are scaled by $\sqrt{\|M^*\|_F/\sqrt{mn}}/\sqrt{k}$, using random seed $1000 + 10 \cdot \mathrm{seed} + k$. Plain full-batch gradient descent takes 200 steps at $\eta = 0.1$. The seven checkpoints are $t \in \{0, 5, 10, 20, 50, 100, 200\}$, giving 42 comparisons. At each checkpoint the pullback coordinate solve is warm-started from the preceding checkpoint and returns a feasible reference S^*.

After reanchoring, average the two top-mode Grams and take the projector $\Pi_0 = NN^\top$ onto its numerical kernel (the inferred free subspace), with active-rank threshold $10^{-10}\max(1, \lambda_{\max}(P))$. Let λ and γ be the top eigenvalue and top gap, and let δ_{box} be the transported budget margin from Appendix D.2. The radius rule is

$$a = \min\{0.1, \delta_{\mathrm{box}}/2, \gamma/[2(\lambda + \gamma)]\}, \qquad S_\pm = (S^*)^{1/2}(I \pm a\Pi_0)(S^*)^{1/2}.$$

The inferred active rank is one at all checkpoints; relative gaps exceed 2.9×10^{-3}, margins exceed 0.30, and mode-balance residuals are below 10^{-6}. The two candidates change the floating-point pullback value by a relative amount below 5×10^{-15}. Exact candidate-value bounds, rather than these numerical geometry checks, establish the reported relative spectral excess. Those bounds are deliberately conservative, rigorous enclosures that sit eight to ten orders of magnitude above the floating-point indication; every reported excess is such an upper bound.

For each sign we compute the response and exact product step using (35). The uniform endpoint with the larger pairing is determined by (15): 26 checkpoints use plus and 16 use minus. Direct pairing and the closed form agree to relative error below 4×10^{-12}. The response-change range in Table 5 covers both signs; the gain column is the extra one-step decrease at that fixed endpoint, divided by the reference one-step decrease. The realized column divides this extra decrease by its first-order prediction. At $\eta = 0.1, 0.01, 0.001$, the realized fractions range over $[0.65, 1.02]$, $[0.96, 1.002]$, and $[0.996, 1.0002]$, respectively. These finite-step comparisons keep all coordinates and gradients fixed.

E.4 Block-extremum comparison at the same checkpoints

The block-extremum comparison keeps each of the 42 stored products, gradients, references and radii unchanged. With the inferred basis N, form A_D, B_D, take $l = 1 - a, h = 1 + a$, and evaluate (11). The original-coordinate candidate is $S_{\mathrm{block}} = (S^*)^{1/2}T_{Z_*}(S^*)^{1/2}$. The pairing comparator is the better of the original uniform endpoints. The one-step consistency comparator is the same endpoint with the larger pairing from (15); it is fixed before inspecting finite-step loss. All finite-step losses use the same $\eta = 0.1$.

A strict numerical increase uses a threshold of 10^{-12} after division by $G_D(I)$ for pairing, or by the reference one-step decrease for loss comparisons. At rank two the one-dimensional block reduces to the uniform signs, with differences at roundoff level. At rank four, 20 of 21 matrices C_D have eigenvalues of both signs and the same 20 improve pairing; the exception is seed 43, step 50. The extra pairing has median 0.280572% and maximum 0.981268% of $G_D(I)$. The extra loss decrease

Table 5: Spectral precision and response freedom along gradient descent ($\eta = 0.1$). Checkpoint columns report the largest certified pullback spectral excess of S^*, $S_\pm$, the response change at both signs, the extra one-step decrease at the uniform endpoint with the larger pairing in (15), and the realized fraction of its first-order prediction. The last two columns are per-head medians over the seven chosen uniform-endpoint candidates (Section 6 pools all 42 states instead: median 1.36%, maximum 4.07% for the two-rate bound): the minimum common-rate matching error and the lower bound for two scalar factor rates. Both are normalized by the reference finite-step displacement norm (Appendix E.5); the matching columns quantify displacement components that scalar rates cannot reproduce.

		Checkpoints				Finite-step matching (%)	
Seed	Rank	spectral excess (10^{-6})	response (%)	gain (%)	realized	common	two-rate LB
17	2	4.2	0.10–3.38	0.03–1.13	0.93–1.01	1.570	1.316
17	4	1.2	0.33–2.13	0.01–0.26	0.66–1.00	1.071	0.767
29	2	0.5	0.07–3.12	0.02–1.50	0.98–0.99	2.359	2.042
29	4	6.7	1.02–4.25	0.25–0.95	0.98–0.99	2.615	1.959
43	2	1.1	0.47–4.94	0.05–1.40	0.82–1.00	1.539	1.244
43	4	2.1	1.16–4.73	0.19–1.66	0.93–1.00	2.057	1.769

relative to the reference decrease has median 0.294426% and maximum 0.974333%. For rank four, the tested radii themselves have median 0.0395108 and range $[0.0050460, 0.0746733]$. Dividing the extra pairing by the actual width $2a$ at each checkpoint gives median 4.98097% and maximum 17.0316% per unit width. These are the training-checkpoint radii; the smaller radii in the separate stationary mechanism controls of Appendix E.6 are not used in this normalization.

Pairing and one-step loss use the floating-point extremal representatives. Their 40-bit dyadic roundings each receive a separate spectral witness, not an original uniform-sign witness. The largest observed coordinate rounding error is below 9.92×10^{-13} in operator norm. Recomputing the global lower bound, candidate upper bound and stored-loss bridge gives a maximum certified pullback spectral excess of 6.697323×10^{-6}, rounded upward, on $\mathcal{Q}_2$. The common certification method and its scope are specified in Appendix F.

E.5 Finite-step scalar-rate controls

Using the same candidates and originally selected uniform signs, form A_0, B_0, R_D from Appendix C.4. The common-rate column of Table 5 is

$$100 \frac{\min_{s \geq 0} \|\Delta_T - \Delta_0(s,s)\|_F}{\|\Delta_0(\eta,\eta)\|_F},$$

computed by evaluating all nonnegative real stationary points and the endpoint of the quartic squared distance. The two-rate column is $100\eta \| \operatorname{Proj}_{\mathcal{L}_D^\perp} \mathcal{P}_T(D)\|_F / \|\Delta_0(\eta,\eta)\|_F$, the lower bound in (41), not the optimized two-rate distance. Both table entries are medians over the seven chosen uniform-endpoint candidates per head.

All 84 sign-specific results are indexed by split, rank, step and sign. Normalized-column SVD gives the projection, with an independent QR check. The three-column spans have smallest singular value above 0.45, and reference displacement norms exceed 1.3×10^{-3}. These controls distinguish coordinate effects from scalar-rate matching; they do not compare the best losses attainable by other updates.

E.6 Local geometry and zero-leakage controls

The six analytic products M^* provide separate mechanism controls. Starting with their SVD-balanced factors, refine the pullback coordinates in directions of positive numerical coordinate curvature. The active-rank threshold is as above; coordinate-Hessian nullity uses $10^{-8} \max(1, \|\mathcal{A}\|_{\mathrm{op}})$, where $\mathcal{A}$ is the numerical coordinate Hessian of f_H at the reference. The resulting active ranks and nullities are reported in Table 6. These numerical refinements are not the training-checkpoint references.

Table 6: Active rank and coordinate-Hessian nullity after post-hoc local refinement of the six digit-head pullback candidates. Numerical ranks are compared with the ideal local-dimension formula using the thresholds in the text. These static controls use different references from the training checkpoints.

Seed	Model rank	Active rank	Numerical nullity	Predicted dimension
17	2	1	1	1
17	4	1	6	6
29	2	1	1	1
29	4	1	6	6
43	2	1	1	1
43	4	1	6	6

Table 7: Fixed-batch response diagnostic on the six stationary models. Each model uses 22 disjoint batches (21 of size 64 and one of size 3) and the same two coordinates chosen from its spectral gap and original budget. Response percentages are the larger of the two changes relative to the anchor update. The last columns give median lower/observed and observed/upper ratios from (39) and (37), evaluated as numerical diagnostics. Batches are within-model repeats, and the six models are dependent fits to one dataset.

Seed	Rank	Radius	Median (%)	Range (%)	Lower/obs.	Obs./upper
17	2	0.01914	1.296	0.373–1.604	0.9976	0.00460
17	4	0.00446	0.342	0.222–0.397	0.9996	0.00078
29	2	0.02556	1.749	1.250–2.192	0.9966	0.00605
29	4	0.00297	0.221	0.178–0.249	0.9997	0.00051
43	2	0.02169	1.340	0.681–1.923	0.9978	0.00508
43	4	0.00346	0.259	0.217–0.313	0.9996	0.00060

Fixed-batch responses. At each refined reference, fix the same two uniform coordinates from the radius rule before evaluating gradients. Split the training rows into 22 consecutive batches: 21 of size 64 and one of size 3. Each batch gradient includes the unchanged ridge term. Also evaluate the full-batch gradient and the projected zero-leakage control $D_{\mathrm{safe}} = (I - \Pi_L)D(I - \Pi_R)$ from Appendix C.4. All 276 gradients and controls are kept, at both signs. A relative response is reported only when the anchor norm exceeds 10^{-12}; otherwise use absolute units.

Every mini-batch anchor norm exceeds that threshold. The observed changes range from 0.178% to 2.192%; the largest absolute response change for a projected control is below 6.4×10^{-17}. Direct factor steps agree with (35) to absolute error below 8.3×10^{-16} at $\eta = 0.01$. Table 7 compares the observed changes with (39) and (37): the lower expression is close, whereas the uniform whole-budget upper bound is conservative. Those ratios use numerical geometry, not exact response certificates. At factor stationarity, the weighted batch-gradient average has zero response at every coordinate, so the individual batch differences reflect gradient-noise geometry rather than a nonzero mean step.

E.7 Longer-trajectory observations

The supplementary material also includes longer runs of the original fixed-coordinate, anchor, plus and minus rules, as well as fixed-coordinate SGD tests. The 200-step refactorizing runs depend on which representative the solver returns in a nearly flat set. Their effects change sign across heads and when the coordinates are re-solved; one re-solved near-repeated case exceeds the prescribed balance tolerance and is flagged as such. The fixed-coordinate SGD tests likewise do not yield a consistent ranking from one-step noise changes; they cycle the 22 batches in a fixed order with batch-mean gradients and one step size, which gives the three-example batch more weight, so their conclusions are specific to that protocol. These results establish no stable long-term advantage. They are reported in full rather than selected by outcome, but are not evidence about repeated application of the block extremum, whose reported comparison is checkpoint-wise.

E.8 TWO-SIDED CURVATURE: THE SAME HEADS UNDER SOFTMAX CROSS-ENTROPY

The squared-loss curvature above is one-sided, and Appendix C.2 shows that its top mode then has a rank-one active latent Gram; the exact free subspace of Section 6.2 is a consequence of that loss family. To test whether the mechanism depends on it, the six heads are retrained with fixed-feature softmax cross-entropy on the same PCA features, one-hot targets and ridge 10^{-2}, from the same seeded initializations, with 200 plain factor-gradient steps at $\eta = 0.1$ (the largest of $0.1, 0.05, 0.02, 0.01$ with a monotone loss). The supplied curvature is the exact cross-entropy pullback $H = J^*\mathcal{C}J$ with $\mathcal{C} = \frac{1}{s}\sum_i (x_i x_i^\top) \otimes (\operatorname{diag} p_i - p_i p_i^\top) + \rho I$, with $\rho = 10^{-2}$ and row-major vectorization of M. This curvature is PSD and two-sided. The heads, loss, training rule, checkpoints, tolerances and outcome rules were fixed before the run. After the run, the analysis added the residual correction of Appendix A.5, excluded the near-repeated checkpoints from the single-mode statistics, and labeled the 1% thresholds as idealized extrapolations, without changing any threshold, seed, checkpoint or outcome rule. Every reported quantity can be recomputed from the stored coordinates without a solver. At the same seven checkpoints per head the largest-eigenvalue coordinate is solved on $\mathcal{Q}_2$ with the solver of Appendix E, the factors are reanchored, and the top mode is analyzed as before.

Table 8: Two-sided curvature: the six heads retrained with fixed-feature softmax cross-entropy and the exact cross-entropy pullback as the supplied curvature (Appendix E.8; seven checkpoints per head). Rep. counts checkpoints whose top eigenvalue is near-repeated (relative gap below 10^{-5}); there the single-mode quantities depend on the eigenvector chosen in the top eigenspace, so the remaining columns except the trace columns are statistics over the simple-top checkpoints of the head only. $p_{\min}/p_{\max}$ is the eigenvalue ratio of the active latent Gram, which has full rank at every simple-top checkpoint. Spread is the relative response change along the softest direction of the Gram within the coordinate budget, with relative objective deviation at most 7×10^{-6} or 10^{-3} from the solved reference (median / max). $\lambda_{\max}$ bound/obs is the median ratio of the largest-eigenvalue bound of Theorem 1, with the optimality-residual correction of Appendix A.5 for the solved reference, to the observed change at the 7×10^{-6} tolerance; the bound contains the observed change at all 34 simple-top checkpoints. The trace columns (all seven checkpoints) give the median response difference between the largest-eigenvalue coordinate and the closed-form trace optimum relative to the latter's response and the median trace bound/observed ratio; all 42 trace bounds contain the observed difference. Floating-point diagnostics at the stored references.

Seed	Rank	rep.	$p_{\min}/p_{\max}$	spread $7{\cdot}10^{-6}$ (%)	spread 10^{-3} (%)	$\lambda_{\max}$ bound/obs	trace obs/ref	trace bound/obs
17	2	1	$10^{-3.4}$–$10^{-0.1}$	4.69 / 6.7	26.1 / 34.7	9.9	10.8%	6.5
17	4	0	$10^{-6.2}$–$10^{-3.0}$	11.73 / 26.3	12.7 / 26.3	26.8	32.6%	7.7
29	2	0	$10^{-4.7}$–$10^{-0.5}$	17.89 / 26.9	41.7 / 46.5	6.6	11.8%	6.2
29	4	5	$10^{-5.8}$–$10^{-5.1}$	24.37 / 34.3	25.9 / 34.4	17.8	38.9%	7.7
43	2	0	$10^{-5.6}$–$10^{-0.1}$	24.15 / 33.3	28.8 / 35.8	12.7	8.4%	7.6
43	4	2	$10^{-6.2}$–$10^{-3.3}$	45.48 / 53.9	48.4 / 59.8	14.0	15.7%	12.4

Of the 42 checkpoints, 34 have a simple top eigenvalue (relative gap at least 10^{-5}; in fact at least 2.6×10^{-3}) and eight are near-repeated (relative gap below 10^{-5}). At a near-repeated top the single-mode quantities of Propositions 4–5 depend on which member of the top eigenspace the eigensolver returns. Exact optimality requires a certifying density supported within the top eigenspace, together with budget-normal multipliers (Theorem 3); full support is required only for the repeated-mode local-set formula. Thus a nonzero balance residual of the returned member does not rule out optimality (at these eight checkpoints the residual correction of Appendix A.5 is 1.2–2200 times the objective gap), and even a balanced member with partial top-space support can leave a one-sided local optimizer set (the example of Appendix C). These eight checkpoints are evaluated for the returned member, flagged as member-dependent, and excluded from the statistics below and from the recomputation checks; certifying them would need the repeated-mode treatment of Appendix C.

At the 34 simple-top checkpoints the active latent Gram has full rank, so under exact simple-top geometry the largest-eigenvalue coordinate would be unique (Proposition 4, or Appendix A with $\mathsf{P}_\Phi = \lambda P \succ 0$ for the one reference that touches the budget boundary); the Gram is weakly active instead: $p_{\min}/p_{\max}$ ranges from 5.8×10^{-7} to 0.89 (median 1.9×10^{-4} at rank two, 2.0×10^{-6} at rank four). Along the softest direction of the Gram, coordinates whose objective deviates from the solved reference by at most 7×10^{-6} change the response by 0.06%–54% (median 6.6% at rank

two, 18% at rank four). The spread is largest early in training (medians 23%–26% at $t \le 20$) and is 0.06%–4.4% at $t = 200$, where only four heads have a simple top and their Grams are better conditioned; at deviation 10^{-3} it is 0.8%–60%. The spread moves to the largest admissible $|t|$ on each side of the reference and reports the larger of the two response changes, relative to the reference response (not to the loss or accuracy). At the largest spread the reference response has norm 0.124 and the change has norm 0.067.

The stored references are feasible floating-point solutions whose top-mode balance residual $\|X^\top X - YY^\top\|_F$, for the unit mode $q = (X, Y)$ of Proposition 4, is below the protocol tolerance 10^{-4} at every simple-top checkpoint (largest 3.8×10^{-5}). All but one lie inside the budget with margin at least 1.7×10^{-3}; the reference with the largest spread (seed 43, rank four, $t = 5$) touches the budget boundary (margin 7×10^{-16}) with its top mode balanced to 2.6×10^{-8}. These are not exact optima, so the certificate is evaluated in the form of Appendix A.5: for the support $\Xi = qq^\top$ of the computed top eigenvector (whose own floating-point error is not bounded here) and the zero normal, which is admissible at every feasible reference, the residual is $\eta_{\mathrm{opt}} = \|P_V - P_U\|_F = \lambda\|X^\top X - YY^\top\|_F$, and with $T = I + tvv^\top$ the objective gap is replaced by $F_\Phi(T) - F_\Phi(I) + \frac{1}{2}\eta_{\mathrm{opt}}(|t| + |t/(1+t)|)$. The correction is below 10% of the gap at 27 of the 33 checkpoints with a positive gap and reaches 1.9 times the gap at one; at the remaining checkpoint the solved reference is suboptimal by 2.8×10^{-7} along the softest direction, and the correction alone yields a positive corrected gap. With it, the bound of Theorem 1 contains the observed change at all 34 simple-top checkpoints (median bound/observed 13, range 6.1–117). The relative deviation at which an exact-reference certificate would guarantee a 1% response tolerance is 10^{-11}–3×10^{-6} over the 34 checkpoints (1.6×10^{-6}–3×10^{-6} at $t = 200$ for three of the four simple-top heads). This is an idealized extrapolation: the uniform residual term $\eta_{\mathrm{opt}} d_{\mathcal{Q}}$ of Appendix A.5, with the explicit endpoint bound on $d_{\mathcal{Q}}$, alone exceeds the corresponding objective gap at 30 of the 34 stored references, so a 1% guarantee from this certificate would need references balanced far below the 10^{-4} tolerance, a demanding target for this solver. The trace certificate at its closed-form interior optimum, which needs no eigenvector and applies at all 42 checkpoints, behaves as in Table 1: the response of the largest-eigenvalue coordinate differs from that of the trace optimum by 4%–57% at objective excess 0.4%–20%, and the bound contains the observed difference in all 42 cases, with median ratio 7.7 (4.8–18.5).

These references carry a double-precision lower bound, not a rational certificate: the solver's relative gap is below 10^{-4} at 38 checkpoints but exceeds the 7×10^{-6} deviation band at five (largest 6.1×10^{-4}), so the band is measured from the solved reference rather than from the true optimum; at the checkpoint with the largest spread (54%) the gap is 1.8×10^{-8}. The returned reference is itself solver-dependent: re-solving on the same hardware with multi-threaded linear algebra returns different references (entries differing by more than 10^{-6}) at 14 of the 42 checkpoints, all at rank four; the trajectories, top eigenvalues and trace optima agree, the corrected bound contains the observed change at every re-solved simple-top checkpoint, and individual spreads at 7×10^{-6} change by up to a factor of 2.2. Recomputation from the stored coordinates is therefore the deterministic check of the reported numbers, and re-solved references are checked only for the solver-independent quantities and the containment property. The exact candidate-value certificates of Appendix F are not repeated here. The comparison adds no training claim.

E.9 Dependence on the budget radius

All coordinate problems use $R = 2$. Because f_H and the full-support objectives are convex on the whole SPD cone, a stationary reference stays optimal on every budget $\mathcal{Q}_{R'}$ that contains it, with the same active Gram, free subspace, coverage matrix and leakages. After reanchoring, R' then enters only through $\delta_{\mathrm{box}}(R') = \min\{1 - 1/(R'^2 s_{\min}),\ R'^2/s_{\max} - 1\}$, with $s_{\min}, s_{\max}$ the extreme eigenvalues of the reference, and through $C_{\mathcal{Q}}$, which is proportional to R' because $u_0 \propto R'^2$ and $\ell_0 \propto R'^{-2}$. At the frozen checkpoints:

- Squared loss, largest eigenvalue (Sections 6.2 and 6.3): the radius a is set by the gap term at 41 checkpoints and by the cap 0.1 at one, never by the budget. Every reference stays interior and every radius unchanged for $R' \ge 1.80$, so the references, free subspaces, leakages, uniform endpoints, response changes, block extrema and two-rate controls are identical for every $R' \ge 1.80$.

- Full-support certificates (Section 6.1): the references stay interior and the comparators feasible for $R' \geq 1.93$. There only $C_{\mathcal{Q}}$ changes, by the factor $R'/2$, so every bound scales linearly with R'; at $R' = 1.93$ the smallest bound/observed ratio would be 4.6.
- Cross-entropy (Section 6.5): at 28 of the 34 simple-top checkpoints the spread at 7×10^{-6} is set by the tolerance and is unchanged for every $R' \geq 2$; at the other six the budget stops the search first, so those spreads depend on R'.

These are floating-point statements about the frozen references and candidates; re-solving on another budget may return different references, and the rational witnesses certify $R = 2$ only.

E.10 Certified stopping

At a feasible iterate, reanchored to I, take the gradient of the criterion as the certifying subgradient and the zero normal. Every minimizer of the criterion on the original budget then has response within $C_{\mathcal{Q}}\sqrt{\kappa_\Phi(D)\eta_{\mathrm{opt}}d_{\mathcal{Q}}}$ of the iterate's response, and within the linear bound of Appendix A.5 when $\mathsf{P}_\Phi \succ 0$. The smaller bound divided by the iterate's response norm is a certified relative tolerance τ_{cert}. At a target τ the rule stops when $\tau_{\mathrm{cert}} \leq \tau$, continues while $\tau < \tau_{\mathrm{cert}} < \infty$, and reports no guarantee when $\kappa_\Phi(D) = \infty$.

For the Schatten-2 criterion at the 42 squared-loss checkpoints, the solver of Appendix E.2 is started from the current coordinates $S = I$ and stopped after each point of a fixed schedule of 1 to 200 iterations; infeasible iterates certify nothing. The rule certifies $\tau = 10\%$ at all 42 checkpoints after a median of 15 scheduled iterations and $\tau = 1\%$ after a median of 19.5, never more than 39 (about 20 objective evaluations, roughly 10 ms per checkpoint). At the stopping iterate the actual relative distance to the response of a converged minimizer is at most 6.6×10^{-4} for 10% and 3.1×10^{-5} for 1%: the rule is safe and conservative by a median factor of about 500. The uniform form alone certifies 5% at all 42 checkpoints but 1% at only 19; at the final iterates the linear form certifies a median of 5.3×10^{-5} and at most 5.8×10^{-4}.

For the largest eigenvalue under squared loss the leakage of Section 6.2 is nonzero at every checkpoint, so $\kappa_\Phi(D) = \infty$ and the rule reports no guarantee at any solver accuracy; the 4.9% spread among certified near-optimal coordinates shows that stopping there is indeed unsafe for the update. Under cross-entropy, at the 34 stored simple-top references of Appendix E.8, the rule certifies 1% at 16, 10% at 19, and nothing below 50% at the other 15, where the weakly active Gram makes $\kappa_\Phi(D)$ large; there it asks for a more accurate reference. At $t = 200$ it certifies 1% at all four simple-top heads. The certificate is exact given an iterate; the reported quantities are floating-point evaluations of it.

F Candidate-value certificates and verification

Floating-point optimization proposes coordinates and density matrices; exact rational arithmetic verifies their budget feasibility and spectral values. We distinguish a bound on a chosen coordinate from the global lower bound needed to measure its relative spectral excess $\mathrm{Exc}_{H,R}(S) = f_H(S)/\mathfrak{F}_R(H) - 1$ (Section 2). For $0 < \underline{\mathfrak{F}} \leq \mathfrak{F}_R(H) \leq \overline{\mathfrak{F}}$ and $0 \leq c_L \leq f_H(S_c) \leq c_U$,

$$\mathrm{Exc}_{H,R}(S_c) \in [c_L/\overline{\mathfrak{F}} - 1,\ c_U/\underline{\mathfrak{F}} - 1]. \tag{44}$$

An upper endpoint within the desired tolerance proves it; a lower endpoint above it rules it out. Otherwise the bounds are inconclusive.

F.1 A density lower bound over the whole budget

Write a PSD finite operator as $\widehat{H} = \mathsf{B}\mathsf{B}^\top$ (the root factor B is unrelated to the certificate B of Section 3) and put $L(S) = \mathsf{B}^\top \mathcal{W}(S)\mathsf{B}$. This is isospectral to its weighted factor operator. A density $\Omega \succeq 0$, $\operatorname{tr}\Omega = 1$, gives

$$\phi(S) = \operatorname{tr}(\Omega L(S)) = \operatorname{tr}(AS^{-1}) + \operatorname{tr}(D_0 S) \leq f_{\widehat{H}}(S), \qquad A, D_0 \succeq 0.$$

At a feasible reference S_c, let $g = -S_c^{-1}AS_c^{-1} + D_0$. Convexity gives $\phi(T) \geq \phi(S_c) + \mathrm{tr}[g(T - S_c)]$ for every feasible T. Minimizing this affine minorant on the spectral box gives

$$L_{\mathrm{aff}} = \phi(S_c) - \mathrm{tr}(gS_c) + R^{-2} \sum_{\lambda_j(g) \geq 0} \lambda_j(g) + R^2 \sum_{\lambda_j(g) < 0} \lambda_j(g) \leq \mathfrak{F}_R(\widehat{H}). \tag{45}$$

No exact top-eigenspace membership is required for the density. To avoid certifying the eigenvalues of g, the rational verifier uses

$$L_{\mathrm{rat}} = \phi(S_c) - k(R^2 - R^{-2}) \max_i \sum_j |g_{ij}| \leq \mathfrak{F}_R(\widehat{H}). \tag{46}$$

Indeed, $\|T - S_c\|_{\mathrm{op}} \leq R^2 - R^{-2}$ and $|\mathrm{tr}[g(T - S_c)]| \leq \|g\|_* \|T - S_c\|_{\mathrm{op}}$, while $\|g\|_* \leq k\|g\|_{\mathrm{op}} \leq k \max_i \sum_j |g_{ij}|$.

F.2 Feasibility and candidate spectral bounds

Binary64 inputs are interpreted as exact dyadic rationals. Candidate coordinates and witness matrices are rounded to denominator 2^{40}. The verifier uses no floating-point eigensolver. Exact LDL pivots check $R^{-2}I \prec S \prec R^2 I$ for the reported candidates. A rational root B makes $\widehat{H} = \mathsf{B}\mathsf{B}^\top$ exactly PSD, and a nonzero rational C makes $\Omega = CC^\top / \|C\|_F^2$ exactly PSD and trace one. The partial Grams, the minorant and (46) are evaluated in rational arithmetic.

For a proposed rational upper endpoint u, the verifier checks strict positive diagonal dominance of

$$Q^\top Q, \qquad uQ^\top Q - Q^\top L(S) Q,$$

where Q is a rational approximate eigenbasis. The first matrix proves that Q is invertible; the second implies $Q^\top(uI - L(S))Q \succ 0$, hence $\lambda_{\max}(L(S)) < u$. Orthogonality of Q is not assumed. A normalized rational Rayleigh quotient supplies the candidate lower endpoint. All comparisons become integer comparisons after clearing positive denominators.

F.3 Transferring operator approximation errors

For a symmetric raw matrix H_{raw}, define $\psi_{H_{\mathrm{raw}}}(S) = \|\mathcal{W}(S)^{1/2} H_{\mathrm{raw}} \mathcal{W}(S)^{1/2}\|_{\mathrm{op}}$. If $\|H_{\mathrm{raw}} - \widehat{H}\|_{\mathrm{op}} \leq \epsilon_H$, then

$$\sup_{S \in \mathcal{Q}_R} |\psi_{H_{\mathrm{raw}}}(S) - f_{\widehat{H}}(S)| \leq R^2 \epsilon_H. \tag{47}$$

This follows from the triangle inequality and $\|\mathcal{W}(S)\|_{\mathrm{op}} \leq R^2$. The exact row-sum bound $\epsilon_H = \max_i \sum_j |(H_{\mathrm{raw}} - \mathsf{B}\mathsf{B}^\top)_{ij}|$ includes root-rounding error. Both candidate and optimum endpoints are expanded by $R^2\epsilon_H$; negative roundoff-scale eigenvalues of the raw matrix are not discarded. For a PSD target the operator norm is the largest eigenvalue. Convexity is applied only to $\widehat{H}$, never to an indefinite raw operator.

With a positive denominator and nonnegative candidate lower bound, (44) applies after transfer. More generally, take the minimum and maximum of the four endpoint ratios, or declare the result inconclusive if the denominator is not bounded away from zero. Negative intermediate spectral-excess lower bounds may be retained.

F.4 From stored features to the exact pullback

Let X_d, U_d, V_d be the exact dyadic values of the stored arrays and Y_d the exact one-hot labels. The specified stored-feature loss is

$$\ell_d(M) = \frac{\|X_d M - Y_d\|_F^2}{2s} + \frac{\|M\|_F^2}{200}, \qquad P_d = X_d^\top X_d / s + I/100.$$

The verifier evaluates P_d exactly and constructs

$$H_{\mathrm{pb},d}(A, B) = \big(P_d(AV_d + U_d B)V_d^\top,\ U_d^\top P_d(AV_d + U_d B)\big).$$

This is PSD because $P_d \succ 0$. With the corresponding stored raw matrix, compute $\zeta = \max_i \sum_j |(H_{\mathrm{pb},d} - H_{\mathrm{pb,raw}})_{ij}|$ in rational arithmetic and expand all endpoints once more by $R^2\zeta$. Thus the final bounds concern the pullback of the loss actually defined by the stored features and factors, not just a matrix produced by the numerical program. No stationarity assumption is used. The checkpoint bridges are below 9×10^{-16}. Separate reconstruction checks verify the splits, labels and numerical PCA preprocessing; they are not claims of exact symbolic PCA.

The references, uniform signs and block candidates each have their own candidate-value witnesses. The block comparison recomputes the global lower and candidate upper endpoints before applying the same transfer. Its 42 witnesses are verified in exact rational arithmetic and certify relative spectral excess below 7×10^{-6}.

F.5 Scope of the verification

The proofs establish the stated universal results. Rational witnesses certify specific rounded candidate values and budget feasibility; numerical checks test implementations and compare pairing or losses. Neither finitely many numerical tests nor a value certificate proves the exact local optimizer-set hypotheses at a floating-point reference.

The supplementary tests cover noncommuting budgets and blocks, boundary optima, rank-deficient factors, singular spectral support, approximate optimality, least-squares and softmax compatibility, repeated-mode certificate choice, and budget reanchoring. Recomputation from the stored results checks all 42 block comparisons and the independent scalar-rate projections, and deliberately corrupted witnesses are rejected. The dataset supplies correlated checkpoints and within-model controls, not independent experiments for each evaluated gradient.

G Closest mathematical comparisons

Table 9: Which claims are prior and which are made here. The first six rows are established by the cited work and are not contributions of this paper.

Claim	Status
Equivalent factor pairs give different Euclidean updates	prior (Dinh et al., 2017; Arora et al., 2018; Yen et al., 2025)
Balancing or invariant updates improve conditioning	prior (Castin et al., 2026; Bogachev et al., 2026)
A prescribed per-step refactorization is consistent	prior (Zhang et al., 2025)
Spectral subdifferentials, optimality certificates, error bounds	prior (Overton, 1992; Shapiro & Fan, 1995; Lewis & Overton, 1996; Lewis, 1999; Overton & Womersley, 1993; Rebjock & Boumal, 2025; Li, 2026)
Closed-form factor gradients matching a target product update	prior (Wang et al., 2025)
Dimension counts of optimal faces; conserved quantities of factor flows	prior (Pataki, 1998; Kunin et al., 2021; Marcotte et al., 2023)
Spectral-objective error $\to$ gradient-dependent product-response bound on the original budget; full-support identification (Thm. 1, Cor. 2)	this paper
Local geometry and dimension of the set of *all* $\lambda_{\max}$-optimal coordinates (Prop. 4)	this paper
Zero leakage as necessary and sufficient for a fixed response (Prop. 5)	this paper
Closed-form extremal response pairing on a certified block with exact endpoint gain (Thm. 6)	this paper

Diagonal support versus latent subspaces. For diagonal similarity scalings $\|e^D A e^{-D}\|_{\mathrm{op}}$, Balakrishnan & Boyd (1995, Theorems 3.1 and 3.3) relate optimality to componentwise matching of an isolated top singular-vector pair and, under irreducibility, uniqueness to its nonzero components. Support-based nonuniqueness is therefore not new. Here the reciprocal factor action matches matrix

Grams $X^\top X = YY^\top = P$. Equality in the certifying subgradient forces $(T - I)P^{1/2} = 0$; full active support makes this condition locally sufficient. For a simple top mode it also identifies the coordinate-Hessian kernel. These directions need not be diagonal: $P = \frac{1}{4}\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ has no zero diagonal entry but has a one-dimensional kernel, while no nonzero diagonal Δ satisfies $\Delta P = 0$. Ghadimi et al. (2025, Proposition 9) give a related sufficient condition for continuous optima of condition-number-optimal diagonal scalings. That objective and its necessity question are different from the largest-eigenvalue result here.

Optimality information versus response information. Spectral subdifferentials and normal-cone conditions are standard (Lewis, 1999; Overton & Womersley, 1993). The structured-metric certificates of Li (2026, Theorems 3.6 and 5.9) include log-spectral objectives and Kronecker partial-trace conditions, and its residual certificates bound point error, objective gap and thresholds (Corollary 5.13, Theorem 5.18); they control the metric, not a response to a fixed gradient. The response theorem uses a PSD certifying subgradient and the original and inverse budget normal cones to form P_Φ, then compares it with Q_D to bound a specified product response. Its contribution is this error transfer, not a new KKT rule or the square-root error order (Rebjock & Boumal, 2025). The PSD domination argument does not cover indefinite subgradients such as those used for the condition number.

Prescribed rules versus sets of near-optimal coordinates. RefLoRA characterizes the minimizers of its refactorization objective, which need not be unique, and establishes consistency of its prescribed rule (Zhang et al., 2025, Theorems 3 and 6); balanced and invariant methods also choose representation-consistent updates (Castin et al., 2026; Yen et al., 2025). A rule can select a consistent representative even when another spectral objective has near-minimizers with different responses. Our theorem addresses that set-valued question for a fixed gradient and budget. LoRA-Pro chooses the factor gradients in closed form so that the induced product update best approximates the full fine-tuning gradient (Wang et al., 2025); it optimizes one update rule, whereas the question here is how much the set of spectrally near-optimal coordinates lets a fixed gradient's update vary. Symmetry teleportation maximizes the gradient norm along a loss-invariant orbit (Zhao et al., 2022). The block theorem solves that pairing objective on an entire certified spectral block, using the inverse chord and a gap-based radius, and states when it strictly exceeds uniform endpoint choices. It is not an optimizer over the whole orbit.

Dimension counts, conserved quantities and quotient geometry. The count $(k-r)(k-r+1)/2$ of Proposition 4 is the dimension of the symmetric matrices on $\ker P$. The same arithmetic governs faces of the semidefinite cone and the multiplicity of optimal eigenvalues (Pataki, 1998), and the number of independent conserved quantities, such as $U^\top U - VV^\top$, of gradient flows on factorized models (Kunin et al., 2021; Marcotte et al., 2023). Those results count invariants of a semidefinite program or of a dynamical system; ours is the dimension of the local optimizer set of a reciprocal SPD coordinate objective, and its proof goes through the active latent Gram rather than a constraint count. Quotient and embedded geometries of fixed-rank matrices (Vandereycken, 2013; Absil et al., 2013; Mishra et al., 2014) remove the gauge altogether; the question here arises only when a gauge is kept and chosen by a criterion.