SPECTRAL REUSE UNDER SQUARING: A SHARP LOCALITY THRESHOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Cached curvature responses support reliable update-length decisions only with an inverse-square guarantee; a quadratic-form guarantee alone is insufficient. We identify a sharp transition governed by spectral locality: for a Loewner approximation factor r and log-spectral coupling width L, the optimal inverse-amplification norm over all dimensions and condition numbers is finite exactly when $(r-1)e^{L/2}/(r+1) < 1$. We derive this optimal norm and construct an asymptotically sharp chain whose amplification grows linearly with dimension at the threshold and exponentially above it. Numerical scans of finite chains show all three regimes. For unrestricted coupling, an exact square-distortion law gives the complementary condition $\epsilon\sqrt{\kappa} \to 0$, where $r = 1 + \epsilon$ and κ bounds the reference condition number. Together, these bounds give an update-length interval that accounts for asymmetric inputs, damping, and nonzero spectral tails. For example selection on synthetic logistic data with local spectral coupling, this interval reduces Hessian–vector products by a factor of 3.1 (595.8 to 189.3 per batch) relative to global exact screening. It also reduces selection time by 24% (1.68 to 1.28 ms). On 1024-dimensional digits features, global exact screening reduces total time from 2.25 to 1.98 s over 32 training steps with cache refreshes. Direct Cholesky is faster in the evaluated cold-start and banded settings.

1 INTRODUCTION

Curvature reuse can accelerate a new linear solve or avoid it by using information from a cached solution. The latter role matters when a learning procedure repeatedly selects examples that produce a large local parameter response. Given a positive-definite curvature matrix H and example gradient g_i, this response is $H^{-1}g_i$. If the procedure selects examples whose $\|H^{-1}g_i\|_2$ exceeds a threshold, it must evaluate many responses as training changes H. Cached responses can avoid these solves, but an incorrect length comparison changes the selected set and the subsequent updates. We therefore ask when cached curvature information can reliably support these decisions, beyond its use as a solver preconditioner.

The practical issue is how to decide that a new solve is unnecessary. A fixed refresh schedule treats all curvature changes alike, whereas the effect on a candidate also depends on its distance from the selection threshold. A valid interval for the current response length can certify many decisions from the cache and reserve new solves for the remaining candidates. Such an interval must account for drift in every direction: a small error in curvature energy need not imply a small error in the Euclidean length used by the selector.

A standard spectral guarantee does not directly answer this question. For cached and current curvature matrices A, B, the Loewner sandwich $r^{-1}A \preceq B \preceq rA$ preserves every quadratic form within factor r. Update lengths instead involve $\|B^{-1}g\|_2^2 = g^\top B^{-2}g$. If A and B commute, squaring simply changes r to r^2; otherwise, distortion can grow with the condition number, as studies of squared-system preconditioning show (Braess & Peisker, 1986; Wathen, 2022). Indeed, $C = A^{-1/2}BA^{-1/2}$ and $M = BA^{-1}$ have the same eigenvalues, but the squared guarantee depends on the singular values of M. Thus a well-conditioned preconditioned system can coexist with unreliable Euclidean update lengths. The missing information is how the relative matrix couples directions with different curvature scales.

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.

We ask: *which restrictions on this mixing permit a uniform reuse guarantee as dimension and global conditioning grow?* We measure mixing in the eigenbasis of A: log-spectral width L allows coupling only when the larger-to-smaller eigenvalue ratio is at most e^L. This condition restricts where couplings occur in the spectrum, rather than only their sparsity or rank. Even local couplings can propagate across the whole spectrum after inversion. A uniform bound depends on whether inverse entries decay fast enough to offset amplification across scales.

Classical inverse-decay and weighted-resolvent arguments explain this competition (Demko et al., 1984; Combes & Thomas, 1973; Janas et al., 2018). Chebyshev expansion already gives a finite bound when $(r-1)e^{L/2}/(r+1) < 1$. Our central result determines the optimal inverse operator norm over all finite spectral placements and multiplicities. It also proves that no uniform bound exists beyond this regime and gives matching growth rates at and above the boundary. A tighter bound can resolve a threshold decision that a looser bound leaves unresolved, which makes the optimal constant useful for reuse. We make three contributions:

- **An optimal local bound and its failure boundary.** We determine the dimension-independent inverse norm and construct a chain approaching it below the threshold, with linear and exponential growth at and above the threshold (section 4, theorem 4.1).
- **An exact unrestricted calibration.** Without locality, we give the worst-case square distortion and an attaining two-dimensional pair. This law identifies the exact constants and the vanishing-error condition $\epsilon\sqrt{\kappa} \to 0$ for $r = 1 + \epsilon$ (section 5, theorem 5.1 and proposition 5.2).
- **A decision-preserving reuse interval.** We combine both bounds with asymmetric drift, damping, and weighted spectral tails, then use residual refinement for unresolved candidates (corollary 5.3 and section 6).
- **Empirical validation and cost accounting.** We test the mathematical transition and the resulting reuse decisions in logistic training with frozen features. With known local support, the method reduces HVPs by a factor of 3.1 and selection time by 24% relative to global exact screening. Against the closer Chebyshev baseline, the HVP reductions are 2.1% and 5.5% for adjacent and wider coupling designs. Dense-feature experiments identify when tail construction and cache setup offset these savings, yielding a cost criterion that separates a stronger guarantee from an actual runtime benefit (proposition 6.1).

2 Related Work

Spectral approximation and squaring. Loewner approximation is central to graph sparsification and preconditioned solvers (Spielman & Teng, 2011; Spielman & Srivastava, 2011). Braess & Peisker (1986) studied squared-system preconditioning, and Wathen (2022) reviews its relation to normal equations. Gratton et al. (2018) bound weighted-Gram spectra through relative residuals and conditioning. Our exact law characterizes the full two-sided Loewner interval and provides an attaining pair. Its proof uses Kantorovich–Wielandt projection geometry and the canonical form of quadratic operators (Marshall & Olkin, 1990; Tso & Wu, 1999; Gau et al., 2008). Appendix J gives the explicit comparison with these results.

Locality and inverse decay. Demko et al. (1984) establish inverse-entry decay for banded matrices, including the sharp Toeplitz decay exponent. Benzi & Razouk (2007); Benzi & Boito (2014) extend decay estimates to functions of sparse matrices and matrices over C^*-algebras. Weighted conjugation and resolvent estimates are central to the Combes–Thomas method (Combes & Thomas, 1973) and its block-Jacobi extensions (Janas et al., 2018). Our relative inverse $A^{1/2}C^{-1}A^{-1/2}$ is a weighted resolvent with weight $\frac{1}{2}\log A$. Chebyshev approximation and classical band-limited conjugation (Aleksandrov & Peller, 2010) give a uniform inverse bound from $[r^{-1}, r]$ and log-spectral width L. We optimize the weighted inverse operator norm uniformly over dimensions, spectral placements, and multiplicities. The attaining chain proves necessity and gives the growth rates at and above the threshold, linking entry decay to the optimal operator-norm bound. Proposition J.1 quantifies the gap between the Chebyshev bound and the optimum. The distinction is between decay of individual inverse entries and control of the largest response to an arbitrary input vector. Entrywise bounds require an additional summation step, whose constants can depend on the arrangement and multiplicity of spectral values. Our upper bound treats these configurations together, while a sequence

of finite matrices approaches the same constant. This combination establishes a uniform optimum, rather than a bound for one prescribed chain or spectral grid.

Curvature reuse. K-FAC, Shampoo, and SOAP reuse curvature factors or their decompositions across updates (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2025). Shampoo evaluates delayed root updates every 20–100 steps; SOAP studies decomposition frequencies from 1 to 100. Given Loewner drift bounds for a symmetric positive-definite (SPD) update operator, our exact law converts an update-length tolerance into a refresh condition (corollary A.6). Inverse-curvature products also appear in data attribution (Kwon et al., 2024). Scalar weights control logistic Hessian variation (Bach, 2010). We combine this comparison with the square law to determine whether update lengths exceed a radius across checkpoints.

3 Spectral Reuse and the Relative Operator

We model one reuse decision by a cached curvature matrix A, a current curvature matrix B, and a current gradient $g \in \mathbb{R}^n$. Both matrices belong to $\mathbb{S}^n_{++}$, the real symmetric positive-definite matrices. The cached response is $u_0 = A^{-1}g$, and the response that a new solve would return is $u = B^{-1}g$. We seek an interval for $\|u\|_2$ using $\|u_0\|_2$ and matrix-level drift information, without computing u for each candidate. If the interval lies on one side of a selection threshold, reuse gives the same decision as a new solve. The gradient is the same in both responses; section 6 explains how frozen features make the cached response available even when gradient scalars change during training.

A uniform interval requires control of $g^\top B^{-2}g / g^\top A^{-2}g$ for every nonzero g. The following spectral factor expresses both the available curvature comparison and the required comparison of inverse squares in one notation. For $A, B \in \mathbb{S}^n_{++}$, define

$$\phi(A, B) = \max\Big\{\Big\|B^{-1/2}AB^{-1/2}\Big\|_2, \Big\|A^{-1/2}BA^{-1/2}\Big\|_2\Big\}. \tag{1}$$

This factor is symmetric and satisfies

$$\phi(A, B) \le r \quad \Longleftrightarrow \quad r^{-1}A \preceq B \preceq rA. \tag{2}$$

To quantify what the input factor alone guarantees, define the worst-case square distortion

$$\mathcal{W}_n(r, \kappa) = \sup_{\substack{A,B\in\mathbb{S}^n_{++} \\ \phi(A,B)\le r,\ \kappa_2(A)\le\kappa}} \phi(A^2, B^2). \tag{3}$$

Inversion preserves ϕ, so the same $\mathcal{W}_n$ controls A^{-2} and B^{-2}. We study the joint limit of improving input accuracy $r = 1 + \epsilon$ and increasing conditioning κ.

The relative operator separates this length question from the usual eigenvalue comparison. Define

$$C = A^{-1/2}BA^{-1/2}, \qquad M = BA^{-1} = A^{1/2}CA^{-1/2}. \tag{4}$$

The symmetric matrix C and relative operator M have the same positive eigenvalues. Their relation gives the central identity.

Proposition 3.1 (Relative-operator identities)**.** *For every $A, B \in \mathbb{S}^n_{++}$,*

$$\phi(A, B) = \max\{\rho(M), \rho(M^{-1})\}, \tag{5}$$

$$\phi(A^2, B^2) = \max\{\|M\|_2^2, \|M^{-1}\|_2^2\}. \tag{6}$$

Proof. The first identity follows by similarity. For the second, use $A^{-1}B^2A^{-1} = M^\top M$ and $B^{-1}A^2B^{-1} = (M^{-1})^\top M^{-1}$. □

For commuting pairs, M is symmetric positive definite and $\phi(A^2, B^2) = \phi(A, B)^2$. In the general case, the similarity in equation (4) creates a gap between eigenvalues and singular values that governs distortion.

4 A PHASE TRANSITION UNDER SPECTRAL LOCALITY

The relative-operator identity makes cross-scale mixing the next question. In an eigenbasis of A, a perturbation entry between λ_i and λ_j receives the first-order multiplier $\sqrt{\lambda_i/\lambda_j} + \sqrt{\lambda_j/\lambda_i}$ (theorem D.1). This multiplier motivates a locality measure based on distances between logarithmic eigenvalues.

We say that $C = A^{-1/2}BA^{-1/2}$ has log-spectral width $L \geq 0$ if

$$\widehat{C}_{ij} = 0 \quad \text{whenever} \quad |\log \lambda_i - \log \lambda_j| > L. \tag{7}$$

Repeated local couplings propagate across the spectrum after inversion. The competition between this propagation and inverse decay gives an exact threshold.

We measure locality in the reference eigenbasis; coordinate sparsity alone does not imply it. The threshold concerns a uniform bound over growing dimensions, not invertibility of any fixed SPD pair.

Theorem 4.1 (Optimal local inverse bound)**.** *Let $\mathcal{I}(r, L)$ be the supremum of $\|B^{-1}A\|_2$ over all finite dimensions $n \geq 2$ and all SPD pairs satisfying $\phi(A, B) \leq r$ and equation* (7)*. Set*

$$c = \frac{r + r^{-1}}{2}, \qquad h = \frac{r - r^{-1}}{2}, \qquad \alpha = c - h\cosh(L/2).$$

Then

$$\mathcal{I}(r, L) = \begin{cases} \alpha^{-1}, & \alpha > 0, \\ +\infty, & \alpha \leq 0. \end{cases} \tag{8}$$

For $r > 1, L > 0$, the finite regime is equivalent to $\frac{r-1}{r+1}e^{L/2} < 1$, or $r < \coth(L/4)$. For the chain in equation (37)*, the norm converges to α^{-1} below this threshold, is $\Theta(n)$ at the threshold, and is $\Theta(((r-1)e^{L/2}/(r+1))^n)$ above it, for fixed r, L.*

Proof idea and sharpness. Balancing inverse decay against cross-scale amplification, band-limited conjugation bounds the symmetric part of BA^{-1} below by αI, giving $\|B^{-1}A\|_2 \leq \alpha^{-1}$ when $\alpha > 0$. For sharpness, take geometrically spaced eigenvalues of A and a tridiagonal Toeplitz relative matrix C. The inverse kernel decays at rate $q = (r-1)/(r+1)$ per step, while similarity amplifies each step by $e^{L/2}$. The chain approaches α^{-1} when $qe^{L/2} < 1$. Its inverse norm grows linearly with dimension when $qe^{L/2} = 1$ and exponentially when $qe^{L/2} > 1$. Appendix B gives the full proof.

The result is uniform over dimension and conditioning. For fixed L and $r = 1 + \epsilon$, it yields $\phi(A^2, B^2) = 1 + O_L(\epsilon)$. Thus sufficiently local perturbations permit accurate reuse even as the global condition number grows. Figure 1 shows the transition and the role of spectral support. Appendix D extends the analysis to soft spectral decay.

Comparison with weighted inverse bounds. For comparison, write $q = (r-1)/(r+1)$ and $t = L/2$. Chebyshev expansion of C^{-1}, followed by band-limited conjugation, gives a uniform bound whenever $qe^t < 1$:

$$G_{\mathrm{Ch}}(r, L) = \frac{1 + qe^t}{1 - qe^t}, \qquad \frac{G_{\mathrm{Ch}}(r, L)}{\mathcal{I}(r, L)} = 1 + h\sinh t \in [1, 2). \tag{9}$$

Both bounds have the same finite regime over arbitrary spectral placements. At $r = 5/2$, $L = \log 4$, the Chebyshev bound is 13 and the optimal norm is $80/11$. Proposition J.1 gives the derivation. For a cached update length q_0, the optimal upper test rejects a candidate when $q_0 < \tau/\mathcal{I}(r, L)$; the Chebyshev test requires $q_0 < \tau/G_{\mathrm{Ch}}(r, L)$.

Exponential conjugation and the scalar-chain decay rate are classical (Combes & Thomas, 1973; Demko et al., 1984; Janas et al., 2018). The exact Hermitian-part estimate in the block-Jacobi proof also recovers the sharp scalar-chain norm. Proposition J.3 relates that argument to the optimal norm over the full class here.

5 THE UNRESTRICTED REUSE BOUNDARY

Without a locality constraint, conditioning determines the optimal worst-case distortion for every $r, \kappa \geq 1$.

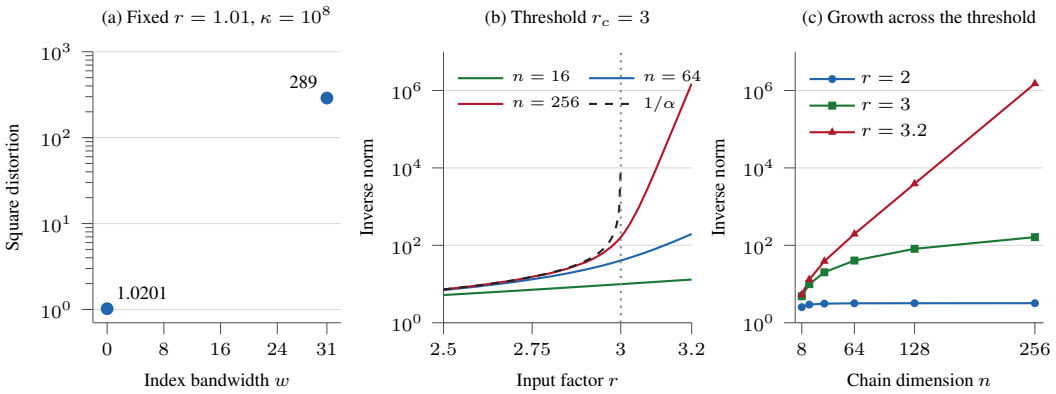

Figure 1: Spectral coupling and the reuse threshold. (a) Recorded median square distortions at the two endpoint bandwidths, with 120 draws each and fixed $n = 32$, $r = 1.01$, $\kappa = 10^8$. (b) Direct finite-chain inverse norms across $r_c = \coth(L/4) = 3$ for $L = \log 4$; the dashed curve is the uniform optimum $1/\alpha$ below r_c. (c) The inverse norm remains bounded, grows linearly, and grows exponentially below, at, and above r_c, respectively.

Theorem 5.1 (Exact square distortion)**.** *For $n \geq 2$, let*

$$z_{r,\kappa} = \frac{(r - r^{-1})(\sqrt{\kappa} + \kappa^{-1/2})}{4}.$$

Then

$$\mathcal{W}_n(r, \kappa) = \left(\sqrt{1 + z_{r,\kappa}^2} + z_{r,\kappa}\right)^2 = \exp(2 \operatorname{arsinh} z_{r,\kappa}). \tag{10}$$

The following two-dimensional pair attains equality:

$$A = \begin{pmatrix} \kappa & 0 \\ 0 & 1 \end{pmatrix}, \qquad P = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad C = r^{-1}I + (r - r^{-1})P, \qquad B = A^{1/2} C A^{1/2}. \tag{11}$$

Identical scalar blocks extend this construction to every $n > 2$.

Proof idea. Convexity reduces the proof to projection geometry: for fixed $S = A^{1/2}$, the norm $\|SCS^{-1}\|_2$ is convex in C. The extreme points of $[r^{-1}I, rI]$ are $r^{-1}I + (r - r^{-1})P$ for orthogonal projections P. The norm $\|SPS^{-1}\|_2$ is at most $(\sqrt{\kappa} + \kappa^{-1/2})/2$; equality holds at condition number κ by equal mixing of the extreme eigenspaces. A two-dimensional singular-value calculation gives equation (10), and the same argument for C^{-1} completes the bound. Appendix A gives the full proof.

Proposition 5.2 (Accuracy and conditioning)**.** *Every feasible pair satisfies*

$$\phi(A^2, B^2) \leq \left(1 + (r - 1)\sqrt{\kappa}\right)^2 \leq (1 + (r - 1)\kappa)^2. \tag{12}$$

Uniformly over $n \geq 2$, $\kappa \geq 1$, and $0 < \epsilon \leq 1/2$,

$$\mathcal{W}_n(1 + \epsilon, \kappa) - 1 = \Theta\left(\epsilon\sqrt{\kappa} + \epsilon^2\kappa\right). \tag{13}$$

Along any sequence with $\epsilon \downarrow 0$, the worst-case error vanishes if and only if $\epsilon\sqrt{\kappa} \to 0$.

Where the exact constant matters. Let $G = \mathcal{W}_n(1 + \epsilon, \kappa) - 1$ and $E = (1 + \epsilon\sqrt{\kappa})^2 - 1$ be the exact and elementary excesses. As $\epsilon \downarrow 0$ and $\kappa \to \infty$, with $x = \epsilon\sqrt{\kappa}$ fixed, their ratio tends to

$$\frac{2 + x}{\sqrt{1 + x^2/4} + x/2}. \tag{14}$$

The limiting ratio decreases from 2 at $x = 0$ to 1 as $x \to \infty$ (corollary A.5). Thus the relative improvement is largest in the small-error reuse regime. At $\epsilon = 0.01$ and $\kappa = 10^2, 10^4, 10^6$, E/G equals 1.99, 1.87, and 1.20. For 500 random extreme points at each configuration with $n = 32$, the median observed excess divided by G is 0.591, 0.329, and 0.132; the attaining pair gives a ratio of one (figure 2 and appendix N).

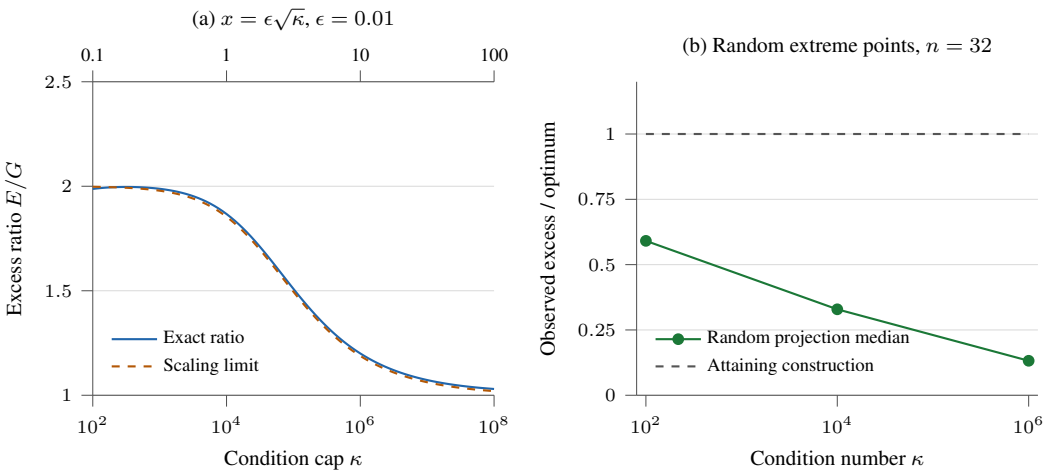

Figure 2: The exact constant at $\epsilon = 0.01$. (a) Elementary-to-optimal excess ratio; the upper axis shows $x = \epsilon\sqrt{\kappa}$ and the dashed curve is equation (14). (b) Median observed excess relative to the optimum for 500 random extreme points per condition number, with $n = 32$. The explicit attaining pair has ratio one.

From a threshold to a reusable interval. Both regimes constrain the same update length. The following consequence combines them and includes nonzero entries outside the local band.

Corollary 5.3 (One interval for global and local reuse)**.** *Suppose $\ell A \preceq B \preceq uA$, $0 < \ell \le u$, and the condition number of at least one matrix is at most κ. Set $c = (\ell + u)/2$, $h = (u - \ell)/2$, and write $C - cI = D_L + R$, where D_L is symmetric and has log-spectral width L. Let $s \ge \|R\|_2$ and $\eta \ge \|A^{1/2} R A^{-1/2}\|_2$. Define*

$$f = \sqrt{\ell u\, \mathcal{W}_n(\sqrt{u/\ell}, \kappa)},$$
$$\Gamma = c + (h + s)e^{L/2} + \eta, \qquad \alpha_L = c - (h + s)\cosh(L/2) - \eta. \tag{15}$$

For $q_0 = \|A^{-1}g\|_2$ and $q = \|B^{-1}g\|_2$,

$$q_0 \max\{f^{-1}, \Gamma^{-1}\} \le q \le q_0 \min\{f/(\ell u), b_L\}, \qquad b_L = \begin{cases} \alpha_L^{-1}, & \alpha_L > 0, \\ +\infty, & \alpha_L \le 0. \end{cases} \tag{16}$$

Proof. We center the input interval to obtain the unrestricted endpoints $(f^{-1}, f/(\ell u))$ (corollary F.5). Since $\|C - cI\|_2 \le h$, the local part satisfies $\|D_L\|_2 \le h + s$. The conjugation argument in theorem 4.1 bounds the norm of BA^{-1} by Γ and its Hermitian part below by $\alpha_L I$. Transposition gives the update bounds, and their intersection gives the stated interval. □

For a banded relative matrix, $s = \eta = 0$, and the local upper endpoint recovers the sharp inverse bound. For a measured tail, its unweighted and weighted norms enter separately, with $\|D_L\|_2 \le h + s$ controlling the truncated matrix. An initial error radius e_0 replaces q_0 by $\max\{0, q_0 - e_0\}$ in the lower endpoint and $q_0 + e_0$ in the upper. Without structural information, the unrestricted interval applies.

The centered Neumann argument instead uses the sufficient gap $c - (h + s)e^{L/2} - \eta$. The Chebyshev bound extends to asymmetric intervals and weighted tails through a resolvent perturbation bound (proposition J.2). All three local screens intersect their bounds with the global interval.

6 Application: Curvature-Sensitive Example Selection

We apply these bounds to a frozen-feature logistic head with damped curvature

$$H_t = \lambda I + \frac{1}{n}\sum_i w_{t,i} x_i x_i^\top, \qquad \lambda > 0.$$

At a base checkpoint, cache the factorization of H_0 and the solutions $H_0^{-1}x$ for a fixed candidate pool. A candidate gradient is a scalar multiple of its frozen feature vector, so we obtain its cached update $H_0^{-1}g_i$ by rescaling. In the local quadratic model, a change of δ in the loss weight of an example gives

$$\arg\min_v \left\{ \tfrac{1}{2} v^\top H_t v + \delta g_i^\top v \right\} = -\delta H_t^{-1} g_i. \tag{17}$$

We retain candidates with sensitivity $\|H_t^{-1}g_i\|_2 > \tau$. Their gradients determine the next update with a fixed training preconditioner.

From weights to update lengths. We bound curvature drift through the weight ratios $\ell = \min\{1, \min_i w_{t,i}/w_{0,i}\}$ and $u = \max\{1, \max_i w_{t,i}/w_{0,i}\}$. The positive rank-one summands give $\ell H_0 \preceq H_t \preceq uH_0$, so $r = \max\{u, \ell^{-1}\}$ is a valid symmetric input factor. For screening, we retain $[\ell, u]$ and tighten it using shared damping:

$$\ell_\lambda = \frac{\lambda + \ell L_0}{\lambda + L_0}, \qquad u_\lambda = \frac{\lambda + uL_0}{\lambda + L_0}, \qquad L_0 \geq \|H_0 - \lambda I\|_2. \tag{18}$$

Cached matrix and feature norms also give a condition-number bound $\bar{\kappa}$ (proposition F.4). These quantities require $O(n)$ work once the logits are available.

With $a = \sqrt{\ell_\lambda u_\lambda}$ and $\rho = \sqrt{u_\lambda/\ell_\lambda}$, set $f = a\sqrt{\mathcal{W}_d(\rho, \bar{\kappa})}$. The unrestricted update interval is

$$\frac{\|H_0^{-1}g\|_2}{f} \leq \|H_t^{-1}g\|_2 \leq \frac{f}{\ell_\lambda u_\lambda}\|H_0^{-1}g\|_2. \tag{19}$$

When a spectral width and tail bound are available, intersect this interval with the local one in equation (16). In a locally supported design with diagonal base covariance, the support gives L directly. Globally supported rows contribute an explicit tail Gram matrix; its construction and norm bounds are part of the screening cost (appendix K).

Screening and refinement. If the resulting interval lies entirely on one side of τ, return the corresponding decision. Otherwise evaluate the target residual of the cached update and run H_0-preconditioned conjugate gradients (PCG) until a residual interval separates the threshold. For a cached lower bound $m_0 \leq \lambda_{\min}(H_0)$, with $m_0 \geq \lambda$, the weight comparison gives

$$H_t \succeq \ell H_0 + (1-\ell)\lambda I \succeq \mu_t I, \qquad \mu_t = \lambda + \ell(m_0 - \lambda). \tag{20}$$

For any iterate v, the residual error radius is $e_t(v) = \|g - H_t v\|_2/\mu_t$, giving the length interval $[\max\{0, \|v\|_2 - e_t(v)\}, \|v\|_2 + e_t(v)]$. Cached solve errors use m_0 in the denominator.

The similarity baseline replaces f by $a(1 + (\rho - 1)\sqrt{\bar{\kappa}})$ and uses the same damping, cache, residual test, and solver. The exact law resolves additional candidates when its narrower interval separates the threshold. Each returned decision agrees with the target solve, so the screen preserves the retained set and the subsequent deterministic training trajectory. Appendix F proves the batch geometry and input bounds.

Computational cost. For n examples, dimension d, and Q candidates, a dense cache costs $O(nd^2 + d^3 + Qd^2)$ and stores $O(d^2 + dQ)$ numbers. Dense HVPs cost $O(d^2)$ after curvature construction; matrix-free HVPs cost $O(nd)$. A cached eigenbasis gives diagonal preconditioning and lets the local screen reuse the target Gram matrix, adding an $O(d^2)$ tail reduction per step. Dense direct selection pays $O(nd^2 + d^3 + Qd^2)$ per step; tridiagonal factorization and solves cost $O(d(1 + Q))$ after band construction. Appendix K gives the full cost accounting.

Proposition 6.1 (When reuse reduces runtime)**.** *Consider a cache window of R batches and an additive cost model. Let S_m be setup cost, C_m the mean per-batch cost outside iterative products, v the common cost per HVP and associated preconditioning work, and h_m mean HVPs per batch, for local and global methods $m \in \{L, G\}$ sharing a solver and representation. Then $T_m = S_m + R(C_m + vh_m)$, and*

$$T_L < T_G \quad \Longleftrightarrow \quad v(h_G - h_L) > C_L - C_G + \frac{S_L - S_G}{R}. \tag{21}$$

If direct selection has setup S_D and mean batch cost C_D, then

$$T_L < T_D \quad \Longleftrightarrow \quad R(C_D - C_L - vh_L) > S_L - S_D. \tag{22}$$

The differences in total cost give both equivalences directly. For dense spectral reuse, $C_L - C_G$ includes the $O(d^2)$ tail reduction. For known local support it includes only scalar bound evaluation. These conditions separate the value of a tighter interval from the cost of obtaining it and the number of batches sharing the cache.

7 EXPERIMENTS

Our experiments address three questions: do finite chains follow the predicted transition, do sharper intervals reduce solver work while preserving decisions, and when do these reductions lower total runtime? We report Hessian–vector products (HVPs) and runtime, with test accuracy and binary cross-entropy for the trained classifier. One HVP is one application of the target curvature to one vector; runtime includes curvature construction and solver work.

7.1 SPECTRAL SUPPORT AND THE PHASE TRANSITION

To isolate the transition, fix $r = 1.01$, $n = 32$, and $\kappa = 10^8$. Widening the relative matrix from diagonal to full support changes median square distortion from 1.0201 to 289 over 120 draws per bandwidth (figure 1a). For the attaining chain with $L = \log 4$, we scan 24 input factors across $r_c = 3$, including 2.995, 3, and 3.005, at $n = 16, 64, 256$. Direct banded solves and singular values give the measured inverse norms. At $n = 256$, these are 36.33, 162.74, and 1.50×10^6 for $r = 2.9, 3, 3.2$. At criticality, increasing n from 128 to 256 changes the norm from 81.25 to 162.74; at $r = 2$ it approaches 3.2. Appendix N gives the construction, scan, and corresponding preconditioned-solve HVP counts.

7.2 SPECTRAL LOCALITY

We next test locality on synthetic 64-dimensional logistic data with diagonal base covariance and rows supported on adjacent coordinate pairs, with covariance condition numbers 10^2, 10^4, and 10^6. The row support determines the log-spectral width in $O(d)$ time. Across five seeds, five quantiles, and eight steps, locality reduces HVPs from 595.8 for the global exact screen to 189.3 per batch (68.2%), versus 204.4 for local Neumann. In the four-screen comparison, selection takes 1.28, 1.68, and 1.30 ms for locality, global exact, and local Neumann; banded direct selection takes 0.76 ms. The Chebyshev screen uses 193.4 HVPs per batch on the adjacent-coupling design, compared with 189.3 for the optimal local bound. With wider coupling at fixed global conditioning, these counts are 278.9 and 263.6. Table 5 includes nonzero tails and runtimes. Tail ablations and cumulative runtime are in figures 4 and 5.

The selection-time reduction over global exact is 24%. The additional HVP reductions over Chebyshev are 2.1% and 5.5%. For each seed, we average differences over its 120 matched batches. Relative to Chebyshev, locality saves 2.80–5.39 HVPs per batch on adjacent coupling and 13.10–17.68 on wider coupling. These selection-time savings are positive for all five seeds in both designs; table 6 reports every paired difference.

For dense random features ($d = 128$, $Q = 512$), we diagonalize H_0 and work in its eigenbasis. The spectral global, local, and Chebyshev screens use the projected features and diagonal preconditioner. Each local screen reuses its target Gram matrix to bound entries outside $L \in \{\log 2, \log 4, \log 8\}$. At $L = \log 2$, locality uses 424.0 HVPs per batch, compared with 452.1 for the spectral global screen and 425.9 for Chebyshev. Including initialization and 32 selection and training steps, total runtime is 122.8 ms for locality, 120.0 ms for the spectral global screen, and 143.3 ms for direct selection. The implementation with separate target and tail Gram matrices takes 175.8 ms. At $d = 512$ and $L = \log 2$, locality takes 1163.1 ms, compared with 1127.1 ms for the spectral global screen and 1361.0 ms for direct selection. The spectral global screen has the lowest mean runtime at both dimensions. Tables 14 and 15 report all methods.

7.3 UNRESTRICTED REUSE

We test unrestricted reuse on a 65-parameter digits parity head. Five seeds, three damping levels, five threshold quantiles, and eight training steps give 600 batches of 256 candidates. The exact and similarity screens use the asymmetric interval and PCG refinement in appendix K. The exact bound

reduces HVPs relative to the similarity bound from 192.1 to 173.8 per batch (9.5%) and selection time from 1.36 to 1.34 ms. Direct Cholesky takes 1.23 ms. The symmetric-interval and threshold ablations are in appendix K.

We then use random Fourier features of the same digits (Rahimi & Recht, 2007), with dimensions $d \in \{128, 512, 1024\}$ and candidate counts $Q \in \{128, 512\}$. Each method runs 32 selection and training steps. Timing starts with base curvature construction and includes calibration, candidate caching, screening, refinement, and training updates. We compare dense and matrix-free implementations of exact screening, similarity screening, and PCG, together with direct Cholesky. For each seed, we perform one warm-up followed by five timed runs.

Table 1: Total time (seconds) for initialization and 32 selection and training steps on random features. MF denotes matrix-free HVPs. Entries are mean ± standard deviation over five seeds, using the median of five runs per seed. The cache stays fixed for 32 steps. All solver implementations and cache-refresh comparisons are in appendix L.

d	Q	Exact dense	Exact MF	Similarity MF	PCG MF	Direct
128	128	0.088 ± 0.004	0.156 ± 0.030	0.164 ± 0.026	0.164 ± 0.028	0.081 ± 0.002
128	512	0.129 ± 0.007	0.392 ± 0.040	0.426 ± 0.040	0.486 ± 0.051	0.155 ± 0.006
512	128	0.796 ± 0.022	0.618 ± 0.044	0.631 ± 0.039	0.644 ± 0.045	0.914 ± 0.030
512	512	1.190 ± 0.078	1.521 ± 0.171	1.603 ± 0.164	1.769 ± 0.135	1.426 ± 0.030
1024	128	2.525 ± 0.069	1.316 ± 0.139	1.375 ± 0.108	1.417 ± 0.090	3.169 ± 0.054
1024	512	3.479 ± 0.222	2.845 ± 0.361	3.055 ± 0.330	3.336 ± 0.297	4.388 ± 0.088

Table 1 shows how dimension and candidate count change the relative cost of direct, dense iterative, and matrix-free solvers. At $d = 1024$, $Q = 128$, matrix-free exact selection takes 1.32 s, matrix-free PCG takes 1.42 s, and direct selection takes 3.17 s. The matrix-free similarity screen takes 1.38 s; its HVP count changes from 137.9 to 124.0 per batch with the exact bound. At $Q = 512$ with refreshes every eight steps, matrix-free exact selection takes 1.98 s, compared with 2.25 s for similarity, 3.49 s for PCG, and 4.41 s for direct selection. Figure 6 shows cumulative runtime, including initialization and refreshes; table 10 compares both refresh intervals.

Operating regimes. The costs in proposition 6.1 explain three outcomes. Known local support reduces both HVPs and selection time relative to global screening, although banded direct solves remain fastest. For dense features at $d = 128, Q = 512$, the local screen saves 6.2% of HVPs over spectral global screening but takes 2.3% longer in total; the added tail reduction exceeds the saved solver work. Repeated dense selection amortizes the cache at larger dimensions, whereas direct Cholesky is faster at $d = 128, Q = 128$ and in the single-batch frozen-feature experiments (table 16).

Scope. The guarantees concern update lengths in fixed parameter coordinates, with positive-definite curvature and valid input and tail bounds. The method preserves decisions when the certified interval separates the threshold and resolves remaining candidates by refinement or a direct solve. The experiments test this mechanism with fixed feature maps and candidate pools, including frozen language-model representations in appendix M. These experiments measure computational savings for the same selector, whose predictive accuracy remains unchanged.

8 Conclusion

Log-spectral locality gives a sharp transition for curvature reuse: dimension-independent inverse amplification below the threshold, linear growth at criticality, and exponential growth above it. The complementary unrestricted law identifies the required input accuracy and supplies the optimal global interval. The intersection of these intervals preserves example-selection decisions and reduces solver work. For locally supported curvature, the method reduces both HVP counts and selection time relative to global screening. For dense curvature, tail evaluation costs and the number of batches that share a cache determine whether these savings reduce total runtime. An important next step is to estimate spectral tails without forming a dense relative matrix, so that structured reuse can remain affordable when the curvature operator is available only through matrix–vector products.

AI USE STATEMENT

We used generative AI to assist with experimental code, improve the language and presentation of the manuscript, and support mathematical derivations and proof revisions. For the theoretical analysis, we used AI to audit derivations for algebraic errors, missing assumptions, and inconsistencies between claims and proofs. AI also assisted with the Chinese translation prepared for reading. Verification materials include the full proofs, numerical checks of the stated formulas and finite-chain construction, and comparisons of selection decisions with direct solves. These checks support review of the AI-assisted work; numerical checks do not replace proof verification. The authors take responsibility for the final claims, proofs, code, experiments, references, and text.

ETHICS STATEMENT

The experiments use public digits and BoolQ datasets and a public Pythia checkpoint. The references credit the datasets and model. The evaluated tasks are curvature-based example selection in logistic training and frozen-feature checkpoint reuse.

REPRODUCIBILITY STATEMENT

The appendix provides the proofs, including the exact global law in appendix A and the optimal local bound in appendix B. Appendices K to N specify data processing, splits, seeds, damping, solver rules, and timing. Appendix O lists experiment entry points and distinguishes available checks from historical artifacts that require additional raw records for full regeneration. Anonymous supplementary code includes the experiment entry points and a runnable verification entry point. The figures use the stated formulas and reported experimental summaries.

REFERENCES

A. B. Aleksandrov and V. V. Peller. Functions of perturbed unbounded self-adjoint operators. Operator Bernstein type inequalities. *Indiana University Mathematics Journal*, 59(4):1451–1490, 2010.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. The experiments use the 1,797-example subset bundled with scikit-learn.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Michele Benzi and Paola Boito. Decay properties for functions of matrices over C^*-algebras. *Linear Algebra and its Applications*, 456:174–198, 2014.

Michele Benzi and Nader Razouk. Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices. *Electronic Transactions on Numerical Analysis*, 28:16–39, 2007.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

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

D. Braess and P. Peisker. On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning. *IMA Journal of Numerical Analysis*, 6(4):393–404, 1986.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

J. M. Combes and L. Thomas. Asymptotic behaviour of eigenfunctions for multiparticle Schrödinger operators. *Communications in Mathematical Physics*, 34:251–270, 1973.

Stephen Demko, William F. Moss, and Philip W. Smith. Decay rates for inverses of band matrices. *Mathematics of Computation*, 43(168):491–499, 1984.

Hwa-Long Gau, Chin-Ying Huang, and Pei Yuan Wu. Numerical radius inequalities for square-zero and idempotent operators. *Operators and Matrices*, 2(1):137–141, 2008.

Serge Gratton, Selime Gürol, Ehouarn Simon, and Philippe L. Toint. A note on preconditioning weighted linear least-squares, with consequences for weakly constrained variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(712):934–940, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1842–1850, 2018.

Jan Janas, Sergey Naboko, and Luis O. Silva. Green matrix estimates of block Jacobi matrices I: Unbounded gap in the essential spectrum. *Integral Equations and Operator Theory*, 90(4), 2018. Article 49.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations*, 2024.

Bas Lemmens and Mark Roelands. Unique geodesics for Thompson's metric. *Annales de l'Institut Fourier*, 65(1):315–348, 2015.

A. W. Marshall and I. Olkin. Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40:89–93, 1990.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numerical Algorithms*, 42(3–4):309–323, 2006.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

S.-H. Tso and Pei Yuan Wu. Matricial ranges of quadratic operators. *Rocky Mountain Journal of Mathematics*, 29(3):1139–1152, 1999.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

Andy Wathen. Some comments on preconditioning for normal equations and least squares. *SIAM Review*, 64(3):640–649, 2022.

A Proofs for the Exact Minimax Theorem

The exact law in theorem 5.1 rests on two reductions: the Loewner interval is reduced to its projection-valued extreme points, and the resulting similarity transform is controlled by projection geometry. For $n = 1$, scalar comparison already gives $\mathcal{W}_1(r, \kappa) = r^2$. For $n \geq 2$, we begin with the geometric ingredient that introduces the condition-number dependence.

Lemma A.1 (Similarity-transformed orthogonal projections)**.** *Let $S \in \mathbb{S}^n_{++}$, $n \geq 2$, and $s = \kappa_2(S)$. Then*

$$\max_{P=P^\top=P^2} \left\|SPS^{-1}\right\|_2 = \frac{s + s^{-1}}{2}. \tag{23}$$

Proof. The cases $P = 0$ and $P = I$ have norms zero and one, respectively. For a nontrivial P, let $\mathcal{L} = \mathrm{range}(P)$. The idempotent $Q = SPS^{-1}$ projects onto $S\mathcal{L}$ along $S\mathcal{L}^\perp$. Let θ be the smallest angle between these subspaces. The standard oblique-projection identity then gives $\|Q\|_2 = 1/\sin\theta$ (Szyld, 2006).

For $u \in \mathcal{L}$ and $v \in \mathcal{L}^\perp$, $u^\top v = 0$ and

$$\frac{|\langle Su, Sv\rangle|}{\|Su\|_2 \|Sv\|_2} = \frac{|u^\top S^2 v|}{\sqrt{u^\top S^2 u}\sqrt{v^\top S^2 v}}. \tag{24}$$

We give a short proof of the Wielandt inequality needed here. Compress S^2 to the span of normalized u and v. If the compression has eigenvalues $\mu_1 \leq \mu_2$, then $\lambda_{\min}(S^2) \leq \mu_1 \leq \mu_2 \leq \lambda_{\max}(S^2)$. After a planar rotation, the squared ratio in equation (24) is

$$\frac{(\mu_2 - \mu_1)^2 t}{\mu_1\mu_2 + (\mu_2 - \mu_1)^2 t}, \qquad 0 \leq t \leq \frac{1}{4}.$$

This expression increases with t and the eigenvalue ratio. Since $\lambda_{\max}(S^2)/\lambda_{\min}(S^2) = s^2$, the ratio in equation (24) is at most

$$\frac{s^2 - 1}{s^2 + 1}.$$

This is the classical Wielandt form of the Kantorovich inequality. Therefore

$$\sin\theta \geq \sqrt{1 - \left(\frac{s^2 - 1}{s^2 + 1}\right)^2} = \frac{2s}{s^2 + 1},$$

which yields the upper bound in equation (23).

To attain equality, take unit eigenvectors $u_{\max}, u_{\min}$ of S for its largest and smallest eigenvalues. Let P project onto $(u_{\max} + u_{\min})/\sqrt{2}$. Restrict S to their span and divide it by its smallest eigenvalue to obtain $S = \mathrm{diag}(s, 1)$. A direct rank-one calculation yields

$$\left\|SPS^{-1}\right\|_2 = \frac{1}{2}\sqrt{s^2 + 1}\sqrt{s^{-2} + 1} = \frac{s + s^{-1}}{2}.$$

□

Lemma A.2 (An operator interval under similarity)**.** *Fix $S \in \mathbb{S}^n_{++}$, $n \geq 2$, let $s = \kappa_2(S)$ and $\chi_s = (s + s^{-1})/2$. For $0 < \ell < u$,*

$$\sup_{\ell I \preceq C \preceq uI} \left\|SCS^{-1}\right\|_2 = f_{\ell,u}(\chi_s), \tag{25}$$

where

$$f_{\ell,u}(x) = \frac{1}{2}\left[\sqrt{4\ell u + (u - \ell)^2 x^2} + (u - \ell)x\right]. \tag{26}$$

Consequently, if $\kappa_2(S) \leq q$, the left side of equation (25) *is at most $f_{\ell,u}((q + q^{-1})/2)$. Some S with $\kappa_2(S) = q$ attains this uniform bound.*

Proof. The map $C \mapsto \left\|SCS^{-1}\right\|_2$ is convex and continuous. The operator interval $\mathcal{C} = \{C = C^\top : \ell I \preceq C \preceq uI\}$ is compact and convex. Its extreme points are precisely

$$C = \ell I + (u - \ell)P, \tag{27}$$

where P is an orthogonal projection. Indeed, after normalizing to $0 \preceq D \preceq I$, any eigenvalue of D strictly between zero and one can be perturbed in both directions, so D is not extreme. Conversely, if $D = P$ is a projection and $P = (D_1 + D_2)/2$ with $0 \preceq D_i \preceq I$, restriction to the range and nullspace of P forces $D_1 = D_2 = P$.

A convex function on a compact convex set attains its maximum at an extreme point. To see this, express a maximizer as a convex combination of extreme points and apply convexity. We therefore only need to analyze equation (27). The cases $P = 0$ and $P = I$ give norms ℓ and u, respectively. For nontrivial P, set $Q = SPS^{-1}$. Relative to $\mathrm{range}(Q) \oplus \mathrm{range}(Q)^\perp$,

$$Q = \begin{pmatrix} I & X \\ 0 & 0 \end{pmatrix}, \qquad \|Q\|_2^2 = 1 + \|X\|_2^2. \tag{28}$$

Hence

$$\ell I + (u - \ell)Q = \begin{pmatrix} uI & (u-\ell)X \\ 0 & \ell I \end{pmatrix}. \tag{29}$$

This is the classical canonical model for a quadratic operator: see Tso & Wu (1999) and Gau et al. (2008, Proposition 1(b,c)). Apply the singular value decomposition to X. Orthogonal transformations reduce equation (29) to scalar blocks and two-dimensional blocks

$$T_\gamma = \begin{pmatrix} u & (u-\ell)\gamma \\ 0 & \ell \end{pmatrix}, \tag{30}$$

where the largest γ equals $\sqrt{k^2 - 1}$ with $k = \|Q\|_2$. Let $\sigma_1 \geq \sigma_2$ be the singular values of this largest block. Then

$$\begin{aligned} \sigma_1\sigma_2 &= \ell u, \\ \sigma_1^2 + \sigma_2^2 &= u^2 + \ell^2 + (u-\ell)^2(k^2 - 1) \\ &= 2\ell u + (u-\ell)^2 k^2. \end{aligned}$$

Therefore

$$\sigma_1 + \sigma_2 = \sqrt{4\ell u + (u-\ell)^2 k^2}, \qquad \sigma_1 - \sigma_2 = (u - \ell)k,$$

and $\sigma_1 = f_{\ell,u}(k)$. This value is at least u and increases with k. By lemma A.1, the largest possible k for the fixed S is χ_s, and equality is attainable. This proves equation (25). Monotonicity in $s \geq 1$ proves the uniform statement. □

Proof of theorem 5.1. If $r = 1$, then $\phi(A, B) = 1$ forces $A = B$, and the result is immediate. Hence assume $r > 1$.

Let

$$S = A^{1/2}, \qquad C = A^{-1/2}BA^{-1/2}.$$

The input constraints imply $r^{-1}I \preceq C \preceq rI$ and $\kappa_2(S) = \sqrt{\kappa_2(A)} \leq q = \sqrt{\kappa}$. Moreover,

$$BA^{-1} = SCS^{-1}, \qquad AB^{-1} = SC^{-1}S^{-1},$$

and C^{-1} belongs to the same operator interval. Apply lemma A.2 to C and C^{-1} with $\ell = r^{-1}$ and $u = r$. Since $\ell u = 1$,

$$\begin{aligned} \max\{\left\|BA^{-1}\right\|_2, \left\|AB^{-1}\right\|_2\} &\leq f_{r^{-1},r}\left(\frac{q + q^{-1}}{2}\right) \\ &= \sqrt{1 + z_{r,\kappa}^2} + z_{r,\kappa}. \end{aligned}$$

We square this inequality and apply proposition 3.1 to obtain the upper bound.

The pair in equation (11) attains equality in dimension two. The matrix C has eigenvalues r and r^{-1}, so $\phi(A, B) = r$, while $\kappa_2(A) = \kappa$. The projection P attains equality in lemma A.1, as do all subsequent inequalities in lemma A.2. For $n > 2$, take direct sums of A, B, C with identical scalar blocks I_{n-2}. This extension preserves both the constraints and the extremal norm. □

Proof of proposition 5.2. Both $C - I$ and $C^{-1} - I$ have norm at most $r - 1$. Similarity by $A^{1/2}$ and the triangle inequality prove the first bound; $\sqrt{\kappa} \le \kappa$ proves the second. For $r = 1 + \epsilon$, $z_{r,\kappa}$ is bounded above and below by universal positive multiples of $x = \epsilon\sqrt{\kappa}$. Since $\mathcal{W}_n - 1 = 2z(\sqrt{1+z^2} + z) = \Theta(z + z^2)$, the joint rate and the equivalence follow. □

Corollary A.3 (Exact input accuracy for a target distortion)**.** *For $n \ge 2$ and $\beta \ge 0$, a uniform output factor at most $1 + \beta$ holds if and only if*

$$r \le \sqrt{1 + c_{\beta,\kappa}^2} + c_{\beta,\kappa}, \qquad c_{\beta,\kappa} = \frac{\beta}{\sqrt{1+\beta}(\sqrt{\kappa} + \kappa^{-1/2})}. \tag{31}$$

Proof. Solve $2\,\mathrm{arsinh}\, z_{r,\kappa} \le \log(1 + \beta)$ for $r \ge 1$. Necessity follows from the attaining pair. □

Proposition A.4 (Elementary bound versus the exact law)**.** *For $r > 1$, $\kappa \ge 1$, and $n \ge 2$, put $G(r,\kappa) = \mathcal{W}_n(r,\kappa) - 1$ and $E(r,\kappa) = (1 + (r-1)\sqrt{\kappa})^2 - 1$. Then*

$$1 \le \frac{E(r,\kappa)}{G(r,\kappa)} \le C(r) := \frac{1 - r^2 + \sqrt{r^4 + 14r^2 + 1}}{2} < 4. \tag{32}$$

The constant four is optimal uniformly in (r,κ), and

$$\limsup_{\epsilon \downarrow 0} \sup_{\kappa \ge 1} \frac{E(1+\epsilon,\kappa)}{G(1+\epsilon,\kappa)} = 2. \tag{33}$$

Proof. The lower bound is proposition 5.2. Set $q = \sqrt{\kappa}$, $x = (r-1)q \ge r - 1$, and $c = (r+1)/(4r)$. Since $z_{r,\kappa} \ge cx$ and $G = 2z_{r,\kappa}(\sqrt{1 + z_{r,\kappa}^2} + z_{r,\kappa})$,

$$\frac{E}{G} \le f(x) := \frac{x+2}{2c(\sqrt{1 + c^2x^2} + cx)}.$$

The logarithmic derivative is $1/(x+2) - c/\sqrt{1+c^2x^2}$, which is nonpositive because $4c^2(x+1) \ge (r+1)^2/(4r) \ge 1$. Thus $f(x) \le f(r-1) = C(r)$. The positive root $C(r)$ solves $C(r)^2 + (r^2 - 1)C(r) - 4r^2 = 0$; comparison with four gives $C(r) < 4$. The choice $r \to \infty$ with $\kappa = r^2$ gives $E/G \to 4$. Finally $C(r) \to 2$ as $r \downarrow 1$, and the sequence $r = 1 + \epsilon, \kappa = \epsilon^{-1}$ has $E/G \to 2$. □

At finite parameters $r = 1.5, \kappa = 19$, the ratio is approximately 2.03677. For fixed $r > 1$, its limit as $\kappa \to \infty$ is $4r^2/(r+1)^2$.

Corollary A.5 (Relative gain in the small-error regime)**.** *Let $\epsilon_j \downarrow 0$, $\kappa_j \to \infty$, and $x_j = \epsilon_j\sqrt{\kappa_j} \to x_0 \in [0, +\infty]$. For the excesses in proposition A.4,*

$$\frac{E(1+\epsilon_j,\kappa_j)}{G(1+\epsilon_j,\kappa_j)} \longrightarrow F(x_0), \qquad F(x) = \frac{2+x}{\sqrt{1 + x^2/4} + x/2}, \quad F(0) = 2, \quad F(+\infty) = 1.$$

The function F is strictly decreasing on $[0, +\infty)$.

Proof. Set $z = z_{1+\epsilon,\kappa}$ and $x = \epsilon\sqrt{\kappa}$. Then

$$\frac{z}{x} = \frac{2+\epsilon}{4(1+\epsilon)}(1 + \kappa^{-1}) \longrightarrow \frac{1}{2}, \qquad \frac{E}{G} = \frac{2+x}{(2z/x)(\sqrt{1+z^2} + z)}.$$

These identities give the limit for finite x_0. Division of the last denominator by x gives the limit for $x_0 = +\infty$. Finally $(\log F)' = 1/(2+x) - 1/\sqrt{4+x^2} < 0$ for $x > 0$. □

The joint limit locates the factor-two gain in the ill-conditioned regime with vanishing distortion. At a fixed condition number, the corresponding $\epsilon \downarrow 0$ limit is $2\kappa/(\kappa+1)$.

Corollary A.6 (Refresh from an update-length tolerance)**.** *Let H_s, H_t be SPD curvature operators, with $\phi(H_s, H_t) \le r_{s,t}$ and $\kappa_2(H_s) \le \bar{\kappa}_s$. For $\beta > 0$, all squared update lengths obey*

$$(1+\beta)^{-1}\|H_s^{-1}g\|_2^2 \le \|H_t^{-1}g\|_2^2 \le (1+\beta)\|H_s^{-1}g\|_2^2$$

whenever

$$r_{s,t} \le \sqrt{1 + c_{\beta,\bar{\kappa}_s}^2} + c_{\beta,\bar{\kappa}_s}, \qquad c_{\beta,\bar{\kappa}_s} = \frac{\beta}{\sqrt{1+\beta}(\sqrt{\bar{\kappa}_s} + \bar{\kappa}_s^{-1/2})}. \tag{34}$$

This is the largest admissible factor using only $r_{s,t}$ and $\bar{\kappa}_s$ uniformly over SPD pairs.

Proof. Apply corollary A.3 to the squared geometry and use inversion invariance of ϕ. The attaining pair proves necessity for a uniform guarantee from the two scalar bounds. □

The factor $r_{s,t}$ gives a refresh rule: retain the cache while equation (34) holds and rebuild it when the test fails. Known locality can extend this window using equation (16). For a factored SPD preconditioner, apply the comparison to the composed operator acting on the vectorized gradient; its factors and matrix powers specify that operator. The guarantee controls update lengths for a given gradient and tolerance.

B Proof of the Locality Threshold

We now prove the local result that complements the unrestricted minimax law. The argument first establishes a dimension-free upper bound and then uses one finite-chain family to show sharpness in all three regimes.

Proof of theorem 4.1. If $r = 1$, then $A = B$ and the optimal norm is one. If $L = 0$, the relative matrix commutes with A, so the optimal norm is r. Both agree with equation (8). For the rest of the proof, assume $r > 1$ and $L > 0$.

Uniform upper bound. We first turn spectral locality into an analytic bound under conjugation. Write $X = \log A$, $D = C - cI$, and $F(z) = e^{izX} D e^{-izX}$. The interval $r^{-1}I \preceq C \preceq rI$ implies $\|D\|_2 \le h$. Every nonzero matrix entry of F has a frequency in $[-L, L]$, and $\|F(t)\|_2 = \|D\|_2 \le h$ for real t. Apply the classical scalar Bernstein inequality for bounded entire functions of exponential type to $u^* F(z) v$, for unit complex vectors u, v, and iterate the inequality. The supremum over u, v then gives

$$\|F^{(k)}(0)\|_2 \le hL^k \qquad (k \ge 0). \tag{35}$$

Aleksandrov & Peller (2010) recall the scalar inequality and its Banach-valued extension. For real t, $F(t)$ is Hermitian, so all its derivatives at zero are Hermitian. The bound in equation (35) ensures absolute convergence of the Taylor expansion at $-i/2$. Thus, for $N = A^{1/2} D A^{-1/2}$, we obtain

$$\begin{aligned} \|N\|_2 &\le h e^{L/2}, \\ \left\| \frac{N + N^*}{2} \right\|_2 &= \left\| \sum_{j=0}^{\infty} \frac{(-1)^j F^{(2j)}(0)}{2^{2j}(2j)!} \right\|_2 \le h \cosh(L/2). \end{aligned} \tag{36}$$

Thus $M = BA^{-1} = cI + N$ has Hermitian part at least αI. If $\alpha > 0$, every unit vector x satisfies $\|Mx\|_2 \ge \mathrm{Re}(x^* M x) \ge \alpha$. Consequently $\|M^{-1}\|_2 \le \alpha^{-1}$. Transposition gives $\|B^{-1}A\|_2 = \|M^{-1}\|_2$.

Sharpness below the threshold. To see that the upper bound cannot be improved, we construct a chain that approaches it as the dimension grows. Suppose $\alpha > 0$. Set $t = L/2$, let T_n be the lower shift matrix, and define

$$A_n = \mathrm{diag}(1, e^L, \dots, e^{(n-1)L}), \quad C_n = cI - \frac{h}{2}(T_n + T_n^\top), \quad B_n = A_n^{1/2} C_n A_n^{1/2}. \tag{37}$$

Since $\|T_n + T_n^\top\|_2 < 2$, C_n has spectrum in $[c - h, c + h] = [r^{-1}, r]$, and its support has width L. The relative matrix is

$$M_n = B_n A_n^{-1} = cI - \frac{h}{2}(e^t T_n + e^{-t} T_n^\top).$$

For $x_n = (1, \dots, 1)^\top/\sqrt{n}$, all interior components of $M_n x_n$ equal αx_n. The two endpoint discrepancies give

$$\|M_n x_n - \alpha x_n\|_2 = \frac{h}{2\sqrt{n}}\sqrt{e^{2t} + e^{-2t}} \longrightarrow 0.$$

Hence $\limsup_n s_{\min}(M_n) \le \alpha$, while the upper-bound argument gives $s_{\min}(M_n) \ge \alpha$. For $\alpha > 0$, $\|M_n^{-1}\|_2 \to \alpha^{-1}$.

Critical and supercritical divergence. The same chain also reveals what fails once the positive lower bound disappears. Now consider $\alpha \le 0$ and put $q = (r-1)/(r+1) \in (0, 1)$. Then

$$c = \frac{1+q^2}{1-q^2}, \quad h = \frac{2q}{1-q^2}, \quad \alpha = \frac{(1-qe^t)(1-qe^{-t})}{1-q^2}. \tag{38}$$

For $i \ge j$, direct solution of the three-term recurrence with zero boundary values gives

$$(C_n^{-1})_{ij} = q^{i-j}\frac{(1-q^{2j})(1-q^{2(n+1-i)})}{1-q^{2(n+1)}}. \tag{39}$$

The entries with $i < j$ follow by symmetry. To verify the formula, substitute it into $cv_i - \frac{h}{2}(v_{i-1} + v_{i+1}) = \mathbf{1}_{\{i=j\}}$, using $v_0 = v_{n+1} = 0$ and the values of c, h in equation (38). This is the finite-chain version of the classical tridiagonal Toeplitz decay construction (Demko et al., 1984). Diagonal similarity therefore gives, for $i \ge j$,

$$(M_n^{-1})_{ij} = (qe^t)^{i-j}\frac{(1-q^{2j})(1-q^{2(n+1-i)})}{1-q^{2(n+1)}}. \tag{40}$$

Since $qe^{-t} < 1$, $\alpha \le 0$ is equivalent to $qe^t \ge 1$. If $qe^t > 1$, the bottom-left entry alone implies

$$\|M_n^{-1}\|_2 \ge (qe^t)^{n-1}\frac{(1-q^2)^2}{1-q^{2(n+1)}} \longrightarrow \infty.$$

The entrywise upper bound by a constant times $(qe^t)^{i-j}$ also bounds the Frobenius norm by a constant times $(qe^t)^n$. Thus growth is exponential in n. If $qe^t = 1$, every lower-triangular entry is at least $a_0 = (1-q^2)^2 > 0$, and all entries are nonnegative. Hence

$$\|M_n^{-1}x_n\|_2 \ge a_0\sqrt{n^{-1}\sum_{i=1}^{n} i^2} = \Omega(n).$$

All entries are also uniformly bounded, so the Frobenius norm is $O(n)$. The critical inverse norm is therefore $\Theta(n)$. This proves divergence when $\alpha \le 0$ and completes the theorem. □

Assume $\alpha > 0$. The proof bounds $\|BA^{-1}\|_2$ by $\Gamma = c + he^{L/2}$. For $U = A^{-1}G$ and $\widetilde{U} = B^{-1}G$, $\widetilde{U} = B^{-1}AU$, while $U = A^{-1}B\widetilde{U}$. The inverse bound and the transposed forward bound give

$$\Gamma^{-2}U^\top U \preceq \widetilde{U}^\top\widetilde{U} \preceq \alpha^{-2}U^\top U.$$

The upper coefficient is sharp across dimensions by taking a maximizing right-hand side for the chain; the displayed Γ is a sufficient forward coefficient. Likewise $\|\widetilde{U}v - Uv\|_2 \le (c-1+he^{L/2})\alpha^{-1}\|Uv\|_2$.

Finite-chain values. To make the three asymptotic regimes concrete, we evaluate the explicit inverse kernel at representative finite dimensions. It gives $\|B_n^{-1}A_n\|_2$ on $L = \log 4$, $r \in \{2, 3, 4\}$, and $n \in \{8, 16, 32, 64, 128\}$. At $n = 128$, the norms are 3.19423, 81.2517, and 2.79316×10^{10}, respectively. The first value approaches the finite limit 3.2, while the other two illustrate growth at and above the threshold.

C Residual and Eigenspace Bounds

The sharp global and local results give worst-case guarantees from coarse matrix information. This section records two complementary diagnostics that describe the distortion of a particular pair: a relative residual and the obliqueness of its invariant subspaces.

Let $d_T(A, B) = \log \phi(A, B)$ denote the Thompson distance (Lemmens & Roelands, 2015). The relative-operator identities give the nonnormality factor

$$\eta(A, B) = \frac{\max\{\|M\|_2, \|M^{-1}\|_2\}}{\max\{\rho(M), \rho(M^{-1})\}} \geq 1, \tag{41}$$

and hence

$$\begin{aligned} \phi(A^2, B^2) &= \phi(A, B)^2 \eta(A, B)^2, \\ d_T(A^2, B^2) &= 2d_T(A, B) + 2\log \eta(A, B). \end{aligned} \tag{42}$$

In the distance identity, the first term is the scalar contribution; the second records the extra noncommutative loss. If A and B commute, M is SPD and $\eta = 1$. Moreover, M is normal if and only if A and B commute: its real spectrum makes a real normal M symmetric, and $BA^{-1} = A^{-1}B$ is equivalent to $AB = BA$.

Two complementary quantities separate closeness to the identity from the nonorthogonality of invariant subspaces. Define the two-sided relative residual by

$$\nu(A, B) = \max\{\|BA^{-1} - I\|_2, \|AB^{-1} - I\|_2\}. \tag{43}$$

The triangle inequality and proposition 3.1 give

$$\phi(A^2, B^2) - 1 \leq 2\nu(A, B) + \nu(A, B)^2. \tag{44}$$

This residual bound supplements the input Loewner factor.

Let $M = BA^{-1}$ have distinct eigenvalues with Riesz projections $Q_1, \ldots, Q_s$, and define

$$\Pi(M) = \max_{J \subseteq \{1,\ldots,s\}} \left\| \sum_{j \in J} Q_j \right\|_2 .$$

The maximum includes the full index set, so $\Pi(M) \geq 1$ even when M is scalar.

Theorem C.1 (Eigenspace obliqueness bound)**.** *If $\phi(A, B) \leq r$ and $\kappa_2(A) \leq \kappa$, then*

$$\sqrt{\phi(A^2, B^2)} \leq r^{-1} + (r - r^{-1})\Pi(BA^{-1}), \tag{45}$$

and

$$1 \leq \Pi(BA^{-1}) \leq \frac{\sqrt{\kappa} + \kappa^{-1/2}}{2}. \tag{46}$$

Moreover, $\Pi(BA^{-1}) = 1$ if and only if $AB = BA$.

The following proof uses a telescoping spectral decomposition. The upper bound in equation (46) is sharp.

Proof of theorem C.1. Let the distinct eigenvalues of $M = BA^{-1}$ be $r^{-1} \leq \lambda_1 < \cdots < \lambda_s \leq r$, with Riesz projections $Q_1, \ldots, Q_s$. Define $R_j = \sum_{\ell=j}^{s} Q_\ell$. Spectral decomposition and telescoping give

$$M = \lambda_1 I + \sum_{j=2}^{s} (\lambda_j - \lambda_{j-1}) R_j. \tag{47}$$

All coefficients are nonnegative, hence

$$\|M\|_2 \leq \lambda_1 + (\lambda_s - \lambda_1)\Pi(M) \leq r^{-1} + (r - r^{-1})\Pi(M).$$

The inverse M^{-1} has the same Riesz projections in reverse eigenvalue order, so the same bound applies to $\|M^{-1}\|_2$. Proposition 3.1 then gives equation (45).

To prove equation (46), recall $M = SCS^{-1}$ with $S = A^{1/2}$ and symmetric C. Every sum of Riesz projections of M has the form SPS^{-1} for an orthogonal spectral projection P of C. Lemma A.1 gives $\Pi(M) \le (\sqrt{\kappa} + \kappa^{-1/2})/2$. A nonzero idempotent has norm at least one, which gives the lower bound. If $AB = BA$, then M is symmetric and all its spectral projections are orthogonal, so $\Pi(M) = 1$. Conversely, if $\Pi(M) = 1$, every nonzero Riesz projection has norm one and is therefore an orthogonal projection. The spectral decomposition of M is then symmetric, which implies $BA^{-1} = A^{-1}B$ and hence $AB = BA$. □

D Additional Structural Results

The main text focuses on hard spectral locality and the unrestricted boundary. Here we place those results beside related structural bounds, then derive a first-order directional law and a softer band-decay condition. Table 2 summarizes the roles of these assumptions.

D.1 Directional amplification

The global minimax law identifies the worst direction over all admissible pairs. To see which perturbation directions are amplified locally, we differentiate the input and output distortions at a fixed matrix A.

Theorem D.1 (Directional condition number)**.** *Let $B_t = A + tH$ with $H = H^\top$, and set $E = A^{-1/2}HA^{-1/2} \neq 0$. As $t \downarrow 0$,*

$$\phi(A, B_t) - 1 = t\,\|E\|_2 + O(t^2), \tag{48}$$

$$\phi(A^2, B_t^2) - 1 = t\,\|\mathcal{L}_A(E)\|_2 + O\left(t^2 \left\|A^{1/2}EA^{-1/2}\right\|_2^2\right), \tag{49}$$

where

$$\mathcal{L}_A(E) = A^{1/2}EA^{-1/2} + A^{-1/2}EA^{1/2}. \tag{50}$$

Thus the directional amplification is $\|\mathcal{L}_A(E)\|_2 \,/\, \|E\|_2$.

If $A = U\,\mathrm{diag}(\lambda_1, \dots, \lambda_n)U^\top$ and $\widehat{E} = U^\top EU$, then

$$\left(U^\top \mathcal{L}_A(E) U\right)_{ij} = \left(\sqrt{\frac{\lambda_i}{\lambda_j}} + \sqrt{\frac{\lambda_j}{\lambda_i}}\right)\widehat{E}_{ij}. \tag{51}$$

Entries within a narrow spectral band receive a uniformly bounded multiplier. For $\kappa_A = \kappa_2(A)$, an entry that couples $\lambda_{\max}$ to $\lambda_{\min}$ receives $\sqrt{\kappa_A} + \kappa_A^{-1/2}$. The choice $E = u_{\max}u_{\min}^\top + u_{\min}u_{\max}^\top$ attains this coefficient and matches the first-order expansion of theorem 5.1.

Proof of theorem D.1. For sufficiently small positive t, B_t is positive definite and

$$A^{-1/2}B_tA^{-1/2} = I + tE.$$

Table 2: Bounds under different structural assumptions. The first row is exact for commuting pairs; the second is the unrestricted minimax result.

Setting	Structural quantity	Output guarantee
Commuting A, B	$AB = BA$	$\phi(A^2, B^2) = \phi(A, B)^2$
Unrestricted worst case	$\kappa_2(A) \le \kappa$	Exact equation (10); excess $\Theta(\epsilon\sqrt{\kappa} + \epsilon^2\kappa)$
Elementary similarity bound	$\kappa_2(A) \le \kappa$	$\phi(A^2, B^2) \le (1 + (r-1)\sqrt{\kappa})^2$; same joint order
Relative residual	$\nu(A, B)$	$\phi(A^2, B^2) - 1 \le 2\nu + \nu^2$
Oblique eigenspaces	$\Pi(BA^{-1})$	$\sqrt{\phi(A^2, B^2)} \le r^{-1} + (r - r^{-1})\Pi$
Local spectral mixing	$\mathrm{Mix}_{A,b}(E)$	$\phi(A^2, B^2) \le (1 - \epsilon\,\mathrm{Mix})^{-2}$
Hard log-spectral locality	Width L, $\alpha > 0$	$\phi(A^2, B^2) \le \max\{\Gamma^2, \alpha^{-2}\}$; exact inverse threshold in theorem 4.1

The two sides of equation (1) have first-order coefficients $\lambda_{\max}(E)$ and $-\lambda_{\min}(E)$, respectively. Their maximum is $\|E\|_2$, proving equation (48).

Let $N = A^{1/2}EA^{-1/2}$. Then $B_tA^{-1} = I + tN$. By proposition 3.1,

$$\phi(A^2, B_t^2) = \max\{\|I + tN\|_2^2, \left\|(I + tN)^{-1}\right\|_2^2\}.$$

For the forward term,

$$(I + tN)^\top(I + tN) = I + t(N + N^\top) + t^2N^\top N.$$

The inverse term has the same first-order expansion with the sign reversed. Eigenvalue perturbation at the identity therefore gives the first-order coefficient

$$\max\{\lambda_{\max}(N + N^\top), -\lambda_{\min}(N + N^\top)\} = \left\|N + N^\top\right\|_2.$$

Since $N + N^\top = \mathcal{L}_A(E)$, equation (49) follows. Finally, the eigenbasis of A gives the entrywise formula in equation (51). □

D.2 Soft logarithmic-band conditions

Hard bandwidth is convenient for the sharp threshold, but applications may have small rather than exactly zero long-range entries. We therefore replace a strict cutoff by a weighted sum across logarithmic bands. Fix $b > 1$ and let

$$P_k = \mathbf{1}_{[b^k, b^{k+1})}(A), \qquad E^{(d)} = \sum_k P_{k+d}EP_k, \tag{52}$$

where $E = A^{-1/2}(B - A)A^{-1/2}$. For $E \neq 0$ define

$$\mathrm{Mix}_{A,b}(E) = \frac{b^{1/2}}{\|E\|_2}\left[\left\|E^{(0)}\right\|_2 + \sum_{d\geq 1}(b^{d/2} + b^{-d/2})\left\|E^{(d)}\right\|_2\right], \tag{53}$$

and set $\mathrm{Mix}_{A,b}(0) = 0$.

Theorem D.2 (Log-spectral-band stability)**.** *For symmetric E,*

$$\left\|A^{1/2}EA^{-1/2}\right\|_2 \leq \mathrm{Mix}_{A,b}(E)\,\|E\|_2. \tag{54}$$

If $\phi(A, B) \leq 1 + \epsilon$ and $K = \mathrm{Mix}_{A,b}(E)$ with $K\epsilon < 1$, then

$$\phi(A^2, B^2) \leq (1 - K\epsilon)^{-2}. \tag{55}$$

For $K\epsilon \leq 1/2$, the output excess is at most $6K\epsilon$.

If $E^{(d)} = 0$ for $|d| > w$, contractivity of block extraction gives

$$\mathrm{Mix}_{A,b}(E) \leq b^{1/2}\left[1 + \sum_{d=1}^{w}(b^{d/2} + b^{-d/2})\right]. \tag{56}$$

For fixed b, w, this bound is independent of the global condition number. More generally, the decay condition

$$\left\|E^{(d)}\right\|_2 \leq C\,\|E\|_2\; b^{-(1/2+\gamma)d} \qquad (d \geq 1) \tag{57}$$

with $C \geq 0, \gamma > 0$ gives $\mathrm{Mix}_{A,b}(E) = O_{b,C,\gamma}(1)$. Its exponent $1/2$ balances the amplification across logarithmic bands.

Proof of theorem D.2. For each integer d, set

$$T_d = \sum_k A^{1/2}P_{k+d}EP_kA^{-1/2}.$$

For fixed d, the source spaces $P_k\mathbb{R}^n$ are mutually orthogonal, as are the target spaces $P_{k+d}\mathbb{R}^n$. Hence the norm of this block shift is the maximum block norm. On the target band, $\left\|A^{1/2}P_{k+d}\right\|_2 \le b^{(k+d+1)/2}$; on the source band, $\left\|P_kA^{-1/2}\right\|_2 \le b^{-k/2}$. It follows that

$$\|T_d\|_2 \le b^{(d+1)/2}\left\|E^{(d)}\right\|_2. \tag{58}$$

The sum over the finitely many nonzero band offsets gives

$$\left\|A^{1/2}EA^{-1/2}\right\|_2 \le b^{1/2}\sum_{d\in\mathbb{Z}} b^{d/2}\left\|E^{(d)}\right\|_2.$$

If E is symmetric, then $E^{(-d)} = (E^{(d)})^\top$ and the two norms agree. We pair positive and negative d to obtain equation (54).

Now let $E = A^{-1/2}(B-A)A^{-1/2}$. The condition $\phi(A,B) \le 1+\epsilon$ implies $\|E\|_2 \le \epsilon$. Set $N = A^{1/2}EA^{-1/2}$, so $BA^{-1} = I+N$ and $\|N\|_2 \le K\epsilon$. If $K\epsilon < 1$, then

$$\begin{aligned}\left\|BA^{-1}\right\|_2 &\le 1+K\epsilon,\\ \left\|AB^{-1}\right\|_2 &= \left\|(I+N)^{-1}\right\|_2\\ &\le (1-K\epsilon)^{-1}.\end{aligned}$$

The second bound dominates the first. Proposition 3.1 yields equation (55). For $x = K\epsilon \le 1/2$, $(1-x)^{-2}-1 \le 6x$, proving the final statement. □

The block extraction used in equation (56) is contractive: $\left\|E^{(d)}\right\|_2 \le \|E\|_2$. Indeed, for any x, orthogonality of target blocks gives

$$\begin{aligned}\left\|E^{(d)}x\right\|_2^2 &= \sum_k \|P_{k+d}EP_kx\|_2^2\\ &\le \|E\|_2^2\sum_k \|P_kx\|_2^2\\ &\le \|E\|_2^2\,\|x\|_2^2.\end{aligned}$$

E THE ISOSPECTRAL ROTATION

The minimax equality pair changes the eigenvalues as well as the eigendirections, which leaves open whether spectral change is essential. It is not: an isospectral rotation shows that a nearly aligned pair can attain the same worst-case order. Let

$$\begin{aligned}A &= \begin{pmatrix}\kappa & 0\\ 0 & 1\end{pmatrix}, \qquad & B &= R_\theta A R_\theta^\top,\\ R_\theta &= \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}, \qquad & q &= \sqrt{\kappa}.\end{aligned} \tag{59}$$

Proposition E.1 (Isospectral rotation)**.** *For the pair in equation* (59)*,*

$$\log\phi(A,B) = 2\,\mathrm{arsinh}\left(\frac{q-q^{-1}}{2}|\sin\theta|\right), \tag{60}$$

$$\log\phi(A^2,B^2) = 2\,\mathrm{arsinh}\left(\frac{q^2-q^{-2}}{2}|\sin\theta|\right). \tag{61}$$

Consequently,

$$\log\phi(A^2,B^2) = 2\,\mathrm{arsinh}\left[(q+q^{-1})\sinh\left(\tfrac{1}{2}\log\phi(A,B)\right)\right]. \tag{62}$$

If $\kappa > 1$ and $\phi(A,B) = 1+\epsilon$, then

$$|\sin\theta| = \frac{\epsilon}{\sqrt{1+\epsilon}(\sqrt{\kappa}-\kappa^{-1/2})}. \tag{63}$$

Thus $|\theta| \asymp \epsilon/\sqrt{\kappa}$ for large κ, even though the output excess has order $\epsilon\sqrt{\kappa}+\epsilon^2\kappa$. The rotation family is order-optimal; equation (11) attains the exact minimax value.

Proof of proposition E.1. Let $C = A^{-1/2}BA^{-1/2}$. Since A and B are isospectral, $\det C = 1$. Direct multiplication gives

$$\operatorname{tr} C = 2 + (q - q^{-1})^2 \sin^2 \theta. \tag{64}$$

Set $u = (q - q^{-1})|\sin\theta|/2$. Then

$$\operatorname{tr} C = 2 + 4u^2 = 2\cosh(2\operatorname{arsinh} u).$$

The two eigenvalues of C are reciprocal, so the larger equals $\exp(2\operatorname{arsinh} u)$, which proves equation (60). For A^2 and B^2, the same calculation replaces q by q^2 and gives equation (61). Finally, $q^2 - q^{-2} = (q + q^{-1})(q - q^{-1})$, which gives equation (62).

If $\phi(A, B) = 1 + \epsilon$, then

$$\sinh\left(\frac{1}{2}\log(1+\epsilon)\right) = \frac{\epsilon}{2\sqrt{1+\epsilon}}.$$

Substitution into equation (60) yields equation (63). □

F Curvature Reuse: Geometry, Residuals, and Sampling Budgets

This section supplies the technical details behind the selection rule in section 6. We move from batch geometry to computable residual certificates, then derive checkpoint bounds and a sampling budget that produce the matrix inputs required by the main theory.

F.1 Batch update geometry

We begin by lifting the single-vector guarantee to an entire batch of candidate directions. Throughout this section $H = F + \lambda I$ is a fixed SPD reference in fixed parameter coordinates, $F \succeq 0$, and $\lambda > 0$. Examples include damped empirical-Fisher and generalized Gauss–Newton matrices.

For $G = [g_1, \ldots, g_m]$, define $U = H^{-1}G$, $\widetilde{U} = \widetilde{H}^{-1}G$, and $K = U^\top U = G^\top H^{-2} G$. The same inverse-square bound controls the complete update Gram matrix.

Corollary F.1 (Update geometry and decision margins)**.** *Let $H, \widetilde{H} \in \mathbb{S}^d_{++}$, $d \geq 2$, $\phi(H, \widetilde{H}) \leq r$, and $\kappa_2(H) \leq \kappa$. For any $G \in \mathbb{R}^{d\times m}$, set $\widetilde{K} = G^\top \widetilde{H}^{-2} G$ and $W = \mathcal{W}_d(r, \kappa)$. Then*

$$W^{-1}K \preceq \widetilde{K} \preceq WK. \tag{65}$$

In particular, every batch a satisfies $W^{-1}\|Ua\|_2^2 \leq \|\widetilde{H}^{-1}Ga\|_2^2 \leq W\|Ua\|_2^2$. If $\widetilde{s}_i = \|\widetilde{H}^{-1}g_i\|_2^2$, the true squared length lies in $[\widetilde{s}_i/W, W\widetilde{s}_i]$; hence $\widetilde{s}_i > W^2\widetilde{s}_j$ implies $\|u_i\|_2 > \|u_j\|_2$.

Proof of corollary F.1. Inversion is an isometry for ϕ, and $\phi(H^{-2}, \widetilde{H}^{-2}) = \phi(H^2, \widetilde{H}^2)$. The minimax theorem therefore gives $W^{-1}H^{-2} \preceq \widetilde{H}^{-2} \preceq WH^{-2}$. Congruence by the possibly rectangular G proves equation (65). For every a, the resulting quadratic form equals the squared length of the corresponding linearized batch update. The choice $a = e_i - e_j$ gives all pairwise squared distances. Since H and $\widetilde{H}$ are invertible, both Gram kernels equal $\ker G$, including zero-update batches. The length intervals follow by rearrangement. Finally, $\|u_i\|_2^2 \geq \widetilde{s}_i/W > W\widetilde{s}_j \geq \|u_j\|_2^2$ proves the ordering. □

For a threshold $\tau > 0$, $W\widetilde{s}_i \leq \tau^2$ implies that the reference update stays within radius τ, and $\widetilde{s}_i/W > \tau^2$ implies that it exceeds the radius. Overlapping intervals remain unresolved. The constant W is sharp over unrestricted $H, \widetilde{H}, G$: take a minimax equality pair and $G = I$. A fixed task subspace can have substantially smaller distortion.

F.2 Exact relative vector error

The Gram bound controls all update lengths simultaneously, but it does not directly quantify how far an approximate update vector moves. The next result gives the exact worst-case relative vector error under the same matrix assumptions.

Theorem F.2 (Exact relative update error)**.** *Let $H, \widetilde{H} \in \mathbb{S}^d_{++}$, $d \geq 2$, $\phi(H, \widetilde{H}) \leq r$, and $\kappa_2(H) \leq \kappa$. Put $a = r - 1$, $b = 1 - r^{-1}$, and $\chi = (\sqrt{\kappa} + \kappa^{-1/2})/2$. Then*

$$D_*(r, \kappa) = \frac{(a+b)\chi + \sqrt{(a-b)^2 + (a+b)^2(\chi^2 - 1)}}{2} \tag{66}$$

is the exact supremum of $\|\widetilde{H}^{-1}H - I\|_2$ over these constraints. For $U = H^{-1}G$ and $\widetilde{U} = \widetilde{H}^{-1}G$, every batch satisfies $\|\widetilde{U}c - Uc\|_2 \leq D_*(r, \kappa)\|Uc\|_2$. The pair in equation (11) *attains the supremum.*

Proof of theorem F.2. The case $r = 1$ has $\widetilde{H} = H$ and $D_* = 0$. For $r > 1$, let $S = H^{1/2}$ and $C = H^{-1/2}\widetilde{H}H^{-1/2}$. Transposition preserves the norm, so

$$\|\widetilde{H}^{-1}H - I\|_2 = \|H\widetilde{H}^{-1} - I\|_2 = \|S(C^{-1} - I)S^{-1}\|_2.$$

The symmetric matrix $D = C^{-1} - I$ ranges over $[-bI, aI]$. For fixed S, convexity reduces maximization to $D = -bI + (a+b)P$ with P an orthogonal projection, exactly as in lemma A.2. For a nontrivial projection $Q = SPS^{-1}$, let $k = \|Q\|_2$. The orthogonal block decomposition used there reduces the largest norm to that of the classical quadratic-operator block (Gau et al., 2008, Proposition 1(c)),

$$T_k = \begin{pmatrix} a & (a+b)\sqrt{k^2 - 1} \\ 0 & -b \end{pmatrix}.$$

Its singular values obey $\sigma_1\sigma_2 = ab$ and $\sigma_1^2 + \sigma_2^2 = (a+b)^2k^2 - 2ab$, hence

$$\|T_k\|_2 = \frac{(a+b)k + \sqrt{(a+b)^2k^2 - 4ab}}{2}.$$

This expression is increasing for $k \geq 1$. By lemma A.1, $k \leq \chi$, with equality for a 45° projection between extreme eigenscales. Trivial projections have norm at most a, which is the value of the displayed formula at $k = 1$. Set $k = \chi$ and write the radicand as $(a-b)^2 + (a+b)^2(\chi^2 - 1)$ to obtain the upper bound.

For the pair in equation (11), the inverse relative matrix is $C^{-1} = r^{-1}I + (r - r^{-1})(I - P)$. The complementary 45° projection also has obliqueness χ, so equality holds. Direct sums with identical scalar blocks establish attainability in every $d > 2$ without changing the condition number. Finally $\widetilde{U} - U = (\widetilde{H}^{-1}H - I)U$ proves the batch bound. An extremal right singular vector of $\widetilde{H}^{-1}H - I$ gives equality for a single update by setting its gradient to H times that vector. □

The same signed-interval argument applies to $C - I$. Therefore $\sup \nu(H, \widetilde{H}) = D_*(r, \kappa)$ and the stronger two-sided vector statement holds:

$$\|\widetilde{U}c - Uc\|_2 \leq D_*(r, \kappa) \min\{\|Uc\|_2, \|\widetilde{U}c\|_2\}. \tag{67}$$

Here $D_*(r, 1) = r - 1$. Moreover, the signed-interval inclusions $[-bI, bI] \subseteq [-bI, aI] \subseteq [-aI, aI]$ give

$$(r-1)\sqrt{\kappa}/r \leq D_*(r, \kappa) \leq (r-1)\sqrt{\kappa}.$$

The two symmetric intervals have exact similarity norm $b\sqrt{\kappa}$ and $a\sqrt{\kappa}$, attained by an off-diagonal involution between the extreme eigenscales. Consequently $D_*(1 + \epsilon, \kappa) = \epsilon\sqrt{\kappa}[1 + O(\epsilon)]$, with a relative remainder uniform in κ. The vector law bounds update error; the square law gives the optimal simultaneous multiplicative bound for all squared update lengths.

F.3 Residual bounds on the task subspace

The preceding laws use only matrix-level drift and are therefore uniform over every right-hand side. When the relevant gradients span a small known subspace, measured solve residuals can give a sharper certificate.

Proposition F.3 (Task-subspace residual bound)**.** *Let $H \succeq \mu I$ with $\mu > 0$, let $G_0 \in \mathbb{R}^{d \times k}$ have full column rank, and let $V \in \mathbb{R}^{d \times k}$ be computed basis updates with full column rank. Write the thin QR factorization $V = QT$, where $Q^\top Q = I$ and T is invertible, and define*

$$R = G_0 - HV, \qquad \delta_{\mathcal{G}} = \frac{\left\|RT^{-1}\right\|_2}{\mu}. \tag{68}$$

If $\delta_{\mathcal{G}} < 1$, then

$$(1-\delta_{\mathcal{G}})^2 V^\top V \preceq G_0^\top H^{-2} G_0 \preceq (1+\delta_{\mathcal{G}})^2 V^\top V. \tag{69}$$

For $G = G_0 C$, the same inequalities hold after congruence by C, with computed updates defined by VC.

Proof. Put $U_0 = H^{-1}G_0$. The residual identity gives $U_0 - V = H^{-1}R$. For every $b \in \mathbb{R}^k$,

$$\|(U_0 - V)b\|_2 \le \mu^{-1}\|RT^{-1}\|_2 \|Tb\|_2 = \delta_{\mathcal{G}}\|Vb\|_2.$$

The triangle and reverse-triangle inequalities bound $\|U_0 b\|_2$ between $(1-\delta_{\mathcal{G}})\|Vb\|_2$ and $(1+\delta_{\mathcal{G}})\|Vb\|_2$. Their squares give the Loewner sandwich, and the substitution $b = Ca$ gives the batch claim. □

For a damped PSD curvature, $\mu = \lambda$ is available without estimating the smallest eigenvalue. The check requires k applications of H, a thin QR factorization, and the norm of a $d \times k$ matrix, with dense postprocessing cost $O(dk^2 + k^3)$. The Frobenius norm supplies a computable upper bound for the spectral norm.

The result also applies when V approximately solves $\widetilde{H}V = G_0$: the measured residual then includes both surrogate and solve errors. The batch guarantee uses the same linear map C for the reference and computed updates. For rank-deficient G, we choose a basis and use this common map to represent both sets of updates. For $\delta_{\mathcal{G}} < 1$, the interval for a reference squared length is $[\ell\widehat{s}, u\widehat{s}]$, where $\ell = (1-\delta_{\mathcal{G}})^2$ and $u = (1+\delta_{\mathcal{G}})^2$. Consequently $u\widehat{s} \le \tau^2$ implies a below-threshold update, $\ell\widehat{s} > \tau^2$ implies an above-threshold update, and $\ell\widehat{s}_i > u\widehat{s}_j$ determines the ordering. We can intersect the residual and square-law intervals when both apply to the computed updates. Numerical tolerances account for factorization and norm-evaluation error.

F.4 Curvature bounds between checkpoints

We next connect these abstract certificates to quantities available during training. For a frozen feature map, write $x_i \in \mathbb{R}^d$ for its training features, including an intercept coordinate if used. A binary logistic head has $w_i(\theta) = \sigma(x_i^\top\theta)(1-\sigma(x_i^\top\theta)) > 0$. Positive weights on fixed rank-one factors give the following Loewner bounds, which are related to logistic Hessian variation (Bach, 2010).

Proposition F.4 (Curvature bounds from weight ratios)**.** *Let $F_j = n^{-1}\sum_i w_{j,i} x_i x_i^\top$ and $H_j = F_j + \lambda I$ for $j \in \{0, t\}$, where $w_{j,i} > 0$ and the damping $\lambda > 0$ is unchanged. Define*

$$\ell = \min\{1, \min_i w_{t,i}/w_{0,i}\}, \qquad u = \max\{1, \max_i w_{t,i}/w_{0,i}\}, \qquad r_{\mathrm{wt}} = \max\{u, \ell^{-1}\}, \tag{70}$$

$$u_F = \max_i w_{t,i}/w_{0,i}, \qquad L_0 = \|F_0\|_\infty, \qquad T_t = \frac{1}{n}\sum_i w_{t,i}\|x_i\|_2^2,$$

$$\bar{\kappa} = \min\left\{1 + \frac{L_0}{\lambda}, 1 + \frac{\min\{u_F L_0, T_t\}}{\lambda}\right\}. \tag{71}$$

Here $\|\cdot\|_\infty$ is the maximum absolute row sum. Then $\ell H_0 \preceq H_t \preceq u H_0$. For $W = \mathcal{W}_d(r_{\mathrm{wt}}, \bar{\kappa})$, every gradient g satisfies

$$\frac{\|H_0^{-1}g\|_2}{\sqrt{W}} \le \|H_t^{-1}g\|_2 \le \sqrt{W}\,\|H_0^{-1}g\|_2. \tag{72}$$

Proof. A comparison of the positive-semidefinite (PSD) rank-one terms gives $\ell F_0 \preceq F_t \preceq u F_0$. Since $\ell \le 1 \le u$, the common damping preserves the claimed sandwich. Both H_j have smallest eigenvalue at least λ. Symmetry implies $\|F_0\|_2 \le \|F_0\|_\infty = L_0$. Positive semidefiniteness and the sharper undamped comparison $F_t \preceq u_F F_0$ give $\|F_t\|_2 \le \min\{u_F L_0, \mathrm{tr}\, F_t\} = \min\{u_F L_0, T_t\}$. The two terms in equation (71) therefore bound the condition numbers of H_0 and H_t, respectively. Because ϕ is symmetric, we can apply the square-distortion bound with either matrix as reference. Monotonicity in the conditioning cap allows us to use the smaller bound. Inversion and evaluation on g prove equation (72). □

The logits determine the curvature weights. With cached L_0 and feature row norms, equations (70) and (71) cost $O(n)$ to evaluate. The logit calculation costs $O(nd)$. Cache setup includes forming and factoring H_0 and solving for the candidate directions.

Retaining the full input interval. The preceding symmetric factor is convenient but loses useful one-sided information. In particular, r_{wt} discards the location of the interval $[\ell, u]$. If we retain this information, the square law gives the optimal lower and upper coefficients separately.

Corollary F.5 (Asymmetric update-geometry bound)**.** *Let $0 < \ell \le u$, $\ell H_0 \preceq H_t \preceq uH_0$, and suppose the condition number of at least one of H_0, H_t is at most κ. Set $a = \sqrt{\ell u}$, $\rho = \sqrt{u/\ell}$, and $W_c = \mathcal{W}_d(\rho, \kappa)$. For $U_j = H_j^{-1}G$,*

$$\frac{1}{a^2 W_c} U_0^\top U_0 \preceq U_t^\top U_t \preceq \frac{W_c}{a^2} U_0^\top U_0. \tag{73}$$

Each coefficient is optimal over the stated class for $d \ge 2$.

Proof. The rescaled matrix H_t/a satisfies $\rho^{-1}H_0 \preceq H_t/a \preceq \rho H_0$. This rescaling preserves the condition number, so the square law applies with either matrix as reference. Inversion and congruence give the result. The equality pair in equation (11), with $r = \rho$ and its second matrix multiplied by a, attains both operator coefficients; corresponding singular vectors attain the individual length endpoints. □

Equivalently, the norm interval has endpoints $\|H_0^{-1}g\|_2/f_{\ell,u}(\chi)$ and $f_{\ell,u}(\chi)\|H_0^{-1}g\|_2/(\ell u)$, where $\chi = (\sqrt{\kappa} + \kappa^{-1/2})/2$ and $f_{\ell,u}$ is defined in equation (26). Monotonicity under inclusion of operator intervals shows that this interval is contained in the symmetric-r_{wt} interval. If $H_t = aH_0$, it recovers the exact rescaling even when $a \neq 1$. The endpoints are optimal for the specified Loewner interval and conditioning bound.

The shared damping tightens the interval at little additional cost. With L_0 from equation (71), define

$$\ell_\lambda = \frac{\lambda + \ell L_0}{\lambda + L_0}, \qquad u_\lambda = \frac{\lambda + uL_0}{\lambda + L_0}. \tag{74}$$

Then $\ell_\lambda H_0 \preceq H_t \preceq u_\lambda H_0$. Indeed, $H_0 \preceq (\lambda + L_0)I$ and $\ell H_0 + (1-\ell)\lambda I \preceq H_t \preceq uH_0 - (u-1)\lambda I$ imply this sandwich. Both refinements use scalar quantities. The corresponding centered similarity bound replaces W_c by $(1 + (\rho - 1)\sqrt{\kappa})^2$ in equation (73). Table 16 evaluates the symmetric intervals from proposition F.4.

Screening and iterative refinement. Once the checkpoint interval is available, it can be used as the first stage of an exact decision procedure. The screening step determines whether $\|H_t^{-1}g\|_2$ exceeds a fixed radius τ. First compare the interval in equation (72) with τ. For a computed cached update v_0, an error radius $e_0 = \|g - H_0 v_0\|_2/m_0$ changes this interval to

$$\left[\frac{\max\{0, \|v_0\|_2 - e_0\}}{\sqrt{W}}, \sqrt{W}(\|v_0\|_2 + e_0)\right].$$

If the interval straddles τ, compute a current residual. For any iterate v, the standard residual identity gives the interval

$$\left[\max\{0, \|v\|_2 - e_t(v)\},\ \|v\|_2 + e_t(v)\right], \qquad e_t(v) = \frac{\|g - H_t v\|_2}{\mu_t}. \tag{75}$$

For unresolved candidates, we use H_0-preconditioned conjugate gradients and stop when a residual interval separates the threshold. An interval wholly below τ rejects the candidate; one wholly above retains it. At the iteration limit, a direct solve determines the remaining decisions. Here m_0 and μ_t are the eigenvalue lower bounds in equation (20). The PCG baseline uses the same warm start, preconditioner, and stopping rule. The similarity baseline uses the elementary constant in place of W.

Threshold separation. We can now state precisely when the exact constant settles a candidate that the similarity bound leaves unresolved. For a nonzero gradient and an exact cached solve, define the log-margin $m_g = |\log(\|H_0^{-1}g\|_2/\tau)|$, where $\tau > 0$ is the radius. The norm interval with multiplicative factors e^{-b}, e^b, $b \ge 0$, strictly separates the threshold exactly when $m_g > b$. The exact law therefore screens the following additional candidates relative to the similarity bound:

$$\tfrac{1}{2}\log \mathcal{W}_d(r, \bar{\kappa}) < m_g \le \log\big(1 + (r-1)\sqrt{\bar{\kappa}}\big). \tag{76}$$

Each additional screened candidate eliminates its initial target residual evaluation and subsequent PCG iterations.

F.5 SIGNED INFLUENCE SCORES REQUIRE ADDITIVE MARGINS

Update lengths are nonnegative quadratic quantities, whereas signed influence scores can be small because of cancellation. Their reuse therefore requires an additive rather than multiplicative margin. Define $t_{qi} = g_q^\top H^{-1} g_i$ and $\widetilde{t}_{qi} = g_q^\top \widetilde{H}^{-1} g_i$. The sign convention of an influence score changes neither the following error bound nor its margin requirement.

Proposition F.6 (Cross-score error)**.** *If $\phi(H, \widetilde{H}) \le r$, then*

$$|\widetilde{t}_{qi} - t_{qi}| \le (r-1)\sqrt{g_q^\top H^{-1} g_q}\,\sqrt{g_i^\top H^{-1} g_i}. \tag{77}$$

For the difference of two scores, replace g_i by $g_i - g_j$ on the right-hand side. The same bound holds with both energies evaluated using $\widetilde{H}^{-1}$ instead of H^{-1}.

Proof. Inversion gives $r^{-1} I \preceq H^{1/2} \widetilde{H}^{-1} H^{1/2} \preceq rI$. The difference from I has spectral norm at most $r-1$. Apply the bilinear operator-norm inequality to $H^{-1/2} g_q$ and $H^{-1/2} g_i$. Linearity gives the score-difference version. The same argument with H and $\widetilde{H}$ exchanged gives the surrogate-energy version. □

An estimated score difference whose magnitude exceeds this error radius has the correct sign. Inversion alone gives no relative guarantee for a cross-score near zero. For example, take $H = I_2$, $\widetilde{H}^{-1} = \left(\begin{smallmatrix} 1 & -t \\ -t & 1 \end{smallmatrix}\right)$, $0 < t < 1$, $g_q = e_1$, and $g_i = (t/2, 1)^\top$. The reference score is $t/2$ and the approximate score is $-t/2$, although $\phi(H, \widetilde{H}) = (1-t)^{-1} \to 1$. This distinguishes signed attribution scores from the update-length and distance geometry controlled by inverse squaring.

F.6 AN EXPLICIT RIDGE-LEVERAGE SAMPLING BUDGET

The previous guarantees are deterministic once a Loewner approximation is available. To connect that requirement to a concrete randomized construction, the following application uses a standard matrix Bernstein inequality (Tropp, 2012, Theorem 6.1). It supplies a sufficient budget for a specified sampling construction, complementing the deterministic minimax input threshold.

Proposition F.7 (Curvature sampling for a target geometry)**.** *Let $a_i \in \mathbb{R}^d$ with $d \ge 2$, $F = \sum_{i=1}^N a_i a_i^\top$, $H = F + \lambda I$ with $\lambda > 0$, and $\kappa_2(H) \le \kappa$. Set*

$$\tau_i = a_i^\top H^{-1} a_i, \qquad D = d_{\mathrm{eff}} = \sum_i \tau_i = \operatorname{tr}(F H^{-1}).$$

If $D > 0$, sample $i_1, \dots, i_s$ independently with replacement using $p_i = \tau_i / D$ on the nonzero terms, and form

$$\widetilde{H} = \lambda I + \frac{1}{s} \sum_{j=1}^{s} \frac{a_{i_j} a_{i_j}^\top}{p_{i_j}}. \tag{78}$$

For $0 < \delta < 1$ and $0 < \eta < 1$, a sufficient sample count for $(1-\delta)H \preceq \widetilde{H} \preceq (1+\delta)H$ with probability at least $1 - \eta$ is

$$s \ge \left\lceil \frac{2D(1+\delta/3)}{\delta^2} \log \frac{2d}{\eta} \right\rceil. \tag{79}$$

In particular, for a target $\beta > 0$, define

$$c_{\beta,\kappa} = \frac{\beta}{\sqrt{1+\beta}(\sqrt{\kappa} + \kappa^{-1/2})}, \qquad r_* = \sqrt{1 + c_{\beta,\kappa}^2} + c_{\beta,\kappa}, \qquad \delta_* = 1 - r_*^{-1}. \tag{80}$$

For $\delta = \delta_*$, the count in equation (79) guarantees

$$(1+\beta)^{-1} G^\top H^{-2} G \preceq G^\top \widetilde{H}^{-2} G \preceq (1+\beta) G^\top H^{-2} G \tag{81}$$

simultaneously for every matrix G, on the same event. If $F = 0$, use $\widetilde{H} = H$ without sampling.

Proof. Put $b_i = H^{-1/2}a_i$, $B = \sum_i b_i b_i^\top \preceq I$, and $Y_j = b_{i_j} b_{i_j}^\top / p_{i_j}$. Then $\mathbb{E}Y_j = B$ and $\|Y_j\|_2 = D$. Since $0 \preceq B \preceq DI$ as well, the centered variable $X_j = Y_j - B$ has $\mathbb{E}X_j = 0$ and $\|X_j\|_2 \le D$. The rank-one identity $Y_j^2 = DY_j$ gives

$$\mathbb{E}X_j^2 = DB - B^2 \preceq DI.$$

Apply the bounded matrix Bernstein inequality to both X_j and $-X_j$. The union bound yields

$$\Pr\left\{\left\|H^{-1/2}(\widetilde{H} - H)H^{-1/2}\right\|_2 \ge \delta\right\} \le 2d\exp\left[-\frac{s\delta^2}{2D(1+\delta/3)}\right]. \tag{82}$$

This proves equation (79). On the resulting event, $r = (1-\delta)^{-1}$ is a valid symmetric Loewner factor because $1 + \delta \le (1-\delta)^{-1}$. For $\delta = \delta_*$, corollary A.3 gives $\mathcal{W}_d(r_*, \kappa) = 1 + \beta$; corollary F.1 finishes the proof. □

For numerical evaluation, we can use $\delta_* = 2c_{\beta,\kappa}/(1 + c_{\beta,\kappa} + \sqrt{1 + c_{\beta,\kappa}^2})$. As $\beta \downarrow 0$, the unrounded budget is asymptotic to $2D(\sqrt{\kappa} + \kappa^{-1/2})^2\beta^{-2}\log(2d/\eta)$. For $0 < \beta \le 1$, the integer count is $O(1 + D\kappa\beta^{-2}\log(d/\eta))$. The elementary similarity-aware bound has the same order, with a high-accuracy budget ratio tending to $[2\kappa/(\kappa+1)]^2$ relative to the exact law. Thus the exact expression improves the allocation constant, while the coarser κ-based bound loses an additional condition-number factor.

Exact leverage scores require H^{-1}. If upper bounds $\widehat{\tau}_i \ge \tau_i$ are available, the same proof uses $p_i = \widehat{\tau}_i/T$ and replaces D by $T = \sum_i \widehat{\tau}_i$: then $\|Y_j\|_2 \le T$ and $\mathbb{E}X_j^2 \preceq TB$. For example, $\widehat{\tau}_i = \|a_i\|_2^2/\lambda$ is available without an inverse. When the sample count exceeds N, the full sum provides the exact curvature.

G Proofs for Transformation Consequences

The square map is the source of the main instability, but several related transformations either preserve the Loewner factor or reduce to the same minimax law. We record these consequences before specializing them to regularized graph operators.

For a transformation F, let $\mathcal{W}_{F,n}(r, \kappa)$ be the supremum of $\phi(F(A), F(B))$ over $\phi(A, B) \le r$ and $\kappa_2(A) \le \kappa$. The square case is $\mathcal{W}_n$ from equation (3).

G.1 Loewner-preserving transformations

We first separate transformations that are stable directly at the input level from those that require the square-distortion theorem. Congruence is an isometry for ϕ: for every invertible S,

$$\phi(SAS^\top, SBS^\top) = \phi(A, B). \tag{83}$$

Inversion is also an isometry,

$$\phi(A^{-1}, B^{-1}) = \phi(A, B), \tag{84}$$

because it reverses both Loewner inequalities.

The stable power interval is larger than these isometries.

Proposition G.1 (Loewner–Heinz stable power interval)**.** *If $A, B \in \mathbb{S}_{++}^n$ and $\phi(A, B) \le r$, then*

$$\phi(A^p, B^p) \le r^{|p|} \qquad \textit{for every } -1 \le p \le 1. \tag{85}$$

Positive regularization is nonexpansive. If $\phi(A, B) \le r$ and $\lambda \ge 0$, then

$$\phi(I + \lambda A, I + \lambda B) \le r. \tag{86}$$

For positive-semidefinite inputs, the assumption $r^{-1}A \preceq B \preceq rA$ gives the same bound after regularization.

Proof of proposition G.1. For $0 \le p \le 1$, apply the Loewner–Heinz theorem (Bhatia, 1997) to $r^{-1}A \preceq B \preceq rA$ and use homogeneity. For $-1 \le p < 0$, first use inversion invariance and then apply the positive-power case to A^{-1} and B^{-1}. □

Proof of equation (86). The inequality $B \preceq rA$ implies $I+\lambda B \preceq I+r\lambda A \preceq r(I+\lambda A)$; the lower inequality follows in the same way. The same sandwich argument applies to positive-semidefinite inputs satisfying $r^{-1}A \preceq B \preceq rA$.

Corollary G.2 (Regularized inverse-square transformation)**.** *Let $n \geq 2$, $A, B \in \mathbb{S}^n_{++}$, $\phi(A, B) \leq r$, and $\kappa_2(I + \lambda A) \leq \kappa_\lambda$ for $\lambda \geq 0$. Then*

$$\phi\big((I + \lambda A)^{-2}, (I + \lambda B)^{-2}\big) \leq \mathcal{W}_n(r, \kappa_\lambda). \tag{87}$$

For $r = 1 + \epsilon \leq 3/2$, the corresponding worst-case excess is $O(\epsilon\sqrt{\kappa_\lambda} + \epsilon^2\kappa_\lambda)$.

Proof of corollary G.2. Set $C = I + \lambda A$ and $D = I + \lambda B$. Equation (86) gives $\phi(C, D) \leq r$. Apply theorem 5.1 to C, D and use inversion invariance: $\phi(C^{-2}, D^{-2}) = \phi(C^2, D^2)$. □

G.2 Graph operators

These transformation rules apply directly to regularized graph filters. Inversion preserves ϕ, so theorem 5.1 controls inverse squares with the same constant. Positive shifts preserve an input Loewner sandwich and improve conditioning; therefore corollary G.2 applies the square law to $(I + \lambda A)^{-2}$ using $\kappa_2(I + \lambda A)$. These guarantees follow directly from the square law. For graph diffusion, a shared invariant subspace sharpens the cap further.

Corollary G.3 (Removing the graph constant mode)**.** *Let $L, \widetilde{L}$ be undirected weighted graph Laplacians on the same $n \geq 3$ vertices, with L connected, and let $\lambda \geq 0$. Set $Q = I + \lambda L$, $\widetilde{Q} = I + \lambda\widetilde{L}$, $r = \phi(Q, \widetilde{Q})$, and*

$$\kappa_\perp = \frac{1 + \lambda\lambda_{\max}(L)}{1 + \lambda\lambda_2(L)}.$$

Then

$$\phi(Q^{-2}, \widetilde{Q}^{-2}) \leq \mathcal{W}_{n-1}(r, \kappa_\perp). \tag{88}$$

Proof of corollary G.3. Both shifted Laplacians are the identity on $\mathrm{span}\{\mathbf{1}\}$ and preserve $\mathbf{1}^\perp$. Their common one-dimensional block contributes distortion one, so r is the distortion of the restrictions to $\mathbf{1}^\perp$. The condition number of Q on this subspace is exactly $\kappa_\perp$. Apply theorem 5.1 in dimension $n - 1$ and then use inversion invariance. □

H Temporal Accumulation for Graph Time Series

The preceding graph result concerns one application of a diffusion operator. In a time series, the same approximation is applied repeatedly, so we next track how its error accumulates through a contractive recurrence.

Proposition H.1 (Temporal accumulation)**.** *For a fixed vertex set, let $P_t = (I + \lambda L_t)^{-2}$ and $\widetilde{P}_t = (I + \lambda\widetilde{L}_t)^{-2}$, and write $R_t = \phi(P_t, \widetilde{P}_t)$. Consider*

$$\begin{aligned} H_t &= \sigma(P_t H_{t-1} W + X_t U), \\ \widetilde{H}_t &= \sigma(\widetilde{P}_t \widetilde{H}_{t-1} W + X_t U), \end{aligned} \tag{89}$$

where σ is 1-Lipschitz in Frobenius norm and $\|W\|_2 \leq \alpha$. With $\Delta_t = \left\|H_t - \widetilde{H}_t\right\|_{\mathrm{F}}$,

$$\Delta_t \leq \alpha\Delta_{t-1} + \alpha(R_t - 1)\left\|\widetilde{H}_{t-1}\right\|_{\mathrm{F}}, \tag{90}$$

$$\Delta_t \leq \alpha^t\Delta_0 + \alpha\sum_{s=1}^{t}\alpha^{t-s}(R_s - 1)\left\|\widetilde{H}_{s-1}\right\|_{\mathrm{F}}. \tag{91}$$

Proof. Every P_t is positive definite with $\|P_t\|_2 = 1$. The Loewner sandwich defining R_t, together with $P_t \preceq I$, gives $\left\|P_t - \widetilde{P}_t\right\|_2 \leq R_t - 1$. Apply the Lipschitz property, add and subtract $P_t\widetilde{H}_{t-1}W$, and use submultiplicativity to obtain equation (90). Iteration gives equation (91). □

I Coordinate Geometry and Rank-One Perturbations

Two short observations delimit the scope of the preceding guarantees. The first concerns the coordinate system in which update lengths are measured; the second shows that low perturbation rank does not replace spectral locality.

Coordinate geometry. We first distinguish matrix comparison from the metric used to measure an update. The input factor ϕ is invariant under invertible congruence. Euclidean update lengths are measured in a fixed parameter coordinate system; a change of metric replaces the identity in the corresponding quadratic form. Orthogonal reparameterizations preserve these lengths.

Rank-one cross-scale mixing. Low rank by itself also does not prevent amplification across distant scales. With $A = \mathrm{diag}(\kappa, 1)$ and the projection P from equation (11), take $C = I + \epsilon P$ and $B = A^{1/2} C A^{1/2}$. Then $B - A$ has rank one, $\phi(A, B) = 1 + \epsilon$, and

$$\|BA^{-1}\|_2 \geq \epsilon(\sqrt{\kappa} + \kappa^{-1/2})/2 - 1.$$

Thus a rank-one perturbation can produce unbounded amplification as $\kappa \to \infty$ at fixed $\epsilon > 0$. The support condition in equation (7) directly restricts this coupling in the reference eigenbasis.

J Comparison with the Closest Matrix Results

This section makes the comparison sketched in section 2 explicit. We first compare constants and admissible tail ranges, then separate the classical ingredients used by the global and local proofs.

Proposition J.1 (Chebyshev comparison on the full local class)**.** *Under the assumptions of theorem 4.1, let $r > 1$, $t = L/2$, and $q = (r-1)/(r+1)$. If $qe^t < 1$, then*

$$\|B^{-1}A\|_2 \leq G_{\mathrm{Ch}}(r, L) = \frac{1 + qe^t}{1 - qe^t}.$$

This bound and the optimal norm have ratio $G_{\mathrm{Ch}}/\mathcal{I} = 1 + h \sinh t \in [1, 2)$. The supremum of this ratio is two, approached as $qe^t \uparrow 1$ at fixed r.

Proof. Set $X = \log A$ and $D = (cI - C)/h$. Then D is symmetric, $\|D\|_2 \leq 1$, and its log-spectral width is L. For the degree-k Chebyshev polynomial T_k, functional calculus gives $\|T_k(D)\|_2 \leq 1$. Every product of k supported factors couples log-eigenvalues at distance at most kL, so $T_k(D)$ has width at most kL. The band-limited estimate used in equations (35) and (36) gives

$$\|e^{X/2} T_k(D) e^{-X/2}\|_2 \leq e^{kt}.$$

The Chebyshev Poisson identity and $c = (1 + q^2)/(1 - q^2)$, $h = 2q/(1 - q^2)$ yield

$$C^{-1} = I + 2 \sum_{k=1}^{\infty} q^k T_k(D).$$

The weighted series converges absolutely when $qe^t < 1$. The sum of these norm bounds gives G_{Ch}. By equation (38),

$$G_{\mathrm{Ch}}(c - h\cosh t) = \frac{(1 + qe^t)(1 - qe^{-t})}{1 - q^2} = 1 + h \sinh t.$$

The condition $t < \log(1/q)$ is equivalent to $h \sinh t < 1$, which also proves the limiting ratio. □

Proposition J.2 (Chebyshev screening with asymmetric inputs and tails)**.** *Use the decomposition and bounds in corollary 5.3, and put $a = h + s$. If $a < c$, define*

$$\delta = \sqrt{c^2 - a^2}, \qquad q_a = \frac{a}{c + \delta}, \qquad G_0 = \frac{1 + q_a e^{L/2}}{\delta(1 - q_a e^{L/2})}.$$

When $q_a e^{L/2} < 1$ and $\eta G_0 < 1$, an upper multiplier for the update-length ratio is

$$b_{\mathrm{Ch}} = \frac{G_0}{1 - \eta G_0}.$$

In the remaining cases, set $b_{\mathrm{Ch}} = +\infty$. The Chebyshev screen replaces b_L by b_{Ch} in equation (16)*, retaining its lower endpoint.*

Proof. The local matrix $C_L = cI + D_L$ has spectrum in $[c-a, c+a]$. The Chebyshev identity in proposition J.1, with the scale factor δ, gives $\|A^{1/2}C_L^{-1}A^{-1/2}\|_2 \le G_0$. Write $N_L = A^{1/2}C_LA^{-1/2}$ and $E = A^{1/2}RA^{-1/2}$. The resolvent identity $(N_L+E)^{-1} = (I+N_L^{-1}E)^{-1}N_L^{-1}$ and $\|N_L^{-1}E\|_2 \le \eta G_0 < 1$ give b_{Ch}. If $a = 0$, the same formula follows from $C_L = cI$ and $q_a = 0$. □

Effect of weighted tails. The constant comparison extends to the nonzero tails used in the screening rule, where the two arguments tolerate different perturbation sizes. The local bound is finite for $\eta < c - a\cosh(L/2)$, whereas the Chebyshev perturbation bound requires $\eta < G_0^{-1}$. The larger admissible tail range can retain a dimension-independent upper bound when $b_{\mathrm{Ch}} = +\infty$. For example, let $n \ge 4$, $A_n = \mathrm{diag}(1, 4, \dots, 4^{n-1})$, and let T_n be the lower shift. Set

$$C_n = \frac{29}{20}I - \frac{13}{25}(T_n + T_n^\top) + \frac{1}{100}(e_1e_4^\top + e_4e_1^\top), \qquad B_n = A_n^{1/2}C_nA_n^{1/2}.$$

The off-diagonal tridiagonal term has norm at most $26/25$ and the last term has norm $1/100$, so $\frac{2}{5}I \preceq C_n \preceq \frac{5}{2}I$. At $L = \log 4$, the last term is the tail, with $s = 1/100$ and $\eta = 2/25$. Thus $c = 29/20$, $h = 21/20$, and $a = h + s = 53/50$, giving

$$b_L = \left(\frac{29}{20} - \frac{53}{50}\frac{5}{4} - \frac{2}{25}\right)^{-1} = \frac{200}{9}, \qquad G_0 = 8 + \frac{636}{\sqrt{9789}} > \frac{25}{2}.$$

Consequently $\eta G_0 > 1$ and $b_{\mathrm{Ch}} = +\infty$ for these input bounds, while b_L is uniform in n and $\kappa_2(A_n) = 4^{n-1}$.

Squaring and relative residuals. We next compare the unrestricted law with results stated in terms of relative residuals. Braess & Peisker (1986) and Wathen (2022) identify the matrix-squaring obstruction in preconditioning. Gratton et al. (2018, Theorem 2.1) consider the spectrum of

$$(\widetilde{\mathsf{A}}^{-1}\mathsf{W}\widetilde{\mathsf{A}}^{-T})(\mathsf{A}^\top\mathsf{W}^{-1}\mathsf{A}),$$

where W is SPD and $\mathsf{A}, \widetilde{\mathsf{A}}$ are nonsingular. Their enclosure is centered at one with radius $(1 + \kappa_2(\mathsf{W}))e + \kappa_2(\mathsf{W})e^2$, where $e = \|\mathsf{A}\widetilde{\mathsf{A}}^{-1} - I\|_2$.

In our notation, choose $(\mathsf{A}, \widetilde{\mathsf{A}}, \mathsf{W}) = (C, I, A^{-1})$. The resulting matrix $A^{-1}CAC$ is similar to $A^{-1/2}CACA^{-1/2} = M^\top M$. The same substitution with C^{-1} controls the inverse relative operator. Since both residuals are at most $r-1$, their theorem gives the bound

$$\phi(A^2, B^2) \le 1 + (1+\kappa)(r-1) + \kappa(r-1)^2. \tag{92}$$

The upper enclosure for both relative operators gives the two-sided factor. The elementary similarity bound in equation (12) is no larger, since $2\sqrt{\kappa} \le 1 + \kappa$. The exact law improves its constant and supplies a matching pair and an explicit input-accuracy budget, while retaining the same joint asymptotic order.

Classical ingredients of the exact constant. The exact formula sharpens these residual bounds by combining two classical geometric facts. Kantorovich–Wielandt geometry controls the angle distortion induced by an SPD similarity, while the quadratic-operator canonical form reduces an affine projection to two-dimensional triangular blocks (Marshall & Olkin, 1990; Tso & Wu, 1999; Gau et al., 2008). Lemmas A.1 and A.2 state and prove these steps. The square law is their specialization to the reciprocal interval $[r^{-1}, r]$; the vector-error law uses the signed interval $[r^{-1} - 1, r - 1]$. The asymmetric bound follows by scalar centering. These matrix identities supply the exact constants for the specified operator intervals.

Inverse decay and the optimal norm bound. For the local theory, the corresponding starting point is decay of the inverse away from the spectral band. Demko et al. (1984) obtain inverse-entry decay from spectral bounds and exhibit a tridiagonal Toeplitz example attaining the decay exponent. For our reciprocal interval the decay parameter is $q = (r-1)/(r+1)$; similarity across adjacent log-eigenscales multiplies it by $e^{L/2}$. This explains the threshold mechanism. Theorem 4.1 additionally supplies the optimal uniform operator-norm bound $(c - h\cosh(L/2))^{-1}$ for arbitrary finite log-spectral configurations and treats the critical boundary explicitly. Band-limited conjugation gives the upper bound for arbitrary spectral configurations. The chain establishes sharpness and the boundary growth rates.

Weighted-resolvent methods. Weighted resolvent estimates offer another route to this decay mechanism. With $X = \log A$, we have $A^{1/2}C^{-1}A^{-1/2} = e^{X/2}C^{-1}e^{-X/2}$. Exponential conjugation underlies resolvent decay estimates (Combes & Thomas, 1973). The following specialization compares the published block-Jacobi rate with the sharp chain used in theorem 4.1.

Proposition J.3 (Optimized block-Jacobi comparison)**.** *For $r > 1$, let C be the half-line Toeplitz Jacobi operator with diagonal $c = (r + r^{-1})/2$ and off-diagonal entries $-a$, where $a = (r - r^{-1})/4$. Its lower spectral edge is $b = c - 2a = r^{-1}$. Set $\psi(t) = t^2 e^t$ for $t \geq 0$. At spectral parameter zero, the supremum of the decay rates given by Janas et al. (2018, Theorem 3.1), optimized over its cutoff $\delta > 0$ and slack $\varepsilon \in (0, 1)$, is*

$$s_{\mathrm{J}}(r) = \psi^{-1}\left(\frac{4}{r^2 - 1}\right). \tag{93}$$

The sum of this Green-kernel estimate therefore bounds the inverse after exponential weighting by $t = L/2$ whenever

$$(r^2 - 1)(L/2)^2 e^{L/2} < 4. \tag{94}$$

The exact Hermitian-part estimate in their equation (3.9), before the relaxation in equation (3.11), instead gives

$$\left\| e^{tN} C^{-1} e^{-tN} \right\|_2 \leq \left[c - 2a \cosh t \right]^{-1} \quad \text{when } c > 2a \cosh t, \tag{95}$$

where $N = \mathrm{diag}(0, 1, \ldots)$; the same bound holds for finite compressions. This is the sharp scalar-chain bound.

Proof. In the notation of their theorem, the off-diagonal block norms equal a, the spectral gap is b, and $\varphi_\delta(a) = 1/\sqrt{\max\{\delta, a\}}$. The exponent per index step is therefore

$$s(\delta, \varepsilon) = \sqrt{\frac{\delta}{\max\{\delta, a\}}}\, \psi^{-1}\left(\frac{(1 - \varepsilon)b}{\delta}\right).$$

Write $B = (1 - \varepsilon)b$. For $\delta \geq a$, this expression decreases with δ. For $\delta \leq a$, put $v = \psi^{-1}(B/\delta)$; then $s = \sqrt{B/a}\, e^{-v/2}$, which increases with δ. The maximum occurs at $\delta = a$, and the limit $\varepsilon \downarrow 0$ gives equation (93). For any $t < s_{\mathrm{J}}$, we can choose positive slack with $s > t$; the weighted entries then have a summable geometric bound. Since ψ is increasing, this is equation (94).

For this chain, the Hermitian-part perturbation has the exact bound $2a(\cosh t - 1)$. Subtract this bound from $b = c - 2a$ to obtain the lower bound $c - 2a \cosh t$ on the Hermitian part. This proves equation (95), first for truncated exponential weights and then in the limit. The published rate uses the relaxation $2(\cosh t - 1) \leq t^2 e^t$. □

At $r = 5/2$ and $L = \log 4$, the left side of equation (94) is $\frac{21}{2}(\log 2)^2 > 4$, whereas $c - 2a \cosh(L/2) = 11/80 > 0$. This separates the optimized published rate from the finite regime. The Chebyshev comparison gives the finite bound 13 at these parameters. Equation (95) identifies the overlap of the proof methods: Hermitian-part conjugation gives the exact scalar-chain constant. Theorem 4.1 optimizes the operator norm over all finite spectral placements and multiplicities under a global interval and a log-spectral support bound. The band-limited argument controls this entire class. The chain approaches the finite supremum and exhibits the growth rates at and above the threshold. Corollary 5.3 applies this optimal norm in the same update interval as the unrestricted law.

Curvature decisions. Finally, these matrix comparisons enter the application through computable residual and curvature bounds. Residual-to-error inequalities and logistic Hessian variation (Bach, 2010) provide computable inputs to the square law. The resulting intervals determine update-radius decisions with cached feature directions.

K Curvature-Sensitive Selection Protocol

This section gives the complete experimental protocol behind sections 6 and 7. We first state the cost model and shared selection algorithm, then describe the datasets, baselines, timing, and paired comparisons.

Cost accounting. We begin with the operations that differ across methods, because HVP counts alone do not capture their total cost. For n training examples, dimension d, and Q candidates, dense cache construction costs $O(nd^2 + d^3 + Qd^2)$ and stores $O(d^2 + dQ)$ numbers in addition to the features. Calibration adds solves on its example pool. At each training step, logits, candidate rescaling, and input bounds cost $O(nd + dQ)$. Dense HVPs cost $O(d^2)$ after an $O(nd^2)$ curvature construction; matrix-free HVPs cost $O(nd)$ directly from the features. The cached-factor preconditioner costs $O(d^2)$ per vector. Dense spectral tails require an $O(d^3 + nd^2)$ eigendecomposition and feature projection at cache setup, storing an additional $O(nd + d^2)$ numbers. In this eigenbasis, a diagonal preconditioner costs $O(d)$, and we reuse the target Gram matrix for the tail bound. This bound adds $O(d^2)$ work per step. With known local support, only globally supported rows contribute to the tail Gram matrix. Dense direct selection costs $O(nd^2 + d^3 + Qd^2)$ per step; for tridiagonal curvature, factorization and the Q solves cost $O(d(1+Q))$ after constructing the bands. The selected-set update in Algorithm 1 is shared by all methods.

Algorithm 1 Curvature-sensitive example selection at training step t

Inputs: fixed candidate pool $\mathcal{P}$, threshold τ, base factorization and cached directions $H_0^{-1}x_i$, current parameter θ_t, and fixed training preconditioner P_0.

1. Compute probabilities $p_{t,i}$, curvature weights $w_{t,i}$, and candidate gradients $g_{t,i} = (p_{t,i} - y_i)x_i$. Rescale the cached directions by $p_{t,i} - y_i$ and their error bounds by $|p_{t,i} - y_i|$.
2. Compute $\ell_\lambda, u_\lambda, \bar{\kappa}$ and the global interval in equation (19). Intersect with equation (16) when local and tail bounds are available.
3. Retain candidates whose lower endpoint exceeds τ and reject those whose upper endpoint is below τ.
4. For remaining candidates, evaluate the target residual and refine with PCG until the residual interval separates τ. Use a direct solve at the iteration limit.
5. Let S_t be the retained set and update $\theta_{t+1} = \theta_t - \gamma P_0^{-1}(|S_t|^{-1}\sum_{i\in S_t} g_{t,i} + \lambda_{\text{train}}\theta_t)$. Use a zero data-gradient term when S_t is empty.

Selection and training. With the shared costs specified, we now describe the trajectory on which all selectors are compared. At a base parameter θ_0, form H_0, cache solutions on a fixed 256-example candidate pool, and calibrate τ at a quantile of $\|H_0^{-1}g_i\|_2$ on calibration examples. We use quantiles $\{0.1, 0.25, 0.5, 0.75, 0.9\}$ and seeds $\{7, 17, 29, 43, 61\}$. After one full-training-set gradient step, each of eight consecutive checkpoints selects $S_t = \{i : \|H_t^{-1}g_i\|_2 > \tau\}$ from the fixed pool and updates

$$\theta_{t+1} = \theta_t - \gamma P_0^{-1}\left(\frac{1}{|S_t|}\sum_{i\in S_t} g_i + \lambda_{\text{train}}\theta_t\right).$$

The data-gradient term is zero when S_t is empty. $P_0 = I$ on digits and $P_0 = H_0$ in the local design. The cache and threshold stay fixed over the eight checkpoints. At each checkpoint, all selectors match the set returned by direct solves. The selected set drives the training update, after which we evaluate test accuracy and binary cross-entropy.

Digits pixels. The first benchmark has dense, low-dimensional curvature and no assumed local structure. We use the 1,797-example UCI digits dataset (Alpaydin & Kaynak, 1998) to predict parity from 64 pixels and an intercept. A stratified split permutes each digit class with seed 2027, assigns $\lfloor 0.6n_c \rfloor$ examples to training and $\lfloor 0.2n_c \rfloor$ to calibration, and uses the remainder for testing. This gives 1,074 training, 355 calibration, and 368 test examples. We divide pixels by 16 and center them using training means. We initialize the head from $\mathcal{N}(0, 0.01^2 I)$ and take 80 full-batch gradient steps with training regularization 0.01. The step size is $0.5/(\|X^\top X/n\|_\infty/4 + 0.01)$. Curvature damping is $\lambda/\|F_0\|_\infty \in \{0.01, 0.1, 1\}$, giving 75 eight-checkpoint trajectories and 600 candidate batches.

Synthetic logistic data. The second benchmark is constructed so that spectral locality is known exactly while global conditioning can vary. The 64-dimensional logistic design contains six copies of all four sign combinations on each adjacent coordinate pair, for 1,512 rows. Normalize each column to unit empirical variance and multiply it by σ_j, where σ_j^2 are geometrically spaced between K^{-1} and one, with $K \in \{10^2, 10^4, 10^6\}$. The balanced signs make the empirical covariance exactly

diagonal. At $\theta_0 = 0$, $F_0 = \frac{1}{4}\operatorname{diag}(\sigma_j^2)$; at every target parameter, the curvature remains tridiagonal. Consequently the relative matrix has width

$$L = \max_j \left|\log(\lambda + \sigma_{j+1}^2/4) - \log(\lambda + \sigma_j^2/4)\right|.$$

The feature support and cached diagonal entries determine this width. We use $\lambda = 10^{-8}\|F_0\|_\infty$, training regularization 10^{-8}, and preconditioned step size 0.1. Teacher parameters satisfy $\theta_j^\star = z_j/(\sigma_j\sqrt{64/2})$, $z_j \sim \mathcal{N}(0,1)$. Training, calibration, and test labels are independent Bernoulli draws with probability $\operatorname{sigmoid}(x_i^\top \theta^\star)$ on the same covariate design. The five seeds, three covariance condition numbers, and five quantiles give 75 trajectories per design variant.

To vary the locality width at the same global condition number, a second design interleaves the lower and upper halves of each consecutive 16-element block of the ordered variances. Adjacent coordinates then differ by at most eight spectral index steps instead of one. This design changes the known log-spectral width while preserving tridiagonal curvature in the feature coordinates.

Nonzero long-range tails. To test whether the local certificate remains useful away from exact bandedness, we add a controlled globally supported component. For the ordered-variance design, append the 64 rows of a Hadamard matrix scaled by σ_j. Scale the two row groups so that their covariance contributions are $(1-\delta)\operatorname{diag}(\sigma_j^2)$ and $\delta\operatorname{diag}(\sigma_j^2)$, respectively, for $\delta \in \{0.001, 0.01\}$. The base covariance stays diagonal. At a target checkpoint, only the globally supported rows contribute outside the known band. Form their weighted Gram matrix, transform by the cached diagonal $H_0^{-1/2}$, and retain the entries outside that band to obtain R. Use

$$s = \|R\|_\infty, \qquad \eta = \sqrt{\|H_0^{1/2} R H_0^{-1/2}\|_1 \|H_0^{1/2} R H_0^{-1/2}\|_\infty}.$$

These quantities bound the two required spectral norms from above. The bound $\|D_L\|_2 \le h + s$ controls the truncated local matrix. Selection time includes forming the tail Gram matrix, at $O(n_{\mathrm{tail}} d^2)$ cost, and computing its norm bounds, at $O(d^2)$ cost.

Baselines. All datasets use the same refinement logic, so the comparisons isolate the quality and cost of the initial screen. Exact and Similarity use the asymmetric damping interval in equation (18), with their respective norm bounds. Their symmetric variants use $r = \max\{u, \ell^{-1}\}$. The locality screen intersects the global exact interval with equation (16). The local Neumann baseline uses the same intersection, input information, and tail computation, replacing $\cosh(L/2)$ by $e^{L/2}$ in the lower bound on the Hermitian part. The local Chebyshev baseline uses proposition J.2, with the global interval and forward local bound unchanged. All screens and PCG share the cached directions, preconditioner, and residual stopping rule. For each cached matrix, use the Gershgorin lower bound

$$m_0 = \max\left\{\lambda, \min_j \left((H_0)_{jj} - \sum_{k\neq j} |(H_0)_{jk}|\right)\right\}.$$

The implementation subtracts an $8d\,\epsilon_{\mathrm{mach}}\|H_0\|_\infty$ rounding margin before taking the maximum with λ. For a diagonal base matrix this recovers its smallest diagonal entry up to this margin. In the synthetic design, $m_0/\lambda = 1 + 10^8/K$ before the margin. All iterative methods use $\mu_t = \lambda + \ell(m_0 - \lambda)$ and cached error radii divided by m_0. We form curvature only when an unresolved candidate first requires a matrix–vector product. The direct baseline uses dense Cholesky for digits and the tail variants, and a tridiagonal positive-definite solver for the tridiagonal designs. HVP implementations exploit the same known band structure.

Timing. We use a common timing protocol to separate algorithmic differences from warm-up and ordering effects. Experiments use float64, one BLAS thread, NumPy 2.2.6, and SciPy 1.15.3 on an Intel Xeon Platinum 8163 CPU at 2.50 GHz. At every checkpoint, we warm up each method once without timing, then perform five timed repetitions in shuffled order. We report the median runtime for each method. Selection time includes logits, gradient rescaling, input and tail bounds, lazy curvature construction, residuals, and solves. We record cache construction and the shared training update separately. For total runtime, `experiments/cache_costs.py` measures base calibration for the direct method and calibration plus candidate-direction caching for reuse methods,

with five shuffled repetitions per configuration. Figure 4 adds this one-time cost to the eight recorded selector calls and training updates after the shared initial full-data step. Both totals include the base calibration used to define the threshold. The entry point is `experiments/p1_study.py`; it writes per-checkpoint CSV records and the parameters and selected sets for each trajectory. Table 3 and figure 3 summarize these records.

Table 3: HVPs and selection time per 256-candidate batch using equation (20). Entries are mean ± standard deviation over five seeds, averaging checkpoints, quantiles, and damping levels or covariance condition numbers within each seed. For each checkpoint, runtime is the median of five measurements and includes input bounds, curvature construction, and refinement. Each design contributes 600 batches.

Dataset	Method	HVPs	Selection time (ms)
Digits	Similarity (symmetric)	221.8 ± 9.9	1.40 ± 0.04
	Exact (symmetric)	206.3 ± 10.7	1.38 ± 0.05
	Similarity	192.1 ± 9.6	1.36 ± 0.05
	Exact	173.8 ± 9.4	1.34 ± 0.06
	PCG	336.4 ± 7.2	1.50 ± 0.06
	Direct	—	1.23 ± 0.04
Synthetic	Similarity	608.5 ± 35.4	1.70 ± 0.06
	Exact	595.8 ± 34.8	1.70 ± 0.06
	Local Neumann	204.4 ± 35.5	1.31 ± 0.06
	Locality	189.3 ± 28.3	1.29 ± 0.05
	PCG	661.5 ± 34.3	1.75 ± 0.06
	Banded direct	—	0.76 ± 0.01

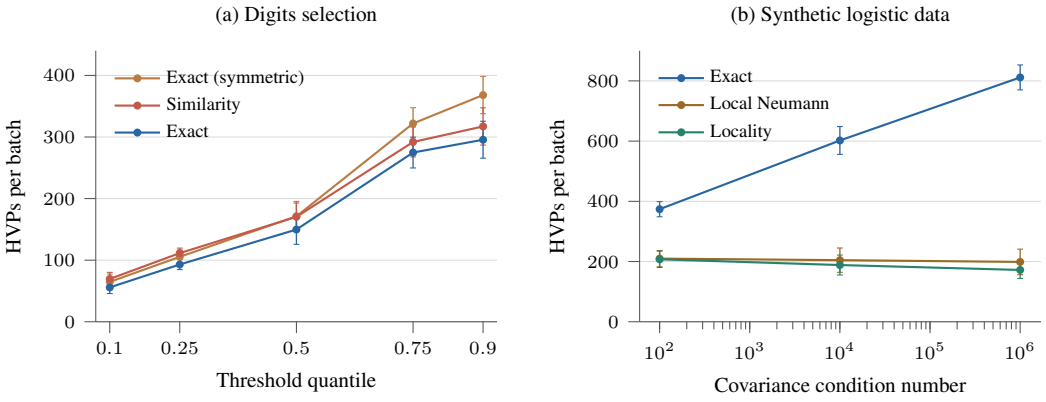

Figure 3: HVPs per 256-candidate batch. Left: threshold quantiles on digits, averaged across training steps and damping levels. Right: covariance condition numbers on synthetic logistic data, averaged across steps and quantiles. Error bars are standard deviations over five seeds.

Table 4: Final digits test performance after eight selected-example updates. Entries are mean ± standard deviation over five seeds, averaging the three damping levels within each seed.

Threshold quantile	Test accuracy (%)	Cross-entropy
0.1	89.49 ± 0.32	0.280 ± 0.005
0.25	89.58 ± 0.28	0.280 ± 0.005
0.5	89.98 ± 0.53	0.283 ± 0.006
0.75	90.13 ± 0.71	0.311 ± 0.011
0.9	87.32 ± 1.12	0.362 ± 0.013

Chebyshev comparison. The main local comparison isolates the gain from the sharp constant over a standard weighted inverse bound. We add the Chebyshev screen to the adjacent-

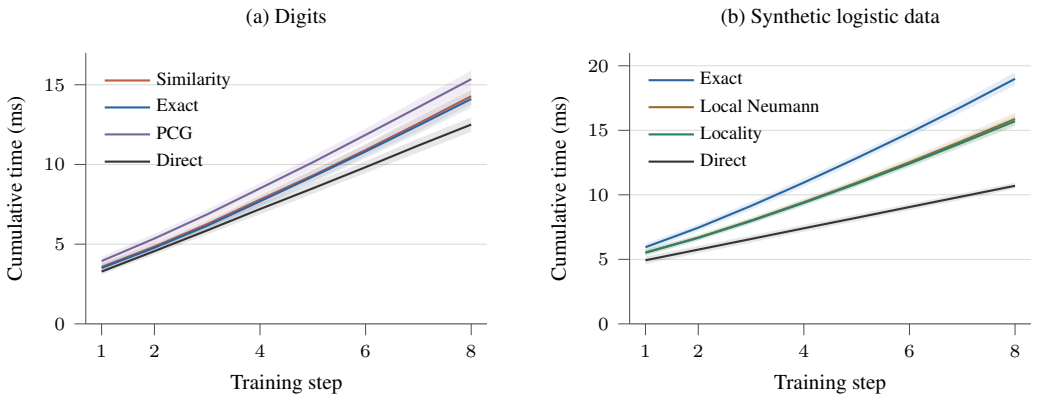

Figure 4: Cumulative selection and training time for the small designs, including base calibration and candidate caching. Lines and bands show mean and standard deviation over five seeds, averaging quantiles and damping levels or covariance condition numbers within each seed. Direct selection uses dense Cholesky on digits and a tridiagonal solver on synthetic data.

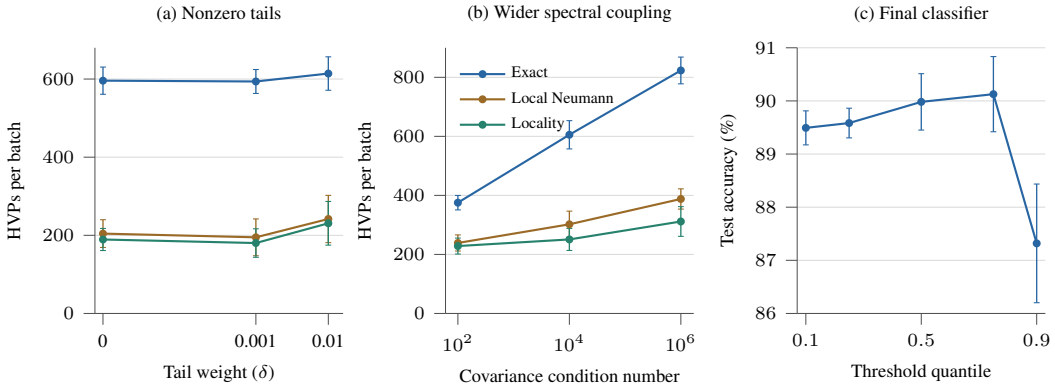

Figure 5: Selection ablations using five seeds. Left: HVPs with nonzero long-range tails at fixed base covariance. Middle: HVPs with wider log-spectral coupling at fixed global condition numbers. Right: final digits test accuracy after eight training updates. Error bars are standard deviations over seeds.

coupling, wider-coupling, and two nonzero-tail designs. Each design uses five seeds, three covariance condition numbers, five threshold quantiles, and eight steps. All methods use equation (20), one warm-up call, and five shuffled timing repetitions at every checkpoint. The entry point is `experiments/p1_study.py`, using `-dataset local -chebyshev -output p1_chebyshev`; `-stride 8` selects the wider-coupling design.

Table 5: Local screening with the Chebyshev baseline. HVPs and selection time are per 256-candidate batch. Entries are mean ± standard deviation over five seeds, averaging condition numbers, quantiles, and steps within each seed. Runtime is the median of five measurements per checkpoint and includes the tail calculation on the nonzero-tail designs.

Design	Method	HVPs / batch	Time (ms)
Adjacent coupling	Global exact	595.8 ± 34.8	1.68 ± 0.06
	Local Neumann	204.4 ± 35.5	1.30 ± 0.06
	Local Chebyshev	193.4 ± 29.4	1.29 ± 0.05
	Locality	189.3 ± 28.3	1.28 ± 0.05
	PCG	661.5 ± 34.3	1.74 ± 0.05
	Banded direct	—	0.76 ± 0.02
Wider coupling	Global exact	601.3 ± 37.8	1.67 ± 0.10
	Local Neumann	309.6 ± 34.4	1.38 ± 0.09
	Local Chebyshev	278.9 ± 37.6	1.36 ± 0.09
	Locality	263.6 ± 36.7	1.34 ± 0.08
	PCG	666.1 ± 37.8	1.73 ± 0.09
	Banded direct	—	0.76 ± 0.04
Tail weight 0.001	Global exact	593.8 ± 30.6	1.88 ± 0.09
	Local Neumann	194.9 ± 47.1	1.62 ± 0.10
	Local Chebyshev	184.3 ± 37.7	1.61 ± 0.09
	Locality	180.2 ± 36.5	1.60 ± 0.09
	PCG	662.6 ± 28.4	1.93 ± 0.09
	Direct	—	1.35 ± 0.03
Tail weight 0.01	Global exact	614.1 ± 42.8	1.93 ± 0.11
	Local Neumann	241.8 ± 60.5	1.71 ± 0.12
	Local Chebyshev	235.4 ± 57.3	1.70 ± 0.11
	Locality	230.8 ± 55.9	1.70 ± 0.13
	PCG	681.5 ± 40.8	1.98 ± 0.11
	Direct	—	1.34 ± 0.02

Paired comparison. Aggregate means can hide whether an improvement is consistent across random seeds, so we also report matched differences. All four screens use the same seed, candidate pool, threshold quantile, curvature state, and residual solver. For each seed, we average the baseline-minus-locality difference over three condition numbers, five quantiles, and eight checkpoints (120 matched batches). Table 6 reports the five seed averages, their mean, and their minimum and maximum. Positive values indicate lower cost for locality. Because the optimal interval lies inside each baseline interval, it requires refinement for no more candidates. The paired differences quantify the resulting savings.

L Random-Feature Experiments

The dense random-feature study complements the structured synthetic design by varying dimension, candidate count, and cache-refresh frequency. This section specifies the representation and timing protocol before describing the two implementations of spectral-tail reuse.

Features and training. We first construct a shared feature representation and pretrained head for all solver comparisons. We use the same digits parity labels, fixed data split, and pixel normalization as in appendix K. For a radial basis function (RBF) kernel $k(x, z) = \exp(-\gamma\|x - z\|_2^2)$, random Fourier features (Rahimi & Recht, 2007) take the form

$$\varphi_j(x) = \sqrt{\frac{2}{d-1}}\cos(\omega_j^\top x + b_j), \quad \omega_j \sim \mathcal{N}(0, 2\gamma I), \quad b_j \sim \mathrm{Uniform}[0, 2\pi].$$

Table 6: Paired baseline-minus-locality differences for each seed in the four-screen study. HVPs and selection time are averaged per batch. The last column gives the mean and range across the five seed averages.

Design	Baseline	7	17	29	43	61	Mean [min, max]
HVP difference per batch							
Adjacent	Global exact	404.03	405.12	425.87	395.70	401.98	406.54 [395.70, 425.87]
Adjacent	Neumann	8.05	18.10	16.54	7.34	25.40	15.09 [7.34, 25.40]
Adjacent	Chebyshev	3.00	4.17	5.01	2.80	5.39	4.07 [2.80, 5.39]
Wider	Global exact	352.77	339.70	348.07	324.68	323.53	337.75 [323.53, 352.77]
Wider	Neumann	51.58	51.60	42.51	43.02	41.35	46.01 [41.35, 51.60]
Wider	Chebyshev	14.43	17.68	14.70	13.10	16.65	15.31 [13.10, 17.68]
Selection-time difference (ms)							
Adjacent	Global exact	0.4004	0.4046	0.4052	0.3836	0.3952	0.3978 [0.3836, 0.4052]
Adjacent	Neumann	0.0036	0.0265	0.0035	0.0024	0.0295	0.0131 [0.0024, 0.0295]
Adjacent	Chebyshev	0.0009	0.0105	0.0024	0.0005	0.0041	0.0037 [0.0005, 0.0105]
Wider	Global exact	0.3148	0.3576	0.3320	0.3156	0.3503	0.3341 [0.3148, 0.3576]
Wider	Neumann	0.0380	0.0393	0.0321	0.0414	0.0567	0.0415 [0.0321, 0.0567]
Wider	Chebyshev	0.0065	0.0206	0.0205	0.0060	0.0250	0.0157 [0.0060, 0.0250]

Set γ to the reciprocal of twice the median pairwise squared distance among 256 sampled training examples. Center the features by training means and append an intercept to obtain dimension d. Seeds $\{7, 17, 29, 43, 61\}$ vary frequencies, phases, and candidate pools; their prefixes are shared across dimensions and pool sizes. We initialize each head at zero and take 80 full-batch gradient steps with regularization 0.001 and step size $0.5/(\|X^\top X/n\|_\infty/4 + 0.001)$. Feature construction and these pretraining steps precede the timed selection and training steps.

Initialization and reuse. After pretraining, the timed portion begins with the cache that each method needs for selection. We start timing before constructing H_0 with $\lambda = 0.01\|F_0\|_\infty$ and setting τ to the median update length on the 355 calibration examples. Reuse methods also cache candidate directions and their residual error bounds. After one timed full-data gradient step, each method executes 32 selected-example updates with $P_0 = I$. We check that selected sets and parameter trajectories agree with direct selection, with the comparisons outside the timer. The threshold and damping stay fixed over the 32 steps. For a refresh interval $R = 8$, rebuild the factorization, cached directions, and input bounds before steps 9, 17, and 25; $R = 32$ keeps the initial cache for all 32 steps. The cumulative runtime includes every refresh.

Solvers and timing. We then vary only the screening bound and linear-algebra implementation. The exact and similarity screens use the asymmetric damping interval, Gershgorin residual lower bound, cached directions, and PCG refinement. PCG starts from the same cached directions. Each method has a dense implementation that forms curvature only when needed, and a matrix-free implementation that applies

$$H_t V = X^\top \big(w_t \odot (XV)\big)/n + \lambda V.$$

The base Cholesky factor supplies the preconditioner in both implementations. Direct selection uses a fresh dense Cholesky factorization at each step. HVP counts include applications to PCG directions and fresh residual evaluations, with each column counted separately. The environment matches appendix K. We perform one warm-up run for each method, followed by five timed runs in shuffled order. We take the median runtime per seed and report its mean and standard deviation across seeds. The timer covers initialization, all selection calls, and all updates; test predictions and reference comparisons occur after the timer. The full grid contains $3 \times 2 \times 5 = 30$ configurations; the refresh comparison adds five at $d = 1024$, $Q = 512$. `experiments/kernel_study.py` records every step of every run, together with reference trajectories, feature parameters, and source hashes.

Separate target and tail matrices. The first locality implementation forms the target and tail quantities separately, making the additional dense work explicit. For $d = 128$, $Q = 512$, we additionally evaluate $L \in \{\log 2, \log 4, \log 8\}$ on the full dense representation for each seed. At cache setup, compute $H_0 = U\Lambda U^\top$ and project the features to $Y = XU$, and include both operations in

Table 7: HVPs and runtime at $d = 128$. HVPs are per candidate batch; setup includes calibration and, for iterative methods, candidate caching. Total time includes setup and 32 selection and training steps. Entries are mean $\pm$ standard deviation over five seeds.

d	Q	Method	HVPs / batch	Setup (ms)	Total (s)
128	128	Exact dense	142.2 ± 31.5	5.61 ± 0.48	0.088 ± 0.004
		Exact MF	142.2 ± 31.5	5.11 ± 0.25	0.156 ± 0.030
		Similarity dense	153.3 ± 26.5	5.25 ± 0.37	0.087 ± 0.009
		Similarity MF	153.3 ± 26.5	5.35 ± 0.31	0.164 ± 0.026
		PCG dense	166.2 ± 21.5	5.39 ± 0.38	0.088 ± 0.008
		PCG MF	166.2 ± 21.5	5.34 ± 0.11	0.164 ± 0.028
		Direct	—	4.46 ± 0.56	0.081 ± 0.002
128	512	Exact dense	466.0 ± 49.4	8.12 ± 0.24	0.129 ± 0.007
		Exact MF	466.0 ± 49.4	8.06 ± 0.15	0.392 ± 0.040
		Similarity dense	519.8 ± 42.6	7.91 ± 0.36	0.134 ± 0.009
		Similarity MF	519.8 ± 42.6	8.09 ± 0.32	0.426 ± 0.040
		PCG dense	596.2 ± 22.5	7.93 ± 0.26	0.143 ± 0.009
		PCG MF	596.2 ± 22.5	7.98 ± 0.29	0.486 ± 0.051
		Direct	—	4.43 ± 0.31	0.155 ± 0.006

Table 8: HVPs and runtime at $d = 512$, using the protocol in table 7.

d	Q	Method	HVPs / batch	Setup (ms)	Total (s)
512	128	Exact dense	128.9 ± 14.5	50.72 ± 1.78	0.796 ± 0.022
		Exact MF	128.9 ± 14.5	48.89 ± 1.75	0.618 ± 0.044
		Similarity dense	141.0 ± 12.7	47.70 ± 1.87	0.803 ± 0.016
		Similarity MF	141.0 ± 12.7	47.29 ± 1.19	0.631 ± 0.039
		PCG dense	155.5 ± 9.6	47.91 ± 2.86	0.818 ± 0.014
		PCG MF	155.5 ± 9.6	47.67 ± 2.46	0.644 ± 0.045
		Direct	—	38.98 ± 1.34	0.914 ± 0.030
512	512	Exact dense	468.3 ± 47.8	69.59 ± 2.00	1.190 ± 0.078
		Exact MF	468.3 ± 47.8	70.78 ± 2.38	1.521 ± 0.171
		Similarity dense	521.8 ± 42.5	69.90 ± 1.26	1.225 ± 0.073
		Similarity MF	521.8 ± 42.5	69.87 ± 2.73	1.603 ± 0.164
		PCG dense	590.8 ± 24.4	67.27 ± 1.92	1.275 ± 0.047
		PCG MF	590.8 ± 24.4	72.09 ± 4.89	1.769 ± 0.135
		Direct	—	38.48 ± 0.98	1.426 ± 0.030

Table 9: HVPs and runtime at $d = 1024$, using the protocol in table 7.

d	Q	Method	HVPs / batch	Setup (ms)	Total (s)
1024	128	Exact dense	124.0 ± 17.1	152.02 ± 4.11	2.525 ± 0.069
		Exact MF	124.0 ± 17.1	152.69 ± 4.15	1.316 ± 0.139
		Similarity dense	137.9 ± 14.6	151.04 ± 2.71	2.580 ± 0.066
		Similarity MF	137.9 ± 14.6	154.02 ± 6.03	1.375 ± 0.108
		PCG dense	153.8 ± 9.7	151.01 ± 6.97	2.610 ± 0.066
		PCG MF	153.8 ± 9.7	149.60 ± 5.51	1.417 ± 0.090
		Direct	—	126.15 ± 3.70	3.169 ± 0.054
1024	512	Exact dense	438.0 ± 53.5	213.16 ± 6.95	3.479 ± 0.222
		Exact MF	438.0 ± 53.5	212.29 ± 5.33	2.845 ± 0.361
		Similarity dense	496.5 ± 44.4	209.69 ± 5.75	3.660 ± 0.260
		Similarity MF	496.5 ± 44.4	210.22 ± 8.94	3.055 ± 0.330
		PCG dense	574.4 ± 21.6	216.05 ± 6.71	3.910 ± 0.179
		PCG MF	574.4 ± 21.6	210.29 ± 8.06	3.336 ± 0.297
		Direct	—	123.21 ± 4.27	4.388 ± 0.088

Table 10: Cache-refresh comparison at $d = 1024$, $Q = 512$ over 32 steps. $R = 8$ includes three refreshes; $R = 32$ uses the initial cache throughout. Direct selection forms and factors curvature at every step in both configurations. Entries are mean $\pm$ standard deviation over five seeds.

	$R = 8$		$R = 32$	
Method	HVPs / batch	Total (s)	HVPs / batch	Total (s)
Exact dense	140.4 ± 27.7	3.022 ± 0.180	438.0 ± 53.5	3.479 ± 0.222
Exact MF	140.4 ± 27.7	1.983 ± 0.214	438.0 ± 53.5	2.845 ± 0.361
Similarity dense	212.4 ± 30.0	3.205 ± 0.159	496.5 ± 44.4	3.660 ± 0.260
Similarity MF	212.4 ± 30.0	2.255 ± 0.216	496.5 ± 44.4	3.055 ± 0.330
PCG dense	528.4 ± 6.5	4.026 ± 0.165	574.4 ± 21.6	3.910 ± 0.179
PCG MF	528.4 ± 6.5	3.487 ± 0.212	574.4 ± 21.6	3.336 ± 0.297
Direct	—	4.408 ± 0.158	—	4.388 ± 0.088

Table 11: Final test performance after 32 selected-example updates on random features. Entries are mean $\pm$ standard deviation over five seeds.

d	Q	Test accuracy (%)	Cross-entropy
128	128	89.78 ± 1.10	0.381 ± 0.022
128	512	89.89 ± 1.09	0.379 ± 0.024
512	128	90.22 ± 1.62	0.376 ± 0.010
512	512	90.92 ± 0.68	0.375 ± 0.012
1024	128	90.43 ± 1.08	0.376 ± 0.005
1024	512	90.65 ± 0.81	0.377 ± 0.007

the measured runtime. The spectral global, local Neumann, and locality screens share the eigenvalue lower bound and condition-number bound, including the rounding margin used for residual stopping. An additional global exact screen uses the inexpensive norm bounds and skips the eigendecomposition. At each target step, form

$$C = \Lambda^{-1/2} \left(Y^\top \operatorname{diag}(w_t) Y/n + \lambda I \right) \Lambda^{-1/2}, \qquad R_{ij} = C_{ij}\, \mathbf{1}\{|\log \Lambda_{ii} - \log \Lambda_{jj}| > L\}.$$

Use $s = \|R\|_\infty$ and $\eta = \sqrt{\|\Lambda^{1/2} R \Lambda^{-1/2}\|_1 \|\Lambda^{1/2} R \Lambda^{-1/2}\|_\infty}$ in corollary 5.3. HVPs and direct solves use the full target curvature. Every iterative method has dense and matrix-free implementations, timed separately on each configuration. We evaluate the three widths on the same feature matrices and training trajectories to keep global conditioning fixed. Table 12 includes the eigendecomposition, projection, tail construction, and norm bounds in total time. `experiments/kernel_tail_costs.py` separately times the tail construction and its two norm bounds at the saved checkpoints, using one warm-up and five repetitions per checkpoint. `experiments/kernel_dense_locality.py` executes the dense implementations with the same features, trajectories, and repetition counts.

Spectral-basis reuse. To remove duplicated matrix construction, the revised implementation works throughout in the cached spectral basis. It computes $H_0 = U \Lambda U^\top$ and $Y = XU$ at cache setup. We rotate cached directions by $U^\top$; candidate gradients use the corresponding rows of Y, and PCG uses the diagonal preconditioner Λ. Training updates and logits remain in the original feature coordinates. Orthogonal invariance preserves update lengths and the selection threshold.

We form each target curvature once as $K_t = Y^\top \operatorname{diag}(w_t) Y/n + \lambda I$. The local screens use this matrix for both HVPs and tail bounds. At cache setup, precompute the mask $M_{ij} = \mathbf{1}\{|\log \Lambda_{ii} - \log \Lambda_{jj}| > L\}$ and $v_i = \Lambda_{ii}^{-1/2}$. For $T = |K_t| \odot M$, the two required bounds are

$$s = \|v \odot (Tv)\|_\infty, \qquad \eta = \sqrt{\|T(v \odot v)\|_\infty \|(v \odot v) \odot (T^\top \mathbf{1})\|_\infty}.$$

These operations add $O(d^2)$ work to the target Gram matrix.

The spectral global, Neumann, Chebyshev, locality, and spectral PCG methods all use these projected features and diagonal preconditioner. The norm-based global and PCG baselines use the original coordinates and cached Cholesky factor. Direct selection uses fresh Cholesky factorizations. "Locality

Table 12: Spectral locality on digits random features at $d = 128$, $Q = 512$, over 32 steps, with separate target and tail matrices. "Exact spectral" and the two local screens use eigenvalue bounds; "Exact" uses cached matrix norms. Total time includes initialization and the spectral computations for each method. HVP counts agree between dense and matrix-free implementations. Direct selection uses Cholesky in both timing series. Entries are mean ± standard deviation over five seeds.

L	Method	HVPs / batch	Dense time (ms)	MF time (ms)
$\log 2$	Exact	466.0 ± 49.4	124.75 ± 7.84	365.99 ± 43.92
	Exact spectral	452.1 ± 49.8	127.99 ± 7.75	361.17 ± 44.10
	Local Neumann	425.9 ± 55.0	182.76 ± 8.70	406.31 ± 46.92
	Locality	424.0 ± 56.3	179.74 ± 10.25	407.15 ± 40.86
	PCG	596.2 ± 22.5	132.78 ± 8.18	533.45 ± 39.01
	Direct	—	147.76 ± 5.34	145.98 ± 5.21
$\log 4$	Exact	466.0 ± 49.4	124.90 ± 7.18	364.34 ± 36.49
	Exact spectral	452.1 ± 49.8	125.62 ± 5.50	367.66 ± 44.73
	Local Neumann	438.0 ± 54.5	180.69 ± 7.92	407.21 ± 40.18
	Locality	435.9 ± 56.5	180.45 ± 9.73	414.51 ± 52.44
	PCG	596.2 ± 22.5	132.37 ± 3.71	531.23 ± 23.66
	Direct	—	146.22 ± 6.48	146.59 ± 6.06
$\log 8$	Exact	466.0 ± 49.4	122.60 ± 3.02	362.25 ± 38.44
	Exact spectral	452.1 ± 49.8	127.11 ± 8.47	364.21 ± 41.69
	Local Neumann	450.0 ± 50.9	181.99 ± 6.75	415.32 ± 48.17
	Locality	449.8 ± 51.2	176.58 ± 6.35	423.55 ± 33.94
	PCG	596.2 ± 22.5	132.28 ± 6.76	527.55 ± 27.67
	Direct	—	145.16 ± 1.44	146.56 ± 4.01

Table 13: Time to compute spectral-tail bounds per checkpoint with a separate projected Gram matrix. Entries are mean ± standard deviation over five seeds, averaging the 32 checkpoints within each seed. Time includes Gram construction, band masking, and both norm bounds.

L	Time (ms)
$\log 2$	1.59 ± 0.07
$\log 4$	1.60 ± 0.05
$\log 8$	1.64 ± 0.04

separate" uses the original implementation: separate Gram matrices in the original and spectral coordinates, with the original-coordinate Cholesky preconditioner.

We evaluate all three widths at $d = 128$, and use $L = \log 2$ for the dimension comparison at $d = 512$. Candidate count is $Q = 512$ and the cache is fixed over 32 steps. Each of five seeds has one warm-up run and five timed runs per method in shuffled order. Timing begins before calibration and cache setup and ends after the last selected-example update. The timer includes eigenvectors, feature projection, direction rotation, tail calculation, and refinement. We compare selected sets and parameter trajectories with direct selection outside the timer. `experiments/locality_reuse_study.py` records the runs.

Table 14: Spectral-basis reuse at $d = 128$ and $Q = 512$, including initialization and 32 selection and training steps. "Locality separate" forms both Gram matrices and uses the original-coordinate preconditioner. Entries are mean $\pm$ standard deviation over five seeds, using the median of five runs per seed.

L	Method	HVPs / batch	Total time (ms)
$\log 2$	Exact	466.0 ± 49.4	123.16 ± 4.11
	Exact spectral	452.1 ± 49.8	120.04 ± 4.77
	Local Neumann	425.9 ± 55.0	122.39 ± 2.99
	Local Chebyshev	425.9 ± 55.0	122.98 ± 8.07
	Locality	424.0 ± 56.3	122.77 ± 6.52
	Locality separate	424.0 ± 56.3	175.82 ± 10.30
	PCG	596.2 ± 22.5	127.31 ± 6.54
	PCG spectral	595.9 ± 22.4	129.85 ± 2.78
	Direct	—	143.29 ± 3.01
$\log 4$	Exact	466.0 ± 49.4	124.37 ± 7.56
	Exact spectral	452.1 ± 49.8	126.01 ± 4.91
	Local Neumann	438.0 ± 54.5	124.59 ± 6.96
	Local Chebyshev	438.0 ± 54.5	126.67 ± 6.00
	Locality	435.9 ± 56.5	128.08 ± 5.29
	Locality separate	435.9 ± 56.5	177.36 ± 10.67
	PCG	596.2 ± 22.5	134.00 ± 4.23
	PCG spectral	595.9 ± 22.4	132.25 ± 4.29
	Direct	—	143.67 ± 2.82
$\log 8$	Exact	466.0 ± 49.4	122.35 ± 6.88
	Exact spectral	452.1 ± 49.8	121.08 ± 6.75
	Local Neumann	450.0 ± 50.9	124.96 ± 8.15
	Local Chebyshev	450.0 ± 50.9	127.31 ± 4.26
	Locality	449.8 ± 51.2	124.68 ± 7.05
	Locality separate	449.8 ± 51.2	179.84 ± 6.80
	PCG	596.2 ± 22.5	129.99 ± 6.56
	PCG spectral	595.9 ± 22.4	129.71 ± 3.05
	Direct	—	145.18 ± 5.66

Table 15: Spectral-basis reuse at $d = 512$, $Q = 512$, and $L = \log 2$, using the protocol in table 14.

L	Method	HVPs / batch	Total time (ms)
$\log 2$	Exact	468.3 ± 47.8	1156.36 ± 77.93
	Exact spectral	444.1 ± 48.6	1127.15 ± 46.75
	Local Neumann	427.6 ± 46.5	1162.21 ± 49.28
	Local Chebyshev	427.6 ± 46.5	1161.45 ± 53.47
	Locality	427.1 ± 46.7	1163.11 ± 44.97
	Locality separate	427.1 ± 46.7	1791.02 ± 69.11
	PCG	590.8 ± 24.4	1244.44 ± 65.46
	PCG spectral	590.8 ± 24.4	1249.48 ± 36.22
	Direct	—	1360.96 ± 22.95

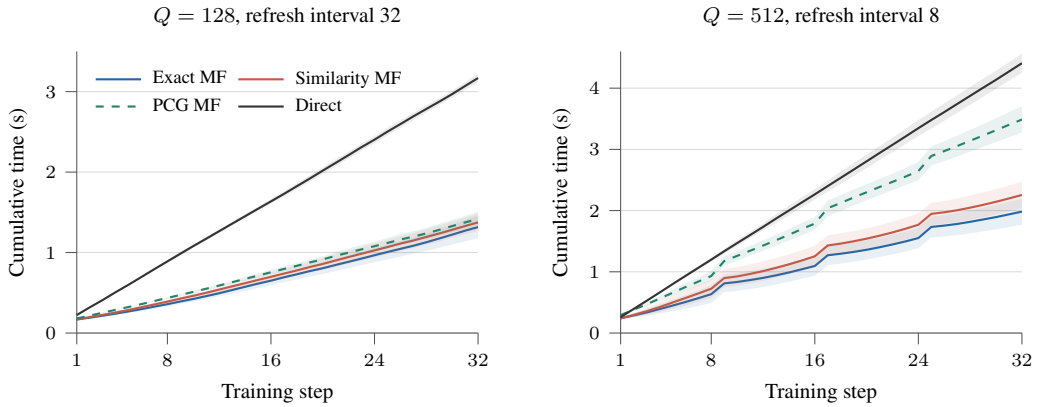

Figure 6: Cumulative selection and training time on 1024-dimensional random features, including base calibration, candidate caching, and refreshes. Iterative curves use matrix-free HVPs; Direct uses dense Cholesky. The cache stays fixed for 32 steps on the left and is refreshed every eight steps on the right. Lines and bands show mean and standard deviation over five seeds, after taking the median of five runs per seed.

M Frozen-Feature Checkpoint Reuse

The preceding experiments update one frozen-feature head over a short cache window. We also test reuse across more widely separated checkpoints and across two feature sources. The comparison uses frozen multilayer perceptron (MLP) features on digits and Pythia-70M representations (Biderman et al., 2023) of BoolQ (Clark et al., 2019), with head dimensions 129 and 513. It evaluates symmetric input intervals, two base checkpoints, three checkpoint intervals, three damping levels, and five seeds per dataset. Each of the 90 settings contains 128 candidate gradients.

Table 16: Frozen-feature checkpoint reuse with symmetric input intervals. HVPs and runtimes are totals over 90 settings per dataset. Total time adds cache construction for each setting to selection time.

Dataset	Method	HVPs	Selection time (ms)	Total time (ms)
Digits	Exact	2,332	37.14	108.16
	Similarity	2,886	37.82	108.83
	PCG	12,144	66.97	130.55
	Cholesky	—	49.57	49.57
BoolQ	Exact	906	73.04	706.26
	Similarity	1,470	84.42	717.63
	PCG	11,544	182.14	766.94
	Cholesky	—	529.61	529.61

Relative to similarity, HVPs decrease by 19.2% and 38.4%, and selection time decreases by 1.8% and 13.5%. Relative to PCG, HVPs decrease by 80.8% and 92.2%. Exact has the lowest selection time, and Cholesky has the lowest total runtime on both datasets.

Representations. We first describe the two fixed representations on which only the final binary head is trained. The digits experiment uses a stratified 60/20/20 training/calibration/test split. A one-hidden-layer MLP with 128 tanh units learns the ten-class task. We then freeze its hidden representation and fit a binary parity head. The language experiment freezes Pythia-70M (Biderman et al., 2023) and fits a binary head on BoolQ (Clark et al., 2019). A stratified subsample selects 1,024 training and 256 calibration examples from the original training split and 512 test examples from the original validation split. We tokenize the question, passage, and answer prefix to at most 128 tokens. The mean final-layer state over non-padding positions gives 512 features.

For both datasets, we center features using training means and divide them by the root mean square feature norm on training data, then add an intercept. The head dimensions are 129 and 513. We minimize mean logistic loss with $0.01\|\theta\|_2^2/2$ regularization by full-batch gradient descent. A training-covariance norm bound determines the fixed step size. Seeds $\{7, 17, 29, 43, 61\}$ vary head initialization and candidate order on a shared data split and representation.

Checkpoint grid. With the representations fixed, we vary both the age of the cache and the amount of damping. Base checkpoints are training steps 20 and 80. Target checkpoints are 1, 4, and 16 training steps after each base checkpoint. Damping is fixed within each pair at $\lambda/\|F_0\|_\infty \in \{0.01, 0.1, 1\}$. This gives $2 \times 3 \times 3 \times 5 = 90$ settings per dataset. For each base checkpoint and damping, the squared radius is the median cached squared update length on calibration examples. It stays fixed across the three target checkpoints. Each setting evaluates 128 held-out candidate gradients.

Cache and methods. Each method receives the same base information, so differences arise only from screening and refinement. The cache stores the base curvature factorization and its solutions on the fixed candidate feature directions. At a target checkpoint, current gradient scalars rescale these solutions. The three iterative methods use this common cache and preconditioner:

1. *Exact*: screen with $\mathcal{W}_d(r, \bar{\kappa})$, then evaluate target residuals and apply PCG to unresolved candidates.
2. *Similarity*: use $(1 + (r-1)\sqrt{\bar{\kappa}})^2$ in the same screening and refinement procedure.
3. *PCG*: start with target residuals and use the same residual stopping rule.

The direct baseline forms the current damped curvature and uses dense Cholesky on both datasets.

We compute the input factor and conditioning bound from logistic weights, cached matrix norms, feature row norms, and damping (proposition F.4). Cached solve residuals enlarge the initial interval. For the residual and PCG stages, each proposed decision uses a fresh target residual with $H_t \succeq \lambda I$. We use a direct solve if PCG reaches the iteration limit. Computations use float64.

Metrics and timing. We count work at the candidate level and include cache construction only in the total-time comparison. One HVP is one target matrix–vector product; a batch of 128 vectors counts as 128 HVPs. Selection time includes target logits, gradient rescaling, screening, residual evaluation, and PCG. Total time adds base-matrix construction, factorization, and cached direction solves for each setting. Feature extraction, head training, and calibration precede timing. We time each method once per setting on CPU. Iterative methods run in a shuffled order, followed by direct Cholesky. The table reports summed runtimes and HVPs across the 90 settings. After the iterative runs, direct solutions supply the reference threshold decisions.

Results. Under this protocol, all methods preserve the reference decisions but pay different screening and setup costs. Each dataset contributes 90 batches of 128 candidate gradients. All four methods select the same candidates, and every iterative solve finishes before the iteration limit. Exact, Similarity, and PCG each use 313 PCG iterations on digits and 12 on BoolQ. Additional screening therefore saves initial residual evaluations.

N Numerical Illustrations of Spectral Mixing

These numerical checks separate three aspects of the theory: the gap between exact and elementary global bounds, the effect of spectral support at fixed input accuracy, and the finite-dimensional behavior near the locality threshold.

N.1 Exact and observed distortion

We begin with the unrestricted setting, comparing the exact worst case with both the elementary bound and randomly oriented extreme points. At $\epsilon = 0.01$ and $\kappa = 10^2, 10^4, 10^6$, the elementary similarity excess E is 1.99, 1.87, and 1.20 times the optimal excess $G = \mathcal{W}_n(1+\epsilon, \kappa) - 1$. The corresponding scaled accuracies $x = \epsilon\sqrt{\kappa}$ are 0.1, 1, and 10. Figure 2a plots the exact ratio and its joint limit from corollary A.5 on a linear vertical axis.

For every $n \in \{2, 8, 32\}$ and $\kappa \in \{10^2, 10^4, 10^6\}$, we sample 500 random extreme points with geometrically spaced eigenvalues of A and $C = r^{-1}I + (r - r^{-1})P$, where P is a Haar-random orthogonal projection of rank $\lfloor n/2 \rfloor$ and $B = A^{1/2}CA^{1/2}$. At $n = 32$, the median $(\phi(A^2, B^2) - 1)/G$ is 0.591, 0.329, and 0.132 as κ increases. The construction in equation (11) attains a ratio of one. The gap measures the effect of cross-scale alignment among extreme points of the same Loewner interval.

N.2 Fixed-parameter spectral support

We next hold the Loewner factor and global conditioning fixed while changing only where the relative matrix can couple the spectrum. Fix $n = 32$, $r = 1.01$, and $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with geometrically spaced eigenvalues from 1 to 10^8. For bandwidth w, let G_w be a symmetric Gaussian matrix masked to $|i - j| \le w$, with extreme eigenvalues $\gamma_{\min} < \gamma_{\max}$. Define

$$C_w = r^{-1}I + (r - r^{-1})\frac{G_w - \gamma_{\min} I}{\gamma_{\max} - \gamma_{\min}}, \qquad B = A^{1/2}C_w A^{1/2}. \tag{96}$$

The normalization preserves bandwidth and fixes $\phi(A, B) = r$. Geometric eigenvalue spacing makes index bandwidth proportional to log-spectral width. For each $w \in \{0, 1, 2, 4, 8, 16, 31\}$, we use 120 independent draws. The median $\phi(A^2, B^2)$ is 1.0201 at $w = 0$ and 289 at $w = 31$. The singular-value identity equation (6) evaluates these quantities directly from the relative operators.

At $r = 1.02$ and $\kappa = 10^4$, commuting, adjacent-scale, and extreme-scale constructions give inverse-square distortion 1.0404, 1.04236, and 5.74878781, respectively. The extreme-scale construction attains $\mathcal{W}_n(r, \kappa)$. These examples illustrate how the location of a perturbation in the reference spectrum changes the output at fixed input accuracy.

N.3 Scanning across the locality threshold

Finally, we follow the sharpness construction across the predicted boundary and track its dependence on dimension. For the chain in equation (37), fix $L = \log 4$, so $r_c = \coth(L/4) = 3$. We form the tridiagonal relative matrix $M_n = cI - \frac{h}{2}(2T_n + \frac{1}{2}T_n^\top)$ directly, solve $M_n X = I$ with a banded linear solver, and compute the largest singular value of X. The analytical comparison uses α^{-1} below the threshold. The scan has $n \in \{16, 64, 256\}$ and 24 values of r between 1.2 and 3.2, with spacing 0.005 immediately around the threshold. The separate dimension scan uses $n \in \{8, 16, 32, 64, 128, 256\}$ and $r \in \{2, 3, 3.2\}$. All computations use double precision.

At $n = 256$ and $r = 2$, the measured norm is 3.19853, approaching the optimal uniform value 3.2. At $r = 2.9$, the measured norm is 36.3314 and the uniform bound is 39.3220. At $r = 3$, the norms for $n = 64, 128, 256$ are 40.5060, 81.2517, and 162.7401. At $r = 3.2$, they are 196.058, 3902.17, and 1.50443×10^6. Each finite-dimensional curve remains finite at r_c. As dimension increases, the curves distinguish convergence to $1/\alpha$ from linear and exponential growth (figure 1b,c).

Preconditioned-solve work. The inverse-amplification transition need not coincide with a difficult preconditioned solve. To show the distinction, we also solve $C_n z = b$ for a unit Gaussian right-hand

side using conjugate gradients with relative residual tolerance 10^{-10}. This is the A_n-preconditioned B_n system in whitened coordinates. The five seeds are $7, 17, 29, 43, 61$, and each right-hand side is shared across the r scan. One multiplication by C_n corresponds to one target-curvature product under this change of coordinates. Figure 7 reports HVPs, with mean and standard deviation over the five seeds. The bounded condition number $\kappa_2(C_n) \le r^2$ keeps solve work small across r_c; the phase transition concerns Euclidean inverse amplification after similarity. Screening costs additionally depend on threshold margins and the cost criterion in proposition 6.1.

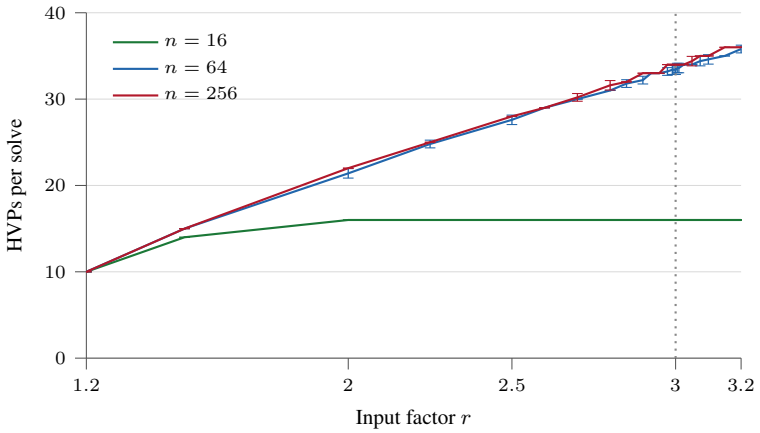

Figure 7: HVPs for the preconditioned chain solves at relative residual tolerance 10^{-10}. Error bars are standard deviations across five paired seeds. The dotted line marks the inverse-amplification threshold $r_c = 3$.

O Reproducibility Details

This final section maps each reported result to its experiment entry point and saved artifact. It also distinguishes checks available in the anonymous supplement from historical records needed for full regeneration.

Code and result provenance. We begin with the contents and limitations of the supplied snapshot. The anonymous code supplement contains the main experiment entry points, dependencies, and numerical checks. Run `run_checks.py--help` for verification commands and experiment entry points. The script `scripts/prepare_digits_input.py` reconstructs the public digits input from the copy bundled with scikit-learn. The supplement contains code only, without datasets, model weights, cached features, or results. The paths below identify the historical artifacts used for the reported figures. Reconstruction of those artifacts still requires the original raw records and plotting scripts, which are absent from this snapshot. The optional native-tail optimization branch is excluded because its C source is unavailable; the reported main-text comparisons use the included NumPy implementation.

The selection experiments run through `experiments/p1_study.py`, with common screening and refinement routines in `experiments/screening.py`. Results in `results/p1_strong` contain the digits, tridiagonal, and nonzero-tail trajectories; `results/p1_strong_width8` contains the wider-coupling ablation. The CSV files record every checkpoint, threshold quantile, seed, method, selected-example count, HVPs, and runtime. Compressed arrays contain the features, labels, base parameters, checkpoint parameters, and selected sets. `scripts/plot_p1_results.py` generates the selection tables and figures. The random-feature experiments use `experiments/kernel_study.py`, with scaling and refresh results in `results/p1_kernel_scale`. The locality results are in `results/p1_kernel_locality` and `results/p1_kernel_locality_dense`; the latter are generated by `experiments/kernel_dense_locality.py`. `experiments/kernel_tail_costs.py` records the tail-bound computation time, and `scripts/plot_kernel_results.py` generates the random-feature tables, figure, and numerical macros from these saved records. The Chebyshev comparisons are in `results/`

p1_chebyshev and results/p1_chebyshev_width8. experiments/locality_reuse_study.py generates the spectral-basis results in results/p1_locality_reuse and results/p1_locality_reuse_512. scripts/plot_locality_revision.py generates their tables and numerical macros in results/p1_locality_revision. The source preceding these implementation changes is archived in results/source_round5. The frozen-feature checkpoint protocol is given in appendix M.

Historical formula and chain checks used scripts/verify_revision.py. The scan in experiments/phase_transition_study.py records norms, solve counts, dimensions, input factors, and settings in results/phase_transition. From screening records, scripts/report_phase_revision.py produces paired seed differences and the summaries in results/phase_revision used for figure 2 and figure 1a. The historical scripts/inline_paper_plots.py embeds saved plot coordinates, uncertainty bands, tables, and numerical macros in the source.