WHEN ARE SAFETY ANCHORS ENOUGH? AVAILABILITY AND CERTIFIED PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety anchors constrain fine-tuning, but a useful guarantee must connect available example gradients to the updates a provider actually permits. We study preservation under a single global parameter update chosen from an enforced region. A spectral nonlinear risk gate turns available-span geometry into explicit pool-expansion, constraint-budget, and abstention decisions. For noncentered Gaussian gradients, an exact scalar representation connects systematic drift to the same spectral feasible families. A complementary pointwise envelope supports conditional flip-rate calibration through fixed-sequence testing. We characterize why this envelope can be vacuous even when every global update is safe, and why average-risk control can permit complete failure on a rare category. Experiments test unplanted 128-dimensional pools, conflicting nonlinear tasks, disjoint handwritten-digit splits, and global curvature bounds through three hidden layers. The results distinguish larger retained-direction displacements from improved task progress: anchors can enable a prescribed protected direction yet greatly limit progress on a conflicting task. On real classifiers, a 2% conditional flip ceiling gives modest task gains and shows clear calibration and curvature limits. The contribution is an anchor-specific availability and certification analysis, with explicit successful and unsuccessful operating regimes.

1 INTRODUCTION

An anchor set can preserve its selected outputs yet miss an important direction, prevent useful adaptation, or lose protection after a nonlinear parameter change. These are different failure modes. Counting examples alone cannot distinguish them.

Safety-sensitive projection methods already address alignment damage from fine-tuning and low-rank modifications (Qi et al., 2024a; Wei et al., 2024; Hsu et al., 2024; Zhang et al., 2025; 2026). The decision studied here is narrower: given an available gradient pool and a specified risk ceiling, which update regions can a provider justify? The provider may accept a region, acquire examples from a missing category, or abstain. An independent trusted optimizer, which evaluates only its final model, solves a weaker problem and remains an important comparator.

Two certification routes answer different parts of this decision. A *global-update* spectral bound keeps the same displacement across protected inputs. Its threshold directly connects to the exact available-span feasible family. A *pointwise* envelope allows the adverse displacement to differ across inputs; it is easier to calibrate, but can be much more conservative. Neither route uniformly dominates. We distinguish these routes: a dimension-free binomial interval does not guarantee a useful certificate in every dimension.

This distinction also changes how we evaluate benefit. A larger anchored radius need not contain a smaller unanchored ball. We therefore report task-bank support and achieved progress, certify new errors among base-correct examples, and test both global and category-targeted attacks. The experiments include difficult cases in which a certificate abstains or anchoring reduces useful learning. Proofs appear in the appendices.

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.

2 CONTRACT, QUANTITIES, AND PRIOR WORK

What is preserved. A fixed base model has a specified scalar decision margin $m_0(x) = m_{\theta_0}(x)$ and gradient $g(x) = \nabla m_{\theta_0}(x)$ in p trainable Euclidean coordinates. For the main preservation contract, $\nu = Q(\cdot \mid m_0 > 0)$: risk is the fraction of previously correct examples made incorrect. For a named group A, use $Q(\cdot \mid m_0 > 0, A)$ instead. Absolute error is reported separately. A guarantee on ν does not cover arbitrary post hoc subgroups.

Enforced total displacement. The provider fixes

$$\mathcal{K}(J, \rho, \zeta) = \{\Delta : \|\Delta\| \le \rho,\ \|J\Delta\| \le \zeta\}, \qquad R_\nu(\mathcal{K}) = \sup_{\Delta \in \mathcal{K}} \nu(m_{\theta_0+\Delta} \le 0). \tag{1}$$

A user may choose a task, optimizer, stopping time, or later update, but every total displacement remains measured from the same base model. One deployed model is chosen before a fresh protected input is drawn. Resetting the radius at each step or permitting an input-dependent model is a different contract. A trusted final-model evaluator can allow more learning because it need not protect every alternative update.

Literal and aggregate constraints. Available gradients form rows of G. Literal anchors use $J = G_I$ and bound individual first-order margin changes. Aggregate constraints use $J = FG$ and preserve weighted combinations, allowing individual changes to cancel. Aggregate rank is not a count of preserved outputs. One example can supply multiple independent rows when several outputs are constrained. We state which semantics each experiment uses.

Safety geometry and useful progress. For exact enforcement let P project onto $\ker J$. Define

$$C = \mathbb{E}_\nu g g^\top, \quad \kappa_C(P) = \lambda_{\max}(PCP), \quad \mathcal{U}_d(P) = \|Pd\| \,/\, \|d\|, \quad S_{\mathcal{D}}(P, \rho) = \rho \min_{d \in \mathcal{D}} \mathcal{U}_d(P). \tag{2}$$

The spectral residual controls worst squared linear drift. The last quantity is the minimum available first-order progress over unit-normalized task directions in a declared bank. With soft enforcement it is a conservative support measure because the exact free ball remains allowed. It does not predict final accuracy when the task gradient changes.

Preservation methods and geometry. Our analysis complements safety projection, representation preservation, and hardening methods (Hsu et al., 2024; Zhang et al., 2025; 2026; Lu et al., 2026; Huang et al., 2024a;b;c; Rosati et al., 2024; Zou et al., 2024). Systematic refusal directions and safety-landscape analyses motivate treating means and displacement scale explicitly (Arditi et al., 2024; Qi et al., 2024b; Peng et al., 2024). Local safety–capability and curvature analyses study related trade-offs (Chen et al., 2025; Rosati et al., 2026). Elastic Weight Consolidation (EWC), Gradient Episodic Memory (GEM), Averaged GEM (A-GEM), Orthogonal Gradient Descent (OGD), Gradient Projection Memory (GPM), and sketched OGD (SketchOGD) preserve information through penalties or restricted updates (Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020; Saha et al., 2021; Min et al., 2025); OGD theory, gradient surgery, and AlphaEdit give related generalization, task, and null-space perspectives (Bennani et al., 2020; Yu et al., 2020; Fang et al., 2025). Our Schur and inertia arguments build on classical indefinite geometry, quadratic optimization, matrix approximation, and design (Gohberg et al., 2005; Polyak, 1998; Ye and Zhang, 2003; Halko et al., 2011; Boutsidis et al., 2011; Pukelsheim, 2006).

Certification and risk control. Weight-perturbation certification and Lipschitz-margin methods already bound model-output changes (Weng et al., 2020; Tsai et al., 2021; Tsuzuku et al., 2018). Our pointwise term specializes this approach to anchor-constrained parameter regions. Selecting a radius by valid hypothesis tests is Learn then Test (LTT), an existing calibration principle (Angelopoulos et al., 2022; Bates et al., 2021). The anchor-specific objects are the realizable constraint family, its global-risk threshold, and the distinction between population preservation and pointwise robustness.

3 CONNECTING AVAILABLE ANCHORS TO A GLOBAL-RISK GATE

A spectral gate does not need to replace $\kappa_C(P)$ by the larger trace of PCP. The key is to apply the probabilistic bound to each fixed global update before taking the supremum.

Theorem 1 (Spectral nonlinear preservation)**.** *Suppose $C \preceq \overline{C}$, $\nu(m_0 \le \gamma) \le b$ for $\gamma > 0$, and every protected margin has Hessian operator norm at most L throughout the radius-ρ ball. Then exact-anchor updates obey*

$$R_\nu(\mathcal{K}(J,\rho,0)) \le \min\left\{1, b + \frac{(\rho\sqrt{\kappa_{\overline{C}}(P)} + L\rho^2/2)^2}{\gamma^2}\right\}. \tag{3}$$

If the statistical assumptions hold on one confidence event, the bound holds there simultaneously for all projectors, including those selected using the estimates.

Taylor's theorem, Markov's inequality, and the L_2 triangle inequality prove the result (Appendix B). For soft enforcement the same bound uses the worst quadratic drift over $\mathcal{K}$, computable by the classical two-constraint scalar dual. The ingredients are standard; their operational consequence is a threshold on exactly the quantity characterized by available-span geometry.

For a risk ceiling $\epsilon > b$ and nonnegative slack, define

$$\tau = \left(\frac{\gamma\sqrt{\epsilon - b} - L\rho^2/2}{\rho}\right)^2. \tag{4}$$

A negative unsquared slack triggers abstention. Otherwise the spectral gate passes exactly when $\kappa_{\overline{C}}(P) \le \tau$.

Let S span the available row space, T its orthogonal complement, and $q = \dim S$. Every aggregate design leaves T free. Write $H = \overline{C} - \tau I$, $A = T^\top H T$, and, when $A \preceq 0$, let E span the vectors in S whose coupling into $\ker A$ vanishes. Put $h = \dim E$ and

$$M = E^\top H E - E^\top \overline{C} T A^\dagger T^\top \overline{C} E.$$

Theorem 2 (Available-span gate feasibility)**.** *The spectral gate is achievable by aggregate rows $J = FG$ if and only if $A \preceq 0$. Its free spaces are exactly $T \oplus EV$ with $V^\top V = I$ and $V^\top M V \preceq 0$. The minimum aggregate rank and the one-task utility ceiling are*

$$k_S = q - h + n_+(M), \qquad B_S(d,\tau) = \frac{\sqrt{\|T^\top d\|^2 + b(M, E^\top d)^2}}{\|d\|}, \tag{5}$$

where $b(M,e) = \max_{\|v\|=1, v^\top M v \le 0} |e^\top v|$, or zero if no safe unit vector exists. The ceiling is attained at rank k_S.

The generalized Schur reduction and maximal-nonpositive extension prove the characterization, including singular cases (Appendix A). Thus the floor $\lambda_{\max}(T^\top \overline{C} T)$ decides whether to expand the pool, k_S screens rank budgets, and B_S screens local task requirements. We implement the attaining aggregate construction, not only pool principal component analysis (PCA). Literal subsets form a stricter family.

Failing this sufficient gate does not prove actual risk is high. For a coherent missing category, Appendix C.3 gives a stronger witness: an unavoidable projected mean can produce one allowed displacement that violates the entire category. Section 6 evaluates both types of decision.

4 POINTWISE CALIBRATION AND ITS UNAVOIDABLE GAP

The spectral route is attractive for dispersed gradients, but estimating an upper matrix can be expensive. A pointwise envelope avoids that matrix estimate by certifying a stronger event.

For each fixed (J,ζ) let Q_J span its row space, $P = I - Q_J Q_J^\top$, and s be its smallest nonzero singular value. For an input, set $a = \|Pg\|$, $b_J = \left\|Q_J^\top g\right\|$, $z = \min(\rho, \zeta/s)$, and

$$w = \min\{z, \rho b_J / \|g\|\}, \qquad h_\rho(x) = a\sqrt{\rho^2 - w^2} + b_J w. \tag{6}$$

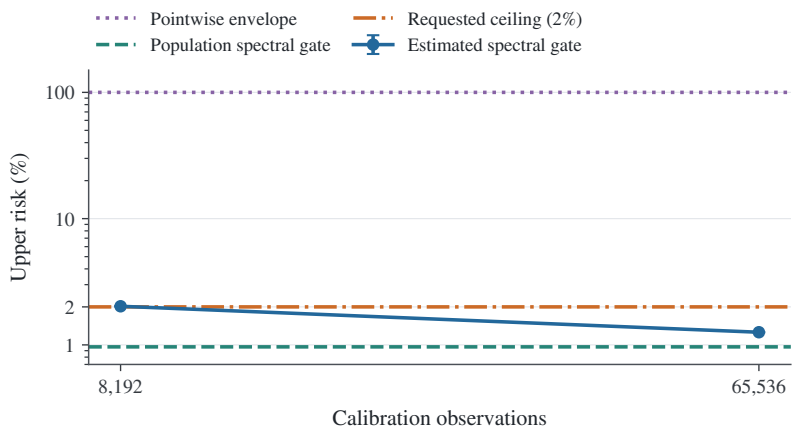


Figure 1: One global update versus input-wise perturbations. In the diffuse-gradient study, the estimated spectral gate passes a 2% risk target at 65,536 samples while the exact pointwise envelope remains one. The population spectral line is a reference, not the estimate used for certification.

The adverse linear change is at most h_ρ. With a valid bound $L(x, \rho)$ over the entire ball, a previously correct example can flip only if

$$e_\rho(x) = \mathbf{1}\{0 < m_0(x) \le h_\rho(x) + L(x, \rho)\rho^2/2\} = 1. \tag{7}$$

Zero gradients and unconstrained designs have the natural limiting definitions. Near-redundant rows weaken protection through s. The support uses the smallest singular value rather than the full ellipsoid, making its additional conservatism measurable.

Theorem 3 (Ordered conditional preservation calibration)**.** *Fix design/tolerance chains, groups, a radius order, and a risk ceiling ϵ independently of calibration. In each chain, test envelope risk on retained base-correct calibration examples from the smallest radius upward, stopping at the first failure. Allocate α_c to each chain/group with $\sum_c \alpha_c \le \delta$. With probability at least $1 - \delta$, every accepted region has conditional flip risk at most ϵ, including after data-dependent selection of its radius and subsequent updates within that region.*

This is fixed-sequence testing in the LTT framework (Angelopoulos et al., 2022). The proof charges any erroneous acceptance to the first true null in its chain. It does not divide α_c by the number of radii. We report the guaranteed ceiling ϵ, not a raw binomial endpoint at the selected radius; the latter is not automatically a simultaneous confidence curve. Bonferroni across radii remains a valid, more conservative comparator (Appendix C).

A statistical improvement cannot fix a quantifier gap. The envelope upper-bounds $\sup_\Delta \mathbb{E}[\mathbf{1}\{\text{flip under } \Delta\}]$ by $\mathbb{E}[\sup_\Delta \mathbf{1}\{\text{flip under } \Delta\}]$. For isotropic Gaussian gradients in a free space of dimension r, constant margin, and zero curvature, these exact risks are respectively

$$\Phi(-t) \quad \text{and} \quad \Pr(\chi_r \ge t), \qquad t = \gamma/(\rho\sigma). \tag{8}$$

At $t = c\sqrt{r}$ with $0 < c < 1$, the first tends exponentially to zero and the second exponentially to one. A unit-sphere example at $p = 128$, $\rho = 1$, $\gamma = 0.9$ has envelope mass one but spectral risk bound 0.97%. No amount of envelope calibration resolves that discrepancy.

Four distinct losses of informativeness. The certificate has four sources of conservatism: the global-to-pointwise quantifier change, the smallest-singular-value support relaxation, the Taylor-radius remainder, and finite-sample calibration. The first remains even with exact support and zero curvature. The final confidence interval has no explicit ambient-dimension factor, but geometry, gradient estimation, and curvature can still make the certified decision useless.

5 SYSTEMATIC DRIFT, TASK BANKS, AND DEEPER MODELS

An exact bridge for a nonzero gradient mean. In the affine margin model $m_\Delta = \gamma + g^\top \Delta$, if $g \sim N(\mu, \Sigma)$, $\gamma > 0$, and Σ is positive definite on the free space, its Gaussian risk ceiling $\epsilon < 1/2$ is equivalent to

$$\rho \sup_{v=Pv, \|v\|=1} \{|\mu^\top v| + z_{1-\epsilon}\sqrt{v^\top \Sigma v}\} \le \gamma.$$

For $C_\lambda = (1 + 1/\lambda)\mu\mu^\top + (1 + \lambda)z_{1-\epsilon}^2\Sigma$, the sharper identity is

$$\left[\sup_{v=Pv, \|v\|=1} \{|\mu^\top v| + z_{1-\epsilon}\sqrt{v^\top \Sigma v}\}\right]^2 = \inf_{\lambda>0} \kappa_{C_\lambda}(P). \tag{9}$$

Young's inequality gives one direction. At the scalar minimum, a top-eigenspace vector makes Young's inequality tight, proving the reverse (Appendix D.2). The continuous union of available-span frontiers for C_λ therefore describes noncentered Gaussian feasibility, with a separately handled mean-null boundary. A finite λ grid remains only a sufficient search. Removing the mean is not universally optimal. Independent trust-region and spectral-dual calculations agree in 243 tested cases, including singular cases.

Several tasks require an actual bank calculation. Extending a safe free space to maximum dimension improves every task's utility, but does not attain every separate one-task ceiling. We give a semidefinite upper bound and feasible-subspace lower bounds for the max–min task-bank objective. For a prescribed free line, four tasks give squared utility $(1-1/\sqrt{2})/2 \approx 0.15$ while the semidefinite program (SDP) permits 0.5. Deciding prescribed-line bank feasibility is hard for nondeterministic polynomial time (NP-hard) by a direct PARTITION reduction; this does not establish hardness when larger free spaces are freely allowed. Appendix D.1 states the formulation, reduction, and scope. We do not substitute separate $B_S(d_i, \tau)$ values for a joint task guarantee.

Availability under sampling and perturbation. For matched Gaussian gradient proposals, the forced-free floor is the squared residual of a randomized range finder. The Halko–Martinsson–Tropp (HMT) bound gives a pool-size benchmark based on spectral tails (Halko et al., 2011). Proposal mismatch can instead leave a direction inaccessible at every sample size. Floors are 1-Lipschitz in covariance operator error; feasible families are bracketed by thresholds $\tau \pm \eta$. Rank and utility boundaries can still jump near singular Schur complements. We use upper matrices and check realized constraints rather than treating a noisy zero eigenvalue as exact (Appendix D.3).

Curvature beyond a single hidden layer. For a deep hyperbolic tangent (tanh) network with frozen readout and *all* hidden affine matrices trainable, we propagate first-derivative vectors and second-derivative quadratic forms in layerwise perturbation norms. Cross-layer terms are included. The resulting global Hessian bound holds throughout the parameter ball, not merely at sampled points. The depth study applies the same compositional formula to depths one through three. It shows severe radius contraction and abstentions rather than assuming the shallow certificate scales to transformers. The exact recurrence and proof are in Appendix E.

6 OPERATIONAL GEOMETRY AND CERTIFICATE CONSERVATISM

Design choices precede the relevant calibration; test data never select constraints or attack restarts. Reported standard errors (SEs) describe variation over fitted seeds or splits, not an independent proof of population calibration.

A theorem-derived design changes a passing decision. In a 128-coordinate signed-categorical model, candidates are actual sampled gradients, not planted oracle normals. A diagonal upper covariance comes from 65,536 independent observations. A known quadratic remainder gives $L = 0.02$; the requested radius is one, risk ceiling 5%, and retained task utility at least 0.76. We compare the attaining aggregate construction with pool PCA at the same rank and with literal enforcement of the entire pool.

At pool size 256, the frontier construction passes both risk and utility in **13/20** matched-proposal trials; pool PCA and literal-full pass in **0/20**. At sizes 64 and 128 it passes 2/20 and 4/20. These are sample frequencies, not structural impossibility claims. If the proposal excludes the leading deployment mode, every tested size is rejected by the pool floor. The theorem supplies the reason to change the proposal rather than merely repeat selection. Aggregation has different output-preservation semantics, as specified in Section 2.

Table 1: Targeted preservation at a 2% ceiling. Bounds are means over 20 calibration draws. The attack rate is exact in the specified categorical population, not an empirical optimum on a neural network.

Eligible pool	Avg. envelope bound	Rare-group bound	Attack on group	Task support
Missing rare mode	1.07%	100%	100%	0.30
Add its true gradient	$4.80 \times 10^{-3}\%$	0.71%	0%	0.30

Average risk can hide a missing category. For $m_\Delta(j) = 0.25 + \Delta_j - 0.01\,\|\Delta\|^2$, let mode zero have mass 1% and let the proposal supply only modes one through nine. At radius 0.3, the average spectral gate and average envelope pass, while the mode-zero floor forces the group gate to abstain. A single allowed displacement flips *every* rare example, matching the lower-bound witness of Appendix C.3. Acquiring one genuine rare-category gradient changes the conditional decision to acceptance without reducing the declared future task support (Table 1). A direct bank gate also detects failure; the geometry identifies the needed pool change.

A dispersed regime reverses which gate is useful. For uniform sphere gradients in dimension 128, margin 0.9, and radius one, the exact global risk is 6.22×10^{-48} while the exact envelope is one. Full covariance estimation gives mean spectral upper bounds of 2.02% at $n = 8{,}192$ and 1.26% at $n = 65{,}536$ over five seeds. Thus the estimated spectral gate passes a 2% ceiling at the larger sample even though envelope calibration cannot. This is a real matrix-estimation experiment, not substitution of the known population covariance.

7 Conflicting tasks and depth limits

Conflicting tasks on the protected distribution. We adapt all 2,048 hidden weights and biases of a 128-unit tanh network with frozen readout. The protected population is concentrated near one input direction. Every task uses that same distribution and asks for a downward output shift of 0.4, 1, or 2. Independent calibration on 8,192 inputs selects a region at a 2% conditional flip ceiling before 300 optimization steps. Ten seeds and held-out task inputs are used; methods have identical step counts and learning rates. Anchor rows use trace selection, orthogonal–triangular (QR) factorization, or random selection.

Table 2: Larger anchored radii can limit useful progress on conflicting tasks. Gain is held-out loss reduction; errors are observed flips among base-correct inputs. The first four rows have a 2% certified ceiling. All seeds select the shown radii.

Design	Radius	Task-bank support	Gain, shift 2	Flip rate
No anchors	0.75	0.75	$63.21 \pm 0.25\%$	0%
Trace, four	2.00	0.09	$17.26 \pm 0.44\%$	0%
QR, four	2.00	0.10	$17.88 \pm 0.34\%$	0%
Random, four	2.00	0.20	$24.92 \pm 0.94\%$	0%
Uncertified ball	2.00	2.00	$(99.95 \pm 2.00 \times 10^{-3})\%$	100%

At shift two, anchoring permits longer displacements but obstructs the task itself (Table 2). At shift one, the unanchored certified design gains 95.42%, compared with 8.64% for trace anchors. A rule maximizing base-point task support chooses the unanchored region in every run, while a radius-only rule can choose a much less useful anchored design. These decisions use gradients and accepted regions, not held-out gains. The zero observed flips do not establish zero population risk; the operational guarantee remains the predeclared 2% ceiling.

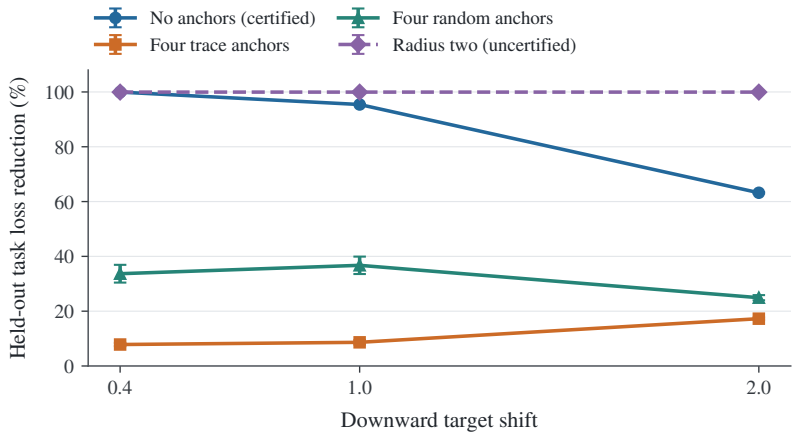


Figure 2: Conflicting tasks distinguish radius from utility. Anchored regions permit radius two but preserve little progress on downward-shift tasks drawn from the protected distribution. The uncertified comparator flips all protected test inputs at shift two. Error bars are SEs over ten seeds.

A valid deeper-network bound can be practically small. We train 27 real digit-parity models: depths one, two, and three; widths 16, 32, and 64; three seeds. Every hidden affine matrix is trainable in the certificate, and readouts are frozen. The study applies the global compositional Hessian bound to nine radii and calibrates conditional flips on disjoint data.

Table 3: Depth study at a 2% flip ceiling; means include abstentions as zero. The same compositional bound is used at every depth; the main shallow experiments use a tighter one-layer formula.

Hidden depth	Models accepting a radius	Mean radius (10^{-3})	Range (10^{-3})
One	$8/9$	15.28	0–40
Two	$5/9$	1.67	0–5
Three	$3/9$	1.11	0–5

The bound includes cross-layer derivatives and requires no block-diagonal Hessian. It nevertheless becomes too conservative to authorize meaningful displacements in many deeper fits. This is evidence about the present certificate's limits, not a claimed transformer-ready method.

8 Disjoint real-data preservation and adversarial updates

Protocol. For each of ten fits on real handwritten digits (Scikit-learn developers, 2026), we use disjoint sets of 600 training, 200 design, 600 calibration, and 397 test images. The binary task is digit parity. A 64-unit tanh model has 4,160 adaptable hidden parameters; its readout is frozen after training. Calibration conditions on base correctness and selects radii at a **2% new-error ceiling**. Three tasks classify images rotated by 15, 30, or 45 degrees, with 300 adaptation steps each. All final task measurements use the untouched test images.

Calibration assumes independent and identically distributed (iid) images. It does not establish new-writer generalization. Repeated splits reuse images across fits; SEs describe split and training variability. The data form a small real benchmark; they do not measure language-model refusal or safety.

What the anchors buy. The larger radii do not translate to greater worst-task support or a clear gain advantage (Table 4). A larger anchored radius can produce an incomparable region: it can exclude directions permitted by the smaller ball. The mean base test error is 4.48%; final absolute errors range from 4.41% to 4.48%, and are reported separately from certified flips. For example, a 2% conditional flip bound implies the population degradation inequality $R_{\text{new}} - R_{\text{base}} \leq 0.02(1 - R_{\text{base}})$, not an absolute-error ceiling.

Table 4: Real preservation under disjoint calibration/test splits. Radii and task support are shown in units of 10^{-2}. Gain and flip rate are averaged over the three tasks within each fit, then over ten fits; gain errors are seed-level SEs. Every fit accepts a global-flip region. Each accepted region has the stated ceiling.

Design	Radius mean [min,max]	$10^2 S_{\mathcal{D}}$	Task gain	Flips	Bound
No anchors	$4.50\ [2.50, 10]$	4.50	$2.27 \pm .43\%$	0.17%	2%
Trace 8	$10.00\ [5, 20]$	3.70	$2.13 \pm .39\%$	0.04%	2%
Trace 16	$13.00\ [5, 30]$	4.40	$2.40 \pm .40\%$	$9.00 \times 10^{-3}\%$	2%
QR 8	$10.50\ [5, 20]$	4.20	$2.37 \pm .41\%$	$9.00 \times 10^{-3}\%$	2%
Random 8	$10.50\ [5, 20]$	4.40	$2.41 \pm .48\%$	0.05%	2%

A stronger trusted comparator. We also train unanchored final models on a fixed radius grid, then evaluate each model's actual calibration flips with Bonferroni across tasks and radii. All 30 task/fit pairs find a passing final model; mean task gain is 6.50% and observed flips 0.35%. This method certifies the chosen final model, not the alternatives in its parameter ball. Its weaker contract allows more benign learning.

Calibration and group resolution. Fixed-sequence testing increases mean global-flip radius from 2.50 to 4.50 without anchors and from 7.00 to 13.00 with 16 trace anchors (all in units of 10^{-2}), relative to Bonferroni across the same 12 radii. But demanding the same 2% guarantee separately on even and odd base-correct images leaves only $0/10$ unanchored, $1/10$ QR/random, and $2/10$ trace fits with a positive certified radius. The retained samples per group are small. We report abstentions as unchanged models rather than relaxing the risk target after seeing results.

9 Attack evidence, limits, and implications

Attacks test enforcement, not certificate tightness. We optimize a global-margin attack and a targeted-even-margin attack on base-correct training images, each with three starts and 250 steps. Restarts are chosen by their training objective. Test and calibration images are never used in attack optimization. For each fit we report the larger test flip rate attained by the two predeclared objectives, then average across fits.

Table 5: Strongest tested global flip rate per fit from two attack objectives. Outside-ball columns are unanchored and uncertified; radius-two results reuse the same attack control across designs and are not independent trials.

Design	Inside region	Bound	Same-radius ball	Radius-two ball
No anchors	0.16%	2%	0.16%	29.07%
Trace 8	0.05%	2%	0.37%	29.07%
Trace 16	0.08%	2%	0.53%	29.07%
QR 8	0.08%	2%	0.37%	29.07%
Random 8	0.11%	2%	0.45%	29.07%

The unconstrained radius-two attack can damage preservation, while tested inside-region attacks remain below the certified ceiling. Same-radius unanchored attacks also stay below 2% empirically. Consequently these attacks neither prove the anchors are necessary at the selected radii nor approximate the worst population risk tightly. In the targeted-even objective, the large unanchored ball attains a mean worst-tested even-group flip rate of 58.91%. Conditional guarantees are therefore a distinct requirement, not an implication of an average ceiling.

What is established. The spectral gate makes the availability theorem operational for a defined sufficient risk criterion. The pool experiment constructs an attaining aggregate rather than assuming ideal normals are observed. The rare-category experiment shows a concrete missing-pool decision and an actual targeted-failure witness. The scalar mean representation extends the spectral feasible-family connection to systematic Gaussian drift. The real-data experiments test conditional preservation, ordered calibration, and attacks with honest task-support comparisons.

What remains limited. The available-family argument uses classical Schur complements and indefinite subspaces; the calibration uses existing risk-control tools. Literal subset optimization and finite-bank optimization are not solved by one-task ceilings. The generic learned-Jacobian experiments use the envelope, not a claim that the available-span theorem predicts their winning selector. Large-model spectral estimation can be expensive, and pointwise envelopes can be uninformative even with unlimited data. The deep study demonstrates that valid global curvature bounds alone do not imply useful radii. Local tangent descriptions (Jacot et al., 2018; Chizat et al., 2019) and a Low-Rank Adaptation (LoRA) parameterization (Hu et al., 2021) do not resolve that problem.

The guarantee assumes a fixed trusted base, enforced total displacement, the declared protected distribution, and calibration independent of design. It is a per-model population statement, not zero lifetime failures or protection against every hidden category. Data-dependent group discovery needs fresh evidence. Our classifier study is small; no pretrained large language model (LLM), attention-block, or useful transformer-curvature experiment is provided. These limitations constrain deployment claims even when every numerical check passes.

Turning a diagnosis into a deployment decision. First specify a conditional preservation target and any groups requiring separate protection. Obtain a valid curvature bound at the proposed radius. With a simultaneous covariance upper matrix, translate the risk target into τ and check the available floor, rank budget, and joint task requirement. A missing-mode floor motivates changing the proposal; a large statistical error motivates additional independent evidence. If the pointwise envelope is structurally too large, more calibration cannot fix it, but the global spectral gate may. When several regions pass, compare retained task support and achieved progress, not radius alone. A trusted final-model study is a legitimate alternative when future adversarial updates are outside the intended use.

Lower ceilings have a measurable evidence cost. With 20 designs and ten protected groups sharing total error .05, even zero observed envelope events require at least 826 retained calibration examples *per group* to pass a 1% ceiling. At 0.1% the requirement rises to 8,290 per group. These are best-case binomial requirements, before curvature or support conservatism. Rare categories require more total iid draws or a justified conditional sampling plan. Selecting an unnamed vulnerable slice after seeing the calibration data is not covered by the original group guarantee.

10 Conclusion

Safety anchors should be judged by which directions they can constrain, which future tasks remain feasible, and which risk statement the evidence supports. Spectral and pointwise certificates expose different failure modes. A useful design may require changing the candidate pool, changing the preservation contract, or declining adaptation—not simply increasing anchor count or radius.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv preprint arXiv:2110.01052*, version 5, 2022. https://arxiv.org/abs/2110.01052v5.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. *arXiv preprint arXiv:2406.11717*, 2024. https://arxiv.org/abs/2406.11717.

A. I. Barvinok. Problems of Distance Geometry and Convex Properties of Quadratic Maps. *Discrete & Computational Geometry*, 13:189–202, 1995. https://doi.org/10.1007/BF02574037.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6), article 43, pp. 1–34, 2021. https://doi.org/10.1145/3478535.

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation Guarantees for Continual Learning with Orthogonal Gradient Descent. *arXiv preprint arXiv:2006.11942*, 2020. https://arxiv.org/abs/2006.11942.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-Optimal Column-Based Matrix Reconstruction. *arXiv preprint arXiv:1103.0995*, 2011. https://arxiv.org/abs/1103.0995.

Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004. https://web.stanford.edu/~boyd/cvxbook/.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient Lifelong Learning with A-GEM. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1812.00420.

Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models. *arXiv preprint arXiv:2503.20807*, 2025. https://arxiv.org/abs/2503.20807.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, 2019. https://arxiv.org/abs/1812.07956.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. https://doi.org/10.1093/biomet/26.4.404.

Chandler Davis and W. M. Kahan. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. https://doi.org/10.1137/0707001.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models. In *International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/29c8c615b3187ee995029284702d3f43-Abstract-Conference.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal Gradient Descent for Continual Learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of Proceedings of Machine Learning Research (PMLR), pp. 3762–3773, 2020. https://proceedings.mlr.press/v108/farajtabar20a.html.

Israel Gohberg, Peter Lancaster, and Leiba Rodman. *Indefinite Linear Algebra and Applications*. Birkhäuser, 2005. https://doi.org/10.1007/b137517.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011. https://arxiv.org/abs/0909.4061.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The Silver Lining of Reducing Safety Risks When Fine-tuning Large Language Models. *arXiv preprint arXiv:2405.16833*, 2024. https://arxiv.org/abs/2405.16833.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. https://arxiv.org/abs/2106.09685.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Advances in Neural Information Processing Systems*, 2024. https://arxiv.org/abs/2402.01109.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Advances in Neural Information Processing Systems*, 2024. https://arxiv.org/abs/2405.18641.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation. *arXiv preprint arXiv:2409.01586*, 2024. https://arxiv.org/abs/2409.01586.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018. https://arxiv.org/abs/1806.07572.

Richard M. Karp. Reducibility among Combinatorial Problems. In *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. https://doi.org/10.1007/978-1-4684-2001-2_9.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming Catastrophic Forgetting in Neural Networks. *arXiv preprint arXiv:1612.00796*, 2017. https://arxiv.org/abs/1612.00796.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. *arXiv preprint arXiv:1706.08840*, 2017. https://arxiv.org/abs/1706.08840.

Guoxin Lu, Letian Sha, Qing Wang, Peijie Sun, Hao Zhou, Hua Dai, and Fu Xiao. Safety Anchor: Defending Harmful Fine-tuning via Geometric Bottlenecks. *arXiv preprint arXiv:2605.05995*, 2026. https://arxiv.org/abs/2605.05995.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. In *Proceedings of the Conference on Learning Theory*, 2009. https://arxiv.org/abs/0907.3740.

Youngjae Min, Benjamin Wright, Jeremy Bernstein, and Navid Azizan. SketchOGD: Memory-Efficient Continual Learning. *arXiv preprint arXiv:2305.16424*, version 3, 2025. https://arxiv.org/abs/2305.16424v3.

Gábor Pataki. On the Rank of Extreme Matrices in Semidefinite Programs and the Multiplicity of Optimal Eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998. https://doi.org/10.1287/moor.23.2.339.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the Safety Landscape: Measuring Risks in Finetuning Large Language Models. *arXiv preprint arXiv:2405.17374*, 2024. https://arxiv.org/abs/2405.17374.

Imre Pólik and Tamás Terlaky. A Survey of the S-Lemma. *SIAM Review*, 49(3):371–418, 2007. https://doi.org/10.1137/S003614450444614X.

Boris T. Polyak. Convexity of Quadratic Transformations and Its Use in Control and Optimization. *Journal of Optimization Theory and Applications*, 99:553–583, 1998. https://doi.org/10.1023/A:1021798932766.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics, SIAM, 2006. https://doi.org/10.1137/1.9780898719109.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.03693.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *arXiv preprint arXiv:2406.05946*, 2024. https://arxiv.org/abs/2406.05946.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation Noising: A Defence Mechanism Against Harmful Finetuning. *arXiv preprint arXiv:2405.14577*, 2024. https://arxiv.org/abs/2405.14577.

Domenic Rosati, Ali Dadsetan, Hong Huang, Xijie Zeng, Hassan Chowdhry, Subhabrata Majumdar, Hassan Sajjad, and Frank Rudzicz. Distribution-Specific Curvature Control with Finite-Sample Guarantees for Open-Weight Safety. *arXiv preprint arXiv:2607.22929*, 2026. https://arxiv.org/abs/2607.22929.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient Projection Memory for Continual Learning. In *International Conference on Learning Representations*, 2021. https://arxiv.org/abs/2103.09762.

Scikit-learn developers. load_digits: Optical Recognition of Handwritten Digits Dataset. Software documentation, accessed September 2026. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. https://doi.org/10.1007/s10208-011-9099-z.

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *arXiv preprint arXiv:1501.01571*, 2015. https://arxiv.org/abs/1501.01571.

Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen. Formalizing Generalization and Adversarial Robustness of Neural Networks to Weight Perturbations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. https://proceedings.neurips.cc/paper/2021/hash/a3ab4ff8fa4deed2e3bae3a5077675f0-Abstract.html.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. https://arxiv.org/abs/1802.04034.

Per-Åke Wedin. Perturbation Bounds in Connection with Singular Value Decomposition. *BIT*, 12:99–111, 1972. https://doi.org/10.1007/BF01932678.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. In *Proceedings of the International Conference on Machine Learning*, volume 235 of PMLR, pp. 52588–52610, 2024. https://proceedings.mlr.press/v235/wei24f.html.

Tsui-Wei Weng, Pu Zhao, Sijia Liu, Pin-Yu Chen, Xue Lin, and Luca Daniel. Towards Certificated Model Robustness Against Weight Perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):6356–6363, 2020. https://doi.org/10.1609/aaai.v34i04.6105.

Yinyu Ye and Shuzhong Zhang. New Results on Quadratic Minimization. *SIAM Journal on Optimization*, 14(1):245–267, 2003. https://doi.org/10.1137/S105262340139001X.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. *arXiv preprint arXiv:2001.06782*, 2020. https://arxiv.org/abs/2001.06782.

Bingjie Zhang, Yibo Yang, Zhe Ren, Dandan Guo, Jindong Gu, Philip Torr, and Bernard Ghanem. A Guardrail for Safety Preservation: When Safety-sensitive Subspace Meets Harmful-resistant Nullspace. *arXiv preprint arXiv:2510.14301*, 2025. https://arxiv.org/abs/2510.14301.

Jiawen Zhang, Yangfan Hu, Kejia Chen, Lipeng He, Jiachen Ma, Jian Lou, Dan Li, Jian Liu, Xiaohu Yang, and Ruoxi Jia. Understanding and Preserving Safety in Fine-Tuned LLMs. *arXiv preprint arXiv:2601.10141*, 2026. https://arxiv.org/abs/2601.10141.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers. *arXiv preprint arXiv:2406.04313*, 2024. https://arxiv.org/abs/2406.04313.

A AVAILABLE-SPAN GEOMETRY: COMPLETE PROOFS

All projectors are Euclidean. An empty matrix has rank zero. A maximum over safe unit vectors is defined as zero when no such vector exists. For a symmetric matrix M, $n_+(M)$ counts its strictly positive eigenvalues, and $M^\dagger$ denotes its Moore–Penrose inverse. Numerical tolerances are implementation choices; the statements below use exact signs.

A.1 GENERALIZED SCHUR REDUCTION

Let $S \in \mathbb{R}^{p\times q}$ have orthonormal columns spanning the candidate-gradient row space, and let T span its orthogonal complement. Write $H = C - \tau I$ and $A = T^\top HT$. Every aggregate constraint matrix $J = FG$ has row space contained in $\mathrm{range}(S)$. Therefore its kernel contains $\mathrm{range}(T)$ and has the unique orthogonal decomposition

$$\ker J = \mathrm{range}(T) \oplus \mathrm{range}(SN_0)$$

for some orthonormal N_0. Conversely, every such subspace is the kernel of aggregate rows: its orthogonal complement lies in $\mathrm{range}(S) = \mathrm{row}(G)$, so a basis of its normals can be expressed as rows of FG.

If $A \not\preceq 0$, a forced-free vector already has Rayleigh quotient above τ, making protection impossible. Suppose $A \preceq 0$. Let Z span $\ker A$. For a candidate direction $v \in \mathrm{range}(S)$, the quadratic form on $t + v$ contains

$$t^\top Ht + 2t^\top Cv + v^\top Hv.$$

Taking $t = \alpha Tz$ for $z \in \ker A$ shows that nonpositivity for every $\alpha \in \mathbb{R}$ requires $z^\top T^\top Cv = 0$. Consequently all additional free directions must lie in

$$\mathcal{E} = \{v \in \mathrm{range}(S) : Z^\top T^\top Cv = 0\}.$$

Choose an orthonormal basis E of $\mathcal{E}$, of dimension h. The coupling condition implies $T^\top CE \in \mathrm{range}(A)$. For any matrix V with orthonormal columns in $\mathbb{R}^h$, completing the square gives

$$\begin{aligned}&\begin{bmatrix}t\\u\end{bmatrix}^\top \begin{bmatrix}A & T^\top CEV\\ V^\top E^\top CT & V^\top E^\top HEV\end{bmatrix}\begin{bmatrix}t\\u\end{bmatrix}\\ &= \big(t + A^\dagger T^\top CEVu\big)^\top A\big(t + A^\dagger T^\top CEVu\big) + u^\top V^\top MVu,\end{aligned}$$

where $M = E^\top HE - E^\top CTA^\dagger T^\top CE$. Thus the free space $\mathrm{range}(T) \oplus \mathrm{range}(EV)$ is safe if and only if $V^\top MV \preceq 0$. This proves the feasible-family characterization, including the singular case. It is an application of the generalized Schur complement, rather than a new block-matrix identity.

A.2 MINIMUM RANK AND TASK UTILITY

A nonpositive subspace for M has dimension at most $h - n_+(M)$, by inertia. A subspace of that dimension exists, and every nonpositive subspace extends to one. For completeness, diagonalize M into positive, negative, and zero blocks. After scaling nonzero eigenvalues, the form is $\|x\|^2 - \|y\|^2$. A nonpositive subspace is a graph $x = Ky$ over a subspace of the negative coordinates, with $\|K\| \le 1$, together with allowed zero coordinates. Extending K by zero on the orthogonal complement of its domain and including all zero coordinates gives a nonpositive subspace of maximum dimension. This is the standard maximal-nonpositive extension theorem of indefinite linear algebra (Gohberg et al., 2005).

The maximum dimension of the free space is $(p-q) + (h - n_+(M))$. Subtracting from p yields the minimum aggregate-constraint rank

$$k_S(\tau) = q - h + n_+(M).$$

For $d \neq 0$, the squared projected task length is

$$\|P_{T\oplus EV}d\|^2 = \|T^\top d\|^2 + \|VV^\top E^\top d\|^2.$$

The second term cannot exceed $b(M, E^\top d)^2$. Indeed, when the projection is nonzero, its normalization is a safe unit vector with inner product equal to its norm. Conversely, a maximizing safe

unit vector extends to a maximum-dimensional nonpositive subspace. That extension cannot lower its task projection; the same upper bound then forces equality. This proves both the utility ceiling and its attainment at rank $k_S(\tau)$. If there are no additional safe vectors, the T component alone is retained. If T is empty, the result reduces to the unrestricted oracle. If $G = 0$, all directions are forced free and feasibility is exactly $C \preceq \tau I$.

A.3 Task families: what extends and what fails

Let $\mathcal{D}$ be any set of nonzero task directions, and let its utility be $\inf_{d\in\mathcal{D}} \|Pd\|/\|d\|$. Extending a safe free space to a maximum-dimensional one increases its projector in positive-semidefinite order. Hence *every* task's utility weakly increases. Therefore, if any aggregate design satisfies specified lower bounds for a task family, one of minimum rank $k_S(\tau)$ does so. This rank-extension statement does not imply that the separate one-task utility maxima can be attained together.

For an explicit separation, take $C = \mathrm{diag}(1, 0)$, $\tau \in (0, 1)$, and one allowed free line $v = (\sin\alpha, \cos\alpha)$. Safety requires $\sin^2\alpha \le \tau$. For tasks e_1, e_2, the respective individual ceilings are $\sqrt{\tau}$ and 1, whereas

$$\max_{v_1^2\le\tau} \min(|v_1|, |v_2|) = \sqrt{\min(\tau, 1/2)}.$$

At $\tau = 1/4$, the target pair $(1/2, 0.9)$ is infeasible because the squared task projections sum to one. For the *whole* two-dimensional task subspace, every rank-one free space has worst-task utility zero. More generally, for an orthonormal task basis D, its uniform utility is $\sqrt{\lambda_{\min}(D^\top PD)}$. A finite task bank and its full span are consequently different utility contracts.

A.4 Literal anchors, aggregate anchors, and discovery

Literal selection uses $J = G_I$ for an index subset I. Its feasible projectors form a subset of the available-span family. Therefore the available-span ceiling is an upper bound, not a realization guarantee for literal selection. If $\|J\Delta\| \le \zeta$ for literal rows, each selected scalar margin's linear change is at most ζ. For $J = FG$, the constraint bounds weighted combinations of those changes. For instance, $G = I_2$, $F = (1, 1)$ permits $\Delta = (t, -t)$ and arbitrarily opposite individual margin changes subject only to the radius. Aggregate rows must be described and enforced as aggregate statistics. Normalizing or whitening them changes the meaning of a numerical tolerance.

Low rank alone does not imply that literal gradients approximate prescribed normals. For a direction uniform on the unit sphere in $\mathbb{R}^p$, the probability of lying within Euclidean distance $\eta < \sqrt{2}$ of either sign of a fixed unit vector equals

$$I_{1-(1-\eta^2/2)^2}\big((p-1)/2, 1/2\big),$$

where I is the regularized incomplete beta function. For fixed small η this decays exponentially with p. The available-span theorem avoids this cap assumption for aggregates, but does not remove the literal-selection gap.

For orthogonal categorical gradients with probabilities $p_1, \ldots, p_r$, all-mode discovery after n iid draws has exact probability

$$\sum_{A\subseteq[r]} (-1)^{|A|} \left(1 - \sum_{j\in A} p_j\right)^n.$$

It is at most $1 - (1 - p_j)^n$ for every required mode j. This is a bound on the population probability, not a deterministic upper bound on a finite Monte Carlo proportion.

B Global-update certificates and operational thresholds

The distribution ν in this appendix is fixed before calibration. It may be the protected distribution, the distribution conditional on a positive base margin, or the latter distribution further conditioned on a predeclared group. All second moments and probabilities in a single certificate use the same ν. We assume measurability and finite second moments. A zero displacement has zero flip risk under the base-correct conditional distribution.

B.1 Spectral nonlinear gate: proof of Theorem 1

Let $r_x(\Delta) = m_{\theta_0+\Delta}(x) - m_0(x) - g(x)^\top \Delta$. A Hessian operator-norm bound L throughout the radius-ρ ball implies

$$|r_x(\Delta)| \le L \|\Delta\|^2 / 2 \le L\rho^2/2.$$

On the event $m_0(x) > \gamma$, a violation implies $|g(x)^\top \Delta| + L\rho^2/2 \ge \gamma$. For each fixed permitted global displacement, Markov's inequality and the triangle inequality in $L_2(\nu)$ give

$$\begin{aligned}
\nu(m_{\theta_0+\Delta} \le 0) &\le b + \frac{\mathbb{E}_\nu(|g^\top \Delta| + L\rho^2/2)^2}{\gamma^2} \\
&\le b + \frac{(\sqrt{\Delta^\top C \Delta} + L\rho^2/2)^2}{\gamma^2} \\
&\le b + \frac{(\rho\sqrt{\kappa_{\overline{C}}(P)} + L\rho^2/2)^2}{\gamma^2}.
\end{aligned}$$

Here $b \ge \nu(m_0 \le \gamma)$ and $C \preceq \overline{C}$. The last bound is independent of Δ, so taking its supremum requires no union bound over displacements and no interchange of expectation and supremum. In particular, the trace of PCP is unnecessary. If b and $\overline{C}$ are simultaneously valid confidence bounds, the deterministic reasoning holds on their common event, even if P was chosen using those data. Allocate failure probability to both bounds before use.

Solving the displayed sufficient inequality for $\kappa_{\overline{C}}(P)$ yields the threshold in Equation (4). A negative numerator before squaring means no positive spectral allowance; it must not be made positive by squaring. Equality at zero requires zero linear drift. If $\rho = 0$, handle the unchanged model directly. Failure to meet this sufficient gate is not a lower bound on actual risk.

Imperfect enforcement. For $\mathcal{K}(J, \rho, \zeta)$ define

$$W(\overline{C}, J, \rho, \zeta) = \sup_{\Delta \in \mathcal{K}} \Delta^\top \overline{C} \Delta.$$

Replacing $\rho\sqrt{\kappa}$ by $\sqrt{W}$ in the proof gives

$$R_\nu(\mathcal{K}) \le \min\left\{1, b + \frac{(\sqrt{W} + L\rho^2/2)^2}{\gamma^2}\right\}. \tag{10}$$

For positive budgets, W equals the scalar dual in Appendix F, a specialization of homogeneous two-constraint quadratic optimization (Ye and Zhang, 2003; Polyak, 1998). This is an exact quadratic-drift calculation, not an exact nonlinear violation probability. Alternatively, a bound ℓ on the L_2 contribution of the constrained component gives the cheaper numerator $\rho\sqrt{\kappa} + \ell + L\rho^2/2$. Neither form requires a trace in the free component.

B.2 A statistical upper matrix without ignoring unseen directions

For bounded gradients $\|g\| \le R$, set $Y = gg^\top - C$. Since $(gg^\top)^2 = \|g\|^2 gg^\top$, we have

$$\mathbb{E}Y^2 = \mathbb{E}(gg^\top)^2 - C^2 \preceq R^2 C, \qquad \|Y\| \le R^2.$$

Matrix Bernstein therefore gives the sufficient operator error

$$a_n(c) = \sqrt{\frac{2R^2 c \log(2p/\delta)}{n}} + \frac{2R^2 \log(2p/\delta)}{3n}, \qquad c \ge \|C\|, \tag{11}$$

with probability at least $1 - \delta$ (Tropp, 2012; 2015). The linear term is a conservative inversion of the Bernstein tail. This uses variance $R^2\|C\|$, not R^4. To obtain an observable bound, let $\hat{c} = \left\|\widehat{C}\right\|$, $a = \sqrt{2R^2 \log(2p/\delta)/n}$, and $b_0 = 2R^2 \log(2p/\delta)/(3n)$. On the same event the unknown error e obeys $e \le a\sqrt{\hat{c} + e} + b_0$. Its positive-root upper bound is

$$e \le b_0 + \frac{a^2 + \sqrt{a^4 + 4a^2(\hat{c} + b_0)}}{2}.$$

Consequently $\overline{C} = \widehat{C} + eI$ is an observable positive semidefinite (PSD) upper matrix. Intrinsic-dimension versions can reduce the logarithmic factor when an effective-rank bound is available; they do not remove the $R^2/\|C\|$ scaling. In the unit-sphere experiment, $R^2 = 1$ while $\|C\| = 1/p$.

A reduced estimate must also control its unobserved tail. Let V be an independently fixed orthonormal q-column basis, $z = V^\top g$, and $r = (I - VV^\top)g$. Suppose $\mathbb{E}zz^\top \preceq \overline{C}_z$ and $\mathbb{E}\|r\|^2 \le \bar{v}$ hold jointly. For every $a > 0$ and every vector u, Young's inequality gives

$$\mathbb{E}(u^\top g)^2 \le (1 + 1/a)u^\top V\overline{C}_z V^\top u + (1 + a)\bar{v}\|u\|^2.$$

Thus

$$C \preceq (1 + 1/a)V\overline{C}_z V^\top + (1 + a)\bar{v}I. \tag{12}$$

A q-dimensional matrix bound and a scalar tail-energy bound suffice. Choosing a after obtaining the common event is valid because the inequality holds for every $a > 0$ on that event. Simply treating the span of observed gradients as the entire population support is not valid. Scalar tail energy can be bounded by bounded-variable concentration or empirical Bernstein (Maurer and Pontil, 2009). This reduced-covariance bound is a theoretical result; it has not been evaluated on large models.

For the known signed one-hot model, $g = se_j$, C is diagonal with entries p_j. Simultaneous one-sided binomial bounds at level δ/p give a diagonal $\overline{C}$. This exploits known population support and is not a procedure for arbitrary Jacobians. The geometry pipeline uses $n = 65{,}536$ and the rare-slice pipeline $n = 100{,}000$; the sphere study instead estimates a full covariance.

B.3 The pointwise quantifier gap

For a fixed region, let $e(x) = \mathbf{1}\{\exists\Delta \in \mathcal{K} : m_{\theta_0+\Delta}(x) \le 0\}$. Always

$$\sup_{\Delta\in\mathcal{K}} \nu(m_{\theta_0+\Delta} \le 0) \le \mathbb{E}_\nu e.$$

The left side chooses one model before the next protected input; the right side can choose a different model for every input. This gap is distinct from the approximation of e and from finite-sample calibration.

For a radius-ρ ball in an r-dimensional free space, constant margin γ, zero curvature, and isotropic Gaussian projected gradients $N(0, \sigma^2 I_r)$, rotational invariance gives

$$R_{\text{global}} = \Phi(-t), \qquad R_{\text{envelope}} = \Pr(\chi_r \ge t), \qquad t = \gamma/(\rho\sigma).$$

The first follows from the worst update norm ρ; the second follows by minimizing the linear margin separately at each example. At $t = c\sqrt{r}$, $0 < c < 1$, a Gaussian Chernoff bound gives $R_{\text{global}} \le e^{-c^2 r/2}$. A chi-square lower-tail Chernoff calculation gives

$$R_{\text{envelope}} \ge 1 - \exp\{-r(c^2 - 1 - \log c^2)/2\}.$$

To derive the latter, apply Markov to $e^{-a\chi_r^2}$, use its moment generating function $(1 + 2a)^{-r/2}$, and optimize $a = (c^{-2} - 1)/2$. Thus calibration alone cannot close an exponentially large gap.

For unit-sphere gradients in dimension p, the exact global risk is

$$\tfrac{1}{2}\Pr\{\text{Beta}(1/2, (p-1)/2) \ge (\gamma/\rho)^2\}$$

when $0 < \gamma < \rho$. The exact pointwise envelope is one, whereas the population spectral bound is $\rho^2/(p\gamma^2)$. These quantities have no curvature or singular-value approximation. The finite-data experiment measures the additional matrix-estimation cost separately.

C Preservation calibration and targeted failure

C.1 Constraint-aware support and Taylor envelope

Let Q_J span $\text{row}(J)$, $P = I - Q_J Q_J^\top$, and s be the smallest nonzero singular value. For $\Delta \in \mathcal{K}$ its constrained length is at most $z = \min(\rho, \zeta/s)$. Decompose g into lengths $a = \|Pg\|$ and

$b = \left\| Q_J^\top g \right\|$. At a fixed constrained length w, Cauchy–Schwarz bounds the adverse linear drift by $a\sqrt{\rho^2 - w^2} + bw$. Maximizing over $0 \le w \le z$ gives

$$w_* = \min\{z, \rho b / \|g\|\}, \qquad h = a\sqrt{\rho^2 - w_*^2} + bw_*.$$

For $g = 0$ set $h = 0$; for no constraints set $h = \rho \|g\|$. This is exact for the relaxed constrained-component ball; it can exceed the support of $\|J\Delta\| \le \zeta$ when singular values differ. A Hessian bound $L(x)$ on the entire radius ball yields the preservation envelope

$$e_{\rho,J}(x) = \mathbf{1}\{0 < m_0(x) \le h_{\rho,J}(x) + L(x)\rho^2/2\}.$$

Among base-correct examples, every actual flip for every permitted displacement is contained in this event. With a nondecreasing bound $L(x, \rho)$, the envelope is monotone in ρ. Exact support also is monotone because the regions are nested for a fixed (J, ζ). The smallest-singular-value relaxation preserves that nesting.

C.2 Conditional sampling and fixed-sequence testing: proof of Theorem 3

A base model, groups, and design chains are fixed independently of iid calibration data. In a group, retain only samples with $m_0 > 0$. Conditional on their count N, these samples are iid from $\nu = Q(\cdot \mid m_0 > 0, \text{group})$. For each radius the envelope count is binomial under that conditional distribution. Empty retained samples cannot certify a positive radius.

Fix a target ϵ before testing. Each design chain tests radii from smallest to largest at its allocated error level α_c, accepting while its binomial test rejects $H_{0,r} : \mathbb{E}_\nu e_r > \epsilon$, and stopping at its first failure. An equivalent implementation accepts when the ordinary one-sided Clopper–Pearson (CP) endpoint at level α_c is at most ϵ (Clopper and Pearson, 1934). If any false rejection occurs in the chain, its first true null is rejected. The probability of rejecting that particular true null is at most α_c, regardless of dependence between tests. Therefore a union bound across chains with $\sum_c \alpha_c \le \delta$ controls the familywise error. This is the fixed-sequence Learn-then-Test argument (Angelopoulos et al., 2022), applied to nested anchor-region envelopes. It is also consistent with risk-controlling prediction-set methodology (Bates et al., 2021).

On this event all accepted regions have population flip risk at most the predeclared ϵ; selection among their radii and designs is valid. The same deterministic containment permits later adaptive update trajectories inside the selected region. Groupwise allocations give the corresponding guarantee for each named group. If no radius passes, leaving the base model unchanged has exactly zero flip risk.

What is and is not a confidence bound. Fixed-sequence testing certifies the chosen risk *at* ϵ. The numerical Clopper–Pearson endpoint at a data-selected radius is not automatically a simultaneous or selection-valid confidence curve. We therefore report ϵ as the deployed upper bound and retain endpoints only as test diagnostics. Bonferroni across all radii produces a simultaneous numerical curve, at a statistical cost. Testing several risk ceilings or selecting groups after calibration requires additional correction or fresh data. The experiment uses a single primary ceiling, 2%.

Calibration cost for groups. At zero observed envelope events, acceptance at level α_c requires

$$N \ge \left\lceil \frac{\log \alpha_c}{\log(1 - \epsilon)} \right\rceil.$$

This is a requirement per retained base-correct group. Rare groups require a correspondingly larger total sample, or valid conditional sampling. Group guarantees do not cover arbitrary, subsequently chosen slices. Risk at most ϵ on average can coexist with failure on every example of a group of mass below ϵ.

C.3 A forced-free targeted-failure witness

Let T span the orthogonal complement of the candidate-gradient row space. Suppose on a protected slice A of mass $w > 0$ that $m_0(x) \le M$, $\|g(x) - \mu\| \le \eta$, and the Hessian operator norm is at most L throughout the radius-ρ ball. Put $b = \|P_T \mu\| > 0$. If

$$M - \rho(b - \eta) + L\rho^2/2 \le 0, \tag{13}$$

then every aggregate design based on that pool allows a single displacement causing a violation on all of A.

Indeed, choose $\Delta_* = -\rho P_T \mu / b$. It is in $\ker G$, hence in $\ker(FG)$ for every aggregate matrix F, and has norm ρ. On the slice,

$$g(x)^\top \Delta_* \le -\rho b + \rho\eta, \qquad m_{\theta_0+\Delta_*}(x) \le M - \rho(b-\eta) + L\rho^2/2 \le 0.$$

Thus $R_Q(\mathcal{K}) \ge w$ and the slice-conditional risk is one. This is an actual risk lower bound, unlike merely failing a sufficient spectral gate. It applies to soft constraints as well, because the witness has zero anchor residual. It requires a coherent slice, not just large covariance. In the categorical experiment, $M = 0.25$, $\rho = 0.3$, $b = 1$, $\eta = 0$, and $L = 0.02$, giving a strictly negative upper margin of approximately -0.05.

D Task banks, mean drift, and stable pool decisions

D.1 Computing a finite task-bank frontier

For normalized tasks u_i, a rank-one free space has squared utility $\min_i (u_i^\top v)^2$. Its semidefinite relaxation is

$$\max_{X \succeq 0, t} \; t \quad \text{s.t.} \quad \operatorname{tr} X = 1, \quad \operatorname{tr}(CX) \le \tau, \quad u_i^\top X u_i \ge t \; (\forall i). \tag{14}$$

Every feasible line yields a feasible $X = vv^\top$. A rank-one optimal X certifies exactness; fractional solutions only upper-bound achievable utility. Standard semidefinite duality (Boyd and Vandenberghe, 2004) gives the readily computed upper bound

$$\inf_{w \in \Delta_m, \eta \ge 0} \left[\lambda_{\max}\left(\sum_i w_i u_i u_i^\top - \eta C \right) + \eta\tau \right].$$

Any feasible w, η gives an upper bound, even if the numerical optimizer stops early. Unlike the one-task quadratic program, this SDP need not admit a rank-one optimum.

For $C = \operatorname{diag}(2, 0)$, $\tau = 1$, and tasks $e_1, e_2, (e_1+e_2)/\sqrt{2}, (e_1-e_2)/\sqrt{2}$, the exact squared max–min is $(1 - 1/\sqrt{2})/2$. To see this, write a feasible unit vector at angle $\theta \in [\pi/4, 3\pi/4]$ modulo sign; the nearest midpoints between the four equally spaced task lines maximize their minimum squared projection, giving $\theta = 3\pi/8$ or $5\pi/8$. The SDP instead achieves $1/2$ with $X = I/2$; the first two task constraints and $\operatorname{tr} X = 1$ show this is optimal. The implementation also checks a 100,001-point angle grid.

For a specified free dimension m, the general projector problem admits the relaxation

$$\max t: \quad 0 \preceq P \preceq I, \quad \operatorname{tr} P = m, \quad PT = T, \quad u_i^\top P u_i \ge t, \quad \begin{bmatrix} \tau I & PC^{1/2} \\ C^{1/2}P & I \end{bmatrix} \succeq 0.$$

For a projector the last condition is exactly $PCP \preceq \tau I$. A projector-valued optimum is an exact certificate. Otherwise it is only an upper bound; feasible subspaces supply lower bounds. The available-span theorem reduces the search dimension and allows extension to the maximum safe free dimension, but does not solve every bank optimization.

A precise hardness statement. With a prescribed one-dimensional free space, deciding finite-bank feasibility is NP-hard. Reduce PARTITION (Karp, 1972) with positive integer weights $a_1, \dots, a_p$ to $C = aa^\top$, $\tau = 0$, tasks e_i, and required squared utilities $1/p$. A unit vector meeting every task constraint has $v_i^2 \ge 1/p$; summing forces $v_i = s_i/\sqrt{p}$, $s_i \in \{-1, 1\}$. Its safety constraint is $(a^\top v)^2 = 0$, exactly the partition equation $\sum_i a_i s_i = 0$. All input quadratic data are rational. This proves NP-hardness for the prescribed-line formulation, not for the distinct problem in which any larger safe free space is permitted.

The relaxation can even accept an infeasible such bank. With $a = (4, 3, 2)$ no signed sum is zero. Let unit vectors $q_1 = (1, 0)$, $q_2 = (-7/8, \sqrt{15}/8)$, and $q_3 = -(4q_1 + 3q_2)/2$. Their Gram matrix divided by three has diagonal $1/3$, trace one, and $a^\top X a = 0$, so it is a feasible fractional witness. Numerical tests check both this witness and the discrete reduction.

D.2 Exact scalar representation for Gaussian mean drift

The following equality sharpens treating Young's inequality only as a sufficient condition. Let $A, B \succeq 0$ on a fixed nonzero free space and set

$$H^2 = \max_{\|v\|=1} \left(\sqrt{v^\top A v} + \sqrt{v^\top B v}\right)^2.$$

Then

$$H^2 = \inf_{\lambda>0} \lambda_{\max}\big((1+1/\lambda)A + (1+\lambda)B\big). \tag{15}$$

If either matrix is zero, the appropriate boundary limit proves equality. Otherwise the scalar objective is convex and diverges at both ends, so it has an interior minimum λ_*. Let E be its top eigenspace. The one-sided derivatives of a largest eigenvalue are the extreme Rayleigh quotients of the derivative restricted to E. At an interior minimum, zero lies between them. Because a Rayleigh quotient on E takes every value between its extreme eigenvalues, there is a unit $v \in E$ with

$$v^\top(-A/\lambda_*^2 + B)v = 0.$$

For that vector, Young's inequality is an equality: $v^\top A v = \lambda_*^2 v^\top B v$. Its squared sum of square roots therefore equals the scalar minimum. The reverse inequality holds for every vector by Young's inequality. This proves the identity, including repeated top eigenvalues. The argument uses standard convex spectral variation, not an assumption that a numerical eigensolver chooses the needed eigenvector.

For the affine model $m_\Delta = \gamma + g^\top \Delta$ with $g \sim N(\mu, \Sigma)$, $\gamma > 0$, and Σ positive definite on the free space, put $A = (P\mu)(P\mu)^\top$ and $B = z_{1-\epsilon}^2 P\Sigma P$, restricted to that space. Equation (15) gives

$$\sup_{\|v\|=1, v=Pv} \{|\mu^\top v| + z_{1-\epsilon}\sqrt{v^\top \Sigma v}\}^2 = \inf_{\lambda>0} \kappa_{C_\lambda}(P), \quad C_\lambda = (1+1/\lambda)\mu\mu^\top + (1+\lambda)z_{1-\epsilon}^2\Sigma.$$

Thus the *continuous union* of the available-span spectral feasible families at threshold γ^2/ρ^2 gives the exact noncentered Gaussian design family, with one boundary case. If $P\mu = 0$, the infimum may occur only as $\lambda \downarrow 0$; handle this family directly with $P\mu = 0$ and $z^2\kappa_\Sigma(P) \le \gamma^2/\rho^2$. When aggregate normals must lie in S, that family requires $\mu \in S$ unless $\mu = 0$. Impose the mean constraint and apply the available-span theorem in $\mu^\perp$, counting its rank cost. A finite λ grid searches a sufficient subset of the continuous family, not the exact optimum. Minimum rank and maximum utility across different λ values need not occur at the same value; feasible rank–utility pairs are the union of the per-λ frontiers.

Independent numerical check. The support also equals $\max_{\|w\|\le 1} \left\|\mu + z\Sigma^{1/2}w\right\|$. In the eigenbasis of $B = z^2\Sigma$, let λ_i be eigenvalues and $b_i = \sqrt{\lambda_i}(U^\top\mu)_i$. Its square is the trust-region dual

$$\inf_{t \ge \lambda_{\max}(B)} \left[\|\mu\|^2 + t + \sum_i \frac{b_i^2}{t - \lambda_i}\right],$$

with the usual range condition at a singular endpoint. We implement both this secular equation and the unrelated scalar largest-eigenvalue minimization. The 243-case study includes zero means and singular hard cases. It checks implementation, while the proof above establishes the identity.

D.3 Pool floors, randomized range finding, and perturbation stability

For matched Gaussian proposals $g = C^{1/2}\omega$, $\omega \sim N(0, I)$, a pool of q columns is $Y = C^{1/2}\Omega$. Its forced-free floor is exactly

$$f_S(C) = \left\|(I - P_Y)C^{1/2}\right\|_{\mathrm{op}}^2.$$

With $q = k + \ell$ and $\ell \ge 2$, the classical randomized range-finder bound gives

$$\mathbb{E}\sqrt{f_S(C)} \le \left(1 + \sqrt{\frac{k}{\ell - 1}}\right)\sqrt{\lambda_{k+1}(C)} + \frac{e\sqrt{k+\ell}}{\ell}\sqrt{\sum_{j>k}\lambda_j(C)}. \tag{16}$$

This is Theorem 10.6 of Halko et al. (2011) applied to $C^{1/2}$. The bound is on the expected *square root* of the floor; squaring its expectation does not upper-bound the expected floor. For example Markov's inequality gives $\Pr(f_S > \tau) \le \mathbb{E}\sqrt{f_S}/\sqrt{\tau}$, a valid but often loose pool-size criterion. Small tails make oversampling useful; heavy tails require more. A mismatched proposal that omits a deployment-relevant mode need not benefit from any number of additional draws. This distinction motivates the targeted expansion experiment rather than claiming Gaussian range finding as new theory.

For a fixed pool, f_S is 1-Lipschitz in operator norm. More generally, $|\kappa_C(P) - \kappa_{C'}(P)| \le \|C - C'\|$ for every projector. Hence if $\|C - C'\| \le \eta$, the feasible families obey

$$\mathcal{F}_{C'}(\tau - \eta) \subseteq \mathcal{F}_C(\tau) \subseteq \mathcal{F}_{C'}(\tau + \eta).$$

The corresponding utility ceilings are bracketed in the same order, and minimum ranks in the reverse order. Separated eigenspaces admit the familiar perturbation controls of Davis and Kahan (1970); Wedin (1972). At a singular Schur boundary, ranks and feasibility may jump; the reduced inverse need not vary smoothly. For instance $C = \left[\begin{smallmatrix} 1 & .2 \\ .2 & .5 \end{smallmatrix}\right]$, pool normal e_2, and $\tau = 1$ leave e_1 forced free. An arbitrarily small positive perturbation of C_{11} makes the threshold infeasible. Robust decisions should use an upper matrix or a threshold margin, not a pseudoinverse computed from an uncertain zero eigenvalue.

E A GLOBAL CURVATURE BOUND WITH MULTIPLE HIDDEN LAYERS

Consider $h_0 = x$, $h_l = \tanh(W_l[h_{l-1}; 1])$ for $l = 1, \dots, D$, and $f = a^\top h_D + b$ with frozen readout. All entries of all hidden affine matrices are trainable. The Euclidean parameter norm is their joint Frobenius norm. We derive a valid bound on the Hessian throughout $\sum_l \left\|W_l - W_l^0\right\|_F^2 \le \rho^2$.

For a unit parameter direction, let r_l be its layerwise Frobenius norms, so $\sum_l r_l^2 = 1$. Put $c_2 = 4/(3\sqrt{3})$, $s_l = \left\|W_l^0\right\|_{\mathrm{op}} + \rho$, $A_0 = \sqrt{\|x\|^2 + 1}$, $v_0 = 0 \in \mathbb{R}^D$, and $K_0 = 0$. Recursively define

$$\begin{aligned} u_l &= s_l v_{l-1} + A_{l-1} e_l, & v_l &= u_l, \\ K_l &= c_2 u_l u_l^\top + e_l v_{l-1}^\top + v_{l-1} e_l^\top + s_l K_{l-1}, & A_l &= \sqrt{\mathrm{width}(h_l) + 1}. \end{aligned} \tag{17}$$

Then a global Hessian bound is

$$L_D(x, \rho) = \|a\|_2 \lambda_{\max}(K_D). \tag{18}$$

The readout norm is its Euclidean norm.

To prove it, differentiate twice along the unit direction. With $z_l = W_l[h_{l-1}; 1]$, the first derivative obeys $\|z_l'\| \le s_l \left\|h_{l-1}'\right\| + r_l A_{l-1} \le u_l^\top r$. The second derivative satisfies $\|z_l''\| \le s_l \left\|h_{l-1}''\right\| + 2 r_l \left\|h_{l-1}'\right\|$. Since $|\tanh'| \le 1$, $|\tanh''| \le c_2$, and $\|z_l' \odot z_l'\|_2 \le \|z_l'\|_2^2$, induction gives $\|h_l''\| \le r^\top K_l r$. Finally $|f''| \le \|a\|_2 r^\top K_D r \le \|a\|_2 \lambda_{\max}(K_D)$. The proof accounts for cross-layer terms; it does not rely on a block-diagonal full Hessian.

For one hidden layer a sharper architecture-specific bound is $L(x) = c_2 \max_j |a_j| (1 + \|x\|^2)$. We use that tighter formula in the main shallow experiments, and the same compositional formula across every depth in the depth study. At width 128 with $a_j = 1/\sqrt{128}$ and unit-norm inputs, the correct shallow expression is $8/(3\sqrt{3}\sqrt{128}) \simeq 0.14$. The depth study trains one-, two-, and three-hidden-layer models and calibrates nine radius choices for each. It measures the useful-radius collapse of a valid bound rather than substituting local sampled curvature for the global assumption. No corresponding useful bound for an attention or transformer block is established here.

F CLASSICAL BENCHMARKS AND SYSTEMATIC DRIFT

F.1 THE UNRESTRICTED LOCAL BENCHMARK

For $C \succeq 0$, a projector P onto the free-update space satisfies

$$\sup_{\Delta = P\Delta, \|\Delta\| \le \rho} \mathbb{E}(g^\top \Delta)^2 = \rho^2 \lambda_{\max}(PCP).$$

This is the Rayleigh variational identity. With a nonzero task direction d, maximizing its first-order progress over the same ball gives $\rho\|Pd\|$. Let

$$B(C, d, \tau) = \max_{\|v\|=1,\ v^\top Cv \le \tau} \frac{|d^\top v|}{\|d\|}.$$

When the safe unit-vector set is empty, define $B = 0$. Inertia requires at least $n_+(C - \tau I)$ scalar constraints. The maximal-nonpositive extension argument in Appendix A shows that, for one task, this rank and the ceiling B can be attained simultaneously. This is an optimality benchmark obtained from classical linear algebra, not a new preservation mechanism. The available-span result identifies the extra restrictions imposed by actual candidate gradients.

For completeness, writing $u = d/\|d\|$, the standard rank-one semidefinite relaxation gives

$$B^2 = \max_{X \succeq 0, \mathrm{tr}\, X = 1, \mathrm{tr}(CX) \le \tau} u^\top X u = \inf_{\eta \ge 0} \{\lambda_{\max}(uu^\top - \eta C) + \eta\tau\}.$$

For an interior feasible threshold, strong conic duality holds and a rank-one optimum exists by the two-constraint rank bound (Barvinok, 1995; Pataki, 1998; Ye and Zhang, 2003). At the minimum-eigenvalue boundary, restrict to that eigenspace directly. A subthreshold has no safe unit vector. These boundary conventions are implemented explicitly.

If $u^\top Cu > \tau$ and $B > 0$, every maximizing v has $v^\top Cv = \tau$. Otherwise it is interior relative to the sphere and a local maximum of $|u^\top v|$ must be $v = \pm u$, contradicting the unsafe-task assumption. A subspace attaining utility B has $Pd/\|Pd\|$ as one such maximizer. Its task-progress-maximizing update therefore attains the subspace's Rayleigh threshold. This explains why a robust subspace need not improve the particular benign update over a trusted pointwise optimizer.

In a two-band model $C = \mathrm{diag}(hI_r, \ell I_{p-r})$, $\ell \le \tau < h$, write the task's high-space length as c. Put $a = (\tau - \ell)/(h - \ell)$. Then $B = 1$ if $c^2 \le a$ and otherwise

$$B = c\sqrt{a} + \sqrt{1 - c^2}\sqrt{1 - a}.$$

PCA protection retains $\sqrt{1 - c^2}$, which is lower in the latter case. This comparison concerns the most useful protected *subspace*; it is not a superiority statement for benign single-step learning.

F.2 Approximate enforcement as a classical quadratic optimization problem

Let $C \succeq 0$, $\rho > 0$, $\zeta > 0$. Homogeneous quadratic optimization with the two inequalities $\|\Delta\|^2 \le \rho^2$ and $\Delta^\top J^\top J\Delta \le \zeta^2$ admits an exact SDP relaxation (Ye and Zhang, 2003, Section 2.2); see also Polyak (1998); Pólik and Terlaky (2007). Applying that result and eliminating the norm-constraint multiplier yields

$$\max_{\|\Delta\| \le \rho, \|J\Delta\| \le \zeta} \Delta^\top C\Delta = \min_{b \ge 0} \{\rho^2[\lambda_{\max}(C - bJ^\top J)]_+ + b\zeta^2\}.$$

To see the scalar elimination, the SDP dual minimizes $a\rho^2 + b\zeta^2$ subject to $a, b \ge 0$ and $aI + bJ^\top J \succeq C$. At a fixed b, its smallest feasible a is exactly the displayed positive eigenvalue part. The SDP is compact and has a rank-one extreme optimum: on a rank-r face, at most two active scalar constraints leave a nonzero perturbation when $r(r+1)/2 > 2$. Strong duality follows from positive budgets. If $\rho = 0$, the value is zero. If $\zeta = 0$, restrict directly to $\ker J$; the same scalar expression may have only an infimum rather than a finite minimizer. These are application formulas, not separate general optimization advances.

Conditioning enters through the smallest nonzero singular value. In two dimensions, rows $e_1^\top$ and $(\cos\alpha, \sin\alpha)$ span the same space for every $\alpha > 0$ but become nearly redundant as $\alpha \downarrow 0$. Updates along the weak right singular vector can have length $\min(\rho, \zeta/\sigma_{\min}(J))$. Thus the soft-constraint guarantee can deteriorate while row count, rank, row norms, and exact-kernel coverage remain unchanged.

F.3 Noncentered Gaussian threshold

In the affine margin model, suppose $g \sim N(\mu, \Sigma)$, $m_0 = \gamma > 0$, and Σ is positive definite on the nonzero free space L. For a displacement Δ, the exact risk is

$$\Phi\left(\frac{-\gamma - \mu^\top \Delta}{\sqrt{\Delta^\top \Sigma \Delta}}\right).$$

For a unit vector $v \in L$, choose the sign making the mean drift adverse. At displacement length $r > 0$, the standardized threshold becomes $(|\mu^\top v| - \gamma/r)/\sqrt{v^\top \Sigma v}$, which increases with r. Therefore the worst sign and radius are $-\operatorname{sign}(\mu^\top v)$ and ρ. For $\epsilon \in (0, 1/2)$, monotonicity of Φ proves

$$\sup_{\Delta \in L, \|\Delta\| \le \rho} Q(m_0 + g^\top \Delta \le 0) \le \epsilon \iff \rho \sup_{\substack{v \in L \\ \|v\|=1}} \left(|\mu^\top v| + z_{1-\epsilon} \sqrt{v^\top \Sigma v} \right) \le \gamma.$$

The zero space has risk zero. For covariance-null directions the deterministic condition must instead be strict, $\rho|\mu^\top v| < \gamma$, since a zero margin counts as a violation. The experiments use positive-definite covariance.

Removing the mean direction is not always optimal. With $\mu = 0.1e_2$, $\Sigma = \operatorname{diag}(4, 0.01)$, $\rho = \gamma = 1$, protecting the mean leaves e_1 free and risk $\Phi(-0.5) \approx 0.31$. Protecting e_1 instead leaves worst risk $\Phi(-9)$, despite its nonzero mean. Which direction matters depends on both the mean and covariance. The second moment $C = \Sigma + \mu\mu^\top$ still supports a valid second-moment bound, but it does not determine exact risk alone.

F.4 Total displacement and changing tasks

A per-step radius is not a deployment-region constraint. In a one-dimensional linear example with $m_0 = 1$, $g = -1$, and successive steps $\delta_t = r > 0$, each step can be small while the total displacement tr eventually violates the margin. A fixed-origin radius/enforcement contract avoids this failure. The certificate in Theorem 3 is uniform over every allowed displacement, so its safety statement survives changes in task direction, choice of optimizer, and adaptive stopping. The local utility diagnostics need not survive those changes: they describe the task bank at the base point. A new, unanticipated task can be constrained to near-zero progress even while the risk guarantee remains valid.