# SHRINKAGE IN ZEROTH-ORDER OPTIMIZATION: ESTIMATION RISK AND EXPECTED DESCENT

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

## ABSTRACT

Reducing the estimation error of a zeroth-order (ZO) update need not improve optimization after its step size is chosen. We study this distinction for anchored ridge shrinkage under exact finite batches of random directions. Optimally rescaled Euclidean risk and optimized isotropic quadratic descent rank updates identically. For iid sphere updates, an explicit Stein kernel gives risk bounds at a specified nonzero shrinkage strength. With the direction count fixed at $m \geq 3$ and dimension sufficiently large, the same zero-anchor rule has certified risk reduction and strictly worse optimized descent. We quantify this relative loss and give explicit finite-dimensional enclosures. The nonlinear coefficient remains random in the limit, making effective steps vary across updates. In a specified curvature family with bounded anchors, the rule can beat Raw while remaining asymptotically inferior to the fixed mixture $(X + a)/2$. Independent simulations test these predictions. At frozen OPT-125M checkpoints, expanded step searches yield interior fitting-loss minima in all forty selections, yet the nonlinear rule has no consistent advantage. Independent profiling predicts some coefficient effects at small steps, with reduced predictive agreement at the full selected step. A local audit obtains informative risk and optimization bounds but no final endpoint confirmation in six budget–curvature comparisons. The results separate estimation risk, expected descent after choosing a step size, and statistical confirmation.

## 1 INTRODUCTION

Zeroth-order (ZO) optimization estimates updates from function values, trading derivative information for queries. This tradeoff has become useful for adapting language models with limited memory: MeZO performs fine-tuning with forward passes, and subsequent methods improve its estimates through variance reduction and structured perturbations (Malladi et al., 2023; Gautam et al., 2024; Chen et al., 2025; Yu et al., 2025). Another approach shrinks a noisy gradient toward historical information, as in the recent SR-Adam preprint (Arashi & Amintoosi, 2026). These developments motivate a precise question: when does a reduction in gradient estimation error translate into a better optimization update?

The answer depends on what changes alongside the error. Standard descent analyses involve both the mean update and its second moment (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017), and local curvature can matter more than ambient dimension (Malladi et al., 2023). For unbiased estimators of the same gradient, lower total variance improves optimal expected descent on an isotropic quadratic. Shrinkage also changes the mean, so its bias and curvature cost must be considered jointly. Even a deterministic rescaling exposes this issue: multiplying every update by a positive constant changes its squared error but can be absorbed into the learning rate. The minimum Euclidean error over fixed rescalings is exactly a constant minus twice the best expected decrease on an isotropic quadratic. A comparison must therefore distinguish lower risk at the original scale from greater progress after selecting the step size.

We study this distinction for ridge shrinkage toward a fixed anchor, holding the update distribution and target fixed at one state. Under iid sphere directions, an explicit Stein kernel yields a risk bound for the actual finite average of ZO estimates. For a fixed number of directions $m \geq 3$ and sufficiently large dimension, this bound is negative while the same shrinkage rule yields lower expected descent



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.

than Raw after each method chooses its optimal fixed step. The two conclusions use exactly the same shrinkage strength.

The mechanism is that increasing dimension does not eliminate coefficient randomness when the number of queried directions stays fixed. A finite collection of gradient projections continues to control both update energy and the shrinkage coefficient. With zero anchor, the effective step $\eta q(X)$ varies across updates without improving signal relative to curvature cost. An anchor can add useful directional information, whose value we separate from the effect of the mixing coefficient through explicit curvature comparisons and the fixed mixture $(X + a)/2$.

**Contributions.** First, we derive an explicit Stein risk bound for iid sphere ZO estimates and prove that, for fixed $m \geq 3$ in sufficiently large dimensions, one nonlinear shrinkage rule reduces risk while worsening expected descent even after separate step optimization. We quantify the limiting relative loss and give explicit bounds in finite dimensions.

Second, we characterize how anchor alignment and transverse curvature govern the comparison. A criterion based on four moments describes the sign of the optimization gain across a curvature family, while an explicit limit distinguishes the value of the anchor from that of nonlinear mixing: the fixed mixture $(X + a)/2$ outperforms the nonlinear rule in the stated regime.

Third, we evaluate these mechanisms with independent simulations, expanded step searches at OPT-125M checkpoints, and a complete local audit with finite-sample confidence bounds. The simulations test quantitative predictions, the checkpoint study separates coefficient effects from finite-difference and Taylor errors, and the audit measures the cost of turning population comparisons into statistically confirmed decisions.

## 2 Related work

Recent ZO fine-tuning research evaluates how gradient estimates translate into progress under memory and query constraints. MeZO and subsequent benchmarking establish this practical setting (Malladi et al., 2023; Zhang et al., 2024), building on classical analyses of estimation error and query complexity (Duchi et al., 2015; Nesterov & Spokoiny, 2017; Shamir, 2017). Our question concerns the criterion used to compare updates within that setting. A lower squared error can reflect a change of scale that disappears after adjusting the learning rate. We examine what remains after each candidate selects its own fixed step, connecting statistical risk to an explicit optimization comparison.

One route to better updates improves their construction through variance reduction (Gautam et al., 2024; Rando et al., 2025) or unbiased estimation at finite perturbation radii (Ma & Huang, 2025b). These methods change the information or function evaluations used to form an estimate. Shrinkage instead transforms an existing estimate and generally changes its mean as well as its variance. Our exact quadratic model removes finite-difference bias and holds the queried update law fixed. This intervention separates the optimization effect of shrinkage from improvements due to additional information or more accurate finite differences, allowing the risk and descent claims to be tested on the same candidate.

Another route changes the geometry of estimation. Orthogonal directions and directionally aligned perturbations alter the sampling law (Kozak et al., 2023; Ma & Huang, 2025a); sparse and low-rank methods restrict or structure the perturbed parameter space (Liu et al., 2025; Chen et al., 2025; Yu et al., 2025). Curvature also guides optimizer design in HiZOO (Zhao et al., 2025) and enters MeZO's analysis through local effective rank (Malladi et al., 2023). We use geometry to answer a different question: how can the same gradient, update distribution, and estimation risk produce opposite descent comparisons? Varying curvature while preserving those quantities isolates alignment and directional energy. The resulting four moments summarize an update's response to curvature; they leave the query space unchanged and incur a separate estimation cost.

Using additional directional information has a longer history. Guided evolutionary strategies exploit surrogate gradients and explicitly study a bias–variance tradeoff (Maheswaranathan et al., 2019); Addax combines first-order and zeroth-order information (Li et al., 2025). The recent SR-Adam preprint uses shrinkage toward historical momentum (Arashi & Amintoosi, 2026). Our anchored comparison shares the premise that an imperfect auxiliary direction can be useful. It distinguishes the value of that direction from the value of a nonlinear mixing coefficient by giving fixed and nonlinear

mixtures the same anchor and separately optimizing their steps. This control identifies whether the coefficient improves the use of available information, beyond simply introducing the anchor or reducing the update norm.

The statistical basis is quadratic-loss shrinkage (James & Stein, 1961; Baranchik, 1970) and its extension beyond Gaussian noise through Stein kernels (Fathi et al., 2022). We specialize this framework to a finite average of iid sphere ZO estimates, retaining the kernel fluctuations that matter when the number of directions is small. The contribution is the connection between this risk analysis and optimization at the same shrinkage strength: a negative risk bound coexists with persistently lower expected descent after step optimization. The anchored analysis then specifies how additional alignment and transverse curvature modify that comparison. This explains why a valid risk certificate and a useful optimization decision require different information.

## 3 Comparing estimation risk and expected descent

A shrinkage rule may estimate the gradient more accurately or merely suit the chosen learning rate. We separate these effects by optimizing a fixed step for each candidate at one state. A quadratic objective lets us change curvature without changing the update distribution.

### 3.1 A controlled model of the update

Fix a nonzero gradient $g$ in a Euclidean space of dimension $r$ and consider

$$F_h(v) = F_h(0) + g^\top v + \tfrac{1}{2} v^\top H_h v, \qquad H_h \succ 0. \tag{1}$$

A two-sided difference is exactly $g^\top u$, independently of $H_h$. Thus varying $H_h$ changes the consequence of an update while preserving its estimation error. The estimators with sphere and Gaussian directions are

$$X_D = \frac{1}{m}\sum_{i=1}^{m} Z_{D,i}, \qquad Z_{\mathrm{S}} = r(g^\top u)u, \quad Z_{\mathrm{G}} = (g^\top z)z, \quad u \sim \mathrm{Unif}(\mathbb{S}^{r-1}),\ z \sim N(0, I_r). \tag{2}$$

The summands are iid conditional on the state, with $\mathbb{E}X_D = g$. We use the sphere law unless stated otherwise; orthogonal directions or control variates would change this distribution.

We study shrinkage toward an anchor because it combines a change of scale with a possible change of direction. For $a, \lambda, \tau$ fixed before drawing $X$, define

$$S(X) = S^{+}_{\lambda,\tau,a}(X) = a + q(X)(X - a), \qquad q(X) = 1 - \frac{\lambda}{\|X - a\|^2 + \tau^2}, \quad 0 \le \lambda \le \tau^2, \tag{3}$$

where $\tau > 0$ makes the denominator positive. The coefficient retains more of $X$ as its distance from $a$ grows, while the ridge bounds shrinkage near the anchor. The positive-part convention is inactive in this range. Only $q(X)$ depends on the new update. Applying this map requires no further function values; obtaining the anchor and parameters is a separate cost.

### 3.2 Accounting for the choice of step size

Estimation risk measures proximity to $g$:

$$\mathcal{R}(Y) = \mathbb{E}\|Y - g\|^2, \qquad \Delta_R = \mathcal{R}(S) - \mathcal{R}(X), \tag{4}$$

To measure optimization progress, optimize one deterministic step for each candidate:

$$\mathcal{D}_H(Y) = \sup_{\eta \ge 0} \left\{ \eta \mathbb{E}[g^\top Y] - \tfrac{1}{2}\eta^2 \mathbb{E}[Y^\top H Y] \right\} = \frac{(\mathbb{E}[g^\top Y])_+^2}{2\mathbb{E}[Y^\top H Y]}. \tag{5}$$

We set $\mathcal{D}_H(0) = 0$ and write $\Delta^*_D(H) = \mathcal{D}_H(S) - \mathcal{D}_H(X)$. Negative $\Delta_R$ and positive $\Delta^*_D$ favor shrinkage. This comparison removes a constant scaling difference while retaining any effect of the random coefficient $q(X)$.

To see why this matters, minimize risk over the same scalar freedom. Completing the square yields

$$\mathcal{R}_{\mathrm{sc}}(Y) := \inf_{c\geq 0} \mathbb{E}\|cY - g\|^2 = \|g\|^2 - 2\mathcal{D}_I(Y). \tag{6}$$

For sphere Raw, the risk-minimizing constant is $q_0 = m/(m + r - 1)$, obtained without the gradient direction or extra queries. Although $q_0 X$ can greatly reduce risk, it has the same optimal expected descent as $X$: reciprocal step rescaling produces identical displacements. It is therefore an essential control for interpreting nonlinear shrinkage. For general $H \succ 0$, the corresponding risk identity uses target $H^{-1}g$ and the $H$ norm (Appendix D.1).

Equation (5) is exact on a quadratic; $H = LI$ also gives a smoothness lower bound for an $L$-smooth objective. Section 6 measures network loss at independently fitted steps. We first use the exact model to determine what risk reduction implies.

## 4 Risk reduction with lower expected descent

To establish a separation, the risk and descent comparisons must concern the same candidate. We bound its risk under the sphere law, identify how coefficient randomness affects descent, and test that effect in finite dimensions.

### 4.1 Certifying risk reduction under the sphere law

With few directions, the fluctuations controlling $q(X)$ need not resemble Gaussian noise. Spherical integration by parts expresses the risk difference through covariance structure and a kernel fluctuation penalty, retaining the actual update law. The starting quantities are the covariance and its gap, with $s = \|g\|^2$:

$$\Gamma = \frac{r}{r+2} sI + \frac{r-2}{r+2} gg^\top, \qquad \Sigma = \Gamma/m, \qquad A = \operatorname{tr}\Sigma - 2\|\Sigma\|_{\mathrm{op}} = \frac{(r-1)(r-2)}{m(r+2)} s. \tag{7}$$

The gap $A$ measures how much total variance remains beyond twice the largest covariance eigenvalue. The centered single-direction coordinates are degree-two spherical harmonics, giving a latent Stein kernel $K$ with $\mathbb{E}K = \Sigma$ and explicit trace and Frobenius fluctuations.

**Theorem 4.1** (Direct sphere risk bound). *For the sphere law (2), let $r \geq 3$, $m \geq 1$, and $\kappa = \|\Sigma\|_{\mathrm{op}}$. Define*

$$\sigma_{\mathrm{tr}}^2 = s^2 \frac{(r-2)^2(r-1)}{2(r+2)m^3}, \tag{8}$$

$$\sigma_F^2 = \frac{s^2}{m^3}\left[\frac{r^2}{4} + \frac{3r(r-2)}{4(r+2)} - \operatorname{tr}(\Gamma_1^2)\right], \quad \Gamma_1 = \frac{r}{r+2} I + \frac{r-2}{r+2} e_1 e_1^\top.$$

*For fixed $a, \lambda, \tau$ satisfying $0 < \lambda \leq \min\{\tau^2, 2A\}$, set $Q = \|g - a\|^2 + \operatorname{tr}\Sigma + \tau^2$. Then*

$$\Delta_R \leq B_{\mathrm{St}} := \frac{\lambda(\lambda - 2A)}{Q} - \frac{\lambda\tau^2(\lambda + 4\kappa)}{Q^2} + \frac{\lambda}{\tau^2}(\sigma_{\mathrm{tr}} + \sigma_F). \tag{9}$$

The first two terms supply improvement from covariance structure and the ridge; the last pays for kernel fluctuations. The proof retains the dependence between $K$ and $X$ (Appendix A). We take $\lambda = \tau^2 = A$ below, preserving a nonvanishing contraction. Unknown target information enters through $\|g\|^2$ and $g^\top a$, calibrated using queries in Appendix B. We now examine the same candidate's descent.

### 4.2 Why step optimization does not remove the loss

A fixed step absorbs a deterministic coefficient but cannot undo variation in $q(X)$. To isolate this effect, first take $m = 1$, $g = e_1$, $a = 0$, and $H = I$. Since $\|X\|^2 = rX_1$, Raw's step $1/r$ is optimal for every sample. Define $\mathbb{E}_w f = \mathbb{E}[\|X\|^2 f]/\mathbb{E}\|X\|^2$. Then

$$\frac{\mathcal{D}_I(qX)}{\mathcal{D}_I(X)} = \frac{(\mathbb{E}_w q)^2}{\mathbb{E}_w q^2} = \frac{1}{1 + \mathrm{CV}_w(q)^2}.$$

Table 1: Prespecified limits and independent sphere observations at $r = 4096$. Risk ratios use unscaled Raw; achieved-decrease ratios use independently fitted fixed steps. Rounded empirical values are not confidence bounds.

| | Risk ratio | | Decrease ratio | |
|---|---|---|---|---|
| $m$ | Limit $b_m$ | Evaluation | Limit $a_m^2/b_m$ | Evaluation |
| 1 | 0.467039 | 0.467457 | 0.920514 | 0.920364 |
| 4 | 0.328113 | 0.327891 | 0.937713 | 0.937339 |
| 8 | 0.292523 | 0.293925 | 0.957560 | 0.956929 |
| 16 | 0.272285 | 0.274523 | 0.974551 | 0.973667 |

The ridge coefficient is nonconstant, so the ratio is strictly below one, even though squared error falls sample by sample for $r > 2$ (Theorem C.1). The loss comes from variable effective steps rather than a poorly chosen nominal rate.

For multiple directions, the question is whether averaging eliminates this variation as dimension grows. Keep $m$ fixed, $g = e_1$, $a = 0$, and $\lambda_r = \tau_r^2 = A_r = (r-1)(r-2)/(m(r+2))$. With $W \sim \chi_m^2$ and $P = W/m$, define

$$q_\infty = \frac{P}{P+1}, \qquad a_m = \mathbb{E}[Pq_\infty], \quad b_m = \mathbb{E}[Pq_\infty^2], \quad c_m = \mathbb{E}[(1-q_\infty)^2]. \tag{10}$$

**Theorem 4.2** (Risk and descent with a fixed number of directions)**.** *For iid sphere updates as $r \to \infty$ with $m$ fixed,*

$$\frac{\mathcal{R}(S_r)}{\mathcal{R}(X_r)} \to b_m < 1, \qquad \frac{\mathcal{D}_I(S_r)}{\mathcal{D}_I(X_r)} \to \frac{a_m^2}{b_m} < 1. \tag{11}$$

*At the same parameters, $B_{\mathrm{St},r}/\mathcal{R}(X_r) \to -3/4 + (1/\sqrt{2} + 1/2)/\sqrt{m} < 0$ for every integer $m \geq 3$. Thus sufficiently large finite dimensions admit certified risk improvement and worse optimized descent for the same candidate. Appendix D.3 gives explicit $O_m(r^{-1/2})$ ratio brackets, not only an existence limit.*

The projected energy remains random: under a common coupling, $(X_r)_1 \to P$ and $\|X_r\|^2/r \to P/m$, so $q(X_r) \to P/(P+1)$. Weighting the limiting law by $P$ makes $a_m^2/b_m$ a squared mean divided by a second moment. Strict Cauchy–Schwarz yields the loss. Since $\mathcal{D}_I(X_r) = m/[2(r+m-1)]$, the loss persists relatively while absolute progress is of order $1/r$. Appendix D.4 gives conservative sufficient dimensions.

### 4.3 Testing the predicted loss in finite dimensions

The limit predicts both the risk ratio and the decrease ratio, allowing a quantitative test of the mechanism. We evaluate 40 combinations of direction law, dimension, and direction count, with 32,768 fitting and 131,072 independent evaluation updates each. The limits were specified before this study. Raw and its fixed rescaling achieve identical decreases after fitting. For sphere $r = 64$, nonlinear-to-Raw decrease ratios are $0.9206$ at $m = 4$ and $0.9429$ at $m = 16$; at $r = 4096$ they approach the predictions in Table 1. Gaussian results and sampler checks appear in Appendix H.

The scale control explains the statistical gain. At $r = 64, m = 16$, evaluated risks are $3.9371$ for Raw, $1.6037$ for nonlinear, and $0.7972$ for $q_0 X$. Constant rescaling improves risk more and preserves Raw's attainable descent. A nonzero anchor introduces directional information beyond this scaling comparison, which we examine next.

## 5 Anchor information and nonlinear mixing

An anchor contributes $(1-q)P_{X^\perp}a$ outside the Raw direction, which step rescaling cannot reproduce. To separate its value from that of the nonlinear coefficient, we change curvature while preserving the update law and compare fixed mixtures using the same anchor.

### 5.1 Separating the anchor from its mixing coefficient

We first extend the limit to an anchor with parallel and perpendicular components. To keep the curvature cost of a bounded perpendicular anchor visible as dimension grows, amplify curvature in that direction. Retain $g = e_1$, fixed $m$, and $\lambda_r = \tau_r^2 = A_r$, and set

$$a_r = \gamma e_1 + \beta e_2, \qquad H_r = I + \zeta r e_2 e_2^\top, \qquad \gamma \in \mathbb{R},\ \beta, \zeta \geq 0,$$

with all three constants fixed. Theorem 4.2's calculation then gives

$$\frac{\mathcal{D}_{H_r}(S_r)}{\mathcal{D}_{H_r}(X_r)} \longrightarrow \frac{(1+\zeta)a_m^2(1+\gamma)_+^2}{(1+\zeta)b_m + m\zeta\beta^2 c_m}. \tag{12}$$

The parallel component changes alignment; the perpendicular component adds curvature energy. The risk ratio remains $b_m$, so it cannot distinguish the resulting optimization signs. For $\gamma > -1$, put $K_m = a_m^2(1+\gamma)^2 - b_m$. The limiting advantage over Raw is positive exactly when

$$(1+\zeta)K_m > m\zeta\beta^2 c_m.$$

A positive critical value of $\zeta$ therefore separates opposite descent outcomes with the same leading risk reduction.

To identify the coefficient's contribution, use the fixed half mixture $Y_r = (X_r + a_r)/2$. For $\gamma > -1$,

$$\frac{\mathcal{D}_{H_r}(S_r)}{\mathcal{D}_{H_r}(Y_r)} \longrightarrow \frac{a_m^2[(1+\zeta) + m\zeta\beta^2]}{b_m(1+\zeta) + m\zeta\beta^2 c_m} < 1. \tag{13}$$

The inequalities $a_m^2 < b_m$ and $a_m^2 < c_m$ make the ratio strictly smaller than one. Thus the anchor can help while this nonlinear rule uses it less effectively than a fixed coefficient. This comparison concerns bounded anchors and $H_r = I + \zeta r e_2 e_2^\top$; explicit finite-dimensional error brackets currently cover the zero-anchor case. Appendix D.2 gives the proofs and Gaussian extension.

### 5.2 Predicting curvature effects from four moments

At a given dimension, the same mechanism depends on alignment and curvature energy. Since curvature weights fixed second moments, only four quantities are needed. Take $g = Ge_1$, $G > 0$, $a \in \operatorname{span}(e_1, e_2)$, and $H(h) = \operatorname{diag}(h_1, h_2, h_o I_{r-2})$ with $h_j > 0$. Define

$$\alpha = \mathbb{E}S_1, \quad u_1 = \mathbb{E}S_1^2, \quad u_2 = \mathbb{E}S_2^2, \quad u_o = \mathbb{E}\|S_{3:r}\|^2.$$

The first measures alignment; the other three record energy along the gradient, along the transverse anchor component, and in the remaining coordinates. Raw's corresponding constants are

$$C_1 = 1 + \frac{2(r-1)}{m(r+2)}, \quad C_2 = \frac{r}{m(r+2)}, \quad C_o = (r-2)C_2, \quad C(h) = \sum_j h_j C_j. \tag{14}$$

**Proposition 5.1** (A sign criterion valid across curvatures)**.** *Put* $u(h) = \sum_j h_j u_j > 0$ *and* $\kappa_j = (\alpha_+)^2 C_j - u_j$. *Then*

$$\Delta_D^*(h) = \frac{G^2 \sum_j h_j \kappa_j}{2C(h)u(h)}, \qquad \Delta_D^*(h) > 0 \iff \sum_j h_j \kappa_j > 0. \tag{15}$$

*The four update moments do not depend on* $h$. *Under* $\alpha > 0$, *a negative* $\kappa_2$ *gives the transverse-curvature threshold* $(h_1\kappa_1 + h_o\kappa_o)/(-\kappa_2)$ *when it is positive. Gaussian directions obey the same statement with their corresponding raw block constants. For a general positive-definite curvature matrix, one additional plane cross moment* $\mathbb{E}S_1S_2$ *is needed; the complement enters through its average trace.*

Because the moments remain fixed as $h$ varies, the criterion predicts a change in descent without changing gradient, risk, or sampled updates. A controlled study froze fourteen positive critical curvatures and tested $0.8$ and $1.2$ times each value with new fitting and evaluation pools. All 28 signs matched, each beyond three paired standard errors: pointwise checks of this intervention, rather than a simultaneous confidence region.

Finite-dimensional comparisons also clarify the role of the control. At $r = 64$, nonlinear sometimes beats the fixed half mixture, whereas a mixture with jointly selected coefficient and step wins all 220 tested comparisons against the selected nonlinear parameters. That control has an additional fitted degree of freedom, and configurations share samples (Appendix G). The relevant distinction is therefore between access to an anchor and how its coefficient uses that information.

Table 2: Independent network evaluation after step-range confirmation. Each method selects its own source-loss step. Positive differences favor nonlinear (N). Each row contains eight fixed, dependent settings; target uses the source-selected update and step.

| Evaluation | Contrast | + / − / unresolved | Mean decrease difference |
|---|---|---|---|
| Source | N–Raw | 2 / 5 / 1 | −0.021877 |
| Source | N–fixed mean | 0 / 7 / 1 | −0.012967 |
| Source | N–joint fixed blend | 0 / 7 / 1 | −0.031929 |
| Target | N–Raw | 1 / 4 / 3 | −0.008758 |
| Target | N–fixed mean | 3 / 0 / 5 | +0.006917 |
| Target | N–joint fixed blend | 1 / 2 / 5 | +0.001277 |

### 5.3 Identifying a useful nonlinear correction

To turn this distinction into a diagnostic, compare $S$ with a fixed mixture $S_F = a + \bar{q}(X - a)$ and write $S = S_F + U$. This isolates exactly what the nonlinear coefficient adds. For $H \succ 0$, let $p_F = \mathbb{E}[g^\top S_F] > 0$, $v_F = \mathbb{E}[S_F^\top H S_F] > 0$, and

$$d = \mathbb{E}[g^\top U], \qquad b = \mathbb{E}[S_F^\top H U], \qquad c = \mathbb{E}[U^\top H U].$$

These terms describe added alignment, its Hessian cross term with the fixed mixture, and the correction's curvature energy. When the updated alignment and energy are positive, separate step optimization favors $S$ precisely when

$$v_F(2p_F d + d^2) > p_F^2(2b + c). \tag{16}$$

Expanding the two optimized decreases gives this condition (Appendix C). If alignment does not increase and the cross term is nonnegative, the correction cannot help. In general, its alignment and energy effects must be evaluated together; coefficient variability alone does not determine the outcome.

The conditional-weighting and linear-projection benchmarks in Appendices D.5 and D.6 formalize how additional information can guide the coefficient. The three correction moments also predict its effect at a common finite step, connecting this analysis to actual network losses below.

## 6 Testing the mechanisms on network updates

A network adds finite-difference bias and error in the local quadratic model. We test which candidate gives lower actual loss after selecting its own step, then use a common step to test whether the correction moments predict the coefficient's effect.

### 6.1 Comparing actual loss after step selection

A contracted update may need a larger nominal step, so a truncated search can obscure the comparison in Section 3. We expand the search at four OPT-125M/SST-2 checkpoints (Zhang et al., 2022; Socher et al., 2013; Wang et al., 2018), taken at steps 16 and 64 of two short Raw sphere-ZO trajectories. Two source minibatches per state are paired with nonoverlapping target minibatches. Source-generated updates and selected steps are applied unchanged to target, fixing a conditional transfer comparison.

The historical EMA anchor, ridge, batches, and earlier mean coefficient remain frozen. Controls use that mean coefficient or jointly select a fixed coefficient and step with the same anchor. Each source setting uses 64 fitting, 64 profiling, and 128 new evaluation updates at $m = 4$. All 40 fitted steps are interior minima on the expanded grids, with at least two finite points to their right; the original grid had 31 boundary selections. Steps are frozen before evaluation. Table 2 uses prespecified multiplicity-adjusted approximate Student-$t$ intervals across these dependent settings.

Fixed mixtures win most resolved source comparisons. On target, nonlinear sometimes improves on fixed mean retention, but has no consistent advantage over Raw. These relative comparisons concern the recorded minibatches, rather than absolute target loss or generalization accuracy. The expanded grids remove the observed boundary problem, while the new fitting and evaluation pools also affect comparisons with the exploratory study. The next test holds steps equal to isolate the coefficient mechanism.

### 6.2 Predicting coefficient effects on independent updates

For $S_F = a + \bar{q}(X - a)$ and $U = S - S_F$, Section 5's correction moments give the following quadratic prediction at a common step:

$$\Delta Q(\eta) = \eta \mathbb{E}[g^\top U] - \frac{\eta^2}{2}\left(2\mathbb{E}[S_F^\top H U] + \mathbb{E}[U^\top H U]\right). \tag{17}$$

This expression remains meaningful for an indefinite local Hessian, unlike unrestricted optimization of a positive quadratic. Actual loss differs by the two endpoint residuals (Appendix I.4). Estimating the moments on a separate profiling pool and freezing the prediction before evaluation tests whether the mechanism predicts new updates.

At fractions $1/8, 1/4, 1/2, 1$ of the selected nonlinear step, predicted and evaluated point-estimate signs match in $12, 13, 15, 12$ of sixteen source/target contexts. Both approximate intervals exclude the prespecified indifference region with the same sign in only $6, 6, 5, 2$ contexts. The four full-step reversals have unresolved prediction intervals. Agreement at small steps therefore gives only partial support for the mechanism at the selected step.

Matched-direction diagnostics separate two approximation errors. Finite-difference and linearized updates differ by 17.39–39.59% relative RMSE across the eight source settings. For the same nonlinear updates, the absolute mean endpoint residual, summarized by its median across contexts, rises from $0.000281$ at $1/8$ step to $0.016134$ at full step. These measured residuals provide no uniform Hessian-Lipschitz bound. Independent prediction additionally incurs moment-estimation error. To examine that statistical obstacle alone, we return to an exact quadratic and ask whether finite samples suffice for a confirmed decision.

## 7 Confirming a local improvement with finite samples

The four-moment criterion is useful only if its estimation cost permits a decision. Returning to the quadratic removes network approximation errors: can risk reduction and an optimization comparison both be confirmed at one fixed state within a finite budget?

### 7.1 From population comparisons to selected steps

The protocol separates risk calibration $\mathcal{C}$, independent moment profiling $\mathcal{O}$, step fitting $\mathcal{F}$, and independent endpoint evaluation $\mathcal{E}$. Calibration bounds the unknown scale in the risk inequality; profiling predicts descent across curvatures; fitting fixes implementable steps; evaluation checks their improvement without reusing selection samples. Simultaneous intervals for bounded means (Hoeffding, 1963; Maurer & Pontil, 2009) give positive and negative regions for the curvature criterion. Because all curvatures use the same moments, one confidence event supports the whole specified family.

A population comparison still has to survive step fitting. For $a_Y = \mathbb{E}[g^\top Y] > 0$ and $v_Y = \mathbb{E}[Y^\top H Y] > 0$,

$$\Delta_D^\eta = \Delta_D^* - \mathrm{reg}_S + \mathrm{reg}_X, \qquad \mathrm{reg}_Y = \tfrac{1}{2} v_Y (\widehat{\eta}_Y - a_Y/v_Y)^2.$$

Thus a positive prediction needs a bound on shrinkage's fitting regret, whereas a negative prediction needs one on Raw's. Theorem E.1 combines these guards with the risk event and an independent endpoint interval. Requiring both a prediction and endpoint confirmation distinguishes the population criterion from its empirical resolution at the chosen budget.

### 7.2 The cost of endpoint confirmation

The completed audit uses three nested calibration/profiling budgets at two curvatures. Risk calibration hides the true scale; profiling uses the quadratic model's known coordinates and Raw moments. All three risk bounds are negative. Both larger budgets separate the two optimization signs and pass their regret guards; the positive case also passes at the smallest budget. Nevertheless, every empirical-Bernstein endpoint interval contains zero (Table 3).

Table 3: Independent endpoint confirmation in the completed local audit. Values are approximate summaries of the reported intervals; none of six budget/curvature instances passes the final endpoint condition. These are not the network's approximate $t$ intervals.

| Transverse curvature | Endpoint mean | Valid endpoint interval | Final acceptance |
|---|---|---|---|
| $h_2 = 1$ | $+0.04699$ | $[-0.00891,\ 0.10289]$ | 0 / 3 |
| $h_2 = 16$ | $-0.02011$ | $[-0.27780,\ 0.23758]$ | 0 / 3 |

The six rows share fitting and endpoint pools; only the endpoint pool's independence from the earlier stages is used for confirmation. Larger profiling budgets therefore sharpen the prediction without increasing endpoint resolution. A direct endpoint estimator resolves the positive case at the two higher matched *per-curvature* query budgets, but not the negative case. Reusing moments across curvatures changes the cost of the whole family, so costs are reported separately in Appendix H.2. At these budgets the final interval, rather than the sign of the moment criterion, prevents complete confirmation. This adds a statistical requirement to the earlier distinction between lower estimation risk and better expected descent.

# 8 Discussion and Conclusion

## 8.1 Discussion

The appropriate comparison depends on the intended use of an estimate. Squared error at a specified scale is meaningful for gradient estimation; optimization also permits a choice of step size and depends on directional curvature. Our analysis separates these effects through fixed rescaling, mixtures with the same anchor, and the alignment and energy of the remaining nonlinear correction. These controls make improvements attributable to the information a method adds and how it uses that information.

The analysis also indicates what a useful nonlinear coefficient would need to learn. Equation (16) tests its added alignment jointly with the change in curvature energy, including the cross term with the fixed mixture. For scalar reweighting, Appendix D.5 identifies conditional alignment relative to conditional energy as the appropriate prediction target. Norm variation is useful only when it informs that quantity. This suggests designing coefficients around information about progress along the sampled direction, and comparing them with fixed mixtures before incurring the cost of further calibration. The required information, its estimation accuracy, and the resulting gain should be assessed together.

The experiments identify where this analysis becomes harder to use. Independent simulations support the quantitative limits, but the checkpoint comparisons show no consistent nonlinear advantage after expanding the step range. Profiling predictions become less reliable at larger steps, and the local audit fails to confirm its predicted signs at the endpoints. Further progress requires controlling approximation error over useful displacements and reducing the cost of statistical confirmation. Reusing moments across configurations is promising only when their acquisition cost and transfer validity support that use.

## 8.2 Conclusion

Reducing estimation risk is insufficient to justify nonlinear shrinkage as an optimization improvement. For the specified sphere rule with a fixed number of directions $m \geq 3$, we prove certified risk reduction together with worse expected descent in sufficiently large dimensions, even after separate step optimization. Persistent coefficient randomness explains this loss. An anchor can improve the direction, yet a fixed mixture uses its information more effectively in the studied asymptotic regime. These results establish a concrete evaluation principle: first account for scale and anchor information, then assess the nonlinear correction through alignment, curvature energy, and independently evaluated progress. This connects the statistical benefit of shrinkage to the conditions needed for useful optimization.

## AI USE STATEMENT

We used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for provide critical ingredients for proving mathematical claims, generate synthetic data sets, propose or refine hypotheses, assist with translation. We have reviewed all AI-assisted work, including checking ideas, checking mathematical claims, testing code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

M. Arashi and M. Amintoosi. Stein-rule shrinkage for stochastic gradient estimation in high dimensions, 2026.

A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 41(2):642–645, 1970. doi: 10.1214/aoms/1177697104.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9BiVepgmWW.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Max Fathi, Larry Goldstein, Gesine Reinert, and Adrien Saumard. Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. *The Annals of Statistics*, 50(5):2737–2766, 2022. doi: 10.1214/22-AOS2208.

Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15180–15208. PMLR, 2024. URL https://proceedings.mlr.press/v235/gautam24a.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 361–379. University of California Press, 1961. URL https://projecteuclid.org/euclid.bsmsp/1200512173.

David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199:1179–1219, 2023. doi: 10.1007/s10107-022-01866-9.

Zeman Li, Xinwei Zhang, Peilin Zhong, Yuan Deng, Meisam Razaviyayn, and Vahab Mirrokni. Addax: Utilizing zeroth-order gradients to improve memory efficiency and performance of SGD for fine-tuning language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QhxjQOMdDF.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1e5c2efbddc02c1d971e2f19ccdb07d0-Abstract-Conference.html.

Shaocong Ma and Heng Huang. Revisiting zeroth-order optimization: Minimum-variance two-point estimators and directionally aligned perturbations. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=ywFOSIT9ik.

Shaocong Ma and Heng Huang. On the optimal construction of unbiased gradient estimators for zeroth-order optimization. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=rVT1GK60Nt.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4264–4273. PMLR, 2019. URL https://proceedings.mlr.press/v97/maheswaranathan19a.html.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017. doi: 10.1007/s10208-015-9296-2.

Marco Rando, Cheik Traoré, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. A structured proximal stochastic variance reduced zeroth-order algorithm, 2025.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017. URL https://jmlr.org/papers/v18/16-632.html.

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

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

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4475–4485, October 2025.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models, 2022.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ad.html.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W. Tsang. Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6bf82cc56a5fa0287c438baa8be65a70-Paper-Conference.pdf.

**Guide to the appendices.** Appendices A–B establish the statistical bounds and calibration; Appendices C–D give the optimization proofs and limits with a fixed number of directions, including all finite-dimensional enclosures. Appendix E states and proves the full local confidence result, and Appendix F gives the direct extension to Gaussian directions. The retained controlled studies are in Appendix G. Appendix H documents the completed follow-up experiments. Appendix I preserves the earlier exploratory checkpoint protocol and diagnostics so that old outcomes, recovered records, and independent confirmation remain distinguishable.

# A Direct risk calculations and proofs

## A.1 Covariance and residual fourth moments

Write $s = \|g\|^2$, $\xi = Z - g$, $\Gamma = \operatorname{Cov} Z$, $t = \operatorname{tr}\Gamma$, $v = \operatorname{tr}(\Gamma^2)$, $\mu_4 = \mathbb{E}\|\xi\|^4$, and $b_3 = \mathbb{E}[\xi\|\xi\|^2]$. The exact one-direction formulas are

$$\Gamma_{\mathrm{S}} = \frac{r}{r+2} sI + \frac{r-2}{r+2} gg^\top, \qquad \Gamma_{\mathrm{G}} = sI + gg^\top, \tag{18}$$

$$\mu_{4,\mathrm{S}} = \frac{(r-1)(3r^2-7r+6)}{r+2} s^2, \qquad \mu_{4,\mathrm{G}} = (3r^2+20r+37)s^2, \tag{19}$$

$$b_{3,\mathrm{S}} = \frac{2(r-2)(r-1)}{r+2} sg, \qquad b_{3,\mathrm{G}} = 2(r+3)sg. \tag{20}$$

For $W = X - g = m^{-1}\sum_i \xi_i$, independence and centering give

$$\mathbb{E}\|W\|^4 = \frac{\mu_4 + (m-1)(t^2+2v)}{m^3}, \qquad \mathbb{E}[W\|W\|^2] = b_3/m^2. \tag{21}$$

Consequently, with $k = g - a$,

$$\mu = t/m + \|k\|^2, \tag{22}$$

$$\mu_2 = \frac{\mu_4 + (m-1)(t^2+2v)}{m^3} + \frac{4}{m} k^\top \Gamma k + \frac{4}{m^2} k^\top b_3 + \frac{2t}{m}\|k\|^2 + \|k\|^4. \tag{23}$$

These expressions apply to actual two-sided quadratic ZO updates for each direction family.

*Proof.* The identities $\mathbb{E}uu^\top = I/r$ and $\mathbb{E}[(g^\top u)^2 uu^\top] = (sI + 2gg^\top)/(r(r+2))$ yield the sphere covariance; Gaussian fourth moments give $\mathbb{E}[(g^\top z)^2 zz^\top] = sI + 2gg^\top$. For $g = \sqrt{s}e_1$, sphere directions satisfy

$$\|Z - g\|^2/s = 1 + r(r-2)u_1^2.$$

Squaring and using $\mathbb{E}u_1^4 = 3/[r(r+2)]$ gives the sphere fourth moment. Rotational symmetry makes $b_3$ parallel to $g$; its coefficient follows from

$$\mathbb{E}[(ru_1^2 - 1)(1 + r(r-2)u_1^2)] = 2(r-2)(r-1)/(r+2).$$

For Gaussian directions let $T = z_1^2$, $U = \sum_{j>1} z_j^2$. Then $T, U$ are independent and $\|Z-g\|^2/s = (T-1)^2 + TU$. The expectations $\mathbb{E}(T-1)^4 = 60$, $\mathbb{E}[T(T-1)^2] = 10$, $\mathbb{E}T^2 = 3$, $\mathbb{E}U = r-1$, and $\mathbb{E}U^2 = r^2 - 1$ give the fourth moment. For the third vector moment, the first component is proportional to

$$\mathbb{E}[(T-1)^3] + \mathbb{E}[(T-1)T]\mathbb{E}U = 8 + 2(r-1) = 2(r+3).$$

All transverse components have zero expectation by sign symmetry.

Expand $\|\sum_i \xi_i\|^4$. Terms involving an index appearing just once vanish. The remaining terms are $m\mu_4$ and $m(m-1)(t^2+2v)$. Dividing by $m^4$ gives the first part of (21). In the third-order vector expansion only the terms with three equal indices survive, giving $b_3/m^2$. Finally expand $(\|W\|^2 + 2k^\top W + \|k\|^2)^2$. The terms are $\mathbb{E}\|W\|^4$, $4k^\top \Gamma k/m$, $4k^\top b_3/m^2$, $2t\|k\|^2/m$, and $\|k\|^4$, proving (23). □

### A.2 A COMPLEMENTARY RESIDUAL-MOMENT INEQUALITY

**Lemma A.1** (Direct residual-moment bound). *Let $\mathbb{E}X = g$ and $\mathbb{E}\|X-a\|^4 < \infty$. Put $k = g - a$, $\mu = \mathbb{E}\|X-a\|^2 > 0$, $\mu_2 = \mathbb{E}\|X-a\|^4$, $c = \lambda/\tau^2 \in (0,1]$, and*

$$L_0 = \frac{\mu^2}{\mu_2 + \tau^2\mu}, \qquad \alpha_0 = \|k\|/\tau.$$

*Then*

$$\Delta_R \le B_{\mathrm{res}} := \lambda[-(2-c)L_0 - cL_0^2 + \alpha_0]. \tag{24}$$

*Define $f(t) = -(2-c)t - ct^2 + 2\alpha_0\sqrt{t(1-t)}$. If $f'(L_0) \le 0$, the additional bound $\Delta_R \le \lambda f(L_0)$ holds. If that condition fails, substituting $L_0$ into $f$ is not justified. The case $\mu = 0$ has $X = a = g$ almost surely and $\Delta_R = 0$.*

**Proof.** Put $y = X - a$, $R = \|y\|^2$, $h = R + \tau^2$, and $T = R/h$. Direct expansion yields

$$\frac{\Delta_R}{\lambda} = -2\mathbb{E}\frac{R}{h} + \lambda\mathbb{E}\frac{R}{h^2} + 2\mathbb{E}\frac{k^\top y}{h} = -(2-c)\mathbb{E}T - c\mathbb{E}T^2 + 2\mathbb{E}\frac{k^\top y}{h}.$$

By Cauchy–Schwarz applied to $\sqrt{R/h}$ and $\sqrt{Rh}$, $\mu^2 \le \mathbb{E}T\,\mathbb{E}(Rh)$, hence $\mathbb{E}T \ge L_0$. Also $\mathbb{E}T^2 \ge (\mathbb{E}T)^2$. Since $\sqrt{R}/(R+\tau^2) \le 1/(2\tau)$, $k^\top y/h \le \|k\|/(2\tau)$. These prove (24). The sharper pointwise bound is $k^\top y/h \le \alpha_0\sqrt{T(1-T)}$. The function $f$ in the lemma is concave on $(0,1)$, continuously extended to its endpoints. Therefore $\Delta_R/\lambda \le \mathbb{E}f(T) \le f(\mathbb{E}T)$. If $f'(L_0) \le 0$, concavity makes $f$ nonincreasing on $[L_0, 1]$, so $f(\mathbb{E}T) \le f(L_0)$. For $\mu > 0$, finite $\mu_2$ and $\tau > 0$ ensure $0 < L_0 < 1$. If $\mu = 0$, $X = a$ almost surely and $g = a$, which proves the exceptional case.

### A.3 RELATION TO THE EXISTING STEIN FRAMEWORK

Fathi et al. (2022, Theorem 3.2) derive a covariance-gap risk term plus a discrepancy controlled by Stein-kernel trace fluctuations and matrix deviations. Their unregularized map has an inverse-squared norm, with corresponding test-function and inverse-moment requirements. Their Example 3.3 treats translated spherical noise by a sphere–ball zero-bias coupling; it is not the quadratic map $u \mapsto r(g^\top u)u$ used here. We do not claim the covariance gap, the Stein risk identity, non-Gaussian shrinkage, or the existence of spherical examples as new.

The specialization here is explicit: the centered ZO coordinates are degree-two harmonics; their tangential derivatives yield a latent kernel whose spectrum and finite-batch trace/Frobenius fluctuations can be computed. Translation to an anchor preserves the noise covariance. A positive ridge supplies bounded derivatives for the residual test, and keeping the negative ridge term plus the centered trace fluctuation gives the displayed bound at fixed shrinkage strength. Conditioning the latent kernel on $X$ would give a kernel expressed in $X$ and cannot increase its squared discrepancies; working with the latent version makes its moments explicit without assuming independence from $X$.

This is a law-specific computable instance of the general strategy, not a claim that the general framework is weaker. The theorem for a fixed number of directions links this instance to a loss in optimal expected descent at the same shrinkage strength. The reported direct-bound checks evaluate these formulas; they are not an empirical comparison against every available Stein refinement.

### A.4 PROOF OF THEOREM 4.1

First take $g = e_1$. Each centered coordinate $Z_j - g_j = ru_1u_j - g_j$ is a degree-two spherical harmonic, so $-\Delta_{\mathbb{S}}(Z_j - g_j) = 2r(Z_j - g_j)$. Integration by parts on the sphere and the chain rule imply, for a smooth scalar test $f$,

$$\mathbb{E}[(Z_j - g_j)f(Z)] = \sum_k \mathbb{E}[T_{jk}(u)\partial_k f(Z)], \qquad T_{jk}(u) = \frac{\nabla_{\mathbb{S}}Z_j \cdot \nabla_{\mathbb{S}}Z_k}{2r}.$$

The tangential derivatives $\nabla_{\mathbb{S}}Z_j = r(u_1e_j + u_je_1 - 2u_1u_ju)$ give

$$T(u) = \frac{r}{2}\{u_1^2 I + uu^\top + u_1(e_1u^\top + ue_1^\top) - 4u_1^2uu^\top\}. \tag{25}$$

For a batch, applying this identity to each independent direction gives a latent kernel $K = m^{-2}\sum_i T(u_i)$. Its two factors $1/m$ arise from the centered average and the chain rule. In particular,

$$\mathbb{E}[(X-g)^\top f(X)] = \mathbb{E}\sum_{jk} K_{jk}\partial_k f_j(X), \qquad \mathbb{E}K = \Gamma/m = \Sigma.$$

The second statement follows by choosing linear test functions. No independence of $K$ and $X$ is asserted or needed.

The eigenvalues of $T(u)$ are $r/2$, zero, and $ru_1^2/2$ with multiplicity $r-2$. Indeed, on $\mathrm{span}(u, e_1)$ the matrix after dividing by $r/2$ has a representation

$$\begin{pmatrix} 1-u_1^2 & u_1\sqrt{1-u_1^2} \\ u_1\sqrt{1-u_1^2} & u_1^2 \end{pmatrix},$$

whose eigenvalues are one and zero. The complement is scalar. Thus

$$\mathrm{tr}\, T = \tfrac{r}{2}[1+(r-2)u_1^2], \qquad \|T\|_F^2 = \tfrac{r^2}{4}[1+(r-2)u_1^4].$$

With $D = K - \Sigma$, independence across directions gives

$$\mathbb{E}(\mathrm{tr}\, D)^2 = \frac{\mathrm{Var}(\mathrm{tr}\, T)}{m^3}, \qquad \mathbb{E}\|D\|_F^2 = \frac{\mathbb{E}\|T\|_F^2 - \|\Gamma\|_F^2}{m^3}.$$

Using the sphere moments gives precisely the stated $\sigma_{\mathrm{tr}}^2, \sigma_F^2$. For general $g$, rotate and scale by $\|g\|$; the kernel and its fluctuation standard deviations scale by $s$, and the variances by $s^2$. Equivalently the general single-direction kernel is

$$\tfrac{r}{2}\{(g^\top u)^2 I + suu^\top + (g^\top u)(gu^\top + ug^\top) - 4(g^\top u)^2 uu^\top\}.$$

For $y = X - a$, $R = \|y\|^2$, $h = R + \tau^2$, apply the Stein identity to $y/h$. The risk difference is exactly

$$\Delta_R = \mathbb{E}\left[\lambda^2 R/h^2 - 2\lambda\,\mathrm{tr}\, K/h + 4\lambda y^\top K y/h^2\right].$$

Since $y^\top \Sigma y \le \kappa R$, the deterministic covariance part is at most

$$\frac{\lambda(\lambda - 2A)}{h} - \frac{\lambda\tau^2(\lambda + 4\kappa)}{h^2}.$$

Both coefficients are nonpositive when $\lambda \le 2A$. Jensen's inequalities for $h^{-1}, h^{-2}$ therefore bound its expectation from above by the corresponding expression with $h$ replaced by $Q = \mathbb{E}h$.

For the trace fluctuation use $\mathbb{E}\,\mathrm{tr}\, D = 0$ and the range $0 \le 1/h \le 1/\tau^2$:

$$|2\lambda\mathbb{E}(\mathrm{tr}\, D/h)| \le 2\lambda\sigma_{\mathrm{tr}}\sqrt{\mathrm{Var}(1/h)} \le \lambda\sigma_{\mathrm{tr}}/\tau^2.$$

For the remaining term use Frobenius Cauchy–Schwarz and $R/h^2 \le 1/(4\tau^2)$:

$$|4\lambda\mathbb{E}(y^\top D y/h^2)| \le 4\lambda\sqrt{\mathbb{E}\|D\|_F^2}\sqrt{\mathbb{E}\|yy^\top/h^2\|_F^2} \le \lambda\sigma_F/\tau^2.$$

The support is compact and the ridge is positive, which justify the test functions. Combining the bounds proves the theorem.

## B Scalar calibration for the direct bound

For general $g$, the target enters the sphere bounds through $s = \|g\|^2$ and $b = g^\top a$, in addition to the known $\|a\|^2$. On an exact quadratic, one pair of function queries along $a$ gives $b$; zero anchor needs no such pair. At the fixed state, let $\widehat{s} = n_C^{-1}\sum_i r(g^\top u_i)^2$. With $x = \log(2/\delta_C)$ and $n_C > 4x$,

$$s_L = \frac{\widehat{s}}{1 + 2\sqrt{x/n_C} + 2x/n_C}, \qquad s_U = \frac{\widehat{s}}{1 - 2\sqrt{x/n_C}} \tag{26}$$

cover $s$ with probability at least $1 - \delta_C$. The proof uses the convex-order domination of $r\,\mathrm{Beta}(1/2, (r-1)/2)$ by $\chi_1^2$, controlling positive and negative exponential moments. It is a scalar norm-concentration argument; the full procedure is not dimension-free.

**Corollary B.1** (Finite-calibration sphere risk certificate)**.** *Suppose a simultaneous calibration event gives $s \in [s_L, s_U]$ and $b \geq b_L$. Let $A_1, \kappa_1, \sigma_{\mathrm{tr},1}, \sigma_{F,1}$ denote the constants in Theorem 4.1 at unit target scale, and set*

$$A_L = s_L A_1, \quad \kappa_L = s_L \kappa_1, \quad Q_U = s_U(1 + (r-1)/m) + \|a\|^2 - 2b_L + \tau^2.$$

*When $Q_U > 0$ and $0 < \lambda \leq \min\{\tau^2, 2A_L\}$, the selected fixed map on a fresh update satisfies, on that event,*

$$\Delta_R \leq U_R := \frac{\lambda(\lambda - 2A_L)}{Q_U} - \frac{\lambda\tau^2(\lambda + 4\kappa_L)}{Q_U^2} + \frac{\lambda}{\tau^2} s_U(\sigma_{\mathrm{tr},1} + \sigma_{F,1}). \tag{27}$$

*A negative $U_R$ certifies conditional risk reduction. If $\lambda \leq 2A_L$ is not established, this displayed certificate is inapplicable, not a proof that the candidate is harmful.*

For general $g$, let $s = \|g\|^2$, $b = g^\top a$, and $v_a = \|a\|^2$. Then $\|g - a\|^2 = s + v_a - 2b$ and

$$(g-a)^\top \Gamma (g-a) = \frac{r}{r+2} s(s + v_a - 2b) + \frac{r-2}{r+2}(s-b)^2.$$

The residual-moment expression uses only these two target scalars as well. This is not an assumption that they are free. In an exact quadratic, $b = [F(\delta a) - F(-\delta a)]/(2\delta)$ at a cost of two queries. For noisy or nonquadratic values, a separate interval for $b$ is required.

### B.1 Proof of the scale interval

For $s > 0$, $r(g^\top u)^2/s$ has law $rB$, with $B \sim \mathrm{Beta}(1/2, (r-1)/2)$. Take an independent $R_\chi \sim \chi^2_r$. Then $R_\chi B \sim \chi^2_1$ and $rB = \mathbb{E}[R_\chi B \mid B]$. Conditional Jensen for the convex function $e^{t\cdot}$ gives

$$\mathbb{E}e^{trB} \leq (1-2t)^{-1/2}, \qquad t < 1/2,$$

including negative $t$. For the centered variable, expansion of $-\log(1-2t)$ gives $\log \mathbb{E}e^{t(rB-1)} \leq t^2/(1-2t)$ for $0 < t < 1/2$ and $\log \mathbb{E}e^{-t(rB-1)} \leq t^2$ for $t > 0$. For an average of $n_C$ independent copies, Chernoff optimization gives

$$\Pr\{\widehat{s}/s > 1 + 2\sqrt{x/n_C} + 2x/n_C\} \leq e^{-x}, \qquad \Pr\{\widehat{s}/s < 1 - 2\sqrt{x/n_C}\} \leq e^{-x}.$$

For completeness, the upper-tail exponent follows by substituting $t = \sqrt{x/n_C}/(1 + 2\sqrt{x/n_C})$ into the bound for the centered sum at level $2\sqrt{n_C x} + 2x$; the lower-tail choice is $t = \sqrt{x/n_C}$. The lower-tail exponent is $-x$ and the upper-tail exponent is at most $-x$. Inverting on their intersection proves (26). The case $s = 0$ gives zero observations. Squared slopes along Gaussian directions have exactly the dominating $\chi^2_1$ law, so the interval applies there too; the direct sphere risk theorem is not thereby extended to Gaussian updates.

### B.2 Proof of Corollary B.1

On the calibration event, $A \geq A_L$, $\kappa \geq \kappa_L$, $Q \leq Q_U$, and $\sigma_{\mathrm{tr}} + \sigma_F \leq s_U(\sigma_{\mathrm{tr},1} + \sigma_{F,1})$. The first numerator $\lambda(\lambda - 2A_L)$ is nonpositive by the stated restriction. Replacing each negative coefficient by a larger one, and each denominator by its upper bound, yields an upper bound term by term. The positive fluctuation term is bounded with $s_U$. This proves (27). Condition on the entire calibration output before applying the result to a fresh update. If $b$ is exact, its measurement incurs no statistical failure; otherwise the joint interval must allocate failure to $s$ and $b$. The same scale event can support multiple fixed anchors with exact projections. A stochastic anchor-projection interval valid only for a prespecified anchor cannot justify adaptive selection of that anchor from its own noisy data.

**Residual-moment calibration is separate.** The reported finite-calibration experiments apply only this Stein certificate. To calibrate Lemma A.1, one must propagate valid intervals through its residual moments and verify its derivative condition uniformly, or use a valid supremum over a simultaneous parameter confidence set. We do not silently substitute point estimates into that lemma or relabel the known-parameter 33/36 count as calibrated acceptance.

# C STRICT SEPARATION, CURVATURE, AND NONLINEAR COMPARISONS

## C.1 AN EXACT SINGLE-DIRECTION SEPARATION

**Theorem C.1** (Samplewise risk improvement with strictly worse tuned descent)**.** *Let $g = e_1$, $r > 2$, $m = 1$, $X = ru_1u$, and $H = I$. For a measurable factor $q(X) \in [0, 1]$,*

$$\|qX - g\|^2 - \|X - g\|^2 = (1 - q)X_1[2 - r(1 + q)] \le 0. \tag{28}$$

*If $\Pr(q < 1) > 0$, the expected inequality is strict. For any nonnegative radial factor with $0 < \mathbb{E}[q^2\|X\|^2] < \infty$, define $\mathbb{E}_w f = \mathbb{E}[\|X\|^2 f]/\mathbb{E}\|X\|^2$. Then*

$$\frac{\mathcal{D}_I(qX)}{\mathcal{D}_I(X)} = \frac{(\mathbb{E}_w q)^2}{\mathbb{E}_w q^2} = \frac{1}{1 + \mathrm{CV}_w(q)^2} \le 1, \qquad \mathcal{D}_I(X) = \frac{1}{2r}. \tag{29}$$

*Equality holds precisely when $q$ is constant under the weighted law. The zero-anchor ridge factor with $0 < \lambda \le \tau^2$ is nonconstant and less than one almost surely. It therefore gives strict samplewise risk improvement and strictly lower expected descent after step optimization.*

**Proof.** With $R_0 = \|X\|^2 = rX_1$, squared-error expansion gives

$$\|qX - g\|^2 - \|X - g\|^2 = (q^2 - 1)R_0 - 2(q - 1)X_1 = (1 - q)X_1[2 - r(1 + q)].$$

For $r > 2$, the bracket is strictly negative, and $X_1 > 0$ almost surely. This proves the risk claim. The quadratic expected decrease from a fixed step is $\eta\mathbb{E}(qX_1) - \eta^2\mathbb{E}(q^2R_0)/2$, yielding

$$\mathcal{D}_I(qX) = \frac{(\mathbb{E}qR_0)^2}{2r^2\mathbb{E}q^2R_0}.$$

Cauchy–Schwarz gives $(\mathbb{E}qR_0)^2 \le \mathbb{E}R_0\mathbb{E}q^2R_0$, with equality precisely for a constant factor under the energy-weighted law. Since $\mathbb{E}R_0 = r$, this proves (29). For ridge $q(R_0) = 1 - \lambda/(R_0 + \tau^2)$, $R_0$ is nondegenerate and the factor is strictly increasing; it is nonconstant under that law. This is a strict finite-parameter result, not a large-ridge limit.

More generally, $\|X - g\|^2 = 1 + (r - 2)X_1 \ge 1$. If a fixed anchor satisfies $\|a - g\| \le 1$, convexity of squared distance proves samplewise risk nonincrease for any measurable convex mixing factor. This condition concerns risk only and supplies neither an anchor-construction procedure nor an optimization guarantee.

## C.2 THE CORRECTION-MOMENT COMPARISON

Let $V = (X - a)/(\|X - a\|^2 + \tau^2)$, so $S = X - \lambda V$. Define

$$p = \mathbb{E}[g^\top X] > 0, \quad M = \mathbb{E}[X^\top HX], \quad \mathsf{A} = \mathbb{E}[g^\top V], \quad \mathsf{B} = \mathbb{E}[X^\top HV], \quad \mathsf{C} = \mathbb{E}[V^\top HV].$$

On the domain with positive expected alignment $p - \lambda\mathsf{A} > 0$ and with positive updated curvature energy,

$$\Delta_D^*(H) > 0 \quad \Longleftrightarrow \quad 2(p^2\mathsf{B} - pM\mathsf{A}) > \lambda(p^2\mathsf{C} - M\mathsf{A}^2). \tag{30}$$

The nonlinear denominator remains inside all expectations. Since $\|V\| \le 1/(2\tau)$, the correction moments exist when $X$ has a finite second moment. Direct expansion gives

$$\mathbb{E}[g^\top S] = p - \lambda\mathsf{A}, \qquad \mathbb{E}[S^\top HS] = M - 2\lambda\mathsf{B} + \lambda^2\mathsf{C}.$$

On the stated domain with positive expected alignment, cross multiplication in (5) gives

$$\Delta_D^*(H) = \frac{\lambda\{2(p^2\mathsf{B} - pM\mathsf{A}) - \lambda(p^2\mathsf{C} - M\mathsf{A}^2)\}}{2M(M - 2\lambda\mathsf{B} + \lambda^2\mathsf{C})},$$

which proves (30). The risk condition instead is

$$\Delta_R = -2\lambda\mathbb{E}[(X - g)^\top V] + \lambda^2\mathbb{E}\|V\|^2 < 0.$$

The different weights explain why it is not equivalent to the optimization condition. For a fixed ridge, the correction moments are fixed while $\lambda$ varies. They cannot be treated as fixed along a path that changes the ridge simultaneously.

### C.3 Proof of Proposition 5.1

Rotations fixing the plane $\mathrm{span}(e_1, e_2)$ preserve the update law and the anchored map. Therefore transverse means and plane–complement second cross moments vanish, and the complementary second-moment block is scalar. For a diagonal curvature family,

$$\mathbb{E}[g^\top S] = G\alpha, \quad \mathbb{E}[S^\top H(h) S] = u(h), \quad \mathbb{E}[g^\top X] = G^2, \quad \mathbb{E}[X^\top H(h) X] = G^2 C(h).$$

Substitute in (5) to obtain

$$\mathcal{D}_H(S) = \frac{G^2(\alpha_+)^2}{2u(h)}, \qquad \mathcal{D}_H(X) = \frac{G^2}{2C(h)}.$$

Their difference is (15). The transverse threshold follows by solving a linear inequality. For Gaussian directions the block constants are $C_1 = 1 + 2/m$, $C_2 = 1/m$, $C_o = (r-2)/m$. For a general positive-definite $H$, the updated curvature energy becomes

$$H_{11}u_1 + 2H_{12}u_{12} + H_{22}u_2 + \frac{\mathrm{tr}(H_{3:r,3:r})}{r-2}u_o, \quad u_{12} = \mathbb{E}S_1 S_2.$$

The raw plane cross moment is zero. Thus its sign functional adds $-2H_{12}u_{12}$ and uses the complementary mean curvature. Curvature information itself must be known or measured; symmetry does not make it observable without cost.

### C.4 The fixed-blend comparisons

**Exact comparison with a specified fixed blend.** For a frozen $S_F = a + \bar{q}(X - a)$, write $S = S_F + U$, $\alpha_F = \mathbb{E}[g^\top S_F] > 0$, $M_F = \mathbb{E}[S_F^\top H S_F]$, $d = \mathbb{E}[g^\top U]$, $b_F = \mathbb{E}[S_F^\top H U]$, and $c_U = \mathbb{E}[U^\top H U]$. When $\alpha_F + d > 0$, nonlinear shrinkage improves the tuned value exactly when

$$M_F(2\alpha_F d + d^2) > \alpha_F^2(2b_F + c_U). \tag{31}$$

For $S = S_F + U$, alignment is $\alpha_F + d$ and curvature energy is $M_F + 2b_F + c_U$. Cross multiplication of the tuned values proves (31). If $d \le 0$ but $\alpha_F + d > 0$, the left side is nonpositive; if also $b_F \ge 0$, the right side is nonnegative. Improvement is then impossible. If $\bar{q} = \mathbb{E}q$, $\mathbb{E}U = \mathrm{Cov}(q, X)$; with an independently estimated frozen $\bar{q}$ one must include the difference between that value and $\mathbb{E}q$.

For a perfect anchor $a = g$, let $R = \|X - g\|^2$, $S_F = g + \mathbb{E}[q](X - g)$. The ridge retention $q(R)$ increases with $R$ and is nonnegative. Then

$$\mathcal{R}(S) - \mathcal{R}(S_F) = \mathrm{Cov}(q^2, R) + \mathrm{Var}(q)\mathbb{E}R \ge 0.$$

The covariance is nonnegative because $q^2$ is an increasing function of $R$: for an independent copy $R'$, it equals one half of $\mathbb{E}[(q(R)^2 - q(R')^2)(R - R')]$. Nondegenerate ridge retention makes the comparison strict when the residual is nonzero. This is a risk statement at matched mean retention, not an optimization ordering for all nonlinear rules.

A fixed blend with a nonnegative step has displacement $cX + da$ for $c, d \ge 0$. Its expected quadratic decrease is

$$cp + d\, g^\top a - \tfrac{1}{2}\{c^2 M + 2cd\, \mathbb{E}[X^\top H a] + d^2 a^\top H a\}.$$

The quadratic coefficient matrix is positive semidefinite because it is the second moment of the two update components under $H$. Interior stationary points and the two nonnegative coordinate boundaries solve this two-dimensional concave problem; singular degeneracies can be handled by the corresponding flat directions. For a nonzero solution $\eta = c + d$ and $q = c/(c + d)$. The boundary $q = 0$ is anchor-only, so perfectly aligned anchors on eigen-directions can provide an especially strong benchmark. This is an additional fitted degree of freedom relative to a single fixed nonlinear candidate, and should be reported as such.

### C.5 Target mismatch and finite-step transfer

At a nonquadratic fixed state, use the actual update law, not linearized moments without error. The following identities are not direct-law certification of the checkpoint experiments. For $e = \theta - g$ and $b = \mathbb{E}S - \theta$, where $\mathbb{E}X = \theta$,

$$\mathbb{E}(\|S - g\|^2 - \|X - g\|^2) = \mathbb{E}(\|S - \theta\|^2 - \|X - \theta\|^2) + 2\langle e, b\rangle. \tag{32}$$

This follows by expanding both squared errors; the common $\|e\|^2$ cancels. A squared bias penalty alone does not transfer the risk difference.

For a shared random minibatch, this identity first conditions on the batch. Averaging leaves $2\mathbb{E}\langle\theta - g, b\rangle$, which need not vanish. Independent direction pools sharing that batch do not create independent information about the population target. The update coordinates and metric must also agree: for a nonorthonormal embedding $P$, parameter-space error is $\|P(Y-g)\|^2$, not unweighted coefficient error. None of the exact-law results asserts validity after changing the target, coordinate system, anchor, or update distribution.

If the Hessian is $L_H$-Lipschitz on every realized update segment and both third update moments are finite, Taylor's theorem for frozen steps gives

$$\Delta_D^{\text{actual}} \geq \Delta_D^{\text{quad}} - \frac{L_H}{6}\{\eta_S^3\mathbb{E}\|S\|^3 + \eta_X^3\mathbb{E}\|X\|^3\}. \tag{33}$$

There is an analogous upper bound with a plus sign. For bounded sphere updates, a containing ball can be specified; for unbounded Gaussian updates a purely local assumption does not cover all segments, so a global condition or a tail allowance is needed. This does not provide stage drift control. Neither the direct sphere risk theorem nor the joint theorem is extended to actual nonquadratic estimators by this inequality alone. In the checkpoint study (Appendix I), the residual is evaluated at tested endpoints but no valid $L_H$ is estimated. The recorded ReLU model and finite-radius directional errors preclude assuming the smooth-query hypotheses on all relevant segments.

## D Optimal rescaling and limits with a fixed number of directions

This section proves the comparisons after optimizing the step size in Sections 4–5. Throughout, expectations refer to a fixed state. The gradient $g$ and positive-definite matrix $H$ are deterministic; the update law has finite second moment, and $\mathcal{R}(Y) = \mathbb{E}\|Y-g\|^2$. Whenever a candidate is selected from data, these arguments are conditional on the information that freezes that candidate, before drawing the update. No result in this section treats a fitted step as a population optimum or an observed endpoint remainder as a smoothness bound.

### D.1 The metric after optimizing a deterministic scale

For any update $Y$, set $p_Y = \mathbb{E}[g^\top Y]$ and $v_{H,Y} = \mathbb{E}[Y^\top HY]$. If $v_{H,Y} > 0$, define

$$\mathcal{D}_H(Y) = \sup_{\eta\geq 0}\{\eta p_Y - \eta^2 v_{H,Y}/2\} = \frac{(p_Y)_+^2}{2v_{H,Y}}.$$

If $Y = 0$ almost surely, define $\mathcal{D}_H(Y) = 0$.

**Proposition D.1** (Risk after optimal rescaling). *For $H \succ 0$ and any square-integrable $Y$,*

$$\inf_{c\geq 0}\mathbb{E}\|cY - g\|^2 = \|g\|^2 - 2\mathcal{D}_I(Y), \tag{34}$$

$$\inf_{c\geq 0}\mathbb{E}\|cY - H^{-1}g\|_H^2 = g^\top H^{-1}g - 2\mathcal{D}_H(Y). \tag{35}$$

*The zero-update case uses the convention above. Thus isotropic optimized descent and Euclidean gradient risk after optimal rescaling rank update laws identically. For general $H$, the corresponding target is $H^{-1}g$ in the $H$ metric, not $g$ in the Euclidean metric.*

*Proof.* Expand the second objective as $c^2 v_{H,Y} - 2cp_Y + g^\top H^{-1}g$. Its minimum over $c \geq 0$ is attained at $(p_Y)_+/v_{H,Y}$ when the denominator is positive. If the denominator vanishes, positive definiteness implies $Y = 0$ almost surely and $p_Y = 0$. Set $H = I$ for the first identity. □

**Proposition D.2** (Fixed-scale and fixed-anchor risk baselines). *Suppose $\mathbb{E}X = g$ and $V = \mathbb{E}\|X - g\|^2$. For a deterministic scalar $q$,*

$$\mathcal{R}(qX) = (q-1)^2\|g\|^2 + q^2 V.$$

*For $g \neq 0$, the optimum over $q \geq 0$ is $q_0 = \|g\|^2/(\|g\|^2 + V)$. In the exact ZO models with iid sphere or Gaussian directions, respectively,*

$$q_0^{\mathrm{S}} = \frac{m}{m+r-1}, \qquad q_0^{\mathrm{G}} = \frac{m}{m+r+1}.$$

*These constants require only the direction family, $r$, and $m$. For a fixed anchor, put $d_a^2 = \|a-g\|^2$. Then*

$$\mathcal{R}(qX + (1-q)a) = q^2 V + (1-q)^2 d_a^2, \qquad q_a^* = \frac{d_a^2}{V + d_a^2}, \quad 0 \leq q \leq 1.$$

*When $V + d_a^2 = 0$ every mixture has zero risk. The anchored optimum generally requires target information and is not a free baseline. None of these exact-family formulas is a risk guarantee for nonlinear finite differences on a network.*

*Proof.* Write $X = g + N$ with $\mathbb{E}N = 0$. Cross terms with deterministic vectors vanish in the squared norm. Differentiating the resulting scalar quadratics proves the optima; insert $V = (r-1)\|g\|^2/m$ for sphere directions and $(r+1)\|g\|^2/m$ for Gaussian directions. □

### D.2 Explicit high-dimensional limits at a fixed direction count

Fix an integer $m \geq 1$. For every $r \geq 3$, let $g = e_1, u_1, \ldots, u_m$ be independent uniform directions in $\mathbb{S}^{r-1}$, and

$$X_r = \frac{1}{m}\sum_{i=1}^{m} r u_{i1} u_i, \qquad A_r = \frac{(r-1)(r-2)}{m(r+2)}, \qquad a_r = \gamma e_1 + \beta e_2,$$

where $\gamma \in \mathbb{R}$, $\beta \geq 0$ are fixed as $r$ varies. Use the candidate with parameters

$$q_r = \frac{\|X_r - a_r\|^2}{\|X_r - a_r\|^2 + A_r}, \qquad S_r = a_r + q_r(X_r - a_r),$$

so $\lambda_r = \tau_r^2 = A_r$. For a fixed $\zeta \geq 0$, set

$$H_r = I_r + \zeta r e_2 e_2^\top.$$

The case $\zeta = 0$ is isotropic. When $\zeta > 0$, the specified transverse curvature grows with dimension; this is not a theorem for all Hessian sequences. Both the anchor coordinates and direction count remain fixed.

Let $W \sim \chi^2_m$, $P = W/m$, and $q_\infty = P/(P+1)$. Define three positive scalar constants

$$a_m = \mathbb{E}[P q_\infty], \qquad b_m = \mathbb{E}[P q_\infty^2], \qquad c_m = \mathbb{E}[(1-q_\infty)^2]. \tag{36}$$

Here $a_m$ is a scalar coefficient, whereas $a_r$ denotes the anchor.

**Theorem D.3** (Fixed-batch risk, curvature, and fixed-blend limits)**.** *In the setting above,*

$$\frac{\mathcal{R}(S_r)}{\mathcal{R}(X_r)} \longrightarrow b_m < 1, \tag{37}$$

$$\frac{\mathcal{D}_{H_r}(S_r)}{\mathcal{D}_{H_r}(X_r)} \longrightarrow \frac{(1+\zeta)a_m^2(1+\gamma)_+^2}{(1+\zeta)b_m + m\zeta\beta^2 c_m}. \tag{38}$$

*The constants satisfy*

$$\begin{aligned} &\frac{1}{2} < a_m < 1, \qquad a_m^2 < b_m < 1, \qquad c_m > a_m^2, \\ &b_m = (m+2)(a_m - \tfrac{1}{2}), \qquad c_m = \frac{m}{2} - (m-1)a_m. \end{aligned} \tag{39}$$

*If $\gamma > -1$, the fixed half blend $Y_r = (X_r + a_r)/2$, with its step optimized separately, obeys*

$$\frac{\mathcal{D}_{H_r}(S_r)}{\mathcal{D}_{H_r}(Y_r)} \longrightarrow \frac{a_m^2[(1+\zeta) + m\zeta\beta^2]}{b_m(1+\zeta) + m\zeta\beta^2 c_m} < 1. \tag{40}$$

*Finally, the direct sphere Stein upper bound at these same parameters satisfies*

$$\frac{B_{\mathrm{St},r}}{\mathcal{R}(X_r)} \longrightarrow -\frac{3}{4} + \frac{1/\sqrt{2}+1/2}{\sqrt{m}}. \tag{41}$$

*For every fixed integer $m \geq 3$ the last limit is strictly negative. For any fixed $\gamma > -1$, sufficiently large finite $r$ therefore admits certified uncalibrated risk improvement while the same nonlinear candidate is worse than the fixed half blend; for zero anchor it is also worse than Raw after step optimization.*

*Proof.* **Joint law.** Couple the directions by $u_i = z_i/\|z_i\|$ with independent infinite sequences of standard normal coordinates $z_{ij}$. For each fixed $i$ and pair $i \neq j$, the strong law gives

$$\|z_i\|^2/r \to 1, \qquad r^{-1}\sum_{k=1}^{r} z_{ik}z_{jk} \to 0.$$

Consequently, with $t_i = \sqrt{r}\, u_{i1}$,

$$(X_r)_1 = \frac{1}{m}\sum_i t_i^2 \longrightarrow P, \qquad (X_r)_2 \longrightarrow \frac{1}{m}\sum_i z_{i1}z_{i2}.$$

Also

$$\frac{\|X_r\|^2}{r} = \frac{1}{m^2}\sum_i t_i^2 + \frac{2}{m^2}\sum_{i<j} t_i t_j u_i^\top u_j \longrightarrow \frac{P}{m}.$$

Conditionally on $z_{11}, \ldots, z_{m1}$, the limiting second coordinate is normal with variance $P/m$. It can therefore be written $\sqrt{P/m}\, Z$, where $Z \sim N(0,1)$ is independent of $P$. Since $\gamma, \beta$ are fixed, subtracting $a_r$ changes the normalized squared norm by a quantity tending to zero. Finally $A_r/r \to 1/m$, so

$$q_r \longrightarrow \frac{P}{P+1} = q_\infty.$$

**Moment convergence.** The following bounds justify the expectations and exclude a tail-based failure of convergence. By convexity of the fourth power of the norm,

$$\mathbb{E}\left(\frac{\|X_r\|^2}{r}\right)^2 \leq \frac{\mathbb{E}\|r u_{11}u_1\|^4}{r^2} = \frac{3r}{r+2} \leq 3.$$

Likewise $\mathbb{E}|(X_r)_1|^4 \leq 105$ and $\mathbb{E}|(X_r)_2|^4 \leq 9$, since

$$r^4\mathbb{E}u_{11}^8 = \frac{105r^3}{(r+2)(r+4)(r+6)} \leq 105, \quad r^4\mathbb{E}u_{11}^4u_{12}^4 = \frac{9r^3}{(r+2)(r+4)(r+6)} \leq 9.$$

The factors $q_r$ lie in $[0,1]$, and the anchor is uniformly bounded. Hence the first-coordinate means, squared second coordinates, and squared norms divided by $r$ used below are uniformly integrable. It follows that

$$\begin{aligned} \mathbb{E}(S_r)_1 &\to \mathbb{E}[Pq_\infty] + \gamma\mathbb{E}[1-q_\infty], \\ r^{-1}\mathbb{E}\|S_r\|^2 &\to b_m/m, \\ \mathbb{E}(S_r)_2^2 &\to b_m/m + \beta^2 c_m. \end{aligned}$$

For the last equality the cross term vanishes because $Z$ is independent and centered. For Raw, the corresponding limits are $1$, $1/m$, and $1/m$.

**Scalar identities.** Since $\mathbb{E}P = 1$,

$$a_m = \mathbb{E}\frac{P^2}{P+1} = \mathbb{E}\frac{1}{P+1} = \mathbb{E}[1-q_\infty].$$

Strict Jensen gives $a_m > 1/2$. The probability law weighted by $P$ has total mass one, and $q_\infty$ is nonconstant under it. Thus strict Cauchy–Schwarz gives $a_m^2 < b_m$; $0 < q_\infty < 1$ gives $b_m < 1$. Under

the original law, $1-q_\infty$ is nonconstant with mean $a_m$, so $c_m > a_m^2$. The Gamma integration-by-parts identity

$$\mathbb{E}[(P-1)f(P)] = \frac{2}{m}\mathbb{E}[Pf'(P)]$$

for $f(p) = (p+1)^{-1}$ yields $1-2a_m = -(2/m)(a_m - c_m)$, or $c_m = m/2-(m-1)a_m$. Polynomial division in $P^3/(P+1)^2$ then gives $b_m = -1+3a_m - c_m = (m+2)(a_m - 1/2)$. Boundary terms vanish because $f$ and $f'$ are bounded and the Gamma density has an exponential tail and shape $m/2 > 0$.

**Risk and descent.** We have $\mathcal{R}(X_r) = (r-1)/m$. The squared norm term in $\mathcal{R}(S_r)$ is of order $r$, while $\mathbb{E}(S_r)_1$ and $\|g\|^2$ stay bounded; this proves the risk limit. The transformed alignment tends to $a_m(1+\gamma)$, whereas

$$r^{-1}\mathbb{E}[S_r^\top H_r S_r] \to (1+\zeta)b_m/m + \zeta\beta^2 c_m.$$

Raw's normalized curvature energy tends to $(1+\zeta)/m$. Substitution in the separately optimized quadratic values proves (38), including nonpositive limiting alignment through the positive part. For $Y_r = (X_r + a_r)/2$, its alignment is $(1+\gamma)/2$ and normalized curvature energy tends to $[(1+\zeta)/m + \zeta\beta^2]/4$. If $\gamma > -1$, division gives (40). The two strict variance inequalities prove its strict upper bound by one.

**The risk certificate.** At unit gradient scale let $\kappa_r = 2(r-1)/(m(r+2))$, $Q_r = \|g - a_r\|^2 + (r-1)/m + A_r$, and use the exact sphere fluctuations

$$\sigma^2_{\mathrm{tr},r} = \frac{(r-2)^2(r-1)}{2(r+2)m^3}, \qquad \sigma^2_{F,r} = \frac{1}{m^3}\left[\frac{r^2}{4} + \frac{3r(r-2)}{4(r+2)} - \mathrm{tr}(\Gamma^2_{1,r})\right].$$

Here $\mathrm{tr}(\Gamma^2_{1,r}) = 4(r-1)^2/(r+2)^2 + (r-1)r^2/(r+2)^2 = O(r)$. The bound at $\lambda = \tau^2 = A_r$ is

$$B_{\mathrm{St},r} = -A_r^2/Q_r - A_r^2(A_r + 4\kappa_r)/Q_r^2 + \sigma_{\mathrm{tr},r} + \sigma_{F,r}.$$

After division by $(r-1)/m$, the first two limits are $-1/2$ and $-1/4$; the fluctuation limits are $1/\sqrt{2m}$ and $1/(2\sqrt{m})$. This proves (41). At $m = 3$, $(1/\sqrt{2} + 1/2)^2 = 3/4 + 1/\sqrt{2} < 27/16$, proving negativity there and for all larger $m$. Strict limiting inequalities imply the asserted existence of sufficiently large finite dimensions. No dimension threshold for a particular neural checkpoint follows. □

**Corollary D.4** (Explicit limiting anchor and curvature regions)**.** *Let $\gamma > -1$ and define $K_m = a_m^2(1+\gamma)^2 - b_m$. The nonlinear candidate's limiting advantage over Raw is positive exactly when*

$$(1+\zeta)K_m > m\zeta\beta^2 c_m.$$

*If $K_m > 0$ and $m\beta^2 c_m > K_m$, it is positive below and negative above*

$$\zeta_* = \frac{K_m}{m\beta^2 c_m - K_m} > 0. \tag{42}$$

*For every fixed curvature on either side, the corresponding strict sign holds for all sufficiently large finite dimensions. Both sides still have the same limiting uncalibrated risk ratio $b_m$, and for $m \geq 3$ both have negative direct risk certificates eventually. Nevertheless the fixed half blend beats the nonlinear candidate on either side asymptotically. These comparisons hold for bounded anchors, fixed $m$, and $H_r = I + \zeta r e_2 e_2^\top$.*

*Proof.* The denominator in (38) is positive. Subtracting one from that ratio leaves the sign of $(1+\zeta)K_m - m\zeta\beta^2 c_m$. Solve its affine inequality and apply the strict limit and half-blend results. □

**Isotropic scale and absolute effect size.** At zero anchor, $G_m = a_m^2/b_m < 1$. Raw's free risk-optimal scale gives $\mathcal{R}(q_0 X_r) = (r-1)/(m+r-1) \to 1$, while $\mathcal{R}(S_r) \sim r b_m/m$. Thus the nonlinear candidate's large risk reduction relative to unscaled Raw is not competitive with this simple scale baseline. In contrast, after optimal rescaling both risks tend to one, and their difference is of order $1/r$. Indeed $\mathcal{D}_I(X_r) = m/[2(r+m-1)]$. The ratio loss is nonvanishing at fixed $m$, but the absolute one-step loss is not claimed to stay bounded away from zero as $r$ grows.

**extension to Gaussian directions.** For Gaussian directions replace $A_r$ by $(r-3)/m$ and take $r > 3$. The exact representation

$$(X_r)_1 = P, \qquad (X_r)_2 = \sqrt{P/m}\, Z, \qquad \|X_r\|^2/r = \{W^2 + WV_{r-1}\}/(m^2 r),$$

uses $V_{r-1} \sim \chi^2_{r-1}$ independent of $W$. For the joint second-coordinate statement, write $V_{r-1} = Z^2 + V_{r-2}$ with $Z, V_{r-2}, W$ independent. The same limits and half-blend conclusion follow. Uniform integrability follows from Gaussian fourth moments (and the same bounded shrinkage factors). The sphere certificate does not transfer to this family. At zero anchor, the separate Gaussian direct bound yields strict risk improvement at the original strength when $m > 8$; for bounded fixed anchors its negative leading coefficient also persists for sufficiently large $r$ when $m > 8$.

### D.3 A FINITE-DIMENSIONAL REMAINDER FOR THE ZERO-ANCHOR COMPARISON

The preceding limit is supplemented by an explicit, conservative error bound; no fitted dimension threshold is needed. Retain the sphere law, zero anchor, and $\lambda = \tau^2 = A_r$. For $r \geq 3$ define

$$d_r = \frac{5r-2}{r(r+2)}, \tag{43}$$

$$t^2_{r,m} = \frac{m-1}{mr^2}\left[\frac{2r^2(r+8)}{(r+2)^2} + \frac{12(m-2)r}{r+2} + (m-2)(m-3)\right], \tag{44}$$

$$e_{r,m} = t_{r,m} + \sqrt{\frac{6}{m(r+2)}}, \tag{45}$$

$$e_a = e_{r,m} + d_r, \qquad e_b = e_{r,m} + 2d_r. \tag{46}$$

Set $t_{r,1} = 0$. The expression for $m \geq 2$ is nonnegative because it is the exact second moment identified in the proof.

**Proposition D.5** (Explicit finite-$r$ ratio brackets)**.** *Define*

$$A^{\text{eff}}_{r,m} = \mathbb{E}[q_r (X_r)_1], \qquad B^{\text{eff}}_{r,m} = (m/r)\mathbb{E}[q_r^2 \|X_r\|^2].$$

*Then*

$$|A^{\text{eff}}_{r,m} - a_m| \leq e_a, \qquad |B^{\text{eff}}_{r,m} - b_m| \leq e_b. \tag{47}$$

*If $b_m > e_b$, then*

$$\frac{r+m-1}{r}\frac{(a_m - e_a)_+^2}{b_m + e_b} \leq \frac{\mathcal{D}_I(S_r)}{\mathcal{D}_I(X_r)} \leq \frac{r+m-1}{r}\frac{(a_m + e_a)^2}{b_m - e_b}. \tag{48}$$

*Also*

$$\frac{\mathcal{R}(S_r)}{\mathcal{R}(X_r)} \leq \frac{r(b_m + e_b) - 2m(a_m - e_a)_+ + m}{r-1}. \tag{49}$$

*Thus $B_{\text{St},r} < 0$ together with an upper endpoint in (48) below one gives a finite-dimensional, same-parameter risk certificate and optimized-descent deterioration. The bounds are $O_m(r^{-1/2})$ perturbations; they are sufficient, not tight dimension thresholds.*

*Proof.* Put $p = (X_r)_1$ and $T = \|X_r\|^2/r$. If $Z_i = r u_{i1} u_i$, the exact identity

$$\delta := p - mT = -\frac{2}{mr}\sum_{i<j} Z_i^\top Z_j$$

follows because $\|Z_i\|^2 = r(Z_i)_1$. For $M = \mathbb{E} Z_i Z_i^\top = r(I + 2e_1 e_1^\top)/(r+2)$, independence gives

$$\mathbb{E}(Z_i^\top Z_j)^2 = \operatorname{tr}(M^2) = \frac{r^2(r+8)}{(r+2)^2}, \quad \mathbb{E}[(Z_i^\top Z_j)(Z_i^\top Z_k)] = \frac{3r}{r+2}$$

for distinct $i, j, k$. Products of disjoint pairs have mean one. There are $m(m-1)/2$ edges, $3\binom{m}{3}$ unordered overlapping edge pairs, and $3\binom{m}{4}$ disjoint pairs. Expanding the square proves $\mathbb{E}\delta^2 = t^2_{r,m}$, including zero at $m = 1$.

Independently attach $R_i \sim \chi^2_r$ to each sphere direction. Then $\sqrt{R_i}u_i$ are independent standard Gaussian vectors. Their Gaussian-update first coordinate $p_G$ has law $P = W/m$ and

$$p_G - p = \frac{1}{m}\sum_i (R_i/r - 1)(Z_i)_1.$$

Conditionally centered independent radii give

$$\mathbb{E}(p_G - p)^2 = \frac{1}{m}\frac{2}{r}\mathbb{E}(Z_i)_1^2 = \frac{6}{m(r+2)}.$$

This coupling is used to bound moments, not to replace the sphere update in the risk theorem.

Write $k = A_r/r$ and $\ell = mk = 1 - d_r > 0$. For $p, T \geq 0$,

$$\left| p\frac{T}{T+k} - \frac{p^2}{p+\ell} \right| = \frac{pk|T - p/m|}{(T+k)(p/m+k)} \leq |\delta|.$$

The map $p^2/(p+\ell)$ is one-Lipschitz in $p \geq 0$ and in $\ell > 0$. Hence comparison with $p_G$ and $\ell = 1$, followed by Cauchy–Schwarz, gives the first bound in (47).

For the second, the function $f_k(T) = T^3/(T+k)^2$ has derivative $T^2(T+3k)/(T+k)^3 \in [0,1]$. Therefore

$$|mf_k(T) - mf_k(p/m)| \leq |\delta|, \qquad mf_k(p/m) = \frac{p^3}{(p+\ell)^2}.$$

The latter function is one-Lipschitz in $p$ and two-Lipschitz in $\ell$. The same coupling gives the second bound. The exact descent ratio is

$$\frac{\mathcal{D}_I(S_r)}{\mathcal{D}_I(X_r)} = \frac{r+m-1}{r}\frac{(A^{\text{eff}}_{r,m})^2}{B^{\text{eff}}_{r,m}}.$$

Its numerator is nonnegative because $p, q_r \geq 0$, and its denominator is positive. Interval division proves (48). Finally $\mathcal{R}(S_r) = rB^{\text{eff}}_{r,m}/m - 2A^{\text{eff}}_{r,m} + 1$, which proves (49). □

### D.4 Deterministic finite-dimensional enclosures

The constants can be enclosed without Monte Carlo:

$$a_m = \int_0^\infty e^{-x}(1+2x/m)^{-m/2}\,dx, \quad b_m = (m+2)(a_m - 1/2), \quad c_m = m/2 - (m-1)a_m. \tag{50}$$

The integral follows by applying Tonelli to $(1+P)^{-1} = \int_0^\infty e^{-x(1+P)}dx$ and the Gamma Laplace transform. The supplied script encloses it with interval arithmetic and composite Simpson integration on $[0, 32]$ with 4096 equal subintervals. With $k = m/2$, the fourth derivative of $f(x) = e^{-x}(1 + x/k)^{-k}$ satisfies

$$|f^{(4)}(x)| \leq C_{4,k} := \sum_{j=0}^{4}\binom{4}{j}\frac{(k)_j}{k^j}, \qquad x \geq 0.$$

This follows by the product rule; every term's absolute value is largest at zero. The Simpson error is at most $32(32/4096)^4 C_{4,k}/180$, and the remaining tail is between zero and $e^{-32}$. Interval operations propagate these deterministic allowances through the displayed formulas. No Monte Carlo error is treated as an outward enclosure.

Table 4 reports the corresponding limits and finite-dimensional certificates.

Table 4: Deterministic limits for risk and expected descent and sufficient finite-dimensional examples for the same zero-anchor candidate. Every finite row also has $B_{\text{St}} < 0$. Limit entries are rounded; ratio bounds use outward enclosures, not Monte Carlo or a sample-split audit.

| $m$ | $b_m$ (approx.) | $a_m^2/b_m$ (approx.) | Sufficient $r$ | Ratio upper bound |
|---|---|---|---|---|
| 4 | 0.328113 | 0.937713 | 131072 | $< 0.981347$ |
| 8 | 0.292523 | 0.957560 | 262144 | $< 0.987756$ |
| 16 | 0.272285 | 0.974551 | 524288 | $< 0.995074$ |

At these dimensions the respective Stein upper bounds are less than $-4797.66$, $-10590.53$, and $-14686.68$. They concern unit-gradient, exact quadratic laws; the large absolute risk scale is $\mathcal{R}(X_r) = (r-1)/m$. These examples certify finite nonzero ratio losses, not useful neural optimization gains or sharp required dimensions.

For $\gamma = 1/2$, $\beta = \sqrt{3}/2$, the leading-order expression $1 + 64\zeta_*$ predicts transverse curvature values approximately $37.2080$ at $m = 4$ and $8.04007$ at $m = 16$. The previously reported sphere thresholds are $38.20499$ and $8.20470$. This is a post hoc asymptotic comparison to the already observed finite-$r$ data, not an independent prediction or a finite-$r$ confidence interval.

### D.5 When can a scalar weight improve expected descent?

**Proposition D.6** (Conditional signal-to-curvature weighting)**.** *Let $\mathcal{T}$ be information computed from a square-integrable update $X$ with $\mathbb{E}[X^\top HX] > 0$. Define $a(\mathcal{T}) = \mathbb{E}[g^\top X \mid \mathcal{T}]$ and $b(\mathcal{T}) = \mathbb{E}[X^\top HX \mid \mathcal{T}]$. Over nonnegative $\mathcal{T}$-measurable weights $q$ with finite nonzero output energy,*

$$\sup_q \mathcal{D}_H(qX) = \frac{1}{2}\mathbb{E}\left[\frac{a(\mathcal{T})_+^2}{b(\mathcal{T})}\right], \tag{51}$$

*with $0/0 = 0$ on $b = 0$. If the right side is positive, $q = a_+/b$ attains it; multiplication by any positive constant leaves its value unchanged. The supremum is also the same when weights are restricted to $0 \le q \le 1$, although an optimum need not then be attained. These are information-specific oracle benchmarks, not a cost-free implementable selector.*

*Proof.* Conditional Cauchy–Schwarz gives $a^2 \le (g^\top H^{-1} g)b$, so $a = 0$ on $b = 0$ and the expectation on the right is finite. For $q \ge 0$,

$$(\mathbb{E}qa)_+ \le \mathbb{E}qa_+ \le (\mathbb{E}q^2 b)^{1/2}\left(\mathbb{E}\frac{a_+^2}{b}\right)^{1/2}.$$

Equality holds for $q = a_+/b$, whose curvature energy and alignment both equal $\mathbb{E}[a_+^2/b]$. For bounded weights set $q_n = \min\{a_+/b, n\}/n$. Scale invariance allows removal of the final factor $1/n$; monotone convergence of alignment and energy proves the same supremum. If the right side is zero every nonnegative weight has nonpositive alignment and zero optimized decrease. □

When $b > 0$, define the probability law $d\nu = b\,dP/\mathbb{E}b$ and $z = a/b$. If $a \ge 0$ almost surely, the ratio to the optimum is the squared $L^2(\nu)$ cosine between $q$ and $z$ (for nonzero quantities). In particular, if $z$ is a positive constant, every nonconstant $q$ reduces optimized descent by $1/(1 + \mathrm{CV}_\nu(q)^2)$. Sample dependence is useful only when it predicts conditional signal relative to conditional curvature cost. The single-direction identity and the limit with a fixed number of directions instantiate a setting in which radial variation does not supply that information. This Cauchy–Schwarz characterization provides a benchmark for the selected weighting rule.

### D.6 What nonlinear information remains beyond a fixed anchor mixture?

Use the Hilbert inner product $\langle Y, Z\rangle_H = \mathbb{E}[Y^\top HZ]$ on square-integrable random vectors, modulo almost-sure equality. Let $\mathcal{V} = \mathrm{span}\{X, a\}$ with deterministic real coefficients, and let $S$ be any square-integrable nonlinear candidate. Define $P = \Pi_{\mathcal{V}} S$, $U = S - P$, $t = H^{-1}g$, and $V_* = \Pi_{\mathcal{V}} t$. Finite-dimensional spans are closed even if the generators are linearly dependent.

**Proposition D.7** (Residual alignment beyond the linear mixture space)**.** *For $J(Y) = \mathbb{E}[g^\top Y] - \mathbb{E}[Y^\top HY]/2$,*

$$\sup_{Y \in \mathrm{span}(\mathcal{V}, S)} J(Y) - \sup_{Y \in \mathcal{V}} J(Y) = \frac{\ell_U^2}{2v_U}, \quad \ell_U = \mathbb{E}[g^\top U], \quad v_U = \mathbb{E}[U^\top HU], \tag{52}$$

*where the ratio is zero if $U = 0$ almost surely. More precisely, for every $\eta \ge 0$,*

$$J(\eta S) - J(V_*) = -\tfrac{1}{2}\|\eta P - V_*\|_H^2 + \eta\ell_U - \tfrac{1}{2}\eta^2 v_U. \tag{53}$$

*Thus the actual candidate has no advantage over the best unrestricted fixed linear mixture when $\ell_U \le 0$. In general its advantage is at most $(\ell_U)_+^2/(2v_U)$. If $V_*$ belongs to the nonnegative cone $\{cX + da : c, d \ge 0\}$, these comparisons also apply to the jointly tuned feasible fixed blend. Otherwise the unrestricted result alone does not compare against that cone.*

*Proof.* Completing the square gives $J(Y) = \|t\|_H^2/2 - \|Y - t\|_H^2/2$. Its maximizer over $\mathcal{V}$ is $V_*$. If $U \neq 0$, then $\operatorname{span}(\mathcal{V}, S) = \mathcal{V} \oplus \operatorname{span}\{U\}$ with orthogonal summands. Projecting $t$ on the additional component gives squared projection norm $\langle t, U\rangle_H^2/\|U\|_H^2 = \ell_U^2/v_U$, proving (52). For a candidate on the ray $\eta S = \eta P + \eta U$, the orthogonal decomposition and $\langle t, P\rangle_H = \langle V_*, P\rangle_H$ give (53). Maximizing only the final scalar quadratic over $\eta \geq 0$ gives the upper bound. If $V_*$ is feasible for the cone, the cone and unrestricted optima coincide; no such conclusion follows without feasibility. If $U = 0$, the exact formula reduces to a nonpositive squared distance. □

The projection can be computed from the Gram matrix of $X, a$ using a pseudoinverse, with its sampling and curvature-information cost recorded. It is a diagnostic upper envelope for the information added by $S$, not a claim that the selected ridge attains the enlarged-space optimum. The strict half-blend result above is stronger within its particular ZO limit: it compares against a fixed coefficient $1/2$, with the same optimization over a single step size and no extra coefficient search.

# E Finite-sample local joint inference

This appendix proves the finite-sample local joint result. The setting is one fixed deterministic quadratic objective, or a controlled family sharing its gradient at that state. The protocol joins a risk statement, a prediction, fitted steps, and a fresh evaluation at that state. The confidence event is over the finite audit pools, not over a moving training trajectory.

## E.1 Complete local statement

The preceding conditions involve population quantities. We now connect risk calibration $\mathcal{C}$, independent moment profiling $\mathcal{O}$, step fitting $\mathcal{F}$, and independent paired evaluation $\mathcal{E}$ at one fixed quadratic state. The completed quadratic audit is reported in Section 7 and Appendix H.2; the network experiment uses a different, approximate inferential procedure.

On a valid scale event $s \leq s_U$, use $B_X = r\sqrt{s_U}$ and $B_S = \min\{\max(B_X, \|a\|), B_X + \lambda/(2\tau)\}$. For $K$ candidates frozen before $N$ independent batched updates, the four profiling variables have range widths $2B_S, B_S^2, B_S^2, B_S^2$. A variable of width $W$ has simultaneous radius $W\sqrt{\log(8K/\delta_M)/(2N)}$. The resulting intervals give

$$\kappa_{j,L} = (\alpha_L)_+^2 C_j - u_{j,U}, \qquad \kappa_{j,U} = (\alpha_U)_+^2 C_j - u_{j,L}.$$

Their positive-weight sums bound the sign functional for *every* $h > 0$, not only a tested grid. When expected alignment and energy are positive, signed division gives lower bounds $\gamma_+(h)$ or $\gamma_-(h)$ on the absolute optimal-step advantage. Appendix E.2 specifies these bounds, support intersections, and abstention conditions; Appendix E.5 provides a sharper paired-statistic alternative.

Fitted steps also need control. Completing the square gives regret $\operatorname{reg}_Y = \frac{1}{2} v_Y(\widehat{\eta}_Y - a_Y/v_Y)^2$. Valid alignment and energy intervals bound this by $\mathfrak{r}_Y$ for any subsequently frozen step. The achieved contrast is $\Delta_D^\eta = \Delta_D^* - \operatorname{reg}_S + \operatorname{reg}_X$. Thus positive predictions require $\mathfrak{r}_S \leq \gamma_+/2$, whereas negative predictions require $\mathfrak{r}_X \leq \gamma_-/2$. These guards yield achieved margins of at least half the corresponding predicted margin; fitting is not assumed to recover the population optimum.

**Theorem E.1** (Local joint confidence and independent sign confirmation)**.** *Fix the quadratic sphere model. Use simultaneous risk-calibration and profiling events with failure probabilities $\delta_C, \delta_M$. Candidates must be frozen before profiling. Retain only valid $U_R < 0$ candidates, select at most $J$ candidate–curvature comparisons in the signed regions, and freeze nonnegative steps satisfying the preceding regret guards before opening $\mathcal{E}$. Let $\gamma_\eta > 0$ bound the achieved absolute margin.*

*For independent evaluation pairs, put $Z_e = F_h(-\widehat{\eta}_X X_e) - F_h(-\widehat{\eta}_S S(X_e))$ and use*

$$C_E = \sum_{Y \in \{X,S\}} \left[\widehat{\eta}_Y \sqrt{s_U} B_Y + \tfrac{1}{2}\widehat{\eta}_Y^2 \|H(h)\|_{\mathrm{op}} B_Y^2\right], \quad r_E = C_E \sqrt{\frac{2\log(2J/\delta_E)}{n_E}}.$$

*With probability at least $1 - \delta_C - \delta_M - \delta_E$, all emitted risk and achieved-sign statements are correct, and every $[\bar{Z} - r_E, \bar{Z} + r_E]$ covers its frozen-step expectation. It excludes zero with the predicted sign when $2r_E < \gamma_\eta$; a sufficient size is*

$$n_E > 8C_E^2 \log(2J/\delta_E)/\gamma_\eta^2. \tag{54}$$

The subsequent subsections prove the statement, explicit sample bounds, and a version without known gradient-aligned coordinates. Curvatures may be selected after profiling because the band is uniform; candidates cannot be changed without new control. This is simultaneous validity, not posterior coverage conditional on acceptance. Practical acceptance rates, unknown curvature acquisition, and reuse along a moving trajectory remain separate questions.

### E.2 Explicit profiling bands and fitting guards

On the calibration event $s \le s_U$, sphere updates satisfy $\|X\| \le B_X = r\sqrt{s_U}$. Since the shrinker is a convex blend and its displacement is at most $\lambda/(2\tau)$, a valid output bound is

$$B_S = \min\{\max(B_X, \|a\|),\ B_X + \lambda/(2\tau)\}. \tag{55}$$

Freeze $K$ candidates before $N$ iid batched updates in $\mathcal{O}$. For each candidate profile $S_1, S_1^2, S_2^2, \|S_{3:r}\|^2$, whose interval widths are $2B_S, B_S^2, B_S^2, B_S^2$. Put $P = 4K$. For a profiled variable of width $W_j$, use

$$e_j = W_j\sqrt{\frac{\log(2P/\delta_M)}{2N}}. \tag{56}$$

These give simultaneous intervals $\alpha \in [\alpha_L, \alpha_U]$ and $u_j \in [u_{j,L}, u_{j,U}]$ with conditional failure at most $\delta_M$. A specified empirical-Bernstein alternative with constants, and a paired-correction construction that cancels shared raw energy, are in Appendix E. Choosing the smaller of independently valid confidence radii without allocating confidence to both is not permitted.

Set

$$\begin{aligned} \kappa_{j,L} &= (\alpha_L)_+^2 C_j - u_{j,U}, \qquad \kappa_{j,U} = (\alpha_U)_+^2 C_j - u_{j,L}, \\ \Lambda_L(h) &= \sum_j h_j \kappa_{j,L}, \qquad \Lambda_U(h) = \sum_j h_j \kappa_{j,U}. \end{aligned} \tag{57}$$

These bands hold simultaneously for *every* positive $h$, without a union bound over a curvature grid. In the region with positive expected alignment $\alpha_L > 0$ with $u_L(h) > 0$, define

$$\gamma_+(h) = \frac{s_L \Lambda_L(h)}{2C(h)u_U(h)} \qquad \text{if } \Lambda_L(h) > 0, \tag{58}$$

$$\gamma_-(h) = \frac{-s_L \Lambda_U(h)}{2C(h)u_U(h)} \qquad \text{if } \Lambda_U(h) < 0. \tag{59}$$

Then $\Delta_D^*(h) \ge \gamma_+(h)$ in the first region and $\Delta_D^*(h) \le -\gamma_-(h)$ in the second. For any prespecified tolerance $\gamma_0 > 0$, retain only curvatures with $\gamma_+(h) \ge \gamma_0$ or $\gamma_-(h) \ge \gamma_0$ to obtain positive- and negative-margin regions. A region can be empty when the uncertainty is too large. Strict signs are not inferred from a point estimate alone. A seven-statistic, fixed-projector extension in Appendix E.6 avoids assuming the gradient-aligned coordinates are observable.

**Step-fitting guards.** For any law with alignment $a_Y > 0$ and curvature energy $v_Y > 0$, the loss of expected decrease from using a fitted step is exactly

$$\mathcal{D}_H(Y) - d_H(Y; \widehat{\eta}_Y) = \tfrac{1}{2} v_Y (\widehat{\eta}_Y - a_Y/v_Y)^2. \tag{60}$$

If intervals give $a_Y \in [a_{Y,L}, a_{Y,U}]$, $v_Y \in [v_{Y,L}, v_{Y,U}]$ with positive lower endpoints, then $\eta_Y^* \in I_Y = [a_{Y,L}/v_{Y,U}, a_{Y,U}/v_{Y,L}]$ and

$$\mathfrak{r}_Y = \tfrac{1}{2} v_{Y,U} \max_{t \in \partial I_Y} (\widehat{\eta}_Y - t)^2$$

is a valid regret bound for *any* later frozen step. For the aligned sphere raw law $\eta_X^* = 1/C(h)$ exactly. An independent fitting pool can be used, but no fitting quality is assumed unless these guards pass. For a positive prediction require $\mathfrak{r}_S \le \gamma_+/2$; for a negative prediction require $\mathfrak{r}_X \le \gamma_-/2$. The achieved difference at the frozen steps then has the predicted sign and absolute margin at least $\gamma_+/2$ or $\gamma_-/2$, respectively.

### E.3 Conditional confidence intervals for bounded means

**Lemma E.2** (Simultaneous confidence intervals for bounded means)**.** *Condition on all information preceding a profiling pool. Let $Z_{ij}$, $i = 1, \dots, N$, $j = 1, \dots, P$, be iid in $i$; dependence across $j$ is unrestricted. Suppose $Z_{ij} \in [\ell_j, u_j]$ almost surely with $W_j = u_j - \ell_j$ fixed under the conditioning. Let $\bar{Z}_j = N^{-1} \sum_i Z_{ij}$. Then with conditional probability at least $1 - \delta$, all means $\mu_j = \mathbb{E} Z_{ij}$ obey*

$$|\bar{Z}_j - \mu_j| \le W_j \sqrt{\frac{\log(2P/\delta)}{2N}}. \tag{61}$$

*For $N \ge 2$, an alternative simultaneous radius is*

$$e_j^{\mathrm{EB}} = \sqrt{\frac{2\widehat{v}_j \log(4P/\delta)}{N}} + \frac{7W_j \log(4P/\delta)}{3(N-1)}, \qquad \widehat{v}_j = \frac{1}{N-1} \sum_i (Z_{ij} - \bar{Z}_j)^2. \tag{62}$$

*The concentration rule must be selected before the pool, or confidence must be allocated across all rules used. No independence of statistics for different candidates or curvatures is needed.*

*Proof.* For a bounded scalar variable, the log moment-generating function of its centered version has second derivative equal to a variance under exponential tilting. Its support still has width $W$, so this variance is at most $W^2/4$. Integrating twice from zero gives $\log \mathbb{E} e^{t(Z - \mathbb{E}Z)} \le t^2 W^2/8$. Independence in $i$, exponential Markov, and optimization in $t$ yield $\Pr(|\bar{Z} - \mathbb{E}Z| > e) \le 2 \exp(-2Ne^2/W^2)$. A union bound over $P$ variables proves (61). For $W = 0$ the assertion is deterministic.

For (62), rescale each variable to $[0, 1]$ and apply Theorem 4 of Maurer & Pontil (2009) to it and its negative (equivalently one minus the rescaled variable), each at failure probability $\delta/(2P)$. The theorem uses the unbiased sample variance and logarithm $\log(2/\delta')$, yielding $\log(4P/\delta)$. Rescaling and the union bound give the stated radius. To use the minimum of the Hoeffding and Bernstein radii, apply each with failure allocation $\delta/2$, rather than using both at $\delta$. □

The range bound may depend on an independent calibration output. It is then valid conditional on that output only on its coverage event. The total failure calculation must include the calibration event's complement. A sample maximum from the same profiling pool is not an a priori range bound.

### E.4 Four-moment bands in gradient-aligned coordinates

Fix coordinates with $g = Ge_1$, $G > 0$, and $a$ in the first two coordinates. Conditional on a successful scale calibration, $G^2 \in [s_L, s_U]$. Equation (55) follows because $\|Z_i\| \le rG$, averaging preserves the norm bound, and

$$S = q(X)X + (1 - q(X))a, \quad 0 \le q(X) \le 1, \quad \|S - X\| \le \frac{\lambda}{2\tau}.$$

For $K$ frozen candidates, apply Lemma E.2 to their $4K$ statistics. Intersect every energy interval with $[0, B_S^2]$ and the alignment interval with $[-B_S, B_S]$. Empty intersections cause abstention, not an assertion. On the moment coverage event, the intervals contain the corresponding population means.

Because $t \mapsto (t_+)^2$ is nondecreasing, for each block

$$(\alpha_L)_+^2 C_j - u_{j,U} \le (\alpha_+)^2 C_j - u_j \le (\alpha_U)_+^2 C_j - u_{j,L}.$$

Multiplying by arbitrary positive $h_j$ and summing gives (57) simultaneously for the whole positive curvature cone. This is finite-dimensional propagation of simultaneous intervals, not a separate test for each $h$.

Let $\Lambda(h) = \sum h_j \kappa_j$. Since

$$\Delta_D^*(h) = \frac{s\Lambda(h)}{2C(h)u(h)},$$

$\Lambda_L(h) > 0$ implies $\Delta_D^*(h) \ge s_L \Lambda_L(h)/(2C(h)u_U(h))$. If $\Lambda_U(h) < 0$, its numerator is negative; using $s \ge s_L$ and $u(h) \le u_U(h)$ instead gives $\Delta_D^*(h) \le s_L \Lambda_U(h)/(2C(h)u_U(h))$. This proves the signed margins in (59). The $s_L$ in the negative upper bound is intentional. A large denominator makes a negative ratio less negative, as required for an upper bound.

For general symmetric $H$ in the same aligned coordinate system, include $S_1S_2$ as a fifth statistic. Its support lies in $[-B_S^2/2, B_S^2/2]$. Intervals for this moment, combined with the sign of $H_{12}$ and the nonnegative diagonal entries of positive definite $H$, give a simultaneous sign band. The resulting interval calculation is not simply the positive-weight sum in (57). The main statement restricts to the diagonal curvature family for clarity.

**An explicit sufficient profiling size.** For one candidate, let $B = B_S$, $\ell = \log(2P/\delta_M)$, and suppose the population signal satisfies $\alpha > 0$, the curvature energy is $u(h) > 0$, and the sign-functional margin is $|\Lambda(h)| \geq \zeta > 0$. The Hoeffding intervals, intersected with their support, obey on their joint event

$$\Lambda(h) - \Lambda_L(h), \quad \Lambda_U(h) - \Lambda(h) \leq 2B^2[4C(h) + h_1 + h_2 + h_o]\sqrt{\ell/(2N)}.$$

Indeed, a true mean and either confidence endpoint differ by at most twice its radius; $x \mapsto (x_+)^2$ is $2B$-Lipschitz on $[-B, B]$. Thus a sufficient size for a half-margin signed region is

$$N \geq \frac{8B^4[4C(h) + h_1 + h_2 + h_o]^2\ell}{\zeta^2}.$$

Also, $N \geq 32B^2\ell/\alpha^2$ gives $\alpha_L \geq \alpha/2$, and $N \geq 8B^4(h_1 + h_2 + h_o)^2\ell/u(h)^2$ gives $u_L(h) \geq u(h)/2$. These are sufficient population-margin statements, not instructions to substitute a selected sample mean for an unknown margin. They expose why worst-case range bounds can require many samples. The empirical-Bernstein and paired-correction alternatives address some conservatism but do not remove acquisition costs.

### E.5 A PAIRED-CORRECTION ALTERNATIVE

The full energy ranges can be conservative. The raw distribution supplies information that should not be counted as independent uncertainty twice. Let $D_0 = \lambda/(2\tau)$ and profile the four paired statistics

$$Z_0 = (S - X)_1, \qquad Z_j = \|P_jS\|^2 - \|P_jX\|^2, \quad j \in \{1, 2, o\}. \tag{63}$$

Here $P_1, P_2, P_o$ are the three orthogonal block projections. Their widths can be bounded by

$$W_0 = 2D_0, \qquad W_j = 2D_E, \qquad D_E = 2B_XD_0 + D_0^2. \tag{64}$$

Indeed, $|Z_0| \leq D_0$ and $|\|P_j(X + (S - X))\|^2 - \|P_jX\|^2| \leq 2B_XD_0 + D_0^2$. Apply either specified concentration rule to these statistics, obtaining $d := \mathbb{E}(S - X)_1 \in [d_L, d_U]$ and $w_j := \mathbb{E}Z_j \in [w_{j,L}, w_{j,U}]$. Let $G_L = \sqrt{s_L}$ and $G_U = \sqrt{s_U}$.

**Proposition E.3** (Paired bands without duplicating raw energy)**.** *If $G_L + d_L > 0$, define*

$$t_L = d_L^2 + 2d_L \begin{cases} G_L, & d_L \geq 0, \\ G_U, & d_L < 0, \end{cases} \tag{65}$$

$$t_U = d_U^2 + 2d_U \begin{cases} G_U, & d_U \geq 0, \\ G_L, & d_U < 0. \end{cases} \tag{66}$$

*Then simultaneous valid bands are*

$$\alpha \in [G_L + d_L, G_U + d_U], \tag{67}$$

$$u_j \in [\max\{0, s_LC_j + w_{j,L}\},\ s_UC_j + w_{j,U}], \tag{68}$$

$$\kappa_j \in [t_LC_j - w_{j,U},\ t_UC_j - w_{j,L}]. \tag{69}$$

*Energy upper endpoints may also be intersected with $B_S^2$. These intervals can replace the full-moment intervals in the local joint theorem. Select the representation before profiling, or allocate confidence to both representations if combining them.*

*Proof.* The mean alignment is $\alpha = G + d$ and the block energy is $u_j = sC_j + w_j$. Thus

$$\kappa_j = \alpha^2C_j - u_j = (2Gd + d^2)C_j - w_j.$$

Over the rectangle $G \in [G_L, G_U]$, $d \in [d_L, d_U]$, the derivative with respect to $d$ is $2(G + d) > 0$ by the hypothesis. The minimum therefore occurs at $d_L$ and the maximum at $d_U$. At a fixed $d$, the expression is linear in $G$ with sign determined by $d$, which gives (66). The other formulas follow by interval propagation. Correlations among the estimates do not invalidate containment; the formulas do not assume independent confidence endpoints. □

This construction can reduce the range scale when the correction is small. The mathematical construction alone does not establish passage at a given sample size; the completed audit reports each stage separately in Appendix H.2. In particular, $B_X$ still depends on the ambient dimension and the calibrated target scale. An empirical-Bernstein radius may exploit much smaller observed variance, but its range term and confidence accounting remain.

### E.6 An extension without an observed gradient direction

The four-moment theorem uses the gradient-aligned coordinates available in a mechanism study. The following version avoids treating that coordinate information as free. Fix orthogonal projections $P_1, \dots, P_J$ before profiling, with $\sum_j P_j = I$, and consider the known curvature family $H(h) = \sum_j h_j P_j$, $h_j > 0$. It need not align with $g$. Define

$$a_S = \mathbb{E}[g^\top S], \quad c_j = \mathbb{E}\|P_j X\|^2, \quad u_j = \mathbb{E}\|P_j S\|^2, \quad s = \|g\|^2.$$

The raw alignment is $s$. On a deterministic quadratic, each sample $g^\top S$ is observable by an additional pair $[F(\epsilon S) - F(-\epsilon S)]/(2\epsilon)$; full recovery of $g$ is unnecessary. The observed variable lies in $[-\sqrt{s_U}B_S, \sqrt{s_U}B_S]$ on the calibration event. Raw and transformed block energies lie in $[0, B_X^2]$ and $[0, B_S^2]$. Thus $1 + 2J$ confidence intervals for bounded means per candidate can be obtained by Lemma E.2. For $J = 3$, this is seven statistics instead of four, and the slope pair must be charged.

Let $a_S \in [a_L, a_U]$, $c_j \in [c_{j,L}, c_{j,U}]$, and $u_j \in [u_{j,L}, u_{j,U}]$. Write $c_L(h) = \sum h_j c_{j,L}$ and similarly for the other energies. The exact numerator of the tuned difference is

$$F(h) = (a_S)_+^2 c(h) - s^2 u(h).$$

It has simultaneous linear bands

$$F_L(h) = (a_L)_+^2 c_L(h) - s_U^2 u_U(h), \tag{70}$$

$$F_U(h) = (a_U)_+^2 c_U(h) - s_L^2 u_L(h). \tag{71}$$

If $c_L(h), u_L(h) > 0$, $F_L(h) > 0$ gives a lower bound $F_L(h)/(2c_U(h)u_U(h))$ on the tuned advantage, and $F_U(h) < 0$ gives the upper bound $F_U(h)/(2c_U(h)u_U(h))$. The proof is interval containment followed by the same signed division as before. These intervals can be less sharp because they do not use the special raw block formula.

This is a function-value route for one fixed quadratic and a prescribed curvature representation. Curvature coefficients must themselves be known or measured. An arbitrary unknown Hessian does not become observable from seven statistics, and different points cannot be pooled as iid observations of one fixed gradient. For a finite list of individual curvatures, directly profiling each $Y^\top HY$ is also possible, with the corresponding confidence and acquisition costs.

### E.7 Step-fitting regret and the achieved-sign guards

Let $d_H(Y;\eta) = \eta a_Y - \eta^2 v_Y/2$. If $a_Y > 0$, completing the square gives

$$d_H(Y;\eta) = \frac{a_Y^2}{2v_Y} - \frac{v_Y}{2}(\eta - a_Y/v_Y)^2.$$

If $a_Y \in [a_L, a_U]$ and $v_Y \in [v_L, v_U]$ with positive lower endpoints, then $\eta_Y^* \in [a_L/v_U, a_U/v_L]$. For every later chosen $\widehat{\eta}_Y$, the largest squared distance from that interval is attained at an endpoint. This proves the bound $\mathfrak{r}_Y$ in the complete statement above. For the aligned sphere law, $a_X = s$, $v_X = sC(h)$, so its optimal step is exactly $1/C(h)$ regardless of the unknown scale, and

$$\mathfrak{r}_X = \tfrac{1}{2} s_U C(h)(\widehat{\eta}_X - 1/C(h))^2. \tag{72}$$

For the transformed law in aligned coordinates, use $a_S \in [\sqrt{s_L}\alpha_L, \sqrt{s_U}\alpha_U]$ when $\alpha_L > 0$, and $v_S \in [u_L(h), u_U(h)]$. In the coordinate-free extension use $a_X \in [s_L, s_U]$, $v_X \in [c_L(h), c_U(h)]$, and its separate signal/energy intervals for $S$.

Let $\mathrm{reg}_Y = \mathcal{D}_H(Y) - d_H(Y;\widehat{\eta}_Y) \geq 0$. The achieved comparison obeys the identity

$$\Delta_D^\eta = \Delta_D^* - \mathrm{reg}_S + \mathrm{reg}_X. \tag{73}$$

Therefore a positive population-optimum margin needs control of $S$'s fitting regret; a negative margin needs control of raw's fitting regret. Controlling the wrong method would not suffice. If $\Delta_D^* \geq \gamma_+ > 0$, then $\Delta_D^\eta \geq \gamma_+ - \mathfrak{r}_S$. If $\Delta_D^* \leq -\gamma_- < 0$, then $\Delta_D^\eta \leq -\gamma_- + \mathfrak{r}_X$. The half-margin guards in the complete statement follow. They remain valid if the steps are selected adaptively using any non-evaluation data, since the profile intervals control every chosen step. An independent fitting pool is a protocol choice that matches the intended experiment, not an unproved assumption that fitting recovers the optimum. Alternatively, an interval midpoint step can be chosen without a fitting pool and checked by the same regret bound.

### E.8 Independent evaluation and the joint event

After calibration, profiling, selection, and fitting, condition on all of those data. Every evaluated candidate, curvature and step is now fixed. The evaluation pairs $X_e, S(X_e)$ are iid across $e$. For each $Y \in \{X, S\}$,

$$|F_h(-\widehat{\eta}_Y Y) - F_h(0)| \leq \widehat{\eta}_Y \sqrt{s_U} B_Y + \tfrac{1}{2}\widehat{\eta}_Y^2 h_{\max} B_Y^2.$$

Thus $Z_e \in [-C_E, C_E]$ and the Hoeffding radius for width $2C_E$ is exactly $r_E$ in Theorem E.1. A union bound over the selected finite list of at most $J$ comparisons gives simultaneous coverage at conditional failure at most $\delta_E$. The list may have been selected from a continuum of curvatures using the uniform profiling event. It cannot be selected using the evaluation values themselves.

For a predicted positive achieved margin $\Delta_D^\eta \geq \gamma_\eta$, on the coverage event,

$$\bar{Z} - r_E \geq \Delta_D^\eta - 2r_E \geq \gamma_\eta - 2r_E > 0.$$

The negative case uses $\bar{Z} + r_E \leq -\gamma_\eta + 2r_E < 0$. Solving $2r_E < \gamma_\eta$ gives (54). This factor of two is necessary to guarantee that the confidence interval itself, not only the sample mean, excludes zero. This is a sufficient, often conservative power bound, not a promised experimental sample count.

Let $\mathcal{G}_C$ be the simultaneous calibration event. For every calibration realization in $\mathcal{G}_C$, the bounds on the profiling variables hold and the conditional profiling failure is at most $\delta_M$. For every preceding-data realization in the calibration and profiling good events, the evaluation failure is at most $\delta_E$. Integrating these conditional statements gives

$$\Pr(\mathcal{G}_C \cap \mathcal{G}_M \cap \mathcal{G}_E) \geq 1 - \delta_C - \delta_M - \delta_E.$$

On this event, Corollary B.1, the uniform curvature bands, the fitting-regret identities, and the evaluation interval statements all hold. This proves Theorem E.1. No bound on the fitting pool's failure is needed: a poor fit merely fails the explicit guard. This is an unconditional simultaneous validity statement for the audit protocol, not a statement that the conditional probability of correctness given acceptance equals the same confidence level.

**Tighter deterministic evaluation range.** Pairing may reduce the range. Let $D_0 = \lambda/(2\tau)$ and

$$B_\Delta = \min\{|\widehat{\eta}_S - \widehat{\eta}_X| B_X + \widehat{\eta}_S D_0, |\widehat{\eta}_S - \widehat{\eta}_X| B_S + \widehat{\eta}_X D_0\}.$$

For displacements $d_S = \widehat{\eta}_S S$, $d_X = \widehat{\eta}_X X$, the difference satisfies

$$|Z_e| \leq \|d_S - d_X\| \left[\|g\| + \tfrac{1}{2} h_{\max}(\|d_S\| + \|d_X\|)\right].$$

Hence one may replace $C_E$ by the smaller of its value in Theorem E.1 and $B_\Delta[\sqrt{s_U} + h_{\max}(\widehat{\eta}_S B_S + \widehat{\eta}_X B_X)/2]$. Both are deterministic valid ranges fixed before evaluation, so taking their minimum incurs no extra probability allocation. An empirical-Bernstein evaluation interval may instead be used with the proved constants, but the explicit power formula (54) refers to the Hoeffding choice. The realized Bernstein interval can be checked without asserting that its radius was known in advance.

### E.9 When the local protocol is nonempty

**Corollary E.4** (Finite audit under strict population margins)**.** *Fix a finite candidate list and a finite set of positive and negative curvature comparisons in the quadratic sphere model. Suppose each selected candidate has $B_{\mathrm{St}} < 0$, $0 < \lambda < 2A$, positive transformed alignment and curvature energy, and its selected population tuned differences have nonzero margins of the stated signs. Suppose $s > 0$ and the anchor projection is measured exactly. Then sufficiently large finite calibration and profiling*

*sample sizes make the risk and signed-region tests pass on their joint confidence event. Selecting interval-midpoint steps (or any independent fitted steps passing the regret guards) and sufficiently many independent evaluation pairs yields the corresponding joint risk and achieved-sign conclusions at the stated total confidence level.*

*Proof.* On the scale event, the multiplicative interval width tends to zero as $n_C$ grows, uniformly over its possible endpoints relative to the fixed positive $s$. Thus $s_L, s_U \to s$, and the calibrated direct expression converges to the strictly negative $B_{\mathrm{St}}$; the strict parameter condition also eventually holds. Candidate norms and output ranges are bounded uniformly for all sufficiently large such $n_C$. Hoeffding profiling widths then go to zero uniformly as $N$ grows. Each signed functional, positive denominator and alignment has a strict margin for the finite list, so the corresponding interval tests eventually pass. The step intervals shrink to positive singleton optima, and midpoint-step regret bounds tend to zero. Hence the half-margin guards eventually pass as well. All retained achieved margins are then bounded away from zero. Their finite pre-evaluation range bounds and (54) give a finite sufficient evaluation size. The total confidence event is that of Theorem E.1. □

This corollary assumes strict direct-bound and tuned-descent margins at the same parameters. It does not prove those margins for a particular empirically favorable anchor, or that the sufficient audit sizes are economical. It removes an asymptotic-probability ambiguity: when the premises hold, a fixed-state finite joint statement is possible without changing the candidate toward identity. The new finite-sample bands have not been retrospectively applied to the reported 28 sign checks.

### E.10 Information and query costs

Count $n_C$ calibration directions separately from the $N$ independent *batched* profiling updates. One profiling sample costs $2m$ function queries to construct; $Nm$ directions are not $Nm$ independent samples of the nonlinear update statistic. A nonzero fixed anchor projection adds two queries in a quadratic, and multiple anchors may share the scale pool. The coordinate-free slope statistic adds two queries per transformed update. Step fitting adds its construction and objective/profiling queries. Evaluation adds construction of each fresh $X_e$ plus the two frozen-endpoint objective values per comparison. Sharing updates across anchors or counterfactual curvatures saves construction cost but creates dependence, which the union-bound protocol explicitly allows.

In a known quadratic simulator, moment coordinates and curvatures are available algebraically. That is a mechanism experiment, not an oracle-cost-free training procedure. In a function-value-only setting, obtaining the three correction moments can use six values along $\pm X, \pm V, \pm(X+V)$ plus a shared base value; direct comparison of two frozen endpoints needs two values. Reuse across many predictions may motivate the moment route, but no lower net acquisition cost has been established. The earlier pointwise curvature checks are distinct from the later complete audit in Appendix H.2. The latter executes all stages but obtains no final endpoint acceptance under its frozen bounds.

## F A direct risk bound for Gaussian directions

The input directions in this appendix are Gaussian, but the update is not the Gaussian proxy used in a comparison argument. Normalize to $g = e_1$ and take $r > 3$; a general nonzero scale can be recovered by scaling all update and anchor coordinates and the squared-norm parameters consistently.

**Proposition F.1** (Direct bound from the law with Gaussian directions)**.** *Let $X = m^{-1}\sum_i z_{i1} z_i$ for iid $z_i \sim N(0, I_r)$ and $m \geq 1$. Fix $a$, $\tau > 0$, $0 < \lambda \leq \tau^2$. Write $a_\perp = a - a_1 e_1$ and*

$$d_a = \sqrt{4a_1^2 + \|a_\perp\|^2}, \qquad \ell_a = 1 - 4/m - d_a/(m\tau), \tag{74}$$

$$c_G = \lambda\left[\lambda + \frac{2(\|a\| - a_1)}{m} - \frac{2(r-3)}{m}\ell_a\right], \quad Q_a = \|g - a\|^2 + (r+1)/m + \tau^2. \tag{75}$$

*Then*

$$\Delta_R \leq \begin{cases} c_G/Q_a, & c_G < 0, \\ c_G/\tau^2, & c_G \geq 0. \end{cases} \tag{76}$$

*In particular, $a = 0$ and $\lambda = \tau^2 = (r-3)/m$ give a strictly negative bound whenever $m > 8$.*

*Proof.* Let $Q_\chi = \sum_i z_{i1}^2$. Conditioning on all first coordinates, the transverse coordinates are centered Gaussian with covariance $Q_\chi I/m^2$. Hence the exact representation is

$$x := X_1 = Q_\chi/m, \qquad X_\perp = \sqrt{x/m}\, z_\perp, \qquad Q_\chi \sim \chi^2_m, \quad z_\perp \sim N(0, I_{r-1}),$$

with independent $x, z_\perp$. The Gamma integration-by-parts identity is

$$\mathbb{E}[(x-1)w(x)] = \frac{2}{m}\mathbb{E}[xw'(x)].$$

When differentiating $f_1(x, \sqrt{x/m}z_\perp)$, the derivative includes $\sum_{j>1} X_j \partial_j f_1/(2x)$. Conditional Gaussian integration by parts for the transverse components then yields a latent, generally nonsymmetric kernel

$$K_{11} = 2x/m, \quad K_{1j} = X_j/m, \quad K_{jj} = x/m \ (j > 1),$$

with other entries zero, satisfying $\mathbb{E}[(X-g)^\top f(X)] = \mathbb{E}\sum_{ij} K_{ij}\partial_j f_i(X)$. It is not a covariance matrix and is not asserted positive semidefinite. For $y = X - a$, $R = \|y\|^2$, $h = R + \tau^2$, the same risk expansion gives

$$\Delta_R = \mathbb{E}[\lambda^2 R/h^2 - 2\lambda \operatorname{tr} K/h + 4\lambda y^\top K y/h^2].$$

Here

$$\operatorname{tr} K = (r+1)x/m, \qquad y^\top K y = \{2xR - a_1\|y_\perp\|^2 + y_1 a_\perp^\top y_\perp\}/m.$$

The anchor quadratic form has largest eigenvalue $c_a = (\|a\| - a_1)/2 \geq 0$; this follows from the plane matrix $\left(\begin{smallmatrix} 0 & \|a_\perp\|/2 \\ \|a_\perp\|/2 & -a_1 \end{smallmatrix}\right)$, and its remaining transverse eigenvalues are no larger. Therefore $y^\top K y \leq (2x + c_a)R/m$. Since $x, c_a \geq 0$ and $R/h \leq 1$,

$$\Delta_R \leq \lambda \mathbb{E}\frac{\lambda + 4c_a/m - 2(r-3)x/m}{h}.$$

The variables $x$ and $h$ are not independent. Conditional on $z_\perp$,

$$xh'(x) = 2R - \|y_\perp\|^2 + 2a_1 y_1 + a_\perp^\top y_\perp \leq 2R + d_a\sqrt{R} \leq (2 + d_a/(2\tau))h.$$

Apply the Gamma identity to $w = 1/h$ to obtain

$$\mathbb{E}(x/h) = \mathbb{E}(1/h) - \frac{2}{m}\mathbb{E}[xh'/h^2] \geq \ell_a \mathbb{E}(1/h).$$

Because $r > 3$, the coefficient multiplying $\mathbb{E}(x/h)$ is negative, so the substitution gives $\Delta_R \leq c_G \mathbb{E}(1/h)$. If $c_G < 0$, Jensen gives the upper bound $c_G/\mathbb{E}h = c_G/Q_a$. If $c_G \geq 0$, use $1/h \leq 1/\tau^2$. The two branches cannot be exchanged.

All identities hold for integer $m \geq 1$. The positive ridge controls denominators; Gamma boundary terms vanish at zero as $x^{m/2}$ times a bounded function, and at infinity by the exponential tail. The $x^{-1/2}$ derivative term becomes integrable after multiplication by $x$. For zero anchor, $d_a = c_a = 0$, and substituting $\lambda = \tau^2 = (r-3)/m$ gives $c_G = \lambda^2(8/m - 1)$. This is negative for $m > 8$, proving the final assertion. □

For the reported 36 candidates with Gaussian directions, this proposition alone gives nine negative bounds, all at $m = 16$. Combining it with the residual-moment bound gives 31. These are different counts. The proposition does not include nonlinear Gaussian finite-difference errors, stochastic target noise, or a trained-model guarantee.

## G Additional calibration and prediction results for the quadratic models

The tables below retain the complete values from the controlled studies discussed in Sections 4 and 5. They are separate from the new nonquadratic checkpoint measurements. The scale-calibration values are medians over 64 repetitions at nominal confidence parameter 0.05, not simultaneous confidence claims. The known-parameter direct-bound count of 33/36 is not a calibrated success count.

Table 5: Median calibrated risk upper bound, sphere $r = 64, m = 16$.

| Calibration directions | $a = 0$ | $a = e_1$ | $a = e_2$ |
|---|---|---|---|
| 128 | 0.7851 | 0.6412 | 0.8986 |
| 256 | 0.2092 | −0.0012 | 0.3712 |
| 512 | −0.1757 | −0.4355 | 0.0238 |
| 1024 | −0.4658 | −0.7702 | −0.2347 |
| 2048 | −0.6711 | −1.0062 | −0.4177 |

The predicted thresholds were frozen using the earlier pool. Each of fourteen positive thresholds was tested at 0.8 and 1.2 times its value with independent fitting and evaluation pools. The new-pool moment thresholds below are checks, not values used to choose the tested curvatures. Neither the pointwise three-SE criterion nor these thresholds executes the local joint confidence protocol.

Table 6: Transverse-curvature thresholds for $\rho = 1, \phi = 60°$, other curvatures 1. The new-pool threshold is a post hoc check, not used to choose evaluation curvatures.

| Law | $m$ | Frozen prediction | New-pool moment threshold |
|---|---|---|---|
| Gaussian | 4 | 38.32869 | 38.25056 |
| Gaussian | 16 | 8.31123 | 8.27102 |
| Sphere | 4 | 38.20499 | 38.37634 |
| Sphere | 16 | 8.20470 | 8.20270 |

### G.1 Known-parameter risk checks and the scale reference

The original controlled grid uses two iid direction laws, $r \in \{16, 64, 256\}$, $m \in \{1, 4, 8, 16\}$, and $a \in \{0, e_1, e_2\}$, with $g = e_1$ and $\lambda = \tau^2 = A(\Sigma)$. Each of 72 fixed candidates uses 131,072 independent full-direction updates as a risk reference. All empirical risk differences are negative beyond their pointwise three-standard-error thresholds. The residual inequality alone gives 31 negative bounds among 36 sphere candidates; combining it with the sphere Stein inequality gives 33. The unproved cases are $r = 16, a = e_2, m = 4, 8, 16$, with upper bounds $0.16245, 0.16788, 0.07060$. Gaussian residual and combined counts are 29 and 31 of 36. These are known-parameter inequality counts, not finite-calibration success rates.

Table 7: Sphere risk comparison at $r = 64, m = 16$, without retuning the candidate. Parentheses contain Monte Carlo standard errors; the other columns evaluate proved inequalities.

| Anchor | MC risk difference (SE) | Residual bound | Stein bound |
|---|---|---|---|
| 0 | −2.33544 (0.00198) | −0.97329 | −1.18532 |
| $e_1$ | −2.79668 (0.00180) | −2.62546 | −1.60654 |
| $e_2$ | −2.01422 (0.00205) | −0.60203 | −0.87019 |

At unit target scale and $r = 64, m = 16$, exact Raw risk is 3.9375, and the free scale baseline has risk $63/79 \approx 0.79747$. The old nonlinear difference in Table 7 implies risk approximately 1.60206 (MC SE 0.00198). The new study's independent estimates in Table 9 are separate observations, not corrections to those old samples.

### G.2 Finite-dimensional anchor controls

The refined region study uses $r = 64$ with sphere/Gaussian laws and $m = 4, 16$, and 425 anchors per grid. The anchor is a controlled quantity, not a free estimate of useful history. Each grid uses independent step-fitting and evaluation pools, shared across anchors and counterfactual curvatures. For Gaussian $m = 16$, $\rho = 1$, $\phi = 60°$, the risk difference is $-2.42668$ both at $H = I$ and with only transverse curvature changed to 16. The evaluated descent differences are $+0.044906 \pm 0.000069$ and $-0.018170 \pm 0.000083$, respectively (standard errors).

The stronger-control comparison retains 48 selected anchor/law/batch configurations and 220 anchor–curvature comparisons. The joint fixed-mixture control solves a two-coefficient nonnegative quadratic problem on the same fitting pool. It has an extra fitted degree of freedom relative to the fixed nonlinear candidate; some collinear anchors make its anchor-only boundary especially strong. The entries are dependent pointwise comparisons, not a generalization rate or a comparison against every nonlinear parameter choice.

Table 8: Pointwise comparisons of the tested fixed nonlinear candidate against controls. Counts use a three-paired-SE threshold, are dependent across configurations, and are not task-generalization rates.

| Control | Nonlinear better | Control better | Unseparated |
|---|---|---|---|
| Fixed fitted mean retention | 23 | 194 | 3 |
| Fixed half mixing | 41 | 179 | 0 |
| Jointly optimized fixed mixing and step | 0 | 220 | 0 |
| Radial norm-matched diagnostic | 158 | 45 | 17 |

The 41 observations favoring nonlinear over fixed-half mixing remain part of the evidence. They do not conflict with the restricted sufficiently-large-$r$ theorem, and they prevent replacing that theorem with a universal finite-dimensional ordering. A radial norm-matched diagnostic additionally shows that equal update norm does not isolate anchor direction or nonlinear value; computing that norm requires the nonlinear candidate itself.

## H Completed simulations, local audit, and network comparisons

The completed follow-up comprises three studies. A evaluates scale baselines and predictions at fixed direction count, B evaluates local joint inference on a specified quadratic, and C expands the network step search and tests predictions from independent profiling. Earlier controlled studies and checkpoint outcomes informed the design. The new pools are kept separate from those exploratory observations; replayed records and reconstructed checkpoints do not add samples.

The methods and tables below were checked against the frozen configuration, executable code, saved selections, and per-update records. The manuscript evidence extract contains the numerical tables and complete final grids, with source hashes and a script that reconstructs the principal summaries. The full experimental directory retains the raw records and execution scripts identified in Table 16.

### H.1 A: scale baselines and predictions at fixed direction count

The grid uses $r \in \{16, 64, 256, 1024, 4096\}$ and $m \in \{1, 4, 8, 16\}$ for iid sphere and iid Gaussian directions, with $g = e_1$, $H = I$, and zero anchor as the main comparison. Each of the 40 cells has 32,768 fitting and 131,072 independent evaluation updates, totaling 6,553,600 updates. At $r = 64, m = 4, 16$, additional anchors $e_1, e_2, (e_1 + \sqrt{3}e_2)/2$ reuse the same pools. We use $\lambda = \tau^2 = (r-1)(r-2)/(m(r+2))$ for sphere directions and $(r-3)/m$ for Gaussian directions throughout each cell.

The candidates are Raw, its free MSE-optimal fixed rescaling, and nonlinear shrinkage. Nonzero-anchor cells additionally include a fixed mean-retention mixture, a fixed half mixture, a mixture with jointly fitted coefficient and step, and the fixed MSE coefficient that uses known target and covariance information. The mean retention is estimated on the fitting pool. Joint fitting solves the nonnegative quadratic in $(\eta q, \eta(1-q))$, considering the interior solution and both axes. Each candidate's nonnegative step is fixed using the fitting pool, then evaluated on the independent pool.

Two distinct decrease statistics are stored. For a frozen fitted step $\widehat{\eta}$, the evaluation mean is $\widehat{\eta}\overline{Y_1} - \widehat{\eta}^2\overline{\|Y\|^2}/2$. This is the quantity in Tables 1 and 9. The archive also stores $(\overline{Y_1})_+^2/(2\overline{\|Y\|^2})$, a plug-in numerical reference for the population optimum obtained from evaluation moments. For sphere $r = 4096$ and $m = 1, 4, 8, 16$, its N/Raw ratios are $0.920373$, $0.937362$, $0.956934$, and $0.973667$, respectively. These are different estimators, rather than different roundings of the independently fitted ratios. Risk after the fitted rescaling equals one minus twice its achieved decrease, with maximum

Table 9: Independent zero-anchor sphere evaluation at $r = 64$. Risks refer to the displayed update before step fitting. The last column uses steps selected on the independent fitting pool.

| $m$ | Raw risk | Nonlinear risk | Free fixed-scale risk | N/Raw achieved decrease |
|---|---|---|---|---|
| 4 | 15.6997 | 5.9065 | 0.9405 | 0.9206 |
| 16 | 3.9371 | 1.6037 | 0.7972 | 0.9429 |

stored floating-point discrepancy $3.33 \times 10^{-16}$ across the 204 candidate rows. Raw and its free fixed rescaling agree in achieved decrease to $8.33 \times 10^{-17}$ across the 40 cells.

**Exact reduced simulation.** We retain $(X_1, X_2, \|X_{3:r}\|^2)$, sufficient for the specified anchors and curvature blocks. Conditional on the first two coordinates of a sphere direction, its remaining orientation is independent and uniform in dimension $r - 2$. If the accumulated residual sum has squared norm $L$, adding a residual vector of length $b = (r/m)|u_1|\sqrt{1 - u_1^2 - u_2^2}$ gives

$$L' = L + b^2 + 2\sqrt{L}\, bT, \tag{77}$$

where $T$ is an independent first coordinate of a uniform sphere vector in dimension $r - 2$. Rotational invariance and induction preserve the joint law of the retained coordinates; the coefficient sign is absorbed by symmetry. For Gaussian directions, $X_1 = W/m$ with $W \sim \chi^2_m$; conditional on $X_1$, $X_2$ is normal with variance $X_1/m$ and the remaining squared norm is $(X_1/m)\chi^2_{r-2}$, independently of $X_2$. A separate full-direction implementation normalizes each sphere draw over all $r$ coordinates. Forty prespecified full/reduced smoke comparisons, covering five $(r, m)$ pairs, two laws and four summaries, all met the standardized-difference threshold 6; the maximum was 2.502. Exactness follows from the conditional-law argument, while this check tests its implementation.

**Stein-bound ablation.** The 26 saved rows compare the same candidate under three bounds. With $t = \operatorname{tr}\Sigma$, $Q = \|g - a\|^2 + t + \tau^2$, and the centered kernel fluctuations from Theorem 4.1, the implemented expressions are

$$U_0 = \frac{\lambda(\lambda - 2A)}{Q} + \frac{\lambda}{\tau^2}\left(2\sqrt{\sigma_{\mathrm{tr}}^2 + t^2} + \sqrt{\sigma_F^2 + \|\Sigma\|_F^2}\right), \tag{78}$$

$$U_1 = \frac{\lambda(\lambda - 2A)}{Q} + \frac{\lambda}{\tau^2}(\sigma_{\mathrm{tr}} + \sigma_F), \tag{79}$$

$$U_2 = U_1 - \frac{\lambda\tau^2(\lambda + 4\kappa)}{Q^2}. \tag{80}$$

Under the theorem's parameter conditions, $U_0$ bounds the kernel terms through uncentered second moments, $U_1$ uses exact centered fluctuations and the centered reciprocal-denominator estimate, and $U_2 = B_{\mathrm{St}}$ also retains the negative ridge term. The calculation in Appendix A gives these successive bounds without changing $\lambda$ or $\tau$. These are our adaptations of the Stein argument described in Appendix A.3. Table 10 gives the zero-anchor $r = 64$ cases; all 26 rows are included in the evidence extract.

Table 10: Successive Stein bounds for one fixed sphere candidate per row, $r = 64, a = 0$. Negative values certify risk reduction relative to Raw.

| $m$ | $\lambda = \tau^2$ | Coarse $U_0$ | Exact fluctuations $U_1$ | Retained ridge $U_2$ |
|---|---|---|---|---|
| 4 | 14.795455 | 30.764048 | 2.379810 | −1.294854 |
| 16 | 3.698864 | 7.104687 | −0.419287 | −1.185324 |

### H.2 B: FINITE-SAMPLE LOCAL AUDIT

The fixed candidate uses sphere directions with $r = 64$, $m = 16$, $g = e_1$, $a = (e_1 + \sqrt{3}e_2)/2$, and $\lambda = \tau^2 = 6362/1666$. The objectives have $H = \mathrm{diag}(1, h_2, 1, \ldots, 1)$ at $h_2 = 1$ and 16. This candidate and both curvatures were chosen using the earlier exploration. Table 11 lists the three

planned calibration/profiling budgets; fitting and evaluation use $n_F = 32768$ and $n_E = 262144$ at every budget.

Risk calibration receives the sampled direction slopes and a symmetric query pair along $a$, without receiving the true scale. It constructs the interval in (26), measures $b = g^\top a = 1/2$, and evaluates (27) after checking $\lambda \le 2A_L$ and $Q_U > 0$. The optimization stages use the known coordinates and Raw moments of this quadratic mechanism, including $\mathbb{E}X_1 = 1$ and $C(h) = 1 + (r-1)/m + (h_2 - 1)r/[m(r+2)]$. Thus the hidden-scale calibration test is separate from the information supplied to optimization profiling.

Profiling uses the four paired corrections in (63). For this $\lambda = \tau^2$ candidate, the implemented ranges are sharper than the general norm envelope:

$$Z_0 \in [-\sqrt{\lambda}/2, \sqrt{\lambda}/2], \tag{81}$$

$$Z_j \in [-2\lambda - |a_j|\sqrt{\lambda},\ |a_j|\sqrt{\lambda}], \quad j = 1, 2, \tag{82}$$

$$Z_o \in [-2\lambda, 0]. \tag{83}$$

To see this, write $y = X - a$. The energy correction in coordinate $j$ is $(q^2-1)y_j^2 - 2(1-q)a_j y_j$, whose first term lies in $[-2\lambda, 0]$ and whose second term has magnitude at most $|a_j|\sqrt{\lambda}$. In the orthogonal complement the anchor is zero. The empirical-Bernstein radius is (62), with unbiased variance and intervals intersected with these known supports. From the correction means we obtain $\alpha = 1 + \mathbb{E}Z_0$ and $u(h) = C(h) + \mathbb{E}Z_1 + h_2\mathbb{E}Z_2 + \mathbb{E}Z_o$. The same four-moment event supports the positive-curvature family.

The total failure allocation is $\delta_C + \delta_M + \delta_E = 0.01 + 0.02 + 0.02 = 0.05$. Each of three calibration looks receives $\delta_C/3$, each of four profiling coordinates at each look receives $\delta_M/12$, and each endpoint interval, including the direct comparator, receives $\delta_E/12$ across three budgets and two curvatures. The calibration and profiling budgets are nested prefixes of their respective pools. One fitting pool and one evaluation pool are shared across budgets and curvatures. The four pools use independent seeds $[20260926, 201]$ through $[20260926, 204]$; direct comparison at budget index $j = 0, 1, 2$ uses $[20260926, 205, j]$. These are six dependent budget–curvature rows, with only two distinct main endpoint intervals.

Raw uses its exact step $1/C(h)$. The nonlinear step is fitted on $\mathcal{F}$, giving $0.3707253279$ at $h_2 = 1$ and $0.1636985140$ at $h_2 = 16$; the corresponding Raw steps are $0.2025316456$ and $0.1710398445$. These steps remain fixed across budgets. The implemented regret bound for nonlinear shrinkage maximizes $u\widehat{\eta}_S^2/2 - \widehat{\eta}_S\alpha + \alpha^2/(2u)$ over the four vertices of the positive-energy $(\alpha, u)$ interval rectangle. Convexity makes this a valid upper bound. Raw has zero fitting regret. With the known unit gradient, the implementation obtains its margins by directly bounding the two optimized values:

$$\widetilde{\gamma}_+ = \frac{1}{2}\left(\frac{(\alpha_L)_+^2}{u_U(h)} - \frac{1}{C(h)}\right), \qquad \widetilde{\gamma}_- = \frac{1}{2}\left(\frac{1}{C(h)} - \frac{(\alpha_U)_+^2}{u_L(h)}\right), \tag{84}$$

with positive energy bounds. These are the executed known-scale bounds, distinguished from the scale-interval propagation in (59). Positive predictions require $\widetilde{\gamma}_+ > 0$ and $\mathfrak{r}_S \le \widetilde{\gamma}_+/2$; negative predictions require $\widetilde{\gamma}_- > 0$ and $\mathfrak{r}_X = 0 \le \widetilde{\gamma}_-/2$.

Table 11: Executed local-audit budgets and outcomes. Both signed columns include the appropriate fitting-regret guard. Fitting and evaluation remain fixed at 32,768 and 262,144 updates, respectively; rows share nested C/O prefixes and the same F/E pools.

| C directions | O updates | Risk upper bound | $h_2 = 1$ | $h_2 = 16$ | Final E accepts |
|---|---|---|---|---|---|
| 4,096 | 4,096 | −0.7266 | Pass | Abstain | 0 / 2 |
| 16,384 | 65,536 | −0.9658 | Pass | Pass | 0 / 2 |
| 65,536 | 1,048,576 | −1.0893 | Pass | Pass | 0 / 2 |

All candidate and step decisions are saved before generating the evaluation pool. For the frozen steps, endpoint differences are bounded by

$$|Z_e| \le M_E := r(\widehat{\eta}_S + \widehat{\eta}_X) + \tfrac{1}{2}h_2 r^2(\widehat{\eta}_S^2 + \widehat{\eta}_X^2), \tag{85}$$

using $\|X\|, \|S\| \le r$. Evaluation applies the allocated empirical-Bernstein interval on $[-M_E, M_E]$. The intervals in Table 3 include zero. The actual interval exclusion check, rather than the Hoeffding sufficient sample-size condition, is the final acceptance rule. The vectorized implementation generates O and F before acquiring C, computes the risk and moment bounds, freezes decisions, then opens E. Its recorded cost includes all stages, including rows that abstain.

The direct comparator evaluates the same frozen endpoints on its own updates. Per curvature, the accounting convention assigns $2n_C + 2 + 38N_O + 32n_F + 34n_E$ oracle-equivalent queries to the audit, and gives the comparator $\lfloor \text{budget}/34 \rfloor$ updates. The three charges are 10,125,314, 12,484,610, and 49,938,434, corresponding to 297,803, 367,194, and 1,468,777 direct updates. These are an accounting model for arithmetic simulation; the CPU studies make no neural-model calls. With two-curvature reuse, the audit's charges are 10,649,602, 13,008,898, and 50,462,722, compared with 10,720,908, 13,218,984, and 52,875,972 for the direct runs. The per-curvature budget match therefore does not imply an equal-budget comparison for the whole family.

Table 12: Endpoint intervals at the largest per-curvature budget. The direct comparator receives more endpoint samples, while the audit allocates most of this budget to profiling.

| Curvature | Audit endpoint interval | Direct-comparator interval |
|---|---|---|
| $h_2 = 1$ | $[-0.00891,\ 0.10289]$ | $[0.03698,\ 0.05702]$ |
| $h_2 = 16$ | $[-0.27780,\ 0.23758]$ | $[-0.06608,\ 0.02602]$ |

The direct interval also excludes zero for $h_2 = 1$ at the middle budget, but none of its $h_2 = 16$ intervals does. Because F and E remain fixed, the main endpoint intervals repeat across the three budgets even as profiling tightens. This identifies the consequence of the chosen allocation and support envelope in this run.

### H.3 C: NETWORK STEP SEARCH AND INDEPENDENT EVALUATION

All eight original source settings are retained, with the same four checkpoints, model/tokenizer/data snapshots, source and target batches, historical anchors and ridge values, and original fitted mean-retention values. No trajectory is retrained. The inherited Engine, two-label loss, estimator radius and EMA definitions are specified in Appendix I.2. C1 reuses the 16 old fitting updates (four directions each) to study the step range. C2 uses 64 new fitting updates, C3 uses 64 independent profiling updates, and evaluation uses 128 new updates per source setting. Stream identifiers are 720000, 730000 and 740000, respectively, plus checkpoint step, with base seed $20260926 + \text{trajectory index}$ and index $10000\,\text{pair} + \text{sample}$. All new seeds were checked against one another and the original training/fitting/evaluation manifest.

The common initial grid is

$$\mathcal{G}_0 = 10^{-6}\{0, 2.5, 5, 10, 20, 40, 80, 160, 320, 640, 1280\}. \tag{86}$$

C1 permits at most two additional doublings if a candidate keeps improving at the boundary or selects the largest step, up to $0.00512$. It then performs one refinement: for each positive selected step, insert geometric midpoints with its positive adjacent grid points. C2 fits on that refined grid and permits one further doubling under the same condition and cap, before evaluation. No extra doubling was used in this run; all final grids end at $0.00128$. The exact refinement sets are given below and are also saved in the frozen records.

Every candidate minimizes mean source endpoint loss. Joint fixed mixing searches $q \in \{0, 0.25, 0.5, 0.75, 1\}$ with the step. Exact ties choose the first minimum in ascending step order for ordinary candidates, and increasing $q$ followed by increasing step for joint selection. Fixed mean mixing retains each setting's original $\bar{q}$. Source-selected parameters are applied unchanged on target, which never enters selection. Table 14 lists all forty selected steps.

A selected grid minimum is classified as interior when at least two finite grid points lie to its right. For joint fixed mixing, this check uses the minimum over $q$ at each step. Otherwise, a plateau requires the last three finite mean losses to span at most $\max\{10^{-5}, 10^{-4}|F(w)|\}$. Remaining cases are labeled as improving, censored by a safety/numerical boundary, or unstable. An endpoint is ineligible if its RMS parameter displacement exceeds $0.1$, any individual loss exceeds 50, or an OOM or nonfinite

result occurs. It is retained with its status, and fitting assigns it infinite loss. Uncaught case exceptions terminate that case. The 600-second elapsed-time check occurs after fitting and after profiling, before the next stage. All recorded fitting endpoints in this run were finite; their largest RMS displacement and individual loss were $0.084085$ and $7.73947$.

Table 13: Exploratory and new sample pools. The forty selections are five candidates in eight source settings. New fitting, profiling and evaluation pools are mutually independent; settings share checkpoint and batch structure.

| Quantity | Earlier exploration | New study |
|---|---:|---:|
| Fitting updates per setting | 16 | 64 |
| Separate profiling updates per setting | — | 64 |
| Evaluation updates per setting | 32 | 128 |
| Upper-boundary step selections | 31 / 40 | 0 / 40 |
| Interior minima with two finite points to the right | Not established for all | 40 / 40 |
| Interval convention | Pointwise 3SE | Adjusted approximate Student-$t$ |

Table 14: All selected source-loss steps, in units of $10^{-5}$. Target uses the same parameters. Every selection is interior under the two-finite-points criterion. The last column is the jointly selected fixed mixing coefficient.

| Seed / step / pair | Raw | Nonlinear | Fixed mean | Joint fixed | Radial | Joint $q$ |
|---|---:|---:|---:|---:|---:|---:|
| 9021 / 16 / 0 | 11.31371 | 22.62742 | 22.62742 | 32 | 22.62742 | 0.25 |
| 9021 / 16 / 1 | 11.31371 | 16 | 22.62742 | 22.62742 | 22.62742 | 0.25 |
| 9021 / 64 / 0 | 4 | 5.65685 | 8 | 16 | 5.65685 | 0.25 |
| 9021 / 64 / 1 | 5.65685 | 5.65685 | 8 | 16 | 5.65685 | 0.25 |
| 9022 / 16 / 0 | 5.65685 | 5.65685 | 11.31371 | 5.65685 | 8 | 1 |
| 9022 / 16 / 1 | 5.65685 | 5.65685 | 11.31371 | 5.65685 | 8 | 1 |
| 9022 / 64 / 0 | 4 | 5.65685 | 8 | 22.62742 | 5.65685 | 0.25 |
| 9022 / 64 / 1 | 4 | 5.65685 | 8 | 5.65685 | 5.65685 | 0.75 |

In units of $10^{-5}$, C1 adds $\{4\sqrt{2}, 8\sqrt{2}, 16\sqrt{2}, 32\sqrt{2}\}$ for both seed-9021 step-16 settings, $\{2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}, 16\sqrt{2}\}$ for seed-9021 step-64 pair 0, both seed-9022 step-16 settings, and seed-9022 step-64 pair 0. It adds $\{4\sqrt{2}, 8\sqrt{2}\}$ for seed-9021 step-64 pair 1 and $\{2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}\}$ for seed-9022 step-64 pair 1. C1 and C2 use identical final grids in every setting.

The main comparison has 64 paired contrasts: eight source settings, two evaluation batches, and four controls against nonlinear shrinkage. For each contrast, with $n = 128$, the interval is

$$\bar{z} \pm t_{n-1,\,1-0.05/(2\cdot 64)} \frac{s_z}{\sqrt{n}}, \tag{87}$$

where $s_z$ is the sample standard deviation of paired differences. Its sign is resolved when the interval excludes zero. These are Bonferroni-adjusted approximate Student-$t$ intervals. The network source/target settings are dependent, and the interpretation remains conditional on their recorded states and batches.

For the scale implementation check, Raw at $\eta = 10^{-5}$ is compared with $cX$ at $\eta/c$, using the first old fitting update in each setting. At $c = 0.5$, parameter and loss differences are zero in FP32. At $c = q_{\mathrm{MSE}} = 4/(D+3) = 3.19388565 \times 10^{-8}$, the maximum direct parameter discrepancy is $1.6194 \times 10^{-6}$ in Euclidean norm and $5.9605 \times 10^{-8}$ in any coordinate; the maximum per-example loss discrepancy is $9.5367 \times 10^{-7}$. The algebraically equivalent displacement has zero parameter discrepancy. Risk optimality of this coefficient belongs to the linearized sphere model.

Relative to the old selected-step table, N–Raw changes point-estimate sign in 4/16 contexts and N–fixed-mean in 8/16. Search range and fitting/evaluation randomness both change, so this comparison does not isolate the causal effect of extending the grid. The old target batches are retained. The old risk reduction range of 43.0–74.6% remains an exploratory statistic in Appendix I.

### H.4 C: INDEPENDENT PROFILING AND MEASURED APPROXIMATION ERRORS

Profiling estimates the two coefficients of (17) using gradients and HVPs of the same two-label objective. It freezes the predictions before evaluation. Nonlinear and fixed mean mixing use the same step in this intervention, at fractions $1/8, 1/4, 1/2, 1$ of the selected nonlinear step. This differs from the main table, where each method selects its own step.

There are 64 profiling predictions and 64 corresponding endpoint contrasts. Each of these two families separately uses (87), with $n = 64$ for profiling and $n = 128$ for endpoints. Their critical values are approximately 3.53061 and 3.44240. A mechanism interval overlapping $[-10^{-5}, 10^{-5}]$ is unresolved. The main contrasts, profiling predictions, and mechanism endpoints each receive a separate nominal failure allocation 0.05; the three families do not form one joint 95% statement. Table 15 counts agreement under these approximate intervals.

Table 15: Independent coefficient predictions over sixteen dependent source/target contexts. The final column requires both separately adjusted approximate intervals to exclude $[-10^{-5}, 10^{-5}]$ with the same sign.

| Step fraction | Point-estimate sign matches | Both intervals resolve the same sign |
|---|---|---|
| $1/8$ | 12 / 16 | 6 / 16 |
| $1/4$ | 13 / 16 | 6 / 16 |
| $1/2$ | 15 / 16 | 5 / 16 |
| 1 | 12 / 16 | 2 / 16 |

The four full-step point-estimate reversals occur at source settings 9021/step16/pair0, 9021/step64/pair1, 9022/step16/pair0, and 9022/step64/pair0. Their profiling intervals were unresolved. The first selected step is approximately $2.26274 \times 10^{-4}$; the other three are $5.65685 \times 10^{-5}$.

For matched directions, the finite-difference and exact linearized updates retain the same anchor and ridge. The relative error

$$\left( \frac{\sum_i \|X_{\mathrm{FD},i} - X_{\mathrm{lin},i}\|^2}{\sum_i \|X_{\mathrm{lin},i}\|^2} \right)^{1/2} \tag{88}$$

ranges from 17.39% to 39.59% across the eight source settings. Separately, for each of sixteen source/target contexts, we average the difference between actual and quadratic-predicted nonlinear decrease over its 64 profiling updates, then take its absolute value. The median across contexts is $0.000280588$, $0.00107154$, $0.00295642$, and $0.0161336$ at the four step fractions. This is the absolute value of the mean residual. Comparing an independent profiling prediction with a new evaluation mean additionally introduces moment-estimation error. Neither calculation estimates a uniform Hessian-Lipschitz envelope.

### H.5 EXECUTION RECORDS, COSTS, AND REPRODUCIBILITY

The original package passed its 285-file and seven-frozen-file hash checks twice. Eleven old tables replayed with maximum absolute discrepancy $4.44 \times 10^{-16}$. The new A/B replay matched all four primary saved tables exactly. Original code and vendor modules were retained. B had a serialization failure before evaluation, repaired and replayed with the same seeds; the failed attempt's complete timing was not captured.

Network execution reused the hash-checked checkpoints in FP32 eval mode with eager attention and disabled TF32/cache. Two SciPy dependency failures occurred after profiling records were written and before evaluation. Continuation loaded frozen selections and saved moments, reconstructed and froze predictions, then collected the original unopened evaluation pools without refitting. The isolated SciPy 1.15.3 installation left the core torch/transformers/numpy environment unchanged. Failure and continuation records are retained separately.

A and B recorded successful kernel-stage CPU wall times of 9.60 and 7.25 seconds, including sampling and raw-record output, with zero neural-model calls. The network ledger records 1,883 allocated GPU seconds, including continuation and direction replay, 115,002 ordinary forward queries, 136 gradient forward/backward evaluations, and 2,432 HVPs. Peak allocated memory

is 34,977,952,256 bytes. Queue waits overlap and are reported separately from execution time. Synthetic oracle-equivalent accounting and measured network costs are distinct quantities.

Table 16: Execution artifacts used to verify the manuscript. The source archive contains a curated extract under `evidence/verified_followup`; full raw updates and runners remain in the experiment directory.

| Purpose | Files and records |
|---|---|
| Configuration | `effective_settings.json`, frozen per-case grids/selections and pool manifests, source SHA256 manifest |
| Scale and limits | `scale_baselines.csv`, `fixed_batch_limits.csv`, `A_verified_ratios.csv`, `stein_ablation.csv` |
| Joint audit | `joint_audit.csv`, `joint_budget.json`, `joint_family_costs.csv`, `B_frozen_decisions.json` |
| Network results | `network_selection_8x5.csv`, `network_contrasts.csv`, `network_mechanism_predictions.csv`, `network_error_split.csv` |
| Raw execution | Per-case `C1_all_endpoints.csv`, `C2_fit_all_endpoints.csv`, `profile_moments.csv`, `C2_evaluation.csv`, `C3_endpoints.csv`; CPU sufficient-coordinate `raw/*.csv.gz` |
| Costs and replay | `query_gradient_HVP_ledger.csv`, `total_costs.json`; `replay_tables.py` in the extract, `replay_cpu.py` and `analyze_next.py` in the full experiment directory |

The evidence manifest distinguishes unchanged copies of execution outputs from editorial extracts and recomputed ratios. The extract permits numerical table reconstruction from saved summaries. Replaying raw updates requires the full experiment directory, and network replay additionally requires the pinned model/checkpoint assets. Operational account paths are excluded from the manuscript extract.

# I Exploratory checkpoints and inherited execution details

This appendix preserves the earlier exploratory fixed-checkpoint study. The independent follow-up with expanded step searches is reported in Section 6 and Appendix H.3. Both use an actual non-quadratic network loss rather than the exact quadratic update law; their sample pools and inferential conventions must not be merged. Its purpose is to examine separation between risk and descent and the reach of the local diagnostics, not to establish a new training optimizer. All reported loss differences use cross-entropy (CE); positive descent contrasts favor the nonlinear candidate. The gradients and Hessian–vector products (HVPs) are paid diagnostic information, not free inputs to a proposed ZO optimizer.

## I.1 Recorded protocol, conditioning, and estimands

Two short full-parameter iid sphere ZO trajectories on OPT-125M/SST-2 use seeds 9021 and 9022. Checkpoints at steps 16 and 64 give four frozen states. Two additional source minibatches per state are paired with nonoverlapping target minibatches, yielding eight source settings and sixteen source/target contexts. Each source setting uses 16 fitting updates and 32 independent evaluation updates at $m = 4$. The update is constructed from the source minibatch and then paired across source and target evaluation. The historical anchor is a bias-corrected EMA of past raw updates; no current diagnostic gradient is used to construct it.

Audit samples are reused across checkpoints and trajectory seeds. Independence of fitting and evaluation directions does not make the sixteen contexts independent tasks. Reported directional standard errors condition on checkpoint, minibatch, and the parameters already selected. A positive/negative count means the paired mean is above/below three of its directional standard errors; an unresolved contrast falls between those thresholds. These are descriptive, pointwise screens, not multiplicity-adjusted significance or confidence-controlled regions.

The five evaluated candidates are raw, nonlinear anchored shrinkage, a fixed blend at the independently fitted mean retention, a fixed blend with mixing coefficient and step selected jointly, and a radial

norm-matched diagnostic. Each candidate selects its step using actual source loss on the fitting pool. Of forty selections, thirty-one reach the frozen grid's upper endpoint $8 \times 10^{-5}$. The grid was not expanded after seeing results. Zero step is a hyperparameter candidate, not risk-triggered fallback. Consequently, the study evaluates the selected steps in this finite range rather than the theoretical supremum in (5). Additional coefficient-search queries were charged, but the comparisons are not cost-matched optimizer tests.

For an evaluation minibatch $B$, let $F_B$ be its loss at checkpoint $w$, and let $g_B$ be its autodifferentiation gradient. The reported risk and decrease are

$$R_B(Y) = \mathbb{E}_{\text{dir}}\|Y - g_B\|^2, \qquad D_B(Y;\eta_Y) = \mathbb{E}_{\text{dir}}[F_B(w) - F_B(w - \eta_Y Y)], \tag{89}$$

$$\Delta_{R,B} = R_B(S) - R_B(X), \qquad \Delta_{D,B} = D_B(S;\eta_S) - D_B(Y;\eta_Y). \tag{90}$$

Here $Y$ denotes the control in the last expression, and all expectations fix the source-generated update law and preceding choices. Risks refer to unscaled estimators, not displacements $\eta_Y Y$. In a target context, $g_B$ changes but the source update does not: this is a transfer diagnostic, not an assertion that $X$ is unbiased for $g_B$. The source finite-radius mean need not equal its autodifferentiation gradient either.

### I.2 Recovered execution specification and evidence provenance

The companion material identifies the actual runner as `run_checkpoint.py`, using its frozen `vendor/run_lm.py` engine. The following values are taken from the executed configurations and implementation.

Table 17: Exact settings of the original frozen-checkpoint study. The completed material recovery does not expand the original samples or step range.

| Item | Actual setting |
|---|---|
| Model | OPT-125M, revision `27dcfa74d334bc871f3234de431e71c6eeba5dd6`; $D = 125239296$ parameters in `list(model.parameters())` order. |
| Precision | FP32, eval mode, eager attention, cache disabled, TF32 disabled; ReLU from the pinned model configuration. |
| Environment | Python 3.10.20, PyTorch 2.11.0+cu130, Transformers 5.6.2, NumPy 2.2.6; NVIDIA RTX 6000 Ada Generation. |
| Data | SST-2 train snapshot `bcdcba79d07bc864c1c254ccfcedcce55bcc9a8c`; 1200 permuted rows with seed 20260971. Exact row schedules are supplied. |
| Input/target | `Text: <sentence>\nSentiment:`; maximum length 96, retaining the first token and final 95 when truncated; left padding and mask. The two single-token labels are `negative` (2430) and `positive` (1313). |
| Loss | Cross-entropy normalized over those two logits, mean over eight examples, not full-vocabulary language model loss. |
| Directions | Four independent full-vector Gaussian draws normalized individually to the unit sphere, without orthogonalization. |
| Difference radius | $\delta = 0.001\sqrt{D}$ on each unit direction; the full-vector radius is not $0.001$. |
| Trajectory | Raw update rate $10^{-5}$; two seeds 9021/9022; centers and history saved after updates 16 and 64. |
| History | EMA coefficients $0.9/0.1$, bias-corrected by $1 - 0.9^t$; anchor is the vector EMA and $\lambda = \tau^2$ is the squared-Raw-norm EMA. |
| Search | $\eta \in \{0, 2.5, 5, 10, 20, 40, 80\} \times 10^{-6}$; joint blend $q \in \{0, 0.25, 0.5, 0.75, 1\}$. Actual source-loss selection on 16 updates. |
| Evaluation | 32 new source-generated updates per source setting; paired source/target loss and full/half-step endpoints. |

For unit $u_i$, the implemented estimator is

$$X = \frac{D}{m} \sum_{i=1}^{m} \frac{F_B(w + \delta u_i) - F_B(w - \delta u_i)}{2\delta} u_i.$$

The original history starts at zero and follows

$$h_t = 0.9h_{t-1} + 0.1X_t, \qquad v_t = 0.9v_{t-1} + 0.1\|X_t\|^2,$$
$$a_t = h_t/(1-0.9^t), \qquad \lambda_t = \tau_t^2 = v_t/(1-0.9^t).$$

Here $X_t$ is formed at $w_{t-1}$ and the saved center after the $t$-th update is $w_t = w_{t-1} - 10^{-5}X_t$. That past-only state is frozen for its diagnostic cases. Candidate residual squared norms are computed as $\max\{\|X\|^2 + \|a\|^2 - 2X^\top a, 0\}$ for floating-point protection. A fixed-mean blend uses the mean of the sixteen fitting-pool retention factors; the transverse intervention retains each original nonlinear factor rather than recomputing it after changing the anchor.

Each ordinary candidate evaluates seven steps; the joint blend evaluates 35 $(q, \eta)$ pairs. Exact ties select the first minimum: lower $\eta$ for ordinary candidates, and lower $q$ followed by lower $\eta$ for the joint grid. Zero steps share the base query, $q = 0$ endpoints share an anchor-only cache, and $q = 1$ shares Raw queries. These caches change cost, not the empirical selection objective. A candidate endpoint is always applied from the frozen center and the parameters are restored afterwards. Nonfinite values or failed consistency checks terminate the run rather than deleting a sample. The inherited configuration entries `tau=1`, `budget=2048`, and several unrelated controller fields do not govern this experiment; actual ridge and costs come from the equations and ledgers above.

The RNG map is `SeedSequence([seed,stream,index])`. Training batches use stream 20 and training directions stream 30, indexed by step. Fitting and evaluation use streams $100 + t$ and $200 + t$, with index $10000\,\text{pair} + \text{sample}$. Training uses the first 1000 selected rows. For pair $j = 0, 1$, the source uses selected positions $1000 + 16j$ through $1007 + 16j$, and target uses the next eight; original row identities and ordering are recorded in the companion schedule. Source/target batches are not new independent tasks across checkpoints.

**Original records versus recovery.** The original archive contains 504 fitting-grid means and 6656 update-level endpoint records. A later replay supplies 8064 per-fitting-update records and 6656 per-example endpoint rows (eight losses per row), explicitly marked as reconstructed. It reports exact tensor equality for the four parameter/anchor states and equality of ridge scalars, all original grid means and endpoints, and specified vector/derivative checks. This is not byte equality of newly serialized checkpoint containers. Original containers and tensor hashes are recorded separately. No new candidate or independent statistical sample is added by this recovery.

The original complete terminal transcript, full pre-run environment lock, and original per-fitting-update CE records were not saved. Equivalent commands, post-run environment snapshots and checked recovery records are provided as such, not represented as contemporaneous evidence. The lightweight archive includes tokenizer and Arrow data; pinned model weights and four large checkpoints are retrieved or reconstructed separately. Offline replay rebuilds eleven numerical tables from the records. The paper source bundle contains a non-identifying configuration extract and replay summary; the original execution companion remains separate because its operational paths require anonymization before public distribution.

### I.3 Selected-step comparisons and minibatch transfer

The nonlinear candidate reduces gradient-referenced update risk by 43.0–74.6% in all sixteen contexts, each beyond the negative pointwise 3SE threshold. Ten nonlinear–raw actual-descent differences have a negative point estimate, six beyond the negative threshold. Table 18 gives all three principal control comparisons. The target contexts reuse the source-generated update and the source-selected step.

Source loss decrease does not transfer reliably to the paired target minibatch in this study. The norm-matched diagnostic has exactly the nonlinear estimator norm on each update but retains the raw direction; each method separately selects its step. Nonlinear versus this diagnostic gives 3/2/3 positive/negative/unresolved source contrasts and 1/4/3 target contrasts. Thus norm reduction alone does not explain all observed signs. This diagnostic requires the nonlinear norm and is not a free practical baseline.

Table 18: Full-step contrasts in the checkpoint study. Positive differences favor nonlinear shrinkage. Each row contains eight conditional settings; counts are pointwise and dependent across rows.

| Control | Batch | Positive / negative / unresolved | Mean $\Delta_D$ (CE) |
|---|---|---|---|
| Raw | source | 0 / 4 / 4 | −0.03287 |
| Raw | target | 1 / 2 / 5 | −0.00647 |
| Fixed fitted mean retention | source | 1 / 5 / 2 | −0.01293 |
| Fixed fitted mean retention | target | 1 / 1 / 6 | −0.01195 |
| Jointly selected fixed blend | source | 0 / 5 / 3 | −0.03438 |
| Jointly selected fixed blend | target | 1 / 2 / 5 | −0.00547 |

Table 19: Mean actual decrease at source-selected steps. Negative target values mean that target loss increases on average; they are not downstream test-accuracy measurements.

| Update | Source decrease | Target decrease |
|---|---|---|
| Raw | 0.08107 | −0.01499 |
| Nonlinear anchored | 0.04820 | −0.02146 |
| Fixed fitted mean retention | 0.06113 | −0.00951 |
| Jointly selected fixed blend | 0.08258 | −0.01598 |

### I.4 Fixed-step identities and observed residuals

Distinguish independently selected-step comparisons from mechanism interventions at the same step. For a fixed evaluation batch, use its recorded gradient $g_B$ and symmetric local second-order form $H_B$ supplied by HVPs. Define the endpoint residual

$$\varepsilon_B(Y,\eta) = F_B(w-\eta Y) - F_B(w) + \eta g_B^\top Y - \tfrac{1}{2}\eta^2 Y^\top H_B Y. \tag{91}$$

This is an exact definition for the measured endpoints. It requires neither a positive definite $H_B$ nor a Hessian-Lipschitz assumption; without further assumptions it supplies no bound at unmeasured endpoints.

Freeze $\bar{q}$ using the fitting pool, and write $S_F = a + \bar{q}(X-a)$ and $U = (q(X)-\bar{q})(X-a)$, so $S = S_F + U$. At the *same* nonlinear-selected step $\eta$, algebra gives

$$D_B(S;\eta) - D_B(S_F;\eta) = \eta\mathbb{E}[g_B^\top U] - \frac{\eta^2}{2}\{2\mathbb{E}[S_F^\top H_B U] + \mathbb{E}[U^\top H_B U]\} \tag{92}$$

$$+ \mathbb{E}[\varepsilon_B(S_F,\eta) - \varepsilon_B(S,\eta)]. \tag{93}$$

It follows by subtracting the two definitions (91) and expanding the quadratic form; no independence of $q$ and $X$ is imposed. The first two terms are the quadratic prediction, and the last term is the observed residual difference. Their algebraic sum matching the observed contrast is not by itself predictive evidence. The diagnostic tests whether omitting the last term preserves the sign and approximate magnitude.

An independently fitted $\bar{q}$ need not equal the population $\mathbb{E}q$, so $\mathbb{E}U$ includes a coefficient mean shift, not only $\mathrm{Cov}(q,X)$. If a valid Hessian-Lipschitz condition holds on every segment, (33) bounds the last term; this checkpoint study does not establish that condition.

### I.5 Coefficient diagnostics: conditional positive and negative mechanisms

At the full nonlinear step, the coefficient quadratic prediction and actual contrast agree in point-estimate sign in 16/16 contexts, including all six contrasts beyond the pointwise 3SE screen. Source counts are 1/4/3 and target counts 0/1/7. The observed standard deviation of $q$ is 0.123–0.183. These contrasts use the same evaluation updates for the HVP diagnostic and endpoint comparison, and are not forecasts frozen on a different pool. They also differ from the separately selected-step fixed-mean control in Table 18.

Table 20 gives two source examples with both positive and negative effects.

Table 20: Source coefficient diagnostics at the same nonlinear step (CE units). The observed residual difference is shown separately; it is not a certified remainder bound.

| Seed / step / pair | Alignment | Curvature | Residual difference | Actual contrast |
|---|---|---|---|---|
| 9021 / 16 / 1 | +0.007271 | −0.004588 | +0.000121 | +0.002804 |
| 9022 / 64 / 0 | +0.046010 | −0.080148 | +0.002794 | −0.031344 |

The second example improves first-order alignment, but the Hessian cross term and the curvature energy of the correction outweigh it. Hence "larger coefficient fluctuations are always worse" is not justified. Correction curvature energy is positive in all sixteen observations, but this is an empirical fact, not a consequence of a globally positive Hessian. Raw and nonlinear empirical mean alignment and curvature energy are both positive in 15/16 contexts; the empirical score formed from the three correction moments is negative in 12/15. This is a plug-in local-model check, not a population certificate or a real-loss optimum.

### I.6 Transverse-anchor intervention and prediction reversals

The source-gradient/anchor cosine lies in $[-3.59 \times 10^{-4}, 1.68 \times 10^{-4}]$. Let $g_s$ be the source gradient and decompose $a = a_{\parallel} + a_{\perp}$ relative to it. The intervention freezes the original samplewise $q(X)$ and step, and replaces only the anchor contribution:

$$S_{\parallel} = q(X)X + (1 - q(X))a_{\parallel}, \qquad T = (1 - q(X))a_{\perp}, \qquad S = S_{\parallel} + T.$$

For the source loss, $g_s^{\top} T = 0$ samplewise, so the first-order term is unchanged. The predicted contrast is $-\eta^2 \mathbb{E}[2S_{\parallel}^{\top} H_s T + T^{\top} H_s T]/2$, plus the corresponding endpoint residual difference for actual loss. Orthogonality to $g_s$ does not remove its Hessian cross term. Nor does it imply orthogonality to the target gradient. The intervention uses the exact source gradient and cannot be presented as a cost-free optimizer.

In six of eight source settings the absolute cross term exceeds the pure transverse term. Gradient–anchor transverse-plane projections have a negative eigenvalue in ten of sixteen evaluation contexts. Origin quadratic predictions agree with actual source signs in only 4/8 contexts, or 3/5 among resolved contrasts. Two resolved reversals are listed in Table 21. These effects are mostly of order $10^{-3}$,

Table 21: Source transverse-anchor effects with samplewise retention and step held fixed. Observed residual differences reverse the quadratic prediction.

| Seed / step / pair | Quadratic prediction | Residual difference | Actual contrast |
|---|---|---|---|
| 9021 / 16 / 0 | −0.003512 | +0.005360 | +0.001848 |
| 9021 / 64 / 0 | +0.000416 | −0.000837 | −0.000421 |

smaller than many coefficient and nonlinear–raw contrasts, allowing residuals to dominate. At half steps, only 3/5 resolved transverse contrasts agree. The study therefore does not establish that origin transverse curvature explains historical-anchor utility on the real network. It identifies a boundary of the controlled quadratic explanation.

### I.7 Finite-radius, finite-step, and smoothness limits

For the separately selected-step nonlinear–raw comparison, the origin quadratic model agrees with actual signs in 13/16 point estimates and all 7/7 contrasts beyond the 3SE screen. This is distinct from the 16/16 common-step coefficient statistic. The absolute residual-difference median ratios at half versus full step are 0.337 for nonlinear–raw, 0.255 for the coefficient intervention, and 0.455 for the transverse intervention. Residuals shrink in 13/16, 13/16, and 5/8 contexts, respectively, not uniformly with cubic scaling. Limited symmetric endpoint probes give groupwise $\ell_2$ relative curvature errors of approximately 23–52% against origin HVPs at full step; halving the step does not improve every group.

Two-point ZO slopes have 18.3–41.3% relative RMSE against exact current directional derivatives. In eight extra checked directions per setting, halving the finite-difference radius reduces this error,

but it does not make the current estimator an exact linearized sphere update. The cached model configuration records ReLU activation, and all checks use FP32 eval mode. A smooth Hessian on every update segment cannot be assumed. Activation switching was not measured, and the residuals are not attributed exclusively to it or to numerical precision. No valid $L_H$ envelope was estimated, so the nonquadratic sufficient bounds remain a framework rather than an observed certificate.

The experiments used full gradients and HVPs for paid diagnosis. They establish neither a cost advantage over direct endpoint-loss selection nor stable reuse across new minibatches or subsequent states. These exploratory findings are limited to their original sphere-query, historical-anchor protocol and finite grid. The later results from expanded step searches and independent profiling supersede them as the principal neural evidence, without adding recovered rows as new samples.