AGOZO: ACTIVATION-GUIDED ORTHOGONALIZED ZEROTH-ORDER OPTIMIZATION FOR LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (ZO) fine-tuning reduces memory use by replacing backpropagation with forward evaluations, but limited queries produce noisy gradient estimates. We propose Activation-Guided Orthogonalized ZO (AGOZO), which combines activation-guided queries with gradient orthogonalization. AGOZO uses input activations to select a low-rank query subspace, estimates the gradient within that subspace, and replaces the estimate's nonzero singular values by one before mapping it back to the full weight matrix. Our theory quantifies how much gradient information this operation preserves. Assuming exact directional derivatives along Gaussian perturbations, we derive upper and lower bounds on the update's average alignment with the true gradient. These bounds agree up to constant factors when the subspace rank and query budget are small relative to the layer's output dimension. We also characterize when orthogonalization provides a stronger one-step loss-reduction guarantee than raw activation-guided ZO under a common smoothness assumption. This comparison considers growing matrix dimensions while keeping the number of queries fixed. We further quantify how sharing queries across layers weakens each layer's alignment. Across three language-model backbones and nine tasks, experiments with three seeds and a matched budget of 40,000 forward evaluations show that AGOZO achieves the highest reported mean validation score or a rounded tie on 16 of 27 model–task pairs.

1 INTRODUCTION

Zeroth-order (ZO) fine-tuning estimates update directions from forward evaluations alone and thus avoids the memory cost of backpropagation (Malladi et al., 2023; Zhang et al., 2024). The price is statistical: for a weight matrix $W \in \mathbb{R}^{m\times n}$, an isotropic query spreads its signal over mn coefficients, so the estimator variance grows with the dimension (Duchi et al., 2015; Nesterov & Spokoiny, 2017; Liu et al., 2020). With only a few queries per step, the goal is therefore not to recover the gradient but to extract a useful matrix update from a small query budget.

Activation-guided ZO (AGZO) (Lin et al., 2026) decides *where* to query. For a linear layer, the gradient factors as $G = QH^\top$, where $H \in \mathbb{R}^{n\times b}$ contains the input activations and Q the upstream signals, so the forward pass reveals a subspace that contains the row space of G. Restricting perturbations to the form $RA^\top$, for an orthonormal activation basis $A \in \mathbb{R}^{n\times r}$, reduces the coefficient dimension from mn to mr. The *activation coverage ratio* $\rho_A(G) = \|GA\|_F^2/\|G\|_F^2$ measures the gradient energy that the basis retains; a fresh random subspace of the same rank retains r/n on average.

The basis does not determine *how* to use the estimated coefficients. Write $S = GA$ and let $\widehat{S}$ be its ZO estimate. AGZO steps along $\widehat{S}$. We propose *Activation-Guided Orthogonalized ZO* (AGOZO), which steps along

$$O = \mathrm{msign}(\widehat{S})A^\top.$$

For $\widehat{S} = U\widehat{\Sigma}V^\top$, the polar factor $\mathrm{msign}(\widehat{S}) = UV^\top$ keeps the estimated singular directions and sets all nonzero singular values to one. With an exact gradient, this is steepest descent under an operator-norm constraint, the principle behind Muon (Jordan et al., 2024; Bernstein & Newhouse,

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.

2024); ZO-Muon applies it to ZO estimates in a random subspace (Lang et al., 2026). With a few queries, however, most of $\widehat{S}$ is noise, and flattening its singular values need not preserve the descent signal. This paper asks when it does.

Our first result quantifies the retained signal. For the ideal Gaussian estimator, the expected polar alignment $\mathbb{E}\langle S, \mathrm{msign}(\widehat{S})\rangle$ is positive, grows with the query count q, and admits a spectrum-sensitive lower bound. For tall matrices and few queries, it scales as

$$\|S\|_F \sqrt{\frac{q}{m+r}},$$

with an additive dimension denominator, and a matching upper bound shows that this scale is tight up to universal constants. The bound also resolves the spectral modes: stronger modes saturate sooner, at a contribution linear in their singular value, while weaker modes keep growing (Figure 1, left). The guarantee concerns total expected alignment, not the recovery of each noisy singular vector.

Under spectral smoothness with constant L_*, this alignment yields a one-step descent certificate with two separate factors:

$$\text{certified decrease} \gtrsim \frac{\|G\|_F^2}{L_*} \underbrace{\rho_A(G)}_{\text{retained signal}} \underbrace{\frac{q}{m+r}}_{\text{polar alignment squared}}.$$

The operator-norm second moment of raw AGZO, however, has the same additive order, so the rate alone does not isolate the effect of orthogonalization. We therefore compare the two updates under the *same* upper-curvature model, each with its best scalar step.

Our second result isolates an effect specific to few queries. Holding q fixed while the matrix dimensions grow at a fixed aspect ratio, we derive the limiting ratio of the two certificates explicitly. Orthogonalization removes the spread of the noisy singular values but discards amplitude information. Under spectral smoothness, the first effect dominates for favorable aspect ratios, so the AGOZO certificate exceeds the raw AGZO certificate while the gradient is still poorly estimated. In the isolated-query model, no scalar rescaling of the raw estimate closes this gap, not even an oracle rescaling that knows the gradient. A mixed activation/spectral curvature model shows which curvature makes this redistribution favorable. All comparisons concern one-step certificates for general smooth, possibly nonconvex losses; they do not claim that one optimizer dominates another. First-order analyses relate the benefit of spectral updates to curvature and gradient-rank conditions (Davis & Drusvyatskiy, 2025); we study the noisy ZO estimator at a fixed, small query budget.

Contributions. (i) We propose AGOZO, which orthogonalizes the ZO estimate after activation-guided dimension reduction (Section 2). (ii) We prove a spectrum-sensitive polar-alignment law with additive few-query scale $\sqrt{q/(m+r)}$, make activation coverage and shared-query signal penalties explicit, and turn the law into one-step comparisons with ZO baselines (Section 4). (iii) We derive a fixed-query certificate ratio that separates noise-spectrum flattening from scalar normalization and identifies the curvature under which flattening helps (Section 4.3). (iv) Across three backbones and nine tasks, AGOZO attains the best or tied-best ZO accuracy on 16 of 27 model–task pairs (Section 5). Throughout, we separate the idealized analysis from the Newton–Schulz implementation. Appendices A and B discuss limitations and related work.

2 The AGOZO Method

AGOZO is defined blockwise for each target linear layer; we describe one block $W \in \mathbb{R}^{m\times n}$. Algorithm 3 in Appendix K applies the same construction to every target layer and evaluates all layerwise perturbations jointly, with one finite difference per query. At step t, the clean forward pass produces the input activations $H_t \in \mathbb{R}^{n\times b}$ of the layer. The orthonormal basis $A_t \in \mathbb{R}^{n\times r}$, $A_t^\top A_t = I_r$, consists of the top r left singular vectors of H_t, computed in practice by randomized subspace iteration (Algorithm 1):

$$A_t = \textsc{SubspaceExtract}(H_t, r, K).$$

Because the layer gradient factors as $G_t = Q_t H_t^\top$, its row space satisfies $\mathrm{row}(G_t) \subseteq \mathrm{col}(H_t)$, and A_t spans the leading directions of this activation subspace. We call $S_t := G_t A_t \in \mathbb{R}^{m\times r}$ the

activation-coordinate gradient and $\mathbb{R}^{m\times r}$ the *activation-coordinate space*: each R in this space parameterizes the layer perturbation $RA_t^\top$. Once A_t is formed, only this rank-r basis, not H_t, is kept across the ZO queries.

With A_t fixed, AGOZO samples perturbations in activation-coordinate space: for $i = 1, \dots, q$, it draws $R_{t,i} \overset{\text{iid}}{\sim} N(0, I_{m\times r})$, lifts it to $\Delta_{t,i} = R_{t,i}A_t^\top$, and computes the one-sided finite difference on the minibatch that formed A_t:

$$y_{t,i} = \frac{F(W_t + \mu\Delta_{t,i}) - F(W_t)}{\mu}.$$

For a smooth objective, as $\mu \to 0$,

$$y_{t,i} = \langle G_t, \Delta_{t,i}\rangle + O(\mu\|\Delta_{t,i}\|_F^2) = \langle G_tA_t, R_{t,i}\rangle + O(\mu\|\Delta_{t,i}\|_F^2) = \langle S_t, R_{t,i}\rangle + O(\mu\|\Delta_{t,i}\|_F^2).$$

The activation-coordinate ZO estimator is

$$\widehat{S}_t = \frac{1}{q}\sum_{i=1}^{q} y_{t,i}R_{t,i} \in \mathbb{R}^{m\times r}, \tag{1}$$

and the Gaussian moment identity $\mathbb{E}[\langle S_t, R_{t,i}\rangle R_{t,i}] = S_t$ makes $\widehat{S}_t$ unbiased for $S_t = G_tA_t$ up to finite-difference error.

Algorithm 3 perturbs all target layers at once, with independent Gaussian coefficient matrices $R_{\ell,i}$, so each forward evaluation returns one shared finite difference with first-order expansion

$$y_i = \sum_{\ell}\langle G_{\ell,t}A_{\ell,t}, R_{\ell,i}\rangle + O(\mu\|\Delta_i\|_F^2).$$

Conditional on the batch and bases, $\mathbb{E}[y_iR_{\ell,i}] = G_{\ell,t}A_{\ell,t} + O(\mu)$: cross-layer terms vanish in expectation by independence, but they add variance and perturb the nonlinear polar map. Section 4.3 quantifies this interference.

AGOZO then orthogonalizes the estimate in activation-coordinate space and lifts the result to the layer:

$$W_{t+1} = W_t - \eta_t\ \mathrm{msign}(\widehat{S}_t)A_t^\top. \tag{2}$$

Here $\mathrm{msign}(X) := UV^\top$ is the partial polar factor of a matrix with compact SVD $X = U\Sigma V^\top$; equivalently, $\mathrm{msign}(X) = X(X^\top X)^{-1/2}$ when X has full column rank. Because $A_t^\top A_t = I_r$, right lifting commutes with the polar factor,

$$\mathrm{msign}(\widehat{S}_tA_t^\top) = \mathrm{msign}(\widehat{S}_t)A_t^\top,$$

so AGOZO computes the polar direction of the lifted estimator while applying the matrix sign to an $m \times r$ matrix instead of an $m \times n$ one. The implementation approximates msign by Newton–Schulz iteration (Algorithm 2); the theory uses the exact polar factor unless stated otherwise. Experiments tune the step size η_t, and the theory sets it to maximize the descent certificate.

3 Geometric Foundations

Two deterministic facts describe what AGOZO approximates and how much gradient information the activation-coordinate space retains: the noiseless limit of the update (Section 3.1) and the coverage factor that enters the descent rate (Section 3.2).

3.1 Oracle Limit: A Right-Projected First-Order Polar Direction

If the activation basis spans the top right singular directions of G, the matrix sign in activation coordinates recovers the truncated first-order polar update.

Proposition 3.1. *Let $G \in \mathbb{R}^{m\times n}$ have compact SVD $G = U\Sigma V^\top$, and let $1 \le r \le \mathrm{rank}(G)$, with $V_r \in \mathbb{R}^{n\times r}$ containing its top r right singular vectors. Set $A = V_r$, and let $G_r := U_r\Sigma_rV_r^\top$ be the rank-r truncation of G. Then*

$$\mathrm{msign}(GA)A^\top = U_rV_r^\top = \mathrm{msign}(G_r).$$

For a general orthonormal A, Lemma J.3 gives $\mathrm{msign}(GA)A^\top = \mathrm{msign}(GAA^\top)$, which in general differs from $\mathrm{msign}(G)AA^\top$: projection and polar normalization do not commute. Proposition 3.1 is an oracle statement, because AGOZO observes only the ZO estimate $\widehat{S}_t$ of $S_t = G_t A_t$. The update $\mathrm{msign}(\widehat{S}_t)A_t^\top$ is thus a noisy approximation of the polar direction of the right-projected gradient, and the central question is whether it stays aligned with S_t. Section 4 answers it.

3.2 Activation coverage ratio (ACR)

For a linear layer on a minibatch, the gradient factors as

$$G = QH^\top, \qquad H \in \mathbb{R}^{n\times b}, \qquad Q \in \mathbb{R}^{m\times b}, \tag{3}$$

with input activations H and upstream signals Q, so $\mathrm{row}(G) \subseteq \mathrm{col}(H)$ (Lin et al., 2026). Activation-guided right bases exploit this structural constraint, which data-independent perturbation spaces ignore.

The *activation coverage ratio* (ACR) quantifies the resulting projection bias:

$$\rho_A(G) := \frac{\|GA\|_F^2}{\|G\|_F^2} = 1 - \frac{\left\|G(I - AA^\top)\right\|_F^2}{\|G\|_F^2} \in [0,1]. \tag{4}$$

It equals one when the row space of G lies in $\mathrm{col}(A)$ and decreases as gradient energy leaves the activation-coordinate space. Because $\rho_A(G_t)$ enters the descent certificate multiplicatively, it measures the price of reducing the right dimension from n to r. Proposition 3.2 expresses ρ_A through the activation spectrum and the upstream-signal energy.

Proposition 3.2 (Exact activation capture identity)**.** *Suppose $G = QH^\top \neq 0$ and let $H = U\Sigma V^\top$ be the compact SVD, with $\Sigma = \mathrm{diag}(s_1, \ldots, s_s)$ and $1 \le r \le s$. Let $A = U_r$ be the top-r left singular basis of H, and define $b_i := \|Qv_i\|_2^2$. Then*

$$\rho_A(G) = \frac{\sum_{i=1}^r s_i^2 b_i}{\sum_{i=1}^s s_i^2 b_i}. \tag{5}$$

If $A_{\mathrm{rand}} \in \mathbb{R}^{n\times r}$ is a Haar-random orthonormal basis independent of G, then

$$\mathbb{E}\,\rho_{A_{\mathrm{rand}}}(G) = \frac{r}{n}. \tag{6}$$

Proof. The first identity follows from $G = QV\Sigma U^\top$, giving $\|G\|_F^2 = \sum_{i=1}^s s_i^2 \|Qv_i\|_2^2$ and $\|GU_r\|_F^2 = \sum_{i=1}^r s_i^2 \|Qv_i\|_2^2$. For the random basis, rotational invariance gives $\mathbb{E}[A_{\mathrm{rand}}A_{\mathrm{rand}}^\top] = (r/n)I_n$, so $\mathbb{E}\,\|GA_{\mathrm{rand}}\|_F^2 = (r/n)\,\|G\|_F^2$. □

Coverage is thus high when the leading activation directions carry most of the spectrum-weighted upstream signal, whereas a fresh random right subspace of rank r retains only r/n of the gradient energy in expectation. We measure $\rho_A(G_t)$ with backward-pass gradients during AGOZO training: on 13 model–task pairs, the median coverage of the activation bases is 0.49–0.89, whereas a random rank-16 subspace retains $r/n \le 0.004$ (Appendix N). Appendix D shows coverage across ranks for OPT-13B on SST-2 (Figure 2) and interprets (5) structurally.

4 Main theoretical results

Three factors determine AGOZO's local progress: activation coverage, matrix-sign alignment, and the optimization geometry. Given the coverage $\rho_A(G)$, the question is whether the orthogonalized ZO estimate stays aligned with $S = GA$. Section 4.1 bounds this alignment, Section 4.2 turns it into one-step descent certificates for AGOZO and the ZO baselines, and Section 4.3 isolates what orthogonalization changes when raw AGZO already has the same few-query dimension order.

Setting. Unless stated otherwise, F is a fixed differentiable objective of one block, all other blocks are frozen, and queries return exact directional derivatives. Expectations condition on the iterate, the basis, and, for a minibatch loss, the batch; the basis may depend on this information but not on the fresh Gaussian probes. Descent on the sampled batch does not by itself imply descent of the full-data loss when the basis and the update use the same batch. A one-sided implementation spends $q+1$ forward evaluations, including the clean one, on q directional queries.

4.1 MATRIX-SIGN ALIGNMENT IN ACTIVATION COORDINATES

Fix a nonzero matrix $S \in \mathbb{R}^{m\times r}$. The idealized activation-coordinate ZO estimator is

$$\widehat{S} := \frac{1}{q}\sum_{i=1}^{q}\langle S, R_i\rangle R_i, \qquad R_i \overset{\text{iid}}{\sim} N(0, I_{m\times r}). \tag{7}$$

AGOZO's descent is controlled by the polar alignment $\mathbb{E}\langle S, \operatorname{msign}(\widehat{S})\rangle$. Theorem 4.1 bounds it from below in a spectrum-sensitive way and from above at the matching few-query scale; Appendix E gives the proofs and further discussion.

Theorem 4.1 (Spectrum-sensitive matrix-sign alignment)**.** *Let $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ be the singular values of $\bar{S} := S/\|S\|_F$, so $\sum_j \sigma_j^2 = 1$. Assume $m \geq C_0 r$, where $C_0 = 16$ suffices (Appendix I.2). Then there exist universal constants $c, C > 0$ such that*

$$\mathbb{E}\langle S, \operatorname{msign}(\widehat{S})\rangle \geq c\|S\|_F \sum_{j=1}^{r}\sigma_j \Phi_C\left(\sigma_j\sqrt{\frac{q}{m+r}}\right), \qquad \Phi_C(x) := \frac{x}{\sqrt{1+Cx^2}}. \tag{8}$$

Moreover,

$$\mathbb{E}\langle S, \operatorname{msign}(\widehat{S})\rangle \leq \min\left\{C\|S\|_F\sqrt{\frac{q}{m+r}},\ \|S\|_*\right\}. \tag{9}$$

The saturation function Φ_C is linear for small arguments and tends to a constant for large ones. In the lower bound, mode j therefore first contributes $\sigma_j^2\sqrt{q/(m+r)}$ and saturates near $q_j^\star \asymp (m+r)/\sigma_j^2$. Since $\Phi_C(x) \asymp \min\{x, 1\}$, the lower bound implies

$$\mathbb{E}\langle S, \operatorname{msign}(\widehat{S})\rangle \gtrsim \|S\|_F \sum_{j=1}^{r}\sigma_j \min\left\{\sigma_j\sqrt{\frac{q}{m+r}},\ 1\right\}. \tag{10}$$

Stronger modes saturate first in this bound, which keeps growing through the weaker ones. The guarantee concerns total alignment, not the recovery of individual singular vectors.

Keeping σ_1 in the bound sharpens the uniform rate through the stable rank $\operatorname{sr}(S) := \|S\|_F^2/\|S\|_{\text{op}}^2 = 1/\sigma_1^2$.

Corollary 4.2. *If $m \geq 16r$, then for every $q \geq 1$*

$$\mathbb{E}\langle S, \operatorname{msign}(\widehat{S})\rangle \geq \frac{c}{\sqrt{1 + Cq/((m+r)\operatorname{sr}(S))}}\|S\|_F\sqrt{\frac{q}{m+r}}.$$

Together with (9), the rate is tight up to universal constants whenever $q \lesssim (m+r)\operatorname{sr}(S)$. The proof uses $\sigma_j^2 \leq 1/\operatorname{sr}(S)$ in the denominator of (8). Appendix I.2 gives explicit finite-dimensional constants.

The raw estimator, by comparison, is unbiased with Frobenius second moment

$$\mathbb{E}\|\widehat{S}\|_F^2 = \left(1 + \frac{mr+1}{q}\right)\|S\|_F^2. \tag{11}$$

The leading 1 is the squared mean, $\|\mathbb{E}\widehat{S}\|_F^2 = \|S\|_F^2$, and $(mr+1)/q$ is the variance relative to $\|S\|_F^2$: mr counts the coefficient entries, the $+1$ comes from the Gaussian fourth moment along S, and $1/q$ reflects averaging over independent directions.

Descent depends on both alignment and update size. Equation (11) measures the total squared singular-value energy, whereas spectral smoothness charges only the largest squared singular value, for which Proposition F.1 gives

$$\mathbb{E}\|\widehat{S}\|_{\text{op}}^2 \lesssim \|S\|_{\text{op}}^2 + \frac{m+r}{q}\|S\|_F^2.$$

Raw AGZO therefore also has an additive $m+r$ scale. To compare the two updates at the same root-mean-square operator norm, Figure 1 divides the raw alignment by $\sqrt{\mathbb{E}\|\widehat{S}\|_{\text{op}}^2}$; the polar update already has unit operator norm. Normalizing only the raw update by (11) would mix norm budgets.

The dotted lines in Figure 1 (left) mark the large-q limits, $\|S\|_* \approx 2.179$ for polar and $\|S\|_F^2/\|S\|_{\text{op}} \approx 1.352$ for raw; our descent comparison keeps q small. For every shape and fixed $S \neq 0$, the mean polar alignment is positive, nondecreasing in q, and tends to $\|S\|_*$ (Proposition F.2); only the quantitative bound requires $m \geq 16r$.

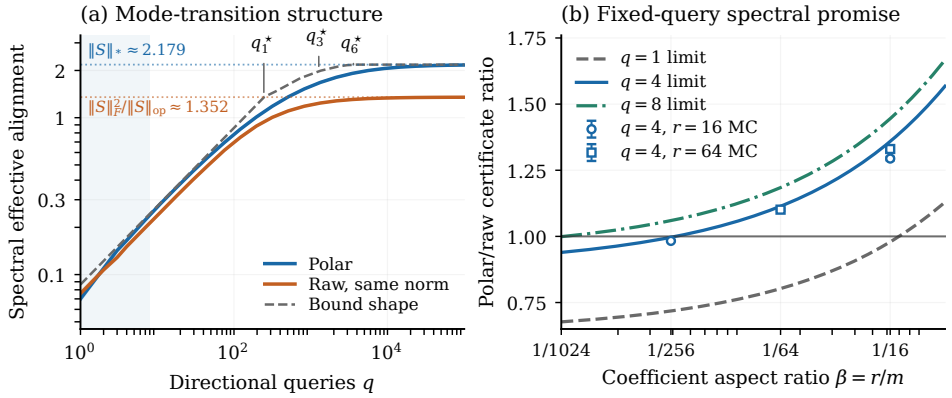


Figure 1: Few-query alignment and the fixed-query certificate ratio. *Left:* normalized alignment $\mathbb{E}\langle S, U\rangle / \sqrt{\mathbb{E}\|U\|_{\mathrm{op}}^2}$ of the polar and raw updates, shown against the hard-saturation bound shape (dashed, constants omitted); polar alignment grows as $\sqrt{q}$ at few queries; $q_j^\star = (m+r)/\sigma_j^2$ marks bound transitions, not measured recovery thresholds. Setup: $m = 128$, $r = 6$, $\|S\|_F = 1$, $\sigma_j^2 \propto j^{-1.5}$; 4096 samples per budget, two-standard-error bands; shading marks $q \le 8$. *Right:* the limit $G_q(\beta)$ of (13) increases with the aspect ratio β and with q; values above one favor the polar certificate. Markers are $q = 4$ Gaussian Monte Carlo (MC) estimates at two ranks, with two-standard-error bars. Appendix G.5 specifies both panels and all finite checks.

4.2 One-step descent and comparison with ZO baselines

A smoothness model turns alignment into a descent guarantee. For $\nu \in \{F, \mathrm{op}\}$, with $L_{\mathrm{op}} := L_*$, assume

$$F(W + \Delta) \le F(W) + \langle G, \Delta\rangle + \tfrac{1}{2} L_\nu \|\Delta\|_\nu^2$$

on the support of the analyzed displacements; this covers general smooth, nonconvex losses. For $W^+ = W - \eta U$, where U has positive mean alignment and a finite second moment of $\|U\|_\nu$, a step size that is deterministic given the conditioning information yields the *descent certificate*

$$\mathbb{E}[F(W) - F(W^+)] \ge \mathcal{D}_\nu(U) := \frac{(\mathbb{E}\langle G, U\rangle)^2}{2L_\nu \mathbb{E}\|U\|_\nu^2}. \tag{12}$$

This step maximizes the smoothness lower bound and may depend on the unknown gradient (Appendix C).

Theorem 4.3 (Few-query one-step progress factors)**.** *Let $G = \nabla F(W) \neq 0$. For each ideal isolated-block estimator in Table 1, condition on the iterate and any orthonormal basis, independently of the fresh probes, and use the indicated smoothness norm. If its estimated matrix has shape $a \times b$, assume $q \lesssim a + b$; each polar row additionally requires $\max(a, b) \ge 16 \min(a, b)$. Then an oracle scalar step size, deterministic conditional on this information, gives*

$$\mathbb{E}[F(W) - F(W^+)] \gtrsim \Gamma \, \|G\|_F^2,$$

with Γ given in the table.

Here $\rho_P = \|GP\|_F^2/\|G\|_F^2$ and $\rho_P^{\mathrm{L}} = \|P^\top G\|_F^2/\|G\|_F^2$ are the gradient-energy fractions retained by rank-r_P right and left orthonormal bases; for fresh Haar bases independent of G, their expectations are r_P/n and r_P/m. Appendix E.2 gives the proofs and the remaining crossover conditions.

Table 1 separates retained signal from query dimension. Under a common L_*, the AGOZO certificate scales better than that of full-space polar ZO when $\rho_A \gtrsim (m+r)/(m+n)$, and better than the mean certificate of a fresh Haar right subspace of the same rank when $\rho_A \gtrsim r/n$. Setting the hidden constants equal and using the ranks of our experiments, measured coverage exceeds both random-subspace thresholds on all 59,520 layer measurements and the full-space threshold on 73.5% of them (Appendix N). These conditions compare scaling certificates, not actual loss decreases, and

Table 1: Few-query one-step descent factors from Theorem 4.3, with universal constants suppressed. The estimated matrix determines the shape (a, b) in the theorem. The norm column specifies the smoothness model, with common L_F or L_* across its rows; these constants are not interchangeable. For fresh rank-r_L Gaussian-product LOZO probes, $\kappa_{\mathrm{LOZO}} = mn + 1 + 2(m+n+1)/r_L$.

Method	Estimated matrix	Norm	Descent factor Γ
MeZO	G	Frobenius	$q/(L_F mn)$
LOZO, fresh rank r_L	G	Frobenius	$q/(L_F \kappa_{\mathrm{LOZO}})$
AGZO, raw	GA	Frobenius	$\rho_A q/(L_F mr)$
AGZO, raw	GA	Operator	$\rho_A q/[L_*(m+r)]$
ZO-Muon, full space	G	Operator	$q/[L_*(m+n)]$
ZO-Muon, random right	GP	Operator	$\rho_P q/[L_*(m+r_P)]$
ZO-Muon, random left	$P^\top G$	Operator	$\rho_P^{\mathrm{L}} q/[L_*(n+r_P)]$
AGOZO	GA	Operator	$\rho_A q/[L_*(m+r)]$

full-space matrices with aspect ratio below 16, including all measured full-space shapes, lie outside the finite-dimensional polar guarantee.

What does the smoothness ratio imply? Against the Frobenius certificate of raw AGZO, coverage cancels and the proxy criterion becomes $L_F/L_* \gtrsim (m+r)/(mr)$. For the smallest positive global smoothness constants, norm equivalence gives $1/\min(m,n) \le L_F/L_* \le 1$. Both activation-guided updates lie in the smaller space $\{\Delta = BA^\top\}$, where the minimal positive constants $L_{F,A}$ and $L_{*,A}$ satisfy

$$\frac{1}{r} \le \frac{L_{F,A}}{L_{*,A}} \le 1, \qquad R_{\mathrm{cross}} := \frac{L_{F,A}}{L_{*,A}} \frac{mr}{m+r} \in \left[\frac{m}{m+r}, \frac{mr}{m+r}\right].$$

The ratio R_{cross} compares the AGOZO/operator and AGZO/Frobenius factors on this common space. Its bounds follow from $\|\Delta\|_{\mathrm{op}}^2 \le \|\Delta\|_F^2 \le r\|\Delta\|_{\mathrm{op}}^2$ and hold for minimal constants only, not for arbitrary choices or restricted/full-space pairings (Appendix E.2). For $m \ge 16r$, $R_{\mathrm{cross}} \ge 16/17$; if $L_{*,A} \approx L_{F,A}$, R_{cross} is of order r.

These cross-norm factors suggest comparable or better scaling for AGOZO, but they do not isolate a matrix-sign benefit: universal prefactors are suppressed, and raw AGZO also admits $L_{*,A}$ through its operator-norm moment bound. Under this *common* geometry, both updates scale as $\rho_A q/[L_{*,A}(m+r)]$. Section 4.3 fixes this geometry and asks whether redistributing the noisy singular values improves the certificate.

4.3 What orthogonalization changes beyond the common rate

Fix the activation basis, the queries, and the curvature information. Write $f(B) = F(W + BA^\top)$, so that $\nabla f(0) = S$, and apply (12) under a common operator-norm upper model on this subspace. With $a_q = \mathbb{E}\langle S, \mathrm{msign}(\widehat{S})\rangle$ and $M_q = \mathbb{E}\|\widehat{S}\|_{\mathrm{op}}^2$, the two certificates are $\mathcal{D}_{\mathrm{polar}} = a_q^2/(2L_*)$ and $\mathcal{D}_{\mathrm{raw}} = \|S\|_F^4/(2L_* M_q)$.

Theorem 4.4 (Fixed-query spectral redistribution)**.** *Fix an integer $q \ge 1$. Let $m, r \to \infty$ with $r/m \to \beta \in (0,1)$, and use the ideal estimator (7). For arbitrary nonzero gradients S,*

$$\frac{\mathcal{D}_{\mathrm{polar}}}{\mathcal{D}_{\mathrm{raw}}} \longrightarrow G_q(\beta) := \kappa_q \mu_\beta^2 (1+\sqrt{\beta})^2, \qquad \kappa_q = \frac{(\mathbb{E}\sqrt{\chi_q^2})^2}{q}. \tag{13}$$

Here μ_β is the limiting mean singular value of a standard Gaussian $m \times r$ matrix divided by $\sqrt{m}$. The same limit holds if the denominator is the best certificate among all nonnegative sample-dependent scalar rescalings of $\widehat{S}$, including ones that know S. The convergence is uniform over $S/\|S\|_F$.

Appendix G proves the theorem and gives the integral for μ_β. The mechanism is *noise-spectrum flattening*, not high-query recovery of the true gradient: the factor $\mu_\beta^2(1+\sqrt{\beta})^2$ measures the gain from redistributing the noisy singular values, and $\kappa_q < 1$ the loss from discarding query amplitudes.

Figure 1 (right) plots $G_q(\beta)$ for $q = 1, 4, 8$. Values above one favor the polar certificate; values below one, which occur for thin shapes at small q, favor raw. The markers estimate the exact finite-size ratio $a_q^2 M_q / \|S\|_F^4$ for synthetic spiked gradients. At $\beta = 1/16$, the $q = 4$ limit is 1.359, and finite checks at 256×16 and 1024×64, which share this aspect ratio, give 1.294 and 1.330; Appendix G.5 lists the gradients, sample counts, and uncertainty. These are certificate comparisons, not training curves or finite-size guarantees, and, unlike the checks, the theorem holds for every nonzero gradient.

Which curvature favors flattening? The comparison extends to the common upper model $\Psi(B) = \operatorname{tr}(BCB^\top) + \lambda \|B\|_{\mathrm{op}}^2$, with $C \succeq 0, \lambda \geq 0, \operatorname{tr} C + \lambda > 0$. Writing $\omega = \lambda / (\operatorname{tr} C + \lambda)$ for its limiting spectral fraction, the certificate ratio becomes

$$G_q(\beta, \omega) = \kappa_q \mu_\beta^2 \big[1 + \omega(2\sqrt{\beta} + \beta)\big]. \tag{14}$$

In this limit, the polar certificate is larger at small q when $G_q(\beta, \omega) > 1$; at $q = 4$ and $\beta = 1/16$, this holds for $\omega > \omega^\star \approx 0.266$ (Appendix G.3). Pure activation curvature ($\omega = 0$) favors raw in this limit, so the spectral share of the curvature matters.

4.4 Shared queries across layers

The preceding analysis isolates one block, whereas Algorithm 3 perturbs all target layers jointly. To quantify how this affects alignment, we use ideal directional derivatives and fix the iterate, batch, and bases. Write $S_\ell = G_\ell A_\ell$. Blocks use independent Gaussian probes, so the other blocks contribute Gaussian noise independent of block ℓ's probes. Theorem F.3 gives

$$\theta_\ell := \frac{\|S_\ell\|_F^2}{\sum_k \|S_k\|_F^2}, \qquad \mathbb{E}\langle S_\ell, \operatorname{msign}(\widehat{S}_\ell)\rangle \asymp \|S_\ell\|_F \sqrt{\frac{\theta_\ell q}{m_\ell + r_\ell}},$$

for $S_\ell \neq 0$, $m_\ell \geq 16 r_\ell$, and $q \lesssim m_\ell + r_\ell$, with the sum over all independently Gaussian-perturbed parameter groups. Block ℓ has the same alignment scaling as an isolated block with an effective query count of $\theta_\ell q$. For L equally energetic blocks, this gives a $1/\sqrt{L}$ alignment penalty relative to isolated queries, with $q + 1$ forward evaluations serving all layers. ACR measures projection loss within a block; θ_ℓ measures signal sharing *across* blocks.

At fixed q and $\theta_\ell > 0$, sharing queries multiplies both raw and polar blockwise certificates by θ_ℓ to leading order in the fixed-aspect limit. Under the same curvature model, their limiting ratio $G_q(\beta, \omega)$ is unchanged (Appendix G.4). Joint descent requires the joint curvature bound of Appendix F.3. The scalar-oracle comparison has been proved only for isolated queries.

5 Experiments

Setup. We fine-tune OPT-1.3B, OPT-13B (Zhang et al., 2022), and Llama-3-8B on nine tasks: five classification, two multiple-choice, and two generation tasks (Table 2). The ZO baselines are MeZO (Malladi et al., 2023), LOZO (Chen et al., 2025), subspace MeZO (S-MeZO), ZO-Muon (Lang et al., 2026), and AGZO (Lin et al., 2026); all methods use a matched budget of 40k forward queries. AGOZO uses $q = 4$ queries per step for 8k steps and activation rank $r = 16$ (64 on five OPT cells); Appendix L gives further details.

Accuracy. Table 2 reports the mean and standard deviation of each run's best validation score over three training seeds. AGOZO attains the best or tied-best ZO result on 16 of the 27 model–task pairs, and ZO-Muon on the remaining 11. AGOZO is best or tied-best on both generation tasks and on ReCoRD for every backbone, and its largest margin over ZO-Muon is 8.4 points (OPT-13B, CB); ZO-Muon leads on WiC and COPA for every backbone and on four of the five classification tasks for Llama-3-8B. Several margins in both directions lie within one seed standard deviation. AGOZO also outperforms AGZO on all 27 pairs. The two use the same type of activation subspace but differ in queries per step (four versus one) and in the tuned learning rate, so this comparison does not isolate the matrix-sign step; Appendix O gives the 2×2 subspace/update ablation and documents these differences.

Cost. AGOZO's total training time is 0.92–1.05× that of ZO-Muon on every model–task pair, basis extraction takes 0.7–1.4% of its step time in a four-setting profile, and within each pair the peak

Table 2: Fine-tuning results over **three training seeds** (mean ± std of best dev accuracy) at a matched 40k-forward-query budget. All rows are zeroth-order methods; the best ZO result for each task is in bold.

Task	SST-2	RTE	CB	BoolQ	WiC	COPA	ReCoRD	SQuAD	DROP
Task type	classification					multiple choice		generation	
OPT-1.3B									
MeZO	75.2 ± 9.8	58.7 ± 1.0	64.3 ± 1.5	64.0 ± 2.1	56.0 ± 1.0	72.0 ± 0.8	69.5 ± 0.8	31.1 ± 1.6	14.3 ± 1.8
LOZO	80.2 ± 1.5	56.4 ± 1.4	65.5 ± 1.7	63.1 ± 2.2	56.6 ± 1.1	74.0 ± 1.4	68.8 ± 1.0	28.8 ± 3.6	13.3 ± 1.1
S-MeZO	60.7 ± 3.4	57.8 ± 1.1	66.1 ± 4.4	62.9 ± 2.7	56.4 ± 0.6	73.7 ± 0.5	68.7 ± 1.1	29.1 ± 0.5	12.8 ± 0.7
ZO-Muon	86.1 ± 1.1	62.8 ± 0.8	70.2 ± 1.7	67.8 ± 1.1	**58.3 ± 1.2**	**79.0 ± 1.4**	**71.9 ± 0.2**	74.5 ± 1.1	25.1 ± 0.4
AGZO	79.5 ± 5.3	54.6 ± 1.2	60.7 ± 2.5	62.8 ± 2.3	54.9 ± 1.2	72.3 ± 2.4	68.7 ± 0.3	36.0 ± 1.2	16.6 ± 1.1
AGOZO	**90.0 ± 0.7**	**64.5 ± 1.0**	**73.2 ± 1.5**	**69.0 ± 0.6**	57.5 ± 2.1	77.3 ± 2.5	**71.9 ± 0.5**	**77.3 ± 1.1**	**25.7 ± 1.3**
OPT-13B									
MeZO	73.4 ± 2.1	57.6 ± 0.2	67.3 ± 4.7	67.3 ± 1.3	56.4 ± 0.9	81.3 ± 1.2	79.2 ± 0.9	54.8 ± 1.7	21.5 ± 0.9
LOZO	63.1 ± 0.6	57.0 ± 0.3	58.9 ± 2.9	64.4 ± 2.0	56.0 ± 0.8	80.7 ± 1.2	78.8 ± 0.8	45.2 ± 3.5	15.5 ± 0.3
S-MeZO	63.3 ± 0.6	56.9 ± 0.7	59.5 ± 4.5	65.0 ± 2.3	55.5 ± 1.6	81.0 ± 0.8	78.6 ± 0.8	46.3 ± 3.3	15.9 ± 0.5
ZO-Muon	**91.4 ± 1.4**	65.9 ± 2.4	69.6 ± 0.0	**73.0 ± 1.4**	**58.9 ± 3.0**	**84.7 ± 0.5**	81.5 ± 1.2	74.3 ± 1.8	28.0 ± 0.6
AGZO	68.4 ± 4.1	55.8 ± 1.6	64.3 ± 3.9	66.5 ± 2.9	52.3 ± 1.1	80.3 ± 0.5	79.3 ± 1.1	62.6 ± 2.2	24.8 ± 0.5
AGOZO	91.3 ± 0.9	**69.0 ± 2.1**	**78.0 ± 1.7**	72.6 ± 1.4	57.9 ± 1.8	84.3 ± 0.5	**81.8 ± 1.3**	**79.6 ± 0.2**	**29.7 ± 1.0**
Llama-3-8B									
MeZO	70.4 ± 2.4	58.8 ± 1.6	66.1 ± 3.9	64.4 ± 1.6	55.5 ± 1.0	85.3 ± 0.5	72.1 ± 7.7	71.2 ± 2.1	35.9 ± 3.3
LOZO	62.1 ± 4.9	57.5 ± 0.7	64.3 ± 2.9	62.5 ± 2.8	53.3 ± 0.8	85.3 ± 0.9	72.2 ± 6.8	59.2 ± 1.2	31.9 ± 2.7
S-MeZO	63.0 ± 4.0	60.6 ± 1.3	67.9 ± 2.9	63.3 ± 2.1	56.4 ± 0.7	86.0 ± 0.8	70.9 ± 6.8	69.0 ± 1.4	36.7 ± 1.6
ZO-Muon	**84.5 ± 0.5**	**72.4 ± 1.9**	**81.5 ± 1.7**	70.6 ± 1.8	**61.1 ± 3.0**	**93.0 ± 0.0**	75.7 ± 6.5	89.6 ± 0.5	54.0 ± 1.4
AGZO	75.6 ± 5.3	59.5 ± 0.7	64.3 ± 1.5	65.2 ± 2.0	54.5 ± 0.7	85.7 ± 0.5	71.0 ± 7.8	72.6 ± 2.1	35.7 ± 4.1
AGOZO	84.2 ± 2.4	67.0 ± 2.2	78.6 ± 0.0	**71.3 ± 0.2**	56.1 ± 0.5	91.7 ± 0.9	**75.9 ± 6.8**	**91.1 ± 1.2**	**54.3 ± 0.8**

memory of the six methods differs by at most 2.6 GB (Appendix M). In an earlier single-run OPT-13B study (Appendix L), AGOZO reaches low training loss earlier than the other ZO baselines and reaches 85% SST-2 accuracy in 600 steps (440 s), versus 3.2k–7.3k steps (1652–2089 s) for the other ZO methods.

6 Conclusion

AGOZO combines activation-guided queries with orthogonalization to construct matrix updates from few forward evaluations. For ideal Gaussian directional queries, we derive upper and lower bounds on expected gradient alignment that agree up to constants in the analyzed few-query regime. We also identify when orthogonalization gives a stronger one-step descent guarantee than raw AGZO under the same smoothness assumption. This comparison considers matrix dimensions growing proportionally with a fixed query count. The advantage depends on matrix shape and curvature. For isolated queries, it can persist even against the best scalar rescaling of the raw estimate. We further quantify how subspace projection and query sharing limit the available gradient signal. Empirically, AGOZO achieves the highest reported mean validation score or a rounded tie among the compared ZO methods on 16 of 27 model–task pairs at a matched query budget. These results assess downstream performance, whereas the theory concerns idealized one-step training-loss reduction.

AI use statement

Generative AI tools were used to assist with writing and editing the text, implementing and running experiments (including hyperparameter tuning), producing figures, checking mathematical claims and proofs, polishing the presentation of human-written proofs, suggesting experimental designs, and literature discovery. We have reviewed all AI-assisted work, verified the code and the proofs, and take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

The smoothness assumptions used by our analysis are stated in Appendix C, and complete proofs of all theoretical results are given in Appendices E–I together with the auxiliary lemmas. Pseudocode for

AGOZO is given in Appendix K, and models, datasets, baselines, query budgets and hyperparameters are listed in Appendix L, which also links an anonymized code repository.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328. Association for Computational Linguistics, 2021.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: 16th Annual Workshop on Optimization for Machine Learning*, 2024.

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*, pp. 62581–62607, 2025.

Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, 2023. doi: 10.1109/OJCSYS.2023.3315088.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025. URL https://arxiv.org/abs/2512.04299.

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.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. arXiv:1812.04754 [cs.LG], December 2018. URL https://arxiv.org/abs/1812.04754.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26740–26756. PMLR, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199(1):1179–1219, 2023.

Yicheng Lang, Changsheng Wang, Yihua Zhang, Mingyi Hong, Zheng Zhang, Wotao Yin, and Sijia Liu. Powering up zeroth-order training via subspace gradient orthogonalization. *arXiv preprint arXiv:2602.17155*, 2026.

Wei Lin, Qingyu Song, and Hong Xu. The multi-query paradox in zeroth-order optimization. *arXiv preprint arXiv:2509.15552*, 2025.

Wei Lin, Yining Jiang, Qingyu Song, Qiao Xiang, and Hong Xu. AGZO: Activation-guided zeroth-order optimization for LLM fine-tuning. In *International Conference on Machine Learning*, 2026.

Andy Zeyi Liu, Elliot Paquette, and John Sous. Evolution of the spectral dimension of transformer activations. In *OPT 2025: Optimization for Machine Learning*, 2025a.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O. Hero, and Pramod K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020. doi: 10.1109/MSP.2020.3003837.

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*, 2025b.

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.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Ryota Nozawa, Pierre-Louis Poirion, and Akiko Takeda. Zeroth-order random subspace algorithm for non-smooth convex optimization. *Journal of Optimization Theory and Applications*, 204(3):53, 2025.

Elliot Paquette, Ke Liang Xiao, and Yizhe Zhu. Power-law spectrum of the random feature model. *arXiv preprint arXiv:2603.14578*, 2026.

Egor Petrov, Grigoriy Evseev, Aleksey Antonov, Andrey Veprikov, Nikolay Bushkov, Stanislav Moiseev, and Aleksandr Beznosikov. Leveraging coordinate momentum in signsgd and muon: Memory-optimized zero-order. *arXiv preprint arXiv:2506.04430*, 2025.

Michael Rotman and Lior Wolf. Natural statistics of network activations and implications for knowledge distillation. In *IEEE International Conference on Image Processing (ICIP)*, pp. 399–403, 2021. doi: 10.1109/ICIP42928.2021.9506486.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, chapter 5, pp. 210–268. Cambridge University Press, 2012. URL https://arxiv.org/abs/1011.3027.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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, 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. *arXiv preprint arXiv:2205.01068*, 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.

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*, pp. 43496–43520, 2025.

A Limitations and Future Work

Activation projection. AGOZO intentionally discards the component $G(I - AA^\top)$ and optimizes using the retained coefficient gradient GA. This is the right tradeoff when the activation coverage ratio $\rho_A(G)$ is large, but if $\rho_A(G)$ is small, then most of the true gradient lies outside the chosen activation subspace. In that case, matrix-sign orthogonalization can use the retained signal more effectively, but it cannot recover gradient components removed by the projection. A next step is to use the activation coverage ratio $\rho_A(G) = \|GA\|_F^2/\|G\|_F^2$ as an online diagnostic for adaptive rank selection, layerwise basis refresh schedules, or fallback rules that increase r when the activation basis is not capturing enough gradient energy.

Estimator and polar-map stability. The alignment theory focuses on an idealized Gaussian activation-coordinate estimator and the exact polar factor. Practical implementations use finite differences, mini-batches, finite Newton–Schulz iterations, and sometimes tuned matrix-sign polynomials. These details matter because the polar map is not globally Lipschitz near rank-deficient or ill-conditioned matrices: a small estimation error can rotate the polar direction when singular values are too small or poorly separated. Future theory should quantify finite-difference bias, stochastic-batch noise, and Newton–Schulz approximation error. Appendix F.5 gives a conditional transfer bound in terms of the update's alignment defect, but does not estimate that defect for the implemented polynomial.

Optimization model. Our descent theory is local and comparative. It gives one-step certificates under Frobenius or spectral smoothness; comparing lower bounds does not prove that one optimizer's actual decrease exceeds another's. Full-space rate extrapolations outside the tall-matrix assumption are not certified. This level suffices to isolate the mechanism but leaves several extensions open. One important direction is a full training-dynamics theorem with momentum, adaptive learning rates, and changing minibatches and activation bases. Appendix F.3 already handles shared-query interference and joint one-step updates under an explicit blockwise curvature bound. Such a result should track layerwise ranks, layerwise activation bases, and a shared forward-query budget across blocks using the same activation coverage and polar-alignment quantities developed here.

Empirical scope. Our experiments evaluate the practical AGOZO pipeline on memory-efficient LLM fine-tuning tasks. They do not by themselves isolate the fixed-query mechanism: the main raw AGZO comparison uses a different query count, and exact polar, scalar-normalized raw, and implemented Newton–Schulz updates are distinct controls. Theorem 4.4 suggests measuring alignment and update spectra at small matched query counts; its curvature extension also calls for checking which upper model is informative for the layer. The highest-priority training control is raw AGZO with per-layer scalar normalization at the same query count, subspace, and learning-rate tuning budget. Rank and query sweeps can then test where the predicted promise appears without treating a certificate comparison as a performance guarantee.

B Related Work

Zeroth-order fine-tuning and estimator design. Classical zeroth-order (ZO) methods estimate gradients from finite differences of function values and have dimension-dependent variance and query complexity (Duchi et al., 2015; Nesterov & Spokoiny, 2017; Liu et al., 2020). MeZO made this approach practical for memory-efficient language-model fine-tuning, using random-direction ZO updates computed from forward passes only (Malladi et al., 2023), and later benchmarking work clarified the regimes in which such methods are competitive or fragile (Zhang et al., 2024). A first line of recent work improves the ZO estimator while remaining close to the ambient parameter-space formulation: Sparse MeZO perturbs fewer coordinates (Liu et al., 2025b), HiZOO uses Hessian-informed scaling (Zhao et al., 2025), and multi-query methods average or align several directional estimates (Lin et al., 2025). These methods address the variance of the raw finite-difference estimator,

but they do not use the layerwise activation identity $G = QH^\top$ to choose a problem-dependent right subspace, nor do they change the raw vector-like update into a spectrally normalized matrix direction. AGOZO adds both: activations decide where to spend queries, and the polar factor decides how to use the noisy matrix estimate.

Low-rank and random-subspace ZO. Another way to reduce ZO variance is to restrict perturbations to a lower-dimensional subspace. LOZO uses low-rank matrix perturbations for language-model fine-tuning (Chen et al., 2025), while SubZero uses random orthonormal sketches to estimate updates in a lower-dimensional matrix space (Yu et al., 2025). ZO-Muon also begins from a projected-subspace view and uses a random low-rank projection before applying matrix-sign orthogonalization (Lang et al., 2026). Related theory studies ZO with orthogonal random directions or random subspaces in more classical optimization settings (Kozak et al., 2023; Nozawa et al., 2025). These methods reduce the search dimension, but their subspaces are data-independent or only weakly tied to the current batch geometry. In contrast, AGOZO uses the activation basis A as a right subspace for a specific layer and batch. This distinction matters because a random rank-r right subspace captures only r/n of the gradient energy in expectation, whereas the activation capture

$$\rho_A(G) = \frac{\|GA\|_F^2}{\|G\|_F^2}$$

is measurable and can be much larger when the activation span matches the gradient row space. The comparison with random-subspace ZO-Muon is therefore quantitative: at equal coefficient dimension, it hinges on whether $\rho_A(G)$ exceeds the coverage of the random subspace.

Activation-guided ZO. AGZO is the closest predecessor to our method (Lin et al., 2026). It observes that, for a linear layer with input activation matrix H, the first-order gradient factors as $G = QH^\top$. Consequently, perturbing directions orthogonal to the activation span wastes queries for that batch. AGZO uses this observation to replace an mn-dimensional layerwise perturbation by an mr-dimensional activation-coordinate perturbation. AGOZO keeps this activation-guided estimator, so it inherits the same projection bias $G(I - AA^\top)$ and the same need for a large activation coverage ratio. The difference is what happens after the coefficient estimate is formed. AGZO uses the noisy estimate $\widehat{S}$ as a raw coefficient-gradient step, whose Frobenius second moment involves the mr coefficient dimension. AGOZO instead applies $\mathrm{msign}(\widehat{S})$, preserving the estimated left and right singular directions while removing the unreliable singular-value scale. The relevant alignment becomes $\langle GA, \mathrm{msign}(\widehat{S})\rangle$, and our theory shows that this polar alignment is controlled by $m + r$. These statements use different norm constraints; under spectral smoothness the raw update has the same few-query dimension scaling (Proposition F.1). The theoretical distinction from AGZO is therefore not the activation projection itself, but the analysis after projection: we prove spectrum-sensitive expected alignment and, at fixed small q, identify when flattening the noise spectrum gives a larger common-model certificate, including against scalar-normalized raw updates. AGZO answers where to query; AGOZO also answers how to turn the resulting matrix estimate into a descent direction.

Muon, matrix-sign updates, and ZO-Muon. Muon applies a matrix-sign or polar factor to first-order matrix gradients and corresponds to steepest descent under a spectral-norm constraint (Jordan et al., 2024; Bernstein & Newhouse, 2024). This geometry suits matrix layers because it normalizes singular values rather than treating the gradient as an unstructured vector. Extending this idea to ZO is difficult: the ZO estimator is noisy, and the polar map is nonlinear and can rotate substantially when applied to high-dimensional noisy matrices. ZO-Muon identifies this failure mode empirically for full-space ZO orthogonalization and proposes subspace gradient orthogonalization as a remedy (Lang et al., 2026); related direct full-space attempts such as JAGUAR-Muon apply gradient orthogonalization to ZO momentum but incur additional overhead and limited gains (Petrov et al., 2025). AGOZO agrees with the central lesson of ZO-Muon: orthogonalization should be applied after dimension reduction. The distinction is the source and geometry of the reduced space. ZO-Muon uses a random projected subspace, while AGOZO uses an activation-informed right subspace and lifts the result as $\mathrm{msign}(\widehat{S})A^\top$. In the noiseless limit this recovers the polar direction of the projected gradient, $\mathrm{msign}(GA)A^\top$; in the noisy ZO regime, our main results prove spectrum-sensitive alignment bounds for this coefficient-space polar update. This is also a theoretical distinction from ZO-Muon: instead of treating subspace orthogonalization through a random projection alone, our analysis exposes the two

quantities that determine whether the activation-guided version is favorable, namely the activation coverage ratio $\rho_A(G)$ and the few-query polar-alignment term for $\mathrm{msign}(\widehat{S})$. This yields explicit comparison criteria against raw AGZO, random-subspace ZO-Muon, and full-space ZO-Muon.

Low-dimensional structure in fine-tuning and activations. The use of an activation basis is motivated by a broad empirical literature on low-dimensional structure in neural-network training. Intrinsic-dimension and low-rank-gradient studies show that fine-tuning updates often concentrate in small subspaces (Aghajanyan et al., 2021; Gur-Ari et al., 2018; Cosson et al., 2023). Parameter- and optimizer-efficient methods such as LoRA and related low-rank approaches exploit this structure from a first-order or parameter-efficient perspective (Hu et al., 2022; Jaiswal et al., 2025). Separately, empirical studies report rapidly decaying or heavy-tailed activation spectra in modern networks (Rotman & Wolf, 2021; Liu et al., 2025a), and theory for random-feature models gives one mechanism by which power-law spectral structure can persist through nonlinear layers (Paquette et al., 2026). These works support the plausibility of activation capture, but they do not by themselves guarantee descent for a ZO method. Our analysis therefore keeps $\rho_A(G)$ explicit: activation guidance is useful only when it captures enough gradient energy, and orthogonalization is useful only when the noisy polar direction remains aligned with GA.

Summary of the distinction. AGOZO is not simply MeZO with a smaller perturbation space, AGZO with a normalization step, or ZO-Muon with a different projection matrix. Compared with MeZO-style estimators, AGOZO replaces ambient random perturbations by activation-coordinate queries. Compared with LOZO, SubZero, and random-subspace ZO-Muon, it uses a data-dependent right subspace whose quality is measured by ACR. Compared with AGZO, it changes the update geometry from raw coefficient estimation to matrix-sign alignment and proves the resulting query dependence. Compared with first-order Muon and ZO-Muon, it proves that the lifted activation-coordinate polar update is the polar factor of a right-projected gradient in the oracle limit and has positive expected alignment under noisy few-query ZO estimation. The resulting theory separates the progress of AGOZO into activation capture and polar-alignment factors, which gives explicit criteria for comparing local descent certificates with raw, random-subspace, and full-space ZO baselines under the stated assumptions.

When spectral updates are promising. Davis & Drusvyatskiy (2025) compare first-order spectral and Euclidean updates through the gradient's squared nuclear-to-Frobenius ratio and the stable rank of incoming activations. Our activation-curvature model is consistent with that first-order viewpoint. The distinction here is the few-query ZO regime: the true gradient need not be spectrally resolved, and the leading certificate ratio instead depends on Gaussian noise-spectrum shape, query-amplitude loss, and the available curvature model. Theorem G.2 also compares with a gradient-informed scalar oracle, separating spectral redistribution from mere rescaling, while Appendix G.4 treats the raw/polar comparison under shared-query interference.

C Smoothness Assumptions and Descent Lemmas

We use two standard smoothness conditions on the per-layer loss $F\colon \mathbb{R}^{m\times n} \to \mathbb{R}$, one adapted to Frobenius-geometry methods and one adapted to spectral-geometry methods.

Definition C.1 (L_F-smoothness)**.** We say F is L_F-smooth with respect to the Frobenius norm if for all $W, W' \in \mathbb{R}^{m\times n}$,

$$F(W') \le F(W) + \langle \nabla F(W),\, W' - W\rangle + \frac{L_F}{2}\|W' - W\|_F^2.$$

Definition C.2 (L_*-smoothness)**.** We say F is L_*-smooth with respect to the operator norm if for all $W, W' \in \mathbb{R}^{m\times n}$,

$$F(W') \le F(W) + \langle \nabla F(W),\, W' - W\rangle + \frac{L_*}{2}\|W' - W\|_{\mathrm{op}}^2.$$

Lemma C.3 (One-step descent)**.** *Let $G_t = \nabla F(W_t)$ and let*

$$W_{t+1} = W_t - \eta D_t$$

be the update at step t, where $\eta > 0$ is deterministic conditional on the past and D_t may be random. Then the following bounds hold.

1. *(Frobenius smoothness.) If F is L_F-smooth with respect to the Frobenius norm, then*

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \eta\, \mathbb{E}_t\langle G_t, D_t\rangle - \frac{L_F\eta^2}{2}\, \mathbb{E}_t\|D_t\|_F^2.$$

In particular, if there exist deterministic quantities $\alpha_t \geq 0$ and $\beta_t > 0$ such that

$$\mathbb{E}_t\langle G_t, D_t\rangle \geq \alpha_t, \qquad \mathbb{E}_t\|D_t\|_F^2 \leq \beta_t^2,$$

then for every deterministic step size

$$0 \leq \eta \leq \eta_{F,t}^{\max} := \frac{2\alpha_t}{L_F\beta_t^2},$$

we have

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \eta\alpha_t - \frac{L_F\eta^2}{2}\beta_t^2 \geq 0.$$

Moreover, the best deterministic step size within this certified range is

$$\eta_{F,t} = \frac{\alpha_t}{L_F\beta_t^2} = \frac{1}{2}\eta_{F,t}^{\max},$$

which gives

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \frac{\alpha_t^2}{2L_F\beta_t^2}.$$

2. *(Operator-norm smoothness.) If F is L_*-smooth with respect to the operator norm, then*

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \eta\, \mathbb{E}_t\langle G_t, D_t\rangle - \frac{L_*\eta^2}{2}\, \mathbb{E}_t\|D_t\|_{\mathrm{op}}^2.$$

In particular, if there exist deterministic quantities $\gamma_t \geq 0$ and $\nu_t > 0$ such that

$$\mathbb{E}_t\langle G_t, D_t\rangle \geq \gamma_t, \qquad \|D_t\|_{\mathrm{op}} \leq \nu_t \quad \textit{almost surely},$$

then for every deterministic step size

$$0 \leq \eta \leq \eta_{\mathrm{op},t}^{\max} := \frac{2\gamma_t}{L_*\nu_t^2},$$

we have

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \eta\gamma_t - \frac{L_*\eta^2}{2}\nu_t^2 \geq 0.$$

The best deterministic step size within this certified range is

$$\eta_{\mathrm{op},t} = \frac{\gamma_t}{L_*\nu_t^2} = \frac{1}{2}\eta_{\mathrm{op},t}^{\max},$$

yielding

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \frac{\gamma_t^2}{2L_*\nu_t^2}.$$

The descent factors Γ in Table 1 are obtained by substituting the relevant alignment bound into Lemma C.3 and optimizing the scalar step size η.

D Structural interpretation of activation capture

This appendix gives a deterministic interpretation of when the activation coverage ratio (ACR)

$$\rho_A(G) = \frac{\|GA\|_F^2}{\|G\|_F^2}$$

is large. The main text uses $\rho_A(G_t)$ as a measurable quantity in the descent rate; the discussion below explains how high ACR follows from effective-rank activation spectra together with moderate alignment between upstream signals and leading activation directions.

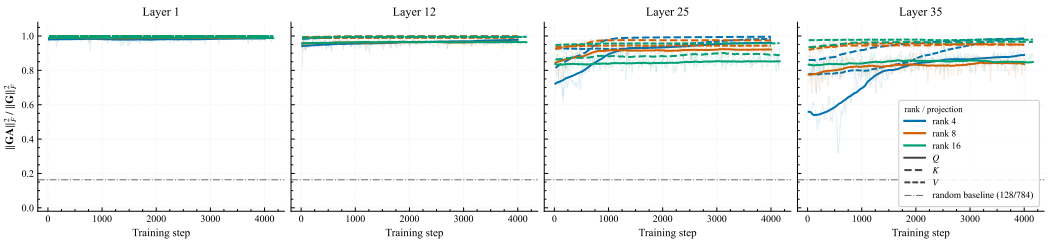


Figure 2: Activation coverage ratio by activation bases during OPT-13B fine-tuning on SST-2.

Heavy-tail activation spectra. A consistent finding in recent work is that hidden activations in trained networks often exhibit rapidly decaying, heavy-tailed singular value spectra. For transformers, Liu et al. (2025a) report power-law decay $s_i(H)^2 \propto i^{-\alpha}$ during training, with exponents that evolve during fine-tuning and correlate with generalization. Earlier observations of power-law activation spectra appear in CNN settings (Rotman & Wolf, 2021). On the theoretical side, Paquette et al. (2026) show that power-law eigenvalue decay in the input covariance can persist across nonlinear random-feature layers. Related observations on layerwise variation in low-rankness across attention and MLP components (Jaiswal et al., 2025) are consistent with the layerwise variation in activation capture observed in Figure 2. More broadly, intrinsic dimension results (Aghajanyan et al., 2021) and the success of LoRA (Hu et al., 2022) reflect the same empirical geometry: fine-tuning updates often concentrate in a low-dimensional subspace.

Assumption D.1 (Effective activation rank)**.** *There exist $r_\star \le n$ and $\epsilon \in [0, 1)$ such that, along the trajectory under consideration,*

$$\frac{\sum_{i \le r_\star} s_i(H_t)^2}{\sum_i s_i(H_t)^2} \ge 1 - \epsilon. \tag{15}$$

Assumption D.1 is implied by a summable power-law spectrum $s_i(H_t)^2 \asymp c\, i^{-\alpha}$ with $\alpha > 1$. In that case,

$$\epsilon = \frac{\sum_{i > r_\star} s_i(H_t)^2}{\sum_i s_i(H_t)^2} \le C(\alpha)\, r_\star^{1-\alpha}.$$

The assumption concerns the activation matrix H_t, not directly the product gradient $G_t = Q_t H_t^\top$; the latter also depends on how the upstream signal Q_t aligns with activation singular directions.

The next corollary makes this dependence explicit. It is intended as a structural decomposition rather than a uniform guarantee: it identifies which properties of (H, Q) govern activation capture.

Corollary D.2 (Activation capture under effective rank)**.** *Suppose $G = QH^\top \ne 0$, $1 \le r_\star \le \mathrm{rank}(H)$, and let $A = U_{r_\star}$ be the top-$r_\star$ left singular basis of $H = U\Sigma V^\top$. With $b_i := \|Qv_i\|_2^2$, define the spectrum-weighted upstream energies on the leading and trailing activation subspaces by*

$$\bar{b}_{\mathrm{top}} := \frac{\sum_{i \le r_\star} s_i^2 b_i}{\sum_{i \le r_\star} s_i^2}, \qquad \bar{b}_{\mathrm{tail}} := \frac{\sum_{i > r_\star} s_i^2 b_i}{\sum_{i > r_\star} s_i^2},$$

and set

$$\tau := \frac{\bar{b}_{\mathrm{tail}}}{\bar{b}_{\mathrm{top}}}.$$

Set $\bar{b}_{\mathrm{tail}} = 0$ when the activation tail is empty. Assume $\bar{b}_{\mathrm{top}} > 0$ and that Assumption D.1 holds with parameter ϵ. Then

$$\rho_A(G) \ge \frac{1 - \epsilon}{1 - \epsilon + \tau\epsilon} \ge 1 - \max\{1, \tau\}\epsilon. \tag{16}$$

Proof. Let $T := \sum_{i=1}^{s} s_i^2$ and $T_\star := \sum_{i=1}^{r_\star} s_i^2$. By Proposition 3.2,

$$\rho_A(G) = \frac{T_\star \bar{b}_{\mathrm{top}}}{T_\star \bar{b}_{\mathrm{top}} + (T - T_\star)\bar{b}_{\mathrm{tail}}} = \frac{T_\star}{T_\star + (T - T_\star)\tau}.$$

Using $T_\star \geq (1-\epsilon)T$ and $T - T_\star \leq \epsilon T$ gives

$$\rho_A(G) \geq \frac{(1-\epsilon)T}{(1-\epsilon)T + \tau\epsilon T} = \frac{1-\epsilon}{1-\epsilon+\tau\epsilon}.$$

Finally the loss of coverage is at most $\tau\epsilon/(1-\epsilon+\tau\epsilon)$. If $\tau \geq 1$, its denominator is at least one; if $0 \leq \tau \leq 1$, the fraction is at most ϵ. This proves the stated simplification. The sharper expression $(1-\epsilon)/(1-\epsilon+\tau\epsilon)$ is the quantity to use in comparisons. □

Corollary D.2 isolates two structural ingredients: the effective rank of the activation matrix, captured by ϵ, and the alignment of the upstream signal with the leading activation directions, captured by τ. If upstream energy is roughly uniform across activation singular directions, then $\tau \approx 1$ and $\rho_A(G) \gtrsim 1-\epsilon$. If upstream energy is itself concentrated on the leading activation directions, then $\tau < 1$ and capture is even larger. Thus, under heavy-tail activation spectra, high activation capture is expected whenever the upstream signal is not adversarially concentrated on the activation tail. This interpretation is consistent with the direct capture measurements reported in Figure 2.

E Proof of main theoretical results

In this section, we're going to use a slightly different notation from the main text. We will use M to denote the true coefficient gradient $G_t A_t$, and $\widehat{H}$ to denote the ZO estimator $\widehat{S}$.

E.1 Matrix-sign alignment

The central technical object of this section is the expected alignment of the orthogonalized ZO estimator with the true coefficient gradient. Throughout this subsection, $M \in \mathbb{R}^{m\times r}$ is a fixed nonzero matrix (we will take $M = G_t A_t$ at a single iterate), and the idealized estimator is

$$\widehat{H} := \frac{1}{q}\sum_{i=1}^{q} \langle M, R_i\rangle R_i, \qquad R_i \overset{\text{iid}}{\sim} N(0, I_{m\times r}). \tag{17}$$

The quantity to control is

$$\mathbb{E}\left\langle M, \operatorname{msign}(\widehat{H})\right\rangle.$$

Bounding it from below pins down a positive descent rate; bounding it from above shows that matrix-sign normalization cannot beat that rate.

Conditional structure. The expression simplifies through a single Gaussian decomposition. Write $\bar{M} := M/\|M\|_F$ and decompose

$$R_i = \alpha_i \bar{M} + \Xi_i, \qquad \alpha_i := \left\langle \bar{M}, R_i\right\rangle \sim N(0,1),$$

where the residual Ξ_i is independent of α_i and centered Gaussian on the orthogonal complement of $\bar{M}$ (in vectorized form). Substituting into the estimator and setting $S := \sum_{i=1}^{q} \alpha_i^2$,

$$\widehat{H} = \frac{\|M\|_F}{q}\left(S\bar{M} + \sum_{i=1}^{q}\alpha_i \Xi_i\right), \qquad S \sim \chi^2_q.$$

After conditioning on $S = s^2$ and using scale invariance of msign, one shows (Lemma I.1) that

$$\mathbb{E}\left\langle M, \operatorname{msign}(\widehat{H})\right\rangle = \|M\|_F\, \mathbb{E}_S[f(\sqrt{S})], \qquad f(s) := \mathbb{E}_\Xi\left\langle \bar{M}, \operatorname{msign}(s\bar{M} + \Xi_{\bar{M}})\right\rangle, \tag{18}$$

where $\Xi_{\bar{M}}$ is a single Gaussian draw with covariance $I - \bar{M}\bar{M}^\top$ in vectorized form.

The q-query problem thus decouples into two pieces. The first is a deterministic *polar-alignment curve* $f(s)$, which measures how well the matrix sign recovers the unit-Frobenius signal $\bar{M}$ from a single noisy observation $s\bar{M} + \Xi_{\bar{M}}$ at signal strength s. The second is a chi-square smoothing of s against $S \sim \chi^2_q$, which encodes the query-budget effect through $\mathbb{E}\sqrt{S} \asymp \sqrt{q}$.

This decomposition also explains where the non-triviality lies. By the symmetry $\Xi_{\bar{M}} \overset{d}{=} -\Xi_{\bar{M}}$ and the antisymmetry $\mathrm{msign}(-X) = -\mathrm{msign}(X)$, we have

$$f(0) = 0 :$$

at zero signal strength, the matrix sign of pure Gaussian noise has no expected alignment with any fixed $\bar{M}$. All of the alignment in (18) therefore comes from the nonlinear bias that the polar map produces once the noise distribution is shifted by $s\bar{M}$, and the rate at which $f(s)$ grows away from zero is what controls the alignment-to-query trade-off. The next two theorems lower-bound this growth — first in the rank-one case as a warmup, then in general — and translate it back into a bound on $\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle$ via the chi-square smoothing in (18). In the few-query regime, the resulting polar alignment scales like $\sqrt{q/(m+r)}$. Its Frobenius-normalized version includes a factor $1/\sqrt{r}$; comparisons with raw updates must retain this normalization.

Rank-one warmup. We begin with the rank-one case, which already exhibits the $\sqrt{q/(m+r)}$ scale and admits an exact ratio-of-chi-squares form.

Theorem E.1 (Rank-one matrix-sign alignment)**.** *Let $M = \sigma uv^\top$ with $\sigma > 0$, $\|u\|_2 = \|v\|_2 = 1$, and let $\widehat{H}$ be the estimator in (17). With independent $X \sim \chi^2_q$ and $Y \sim \chi^2_{m+r-2}$,*

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \geq \|M\|_F\, \mathbb{E}\sqrt{\frac{X}{X+Y}}.$$

Consequently there is a universal constant $c > 0$ such that

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \geq c\|M\|_F \min\left\{\sqrt{\frac{q}{m+r}},\, 1\right\}.$$

The full proof is in Appendix H. By rotational invariance, one reduces to $M = \sigma E_{11}$; a thin QR decomposition of $\widehat{H}$ followed by a reflection cancels all off-diagonal contributions, so the alignment is controlled by the diagonal entry w_{11} of the matrix sign of an upper-triangular Gaussian matrix, which admits an explicit inverse-square-root chi-square bound. This is an explicit lower bound, not an equality for general matrix shapes. It is informative for two reasons: even at $q = 1$ the alignment is strictly positive — the matrix-sign step extracts descent signal from a single query — and the bound interpolates smoothly between a few-query regime ($\sqrt{q/(m+r)}$) and a saturated-signal regime ($\Theta(1)$), which generalizes to the spectrum-sensitive setting below.

General spectrum-sensitive lower bound. For arbitrary M, the lower-bound terms associated with singular values of $\bar{M} := M/\|M\|_F$ have different transition scales: the term for σ_j approaches saturation when $\sigma_j^2 q/(m+r) \gtrsim 1$. The next theorem makes this dependence explicit through a smooth interpolating function; it does not assert individual singular-vector recovery.

Theorem E.2 (General matrix-sign alignment lower bound)**.** *Let $M \in \mathbb{R}^{m\times r}$ be nonzero, and let $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ be the singular values of $\bar{M} = M/\|M\|_F$, so $\sum_j \sigma_j^2 = 1$. Assume $m \geq 16r$. Then there exist universal constants $c, C > 0$ such that*

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \geq c\|M\|_F \sum_{j=1}^{r} \sigma_j\, \Phi_C\left(\sigma_j\sqrt{\frac{q}{m+r}}\right), \qquad \Phi_C(x) := \frac{x}{\sqrt{1+Cx^2}}.$$

The function Φ_C has the two limits

$$\Phi_C(x) \asymp x \ \text{ for } x \leq 1, \qquad \Phi_C(x) \asymp 1 \ \text{ for } x \gg 1.$$

Each singular direction contributes linearly in $\sigma_j\sqrt{q/(m+r)}$ until it saturates, after which it contributes its full nuclear-norm share σ_j. The bound is sensitive to spectral concentration: large modes reach their own ceilings sooner. At fixed Frobenius norm, a flatter spectrum has a larger nuclear-norm ceiling and a larger lower bound; earlier saturation does not mean larger total alignment.

The proof, given in Corollary I.4, combines three pieces. A conditional representation reduces the q-query estimator to a one-parameter shifted-Gaussian polar curve $f(s) = \mathbb{E}\left\langle \bar{M}, \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}})\right\rangle$

at $s \sim \sqrt{\chi^2_q}$. A tall shifted-Gaussian singular-value lemma (Lemma J.1) shows that the shifted model is well-conditioned with constant probability when $m \geq C_0 r$. In fact the proof of that lemma yields the stronger quantitative form

$$\mathbb{P}\big(\sigma_{\min}(X(s)) \geq a\sqrt{m} \text{ and } \|\Xi_{\bar{M}}\|_{\mathrm{op}} \leq b\sqrt{m}\big) \;\geq\; 1 - Ce^{-cm}$$

for universal constants $a, b, c, C > 0$, uniformly in $s \geq 0$ and in $\bar{M}$ with $\|\bar{M}\|_F = 1$. The probability is therefore exponentially close to 1 in the output dimension m and depends on no other problem parameters; A row-leverage bound applied to the SVD-coordinate form of the polar derivative converts these conditioning estimates into a lower bound on $f'(s)$, which integrates and is then chi-square-smoothed back to a bound in q.

Few-query rate and matching upper bound. The regime relevant for ZO methods is $q \ll m + r$, where the spectrum-sensitive bound collapses to a clean Frobenius rate. Corollary I.6 gives, for every $q \geq 1$,

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \geq \frac{c}{\sqrt{1 + Cq/(m+r)}}\,\|M\|_F\sqrt{\frac{q}{m+r}},$$

so the rate $\|M\|_F\sqrt{q/(m+r)}$ holds up to a universal constant for $q \lesssim m + r$ (Appendix I.2). In particular, $q = 1$ gives $\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \geq c\|M\|_F/\sqrt{m+r}$: a single coefficient-space query suffices for strictly positive expected alignment.

The next theorem shows this rate cannot be improved by matrix-sign normalization.

Theorem E.3 (Matching matrix-sign alignment upper bound)**.** *Under the assumptions of Theorem 4.1, there is a universal constant $C > 0$ such that for every nonzero M,*

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \leq \min\left\{C\|M\|_F\sqrt{\frac{q}{m+r}},\; \|M\|_*\right\}.$$

The first term in the minimum matches Corollary 4.2 up to a constant; the second is the trivial spectral–nuclear duality bound $\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \leq \|M\|_*\|\,\mathrm{msign}(\widehat{H})\|_{\mathrm{op}} \leq \|M\|_*$, which dominates once the alignment saturates. Together, Theorem 4.1 and Theorem E.3 pin down the alignment rate to $\Theta\Big(\|M\|_F\sqrt{q/(m+r)}\Big)$ in the few-query regime. The proof of the upper bound, in Appendix I.3, integrates a Cauchy–Schwarz estimate on the polar derivative against an inverse-singular-value moment bound for the shifted-Gaussian model.

Regimes and tightness. The smooth interpolating function Φ_C admits the elementary two-sided estimate $\Phi_C(x) \asymp \min\{x, 1\}$, so Theorem 4.1 can be restated in hard-saturation form: specifically, using Corollary I.5 we get

$$\mathbb{E}\left\langle M, \mathrm{msign}(\widehat{H})\right\rangle \;\geq\; c\|M\|_F\sum_{j=1}^{r}\sigma_j\min\left\{\sigma_j\sqrt{\frac{q}{m+r}},\, 1\right\}. \tag{19}$$

Each mode's term in this lower bound therefore has two regimes, with a transition at the scale $\sigma_j^2\, q/(m+r) \approx 1$:

- **Few-query regime** ($\sigma_j^2\, q/(m+r) \leq 1$). The j-th mode contributes $\sigma_j^2\sqrt{q/(m+r)}$, linear in $\sqrt{q}$. Summing over j and using $\sum_j \sigma_j^2 = 1$ gives Corollary 4.2: under $q \leq c_0(m+r)$ every mode is in this regime simultaneously, and the total alignment is $\gtrsim \|M\|_F\sqrt{q/(m+r)}$. At $q = 1$ this yields a strictly positive alignment of order $\|M\|_F/\sqrt{m+r}$, so a single coefficient-space query already delivers expected descent.
- **Saturated regime** ($\sigma_j^2\, q/(m+r) \gg 1$). The j-th term has order σ_j and no longer grows in the hard-min proxy. When all nonzero modes are saturated, the bound has nuclear-norm scale $\|M\|_*$, matching the spectral–nuclear duality ceiling of Theorem E.3.

Different modes can be in different regimes in the lower bound. The $\sqrt{q/(m+r)}$ rate holds for every spectrum while $q \lesssim (m+r)/\sigma_1^2$; a flatter spectrum extends this interval. For a flat rank-k

spectrum it extends to $q \lesssim k(m+r)$. Combined with the matching upper bound, this pins the alignment rate to $\Theta(\|M\|_F\sqrt{q/(m+r)})$ in the small-query regime and to $\Theta(\|M\|_*)$ once every mode has saturated, with the spectrum-sensitive interpolation between them captured by (19). The lower bound thus depends on the full singular spectrum: after the leading term approaches saturation, weaker terms can continue increasing. This describes total-alignment guarantees, not exact recovery thresholds. The exact expected alignment approaches $\|M\|_*$ as $q \to \infty$ (Proposition F.2).

The contrast with raw AGZO is also instructive. The raw estimator is unbiased, so $\mathbb{E}\langle M, \widehat{H}\rangle = \|M\|_F^2$, but its update norm is large when $q \ll mr$. Therefore the relevant descent-normalized quantity is not the raw inner product itself, but

$$\frac{\mathbb{E}\langle M, \widehat{H}\rangle}{\sqrt{\mathbb{E}\|\widehat{H}\|_F^2}}.$$

For $\|M\|_F = 1$, this quantity approaches 1 as $q \to \infty$. By contrast, the matrix-sign update can align up to the nuclear-norm scale $\|M\|_* = \sum_j \sigma_j$, which is larger than one whenever the target has more than one significant singular direction. These ceilings correspond to different unit balls, so their difference alone does not prove an optimization advantage. The common-geometry comparison is given in Appendix F.1.

Connection to noisy first-order Muon. Although our estimator $\widehat{H}$ comes from zeroth-order queries, the alignment problem above can also be viewed as a stylized noisy-gradient problem. Indeed, the conditional representation

$$\widehat{H} \mid S = s^2 \overset{d}{=} \frac{\|M\|_F s}{q}\left(s\bar{M} + \Xi_{\bar{M}}\right)$$

shows that, up to a positive scale factor, the matrix-sign update is applied to a noisy observation of the normalized coefficient gradient $\bar{M}$. Thus our analysis studies when

$$\mathrm{msign}(s\bar{M} + \Xi_{\bar{M}})$$

remains aligned with $\bar{M}$. This is closely analogous to stochastic first-order Muon, where one applies msign to a minibatch gradient $G + \mathcal{E}$ rather than to the exact gradient G. In that setting, the signal-to-noise ratio of the minibatch gradient plays the role of s, while here the query budget enters through $s \sim \sqrt{\chi_q^2}$. The lower bound therefore gives a clean model for the noise robustness of matrix-sign normalization: terms associated with larger singular values approach saturation first, and the total-alignment guarantee follows a smooth mode-wise saturation law.

The analogy should not be read as a theorem for arbitrary stochastic gradients. Minibatch noise in first-order training is data-dependent and generally anisotropic, whereas the present proof uses the special Gaussian structure of the ZO estimator. The role of this connection is instead interpretive: it explains why the same polar normalization used in first-order Muon can be useful in stochastic settings.

E.2 Baseline one-step certificates and proofs

Interpreting the main-text comparison. Theorem 4.3 and Table 1 collect the few-query progress factors. Here we give the detailed crossover comparisons before proving the estimator-moment bounds that yield them.

The theorem gives local lower-bound certificates for a fixed objective, using step sizes that may depend on the unknown gradient. Comparing two such lower bounds does not order the actual decreases of two methods. Against the Frobenius certificate for classical AGZO, the activation capture cancels, and the ratio of the displayed factors favors AGOZO when

$$\frac{1}{L_*}\frac{q}{m+r} \gtrsim \frac{1}{L_F}\frac{q}{mr}, \qquad \text{i.e.} \qquad \frac{L_F}{L_*} \gtrsim \frac{m+r}{mr}.$$

This comparison changes both the update and the smoothness geometry. When both methods are analyzed under the same spectral smoothness constant, raw AGZO has the same order $q/(m+r)$

Table 3: Certified one-step descent factors Γ, with $\mathbb{E}_t[F(W_t) - F(W_{t+1})] \gtrsim \Gamma\|G\|_F^2$, in the few-query regime and up to universal constants. "Estimated matrix" is the matrix whose ZO estimate drives the update. The shape and oracle-step assumptions are those of Theorem 4.3. The last column compares the displayed scaling proxies, with hidden constants set equal. Random coverage is replaced by its fresh, independent Haar expectation. Here $\kappa_{\mathrm{LOZO}} = mn+1+2(m+n+1)/r_L$, $T_{\mathrm{full}} = \frac{m+r}{m+n}$, $T_{\mathrm{right}} = \frac{(m+r)r_P}{n(m+r_P)}$ and $T_{\mathrm{left}} = \frac{(m+r)r_P}{m(n+r_P)}$.

Method	**Estimated matrix**	**Geometry**	**Descent factor Γ**	**Proxy crossover**
MeZO	$G \in \mathbb{R}^{m\times n}$, raw	Frobenius	$\frac{1}{L_F}\frac{q}{mn}$	$\rho_A \gtrsim \frac{L_*}{L_F}\frac{m+r}{mn}$
LOZO, fresh, rank r_L	G, rank-r_L probes, raw	Frobenius	$\frac{1}{L_F}\frac{q}{\kappa_{\mathrm{LOZO}}}$	$\rho_A \gtrsim \frac{L_*}{L_F}\frac{m+r}{\kappa_{\mathrm{LOZO}}}$
AGZO	$GA \in \mathbb{R}^{m\times r}$, raw	Frobenius	$\frac{1}{L_F}\rho_A\frac{q}{mr}$	$\frac{L_F}{L_*} \gtrsim \frac{m+r}{mr}$
AGZO	$GA \in \mathbb{R}^{m\times r}$, raw	Spectral	$\frac{1}{L_*}\rho_A\frac{q}{m+r}$	Same order
ZO-Muon, full space	$G \in \mathbb{R}^{m\times n}$, msign	Spectral	$\frac{1}{L_*}\frac{q}{m+n}$	$\rho_A \gtrsim T_{\mathrm{full}}$
ZO-Muon, right proj.	$GP \in \mathbb{R}^{m\times r_P}$, msign	Spectral	$\frac{1}{L_*}\rho_P\frac{q}{m+r_P}$	$\rho_A \gtrsim T_{\mathrm{right}}$
ZO-Muon, left proj.	$P^\top G \in \mathbb{R}^{r_P\times n}$, msign	Spectral	$\frac{1}{L_*}\rho_P^{\mathrm{L}}\frac{q}{n+r_P}$	$\rho_A \gtrsim T_{\mathrm{left}}$
AGOZO	$GA \in \mathbb{R}^{m\times r}$, msign	Spectral	$\frac{1}{L_*}\rho_A\frac{q}{m+r}$	–

certificate. Under Frobenius smoothness, AGOZO pays the rank cost r. Neither comparison proves a few-query rank-factor advantage for orthogonalization (Appendix F.1).

Against full-space ZO-Muon, when both estimated matrices satisfy the aspect assumption and a common spectral smoothness bound is used, the displayed factors compare as

$$\rho_A(G)\frac{q}{m+r} \gtrsim \frac{q}{m+n}, \qquad \text{i.e.} \qquad \rho_A(G) \gtrsim \frac{m+r}{m+n}.$$

For nearly square full-space matrices the rate $q/(m+n)$ is an extrapolation, not a consequence of this theorem. The numerical thresholds below also suppress different universal constants. Against random-subspace ZO-Muon with a right or a left projection,

$$\frac{\rho_A(G)}{m+r} \gtrsim \frac{\rho_P(G)}{m+r_P} \quad \text{(right)}, \qquad \frac{\rho_A(G)}{m+r} \gtrsim \frac{\rho_P^{\mathrm{L}}(G)}{n+r_P} \quad \text{(left)},$$

where $\mathbb{E}\rho_P(G) = r_P/n$ and $\mathbb{E}\rho_P^{\mathrm{L}}(G) = r_P/m$ for a fresh Haar-random basis independent of G. A basis reused across updates is generally dependent on the current gradient, so this expectation is then only a reference value. Appendix N reports these dimension-based diagnostics on the measured ρ_A of our runs.

Each certificate follows from the expected descent inequality once alignment and the second moment of the update norm are known. We first give the raw Frobenius and polar spectral certificates. Proposition F.1 adds the raw spectral certificate in Table 1, and Appendix F.1 compares the updates in a common norm. We condition on the current iterate, so $G = \nabla F(W)$ is fixed and work in the $\mu \to 0$ finite-difference limit. Table 3 expands the main-text Table 1 with explicit crossover diagnostics; the derivations for MeZO, AGZO, polar updates, and fresh-sampled LOZO follow.

Raw Gaussian estimators: MeZO and AGZO. For a fixed $M \in \mathbb{R}^{a\times b}$, let $\widehat{M} = q^{-1}\sum_{i=1}^q \langle M, R_i\rangle R_i$ with $R_i \overset{\text{iid}}{\sim} N(0, I_{a\times b})$, and let $D = ab$. Then $\mathbb{E}\widehat{M} = M$, and the Gaussian identity $\mathbb{E}[(g^\top z)^2\|z\|_2^2] = (D+2)\|g\|_2^2$ for $z \sim N(0, I_D)$ gives the second moment of a single query. Writing $g = \|g\|_2 e_1$, it equals $\|g\|_2^2\sum_{i=1}^D \mathbb{E}[z_1^2 z_i^2]$, where each coordinate $i \neq 1$ contributes 1 and the coordinate along g contributes the fourth moment $\mathbb{E}z_1^4 = 3$, i.e. 2 more. Subtracting the

squared mean leaves a per-query variance of $(D+1)\|M\|_F^2$, and averaging q independent queries gives

$$\mathbb{E}\|\widehat{M}\|_F^2 = \Big(1 + \frac{D+1}{q}\Big)\|M\|_F^2. \tag{20}$$

The Frobenius part of Lemma C.3 with $\alpha = \|M\|_F^2$ and β^2 given by (20) yields

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \geq \frac{1}{2L_F}\,\frac{\|M\|_F^2}{1+(D+1)/q} \asymp \frac{1}{L_F}\,\frac{q}{D}\|M\|_F^2 \qquad (q \ll D). \tag{21}$$

MeZO is the case $M = G$, $D = mn$. AGZO perturbs along $R_i A^\top$ and updates along $O_{\mathrm{AGZO}} = \widehat{S}_A A^\top$, where $\widehat{S}_A$ is the estimator of $M = GA \in \mathbb{R}^{m\times r}$. Since $A^\top A = I_r$, we have $\langle G, R_i A^\top\rangle = \langle GA, R_i\rangle$, $\langle G, O_{\mathrm{AGZO}}\rangle = \langle GA, \widehat{S}_A\rangle$ and $\|O_{\mathrm{AGZO}}\|_F = \|\widehat{S}_A\|_F$, so (21) applies with $D = mr$ and $\|M\|_F^2 = \rho_A(G)\|G\|_F^2$.

Matrix-sign estimators: AGOZO and ZO-Muon. The four polar methods share one structure. Let $L \in \mathbb{R}^{m\times a}$ and $R \in \mathbb{R}^{n\times b}$ have orthonormal columns, set $S := L^\top G R \in \mathbb{R}^{a\times b}$, perturb along $LR_iR^\top$, and update along $O := L\,\mathrm{msign}(\widehat{S})\,R^\top$, where $\widehat{S}$ is the Gaussian estimator of S. Then

$$\|O\|_{\mathrm{op}} \leq 1, \qquad \langle G, O\rangle = \langle S, \mathrm{msign}(\widehat{S})\rangle.$$

If $a \geq 16b$ and $q \lesssim a+b$, Corollary 4.2 gives $\mathbb{E}\langle S, \mathrm{msign}(\widehat{S})\rangle \gtrsim \|S\|_F\sqrt{q/(a+b)}$ (both conditions are quantified for our layers in Appendix I.2), and the operator-norm part of Lemma C.3 with $\nu = 1$ yields

$$\mathbb{E}_t[F(W_t) - F(W_{t+1})] \gtrsim \frac{1}{L_*}\,\frac{\|S\|_F^2}{\|G\|_F^2}\,\frac{q}{a+b}\,\|G\|_F^2. \tag{22}$$

If instead $b \geq 16a$, the same bound holds by transposition: $\widehat{S}^\top$ is the Gaussian estimator of $S^\top$ and $\mathrm{msign}(X^\top) = \mathrm{msign}(X)^\top$, so $\langle S, \mathrm{msign}(\widehat{S})\rangle = \langle S^\top, \mathrm{msign}(\widehat{S}^\top)\rangle$, and (22) depends on (a,b) only through $a+b$. The four methods are the following choices:

- *AGOZO*: $L = I_m$, $R = A$, so $S = GA$, $(a,b) = (m,r)$ and $\|S\|_F^2/\|G\|_F^2 = \rho_A(G)$;
- *full-space ZO-Muon*: $L = I_m$, $R = I_n$, so $S = G$ and $(a,b) = (m,n)$;
- *right-projection ZO-Muon*: $L = I_m$, $R = P \in \mathbb{R}^{n\times r_P}$, so $(a,b) = (m, r_P)$ and the ratio is $\rho_P(G) = \|GP\|_F^2/\|G\|_F^2$;
- *left-projection ZO-Muon*, the variant of the original ZO-Muon (Lang et al., 2026): $L = P \in \mathbb{R}^{m\times r_P}$, $R = I_n$, so $S = P^\top G$, $(a,b) = (r_P, n)$ and the ratio is $\rho_P^{\mathrm{L}}(G) = \|P^\top G\|_F^2/\|G\|_F^2$.

In our layers the shape ratio is at least 16 for AGOZO (m/r) and for both random-subspace variants (m/r_P and n/r_P, the left projection using the transposed form). For full-space ZO-Muon the condition reads $\max(m,n) \geq 16\min(m,n)$. The attention and MLP full-space blocks in the reported measurements have aspect ratios at most four, and all fall outside this sufficient condition. For these shapes, the factor $q/(m+n)$ is an extrapolation. Proposition F.2 proves positive alignment for them, but not this quantitative rate.

Random bases. For a Haar-random basis independent of G, $\mathbb{E}[PP^\top] = (r_P/n)I_n$ for $P \in \mathbb{R}^{n\times r_P}$ and $(r_P/m)I_m$ for $P \in \mathbb{R}^{m\times r_P}$, so $\mathbb{E}_P\rho_P(G) = r_P/n$ and $\mathbb{E}_P\rho_P^{\mathrm{L}}(G) = r_P/m$. The two projections therefore give different factors in expectation,

$$\frac{\mathbb{E}\Gamma_{\mathrm{left}}}{\mathbb{E}\Gamma_{\mathrm{right}}} = \frac{r_P/(m(n+r_P))}{r_P/(n(m+r_P))} = \frac{n(m+r_P)}{m(n+r_P)},$$

which equals one only for square blocks; the left projection is favored when $n > m$ and the right projection when $m > n$.

Fresh-sampled LOZO. A rank-r_L LOZO probe is $\Delta = UV^\top$ with $U \in \mathbb{R}^{m\times r_L}$ and $V \in \mathbb{R}^{n\times r_L}$ having i.i.d. $N(0,1)$ entries, and the one-sample estimator is $Z = r_L^{-1}\langle G, \Delta\rangle\Delta$. Since $\mathbb{E}[\Delta_{ij}\Delta_{kl}] = r_L\,\delta_{ik}\delta_{jl}$, $\mathbb{E}Z = G$. The map $G \mapsto \mathbb{E}[\langle G, \Delta\rangle^2\|\Delta\|_F^2]$ is a quadratic form invariant

under $G \mapsto O_1 G O_2^\top$ for orthogonal O_1, O_2, hence a multiple of $\|G\|_F^2$, and we may take $G = e_1 e_1^\top$. With $X = UV^\top$,

$$\mathbb{E}X_{11}^4 = 3r_L^2 + 6r_L, \qquad \mathbb{E}X_{11}^2 X_{1j}^2 = \mathbb{E}X_{11}^2 X_{i1}^2 = r_L(r_L+2),$$
$$\mathbb{E}X_{11}^2 X_{ij}^2 = r_L^2 \qquad \text{for all } i,j \neq 1,$$

and summing over the mn entries gives $\mathbb{E}X_{11}^2\|X\|_F^2 = r_L^2(mn+2) + 2r_L(m+n+1)$. Hence $\mathbb{E}\|Z\|_F^2 = \big(mn + 2 + 2(m+n+1)/r_L\big)\|G\|_F^2$, and averaging q independent samples gives

$$\mathbb{E}\|\widehat{G}_{\mathrm{LOZO}}\|_F^2 = \Big(1 + \frac{\kappa_{\mathrm{LOZO}}}{q}\Big)\|G\|_F^2, \qquad \kappa_{\mathrm{LOZO}} := mn + 1 + \frac{2(m+n+1)}{r_L}.$$

The Frobenius part of Lemma C.3 then gives $\Gamma_{\mathrm{LOZO}} \asymp q/(L_F \kappa_{\mathrm{LOZO}})$ for $q \ll \kappa_{\mathrm{LOZO}}$. This models LOZO with a fresh right factor at every step; its lazy sampling, which keeps V fixed over several steps, is a multi-step effect outside this one-step certificate.

Global and restricted smoothness constants. For the smallest positive global constants in Appendix C, norm equivalence gives

$$L_F \leq L_* \leq \min(m,n) L_F.$$

The sharper rank-r relation requires different, restricted constants. If $L_{F,A}$ and $L_{*,A}$ are the smallest positive constants on the same displacement space $\{\Delta = XA^\top\}$, then

$$L_{F,A} \leq L_{*,A} \leq r L_{F,A}.$$

This follows from $\|\Delta\|_{\mathrm{op}}^2 \leq \|\Delta\|_F^2 \leq r\|\Delta\|_{\mathrm{op}}^2$. It does not bound arbitrary chosen smoothness constants, nor does it allow a restricted $L_{*,A}$ to cancel against a full-space constant. For the isotropic quadratic $F(W) = \|W\|_F^2/2$, $L_{F,A} = 1$ and $L_{*,A} = r$: curvature spread across all coefficient directions gives the largest, not the smallest, ratio $L_{*,A}/L_{F,A}$. Most importantly, raw AGZO can use the same restricted spectral bound as AGOZO, by Proposition F.1.

Crossover conditions. The last column of Table 3 compares dimension-based proxies for the stated certificates. It suppresses method-dependent universal constants and is not a sufficient condition for actual dominance. Against the spectral baselines, L_* cancels and the conditions depend only on ρ_A, the shape of the block and the ranks. Against the Frobenius baselines, the ratio L_*/L_F remains; for MeZO and LOZO we use the exact certified factors $q/(2L_F(q+mn+1))$ and $q/(2L_F(q+\kappa_{\mathrm{LOZO}}))$, which reduce to the table entries for $q \ll mn$. For $r_P = r$, $T_{\mathrm{right}} = r/n$. The comparisons apply to isolated blocks with the stated shape conditions. Shared-query signal fractions can differ across methods and introduce an additional dependence described in Appendix F.3.

F Further consequences and implementation scope

F.1 Comparing raw and polar updates in the same geometry

The different scales of raw Frobenius-normalized alignment and unnormalized polar alignment do not by themselves imply different optimization rates. The following bound makes the spectral comparison explicit.

Proposition F.1 (Raw Gaussian estimation under spectral smoothness)**.** *For any nonzero $S \in \mathbb{R}^{m\times r}$, any $q \geq 1$, and the estimator in (7), there is a universal $C_{\mathrm{raw}} > 0$ such that*

$$\mathbb{E}\|\widehat{S}\|_{\mathrm{op}}^2 \leq C_{\mathrm{raw}} \left(\|S\|_{\mathrm{op}}^2 + \frac{m+r}{q}\|S\|_F^2 \right). \tag{23}$$

Consequently, under a spectral smoothness bound with constant L_*, an oracle deterministic scalar step along $\widehat{S}A^\top$ certifies decrease at least

$$\frac{\|S\|_F^2}{2C_{\mathrm{raw}}L_*} \frac{q}{m+r+q/\operatorname{sr}(S)}. \tag{24}$$

This proposition has no aspect-ratio assumption.

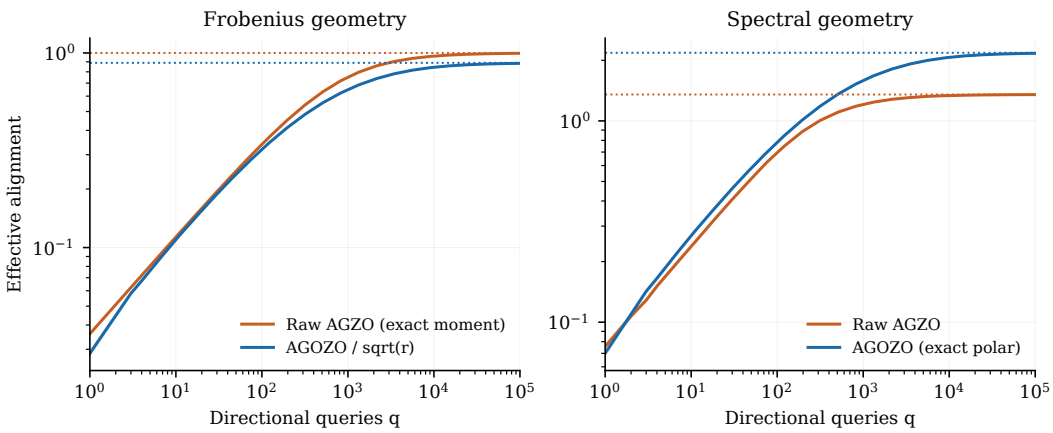


Figure 3: Effective alignment under matched norm constraints within each panel. The gradient, dimensions, and Gaussian samples are the same as in Figure 1, left. In Frobenius geometry the polar update must be divided by $\sqrt{r}$; in spectral geometry it already has unit norm. The red Frobenius curve uses the exact raw second moment; other curves and two-standard-error bands use 4096 samples per budget. This bookkeeping explains why a raw Frobenius curve cannot be compared directly with an unscaled polar curve as a descent advantage.

Proof. The conditional representation, before dropping its scale, is

$$\widehat{S} \overset{d}{=} \frac{T}{q} S + \frac{\|S\|_F \sqrt{T}}{q} \Xi_{\bar{S}}, \qquad T \sim \chi^2_q, \quad T \perp \Xi_{\bar{S}}.$$

Write $\Xi_{\bar{S}} = Z - \langle Z, \bar{S} \rangle \bar{S}$. Gaussian operator-norm bounds give $\mathbb{E}\|Z\|^2_{\mathrm{op}} \le C(m+r)$; this follows, for example, by integrating $\mathbb{P}(\|Z\|_{\mathrm{op}} > \sqrt{m} + \sqrt{r} + t) \le e^{-t^2/2}$. Since $\mathbb{E}\langle Z, \bar{S} \rangle^2 = 1$ and $\|\bar{S}\|_{\mathrm{op}} \le 1$, also $\mathbb{E}\|\Xi_{\bar{S}}\|^2_{\mathrm{op}} \le C(m+r)$. Hence

$$\mathbb{E}\|\widehat{S}\|^2_{\mathrm{op}} \le 2(1 + 2/q)\|S\|^2_{\mathrm{op}} + \frac{2\|S\|^2_F}{q} \mathbb{E}\|\Xi_{\bar{S}}\|^2_{\mathrm{op}},$$

which implies (23). Unbiasedness gives $\mathbb{E}\langle S, \widehat{S} \rangle = \|S\|^2_F$. The expected smoothness inequality uses the second moment of the update norm; an almost-sure norm bound is not needed. Optimize its quadratic in the deterministic step size to obtain (24). □

Thus, in the few-query regime, both raw AGZO and AGOZO have a spectral certificate of order $\|S\|^2_F q/(L_*(m+r))$. Under Frobenius smoothness, $\|\operatorname{msign}(\widehat{S})\|^2_F = r$ almost surely in the tall model, and the polar certificate is instead of order $\|S\|^2_F q/(L_F r(m+r))$, comparable to the raw certificate $\|S\|^2_F q/(L_F mr)$. These are comparisons of certificates, not an ordering of actual loss decreases.

There is a distinct, spectrum-dependent difference in the oracle limit. If S is known exactly, optimizing the same spectral-smoothness quadratic along S or $\operatorname{msign}(S)$ yields, respectively,

$$\mathcal{C}_{\mathrm{raw}} = \frac{\|S\|^4_F}{2L_*\|S\|^2_{\mathrm{op}}}, \qquad \mathcal{C}_{\mathrm{polar}} = \frac{\|S\|^2_*}{2L_*}.$$

Writing $r_*(S) := \|S\|^2_*/\|S\|^2_F$, their ratio is exactly

$$\frac{\mathcal{C}_{\mathrm{polar}}}{\mathcal{C}_{\mathrm{raw}}} = \frac{r_*(S)}{\mathrm{sr}(S)} = \left(\frac{\|S\|_* \|S\|_{\mathrm{op}}}{\|S\|^2_F} \right)^2 \ge 1. \tag{25}$$

The inequality follows from $\sum_j s_j^2 \le s_1 \sum_j s_j$. Equality holds exactly when the nonzero singular values are equal. This distinguishes the high-query spectral benefit from the few-query dimension scaling; it does not establish dominance at every finite query budget.

F.2 Positive, monotone alignment without an aspect restriction

Proposition F.2 (Monotonicity and exact oracle limit). *For any nonzero $S \in \mathbb{R}^{m\times r}$, define $\mathcal{A}_q(S) := \mathbb{E}\langle S, \mathrm{msign}(\widehat{S}_q)\rangle$ using q independent Gaussian directional queries. For all matrix shapes,*

$$0 < \mathcal{A}_q(S) \le \mathcal{A}_{q+1}(S) \le \|S\|_*, \qquad \lim_{q\to\infty} \mathcal{A}_q(S) = \|S\|_*.$$

Proof. For fixed $\Xi = \Xi_{\bar{S}}$, set $h_\Xi(s) = \|s\bar{S} + \Xi\|_*$ and $g_\Xi(s) = \langle \bar{S}, \mathrm{msign}(s\bar{S} + \Xi)\rangle$. The partial polar factor is a subgradient of the nuclear norm, including at rank-deficient matrices. Thus $g_\Xi(s) \in \partial h_\Xi(s)$, and monotonicity of scalar convex subgradients makes g_Ξ nondecreasing. Symmetry of Ξ and oddness of the partial polar map make $f(s) = \mathbb{E}g_\Xi(s)$ odd. In particular $f(s) \ge 0$ for $s \ge 0$.

Strict positivity does not require differentiability. Pair the realizations Ξ and $-\Xi$. Their sum at $s > 0$ is $g_\Xi(s) - g_\Xi(-s) \ge 0$. Optimality of the polar factor in the operator-norm unit ball gives

$$g_\Xi(s) \ge \|\bar{S}\|_* - \frac{2\|\Xi\|_*}{s}.$$

On the event $\|\Xi\|_* < s\|\bar{S}\|_*/4$, the same lower bound holds for both members of the pair and is strictly positive. This event has positive probability for the Gaussian law on $\bar{S}^\perp$ (and probability one when that space is zero-dimensional). Hence $f(s) > 0$.

Couple $T_q = \sum_{i=1}^q z_i^2$, $z_i \overset{\text{iid}}{\sim} N(0,1)$. Then $T_{q+1} \ge T_q > 0$ almost surely, and the conditional representation $\mathcal{A}_q(S) = \|S\|_F \mathbb{E} f(\sqrt{T_q})$ proves positivity and monotonicity. Finally, for any matrix X, polar optimality and norm duality imply

$$0 \le \|S\|_* - \langle S, \mathrm{msign}(X)\rangle \le 2\|X - S\|_*.$$

The strong law gives $\widehat{S}_q \to S$ almost surely. Since the alignment is bounded in absolute value by $\|S\|_*$, dominated convergence proves the limit. Continuity of the polar factor at S is not needed. □

This proposition guarantees a descent signal even for square full-space matrices. It supplies no $q/(m + r)$ quantitative rate for those shapes. It also implies diminishing need for a recovery interpretation: the monotone quantity is total expected alignment, not each noisy singular vector.

F.3 Shared queries: signal energy from other blocks is noise

Unbiasedness of a block estimate does not make its polar direction insensitive to the other perturbed blocks. The following extension quantifies that effect.

Theorem F.3 (Gaussian nuisance signal and shared-query alignment). *Let $S \ne 0$, $R_i \overset{\text{iid}}{\sim} N(0, I_{m\times r})$, and*

$$y_i = \langle S, R_i\rangle + \nu z_i, \qquad z_i \overset{\text{iid}}{\sim} N(0,1), \qquad z_i \perp (R_j)_j, \qquad \nu \ge 0.$$

Define $\widehat{S} = q^{-1}\sum_i y_i R_i$, $v^2 = \|S\|_F^2 + \nu^2$, and $\theta = \|S\|_F^2/v^2 \in (0,1]$. Then

$$\widehat{S} \overset{d}{=} \frac{v\sqrt{T}}{q}\left[\left(\sqrt{\theta T} + \sqrt{1-\theta}\, z\right)\bar{S} + \Xi_{\bar{S}}\right], \tag{26}$$

where $T \sim \chi^2_q$, $z \sim N(0,1)$, and $\Xi_{\bar{S}}$ are independent. Moreover,

$$\mathbb{E}\widehat{S} = S, \qquad \mathbb{E}\|\widehat{S}\|_F^2 = \left(1 + \frac{1}{q}\right)\|S\|_F^2 + \frac{mr}{q}v^2. \tag{27}$$

If $m \ge 16r$, the constants below are universal and

$$\mathbb{E}\langle S, \mathrm{msign}(\widehat{S})\rangle \ge c\|S\|_F \sum_j \sigma_j \Phi_C\left(\sigma_j\sqrt{\frac{\theta q}{m+r}}\right), \tag{28}$$

$$\mathbb{E}\langle S, \mathrm{msign}(\widehat{S})\rangle \le \min\left\{C\|S\|_F\sqrt{\frac{\theta q}{m+r}},\ \|S\|_*\right\}. \tag{29}$$

In particular, for $\theta q \lesssim (m+r)\,\mathrm{sr}(S)$, the alignment is $\Theta(\|S\|_F\sqrt{\theta q/(m+r)})$.

Proof. Apply the Gaussian decomposition of Lemma I.1 to the augmented query $(\mathrm{vec}(R_i), z_i)$, whose unit signal direction is $(\mathrm{vec}(S), \nu)/v$. The covariance of the matrix part of the residual is $I - \theta\,\mathrm{vec}(\bar{S})\,\mathrm{vec}(\bar{S})^\top$. It can therefore be represented as $\Xi_{\bar{S}} + \sqrt{1-\theta}\,z\bar{S}$, giving (26). Wick's identity gives $\mathbb{E}[y_i^2\|R_i\|_F^2] = mr\,v^2 + 2\|S\|_F^2$; subtracting the squared mean and averaging q independent observations proves (27).

Let $f(s) = \mathbb{E}\langle \bar{S}, \mathrm{msign}(s\bar{S} + \Xi_{\bar{S}})\rangle$ and

$$g(a) := \mathbb{E}_z f(a + \tau z), \qquad \tau = \sqrt{1-\theta} \le 1.$$

The tall-model derivative proof and symmetry give, for almost every real u,

$$\frac{c}{\sqrt{m}} \sum_j \frac{\sigma_j^2}{(1 + C(r + u^2\sigma_j^2)/m)^{3/2}} \le f'(u) \le \frac{C}{\sqrt{m+r}}.$$

The integrable derivative bound justifies differentiation of the convolution, and $g(0) = 0$ by symmetry. For the lower bound, restrict to $|z| \le 1$, which has fixed positive probability. For $a \ge 0$ on this event, $(a + \tau z)^2 \le 2a^2 + 2$. Since $m \ge 16r$ and $\sigma_j \le 1$, increasing universal constants gives

$$g'(a) \ge \frac{c}{\sqrt{m}} \sum_j \frac{\sigma_j^2}{(1 + Ca^2\sigma_j^2/m)^{3/2}}.$$

Integrate from zero to a, then average at $a = \sqrt{\theta T}$. Lemma I.3 gives (28). The upper derivative bound yields $g(a) \le Ca/\sqrt{m+r}$; use $\mathbb{E}\sqrt{T} \le \sqrt{q}$ and nuclear–spectral duality for (29). Finally use $\sigma_j^2 \le 1/\mathrm{sr}(S)$ for the few-query statement. □

For the joint perturbations in Algorithm 3, fix all current bases and write $S_\ell = G_\ell A_\ell$. Then

$$V^2 = \sum_k \|S_k\|_F^2, \qquad \nu_\ell^2 = \sum_{k \ne \ell} \|S_k\|_F^2, \qquad \theta_\ell = \frac{\|S_\ell\|_F^2}{V^2}.$$

Any additional independently perturbed Gaussian parameter groups also contribute their signal energy to V^2. The effective query count for block ℓ is $\theta_\ell q$. In the few-query regime its polar alignment is therefore of order

$$\frac{\|S_\ell\|_F^2}{V}\sqrt{\frac{q}{m_\ell + r_\ell}},$$

rather than the isolated-block value $\|S_\ell\|_F\sqrt{q/(m_\ell + r_\ell)}$. For equally energetic blocks this is a factor $1/\sqrt{L}$ per block. Thus layerwise ACR alone does not determine alignment in a joint query.

A joint one-step certificate is possible with a corresponding joint curvature assumption. Suppose, for all relevant block displacements,

$$F(W + \Delta) \le F(W) + \sum_\ell \langle G_\ell, \Delta_\ell\rangle + \frac{1}{2}\sum_\ell L_\ell \|\Delta_\ell\|_{\mathrm{op}}^2.$$

Let a_ℓ be a valid lower bound from (28). No independence of the block updates is needed: a common step size $\eta = (\sum_\ell a_\ell)/(\sum_\ell L_\ell)$ gives decrease at least

$$\frac{(\sum_\ell a_\ell)^2}{2\sum_\ell L_\ell}.$$

Allowing separate oracle steps $\eta_\ell = a_\ell/L_\ell$ gives instead $\sum_\ell a_\ell^2/(2L_\ell)$. These statements concern the fixed objective and exact Gaussian queries; they do not identify the curvature constants or the tuned experimental steps.

F.4 What activation coverage says about rank and basis error

For the exact leading activation basis, put $w_i = s_i^2\|Qv_i\|_2^2$ and $C_r = \sum_{i \le r} w_i$. With the objective, query count, and spectral smoothness constant fixed, the isolated-block few-query proxy is proportional to $C_r/(m+r)$. For adjacent admissible ranks,

$$\frac{C_{r+1}}{m+r+1} > \frac{C_r}{m+r} \iff w_{r+1} > \frac{C_r}{m+r}. \tag{30}$$

The analogous raw Frobenius proxy C_r/r requires $w_{r+1} > C_r/r$. This difference concerns the geometry of the certificate: raw spectral estimation obeys the first rule as well. It is a criterion for a dimension-based proxy, not a theorem that increasing rank improves the realized loss decrease. It also requires upstream-gradient information, so it is not a forward-only rank-selection algorithm as stated.

The exact capture formula uses the exact leading singular subspace of H. For an approximate or stale orthonormal basis $\widetilde{A}$, write $P = AA^\top$ and $\widetilde{P} = \widetilde{A}\widetilde{A}^\top$. Then

$$|\rho_{\widetilde{A}}(G) - \rho_A(G)| = \frac{|\operatorname{tr}(G^\top G(\widetilde{P} - P))|}{\|G\|_F^2} \le \|\widetilde{P} - P\|_{\mathrm{op}}.$$

Thus the general measured ACR remains applicable, but activation singular values alone do not certify a stale or approximate basis. If a stored full-column-rank basis satisfies $\|\widetilde{A}^\top \widetilde{A} - I\|_{\mathrm{op}} \le \delta < 1$, its uncorrected energy $\|G\widetilde{A}\|_F^2$ lies between $1-\delta$ and $1+\delta$ times the energy of the orthogonal projector onto its column space.

F.5 Transferring a certificate to an approximate update

For a fixed block let $D = \operatorname{msign}(\widehat{S})$ be the ideal update and $\widetilde{D}$ the implemented update. For $b > 0$, if

$$\mathbb{E}\langle S, D\rangle \ge a, \qquad \mathbb{E}\|\widetilde{D} - D\|_F \le \delta, \qquad \mathbb{E}\|\widetilde{D}\|_{\mathrm{op}}^2 \le b^2,$$

then Cauchy–Schwarz and the expected descent inequality give an oracle certificate

$$\mathbb{E}[F(W) - F(W - \eta\widetilde{D}A^\top)] \ge \frac{(a - \|S\|_F\delta)_+^2}{2L_* b^2}, \qquad \eta = \frac{(a - \|S\|_F\delta)_+}{L_* b^2},$$

where $x_+ = \max\{x, 0\}$. The tolerance needed to preserve the few-query signal is therefore on the scale $\delta < c\sqrt{q/(m+r)}$, or $\delta < c\sqrt{\theta q/(m+r)}$ for shared queries. This is a conditional error-transfer statement; it does not supply a finite-difference or Newton–Schulz error estimate.

In particular, the implementation's quintic coefficients $(3.4445, -4.7750, 2.0315)$ give the singular-value map $p(x) = 3.4445x - 4.7750x^3 + 2.0315x^5$, for which $p(1) = 0.701$. This finite-iteration spectral filter is not an iteration converging to unit singular values. Its norm and alignment defect must be controlled to transfer the exact-polar certificate; an operator-norm bound of one cannot simply be assumed.

G Fixed-query spectral redistribution under general upper models

This appendix proves Theorem 4.4, including its comparison with scalar-normalized raw updates, and the mixed-curvature extension (14). Throughout, the number of directional queries q is fixed: the asymptotic is in matrix size, not in query count. All expectations condition on the iterate, activation basis, gradient, and curvature information before drawing the queries.

G.1 A common model and exact finite-dimensional certificates

Let $f(B) = F(W + BA^\top)$, $S = \nabla f(0) \ne 0$, $s = \|S\|_F$, and $\bar{S} = S/s$. Suppose

$$f(B) \le f(0) + \langle S, B\rangle + \tfrac{1}{2}\Psi(B), \qquad \Psi(B) = \operatorname{tr}(BCB^\top) + \lambda\|B\|_{\mathrm{op}}^2, \quad C \succeq 0, \quad \lambda \ge 0, \tag{31}$$

where $c := \operatorname{tr} C$ satisfies $c + \lambda > 0$. The upper bound must hold on the support of the analyzed displacements; because raw Gaussian updates are unbounded, a bound only inside an unspecified fixed neighborhood does not suffice. A global bound is a sufficient assumption. Convexity, a quadratic objective, a constant Hessian, and Hessian Lipschitzness are not required. For any random direction U with positive mean alignment and finite, positive $\mathbb{E}\Psi(U)$, optimizing a conditional deterministic step gives

$$\mathbb{E}[f(0) - f(-\eta U)] \ge \eta\mathbb{E}\langle S, U\rangle - \tfrac{1}{2}\eta^2\mathbb{E}\Psi(U), \qquad \mathcal{D}_\Psi(U) := \frac{(\mathbb{E}\langle S, U\rangle)^2}{2\mathbb{E}\Psi(U)}. \tag{32}$$

These are model-optimal, oracle-step certificates, not orderings of actual loss decreases or guarantees for a prescribed training learning rate.

Use the exact isolated estimator (7) and let $D = \operatorname{msign}(\widehat{S})$, $a_q = \mathbb{E}\langle S, D\rangle$, $M_q = \mathbb{E}\|\widehat{S}\|_{\mathrm{op}}^2$, and $h = \operatorname{tr}(SCS^\top)$. For $m \geq r$, $\widehat{S}$ has full column rank almost surely, so $D^\top D = I_r$. Gaussian fourth moments and unbiasedness give the exact certificates

$$\mathcal{D}_{\mathrm{raw}} = \frac{s^4}{2\big[(1+1/q)h + (ms^2/q)c + \lambda M_q\big]}, \qquad \mathcal{D}_{\mathrm{polar}} = \frac{a_q^2}{2(c+\lambda)}. \tag{33}$$

Indeed, for $g = \operatorname{vec}(S)$, $\mathbb{E}\operatorname{vec}(\widehat{S})\operatorname{vec}(\widehat{S})^\top = (1+1/q)gg^\top + (s^2/q)I$. Contracting with the quadratic form induced by C proves the raw cost; the polar cost is exactly $c + \lambda$.

The model includes activation-aware curvature without requiring a Kronecker approximation to a Hessian. If $f(B) = \ell(Z_0 + BK)$, where $K = A^\top H$ and ℓ has a one-sided Frobenius upper bound with constant L, then $C = LKK^\top, \lambda = 0$ is valid even when ℓ is nonconvex. A spectral term may represent additional available upper-curvature information. The comparison must use the same justified model for both updates; selecting a looser spectral term solely to favor one certificate is not evidence of an optimization benefit.

A stronger scalar-normalization control. Write $b = \langle S, \widehat{S}\rangle \geq 0$ and $v = \Psi(\widehat{S}) > 0$ almost surely. Assume $\mathbb{E}(b^2/v) < \infty$, as guaranteed when $m - r > 2$ by the inverse-moment argument below; this holds eventually in the fixed-aspect limit. Among all nonnegative measurable sample-dependent scalar multipliers $t = t(\widehat{S}, S)$ with finite model moments, Cauchy–Schwarz gives

$$\mathcal{D}_{\mathrm{scalar}} := \sup_{t \geq 0} \mathcal{D}_\Psi(t\widehat{S}) = \frac{1}{2}\mathbb{E}\frac{b^2}{v}. \tag{34}$$

In fact, $(\mathbb{E}tb)^2 \leq \mathbb{E}(t^2 v)\mathbb{E}(b^2/v)$, with equality for $t = b/v$. This benchmark includes Frobenius or spectral normalization, clipping, and even scalar choices that know the true gradient. It is still a benchmark for the upper model, not an oracle for the actual nonlinear loss.

G.2 A UNIFORM GAUSSIAN POLAR RESPONSE

For a fixed integer $q \geq 1$, define

$$\mu_q = \mathbb{E}\sqrt{\chi_q^2} = \sqrt{2}\,\frac{\Gamma((q+1)/2)}{\Gamma(q/2)}, \qquad \kappa_q = \frac{\mu_q^2}{q}, \qquad e_\beta = (1+\sqrt{\beta})^2. \tag{35}$$

The Gaussian mean-singular-value limit is

$$\mu_\beta = \frac{2}{\pi}\int_0^\pi \frac{\sin^2 t}{\sqrt{1+\beta-2\sqrt{\beta}\cos t}}\,dt, \qquad 0 < \beta < 1. \tag{36}$$

This is the square-root moment of the Marchenko–Pastur law (Marchenko & Pastur, 1967); its continuous limits are $\mu_0 = 1$ and $\mu_1 = 8/(3\pi)$.

Lemma G.1 (Uniform fixed-query response)**.** *Let $m, r \to \infty$ with $r/m \to \beta \in (0,1)$ and keep q fixed. Uniformly over nonzero $S \in \mathbb{R}^{m\times r}$,*

$$\frac{a_q}{s} = \frac{\mu_q\mu_\beta}{\sqrt{m}} + o(m^{-1/2}), \qquad \frac{M_q}{s^2} = \frac{m}{q}e_\beta(1+o(1)). \tag{37}$$

Proof. Let $T \sim \chi_q^2$ be independent of a standard Gaussian matrix $Z \in \mathbb{R}^{m\times r}$, and put $g = \langle Z, \bar{S}\rangle$. The exact conditional Gaussian representation is

$$\widehat{S} \overset{d}{=} \frac{s\sqrt{T}}{q}Y, \qquad Y = Z + u\bar{S}, \quad u = \sqrt{T} - g, \qquad \langle S, \widehat{S}\rangle = \frac{s^2 T}{q}. \tag{38}$$

It follows by decomposing each query into its component along $\bar{S}$ and its independent orthogonal component. The positive scalar prefactor does not affect the polar factor.

Write $P(Z) = \mathrm{msign}(Z)$ and $h_Z = \langle \bar{S}, DP(Z)[\bar{S}] \rangle$. Orthogonal invariance forces $\mathbb{E}DP(Z)$ to be a scalar multiple of the identity operator on matrix space: independent row and column sign changes remove off-diagonal coordinate entries, and permutations identify the diagonal entries. Gaussian integration by parts determines the scalar:

$$\mathbb{E}h_Z = \frac{\mathbb{E}\|Z\|_*}{mr}, \qquad \mathbb{E}\langle \bar{S}, P(Z) \rangle = 0, \qquad \mathbb{E}(gh_Z) = 0. \tag{39}$$

The last two identities use the oddness of P and g and the evenness of DP. Integration by parts can first be performed on $Z(Z^\top Z + \epsilon I)^{-1/2}$; the inverse-moment bound below justifies passing to the polar limit.

We give remainder control because pointwise differentiability alone would not justify the dimension-dependent limit. Gaussian extreme-singular-value bounds (Vershynin, 2012, Corollary 5.35) and fixed-q chi-square tails imply that, for constants depending only on a compact neighborhood of β,

$$\mathcal{E} = \{ s_{\min}(Z) \geq a\sqrt{m}, \quad \|Z\|_{\mathrm{op}} \leq b_0\sqrt{m}, \quad |u| \leq a\sqrt{m}/2 \} \quad \text{satisfies} \quad \mathbb{P}(\mathcal{E}^c) \leq K_q e^{-c_0 m}.$$

On this event the whole segment $Z + t u\bar{S}$, $0 \leq t \leq 1$, is well conditioned. For completeness, the integral representation

$$P(V) = \frac{2}{\pi} \int_0^\infty V(t^2 I + V^\top V)^{-1}\, dt$$

implies, when $s_{\min}(V) \geq k$ and $\|V\|_{\mathrm{op}} \leq K$,

$$\|D^2 P(V)[E, H]\|_F \leq \left(\frac{3K}{k^3} + \frac{3K^3}{k^5} \right) \|E\|_F \|H\|_F.$$

To see the dimension-free constant, put $R = (t^2 I + V^\top V)^{-1}$ and $B_E = V^\top E + E^\top V$. The second derivative of VR is

$$-ERB_H R - HRB_E R + VRB_H RB_E R + VRB_E RB_H R - VR(E^\top H + H^\top E)R.$$

Its Frobenius norm is at most $[6K/(t^2 + k^2)^2 + 8K^3/(t^2 + k^2)^3]\|E\|_F \|H\|_F$; integration gives the displayed bound. On $\mathcal{E}$ this is $O_\beta(m^{-1})$. Thus

$$\langle \bar{S}, P(Y) \rangle = \langle \bar{S}, P(Z) \rangle + u h_Z + O_\beta(u^2/m) \quad \text{on } \mathcal{E}. \tag{40}$$

The complement is negligible in expectation. Indeed, $|\langle \bar{S}, P(Y) \rangle| \leq \sqrt{r}$ and $|h_Z| \leq 2/s_{\min}(Z)$. For $V = Z^\top Z$, the Schur-complement identity gives $(V^{-1})_{jj} \overset{d}{=} 1/\chi^2_{m-r+1}$, and hence

$$\mathbb{E}s_{\min}(Z)^{-4} \leq \mathbb{E}(\mathrm{tr}\, V^{-1})^2 \leq \frac{r^2}{(m-r-1)(m-r-3)} = O_\beta(1). \tag{41}$$

Here $(\sum_j x_j)^2 \leq r \sum_j x_j^2$ and the chi-square inverse second moment suffice; independence of diagonal entries is not asserted. Since $\mathbb{E}u^4 = O_q(1)$, Hölder's inequality bounds the omitted uh_Z contribution by a constant times $\mathbb{P}(\mathcal{E}^c)^{1/2}$. The polar terms are also exponentially negligible. Taking expectations in (40) therefore yields

$$\frac{a_q}{s} = \mu_q \frac{\mathbb{E}\|Z\|_*}{mr} + O_{\beta,q}(m^{-1}).$$

The Marchenko–Pastur law gives $\mathbb{E}\|Z\|_*/(r\sqrt{m}) \to \mu_\beta$; uniform integrability follows from $\|Z\|_*/(r\sqrt{m}) \leq \|Z\|_F/\sqrt{mr}$. This proves the first assertion uniformly over $\bar{S}$.

For the second, $\big|\|Y\|_{\mathrm{op}} - \|Z\|_{\mathrm{op}}\big| \leq |u|$. Gaussian edge convergence and its moment bounds (Vershynin, 2012, Theorem 5.31 and Corollary 5.35) imply $\mathbb{E}\|Z\|^2_{\mathrm{op}}/m \to e_\beta$. Moreover,

$$\frac{1}{m}\mathbb{E}\left[T\left|\|Y\|^2_{\mathrm{op}} - \|Z\|^2_{\mathrm{op}}\right|\right] \leq \frac{1}{m}\mathbb{E}[T(2\|Z\|_{\mathrm{op}}|u| + u^2)] = O_{\beta,q}(m^{-1/2}).$$

Since T and Z are independent, $\mathbb{E}[T\|Y\|^2_{\mathrm{op}}]/m \to q e_\beta$. Substitute (38) to finish the proof. □

G.3 MIXED CURVATURE AND THE STRONGEST SCALAR CONTROL

Theorem G.2 (Fixed-query comparison under mixed curvature)**.** *Under the assumptions of Lemma G.1, allow deterministic $C = C_{m,r} \succeq 0$ and $\lambda = \lambda_{m,r} \geq 0$ in (31), with $c + \lambda > 0$, and suppose $\lambda/(c+\lambda) \to \omega \in [0,1]$. Then*

$$\frac{\mathcal{D}_{\mathrm{polar}}}{\mathcal{D}_{\mathrm{raw}}}, \quad \frac{\mathcal{D}_{\mathrm{polar}}}{\mathcal{D}_{\mathrm{scalar}}} \longrightarrow \kappa_q \mu_\beta^2 \big[(1-\omega) + \omega e_\beta\big]. \tag{42}$$

No assumption on the spectrum or alignment of S and C is needed. The convergence is uniform over normalized gradients and over the normalized positive-semidefinite curvature matrices at each shape.

Proof. Because $0 \leq h \leq s^2 c$, its contribution to the raw cost in (33) is lower order. Write $d = (1-\omega) + \omega e_\beta \geq 1$. Lemma G.1 gives

$$\mathcal{D}_{\mathrm{raw}} = \frac{s^2 q}{2m(c+\lambda)d}(1+o(1)), \qquad \mathcal{D}_{\mathrm{polar}} = \frac{s^2 \mu_q^2 \mu_\beta^2}{2m(c+\lambda)}(1+o(1)).$$

This proves the first ratio. To establish the scalar comparison rather than only a comparison with a fixed raw step, put $J(Y) = \mathrm{tr}(YCY^\top) + \lambda \|Y\|_{\mathrm{op}}^2$. Equations (34) and (38) give exactly

$$\mathcal{D}_{\mathrm{scalar}} = \frac{s^2}{2} \mathbb{E} \frac{T}{J(Y)}. \tag{43}$$

For $c > 0$, the Gaussian quadratic form obeys

$$\mathbb{E} \frac{\mathrm{tr}(ZCZ^\top)}{mc} = 1, \qquad \mathrm{Var}\left(\frac{\mathrm{tr}(ZCZ^\top)}{mc}\right) = \frac{2\,\mathrm{tr}(C^2)}{mc^2} \leq \frac{2}{m}.$$

Also $\mathrm{tr}(\bar{S}C\bar{S}^\top) \leq c$, so the perturbation by $u\bar{S}$ changes this normalized quadratic form by $O_{\mathbb{P},q}(m^{-1/2})$, uniformly in C/c and $\bar{S}$. Together with the Gaussian edge this proves $J(Y)/[m(c+\lambda)] \to d$ in probability, including when $c = 0$.

An inverse-moment argument is necessary to pass to expectations. Let Q be the orthogonal projection onto $\mathrm{col}(S)^\perp$, of dimension $k = m - \mathrm{rank}(S) \geq m - r$. Then $QY = QZ$. If $\lambda \geq (c+\lambda)/2$,

$$\frac{J(Y)}{c+\lambda} \geq \tfrac{1}{2}\|QZe_1\|^2, \qquad \|QZe_1\|^2 \sim \chi_k^2.$$

Otherwise $c \geq (c+\lambda)/2$, and diagonalizing C gives

$$\frac{J(Y)}{c+\lambda} \geq \frac{\mathrm{tr}(QZCZ^\top Q)}{2c} \overset{d}{=} \tfrac{1}{2}\sum_j w_j X_j, \qquad w_j \geq 0, \quad \sum_j w_j = 1, \quad X_j \overset{\mathrm{iid}}{\sim} \chi_k^2.$$

For any fixed $p > 1$, convexity gives $(\sum_j w_j X_j)^{-p} \leq \sum_j w_j X_j^{-p}$. In both cases, independence of T and Z, and $\mathbb{E}(\chi_k^2)^{-p} = 2^{-p}\Gamma(k/2-p)/\Gamma(k/2)$, show that $m(c+\lambda)T/J(Y)$ has bounded p-th moments for all sufficiently large m, uniformly over the stated choices. Consequently,

$$m(c+\lambda)\mathbb{E}\frac{T}{J(Y)} \longrightarrow \frac{q}{d}.$$

Thus $\mathcal{D}_{\mathrm{scalar}} = \mathcal{D}_{\mathrm{raw}}(1+o(1))$, proving the second ratio. Taking $C = 0, \lambda = L_*$ proves Theorem 4.4. □

The optimal raw scalar in this representation is $t = b/v = q/J(Y)$: the chi-square amplitude T cancels. Since $J(Y)$ concentrates, this oracle is asymptotically a deterministic scalar step. Flattening the matrix's singular-value profile is a different operation from removing that scalar amplitude, which explains why the favorable ratio can survive this control.

Interpreting the criterion. For a justified common model, the strict inequality favoring the polar certificate is

$$\omega > \frac{(\kappa_q \mu_\beta^2)^{-1} - 1}{2\sqrt{\beta} + \beta}, \tag{44}$$

provided the right-hand side is below one. For example, at $q = 4, \beta = 1/16$, the threshold is approximately 0.266322. At $\omega = 0$, the limit $\kappa_q \mu_\beta^2 < 1$ favors raw; at $\omega = 1$, flattening a sufficiently spread noise spectrum can more than compensate for the query-amplitude penalty. This is a condition on the available curvature model, not a free algorithmic choice of ω. It differs from the first-order, gradient-rank versus activation-curvature conditions of Davis & Drusvyatskiy (2025): here the gradient is not resolved, q remains fixed, and the leading distinction survives a scalar oracle.

G.4 Shared queries and what is invariant

Under the nuisance model of Theorem F.3, let $y_i = \langle S, R_i \rangle + \nu z_i$ and $\theta = s^2/(s^2 + \nu^2) \in (0, 1]$. Its conditional representation is

$$\widehat{S} \overset{d}{=} \frac{s\sqrt{T}}{\sqrt{\theta} q} \left[Z + \left(\sqrt{\theta T} + \sqrt{1-\theta}\, z - g \right) \bar{S} \right],$$

where T, Z, z are independent, $z \sim N(0,1)$, and $g = \langle Z, \bar{S} \rangle$. For fixed $\theta > 0$, the response proof applies with $u = \sqrt{\theta T} + \sqrt{1-\theta}\, z - g$, giving

$$\frac{a_q}{s} = \frac{\sqrt{\theta}\mu_q \mu_\beta}{\sqrt{m}} + o(m^{-1/2}), \qquad \frac{M_q}{s^2} = \frac{m}{\theta q} e_\beta (1 + o(1)).$$

The exact raw activation-curvature cost is $(1 + 1/q)h + (ms^2/(\theta q))c$. Hence both raw and polar certificates are multiplied by θ to leading order, and their ratio still tends to (42). This statement does not cover a signal fraction tending to zero with matrix size, nor does it assert scalar-oracle equivalence for shared queries: individual alignments can now be negative, so a gradient-informed scalar oracle can discard adverse samples. A joint multi-block update additionally needs the blockwise upper model stated in Appendix F.3.

Scale robustness is a separate mechanism. If an independent positive random multiplier a changes an estimator to $a\widehat{S}$, with $\mathbb{E}a = 1$ and $\mathbb{E}a^2 = K$, then its fixed-step raw certificate is divided by K, whereas its exact polar direction is unchanged. Frobenius normalization is also invariant to this multiplier. This elementary robustness property motivates a scale-normalized control; it does not replace the spectral-redistribution comparison above, which already allows every scalar normalization in the isolated-query model.

G.5 Numerical checks and scope

Figure 1, left: alignment and its plateaus. Set $m = 128, r = 6$, and let the top $r \times r$ block of S be diagonal, with all other entries zero and

$$\sigma_j = \frac{j^{-3/4}}{(\sum_{k=1}^{6} k^{-3/2})^{1/2}}, \qquad \|S\|_F = 1.$$

For each of 25 rounded, logarithmically spaced budgets from $q = 1$ to 10^5, draw 4096 independent samples using the exact law (38), with NumPy generator seed 20260925. The methods share the same draws at each budget. With $\alpha_i = \langle S, \operatorname{msign}(\widehat{S}_i) \rangle$ and $K_i = \|\widehat{S}_i\|_{\mathrm{op}}^2$, the plotted estimates are $\bar{\alpha}$ for polar and $1/\sqrt{\bar{K}}$ for raw, using the exact raw mean alignment $\mathbb{E}\langle S, \widehat{S} \rangle = 1$. The bands are twice the standard error: the ordinary sample-mean error for polar, and the delta-method error $\mathrm{sd}(K_i)/(2\sqrt{4096}\,\bar{K}^{3/2})$ for raw. The two dotted horizontal lines mark the exact $q \to \infty$ limits

$$A_{\mathrm{polar}}^{\infty} = \|S\|_* \approx 2.1792, \qquad A_{\mathrm{raw}}^{\infty} = \frac{\|S\|_F^2}{\|S\|_{\mathrm{op}}} \approx 1.3522.$$

The first follows from Proposition F.2; the second from $\widehat{S} \to S$ in mean square and unbiasedness. These are effective alignments under a common operator-norm budget, not cosine similarities. The dashed curve is only the shape of (10), with universal constants omitted; the marked $q_j^\star$ are its transition scales. The pale band marks $q \le 8$; the larger budgets illustrate saturation, not the practical query regime advocated here.

Figure 1, right: limits versus finite checks. The three curves evaluate $G_q(\beta)$ for $q = 1, 4, 8$, using the integral (36) without fitting constants. The four markers are Monte Carlo estimates of the *finite-dimensional* certificate ratio $a_q^2 M_q$ at $q = 4$, with $\|S\|_F = 1$ and the common pure spectral model $C = 0, \lambda = 1$. No loss trajectory is simulated. For each matrix shape in Table 4, S has a diagonal top $r \times r$ block, zeros elsewhere, and the spiked spectrum

$$\sigma_1^2 = \tfrac{1}{2}, \qquad \sigma_j^2 = \frac{1}{2(r-1)}, \quad j = 2, \dots, r.$$

Each independent draw uses $T_i \sim \chi_q^2$ and a standard Gaussian $Z_i \in \mathbb{R}^{m \times r}$, with $Y_i = Z_i + (\sqrt{T_i} - \langle Z_i, S \rangle)S$ and $\widehat{S}_i = (\sqrt{T_i}/q)Y_i$, exactly as in (38). The polar factor is computed as $Y_i(Y_i^\top Y_i)^{-1/2}$, checked against a direct SVD, not with the approximate Newton–Schulz polynomial used in training.

Table 4: All finite checks in Figure 1, right. Every row uses $q = 4$ and the spiked unit-Frobenius gradient specified in the text. N counts independent estimator draws, not training steps. Error bars are twice the estimated Monte Carlo standard error; they do not include the finite-size discrepancy from the limit $G_4(\beta)$.

Shape $m \times r$	$\beta = r/m$	N	Finite ratio ± 2 SE	Limit $G_4(\beta)$
4096×16	$1/256$	2048	0.9829 ± 0.0016	0.9965
4096×64	$1/64$	1024	1.1011 ± 0.0049	1.1139
256×16	$1/16$	4096	1.2942 ± 0.0062	1.3589
1024×64	$1/16$	2048	1.3303 ± 0.0081	1.3589

The last two rows share the same horizontal coordinate because both have $\beta = 1/16$; they are distinct sizes and samples, not duplicate points. Their finite ratios differ, and the larger matrix is closer to the common limit in these checks. Circles denote $r = 16$, squares $r = 64$; no horizontal jitter is applied.

Variance reduction and error bars. The right-panel computation uses exact zero-mean control variates to reduce query-amplitude noise. Put $\mu_q = \mathbb{E}\sqrt{\chi_q^2}$, $b = \mu_\beta/\sqrt{m}$, $e = m(1 + \sqrt{\beta})^2$, and define

$$A_i = \langle S, \operatorname{msign}(Y_i) \rangle - b(\sqrt{T_i} - \mu_q), \qquad B_i = \frac{T_i \|Y_i\|_{\mathrm{op}}^2 - e(T_i - q)}{q^2}.$$

Since both subtracted terms have expectation zero, these remain unbiased estimators of a_q and M_q at every finite size; the asymptotic coefficients b, e do not force the estimate toward the limiting curve. The plotted ratio is the plug-in estimate $\widehat{R} = \bar{A}^2 \bar{B}$. Its delta-method standard error is the sample standard deviation of

$$2\bar{A}\bar{B}(A_i - \bar{A}) + \bar{A}^2(B_i - \bar{B})$$

divided by $\sqrt{N}$, retaining the covariance of the paired moments. The right-panel simulations use NumPy generator seed 20260928.

At $256 \times 16, q = 4$, unplotted flat-spectrum and rank-one controls give ratios approximately 1.292 and 1.285, respectively, versus 1.294 for the plotted spiked gradient (4096 samples per case). These checks test the analytic mechanism; they are not LLM fine-tuning results and do not establish the relevant curvature fraction in the experimental models.

A separate mixed-model check at $m = 512, r = 32, q = 4$, with a spiked unit-Frobenius gradient and $C = I_r/(2r), \lambda = 1/2$, estimates the polar/raw certificate ratio as 1.0975 ± 0.0049 and the polar/scalar-oracle ratio as 1.0971 ± 0.0049 (4096 samples, twice the estimated standard error). With $C = I_r/r, \lambda = 0$, the corresponding raw comparison is 0.8725 ± 0.0038. The checks also verify the exact mixed moments, polar cost, scalar-oracle Cauchy–Schwarz inequality, and a globally valid upper bound for a smooth nonconvex test loss. These finite-size examples illustrate both sides of the model-dependent criterion; they do not assert a fitted curvature model for a neural-network layer.

The results concern exact Gaussian directional derivatives and the exact polar map. Finite-difference bias, minibatch changes, the implemented Newton–Schulz polynomial, and non-targeted parameter updates require their own controls. Appendix F.5 states a conditional alignment-defect transfer bound. In particular, the fixed-aspect asymptotic does not supply a certified crossover dimension for a layer

with a fixed small rank, and a very thin layer can favor the raw certificate. The useful conclusion is a small-query, curvature-dependent mechanism and a measurable criterion for promise, not a universal explanation of every observed gain.

H Proof of Theorem E.1

The rank-one theorem gives an explicit lower bound, with equality when one matrix dimension is one. It already shows positive expected alignment at $q = 1$.

Proof. If $r > m$, transpose all matrices; the estimator and polar alignment commute with transposition, and the asserted bound is symmetric in m, r. We may therefore assume $m \geq r$. If $m = r = 1$, the estimator has the same sign as M almost surely, $Y = 0$, and both sides equal $\|M\|_F$. Otherwise, by left-right rotational invariance of the Gaussian queries, assume without loss of generality that

$$u = e_1, \qquad v = e_1, \qquad M = \sigma E_{11},$$

where $E_{11} \in \mathbb{R}^{m\times r}$ has a single nonzero entry at $(1, 1)$. For each i, define

$$x_i := (R_i)_{11}, \qquad s := \left(\sum_{i=1}^{q} x_i^2\right)^{1/2}.$$

Then $s^2 \sim \chi^2_q$ and

$$\langle M, R_i\rangle = \sigma x_i.$$

Write

$$R_i = x_i E_{11} + N_i,$$

where N_i has zero $(1, 1)$ entry, iid $N(0, 1)$ entries elsewhere, and is independent of x_i. Then

$$\widehat{H} = \frac{\sigma}{q}\left(s^2 E_{11} + \sum_{i=1}^{q} x_i N_i\right).$$

Conditional on $x_1, \ldots, x_q$, the matrix $\sum_i x_i N_i$ has independent Gaussian entries with variance s^2 in every entry except $(1, 1)$, which is zero. Hence, conditional on s,

$$\widehat{H} \overset{d}{=} \frac{\sigma s}{q}(sE_{11} + Z),$$

where $Z_{11} = 0$ and all other entries are iid $N(0, 1)$. Since msign is invariant under positive scalar multiplication,

$$\mathrm{msign}(\widehat{H}) \overset{d}{=} \mathrm{msign}(B_s), \qquad B_s := sE_{11} + Z.$$

It remains to lower-bound

$$\mathbb{E}\,\langle E_{11}, \mathrm{msign}(B_s)\rangle\,.$$

Write the columns of B_s as

$$B_s = [g, b_2, \ldots, b_r],$$

where

$$g = (s, \xi_2, \ldots, \xi_m)^\top, \qquad \xi_2, \ldots, \xi_m \overset{\text{iid}}{\sim} N(0, 1),$$

and $b_2, \ldots, b_r \overset{\text{iid}}{\sim} N(0, I_m)$ independently of g. Let $B_s = QT$ be the thin QR decomposition with positive diagonal entries in T. Write

$$Q = [q_1, \ldots, q_r], \qquad T = (t_{ij})_{i,j=1}^{r}.$$

Then

$$q_1 = \frac{g}{\|g\|_2}, \qquad t_{11} = \|g\|_2\,, \qquad t_{1j} = q_1^\top b_j, \quad j = 2, \ldots, r.$$

Conditional on g, $t_{12}, \ldots, t_{1r}$ are iid $N(0, 1)$.

Let

$$W := \mathrm{msign}(T).$$

Since Q has orthonormal columns,

$$\operatorname{msign}(B_s) = Q \operatorname{msign}(T) = QW.$$

Therefore

$$\langle E_{11}, \operatorname{msign}(B_s)\rangle = e_1^\top QW e_1 = \sum_{j=1}^{r} (e_1^\top q_j) w_{j1}. \tag{45}$$

We next show that all cross terms vanish after conditioning on g.

Fix g, hence fix $q_1 = g/\|g\|_2$. If $e_1 \in \operatorname{span}\{q_1\}$, then $e_1^\top q_j = 0$ for all $j \geq 2$. Otherwise define

$$u_0 := \frac{P_{q_1^\perp} e_1}{\left\| P_{q_1^\perp} e_1 \right\|_2}, \qquad H_0 := I_m - 2u_0 u_0^\top.$$

Then $H_0 q_1 = q_1$. If

$$e_1 = a q_1 + b u_0, \qquad a = e_1^\top q_1, \qquad b = \left\| P_{q_1^\perp} e_1 \right\|_2,$$

then for every $v \in q_1^\perp$,

$$e_1^\top H_0 v = -e_1^\top v. \tag{46}$$

Define $B_s' := H_0 B_s$. Since $H_0 g = g$, conditionally on g,

$$B_s' = [g, H_0 b_2, \ldots, H_0 b_r] \overset{d}{=} B_s.$$

If $B_s = QT$, then $B_s' = (H_0 Q)T$ is the corresponding QR factorization, so the T factor and $W = \operatorname{msign}(T)$ are unchanged. Averaging the two equally distributed matrices B_s and B_s' in (45), and using (46), cancels every $j \geq 2$ term. Hence

$$\mathbb{E}[\langle E_{11}, \operatorname{msign}(B_s)\rangle \mid g] = (e_1^\top q_1)\mathbb{E}[w_{11} \mid g] = \frac{s}{\|g\|_2}\mathbb{E}[w_{11} \mid g]. \tag{47}$$

It remains to lower-bound w_{11}. For square invertible T,

$$W = \operatorname{msign}(T) = (TT^\top)^{-1/2} T.$$

The first column of T is $t_{11} e_1$, so

$$We_1 = t_{11}(TT^\top)^{-1/2} e_1.$$

Thus

$$w_{11} = t_{11} e_1^\top (TT^\top)^{-1/2} e_1.$$

For any positive definite matrix A_0 and unit vector x,

$$x^\top A_0^{-1/2} x \geq \frac{1}{\sqrt{x^\top A_0 x}}.$$

Applying this with $A_0 = TT^\top$ and $x = e_1$ gives

$$w_{11} \geq \frac{t_{11}}{\sqrt{e_1^\top TT^\top e_1}} = \frac{\|g\|_2}{\sqrt{\|g\|_2^2 + \sum_{j=2}^{r} t_{1j}^2}}.$$

Conditional on g, $\sum_{j=2}^{r} t_{1j}^2 \sim \chi^2_{r-1}$. Combining this bound with (47),

$$\mathbb{E}[\langle E_{11}, \operatorname{msign}(B_s)\rangle \mid g] \geq s\mathbb{E}\left[\frac{1}{\sqrt{\|g\|_2^2 + \chi^2_{r-1}}} \,\middle|\, g \right].$$

Since

$$\|g\|_2^2 = s^2 + \chi^2_{m-1},$$

and independent chi-square random variables add, after conditioning only on s,

$$\mathbb{E}[\langle E_{11}, \mathrm{msign}(B_s)\rangle \mid s] \geq \mathbb{E}\left[\frac{s}{\sqrt{s^2+Y}}\,\middle|\, s\right], \qquad Y \sim \chi^2_{m+r-2}.$$

Since $s^2 = X \sim \chi^2_q$, with X independent of Y,

$$\mathbb{E}\left\langle E_{11}, \mathrm{msign}(\widehat{H})\right\rangle \geq \mathbb{E}\sqrt{\frac{X}{X+Y}}.$$

Multiplying by $\sigma = \|M\|_{\mathrm{F}}$ proves the first part of the theorem.

For $m + r > 2$, $U = X/(X+Y) \sim \mathrm{Beta}(q/2, (m+r-2)/2)$, and

$$\mathbb{E}[U^{1/2}] = \frac{\Gamma((q+1)/2)\Gamma((q+m+r-2)/2)}{\Gamma(q/2)\Gamma((q+m+r-1)/2)}.$$

Gautschi–Wendel Gamma-ratio bounds give, for $q \geq 2$,

$$\mathbb{E}[U^{1/2}] \geq \sqrt{\frac{q-1}{q+m+r-2}}.$$

For $q = 1$,

$$\mathbb{E}[U^{1/2}] \geq \sqrt{\frac{2}{\pi(m+r-1)}}.$$

Absorbing constants concludes the last part of the theorem. □

I General matrix-sign ZO dot-product lower bound

We now present the spectrum-sensitive analysis. This is the main technical novelty of the alignment theory: it lower-bounds the alignment of a Muonized q-query coefficient estimator with an arbitrary coefficient gradient M.

Let

$$\widehat{H} := \frac{1}{q}\sum_{i=1}^{q} \langle M, R_i\rangle\, R_i, \qquad R_i \in \mathbb{R}^{m\times r}, \qquad (R_i)_{ab} \overset{\text{iid}}{\sim} N(0,1). \tag{48}$$

Assume $M \neq 0$ and define

$$\bar{M} := \frac{M}{\|M\|_{\mathrm{F}}}, \qquad \left\|\bar{M}\right\|_{\mathrm{F}} = 1.$$

Let

$$\sigma_1 \geq \cdots \geq \sigma_r \geq 0$$

be the singular values of $\bar{M}$, so $\sum_{j=1}^{r} \sigma_j^2 = 1$.

Lemma I.1 (Conditional representation)**.** *Let $M \in \mathbb{R}^{m\times r}$ be nonzero and set*

$$\bar{M} := \frac{M}{\|M\|_F}.$$

Let

$$S \sim \chi^2_q.$$

Then

$$\mathbb{E}\langle M, \mathrm{msign}(\widehat{H})\rangle = \|M\|_F \mathbb{E}_S[f(\sqrt{S})],$$

where f is the polar-alignment curve associated with $\bar{M}$:

$$f(s) = \mathbb{E}_\Xi \left\langle \bar{M}, \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}})\right\rangle.$$

Proof. Let

$$m_0 := \mathrm{vec}(\bar{M}), \qquad z_i := \mathrm{vec}(R_i) \sim N(0, I_{mr}),$$

and define

$$\alpha_i := \langle \bar{M}, R_i \rangle = m_0^\top z_i.$$

Then $\alpha_i \sim N(0,1)$, and we can decompose

$$z_i = \alpha_i m_0 + \xi_i,$$

where

$$\xi_i \sim N(0, I_{mr} - m_0 m_0^\top), \qquad \xi_i \perp \alpha_i.$$

Let

$$S := \sum_{i=1}^{q} \alpha_i^2.$$

Then $S \sim \chi^2_q$. Moreover,

$$\sum_{i=1}^{q} \langle M, R_i \rangle R_i = \|M\|_F \sum_{i=1}^{q} \alpha_i R_i.$$

Conditioned on $S = s^2$,

$$\sum_{i=1}^{q} \alpha_i R_i \overset{d}{=} s(s\bar{M} + \Xi_{\bar{M}}),$$

where

$$\mathrm{vec}(\Xi_{\bar{M}}) \sim N(0, I_{mr} - m_0 m_0^\top).$$

Since the matrix sign is invariant under positive scaling,

$$\mathrm{msign}(\widehat{H}) \overset{d}{=} \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}}) \qquad \text{conditioned on } S = s^2.$$

Taking expectation gives the claim. □

I.1 Spectrum-sensitive smooth lower bound

Throughout this section, universal constants may change from line to line.

Theorem I.2 (Spectrum-sensitive smooth polar alignment with explicit aspect term)**.** *There exist universal constants $c, C > 0$ such that the following holds. Let*

$$\bar{M} \in \mathbb{R}^{m \times r}, \qquad \|\bar{M}\|_F = 1, \qquad m \geq 16r.$$

Let $\sigma_1, \ldots, \sigma_r$ be the singular values of $\bar{M}$. Then, for all $s \geq 0$,

$$f(s) := \mathbb{E}_\Xi \left\langle \bar{M}, \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}}) \right\rangle$$

satisfies

$$f(s) \geq c \sum_{j=1}^{r} \sigma_j \Psi_C \left(\frac{s\sigma_j}{\sqrt{m}}, \frac{r}{m} \right),$$

where

$$\Psi_C(x, \rho) := \frac{x}{(1 + C\rho)\sqrt{1 + C\rho + Cx^2}}.$$

Proof. Let

$$X = X(s) := s\bar{M} + \Xi_{\bar{M}}.$$

For almost every realization, X has full column rank. Write its thin SVD as

$$X = U \Lambda V^\top,$$

where

$$U \in \mathbb{R}^{m \times r}, \qquad V \in O(r), \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \qquad \lambda_i > 0.$$

Let $U_\perp$ be an orthonormal basis for $\mathrm{col}(U)^\perp$, and define

$$A := U^\top \bar{M} V, \qquad B := U_\perp^\top \bar{M} V.$$

By Lemma J.4, f is absolutely continuous and, for almost every $s \geq 0$,

$$f'(s) = \mathbb{E}\left[\left\langle \bar{M}, D\,\mathrm{msign}(X)[\bar{M}]\right\rangle\right].$$

The Fréchet derivative of the polar factor gives

$$\left\langle \bar{M}, D\,\mathrm{msign}(X)[\bar{M}]\right\rangle = \sum_{i<j} \frac{(A_{ij}-A_{ji})^2}{\lambda_i+\lambda_j} + \sum_{i=1}^{r}\sum_{k=1}^{m-r} \frac{B_{ki}^2}{\lambda_i}.$$

Since all terms are nonnegative,

$$f'(s) \geq \mathbb{E}\,\mathrm{tr}\left[\bar{M}^\top (I-UU^\top)\bar{M}(X^\top X)^{-1/2}\right].$$

Let

$$K := \bar{M}^\top \bar{M}.$$

On the event $\mathcal{G}_s$ from Lemma J.1,

$$X^\top X \preceq 2s^2K + 2b^2 mI.$$

Since $A \mapsto A^{-1/2}$ is operator-antitone,

$$(X^\top X)^{-1/2} \succeq (2s^2K + 2b^2mI)^{-1/2}$$

on $\mathcal{G}_s$. Therefore

$$f'(s) \geq \mathbb{E}\,\mathrm{tr}\left[\bar{M}^\top (I-UU^\top)\bar{M}(2s^2K + 2b^2mI)^{-1/2}\mathbf{1}_{\mathcal{G}_s}\right].$$

Take the SVD

$$\bar{M} = U_M \Sigma_M V_M^\top, \qquad \Sigma_M = \mathrm{diag}(\sigma_1,\dots,\sigma_r).$$

Then

$$K = V_M \Sigma_M^2 V_M^\top,$$

and hence

$$(2s^2K + 2b^2mI)^{-1/2} = V_M\,\mathrm{diag}\left(\frac{1}{\sqrt{2s^2\sigma_j^2 + 2b^2m}}\right)_{j=1}^{r} V_M^\top.$$

Thus

$$f'(s) \geq \sum_{j=1}^{r} \frac{\sigma_j^2 \mathbb{E}\left[\|(I-UU^\top)u_j\|^2 \mathbf{1}_{\mathcal{G}_s}\right]}{\sqrt{2s^2\sigma_j^2 + 2b^2m}}.$$

By Lemma J.5,

$$\mathbb{E}\left[\|(I-UU^\top)u_j\|^2 \mathbf{1}_{\mathcal{G}_s}\right] \geq \frac{c}{1 + C(s^2\sigma_j^2 + r)/m}.$$

Therefore

$$f'(s) \geq c\sum_{j=1}^{r} \frac{\sigma_j^2}{\sqrt{s^2\sigma_j^2 + m}} \frac{1}{1 + C(s^2\sigma_j^2 + r)/m}.$$

Equivalently, with

$$x_j := \frac{s\sigma_j}{\sqrt{m}}, \qquad \rho := \frac{r}{m},$$

we get

$$f'(s) \geq \frac{c}{\sqrt{m}} \sum_{j=1}^{r} \frac{\sigma_j^2}{\sqrt{1+x_j^2}\,(1 + Cx_j^2 + C\rho)}.$$

After increasing C and decreasing c if necessary,

$$\frac{1}{\sqrt{1+x^2}\,(1+Cx^2+C\rho)} \geq \frac{c}{(1+Cx^2+C\rho)^{3/2}}.$$

Hence, for almost every s,

$$f'(s) \geq \frac{c}{\sqrt{m}} \sum_{j=1}^{r} \frac{\sigma_j^2}{\left(1 + C\frac{s^2\sigma_j^2}{m} + C\frac{r}{m}\right)^{3/2}}.$$

Since $f(0) = 0$ by symmetry and f is absolutely continuous,

$$f(s) = \int_0^s f'(u)\, du.$$

Therefore

$$f(s) \geq \frac{c}{\sqrt{m}} \sum_{j=1}^{r} \sigma_j^2 \int_0^s \frac{du}{\left(1 + C\frac{u^2\sigma_j^2}{m} + C\frac{r}{m}\right)^{3/2}}.$$

For each j with $\sigma_j > 0$, set

$$v = \frac{u\sigma_j}{\sqrt{m}}, \qquad du = \frac{\sqrt{m}}{\sigma_j}\, dv.$$

The $\sigma_j = 0$ terms vanish. Hence

$$f(s) \geq c \sum_{j=1}^{r} \sigma_j \int_0^{s\sigma_j/\sqrt{m}} \frac{dv}{\left(1 + C\frac{r}{m} + Cv^2\right)^{3/2}}.$$

Let

$$A_\rho := 1 + C\rho.$$

Then

$$\int_0^x \frac{dv}{(A_\rho + Cv^2)^{3/2}} = \frac{x}{A_\rho\sqrt{A_\rho + Cx^2}}.$$

Thus

$$f(s) \geq c \sum_{j=1}^{r} \sigma_j \frac{s\sigma_j/\sqrt{m}}{(1 + Cr/m)\sqrt{1 + Cr/m + Cs^2\sigma_j^2/m}}.$$

This is exactly

$$f(s) \geq c \sum_{j=1}^{r} \sigma_j \Psi_C\left(\frac{s\sigma_j}{\sqrt{m}}, \frac{r}{m}\right).$$

□

Lemma I.3 (Chi-square smoothing for the soft saturation law)**.** *Let $S \sim \chi_q^2$, $q \geq 1$, and define*

$$\Phi_C(x) := \frac{x}{\sqrt{1 + Cx^2}}, \qquad x \geq 0,$$

where $C > 0$ is fixed. Then there exists a constant $c_\chi > 0$, depending only on C, such that for every $a \geq 0$,

$$\mathbb{E}\Phi_C(a\sqrt{S}) \geq c_\chi \Phi_C(a\sqrt{q}).$$

Proof. By the Paley–Zygmund inequality applied to S, with $\theta = 1/2$,

$$\mathbb{P}(S \geq q/2) = \mathbb{P}(S \geq \tfrac{1}{2}\mathbb{E}S) \geq \frac{(1-\theta)^2(\mathbb{E}S)^2}{\mathbb{E}S^2} = \frac{1}{4}\frac{q^2}{q(q+2)} \geq \frac{1}{12}.$$

Since Φ_C is increasing,

$$\mathbb{E}\Phi_C(a\sqrt{S}) \geq \mathbb{P}(S \geq q/2)\, \Phi_C(a\sqrt{q/2}) \geq \frac{1}{12}\Phi_C(a\sqrt{q/2}).$$

For every $x \geq 0$,

$$\Phi_C(x/\sqrt{2}) = \frac{x/\sqrt{2}}{\sqrt{1 + Cx^2/2}} \geq \frac{1}{\sqrt{2}}\frac{x}{\sqrt{1 + Cx^2}} = \frac{1}{\sqrt{2}}\Phi_C(x).$$

Therefore

$$\mathbb{E}\Phi_C(a\sqrt{S}) \geq \frac{1}{12\sqrt{2}}\Phi_C(a\sqrt{q}).$$

This proves the lemma with $c_\chi = 1/(12\sqrt{2})$. □

Corollary I.4. *Under the assumptions of Theorem I.2, there exist universal constants $c, C > 0$ such that*

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c\|M\|_F \sum_{j=1}^{r} \sigma_j \Phi_C\left(\sigma_j\sqrt{\frac{q}{m+r}}\right),$$

where

$$\Phi_C(x) := \frac{x}{\sqrt{1+Cx^2}}.$$

Proof. Theorem I.2 and Lemma I.1 give

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c\|M\|_F \mathbb{E}_S\left[\sum_{j=1}^{r} \sigma_j \Psi_C\left(\sigma_j\sqrt{\frac{S}{m}}, \frac{r}{m}\right)\right],$$

where

$$\Psi_C(x,\rho) := \frac{x}{(1+C\rho)\sqrt{1+C\rho+Cx^2}}.$$

Let

$$\rho := \frac{r}{m}, \qquad y_j := \sigma_j\sqrt{\frac{S}{m+r}}.$$

Then

$$\sigma_j\sqrt{\frac{S}{m}} = y_j\sqrt{1+\rho}.$$

Therefore

$$\Psi_C\left(\sigma_j\sqrt{\frac{S}{m}}, \rho\right) = \frac{y_j\sqrt{1+\rho}}{(1+C\rho)\sqrt{1+C\rho+C(1+\rho)y_j^2}}.$$

Since the theorem assumes the tall regime $m \geq C_0 r$, we have

$$0 \leq \rho \leq \frac{1}{C_0}.$$

Hence the factor

$$\frac{\sqrt{1+\rho}}{1+C\rho}$$

is bounded below by a universal positive constant, and

$$1+C\rho+C(1+\rho)y_j^2 \leq C'(1+C'y_j^2)$$

for another universal constant $C' > 0$. Thus

$$\Psi_C\left(\sigma_j\sqrt{\frac{S}{m}}, \frac{r}{m}\right) \geq c\frac{y_j}{\sqrt{1+C'y_j^2}} = c\Phi_{C'}(y_j).$$

Substituting

$$y_j = \sigma_j\sqrt{\frac{S}{m+r}}.$$

Finally applying Lemma I.3 yields the result. □

Corollary I.5 (Hard-saturation)**.** *Under the assumptions of Theorem I.2, there exists a universal constant $c > 0$ such that*

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c\|M\|_F \sum_{j=1}^{r} \sigma_j \min\left\{\sigma_j\sqrt{\frac{q}{m+r}}, 1\right\}.$$

Proof. By Corollary I.4,

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c\|M\|_F \sum_{j=1}^{r} \sigma_j \Phi_C\left(\sigma_j \sqrt{\frac{q}{m+r}}\right).$$

For every fixed $C > 0$,

$$\Phi_C(x) = \frac{x}{\sqrt{1+Cx^2}} \geq c\min\{x, 1\} \qquad \text{for all } x \geq 0.$$

Applying this with

$$x = \sigma_j \sqrt{\frac{q}{m+r}}$$

proves the claim. □

Corollary I.6 (Uniform-in-q form)**.** *Under the assumptions of Theorem I.2, for every $q \geq 1$,*

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq \frac{c}{\sqrt{1 + Cq/((m+r)\operatorname{sr}(M))}} \|M\|_F \sqrt{\frac{q}{m+r}},$$

where $\operatorname{sr}(M) = 1/\sigma_1^2$, with the constants c, C of Corollary I.4. In particular, if $q \leq c_0(m+r)\operatorname{sr}(M)$, the right-hand side is at least $(1 + Cc_0)^{-1/2} c\|M\|_F \sqrt{q/(m+r)}$.

Proof. Let $x := \sqrt{q/(m+r)}$ and $k = \operatorname{sr}(M)$. Since $\sigma_j^2 \leq 1/k$,

$$\Phi_C(\sigma_j x) = \frac{\sigma_j x}{\sqrt{1 + C\sigma_j^2 x^2}} \geq \frac{\sigma_j x}{\sqrt{1 + Cx^2/k}}.$$

Corollary I.4 then gives

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c\|M\|_F \sum_{j=1}^{r} \sigma_j^2 \frac{x}{\sqrt{1 + Cx^2/k}} = \frac{c}{\sqrt{1 + Cq/((m+r)k)}} \|M\|_F \sqrt{\frac{q}{m+r}}.$$

□

I.2 Finite-dimensional constants

This subsection makes the hidden constants in the alignment bound numerical. Table 5 is a check of the theorem's applicability and the size of its guaranteed lower bound, not a comparison of measured optimizer performance. Throughout, $M \neq 0$ is the coefficient gradient being estimated, and $\widehat{H}$ is its ideal Gaussian ZO estimate from q independent directions. The pair (m, r) is its shape after transposition if necessary, with $m \geq r$. Thus these dimensions describe the projected matrix, not necessarily the full weight matrix. The calculation assumes an isolated block, ideal Gaussian directional queries, and the exact matrix sign.

The aspect requirement can be made explicit: $C_0 = 16$ suffices. Lemma J.5 gives the derivative estimate with $C = 40$ and prefactor $c_1 = 1/(24\sqrt{2})$ for every $m \geq 16r$. Integrating gives the stated Ψ_{40} bound. Thus the requirement is a proved finite-dimensional condition, rather than an unspecified asymptotic threshold.

The constants improve for the representative projected shapes in the table. Use a Gaussian deviation parameter $t = 3$ for the row-deleted matrix and define

$$\gamma_{m,r} = \frac{(\sqrt{m-r-1} - \sqrt{r} - 3)^2}{m}, \qquad p_m = 1 - e^{-9/2} - 3e^{-m/2}.$$

For the shapes in Table 5, the expression inside the square is positive, $\gamma_{m,r} \geq 0.389$, and $p_m > 0.988$. The row moment is exactly $\mathbb{E}\|x\|^2 = s^2\sigma_j^2 + r - \sigma_j^2$, so the Markov argument yields

$$C = \max\{1, 2/(p_m\gamma_{m,r})\}, \qquad c_1 = \frac{p_m}{8\sqrt{2}}.$$

The required C's are 3.39, 3.69, and 5.20 for the distinct shapes in the table. Therefore the common choices $C = 6$ and $c_1 = 0.087$ are valid for all listed shapes.

Chi-square smoothing can also use the explicit constant

$$c_\chi(q) := \mathbb{E}\min\{1, \sqrt{T/q}\}, \qquad T \sim \chi^2_q.$$

Indeed $\Phi_C(\lambda x) \geq \min\{1, \lambda\}\Phi_C(x)$. Its values are $c_\chi(1) \approx 0.6313$ and $c_\chi(4) \approx 0.8295$. Since $C \geq 1$, the conversion from Ψ_C to Φ_C gives

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq c_1 c_\chi(q)\, \frac{\sqrt{1 + r/m}}{(1 + Cr/m)^{3/2}}\, \frac{\|M\|_F \sqrt{q/(m+r)}}{\sqrt{1 + Cq/(m+r)}}. \tag{49}$$

The query denominator can be strengthened by replacing q there by $q/\operatorname{sr}(M)$; the table uses the conservative version above, which requires no knowledge of the gradient's stable rank.

Column definitions. To separate the contributions to (49), write

$$a_{m,r} := \frac{\sqrt{1 + r/m}}{(1 + Cr/m)^{3/2}}, \qquad b_{m,r}(q) := \frac{1}{\sqrt{1 + Cq/(m+r)}},$$

$$K_{m,r}(q) := c_1 c_\chi(q)\, a_{m,r} b_{m,r}(q).$$

The *aspect ratio* m/r checks the sufficient condition $m/r \geq 16$. The *shape factor* $a_{m,r}$ records the loss in the proof's coefficient due to the finite matrix aspect ratio; it approaches one as r/m decreases. The *query factor* $b_{m,r}(q)$ is the saturation correction to the leading $\sqrt{q/(m+r)}$ law, not the full dependence on the query count. The *bound constant* includes both factors and the proof constants c_1 and $c_\chi(q)$. Its displayed value $\underline{K} \leq K_{m,r}(4)$ is rounded down so that every row certifies

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq \underline{K}\, \|M\|_F \sqrt{\frac{4}{m+r}}.$$

These constants are analytical lower-bound coefficients, not measured alignments, cosine similarities, or success probabilities.

For AGOZO, $M = GA$, with activation rank r. For ZO-Muon, *left projection* compresses the output dimension by estimating $P^\top G$, transposed to $M = (P^\top G)^\top$ for the table; *right projection* compresses the input dimension by estimating $M = GP$. Here P has orthonormal columns and projection rank 64. The two $(2048, 64)$ rows therefore have identical constants because these depend only on the coefficient-matrix shape. The methods can still retain different gradient signal, which enters through $\|M\|_F$.

Table 5: Explicit alignment lower-bound constants for representative projected shapes, with $q = 4$, $C = 6$, $c_1 = 0.087$, and $c_\chi(4) \approx 0.829548$. The shape and query factors are displayed to three decimal places; the last column is $K_{m,r}(4)$, computed using unrounded factors and rounded down to three decimal places. All rows satisfy the theorem's aspect condition $m/r \geq 16$.

Method / projection	Coefficient shape (m, r)	Aspect ratio m/r	Shape factor $a_{m,r}$	Query factor $b_{m,r}(4)$	Bound constant $\underline{K}$
AGOZO, $r = 16$	$(1024, 16)$	64	0.881	0.989	0.062
AGOZO, $r = 64$	$(2048, 64)$	32	0.785	0.994	0.056
ZO-Muon, left proj.	$(2048, 64)$	32	0.785	0.994	0.056
ZO-Muon, right proj.	$(1024, 64)$	16	0.639	0.989	0.045

Reading a row and the takeaway. For the first row, $m/r = 64$ meets the aspect condition, while $b_{1024,16}(4) \approx 0.989$ shows that the finite-query saturation correction is close to one. Multiplying all four factors gives

$$K_{1024,16}(4) = 0.087\, c_\chi(4)\, a_{1024,16} b_{1024,16}(4) \approx 0.062863.$$

Rounding down to 0.062 gives the concrete guarantee

$$\mathbb{E}\langle M, \operatorname{msign}(\widehat{H})\rangle \geq 0.062\,\|M\|_F\sqrt{\frac{4}{1040}} \approx 0.003845\,\|M\|_F.$$

The table thus verifies that these projected shapes fall within the proved regime and gives an explicit, conservative constant for the additive $m + r$ alignment scale at four queries. It does not predict that empirical alignment will be this small, or that a larger entry implies a better optimizer: the shape, retained gradient norm, and curvature must also be accounted for.

In particular, under spectral smoothness with constant L_*, the descent lemma and an oracle step size turn a row's alignment guarantee into

$$\mathbb{E}[F(W) - F(W^+)] \geq \frac{K^2}{2L_*}\,\frac{4}{m+r}\,\|M\|_F^2.$$

The alignment constant is *squared* in this descent certificate. Consequently, an order comparison that sets all prefactors to one is a scaling diagnostic, not a finite-dimensional guarantee of superiority. For queries shared across blocks, Appendix F.3 must additionally account for the fraction of query signal attributable to the block; the numerical isolated-block constants above are not certified unchanged by that extension. Finally, full-space attention and MLP matrices with aspect ratios between one and four fail the sufficient aspect condition $m/r \geq 16$. This excludes them from this quantitative bound, not from having positive alignment or useful empirical performance.

I.3 A MATCHING UPPER BOUND FOR MATRIX-SIGN ZO ALIGNMENT

We now show that the $\sqrt{q/(m+r)}$ alignment scale in the few-query regime is not an artifact of the lower-bound proof. For the actual coefficient-space estimator used by AGOZO, the same scale is also an upper bound.

Let

$$M \in \mathbb{R}^{m\times r}$$

be fixed and nonzero, and let

$$R_1, \ldots, R_q \overset{\text{iid}}{\sim} N(0, I_{m\times r}).$$

Define the idealized directional-derivative estimator

$$\widehat{H} := \frac{1}{q}\sum_{i=1}^{q}\langle M, R_i\rangle R_i.$$

Since the matrix sign is invariant under positive rescaling, the factor $1/q$ does not affect $\operatorname{msign}(\widehat{H})$. We therefore work with

$$S_q := \sum_{i=1}^{q}\langle M, R_i\rangle R_i.$$

Lemma I.7. *Let $X \in \mathbb{R}^{m\times r}$ have full column rank, and let*

$$\operatorname{msign}(X) := X(X^\top X)^{-1/2}$$

denote its thin polar factor. Then, for every $H \in \mathbb{R}^{m\times r}$,

$$\langle H, D\operatorname{msign}(X)[H]\rangle \leq \frac{2\|H\|_F^2}{\sigma_{\min}(X)}.$$

Proof. By left and right orthogonal invariance, it suffices to work in coordinates where

$$X = \begin{pmatrix}\Lambda\\ 0\end{pmatrix}, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r), \qquad \lambda_i > 0.$$

Then

$$\operatorname{msign}(X) = \begin{pmatrix}I_r\\ 0\end{pmatrix}.$$

Write

$$H = \begin{pmatrix} A \\ B \end{pmatrix}, \qquad A \in \mathbb{R}^{r\times r}, \qquad B \in \mathbb{R}^{(m-r)\times r}.$$

Let

$$D\,\mathrm{msign}(X)[H] = \begin{pmatrix} K \\ L \end{pmatrix}.$$

Differentiating the polar decomposition $X = UP$, with $U = \mathrm{msign}(X)$ and $P = (X^\top X)^{1/2}$, gives

$$L = B\Lambda^{-1},$$

and K is skew-symmetric with

$$K\Lambda + \Lambda K = A - A^\top.$$

Therefore, for $i \neq j$,

$$K_{ij} = \frac{A_{ij} - A_{ji}}{\lambda_i + \lambda_j}, \qquad K_{ii} = 0.$$

Hence

$$\begin{aligned}\langle H, D\,\mathrm{msign}(X)[H]\rangle &= \langle A, K\rangle + \langle B, B\Lambda^{-1}\rangle \\ &= \sum_{i<j} \frac{(A_{ij} - A_{ji})^2}{\lambda_i + \lambda_j} + \sum_{i=1}^{r}\sum_{k} \frac{B_{ki}^2}{\lambda_i}.\end{aligned}$$

Since

$$\lambda_i + \lambda_j \geq \sigma_{\min}(X), \qquad \lambda_i \geq \sigma_{\min}(X),$$

and

$$\sum_{i<j}(A_{ij} - A_{ji})^2 \leq 2\|A\|_F^2,$$

we obtain

$$\langle H, D\,\mathrm{msign}(X)[H]\rangle \leq \frac{2\|A\|_F^2 + \|B\|_F^2}{\sigma_{\min}(X)} \leq \frac{2\|H\|_F^2}{\sigma_{\min}(X)}.$$

□

Theorem I.8 (Upper bound for matrix-sign ZO alignment)**.** *Assume $m \geq 16r$. Then there exists a universal constant $C > 0$ such that, for every nonzero $M \in \mathbb{R}^{m\times r}$,*

$$\mathbb{E}\,\langle M, \mathrm{msign}(S_q)\rangle \leq C\|M\|_F\sqrt{\frac{q}{m+r}}.$$

Moreover, the global bound

$$\mathbb{E}\,\langle M, \mathrm{msign}(S_q)\rangle \leq \|M\|_*$$

also holds. Consequently,

$$\mathbb{E}\,\langle M, \mathrm{msign}(S_q)\rangle \leq \min\left\{C\|M\|_F\sqrt{\frac{q}{m+r}}, \|M\|_*\right\}.$$

Proof. Normalize

$$\bar{M} := \frac{M}{\|M\|_F}, \qquad \|\bar{M}\|_F = 1.$$

Let

$$m_0 := \mathrm{vec}(\bar{M}), \qquad z_i := \mathrm{vec}(R_i) \sim N(0, I_{mr}),$$

and define

$$\alpha_i := \langle \bar{M}, R_i\rangle = m_0^\top z_i.$$

Then $\alpha_i \sim N(0,1)$, and we can decompose

$$z_i = \alpha_i m_0 + \xi_i,$$

where

$$\xi_i \sim N(0, I_{mr} - m_0 m_0^\top), \qquad \xi_i \perp \alpha_i.$$

Let

$$T := \sum_{i=1}^{q} \alpha_i^2.$$

Then $T \sim \chi_q^2$.

Since

$$\langle M, R_i \rangle = \|M\|_F \alpha_i,$$

we have

$$S_q = \|M\|_F \sum_{i=1}^{q} \alpha_i R_i.$$

Conditioned on $T = s^2$,

$$\sum_{i=1}^{q} \alpha_i R_i \overset{d}{=} s(s\bar{M} + \Xi_{\bar{M}}),$$

where

$$\mathrm{vec}(\Xi_{\bar{M}}) \sim N(0, I_{mr} - m_0 m_0^\top).$$

Therefore, by scale invariance of the matrix sign,

$$\mathrm{msign}(S_q) \overset{d}{=} \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}}) \qquad \text{conditioned on } T = s^2.$$

Define the polar-alignment curve

$$\phi_{\bar{M}}(s) := \mathbb{E}_\Xi \left\langle \bar{M}, \mathrm{msign}(s\bar{M} + \Xi_{\bar{M}}) \right\rangle.$$

Then

$$\mathbb{E} \langle M, \mathrm{msign}(S_q) \rangle = \|M\|_F \mathbb{E}_T[\phi_{\bar{M}}(\sqrt{T})].$$

It remains to upper-bound $\phi_{\bar{M}}(s)$. Let

$$X(s) := s\bar{M} + \Xi_{\bar{M}}.$$

By Lemma J.4, the function $\phi_{\bar{M}}$ is absolutely continuous and, for almost every $s \geq 0$,

$$\phi'_{\bar{M}}(s) = \mathbb{E}\left[\left\langle \bar{M}, D\,\mathrm{msign}(X(s))[\bar{M}] \right\rangle\right],$$

where

$$X(s) := s\bar{M} + \Xi_{\bar{M}}.$$

Applying Lemma I.7 with

$$H = \bar{M}, \qquad \|\bar{M}\|_F = 1,$$

gives

$$\left\langle \bar{M}, D\,\mathrm{msign}(X(s))[\bar{M}] \right\rangle \leq \frac{2}{\sigma_{\min}(X(s))}$$

almost surely. Therefore

$$\phi'_{\bar{M}}(s) \leq 2\mathbb{E}\frac{1}{\sigma_{\min}(X(s))}$$

for almost every $s \geq 0$.

We now bound the inverse smallest singular value. Write

$$\Xi_{\bar{M}} = Z - \langle Z, \bar{M} \rangle \bar{M}, \qquad Z_{ab} \overset{\text{iid}}{\sim} N(0,1).$$

Then

$$X(s) = Z + (s - \langle Z, \bar{M} \rangle)\bar{M}.$$

Let

$$k := \mathrm{rank}(\bar{M}) \leq r,$$

and let $Q_\perp \in \mathbb{R}^{m \times (m-k)}$ have orthonormal columns spanning $\mathrm{col}(\bar{M})^\perp$. Since

$$Q_\perp^\top \bar{M} = 0,$$

we have

$$Q_\perp^\top X(s) = Q_\perp^\top Z.$$

Hence, for every $v \in S^{r-1}$,

$$\|X(s)v\|_2 \geq \|Q_\perp^\top X(s)v\|_2 = \|Q_\perp^\top Zv\|_2.$$

Taking the infimum over $v \in S^{r-1}$ gives

$$\sigma_{\min}(X(s)) \geq \sigma_{\min}(Q_\perp^\top Z).$$

By rotational invariance, $Q_\perp^\top Z$ is an $(m-k) \times r$ matrix with iid $N(0,1)$ entries. Since $k \leq r$ and $m \geq C_0 r$, the Gaussian inverse smallest-singular-value moment bound, Lemma J.2, implies

$$\mathbb{E}\frac{1}{\sigma_{\min}(Q_\perp^\top Z)} \leq \frac{C}{\sqrt{m-k}}.$$

Moreover, because $k \leq r$,

$$m - k \geq m - r.$$

Under $m \geq C_0 r$,

$$m - r \geq \frac{C_0 - 1}{C_0 + 1}(m + r).$$

Therefore

$$\mathbb{E}\frac{1}{\sigma_{\min}(X(s))} \leq \frac{C}{\sqrt{m+r}}.$$

Substituting into the preceding bound on $\phi'_{\bar{M}}$, we obtain

$$\phi'_{\bar{M}}(s) \leq \frac{C}{\sqrt{m+r}}$$

for almost every $s \geq 0$.

At $s = 0$,

$$\phi_{\bar{M}}(0) = 0,$$

because

$$\Xi_{\bar{M}} \overset{d}{=} -\Xi_{\bar{M}}, \qquad \operatorname{msign}(-X) = -\operatorname{msign}(X).$$

Thus, by absolute continuity,

$$\phi_{\bar{M}}(s) = \int_0^s \phi'_{\bar{M}}(u)\, du \leq C\frac{s}{\sqrt{m+r}}.$$

Returning to (I.3),

$$\mathbb{E}\,\langle M, \operatorname{msign}(S_q)\rangle \leq C\|M\|_F \mathbb{E}_T\left[\frac{\sqrt{T}}{\sqrt{m+r}}\right].$$

Since $T \sim \chi_q^2$, Jensen's inequality gives

$$\mathbb{E}\sqrt{T} \leq \sqrt{\mathbb{E}T} = \sqrt{q}.$$

Therefore

$$\mathbb{E}\,\langle M, \operatorname{msign}(S_q)\rangle \leq C\|M\|_F\sqrt{\frac{q}{m+r}},$$

which proves (I.8).

Finally, since $\|\operatorname{msign}(S_q)\|_{\mathrm{op}} \leq 1$ almost surely, spectral–nuclear duality gives

$$\langle M, \operatorname{msign}(S_q)\rangle \leq \|M\|_* \,\|\operatorname{msign}(S_q)\|_{\mathrm{op}} \leq \|M\|_*.$$

Taking expectation proves (I.8), and combining the two bounds gives (I.8). □

Tightness and regimes for general rank. For general-rank M, the upper bound

$$\mathbb{E}\langle M, \mathrm{msign}(S_q)\rangle \le \min\left\{C\|M\|_F\sqrt{\frac{q}{m+r}}, \|M\|_*\right\}$$

should be interpreted as a two-regime envelope. The two terms are equal when

$$q \asymp (m+r)\left(\frac{\|M\|_*}{\|M\|_F}\right)^2.$$

Equivalently, if we define the nuclear effective rank

$$r_*(M) := \left(\frac{\|M\|_*}{\|M\|_F}\right)^2,$$

then the crossover occurs around

$$q \asymp (m+r)r_*(M).$$

For

$$q \ll (m+r)r_*(M),$$

the query-limited term is smaller, and the upper bound becomes

$$\mathbb{E}\langle M, \mathrm{msign}(S_q)\rangle \lesssim \|M\|_F\sqrt{\frac{q}{m+r}}.$$

In the few-query subregime where all normalized singular modes satisfy

$$\sigma_j\sqrt{\frac{q}{m+r}} \lesssim 1,$$

our smooth lower bound gives

$$\mathbb{E}\langle M, \mathrm{msign}(S_q)\rangle \gtrsim \|M\|_F\sqrt{\frac{q}{m+r}},$$

so the query-limited upper bound is tight up to universal constants. On the other hand, when

$$q \gg (m+r)r_*(M),$$

the nuclear-norm ceiling is the smaller term in the upper bound. If q is large enough that every nonzero normalized singular mode is saturated, namely

$$\sigma_j\sqrt{\frac{q}{m+r}} \gtrsim 1 \qquad \text{for all } j \text{ with } \sigma_j > 0,$$

then the smooth lower bound yields

$$\mathbb{E}\langle M, \mathrm{msign}(S_q)\rangle \gtrsim \|M\|_F\sum_j \sigma_j = \|M\|_*,$$

matching the nuclear-norm upper bound. Between these two extremes lies a partial-saturation regime: some large singular modes have already reached their nuclear-norm contribution, while smaller modes remain query-limited. In this middle regime, the lower bound resolves the mode-by-mode behavior through

$$\|M\|_F\sum_j \sigma_j\Phi\left(\sigma_j\sqrt{\frac{q}{m+r}}\right),$$

whereas the upper bound only gives the coarser envelope involving $\|M\|_F$ and $\|M\|_*$. Thus, for general spectra, our bounds are tight in the fully unsaturated few-query regime and in the fully saturated high-query regime. In the intermediate partial-saturation range the lower bound is spectrum-sensitive, but this envelope does not prove its sharpness.

J AUXILIARY LEMMAS

Lemma J.1 (Tall shifted-Gaussian singular-value lemma)**.** *Let $\|\bar{M}\|_F = 1$, $m \geq 16r$, and $\Xi_{\bar{M}} = Z - \langle Z, \bar{M}\rangle \bar{M}$, with Z iid Gaussian. For $X(s) = s\bar{M} + \Xi_{\bar{M}}$, one may take $a = 2/5$, $b = 4$, and $p_0 = 1/3$ in*

$$\mathbb{P}\left(\sigma_{\min}(X(s)) \geq a\sqrt{m},\ \|\Xi_{\bar{M}}\|_{\mathrm{op}} \leq b\sqrt{m}\right) \geq p_0. \tag{50}$$

The event can be chosen to hold simultaneously for every real s, and its probability is at least $1 - e^{-m/32} - 3e^{-m/2}$.

Proof. Let Q span $\mathrm{col}(\bar{M})^{\perp}$ and $k = \mathrm{rank}(\bar{M}) \leq r$. Then $Q^{\top}X(s) = Q^{\top}Z$ is independent of s, and $\sigma_{\min}(X(s)) \geq \sigma_{\min}(Q^{\top}Z)$. The Gaussian singular-value inequality, with $t = \sqrt{m}/4$, gives

$$\sigma_{\min}(Q^{\top}Z) \geq \sqrt{m-k} - \sqrt{r} - \frac{\sqrt{m}}{4} \geq \frac{2}{5}\sqrt{m}$$

with probability at least $1 - e^{-m/32}$. The last inequality uses $k \leq r \leq m/16$. Also $\|\Xi_{\bar{M}}\|_{\mathrm{op}} \leq \|Z\|_{\mathrm{op}} + |\langle Z, \bar{M}\rangle|$. The usual upper Gaussian bound gives $\|Z\|_{\mathrm{op}} \leq 3\sqrt{m}$ except on an event of probability $e^{-m/2}$; the scalar term is at most $\sqrt{m}$ except on an event of probability $2e^{-m/2}$. A union bound proves the claim, since $m \geq 16$ makes the stated probability larger than $1/3$. □

Lemma J.2 (Gaussian inverse smallest-singular-value moment)**.** *Let $W \in \mathbb{R}^{p\times r}$ have iid $N(0,1)$ entries and $p > r \geq 1$. Then the following finite-dimensional bound holds:*

$$\mathbb{E}\frac{1}{\sigma_{\min}(W)} \leq \frac{e\sqrt{p}}{p-r}.$$

In particular, if $p \geq (1+\delta)r$, then the right side is at most $e(1+\delta)/(\delta\sqrt{p})$.

Proof. For $r \geq 2$, apply the Gaussian pseudoinverse bound of Halko et al. (2011, Proposition A.4) to $W^{\top}$: $\|(W^{\top})^{\dagger}\|_{\mathrm{op}} = 1/\sigma_{\min}(W)$. For $r = 1$, the singular value has the χ_p distribution and

$$\mathbb{E}\chi_p^{-1} = 2^{-1/2}\frac{\Gamma((p-1)/2)}{\Gamma(p/2)}.$$

For $p \geq 3$, this is at most $(\mathbb{E}\chi_p^{-2})^{1/2} = 1/\sqrt{p-2} \leq e\sqrt{p}/(p-1)$; the case $p = 2$ follows directly from the displayed gamma ratio. Finally $p - r \geq \delta p/(1+\delta)$ proves the aspect-gap form. □

Lemma J.3 (Right lifting commutes with the polar factor)**.** *Let $H \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{n\times r}$ satisfy $A^{\top}A = I_r$. Then*

$$\mathrm{msign}(HA^{\top}) = \mathrm{msign}(H)A^{\top}. \tag{51}$$

Proof. Let $H = U\Sigma V^{\top}$ be the compact SVD. Since $A^{\top}A = I_r$, the matrix AV also has orthonormal columns. Therefore

$$HA^{\top} = U\Sigma(AV)^{\top}$$

is a compact SVD of $HA^{\top}$. Hence

$$\mathrm{msign}(HA^{\top}) = U(AV)^{\top} = UV^{\top}A^{\top} = \mathrm{msign}(H)A^{\top}.$$

□

Lemma J.4 (Differentiating the polar-alignment curve)**.** *Let $\|\bar{M}\|_F = 1$, $m \geq 16r$, and $X(s) = s\bar{M} + \Xi_{\bar{M}}$. Then $f(s) = \mathbb{E}\langle \bar{M}, \mathrm{msign}(X(s))\rangle$ is absolutely continuous on compact real intervals and*

$$f'(s) = \mathbb{E}\langle \bar{M}, D\,\mathrm{msign}(X(s))[\bar{M}]\rangle$$

for almost every real s.

Proof. Write $X(s) = Z + (s - \langle Z, \bar{M} \rangle)\bar{M}$, with Z iid Gaussian, and let Q have orthonormal columns spanning $\mathrm{col}(\bar{M})^{\perp}$. For $k = \mathrm{rank}(\bar{M}) \le r$, $Q^{\top}Z$ is an $(m-k) \times r$ iid Gaussian matrix. Almost surely it has full column rank, and simultaneously for every real s,

$$\sigma_{\min}(X(s)) \ge \sigma_{\min}(Q^{\top}Z) > 0.$$

Thus the sample alignment curve is continuously differentiable for all s, including when the signal $\bar{M}$ is rank deficient. Lemma I.7 bounds its derivative by $2/\sigma_{\min}(Q^{\top}Z)$, independently of s. This random bound is integrable by Lemma J.2. Apply the fundamental theorem of calculus to each sample path and use Fubini on any compact interval. The resulting integral identity proves absolute continuity and the derivative formula. This uniform projection argument avoids a regularization or a separate rank-deficient exceptional set for each value of s. □

Lemma J.5 (Soft row-leverage lower bound)**.** *Under the assumptions of Lemma J.4, let* $\bar{M} = U_M \Sigma_M V_M^{\top}$*, and let* $U(s)$ *span* $\mathrm{col}(X(s))$*. For a left singular direction* u_j *with singular value* σ_j*, put* $\ell_j(s) = \|(I - U(s)U(s)^{\top})u_j\|^2$*. For the event* $\mathcal{G}_s$ *in Lemma J.1, with* $a = 2/5$ *and* $b = 4$*,*

$$\mathbb{E}[\ell_j(s)\mathbf{1}_{\mathcal{G}_s}] \ge \frac{1/6}{1 + 40(s^2\sigma_j^2 + r)/m}.$$

Proof. Rotate u_j to the first coordinate and write $X(s) = (x^{\top}; Y)$, where $x = X(s)^{\top}u_j$. On the almost-sure full-rank set for Y, Sherman–Morrison gives

$$\ell_j(s) = \frac{1}{1 + x^{\top}(Y^{\top}Y)^{-1}x}.$$

Project Y onto the orthogonal complement of the row-deleted signal. The resulting iid Gaussian matrix has at least $m - r - 1$ rows and r columns. With $t = \sqrt{m}/4$, its smallest singular value is at least

$$\sqrt{m - r - 1} - \sqrt{r} - \frac{\sqrt{m}}{4} \ge \frac{2}{5}\sqrt{m}.$$

Indeed $m - r - 1 \ge m - 2r$ and $r/m \le 1/16$. Consequently the event

$$\mathcal{H}_{j,s} = \{Y^{\top}Y \succeq (4/25)mI,\ \|\Xi_{\bar{M}}\|_{\mathrm{op}} \le 4\sqrt{m}\}$$

has probability at least $1 - e^{-m/32} - 3e^{-m/2} > 1/3$, and $\mathcal{H}_{j,s} \subseteq \mathcal{G}_s$. On this event $\ell_j(s) \ge 1/(1 + 25\|x\|^2/(4m))$. The row moment is exactly

$$\mathbb{E}\|x\|^2 = s^2\sigma_j^2 + r - \sigma_j^2 \le s^2\sigma_j^2 + r =: D_j.$$

Markov's inequality gives $\mathbb{P}(\|x\|^2 > 6D_j) \le 1/6$. Intersecting with $\mathcal{H}_{j,s}$ leaves probability at least $1/6$, without assuming independence of the two events. On this intersection $\ell_j(s) \ge 1/(1 + (75/2)D_j/m) \ge 1/(1 + 40D_j/m)$. This proves the bound, including when $\sigma_j = 0$. □

K Algorithms

This appendix gives the implementation-level pseudocode corresponding to Section 2. Algorithm 1 constructs the activation basis A from the current batch activations using randomized subspace iteration. Algorithm 2 gives the Newton–Schulz routine used in experiments to approximate the polar factor without computing an SVD. Algorithm 3 combines these components into the full multi-layer AGOZO update.

The pseudocode is written to emphasize the forward-only memory profile. A clean forward pass first records the layerwise activations needed to form the low-rank bases. After the bases are extracted, the full activation matrices do not need to be kept across the q finite-difference queries. Each query perturbs all target linear layers jointly, evaluates one scalar loss difference, and then regenerates the layerwise random matrices from stored seeds to restore the weights and accumulate the activation-coordinate estimators. Thus one finite-difference query contributes to every target layer while retaining only rank-r_ℓ bases and coefficient matrices for those layers.

For clarity, the layerwise dimensions in Algorithm 3 are as follows. If $W_\ell \in \mathbb{R}^{d_{\ell,\mathrm{out}} \times d_{\ell,\mathrm{in}}}$, then $H_\ell \in \mathbb{R}^{d_{\ell,\mathrm{in}} \times b}$, $A_\ell \in \mathbb{R}^{d_{\ell,\mathrm{in}} \times r_\ell}$, $R_{\ell,i} \in \mathbb{R}^{d_{\ell,\mathrm{out}} \times r_\ell}$, and $\widehat{S}_\ell \in \mathbb{R}^{d_{\ell,\mathrm{out}} \times r_\ell}$. The lifted perturbation $R_{\ell,i}A_\ell^{\top}$ therefore has the same shape as W_ℓ. The stored seeds are an implementation device: they allow the same random matrices to be regenerated after the perturbed forward pass, so the algorithm need not store all q perturbation matrices.

Algorithm 1 SUBSPACEEXTRACT(H, r, K): activation-informed basis via power iteration

Require: Activation matrix $H \in \mathbb{R}^{n\times b}$, target rank r, number of power-iteration steps K
Ensure: Orthonormal activation basis $A \in \mathbb{R}^{n\times r}$

Sample Gaussian test matrix $\Omega \in \mathbb{R}^{b\times r}$
$Y \leftarrow H\Omega$
for $k = 1, \ldots, K$ **do**
    $[Q, \cdot] \leftarrow \mathrm{qr}(Y)$ ▷ orthonormalize columns
    $Y \leftarrow H(H^\top Q)$
end for
$[Q, \cdot] \leftarrow \mathrm{qr}(Y)$
return $A \leftarrow Q$

Algorithm 2 $\mathrm{NS}_k^{(a,b,c)}(B; \varepsilon)$: Newton–Schulz approximation to $\mathrm{msign}(B)$

Require: Matrix $B \in \mathbb{R}^{p\times q}$, steps k, coefficients (a, b, c), stabilizer $\varepsilon > 0$
Ensure: Approximate polar factor $D \approx \mathrm{msign}(B)$

$\mathrm{transpose} \leftarrow \mathbf{1}\{p > q\}$
if transpose **then**
    $X \leftarrow B^\top$
else
    $X \leftarrow B$
end if
$X \leftarrow X/(\|X\|_F + \varepsilon)$ ▷ normalization for stability
for $\ell = 1, \ldots, k$ **do**
    $A_\ell \leftarrow XX^\top$
    $P_\ell \leftarrow bA_\ell + cA_\ell^2$
    $X \leftarrow aX + P_\ell X$
end for
if transpose **then**
    $D \leftarrow X^\top$
else
    $D \leftarrow X$
end if
return D

L EXPERIMENTAL DETAILS

Anonymous repo of the paper is in the following Link.

Models and datasets Table 2 evaluates OPT-1.3B and OPT-13B (Zhang et al., 2022), together with Llama-3-8B, on the classification tasks SST-2, RTE, CB, BoolQ, and WiC, the multiple-choice tasks COPA and ReCoRD (all from GLUE and SuperGLUE (Wang et al., 2019)), and the generation tasks SQuAD and DROP.

Baselines and evaluations We benchmark AGOZO against the following ZO baselines: MeZO (Malladi et al., 2023), LOZO (Chen et al., 2025), Subspace-MeZO, ZO-Muon (Lang et al., 2026), and AGZO (Lin et al., 2026).

Implementation details We perform hyperparameter tuning for each method to maximize downstream validation performance. Final hyperparamters are reported in tables 6, and 7. For projected or subspace-based methods, the rank r is the dimension of the restricted search space, and the interval v records how often the random projection or activation basis is refreshed. For ZO-Muon and AGOZO, we report the number of optimization steps separately for different directional-query counts N_q. With the one-sided estimator and one clean evaluation per step, the forward count is $(N_q + 1)$ times the number of steps. Thus 8k steps at $N_q = 4$ use 40k forwards. Projection matrices for LOZO, Subspace-MeZO, and ZO-Muon, as well as the activation basis used by AGOZO, are sampled lazily using QR decomposition and then reused according to the listed refresh interval.

Algorithm 3 AGOZO: activation-guided Muonized ZO update

Require: Weights $W = \{W_\ell\}_{\ell=1}^{L}$, batch B, ranks $\{r_\ell\}$, queries q, radius μ, step size η, power steps K
Require: Newton–Schulz parameters $k_{\mathrm{NS}}, (a, b, c), \varepsilon$
Ensure: Updated weights W^+

1. Forward & subspace extraction

Run a clean forward pass on B and compute $f_0 \leftarrow F(W; B)$.
for linear layer $\ell = 1, \ldots, L$ **do**
  Extract activation basis

$$A_\ell \leftarrow \textsc{SubspaceExtract}(H_\ell, r_\ell, K).$$

  Initialize $\widehat{S}_\ell \leftarrow 0 \in \mathbb{R}^{d_{\ell,\mathrm{out}} \times r_\ell}$.
end for

2. In-place ZO coefficient estimation

for $i = 1, \ldots, q$ **do**
  for linear layer $\ell = 1, \ldots, L$ **do**
    Sample seed $s_{\ell,i}$, generate $R_{\ell,i} \sim N(0, I_{d_{\ell,\mathrm{out}} \times r_\ell})$.
    Set $\Delta_{\ell,i} \leftarrow R_{\ell,i} A_\ell^\top$ and perturb $W_\ell \leftarrow W_\ell + \mu \Delta_{\ell,i}$.
  end for
  Compute $f_i^+ \leftarrow F(W; B)$, and set $y_i \leftarrow (f_i^+ - f_0)/\mu$.
  for linear layer $\ell = 1, \ldots, L$ **do**
    Regenerate $R_{\ell,i}$ from $s_{\ell,i}$ and set $\Delta_{\ell,i} = R_{\ell,i} A_\ell^\top$.
    Restore $W_\ell \leftarrow W_\ell - \mu \Delta_{\ell,i}$.
    Accumulate $\widehat{S}_\ell \leftarrow \widehat{S}_\ell + \frac{1}{q} y_i R_{\ell,i}$.
  end for
end for

3. Orthogonalization & lifted update

for linear layer $\ell = 1, \ldots, L$ **do**
  $D_\ell \leftarrow \mathrm{NS}_{k_{\mathrm{NS}}}^{(a,b,c)}(\widehat{S}_\ell; \varepsilon) \approx \mathrm{msign}(\widehat{S}_\ell)$.
  $W_\ell \leftarrow W_\ell - \eta\, D_\ell A_\ell^\top$.
end for
return $W^+ \leftarrow W$

Convergence speed. Figure 4 compares training progress at a budget of 5k forward queries, with each method's learning rate (and rank, where applicable) tuned per cell. On the four OPT-13B cells AGOZO descends fastest from the first few hundred queries and reaches the best baseline's 5k-query loss reduction after 1.5–2.2k queries. On Llama-3-8B it leads BoolQ for the first ≈1k queries before MeZO, LOZO and AGZO catch up.

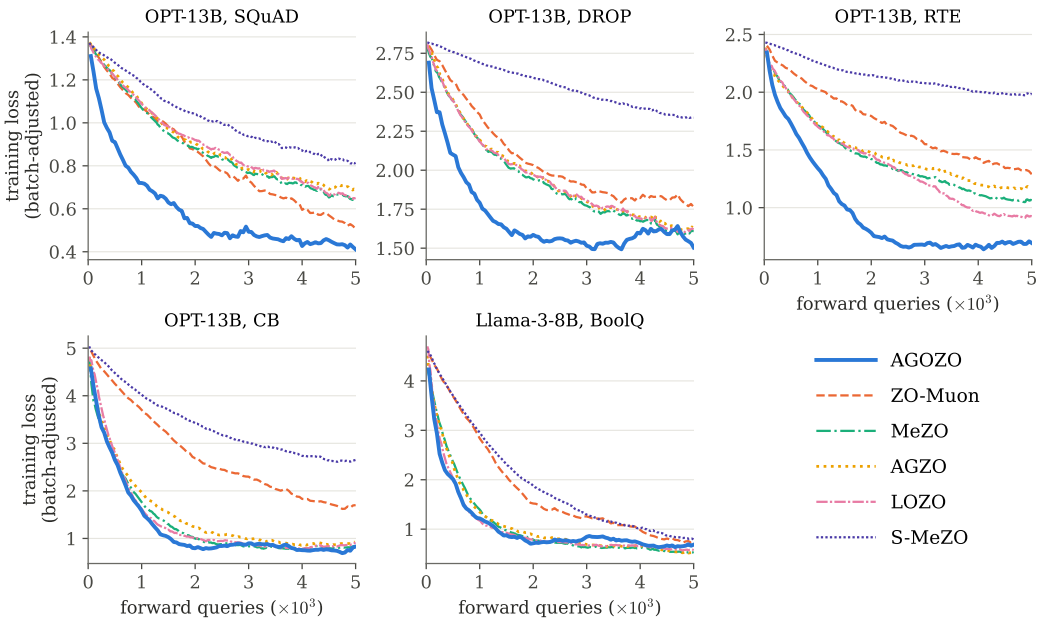


Figure 4: Training progress at a budget of 5k forward queries (one seed). Each curve shows the training loss.

Table 6: Optimizer hyperparameters for the 3-seed experiments. Every zeroth-order method is matched at a budget of ≈ 40k forward queries (steps $\times$ forward passes per step). Rank and refresh refer to the perturbation subspace: a random basis for LOZO, S-MeZO and ZO-Muon, the top right singular vectors of the layer input activations for AGZO and AGOZO.

Method	**LR η**	**Steps**	**Fwd/step**	**Queries**	**Rank**	**Refresh**	**Update**
MeZO	10^{-7} / 5×10^{-8} [a]	20,000	2	40,000	–	–	raw
LOZO	10^{-7} / 5×10^{-8} [a]	20,000	2	40,000	8	100	raw
S-MeZO	10^{-6}	20,000	2	40,000	64	100	raw
ZO-Muon	10^{-2}	8,000	5	40,000	64	100	msign (NS)
AGZO	10^{-7}	20,000	2.01 [b]	40,200	16	100	raw
AGOZO	10^{-3} [c]	8,000	5	40,000	16 [c]	100	msign (NS)

All methods: batch 16, $\epsilon = 10^{-3}$, max. length 2048, 1,000 training examples, full fine-tuning in bfloat16, constant LR, no weight decay. *Estimators:* MeZO, LOZO, S-MeZO and AGZO use a two-point antithetic estimate $(\mathcal{L}(\theta+\epsilon\Delta) - \mathcal{L}(\theta-\epsilon\Delta))/2\epsilon$; ZO-Muon and AGOZO a one-sided estimate averaged over 4 directions, plus one base forward. *AGZO / AGOZO:* 5 power iterations for the activation basis; non-targeted matrices use a Haar basis of rank 8 (refresh 100); the embedding / LM head uses a Gaussian perturbation with a raw update at $\eta = 10^{-7}$. In every method except MeZO, 1-D parameters (biases, LayerNorm) take a fixed 10^{-7} step.

[a] 10^{-7} on SST-2, RTE, CB, BoolQ and WiC for OPT-1.3B and OPT-13B; 5×10^{-8} on all other cells.
[b] Two antithetic forwards per step, plus one forward every 100 steps to refresh the activation basis.
[c] Selected per cell on the ten OPT classification cells (Table 7); $\eta = 10^{-3}$, rank 16 elsewhere.

Table 7: AGOZO learning rate η / activation rank per cell in the 3-seed experiments. On the ten OPT classification cells both were selected on seed 0 from a sweep over $\eta \in [0.8, 8]\times10^{-3}$ (plus 2×10^{-2} on OPT-1.3B CB) and rank $\in \{16, 64\}$; the grid differs slightly by cell. All other cells use $\eta = 10^{-3}$, rank 16.

	Classification					**Multiple choice**		**Generation**	
Model	SST-2	RTE	CB	BoolQ	WiC	COPA	ReCoRD	SQuAD	DROP
OPT-1.3B	2e-3 / 16	2e-3 / 16	8e-3 / 64	2e-3 / 64	6e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16
OPT-13B	2e-3 / 64	2e-3 / 16	2e-3 / 64	2e-3 / 16	2e-3 / 64	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16
Llama-3-8B	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16	1e-3 / 16

M TIME AND MEMORY COMPARISON

We first decompose the per-step cost of AGOZO and ZO-Muon (Section M.1), and then report total training time and peak GPU memory of all methods on all tasks (Section M.2).

M.1 PER-STEP COST BREAKDOWN

We compare the per-step cost of AGOZO with that of ZO-Muon, its closest baseline, on OPT-13B and Llama-3-8B for COPA and ReCoRD. Both methods are run for 600 steps on the same machine with the same zeroth-order estimator (five forward passes per step), and both refresh their subspace every 100 steps: AGOZO recomputes its activation basis and ZO-Muon resamples its random projection. Every region of the step is timed with CUDA-synchronized timers, which together account for 97–99% of the measured step time. Figure 5 reports the amortized cost per step, i.e. the total time of each region over the run divided by the number of steps, so the cost of the refresh steps is spread over the whole run.

The step is dominated by the forward passes and the perturbation of the weights, which every subspace ZO method pays. The forward passes are at parity between the two methods, and so is the perturbation on OPT-13B; on Llama-3-8B AGOZO's perturbation is about 13% more expensive because its untied input and output embeddings are perturbed with full Gaussian noise, a configuration choice rather than part of the method. Activation-basis extraction accounts for 0.7–1.4% of AGOZO's runtime. Newton–Schulz orthogonalization is not specific to AGOZO: ZO-Muon applies it as well, and since AGOZO orthogonalizes an $m \times r$ matrix with $r = 16$ rather than an $r_P \times n$ matrix with $r_P = 64$, it is slightly cheaper for AGOZO in all four settings. Overall, AGOZO's amortized step costs 1.01–1.06× that of ZO-Muon.

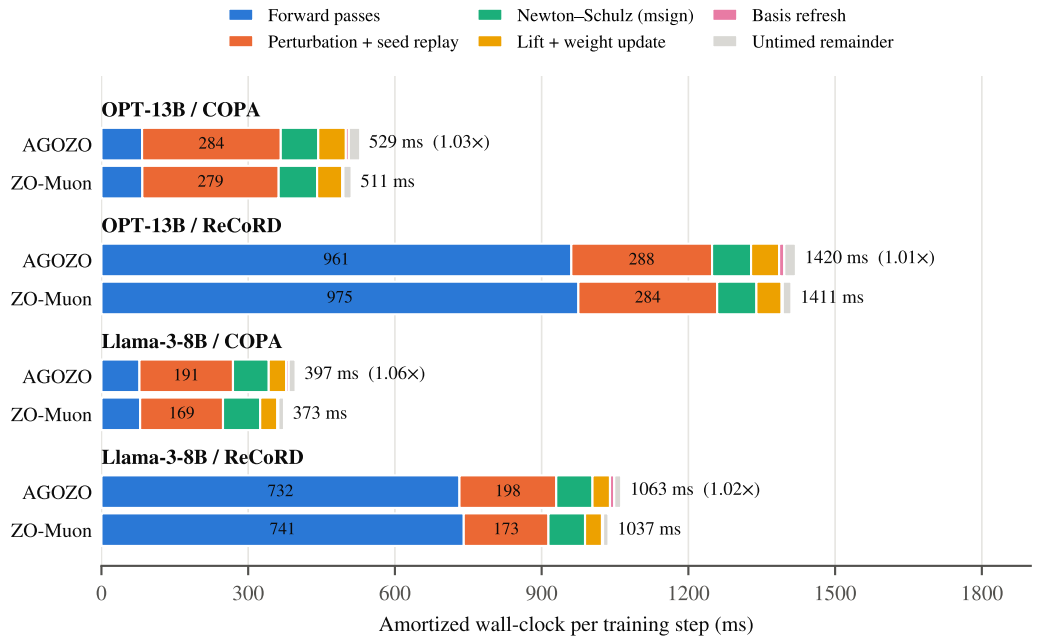


Figure 5: Amortized wall-clock time per training step of AGOZO and ZO-Muon, decomposed by region (600 steps, same hardware, basis refreshed every 100 steps). Numbers inside the bars are milliseconds per step; the ratio is AGOZO relative to ZO-Muon. AGOZO times exclude finite-value assertions left in our implementation for debugging, which are not part of the method. Timers are CUDA-synchronized, so absolute times are inflated relative to uninstrumented runs; the figure is meant for the composition of the step and the difference between the methods.

M.2 TOTAL TRAINING TIME AND PEAK MEMORY

Figures 6 and 7 report the total training time and the peak GPU memory of every method on every model and task, measured on the runs behind Table 2: for each cell we use one run (seed 1) of the reported configuration, and all runs were executed on the same $8\times$H200 node (1 gpu per cell). All methods use the same budget of 40k forward queries. Training time counts optimizer steps only and excludes evaluation, and peak memory is measured during training. Table 8 gives the averages over the nine tasks.

On OPT-13B and Llama-3-8B, MeZO is the slowest method, and AGOZO takes between 0.99 and $1.05\times$ the training time of ZO-Muon on every task. On OPT-1.3B, the two methods that apply Newton–Schulz orthogonalization, ZO-Muon and AGOZO, are the slowest, and AGOZO takes 0.92–$1.00\times$ the time of ZO-Muon. Peak memory is determined by the model and the sequence length of the task rather than by the optimizer: within any model–task pair, the six methods differ by at most 2.6 GB. The largest differences occur on the short-sequence tasks (SST-2, WiC, COPA) of Llama-3-8B, where MeZO, AGZO and AGOZO peak about 2.5 GB above ZO-Muon and S-MeZO.

Table 8: Training time and peak GPU memory averaged over the nine tasks, for the runs behind Table 2 (one run per cell, same $1\times$H200 gpu per cell, 40k forward queries). Per-task values are shown in Figures 6 and 7.

	OPT-1.3B		**OPT-13B**		**Llama-3-8B**	
Method	Time (h)	Mem. (GB)	Time (h)	Mem. (GB)	Time (h)	Mem. (GB)
MeZO	0.47	5.7	3.22	27.6	2.11	24.0
LOZO	0.45	5.7	2.46	27.5	1.64	23.7
S-MeZO	0.46	5.7	2.26	27.4	1.57	23.3
AGZO	0.42	5.7	2.33	27.4	1.66	24.0
ZO-Muon	0.58	5.7	2.46	27.4	1.80	23.3
AGOZO	0.56	5.7	2.49	27.5	1.83	24.0

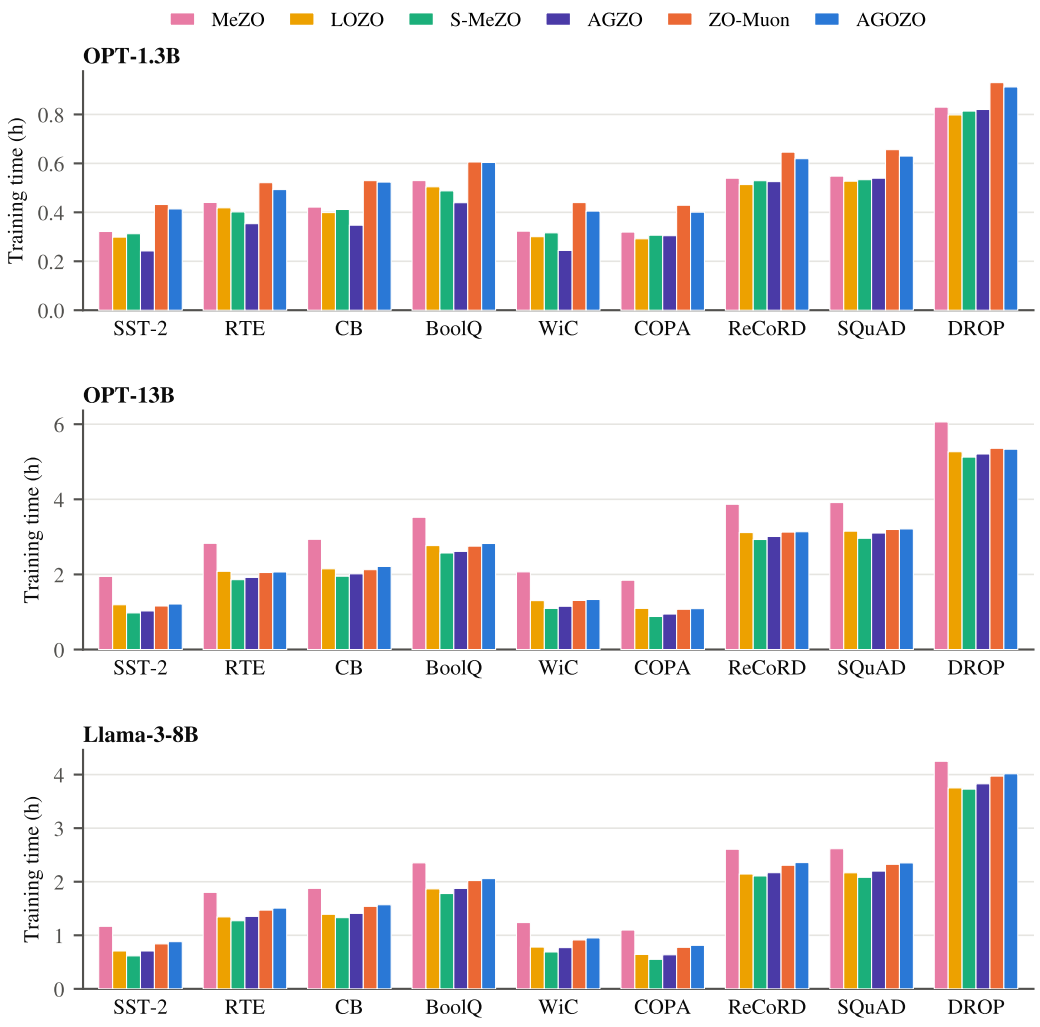


Figure 6: Total training time of each method at a budget of 40k forward queries, for every model and task in Table 2 (one run per cell, same 8×H200 node; evaluation time excluded).

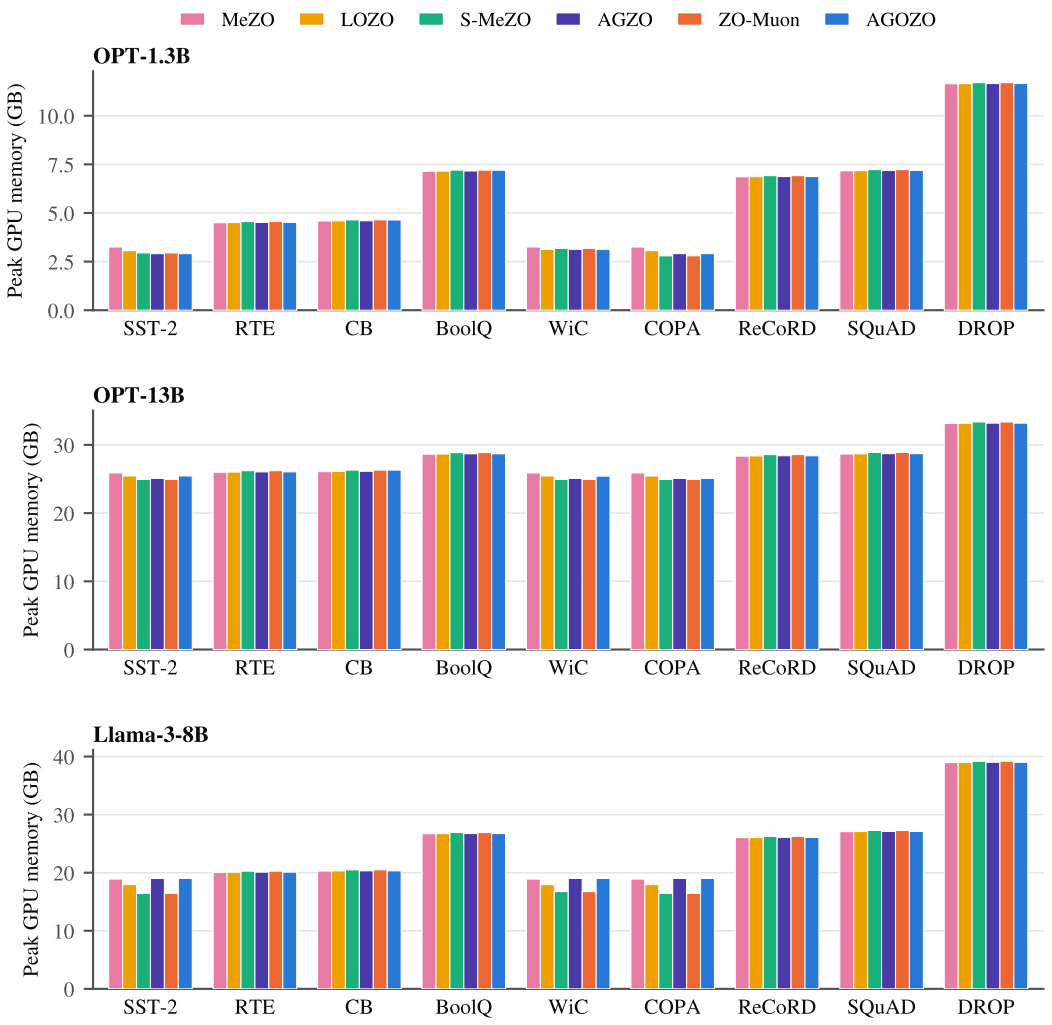


Figure 7: Peak GPU memory during training of each method, for every model and task in Table 2 (same runs as Figure 6).

N MEASURED ACTIVATION COVERAGE AND CROSSOVER CONDITIONS

The dimension factors in Theorem 4.3 motivate coverage diagnostics. Setting the hidden constants equal and replacing random coverage by its fresh, independent Haar expectation gives the thresholds

$$T_{\text{full}} = \frac{m+r}{m+n}, \qquad T_{\text{right}} = \frac{(m+r)\,r_P}{n\,(m+r_P)}, \qquad T_{\text{left}} = \frac{(m+r)\,r_P}{m\,(n+r_P)},$$

against full-space ZO-Muon, right-projection random-subspace ZO-Muon and left-projection random-subspace ZO-Muon, respectively (Appendix E). The smoothness constant L_* is common to AGOZO and the three baselines, so it cancels from each condition. These thresholds compare scaling proxies, not proven orderings of loss decrease. Full-space shapes in these measurements have aspect ratios at most four, outside the sufficient ratio of sixteen. Reused random bases need not be independent of the current gradient, and shared-query signal fractions are not included. Subject to these limitations, the thresholds depend only on the shape $m \times n$ of the layer and on the ranks, which we set to the values used in our experiments, $r = 16$ for AGOZO and $r_P = 64$ for ZO-Muon.

We measure $\rho_A(G) = \|GA\|_F^2 / \|G\|_F^2$ during AGOZO training, with G the true gradient of the layer computed by a backward pass and A the activation basis the optimizer is using at that step. We do this on 13 model–task pairs, all nine tasks for Llama-3-8B and COPA, ReCoRD, SQuAD and DROP for OPT-13B, at every activation-guided layer (240 layers for OPT-13B and 224 for Llama-3-8B) and at 20 checkpoints (every 50 steps of a 1000-step run with $q = 4$ queries per step), for a total of 59,520 measurements. The basis is built in bfloat16 and is orthonormal to $\|A^\top A - I_r\|_F \le 6.2 \times 10^{-3}$, which changes ρ_A by at most about 1% relative.

Table 9 reports the results per model and task. The median coverage is 0.49–0.89, against $r/n \le 0.004$ for a random rank-16 subspace. Both random-subspace thresholds are exceeded on every measurement. The full-space threshold is much larger and is exceeded on 73.5% of the measurements. Table 10 shows that it fails mostly where the geometry makes T_{full} large: for the wide up-projections ($m \gg n$), $T_{\text{full}} \approx 0.8$ and the condition holds on 26–59% of the measurements, whereas for the narrow layers (Llama-3-8B `k_proj`/`v_proj` with $m = 1024$, and the down-projections) it holds on 93–99.5%. For the square attention layers, where $T_{\text{full}} \approx 0.5$, it holds on 39–$99\%$; the lowest value is for OPT-13B `out_proj`, which also has the lowest median coverage (0.44).

Table 10 does not repeat the $> T_{\text{right}}$ and $> T_{\text{left}}$ columns of Table 9, because they equal 100% for every layer type. Its last column shows instead by how much the measured coverage exceeds the left-projection reference threshold: the smallest ratio ρ_A / T_{left} among all measurements of that layer type. A value above one means that even the least favorable measurement lies above the threshold. The smallest value is 2.1, for OPT-13B `k_proj`, and most layer types have a margin between 3 and 26. The margin is large because both thresholds are small: for these shapes T_{left} and T_{right} are within 5% of r_P/n and at most 0.016, whereas the measured coverage is below 0.05 in only 0.04% of the measurements. For the same reason the margins against the right-projection variant are nearly the same as those reported.

Table 9: Measured activation coverage ρ_A and dimension-based crossover diagnostics, per model and task. These suppress universal constants; the full-space rates extrapolate beyond the proved aspect condition. Each measurement is one activation-guided layer at one of 20 checkpoints. The last three columns give the percentage of measurements with ρ_A above the threshold against full-space ZO-Muon (T_{full}), right-projection random-subspace ZO-Muon (T_{right}) and left-projection random-subspace ZO-Muon (T_{left}).

Model	**Task**	**#meas.**	**median** ρ_A	**min** ρ_A	$> T_{\text{full}}$	$> T_{\text{right}}$	$> T_{\text{left}}$
OPT-13B	COPA	4,800	0.789	0.042	82.8%	100%	100%
OPT-13B	ReCoRD	4,800	0.647	0.033	68.8%	100%	100%
OPT-13B	SQuAD	4,800	0.695	0.036	75.5%	100%	100%
OPT-13B	DROP	4,800	0.645	0.027	67.5%	100%	100%
Llama-3-8B	SST-2	4,480	0.883	0.247	91.9%	100%	100%
Llama-3-8B	RTE	4,480	0.790	0.213	81.3%	100%	100%
Llama-3-8B	CB	4,480	0.790	0.482	80.0%	100%	100%
Llama-3-8B	BoolQ	4,480	0.596	0.079	60.4%	100%	100%
Llama-3-8B	WiC	4,480	0.891	0.076	99.4%	100%	100%
Llama-3-8B	COPA	4,480	0.581	0.051	64.8%	100%	100%
Llama-3-8B	ReCoRD	4,480	0.493	0.069	56.9%	100%	100%
Llama-3-8B	SQuAD	4,480	0.621	0.103	63.6%	100%	100%
Llama-3-8B	DROP	4,480	0.554	0.117	62.3%	100%	100%
All		59,520	0.714	0.027	**73.5%**	**100%**	**100%**

Table 10: Dimension-based reference thresholds and measured activation coverage per layer type, pooled over tasks ($r = 16$, $r_P = 64$). $m \times n$ is the weight shape. The last column is the smallest ratio ρ_A / T_{left} over all measurements of that layer type. Threshold exceedance does not certify optimizer dominance.

Model	**Layer**	m	n	T_{full}	T_{right}	T_{left}	**median** ρ_A	$> T_{\text{full}}$	$\min \rho_A / T_{\text{left}}$
OPT-13B	`q_proj`	5120	5120	0.502	0.0124	0.0124	0.832	77.8%	6.7
OPT-13B	`k_proj`	5120	5120	0.502	0.0124	0.0124	0.761	72.8%	2.1
OPT-13B	`v_proj`	5120	5120	0.502	0.0124	0.0124	0.965	99.2%	26.2
OPT-13B	`out_proj`	5120	5120	0.502	0.0124	0.0124	0.443	39.2%	2.6
OPT-13B	`fc1`	20480	5120	0.801	0.0125	0.0124	0.858	58.8%	16.7
OPT-13B	`fc2`	5120	20480	0.201	0.0031	0.0031	0.410	94.1%	9.6
Llama-3-8B	`q_proj`	4096	4096	0.502	0.0154	0.0154	0.674	80.3%	6.5
Llama-3-8B	`k_proj`	1024	4096	0.203	0.0149	0.0156	0.693	99.1%	4.5
Llama-3-8B	`v_proj`	1024	4096	0.203	0.0149	0.0156	0.929	99.5%	3.2
Llama-3-8B	`o_proj`	4096	4096	0.502	0.0154	0.0154	0.777	87.4%	10.1
Llama-3-8B	`gate_proj`	14336	4096	0.779	0.0156	0.0154	0.619	26.3%	7.4
Llama-3-8B	`up_proj`	14336	4096	0.779	0.0156	0.0154	0.642	28.1%	8.3
Llama-3-8B	`down_proj`	4096	14336	0.223	0.0044	0.0045	0.643	93.1%	17.7

O ABLATION: SUBSPACE AND COEFFICIENT UPDATE

AGOZO differs from subspace ZO in two ways: the subspace is chosen from the activations rather than at random, and the coefficient estimate is replaced by its matrix sign. Four of the methods in Table 2 form the corresponding 2×2 design:

	Raw coefficient update	**Matrix-sign update**
Random subspace	S-MeZO	ZO-Muon
Activation subspace	AGZO	AGOZO

Figure 8 shows the task-averaged accuracy of each corner for each model and the four simple effects on every model–task pair, computed from the means reported in Table 2.

The matrix-sign configurations exhibit the larger observed difference. They outperform their raw-update counterparts on all 27 model–task pairs with either subspace, by 11.9 points on average with the random subspace and 10.2 points with the activation subspace, and most on SQuAD, DROP and SST-2. The effect of the activation subspace is smaller and depends on the model. With the raw update, it helps on 13 of the 27 pairs, by 2.2 points on average, mainly on SST-2 and on SQuAD and DROP of the OPT models. With the matrix-sign update, it helps on 15 of the 27 pairs, by 0.6 points on average: $+1.2$ on OPT-1.3B, $+1.9$ on OPT-13B and -1.4 on Llama-3-8B, where ZO-Muon is ahead on RTE, WiC and CB. Since the seed standard deviation of a single method is about 1.4–2.2 points on average (Table 2), differences of a few points on a single pair are not resolved.

The design is not perfectly factorial, and three differences between the corners should be kept in mind. First, the random subspace of S-MeZO and ZO-Muon is a left (output-side) basis of rank $r_P = 64$, whereas AGZO and AGOZO use a right activation basis of rank 16 (64 on five OPT cells). Second, at the common budget of 40k forward queries, the raw-update methods take one query per step for 20k steps and the matrix-sign methods four queries per step for 8k steps, so the update factor also changes how the budget is allocated. Third, each method uses its own learning rate (Appendix L), and AGOZO's was selected per task on the ten OPT classification cells.

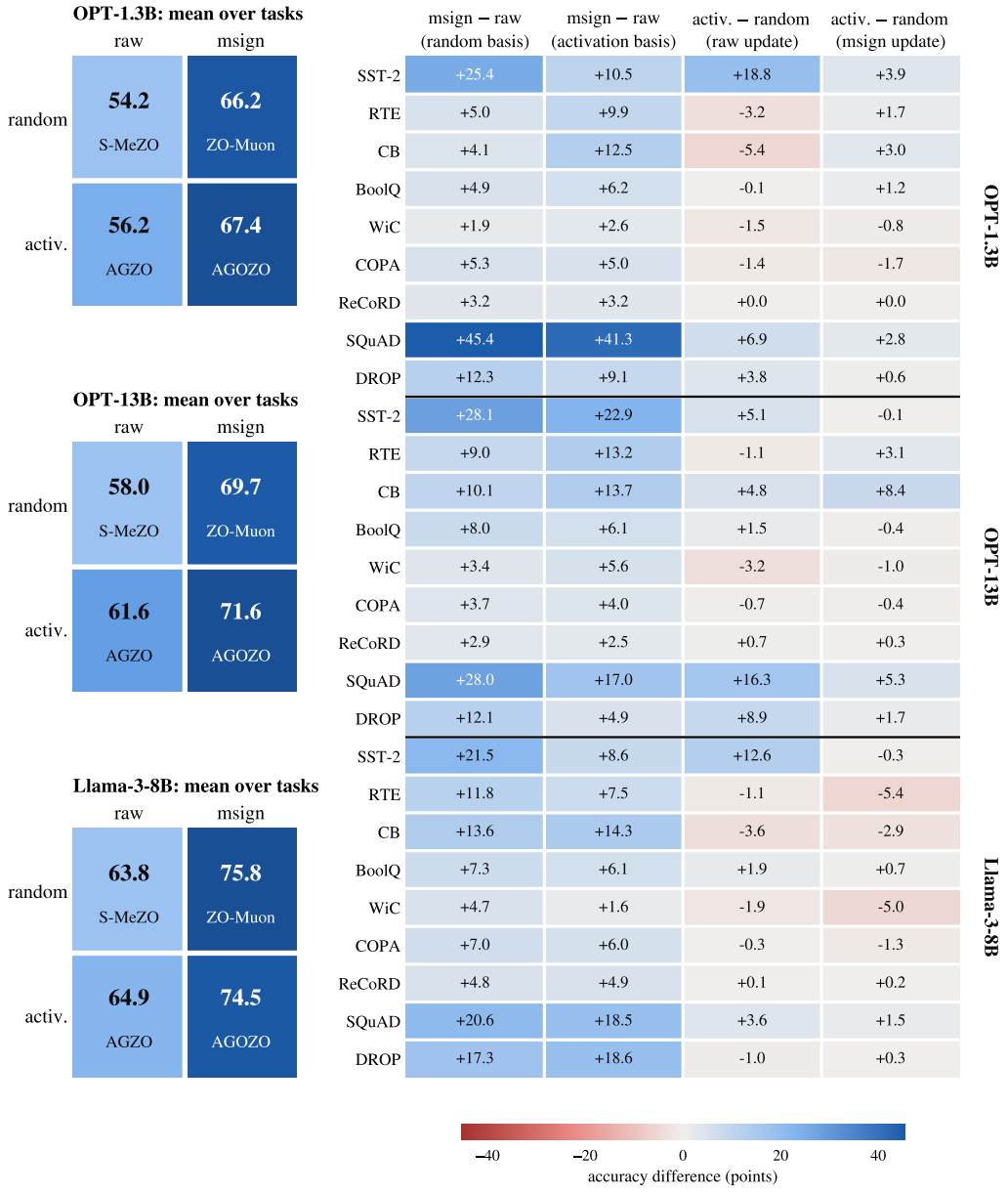


Figure 8: 2×2 ablation of the subspace (random vs. activation) and the coefficient update (raw vs. matrix sign), using the 3-seed means of Table 2. *Left:* mean accuracy over the nine tasks for each corner and model. *Right:* simple effects on every model–task pair, in accuracy points; the first two columns compare the update with the subspace fixed (ZO-Muon−S-MeZO and AGOZO−AGZO), the last two compare the subspace with the update fixed (AGZO−S-MeZO and AGOZO−ZO-Muon).