ESTIMATING TRACES FROM NOISY MATRIX–VECTOR PRODUCTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Trace estimation for curvature, sensitivity, and second-order statistics can involve stochastic products, making the low-rank approximation used for variance reduction noisy. We determine both the optimal query cost and when spending queries on such an approximation improves accuracy. For a d-dimensional positive semidefinite matrix with trace in $[1, 2]$, we allow adaptive unit-vector queries with fresh centered noise of total second moment at most σ^2 and query-direction variance at most $\rho\sigma^2$, where $1/d \le \rho \le 1$. At constant confidence, the minimax cost for relative error ε is $\Theta(\min\{d, \varepsilon^{-1}\} + d^2\rho\sigma^2\varepsilon^{-2})$. Complete correction adds the trace of the realized noisy approximation and subtracts that same approximation in every residual observation. This correction and sufficiently averaged projected estimation attain the rate, with coordinate queries in small dimensions. An exact equal-budget variance identity determines when correction improves on independent sphere sampling. Fixed noise covariance reduces the worst-case noise term to $d\sigma^2\varepsilon^{-2}$. Experiments on an 81,290-parameter CNN reveal a minibatch-quality crossover. Paid decisions and a fresh-product XTrace comparison expose the cost of noisy pilots; model-score comparisons identify settings where direct model access is computationally preferable.

1 INTRODUCTION

At fixed model parameters and a fixed query direction, two computations of the same matrix–vector product can return different vectors. In neural-network curvature diagnostics, resampling a minibatch changes a Fisher or generalized Gauss–Newton product used to summarize average curvature (Martens, 2020; Schneider et al., 2021; Dangel et al., 2023). In Jacobian regularization, sampling inputs or data augmentations changes the Gram product whose trace measures average squared sensitivity (Hoffman et al., 2019). In streaming feature or gradient statistics, fresh samples change the outer-product action used to estimate total second-order energy. These tasks share a fixed target operator, while the computation of each product introduces new randomness.

Writing $H(Z)$ for a sampled matrix and b for the batch size gives a common positive semidefinite (PSD) target and response:

$$A = \mathbb{E}_Z H(Z) \succeq 0, \qquad \mathcal{O}_A(v) = \frac{1}{b}\sum_{\ell=1}^{b} H(Z_\ell)v. \tag{1}$$

Each application supplies a different H. Fresh independent batches give unbiased responses; a known uniform second-moment bound supplies the noise envelope for our accuracy guarantees. Unbiased response compression, such as applying the quantizer of Alistarh et al. (2017) to Av, provides another source of centered, direction-dependent noise. Beyond learning statistics, posterior covariance traces measure uncertainty in A-optimal experimental design (Herman et al., 2020); the same framework applies when their products are computed by fresh unbiased randomized solves with controlled second moments. Appendix G spells out these operator mappings and access requirements.

Trace estimators also randomize the query direction, introducing *probe variance* even with exact products. Low-rank methods reduce it by approximating the operator and sampling the residual trace. Hutch++ thereby improves the constant-confidence relative-error guarantee from $O(\varepsilon^{-2})$ to $O(\varepsilon^{-1})$ exact calls for PSD matrices (Meyer et al., 2021). With noisy products, the initial construction and

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.

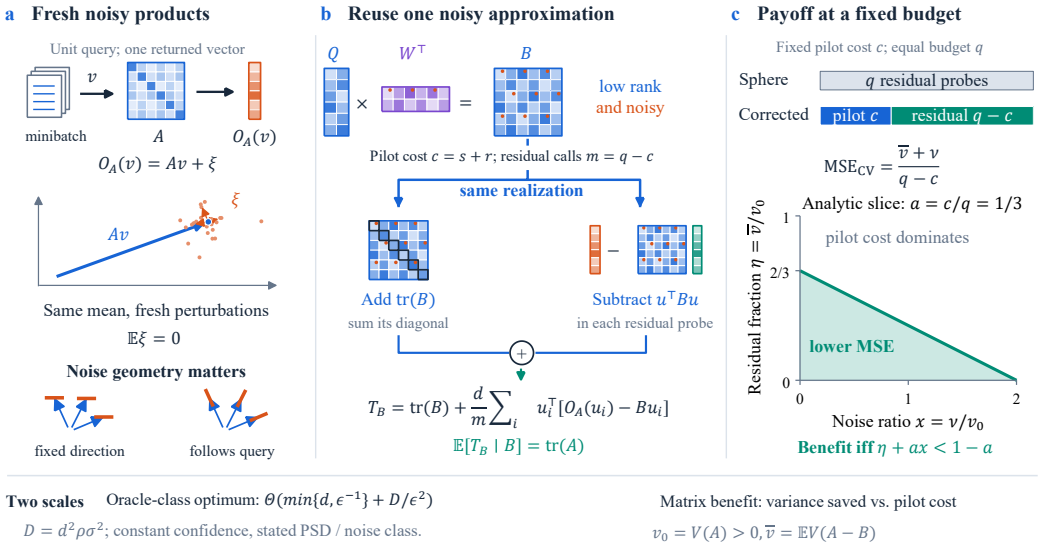


Figure 1: Noisy products, complete correction, and its benefit condition. A sketch of width s yields r basis vectors and a realized approximation B at cost $c = s + r$; the same B is added and subtracted in the estimator. For fixed c and $V(A) > 0$, correction helps when $\eta + ax < 1 - a$, with $a = c/q$, $\eta = \mathbb{E}V(A - B)/V(A)$ and $x = \nu/V(A)$. Panel (c) is the analytic slice $a = 1/3$; panels (a)–(b) are schematic.

measurement queries—the *pilot*—produce an uncertain subspace and trace contribution, or *head*. Measuring the head once can leave an error floor that additional residual probes never remove. The decision is how to spend a limited budget: build and measure a reusable approximation, or use those calls for independent averaging.

We answer two questions: how many noisy products guarantee an accurate trace estimate, and when does a low-rank pilot improve accuracy at a fixed budget? We establish the exact constant-confidence minimax query order for this noisy PSD oracle class, including adaptive algorithms (Theorem 1). For a particular operator and fixed pilot cost, an exact variance identity compares correction with plain sphere sampling, which averages independent random-direction quadratic forms. Correction helps precisely when the pilot removes enough probe variance to repay its query cost and the noise averaging forgone on those calls (Corollary 3).

Complete correction adds the trace of the realized noisy approximation B and subtracts that same B in every residual observation. Its conditional mean is correct, so pilot error contributes to a variance that decreases with further sampling. Sufficient averaging of a projected head also attains the optimal rate. Noise geometry determines the remaining cost: at equal energy, query-aligned noise contributes d times the sphere-observation variance of fixed-covariance noise. We test these operating conditions with fresh-minibatch neural-network curvature, paid pilot decisions, and a fresh-product XTrace comparison. Independent predictions and head and geometry controls connect their behavior to the variance identities; measured time and structural trace estimators expose costs beyond the query count. Figure 1 summarizes the correction and its operating condition.

2 The Query Cost of Fresh Oracle Noise

Let $d \geq 2$, let $A \in \mathbb{R}^{d\times d}$ be an unknown PSD matrix, and write $\tau = \mathrm{tr}(A) \in [1, 2]$. For $0 < \varepsilon < 1/8$, the target is

$$\Pr\{|\widehat{\tau} - \tau| \leq \varepsilon\tau\} \geq 1 - \delta. \tag{2}$$

A unit-vector query returns

$$\mathcal{O}_A(v) = Av + \xi(v; Z), \qquad \|v\|_2 = 1, \tag{3}$$

Table 1: Query orders at $\delta = 0.1$ for $d \geq \varepsilon^{-1}$ and trace in $[1, 2]$. The sphere column is a worst-case upper bound.

Noise envelope	Minimax query order	Sphere guarantee
$0 \leq D \leq \varepsilon$	$\Theta(\varepsilon^{-1})$	$O(\varepsilon^{-2})$
$\varepsilon < D < 1$	$\Theta(D\varepsilon^{-2})$	$O(\varepsilon^{-2})$
$D \geq 1$	$\Theta(D\varepsilon^{-2})$	$O(D\varepsilon^{-2})$

where each seed Z is independent of previous oracle seeds and algorithm randomness. The noise law may depend on A and v. For every unit vector,

$$\mathbb{E}_Z \xi(v; Z) = 0, \qquad \mathbb{E}_Z \left\|\xi(v; Z)\right\|_2^2 \leq \sigma^2, \qquad \mathbb{E}_Z (v^\top \xi(v; Z))^2 \leq \rho\sigma^2. \tag{4}$$

We use known envelopes $\sigma^2 \geq 0$, $1/d \leq \rho \leq 1$, and $D = d^2\rho\sigma^2$. The range of ρ defines the class and ensures $D \geq d\sigma^2$. These envelopes set certified accuracy schedules; fixed-budget estimators and the empirical policy in Section 5 do not require them as inputs. Queries may be adaptive; each returned vector counts as one call. The model permits exact local arithmetic.

Unit normalization fixes the scale at which noise is charged. A product with a nonzero vector z is implemented as $\|z\|_2 \mathcal{O}_A(z/\|z\|_2)$, so its noise energy is bounded by $\|z\|_2^2 \sigma^2$. Repeating a direction produces a new centered response and can reduce variance by averaging. The two moment constraints play separate roles: total energy controls how much a pilot subspace is perturbed, while directional variance controls the scalar information used to estimate the trace.

Theorem 1 (Constant-confidence minimax complexity)**.** *At $\delta = 0.1$, the minimax bounded query count over the matrix and oracle class above is*

$$\Theta\left(\min\{d, \varepsilon^{-1}\} + \frac{D}{\varepsilon^2}\right). \tag{5}$$

Complete correction and an appropriately averaged projected estimator, each with the coordinate fallback in Appendix A, attain this rate. The lower bound includes adaptive algorithms with unrestricted real-valued unit queries.

Why the bound is sharp. For a best rank-k approximation A_k, a noisy Gaussian pilot has expected squared error at most $2\|A - A_k\|_{\mathrm{F}}^2 + 2d\sigma^2$. Correcting it gives risk at most $(4\tau^2/k + 5D)/m$ with m residual queries. Choosing k of order ε^{-1}, or using coordinates when d is smaller, proves the upper bound. For the noise lower bound, the scalar matrices I/d and $(1 + 4\varepsilon)I/d$ with the Gaussian covariance in (20), for every allowed ρ, have conditional KL divergence $8\varepsilon^2/D$ per call. Adaptive testing therefore requires $\Omega(D\varepsilon^{-2})$ queries. A shifted-GOE construction gives the noiseless $\Omega(\min\{d, \varepsilon^{-1}\})$ term using the conditional-uncertainty method of Woodruff et al. (2022). Query schedules and lower bounds appear in Appendices A–B; Appendix C proves the pilot bound.

Independent median amplification multiplies the upper bound by $O(\log(1/\delta))$. The noise lower bound is $D\,\mathrm{kl}(1 - \delta, \delta)/(8\varepsilon^2)$, matching this confidence dependence for the noise term; the optimal noiseless dependence is left open. The trace interval fixes scale: for a known lower bound $\tau \geq t_0 > 0$, rescale the oracle by t_0 and replace σ^2 by σ^2/t_0^2.

Table 1 identifies when the $O(\varepsilon^{-1})$ exact-product scaling survives. A particular matrix can still favor or disfavor a pilot within each regime; the next section gives its exact variance comparison.

3 Turning a noisy pilot into variance reduction

We first analyze complete correction for any pilot matrix B with computable trace and products and $\mathbb{E}\|A - B\|_{\mathrm{F}}^2 < \infty$. This control-variate calculation separates the use of a noisy approximation from its construction. Section 3.2 then supplies a Gaussian pilot that attains the query bound.

3.1 COMPLETE CORRECTION AND ITS EXACT BENEFIT CONDITION

With independent uniform sphere probes $u_1, \dots, u_m$ and fresh oracle seeds, define

$$T_B = \mathrm{tr}(B) + \frac{d}{m} \sum_{i=1}^{m} u_i^\top \{\mathcal{O}_A(u_i) - Bu_i\}. \tag{6}$$

For a fixed matrix and noise law, write

$$\nu = d^2 \mathbb{E}_{u,Z}(u^\top \xi(u; Z))^2 \le D, \tag{7}$$

$$V(M) = \frac{2d}{d+2} \left\| \mathrm{sym}(M) - \frac{\mathrm{tr}(M)}{d} I \right\|_{\mathrm{F}}^2. \tag{8}$$

Here ν is the response-noise variance and $V(M)$ the probe variance of $du^\top Mu$. We use $\mathrm{sym}(M) = (M + M^\top)/2$, allowing nonsymmetric B.

Lemma 2 (Residual identity)**.** *Let the pilot and residual randomness be independent. Then*

$$\mathbb{E}[T_B \mid B] = \tau, \tag{9}$$

$$\mathbb{E}(T_B - \tau)^2 = \frac{\mathbb{E}V(A-B) + \nu}{m} \le \frac{2\mathbb{E}\,\|A-B\|_{\mathrm{F}}^2 + \nu}{m}. \tag{10}$$

Proof. Condition on the pilot transcript. Sphere second and fourth moments give mean $\mathrm{tr}(A-B)$ and variance $V(A-B)$ for $du^\top(A-B)u$. Conditional centering of the oracle error makes its covariance with that quadratic form zero. Independent residual observations then give (10); averaging over the pilot adds no variance because the conditional mean is always τ. □

This control-variate identity puts all pilot error into a variance that decays with m. Reusing the same realized B in both terms is essential to that cancellation.

Corollary 3 (Exact finite-budget benefit criterion)**.** *Let the pilot have deterministic cost $c < q$, put $m = q - c$, and write $v_0 = V(A)$ and $\bar{v} = \mathbb{E}V(A-B)$. At equal total query budget,*

$$\mathrm{MSE}_{\mathrm{CV}} = \frac{\bar{v} + \nu}{q - c}, \qquad \mathrm{MSE}_{\mathrm{sphere}} = \frac{v_0 + \nu}{q}. \tag{11}$$

If $v_0 > 0$, set $a = c/q$, $\eta = \bar{v}/v_0$ and $x = \nu/v_0$. Then

$$\frac{\mathrm{MSE}_{\mathrm{CV}}}{\mathrm{MSE}_{\mathrm{sphere}}} = \frac{\eta + x}{(1-a)(1+x)} < 1 \quad \Longleftrightarrow \quad \eta + ax < 1 - a. \tag{12}$$

The two MSE identities follow by applying Lemma 2 to B and to $B = 0$; rearranging gives the criterion. A concentrated spectrum can make η small, while noise can both degrade the pilot and increase x. Thus η must be evaluated at the actual noise law. For scalar A, $v_0 = 0$: at $\nu > 0$ any positive pilot cost gives an MSE ratio at least $q/(q-c) > 1$, since no probe variance can be removed.

For a fixed pilot law with $\eta < 1$, the same condition gives an explicit break-even budget:

$$q > \frac{c(1+x)}{1-\eta}. \tag{13}$$

The numerator charges for pilot calls and the noise averaging they replace; the denominator is the fraction of probe variance removed. Even $\eta = 0$ requires $q > c(1+x)$; if $\eta \ge 1$, no budget gives a strict improvement. This comparison keeps the pilot law fixed as q varies. Changing its width changes both c and η, so each allocation must be assessed through its own residual quality.

3.2 A GAUSSIAN PILOT THAT ATTAINS THE QUERY BOUND

Choose an integer $k \ge 1$ with $s = 2k + 2 \le d$. For independent Gaussian columns g_j, query $g_j / \|g_j\|_2$ and rescale the responses to obtain

$$Y = AG + E, \qquad e_j = \|g_j\|_2 \, \xi_j. \tag{14}$$

Algorithm 1: Complete correction with a noisy Gaussian pilot

Inputs: unit-vector oracle $\mathcal{O}_A$, width $s = 2k + 2 \le d$, residual count $m \ge 1$.

1. Draw independent $g_1, \dots, g_s \sim N(0, I_d)$ and obtain $y_j = \|g_j\|_2 \mathcal{O}_A(g_j / \|g_j\|_2)$.
2. Form an orthonormal basis $Q = [q_1, \dots, q_r]$ of the columns of $Y = [y_1, \dots, y_s]$.
3. With fresh oracle seeds, obtain $w_j = \mathcal{O}_A(q_j)$ for $j = 1, \dots, r$ and form $W = [w_1, \dots, w_r]$.
4. Draw m independent sphere probes u_i, independently of the pilot, and obtain $h_i = \mathcal{O}_A(u_i)$ with fresh seeds.
5. Return $\widehat{\tau} = \operatorname{tr}(W^\top Q) + \dfrac{d}{m} \displaystyle\sum_{i=1}^{m} \left[u_i^\top h_i - (Q^\top u_i)^\top (W^\top u_i)\right]$.

Query count: $s + r + m \le 2s + m$. For $r = 0$, the head and control terms are zero.

Let Q span $\operatorname{range}(Y)$ with $r \le s$ orthonormal columns. A fresh query per column gives $W = AQ + F$; set

$$B = QW^\top, \qquad \operatorname{tr}(B) = \operatorname{tr}(W^\top Q). \tag{15}$$

The pilot costs $c = s + r \le 2s$ calls and stores only Q, W; if $r = 0$, use $B = 0$.

Algorithm 1 fixes m before drawing the pilot. Choosing $m = q - 2s > 0$ guarantees at most q calls even when the sketch rank varies, and Lemma 2 applies to this fixed residual count.

Streaming the residuals requires $O(ds)$ storage. Each control term is $(Q^\top u_i)^\top (W^\top u_i)$; dense orthogonalization costs $O(ds^2)$ operations and all residual controls cost $O(dsm)$. These local costs accompany the query bound and matter when selecting a pilot width.

Lemma 4 (Pilot error)**.** *For a best rank-k approximation A_k, this pilot satisfies*

$$\begin{aligned} \mathbb{E}\,\|A - B\|_{\mathrm{F}}^2 &\le \left(1 + \frac{k}{k+1}\right) \|A - A_k\|_{\mathrm{F}}^2 + \left(\frac{dk}{k+1} + s\right) \sigma^2 \\ &\le 2\,\|A - A_k\|_{\mathrm{F}}^2 + 2d\sigma^2. \end{aligned} \tag{16}$$

The proof extends Gaussian range finding to normalized noisy queries. Normalization couples each column length to a pseudoinverse weight. For the leading k rows G_1 of G, the exact moment is

$$\mathbb{E} \sum_{j=1}^{s} \|g_j\|_2^2 \left\| \operatorname{row}_j(G_1^\dagger) \right\|_2^2 = \frac{dk}{k+1}. \tag{17}$$

Appendix C derives this moment and Lemma 4. The PSD tail bound $\|A - A_k\|_{\mathrm{F}}^2 \le \tau^2/k$ then gives the upper-bound construction described in Section 2; Appendix A supplies explicit schedules.

3.3 Averaging the projected trace

The alternative decomposition is $\tau = \operatorname{tr}(Q^\top A Q) + \operatorname{tr}((I-P)A(I-P))$, where $P = QQ^\top$. Average t fresh responses per column of Q into $\bar{W}$, and let $z_i = (I - P)u_i$. The projected estimator is

$$T_{\mathrm{proj}} = \operatorname{tr}(Q^\top \bar{W}) + \frac{d}{m} \sum_{i=1}^{m} z_i^\top [\|z_i\|_2 \, \mathcal{O}_A(z_i / \|z_i\|_2)]. \tag{18}$$

Zero projected vectors contribute zero without a call.

Proposition 5 (Averaging the head)**.** *For the Gaussian pilot, this estimator is unbiased and*

$$\mathbb{E}(T_{\mathrm{proj}} - \tau)^2 \le \frac{s\rho\sigma^2}{t} + \frac{4\,\|A - A_k\|_{\mathrm{F}}^2 + 2d\sigma^2 + D}{m}. \tag{19}$$

Suitable t, m, with the coordinate fallback when $s > d$, attain the rate in Theorem 1.

With one head query per column, an independent head error can persist. For $A = I/d$, noise $\xi(v) = \sigma Z v$ and a fixed rank-r pilot, the MSE tends to $r\sigma^2$ as $m \to \infty$. Complete correction or increasing head averaging removes this floor. Appendix A proves Proposition 5 and the exact finite-m expression (Proposition 9).

Table 2: Oracle-noise variance ν in one scaled sphere observation. Each law has total energy σ^2; w is a fixed unit vector and $z \sim N(0,1)$ is refreshed per call.

Noise law	Covariance	ν
Isotropic Gaussian	$\sigma^2 I/d$	$d\sigma^2$
Fixed direction, $\xi = \sigma z w$	$\sigma^2 w w^\top$	$d\sigma^2$
Query direction, $\xi = \sigma z v$	$\sigma^2 v v^\top$	$d^2\sigma^2$
Mixture in (20)	$\Sigma(v)$	$d^2\rho\sigma^2$

4 Noise geometry changes the optimal cost

For the lower-bound construction, each response has covariance

$$\Sigma(v) = \sigma^2 \left[\rho v v^\top + \frac{1-\rho}{d-1}(I - v v^\top) \right]. \tag{20}$$

Its directional variance attains $\rho\sigma^2$ at every query, giving $\nu = D$. Fixed covariance has a smaller average contribution to sphere probes.

Theorem 6 (Fixed-covariance refinement)**.** *If the noise covariance is a fixed Σ for every unit query and $\operatorname{tr}(\Sigma) \le \sigma^2$, then $\nu = d \operatorname{tr}(\Sigma)$. The constant-confidence minimax query complexity over all such fixed-covariance oracles is*

$$\Theta\left(\min\{d, \varepsilon^{-1}\} + \frac{d\sigma^2}{\varepsilon^2} \right). \tag{21}$$

The identity follows from $\mathbb{E} u u^\top = I/d$. Substituting it in the pilot analysis proves the upper bound; the isotropic Gaussian subfamily and the noiseless lower bound match it (Appendices A–B). This is a worst case over the covariance class; a specified covariance can allow better strategies.

Fixed rank-one and isotropic noise therefore have the same sphere-averaged cost at equal energy. The extra factor d arises when the noise direction follows the query.

Every query in (20) carries variance $\rho\sigma^2$ along itself, where the two scalar hard instances differ. Its conditional information remains unchanged after any transcript, giving the adaptive lower bound. Fixed covariance instead gives $\mathbb{E}_u u^\top \Sigma u = \operatorname{tr}(\Sigma)/d$. A sampled-matrix oracle additionally ties the noise $(H - A)v$ to a shared distribution of H. The GGN and Digits studies instantiate this structured access; the latter also admits the exact moments in Appendix F.

5 Experiments: accuracy, decisions, and computational cost

We test the benefit of a noisy pilot on learned curvature operators and on synthetic operators that isolate its operating conditions. The common comparison includes sphere and Rademacher sampling, complete correction, averaged projection, a paid diagnostic policy, and XTrace with fresh products. For every vector-oracle method, all returned vectors count toward the budget q. Correction uses $s = 2\lfloor q/12 \rfloor$ sketch columns; projected head averaging follows Appendix H. Neither these fixed-budget allocations nor the policy receives σ^2, ρ, or the true trace.

We report relative RMSE and gains over the full-budget sphere baseline. RMSE intervals are pointwise descriptive 95% bootstrap intervals. The complete protocols, empirical biases, uncertainty, and all tested configurations are retained in the appendix and supplement.

5.1 Fresh-minibatch neural-network curvature

We train an 81,290-parameter CNN on CIFAR-10 (Krizhevsky, 2009) and freeze it after epochs 1 and 5. Each target is the mean softmax-cross-entropy GGN over the same fixed pool of 256 test images:

$$A = \frac{1}{256} \sum_x J_x^\top [\operatorname{diag}(p_x) - p_x p_x^\top] J_x. \tag{22}$$

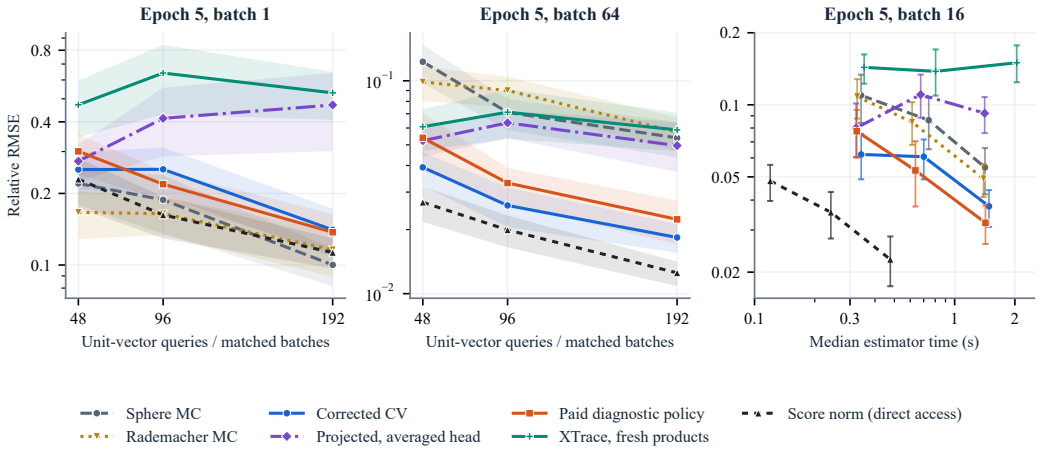


Figure 2: Fresh-minibatch GGN trace estimation at epoch 5. Left and center: relative RMSE at two batch sizes; every vector method pays for all q calls. Right: error versus median synchronized estimator time at $b = 16$, including products and local algebra. Score norm uses direct model/sample access and q matched batch units, with zero matvec calls. Each point uses 50 trials; bands and vertical bars show 95% RMSE bootstrap intervals. Both checkpoints and all batch sizes appear in Appendix J.

Here J_x is the parameter Jacobian of the logits and p_x the model probabilities. Every product draws a new batch of $b \in \{1, 16, 64\}$ images with replacement and uses automatic differentiation, without forming A. Minibatch curvature noise is also studied by Dangel et al. (2023). We use $q \in \{48, 96, 192\}$, 50 trials per setting, and separately computed float64 trace references (Appendix J).

At epoch 5 and $q = 192$, correction loses at $b = 1$: its RMSE is 0.141 versus sphere's 0.100. At $b = 64$, these become 0.0184 and 0.0539, a gain of 2.94 with interval $[2.29, 3.74]$. The corresponding epoch-1 gain is 3.48 $[2.72, 4.39]$. Thus improved product quality reveals a correction benefit for both learned operators (Figure 2). The fixed two-response projected head and fresh-product XTrace do not match correction's improvement as q grows; their full curves are retained in Figure 16.

The model also permits a different estimator: sample a label from p_x and average $\|\nabla_\theta \log p_\theta(y \mid x)\|_2^2$. Its expectation is the same GGN trace (Martens, 2020). This *score-norm* comparator accesses samples and model gradients directly. Both it and a q-query vector estimator process bq examples; their measured costs distinguish these access models. At epoch 5, $b = 16, q = 192$, correction reaches RMSE 0.0377 in a median 1.48 seconds, while score norm reaches 0.0226 in 0.477 seconds. XTrace reaches 0.150 in 2.04 seconds. The query comparison therefore coexists with a cheaper option when model gradients are directly accessible.

5.2 Testing the benefit identity

For $d = 128$, $\tau = 1$, $q = 192$, and $s = 32$, we sweep seven isotropic noise levels from $\nu = 0$ to 10 on power, spiked, and flat spectra. Each point uses 200 trials over five rotations. A separate set of 500 pilots per configuration estimates $\bar{v} = \mathbb{E}V(A - B)$, predicting corrected MSE through (11). Across the 20 defined configurations, the largest relative discrepancy between predicted and measured point estimates is 19.2% (Figure 3).

These independent diagnostic calls test the identity and are outside the estimator budgets. They are not supplied to any method. The prediction captures both ways noise erodes the benefit: larger residual response variance and poorer pilots. The flat control confirms that there is no probe variance for a pilot to remove. The next study charges for the information used to decide whether to use a pilot.

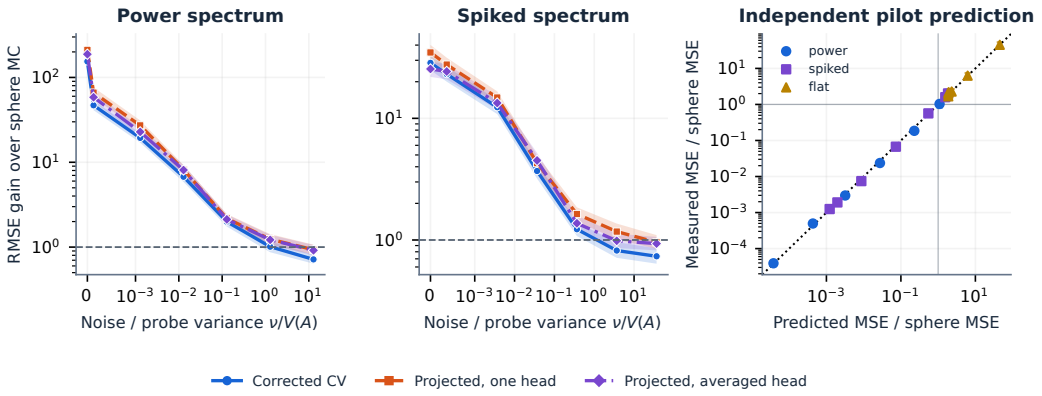


Figure 3: Finite-budget crossover at $q = 192$. Left and center: RMSE gain over sphere sampling, with 95% bootstrap bands. Right: MSE predicted from 500 separate pilots versus measured MSE from 200 estimator trials, each normalized by exact sphere MSE; bars show uncertainty on both axes. The flat/noiseless ratio is undefined and omitted.

5.3 Paying for decisions and comparing sample reuse

Our paid policy constructs B, then uses $h = \max\{4, \lfloor q/12 \rfloor\}$ independent calls to compare the empirical variances of plain and corrected scalar observations from the same responses. It freezes the choice before using the remaining $q - s - r - h$ fresh calls. Both outcomes are conditionally unbiased; choosing plain production retains the cost already spent on the pilot and diagnostics (Appendix I.1).

The paid diagnostic estimates $S_0 = V(A) + \nu$ and $S_B(B) = V(A - B) + \nu$ directly from paired responses. It needs no separate estimate of the worst-direction envelope ρ. Given the remaining m calls, the ideal production decision compares $S_B(B)/m$ with S_0/m; the full-budget sphere comparator has risk S_0/q. Corollary 3 assesses the investment before paying for a pilot of fixed cost. Our paid policy measures the resulting accuracy after both pilot and diagnostic costs are charged.

In the six-method comparison at $q = 192$, isotropic $\nu = 0.01$ gives policy gains of 5.78 on the power spectrum and 3.37 on the spiked spectrum. At $\nu = 1$, they fall to 0.77 and 0.74. On the flat spectrum at $\nu = 0.01$, the policy chooses plain production in 92.0% of runs yet achieves only 0.64 gain: recognizing an unhelpful pilot does not recover its cost. Increasing h from 8 to 32 does not yield a stable accuracy improvement (Appendix I.3).

XTrace (Epperly et al., 2024) uses leave-one-out subspaces and a second shared batch of fresh products. Appendix I.2 gives its implementation and centering argument. It substantially outperforms fixed correction with exact products: on the power spectrum, their gains are approximately $1{,}698$ and 170. At $\nu = 0.01$, these become 5.56 and 5.38; at $\nu = 10$, both lose to sphere sampling. Sample reuse changes the finite-budget comparison, while fresh noise can erase its exact-product advantage.

5.4 Mechanisms and accessible structure

For $A = I/32$, width eight, and query-coupled $\sigma = 0.02$, the one-head RMSE reaches 0.0539 at $q = 8192$, matching the predicted floor 0.0566. Correction and averaged projection reach 0.00721 and 0.00728 (Figure 5). At equal noise energy, $d = 256$, and $q = 128$, $q\,\mathrm{MSE}/\sigma^2$ is 264, 245, and $65{,}729$ for isotropic, fixed-direction, and query-coupled noise, close to the predicted 256, 256, and $65{,}536$ (Figure 6).

On the Digits covariance operator (Figure 7), correction at $q = 192$ changes from RMSE 0.114 versus sphere's 0.106 at $b = 1$ to 0.0142 versus 0.0645 at $b = 64$. Yet batch-one responses also reveal individual sample norms (Lemma 15), giving RMSE 0.0103; direct row access at $b = 64$ gives 0.00136. Product quality, sample work, and available structure therefore belong together in an estimator comparison.

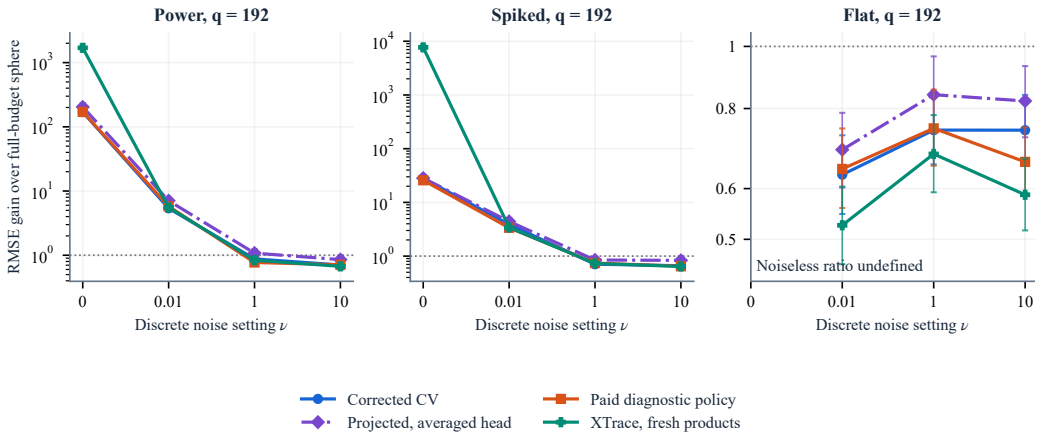


Figure 4: Full-budget comparison with paid decisions and XTrace, at $d = 128, q = 192$. The policy's pilot, diagnostics, and production share the 192 calls. XTrace uses two batches of 96 fresh products. Gains above one improve on sphere sampling. The horizontal axis lists discrete isotropic noise settings; error bars are 95% bootstrap intervals over 200 trials. A flat/noiseless ratio is undefined.

6 Related work

Hutch++ combines low-rank approximation with residual trace estimation (Meyer et al., 2021). Adaptive allocation and Nyström++ (Persson et al., 2022), and leave-one-out sample reuse in XTrace and XNysTrace (Epperly et al., 2024), improve exact-product efficiency. Liu et al. (2025) study Hutch++ for generative-model objectives.

The noiseless query term and Gaussian range-finding framework come from established trace-estimation and randomized approximation methods (Meyer et al., 2021; Halko et al., 2011; Woodruff et al., 2022). Related analyses address deterministic approximation perturbations (Chen et al., 2025) and implicit-product accuracy for Hutchinson estimation (Chen, 2016).

7 Discussion

A noisy low-rank approximation must repay its construction cost and the noise averaging displaced by those queries. Complete correction uses the same realized approximation in both terms, so pilot error affects variance that production queries reduce. The head-floor experiment shows why: additional residual probes cannot repair an independently frozen head error. Correction and head averaging both restore improvement with further sampling.

The investment decision also precedes the production decision. On the flat spectrum, paid diagnostics usually select plain production, yet still lose to full-budget averaging after paying for the pilot. Diagnostics consume observations that could reduce final error. This motivates allocation rules that jointly choose pilot and diagnostic effort, with independent production preserving conditional unbiasedness. Certified stopping requires tail or moment control beyond empirical variance (Appendix D).

Noise geometry determines how expensive this tradeoff becomes. Fixed-direction and isotropic noise have the same sphere-averaged cost at equal energy; following the query raises it by a factor of d. Product-error energy alone therefore misses a determinant of trace accuracy. The curvature crossover shows that improved product quality can make correction worthwhile, while larger minibatches also consume more examples.

Practical selection should start with accessible structure and the resource being budgeted. Model-score access provides a cheaper alternative in the measured curvature comparison; batch-one covariance responses themselves reveal sample norms. These cases motivate comparisons at matched sample or time budgets as well as query budgets.

AI USE STATEMENT

Generative AI tools assisted with mathematical derivations and proof exposition, literature search, experiment design and analysis, implementation, figure preparation, and manuscript revision. Verification included mathematical audits, tests of exact identities and moment formulas, and reproduction checks of recorded experiments. The authors take responsibility for the final text, claims, proofs, code, and figures.

REPRODUCIBILITY STATEMENT

The supplement contains source code, 204,100 retained estimator trials, independent pilot diagnostics, frozen CNN checkpoints and scoring references, dependency versions, and seed and checksum manifests. Appendix H specifies the experimental protocols and information available to each estimator; Appendices A–F contain the proofs and exact moment calculations. Appendices I–J describe the paid decisions, fresh-product XTrace comparison, and neural-network curvature protocols. Empirical figures and tables can be reconstructed from the supplied records.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Z. Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6c340f25839e6acdc73414517203f5f0-Abstract.html.

R. R. Bahadur and Leonard J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956. doi: 10.1214/aoms/1177728077. URL https://www.jstor.org/stable/2237199.

Jie Chen. How accurately should I compute implicit matrix-vector products when applying the Hutchinson trace estimator? *SIAM Journal on Scientific Computing*, 38(6):A3515–A3539, 2016. doi: 10.1137/15M1051506. URL https://jiechenjiechen.github.io/pub/hutchinson.pdf.

Tyler Chen, Feyza Duman Keles, Diana Halikias, Cameron Musco, Christopher Musco, and David Persson. Near-optimal hierarchical matrix approximation from matrix-vector products. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2025. URL https://arxiv.org/abs/2407.04686.

Felix Dangel, Lukas Tatzel, and Philipp Hennig. ViViT: Curvature access through the generalized gauss–newton's low-rank structure. *Transactions on Machine Learning Research*, February 2023. URL https://openreview.net/forum?id=DzJ7JfPXkE.

Ethan N. Epperly, Joel A. Tropp, and Robert J. Webber. XTrace: Making the most of every sample in stochastic trace estimation. *SIAM Journal on Matrix Analysis and Applications*, 45(1):1–23, 2024. doi: 10.1137/23M1548323. URL https://arxiv.org/abs/2301.07825.

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. URL https://arxiv.org/abs/0909.4061.

Elizabeth Herman, Alen Alexanderian, and Arvind K. Saibaba. Randomization and reweighted ℓ_1-minimization for A-optimal design of linear inverse problems. *SIAM Journal on Scientific Computing*, 42(3):A1714–A1740, 2020. doi: 10.1137/19M1267362. URL https://arxiv.org/abs/1906.03791.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with Jacobian regularization. arXiv:1908.02729, 2019. URL https://arxiv.org/abs/1908.02729.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Xinyang Liu, Hengrong Du, Wei Deng, and Ruqi Zhang. Optimal stochastic trace estimation in generative modeling. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4600–4608, 2025. URL https://proceedings.mlr.press/v258/liu25k.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://www.jmlr.org/papers/v21/17-678.html.

Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, 2021. URL https://arxiv.org/abs/2010.09649.

David Persson, Alice Cortinovis, and Daniel Kressner. Improved variants of the Hutch++ algorithm for trace estimation. *SIAM Journal on Matrix Analysis and Applications*, 43(3):1162–1185, 2022. doi: 10.1137/21M1447623. URL https://arxiv.org/abs/2109.10659.

Andres Potapczynski, Luhuan Wu, Dan Biderman, Geoff Pleiss, and John P. Cunningham. Bias-free scalable Gaussian processes via randomized truncations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8609–8619, 2021. URL https://proceedings.mlr.press/v139/potapczynski21a.html.

Frank Schneider, Felix Dangel, and Philipp Hennig. Cockpit: A practical debugging tool for the training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ae3539867aaeec609a4260c6feb725f4-Abstract.html.

scikit-learn developers. The Digits dataset: load_digits. scikit-learn documentation, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html. Accessed September 26, 2026.

David P. Woodruff, Fred Zhang, and Qiuyi Zhang. Optimal query complexities for dynamic trace estimation. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.15219.

A QUERY BOUNDS AND PROJECTED-ESTIMATOR PROOFS

This appendix gives the component bounds and explicit query schedules used in Theorem 1.

A.1 UPPER BOUND AND QUERY SCHEDULE

Theorem 7 (Query upper bound). *Under* (3)–(4)*, an estimator satisfies* (2) *with* $\delta = 0.1$ *using*

$$O\left(\min\{d, \varepsilon^{-1}\} + \frac{D}{\varepsilon^2}\right) \tag{23}$$

queries. Complete control-variate correction, defined in Section 3, and an appropriately averaged projected estimator each attain this bound with the coordinate fallback below. Independent median amplification gives an upper bound equal to (23) *times* $O(\log(1/\delta))$.

Proof. For PSD matrices, $\|A - A_k\|_{\mathrm{F}}^2 \le \tau^2/k$. Combining Lemmas 2 and 4 and $D \ge d\sigma^2$ gives

$$\frac{\mathbb{E}(T_B - \tau)^2}{\tau^2} \le \frac{4/k + 5D}{m}. \tag{24}$$

Take $k = \lceil \varepsilon^{-1} \rceil$. If $d \ge 2k + 2$, use the pilot above and $m = \lceil 40/(k\varepsilon^2) + 50D/\varepsilon^2 \rceil$. Chebyshev's inequality gives failure probability at most 0.1, with the query count in Theorem 7.

If $d < 2k + 2$, query each coordinate $r_0 = \max\{1, \lceil 10d\rho\sigma^2/\varepsilon^2 \rceil\}$ times and sum the averaged diagonal responses. Its variance is at most $d\rho\sigma^2/r_0 \le \varepsilon^2/10$ and its cost is at most $d + 10D/\varepsilon^2$. In this branch, $d = O(\min\{d, \varepsilon^{-1}\})$. These sufficient constants establish the query order. Experiments use a shared fixed-budget allocation, examined through Corollary 3.

Setting $B = 0$ in Lemma 2 gives the ordinary sphere bound $O((1 + D)/\varepsilon^2)$ in Table 1. □

A.2 ADAPTIVE NOISE LOWER BOUND

Theorem 8 (Adaptive noise lower bound). *For* $\sigma > 0$*, any randomized algorithm using at most* q *queries and satisfying* (2) *uniformly over the model class must obey*

$$q \ge \frac{D}{8\varepsilon^2}\ \mathrm{kl}(1 - \delta, \delta), \qquad 0 < \delta < 1/2, \tag{25}$$

where $\mathrm{kl}(a, b) = a\log(a/b) + (1 - a)\log((1 - a)/(1 - b))$. *In particular,* $q = \Omega(D\varepsilon^{-2}\log(1/\delta))$ *for* $\delta \le 0.1$.

Proof. Consider $A_0 = I/d$ and $A_1 = (1 + 4\varepsilon)I/d$. At each unit query v, add an independent Gaussian vector with covariance (20). Both noise constraints hold with equality. The means differ by $4\varepsilon v/d$, so the conditional Gaussian KL divergence is $8\varepsilon^2/D$. At $\rho = 1$, both distributions have the same one-dimensional support and the same formula holds on that support.

Include the algorithm's random seed in its transcript. Given a transcript prefix, its rule for choosing the next query is the same under both hypotheses, so the rule contributes no additional conditional KL. The chain rule gives $\mathrm{KL}(P_0\|P_1) \le 8q\varepsilon^2/D$. A bounded stopping rule can be padded to q calls.

The success intervals $[1 - \varepsilon, 1 + \varepsilon]$ and $[(1 + 4\varepsilon)(1 - \varepsilon), (1 + 4\varepsilon)(1 + \varepsilon)]$ are disjoint. Thresholding a successful trace estimate therefore defines a test with both error probabilities at most δ. Binary data processing yields $\mathrm{KL}(P_0\|P_1) \ge \mathrm{kl}(1 - \delta, \delta)$, proving (25). □

A.3 PROJECTED HEAD AVERAGING AND ITS ERROR FLOOR

Proof of Proposition 5. Condition on Q. The head has the correct mean and variance at most $r\rho\sigma^2/t$. It is independent of the residual. Projected noise contributes at most $D\mathbb{E}\,\|(I - P)u\|_2^4 \le D$ per scaled sample. The quadratic-form contribution is bounded by $2\,\|(I - P)A(I - P)\|_{\mathrm{F}}^2 \le 2\,\|(I - P)A\|_{\mathrm{F}}^2$. The projection part of Lemma 4 bounds its expectation by $2\,\|A - A_k\|_{\mathrm{F}}^2 + d\sigma^2$.

Take $t = \max\{1, \lceil 20s\rho\sigma^2/\varepsilon^2 \rceil\}$ and $m = \lceil 80/(k\varepsilon^2) + 60D/\varepsilon^2 \rceil$. Each term in (19), after division by τ^2, is at most $\varepsilon^2/20$. The cost satisfies $s + st + m \le 2s + 20s^2\rho\sigma^2/\varepsilon^2 + m \le 2s + 20D/\varepsilon^2 + m$. Use the coordinate branch if $s > d$. □

Proposition 9 (Unaveraged head error). *Let $A = I/d$ and $\xi(v; Z) = \sigma Z v$ with independent standard Gaussian Z. Fix a pilot subspace of rank $r \in \{1, \dots, d-1\}$. With one fresh head query per column and m independent residual queries,* (18) *has MSE*

$$r\sigma^2 + \frac{1}{m}\left[\frac{2r(d-r)}{d^2(d+2)} + \frac{\sigma^2 d(d-r)(d-r+2)}{d+2}\right]. \tag{26}$$

Consequently its MSE tends to $r\sigma^2$, whereas (6) *with a fixed finite-energy pilot tends to zero as $m \to \infty$.*

Proof. Each head observation is $1/d + \sigma Z$, giving variance $r\sigma^2$. Write $L = \|(I-P)u\|_2^2$. Its first two moments are $(d-r)/d$ and $(d-r)(d-r+2)/[d(d+2)]$, so $\mathrm{Var}(L) = 2r(d-r)/[d^2(d+2)]$. A scaled residual observation is $L + d\sigma ZL$. Its two terms are uncorrelated and independent of the head. Summing the variances proves the formula. Lemma 2 proves the final assertion. □

A.4 Fixed-covariance query complexity

Proof of Theorem 6. The sphere second moment implies $\nu = d^2 \mathbb{E}_u u^\top \Sigma u = d\,\mathrm{tr}(\Sigma)$. Lemma 4 still applies, giving (24) with D replaced by $d\sigma^2$. In the coordinate branch, the variance of the summed estimate is $\mathrm{tr}(\Sigma)/r_0$. Taking $r_0 = \max\{1, \lceil 10\sigma^2/\varepsilon^2 \rceil\}$ gives (21). For the noise lower bound, apply Theorem 8 to its isotropic construction with $\rho = 1/d$ and $\Sigma = \sigma^2 I/d$, which also satisfies every stated $\rho \geq 1/d$. Lemma 11 supplies the remaining noiseless term. □

B A noiseless adaptive lower bound in the stated PSD class

We give the precise finite-dimensional lower bound used in Theorem 1. The underlying Gaussian observation principle is standard in adaptive matrix-query lower bounds (Woodruff et al., 2022); the shift and constants below ensure PSD matrices and the trace interval without normalizing by an unknown random trace.

Lemma 10 (Unobserved Gaussian block). *Let $G \in \mathbb{R}^{n \times n}$ be a Gaussian orthogonal ensemble matrix: its upper-triangular entries are independent, $G_{ii} \sim N(0, 2)$, and $G_{ij} \sim N(0, 1)$ for $i < j$. After $q < n$ adaptive exact matrix–vector queries, one may augment the observations, if necessary, so that in an observation-measurable orthonormal basis,*

$$G = \begin{bmatrix} K & C^\top \\ C & G_{\mathrm{rem}} \end{bmatrix},$$

where K, C are known and the conditional law of G_{rem} is an independent GOE matrix of size $n - q$. In particular, conditional on the augmented transcript, the unknown contribution to $\mathrm{tr}(G)$ is $N(0, 2(n-q))$.

Proof. Condition on the algorithm's private randomness. Given previous products, the component of a new query in the span of earlier query directions gives an already known product. Subtract that component. If the residual is nonzero, normalize it to obtain a new orthogonal query direction; exact products allow this transformation without another observation. If it is zero, augment the transcript with a product in any measurable orthogonal direction. This can only help the algorithm.

For the first normalized direction, orthogonal invariance of the GOE makes its coordinates those of the first basis vector. Observing its product reveals the first row and column. In that basis, the remaining principal block is independent GOE, since the upper-triangular entries are independent. Inductively, a new direction in the unqueried complement is a function only of already observed entries. Conditional on them it is fixed; rotating the independent GOE block to this direction preserves its law. Revealing its first row and column leaves another independent GOE block. After q such steps the stated decomposition follows. The trace of the remaining block is the sum of $n - q$ independent $N(0, 2)$ entries. This argument allows arbitrary measurable adaptive query rules. □

Lemma 11 (PSD noiseless query lower bound). *For $d \geq 2$ and $0 < \varepsilon < 1/8$, any algorithm that estimates every PSD trace in $[1, 2]$ to relative error ε with probability at least 0.9 requires $\Omega(\min\{d, \varepsilon^{-1}\})$ exact vector queries, even with unrestricted adaptive real-valued queries.*

Proof. First suppose $\varepsilon \le 1/200$ and set $n = \min\{d, \lfloor 1/(100\varepsilon) \rfloor\} \ge 2$. Draw a GOE matrix G of size n and define

$$M = \frac{24\sqrt{n}\, I_n + G}{16n\sqrt{n}}, \qquad A = \mathrm{diag}(M, 0_{d-n}). \tag{27}$$

The active block may be disclosed to the algorithm. A product with A is simulated by one product with G after subtracting the known shift and scaling. Arbitrary components outside the active block reveal no extra information.

Let $\mathcal{E} = \{\|G\|_{\mathrm{op}} \le 8\sqrt{n}\}$. On $\mathcal{E}$, the eigenvalues of M are in $[1/n, 2/n]$, so A is PSD and $\mathrm{tr}(A) \in [1, 2]$. We need a concrete lower bound on $\Pr(\mathcal{E})$ rather than assume every shifted Gaussian matrix is PSD. A $1/4$-net of the unit sphere can be chosen with at most 9^n points by the usual disjoint-ball volume argument. For a symmetric matrix, its operator norm is at most twice the largest absolute quadratic form over this net: approximate a maximizing unit vector and bound the two approximation terms by $2(1/4)\,\|G\|_{\mathrm{op}}$. At each fixed unit vector x, $x^\top G x \sim N(0, 2)$. A Gaussian tail bound and a union bound give

$$\Pr(\mathcal{E}^c) \le 2 \cdot 9^n e^{-4n} = 2e^{-(4-\log 9)n} < 0.055, \qquad n \ge 2. \tag{28}$$

Suppose $q \le n/2$. By Lemma 10, conditional on the augmented transcript, $\mathrm{tr}(A)$ is Gaussian with a known mean and standard deviation

$$s_q = \frac{\sqrt{2(n-q)}}{16n\sqrt{n}} \ge \frac{1}{16n}.$$

Any interval of length 4ε therefore has conditional probability at most its length times the maximum Gaussian density:

$$\Pr\{|\widehat{\tau} - \mathrm{tr}(A)| \le 2\varepsilon \mid \text{transcript}\} \le \frac{64\varepsilon n}{\sqrt{2\pi}} \le 0.256. \tag{29}$$

This also holds after averaging the transcript and the private randomness. In contrast, a uniformly 0.9-successful estimator on the valid PSD class would have probability at least $0.9\Pr(\mathcal{E}) > 0.850$ of both $\mathcal{E}$ and relative success. On that event, relative success implies absolute error at most 2ε. This contradicts (29). Notice that we never condition the Gaussian posterior on $\mathcal{E}$; the contradiction compares unconditional probabilities, avoiding a truncated-Gaussian posterior assumption.

It follows that $q > n/2$. Since $\lfloor 1/(100\varepsilon) \rfloor \ge 1/(200\varepsilon)$, this implies $q \ge (1/400)\min\{d, \varepsilon^{-1}\}$ up to harmless strict integer rounding. If $\varepsilon > 1/200$, zero queries cannot distinguish the disjoint relative-success intervals for I/d and $2I/d$ with success 0.9 on both. Thus $q \ge 1$, which gives the same absolute-constant order because $\min\{d, \varepsilon^{-1}\} < 200$. □

Combining Theorem 7 with Theorem 8 and Lemma 11, using $\max\{a, b\} \ge (a+b)/2$, proves Theorem 1. The same combination with the isotropic noise subfamily gives the lower bound in Theorem 6.

C Noisy pilot proof and weighted Gaussian inverse moment

Work in an eigenbasis of A, and partition $A = \mathrm{diag}(\Lambda_1, \Lambda_2)$ and $G = [G_1^\top, G_2^\top]^\top$, with $G_1 \in \mathbb{R}^{k \times s}$. Since $s > k+1$, G_1 has full row rank almost surely. Let $P = QQ^\top$. The matrix $C = YG_1^\dagger[I_k \;\; 0]$ has its columns in $\mathrm{range}(Y)$, hence

$$\|(I-P)A\|_{\mathrm{F}}^2 \le \|A - C\|_{\mathrm{F}}^2 = \|\Lambda_2\|_{\mathrm{F}}^2 + \left\| \begin{bmatrix} 0 \\ \Lambda_2 G_2 G_1^\dagger \end{bmatrix} + EG_1^\dagger \right\|_{\mathrm{F}}^2. \tag{30}$$

Conditional on all of G, the columns e_j are independent and centered, with $\mathbb{E}[\|e_j\|_2^2 \mid G] \le \sigma^2 \|g_j\|_2^2$. Thus the cross term with the noiseless part vanishes in conditional expectation, and

$$\mathbb{E}[\left\| EG_1^\dagger \right\|_{\mathrm{F}}^2 \mid G] \le \sigma^2 \sum_{j=1}^{s} \|g_j\|_2^2 \left\| \mathrm{row}_j(G_1^\dagger) \right\|_2^2. \tag{31}$$

The standard inverse moment is $\mathbb{E}\left\|G_1^\dagger\right\|_{\mathrm{F}}^2 = k/(s-k-1)$ (Halko et al., 2011). One derivation uses the diagonal entries of $(G_1 G_1^\top)^{-1}$: each is the reciprocal of a Gaussian row's squared distance from the span of the other rows. That squared distance has a χ^2_{s-k+1} law, whose reciprocal mean is $1/(s-k-1)$. Independence of G_2 and G_1 also gives

$$\mathbb{E}\left\|\Lambda_2 G_2 G_1^\dagger\right\|_{\mathrm{F}}^2 = \frac{k}{s-k-1}\,\|\Lambda_2\|_{\mathrm{F}}^2\,. \tag{32}$$

Lemma 12 (General weighted inverse moment)**.** *For $s > k+1$ and $ks > 2$,*

$$\beta_{d,k,s} := \mathbb{E}\sum_{j=1}^{s}\|g_j\|_2^2\left\|\mathrm{row}_j(G_1^\dagger)\right\|_2^2 = \frac{(d-k)k}{s-k-1} + \frac{k(ks-2)/(s-k-1)+2k}{s+2}. \tag{33}$$

Proof. The last $d-k$ coordinates of g_j are independent of G_1. Their contribution is $(d-k)k/(s-k-1)$. For the other coordinates, write the polar decomposition $G_1 = M^{1/2}H$, where $M = G_1 G_1^\top$ and $HH^\top = I_k$. Gaussian orthogonal invariance implies that H is uniformly distributed on the row-Stiefel manifold independently of M. Each column h_j has the distribution of the first k coordinates of a uniform sphere vector in $\mathbb{R}^s$.

For symmetric $U, V \in \mathbb{R}^{k\times k}$, sphere fourth moments give

$$\mathbb{E}_H\sum_{j=1}^{s}(h_j^\top U h_j)(h_j^\top V h_j) = \frac{\mathrm{tr}(U)\,\mathrm{tr}(V) + 2\,\mathrm{tr}(UV)}{s+2}. \tag{34}$$

Since $G_1^\dagger = H^\top M^{-1/2}$, set $U = M, V = M^{-1}$. The remaining contribution is $[\mathbb{E}\{\mathrm{tr}(M)\,\mathrm{tr}(M^{-1})\} + 2k]/(s+2)$.

To evaluate the expectation, let $T = \mathrm{tr}(M) = \|G_1\|_{\mathrm{F}}^2 \sim \chi^2_{ks}$. The radial variable T is independent of M/T. If $Z_0 = \mathrm{tr}((M/T)^{-1})$, then $\mathrm{tr}(M^{-1}) = Z_0/T$. Because $\mathbb{E}T^{-1} = 1/(ks-2)$,

$$\mathbb{E}Z_0 = (ks-2)\mathbb{E}\,\mathrm{tr}(M^{-1}) = \frac{k(ks-2)}{s-k-1}.$$

But $\mathrm{tr}(M)\,\mathrm{tr}(M^{-1}) = Z_0$. Substitution proves (33). At $s = 2k+2$ it simplifies to $dk/(k+1)$. □

Using (30)–(33) with $s = 2k+2$ gives

$$\mathbb{E}\,\|(I-P)A\|_{\mathrm{F}}^2 \le \left(1 + \frac{k}{k+1}\right)\|A - A_k\|_{\mathrm{F}}^2 + \frac{dk}{k+1}\sigma^2. \tag{35}$$

Because A is symmetric, $B = Q(AQ+F)^\top = PA + QF^\top$. The terms $(I-P)A$ and $QF^\top$ are Frobenius-orthogonal pointwise, since $(I-P)Q = 0$. Therefore

$$\|A-B\|_{\mathrm{F}}^2 = \|(I-P)A\|_{\mathrm{F}}^2 + \|F\|_{\mathrm{F}}^2\,.$$

The fresh second batch has $\mathbb{E}[\|F\|_{\mathrm{F}}^2 \mid Q] \le r\sigma^2 \le s\sigma^2$, proving Lemma 4. This argument also covers rank-deficient A and rank-deficient Y.

D Causal noise and unknown-variance proofs

The memoryless model is convenient but not necessary for the constant-confidence rate. Suppose instead that, at time t, the oracle draws from a causal kernel depending on A, the observed history $\mathcal{H}_{t-1}$, and the current query v_t. For every possible history and unit query, assume

$$\mathbb{E}[\xi_t \mid \mathcal{H}_{t-1}, v_t] = 0, \quad \mathbb{E}[\|\xi_t\|_2^2 \mid \mathcal{H}_{t-1}, v_t] \le \sigma^2, \quad \mathbb{E}[(v_t^\top \xi_t)^2 \mid \mathcal{H}_{t-1}, v_t] \le \rho\sigma^2. \tag{36}$$

The kernel has no access to unexposed algorithm randomness; formally it can be sampled using a fresh seed independent of all algorithm seeds. Its distribution and covariance may vary with the history.

Proposition 13 (Causal extension)**.** *The upper bound and constant-confidence minimax rate of Theorems 7 and 1 remain valid under* (36)*. The exact memoryless variance identity need not hold with a single fixed ν, but its upper bound with D does.*

The proof below replaces independence of noise terms by conditional centering and martingale orthogonality. This extension does not allow arbitrary correlated or biased noise.

Conversely, removing every known quantitative noise bound is a substantial change. The following statement is an oracle specialization of the classical rare-contamination obstruction in nonparametric inference (Bahadur & Savage, 1956).

Proposition 14 (No unrestricted variance-adaptive stopping)**.** *Fix $d \geq 2$, $0 < \varepsilon < 1/8$, and $\delta = 0.1$. Over the union of the model classes with all finite σ^2 and $\rho = 1$, there is no algorithm that both stops almost surely on every instance and satisfies* (2) *uniformly, if no noise upper bound or additional distributional restriction is supplied.*

The proof below establishes the claim using a rare response that changes the oracle mean while remaining invisible for arbitrarily many typical calls. Thus unknown-variance stopping cannot be justified merely by replacing σ^2 with an empirical estimate. A useful adaptive theorem must add assumptions, for example a known tail or moment envelope.

D.1 Proofs

Proof of Proposition 13. For the pilot, reveal the entire Gaussian matrix G for purposes of analysis, but not to the oracle. Because the oracle kernel is causal and uses fresh exogenous seeds, its conditional mean and moment bounds remain valid after conditioning on G and the preceding history. The rescaled errors e_j are conditionally centered. For $i < j$, condition on G and the history just before call j; e_i is measurable and $\mathbb{E}[e_j \mid G, \text{history}] = 0$. The weighted cross terms in $\left\| EG_1^\dagger \right\|_{\mathrm{F}}^2$ therefore have zero expectation. Its diagonal terms obey (31). The cross term with the noiseless sketch is also centered. All weighted inverse and projection bounds are unchanged.

The second pilot stage needs only the sum of the squared column norms, bounded by $s\sigma^2$. For complete correction, condition on the pilot and generate each residual sphere probe independently of the current history. The centered residual observations form martingale differences. Their conditional second moments are bounded by $2\|A - B\|_{\mathrm{F}}^2 + D$. Distinct differences are orthogonal in expectation, so averaging gives the same m^{-1} variance upper bound.

For projected estimation, every centered head observation is a martingale difference. Their average has variance at most $r\rho\sigma^2/t$. The residual mean conditional on the complete head history is still the true projected trace; its centered error is orthogonal to the head error. The same conditional argument bounds its variance as in (19). Coordinate queries also use a sum of orthogonal martingale differences. Hence both upper bounds remain valid. Memoryless Gaussian and zero-noise oracles are subclasses, so the lower bounds and constant-confidence minimax rate remain valid. □

This argument concerns a single constant-confidence execution. If causal noise state persists between repetitions, independent-run amplification should not be assumed without justification. One can instead use the conditional failure bound for each fresh algorithm run and a martingale concentration argument, provided the uniform kernel bounds remain true after every preceding run.

Proof of Proposition 14. Assume such an algorithm exists. Run it on the noiseless matrix $A_0 = I/d$. Its stopping time N is almost surely finite and its estimate lies in $I_0 = [1 - \varepsilon, 1 + \varepsilon]$ with probability at least 0.9. There is a finite L such that the event $\{N \leq L, \widehat{\tau} \in I_0\}$ has probability at least 0.85.

Let $A_1 = (1 + 4\varepsilon)I/d$ and $c = 4\varepsilon/d$. For a parameter $\eta > 0$, define a fresh oracle by

$$\mathcal{O}_{A_1}(v) = A_0 v + \frac{c}{\eta} J v, \qquad J \sim \text{Bernoulli}(\eta)$$

independently at every call. Its mean is $A_1 v$ and its centered noise is $c(J/\eta - 1)v$. Its total and directional second moments equal $c^2(1/\eta - 1)$, which is finite for each fixed $\eta > 0$; hence it belongs to the union in the proposition.

Couple the two executions using the same algorithm seed. Until the first $J = 1$, every response and every adaptive query is identical. By a union bound, the probability that a spike occurs in the first L calls is at most $L\eta$. Therefore on A_1 the algorithm stops with output in I_0 with probability at least $0.85 - L\eta$. Choose η so small that this exceeds 0.8. But I_0 is disjoint from the relative-success interval for A_1. The failure probability on this valid finite-variance instance is consequently greater than 0.8, contradicting the required 0.1. □

E PSD TAILS, CONFIDENCE AMPLIFICATION, AND LOCAL COSTS

For ordered PSD eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$,

$$\|A - A_k\|_{\mathrm{F}}^2 = \sum_{j>k} \lambda_j^2 \leq \lambda_{k+1} \sum_{j>k} \lambda_j \leq \frac{\tau^2}{k+1} \leq \frac{\tau^2}{k}.$$

This establishes the tail estimate used in both upper bounds.

For confidence $1 - \delta$, independently repeat a 0.9-success construction an odd number L of times. If the median is outside the target interval, at least $L/2$ repetitions fail. Hoeffding's inequality bounds this probability by $\exp(-2L(1/2 - 0.1)^2) = \exp(-0.32L)$. Thus an odd $L \geq \log(1/\delta)/0.32$ suffices. The lower-bound coefficient is $\mathrm{kl}(1 - \delta, \delta) = (1 - 2\delta)\log((1 - \delta)/\delta)$, which is of order $\log(1/\delta)$ for $\delta \leq 0.1$.

Besides queries, the pilot uses $O(ds)$ storage and $O(ds^2)$ local orthogonalization work. The residual factors cost $O(dsm)$ arithmetic for m probes; streaming the residual samples avoids storing them all. These costs can make a query-efficient estimator unattractive when an individual product is cheap. Our query results do not imply a wall-clock speedup.

F SAMPLE COVARIANCE: EXACT MOMENTS AND ACCESSIBLE STRUCTURE

Let x have finite fourth moment, $H = xx^\top$, $A = \mathbb{E}H$, $\tau = \mathrm{tr}(A)$, and $\mu_4 = \mathbb{E}\,\|x\|_2^4$. A batch-$b$ query averages b independent copies of Hv. Its noise is the average of Δv, where $\Delta = H - A$.

The total second moment obeys

$$\mathbb{E}\,\|\xi(v)\|_2^2 = \frac{1}{b} v^\top \big(\mathbb{E}[\|x\|_2^2\, xx^\top] - A^2\big) v. \tag{37}$$

Thus the operator norm of the PSD matrix in parentheses, divided by b, is a valid σ^2. Taking $\rho = 1$ always gives a valid directional bound.

For a uniform sphere vector independent of Δ,

$$d^2 \mathbb{E}_u (u^\top \Delta u)^2 = \frac{d}{d+2}\big[(\mathrm{tr}\,\Delta)^2 + 2\,\|\Delta\|_{\mathrm{F}}^2\big].$$

Since $\mathbb{E}(\mathrm{tr}\,\Delta)^2 = \mu_4 - \tau^2$ and $\mathbb{E}\,\|\Delta\|_{\mathrm{F}}^2 = \mu_4 - \|A\|_{\mathrm{F}}^2$, the average directional moment is

$$\nu = \frac{d}{b(d+2)} \left(3\mu_4 - \tau^2 - 2\,\|A\|_{\mathrm{F}}^2\right). \tag{38}$$

These equations describe the actual minibatch mechanism rather than an externally added Gaussian perturbation.

At q queries, the direct estimator with the same bq independent samples is $T_{\mathrm{direct}} = (bq)^{-1} \sum_{i,\ell} \|x_{i\ell}\|_2^2$. Its variance is $(\mu_4 - \tau^2)/(bq)$. For plain sphere Monte Carlo,

$$T_{\mathrm{sphere}} = \frac{d}{bq} \sum_{i=1}^{q} \sum_{\ell=1}^{b} (u_i^\top x_{i\ell})^2, \qquad \mathbb{E}[T_{\mathrm{sphere}} \mid \{x_{i\ell}\}] = T_{\mathrm{direct}}.$$

The law of total variance therefore proves $\mathrm{Var}(T_{\mathrm{direct}}) \leq \mathrm{Var}(T_{\mathrm{sphere}})$. The comparison follows from conditioning on the sampled rows and concerns plain sphere sampling. At batch size one, the same norm estimator is also realizable from vector responses, as the following identity shows.

Lemma 15 (Norm recovery from a rank-one response)**.** *Let $y = xx^\top v$ for a unit vector v. When $y \neq 0$,*

$$\|x\|_2^2 = \frac{\|y\|_2^2}{v^\top y}. \tag{39}$$

For v drawn independently and uniformly from the sphere, setting the recovered norm to zero at $y = 0$ gives $\|x\|_2^2$ almost surely. Hence q batch-one responses realize T_{direct}, with variance $(\mu_4 - \tau^2)/q$.

Proof. Writing $a = x^\top v$ gives $v^\top y = a^2$ and $\|y\|_2^2 = a^2 \|x\|_2^2$. If $x \neq 0$, an independent sphere query has $a \neq 0$ almost surely; if $x = 0$, then $y = 0$ and its norm is zero. □

The recorded batch-one direct trials average the sampled norms in code. Equation (39) gives a vector-response implementation that is equal on the same samples, so those trials also describe its error distribution. The plotted curve is reused by this exact equivalence; it is not a new run of a response-recovery implementation.

For larger batches, one aggregated response generally does not determine the average sample norm. For example, let $z \perp v$ and $a \neq 0$. The two-sample batches $(av + z, av - z)$ and (av, av) both return $a^2 v$, while their average squared norms are $a^2 + \|z\|_2^2$ and a^2. Adding zero samples gives the same non-identifiability for any $b > 2$. Direct averaging of individual sample norms therefore uses sample information not generally determined by a single response when $b > 1$. This observation does not assert optimality of any estimator for every structured oracle.

G From application operators to noisy products

The applications in the introduction share a fixed mean operator and fresh randomness in each response. The following mappings identify sufficient conditions for using our results; the cited works establish the operators, trace objectives, or computational primitives.

Sampling an operator. Let $H(Z)$ be a random matrix with $A = \mathbb{E}H(Z) \succeq 0$. For independent batch elements in (1), if $M_2 = \mathbb{E}\,\|H(Z) - A\|_{\mathrm{op}}^2 < \infty$, then

$$\mathbb{E}\xi_b(v) = 0, \qquad \mathbb{E}\,\|\xi_b(v)\|_2^2 = \frac{1}{b}\mathbb{E}\,\|(H(Z) - A)v\|_2^2 \leq \frac{M_2}{b}, \qquad \|v\|_2 = 1. \tag{40}$$

The directional moment is no larger, so $\sigma^2 = M_2/b$ and $\rho = 1$ are valid envelopes. A sharper ρ requires a directional calculation, as in Appendix F. The sufficient query schedules use a known numerical bound on these moments. Model parameters and the sampling distribution stay fixed during one trace estimate, and each call draws a new independent batch.

Curvature, sensitivity, and second-order statistics. For generalized Gauss–Newton curvature, $H(Z) = J_Z^\top C_Z J_Z$ is PSD when the loss Hessian in output space satisfies $C_Z \succeq 0$. Jacobian-vector and transposed-Jacobian-vector operations evaluate its action without forming the matrix. Fisher curvature instead admits $H = ss^\top$, with s a model score sampled under the model distribution; Martens (2020) discusses its relation to generalized Gauss–Newton curvature. Curvature traces summarize training dynamics in Schneider et al. (2021). For input sensitivity, $H(x) = J(x)^\top J(x)$ gives $\mathrm{tr}(A) = \mathbb{E}_x\,\|J(x)\|_{\mathrm{F}}^2$, the average squared-Jacobian penalty studied by Hoffman et al. (2019). Sampling inputs or augmentations defines the distribution in this expectation; the same sampled input must be used for both Jacobian factors in a product. For features or gradients z, $H = zz^\top$ gives $\mathrm{tr}(A) = \mathbb{E}\,\|z\|_2^2$. Centering z at its true mean gives covariance and total variance. These constructions satisfy the response conditions through (40).

Unbiased compression. Let a compressor C satisfy $\mathbb{E}[\mathsf{C}(w) \mid w] = w$ and $\mathbb{E}\,\|\mathsf{C}(w) - w\|_2^2 \leq \omega\,\|w\|_2^2$. Returning $\mathcal{O}_A(v) = \mathsf{C}(Av)$ with fresh compression randomness gives

$$\mathbb{E}\,\|\xi(v)\|_2^2 \leq \omega\,\|A\|_{\mathrm{op}}^2, \qquad \|v\|_2 = 1.$$

The QSGD quantizer supplies these properties (Alistarh et al., 2017, Lemma 3.1). Thus $\rho = 1$ and $\sigma^2 = \omega\,\|A\|_{\mathrm{op}}^2$ suffice; under our trace normalization, 4ω is a known envelope. The response

covariance may vary with Av, so compression need not behave like fixed-covariance noise. This is an application of the compressor to product responses; communication volume is a separate implementation cost.

Posterior covariance and randomized solves. For a Gaussian posterior with fixed precision $K \succ 0$, $A = K^{-1}$ is its covariance and $\mathrm{tr}(A)$ is total posterior variance, the A-optimal design objective (Herman et al., 2020). A full-vector solve with independent randomness fits the oracle when it is unbiased for $K^{-1}v$ and has a uniform second-moment bound. Randomized truncation with survival-probability reweighting is one route to unbiased inverse actions (Potapczynski et al., 2021). Ordinary early termination of a deterministic solver can introduce bias. Our query count treats each solve as one returned vector; its inner work must also enter a computational comparison.

Using additional structure. If individual sample traces or factor norms are directly available, averaging them can be cheaper. Even a batch-one response for $H = zz^\top$ reveals $\|z\|_2^2$ almost surely under a sphere query (Lemma 15). The exact variance identity can assess any pilot satisfying its assumptions, while the general oracle-class lower bound need not be attained within a particular application subclass.

H Additional experiment details

Computational protocol. The synthetic mechanism and Digits studies use float64, Python 3.12.7, NumPy 1.26.4, SciPy 1.13.1, and one BLAS thread on an Intel Core i9-12900H. Their confidence bands use 1,000 percentile-bootstrap resamples; the budgeted-decision and learning studies below use 2,000. The budget sweep uses $d = 128$, $q \in \{24, 48, 96, 192, 384\}$, and isotropic $\nu \in \{0, 0.01, 1\}$, with 200 repetitions balanced over five rotations. Spectra are trace-normalized j^{-2}, four unit eigenvalues followed by 0.01, or all ones. Geometry uses $d \in \{16, 32, 64, 128, 256\}$ and $q \in \{32, 128, 512\}$ with 400 repetitions; raw records also include mixed directions with $\rho \in \{1/d, 0.25, 1\}$. Digits uses the five budget values and $b \in \{1, 8, 64\}$, with 200 repetitions. Oracle and probe streams are separate; true traces and full reference operators are used only for scoring and diagnostics.

Implementation. All query inputs are checked to have unit norm. The homogeneous extension for a nonzero vector z is $\|z\|_2 \, \mathcal{O}_A(z/\|z\|_2)$. Gaussian sketches use that extension. The implementation obtains a rank-revealing orthonormal basis using pivoted QR, with the threshold $\epsilon_{\mathrm{machine}} \max(d, s) \max_j |R_{jj}|$. This numerical rank rule is an implementation choice; the theoretical basis spans the exact range. The correction evaluates $u^\top Bu = (Q^\top u)^\top (W^\top u)$ without forming B.

Each noise draw is fresh, including when a direction is repeated. Isotropic Gaussian noise is $\sigma g/\sqrt{d}$, fixed-direction noise is $\sigma z e_1$, and coupled noise is $\sigma z v$. For the mixed law, we draw an independent scalar Gaussian and project a Gaussian vector onto $v^\perp$, using the weights in (20). Student noise replaces z by $t_3/\sqrt{3}$. Probe randomness and oracle randomness use distinct generators.

Raw data and uncertainty. The main budget suite contains 45,000 estimates; geometry has 30,000; fixed-width head experiments have 9,600; Digits has 15,000; allocation has 3,000. Separate moment diagnostics include 20,000 weighted-inverse samples and 6,000 noisy pilots. The record for every estimator run contains its estimate, true trace, signed relative error, budget, actual query count, runtime, repetition identifier, and configuration. Repetition seeds are a deterministic function of the full configuration. Five fixed rotations in the main suite are balanced across repetitions; bootstrap bands resample the retained errors and should be interpreted as descriptive uncertainty for this design.

Numerical checks verify exact scalar-matrix traces, complete-basis correction, nonsymmetric control variates, total and directional noise moments, and query accounting. Symbolic algebra verifies the simplification of (33). Monte Carlo moment checks do not replace the proof; in particular, small oversampling can produce heavy-tailed inverse-moment samples.

Independent predictions. The transition study adds 16,800 estimator trials and 10,500 independent pilots. Its fixed settings are $d = 128$, $q = 192$, $s = 32$; three spectra; all seven noise levels from 0 through 10; four methods; 200 trials per point; and 500 separate pilots per spectrum/noise pair. The study uses five balanced rotations. Exact random seeds are provided in the supplement. Independent

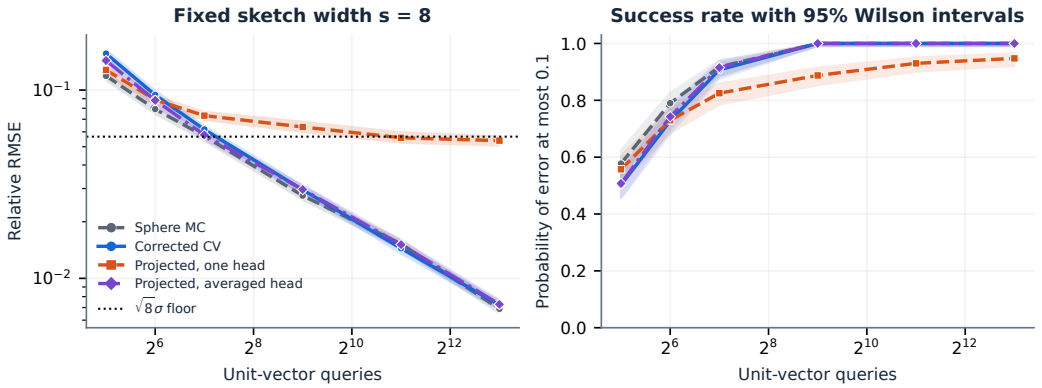


Figure 5: Head uncertainty with a fixed width-eight pilot. A single head observation leaves the predicted RMSE floor; complete correction and increasing head averaging restore decreasing error. The right panel reports success at relative error 0.1 with 95% Wilson intervals.

diagnostic and estimator random streams ensure that pilot information is never supplied to the benchmark methods. Diagnostic calls are additional to their query budgets. Predictions use the exact sphere variance $(V(A) + \nu)/q$.

These mechanism and Digits suites contain 119,400 estimator trials and 36,500 inverse-moment/pilot diagnostics. The budgeted-decision study adds 78,400 estimator trials and the learning-curvature study adds 6,300, giving 204,100 estimator trials in total. Transition gain intervals independently resample the two methods' errors. Pointwise 95% percentile-bootstrap intervals for the prediction and measurement axes use 500 pilots and 200 estimator trials, respectively; analogous gain intervals are computed for Digits. These uncertainty summaries use the retained observations; they do not change oracle trials or estimator settings. The flat/noiseless gain is undefined at the population level and is not interpreted from rounding errors.

Choice of environments. Power and spiked spectra provide distinct ways for a small subspace to reduce probe variance. The flat spectrum is the analytic control $V(A) = 0$. Scalar matrices isolate oracle geometry and head uncertainty. Digits keeps the mean operator fixed while varying product quality through batch size. These controls isolate the mechanisms tested on learned curvature and by paid decisions. The full sweeps below retain the complete tested range.

Fixed-budget allocation. All pilot, head and residual calls count toward the same budget q. The default width is $s = \min\{d, \max\{4, 2\lfloor q/12 \rfloor\}\}$; averaged projection uses $t = \max\{1, \lfloor \lfloor q/3 \rfloor / s \rfloor\}$. Remaining calls estimate the residual. This allocation uses neither the true trace nor the noise level. Experiments spend all calls left after the realized pilot on the residual; Algorithm 1 reserves $2s$ pilot calls to keep its theoretical residual count fixed. Appendix H gives complete protocols, raw-record counts and sweeps.

H.1 Head uncertainty and directional noise

On $A = I/32$, we fix a width-eight pilot and query-coupled noise with $\sigma = 0.02$, then increase the budget from 32 to 8,192. The one-head RMSE approaches the predicted floor $\sqrt{8}(0.02) = 0.0566$: at the largest budget it is 0.0539 with interval $[0.0501, 0.0575]$. Correction and averaged projection reach 0.00721 and 0.00728, respectively (Figure 5). This scalar operator isolates head error because sphere sampling has no matrix-induced variance.

Holding noise energy at $\sigma^2 = 10^{-4}$ instead isolates direction. For $A = I/d$, sphere MSE is exactly ν/q. At $d = 256, q = 128$, measured $q\,\mathrm{MSE}/\sigma^2$ is 264, 245 and 65,729 for isotropic, fixed-direction and query-coupled noise, against predictions 256, 256 and 65,536. Figure 6 gives the dimension sweep and a finite-variance heavy-tail diagnostic.

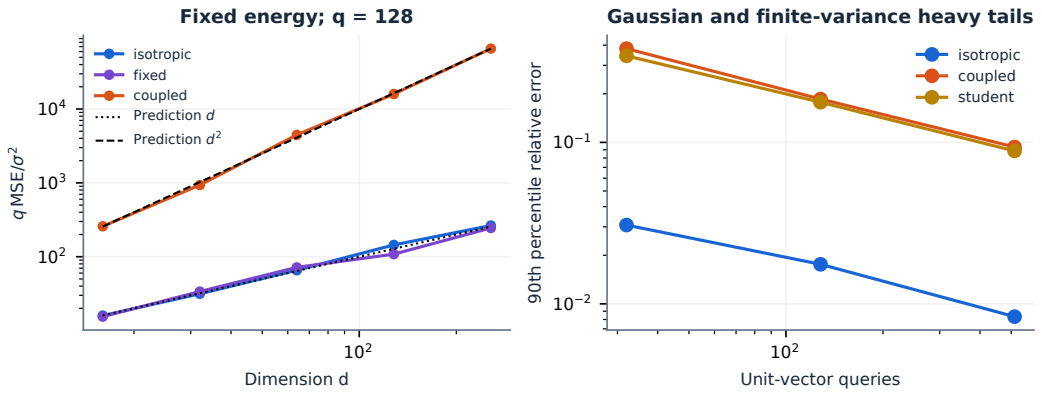


Figure 6: Equal noise energy with different directions. Left: five dimensions and three Gaussian geometries, with 400 trials per point; the d and d^2 lines are exact sphere-variance predictions. Right: empirical 90th-percentile absolute relative errors at $d = 128$; Student-$t_3/\sqrt{3}$ replaces Gaussian query-coupled noise.

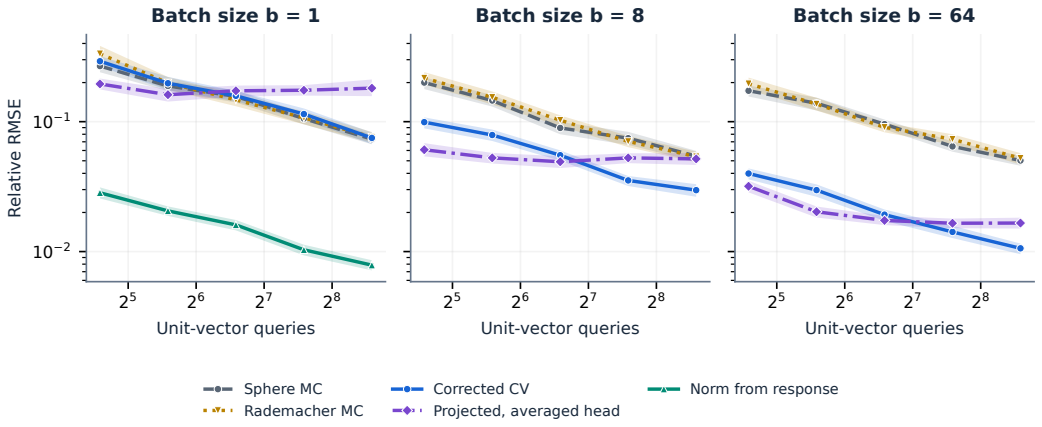


Figure 7: Minibatch quality and estimator structure on Digits; 200 trials per point. Generic vector estimators benefit from improved product quality. At $b = 1$, the sample-norm baseline is also realizable from each complete vector response; its curve reuses the recorded direct-sampling results by the exact identity in Appendix F. Across panels, the sample cost is bq.

H.2 Minibatch quality for a fixed mean operator

For the 1,797 Digits observations with 64 features (scikit-learn developers, 2026), divide pixels by 16 and set $A = \mathbb{E}xx^\top$ under uniform row sampling. Each call returns $b^{-1}\sum_{\ell=1}^{b} x_\ell(x_\ell^\top v)$ from a fresh batch. Increasing b divides the noise moments by b while keeping A fixed (Appendix F). Each query uses b samples, so comparisons across batch sizes have different sample costs.

At $q = 192, b = 1$, correction and sphere sampling have RMSEs 0.114 and 0.106: gain 0.92 with interval $[0.80, 1.07]$. At $b = 64$, they reach 0.0142 and 0.0645, giving gain 4.55 with interval $[3.91, 5.30]$. Thus higher product quality reveals a correction benefit within this estimator comparison.

The response structure also permits a stronger baseline. At $b = 1$, $y = x(x^\top v)$ reveals $\|x\|_2^2 = \|y\|_2^2/(v^\top y)$ almost surely for an independent sphere query. Averaging these values attains the direct sample-norm estimator's error distribution, whose recorded RMSE is 0.0103 at $q = 192$. At $b = 64$, direct access to the sampled rows gives RMSE 0.00136. Appendix F proves the recovery identity and explains why one aggregated batch response need not reveal its average sample norm. Figure 10 shows the full baseline comparison.

At a fixed sample budget $N = bq$, let ν_1 be the batch-one directional noise moment. Since $\nu_b = \nu_1/b$, sphere sampling has

$$\mathrm{MSE}_{\mathrm{sphere}}(N, b) = \frac{V(A) + \nu_1/b}{N/b} = \frac{bV(A) + \nu_1}{N}. \tag{41}$$

Thus averaging more rows per call leaves the oracle-noise contribution unchanged at fixed N, while fewer probe directions increase the matrix-induced contribution. For correction, batch size also changes the pilot quality and the fraction of calls spent constructing it. A gain at fixed query count therefore need not be a gain at fixed sample cost.

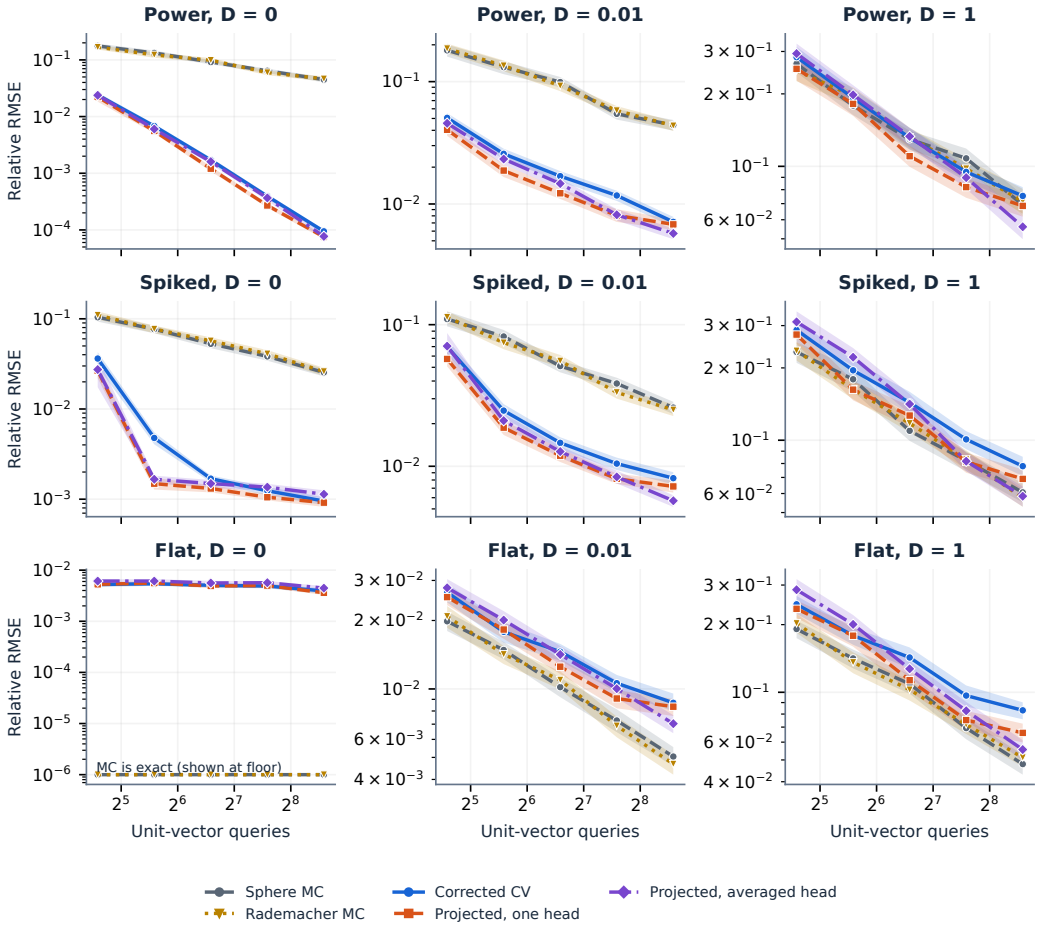


Figure 8: Complete budget sweep: three spectra, three isotropic noise levels, five estimators, five budgets, and 200 trials per point. The flat/noiseless ordinary estimates are exact to rounding and displayed at 10^{-6} for visibility. Axis ranges vary across panels.

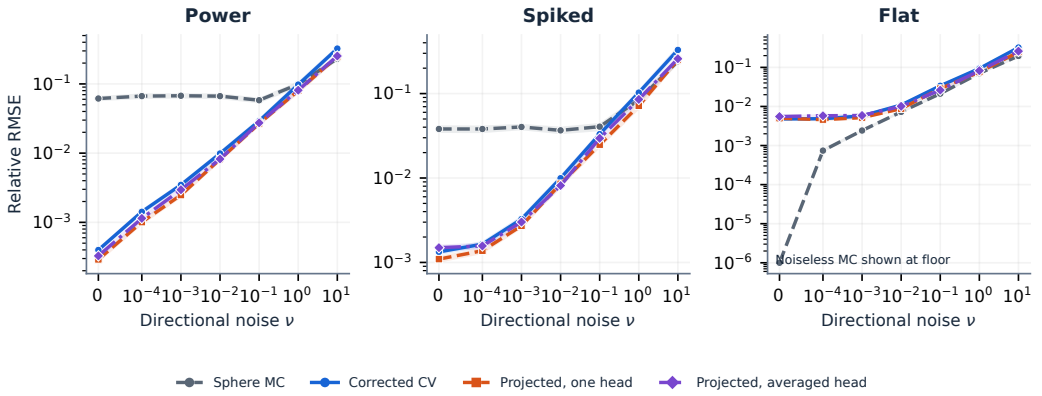


Figure 9: Complete RMSE sweep for the transition study: three spectra, seven noise levels, and four estimators at $q = 192$. Flat/noiseless sphere error is shown at 10^{-6}. These records underlie Figure 3; a ratio to a zero population baseline is undefined.

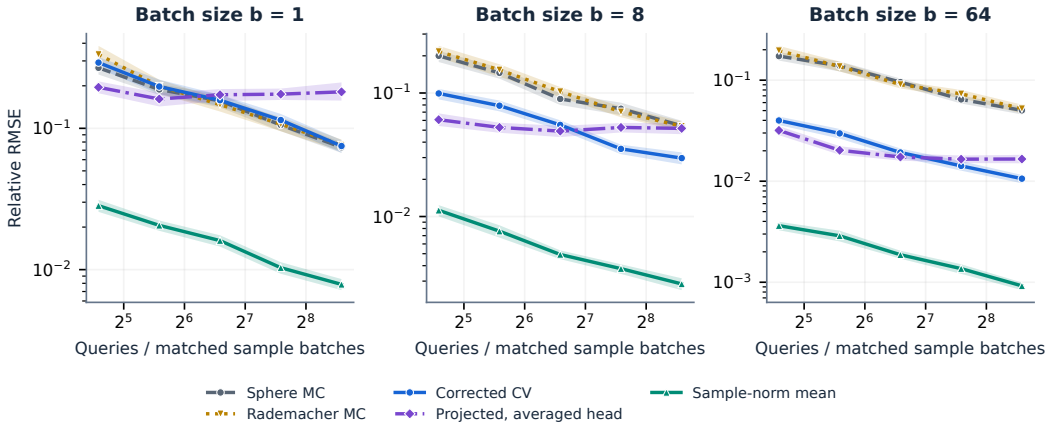


Figure 10: Digits with the sample-norm baseline, using bq independent rows at budget q. At $b = 1$, norm averaging is realizable from vector responses by Lemma 15; its curve reuses the equivalent recorded direct-sampling trials. At $b > 1$, the plotted norm estimator inspects the individual rows. Its horizontal unit is a matched sample batch.

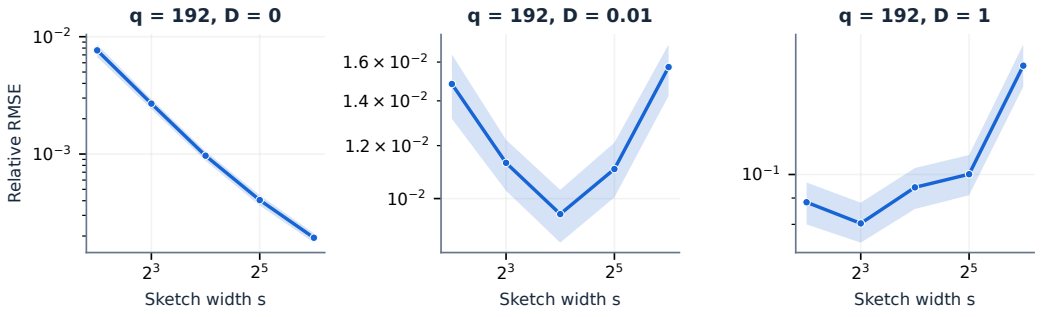


Figure 11: Allocation sensitivity for complete correction on the power spectrum at $q = 192$; 200 repetitions per point. Sketch widths are $4, 8, 16, 32, 64$. The larger-sketch benefit in the noiseless case does not persist monotonically with noise. These are diagnostic configurations, not a tuned policy used in the main plots.

Table 3: Weighted inverse-moment diagnostics: 4,000 draws per configuration. The reference is the exact value in (33). Empirical means provide a numerical check, not a proof or a tail-confidence statement.

d	k	s	Empirical mean	Exact mean
16	1	4	7.818	8.000
32	2	6	21.050	21.333
64	4	10	50.832	51.200
64	8	18	56.592	56.889
32	3	9	19.337	19.309

Table 4: Noisy pilot Frobenius error: 1,000 independent pilots per row, $\sigma = 0.02$. The proved value is an upper bound, not an equality. Student noise uses $t_3/\sqrt{3}$ along the query.

Noise	d	k	Empirical risk	Proved upper bound
isotropic	32	2	0.01416	0.02361
isotropic	64	4	0.01432	0.02690
coupled	32	2	0.01074	0.02361
coupled	64	4	0.00927	0.02690
student	32	2	0.00841	0.02361
student	64	4	0.00756	0.02690

Measured operator moments. For Digits after division by 16, $\tau = 15.014199$, $\mu_4 = 230.527933$, and $\|A\|_{\mathrm{F}}^2 = 110.960575$. At batch size one, (37) gives $\sigma^2 = 52.638770$ and (38) gives $\nu = 236.835372$. Both noise moments divide by b for a batch of size b. These diagnostic quantities are not used to tune the estimators.

Reproduction. The supplement's README gives the exact commands for both verification scripts, the six experimental suites, and the transition study. Use 400 repetitions for geometry and fixed-head suites, and 200 for the other suites. The transition runner uses its fixed protocol counts. The supplied analysis scripts reconstruct all empirical figures and numerical tables from retained raw records. Separate reruns reproduced 1,125 main-suite estimates, all 16,800 transition estimates, and all 10,500 transition pilots exactly, ignoring runtime fields. The conservative theorem schedules and empirical fixed-budget rules are explicitly separated.

I Budgeted decisions and a fresh-product leave-one-out baseline

I.1 A paid diagnostic decision

The control-variate identity in Lemma 2 allows the pilot to be evaluated within the same query budget as the final estimate. Fix a total budget q, a Gaussian sketch width s, and a diagnostic count $h \geq 2$. Construct $B = QW^\top$ as in Section 3.2: the sketch uses s unit-vector calls, and fresh products on the $r = \mathrm{rank}(Y)$ columns of Q use another r calls. Thus the realized pilot charge is $c = s + r \leq 2s$. We require $q \geq 2s + h + 1$, leaving $m = q - s - r - h \geq 1$ fresh production calls for every possible rank. This implementation spends exactly $c + h + m = q$ calls, including the diagnostics.

After the pilot, draw independent sphere directions $u_1, \ldots, u_h$ and obtain one fresh vector response for each. The same response supplies two paired scalar observations,

$$X_i = du_i^\top \mathcal{O}_A(u_i), \qquad C_i = du_i^\top Bu_i, \qquad R_i = \mathrm{tr}(B) + X_i - C_i. \tag{42}$$

Conditional on the pilot, X_i and R_i are unbiased for τ and have variances $S_0 = V(A) + \nu$ and $S_B(B) = V(A - B) + \nu$, respectively. We use their ordinary sample variances, computed from these h paid paired observations. The primary *select* rule sets $\gamma = 1$ if $\widehat{\mathrm{Var}}(R) < \widehat{\mathrm{Var}}(X)$, and sets

$\gamma = 0$ otherwise. As a secondary ablation, the *blend* rule sets

$$\gamma = \Pi_{[0,1]}\left(\frac{\widehat{\mathrm{Cov}}(X, C)}{\widehat{\mathrm{Var}}(C)}\right), \tag{43}$$

with $\gamma = 0$ if the denominator vanishes. Its population counterpart minimizes the variance of $X - \gamma C$ over $[0, 1]$. Neither rule needs a separate estimate of ν, $V(A)$, or $\mathbb{E}V(A - B)$.

Freeze B, γ, and the realized m after the pilot and diagnostics. With independent, fresh production directions u'_j and responses, return

$$\widehat{\tau}_\gamma = \gamma\,\mathrm{tr}(B) + \frac{d}{m}\sum_{j=1}^{m} u_j'^{\top}\{\mathcal{O}_A(u'_j) - \gamma B u'_j\}. \tag{44}$$

Let $\mathcal{F}$ denote the complete pilot and diagnostic transcript. Under the memoryless oracle of Section 2,

$$\mathbb{E}[\widehat{\tau}_\gamma \mid \mathcal{F}] = \tau, \qquad \mathbb{E}[(\widehat{\tau}_\gamma - \tau)^2 \mid \mathcal{F}] = \frac{V(A - \gamma B) + \nu}{q - s - r - h}. \tag{45}$$

Indeed, conditional on $\mathcal{F}$, the coefficient and matrix are fixed, each production summand has mean $\mathrm{tr}(A - \gamma B)$, and its probe and response-noise variances add. The conditional mean does not depend on the diagnostic decision, even when that decision chooses the higher-variance option. Averaging (45) gives the exact risk identity

$$\mathbb{E}(\widehat{\tau}_\gamma - \tau)^2 = \mathbb{E}\left[\frac{V(A - \gamma B) + \nu}{q - s - r - h}\right]. \tag{46}$$

Both r and γ can be random; replacing the denominator by $q - s - \mathbb{E}r - h$ would generally change this quantity. Conditional on the transcript, the production directions must still be independent uniform sphere draws. Choosing nonuniform directions from the pilot would change the relevant second and fourth moments and would require a separate weighting and response-noise analysis.

The decision here is made *after* paying for the pilot. At its realized remaining budget, the conditional choice between complete correction and plain production is $S_B(B) < S_0$; both choices have denominator $q - c - h$. Corollary 3 instead asks, before buying a pilot of fixed cost c and without diagnostics, whether corrected production improves on q plain sphere calls. The conditional MSE given a realized pilot B beats that full-budget benchmark only if $S_B(B)/(q - c - h) < S_0/q$. For the adaptive rule, the comparison with q plain calls uses the unconditional risk in (46); the diagnostics do not provide a free forecast of whether the pilot should have been constructed.

I.2 Fresh-product XTrace

We implement the published leave-one-out XTrace construction of Epperly et al. (2024) using two batches of fresh oracle calls. Let the even budget be $q = 2k$ with $k \le d$. Draw independent uniform sphere directions u_i, set $\omega_i = \sqrt{d}\,u_i$, and use k calls to form $Y_i = \sqrt{d}\,\mathcal{O}_A(u_i)$. Let $Q \in \mathbb{R}^{d\times k}$ be an orthonormal basis for the sketch columns, completed to k columns when the sketch is rank deficient. A second, independent batch of k unit-vector calls gives $Z_j = \mathcal{O}_A(q_j)$ for the columns q_j of Q; write $Z = [Z_1, \ldots, Z_k]$. Thus all $2k = q$ vector responses are charged. Conditional on the first batch, $\mathbb{E}[Z \mid \omega_1, \ldots, \omega_k, Y] = AQ$.

For each held-out index i, let P_i project onto $\mathrm{span}(Y_{-i})$, put $p_i = P_i\omega_i$ and $z_i = (I - P_i)\omega_i$, and define $c_i = Q^\top p_i$. Since $\mathrm{range}(P_i) \subseteq \mathrm{range}(Q)$, $Qc_i = p_i$. The leave-one-out score and final estimate are

$$T_i = \mathrm{tr}(P_i Z Q^\top) + \beta_i z_i^\top (Y_i - Zc_i), \qquad \widehat{\tau}_{\mathrm{XT}} = \frac{1}{k}\sum_{i=1}^{k} T_i. \tag{47}$$

The unnormalized version uses $\beta_i = 1$. The residual-normalized version uses $\beta_i = (d - r_i)/\|z_i\|_2^2$, where $r_i = \mathrm{rank}(Y_{-i})$.

The two-stage construction remains unbiased with fresh centered products. First condition on every first-stage probe and response. Independence of the second batch and $Qc_i = p_i$ give

$$\mathbb{E}_Z[T_i \mid \omega, Y] = \mathrm{tr}(P_i A) + \beta_i z_i^\top (Y_i - Ap_i). \tag{48}$$

Next condition on Y_{-i} and ω_i. The held-out first-stage noise is independent of Y_{-i} and centered, so $\mathbb{E}[Y_i \mid Y_{-i}, \omega_i] = A\omega_i$. The remaining term is $\beta_i z_i^\top A z_i$. Conditional on Y_{-i}, P_i is fixed and ω_i is isotropic. For $\beta_i = 1$, $\mathbb{E}_{\omega_i}[z_i z_i^\top \mid Y_{-i}] = I - P_i$; for residual normalization, rotational symmetry in the $(d - r_i)$-dimensional orthogonal complement gives $\mathbb{E}_{\omega_i}[\beta_i z_i z_i^\top \mid Y_{-i}] = I - P_i$. Hence $\mathbb{E}T_i = \operatorname{tr}(P_i A) + \operatorname{tr}((I - P_i)A) = \tau$, and averaging preserves this mean. The cancellation after the second-batch expectation permits Q to depend on the held-out Y_i while each projector P_i uses only Y_{-i}.

Residual normalization can have an infinite second moment when the leave-one-out codimension is one or two. For a nontrivial projection ($0 < r_i < d$), a spherical probe has $\mathbb{E}\|z_i\|_2^{-2} < \infty$ if and only if $d - r_i > 2$; when $r_i = 0$, $\|z_i\|_2^2 = d$ is fixed. We conservatively choose the unnormalized score in advance whenever $d - k + 1 \leq 2$. This fixed fallback does not depend on a realized residual length; otherwise the implementation uses residual normalization. Rank-deficient sketches still use k second stage calls and the actual leave-one-out subspaces.

The exact-product variance analysis of Epperly et al. (2024) does not directly apply to this fresh-noise oracle. Both Y_i and Z_j contain noise, the sketch subspaces depend on first-stage noise, and the second-stage responses are shared across the leave-one-out scores. Equation (47) establishes the estimator and the argument above establishes its mean under the stated oracle assumptions; its risk in this setting is measured in the budget-matched experiments.

I.3 Budgeted selection and sample-reuse study

This study compares six vector-oracle methods on the same prescribed matrix families: sphere and Rademacher sampling, complete correction, averaged projection, paid selection, and fresh-product XTrace. Matrices have $d = 128$ and trace one. The power, spiked, and flat spectra are as in Appendix H; each configuration has 200 trials, balanced over five rotations with seeds 9100–9104. Budgets are $q = 48, 96, 192$, with sketch widths $s = 8, 16, 32$. The paid policy uses $h = 4, 8, 16$ diagnostic calls, respectively. Averaged projection makes two fresh head calls per basis vector at these budgets. Every method uses exactly q returned vectors, including both XTrace batches and all policy phases.

The isotropic grid uses all three spectra and $\nu \in \{0, 0.01, 1, 10\}$. We also compare query-aligned Gaussian and standardized Student-t_3 noise on the power and flat spectra at $\nu \in \{0.01, 1\}$. These conditions fix the actual sphere-observation noise variance: isotropic noise uses $\sigma^2 = \nu/d$, whereas query-aligned noise uses $\sigma^2 = \nu/d^2$. Thus this grid matches scalar noise rather than total noise energy. The separate geometry experiment in Figure 6 holds energy fixed. There are 360 method/configuration groups and 72,000 estimator trials.

The diagnostic sensitivity study fixes $q = 192$, $s = 32$, power or flat spectra, isotropic $\nu = 0.01$ or 1, and $h \in \{8, 16, 32\}$. It compares binary selection and the clipped control weight with full-budget sphere sampling and fixed correction. The 32 groups contain another 6,400 trials. Its configuration-hashed streams are independent of the main comparison, including where nominal settings coincide. More diagnostics improve the information available for the decision while reducing production calls. For example, on the power spectrum at $\nu = 1$, binary selection gives RMSE 0.1164, 0.1112, and 0.1187 at the three h values. On the flat spectrum they are 0.0869, 0.0935, and 0.1061. This tradeoff is measured as the error of the entire paid policy.

The estimation seed is 20260925. Each complete configuration and repetition determines separate probe and oracle streams by a SHA-256-based rule. Raw records preserve the signed relative error, estimate, trace, actual queries, phase counts, selected coefficient, diagnostic variances, and elapsed time. Empirical bias, MSE, and RMSE are computed from the retained errors. The analysis uses 2,000 percentile-bootstrap replicates, resampling within each fixed rotation and independently across methods for gain ratios. Intervals are pointwise and descriptive; in particular, Student-t_3 noise has no fourth moment, so these empirical MSE intervals are not finite-sample coverage guarantees under the paper's moment assumptions. A flat/noiseless sphere gain is undefined and is excluded from ratios.

The prespecified main figure uses all three spectra at $q = 192$ with isotropic noise. Figures 12–13 retain the complete main grid. Figures 14–15 show both paid rules across the complete diagnostic sweep. No configuration or repetition was selected by its observed gain. The numerical tables accompanying the raw records also retain all bias estimates and uncertainty intervals.

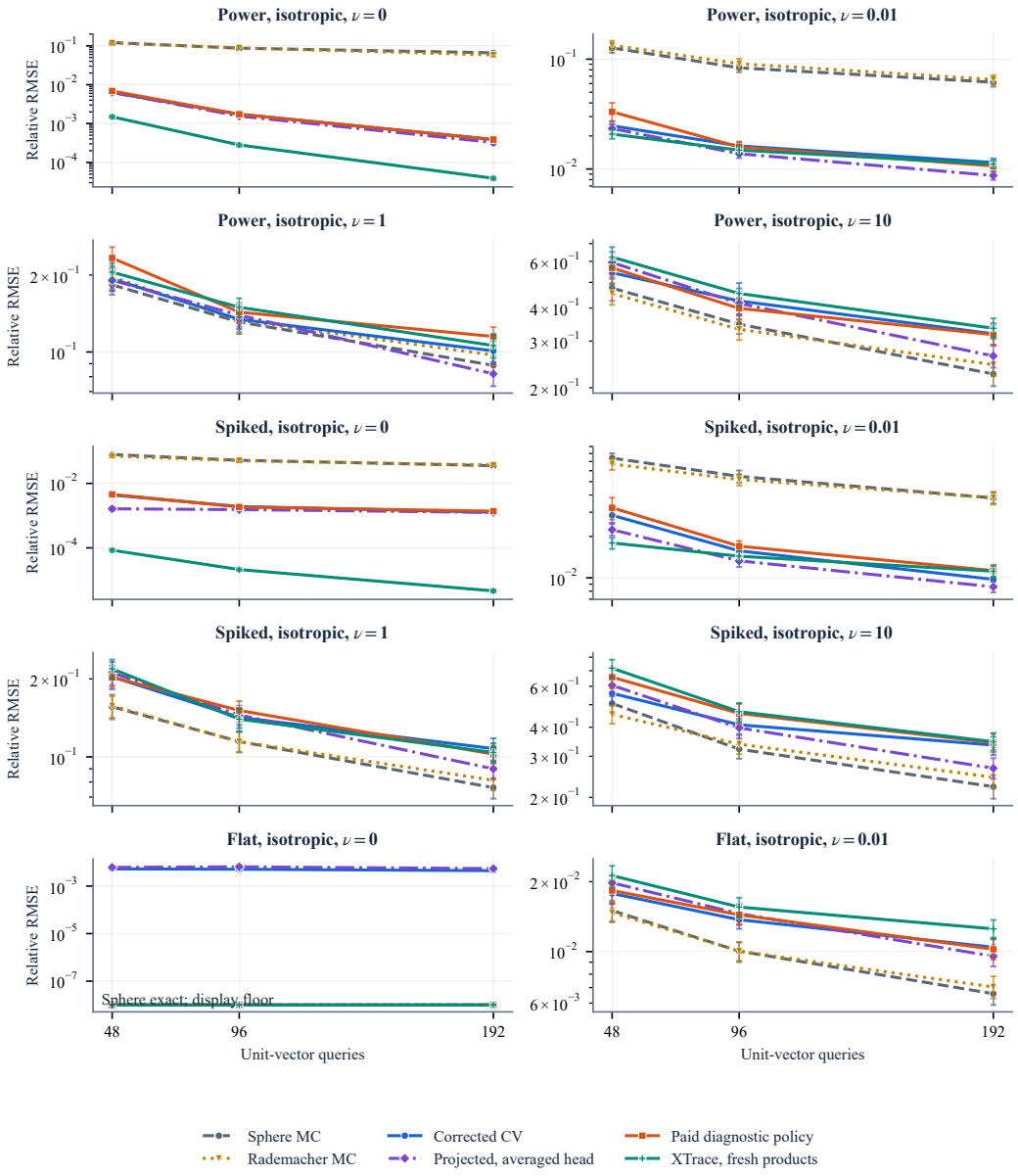


Figure 12: Complete six-method comparison, first ten conditions in the prespecified grid. Each curve uses $q = 48, 96, 192$ and 200 trials per point, with 95% bootstrap error bars. The flat/noiseless sphere, Rademacher, and XTrace errors are exact up to rounding and displayed at a numerical floor of 10^{-8}.

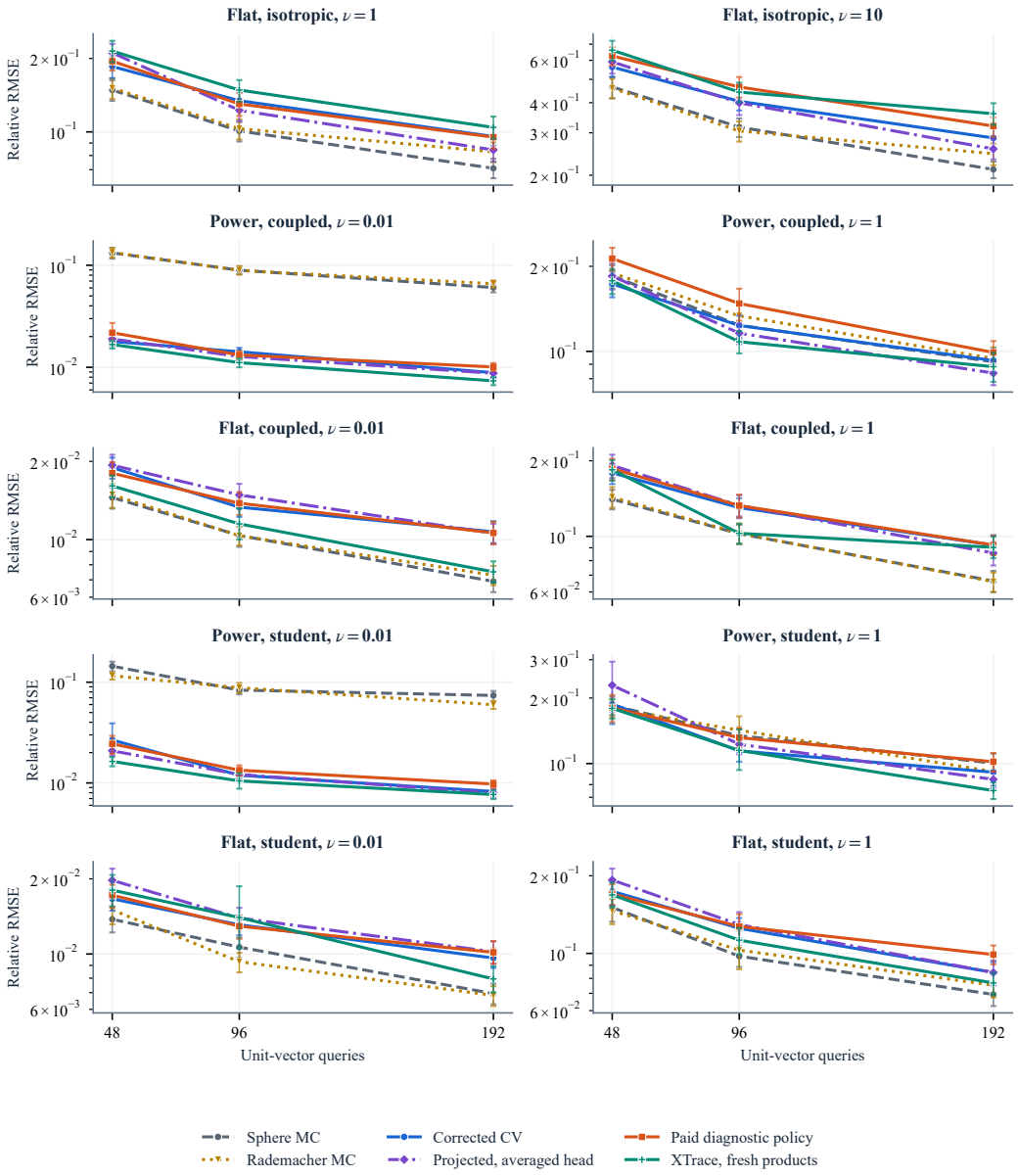


Figure 13: Complete six-method comparison, remaining ten conditions. The coupled and Student noise conditions match ν to the isotropic conditions, with lower total noise energy. The plot retains every tested failure regime; Student-t_3 intervals describe these recorded samples.

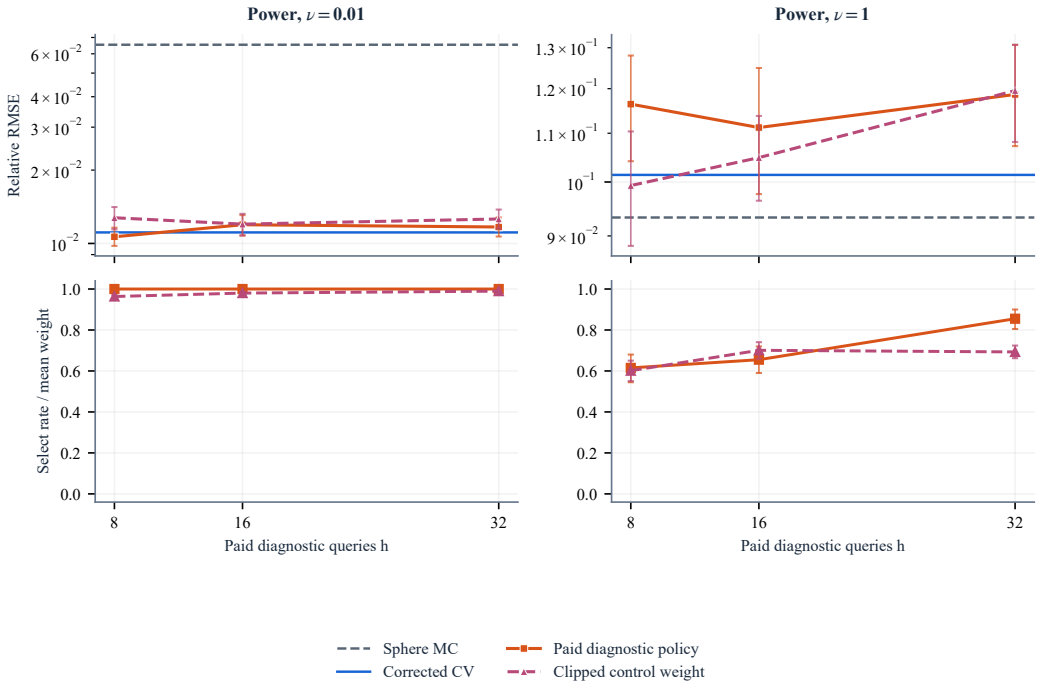


Figure 14: Paid diagnostic sensitivity on the power spectrum, at $q = 192$, $s = 32$. Top: full-budget error, including the h diagnostic calls; horizontal lines show the independently run sphere and fixed-correction baselines. Bottom: correction-selection frequency or mean clipped coefficient. Error bars are 95% bootstrap intervals over 200 trials per point.

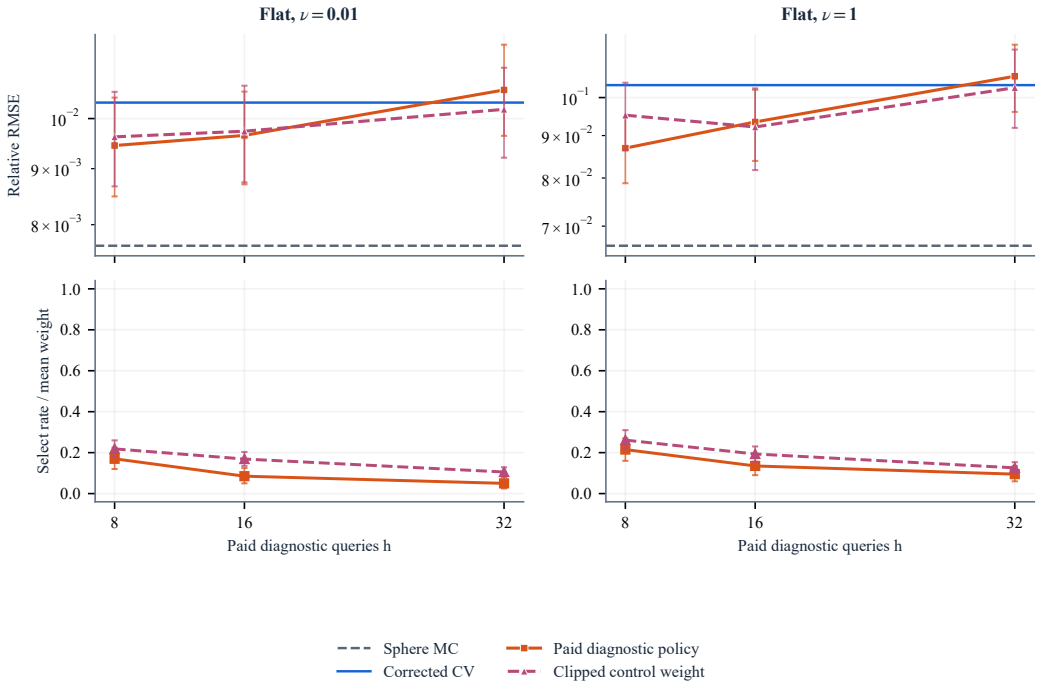


Figure 15: The same diagnostic sweep for the flat spectrum. A more frequent choice of plain production does not recover the pilot and diagnostic calls already spent. Both rules are evaluated against the full 192-query sphere baseline.

J FROZEN NEURAL-NETWORK CURVATURE STUDY

Target and preparation. We study the generalized Gauss–Newton (GGN) trace of a frozen CIFAR-10 classifier (Krizhevsky, 2009; Martens, 2020; Dangel et al., 2023). The network has three 3×3 convolutional layers with 16, 32, and 32 channels, ReLU activations, two 2×2 max-pooling layers, a final 2×2 average-pooling layer, and fully connected layers of width 128 and 10. It has $d = 81{,}290$ trainable parameters and no batch normalization or dropout. We train on all 50,000 official training images with pixels scaled to $[0, 1]$, no augmentation, Adam at learning rate 10^{-3}, batch size 128, and preparation seed 20260925. Checkpoints after epochs 1 and 5 were fixed before trace-estimator runs. A single set of 256 distinct images was selected uniformly without replacement from the official test split, using the recorded seed; it is held fixed across checkpoints and estimators. Parameters remain frozen and the model is in evaluation mode. The official test accuracies of the two checkpoints are 0.4095 and 0.5710.

For pool image x_i, let $f_\theta(x_i) \in \mathbb{R}^{10}$ be the logits, $J_i = \partial f_\theta(x_i)/\partial\theta$, and $p_i = \mathrm{softmax}(f_\theta(x_i))$. The fixed target is

$$A = \frac{1}{256}\sum_{i=1}^{256} G_i, \qquad G_i = J_i^\top C_i J_i, \quad C_i = \mathrm{diag}(p_i) - p_i p_i^\top \succeq 0. \tag{49}$$

This is the softmax cross-entropy GGN, a PSD matrix in parameter space; test labels do not enter its construction. The scoring-only reference uses the ten-row logit Jacobian for each image, without forming a $d \times d$ matrix:

$$\mathrm{tr}(G_i) = \sum_{c=1}^{10} p_{ic}\,\|J_{ic}\|_2^2 - \left\|\sum_{c=1}^{10} p_{ic} J_{ic}\right\|_2^2. \tag{50}$$

Evaluated in float64 and averaged over the pool, this gives reference traces 1278.3843 and 1965.2207 at epochs 1 and 5, respectively. Per-image traces are retained only for scoring, never passed to a vector-oracle estimator.

Fresh product access and structural comparator. For every unit vector v, the oracle samples b pool indices independently with replacement and returns $b^{-1}\sum_{\ell=1}^{b} G_{i_\ell} v$. Every query column, including a repeated direction in a matrix argument, gets a new batch. A Jacobian–vector product, multiplication by the ten-dimensional C_i, and a vector–Jacobian product implement the action without materializing G_i or A. In exact arithmetic its expectation is Av; the finite pool supplies finite second-moment envelopes, with the conservative directional choice $\rho = 1$. Each returned vector costs one oracle query and b sampled images.

We also time a richer-access model-score estimator. For each sampled image it draws an independent *model* label $Y \sim p_i$ and computes $s_{iY} = \nabla_\theta \log p_\theta(Y \mid x_i) = J_i^\top(e_Y - p_i)$. Consequently

$$\mathbb{E}_{Y\sim p_i} s_{iY} s_{iY}^\top = G_i, \qquad \mathbb{E}_{Y\sim p_i}\,\|s_{iY}\|_2^2 = \mathrm{tr}(G_i). \tag{51}$$

Thus its sample mean targets the same trace, whereas using observed CIFAR labels would generally target an empirical-Fisher quantity. The implementation computes per-image gradients during each run, vectorized in chunks of at most 64, without looking up the scoring cache. At nominal budget q it draws bq images in q matched batch units and performs *zero* GGN matrix–vector queries; error is compared together with its measured time and explicitly different access.

Frozen comparison and numerical accounting. The formal grid fixes epochs $\{1, 5\}$, $b \in \{1, 16, 64\}$, $q \in \{48, 96, 192\}$, and 50 independent repetitions of each of seven methods: sphere and random-sign Monte Carlo, complete correction, averaged projection, paid pilot selection, XTrace with fresh products, and model-score norms (6,300 trials). The estimator master seed is 2026092504; a hash of each trial configuration creates separate probe and minibatch streams. Pilot, diagnostic, and production calls are all charged. The prepared checkpoint and pool, protocol, code hashes, per-trial seeds, counts of returned vectors, and sample draws are recorded in the supplied manifests and append-only ledger. The vector methods see only oracle responses, never the reference trace or individual Jacobians.

Table 5: Neural-network GGN relative RMSE over 50 independent trials in each configuration. E is the frozen training epoch, b is the fresh minibatch size, and q is the nominal budget. Sp: sphere Monte Carlo; Rd: random-sign Monte Carlo; CV: complete correction; Pr: averaged projection; Pd: paid selection; XT: XTrace with fresh products. Score uses model-label score-gradient norms, a richer-access comparator with bq sampled images and zero GGN matrix–vector calls; its q denotes matched sampled-batch units. All other columns use q vector responses and bq sampled images. Values are rounded; prose comparisons use unrounded summaries.

E	b	q	Sp	Rd	CV	Pr	Pd	XT	Score
1	1	48	0.204	0.208	0.242	0.331	0.261	0.440	0.206
1	1	96	0.141	0.149	0.160	0.280	0.187	0.490	0.140
1	1	192	0.115	0.108	0.149	0.360	0.142	0.422	0.104
1	16	48	0.135	0.141	0.062	0.089	0.070	0.125	0.062
1	16	96	0.087	0.079	0.040	0.087	0.053	0.119	0.039
1	16	192	0.067	0.066	0.034	0.092	0.030	0.117	0.028
1	64	48	0.107	0.106	0.035	0.044	0.037	0.065	0.020
1	64	96	0.079	0.092	0.018	0.042	0.027	0.074	0.021
1	64	192	0.059	0.058	0.017	0.043	0.018	0.055	0.013
5	1	48	0.220	0.167	0.252	0.273	0.301	0.472	0.230
5	1	96	0.188	0.164	0.253	0.414	0.219	0.642	0.162
5	1	192	0.100	0.116	0.141	0.471	0.137	0.530	0.113
5	16	48	0.110	0.108	0.062	0.080	0.078	0.143	0.048
5	16	96	0.086	0.085	0.061	0.110	0.053	0.138	0.036
5	16	192	0.055	0.049	0.038	0.092	0.032	0.150	0.023
5	64	48	0.123	0.099	0.039	0.052	0.054	0.061	0.027
5	64	96	0.071	0.090	0.026	0.063	0.033	0.071	0.020
5	64	192	0.054	0.057	0.018	0.050	0.022	0.059	0.013

Formal products run in PyTorch float32 and return float64 NumPy arrays; references use float64 throughout. Both CUDA matmul and cuDNN TF32 are disabled, deterministic algorithms are enabled, and one CPU BLAS thread handles local algebra. A pre-specified same-batch comparison at both checkpoints and $b = 1, 16, 64$ found a maximum absolute difference in a scaled trace observation of 2.23×10^{-6} relative to its reference trace, and a maximum relative vector-product difference of 5.52×10^{-6}, between float32 and float64 products. These checks used separate calls before the formal sweep. Timings on an NVIDIA RTX 3060 Laptop GPU synchronize CUDA before and after each estimator; they include probe generation, minibatch sampling, automatic differentiation, QR, and local estimator algebra, or score-gradient computation for the structural comparator. Training, reference construction, model loading, and warm-up are excluded. The frozen software record specifies PyTorch 2.5.1+cu121 and torchvision 0.20.1+cu121; full commands and hashes accompany the supplemental code.

GGN results. Across all 18 configurations, complete correction has higher point-estimate RMSE than sphere sampling at $b = 1$ and lower RMSE at $b = 16, 64$. At epoch 1 and $q = 96$, the sphere-to-CV RMSE gain changes from 0.88 (95% bootstrap interval $[0.67, 1.13]$) at $b = 1$ to 2.15 $[1.63, 2.86]$ at $b = 16$ and 4.26 $[3.03, 5.72]$ at $b = 64$. The fixed projected implementation takes $t = 2$ head repetitions at every budget, rather than the accuracy- and noise-dependent sufficient schedule of the theorem; its observed losses therefore do not test that schedule. Fresh-product XTrace has higher point-estimate RMSE than correction in all 18 configurations, despite its favorable exact-product synthetic control. The model-score comparator uses richer access: at $b = 16$ it has the lowest RMSE in five of six configurations and the lowest median time in all six. For example, at epoch 5, $q = 96$, it takes 0.241 s with RMSE 0.0356, versus 0.637 s and 0.0532 for the lowest-RMSE vector-oracle method in that configuration (paid selection). At $b = 64$, score norms are slower than the fastest vector method in all six configurations; changing minibatch size must therefore be evaluated by both error and computation.

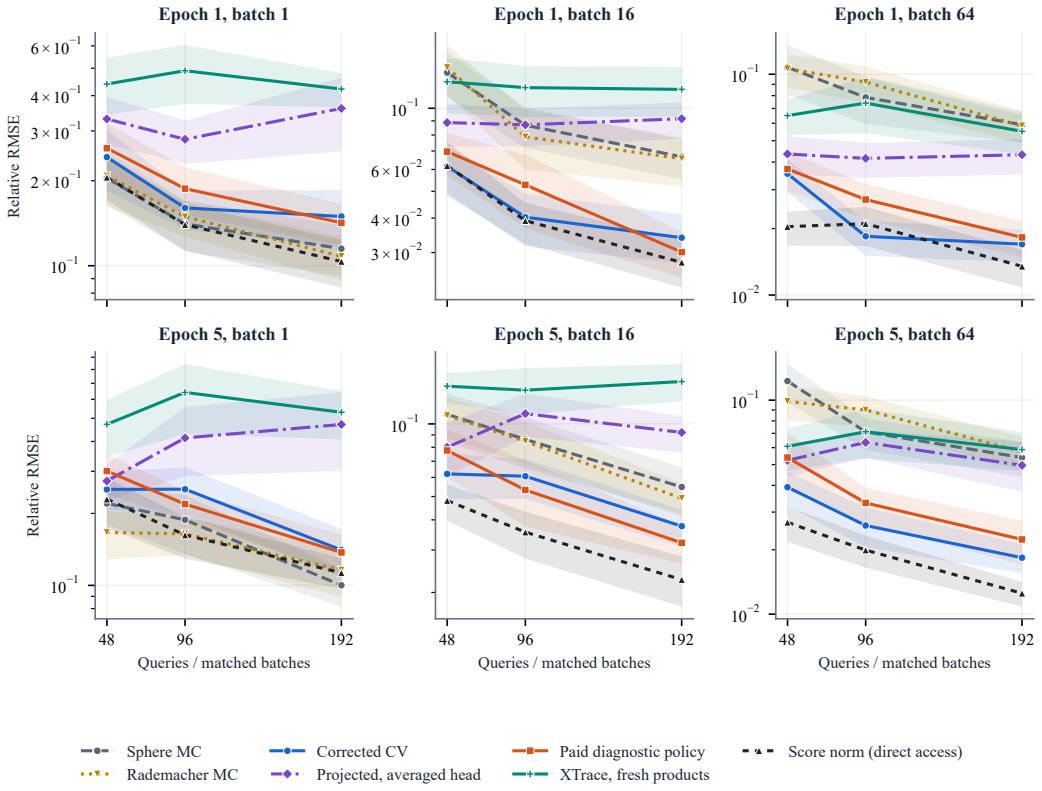


Figure 16: Complete frozen-CNN GGN comparison: both checkpoints, all three batch sizes, and $q = 48, 96, 192$. Each point uses 50 independent trials, with pointwise 95% RMSE bootstrap bands. Each method uses bq sampled-image draws at each point; score norm uses direct model access and matched batch units rather than vector-oracle calls.

Record and replay checks. The three estimator ledgers for paid selection, sensitivity, and learned curvature contain exactly 72,000, 6,400, and 6,300 trials. Independent checks verify their complete configuration grids, separate seeds, source and raw-data hashes, reference traces, error arithmetic, and phase-specific query and sample counts. Replaying repetition zero in all 392 synthetic configurations reproduces every non-time field exactly. A separate replay of all 126 GGN configurations reproduces every non-time field bit for bit on the recorded software and GPU, with zero execution errors; numerical tolerances and differences are retained in the replay report.