Pure Relative Structured Matrix Learning with Square-Root Matvecs

Pahan Dewasurendra
Johns Hopkins University

Abstract

Let $\mathcal{L} \subseteq \mathbb{R}^{n\times m}$ be any known q-dimensional linear matrix family, and let an unknown A be accessible only through products Ax and $A^\top y$. A recent result obtains a nearly optimal approximation from $\mathcal{L}$ in about $\sqrt{q}$ queries, but incurs factor three and additive error proportional to $\|A\|_F$. Whether sublinear query complexity can give pure $(1+\epsilon)$ relative error was left open.

We resolve this problem with a polynomial-time, nonadaptive algorithm. For a Frobenius-orthonormal basis $B_1, \ldots, B_q$, form the basis-invariant partial trace $S_L = \sum_i B_i B_i^\top$. The algorithm queries the leading eigenspace of S_L exactly from the left, then fits the remaining directions from a Gaussian right sketch. If λ_j are the eigenvalues of S_L, its constant-success query complexity is

$$\min_s \left\{ s + C \begin{bmatrix} \max\{1, \lambda_{s+1}\} \log^2(2q) \\ + \lambda_{s+1}/\epsilon \end{bmatrix} \right\}.$$

The transposed profile is also available. Since $\lambda_{s+1} \le q/(s+1)$, every family admits pure relative error with $\widetilde{O}(\sqrt{q/\epsilon})$ queries, independent of ambient dimensions and $\|A\|_F/\mathrm{OPT}$. The key proof is a partial-trace covariance lemma for a Gaussian matrix-valued design. Exact high-leverage queries make its expected Hessian the identity, while the centered regression offset has variance only $O(\lambda_{s+1}\mathrm{OPT}^2/k)$. A median construction amplifies success without estimating residual norms. Finally, a symmetric Wishart lower bound for fixed-sparsity approximation yields $\Omega(\sqrt{q/\epsilon})$ adaptive two-sided queries when $q\epsilon$ is bounded below. Thus the universal rate is optimal up to logarithmic factors in this joint regime. We also show that the transformed profile gives pure relative reducible prediction risk under known input and output covariances. For structured positive-definite precisions, it further controls Gaussian KL error and preconditioned condition number.

1 Introduction

Matrix-vector products are the basic access primitive for implicit Hessians, linear predictors, kernel matrices, and matrix functions. When a matrix A is too large to form, one may still apply $x \mapsto Ax$ and, when an adjoint is available, $y \mapsto A^\top y$. This motivates a statistical question: how many such queries reveal the best structured approximation to A?

This question includes a structured prediction problem. If $Y = AX + \eta$ and only input-output experiments with the unknown linear operator are available, then approximating A under the covariance of X is exactly approximating the reducible prediction risk. It also includes matrix-free curvature compression: automatic differentiation supplies Hessian-vector products without materializing a Hessian, while a structured precision can define a scalable Gaussian approximation or preconditioner. Our results apply agnostically in both cases. The target operator need only be near the proposed structure.

We study a known linear matrix family $\mathcal{L} \subseteq \mathbb{R}^{n\times m}$ of dimension q. Given two-sided matvec access to an arbitrary A, the goal is to return $\widehat{B} \in \mathcal{L}$ such that

$$\|A - \widehat{B}\|_F \le (1+\epsilon) \min_{B\in\mathcal{L}} \|A - B\|_F. \tag{1}$$

The target is agnostic. No stochastic noise model or promise that $A \in \mathcal{L}$ is imposed.

Amsel et al. (2026a) initiated a general theory for this problem. For a finite family $\mathcal{F}$, they give a nearly quadratic improvement from $O(\log|\mathcal{F}|)$ one-sided queries to $\widetilde{O}(\sqrt{\log|\mathcal{F}|})$ two-sided queries. Covering a linear family gives roughly $\sqrt{q}$ queries, but the guarantee is

$$(3+\epsilon)\mathrm{OPT} + \alpha\|A\|_F. \tag{2}$$

They explicitly ask whether every q-dimensional linear family admits pure $(1+\epsilon)$ error with $\widetilde{O}(\sqrt{q}/\operatorname{poly}(\epsilon))$ queries, and whether this can be achieved in polynomial time.

The obstruction is not identifiability. In the realizable case, generic queries often recover a linear family exactly (Halikias and Townsend, 2024). The difficulty is an arbitrary orthogonal residual. A direct Gaussian least-squares fit has an unbiased normal equation, but a few matrix contractions need not embed all q parameter directions. Requiring a conventional subspace embedding returns to order q queries.

Our solution separates the directions that actually obstruct concentration. Let $B_1, \dots, B_q$ be a Frobenius-orthonormal basis of $\mathcal{L}$ and define

$$S_L = \sum_{i=1}^{q} B_i B_i^\top, \qquad S_R = \sum_{i=1}^{q} B_i^\top B_i. \tag{3}$$

These partial traces are basis invariant, positive semidefinite, and have trace q. We query a leading eigenspace of one partial trace exactly. On its orthogonal complement, Gaussian matvecs have both a well-conditioned Hessian and a small regression offset. The first fact costs the remaining spectral leverage times logarithmic factors. The second costs the same leverage divided by ϵ. Optimizing where to split gives square-root complexity.

Contributions. Our principal results are theoretical.

- We give a nonadaptive, polynomial-time algorithm satisfying (1) for every rectangular linear family. Its query complexity is the better of two basis-invariant spectral profiles, one for each query orientation.
- The profile implies a universal $\widetilde{O}(\sqrt{q/\epsilon})$ upper bound. It has no additive error and no dependence on the ambient dimensions or the ratio $\|A\|_F/\mathrm{OPT}$.
- We prove a Gaussian covariance lemma controlled by a partial trace. The proof combines a dimension-free Schatten moment estimate, an exact fourth-moment calculation, and matrix concentration for unbounded summands.
- We convert the adaptive symmetric Wishart lower bound of Amsel et al. (2026b) into $\Omega(\sqrt{q/\epsilon})$ for general q-dimensional linear families when $q\epsilon = \Omega(1)$. This matches our upper bound up to logarithms. A separate GOE construction shows that $\Omega(1/\sqrt{\epsilon})$ queries can be necessary even for a one-dimensional family.

Table 1: General q-dimensional linear families. Logarithmic factors other than $\log(1/\alpha)$ are suppressed.

Result	Queries	Guarantee
Amsel et al.	$\frac{\sqrt{q\log(1/\alpha)}}{\epsilon^2}$	$(3+\epsilon)\mathrm{OPT} + \alpha\|A\|_F$
This paper	$\sqrt{q/\epsilon}$	$(1+\epsilon)\mathrm{OPT}$

- We lift the theorem to covariance-weighted prediction loss without changing its query count. For positive-definite precision matrices, the same output controls both Gaussian KL divergence and preconditioning quality. Circulant, Toeplitz, and known-graph precision families give explicit dimension-adaptive profiles.

Why two sides matter. The partial trace identifies output directions shared by many basis matrices. A left query reveals one such direction exactly, no matter how many coefficients use it. Gaussian right queries then handle the diffuse remainder. Strong one-sided barriers are known for general matrix-free tasks (Halikias et al., 2026). Our estimator makes the complementary roles of the two orientations explicit.

Table 1 separates the guarantee and query dependence of the closest general result. The earlier linear-family bound is obtained by covering a bounded part of $\mathcal{L}$, then running a finite-candidate procedure. Its runtime scales with the cover size. Our algorithm instead solves linear least squares in the original q coefficients.

2 Why a Spectral Split Is Necessary

It is tempting to fit $\mathcal{L}$ using only a right Gaussian sketch, or only a left sketch, then choose whichever orientation has smaller aggregate variance. This cannot prove a square-root theorem. For an orthonormal basis, define the two offset leverages of a residual $R \perp \mathcal{L}$ by

$$V_R(R) = \sum_i \|R^\top B_i\|_F^2, \qquad V_L(R) = \sum_i \|R B_i^\top\|_F^2. \tag{4}$$

A one-sided Gaussian normal equation has total offset variance proportional to the corresponding quantity. One might hope that $\min\{V_R, V_L\} = O(\sqrt{q}\|R\|_F^2)$, or that their product is at most $q\|R\|_F^4$. Both statements are false.

Proposition 1 (Deleted-tangent obstruction)**.** *For every $a, b \geq 1$, there is a linear family of dimension $q = a + b$ and a unit residual $R \perp \mathcal{L}$ such that*

$$V_R(R) = b, \qquad V_L(R) = a.$$

When a and b are comparable, both global orientations have order-q offset variance. Nevertheless, the spectral-

split estimator removes all contamination with one exact query.

Proof. Let $u, e_1, \dots, e_a$ and $v, f_1, \dots, f_b$ be orthonormal systems. Take $R = uv^\top$ and let $\mathcal{L}$ have orthonormal basis

$$\{uf_j^\top : 1 \le j \le b\} \cup \{e_i v^\top : 1 \le i \le a\}. \tag{5}$$

The deleted matrix $uv^\top$ is orthogonal to every basis element. For the first arm, $\|R^\top(uf_j^\top)\|_F = 1$, while the corresponding terms from the second arm vanish. This gives $V_R = b$. The transposed calculation gives $V_L = a$.

The left partial trace is

$$S_L = buu^\top + \sum_{i=1}^{a} e_i e_i^\top.$$

Querying u exactly leaves partial-trace norm one. More strongly, the remaining residual is $(I - uu^\top)R = 0$, so the random-stage offset vanishes. □

This example isolates the role of the deterministic term in (6). It is not included to obtain a conventional subspace embedding. It makes the high-curvature and high-offset directions exact, which leaves a centered problem whose variance is governed by the residual partial trace. A fixed global choice of orientation misses this local structure.

3 Problem and Algorithm

Let $\mathcal{L} \subseteq \mathbb{R}^{n\times m}$ have Frobenius-orthonormal basis $B_1, \dots, B_q$. Write

$$B_\star = P_{\mathcal{L}} A, \qquad R = A - B_\star, \qquad \mathrm{OPT} = \|R\|_F.$$

Then $\langle R, B\rangle_F = 0$ for every $B \in \mathcal{L}$. For the left partial trace S_L, let $\lambda_1^L \ge \lambda_2^L \ge \cdots \ge 0$ be its eigenvalues, padded by zeros. Fix $s \in \{0, \dots, n\}$, let P project onto its first s eigenvectors, set $Q = I - P$, and abbreviate $\lambda = \lambda_{s+1}^L = \|QS_LQ\|_{\mathrm{op}}$.

If $\lambda = 0$, query $A^\top$ on an orthonormal basis of range(P) and minimize $\|P(A-B)\|_F^2$ over $B \in \mathcal{L}$. This recovers $B_\star$ because $QB = 0$ for every $B \in \mathcal{L}$.

Suppose $\lambda > 0$. In addition to the same s left queries, draw $G = [g_1, \dots, g_k] \in \mathbb{R}^{m\times k}$ with independent standard Gaussian entries and issue the k right queries Ag_ℓ. Return the minimum-norm minimizer

$$\widehat{B} \in \arg\min_{B\in\mathcal{L}} \left\{ \|P(A-B)\|_F^2 + \frac{1}{k}\|Q(A-B)G\|_F^2 \right\}. \tag{6}$$

The objective is observable. The left responses give PA, and QAG is obtained from $AG - PAG$. Equation (6) is an ordinary q-variable least-squares problem.

Define the one-orientation profile

$$\mathsf{Q}_L(\epsilon) = \min_{0\le s\le n} \left\{ s + C\left[\max\{1, \lambda_{s+1}^L\} \log^2(2q) + \frac{\lambda_{s+1}^L}{\epsilon} \right] \right\}, \tag{7}$$

where the bracketed term is zero if $\lambda_{s+1}^L = 0$. Define $\mathsf{Q}_R(\epsilon)$ analogously from S_R. The right profile swaps the roles of A and $A^\top$.

Theorem 2 (Pure relative profile)**.** *For every $A \in \mathbb{R}^{n\times m}$, every q-dimensional linear family $\mathcal{L}$, and every $\epsilon \in (0,1)$, the better orientation of* (6) *returns $\widehat{B} \in \mathcal{L}$ satisfying* (1) *with probability at least* 0.9 *using at most*

$$\min\{\mathsf{Q}_L(\epsilon), \mathsf{Q}_R(\epsilon)\}$$

matvec queries. The query vectors are nonadaptive and the estimator runs in time polynomial in the explicit input size of the basis.

For failure probability δ, multiply only the bracketed random-query term in (7) *by $C\log(1/\delta)$, and likewise for the right profile.*

Computation model. The polynomial-time statement uses real arithmetic and an explicit linearly independent basis. A nonorthonormal basis is whitened through its $q \times q$ Frobenius Gram matrix. The partial trace is then formed explicitly, a symmetric eigenproblem is solved, and (6) is a least-squares problem in q unknowns. No eigengap is required: for any computed rank-s projector, the guarantee holds with the directly certifiable residual leverage $\Lambda_L(P)$ in (9). On the covariance event the least-squares Hessian has smallest eigenvalue at least $1/2$. Bit complexity still depends on the conditioning of the supplied basis, which we do not hide inside the query bound.

The leading eigenspace is not merely convenient. It is the best rank-s choice for this estimator. For an arbitrary rank-s orthogonal projector P, define its residual leverage

$$\Lambda_L(P) = \|(I-P)S_L(I-P)\|_{\mathrm{op}}. \tag{8}$$

The same proof as Theorem 2 gives query count

$$s + C\left[\max\{1, \Lambda_L(P)\}\log^2(2q) + \frac{\Lambda_L(P)}{\epsilon}\right]. \tag{9}$$

Proposition 3 (Optimal exact-query subspace)**.** *Among all rank-s orthogonal projectors P,*

$$\min_{\mathrm{rank}(P)=s} \Lambda_L(P) = \lambda_{s+1}^L.$$

Thus the leading eigenspace minimizes both random-query terms in (9).

Proof. For $Q = I - P$, $\Lambda_L(P)$ is the largest Rayleigh quotient of S_L on range(Q). The claim is the Courant–Fischer characterization of λ_{s+1}^L. □

Corollary 4 (Universal square-root bound)**.** *Let* $L_\epsilon = \log^2(2q) + 1/\epsilon$. *Constant-success pure relative error requires at most*

$$O\Big(\sqrt{qL_\epsilon} + \log^2(2q)\Big) = \widetilde{O}\Big(\sqrt{q/\epsilon}\Big) \tag{10}$$

two-sided matvec queries.

Proof. The partial-trace eigenvalues sum to q, so $\lambda_{s+1}^L \le q/(s+1)$. Also $\max\{1, \lambda\} \le 1 + \lambda$. Insert these inequalities into (7) and choose $s = \min\{n, \lfloor\sqrt{qL_\epsilon}\rfloor\}$. If this choice reaches n, the exact queries already cover the row support of $\mathcal{L}$. Otherwise the two s-dependent terms are both $O(\sqrt{qL_\epsilon})$. □

4 Statistical Consequences and Concrete Profiles

The Frobenius objective is not tied to isotropic inputs. Let $\Gamma \succ 0$ and $\Omega \succ 0$ be known input and output weights and define

$$\|M\|_{\Omega,\Gamma}^2 = \|\Omega^{1/2} M \Gamma^{1/2}\|_F^2. \tag{11}$$

Transform the operator and family by $A' = \Omega^{1/2} A \Gamma^{1/2}$ and $\mathcal{L}' = \Omega^{1/2}\mathcal{L}\Gamma^{1/2}$. A right query to A' is implemented by querying A at $\Gamma^{1/2} g$, then multiplying the response by $\Omega^{1/2}$. A left query is implemented analogously. Thus whitening costs no additional oracle calls.

Corollary 5 (Structured prediction risk)**.** *Let* $Y = AX + \eta$, *where* $\mathbb{E}[\eta \mid X] = 0$ *and* $\mathbb{E}[XX^\top] = \Gamma$. *Given two-sided queries to* A, *the transformed profile of Theorem 2 returns* $\widehat{B} \in \mathcal{L}$ *such that*

$$\mathbb{E}\|\Omega^{1/2}(Y - \widehat{B}X)\|_2^2 - \mathsf{N} \le (1+\epsilon) \min_{B \in \mathcal{L}} \Big\{\mathbb{E}\|\Omega^{1/2}(Y - BX)\|_2^2 - \mathsf{N}\Big\}, \tag{12}$$

with probability at least 0.9, *where* $\mathsf{N} = \mathbb{E}\|\Omega^{1/2}\eta\|_2^2$. *The profile uses a basis orthonormal under* (11) *and its corresponding weighted partial traces.*

Proof. The conditional mean assumption gives $\mathbb{E}\|\Omega^{1/2}(Y - BX)\|_2^2 = \mathsf{N} + \|A - B\|_{\Omega,\Gamma}^2$. Apply Theorem 2 to $(A', \mathcal{L}')$ with $\epsilon' = \sqrt{1+\epsilon} - 1 \ge \epsilon/3$, then square its guarantee. Replacing ϵ by ϵ' changes the profile by only a universal constant. □

Table 2: Constant-success profile examples. Tildes hide logarithmic factors.

Family	q	$\|S_L\|_{\mathrm{op}}$	Queries
Circulant	n	1	$\widetilde{O}(1/\epsilon)$
Toeplitz	$2n-1$	H_n	$\widetilde{O}(\log n/\epsilon)$
Graph precision	$n + \|E\|$	$1 + \Delta/2$	$\widetilde{O}((1+\Delta)/\epsilon)$
$a \times b$ block	ab	b	$\min\{a, b\}$ exact

This statement is distribution-free beyond second moments. Gaussian probes are an experimental design chosen by the algorithm, not an assumption on X. The result is therefore an oracle-complexity theorem for the best structured linear predictor under a specified deployment covariance.

For Gaussian uncertainty quantification, the matrix approximation also has a direct distributional consequence.

Corollary 6 (Structured Gaussian precision)**.** *Suppose* $A = A^\top \succeq \mu I$, *every matrix in* $\mathcal{L}$ *is symmetric, and set* $r = (1+\epsilon)\mathrm{OPT}/\mu < 1$. *The output of Theorem 2 is positive definite and satisfies*

$$\mathrm{KL}\Big(N(0, A^{-1}) \,\|\, N(0, \widehat{B}^{-1})\Big) \le \frac{(1+\epsilon)^2 \mathrm{OPT}^2}{4\mu^2(1-r)}, \tag{13}$$

$$\kappa(\widehat{B}^{-1/2} A \widehat{B}^{-1/2}) \le \frac{1+r}{1-r}. \tag{14}$$

Proof. Let $E = A^{-1/2}(\widehat{B} - A)A^{-1/2}$. Then $\|E\|_{\mathrm{op}} \le r$, which proves positive definiteness and (14). If e_i are the eigenvalues of E, the KL divergence is $\frac{1}{2}\sum_i [e_i - \log(1+e_i)]$. For $|e_i| \le r$, the summand is at most $e_i^2/[2(1-r)]$. Finally, $\|E\|_F \le \|A - \widehat{B}\|_F/\mu$. □

Table 2 makes the spectral profile explicit for statistical structures. For real Toeplitz matrices, the normalized diagonal basis gives the ith partial-trace entry $H_n - H_{i-1} + H_{n-1} - H_{n-i}$, so its maximum is H_n. For a known graph precision model, use E_{ii} on vertices and $(E_{ij} + E_{ji})/\sqrt{2}$ on edges. The partial trace is diagonal with entry $1 + \deg(i)/2$. Exact queries can therefore remove hubs before the random stage.

For diagonal matrices, $S_L = I$ and the random stage similarly uses $O(\log^2 q + 1/\epsilon)$ queries, matching specialized dimension-free behavior up to logarithms (Amsel et al., 2026b; Dharangutte and Musco, 2023). For an arbitrary coordinate subspace, the partial traces are row and column degree matrices. Graph degeneracy can give a sharper sparse-only profile (Musco and Ramesh, 2026), while our theorem also covers dense overlapping bases and weighted prediction loss.

Figure 1 is a small falsification check rather than evidence for the theorem. We generated a unit random

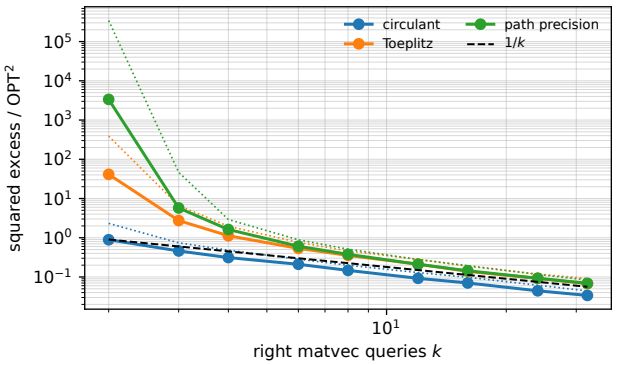

Figure 1: Relative squared excess for three 32×32 families. The dashed line has slope $1/k$.

residual, projected it exactly orthogonal to each family, and solved the right-sketch least-squares problem over 120 independent sketches. The median squared excess follows the predicted $1/k$ behavior once the random design is nonsingular. Dotted curves show the 90th percentile. The supplement gives the complete protocol and seed.

5 Proof of the Upper Bound

We give the main argument and defer moment details and amplification to the supplement. Set $C_i = QB_i$ and define the $q \times q$ Gram matrix

$$K_{ij} = \langle C_i, C_j \rangle_F.$$

Since the original basis is orthonormal, $0 \preceq K \preceq I$. Moreover,

$$\sum_{i=1}^{q} C_i C_i^\top = Q S_L Q \preceq \lambda I. \tag{15}$$

For $g \sim N(0, I_m)$, let $D(g) \in \mathbb{R}^{n \times q}$ have columns $C_i g$.

Lemma 7 (Partial-trace Gaussian covariance)**.** *Suppose $K \preceq I$ and $\sum_i C_i C_i^\top \preceq \tau I$ for $\tau \geq 1$. There is a universal constant C such that, if $g_1, \dots, g_k$ are independent Gaussians and*

$$k \geq C\tau \log^2(2q),$$

then, with probability at least 0.99,

$$\left\| \frac{1}{k} \sum_{\ell=1}^{k} D(g_\ell)^\top D(g_\ell) - K \right\|_{\mathrm{op}} \leq \frac{1}{2}. \tag{16}$$

Proof. The ambient dimensions disappear through Schatten moments. Write $D(g) = \sum_a g_a A_a$, where $A_a = [C_1 e_a, \dots, C_q e_a]$. Direct calculation gives

$$\sum_a A_a A_a^\top = \sum_i C_i C_i^\top \preceq \tau I, \qquad \sum_a A_a^\top A_a = K \preceq I. \tag{17}$$

Both matrices have trace $\operatorname{tr} K \leq q$. Hence, for every integer $p \geq 1$,

$$\operatorname{tr}\left(\sum_a A_a A_a^\top \right)^p \leq \tau^{p-1} q, \qquad \operatorname{tr} K^p \leq q. \tag{18}$$

Rectangular noncommutative Khintchine (Buchholz, 2001), followed by $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_{S_{2p}}$, now gives

$$\left(\mathbb{E} \|D(g)\|_{\mathrm{op}}^{2p} \right)^{1/(2p)} \leq C\sqrt{p\tau}, \qquad p \geq \log(2q). \tag{19}$$

The factors $q^{1/(2p)}$ from (18) are constant. This is the step that removes n and m.

Set $Z = D(g)^\top D(g)$, so $\mathbb{E}Z = K$. For a unit $\theta \in \mathbb{R}^q$, write $C_\theta = \sum_i \theta_i C_i$. Then

$$\theta^\top \mathbb{E} Z^2 \theta = \sum_i \mathbb{E} \langle C_i g, C_\theta g \rangle^2. \tag{20}$$

For a real Gaussian, the identity

$$\mathbb{E}(g^\top M g)^2 = (\operatorname{tr} M)^2 + \operatorname{tr}(M^2) + \operatorname{tr}(MM^\top)$$

and Cauchy–Schwarz yield

$$\begin{aligned} \theta^\top \mathbb{E} Z^2 \theta &\leq \theta^\top K^2 \theta + 2 \sum_i \|C_i^\top C_\theta\|_F^2 \\ &= \theta^\top K^2 \theta + 2 \operatorname{tr}\left(C_\theta^\top \sum_i C_i C_i^\top C_\theta \right) \\ &\leq \theta^\top (K^2 + 2\tau K)\theta. \end{aligned} \tag{21}$$

Consequently,

$$\mathbb{E}(Z - K)^2 \preceq 2\tau K \preceq 2\tau I. \tag{22}$$

For independent copies $X_\ell = Z_\ell - K$, the variance parameter of $\sum_\ell X_\ell$ is at most $2k\tau$. Equation (19) and $\|X_\ell\|_{\mathrm{op}} \leq \|D(g_\ell)\|_{\mathrm{op}}^2 + 1$ also imply

$$\left(\mathbb{E} \max_{\ell \leq k} \|X_\ell\|_{\mathrm{op}}^2 \right)^{1/2} \leq C\tau \log(2qk). \tag{23}$$

Indeed, take the L_{2p} norm with $p = \lceil \log(2qk) \rceil$ and use the union bound in moment form.

The expected-norm inequality of Tropp (2015) applied to the centered $q \times q$ matrices X_ℓ gives

$$\begin{aligned} &\left(\mathbb{E} \left\| \frac{1}{k} \sum_{\ell=1}^{k} X_\ell \right\|_{\mathrm{op}}^2 \right)^{1/2} \\ &\qquad \leq C \sqrt{\frac{\tau \log(2q)}{k}} + \frac{C\tau \log(2q) \log(2qk)}{k}. \end{aligned} \tag{24}$$

For $k = C_0 \tau \log^2(2q)$ and sufficiently large C_0, the right side is at most $1/20$. It decreases for larger k. Markov's inequality proves (16). Supplement A gives the arbitrary-accuracy version. □

The second ingredient controls only the centered offset, which is why the accuracy dependence is $1/\epsilon$ rather than $1/\epsilon^2$.

Lemma 8 (Offset variance)**.** *For any fixed $M \in \mathbb{R}^{n\times m}$, define*

$$\xi_i = \frac{1}{k}\sum_{\ell=1}^{k}\left\{\langle Mg_\ell, C_i g_\ell\rangle - \langle M, C_i\rangle_F\right\}.$$

Under (15)*,*

$$\mathbb{E}\|\xi\|_2^2 \le \frac{2\lambda}{k}\|M\|_F^2. \tag{25}$$

Proof. The real Gaussian quadratic-form identity implies $\mathrm{Var}(\langle Mg, C_i g\rangle) \le 2\|M^\top C_i\|_F^2$. Summing over i gives

$$\sum_i \|M^\top C_i\|_F^2 = \mathrm{tr}\left(M^\top \sum_i C_i C_i^\top M\right) \le \lambda\|M\|_F^2.$$

Independence across the k columns completes the proof. □

Proof of Theorem 2. In orthonormal coordinates of $\mathcal{L}$, the Hessian of (6) is

$$H = H_P + \frac{1}{k}\sum_{\ell=1}^{k} D(g_\ell)^\top D(g_\ell),$$
$$(H_P)_{ij} = \langle PB_i, PB_j\rangle_F.$$

Since $H_P + K = I$, Lemma 7 with $\tau = \max\{1, \lambda\}$ gives $H \succeq I/2$.

At $B_\star$, exact orthogonality $\langle R, B_i\rangle_F = 0$ makes the gradient equal to the centered vector in Lemma 8, with $M = QR$. If d is the coefficient vector of $\widehat{B} - B_\star$, the normal equation is $Hd = \xi$. On the Hessian event,

$$\|\widehat{B} - B_\star\|_F^2 = \|d\|_2^2 \le 4\|\xi\|_2^2.$$

Lemma 8 and Markov's inequality show that $k \ge C\lambda/\epsilon$ makes this at most $2\epsilon\mathrm{OPT}^2$ with constant probability. Finally, orthogonality gives the exact identity

$$\|A - \widehat{B}\|_F^2 = \mathrm{OPT}^2 + \|\widehat{B} - B_\star\|_F^2. \tag{26}$$

Since $\sqrt{1+2\epsilon} \le 1+\epsilon$, the constant-success result follows.

□

We use the metric-median device of Amsel et al. (2026b). Naive confidence amplification would estimate each candidate's residual norm. A Frobenius sketch accurate enough to distinguish $(1+\epsilon)$ error costs $O(1/\epsilon^2)$ extra queries, which would destroy the rate. The following metric argument avoids validation entirely.

Lemma 9 (Validation-free amplification)**.** *Run the Gaussian stage independently an odd number r of times, reusing the same exact responses. For candidate i, let ρ_i be its median Frobenius distance to the r candidates, including itself. Return a candidate minimizing ρ_i. If each candidate satisfies*

$$\|\widehat{B}_i - B_\star\|_F^2 \le \gamma\mathrm{OPT}^2 \tag{27}$$

with probability at least 19/20*, then $r = C\log(1/\delta)$ gives*

$$\|\widehat{B} - B_\star\|_F^2 \le 9\gamma\mathrm{OPT}^2$$

with probability at least $1-\delta$.

Proof. A Chernoff bound gives a strict majority of candidates satisfying (27). Condition on this event. Every good candidate has median radius at most $2\sqrt{\gamma}\mathrm{OPT}$. The selected candidate has no larger radius. Its median ball and the strict majority of good candidates must intersect. The triangle inequality through a candidate in the intersection gives distance at most $3\sqrt{\gamma}\mathrm{OPT}$ from $B_\star$. □

Set $\gamma = 2\epsilon/9$ and apply (26). This proves the confidence statement in Theorem 2. Pairwise distances are computed from the known coefficient vectors, so amplification uses no new kind of oracle query.

6 Lower Bounds and Optimality

The square-root dependence is already necessary for exact recovery. Take the family of all $r \times r$ matrices, so $q = r^2$. Under a continuous Gaussian prior, each adaptive left or right query imposes at most r linear constraints. Fewer than r queries leave a non-atomic posterior on a positive-dimensional affine space. Since pure relative error must be exact when $\mathrm{OPT} = 0$, Yao's principle gives $\Omega(\sqrt{q})$ queries.

Accuracy alone also forces a square-root rate, even for one parameter. This endpoint is useful because the fixed-sparsity construction below requires $q\epsilon$ to be bounded below.

Proposition 10 (One-parameter accuracy lower bound)**.** *For every sufficiently small ϵ, there are an ambient dimension d and a one-dimensional family $\mathcal{L}$ such that every possibly randomized and adaptive two-sided algorithm needs $\Omega(1/\sqrt{\epsilon})$ queries for a constant-probability $(1+\epsilon)$ approximation.*

Proof. Let $\mathcal{L} = \mathrm{span}\{I_d/\sqrt{d}\}$ and draw A from the Gaussian orthogonal ensemble. Thus the law is invariant under $A \mapsto U^\top AU$, its diagonal entries have variance two, and its upper off-diagonal entries have

variance one. Since $A = A^\top$, transpose queries give no extra access.

Fix a deterministic adaptive algorithm. New queries may be orthogonalized against earlier ones because the response on their span is already known. After t orthonormal queries, rotate them to the first t coordinates. The transcript determines the first t rows and columns, while the bottom-right $(d-t)\times(d-t)$ block remains an independent GOE matrix. This remains true for adaptive queries by sequential conditioning and rotational invariance.

The optimal coefficient is $\beta_\star = \langle A, I_d/\sqrt{d}\rangle_F = \mathrm{tr}(A)/\sqrt{d}$. Conditional on the transcript, its unknown part is Gaussian with variance $2(d-t)/d$. If $t \le d/2$, this variance is at least one. Also $\mathbb{E}\|A\|_F^2 = d(d+1)$. With probability at least 0.9, $\mathrm{OPT}^2 \le \|A\|_F^2 \le 10d(d+1)$. Any successful output must satisfy

$$(\widehat{\beta} - \beta_\star)^2 \le ((1+\epsilon)^2 - 1)\mathrm{OPT}^2 \le 3\epsilon \mathrm{OPT}^2. \quad (28)$$

Take $d = \lfloor c/\sqrt{\epsilon} \rfloor$ for a sufficiently small constant c. On the norm event, the right side of (28) is a small universal constant. Gaussian anti-concentration bounds the conditional success probability away from one for every transcript-measurable estimate. Thus $t > d/2 = \Omega(1/\sqrt{\epsilon})$ is necessary. Yao's principle extends the conclusion to randomized algorithms. Supplement C gives explicit constants and the conditioning induction. □

The dependence on ϵ can be coupled to q. The following result is a reparameterization of the adaptive fixed-sparsity lower bound of Amsel et al. (2026b).

Theorem 11 (Joint lower bound)**.** *There are universal constants $c, c_0, c_1 > 0$ such that, for $\epsilon < c_0$ and $q\epsilon \ge c_1$, there is a q-dimensional linear matrix family for which every possibly randomized and adaptive two-sided algorithm using fewer than*

$$c\sqrt{q/\epsilon}$$

queries fails to return a $(1+\epsilon)$ approximation with constant probability.

Proof. Choose a sparsity level u. The cited theorem constructs a symmetric $d \times d$ Wishart hard instance with $d = \Theta(u/\epsilon)$ for any fixed pattern having between constant multiples of u entries in every row and column. It requires $\Omega(u/\epsilon)$ adaptive forward queries. Symmetry makes a transpose query identical to a forward query.

Choose a cyclic u-regular pattern S. Its coordinate family

$$\mathcal{L}_0 = \{B : B = S \circ B\}$$

has dimension

$$q_0 = du = \Theta(u^2/\epsilon). \quad (29)$$

The source lower bound becomes

$$\Omega(u/\epsilon) = \Omega(\sqrt{q_0/\epsilon}). \quad (30)$$

For $q\epsilon$ above a universal constant, choose $u = \lfloor c_2\sqrt{q\epsilon} \rfloor$ with c_2 small enough. The constants in $d = \Theta(u/\epsilon)$ can be chosen so that $c_3 q \le q_0 \le q$. Integer rounding changes only universal constants.

It remains to reach dimension exactly q. Set $p = q - q_0$, let $\mathcal{D}_p$ be the p-dimensional diagonal family on a disjoint block, and define

$$\mathcal{L} = \{B_0 \oplus D : B_0 \in \mathcal{L}_0,\ D \in \mathcal{D}_p\}.$$

Embed a hard matrix as $A' = A_0 \oplus 0$. A query at (x, z) returns $(A_0 x, 0)$, so it is simulated by one query to A_0. The same holds for a transpose query. If an output $\widehat{B}_0 \oplus \widehat{D}$ succeeds, then

$$\begin{aligned}\|A_0 - \widehat{B}_0\|_F^2 &\le \|A_0 - \widehat{B}_0\|_F^2 + \|\widehat{D}\|_F^2 \\ &\le (1+\epsilon)^2 \|A_0 - S \circ A_0\|_F^2.\end{aligned}$$

Restriction to the first block would therefore solve the source problem. Equation (30) and $q_0 \ge c_3 q$ prove the claim. The supplement records the source theorem and rounding explicitly. □

Theorems 2 and 11 match up to logarithmic factors whenever $q\epsilon = \Omega(1)$. We do not claim a matching lower bound in the edge regime $q\epsilon = o(1)$. The exact-recovery and GOE constructions still give separate square-root lower bounds in q and $1/\epsilon$, while the upper theorem remains valid throughout.

7 Related Work and Discussion

Fixed-sparsity approximation is a coordinate-subspace instance of our problem. Gaussian rowwise regression gives $O(s/\epsilon)$ queries for maximum row degree s, with a matching adaptive lower bound (Amsel et al., 2026b). More recent work characterizes exact sparse recovery by graph degeneracy and gives a nearly matching approximation algorithm (Musco and Ramesh, 2026). These methods exploit disjoint coordinate support. Our partial-trace profile applies to arbitrary dense and overlapping basis matrices.

The split between deterministic high leverage and randomized low leverage has an analogue in row-sampled least squares (Larsen and Kolda, 2022, 2026). In that setting, leverage belongs to scalar rows of one design and a sample reveals one response row. Here

one matvec reveals a tensor contraction shared by all q parameters. The partial trace controls the resulting matrix-valued design. Classical active regression needs $\Theta(q/\epsilon)$ scalar labels (Chen and Price, 2019). Sharp sketch-and-solve theory likewise has excess proportional to design rank divided by sketch size (Epperly and Webber, 2026). The two-sided tensor structure is what permits a square-root improvement.

Hierarchical matrix approximation addresses important nonlinear families that are outside our fixed linear-subspace model. The HODLR algorithm of Chen et al. (2025) uses robust recursive peeling and obtains $(1+\epsilon)$ error in $O(k\log^4(n)/\epsilon^3)$ matvecs. The HSS method of Amsel et al. (2026c) obtains an $O(\log(n/k))$ expected squared-error factor with $O(k\log(n/k))$ queries. These specialized guarantees exploit low-rank block geometry that cannot be represented by one q-dimensional linear family. Conversely, they do not give pure relative error for arbitrary dense linear constraints. Recent adaptive matrix-free low-rank work chooses an unknown numerical rank using residual indicators and emphasizes floating-point performance (Smith et al., 2026). It is complementary to our fixed-family agnostic query theorem.

Structured covariance and curvature approximation motivate the weighted and positive-definite consequences in Section 4. We do not claim a sample-based covariance estimator. Corollary 5 instead quantifies how many exact operator experiments recover the best structured prediction rule, while Corollary 6 translates its error into Gaussian and optimization criteria.

Our logarithmic factors come from controlling unbounded Gaussian covariance summands through noncommutative Khintchine and a general expected-norm inequality (Buchholz, 2001; Tropp, 2015). Removing them may require a sharper lower-tail argument tailored to the regularized Hessian. Another open direction is the exact minimax profile outside the regime of Theorem 11. The present result resolves the pure-relative existence and polynomial-time questions while leaving these finer issues.

A Partial-Trace Covariance

This section proves the concentration lemma used in the main paper. We state a slightly more general accuracy form.

Lemma 12 (Partial-trace Gaussian covariance)**.** *Let $C_1, \dots, C_q \in \mathbb{R}^{n\times m}$ satisfy*

$$K = (\langle C_i, C_j\rangle_F)_{i,j} \preceq I, \qquad \sum_{i=1}^{q} C_i C_i^\top \preceq \tau I, \quad (31)$$

where $\tau \geq 1$. For $g \sim N(0, I_m)$, let $D(g) = [C_1 g, \dots, C_q g] \in \mathbb{R}^{n\times q}$. There is a universal constant C such that, for every $\eta \in (0,1)$ and

$$k \geq \frac{C\tau \log^2(2q)}{\eta^2}, \quad (32)$$

independent $g_1, \dots, g_k$ satisfy

$$\left\| \frac{1}{k}\sum_{\ell=1}^{k} D(g_\ell)^\top D(g_\ell) - K \right\|_{\mathrm{op}} \leq \eta \quad (33)$$

with probability at least 0.99*.*

We use the following standard rectangular form of noncommutative Khintchine (Buchholz, 2001). The Gaussian statement follows from the same trace-moment proof as the Rademacher statement. It can also be obtained by self-adjoint dilation.

Lemma 13 (Rectangular Gaussian Khintchine)**.** *For fixed real matrices $A_1, \dots, A_m$ and an integer $p \geq 1$,*

$$\left(\mathbb{E}\left\|\sum_{a=1}^{m} g_a A_a\right\|_{S_{2p}}^{2p}\right)^{1/(2p)} \leq \sqrt{2p-1}\max\left\{ \left\|\left(\sum_a A_a A_a^\top\right)^{1/2}\right\|_{S_{2p}}, \left\|\left(\sum_a A_a^\top A_a\right)^{1/2}\right\|_{S_{2p}} \right\}. \quad (34)$$

Here $\|\cdot\|_{S_r}$ denotes the Schatten r-norm.

Proof of Lemma 12. For each input coordinate $a \in [m]$, define

$$A_a = [C_1 e_a, \dots, C_q e_a] \in \mathbb{R}^{n\times q}.$$

Then $D(g) = \sum_a g_a A_a$, and direct calculation gives

$$\sum_a A_a A_a^\top = \sum_i C_i C_i^\top \preceq \tau I, \qquad \sum_a A_a^\top A_a = K \preceq I. \quad (35)$$

Both matrices in (35) have trace $\operatorname{tr} K \leq q$. Consequently,

$$\operatorname{tr}\left(\sum_a A_a A_a^\top\right)^p \leq \tau^{p-1} q, \quad (36)$$

$$\operatorname{tr} K^p \leq q. \quad (37)$$

Apply Lemma 13, use $\|M\|_{\mathrm{op}} \leq \|M\|_{S_{2p}}$, and take $p \geq \log(2q)$. Since $q^{1/(2p)}$ is bounded by a universal constant and $\tau \geq 1$, we obtain

$$\left(\mathbb{E}\|D(g)\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} \leq C\sqrt{p\tau}. \quad (38)$$

Set $Z = D(g)^\top D(g)$, so $\mathbb{E}Z = K$. We next compute its centered matrix variance. For a unit vector $\theta \in \mathbb{R}^q$, let $C_\theta = \sum_i \theta_i C_i$. Since

$$\theta^\top Z^2 \theta = \sum_{i=1}^{q} \langle C_i g, C_\theta g\rangle^2,$$

the real Gaussian fourth-moment identity gives

$$\begin{aligned}\theta^\top \mathbb{E}Z^2\theta &= \sum_i \left\{\langle C_i, C_\theta\rangle_F^2 + \operatorname{tr}[(C_i^\top C_\theta)^2]\right\} \\ &\quad + \sum_i \|C_i^\top C_\theta\|_F^2 \\ &\leq \theta^\top K^2\theta + 2\sum_i \|C_i^\top C_\theta\|_F^2 \\ &= \theta^\top K^2\theta + 2\operatorname{tr}\left(C_\theta^\top \sum_i C_i C_i^\top C_\theta\right) \\ &\leq \theta^\top (K^2 + 2\tau K)\theta. \end{aligned} \quad (39)$$

The first equality uses $\mathbb{E}(g^\top M g)^2 = (\operatorname{tr} M)^2 + \operatorname{tr}(M^2) + \operatorname{tr}(MM^\top)$. Because $\mathbb{E}(Z-K)^2 = \mathbb{E}Z^2 - K^2$, equation (39) implies

$$\|\mathbb{E}(Z-K)^2\|_{\mathrm{op}} \leq 2\tau. \quad (40)$$

Let $X_\ell = Z_\ell - K$ for independent copies Z_ℓ. The variance parameter of $\sum_\ell X_\ell$ is at most $2k\tau$. We also need the large-deviation parameter

$$L = \left(\mathbb{E}\max_{\ell\leq k}\|X_\ell\|_{\mathrm{op}}^2\right)^{1/2}.$$

For any integer $p \geq \log(2q)$,

$$\begin{aligned}\|\|X_\ell\|_{\mathrm{op}}\|_{L_{2p}} &\leq \|\|D(g_\ell)\|_{\mathrm{op}}^2\|_{L_{2p}} + 1 \\ &= \|\|D(g_\ell)\|_{\mathrm{op}}\|_{L_{4p}}^2 + 1 \leq Cp\tau.\end{aligned}$$

The elementary maximum bound gives

$$\begin{aligned}L &\leq \left(\mathbb{E}\max_{\ell\leq k}\|X_\ell\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} \\ &\leq k^{1/(2p)}\max_{\ell\leq k}\|\|X_\ell\|_{\mathrm{op}}\|_{L_{2p}}.\end{aligned}$$

Taking $p = \lceil \log(2qk) \rceil$ yields

$$L \le C\tau \log(2qk). \tag{41}$$

The expected-norm theorem for independent centered random matrices (Tropp, 2015, Theorem I), applied in dimension q, now gives

$$\left(\mathbb{E} \left\| \frac{1}{k} \sum_{\ell=1}^{k} X_\ell \right\|_{\mathrm{op}}^2 \right)^{1/2} \le C \sqrt{\frac{\tau \log(2q)}{k}} + \frac{C\tau \log(2q) \log(2qk)}{k}. \tag{42}$$

Let $k_0 = C_0 \tau \log^2(2q)/\eta^2$. Since τ may be replaced by $\min\{\tau, q\}$, we may assume $\tau \le q$. At $k = k_0$,

$$\log(2qk) = O(\log(2q) + \log(1/\eta)).$$

The first term of (42) is at most $C\eta/\sqrt{\log(2q)}$. For the second term, use $\eta \log(1/\eta) \le 1/e$. Increasing C_0 makes the root second moment at most $\eta/10$. For $k \ge k_0$, the first term decreases and $\log(2qk)/k$ also decreases. Markov's inequality proves (33) with probability at least 0.99. □

B Upper Bound and Amplification

We first record basis invariance and observability.

Lemma 14 (Basis invariance)**.** *The partial traces $S_L = \sum_i B_i B_i^\top$ and $S_R = \sum_i B_i^\top B_i$ do not depend on the choice of Frobenius-orthonormal basis of $\mathcal{L}$.*

Proof. Any two orthonormal bases are related by an orthogonal matrix U. If $B_i' = \sum_j U_{ij} B_j$, then

$$\sum_i B_i' B_i'^\top = \sum_{j,k} (U^\top U)_{jk} B_j B_k^\top = \sum_j B_j B_j^\top.$$

The right partial trace is identical after transposition. □

Lemma 15 (Observable objective)**.** *Let $P = VV^\top$, where $V \in \mathbb{R}^{n \times s}$ has orthonormal columns. The responses $A^\top V$ and AG determine*

$$\|P(A - B)\|_F^2 + k^{-1} \|Q(A - B)G\|_F^2$$

for every known B.

Proof. The left responses determine $PA = V(A^\top V)^\top$. Therefore PAG and $QAG = AG - PAG$ are known. All terms involving B are known from the basis. □

Lemma 16 (Offset variance)**.** *Suppose $\sum_i C_i C_i^\top \preceq \lambda I$. For fixed M, define*

$$\xi_i = \frac{1}{k} \sum_{\ell=1}^{k} \{ \langle M g_\ell, C_i g_\ell \rangle - \langle M, C_i \rangle_F \}.$$

Then $\mathbb{E}\|\xi\|_2^2 \le 2\lambda \|M\|_F^2/k$.

Proof. For $N = M^\top C_i$, the Gaussian identity $\mathrm{Var}(g^\top N g) = \mathrm{tr}(N^2) + \mathrm{tr}(NN^\top)$ gives

$$\mathrm{Var}(\langle Mg, C_i g \rangle) \le 2 \|M^\top C_i\|_F^2.$$

Summing over i and using independence across ℓ gives

$$\mathbb{E}\|\xi\|_2^2 \le \frac{2}{k} \sum_i \|M^\top C_i\|_F^2 = \frac{2}{k} \mathrm{tr}\left(M^\top \sum_i C_i C_i^\top M \right) \le \frac{2\lambda}{k} \|M\|_F^2.$$

□

Theorem 17 (One-orientation upper bound)**.** *Fix s and let $\lambda = \lambda_{s+1}^L > 0$. For every $\gamma \in (0, 1)$, if*

$$k \ge C \left[\max\{1, \lambda\} \log^2(2q) + \frac{\lambda}{\gamma} \right], \tag{43}$$

then the estimator in the main paper satisfies

$$\|\widehat{B} - B_\star\|_F^2 \le \gamma \mathrm{OPT}^2 \tag{44}$$

with probability at least 19/20.

Proof. Let $C_i = QB_i$ and $D(g) = [C_1 g, \ldots, C_q g]$. Its Gram matrix $K = (\langle C_i, C_j \rangle_F)$ satisfies $K \preceq I$, while

$$\sum_i C_i C_i^\top = Q S_L Q \preceq \lambda I.$$

Set $\tau = \max\{1, \lambda\}$. The Hessian of the objective in the orthonormal coordinates of $\mathcal{L}$ is

$$H = H_P + \frac{1}{k} \sum_{\ell=1}^{k} D(g_\ell)^\top D(g_\ell),$$
$$(H_P)_{ij} = \langle PB_i, PB_j \rangle_F.$$

Because $H_P + K = I$, Lemma 12 with $\eta = 1/2$ implies

$$H \succeq \frac{1}{2} I \tag{45}$$

except with probability 1/100.

Write $R = A - B_\star$. At $B_\star$, the ith normal-equation offset is

$$h_i = \langle PR, PB_i \rangle_F + \frac{1}{k} \sum_{\ell=1}^{k} \langle QRg_\ell, QB_i g_\ell \rangle.$$

Orthogonality gives

$$\langle PR, PB_i\rangle_F + \langle QR, QB_i\rangle_F = \langle R, B_i\rangle_F = 0.$$

Thus h is the centered vector in Lemma 16 with $M = QR$. In particular,

$$\mathbb{E}\|h\|_2^2 \le \frac{2\lambda}{k}\mathrm{OPT}^2. \tag{46}$$

By Markov's inequality and a sufficiently large constant in (43),

$$\|h\|_2^2 \le \frac{\gamma}{4}\mathrm{OPT}^2 \tag{47}$$

except with probability $1/25$.

Let d be the coefficient vector of $\widehat{B} - B_\star$. On (45), the objective is strictly convex and its normal equation is $Hd = h$. Therefore

$$\|\widehat{B} - B_\star\|_F^2 = \|d\|_2^2 \le 4\|h\|_2^2 \le \gamma\mathrm{OPT}^2.$$

A union bound gives success probability at least $1 - 1/100 - 1/25 = 19/20$. □

If $\lambda = 0$, then $QB_i = 0$ for all i. The exact part of the objective has Hessian I and recovers $B_\star$ without random queries. The profile in the main paper follows by taking $\gamma = 2\epsilon$ in Theorem 17 and applying

$$\|A - \widehat{B}\|_F^2 = \mathrm{OPT}^2 + \|\widehat{B} - B_\star\|_F^2.$$

We next prove amplification without estimating OPT or any residual norm.

Lemma 18 (Metric median amplification)**.** *Let $\widehat{B}_1, \dots, \widehat{B}_r$ be independent candidates, where r is odd and each candidate satisfies*

$$\|\widehat{B}_i - B_\star\|_F^2 \le \gamma\mathrm{OPT}^2 \tag{48}$$

with probability at least $19/20$. Let $h = (r+1)/2$. For each i, let ρ_i be the hth smallest value among $\{\|\widehat{B}_i - \widehat{B}_j\|_F : j \in [r]\}$. Return a candidate minimizing ρ_i. There is a universal C such that $r \ge C\log(1/\delta)$ implies

$$\|\widehat{B} - B_\star\|_F^2 \le 9\gamma\mathrm{OPT}^2 \tag{49}$$

with probability at least $1 - \delta$.

Proof. A Chernoff bound shows that at least h candidates satisfy (48) with probability at least $1 - \delta$. Condition on this event. Every good candidate has at least h good candidates within distance $2\sqrt{\gamma}\mathrm{OPT}$, so its radius is at most this value. The selected candidate has radius no larger. Its radius ball contains at least h candidates, and the good set also has at least h members. Since $2h > r$, the two sets intersect. If $\widehat{B}_j$ is in their intersection, then

$$\|\widehat{B} - B_\star\|_F \le \|\widehat{B} - \widehat{B}_j\|_F + \|\widehat{B}_j - B_\star\|_F \le 3\sqrt{\gamma}\mathrm{OPT}.$$

□

Take $\gamma = 2\epsilon/9$ in Lemma 18. Then

$$\|A - \widehat{B}\|_F^2 \le (1 + 2\epsilon)\mathrm{OPT}^2 \le (1+\epsilon)^2\mathrm{OPT}^2.$$

The same exact responses $A^\top V$ are reused for all candidates. Only the Gaussian stage is repeated, proving the confidence statement in the main paper.

C Lower Bounds

C.1 Exact-recovery dimension bound

Proposition 19. *For $q = r^2$, some q-dimensional linear family requires at least r adaptive two-sided queries for any finite multiplicative approximation guarantee with positive worst-case success probability.*

Proof. Take $\mathcal{L} = \mathbb{R}^{r\times r}$ and draw A from a nondegenerate Gaussian distribution on this space. Fix a deterministic adaptive algorithm using $t < r$ queries. Conditional on any transcript up to a given round, the next query vector is fixed. Its response gives at most r scalar linear constraints on the r^2 entries of A, whether the orientation is A or $A^\top$. Sequential Gaussian conditioning shows that after t rounds the conditional law remains Gaussian on an affine space of dimension at least $r^2 - tr > 0$. It is non-atomic. No transcript-measurable output equals A with positive probability.

For every $A \in \mathcal{L}$, the optimum residual is zero. Any finite multiplicative guarantee must output A exactly. Therefore every deterministic algorithm has zero average success under the Gaussian prior. Yao's principle rules out a randomized algorithm with positive worst-case success. □

C.2 One-parameter accuracy bound

We give the deferred-decisions argument and constants for the one-parameter proposition in the main paper. Write $G \sim \mathrm{GOE}_d$ when the diagonal entries of G are independent $N(0,2)$ variables and its entries strictly above the diagonal are independent $N(0,1)$ variables, with symmetry below the diagonal.

Lemma 20 (Adaptive GOE residual)**.** *Fix a deterministic adaptive algorithm that queries a symmetric matrix by matvecs. After t linearly independent queries to $G \sim \mathrm{GOE}_d$, there is a transcript-measurable orthogonal matrix U_t such that, conditionally on the transcript,*

$$U_t^\top G U_t = \Delta_t + \begin{bmatrix} 0_{t\times t} & 0 \\ 0 & H_t \end{bmatrix}, \qquad H_t \sim \mathrm{GOE}_{d-t}, \tag{50}$$

where Δ_t is transcript measurable and H_t is independent of the transcript.

Proof. Repeated queries and components in the span of previous queries reveal no new information, so Gram–Schmidt reduces to orthonormal query directions. We prove the claim by induction. It is immediate for $t = 0$. Suppose it holds after t queries. In the coordinates of (50), the next new query has a nonzero component in the last $d - t$ coordinates. Conditional on the transcript, rotate only those coordinates so that its normalized new component is the first coordinate. Rotational invariance leaves H_t GOE. The response reveals the first row and column of this rotated H_t. By the independent-entry definition of GOE, its remaining principal block is an independent GOE_{d-t-1} matrix. Absorb every revealed entry into Δ_{t+1}. This proves the induction, including when each query is chosen from all earlier responses. □

Proposition 21 (One-parameter accuracy lower bound)**.** *For all sufficiently small $\epsilon > 0$, some one-dimensional linear family requires $\Omega(1/\sqrt{\epsilon})$ adaptive two-sided queries for success probability at least $2/3$ at relative error $1 + \epsilon$.*

Proof. Set $c = 1/100$, let $d = \lfloor c/\sqrt{\epsilon} \rfloor$, and take

$$\mathcal{L} = \mathrm{span}\{I_d/\sqrt{d}\}, \qquad A \sim \mathrm{GOE}_d\,.$$

The input is symmetric, so a transpose query reveals the same response as a forward query. It is enough by Yao's principle to fix a deterministic adaptive algorithm. Its output is $\widehat{\beta} I_d/\sqrt{d}$, while the optimal coefficient is $\beta_\star = \mathrm{tr}(A)/\sqrt{d}$.

After $t \le d/2$ informative queries, Lemma 20 shows that the unknown part of $\beta_\star$, conditionally on the transcript, is

$$\frac{\mathrm{tr}(H_t)}{\sqrt{d}} \sim N\left(0, \frac{2(d-t)}{d}\right).$$

Its conditional variance is at least one. The density of a Gaussian with variance at least one is everywhere at most $1/\sqrt{2\pi}$. Hence, for every transcript-measurable estimate $\widehat{\beta}$ and every $a > 0$,

$$\Pr\big\{|\widehat{\beta} - \beta_\star| \le a \mid \text{transcript}\big\} \le \frac{2a}{\sqrt{2\pi}}. \tag{51}$$

We have $\mathbb{E}\|A\|_F^2 = d(d+1)$, so Markov's inequality gives the event

$$\mathrm{OPT}^2 \le \|A\|_F^2 \le 10d(d+1) \tag{52}$$

with probability at least 0.9. For $\epsilon \le 1$, a successful $(1+\epsilon)$ approximation must obey

$$\begin{aligned}(\widehat{\beta} - \beta_\star)^2 &\le \big((1+\epsilon)^2 - 1\big)\mathrm{OPT}^2 \\ &\le 3\epsilon \mathrm{OPT}^2.\end{aligned}$$

On (52), the right side is at most $30\epsilon d(d+1) \le 60c^2$. Apply (51) with $a = \sqrt{60}c$. The total success probability is at most

$$\frac{1}{10} + \frac{2\sqrt{60}c}{\sqrt{2\pi}} < \frac{1}{6}.$$

For sufficiently small ϵ, we also have $d \ge c/(2\sqrt{\epsilon})$. Therefore success probability $2/3$ requires more than $d/2 = \Omega(1/\sqrt{\epsilon})$ queries under this input distribution. Yao's principle gives the claimed worst-case lower bound for randomized algorithms. □

C.3 Joint accuracy bound

We state the source theorem in the form needed here. It is Theorem 2 of Amsel et al. (2026b), together with the dimension choice in its proof.

Proposition 22 (Fixed-sparsity Wishart lower bound)**.** *Fix $\alpha \in (0,1)$. There are constants $a, b, c_0, c_1 > 0$ such that, for every integer $u \ge 1$ and $\epsilon < c_0$, one can choose*

$$a\frac{u}{\epsilon} \le d \le b\frac{u}{\epsilon} + u \tag{53}$$

and a symmetric Wishart distribution on $d \times d$ matrices with this property. For every mask having between αu and u entries in every row and column, any possibly randomized adaptive algorithm using fewer than $c_1 u/\epsilon$ matvec queries has probability at most $1/25$ of returning a $(1+\epsilon)$ fixed-mask approximation.

The source theorem is stated for forward queries. Its hard matrix is symmetric, so allowing transpose queries does not enlarge the transcript.

Theorem 23 (Joint lower bound for linear families)**.** *There are universal constants $c, c_0, c_2 > 0$ such that, whenever $\epsilon < c_0$ and $q\epsilon \ge c_2$, some q-dimensional linear family requires at least $c\sqrt{q/\epsilon}$ possibly adaptive two-sided queries for constant-probability $(1+\epsilon)$ approximation.*

Proof. For each u and d in Proposition 22, choose a cyclic u-regular mask $S \in \{0,1\}^{d\times d}$. This exists after reducing c_0 if necessary so that $d \ge u$. The coordinate family

$$\mathcal{L}_0 = \{B : B = S \circ B\}$$

has dimension

$$q_0 = du = \Theta(u^2/\epsilon). \tag{54}$$

Proposition 22 gives query lower bound

$$\Omega(u/\epsilon) = \Omega(\sqrt{q_0/\epsilon}). \tag{55}$$

Given q with $q\epsilon$ above a large enough constant, choose $u = \lfloor c_3\sqrt{q\epsilon} \rfloor$ for a sufficiently small universal c_3. Equations (53) and (54) ensure $c_4 q \le q_0 \le q$ for a universal $c_4 > 0$. Rounding changes only constants.

Let $p = q - q_0$. On a disjoint $p \times p$ block, let $\mathcal{D}_p$ be the diagonal coordinate family, and define

$$\mathcal{L} = \{B_0 \oplus D : B_0 \in \mathcal{L}_0,\ D \in \mathcal{D}_p\}.$$

This family has dimension exactly q. Embed a hard input as $A' = A_0 \oplus 0$. Its best approximation in $\mathcal{L}$ is $(S \circ A_0) \oplus 0$, and its optimum residual equals the original one.

Suppose an algorithm succeeds for $\mathcal{L}$. A query to A' with vector (x, z) returns $(A_0 x, 0)$, and a transpose query is identical. Thus every query can be simulated using one query to A_0. The algorithm's output has the form $\widehat{B}_0 \oplus \widehat{D}$. Success implies

$$\begin{aligned}\|A_0 - \widehat{B}_0\|_F^2 &\le \|A_0 - \widehat{B}_0\|_F^2 + \|\widehat{D}\|_F^2 \\ &\le (1+\epsilon)^2 \|A_0 - S \circ A_0\|_F^2.\end{aligned}$$

Hence restriction to the first block solves the source problem. By (55) and $q_0 \ge c_4 q$, the required number of queries is $\Omega(\sqrt{q/\epsilon})$. □

D Profiles for Basic Families

Coordinate subspaces. For a binary mask S, use basis matrices E_{ij} at its nonzero locations. Then

$$\begin{aligned}S_L &= \mathrm{diag}(d_1^{\mathrm{row}}, \dots, d_n^{\mathrm{row}}), \\ S_R &= \mathrm{diag}(d_1^{\mathrm{col}}, \dots, d_m^{\mathrm{col}}).\end{aligned}$$

The spectral split exactly queries selected high-degree rows or columns and uses a Gaussian fit on the remainder.

Full rectangular block. For all matrices supported on an $a \times b$ block, $S_L = bI_a$ on the row block and $S_R = aI_b$ on the column block. Taking $s = a$ in the left profile or $s = b$ in the right profile recovers the block exactly in $\min\{a, b\}$ queries. For $a = b = \sqrt{q}$, this matches Proposition 19.

Row star. For a single active row with q independent entries, $S_L = qee^\top$. One left query in direction e observes the entire structured component and gives exact recovery.

Diagonal family. For the $q \times q$ diagonal family, $S_L = S_R = I_q$. Taking $s = 0$ gives $O(\log^2(2q) + 1/\epsilon)$ queries. The specialized rowwise inverse-Wishart analysis removes the logarithmic term (Amsel et al., 2026b). This example shows that the universal profile is adaptive to diffuse leverage, even though the general matrix-concentration proof is not logarithmically sharp.

Circulant family. Let P_j be the jth cyclic shift permutation. The matrices $B_j = P_j/\sqrt{n}$ form a Frobenius-orthonormal basis, and

$$\sum_{j=0}^{n-1} B_j B_j^\top = \sum_{j=0}^{n-1} \frac{1}{n} I = I.$$

Thus $s = 0$ gives $O(\log^2(2n) + 1/\epsilon)$ queries.

Toeplitz family. For offset $j \in \{-(n-1), \dots, n-1\}$, let T_j be the binary matrix on the jth diagonal and set $B_j = T_j/\sqrt{n - |j|}$. These matrices form an orthonormal basis. Their left partial trace is diagonal. At row i its entry is

$$\begin{aligned}[S_L]_{ii} &= \sum_{j=0}^{n-i} \frac{1}{n-j} + \sum_{h=1}^{i-1} \frac{1}{n-h} \\ &= H_n - H_{i-1} + H_{n-1} - H_{n-i}.\end{aligned}$$

The maximum is H_n, attained at an endpoint. The right partial trace is the reflection of the same diagonal. Taking $s = 0$ therefore gives $O(H_n[\log^2(4n) + 1/\epsilon])$ queries.

Known-graph symmetric precision family. For a graph $G = (V, E)$ on n vertices, take basis E_{ii} for each vertex and $(E_{ij} + E_{ji})/\sqrt{2}$ for each edge. This is Frobenius orthonormal and

$$S_L = S_R = \mathrm{diag}\left(1 + \frac{\deg(1)}{2}, \dots, 1 + \frac{\deg(n)}{2}\right).$$

The no-split profile is controlled by $1 + \Delta/2$. More generally, querying the s highest-degree coordinate directions exactly leaves residual leverage $1 + d_{s+1}/2$, where $d_1 \ge \cdots \ge d_n$ are the degrees.

E Protocol for the Profile Illustration

Figure 1 in the main paper uses 32×32 circulant, Toeplitz, and path precision families with their normalized bases above. For each family, NumPy generator seed 20260729 produces a standard Gaussian matrix. We subtract its exact Frobenius projection onto the family and normalize the remaining residual R to $\|R\|_F = 1$. The structured optimum is set to zero because the coefficient error induced by an orthogonal residual is translation invariant.

For each $k \in \{2, 3, 4, 6, 8, 12, 16, 24, 32\}$, we draw 120 independent standard Gaussian matrices $G \in \mathbb{R}^{32 \times k}$

and solve

$$\widehat{B} \in \arg\min_{B \in \mathcal{L}} \|(R - B)G\|_F^2$$

with NumPy's dense least-squares solver. The reported statistic is $\|\widehat{B}\|_F^2 / \|R\|_F^2$, which equals squared excess over the optimum by orthogonality. Solid curves are medians and dotted curves are 90th percentiles. The supplementary file `reproduce_profile.py` generates the figure using NumPy 2.2.6 and Matplotlib 3.10.9. It uses one CPU process, no external data, and completes in under one minute on a conventional compute node.

References

Amsel, N., Avi, P., Chen, T., Duman Keles, F., Hegde, C., Musco, C., Musco, C., and Persson, D. (2026a). Query efficient structured matrix learning. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 158–194. PMLR.

Amsel, N., Chen, T., Duman Keles, F., Halikias, D., Musco, C., and Musco, C. (2026b). Fixed-sparsity matrix approximation from matrix-vector products. *SIAM Journal on Matrix Analysis and Applications*, 47(2):483–511.

Amsel, N., Chen, T., Duman Keles, F., Halikias, D., Musco, C., Musco, C., and Persson, D. (2026c). Quasi-optimal hierarchically semi-separable matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 47(2):586–621.

Buchholz, A. (2001). Operator Khintchine inequality in non-commutative probability. *Mathematische Annalen*, 319(1):1–16.

Chen, T., Duman Keles, F., Halikias, D., Musco, C., Musco, C., and Persson, D. (2025). Near-optimal hierarchical matrix approximation from matrix-vector products. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2656–2692. SIAM.

Chen, X. and Price, E. (2019). Active regression via linear-sample sparsification. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 663–695. PMLR.

Dharangutte, P. and Musco, C. (2023). A tight analysis of hutchinson's diagonal estimator. In *Symposium on Simplicity in Algorithms*, pages 353–364. SIAM.

Epperly, E. N. and Webber, R. J. (2026). Sharp analysis of sketched least squares and randomized low-rank approximation. *arXiv preprint arXiv:2605.19096*.

Halikias, D., Hochstenbach, M. E., and Townsend, A. (2026). Transpose-free linear algebra. *arXiv preprint arXiv:2606.01335*.

Halikias, D. and Townsend, A. (2024). Structured matrix recovery from matrix-vector products. *Numerical Linear Algebra with Applications*, 31(1):e2531.

Larsen, B. W. and Kolda, T. G. (2022). Practical leverage-based sampling for low-rank tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 43(3):1488–1517.

Larsen, B. W. and Kolda, T. G. (2026). Revisiting approximate leverage score sketching for matrix least squares. *arXiv preprint arXiv:2201.10638*.

Musco, C. and Ramesh, I. (2026). Nearly instance optimal sparse matrix approximation from matrix-vector products. *arXiv preprint arXiv:2606.12179*.

Smith, A. I., Do, E., and Chen, C. (2026). Adaptive, matrix-free low-rank approximation. *arXiv preprint arXiv:2607.06758*.

Tropp, J. A. (2015). The expected norm of a sum of independent random matrices: An elementary approach. *arXiv preprint arXiv:1506.04711*.