FUTURE DELETIONS FROM PAST RETRAINING: EXACT MACHINE UNLEARNING VIA SPARSE RETRAINING ARCHIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact machine unlearning seeks to reproduce retraining outcomes after designated data are removed, but repeated retraining becomes impractical as the number of possible deletion combinations grows exponentially. In this paper, we study the recovery of all deletion targets from *sparse retraining archives* containing models retrained for only a small fraction of deletion requests, without further data access or retraining. For ridge regression, we show that all deletion responses lie in a low-dimensional space and satisfy a shared affine matrix equation when the curvature contributed by deletable data is low rank. This structure yields a convex decoder that recovers unseen retraining targets from archived models. Under independently and uniformly sampled deletion requests, we establish upper and lower bounds showing that, at fixed rank and confidence, the archive size required for exact recovery grows only linearly with the number of deletable groups despite the exponential catalog size. We further derive uniform recovery bounds for noisy archives, characterizing rank-dependent error amplification, and extend the analysis to approximately low-rank deleted curvature through an explicit spectral-tail term. Controlled experiments recover all 4,096 deletion targets from 96 archived models and support the predicted acquisition and error-amplification mechanisms. Further experiments assess noisy recovery in the convex special case and spectral-tail effects.

1 INTRODUCTION

Exact machine unlearning aims to reproduce retraining after designated records are removed (Ginart et al., 2019; Bourtoule et al., 2021; Guo et al., 2020). With M deletable groups, a fixed dataset induces 2^M retraining targets. A *sparse retraining archive* stores deletion subsets and retrained models for only a small fraction of these requests, providing reusable information when downstream users can access model parameters but not records or sufficient statistics. We ask whether this archive and the original model determine unseen targets without further data access or retraining. As a baseline acquisition model, requests are sampled independently and uniformly, with every target defined relative to the same original dataset.

Recovering unseen targets from model parameters is nontrivial because deletion effects are generally nonadditive. Even in ridge regression, joint deletion changes both the curvature and the linear term of the objective, so combining parameter differences through task arithmetic need not reproduce retraining (Ilharco et al., 2023; Yadav et al., 2023). Operator inference fits shared equations from observed solutions (McQuarrie et al., 2023; Wen et al., 2025), but agreement on a finite archive does not by itself guarantee recovery at unseen requests. Inexact archives pose an additional challenge, as small observation errors may be amplified elsewhere in the deletion catalog. These challenges lead to two questions:

Q1 (Acquisition): *How many archived retraining results suffice to recover every deletion target exactly?*
Q2 (Precision): *How accurate must the archived results be to guarantee accurate recovery of every deletion target?*

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.

For ridge regression with rank-r deleted curvature, deletion responses lie in a space of dimension at most r and satisfy a shared affine matrix equation. This structure enables response recovery even with nonunique coefficients (Example A.3). Our contributions are summarized as follows:

- **Acquisition complexity (Q1).** We establish upper and lower bounds on the number of archived retraining results required for exact recovery of the full deletion catalog (Theorem 5.1). At fixed rank and confidence, both bounds scale linearly with the number of deletable groups despite the exponential catalog size.
- **Precision requirements (Q2).** We characterize how bounded archive errors can amplify at unseen deletion targets and how this amplification depends on the deleted-curvature rank (Theorem 5.2). We further extend the guarantee to approximately low-rank deleted curvature through an explicit spectral-tail term (Corollary B.2).
- **Recovery algorithms.** For exact archives, we learn the unknown response space and fit the shared affine equation by convex optimization. For perturbed archives, we develop a robust construction that attains the corresponding precision guarantee.
- **Empirical evaluation.** We conduct controlled low-rank ridge experiments to evaluate all-target recovery, response-space discovery, and the predicted acquisition and precision mechanisms. In our catalog-expansion study, 96 archived models suffice to recover all 4,096 deletion targets.

2 Related Work

Machine unlearning. Exact unlearning exploits additive statistics (Cao & Yang, 2015), deletion-efficient training structures (Ginart et al., 2019; Bourtoule et al., 2021; Brophy & Lowd, 2021), or algorithmic stability (Ullah et al., 2021); ridge heads admit exact updates through sufficient statistics (Quan et al., 2026). Other work studies certified removal (Guo et al., 2020), gradient-based deletion (Neel et al., 2021), generalization and storage (Sekhari et al., 2021), and adaptive requests or query release (Gupta et al., 2021; Ullah & Arora, 2023). We study simultaneous recovery of a fixed ridge learner's deletion targets from independently sampled model archives, without records or sufficient statistics. Our guarantees concern parameter equality or error rather than distributional indistinguishability.

Model composition and deletion interactions. Task arithmetic combines parameter differences (Ilharco et al., 2023), with extensions addressing interference and tangent-space representations (Yadav et al., 2023; Ortiz-Jimenez et al., 2023). Task Simplex Arithmetic ensembles task vectors for unlearning (Dong et al., 2025), while SIFT-Masks enables exact unlearning through sign-constrained finetuning and merging (Kuo et al., 2025). Influence functions approximate training-point effects using derivatives (Koh & Liang, 2017), and group influence analysis captures higher-order deletion effects (Basu et al., 2020). We retain the ridge training rule and infer its nonadditive deletion responses from archived solutions through an exact affine equation.

Operator inference. Operator inference learns reduced operators from observed states (Peherstorfer & Willcox, 2016), including parametric differential and matrix equations (McQuarrie et al., 2023; Wen et al., 2025). Designed snapshots can ensure exact operator identification (Rosenberger et al., 2026). Our decoder adopts the shared-equation perspective and studies the information available in passive random deletion archives. We establish acquisition bounds for recovering every deletion response, including cases where the equation coefficients remain nonunique, and quantify worst-case error amplification under bounded archive perturbations. Both lower bounds are realized by ridge datasets fixed before acquisition.

3 Problem Formulation

3.1 Ridge learning and deletion requests

Let $D = ((x_i, y_i))_{i=1}^N$ be an indexed dataset, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, and consider the model $f_\theta(x) = x^\top \theta$. Let $\mathcal{G} = (G_1, \ldots, G_M)$ consist of disjoint nonempty groups of record indices;

records outside their union are always retained. A deletion request $J \subseteq [M]$ leaves the index set $I_J = [N] \setminus \bigcup_{j \in J} G_j$. For a fixed public regularization parameter $\lambda > 0$, the retraining target is:

$$\theta_J = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{\lambda}{2} \|\theta\|_2^2 + \frac{1}{2} \sum_{i \in I_J} (x_i^\top \theta - y_i)^2 \right\}. \tag{1}$$

The regularization parameter remains fixed across requests. Write $\theta_0 = \theta_\varnothing$ for the original parameters and $t_J = \theta_J - \theta_0$ for the deletion response. All 2^M targets refer to the same fixed dataset and groups.

Define the original and deleted curvature matrices by

$$H = \lambda I_d + \sum_{i=1}^{N} x_i x_i^\top, \qquad H_{\mathrm{del}} = \sum_{j=1}^{M} \sum_{i \in G_j} x_i x_i^\top. \tag{2}$$

Let $\mathcal{C}_{M,r}$ denote the class of instances $(D, \mathcal{G})$ satisfying $\mathrm{rank}(H_{\mathrm{del}}) \le r$, over all $d \ge \max\{1, r\}$, $N \ge M$, and $\lambda > 0$. This class places no uniform bounds on features, labels, or conditioning. The rank condition concerns deleted curvature; H may have full rank. Neither matrix is observed.

3.2 Past retraining observations

Definition 3.1 (Past retraining observations)**.** For a fixed instance, the archive contains the exact original parameters and n deletion requests with possibly perturbed retraining results:

$$\mathcal{O}_n^\eta = \big(\theta_0, (A_i, \widetilde{\theta}_{A_i})_{i=1}^n\big), \qquad \widetilde{\theta}_{A_i} = \theta_{A_i} + e_i, \qquad \|e_i\|_2 \le \eta. \tag{3}$$

Here $\eta \ge 0$, and $\eta = 0$ gives exact observations.

We consider passive acquisition: the instance is fixed before $A_1, \dots, A_n$ are sampled independently and uniformly from $2^{[M]}$. Sampling is with replacement, and repeated requests count toward n. Errors may depend jointly on the instance and the entire mask sequence, with additional randomness allowed, provided the bound holds almost surely. The separately supplied θ_0 remains exact even when a sampled empty request is perturbed.

The processor receives only these observations and the public problem parameters. Requests identify groups without revealing their records. The processor has no access to records, group memberships, sufficient statistics, derivatives, the Hessian, a response basis, or additional model queries. It returns estimates $\widehat{\theta}_J$ for deletion requests $J \subseteq [M]$. Formal conventions for processor families, randomness, and measurability are given in Appendix C.

3.3 Q1: Acquisition complexity

Definition 3.2 (Exact acquisition complexity)**.** For $0 < \delta \le 1/4$, define $n^\star_{\mathrm{exact}}(M, r, \delta)$ as the smallest $n \in \mathbb{N}_0$ for which a processor family satisfies, for every fixed instance in $\mathcal{C}_{M,r}$,

$$\Pr\Big\{ \widehat{\theta}_J = \theta_J \text{ for all } J \subseteq [M] \Big\} \ge 1 - \delta \tag{4}$$

under exact observations $\mathcal{O}_n^0$. Set $n^\star_{\mathrm{exact}} = +\infty$ if no such budget exists.

The guarantee is uniform over admissible d, N, λ and covers requests selected after archive observation.

3.4 Q2: Precision requirements

For precision analysis, let $\mathcal{C}_{M,r,\kappa,B} \subseteq \mathcal{C}_{M,r}$ additionally satisfy

$$H \preceq \kappa \lambda I_d, \qquad \sum_{i=1}^{N} y_i^2 \le \lambda B^2, \tag{5}$$

for public $\kappa \ge 1$ and $B \ge 0$. These spectral and label-energy bounds apply only to the robust class.

Definition 3.3 (Uniform recovery radius)**.** The radius $\mathcal{R}^{\mathrm{hp}}_{n,M,r,\kappa,B}(\eta;\delta)$ is the infimum of $\epsilon \geq 0$ for which a processor family satisfies, for every instance in $\mathcal{C}_{M,r,\kappa,B}$ and every admissible error rule,

$$\Pr\left\{\max_{J\subseteq[M]} \|\widehat{\theta}_J - \theta_J\|_2 \leq \epsilon\right\} \geq 1-\delta. \tag{6}$$

The guarantee is uniform over admissible d, N, λ. Probability includes acquisition and any processor or error randomness; the infimum of an empty set is $+\infty$.

Q1 asks for the smallest observation budget permitting exact recovery; Q2 asks for the smallest recovery radius at a given budget and noise level. Appendix C specifies the decoder conventions.

4 Main Intuition

Low-rank deleted curvature constrains both the response space and the equations governing deletion responses. These constraints yield a three-step recovery method.

Step 1: Identify the response structure. For the deletion responses t_J defined in Section 3.1, define the private group statistics as:

$$H_j = \sum_{i\in G_j} x_i x_i^\top, \qquad b_j = \sum_{i\in G_j} y_i x_i, \qquad a_j = H_j\theta_0 - b_j. \tag{7}$$

The ridge normal equations give:

$$R_J t_J = \sum_{j\in J} a_j, \qquad R_J = H - \sum_{j\in J} H_j \succeq \lambda I_d. \tag{8}$$

The dependence of retained curvature on J makes joint deletion responses generally nonadditive. The rank constraint nevertheless places every response in a common low-dimensional space.

Lemma 4.1 (Response geometry)**.** *For every instance in $\mathcal{C}_{M,r}$,*

$$S := \operatorname{span}\{t_J : J \subseteq [M]\} \subseteq H^{-1}\operatorname{range}(H_{\mathrm{del}}), \qquad \dim S \leq r. \tag{9}$$

The lemma motivates learning a response basis from archived parameter differences. Its proof uses the private statistics only for analysis (Appendix A.1).

Step 2: Fit a shared response equation. Suppose the observed differences span S, and let $V \in \mathbb{R}^{d\times s}$ be an orthonormal basis, where $s = \dim S > 0$. Writing $t_J = Vu_J$, the projected normal equations take the form:

$$\left(B_0 - \sum_{j\in J} C_j\right) u_J = \sum_{j\in J} v_j. \tag{10}$$

Crucially, the coefficients B_0, $\{C_j\}_{j=1}^M$, and $\{v_j\}_{j=1}^M$ are shared across all deletion requests, so archived responses provide constraints on a single response equation rather than on independent targets. One valid coefficient tuple is

$$B_0 = \frac{V^\top H V}{\lambda}, \qquad C_j = \frac{V^\top H_j V}{\lambda}, \qquad v_j = \frac{V^\top a_j}{\lambda}, \tag{11}$$

and satisfies

$$B_0 \succeq I_s, \qquad C_j \succeq 0, \qquad \sum_{j=1}^M C_j \preceq B_0 - I_s. \tag{12}$$

Observed coordinates $u_{A_i} = V^\top t_{A_i}$ make Eq. (10) linear in the unknown coefficients. Fitting these equations under Eq. (12) determines a candidate response function.

Step 3: Recover an unseen deletion target. Given fitted coefficients $\widehat{B}_0, \widehat{C}_j, \widehat{v}_j$, the prediction for a new deletion request J is

$$\widehat{\theta}_J = \theta_0 + V \left(\widehat{B}_0 - \sum_{j \in J} \widehat{C}_j \right)^{-1} \sum_{j \in J} \widehat{v}_j. \tag{13}$$

The semidefinite constraints ensure invertibility at every request. If $S = \{0\}$, every target equals θ_0.

These constraints make every predicted target well defined, but they do not alone ensure correctness at unseen requests. The key is that violations of the shared equation admit low-degree polynomial representations, allowing random archives to enforce the equation at every deletion request. Section 5 quantifies the required archive size and the effect of observation errors.

Response recovery does not require unique coefficients. Example A.3 exhibits distinct coefficient tuples with the same original Hessian and identical responses at every deletion request, demonstrating nonuniqueness beyond a common scaling ambiguity.

5 Theoretical Analysis and Algorithm Design

We first establish acquisition and precision bounds, then give the decoders that attain the upper bounds. Write $p_s = (M+1)s(s+1)/2 + Ms$ for the coefficient count in an s-dimensional response space. Set $q = r + 1$, $N_q = \sum_{j=0}^{\min(q,M)} \binom{M}{j}$, and $0 < \delta \leq 1/4$.

5.1 Acquisition complexity

Theorem 5.1 (Exact acquisition complexity)**.** *There exist universal constants $c_{\mathrm{ex}}, C_{\mathrm{ex}} > 0$ such that, for all $M, r \geq 1$,*

$$n^{\star}_{\mathrm{exact}}(M, r, \delta) \leq C_{\mathrm{ex}}\, 2^r [Mr^2 + \log(1/\delta)]. \tag{14}$$

The exact decoder in Section 5.3, with the batch sizes in Eq. (A.5), achieves this upper bound. Conversely, for $M > r \geq 1$,

$$n^{\star}_{\mathrm{exact}}(M, r, \delta) \geq c_{\mathrm{ex}}\, 2^r [M - r + \log(1/\delta)]. \tag{15}$$

For $r = 0$, every deletion target equals θ_0, so $n^{\star}_{\mathrm{exact}}(M, 0, \delta) = 0$.

At fixed rank and confidence, both bounds scale linearly with M, despite the 2^M deletion requests. The bounds share the factor 2^r but retain a polynomial gap in rank. Appendix A.2 proves sufficiency, while Appendices A.3–A.4 establish necessity through a ridge instance fixed before acquisition.

Acquisition mechanism. Whitening the deleted-curvature space expresses each deletion response as $N(z)/D(z)$ on $z \in \{0,1\}^M$, where $D(z) > 0$ and both polynomials have degree at most r. A nonzero degree-h Boolean polynomial is nonzero on at least a 2^{-h} fraction of the cube (O'Donnell, 2014). Thus, whenever the observed responses span a proper subspace of S, a fresh uniform mask reveals a missing response direction with probability at least 2^{-r}. After S is recovered, any missing direction in the shared coefficient-equation system is detected with probability at least $2^{-(r+1)}$. Rank-growth arguments then yield the batch sizes in Eq. (A.5), after which all feasible fits agree on every deletion response even when their coefficient vectors are not unique.

Acquisition lower bound. The lower-bound construction uses r gate groups and $L = M - r$ terminal groups. A deletion mask reveals terminal information only when all gate groups are deleted, which occurs with probability 2^{-r} under uniform acquisition. The number of informative observations is therefore $\mathrm{Binomial}(n, 2^{-r})$. If their terminal-incidence matrix is rank deficient, multiple terminal parameter choices remain consistent with the same archive while producing different target heads. Averaging over terminal parameters yields, for each processor, a hard instance fixed before acquisition. Appendices A.3–A.4 give the ridge construction and prove Eq. (15).

5.2 PRECISION REQUIREMENTS

Theorem 5.2 (Sparse archive precision). *There exists a universal constant $C_{\mathrm{rob}} > 0$ such that, for the robust class with $r \geq 1$, the observation count*

$$n \geq C_{\mathrm{rob}}\, 9^q M r^2 \log(r+1)[q + \log(Mr/\delta)] \tag{16}$$

implies

$$\mathcal{R}^{\mathrm{hp}}_{n,M,r,\kappa,B}(\eta;\delta) \leq \min\{B, 2\sqrt{2}\kappa^q \sqrt{N_q}\eta\}. \tag{17}$$

For even $M \geq 2$ and all $n, r \geq 1$, an actual-record family with $\kappa = 4$ gives

$$\mathcal{R}^{\mathrm{hp}}_{n,M,r,4,B}(\eta;\delta) \geq \min\left\{\frac{B\Delta_r}{2}, \frac{\eta}{3^q}\left(\frac{2M}{\log(2n/\delta)}\right)^{q/2}\right\}, \tag{18}$$

where $\Delta_r \geq 64^{-q}$ is defined in Eq. (B.28) for the balanced-sign family in Appendix B.4.

At fixed r and δ, the sufficient robust archive size is $O_{r,\delta}(M \log M)$. For $\kappa = 4$ and even $M \to \infty$, budgets satisfying Eq. (16) and $n \leq M^C$ for fixed $C > 1$ yield an amplification rate of $M^{(r+1)/2}$ up to logarithmic factors and rank-dependent constants, whenever both bounds lie on their noise-linear branches. Thus, sparse acquisition can coexist with substantial error amplification. Appendix B gives the proof and a sufficient noise window; Appendix D.3.2 evaluates the regime numerically.

Precision upper bound. We compare the true response with any fit compatible with the perturbed archive. Their combined response space has dimension at most $2r$, so the comparison is independent of the ambient dimension. Substituting the true rational response into the fitted affine equation yields a polynomial residual P of degree at most $q = r + 1$, whose norm is at most $2\kappa\eta$ on every observed mask. Hypercontractivity together with a VC-dimension and epsilon-net argument ensures that uniformly sampled masks detect any residual with sufficiently large L^2 norm (O'Donnell, 2014; Goldberg & Jerrum, 1995; Haussler & Welzl, 1987). Consequently, the observed residual constraints control every compatible residual over the full Boolean cube. A Boolean L^2-to-uniform bound contributes the factor $\sqrt{N_q}$, and the spectral assumptions transfer this residual control to parameter error, yielding Eq. (17). The event depends only on the sampled masks and therefore remains valid for fits and bounded errors chosen after the entire mask sequence is observed.

Precision lower bound. We construct two fixed ridge instances that are nearly indistinguishable on typical archived requests but remain separated at a designated target. For balanced signs $s_j \in \{-1, 1\}$, let $\ell_J = M^{-1}\sum_{j\in J} s_j$. A path-shaped ridge construction makes the parameter gap at request J scale as $|\ell_J|^q$ relative to the target separation γ. Under uniform sampling, sign cancellation keeps $\max_{i\leq n}|\ell_{A_i}|$ of order $\sqrt{\log(2n/\delta)/M}$ with probability at least $1-\delta$. Choosing γ accordingly makes every observed parameter pair lie within 2η, so admissible perturbations can map the two archives to identical midpoint observations while preserving a gap γ at the designated target. No processor can then recover both instances within error below $\gamma/2$, which yields Eq. (18). Both instances are fixed before acquisition.

Approximately low-rank curvature. We extend the precision guarantee to deleted curvature satisfying the public tail bound $\sigma_{r+1}(H_{\mathrm{del}}) \leq \varepsilon$ (Corollary B.2). Projecting the deletable features onto the leading r curvature directions defines a comparison instance whose retrained parameters differ from the original ones by at most $\rho = B(1 + 2\sqrt{\kappa - 1})\sqrt{\varepsilon/\lambda}$ uniformly over targets. The corresponding deletion responses differ by at most 2ρ. Thus, the observed displacements approximate the comparison responses with effective error $\eta + 2\rho$. Applying the robust guarantee gives the all-target bound

$$\min\{B,\, 2\rho + K_r(\eta + 2\rho)\}, \qquad K_r = 2\sqrt{2}\,\kappa^{r+1}\sqrt{N_{r+1}}.$$

The spectral tail thus contributes both approximation error and amplified archive error. The projection is used only in the analysis. The processor requires only the original archive and the public tail bound, not the projection basis. Appendix B.7 gives the full construction and proof.

5.3 Recovery algorithms

Learning the response space. Split exact observations into independent batches of sizes n_1, n_2 specified in Appendix A.2. Ordered Gram–Schmidt on the first batch of displacements yields an orthonormal basis $V \in \mathbb{R}^{d\times s}$.

Fitting the shared equation. For second-batch masks, set $u_i = V^\top t_{A_i}$. If these displacements lie in $\mathrm{range}(V)$, fit symmetric matrices $B_0, C_1, \ldots, C_M \in \mathbb{R}^{s\times s}$ and vectors $v_j \in \mathbb{R}^s$ by solving:

$$\begin{aligned} \min \quad & \|B_0\|_F^2 + \sum_{j=1}^{M} (\|C_j\|_F^2 + \|v_j\|_2^2) \\ \text{subject to} \quad & B_0 \succeq I_s, \quad C_j \succeq 0, \quad \sum_j C_j \preceq B_0 - I_s, \\ & \left(B_0 - \sum_{j\in A_i} C_j\right) u_i = \sum_{j\in A_i} v_j, \quad n_1 < i \le n_1 + n_2. \end{aligned} \tag{19}$$

Answering a deletion request. The fitted coefficients answer requests through Eq. (13) in $O(|J|s^2 + s^3 + ds)$ arithmetic operations. The true projected coefficients establish feasibility once the response space is recovered, and the PSD constraints ensure every target system is invertible.

Robust recovery construction. Observation errors can introduce directions outside the true response space. The robust construction therefore fits coefficients directly in the original d-dimensional space. With $L_J = B_0 - \sum_{j\in J} B_j$ and $w_J = \sum_{j\in J} w_j$, impose

$$\begin{aligned} & I_d \preceq B_0 \preceq \kappa I_d, \quad B_j \succeq 0, \quad \sum_j B_j \preceq B_0 - I_d, \quad \mathrm{rank}\Big[\sum_j B_j, w_1, \ldots, w_M\Big] \le r, \\ & \|L_{A_i}(\widetilde{\theta}_{A_i} - \theta_0) - w_{A_i}\|_2 \le \kappa\eta \quad (1 \le i \le n). \end{aligned} \tag{20}$$

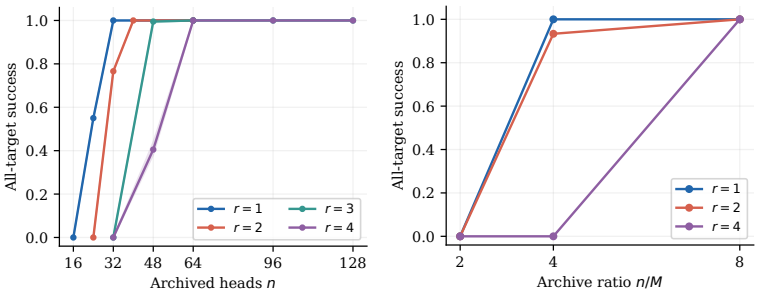
The true normalized statistics ensure feasibility. A minimum-norm feasible tuple predicts $\widehat{\theta}_J = \theta_0 + L_J^{-1} w_J$ and attains the precision bound, with the radius capped at B by returning the zero parameter vector when needed. The rank constraint is automatic when $d = r$; efficient solution of the general nonconvex problem remains open. Appendix C specifies minimizer selection, measurability, lookup, and fallback rules.

6 Experiments

6.1 All-Target Recovery from Sparse Archives

We first test whether the exact decoder can recover the entire deletion catalog from incomplete random archives, including targets that never appear in the archive. Each method receives the same original model parameters and retraining observations under iid uniform deletion masks. We compare the exact decoder with linear mask-to-displacement regression and observation lookup. All methods use exact lookup on sampled targets. An exhaustive retraining oracle evaluates $\max_J \|\widehat{\theta}_J - \theta_J\|_2$, with success threshold $10^{-7}\max\{1, \max_J \|\theta_J\|_2\}$. We refer to satisfaction of this criterion as numerical all-target recovery. We include all solver outcomes and use $n_1 = \lfloor n/3 \rfloor$ and $n_2 = n - n_1$, below the worst-case sufficient budgets. Appendix D provides data construction, seeds, repetition counts, solver settings, and threshold-sensitivity results for all studies.

All-target recovery on real representations. We construct controlled ridge tasks from visual and language representations of CIFAR-10/100, KnowUndo, and TOFU (Krizhevsky, 2009; Radford et al., 2021; Tian et al., 2024; Maini et al., 2024). Gaussian projection to r dimensions enforces the low-rank setting. Visual tasks use binary category targets, while language tasks predict $\log(1 + \text{answer word count})$ from question representations. Figure 1(a) shows a transition to all-target recovery, using paired masks across datasets at ranks one and two. Linear regression and lookup achieve no all-target recovery at the tested budgets, whereas fitting the shared response equation recovers unobserved targets.



(a) Real representations. (b) Catalog expansion.

Figure 1: All-target recovery from sparse archives. (a) Mean success frequencies across four datasets. Bands show their range. (b) Catalog-expansion curves coincide across $M = 8, 10, 12$.

Table 1: All-target success (%) for matched affine fits.

r	n	Structured	Symmetric, no PSD	OpInf-style LS
2	32	70.0	70.0	0
2	48	100	100	100
4	48	45.8	45.8	0
4	64	100	100	53.3

Affine fitting and symmetry ablation. On matched archives, the structured and symmetric fits achieve similar recovery frequencies and outperform unrestricted affine least squares at intermediate budgets (Table 1). In some unsuccessful configurations, the structured fit has substantially smaller maximum errors than the symmetric comparator. A separate comparison with common normalization and endpoint constraints likewise shows similar recovery frequencies and smaller PSD errors in some configurations (Appendix D.2.3).

Response-space discovery and catalog expansion. We embed low-rank deleted curvature in unknown subspaces of larger ambient spaces with a full-dimensional undeletable background. At sufficiently large tested budgets, the decoder recovers every target across these instances. The recovered response spaces are generally not invariant under H, consistent with Lemma 4.1. As the catalog expands, success curves coincide across the tested group counts when budgets are normalized by M (Figure 1(b)). Table 2 pools the tested ranks and shows numerical all-target recovery with a linear archive budget, while most targets remain unobserved. The exact decoder supports on-demand queries, with fitting costs growing with rank and varying less with ambient dimension at fixed small rank over the tested range. Appendix E reports timings, preparation costs, storage, statistics-based comparisons, and amortization thresholds.

6.2 Acquisition and Precision Scaling

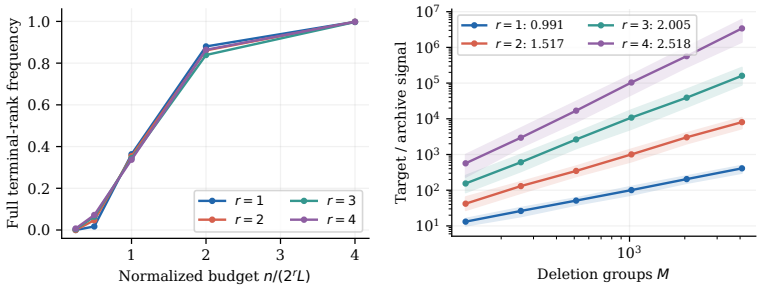
We next examine the mechanisms underlying the acquisition and precision lower bounds in Theorems 5.1 and 5.2. The gate family tests the predicted acquisition scaling, while the balanced-sign family tests the rank-dependent amplification of archive errors.

Gate activation and acquisition. We vary the rank, terminal count $L = M - r$, and observation budget in the gate family, using the normalized scale $n/(2^r L)$. Figure 2(a) shows similar transitions to full terminal-incidence rank across the displayed ranks. This transition is consistent with the predicted acquisition scale: only requests that activate all gates reveal terminal parameters. The diagnostic measures observability in the lower-bound family.

Balanced-sign precision amplification. For the balanced-sign family, we fix the observation budget and vary the number of deletion groups and rank. We compute the ratio of a target's parameter separation to the largest separation over observed requests for each fixed instance pair, using a stable

Table 2: All-target recovery with a linear archive budget.

M	Observations	Unseen targets	Successes	Max. parameter error
8	64	193–204	90/90	6.667×10^{-14}
10	80	943–951	90/90	1.518×10^{-14}
12	96	3999–4003	90/90	8.413×10^{-14}



(a) Gate acquisition. (b) Precision.

Figure 2: Scaling in the lower-bound families. (a) Full terminal-incidence rank frequency. (b) Target-to-archive amplification: medians and 10th–90th percentile bands.

resolvent identity. Figure 2 shows approximately power-law growth in this ratio, with slopes close to the predicted $(r+1)/2$ exponent. Higher rank produces stronger amplification, while the absolute target separation decreases, reflecting the shrinking signal window. Thus, the constructed pairs are much harder to distinguish at sampled masks than at the designated target.

6.3 Recovery under Archive Noise and Spectral Tails

On unprojected CLIP features with full-rank deleted curvature, an archive-only SVD heuristic truncates the response basis. With the larger tested basis and archive budget, its PSD fit improves mean worst unseen-target parameter error and prediction RMSE relative to linear regression on both CIFAR tasks. Recovery remains approximate, and some affine fits perform worse than lookup at the smaller budget. The spectral-tail bound reduces to the trivial radius B in all tested configurations (Appendix D.4.2).

Noisy recovery and controlled spectral tails. When $d = r$, the robust rank constraint is automatic and fitting is convex. In the tested low-dimensional instances, worst all-target errors increase with archive noise. With controlled spectral tails, projection discrepancies follow the predicted $\sqrt{\varepsilon}$ scaling, while a learned-subspace heuristic exhibits increasing recovery error as the tail grows (Figure 5). These experiments evaluate the convex special case and the spectral-tail mechanism.

When $d \gg r$, an SVD-based heuristic improves on the initial-model reference at small noise, but subspace estimation degrades as noise grows and some exploratory stress-test fits exceed the reference error. These results concern the heuristic rather than the theorem's global rank-constrained decoder (Appendix D.4.4).

7 Conclusion

For ridge regression with low-rank deleted curvature, sparse retraining archives enable exact catalog recovery with archive size linear in the number of groups at fixed rank and confidence. Our analysis bounds rank-dependent noise amplification and spectral-tail error, with controlled experiments supporting these mechanisms. Future work includes nonuniform acquisition, efficient high-dimensional robust recovery, and extensions beyond ridge regression.

AI Assistance Disclosure

Generative AI tools were used to assist with language editing and polishing of the manuscript, including improvements to grammar, clarity, concision, and readability. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of the paper.

References

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724, 2020.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy*, pp. 141–159, 2021.

Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1092–1104, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480, 2015.

Michel Coste. An introduction to semialgebraic geometry, n.d. URL https://perso.univ-rennes1.fr/michel.coste/polyens/SAG.pdf. Theorem 1.9, p. 11.

Junhao Dong, Hao Zhu, Yifei Zhang, Xinghua Qu, Yew-Soon Ong, and Piotr Koniusz. Machine unlearning via task simplex arithmetic. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Y. Zou. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3832–3842, 2020.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16319–16330, 2021.

David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. *Discrete & Computational Geometry*, 2:127–151, 1987. doi: 10.1007/BF02187876.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894, 2017.

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.

Kevin Kuo, Amrith Setlur, Kartik Srinivas, Aditi Raghunathan, and Virginia Smith. Exact unlearning of finetuning data via model merging at scale, 2025. URL https://arxiv.org/abs/2504.04626.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024. URL https://arxiv.org/abs/2401.06121.

Shane A McQuarrie, Parisa Khodabakhshi, and Karen E Willcox. Nonintrusive reduced-order models for parametric partial differential equations via data-driven operator inference. *SIAM Journal on Scientific Computing*, 45(4):A1917–A1946, 2023.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132, pp. 931–962, 2021.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. URL https://arxiv.org/abs/2105.10386.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66727–66754, 2023.

Benjamin Peherstorfer and Karen Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215, 2016.

Yijun Quan, Wentai Wu, and Giovanni Montana. Exact federated continual unlearning for ridge heads on frozen foundation models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 114–131, 2026.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Henrik Rosenberger, Benjamin Sanderse, and Giovanni Stabile. Exact operator inference with minimal data. *SIAM Journal on Scientific Computing*, 48(5):A2391–A2416, 2026.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1524–1537, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Enayat Ullah and Raman Arora. From adaptive query release to machine unlearning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 34642–34667, 2023.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142, 2021.

Xuelian Wen, Qiuqi Li, and Juan Zhang. Data-driven model reduction for parameter-dependent matrix equations via operator inference, 2025. URL https://arxiv.org/abs/2511.16033.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115, 2023.

A Exact Recovery: Structure and Acquisition Bounds

We prove Theorem 5.1 using the instance classes and observation interface of Section 3. Each instance is fixed before the iid uniform masks are sampled, and the recovery guarantees hold simultaneously over all targets. Write μ_M for the uniform measure on $\{0,1\}^M$, $t_J = \theta_J - \theta_0$, $s_* = \dim S$, and $p_s = (M+1)s(s+1)/2 + Ms$. The normalized true coefficients are $B_0^* = H/\lambda$, $B_j^* = H_j/\lambda$, and $w_j^* = a_j/\lambda$. Decoder selection and boundary conventions are given in Appendix C.

A.1 Response geometry and rational representation

Proof of Lemma 4.1. Writing $H_j = X_j^\top X_j$ and $b_j = X_j^\top y^{(j)}$ gives $\ker H_j = \ker X_j$, hence $b_j, a_j \in \operatorname{range} H_j$. Positivity gives $\ker H_{\mathrm{del}} = \bigcap_j \ker H_j$, so $\operatorname{range} H_j \subseteq \operatorname{range} H_{\mathrm{del}}$. The normal equations in Eq. (8) therefore imply

$$Ht_J = \sum_{j \in J} (a_j + H_j t_J) \in \operatorname{range} H_{\mathrm{del}}. \tag{A.1}$$

Since H is invertible, $S \subseteq H^{-1} \operatorname{range} H_{\mathrm{del}}$ and $s_* \le r$. □

The same range argument shows that the true normalized tuple satisfies the structural constraints in Eq. (20); the upper spectral constraint is needed only for the robust class. The following representation applies to both true and fitted tuples, including those without a record realization.

Lemma A.1 (Low-rank rational responses)**.** *Suppose symmetric matrices and source vectors satisfy*

$$B_0 \succeq I_d, \quad B_j \succeq 0, \quad \sum_j B_j \preceq B_0 - I_d, \quad k := \operatorname{rank}[\sum_j B_j, w_1, \dots, w_M] \le r. \tag{A.2}$$

For $L(z) = B_0 - \sum_j z_j B_j$ and $w(z) = \sum_j z_j w_j$, the responses $L(z)^{-1} w(z)$ on the Boolean cube span at most k dimensions and have a common representation $N(z)/D(z)$ with $\deg N, \deg D \le k$ and $D(z) = \det L(z)/\det B_0 > 0$. If $B_0 \preceq \kappa I_d$, then $\kappa^{-r} \le D(z) \le 1$ and $\|L(z)^{-1}\|_{\mathrm{op}} \le 1$.

Proof. Let $U = \operatorname{range}[\sum_j B_j, w_1, \dots, w_M]$ and let Z be an orthonormal basis of $B_0^{-1/2} U$. Positivity implies $\operatorname{range} B_j \subseteq U$. Define

$$T_j = Z^\top B_0^{-1/2} B_j B_0^{-1/2} Z, \quad v_j = Z^\top B_0^{-1/2} w_j, \quad F(z) = I_k - \sum_j z_j T_j. \tag{A.3}$$

The whitened pencil acts as $F(z)$ on $\operatorname{range} Z$ and as identity on its orthogonal complement. Since $I_d \preceq L(z) \preceq B_0$, it is positive definite. Consequently,

$$\begin{aligned} L(z)^{-1} w(z) &= B_0^{-1/2} Z F(z)^{-1} \sum_j z_j v_j, \\ D(z) &= \det F(z) = \frac{\det L(z)}{\det B_0}, \\ N(z) &= B_0^{-1/2} Z \operatorname{adj}(F(z)) \sum_j z_j v_j. \end{aligned} \tag{A.4}$$

This proves the span and degree claims. Under the spectral upper bound, $\kappa^{-1} I_k \preceq F(z) \preceq I_k$, so $\kappa^{-k} \le D(z) \le 1$; $L(z) \succeq I_d$ gives the inverse bound. For $k = 0$, all sources vanish and one takes $D = 1, N = 0$. Including the source vectors in U covers fitted sources outside $\operatorname{range} \sum_j B_j$. Multilinearization preserves cube values and does not increase degree. □

For the true tuple this gives $t_J = N_*(1_J)/D_*(1_J)$ with $D_*(1_J) = \det R_J / \det H$. Positivity and the degree bound hold without a spectral upper bound.

A.2 TWO-STAGE ACQUISITION AND EXACT COMPLETION

For exact observations and $r \geq 1$, the decoder uses independent batches of sizes

$$n_1 = \lceil 8 \cdot 2^r [r + \log(2/\delta)] \rceil, \quad n_2 = \left\lceil 8 \cdot 2^{r+1} [p_r + \log(2/\delta)] \right\rceil. \tag{A.5}$$

Additional observations are retained for lookup. The first batch determines the response basis, and the second supplies the coefficient-fitting equations. Appendix C specifies output priorities and fallback rules.

Both acquisition stages use the following detection and rank-growth bounds.

Lemma A.2 (Polynomial detection and rank growth)**.** *A nonzero degree-h polynomial on the Boolean cube is nonzero on a set of mass at least 2^{-h}. Suppose an adapted nonnegative integer-valued deficit is initially at most D, never increases, and decreases with conditional probability at least p whenever it is positive. After n steps its failure probability, also conditional on the initial history, is at most $\Pr\{\mathrm{Bin}(n,p) < D\}$. In particular it is at most $\delta/2$ when $np \geq 2D + 8\log(2/\delta)$.*

Proof. Multilinearize the polynomial. A nonzero constant has full support. Otherwise write it as $P_0(z') + z_M Q(z')$ after choosing a coordinate on which it varies. The nonzero polynomial Q has degree at most $h-1$. At every point where $Q \neq 0$, at least one of $P_0, P_0 + Q$ is nonzero. Induction on cube dimension gives support at least $\frac{1}{2}2^{-(h-1)} = 2^{-h}$.

For the deficit, let X_i indicate a strict drop, counting an artificial success after completion, and set $\alpha_i = \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \geq p$. With independent auxiliary uniforms, thin each success to $Y_i = X_i 1_{\{U_i \leq p/\alpha_i\}}$. Relative to the enlarged filtration, Y_i has conditional success probability exactly p. Iterated conditioning gives their product Bernoulli law, including conditional on $\mathcal{F}_0$. If the deficit remains positive, fewer than D drops occurred, hence $\sum_i Y_i < D$. Finally, for $W \sim \mathrm{Bin}(n,p)$ and $\nu = np$, the exponential Markov inequality with parameter $\log 2$ gives:

$$\Pr(W \leq \nu/2) \leq \exp[-(1 - \log 2)\nu/2] \leq e^{-\nu/8}. \tag{A.6}$$

The stated size condition makes $D \leq \nu/2$ and $e^{-\nu/8} \leq \delta/2$. For $D = 0$ completion is deterministic. □

Discovering the response space. Let S_i be the span of the first i sampled responses. At a history with $S_i \subsetneq S$, choose a target outside S_i and its nonzero orthogonal component v. The polynomial $v^\top N_*(z)$ has degree at most r, vanishes whenever $t_J \in S_i$, and is nonzero at the chosen target because $D_* > 0$. Thus the next independent mask enlarges S_i with conditional probability at least 2^{-r}. The detecting set is measurable: on a fixed instance it is a finite set determined by the observed-span projector. Only its cardinality is used, so a measurable choice of the witness v is unnecessary. Applying Lemma A.2 to $s_* - \dim S_i$, with $s_* \leq r$, shows that the first batch in Eq. (A.5) recovers S with probability at least $1 - \delta/2$.

Acquiring the equation row space. Condition on a successful first batch, and write $x_J = V^\top t_J$. Identify a symmetric coefficient tuple $c = (B_0, C_1, \ldots, C_M, v_1, \ldots, v_M)$ with its p_s independent entries. Define the linear block

$$T_J c = \left(B_0 - \sum_{j \in J} C_j \right) x_J - \sum_{j \in J} v_j. \tag{A.7}$$

Let W be the row span of all T_J and W_i the span of the first i second-batch blocks. The deficit $\dim W - \dim W_i$ is adapted, nonincreasing, and at most $p_s \leq p_r$. If $W_i \subsetneq W$, some coefficient vector c annihilates W_i but has $T_{J_\circ} c \neq 0$ at another target. A nonzero coordinate of

$$\left(B_0 - \sum_j z_j C_j \right) V^\top N_*(z) - D_*(z) \sum_j z_j v_j \tag{A.8}$$

is a degree-$(r+1)$ polynomial. It vanishes whenever all rows of T_J belong to W_i and is nonzero at $J_\circ$. Therefore each fresh mask reduces the deficit with conditional probability at least $2^{-(r+1)}$. Ranks and detecting sets are measurable through finite minors. Lemma A.2 and the second batch size give conditional row-space failure at most $\delta/2$. Together, both stages succeed with probability at least $1 - \delta$, without a union bound over targets.

From equation completeness to response recovery. When the first batch recovers S, projecting the true normalized coefficients gives a feasible tuple:

$$\overline{B}_0 = V^\top H V/\lambda, \qquad \overline{C}_j = V^\top H_j V/\lambda, \qquad \overline{v}_j = V^\top a_j/\lambda. \tag{A.9}$$

Indeed $t_J = V x_J$ implies $(\overline{B}_0 - \sum_{j\in J}\overline{C}_j)x_J = \sum_{j\in J}\overline{v}_j$. When the second batch recovers the equation row space, the sampled and full equation systems have equal kernels, so every feasible fit satisfies $T_J c = 0$ for every target. Its structural constraints give

$$B_0 - \sum_{j\in J} C_j = \left(B_0 - \sum_j C_j\right) + \sum_{j\notin J} C_j \succeq I_s. \tag{A.10}$$

The unique solution is therefore x_J, and Eq. (13) returns θ_J. This argument determines responses without requiring unique coefficients or invariance of S under the retained matrices. When $S = \{0\}$ all displacements and equation blocks vanish, and the constant-head branch is exact. Lookup priorities agree with the recovered heads.

Explicit upper constant. For $M, r \geq 1$, $p_r \leq 4Mr^2$, $r \leq Mr^2$, and $\log(2/\delta) \leq \frac{3}{2}\log(1/\delta)$. Retaining both ceiling terms in Eq. (A.5) gives

$$\begin{aligned} n_1 + n_2 &\leq 8 \cdot 2^r[r + 2p_r + 3\log(2/\delta)] + 2 \\ &\leq 2^r[73Mr^2 + 36\log(1/\delta)] \\ &\leq 80 \cdot 2^r[Mr^2 + \log(1/\delta)]. \end{aligned} \tag{A.11}$$

Thus $C_{\mathrm{ex}} = 80$ is valid. Appendix C establishes that the selected decoder is Borel. If $r = 0$, the source-range argument gives $a_j = 0$ and all heads equal θ_0, proving the zero-observation claim.

A.3 A GATE FAMILY OF ACTUAL RIDGE RECORDS

Let $M > r \geq 1$, $L = M - r$, $d = r$, $\lambda = 1/4$, and $\tau = g = 1/32$. There are r gate groups: a source record $(\sqrt{\tau}e_1, -g/\sqrt{\tau})$ and $r-1$ edge records $(\sqrt{\tau}(e_i + e_{i+1}), 0)$. Terminal group $r + j$ contains $(\sqrt{h_j}e_r, 0)$, where $h_j \in [\tau/(2L), \tau/L]$. Write $\mathcal{G} = \{1, \ldots, r\}$ and

$$C_{\mathrm{gate}} = \tau e_1 e_1^\top + \tau \sum_{i=1}^{r-1}(e_i + e_{i+1})(e_i + e_{i+1})^\top. \tag{A.12}$$

Add r undeletable records with covariance $Q(h) = \frac{3}{4}I_r - C_{\mathrm{gate}} - (\sum_j h_j)e_r e_r^\top$ and linear term ge_1. Such a block is realized by the rows of $Q(h)^{1/2}$ as features and $Q(h)^{-1/2}ge_1$ as labels.

Validity and common public inputs. The inequality $\sum_i (v_i + v_{i+1})^2 \leq 4\|v\|^2$ gives $C_{\mathrm{gate}} \preceq 5\tau I_r$. Since $\sum_j h_j \leq \tau$,

$$\frac{9}{16}I_r \preceq Q(h) \preceq \frac{3}{4}I_r, \qquad H = I_r, \quad b = 0, \quad \theta_0 = 0, \quad R_J \succeq \frac{13}{16}I_r. \tag{A.13}$$

All blocks therefore exist, with the same sample size $N = M + r$ and fixed singleton groups. Feature norms and label magnitudes are at most one: the background label vector has norm at most $4g/3 = 1/24$, and the other bounds follow directly from the records. The source and edge vectors span $\mathbb{R}^r$, because a vector orthogonal to them satisfies $v_1 = 0$ and $v_i + v_{i+1} = 0$ successively. Thus $\operatorname{rank} H_{\mathrm{del}} = r$, also for $r = 1$.

Inactive masks conceal the full parameter vector. Write $t_J^{\mathrm{term}} = \sum_{j:r+j\in J} h_j$. The retained equation is

$$\left[I_r - \tau 1_{\{1\in J\}} e_1 e_1^\top - \tau \sum_{i:i+1\in J}(e_i + e_{i+1})(e_i + e_{i+1})^\top - t_J^{\mathrm{term}} e_r e_r^\top\right]\theta_J = g 1_{\{1\in J\}} e_1, \tag{A.14}$$

where $1 \leq i < r$. If the source gate is absent from J, the head is zero. If an edge gate is absent, its missing off-diagonal entry splits the system into two positive-definite blocks. The block containing e_1 is independent of h, while the block containing e_r has zero right-hand side and zero solution. Hence every coordinate of an inactive head is independent of h. The shared full matrix $H = I_r$ ensures that the changing background covariance cancels the terminal covariance before deletion.

Active masks reveal one selected sum. If $\mathcal{G} \subseteq J$, the head equals $\Psi(\sum_{j:r+j\in J} h_j)$, where

$$\Psi(t) = g(A - te_r e_r^\top)^{-1} e_1, \qquad A = I_r - C_{\text{gate}}, \quad 0 \le t \le \tau. \tag{A.15}$$

These matrices are at least $13I_r/16$. Their path cofactor gives $e_r^\top A^{-1} e_1 = \tau^{r-1}/\det A > 0$, with empty product one for $r = 1$. Set $\alpha = e_r^\top A^{-1} e_r > 0$ and $\beta = e_r^\top A^{-1} e_1 > 0$. Positivity applied to $A^{-1}e_r$ gives $1 - t\alpha > 0$, and the rank-one resolvent identity yields

$$e_r^\top \Psi(t) = \frac{g\beta}{1 - t\alpha}, \qquad t = \frac{1 - g\beta/(e_r^\top \Psi(t))}{\alpha}. \tag{A.16}$$

Thus Ψ is injective and has a continuous inverse on its image. For any fixed mask list, its archive is a Borel function of Th, where T has the terminal-incidence rows of active masks. The target heads at $\mathcal{G} \cup \{r + j\}$ determine each h_j through Eq. (A.16). The number K_n of active rows is $\mathrm{Bin}(n, 2^{-r})$ under iid fair masks.

A.4 Fixed-instance acquisition lower bound

Fix any randomized Borel processor and put the uniform Lebesgue law on $\mathcal{H} = [\tau/(2L), \tau/L]^L$. For a fixed mask list with $K_n < L$, the measurement matrix T has rank less than L. No measurable map of Th and an independent seed can recover h exactly with positive probability under this law. To see this, choose orthogonal coordinates $h = Uu + Vv$ along $(\ker T)^\perp$ and $\ker T$. For each fixed u and seed, $Th = TUu$ is constant in v, so the equality between the proposed reconstruction and h holds for at most one v. The kernel has positive dimension, hence this section has Lebesgue measure zero.

Compose the processor's predictions at $\mathcal{G} \cup \{r + j\}$ with the Borel inverse in Eq. (A.16), extended arbitrarily outside its image. Exact recovery of all targets would recover h, so the preceding argument makes its seed success probability zero for almost every h on each deficient mask list. There are only 2^{Mn} lists. Intersecting these full-measure sets therefore produces a single full-measure set of parameters such that, for every fixed h in it,

$$\Pr\{\text{some target is recovered incorrectly}\} \ge \Pr\{\mathrm{Bin}(n, 2^{-r}) < L\}. \tag{A.17}$$

The resulting hard instance is fixed before the masks and processor seed are sampled, and the argument applies to the complete observed heads.

A processor succeeding with probability at least $1 - \delta$ on every instance must consequently satisfy $\Pr(K_n < L) \le \delta$. With $p = 2^{-r} \le 1/2$,

$$np \ge L \Pr(K_n \ge L) \ge \tfrac{3}{4} L, \qquad (1 - p)^n \le \delta. \tag{A.18}$$

Since $-\log(1 - p) \le 2p$, the second inequality gives $np \ge \frac{1}{2}\log(1/\delta)$. Averaging the two bounds yields

$$n \ge \frac{1}{4}\, 2^r [M - r + \log(1/\delta)]. \tag{A.19}$$

Thus $c_{\text{ex}} = 1/4$ is valid, completing Theorem 5.1.

A.5 Response recovery with nonunique coefficients

Example A.3 (Ambiguity with a fixed original Hessian)**.** Let $M = 3$, $d = r = 2$, $\theta_0 = 0$, and $V = I_2$. There exists a continuum of distinct feasible coefficient tuples, indexed by $h \in [1/2, 1]$ with fixed $B_0 = 3I_2$, that all induce the same response function

$$\widehat{\theta}_J = \tfrac{1}{2}(z_1 + z_2 + z_1 z_2) e_1 + \tfrac{2}{5} z_3 e_2, \qquad z = 1_J \in \{0, 1\}^3. \tag{A.20}$$

Thus, even after fixing the common scaling through B_0, the coefficient-fitting system remains rank deficient although every deletion response is uniquely determined. The responses span both coordinates and exhibit the nonadditive interaction $\theta_{\{1,2\}} - \theta_{\{1\}} - \theta_{\{2\}} = e_1/2$.

For Example A.3, take $M = 3$, $d = r = 2$, $N = 5$, $\lambda = 1$, singleton groups $G_j = \{j\}$, and $c = 3/2 - h$ with $h \in [1/2, 1]$. Use the records

$$\begin{aligned} &(x_1, y_1) = (\sqrt{h}\, e_1, -(3 - h)/(2\sqrt{h})), \quad (x_2, y_2) = (\sqrt{c}\, e_1, -(3 - c)/(2\sqrt{c})), \\ &(x_3, y_3) = (e_2/\sqrt{2}, -\sqrt{2}), \qquad\qquad (x_4, y_4) = (e_1/\sqrt{2}, 9\sqrt{2}/4), \\ &(x_5, y_5) = (\sqrt{3/2}\, e_2, \sqrt{2/3}). \end{aligned} \tag{A.21}$$

Records 4 and 5 are undeletable. The deletable covariance is $\mathrm{diag}(3/2, 1/2)$ and the background covariance is $\mathrm{diag}(1/2, 3/2)$, so $H = 3I_2$. The background linear term $9e_1/4 + e_2$ cancels those of the first three records, giving $b = 0$ and $\theta_0 = 0$. Thus $a_j = -b_j$, and the normalized group coefficients are $C_j = x_j x_j^\top$ and $v_j = -y_j x_j$ for $j = 1, 2, 3$; the coercivity margin is $B_0 - \sum_j C_j = \mathrm{diag}(3/2, 5/2) \succeq I_2$.

For $z = 1_J$, direct solution gives

$$\theta_J = \frac{(3-h)z_1 + (3-c)z_2}{2(3 - hz_1 - cz_2)} e_1 + \frac{z_3}{3 - z_3/2} e_2. \tag{A.22}$$

The first fraction takes values $0, 1/2, 1/2, 3/2$ for $(z_1, z_2) = (0,0), (1,0), (0,1), (1,1)$, while the second equals $2z_3/5$. This proves Eq. (A.20) for every Boolean mask, independently of h. The singleton heads $e_1/2$ and $2e_2/5$ span $\mathbb{R}^2$, and the interaction of groups 1 and 2 is $e_1/2$. Any common left multiplication preserving $B_0 = 3I_2$ is the identity, so varying h changes the coefficients beyond common scaling. The group identifiers stay fixed. At $h = c$, identical numerical records remain distinct indexed records, as permitted by the formulation.

B Precision Bounds and Approximately Low-Rank Curvature

We prove Theorem 5.2 and then extend it to approximately low-rank deleted curvature. Throughout this section, $r \geq 1$, $q = r + 1$, and $0 < \delta \leq 1/4$. The true instance belongs to the robust class in Eq. (5). Errors may depend jointly on the fixed instance and the complete mask sequence, subject only to $\|e_i\|_2 \leq \eta$ almost surely.

B.1 Pair compression and polynomial residuals

For $a \in \{*, f\}$, let $\mathrm{L}_a(z) = B_{0,a} - \sum_j z_j B_{j,a}$, $\mathrm{w}_a(z) = \sum_j z_j w_{j,a}$, and $u_a(z) = \mathrm{L}_a(z)^{-1}\mathrm{w}_a(z)$. Here $*$ denotes the true normalized coefficients and f any tuple satisfying the robust structural constraints. By Lemma A.1, each response family spans a space of dimension at most r. Let W be an orthonormal basis of their combined span, of dimension $m \leq 2r$, and define

$$C_{j,a} = W^\top B_{j,a} W, \quad v_{j,a} = W^\top w_{j,a}, \quad L_a(z) = C_{0,a} - \sum_j z_j C_{j,a}, \quad w_a(z) = \sum_j z_j v_{j,a}. \tag{B.1}$$

For $x_a = W^\top u_a$, the identity $u_a = W x_a$ implies $L_a x_a = w_a$ and $\|u_* - u_f\|_2 = \|x_* - x_f\|_2$. Congruence gives $I_m \preceq L_a(z) \preceq \kappa I_m$ and preserves all coefficient inequalities. It also preserves the augmented rank bound: the columns of $\sum_j C_{j,a}$ and the $v_{j,a}$ lie in

$$W^\top \mathrm{range}[\sum_j B_{j,a}, w_{1,a}, \ldots, w_{M,a}], \tag{B.2}$$

whose dimension is at most r. This projection does not require invariance of the combined span under either pencil.

Reapplying Lemma A.1 to the compressed true tuple gives polynomials of degree at most r satisfying

$$D_*(z) = \frac{\det L_*(z)}{\det C_{0,*}}, \qquad N_*(z) = D_*(z) x_*(z), \qquad \kappa^{-r} \leq D_*(z) \leq 1 \tag{B.3}$$

on the cube. Define the residual of degree at most q as:

$$P(z) = L_f(z) N_*(z) - D_*(z) w_f(z) = D_*(z) L_f(z) (x_*(z) - x_f(z)). \tag{B.4}$$

Since $D_* \geq \kappa^{-r}$ and $L_f \succeq I_m$, this residual bounds the prediction error:

$$\|\theta_J - (\theta_0 + u_f(1_J))\|_2 \leq \kappa^r \|P(1_J)\|_2. \tag{B.5}$$

If the tuple also satisfies the archive residual constraints, then at every sampled mask

$$\|\mathrm{L}_f u_* - \mathrm{w}_f\|_2 \leq \|\mathrm{L}_f(\widetilde{\theta}_{A_i} - \theta_0) - \mathrm{w}_f\|_2 + \|\mathrm{L}_f e_i\|_2 \leq 2\kappa\eta. \tag{B.6}$$

Projecting by $W^\top$ and multiplying by $D_* \leq 1$ yields

$$\|P(1_{A_i})\|_2 \leq 2\kappa\eta \qquad (1 \leq i \leq n). \tag{B.7}$$

Only the projection of the original cross-residual is used; that residual need not lie in the pair span. When $m = 0$, take $D_* = 1$ and $N_* = P = 0$. All identities remain valid.

B.2 A COMMON CLASS OF RESIDUAL THRESHOLD SETS

To control fits selected from a perturbed archive, we bound the VC dimension of a common class containing all sets $\{z : \|P(z)\|_2 > t\}$, for all true/fitted pairs, dimensions, and thresholds $t \geq 0$.

Pad both compressed tuples to dimension $h = 2r$ by adjoining identity blocks to $C_{0,a}$, zero blocks to $C_{j,a}$, and zero coordinates to $v_{j,a}$. Denote the padded pencils by $\overline{L}_a$, their sources by $\overline{w}_a$, and the constant matrices by $\overline{C}_{0,a}$. Padding preserves determinant ratios and residual norms. The adjugate identity expresses the padded residual as

$$(P(z), 0) = \frac{\overline{L}_f(z)\,\mathrm{adj}(\overline{L}_*(z))\overline{w}_*(z) - \det(\overline{L}_*(z))\overline{w}_f(z)}{\det \overline{C}_{0,*}}. \tag{B.8}$$

Since $\det \overline{C}_{0,*} > 0$, the threshold set is defined by one polynomial inequality:

$$\begin{aligned}\Phi(z;\xi) = \|\overline{L}_f(z)\,\mathrm{adj}(\overline{L}_*(z))\overline{w}_*(z) - \det(\overline{L}_*(z))\overline{w}_f(z)\|_2^2 \\ - t^2 \det(\overline{C}_{0,*})^2 > 0.\end{aligned} \tag{B.9}$$

The two symmetric matrix collections, two source collections, and threshold use

$$p = (M+1)h(h+1) + 2Mh + 1 \leq 21Mr^2 \tag{B.10}$$

real parameters. Each matrix or source entry has joint degree at most two in the masks and parameters. The numerator coordinates have degree at most $2h+2$, so Φ has joint degree at most $4h + 4 = 8r + 4$. Allow all real parameter values in this polynomial formula to define the enlarged range class $\mathscr{R}_{M,r}$. The polynomial formula remains defined for singular tuples. The compression basis is absorbed into the freely varying compressed coefficients, so neither the basis nor the ambient dimension contributes additional parameters. The residuals themselves have mask degree at most q; the larger joint degree is used only to bound the VC dimension.

Goldberg–Jerrum's bound for p real parameters, one polynomial predicate, and joint degree $8r + 4$ gives (Goldberg & Jerrum, 1995, Theorem 2.2)

$$\mathrm{VCdim}(\mathscr{R}_{M,r}) \leq 2p\log_2(8e(8r+4)) \leq v := \left\lceil 42Mr^2\log_2(8e(8r+4))\right\rceil. \tag{B.11}$$

The theorem applies first to the polynomial class on real inputs; restricting it to the Boolean cube cannot increase its VC dimension.

B.3 UNIFORM POLYNOMIAL NORM BOUNDS

Lemma B.1 (Vector-valued Boolean norm bounds)**.** *Let P be a Hilbert-valued polynomial of degree at most q on the uniform Boolean cube. Then*

$$\mathbb{E}\|P\|^4 \leq 9^q(\mathbb{E}\|P\|^2)^2, \qquad \max_z \|P(z)\| \leq \sqrt{N_q}\,\|P\|_{L^2(\mu_M)}. \tag{B.12}$$

If P is nonzero on the cube, then

$$\mu_M\{z : \|P(z)\| > \|P\|_{L^2(\mu_M)}/\sqrt{2}\} \geq \varepsilon_q := \frac{1}{4\cdot 9^q}. \tag{B.13}$$

Proof. Multilinearize and replace bits by independent uniform signs, preserving degree. In an orthonormal basis of the finite coefficient span, each scalar coordinate P_a obeys the Bonami bound $\mathbb{E}P_a^4 \leq 9^q(\mathbb{E}P_a^2)^2$ (O'Donnell, 2014, Section 9.1). Therefore

$$\mathbb{E}\|P\|^4 = \sum_{a,b}\mathbb{E}(P_a^2P_b^2) \leq \sum_{a,b}\sqrt{\mathbb{E}P_a^4\mathbb{E}P_b^4} \leq 9^q\left(\sum_a \mathbb{E}P_a^2\right)^2. \tag{B.14}$$

For $X = \|P\|^2$ with positive mean, Cauchy–Schwarz gives

$$\mathbb{E}X \leq \tfrac{1}{2}\mathbb{E}X + \sqrt{\mathbb{E}X^2\,\Pr(X > \mathbb{E}X/2)}. \tag{B.15}$$

Combining with the fourth-moment bound proves the strict small-ball inequality. Finally, the Walsh expansion has at most N_q coefficients and satisfies Parseval's identity. Since each character has absolute value one, Cauchy–Schwarz gives $\|P(z)\| \leq \sum_S \|\widehat{P}_S\| \leq \sqrt{N_q}(\sum_S \|\widehat{P}_S\|^2)^{1/2}$. This includes $q > M$, when $N_q = 2^M$, without a factor depending on the output dimension. □

For a range space of VC dimension at most $v \geq 1$, the iid range-hitting bound of Haussler & Welzl (1987, Corollary 3.7) ensures that every range of mass greater than ε is sampled, with probability at least $1-\delta$, if

$$n > \max\left\{\frac{4}{\varepsilon}\log_2\frac{2}{\delta}, \frac{8v}{\varepsilon}\log_2\frac{8v}{\varepsilon}\right\}. \tag{B.16}$$

Binary logarithms are conservative also under the natural-logarithm convention. The finite uniform domain and sampling with replacement meet its hypotheses. Replacing the actual dimension by the positive upper bound v is valid by monotonicity; a zero-dimensional class has at most one distinct range, whose miss probability is directly bounded by $e^{-n\varepsilon}$.

Apply this result with $\varepsilon = \varepsilon_q/2$. Under

$$n > \max\{32 \cdot 9^q \log_2(2/\delta),\ 64v9^q \log_2(64v9^q)\}, \tag{B.17}$$

there is an event $\mathcal{E}_{M,r,n}$ of probability at least $1-\delta$ on which every range in $\mathscr{R}_{M,r}$ of mass at least ε_q is hit. Using $\varepsilon_q/2$ also covers ranges of mass exactly ε_q. This event depends only on the masks. It is measurable because the underlying cube has only finitely many subsets, despite the real parameterization of the ranges.

Explicit sampling constant. Set $F = Mr^2\log(r+1)$. Since $M, r \geq 1$, $F \geq \log 2 > 1/2$. Using $8r+4 \leq 12(r+1)$, $\log(96e) < 6$, and $\log 2 > 2/3$ gives $\log_2(8e(8r+4)) < 15\log(r+1)$, hence $v \leq 630F + 1 \leq 1000F$. Further, $\log\log(r+1) \leq \log r$, $\log 64000 < 12$, $\log 9 < 3$, and $q \geq 2$ imply

$$\log(64v9^q) \leq 12 + \log M + 3\log r + 3q \leq 9[q + \log(Mr)]. \tag{B.18}$$

Since $1/\log 2 < 2$ and $\delta \leq 1/4$, the two terms in Eq. (B.17) are bounded by

$$\begin{aligned} 64v9^q \log_2(64v9^q) &\leq 1{,}152{,}000\, 9^q F[q + \log(Mr)], \\ 32 \cdot 9^q \log_2(2/\delta) &\leq 256 \cdot 9^q F \log(1/\delta). \end{aligned} \tag{B.19}$$

Both are strictly smaller than $2{,}000{,}000\, 9^q F[q+\log(Mr/\delta)]$. Thus Eq. (16) with $C_{\mathrm{rob}} = 2{,}000{,}000$ implies the required strict condition, including integer rounding.

Uniform error transfer. Fix any sequence in $\mathcal{E}_{M,r,n}$ and any compatible true/fitted pair, including a fit selected after observing the perturbed heads. If $\|P\|_{L^2} > 2\sqrt{2}\kappa\eta$, Lemma B.1 makes the range $\{z : \|P(z)\| > 2\kappa\eta\}$ have mass at least ε_q. It belongs to $\mathscr{R}_{M,r}$, so some sampled residual exceeds $2\kappa\eta$, contradicting Eq. (B.7). Therefore

$$\|P\|_{L^2} \leq 2\sqrt{2}\kappa\eta, \qquad \max_J \|\theta_J - (\theta_0 + u_f(1_J))\|_2 \leq 2\sqrt{2}\kappa^{r+1}\sqrt{N_q}\eta =: E. \tag{B.20}$$

The second inequality follows from Eqs. (B.5) and (B.12). The argument is deterministic on one event that already covers every pair, dimension, and threshold. Hence it remains valid for arbitrary jointly chosen bounded errors; no independence between errors, fitted coefficients, and masks is needed.

The true normalized statistics establish feasibility because their sampled residual is $\mathrm{L}_*(1_{A_i})e_i$ with norm at most $\kappa\eta$. Appendix C supplies the Borel selector. Comparing the ridge objective at θ_J and zero gives $\lambda\|\theta_J\|^2 \leq \sum_i y_i^2 \leq \lambda B^2$. When $E > B$, zero prediction at noninitial targets therefore incurs error at most B; otherwise the fitted predictor incurs error at most E. Initial and zero-error lookup priorities are exact. This one processor family proves Eq. (17) uniformly over public d, N, λ and all admissible error rules.

B.4 Balanced-sign records and target separation

Let $M \geq 2$ be even, fix $s_j \in \{-1, 1\}$ with $\sum_j s_j = 0$, and set $d = r$, $\lambda = 1/4$, $a = 1/4$, and $\tau = g = 1/32$. Let K be the path adjacency matrix on r coordinates, with $K = 0$ for $r = 1$. For $0 \leq h \leq \tau$, prescribe

$$H_j(h) = \frac{aI_r + s_j(\tau K + he_re_r^\top)}{M}, \qquad b_j = -\frac{s_j g e_1}{M}. \tag{B.21}$$

Use r records per group, with feature matrix $H_j(h)^{1/2}$ and label vector $H_j(h)^{-1/2}b_j$, and add r undeletable records with covariance $I_r/2$ and zero labels.

Record realization and normalization. The path satisfies $\|K\|_{\mathrm{op}} \le 2$, by $2|v_i v_{i+1}| \le v_i^2 + v_{i+1}^2$. Hence

$$\frac{5}{32M} I_r \preceq H_j(h) \preceq \frac{11}{32M} I_r, \qquad \sum_j b_j^\top H_j(h)^{-1} b_j \le \frac{32g^2}{5} = \frac{1}{160}. \tag{B.22}$$

All blocks exist. Their feature norms are at most $1/\sqrt{2}$, and individual label magnitudes are at most $g/\sqrt{5/32}$. Balance gives common $N = (M+1)r$, $H_{\mathrm{del}} = aI_r$, $H = I_r$, $b = 0$, and $\theta_0 = 0$. Scaling all labels by $B \ge 0$ scales all heads by B and yields energy at most $B^2/160 \le \lambda B^2$, while $H = 4\lambda I_r$. Thus every scaled instance belongs to the required robust class at $\kappa = 4$ with the same public inputs.

Full-vector path comparison. For $u_J = |J|/M$ and $\ell_J = M^{-1}\sum_{j\in J} s_j$, one has $0 \le u_J \le 1$ and $|\ell_J| \le 1/2$. Before scaling, the retained matrix and head are

$$R_h(u,\ell) = (1-au)I_r - \ell(\tau K + h e_r e_r^\top), \qquad f_h(u,\ell) = g\ell R_h(u,\ell)^{-1} e_1. \tag{B.23}$$

Every such matrix is at least $(3/4 - 3\tau/2)I_r = 45I_r/64 \succeq 2I_r/3$. The resolvent identity and path cofactor give

$$\begin{aligned} f_h - f_0 &= gh\ell^2 R_h^{-1} e_r (e_r^\top R_0^{-1} e_1), \\ e_r^\top R_0^{-1} e_1 &= \frac{(\ell\tau)^{r-1}}{\det R_0}. \end{aligned} \tag{B.24}$$

For the cofactor, deleting row 1 and column r leaves an upper-triangular minor with diagonal $-\ell\tau$; its cofactor sign cancels $(-1)^{r-1}$. At $r = 1$, the minor is empty and its determinant is one. Thus the formulas also cover negative or zero ℓ, without division by ℓ. Coercivity now yields

$$\|f_h(u,\ell) - f_0(u,\ell)\|_2 \le gh\tau^{r-1}(3/2)^q |\ell|^q. \tag{B.25}$$

At the positive-sign target $J_+ = \{j : s_j = 1\}$, $(u,\ell) = (1/2, 1/2)$ and $0 \prec R_h \preceq 59I_r/64 \preceq I_r$. Therefore $\det R_0 \le 1$ and $e_r^\top R_h^{-1} e_r \ge 1$. Taking the terminal coordinate in Eq. (B.24) proves

$$\|f_h(1/2,1/2) - f_0(1/2,1/2)\|_2 \ge gh\tau^{r-1} 2^{-q}. \tag{B.26}$$

Combining these inequalities gives the full-vector comparison

$$\|f_h(u,\ell) - f_0(u,\ell)\|_2 \le 3^q |\ell|^q \|f_h(1/2,1/2) - f_0(1/2,1/2)\|_2. \tag{B.27}$$

Define the endpoint signal by

$$\Delta_r = \|f_\tau(1/2,1/2) - f_0(1/2,1/2)\|_2 \ge g\tau^r 2^{-q} = 64^{-q} > 0. \tag{B.28}$$

Because the retained matrices remain uniformly positive definite, the scaled target gap is continuous in h, from zero to $B\Delta_r$. Every gap in $[0, B\Delta_r]$ is consequently attained by a deterministic h. Equation (B.27) holds for that same choice at every mask.

B.5 A FIXED-PAIR PRECISION LOWER BOUND

Fix $n \ge 1$ and set

$$t = \sqrt{\frac{\log(2n/\delta)}{2M}}, \qquad \gamma = \min\left\{ B\Delta_r, \frac{2\eta}{3^q t^q} \right\}. \tag{B.29}$$

The preceding continuity argument selects two instances, with parameters 0 and h, whose target distance is γ and whose distance at every mask is at most $3^q|\ell_J|^q\gamma$. This pair depends only on the public parameters and the fixed sign catalog, and is chosen before acquisition.

For a fair mask with independent indicators X_j, balance gives $M\ell_A = \sum_j s_j(X_j - 1/2)$. Its moment-generating function obeys

$$\mathbb{E}e^{vM\ell_A} = \prod_j \cosh(v/2) \le e^{Mv^2/8}, \tag{B.30}$$

since $\log\cosh x \le x^2/2$. Exponential Markov with $v = 4t$, applied to both signs, gives $\Pr(|\ell_A| > t) \le 2e^{-2Mt^2}$. A union bound therefore shows that

$$\mathcal{E} = \{\max_{i\le n} |\ell_{A_i}| \le t\}, \qquad \Pr(\mathcal{E}) \ge 1 - \delta. \tag{B.31}$$

If $t \geq 1/2$, the event is the entire sample space, consistently with this bound.

Let $\theta^{(0)}, \theta^{(1)}$ denote the two fixed catalogs. Define the errors on their respective instances by

$$e_i^{(0)} = \tfrac{1}{2} 1_{\mathcal{E}} (\theta^{(1)}_{A_i} - \theta^{(0)}_{A_i}), \qquad e_i^{(1)} = -e_i^{(0)}. \tag{B.32}$$

On $\mathcal{E}$ the sampled distances are at most $3^q t^q \gamma \leq 2\eta$, so these errors have norm at most η; off $\mathcal{E}$ they vanish. They are measurable functions of the finite mask list, admissible under the joint-error model. They can be extended by zero away from the respective fixed instance. On $\mathcal{E}$, both archives report the same midpoint heads, with common masks and exact initial head zero. This identity includes every observed vector coordinate and repeated masks.

Couple the two experiments using the same masks and the same processor seed, drawn independently of the masks. Every randomized processor then gives the same predictions on $\mathcal{E}$. For any $0 \leq \epsilon < \gamma/2$, it cannot be within ϵ of both target heads at J_+, whose distance is γ. If F_b is the all-target failure event on instance b, then

$$1_{F_0} + 1_{F_1} \geq 1_{\mathcal{E}}, \qquad \Pr(F_0) + \Pr(F_1) \geq 1 - \delta. \tag{B.33}$$

One fixed instance has failure probability at least $(1-\delta)/2 > \delta$. The choice of instance depends only on the two marginal failure probabilities, so it is fixed before sampling. Thus every radius strictly below $\gamma/2$ fails the uniform guarantee for every processor. Taking infima proves

$$\mathcal{R}^{\mathrm{hp}}_{n,M,r,4,B}(\eta;\delta) \geq \frac{\gamma}{2} = \min\left\{ \frac{B\Delta_r}{2}, \frac{\eta}{3^q} \left(\frac{2M}{\log(2n/\delta)} \right)^{q/2} \right\}. \tag{B.34}$$

When $B = 0$ or $\eta = 0, \gamma = 0$ and the lower bound follows from nonnegativity. The strict inequality $\epsilon < \gamma/2$ suffices because the recovery radius is an infimum. This completes Theorem 5.2.

B.6 Scaling and the stated precision window

At $\kappa = 4$, a sufficient condition for both bounds in Theorem 5.2 to lie on their noise-linear branches is

$$\eta \leq \min\left\{ \frac{B}{2\sqrt{2}\, 4^q \sqrt{q+1}\, M^{q/2}}, \frac{B\Delta_r 3^q}{2} \left(\frac{\log(2n/\delta)}{2M} \right)^{q/2} \right\}. \tag{B.35}$$

The explicit constants C_{ex}, c_{ex}, and C_{rob} are justified in Appendices A.2, A.4, and B.3, respectively.

For $1 \leq q \leq M$, the product formula for $\binom{M}{q}$ and the bounds $\binom{M}{j} \leq M^j \leq M^q$ give

$$(M/q)^q \leq \binom{M}{q} \leq N_q \leq (q+1)M^q. \tag{B.36}$$

Hence $N_q = \Theta_r(M^q)$ at fixed rank. At $\kappa = 4$, the first term in Eq. (B.35) ensures $2\sqrt{2}\, 4^q \sqrt{N_q}\eta \leq B$, while the second ensures that the lower linear term does not exceed $B\Delta_r/2$. Both bounds therefore use their linear branches in that window, with the common $M^{q/2}$ factor and the exact lower logarithm $[\log(2n/\delta)]^{-q/2}$. For fixed $r, \delta, \kappa = 4$, and even $M \to \infty$, budgets satisfying Eq. (16) and $n \leq M^C$ for a fixed $C > 1$ give $\log(2n/\delta) = O(\log M)$. In this regime, the amplification exponent in M is $q/2$ up to logarithmic factors. The comparison leaves a gap in rank-dependent constants and saturation scales. For larger budgets, the amplification lower bound is governed by its explicit $\log(2n/\delta)$ dependence. For $q > M$, the original theorem remains valid with $N_q = 2^M$. Likewise, for $M \geq 2r$, Theorem 5.1 gives a necessary count at least $2^r[M + \log(1/\delta)]/8$, establishing the stated linear dependence on M at fixed rank and confidence.

B.7 Extension to approximately low-rank deleted curvature

The original instance satisfies the spectral and label-energy bounds in Eq. (5), but its deleted curvature may have rank greater than r. Write $\sigma_{r+1}(H_{\mathrm{del}})$ for its $(r+1)$st largest eigenvalue, with value zero when $r = d$. The processor receives the tail bound $\varepsilon \geq 0$, but not the spectral projector used in the proof.

Corollary B.2 (Recovery with a spectral tail)**.** *Consider the same ridge and archive interface, with $H \preceq \kappa\lambda I_d$ and $\sum_i y_i^2 \le \lambda B^2$. Replace the exact-rank condition by $\sigma_{r+1}(H_{\mathrm{del}}) \le \varepsilon$, for $1 \le r \le d$ and a public $\varepsilon \ge 0$, taking $\sigma_{d+1} = 0$. Define*

$$\rho = B(1 + 2\sqrt{\kappa - 1})\sqrt{\varepsilon/\lambda}, \qquad K_r = 2\sqrt{2}\kappa^{r+1}\sqrt{N_{r+1}}. \tag{B.37}$$

Under Eq. (16), a Borel decoder using only the original archive and public parameters satisfies, for every fixed instance and every admissible error rule,

$$\Pr\Big\{\max_J \|\widehat{\theta}_J - \theta_J\|_2 \le \min\{B,\, 2\rho + K_r(\eta + 2\rho)\}\Big\} \ge 1 - \delta. \tag{B.38}$$

Lemma B.3 (Uniform ridge approximation by spectral projection)**.** *Under the spectral and label-energy bounds above, let $1 \le r \le d$ and $\sigma_{r+1}(H_{\mathrm{del}}) \le \varepsilon$. Choose an orthogonal projector P onto r leading eigenvectors of H_{del}. Replace each deletable feature x_i by Px_i, keeping every label, background record, and group identifier unchanged. The resulting instance has deleted-curvature rank at most r and satisfies the same spectral and label-energy bounds. Its heads $\theta_J^{(r)}$ satisfy*

$$\max_{J \subseteq [M]} \|\theta_J - \theta_J^{(r)}\|_2 \le \rho, \qquad \rho = B\big(1 + 2\sqrt{\kappa - 1}\big)\sqrt{\varepsilon/\lambda}. \tag{B.39}$$

Proof. Write $Q = I_d - P$, and let X_{del} stack the deletable feature rows. Since P is a spectral projector,

$$H_{\mathrm{del}} = PH_{\mathrm{del}}P + QH_{\mathrm{del}}Q, \qquad \|X_{\mathrm{del}}Q\|_{\mathrm{op}}^2 = \|QH_{\mathrm{del}}Q\|_{\mathrm{op}} \le \varepsilon. \tag{B.40}$$

Thus the projected full matrix is

$$H^{(r)} = H - QH_{\mathrm{del}}Q \preceq H \preceq \kappa\lambda I_d. \tag{B.41}$$

It remains at least λI_d, because it is the regularized covariance of the projected records. The projected deletable covariance is $PH_{\mathrm{del}}P$ of rank at most r. The labels, sample size, and disjoint nonempty indexed groups are unchanged, even if some projected records have zero features. Hence the projected instance belongs to the original exact-rank robust class with the same public parameters.

Fix a target J, and let X_T stack only the retained deletable records, with label vector y_T. Background contributions are unchanged. Put $A_T = X_T^\top X_T$. Row restriction and $H_{\mathrm{del}} \preceq H - \lambda I_d$ imply

$$\|X_T Q\|_{\mathrm{op}} \le \sqrt{\varepsilon}, \qquad \|X_T\|_{\mathrm{op}} \le \sqrt{(\kappa - 1)\lambda}, \qquad \|y_T\|_2 \le \sqrt{\lambda}B. \tag{B.42}$$

If b_J and $b_J^{(r)}$ denote the retained linear terms, then

$$\|b_J - b_J^{(r)}\|_2 = \|QX_T^\top y_T\|_2 \le B\sqrt{\lambda\varepsilon}. \tag{B.43}$$

The retained curvature difference obeys

$$\begin{aligned} R_J - R_J^{(r)} &= A_T - PA_TP = QA_T + PA_TQ, \\ \|R_J - R_J^{(r)}\|_{\mathrm{op}} &\le 2\|X_TQ\|_{\mathrm{op}}\|X_T\|_{\mathrm{op}} \le 2\sqrt{(\kappa - 1)\lambda\varepsilon}. \end{aligned} \tag{B.44}$$

For an individual retained subset, the covariance difference can be indefinite; the operator-norm bound handles this case. Comparing the projected ridge objective at its minimizer and zero gives $\|\theta_J^{(r)}\|_2 \le B$. Subtracting the two retained normal equations yields

$$R_J(\theta_J - \theta_J^{(r)}) = (b_J - b_J^{(r)}) - (R_J - R_J^{(r)})\theta_J^{(r)}. \tag{B.45}$$

Since $\|R_J^{-1}\|_{\mathrm{op}} \le 1/\lambda$, the preceding bounds imply Eq. (B.39). The bounds hold with the same constants for every retained subset, including the initial target. □

Decoder and effective observation error. Set $a = 2\rho$, $\zeta = \eta + a$, $K_r = 2\sqrt{2}\kappa^{r+1}\sqrt{N_{r+1}}$, and $E_{\mathrm{app}} = a + K_r\zeta$. The approximate-rank decoder uses the original input archive and initial head. In Eq. (20), replace the residual tolerance $\kappa\eta$ by $\kappa\zeta$, retaining the rank constraint, objective, and lexicographic selection. Initial-target lookup remains exact. When the original error bound is $\eta = 0$, exact lookup at supplied nonempty targets also retains priority, even if $\zeta > 0$. For remaining targets, return zero if $B = 0$ or $E_{\mathrm{app}} > B$; otherwise return $\theta_0 + L_J^{-1}w_J$ from the selected feasible fit. An infeasible fit returns θ_0, and invalid encodings return zero. These conventions define a total Borel rule by Appendix C, since ζ and E_{app} are Borel functions of the public parameters. The corollary uses $r \ge 1$, so no rank-zero branch is invoked.

Proof of Corollary B.2. Fix the original instance and construct its projected counterpart using Lemma B.3, both before acquisition. Define $t_J^{(r)} = \theta_J^{(r)} - \theta_0^{(r)}$. The initial and target approximation bounds give

$$\|t_J - t_J^{(r)}\|_2 \le 2\rho = a \qquad \text{for every } J. \tag{B.46}$$

Consequently the actual input displacements have the representation

$$\widetilde{\theta}_{A_i} - \theta_0 = t_{A_i}^{(r)} + \xi_i, \qquad \xi_i = e_i + (t_{A_i} - t_{A_i}^{(r)}), \qquad \|\xi_i\|_2 \le \eta + a = \zeta. \tag{B.47}$$

These effective errors may depend on the whole mask sequence; the existing uniform residual event permits that dependence. The projected instance's normalized coefficients, using its own initial head $\theta_0^{(r)}$ to define its source vectors, satisfy the rank and spectral constraints. Their sampled residual is $L_{*,r}(1_{A_i})\xi_i$ with norm at most $\kappa\zeta$, proving feasibility of the modified fit on every admissible input.

Under Eq. (16), the mask event in Appendix B.3 has probability at least $1 - \delta$. Apply the deterministic error bound on this event to the projected response family, the effective errors ξ_i, and any feasible fit. Every feasible fitted displacement u_f then obeys

$$\max_J \|u_f(1_J) - t_J^{(r)}\|_2 \le K_r\zeta. \tag{B.48}$$

The argument uses displacement residuals, so the processor needs neither P, the projected records, nor $\theta_0^{(r)}$. Returning the original initial head plus the fitted displacement gives

$$\|\theta_0 + u_f(1_J) - \theta_J\|_2 \le \|u_f(1_J) - t_J^{(r)}\|_2 + \|t_J^{(r)} - t_J\|_2 \le K_r\zeta + a = E_{\mathrm{app}}. \tag{B.49}$$

The original instance still has $\|\theta_J\|_2 \le B$, so the zero-output branch gives the cap B when $E_{\mathrm{app}} > B$. Initial and exact sampled-target overrides incur zero error. Thus one mask event proves the simultaneous bound for every admissible jointly bounded error rule. At $\varepsilon = 0$, the tail vanishes, $a = 0$, and the rule and guarantee reduce to the original robust construction. At $B = 0$, every true head is zero and the specified priorities agree with the zero-output branch. □

Order of the approximation term. The square-root dependence on the spectral tail is necessary for this projection comparison. For $M = 1$, $d = 2$, $r = 1$, $\lambda = B = 1$, and $0 < \varepsilon < 1$, place the two records $(e_1, 0)$ and $(\sqrt{\varepsilon}e_2, 1)$ in the single deletion group, with no background. Then $H_{\mathrm{del}} = \mathrm{diag}(1, \varepsilon)$ and $H \preceq 2I_2$, but

$$\theta_0 = \frac{\sqrt{\varepsilon}}{1+\varepsilon}e_2, \qquad \theta_0^{(r)} = 0. \tag{B.50}$$

Deleting the group makes both heads zero, so the displacement discrepancy also has size $\sqrt{\varepsilon}/(1+\varepsilon)$. This example concerns the projection error, not a minimax lower bound for approximate-rank archive recovery. In the corollary, the tail contribution is amplified by $1 + K_r$; a nontrivial bound below B requires $2(1+K_r)\rho + K_r\eta < B$. The extension uses the same sufficient observation count and rank-constrained fitting problem as the exact-rank robust construction.

C Decoder Conventions and Well-Posedness

C.1 Processor families and public parameters

The public integer parameters satisfy $M, d \ge 1$, $N \ge M$, and $0 \le r \le d$, with $\lambda > 0$; the robust class additionally has $\kappa \ge 1$, $B, \eta \ge 0$. The processor family is indexed by these public parameters. Each processor is a Borel function of the archive and an independent auxiliary seed shared across targets. The constructions below are deterministic and ignore this seed. These conventions give the processor families in Definitions 3.2 and 3.3.

C.2 Selection and output priorities

Invalid input encodings return zero. On valid inputs, the initial target returns θ_0 first; when $\eta = 0$, any nonempty supplied target returns its earliest archived head. For other targets, the exact decoder

uses Eq. (19) with ordered Gram–Schmidt and the batches in Eq. (A.5). An empty span, a second-batch displacement outside that span, or an infeasible fit returns θ_0. Otherwise the minimum-norm coefficients determine Eq. (13).

For the robust decoder, minimize the sum of squared coefficient norms subject to Eq. (20), then select the lexicographically least minimizer in a fixed coordinate order. After the initial and exact-lookup priorities, $B = 0$ returns zero, $r = 0$ returns θ_0, and $E = 2\sqrt{2}\kappa^{r+1}\sqrt{N_{r+1}}\eta > B$ returns zero. Otherwise a feasible fit returns $\theta_0 + L_J^{-1} w_J$; an infeasible fit returns θ_0. Infeasibility can arise on arbitrary input arrays but not on admissible robust archives, whose true normalized coefficients witness feasibility.

C.3 Existence and Borel selection

Fix the dimensions and mask list. The exact fitting problem has closed convex feasible sets. The robust feasible sets are also closed because the rank constraint is closed. Both relations are semialgebraic: PSD constraints use principal minors, rank constraints use vanishing minors, and residual inequalities become polynomial after squaring. In independent symmetric-entry coordinates, the squared Frobenius objective has positive weights one or two and is coercive. Restricting a nonempty feasible set to the objective sublevel through any feasible point yields a compact set. The set of minimizers is therefore nonempty and compact. Strict convexity makes the exact minimizer unique. Successively minimizing the finitely many coordinates over the compact robust minimizing set gives its unique lexicographic minimum.

To verify measurability, write $\mathcal{F}(x, c)$ for a fitting relation and $Q(c)$ for its objective. The minimizer graph is

$$\mathcal{G}_0(x, c) \iff \mathcal{F}(x, c) \text{ and } \neg\exists b\,[\mathcal{F}(x, b) \text{ and } Q(b) < Q(c)]. \tag{C.1}$$

Semialgebraic projection and complement closure (Coste, n.d., Theorem 1.9) make this graph and its feasible-input domain semialgebraic. Each lexicographic restriction similarly excludes the existence of a point with smaller next coordinate in the preceding graph. After all coordinates have been minimized, each feasible input has a unique selected coefficient tuple. The inverse image of an open rational box under either selector is the projection of this graph intersected with the box, hence Borel. Such boxes form a countable basis, proving Borel measurability.

Ordered Gram–Schmidt is Borel on each keep/discard pattern: it tests whether $\|t - VV^\top t\|^2$ is zero and normalizes only on the positive branch. There are finitely many patterns, and the second-batch span test is also Borel. Matrix inversion is continuous on the coercive domain. Finite lookups, public comparisons, and the stated fallbacks preserve measurability. Finally, the countable union over discrete dimensions, masks, and attained span dimensions defines the required total Borel processor family. Neither the existence argument nor the exact selector requires a spectral ceiling.

C.4 Boundary cases

If $r = 0$, the source-range argument gives every $a_j = 0$, hence $\theta_J = \theta_0$ for all J. If $B = 0$, all labels and heads vanish. If the true response span is zero at positive public rank, the exact decoder also returns the correct constant catalog. In the robust proof, $\eta = 0$ forces $P = 0$ on the uniform mask event, so all targets are recovered exactly there; supplied targets remain exact on every outcome. Repeated masks and noisy empty-mask entries do not affect these conclusions because the exact initial head takes priority. The bound $N_q = \sum_{j=0}^{\min(q,M)} \binom{M}{j}$ already includes $q > M$.

D Experimental Details and Additional Results

The protocol below is followed by exact-recovery studies, acquisition and precision diagnostics, and recovery under archive noise and spectral tails. Appendix E reports computational costs and amortization.

D.1 Data and numerical protocol

All recovery tasks use Eq. (1) with $\lambda = 1$. Visual inputs are frozen CLIP ViT-B/32 representations of CIFAR-10/100. Each seed selects 2,000 stratified records. Scalar targets distinguish vehicles

from other CIFAR-10 classes and animals from other CIFAR-100 coarse classes. Eight class-based deletion groups contain 80 records each; remaining records form the undeletable background.

Language representations. Language representations are extracted from frozen Qwen2.5-7B-Instruct and `meta-llama/Llama-2-7b-hf` checkpoints for KnowUndo and TOFU, respectively, without loading fine-tuned adapters (Qwen et al., 2024; Touvron et al., 2023). We use the final hidden state after the last transformer layer and final normalization (layer 28 for Qwen and layer 32 for Llama), selecting the last non-padding prompt token as the pooled representation. The resulting feature dimensions are 3,584 and 4,096, respectively. Inputs contain only questions: Qwen uses its tokenizer's chat template with a generation prompt, whereas Llama uses `[inst] question [/inst]` with special tokens enabled. Sequences are left-padded within each batch. No explicit maximum-length cap or truncation is applied; the maximum observed input lengths, including prompt formatting and special tokens, are 52 tokens for KnowUndo and 42 tokens for TOFU. Extraction uses bfloat16 model weights, and representations are saved in float32.

Language data selection. For KnowUndo, we use the privacy subset. With NumPy random seed 20260907, we sample without replacement 80 deletion candidates from `unlearn_train.json`, restricted to question–answer pairs present in `full.json`, followed by 120 background records from `retention_train.json`; the recovery experiments use the first 80 sampled background records. For TOFU, deletion candidates comprise the first 40 records of `forget01_official.jsonl` and the first 40 records of `retain90.jsonl`; background records are entries 40–119 of `retain90.jsonl`, using zero-based indexing. Each recovery task therefore contains 80 deletion candidates and 80 background records. Candidates are partitioned in their stored order into eight consecutive groups of ten. The scalar target is $\log(1 +$ answer word count$)$, with words counted by whitespace splitting.

Projection and archive sampling. A Gaussian matrix with entries $N(0, 1/r)$ projects the saved features to r dimensions. The real-representation extensions use seeds $101, 202, 303, 404, 505$. Visual seeds change data selection and projection; language seeds change the projection on a shared corpus. Archive repetitions are nested within these fixed instances. Rank-one budgets are $16, 24, 32, 64$; rank-two budgets are $24, 32, 40, 64$; rank-three/four budgets are $32, 48, 64, 96, 128$. There are 16,000 decoder fits in the rank-one/two extension (100 archives per instance/budget) and 4,000 in the higher-rank extension (20 archives). Both include two matched-archive baselines per fit. Rank-one/two masks are paired across datasets; higher-rank masks are independently sampled across datasets.

The numerical decoder uses twice-reorthogonalized Gram–Schmidt with relative threshold 10^{-10}, followed by an SVD representation of normalized homogeneous equation rows at threshold 10^{-10}. An orthonormal symmetric-matrix basis preserves the Frobenius objective. CVXPY 1.7.5 and CLARABEL 0.11.1 solve the fit with absolute gap, relative gap, and feasibility tolerances 10^{-10} and at most 300 iterations. NumPy 2.2.6 and SciPy 1.15.3 support the computations, with one CPU BLAS thread. Sampled-equation residuals and minimum fitted-system eigenvalues are checked before extrapolation; failures trigger the initial-head fallback. Exact sampled-head lookup always takes priority. All solver statuses remain included.

The linear baseline regresses archived displacements on mask indicators, and the lookup baseline returns the initial head on unseen targets. Both override their sampled-target predictions by exact lookup. Exhaustive oracle heads simulate acquisition and evaluate all targets; they are never passed to the decoder beyond sampled entries. Evaluation therefore computes more heads than the acquisition interface supplies. These recovery studies are complemented by the separate timing and storage measurements in Appendix E.

D.2 Exact recovery and affine fitting

D.2.1 Real representations, response geometry, and catalog expansion

Table 3 separates real-representation errors at $n = 64$. Every cell has 100% recovery, compared with 0% for both baselines. Rank-one/two fits contain 15,985 optimal and 15 optimal-inaccurate statuses; higher-rank fits contain 3,825 optimal and 175 optimal-inaccurate statuses. Changing the

success threshold from 10^{-7} to 10^{-8} changes no classification in either extension; changing to 10^{-6} changes one in each.

Table 3: Worst all-target error at $n = 64$. Each rank-one/two cell has 500 archives across five instances, and each rank-three/four cell has 100 across five instances. All cells have 100% numerical recovery.

Dataset	$r = 1$	$r = 2$	$r = 3$	$r = 4$
CIFAR-10	$6.106 \cdot 10^{-16}$	$2.132 \cdot 10^{-14}$	$4.880 \cdot 10^{-14}$	$5.132 \cdot 10^{-14}$
CIFAR-100	$6.661 \cdot 10^{-16}$	$2.803 \cdot 10^{-11}$	$3.105 \cdot 10^{-14}$	$4.593 \cdot 10^{-13}$
KnowUndo	$5.100 \cdot 10^{-16}$	$4.263 \cdot 10^{-14}$	$3.742 \cdot 10^{-14}$	$8.426 \cdot 10^{-14}$
TOFU	$9.541 \cdot 10^{-16}$	$9.748 \cdot 10^{-14}$	$1.081 \cdot 10^{-13}$	$2.409 \cdot 10^{-13}$

All 4,000 higher-rank fits acquire the full response span in the first batch. The second-batch equation rank reaches $p_s - 1$ in exactly the 2,960 successful fits; the other 1,040 fail the all-target criterion. This correspondence is specific to the tested instances; the theorem permits recovery with additional coefficient ambiguity.

Ambient geometry uses $M = 8$, $r = 1, 2, 4$, $d = 8, 32, 64$, three dataset seeds, and 20 archives at each $n = 16, 32, 64, 128$. Deleted records lie in unknown randomly oriented subspaces, with a full-dimensional undeletable background. The oracle projection residual onto $H^{-1}\,\mathrm{range}(H_{\mathrm{del}})$ is at most 4.054×10^{-14}; this basis is not supplied. Of 2,160 fits, 2,113 are optimal and 47 optimal-inaccurate. All 1,080 fits at $n \geq 64$ recover, with maximum error 3.501×10^{-10}.

Catalog expansion uses $M = 8, 10, 12$, $r = 1, 2, 4$, three seeds, ten archives per instance, and budgets $2M, 4M, 8M$. It contains 795 optimal and 15 optimal-inaccurate fits. Every cell fails at $2M$. At $4M$, success is 100%, 93.3%, and 0% for the three ranks, respectively, at each M. Every cell succeeds at $8M$. The stricter 10^{-8} threshold changes no classification in either the geometry or catalog study. Datasets differ across dimensions and catalog sizes, so these comparisons do not isolate a causal effect of either parameter.

D.2.2 Matched affine fitting and symmetry ablation

We compare three shared affine response fits on 36 fixed real-representation tasks: four datasets, ranks $1, 2, 4$, and seeds $101, 202, 303$. Each task uses ten iid uniform archives at each budget $n \in \{24, 32, 48, 64\}$, giving 1,440 archives and 4,320 method evaluations. All methods receive the same initial head, first-batch response basis, second-batch observations, and exact sampled-target lookup. The structured method imposes the PSD constraints of the exact decoder. The symmetric ablation retains symmetric coefficients, replaces the PSD constraints by $\mathrm{tr}(B_0) = s$, and selects minimum coefficient norm. Thus it also changes the scale normalization. The OpInf-style baseline fixes the leading matrix to I_s and fits unrestricted affine coefficients by least squares. This baseline isolates affine fitting under the shared passive archive; it uses neither designed acquisition nor tuned regularization from published OpInf implementations.

The structured method has 195 paired successes where affine least squares fails and one paired failure where it succeeds. Against the symmetric ablation, there are zero wins, one loss, and 1,439 ties. The exceptional archive is KnowUndo, $r = 1$, seed 101, $n = 24$, repetition 4: structured error is 2.554×10^{-5}, while both alternatives have error 3.034×10^{-8}. The alternatives pass the declared 10^{-7} threshold but fail at 10^{-8}. All outcomes are included in the reported statistics. The structured solver returns 1,364 optimal and 76 optimal-inaccurate statuses; no method invokes the singular-target fallback. Errors on failed archives differ substantially: at $(r, n) = (4, 48)$, the largest error is 0.0242 for the structured fit and 621.5 for the symmetric ablation. The results support the symmetric affine representation; they do not isolate a PSD contribution to success probability.

D.2.3 PSD ablation with common normalization

To separate PSD constraints from the normalization change in Appendix D.2.2, we compare fits with common $\mathrm{tr}(B_0) = r$, identical endpoint coercivity constraints, equality fitting, and minimum coefficient norm. Only the PSD constraints on the deletion matrices B_j differ. Intermediate target systems can be indefinite for the symmetric comparator. We test $r \in \{1, 2\}$ and $n \in \{24, 32, 64\}$,

Table 4: All-target success (%) in the matched affine comparison. Each row contains 120 archives nested within 12 fixed tasks.

r	n	Structured	Symmetric, no PSD	OpInf-style LS
1	24	54.2	55.0	55.0
1	32	100	100	100
1	48	100	100	100
1	64	100	100	100
2	24	0	0	0
2	32	70.0	70.0	0
2	48	100	100	100
2	64	100	100	100
4	24	0	0	0
4	32	0	0	0
4	48	45.8	45.8	0
4	64	100	100	53.3

with 24 archives per configuration, totaling 144 paired archives and 288 fits. All 288 solver statuses are optimal.

Success uses the original 10^{-7} scaled all-catalog parameter criterion. The PSD method wins on zero archives, loses on two, and ties on 142. Table 5 shows smaller mean and maximum errors for PSD in some unsuccessful configurations, but no advantage in exact-recovery frequency. Among unsuccessful runs, PSD has median error 0.013906 and maximum 62.749 over 46 runs, versus median 1.7108 and maximum 174.56 over 44 symmetric runs. These method-specific failure sets differ, so their summaries are not a paired comparison. PSD is not uniformly better and can itself have large errors.

Table 5: PSD ablation with common normalization. Each method has 24 runs per configuration.

r	n	Method	Successes	Mean error	Maximum error
1	24	PSD	15/24	0.0035639	0.041222
1	24	Symmetric	16/24	1.3733	31.249
1	32	PSD	24/24	1.2141×10^{-12}	1.5212×10^{-11}
1	32	Symmetric	24/24	5.0167×10^{-13}	9.6952×10^{-12}
1	64	PSD	24/24	1.8133×10^{-13}	1.3621×10^{-12}
1	64	Symmetric	24/24	2.0612×10^{-14}	9.2314×10^{-14}
2	24	PSD	0/24	12.182	62.749
2	24	Symmetric	0/24	34.315	174.56
2	32	PSD	11/24	0.003578	0.020689
2	32	Symmetric	12/24	9.2267	150.74
2	64	PSD	24/24	3.7544×10^{-11}	7.2009×10^{-10}
2	64	Symmetric	24/24	3.0686×10^{-12}	3.4795×10^{-11}

D.3 Acquisition and precision scaling

D.3.1 Constructed acquisition and precision obstructions

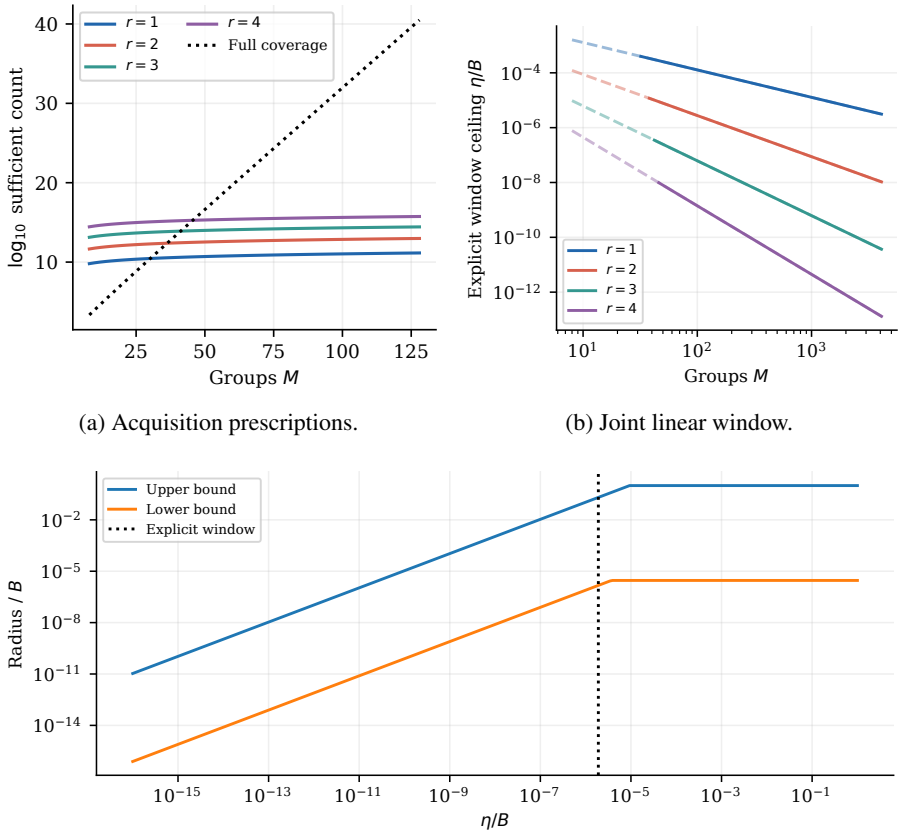
The gate study uses the actual records in Appendix A.3, varying $r = 1, 2, 3, 4$, $L = 2, 4, 8$, and $a = 0.25, 0.5, 1, 2, 4$, with $n = \lceil a2^r L \rceil$ and 1,000 archives per cell. Across all 60,000 archives, the maximum absolute difference between empirical no-activation frequency and $(1 - 2^{-r})^n$ over configurations is 0.035. Active counts and terminal-incidence ranks are distinct diagnostics.

The extended precision study uses $n = 512$, six group counts $M \in \{128, 256, 512, 1024, 2048, 4096\}$, ranks one through four, and 100 archives per configuration. It has 2,400 independent archives; three error levels reuse each archive. Independent $\mathrm{Binomial}(M/2, 1/2)$ positive- and negative-sign inclusion counts exactly reproduce the fair-mask distribution of (u_J, ℓ_J). Stable resolvent identities compute full-vector differences for fixed

parameters $h = 0, 1/32$. Figure 2 shows archive percentiles, not confidence intervals for the median.

At each error level, a sampled pair whose gap is at most 2η is replaced by its midpoint; otherwise each instance supplies its exact parameters. Errors are globally bounded, and the observations coincide when every sampled gap is at most 2η. The empirical log–log slopes of the median target-to-observation separation ratios are $0.991, 1.517, 2.005, 2.518$ for ranks one through four. The corresponding absolute target separations are $3.247 \times 10^{-4}, 5.802 \times 10^{-6}, 1.036 \times 10^{-7}, 1.851 \times 10^{-9}$. These values support the amplification and shrinking-signal observations in Section 6.

D.3.2 Numerical evaluation of the certified precision regime



(a) Acquisition prescriptions.

(b) Joint linear window.

(c) $M = 128$, $r = 2$, $n = n_{\mathrm{rob}}$.

Figure 3: Certified robust regimes at $\kappa = 4$, $B = 1$, and $\delta = 0.05$. (a) Sufficient robust acquisition and full-coverage counts. (b) The explicit joint linear window; solid segments indicate where the robust sufficient count is smaller than the full-coverage prescription. (c) Upper and lower recovery-radius bounds at the sufficient robust budget n_{rob}; the dotted line marks the explicit precision-window ceiling.

At $M = 128$, $\kappa = 4$, $B = 1$, and $\delta = 0.05$, using the explicit choice $C_{\mathrm{rob}} = 2{,}000{,}000$ from Appendix B.3 in Eqs. (16) and (B.35) gives the following nonempty information-theoretic regimes. The table evaluates the theorem's sufficient conditions; it reports no robust-decoder runs.

D.4 Recovery under archive noise and spectral tails

D.4.1 Unprojected deletion-curvature spectra

We measure the deleted curvature before the Gaussian projection, using the same saved features and deletion records as the recovery experiments. Each visual dataset contributes the five record-

Table 6: Robust sufficient counts and explicit precision windows at $M = 128$.

r	Sufficient count	Maximum η/B
1	1.415×10^{11}	9.967×10^{-5}
2	9.465×10^{12}	1.907×10^{-6}
3	2.713×10^{14}	3.770×10^{-8}
4	5.541×10^{15}	7.604×10^{-10}

selection seeds; each language dataset contributes one shared set of 80 deletion records, since its five seeds only change the subsequent projection. No centering, additional normalization, or feature truncation is applied. The underlying records are used only for this diagnostic.

Let X_{del} stack all deletable feature vectors, and let $\gamma_1 \geq \gamma_2 \geq \cdots$ be the eigenvalues of $H_{\mathrm{del}} = X_{\mathrm{del}}^\top X_{\mathrm{del}}$. We compute them by squaring the singular values of X_{del}. The trace effective rank and cumulative-energy dimension are given by:

$$r_{\mathrm{tr}} = \frac{\sum_i \gamma_i}{\gamma_1}, \qquad k_a = \min\left\{ k : \frac{\sum_{i=1}^k \gamma_i}{\sum_i \gamma_i} \geq a \right\}. \tag{D.1}$$

For rank-r spectral truncation, the relative trace tail is $\tau_r = \sum_{i>r} \gamma_i / \sum_i \gamma_i$, and the relative operator-norm error is γ_{r+1}/γ_1. Numerical rank counts singular values exceeding $\max\{n_{\mathrm{del}}, d\}\epsilon_{\mathrm{mach}}\sigma_1(X_{\mathrm{del}})$ in float64 arithmetic. Numerical rank is reported separately from the algebraic-rank assumption in the theory.

Table 7: Unprojected curvature spectra. Ranges cover five visual record selections; each language row represents one shared deletion set. $k_{.90}$ and $k_{.99}$ capture 90% and 99% of curvature trace.

Dataset	Sets	d	Num. rank	r_{tr}	$k_{.90}$	$k_{.99}$	τ_4
CIFAR-10	5	512	512	1.290–1.300	16–18	183–189	0.163–0.168
CIFAR-100	5	512	512	1.294–1.302	22–23	193–197	0.179–0.183
KnowUndo	1	3584	80	1.213	6	44	0.113
TOFU	1	4096	80	1.200	2	32	0.072

Despite trace effective ranks near one, the leading four directions leave 7.2–18.3% of curvature trace, with relative operator-norm errors between 0.0108 and 0.0151. These diagnostics quantify curvature truncation, not the resulting all-target recovery error. Theorems 5.1 and 5.2 use algebraic rank. Corollary B.2 instead requires an operator-norm tail bound relative to λ, together with the spectral and label-energy parameters. The reported relative spectra alone do not establish a nontrivial approximate-rank recovery bound.

D.4.2 Recovery on unprojected CLIP features

We extend the binary CIFAR ridge tasks to the original 512-dimensional frozen CLIP features, retaining their feature normalization without projection. The deleted-curvature numerical rank is 512. The archive processor truncates only its SVD response basis to $s \in \{2, 8\}$ dimensions, where s denotes the fitted dimension rather than the deleted-curvature rank. Budgets are $n \in \{32, 128\}$. Three data seeds and two archive repetitions yield six runs per configuration, with 144 method evaluations: 96 affine fits and 48 linear or lookup baselines. All runs use CPU computation on existing features. Repeated archives are nested within data seeds and are not treated as independent datasets. Test labels are never passed to the archive processor.

Metrics and results. For each run, let $\mathcal{U} = 2^{[M]} \setminus \{A_1, \ldots, A_n\}$ denote unseen requests and define $E_{\mathrm{unseen}} = \max_{J \in \mathcal{U}} \|\widehat{\theta}_J - \theta_J\|_2$. Tables 8 and 9 report its mean, median, and maximum across runs. Prediction RMSE relative to exact retraining is defined per request by

$$\mathrm{RMSE}_J = \left[\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \left(x_i^\top \widehat{\theta}_J - x_i^\top \theta_J \right)^2 \right]^{1/2}. \tag{D.2}$$

We first average RMSE_J arithmetically over $J \in \mathcal{U}$, then average over data seeds and archive repetitions. Squared errors are not pooled across requests before taking the square root. This metric measures prediction fidelity to exact retraining, not error against test labels. The separate test-label metrics are MSE without a square root and classification accuracy, each averaged over unseen targets and then runs.

At $n = 128$, the $s = 8$ PSD fit improves mean E_{unseen} and prediction RMSE over linear mask regression on both tasks, but does not achieve exact catalog recovery. At $n = 32$, some affine fits have errors larger than lookup. High classification accuracy alone is insufficient evidence of recovery: returning the initial model yields accuracy 0.99267 on CIFAR-10 and 0.96600 on CIFAR-100, compared with 0.99168 and 0.96155 for retraining at $n = 128$. These experiments concern binary ridge heads on frozen features, not multiclass classifier retraining or generative knowledge unlearning.

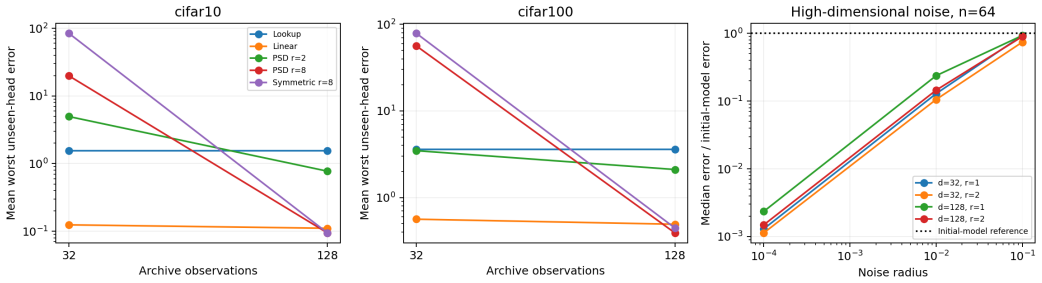


Figure 4: Unprojected-feature recovery (left and middle) and high-dimensional noise (right). Left-panel legends use r for fitted dimension s, not curvature rank. The right panel shows median error ratios at $n = 64$.

The left panels of Figure 4 show means of E_{unseen} and connect the two tested budgets. The right panel is described in Appendix D.4.4. Aggregate curves do not show every run's failure behavior, which is retained in the tables below.

Table 8: CIFAR-10 on unprojected features. Six runs per row.

n	s	Method	Mean E_{unseen}	Median	Maximum	Pred. RMSE
32	–	Linear	0.12429	0.13114	0.14796	0.0019424
32	–	Lookup	1.5488	1.5979	1.6158	0.031494
32	2	PSD	4.9601	0.89295	18.862	0.022756
32	2	Symmetric	24.059	22.365	54.443	0.037478
32	8	PSD	19.834	17.731	47.607	0.019978
32	8	Symmetric	84.202	73.029	145.75	0.062517
128	–	Linear	0.11065	0.11668	0.11941	0.0017076
128	–	Lookup	1.5488	1.5979	1.6158	0.031641
128	2	PSD	0.7723	0.77828	0.81358	0.015987
128	2	Symmetric	0.77227	0.77824	0.81357	0.015985
128	8	PSD	0.094594	0.096954	0.11222	0.0014725
128	8	Symmetric	0.094595	0.096954	0.11222	0.0014726

Measured spectral-tail parameters. For these tasks, $\lambda = 1$, $B = \|y\|_2 = 44.721$ (rounded), $\varepsilon = \sigma_{s+1}(H_{\mathrm{del}})$, and $\kappa = 1 + \|X\|_{\mathrm{op}}^2$. Table 10 evaluates the uncapped expression $2\rho + K_s(\eta + 2\rho)$ at $\eta = 0$, divided by B, using the measured parameters. Every value exceeds one by many orders of magnitude, so the expression reduces to the trivial cap B. In addition, the tested budgets do not satisfy the sufficient sample bound, and the SVD heuristic is not the global rank-constrained decoder of Corollary B.2. Thus the corollary gives no nontrivial precision explanation for these experiments.

Numerical checks. No failed or inaccurate outcome is excluded. Of the 96 full-feature affine fits, 95 return optimal and one optimal-inaccurate status. Each returns a candidate under the declared numerical checks, but optimal-inaccurate is not relabeled optimal. The experiment records retain

Table 9: CIFAR-100 on unprojected features. Six runs per row.

n	s	Method	Mean E_{unseen}	Median	Maximum	Pred. RMSE
32	–	Linear	0.56264	0.55955	0.66792	0.01329
32	–	Lookup	3.6167	3.7018	3.7879	0.097787
32	2	PSD	3.4782	2.3928	6.9085	0.064734
32	2	Symmetric	1055.9	669.55	3457.9	0.32832
32	8	PSD	56.267	55.328	84.703	0.10573
32	8	Symmetric	78.641	55.601	219.75	0.14352
128	–	Linear	0.49193	0.48225	0.52964	0.01181
128	–	Lookup	3.6167	3.7018	3.7879	0.098122
128	2	PSD	2.1078	2.1293	2.2473	0.058135
128	2	Symmetric	2.1087	2.1297	2.2474	0.05814
128	8	PSD	0.39122	0.37345	0.5533	0.0066312
128	8	Symmetric	0.44563	0.33097	0.97048	0.0066146

Table 10: Measured tail parameters and uncapped bound divided by B. Ranges span data seeds.

Dataset	s	ε	κ	Bound$/B$
CIFAR-10	2	12.896–13.286	1515–1516.5	$(5.371\text{–}5.450) \times 10^{13}$
CIFAR-10	8	3.5934–3.6712	1515–1516.5	$(5.691\text{–}5.801) \times 10^{32}$
CIFAR-100	2	8.5839–9.2987	1521.5–1524	$(4.472\text{–}4.639) \times 10^{13}$
CIFAR-100	8	3.5410–3.5740	1521.5–1524	$(5.879\text{–}5.987) \times 10^{32}$

residuals, constraint violations, smallest target eigenvalues, maximum condition numbers, archive singular values, and unseen-target counts. Independent direct record refits agree with enumerated oracle solutions for masks 0, 85, and 255. On separate exact low-rank synthetic instances of ranks one and two, both equality-fit variants recover an omitted all-delete target with maximum catalog error below 5×10^{-16} without lookup. These checks do not certify the heuristic in the unprojected or noisy settings.

D.4.3 Bounded-noise recovery and controlled spectral tails

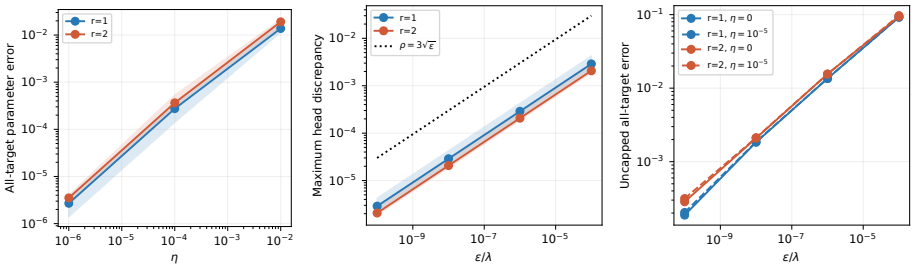
Convex robust special case. We use synthetic instances with $M = 6$, $d = r \in \{1, 2\}$, $\lambda = B = 1$, and $\kappa = 2$. Row and label caps enforce the spectral envelope and label-energy bound. Three data seeds $11, 22, 33$, five archives per instance, $n = 64$, and three noise radii $\eta \in \{10^{-6}, 10^{-4}, 10^{-2}\}$ give 90 fits. Each archived head receives an independent Gaussian direction normalized to norm η; the same masks and directions are reused across noise radii. The joint rank constraint in Eq. (20) is automatic when $d = r$, so the residual-constrained coefficient fit is convex. We evaluate its uncapped predictions on all 64 targets, retain the exact initial head, and do not override predictions by noisy lookup. All 90 statuses are optimal; maximum residual-constraint violation is 4.58×10^{-10}.

Table 11: Worst all-target error over 15 archives per rank and noise radius for the uncapped convex robust fit.

r	$\eta = 10^{-6}$	$\eta = 10^{-4}$	$\eta = 10^{-2}$
1	3.983×10^{-6}	4.197×10^{-4}	1.951×10^{-2}
2	4.867×10^{-6}	5.889×10^{-4}	2.174×10^{-2}

These experiments evaluate the convex special case at budgets below the sufficient theoretical bound, using random bounded errors and uncapped predictions. The uniform adversarial guarantee is established by the theory.

Controlled tail construction. Set $M = 6$, $d = r + 2$, $r \in \{1, 2\}$, and $\lambda = B = 1$, $\kappa = 2$. The 24 deletable rows are partitioned into six groups of four. Construct $X_{\text{del}} = U \operatorname{diag}(\sqrt{\gamma}) V^\top$ from random orthonormal U and orthogonal V, with r leading eigenvalues linearly spaced from 0.2 to 0.1 (the single value 0.2 for $r = 1$), followed by ε and $\varepsilon/2$. The undeletable features are



(a) Convex robust fit ($d = r$). (b) Projection discrepancy. (c) Subspace heuristic ($n = 64$).

Figure 5: Noisy recovery and spectral-tail degradation. (a) median all-target error for the convex robust fit; bands are 10th–90th archive percentiles over 15 archives per point. (b) median uniform head discrepancy between original and projected instances, with ranges over five instances; the dotted line is the analytical bound. (c) median uncapped heuristic error over 15 archives per point at $n = 64$, using exact (solid) and noisy (dashed) observations. Projection is used only for the diagnostic in (b).

$\sqrt{0.1}I_d$, and the full Gaussian label vector is normalized to unit norm. Thus $\sigma_{r+1}(H_{\mathrm{del}}) = \varepsilon$ and $H \preceq 1.3I_d \preceq \kappa I_d$. We vary $\varepsilon \in \{0, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}\}$, using seeds $11, 22, 33, 44, 55$, three archives per instance, $n \in \{32, 64\}$, and $\eta \in \{0, 10^{-5}\}$, for 600 fits. Masks and unit noise directions are paired across tail and noise levels, with the smaller budget a prefix of the larger archive.

Exhaustive retraining evaluates both the original and rank-r projected instances. Over positive tails, the largest ratio of uniform head discrepancy to $\rho = 3\sqrt{\varepsilon}$ is 0.142. Across this range, the discrepancies track the square-root tail scale. At zero tail, remaining discrepancies are numerical roundoff. The comparison projector is used only to compute these diagnostics.

Archive-only heuristic. The fitting procedure receives the original head, noisy archived heads, masks, and public parameters. It estimates an r-dimensional basis from the first $\lfloor n/3 \rfloor$ archived displacements by SVD, then solves a convex coefficient fit in that basis using all observations. It imposes the spectral and PSD constraints and bounds the augmented residual, including the perpendicular observed displacement, by $\kappa(\eta + 2\rho)$. Restricting the fit to an estimated basis can make the problem infeasible. Corollary B.2 therefore does not directly apply to this heuristic.

The 600 fits have 596 optimal, three optimal-inaccurate, and one infeasible-inaccurate statuses. The latter occurs at $r = 2$, seed 11, repetition 0, $n = 32$, $\varepsilon = 0$, $\eta = 10^{-5}$ and returns the initial-head fallback, with all-target error 0.0970; it remains in the results. Returned fits have residual violation at most 7.25×10^{-11} and structural margin at least -5.78×10^{-13}. At $n = 64$ and $\eta = 0$, median uncapped errors increase from 1.87×10^{-4} to 0.0923 for $r = 1$ and from 2.85×10^{-4} to 0.0944 for $r = 2$ as ε increases from 10^{-10} to 10^{-4}. Zero-tail medians are below 2.11×10^{-12}.

The saved results also apply the public zero-output rule when $2\rho + K_r(\eta + 2\rho) > B$, which occurs at $\varepsilon = 10^{-4}$ on this grid. Across both budgets, capped median errors there range from 0.157 to 0.182, compared with uncapped medians from 0.0923 to 0.0970. Exact observed-head lookup is used only when $\eta = 0$, and the initial head is always retained. Figure 5 displays uncapped errors to show the behavior of the fitted predictor. The cap guarantees the elementary radius B on capped cells; the plotted finite-budget heuristic errors are not certified by the stronger spectral-tail expression.

D.4.4 High-dimensional bounded-noise recovery

We test an empirical SVD-based response fit at $d \in \{32, 128\}$, $r \in \{1, 2\}$, and $n \in \{32, 64\}$. The initial grid uses $\eta \in \{0, 10^{-4}, 10^{-2}\}$, with three data seeds and two archive repetitions per configuration, giving 144 fits. An additional 48 fits at $\eta = 0.1$ were added after observing this grid and are explicitly exploratory stress tests. Noise directions are normalized Gaussian draws rather than adversarial errors. The least-squares objective includes a fixed 10^{-8} penalty, contributing to the

noiseless numerical floor. This empirical fit is distinct from both the exact decoder and the theorem's global rank-constrained robust construction.

Let $E_0 = \max_{J \subseteq [M]} \|\theta_0 - \theta_J\|_2$ denote the all-catalog error of returning the initial model. We compute E_{unseen}/E_0 within each run and then report its median. The numerator uses unseen targets, whereas the denominator uses the full catalog. The subspace diagnostic is the sine of the largest principal angle between learned and true response subspaces. Tables 12 and 13 include absolute errors, median ratios, counts above the initial-model reference, and median angle sines.

At $\eta \leq 10^{-2}$, no tested fit exceeds the initial-model reference. At $\eta = 0.1$, 14 of 48 fits exceed it, and median angle sines approach one. At $n = 64$, all four median ratios remain below one, although 3 of 24 individual stress-test fits exceed one. Thus the median curves in Figure 4 do not imply uniform improvement. All 192 statuses are optimal, including the stress tests.

Table 12: High-dimensional noise at $n = 32$. Six runs per row. F counts errors exceeding the initial-model reference. Rows at $\eta = 0.1$ are exploratory.

d	r	η	Median error	Maximum	Median ratio	F	Angle sine
32	1	0	2.022×10^{-9}	2.1447×10^{-8}	5.0843×10^{-8}	0/6	1.7167×10^{-15}
32	1	0.0001	8.2955×10^{-5}	0.00011341	0.0022103	0/6	0.001646
32	1	0.01	0.006974	0.0088523	0.19122	0/6	0.16295
32	1	0.1	0.042758	0.048624	1.042	4/6	0.99026
32	2	0	1.8222×10^{-7}	1.072×10^{-5}	4.3574×10^{-6}	0/6	1.5983×10^{-15}
32	2	0.0001	0.00013986	0.00041123	0.0027832	0/6	0.0026492
32	2	0.01	0.0098191	0.011754	0.18437	0/6	0.25484
32	2	0.1	0.046722	0.059548	0.93914	2/6	0.93902
128	1	0	2.8148×10^{-9}	1.0397×10^{-8}	1.5922×10^{-7}	0/6	8.7471×10^{-15}
128	1	0.0001	7.1694×10^{-5}	0.00010117	0.0037563	0/6	0.0038633
128	1	0.01	0.0068104	0.0083842	0.3567	0/6	0.3595
128	1	0.1	0.019815	0.021606	0.99226	3/6	0.9936
128	2	0	1.1548×10^{-8}	0.00021004	5.1809×10^{-7}	0/6	3.258×10^{-15}
128	2	0.0001	8.7084×10^{-5}	0.000705	0.0032231	0/6	0.00327
128	2	0.01	0.0077787	0.016361	0.29476	0/6	0.31381
128	2	0.1	0.034496	0.036311	0.96907	2/6	0.99741

Table 13: High-dimensional noise at $n = 64$. Six runs per row. F counts errors exceeding the initial-model reference. Rows at $\eta = 0.1$ are exploratory.

d	r	η	Median error	Maximum	Median ratio	F	Angle sine
32	1	0	1.2732×10^{-9}	3.5994×10^{-9}	3.6963×10^{-8}	0/6	1.7543×10^{-15}
32	1	0.0001	4.8133×10^{-5}	5.448×10^{-5}	0.0013008	0/6	0.0012636
32	1	0.01	0.0048501	0.0058721	0.1289	0/6	0.12638
32	1	0.1	0.032051	0.041159	0.93047	1/6	0.9596
32	2	0	6.0783×10^{-8}	4.7002×10^{-7}	8.39×10^{-7}	0/6	1.8494×10^{-15}
32	2	0.0001	6.9837×10^{-5}	7.8275×10^{-5}	0.0011338	0/6	0.0016889
32	2	0.01	0.0059753	0.0075031	0.10562	0/6	0.1733
32	2	0.1	0.037754	0.051177	0.74257	0/6	0.96164
128	1	0	1.687×10^{-9}	3.6474×10^{-9}	7.9114×10^{-8}	0/6	8.7471×10^{-15}
128	1	0.0001	4.5036×10^{-5}	5.4182×10^{-5}	0.0023555	0/6	0.0025471
128	1	0.01	0.0045543	0.0052832	0.23732	0/6	0.24478
128	1	0.1	0.017907	0.019639	0.92879	1/6	0.99816
128	2	0	2.5898×10^{-9}	9.2558×10^{-9}	8.6469×10^{-8}	0/6	1.4195×10^{-15}
128	2	0.0001	5.0806×10^{-5}	5.8222×10^{-5}	0.0014904	0/6	0.0021147
128	2	0.01	0.0049174	0.0057069	0.14501	0/6	0.20471
128	2	0.1	0.031315	0.035875	0.89665	1/6	0.9974

E Archive Costs, Computational Scaling, and Amortization

E.1 Archive provenance and costs

The acquisition model generates every retained-data head from the same original dataset under iid uniform deletion masks. An existing archive must have that same sampling law; ordinary nested deletion logs do not meet it. Let $c(J)$ be the cost of preparing a head, C_{fit} the basis and equation fitting cost, and $c_{\mathrm{query}}(J)$ the fitted-system query cost. For a later query list $J_1, \ldots, J_Q$,

$$C_{\mathrm{prep}} = c(\varnothing) + \sum_{i=1}^{n} c(A_i), \qquad C_{\mathrm{total}} = C_{\mathrm{prep}} + C_{\mathrm{fit}} + \sum_{t=1}^{Q} c_{\mathrm{query}}(J_t). \tag{E.1}$$

Reusing duplicate heads can reduce computation, although each sampled occurrence counts toward n. For a downstream user of an existing archive, preparation is a sunk cost; it remains part of the total system cost. End-to-end savings require future query savings to offset preparation and fitting. The measurements below quantify implementation-specific amortization thresholds for synthetic instances, including archive preparation and statistics-based comparisons. The comparison distinguishes existing and manufactured archives, dense and low-rank statistics, and the role of request caching.

E.2 Instances and measurement protocol

We evaluate 20 configurations with three fixed data seeds each, giving 60 decoder fits in separate processes. All failures and optimal-inaccurate outcomes are retained. These experiments measure implementation costs at the tested archive budgets, not the worst-case sufficient sample bound.

Instances and evaluation coverage. Each deletable group contains four records, and there are $2d$ undeletable background records, for $N = 4M + 2d$. Deleted curvature has rank exactly r, while the background covariance is generally dense. Seeds are 11, 22, and 33. Archive masks are iid uniform with replacement, including duplicate acquisition costs. Budgets n for the same instance use prefix archives. Except for the additional $(M, r, d, n) = (8, 8, 128, 512)$ configuration, each sweep changes one parameter around $(8, 2, 128, 128)$.

For $M \leq 12$, evaluation covers every deletion target. For $M \in \{16, 24, 32\}$, it covers only 256 independent random targets plus the empty and full-deletion endpoints. Success means maximum parameter error over the evaluated set is at most $10^{-7} \max\{1, \max_J \|\theta_J\|_2\}$, with the maximum also taken over that set. Thus the larger-M results do not verify full-catalog recovery. There are 57 successes out of 60 fits: 48/51 under exhaustive evaluation and 9/9 under sampled evaluation. All three failures occur at $(8, 2, 128, 16)$. Archive coverage and query lookup hit rates can be high at small M or large n, so reported query timings should not be interpreted as timings exclusively for unseen requests.

Timing and memory protocol. Runs execute sequentially on an Intel Core Ultra 9 285K with one CPU thread. The parent and decoder subprocesses are pinned to the same available core. We use NumPy 2.2.6, SciPy 1.15.3, CVXPY 1.7.5, and CLARABEL 0.11.1. Timings exclude data synthesis, process startup, imports, and interprocess communication, but include actual numerical computation. Record and statistics queries are timed over three rounds, and archive queries over ten. Tables report medians across three instances, except maximum error, which is the worst across instances. Microsecond differences are not interpreted as statistically significant.

Each decoder process receives only the original head, archived heads, masks, and public parameters. It receives no records, Hessian, or oracle. Total fit time includes response-space discovery, equation construction and SVD, CVXPY modeling and compilation, solving, numerical feasibility checks, and one empty-target output. SDP time includes modeling through solution, whereas solver time is the solver's own reported duration. Target predictions are computed on demand, so catalog enumeration is not included in fit time.

Peak memory is Linux `VmHWM` from `/proc/self/status` in each independently executed address space, including Python, NumPy, CVXPY, and solver runtimes. We also report the increase

over the pre-fit high-water mark, which is not an exact allocation count. All 60 runs use this measurement after preliminary timing revealed that `ru_maxrss` can inherit the parent's historical high-water mark.

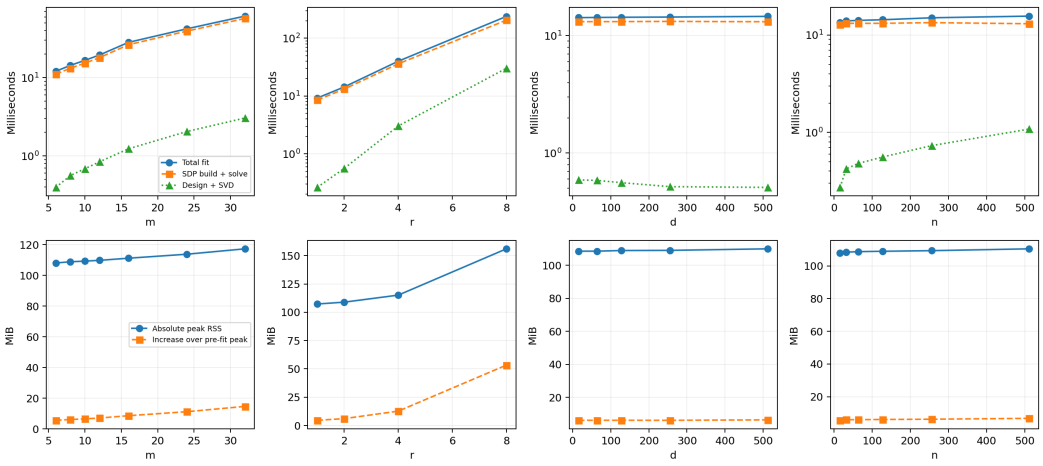


Figure 6: Exact-decoder scaling. Columns vary M, r, d, and n. Top: timing medians with min–max fit bands. Bottom: peak memory and its increase over the pre-fit peak.

E.3 Scaling results

At the reference configuration, median fit time is 14.37 ms, including 13.20 ms for SDP modeling through solving and 0.838 ms reported by the solver. Increasing rank from one to eight raises fit time from 9.22 to 238.43 ms and peak memory from 107.29 to 156.00 MiB. At fixed $r = 2$, changing d from 16 to 512 has a smaller effect over this grid, with fit times from 14.27 to 14.62 ms. Figure 6 and Tables 14–15 include all configurations, including unsuccessful fits.

Table 14: Exact-decoder timing medians (ms). SDP includes modeling through solving.

M	r	d	n	Fit	Design/SVD	SDP	Solver
6	2	128	128	12.048	0.395	11.091	0.518
8	1	128	128	9.222	0.263	8.516	0.111
8	2	16	128	14.268	0.591	13.177	0.813
8	2	64	128	14.314	0.585	13.19	0.777
8	2	128	16	13.476	0.273	12.861	0.643
8	2	128	32	13.989	0.423	13.203	0.833
8	2	128	64	14.156	0.481	13.196	0.83
8	2	128	128	14.367	0.56	13.2	0.838
8	2	128	256	15.049	0.732	13.414	0.847
8	2	128	512	15.65	1.084	13.071	0.822
8	2	256	128	14.432	0.518	13.246	0.835
8	2	512	128	14.621	0.51	13.162	0.834
8	4	128	128	40.018	3.033	36.099	10.822
8	8	128	128	238.43	30.219	205.35	98.616
8	8	128	512	289.94	80.959	204.56	98.815
10	2	128	128	16.643	0.684	15.346	1.279
12	2	128	128	19.495	0.839	18.026	1.877
16	2	128	128	28.311	1.232	26.424	6.709
24	2	128	128	42.134	2.046	39.377	11.141
32	2	128	128	61.29	3.05	57.148	19.682

Table 15: Evaluation success and peak memory (MiB). For $M > 12$, success refers only to sampled targets and endpoints.

M	r	d	n	Successes	Peak	Increase	Worst error
6	2	128	128	3/3	108.14	5.582	8.5294×10^{-14}
8	1	128	128	3/3	107.29	4.6016	2.6019×10^{-14}
8	2	16	128	3/3	108.52	6.0703	3.7144×10^{-15}
8	2	64	128	3/3	108.54	6.082	3.2622×10^{-13}
8	2	128	16	0/3	107.96	5.582	0.19427
8	2	128	32	3/3	108.38	5.9102	2.6554×10^{-14}
8	2	128	64	3/3	108.7	5.9883	2.5134×10^{-14}
8	2	128	128	3/3	108.92	6.0859	2.5047×10^{-14}
8	2	128	256	3/3	109.3	6.2461	2.3104×10^{-14}
8	2	128	512	3/3	110.42	6.7812	2.3097×10^{-14}
8	2	256	128	3/3	109.03	6.0703	3.0855×10^{-14}
8	2	512	128	3/3	109.95	6.332	1.8953×10^{-14}
8	4	128	128	3/3	115.14	12.703	1.5531×10^{-13}
8	8	128	128	3/3	156	53.328	1.4919×10^{-13}
8	8	128	512	3/3	176.15	72.539	1.3771×10^{-13}
10	2	128	128	3/3	109.36	6.5117	4.6423×10^{-14}
12	2	128	128	3/3	109.86	7.0664	5.1103×10^{-14}
16	2	128	128	3/3	111.25	8.5898	7.7335×10^{-13}
24	2	128	128	3/3	113.85	11.188	1.5615×10^{-14}
32	2	128	128	3/3	117.36	14.648	1.1096×10^{-13}

E.4 Preparation and query costs

For $(M, r, n) = (8, 2, 128)$, we compare direct retained-record retraining, dense group-statistics queries, a low-rank statistics oracle, and fitted archive queries. Archive construction by direct retraining includes n actual record refits. Construction from statistics includes statistics setup and n statistics-based head solves. Both count repeated masks. Low-rank statistics preprocessing includes statistics construction, deleted-curvature eigendecomposition, and full-Hessian solves. Full-catalog precomputation includes statistics construction and every target head. The initial head's direct training time is reported separately.

At $d = 128$, direct preparation of the archive takes 19.31 ms, followed by 14.37 ms of fitting, and precomputing all 256 heads takes 17.97 ms. Median single-query times are $148.97\,\mu$s for record refitting, $70.75\,\mu$s for dense statistics, $9.19\,\mu$s for low-rank statistics, and $8.91\,\mu$s for fitted archives. Batched archive timings in Figure 7 and Table 17 are amortized per request rather than single-request latency.

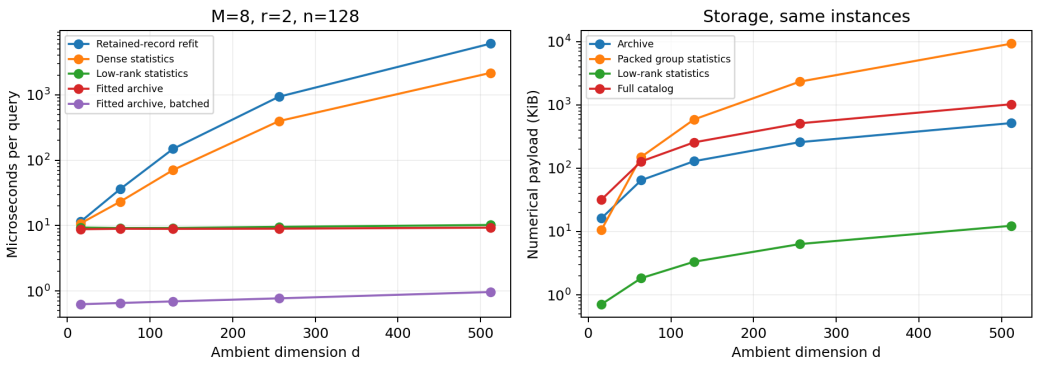


Figure 7: Query costs and storage at $M = 8, r = 2, n = 128$. Batched timings are amortized per request. Storage counts numerical payload rather than process memory.

Table 16: Preparation and fitting medians (ms). Record archive and statistics archive include duplicate masks.

d	Initial	Record archive	Stats setup	Stats archive	LR setup	Fit	Catalog
16	0.087	2.549	0.064	1.476	0.152	14.268	2.774
64	0.187	5.559	0.1	3.005	0.339	14.314	5.975
128	0.357	19.308	0.275	9.03	0.924	14.367	17.972
256	1.316	119.81	2.802	51.994	5.599	14.432	102.53
512	6.493	771.93	11.469	274.9	27.41	14.621	551.22

Table 17: Median query costs (μs). Batched archive values are per request.

d	Records	Dense stats	Low-rank stats	Archive	Batched archive
16	11.454	10.783	9.332	8.803	0.627
64	36.154	23.198	9.165	8.938	0.654
128	148.97	70.75	9.185	8.913	0.693
256	933.11	395.73	9.561	8.97	0.771
512	6050.5	2160.1	10.213	9.3	0.963

E.5 Storage accounting

Payload counts use float64 values and the indicated mask or group-membership encodings, excluding Python object overhead and file compression. Raw-data storage includes all features, labels, and membership codes. An archive includes the initial head, n observed heads, and masks. Packed statistics exploit Hessian symmetry, whereas dense statistics match the arrays used for timings. A global H, b pair alone is insufficient for arbitrary group deletion, so the comparison includes global and per-group statistics.

Ignoring mask bytes, archive and packed-statistics payloads are respectively $8(n+1)d$ and $8(M+1)[d(d+1)/2+d]$ bytes. The archive is smaller exactly when

$$n + 1 < (M + 1)(d + 3)/2. \tag{E.2}$$

This condition must also be compatible with a sufficient archive budget. Substituting the sufficient fixed-rank scale $n = O(2^r M r^2)$ makes the required ambient-dimension scale depend on $2^r r^2$ as well. Raw-data counts use the synthetic sample size $N = 4M + 2d$.

At the reference configuration, the archive occupies 129.12 KiB, packed statistics 589.50 KiB, and the low-rank statistics oracle 3.34 KiB. The fitted basis/operator payload can be much smaller than the archive, but preserving exact observed-head lookup also requires the distinct observed heads and masks. Table 18 distinguishes the raw n-entry archive, fitted operators, and deduplicated lookup payload. Values follow the precision of the supplied summary, so scientific-notation entries are rounded byte counts.

Low-rank statistics oracle. With record access, choose an orthonormal basis U of $\mathrm{range}(H_{\mathrm{del}})$ and define

$$W = H^{-1}U, \qquad G = U^\top W, \qquad C_j = U^\top H_j U, \qquad w_j = C_j U^\top \theta_0 - U^\top b_j. \tag{E.3}$$

Writing $C_J = \sum_{j\in J} C_j$ and $w_J = \sum_{j\in J} w_j$, the oracle returns

$$\theta_J = \theta_0 + W(I - C_J G)^{-1} w_J. \tag{E.4}$$

Indeed $(H - UC_JU^\top)W = U(I - C_JG)$, and the retained system is positive definite. Deployment payload is $O(dr + Mr^2)$, the same order as the fitted archive operators, without random archive acquisition. Every configuration compares this oracle against direct record retraining at absolute and relative tolerances 10^{-9}. This comparator uses record-derived statistics, while the archive decoder uses only observed model parameters.

E.6 AMORTIZATION THRESHOLDS

For archive query time t_a, comparator time t_b, and setup costs C_a, C_b, we compute

$$Q = \max\left\{0, \left\lceil \frac{C_a - C_b}{t_b - t_a} \right\rceil \right\} \tag{E.5}$$

only when evaluated requests pass the recovery criterion and $t_b > t_a$. Existing archives treat acquisition as a sunk cost. Manufactured archives include initial-head training, n record refits, and fitting. Statistics-manufactured archives instead include statistics setup and n statistics solves, while retaining the same direct initial-head training cost. Comparator setup costs are included. Thresholds are computed per instance and summarized across instances, so they need not equal a ratio formed from separately rounded table medians.

At $d = 128$, the reported thresholds against record refitting are 103 queries for an existing archive and 243 for a manufactured archive. Against dense statistics, they are 227 and 545, respectively. These implementation-specific timing intersections assume requests are processed without result caching and ignore timing uncertainty. For thresholds above 2^M, full-catalog precomputation or request caching provides an alternative. Low-rank statistics and archive queries have similar microsecond costs, so their timing intersections are sensitive to measurement variability. For $M > 12$, full-catalog costs are extrapolated from mean query time rather than measured by enumeration.

Table 18: Input and statistics payloads (bytes, at reported precision).

d	Records	Archive	Packed stats	Dense stats	Low-rank stats
16	8768	16640	10944	19584	728
64	83360	66176	1.5437×10^5	2.9952×10^5	1880
128	2.975×10^5	1.3222×10^5	6.0365×10^5	1.1889×10^6	3416
256	1.119×10^6	2.6432×10^5	2.3869×10^6	4.737×10^6	6488
512	4.3349×10^6	5.2851×10^5	9.4925×10^6	1.8911×10^7	12632

Table 19: Fitted operators, deduplicated lookup, and catalog payloads (bytes).

d	Operators	Lookup	Catalog
16	728	13029	32768
64	1880	52326	1.3107×10^5
128	3416	1.0762×10^5	2.6214×10^5
256	6488	1.967×10^5	5.2429×10^5
512	12632	4.3838×10^5	1.0486×10^6

Table 20: Amortization thresholds in queries. E: existing archive. R: record-manufactured archive. S: statistics-manufactured archive. Column suffix identifies the comparator.

d	E/records	R/records	E/stats	R/stats	S/stats
16	5304	6294	7326	8115	8090
64	530	743	987	1393	1212
128	103	243	227	545	382
256	16	147	30	346	176
512	3	132	2	365	138

E.7 NUMERICAL CHECKS AND FAILURES

Solver statuses are 56 optimal and four optimal-inaccurate. The only evaluation failures are $(M, r, d, n) = (8, 2, 128, 16)$ at seeds 11, 22, and 33, all with optimal status, errors 0.19427, 0.13343, and 0.12393, and 239 unseen targets each. They remain included in costs and results. Feasible coefficients at an insufficient archive budget need not recover unseen targets.

The default and on-demand decoder interfaces agree at absolute and relative tolerances 10^{-10} across nine configurations with $r \in \{1, 2, 4\}$ and $n \in \{8, 32, 128\}$, including insufficient archives. The on-demand interface checks group PSD constraints and the full-deletion endpoint rather than only the queried systems. Each formal configuration checks both statistics-query implementations against record retraining. These synthetic-record experiments require no external feature cache or GPU.