Regression Coresets: Attainable Fits, Conditioning, and Reuse

Mantis Discovery

Abstract

Weighted regression coresets may reproduce one least-squares solution while remaining poorly conditioned or failing when the regularization parameter or response changes. We characterize what a fixed weighted subset can preserve in each setting. For a prescribed support, the closure of its weighted least-squares solutions is the union of the bounded faces of a hyperplane arrangement. On supports of size $d + 1$, this set is the convex hull of the d-row interpolants, which permits rational certificates for the optimal fit. A recursive construction also gives new upper bounds for the scalar minimax risk, including $8/5 \le R_5(7) \le 5/3$. For exact coresets, we formulate convex reweighting problems that improve conditioning on a fixed support or bound covariance inflation during construction. Finally, a weighted subset preserves the entire ridge path exactly when it matches the response cross moment and the action of the covariance on a Krylov subspace. Experiments on 16 configurations from three tabular data sources improved every initial condition number, by a median factor of 6.46, but usually increased sensitivity to new labels.

1 Introduction

A weighted subset can reproduce a least-squares fit with very few examples. What can that subset actually do? It may return the right coefficients but have a poorly conditioned covariance. It may preserve one fit but fail after the regularization parameter or labels change. These distinctions matter when a coreset is retained as training data rather than used only to report a fitted model.

Hanneke et al. [1, 2] showed that $2d$ weighted examples reproduce every scalar least-squares fit, while d examples can incur a loss ratio of $d + 1$. Zhang [3] obtained the value at budget $2d - 1$, two further low-dimensional cases, and a harmonic lower bound; the vector-valued threshold is treated in [4]. Moment coresets preserve the full quadratic objective with $O(d^2)$ rows [5, 6]. These support bounds do not by themselves control conditioning or identify the best attainable fit at an intermediate budget.

Theorem 3.1 characterizes the coefficients attainable on a fixed support. For $d + 1$ rows, Theorem 3.2 reduces optimization to a convex quadratic program; a four-row counterexample shows why this reduction does not extend to arbitrary support sizes. Theorem 3.3 then improves upper bounds for $R_d(d + k)$ in part of the open range $k < d/2$.

The conditioning results address a different weakness of exact coresets. Theorem 4.1 optimizes the weights on an existing exact support, while Theorem 4.2 constructs an exact support with a bound on its covariance. The ridge results characterize preservation across all regularization parameters and give generic lower bounds on the number of original observations required. The same Krylov condition preserves gradient dynamics from zero. For changed responses, Theorem 5.9 separates the cost of discarding observations from the additional cost of cancelling the observed residual.

Prior work and techniques. The interpolation formula used at $d + 1$ is due to Giannelli et al. [17]. Positive quadrature and conic compression are established tools [5, 6], as are convex condition-number minimization [23, 24], cutting planes, and column generation in experimental design [18, 19]. We impose residual equations on these conditioning problems so that reweighting preserves the observed fit. Krylov

subspaces already support fast ridge-path computation [7, 8]; here they characterize necessity and the number of original observations that must be retained. Statistical differences between regression sketches [10] and their derivatives [11] motivate the analysis of changed responses. The proofs cite volume sampling and positive bases where those standard results are used [12–14].

2 Regression and exact subsets

Let $X \in \mathbb{R}^{N\times d}$ have rows $x_i^\top$, and let $y \in \mathbb{R}^N$. Write

$$L(w) = \|Xw - y\|^2, \quad H = X^\top X, \quad w_\star = X^+ y, \quad r = Xw_\star - y, \quad E = \|r\|^2.$$

Given nonnegative weights a, not all zero, let w_a be the Euclidean minimum-norm minimizer of $\sum_i a_i(x_i^\top w - y_i)^2$. Its support is $S = \{i : a_i > 0\}$, so rows with zero weight are not retained. We call (S, a) an exact coreset for a specified family of fits if w_a reproduces every coefficient vector or matrix in that family with the original responses. Unless stated otherwise, both S and a remain fixed across the family.

Proposition 2.1 (Selected least squares)**.** *Set $M = X^\top \operatorname{diag}(a) X$, $g = X^\top \operatorname{diag}(a) r$, and $P = M^+ M$. Then*

$$\begin{aligned} w_a - w_\star &= -(I - P)w_\star - M^+ g, \\ L(w_a) - E &= e^\top H e, \qquad e = (I - P)w_\star + M^+ g. \end{aligned} \tag{1}$$

Thus $w_a = w_\star$ exactly when $g = 0$ and $w_\star \in \operatorname{range}(M)$. If $M \succ 0$, cancellation of g alone is sufficient.

Proof. The selected cross moment is $Mw_\star - g$. Apply M^+ and use $X^\top r = 0$ to obtain the two identities. The summands of e lie in the orthogonal spaces $\ker M$ and $\operatorname{range} M$, and $g \in \operatorname{range} M$. □

An invertible feature transformation preserves a unique minimizer but can change the minimum-norm choice from an affine set. We therefore whiten only after establishing that the weighted covariance has full rank. Appendix B gives the minimax reduction and a counterexample for singular supports.

3 Which fits can a weighted subset attain?

Fix a support S and put $V = \operatorname{span}\{x_i : i \in S\}$. The minimum-norm weighted solution lies in V. Ignore zero feature rows and consider the affine hyperplanes

$$\mathcal{H}_i = \{w \in V : x_i^\top w = y_i\}, \qquad i \in S.$$

A face specifies which residuals vanish and the signs of the others. Its closure is a polyhedron in V. The union of the bounded closed faces is the *bounded complex* of the arrangement.

Theorem 3.1 (Attainable coefficients)**.** *Let $\mathcal{W}_S$ be the set of minimum-norm least-squares solutions obtained with strictly positive weights on every row of S. Then*

$$\overline{\mathcal{W}_S} = \bigcup_{F \text{ bounded face of } \{\mathcal{H}_i\}_{i\in S}} \overline{F}. \tag{2}$$

Consequently, for any positive-definite evaluation covariance H, the infimum loss on S equals the minimum obtained by projecting the full coefficient vector onto one of these bounded faces in the H metric.

The statement includes singular supports: the arrangement is formed in their original row span, rather than after a nonorthogonal change of the minimum-norm rule. If $V = \{0\}$, the only attainable coefficient is zero.

The proof is geometric. At a weighted optimum, the normal equations rule out every nonzero recession direction of its face. Conversely, within a bounded face, the signed normals positively span the tangent space. They supply positive cancelling weights there; increasing the weights on the face's equality constraints approaches any prescribed point of that face. Appendix A gives the argument, including boundary points and the minimum-norm condition.

3.1 Supports with $d + 1$ rows

For $|S| = d + 1$ and $\operatorname{rank} X_S = d$, let $\mathcal{B}_S$ contain its nonsingular d-row subsets. Write $w_B = X_B^{-1} y_B$ and $\Delta_j = \det(X_{S\setminus\{j\}})^2$. The determinant formula expressing weighted least squares as a convex combination of interpolants is known [17]. At this particular budget it has a converse.

Theorem 3.2 (Attainable coefficients on $d + 1$ rows)**.** *For a rank-d support S of size $d + 1$,*

$$\overline{\mathcal{W}_S} = \operatorname{conv}\{w_B : B \in \mathcal{B}_S\}. \tag{3}$$

Every strictly positive barycentric vector $(\pi_j)_{\Delta_j>0}$ is realized by the weights $a_j = \Delta_j/\pi_j$ on these indices, with arbitrary positive weights on indices for which $\Delta_j = 0$. Thus the exact infimum on S is the convex quadratic program

$$\min_{\pi\geq 0,\, \mathbf{1}^\top\pi=1} \left\| \sum_{B\in\mathcal{B}_S} \pi_B (w_B - w_\star) \right\|_H^2 . \tag{4}$$

Indeed the known barycentric coefficient of the basis omitting j is proportional to Δ_j/a_j: the product of all selected weights cancels. This proves both directions. Vanishing barycentric coefficients are obtained by limits, so the theorem concerns an infimum when finite weights do not attain the boundary.

For a specified rational instance, enumerating the full-rank $(d + 1)$-row supports and solving (4) is a finite exact procedure. A rational solution of the simplex Karush–Kuhn–Tucker conditions certifies each projection. Singular supports must either be treated by Theorem 3.1 or excluded by the distance bound in Appendix A.1.

The restriction $|S| = d + 1$ matters. For

$$X = \begin{pmatrix} 2 & 0 \\ 2 & -1 \\ -3 & -2 \\ 3 & 3 \end{pmatrix}, \qquad y = (2, -3, 0, -4)^\top,$$

the two basis interpolants $(1, 5)^\top$ and $(8/3, -4)^\top$ have midpoint $w = (11/6, 1/2)^\top$. That midpoint is not even a limit of positive weighted fits on all four rows. Its residual signs admit the nonzero recession direction $(1, 2)^\top$. Thus an arbitrary convex combination of interpolants can fall outside the attainable set once more than one extra row is allowed.

3.2 Upper bounds in the unresolved range

Let $R_d(n)$ be the worst-case ratio for the fixed minimum-norm learner, as defined in Appendix B. The known boundaries are $R_d(d) = d + 1$ and $R_d(2d) = 1$ [1]. Write $n = d + k$ and let $\Lambda(d, k)$ be the harmonic lower bound (24), also obtained by Zhang [3].

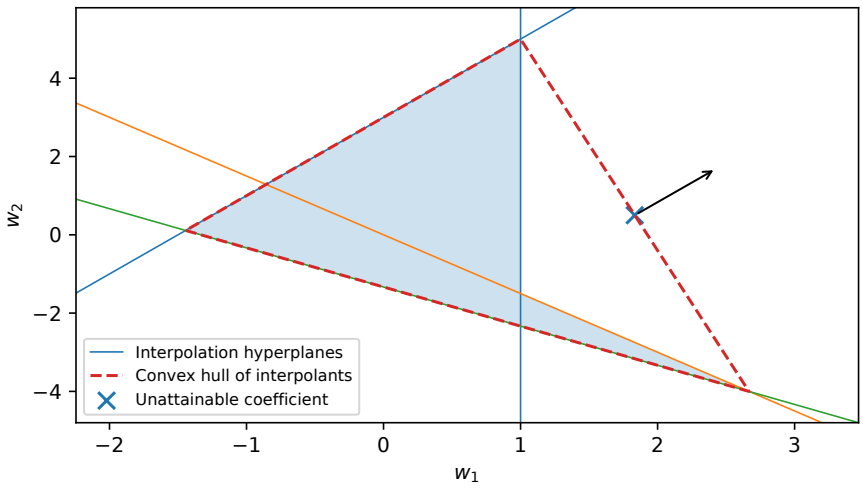


Figure 1: The attainable set need not be convex. Shading marks the bounded faces for the four-row example. The marked midpoint belongs to the convex hull of two interpolants but lies in an unbounded arrangement face; the arrow is a recession direction. It cannot be approached by positive weighted fits on all four rows.

The subspace restriction used for Theorem B.2 also yields an upper bound when $k < d/2$. Define $u_{0,k} = 0$, $u_{d,0} = d$, and $u_{d,k} = 0$ when $k \geq d$. For $1 \leq k < d$, put

$$\begin{aligned} A_{d,k} &= \frac{d-k}{k+1}, \\ C_{d,k} &= \max\left\{\frac{d-k}{d},\ \max_{2\leq s\leq d} u_{d-s,k-1},\ \max_{1\leq z\leq d} u_{d-z,k}\right\}, \\ u_{d,k} &= \min\{A_{d,k}, C_{d,k}, u_{d,k-1}\}. \end{aligned} \tag{5}$$

Use the sharper seeds $u_{3,1} = 2/3$ and $u_{4,1} = 1$ from Zhang [3] in place of this recurrence at those two indices.

Theorem 3.3 (Recursive upper bound)**.** *For every d and $0 \leq k < d$,*

$$1 + \Lambda(d,k) \leq R_d(d+k) \leq 1 + u_{d,k}. \tag{6}$$

The upper construction has full selected rank and approaches the stated bound with finite weights. In particular,

$$\frac{8}{5} \leq R_5(7) \leq \frac{5}{3}, \qquad \frac{5}{3} \leq R_6(8) \leq 2.$$

A positive circuit of rank $s \geq 2$ consumes $s+1$ rows and leaves a quotient of dimension $d-s$ with extra budget $k-1$. Zero-residual directions give the second reduction in (5). If neither reduction applies, the gradients lie on independent stationary lines and the excess is at most $(d-k)/d$. The bound $A_{d,k}$ comes from volume-sampled interpolation constraints followed by exact selection in the remaining subspace. The proof is in Appendix C.

4 Conditioning exact coresets

Exact residual cancellation does not control conditioning. We therefore study two problems: reweighting an existing exact support, and constructing an exact support while controlling its covariance. Both formulations allow scalar or matrix responses.

d	n	Harmonic lower bound	Anchor upper bound	Recursive upper bound
5	7	8/5	2	5/3
6	8	5/3	7/3	2
7	10	11/7	2	5/3
8	11	13/8	9/4	2
9	13	14/9	2	5/3
10	14	8/5	11/5	2

Table 1: All ratios include the optimal loss. Every displayed budget lies below $\lceil 3d/2 \rceil$. The recursion uses Zhang's known values at $(d, k) = (3, 1)$ and $(4, 1)$.

Work in orthonormal prediction coordinates $U = X(X^\top X)^{-1/2}$. Write u_i for its feature vectors, $R = XW_\star - Y$, and $g_i = \mathrm{vec}(u_i R_i^\top)$. Thus $\sum_i g_i = 0$. The weighted covariance is $H_a = \sum_i a_i u_i u_i^\top$. When $H_a \succ 0$, exact reproduction is equivalent to $\sum_i a_i g_i = 0$. We measure conditioning in these fixed coordinates.

4.1 Optimal weighting of an exact support

Suppose a support S already has positive exact weights and full selected rank. Let $\kappa_\star(S)$ be the smallest condition number over all nonnegative exact weightings on S. Scaling all weights leaves the fit and condition number unchanged, giving the convex formulation

$$t_\star = \max_{a \geq 0, t} \{t : tI \preceq H_a \preceq I, \quad G_S^\top a = 0\}, \qquad \kappa_\star(S) = 1/t_\star. \tag{7}$$

This normalization is classical in condition-number minimization [23, 24]. Here the signed residual equations restrict the weights to those preserving the observed fit.

Theorem 4.1 (Refinement without additional rows)**.** *The dual of* (7) *is*

$$D_\star = \min \left\{ \mathrm{tr}\, Q : \begin{array}{l} P, Q \succeq 0, \quad \mathrm{tr}\, P = 1, \\ u_i^\top (Q - P) u_i - g_i^\top \lambda \geq 0 \quad (i \in S) \end{array} \right\}, \tag{8}$$

with $D_\star = t_\star$. For any feasible primal value $t > 0$ and dual value D,

$$\frac{1}{D} \leq \kappa_\star(S) \leq \frac{1}{t}, \qquad \frac{\kappa(H_a)}{\kappa_\star(S)} \leq \frac{D}{t}. \tag{9}$$

The algorithm retains the best feasible weights seen, including the starting weights, so it cannot add rows or worsen the initial condition number. It solves linear programming relaxations and adds Rayleigh quotients from violated eigendirections. Appendix D derives the dual and explains how numerical solutions are repaired before the interval is reported.

For the half-squared weighted loss, gradient descent with step $2/(\lambda_{\max} + \lambda_{\min})$ contracts prediction-coordinate error by at most $(\kappa - 1)/(\kappa + 1)$. A smaller measured condition number improves this contraction bound; the ratio D/t instead measures suboptimality among exact weights on S. Section 5.1 treats changes to the responses.

4.2 Constructing an exact subset with a covariance bound

Choose a fixed nonnegative weighting b with $C = \sum_i b_i u_i u_i^\top \succ 0$. Put $v_i = C^{-1/2} u_i$, $h = -G^\top b$, and $\rho = \mathrm{rank}\, G$. An added nonnegative weighting z satisfying $G^\top z = h$ makes $a = b + z$ exact, while $C \succ 0$

Algorithm 1 Construction followed by reweighting on the resulting support

Require: Full fit, features, reference rows, iteration budgets

Solve $G^\top z = -G^\top b$, $z \geq 0$, minimizing mass or covariance inflation
Set $a = b + z$; verify exact residual cancellation and selected rank
Fix $S = \operatorname{supp}(a)$; normalize a so that $\lambda_{\max}(H_a) = 1$
for each refinement iteration **do**
    Solve the LP relaxation of (7) on S
    Rescale any positive-definite candidate and retain the best feasible weights
    Form and repair the dual (8); retain the smallest upper bound D
    if $D/t \leq 1 + \varepsilon$ **then**
        break
    end if
    Add violated minimum- and maximum-eigenvector Rayleigh tests
end for
Independently refit the returned rows and verify the primal and dual inequalities
return selected observations, weights, and the interval (9)

prevents rank loss. We control the added covariance by solving

$$\tau_\star = \min_{z \geq 0,\, G^\top z = h} \lambda_{\max}\left(\sum_i z_i v_i v_i^\top \right). \tag{10}$$

A feasible value U gives

$$C \preceq H_a \preceq (1 + U)C, \qquad \kappa(C^{-1/2} H_a C^{-1/2}) \leq 1 + U. \tag{11}$$

The dual is

$$\max_{\lambda, Q} \; h^\top \lambda \quad \text{subject to} \quad Q \succeq 0, \quad \operatorname{tr} Q \leq 1, \quad g_i^\top \lambda \leq v_i^\top Q v_i \quad \forall i. \tag{12}$$

Unlike (8), these inequalities range over every original candidate row.

Theorem 4.2 (Support and covariance)**.** *Replace the matrix inequality in* (10) *by q tests*

$$\sum_i z_i \langle P_j, v_i v_i^\top \rangle \leq \eta, \quad P_j \succeq 0, \quad \operatorname{tr} P_j = 1, \quad P_1 = I/d.$$

A basic optimum gives an exact coreset with

$$|\operatorname{supp}(a)| \leq |\operatorname{supp}(b)| + \rho + q - 1. \tag{13}$$

Its actual largest increment eigenvalue U and any feasible dual value L satisfy $L \leq \tau_\star \leq U$, and (11) *holds. If $h = 0$, take $z = 0$ instead.*

For $h \neq 0$, the positive LP variable η accounts for the minus one in (13). With scalar responses and a d-row reference basis, the first relaxation uses at most $2d$ rows, recovering the known threshold [1]. Each additional matrix test raises this support bound by at most one; for m responses use $\rho \leq dm$. The first relaxation minimizes added trace, whereas later tests constrain its distribution across directions.

Cutting planes and column generation are established methods [18, 19]. Here they are subject to the residual equations required for exact reproduction. The construction gives the support bound (13); subsequent reweighting gives the interval (9). Minimizing trace or covariance inflation need not minimize the condition number.

5 Preserving the regularization path

For ridge regression, use positive full-data weights p_i with $\sum_i p_i = 1$ and a response matrix $Y \in \mathbb{R}^{N\times m}$. From this point, unless discussing the unregularized minimax problem, set

$$H = \sum_i p_i x_i x_i^\top \succ 0, \qquad B = \sum_i p_i x_i Y_i^\top, \qquad W(\lambda) = (H + \lambda I)^{-1} B.$$

Replacing p by $a \geq 0$ gives H_a, B_a, and the selected solution $W_a(\lambda)$. Its objective is

$$\frac{1}{2}\sum_i a_i \|W^\top x_i - Y_i\|^2 + \frac{\lambda}{2}\|W\|_F^2.$$

All coordinates, including an intercept when present, are penalized. Unlike unregularized least squares, ridge regression changes when all data weights are rescaled. We distinguish unrestricted nonnegative weights from normalized weights satisfying $\sum_i a_i = 1$.

Define the block Krylov subspace

$$\mathcal{K}(H, B) = \operatorname{span}\{\operatorname{range} B, \operatorname{range}(HB), \ldots, \operatorname{range}(H^{d-1}B)\}, \qquad \kappa = \dim \mathcal{K}(H, B).$$

For a scalar response, b is *cyclic* when $\kappa = d$.

Theorem 5.1 (Path equality)**.** *A fixed nonnegative weight vector a satisfies $W_a(\lambda) = W(\lambda)$ for every $\lambda > 0$ if and only if*

$$B_a = B, \qquad (H_a - H)v = 0 \quad \text{for every } v \in \mathcal{K}(H, B).$$

Equality on a nonempty open interval also suffices. When $\kappa = d$, these conditions reduce to $H_a = H$ and $B_a = B$.

Proof. For large λ, expand the resolvent:

$$(H + \lambda I)^{-1} B = \sum_{j\geq 0} (-1)^j \lambda^{-j-1} H^j B.$$

Equality gives $B_a = B$ and $H_a^j B = H^j B$. Consequently $H_a H^j B = H^{j+1} B$, so $H_a - H$ vanishes on $\mathcal{K}$. Conversely, Cayley–Hamilton makes $\mathcal{K}$ invariant under H; hence the columns of $W(\lambda)$ lie in $\mathcal{K}$ and solve $(H_a + \lambda I)W = B_a$. The selected solution is unique. Equality on an interval extends by the identity theorem for rational functions. □

The condition permits exceptions to full covariance preservation. At uniform weights, the three examples

$$((1, 0), 1), \quad ((0, 1), 0), \quad ((0, 2), 0)$$

have $H = \operatorname{diag}(1/3, 5/3)$ and $b = (1/3, 0)^\top$. Weights $(1/3, 2/3, 0)$ change the second covariance entry but preserve $w(\lambda) = (1/(1 + 3\lambda), 0)^\top$. The path only explores a one-dimensional subspace.

Proposition 5.2 (Support bounds)**.** *Put $D_{d,m} = d(d + 1)/2 + dm$. The moment construction of Maalouf et al. [5], Bayer and Teichmann [6] gives a coreset preserving the ridge path on at most $\min(N, D_{d,m})$ rows with unrestricted nonnegative weights, or $\min(N, D_{d,m} + 1)$ with normalized weights. These two full-dimensional bounds are sharp. For $\kappa \geq 1$, the corresponding upper bounds improve to*

$$d\kappa - \frac{\kappa(\kappa - 1)}{2} + dm \quad \text{and one larger}, \tag{14}$$

or to the rank of the corresponding moment matrix, appending a mass coordinate when weights are normalized.

In a basis $(Q, Q_\perp)$ adapted to $\mathcal{K}$, it suffices to preserve $Q^\top HQ$, $Q_\perp^\top HQ$, and B. Their independent coordinates give (14); conic Carathéodory gives the upper bound. Sharpness requires more than counting coordinates: Appendix E constructs independent moment vectors from actual regression examples whose responses are cyclic.

Corollary 5.3 (Scalar responses and intercepts)**.** *The scalar count is $d(d+3)/2$, or one larger under normalized weights. With a mandatory intercept, the sharp count is $D_{d,m}$ under either weight convention, since covariance preservation already fixes total mass. If $B = 0$, the path is zero: an empty data term suffices with unrestricted weights, while normalized cancellation admits a bound of $dm + 1$ rows.*

Theorem 5.4 (Generic minimum support)**.** *Fix $N \geq d$, $m \geq 1$, and positive full-data weights. Give the free entries of $(X, Y) \in \mathbb{R}^{N\times d} \times \mathbb{R}^{N\times m}$ their usual Euclidean measure, restricted to the open set* $\operatorname{rank} X = d$*. Outside a Lebesgue-null algebraic subset, the minimum support preserving the ridge path is*

$$\min(N, D_{d,m}) \quad \textit{with unrestricted weights}, \qquad \min(N, D_{d,m}+1) \quad \textit{with normalized weights}.$$

With a mandatory intercept, both minima are $\min(N, D_{d,m})$*; here the intercept column is fixed and genericity is over the remaining entries of X and Y.*

The proof combines generic cyclicity with the fact that a positive target moment generically lies outside the span of any smaller set of moment vectors. Both algebraic arguments, including the intercept construction, are in Appendix E. Thus when $d < N \leq D_{d,m}$, generic data admit no proper subset preserving the path, although one scalar fit may need only $d + 1$ examples under row general position and nonzero residuals.

These are lower bounds on retained original examples, not storage in bits. Directly storing H and B uses $D_{d,m}$ scalars. Matching them need not match the objective's constant $\sum_i p_i \|Y_i\|^2$, so path equality alone does not imply equality of objective values or marginal likelihoods.

Proposition 5.5 (Classical interpolation on the path)**.** *If $\kappa \geq 1$, exact agreement at $\kappa + 1$ distinct positive regularization values, with the same selected weights, implies agreement along the entire ridge path.*

Differences of the full solutions at these values span $\mathcal{K}$ through a nonsingular Cauchy matrix. Subtracting their normal equations then gives Theorem 5.1. Appendix E.5 proves this claim and bounds propagation of numerical errors.

A matrix response allows each column to use its own regularization parameter. A fixed positive-definite penalty R is handled by transforming x_i to $R^{-1/2}x_i$. Finite feature maps are also allowed, with d equal to their feature dimension.

Corollary 5.6 (Gradient dynamics)**.** *The path condition in Theorem 5.1 is also necessary and sufficient for equality of the full and selected gradient flows*

$$\dot{W} = B - HW, \qquad \dot{W}_a = B_a - H_a W_a, \qquad W(0) = W_a(0) = 0.$$

Under that condition, gradient descent from zero agrees at every iteration for any common sequence of scalar step sizes. The same is true with a common scalar momentum schedule initialized at zero.

Proof. The derivatives of the flow at zero are $B, -HB, H^2B, \ldots$; equality gives precisely the Krylov condition. Conversely, all iterates and flow directions remain in $\mathcal{K}(H, B)$, on which the selected and full updates agree. Induction gives the discrete statements. □

Thus the same conditions preserve both the ridge path and optimization from zero. The support lower bounds do not require equality of the full quadratic objective.

5.1 Label changes and residual cancellation

We return here to unweighted least squares.

Remark 5.7 (Relabelling). For full-rank X, a fixed support reproduces the full least-squares fit for every $y \in \mathbb{R}^N$ only if it contains every row with $x_i \neq 0$. This remains necessary when weights may be chosen after seeing all the new labels, provided selected examples retain their original labels.

Proof. If row j is omitted and $x_j \neq 0$, take $y = e_j$. All selected labels vanish, so the selected minimum-norm solution is zero for every nonnegative weighting. The full solution is $(X^\top X)^{-1} x_j \neq 0$. Conversely, nonzero-feature rows with their full weights reproduce the fit; zero-feature rows contribute only constants. □

With fixed weights, agreement on an open set of labels implies agreement of the linear coefficient maps.

Sensitivity with fixed weights. Let $U \in \mathbb{R}^{N\times d}$ satisfy $U^\top U = I$ and span the full prediction space. For a selected support of size s, let $Z = U_S$ have full column rank, $A = \operatorname{diag}(a_S)$, and

$$T = (Z^\top A Z)^{-1} Z^\top A, \qquad J = T E_S - U^\top,$$

where E_S restricts labels to S. The matrix J compares selected and full coefficient maps in orthonormal prediction coordinates. Both support and weights are fixed before any label perturbation.

Proposition 5.8 (Sensitivity from the selected design)**.** *One has*

$$JJ^\top = TT^\top - I \succeq (Z^\top Z)^{-1} - I. \tag{15}$$

For a zero-mean perturbation of the observed responses with covariance $\sigma^2 I$, an initially exact coreset has expected squared change $\mathbb{E}\|J\xi\|^2 = \sigma^2 \operatorname{tr}(JJ^\top)$ in the fitted values on the full design. This is not a held-out prediction error. If $h_{\max} = \max_i \|u_i\|^2$, then

$$\operatorname{tr}(JJ^\top) \geq \frac{d^2}{s h_{\max}} - d. \tag{16}$$

Equal row leverage d/N gives $d(N/s - 1)$ and $\|J\|_{\mathrm{op}}^2 \geq N/s - 1$.

This is the Gauss–Markov projection identity [16]; the trace bound follows from Cauchy–Schwarz applied to the covariance eigenvalues.

Exact fitting also requires the selected estimator to annihilate the observed residuals.

Theorem 5.9 (Sensitivity under residual cancellation)**.** *Suppose $TZ = I$ and $TR_S = 0$, where R_S contains selected residual columns. Set*

$$P_R = I - R_S R_S^+, \qquad Q = Z^\top P_R Z, \qquad T_\sharp = Q^{-1} Z^\top P_R.$$

Then $Q \succ 0$ and

$$TT^\top = Q^{-1} + (T - T_\sharp)(T - T_\sharp)^\top. \tag{17}$$

Hence $\operatorname{tr}(JJ^\top) \geq \operatorname{tr}(Q^{-1}) - d$. If $s = d + \operatorname{rank} R_S$, equality holds and the two constraints determine T uniquely.

Proof. If $Qv = 0$, then $Zv \in \operatorname{range} R_S$; applying T gives $v = 0$. The matrix $T_\sharp$ satisfies both constraints and $T_\sharp T_\sharp^\top = Q^{-1}$. For $C = T - T_\sharp$, we have $CZ = 0$, $CR_S = 0$, and $CP_R = C$, so $CT_\sharp^\top = 0$. Expansion gives (17). Also $\operatorname{range} Z \cap \operatorname{range} R_S = \{0\}$. When their dimensions sum to s, they span $\mathbb{R}^s$ and force $C = 0$. □

For a nonzero scalar residual $q = r_S$, put $H_S = Z^\top Z$, $h = Z^\top q$, and $v = q^\top q$. Sherman–Morrison gives

$$\operatorname{tr}(JJ^\top) \geq \underbrace{\operatorname{tr}(H_S^{-1}) - d}_{\text{discarding observations}} + \underbrace{\frac{\|H_S^{-1}h\|^2}{v - h^\top H_S^{-1} h}}_{\text{cancelling the residual}} . \tag{18}$$

The denominator is positive when an exact full-rank coreset exists. Every such scalar coreset on $d + 1$ rows attains equality. For larger supports, $T_\sharp$ is optimal among linear estimators satisfying the constraints, but need not arise from diagonal nonnegative weights. When $q = 0$, Proposition 5.8 applies without the additional term.

Label derivatives depend on the observation-to-coefficient map, not only on its Hessian [11].

6 Experiments

We test the rational calculation for supports of size $d + 1$, reweight exact supports for better conditioning, and measure preservation under reuse. The empirical study has 16 configurations derived from diamonds, RAND insurance, and Texas housing [20–22]. Configurations differ by source, feature map, or split and are not independent datasets. They include linear features, random Fourier features up to dimension 257, and frozen features from five neural networks trained on the training partitions. The largest fit uses 43,152 training observations. Texas housing uses a chronological split. Appendix F gives the protocol and configuration-level results.

Rational instances. Nine rational datasets test the $d + 1$ calculation. Block constructions recover ratios $5/3$ at $(d, n) = (3, 4)$, 2 at $(4, 5)$, and $5/2$ for a two-block instance at $(6, 7)$. Three overlapping constructions give $6/5$, $59/52$, and $413/342$, all below the harmonic lower bound. Applying the recursion through $d = 100$ improves the previous upper bound for 96 parameter pairs beyond the known cases $(d, k) = (3, 1)$ and $(4, 1)$.

Condition refinement on the same support. Each configuration starts from the support obtained by minimizing the added weight subject to the residual equations. Reweighting changes only this support and may remove rows. All 16 condition numbers improve, by a pooled median factor of 6.46, without adding rows. The median factors within diamonds, RAND, and Texas housing are 7.36, 7.29, and 3.83, respectively. The median certified factor relative to the best weighting on the initial support is 1.19; the 257-dimensional configuration retains a factor of 31.91 under its iteration limit. Table 2 and Figure 2 summarize all configurations.

The median number of gradient descent iterations predicted by the spectral recurrence falls by a factor of 6.35. Label sensitivity, however, increases by a median factor of 2.19 and improves in only three configurations. Separate refitting in the original feature coordinates gives a maximum held-out discrepancy of 1.03×10^{-10} from the full fit.

Covariance control during construction. A separate comparison starts every method from the same $2d$ rows. Minimizing the largest eigenvalue of the added covariance improves that objective over minimizing either total added weight or added trace in all 16 configurations. At matched row budgets, however, its median condition number is 1.05 times that of the minimum-weight construction, and it is smaller in only seven configurations. Direct reweighting of a completed exact support is therefore more reliable when condition number is the objective. Only two covariance optimizations reach a 5% primal–dual gap within their cut limits.

Source and features	Tasks	d	Before	After	Factor	Rows	Seconds
Diamonds, linear	3	24	910.2	5.0	1.05	75	5.2
Diamonds, Fourier 64	3	65	65.7	12.8	1.52	166	41.3
Diamonds, Fourier 128	1	129	332.2	92.9	6.34	216	31.9
Diamonds, Fourier 256	1	257	591.0	312.6	31.91	380	89.6
Diamonds, neural 64	2	65	206.9	17.6	1.63	154	46.0
RAND, linear	2	10	29.5	4.5	1.05	28	0.2
RAND, neural 32	2	33	54.8	6.5	1.08	90	11.5
Texas, linear	1	52	32.4	6.4	1.29	128	24.4
Texas, neural 32	1	33	12.9	5.0	1.08	90	8.3

Table 2: Reweighting results, with medians within each feature view. "Factor" is D/t, which bounds the returned condition number relative to the best exact weighting on the initial support. Rows count positive final weights. Configuration-level results appear in Table 4.

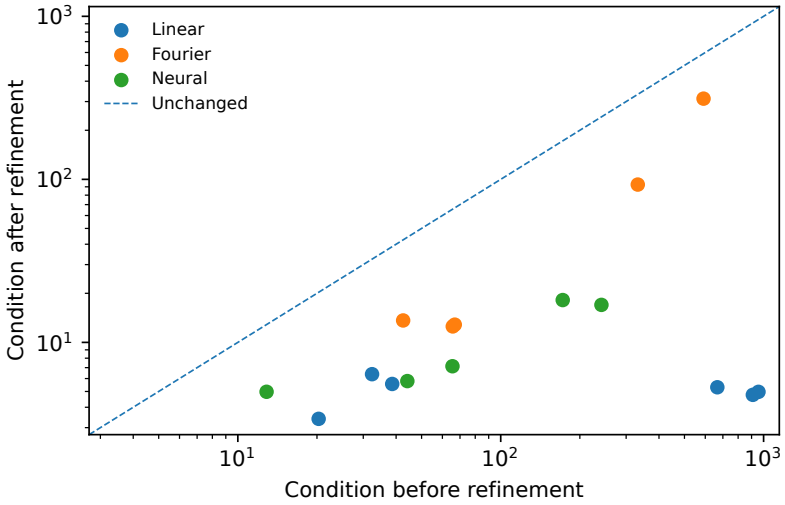


Figure 2: Condition numbers before and after reweighting each exact support. Both axes use orthonormal prediction coordinates; the diagonal denotes no change.

Reuse of the compressed data. Across 21 ridge configurations, weights satisfying Theorem 5.1 have maximum held-out discrepancy 4.75×10^{-12} on the prescribed grid. Weights fitted only at $\lambda = 0.1$ have median maximum path discrepancy 1.97. Across nine configurations with two response columns, preserving only the first response gives median discrepancy 0.682 on the second; preserving both reduces the maximum discrepancy to 4.61×10^{-12}. These values are scaled prediction discrepancies.

Verification. A separate implementation reconstructs the reported fits and covariance spectra from the saved observations and weights, and checks support membership and every dual inequality. The supplement contains the configuration-level results and finite optimization gaps.

7 Limitations

The scalar minimax risk remains unknown for much of the range below $\lceil 3d/2 \rceil$; the recursive theorem supplies bounds rather than the conjectured value. The conditioning optimum in Theorem 4.1 ranges over weights on a fixed support, not over all supports, and finite cutting-plane runs can leave a wide optimality interval. The experiments use three tabular sources and frozen feature maps, so they do not address continued representation training or broad pretrained models. Numerical certificates are checked to the stated tolerances; only the rational calculations use exact arithmetic.

References

[1] S. Hanneke, S. Moran, A. Shlimovich, and A. Yehudayoff. Data Selection for ERMs. *COLT*, PMLR 291:2634–2665, 2025.

[2] S. Hanneke, S. Moran, A. Shlimovich, and A. Yehudayoff. Open Problem: Data Selection for Regression Tasks. *COLT*, PMLR 291:6225–6229, 2025.

[3] G. Zhang. Exact Risk Ratios for Weighted Data Selection in Linear Regression. arXiv:2608.28007, version 1, 28 August 2026. Preprint.

[4] G. Zhang. Exact Recovery Thresholds for Weighted Data Selection in Vector-Valued Linear Regression. arXiv:2608.30254, version 1, 31 August 2026. Preprint.

[5] A. Maalouf, I. Jubran, and D. Feldman. Fast and Accurate Least-Mean-Squares Solvers. *NeurIPS*, 2019.

[6] C. Bayer and J. Teichmann. The proof of Tchakaloff's theorem. *Proceedings of the AMS*, 134(10):3035–3040, 2006.

[7] Y. Wang and M. Pilanci. Sketching the Krylov Subspace: Faster Computation of the Entire Ridge Regularization Path. arXiv:2210.12212, 2022.

[8] T. Chen, C. Huber, E. Lin, and H. Zaid. Preconditioning without a preconditioner: faster ridge-regression and Gaussian sampling with randomized block Krylov subspace methods. *Electronic Transactions on Numerical Analysis*, 65:63–92, 2026. doi:10.1553/etna_vol65s63.

[9] B. Shi and J. Phillips. A Deterministic Streaming Sketch for Ridge Regression. *AISTATS*, PMLR 130:586–594, 2021.

[10] S. Wang, A. Gittens, and M. W. Mahoney. Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging. *ICML*, PMLR 70:3608–3616, 2017.

[11] H. Liao, B. A. Pearlmutter, V. K. Potluru, and D. P. Woodruff. Automatic Differentiation of Sketched Regression. *AISTATS*, PMLR 108:4367–4376, 2020.

[12] M. Dereziński and M. K. Warmuth. Reverse Iterative Volume Sampling for Linear Regression. *JMLR*, 19(23):1–39, 2018.

[13] C. Davis. Theory of positive linear dependence. *American Journal of Mathematics*, 76(4):733–746, 1954.

[14] R. L. McKinney. Positive bases for linear spaces. *Transactions of the AMS*, 103:131–148, 1962.

[15] D. Dadush, L. A. Végh, and G. Zambelli. Rescaling Algorithms for Linear Conic Feasibility. *Mathematics of Operations Research*, 45(2):732–754, 2020.

[16] A. C. Aitken. On Least Squares and Linear Combination of Observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936. doi:10.1017/S0370164600014346.

[17] C. Giannelli, S. Imperatore, L. M. Kreusser, E. Loayza-Romero, F. Mohammadi, and N. Villamizar. A general formulation of reweighted least squares fitting. *Mathematics and Computers in Simulation*, 225:52–65, 2024. doi:10.1016/j.matcom.2024.04.029.

[18] J. E. Kelley, Jr. The Cutting-Plane Method for Solving Convex Programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960. doi:10.1137/0108053.

[19] S. Ahipasaoglu, S. Cipolla, and J. Gondzio. A column generation approach to exact experimental design. arXiv:2507.03210, version 2, 2026.

[20] ggplot2 developers. Prices of over 50,000 round cut diamonds. https://ggplot2.tidyverse.org/reference/diamonds.html. Dataset documentation accessed 6 September 2026.

[21] ggplot2 developers. Housing sales in Texas. https://ggplot2.tidyverse.org/reference/txhousing.html. Source: Texas A&M Real Estate Center. Dataset documentation accessed 6 September 2026.

[22] Statsmodels developers. RAND Health Insurance Experiment Data. https://www.statsmodels.org/stable/datasets/generated/randhie.html. Dataset documentation accessed 6 September 2026.

[23] Z. Lu and T. K. Pong. Minimizing Condition Number via Convex Programming. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1193–1211, 2011. doi:10.1137/100795097.

[24] Z. Qu, W. Gao, O. Hinder, Y. Ye, and Z. Zhou. Optimal Diagonal Preconditioning. arXiv:2209.00809, 2022.

A Proofs for attainable coefficients

Proof of Theorem 3.1. Work in V with its inherited Euclidean metric. Positive weights make the weighted objective strictly convex on V; its unique minimizer there is the minimum-norm minimizer in the original space.

Let w be such a minimizer and F the face containing it in relative interior. Set $I_0 = \{i : x_i^\top w = y_i\}$ and $s_i = \operatorname{sign}(x_i^\top w - y_i)$ for $i \notin I_0$. A recession direction v of $\overline{F}$ satisfies

$$x_i^\top v = 0 \ (i \in I_0), \qquad s_i x_i^\top v \geq 0 \ (i \notin I_0).$$

Taking the inner product of the selected normal equation with v gives $\sum_{i \notin I_0} a_i |x_i^\top w - y_i| s_i x_i^\top v = 0$. All summands are nonnegative and all displayed coefficients are positive. Therefore $x_i^\top v = 0$ for every i, so $v = 0$ since the rows span V. The face is bounded. A finite union of bounded closed faces is closed, which proves one inclusion even after taking the closure of $\mathcal{W}_S$.

Conversely, take w in relative interior of a bounded face. Let $Z = \operatorname{span}\{x_i : i \in I_0\}$ and let P project within V onto $Z^\perp$. Boundedness says

$$\{v \in Z^\perp : s_i (P x_i)^\top v \geq 0 \ \forall i \notin I_0\} = \{0\}.$$

The theorem of alternatives implies that the signed projected normals positively span $Z^\perp$. They therefore admit a dependence with strictly positive coefficients c_i: one may sum, for each normal, a nonnegative representation of its negative. Set $a_i = c_i / |x_i^\top w - y_i|$ for $i \notin I_0$. Their projected gradient at w is zero, and their covariance is positive definite on $Z^\perp$. Give every equality row weight $t > 0$. Its residual at w is zero, and their covariance is positive definite on Z. The block inverse calculation of Lemma B.7 shows that the combined minimizer tends to w as $t \to \infty$. Full rank on V holds for every finite t.

When $Z = \{0\}$ the constructed weights already return w. When $Z = V$, all directions are fixed by equality rows and the same limit applies without a quotient. Points on face boundaries follow by closure; equivalently, use the bounded subface containing each point in relative interior. This proves equality. □

A.1 Certificates for supports of size $d + 1$

The determinant barycenter in Giannelli et al. [17] has coefficients

$$p_B(a) = \frac{\det(X_B)^2 \prod_{i \in B} a_i}{\det(X_S^\top \operatorname{diag}(a) X_S)}.$$

If S has $d + 1$ rows, every basis omits one row and the common product $\prod_{i \in S} a_i$ cancels. Thus $p_{S \setminus j} \propto \Delta_j / a_j$. Choosing $a_j = \Delta_j / \pi_j$ proves Theorem 3.2. Indices with $\Delta_j = 0$ belong to every valid basis; their positive weights cancel from all barycentric ratios. Boundary vectors are limits of strictly positive barycentric vectors.

For rational data, let E_S have columns $w_B - w_\star$ and set $Q = E_S^\top H E_S$. A rational vector π certifies a projection precisely when

$$\pi \geq 0, \quad \mathbf{1}^\top \pi = 1, \quad Q\pi \geq (\pi^\top Q \pi)\mathbf{1}. \tag{19}$$

These are the simplex KKT conditions for a positive-semidefinite quadratic. A rational optimal solution exists on an affinely independent active set. The implementation uses floating arithmetic only to propose that set, then solves and checks (19) over the rationals. The verifier reconstructs every nonsingular basis and checks that every rank-d support of size $d + 1$ occurs in the certificate list. A d-row support is included as a hull vertex of an extension.

For the original minimum-norm learner, singular supports must also be covered. Their outputs lie in proper row spans. Every such span is contained in the span of an independent $(d - 1)$-row subset of the full design. If B contains a basis for one of these larger spans, then

$$\min_{w \in \operatorname{range} B} \|w - w_\star\|_H^2 = \left\|w_\star - B(B^\top H B)^{-1} B^\top H w_\star\right\|_H^2 \tag{20}$$

is an exact lower bound for every selected solution in that span. The translated fixtures verify that each of these lower bounds exceeds the best full-rank projection. This is a finite check on the actual instance, not an assumption that a large translation is sufficient.

For other budgets, Theorem 3.1 gives a finite, generally exponential procedure: enumerate supports and their bounded faces in the original row span, then project onto each face. Rational polyhedral quadratic programs have rational primal and dual certificates. We implement the simpler $d + 1$ case, not a general arrangement enumerator.

A.2 Why the full interpolation hull is not enough

For the four-row example in Section 3, the residual at $w = (11/6, 1/2)^\top$ is

$$(5/3, 37/6, -13/2, 11)^\top.$$

The directional derivatives of the residuals along $v = (1, 2)^\top$ are $(2, 0, -7, 9)^\top$. Their products are nonnegative and three are strictly positive. Thus no strictly positive weighted normal equation can vanish at w. All residuals are nonzero, so the same obstruction holds in a neighborhood. The midpoint belongs to the convex hull of two interpolants but not to the closure of attainable fits.

B Risk ratios for a single scalar fit

Return to $L(w) = \|Xw - y\|^2$, $w_\star = X^+ y$, $r = Xw_\star - y$, and $E = \|r\|^2$. An admissible dataset is a finite pair $D = (X, y)$ with $X \in \mathbb{R}^{N \times d}$ and arbitrary finite N; matrices of rank below d are allowed. Following [1], define

$$R_d(n) = \sup_D \frac{\inf_{|\operatorname{supp}(a)| \leq n} L(w_a)}{E}. \tag{21}$$

The ratio is one when both losses vanish and infinite when only the denominator vanishes. The quantity $R_{d,N}(n)$ restricts the same supremum to datasets with exactly N rows and $\operatorname{rank} X = d$. Multiplying all selected weights by a positive constant does not change w_a.

B.1 Coordinates and singular selections

Whitening can change the Euclidean minimum-norm choice. For

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad y = (1, 4, 0)^\top,$$

the full fit is $(2, 1)^\top$ with $E = 3$. Selecting the first row gives $(1, 0)^\top$ and ratio 3. Whitening, applying Euclidean minimum-norm ERM, and mapping back gives $(1, -1/2)^\top$ and ratio $25/6$. The same warning is made by Zhang [3].

Lemma B.1 (Excluding singular selections)**.** *Fix full-rank X, $r \perp \operatorname{range} X$, and $n \geq d$. Some vector u satisfies, for labels $y^{(\alpha)} = \alpha X u - r$ and full loss L_α,*

$$\sup_{\alpha > 0} \inf_{|\operatorname{supp}(a)| \leq n} [L_\alpha(w_a) - E] = \inf_{\substack{|\operatorname{supp}(a)| \leq n \\ M \succ 0}} g^\top M^{-1} H M^{-1} g,$$

where $H = X^\top X$, $M = X^\top \operatorname{diag}(a) X$, and $g = X^\top \operatorname{diag}(a) r$.

Proof. Choose u outside the finite union of proper spans of feature subsets. Its minimum distance δ to them is positive. Every singular weighted solution belongs to one such span and incurs excess at least $\lambda_{\min}(H)\alpha^2\delta^2$. Full-rank solutions have displacement $-M^{-1}g$, independent of α, and some d-row basis has finite cost. Because there are finitely many supports, a finite α_0 makes every singular solution worse than that basis for $\alpha \geq \alpha_0$. The two infima then agree. □

Thus minimax analysis may restrict to weights whose covariance has full rank, for which the change $v = H^{1/2} w$ gives

$$X^\top X = I, \quad X^\top r = 0, \qquad L(w_a)/E = 1 + \|M^{-1} g\|^2 / E. \tag{22}$$

A rank-deficient full design is first restricted isometrically to its row span. All upper constructions below have full selected rank at finite weights.

B.2 Statements

Write $n = d + k$. For $1 \leq h \leq d$, define

$$\Gamma(d, h) = \min_{\substack{\delta_1 + \cdots + \delta_h = d \\ \delta_j \in \mathbb{N}}} \sum_j \delta_j^{-1} = \frac{h - s}{q} + \frac{s}{q + 1}, \qquad d = qh + s, \quad 0 \leq s < h, \tag{23}$$

and, for $0 \leq k < d$,

$$\Lambda(d, k) = \max_{h = k+1, \ldots, d} \frac{h - k}{\Gamma(d, h)}. \tag{24}$$

Orthogonal simplex blocks give $R_d(d + k) \geq 1 + \Lambda(d, k)$, a bound also reported in [3].

Theorem B.2 (Risk for larger budgets)**.** *For $\lceil 3d/2 \rceil \leq n \leq 2d$,*

$$R_d(n) = 3 - n/d. \tag{25}$$

The upper bound holds for every dataset; equality takes the supremum over datasets.

Theorem B.3 (Deleting one example)**.** *For* $1 \le h \le d$*,*

$$R_{d,d+h}(d+h-1) = 1 + \Gamma(d,h)^{-1}. \tag{26}$$

For $N > 2d$*, the value at budget* $N-1$ *is one.*

Fixed-size and unrestricted values differ: at $(d,n) = (8,10)$, size $N = 11$ gives $13/7$, but the unrestricted lower bound is 2. The conjecture $R_d(d+k) = 1 + \Lambda(d,k)$ remains open in general for $k < d/2$; the cells $(3,4)$ and $(4,5)$ have separate proofs in [3].

Corollary B.4 (Budget for a prescribed ratio)**.** *For* $0 \le \varepsilon \le 1/2$*,*

$$\min\{n : R_d(n) \le 1+\varepsilon\} = \lceil (2-\varepsilon)d \rceil. \tag{27}$$

This concerns infimum values. Finite positive weights achieve $1+\varepsilon+\eta$ *for every* $\eta > 0$*. In particular,* $R_d(2d-1) = 1 + 1/d$*.*

Proof. Theorem B.2 supplies the upper bound. The one-dimensional simplex blocks in the lower construction give $R_d(n) \ge 3 - n/d$ for $d \le n \le 2d$; budgets below d have infinite ratio. These exclude every smaller integer budget. Finite-weight approximation follows from (28). □

B.3 Boundary cases and lower bounds

Lemma B.5. *Some* d*-row basis* S *satisfies* $\|X_S^{-1} r_S\|^2 \le dE$*.*

This is the size-d volume-sampling loss bound of Dereziński and Warmuth [12], used by Hanneke et al. [1].

Lemma B.6. *Every full-rank dataset has an exact coreset with full rank on at most* $d + \operatorname{rank} G \le 2d$ *rows, where* $G_i = r_i x_i$*.*

The $2d$ bound is due to Hanneke et al. [1]. Its rank refinement follows from Theorem 4.2 with a d-row reference basis and one trace constraint.

These give the known boundary values [1, 12, 13]. A regular simplex attains the dE basis excess. For $n < d$, a realizable dataset with a coefficient vector outside every proper selected row span has positive selected loss and zero full loss.

The block calculation and harmonic lower construction are given in [3]. A whitened regular δ-simplex with residual energy E_b has zero excess with all $\delta+1$ vertices and excess δE_b with any δ vertices. Orthogonal blocks with equal charges $\delta_j E_j$ yield (24); Lemma B.1 excludes singular selections. Balanced dimensions minimize the reciprocal sum. These ingredients are cited rather than rederived here.

For $k \ge d/2$, Cauchy–Schwarz gives $\Gamma(d,h) \ge h^2/d$. Since $(h-k)/h^2$ is nondecreasing on $k < h \le d \le 2k$, its largest value occurs at $h = d$. Thus $\Lambda(d,k) = (d-k)/d$, proving the lower bound in Theorem B.2.

B.4 A limiting-weight argument

Lemma B.7 (Subspace restriction)**.** *Let a selected set* T *have positive weights* γ*, zero cross moment* $\sum_{i\in T} \gamma_i r_i x_i = 0$*, and covariance positive definite on* $V = \operatorname{span}\{x_i : i \in T\}$*. Let* P *project onto* $V^\perp$*, and let rows* S' *outside* T *span that quotient after projection. Weighting* T *by* $t\gamma$ *and* S' *by fixed positive weights gives full rank for every* $t > 0$*. As* $t \to \infty$*, its displacement tends to the quotient displacement, with zero component on* V*.*

The projected outside data are whitened and stationary, with labels $\widetilde{y}_i = \langle P w_\star, P x_i \rangle - r_i$ *and energy* $\sum_{i\notin T} r_i^2$*.*

Proof. In $V \oplus V^\perp$, write

$$M_t = \begin{pmatrix} tA + B & C \\ C^\top & D \end{pmatrix}, \quad g_t = \begin{pmatrix} g_V \\ g_\perp \end{pmatrix}, \qquad A \succ 0,\ D \succ 0.$$

A vector killed by both weighted covariances must vanish, proving full rank. The block inverse gives $M_t^{-1} g_t \to (0, D^{-1} g_\perp)$. Because every row in T lies in V, $\sum_{i \notin T} P x_i x_i^\top P = P$ and $\sum_{i \notin T} r_i P x_i = 0$. The specified labels retain these residuals. □

Proposition B.8 (Convergence rate)**.** *With $h = g_V - CD^{-1} g_\perp$ and $q_\infty = D^{-1} g_\perp$,*

$$\|M_t^{-1} g_t - (0, q_\infty)\| \le \frac{\sqrt{1 + \|D^{-1} C^\top\|_{\mathrm{op}}^2}\, \|h\|}{t \lambda_{\min}(A)}. \tag{28}$$

Proof. The first block is $p_t = (tA + B - CD^{-1}C^\top)^{-1} h$; the second is $q_\infty - D^{-1} C^\top p_t$. The Schur complement $B - CD^{-1}C^\top$ is positive semidefinite, so $\|p_t\| \le \|h\| / (t \lambda_{\min} A)$. □

If the right side is δ_t, the squared excess differs from its limit by at most $2\|q_\infty\|\delta_t + \delta_t^2$. Choose finite quotient weights first and then finite t at each induction step. This approaches the bound with full rank, but gives no uniform bound on weight magnitudes.

B.5 Proof for larger budgets

Proof of Theorem B.2. Set $t = 2d - n = d - k \le d/2$. Induct on t, simultaneously over dimensions, to obtain full-rank selections with excess arbitrarily close to at most tE/d. The case $t = 0$ is Lemma B.6. Zero features may be discarded from the construction.

If zero-residual rows span dimension $z > 0$, fix a basis of that span using Lemma B.7. The quotient has dimension $p = d - z$, extra budget k, and defect $t - z$. It is exact if $k \ge p$; otherwise induction applies because $0 < t - z < t$ and $2k \ge p$. Its coefficient $(t - z)/(d - z)$ is at most t/d. A zero-dimensional quotient needs no rows.

Assume now that nonzero features have nonzero residuals. The vectors $v_i = r_i x_i$ sum to zero. Extreme rays of their pointed nonnegative dependence cone are positive circuits, and the all-ones dependence is a sum of such rays. If a circuit has rank $s \ge 2$, fix its $s + 1$ rows. The quotient has dimension $p = d - s$, energy $E' \le E$, extra budget $k - 1$, and defect $t' = t - s + 1$. It is exact when $p = 0$ or $k - 1 \ge p$. Otherwise $0 < t' < t$ and $2(k - 1) - p = 2k - d + s - 2 \ge 0$. Induction gives coefficient

$$\frac{t - s + 1}{d - s} \le \frac{t}{d},$$

because $sk \ge d$. The combined support is at most $(s + 1) + (p + k - 1) = n$.

If every positive circuit has rank one, group the all-ones decomposition by lines. Each line is stationary and contains both orientations. A minimal dependence among distinct lines could be oriented positively by choosing the required sign on each line, creating a circuit of rank at least two. Thus the active lines are independent. There are exactly d, and their rank-one covariance contributions sum to I, which makes them orthogonal.

Let their energies be E_j. Cancel the k largest using two opposite gradients on each line. On each remaining line, some row satisfies $r_i^2 / x_i^2 \le E_j$, by averaging with weights x_i^2 whose sum is one. Keeping that row costs at most E_j. Total support is $2k + t = n$ and excess is at most the sum of the t smallest energies, hence tE/d.

The simplex lower bound proves equality. For full-data rank $r < d$, restrict orthogonally to its row span. Either $n \ge 2r$ gives exactness or the rank-r bound is $3 - n/r \le 3 - n/d$. □

B.6 Proof for deleting one example

Lemma B.9 (Circuit decomposition)**.** *Let nonzero vectors v_i have a positive stationary vector λ and dependence dimension $h \geq 2$. An extreme ray $c \geq 0$ has minimally dependent proper support S. With*

$$\theta = \min_{i:c_i>0} \lambda_i/c_i, \quad \lambda' = \lambda - \theta c, \quad X_0 = \{i : \lambda'_i = 0\},$$

λ' is nonnegative and stationary, and $\varnothing \neq X_0 \subseteq S$.

Proof. The cone is nonzero and pointed. A second dependence within an extreme ray's support permits positive and negative perturbations, contradicting extremality. Its support is therefore minimally dependent and cannot be all indices when $h \geq 2$. Subtracting the maximal feasible multiple exhausts at least one coordinate; coordinates outside S remain positive. □

Lemma B.10. *For $2 \leq h \leq d$,*

$$\Gamma(d,h) = \min_{1\leq\delta\leq d-h+1} \{\delta^{-1} + \Gamma(d-\delta, h-1)\}.$$

Also $\Gamma(d,h)$ is nonincreasing in d, $\Gamma(d,h) \geq h/d$, and for $a, b > 0$, $0 \leq F \leq E$, $\min\{aF, b(E-F)\} \leq E/(a^{-1}+b^{-1})$.

Proof. Separate one partition part for the recurrence; enlarge a part for monotonicity. Each reciprocal is at least $1/d$. The last bound is maximized when the two linear terms agree. □

Proof of Theorem B.3. Balanced simplex blocks with equal charges and a sufficiently large coefficient vector give excess $E/\Gamma(d,h)$ at budget $N-1$. For the upper bound, induct on N, retaining full-rank finite-weight approximations. The case $N = d+1$ follows from Lemma B.5; when $N > 2d$, Lemma B.6 applies.

A zero feature can be omitted. If zero-residual rows are dependent, omit one while retaining spanning and stationarity. Otherwise fix all $z > 0$ such rows. The quotient has dimension $d' = d - z$, fewer points, and the same codimension h. It is exact if $h > d'$; otherwise induction gives excess at most $E/\Gamma(d',h) \leq E/\Gamma(d,h)$. Hence assume all features and residuals are nonzero.

For $h \geq 2$, apply Lemma B.9 to $v_i = r_i x_i$ with $\lambda = \mathbf{1}$. Let the circuit S have rank s and size $s+1$, with exhausted set X_0. Write $E_S = \sum_{i\in S} r_i^2$ and $E_X = \sum_{i\in X_0} r_i^2 \leq E_S$.

First fix the circuit. The outside quotient has dimension $d-s$, codimension $h-1$, and energy $E - E_S$. Deleting one quotient row either gives exactness or, by induction, excess

$$(E - E_S)/\Gamma(d-s, h-1). \tag{29}$$

A zero-dimensional quotient is exact. Total support is at most $N-1$.

Alternatively, fix the remainder $\operatorname{supp}\lambda'$, with span W and $\delta = d - \dim W \leq s$. If $\delta = 0$, it is already exact and omits an exhausted row. Otherwise the projected exhausted rows form a whitened stationary dataset of dimension δ and energy E_X. They span and have a positive dependence, so $|X_0| \geq \delta + 1$. If $|X_0| - 1 \geq 2\delta$, deletion is exact. Otherwise induction bounds its excess by

$$E_X/\Gamma(\delta, |X_0| - \delta) \leq \delta E_X \leq \delta E_S.$$

The deleted row belongs to no retained remainder, so support is again at most $N-1$.

When neither construction is exact, $d - s \geq h - 1$, $1 \leq \delta \leq s$, and $\delta \leq d - h + 1$. Monotonicity and Lemma B.10 give

$$\min\left\{\delta E_S, \frac{E - E_S}{\Gamma(d-\delta, h-1)}\right\} \leq \frac{E}{\delta^{-1} + \Gamma(d-\delta, h-1)} \leq \frac{E}{\Gamma(d,h)}.$$

Choose finite quotient approximations, then combine the two weightings using (28). Full rank and the support bound persist. □

Unit-weight deletion. Sherman–Morrison gives excess $r_i^2 h_i/(1-h_i)^2$ for deleting a row of leverage $h_i < 1$. Averaging with weights $(1-h_i)^2/h_i$ and applying Jensen gives $dE/(N-d)^2$. This calculation makes no claim of optimal weighting.

C The recursive risk bound

We first justify the bound $A_{d,k}$ in (5). Degenerate designs require an inequality rather than the identity used for full-rank minors.

Lemma C.1 (Volume-sampled interpolation constraints)**.** *Let $X^\top X = I$, $X^\top r = 0$, and $E = \|r\|^2$. For $1 \le t \le d$, some independent t-row subset T satisfies*

$$r_T^\top (X_T X_T^\top)^{-1} r_T \le \frac{t}{d-t+1} E.$$

Proof. The weights $D_T = \det(X_T X_T^\top)$ sum to $\binom{d}{t}$. Since $\operatorname{range}(X)$ and $\operatorname{span}\{r\}$ are orthogonal invariant subspaces, the nonzero eigenvalues of $XX^\top + srr^\top$ are 1 with multiplicity d and sE. The sum of all t-row principal minors is therefore

$$\binom{d}{t} + sE\binom{d}{t-1},$$

since its nonzero eigenvalues are d ones and sE. For an independent T, the derivative of its determinant at $s = 0$ is $D_T r_T^\top (X_T X_T^\top)^{-1} r_T$. Singular minors can have a nonnegative, nonzero derivative. Dropping those terms gives an inequality, not an identity. Divide by $\binom{d}{t}$ and take a term no larger than the weighted average. □

Set $t = d - k$ and let u be the minimum-norm solution of $X_T u = r_T$. The full-loss minimizer subject to fitting the selected labels on T is $w_\star - u$, because u lies in the row span of X_T. Its excess is $\|u\|^2$. On $\ker X_T$, the constrained problem is a stationary regression in k dimensions. Select at most $2k$ rows reproducing its optimum [1]; together with T they have size at most $d + k$. Giving T increasingly large weights approaches the constrained optimum. The selected covariance remains full rank. This proves excess at most $A_{d,k}E$ as an infimum.

Proof of Theorem 3.3. Induct on dimension and then on the extra budget. The preceding construction gives $A_{d,k}$, and monotonicity in the row budget gives $u_{d,k-1}$.

For $C_{d,k}$, first remove zero feature rows from the construction. If zero-residual rows span a space of dimension $z > 0$, fix a basis of this space with large weights. Lemma B.7 leaves a stationary quotient in dimension $d - z$, with extra budget k and residual energy at most E. Its excess is bounded by $u_{d-z,k}E$.

Otherwise every remaining feature has nonzero residual. Decompose the positive dependence $\sum_i r_i x_i = 0$ into positive circuits. If any circuit has feature rank $s \ge 2$, fixing its $s + 1$ rows leaves dimension $d - s$ and extra budget $k - 1$, with excess at most $u_{d-s,k-1}E$. If every positive circuit has rank one, the independent-line argument in the proof of Theorem B.2 gives excess at most $(d-k)E/d$. The maximum of these alternatives is $C_{d,k}E$.

At $(d,k) = (3,1), (4,1)$ use the minimax values of Zhang [3], transferred to full-rank selections by Lemma B.1. Every quotient construction has full selected rank at finite weights, and the convergence estimate (28) supplies any prescribed positive slack. Taking infima gives (6). The harmonic lower bound is (24).

For completeness, the recursion is nondecreasing in d at fixed k and nonincreasing in k at fixed d. In dimension, each maximum has an expanding index set, $(d-k)/d$ and $A_{d,k}$ increase, and the minimum of nondecreasing sequences is nondecreasing. The two substituted seeds preserve the adjacent inequalities

directly. In budget, the final minimum enforces monotonicity. A full design of rank $r < d$ can therefore use its rank-r bound with extra budget $d + k - r$, which is at most $u_{d,k}$. This extends the result to the rank-at-most-d definition of R_d. □

The rational calculation through $d = 100$ improves the previous upper bound for 96 parameter pairs beyond the two cited cases. It matches the harmonic lower bound for $k \geq d/2$ and at those cases, but yields no further exact value of the unrestricted ratio.

D Covariance optimization and verification

D.1 Condition refinement on a fixed support

Proof of Theorem 4.1. For any exact positive-definite covariance H_a, rescaling by $1/\lambda_{\max}(H_a)$ makes its eigenvalues lie in $[1/\kappa(H_a), 1]$. Conversely, any feasible $t > 0$ gives $\kappa(H_a) \leq 1/t$. At the optimum the upper eigenvalue can be taken to be one, which proves the reciprocal relation.

The Lagrangian for the maximization problem is

$$t(1 - \operatorname{tr} P) + \operatorname{tr} Q + \sum_{i \in S} a_i \{u_i^\top (P - Q) u_i + g_i^\top \lambda\}.$$

Its supremum over $t \in \mathbb{R}$ and $a \geq 0$ is finite precisely under (8). The positive starting weights, scaled to have largest eigenvalue below one, and a smaller positive t give a strictly feasible primal. Strong duality applies. Zero feature rows can be removed; the remaining weights are bounded on $H_a \preceq I$ by the trace constraint, so the optimum is attained. Weak duality gives (9). Keeping the best rescaled feasible covariance retains the initial reciprocal condition number, and no LP introduces a row outside S. □

For a numerical dual with $P, Q \succeq 0$ and $\operatorname{tr} P = 1$, set

$$\delta = \max_{i \in S : u_i \neq 0} \frac{[u_i^\top (P - Q) u_i + g_i^\top \lambda]_+}{\|u_i\|^2}.$$

Replacing Q by $Q + \delta I$ makes every atom inequality feasible and increases the dual value by $d\delta$. A zero feature has $g_i = 0$ and imposes no constraint. This repair is deterministic; the verifier still checks the repaired matrices and every row in S. The trivial choice $P = Q = I/d$, $\lambda = 0$ supplies the bound $t_\star \leq 1$ when it is stronger than a repaired LP dual. The retained primal and dual may come from different iterations.

The implementation starts from the eigenbasis of the initial covariance and adds violated extreme eigenvectors. LP relaxation values alone are not reported as feasible primal values. Instead, each candidate is evaluated by its full spectrum; positive-definite candidates are normalized and compared with the starting covariance. The certificate is valid after any finite number of cuts. General convex condition-number optimization and diagonal preconditioning precede this application [23, 24]; the residual equations, support restriction, and independent exact-fit replay specify the problem solved here.

D.2 Covariance increments

Duality and Theorem 4.2. Write $D(z) = \sum_i z_i v_i v_i^\top$. The primal is $\min\{\eta : G^\top z = h,\ z \geq 0,\ \eta \geq 0,\ D(z) \preceq \eta I\}$. Its Lagrangian is

$$\eta(1 - \operatorname{tr} Q) + h^\top \lambda + \sum_i z_i \big(v_i^\top Q v_i - g_i^\top \lambda\big), \qquad Q \succeq 0.$$

Minimizing over nonnegative η, z yields (12). A feasible correction exists on the full pool: choose a constant $c > \max_i b_i$ and put $z_i = c - b_i$, using $\sum_i g_i = 0$. These weights are strictly positive, and sufficiently large η

makes the matrix inequality strict. Slater's condition gives strong duality. The optimum is attained: after discarding zero feature rows, the trace inequality bounds every z_i on a bounded objective sublevel set.

Exactness follows from $\sum_i (b_i + z_i) g_i = 0$, and $H_a \succeq C \succ 0$ makes the selected minimizer unique. For a feasible correction, $0 \preceq D(z) \preceq UI$ gives (11) by congruence.

Consider a basic solution of the q-constraint LP. If $h \neq 0$, then $\eta > 0$: the trace constraint would otherwise give $D(z) = 0$, forcing every positively weighted feature and gradient to vanish. There are ρ independent equality constraints and at most q active matrix tests. Including the positive variable η, at most $\rho + q$ variables can be positive at a basic point. Thus z has at most $\rho + q - 1$ positive entries, proving (13). If $h = 0$, use the zero correction.

Any feasible dual satisfies $L \leq \tau_\star$ by weak duality, and the computed correction gives $\tau_\star \leq U$. The multiplicative certificate follows. Best primal and dual bounds may be taken from different iterations. In that case the support count uses the cut count of the retained primal, while the larger lower bound remains valid. □

Termination with exact separation. With exact LP solutions over all candidate rows and $h \neq 0$, the cutting-plane method terminates at any positive relative tolerance. Let $\eta_0 > 0$ be the first trace relaxation value. One has $\eta_0 \leq \tau_\star \leq d\eta_0$, since the trace-optimal added covariance has largest eigenvalue at most its trace. At every iteration $\operatorname{tr} D(z) \leq d\eta \leq d\tau_\star$. If a new eigenvector violates the bound by more than $\varepsilon\eta$, it is separated from every previous test vector by at least $\varepsilon/(2d^2)$, using

$$|v^\top D v - u^\top D u| \leq 2\|D\|_{\mathrm{op}} \|v - u\|.$$

A compact sphere contains only finitely many such separated vectors. The argument proves finite termination for a fixed tolerance but does not give a useful polynomial bound on the number of cuts.

Numerical checks. For each reported bound, a separate implementation recomputes g_i, the weighted covariance, positive semidefiniteness, trace, and the maximum violation $\max_i (g_i^\top \lambda - v_i^\top Q v_i)$. It also refits the weighted rows in the original parameter coordinates. These inequalities use the stated floating-point tolerances; the $d + 1$ calculations use rational arithmetic.

E Proofs for regularization paths

E.1 Generic cyclicity

The standard Krylov determinant argument specifies the exceptional algebraic set. The use of Krylov subspaces for ridge paths is established in [7].

Corollary E.1 (Generic rigidity)**.** *For fixed positive full-data weights and $N \geq d$, a scalar response is cyclic outside a Lebesgue-null algebraic subset of the full-rank (X, y) data space. Thus, generically, preserving one entire ridge path requires preserving H and b.*

Proof. The determinant of $[b, Hb, \ldots, H^{d-1}b]$ is a polynomial in the data entries. It is not the zero polynomial: choose the first d rows as $x_j = \sqrt{j/p_j}\, e_j$, $y_j = 1/\sqrt{p_j j}$ and set any other rows to zero. Then $H = \operatorname{diag}(1, \ldots, d)$ and $b = \mathbf{1}$, giving a nonzero Vandermonde determinant. Its zero set has Lebesgue measure zero. The statement concerns the unrestricted feature model; on a prescribed structured feature family, cyclicity must be checked within that family. □

E.2 Support bounds and sharpness

Sharpness in Proposition 5.2. Choose an orthonormal basis Q of $\mathcal{K}$ and complete it to $(Q, Q_\perp)$. Matching the covariance action on $\mathcal{K}$ requires matching the symmetric block $Q^\top HQ$ and the rectangular block $Q_\perp^\top HQ$. These have $\kappa(\kappa+1)/2 + (d-\kappa)\kappa = d\kappa - \kappa(\kappa-1)/2$ independent coordinates. Append the dm coordinates of B. The target is a positive combination of the moment vectors of the original rows. Conic Carathéodory [5, 6] gives the first rank bound. Appending a coordinate equal to one imposes normalized mass and gives the second. Taking $\kappa = d$ gives $D_{d,m}$.

For sharpness, use the following $D_{d,m}$ original rows, with e_j and f_h the coordinate vectors in feature and response space:

$$(je_j, 0)\ (1 \le j \le d), \qquad (e_i + e_j, 0)\ (i < j), \qquad (e_j, f_h)\ (1 \le j \le d,\ 1 \le h \le m). \tag{30}$$

Their moment vectors $(\mathrm{svec}(x_i x_i^\top), \mathrm{vec}(x_i Y_i^\top))$ are linearly independent. In a dependence, cross-moment coordinates first force all coefficients of the last dm rows to vanish; covariance off-diagonal coordinates then force all pair-row coefficients to vanish; diagonal coordinates force the remaining d coefficients to vanish.

Before uniform normalization, the covariance and cross moments are

$$H_0 = \mathrm{diag}(j^2 + d + m - 2)_{j=1}^{d} + \mathbf{1}\mathbf{1}^\top, \qquad B_0 = \mathbf{1}\mathbf{1}_m^\top.$$

Every diagonal entry of the displayed diagonal matrix is distinct. No eigenvector of H_0 is orthogonal to $\mathbf{1}$: such an eigenvector would be an eigenvector of that diagonal matrix, hence supported on one coordinate, a contradiction. An eigenspace of dimension at least two would contain a vector orthogonal to $\mathbf{1}$, so the eigenvalues are simple. Hence $\mathbf{1}$ is cyclic, and $\mathcal{K} = \mathbb{R}^d$. Positive uniform normalization preserves cyclicity. Theorem 5.1 forces exact moment matching, whose representation is unique because the moment vectors are independent. All $D_{d,m}$ original positive weights are therefore necessary.

For normalized sharpness, append one row $(0, 0)$ and put positive uniform mass on all $D_{d,m} + 1$ rows. Moment matching uniquely fixes the positive masses on the original $D_{d,m}$ rows, while the mass equation forces a positive coefficient on the zero row. Every row is necessary. This last example explains why a mandatory intercept changes the normalized count. □

E.3 The intercept constraint

Put $p = d - 1$ and write features as $x = (1, z)$. For $p \ge 1$, choose $L > 0$ and take the zero-response rows at

$$z = 0, \quad z = \pm Lje_j\ (1 \le j \le p), \quad z = L(ie_i + je_j)\ (i < j).$$

There are $d(d+1)/2$ such rows. Add the dm rows with $z \in \{0, Le_1, 2Le_2, \ldots, pLe_p\}$ and response $Y = f_h$, one copy for each $1 \le h \le m$. Every row has intercept one and all pairs (x, Y) are distinct.

The zero-response covariance rows are linearly independent. Indeed, a quadratic polynomial vanishing at these z values has zero constant coefficient by evaluation at zero; evaluation at each opposite pair kills its corresponding linear and diagonal coefficients; evaluation at each pair-sum then kills the off-diagonal coefficient. The added response rows make the entire $D_{d,m}$-row moment matrix independent: their cross-moment coordinates first force their coefficients to vanish in any dependence, after which the covariance argument applies.

It remains to show cyclicity for some L, since independence alone does not force path preservation to match all moments. Let $v = (1, \ldots, p)^\top$, $\beta = p + m - 1$, and

$$D = D_{d,m}, \quad K = (p+m)\,\mathrm{diag}(1^2, \ldots, p^2) + vv^\top.$$

Before uniform normalization the moments are

$$H_0 = \begin{pmatrix} D & \beta L v^\top \\ \beta L v & L^2 K \end{pmatrix}, \qquad B_0 = \begin{pmatrix} d \\ Lv \end{pmatrix} \mathbf{1}_m^\top.$$

Suppose an eigenvector (a, z) of H_0 is orthogonal to $b_0 = (d, Lv)^\top$. Its top equation gives

$$\lambda a = (D - \beta d)a = ca, \qquad c = \frac{(p+1)(4-p)}{2}.$$

If $a = 0$, then $v^\top z = 0$ and the bottom equation becomes a diagonal eigenvalue equation with distinct entries; z must be a coordinate vector, contradicting $v^\top z = 0$. Thus an orthogonal eigenvector would require the eigenvalue c.

If $c \le 0$, this is impossible because $H_0 \succ 0$. If $c > 0$, choose L sufficiently large. The bottom block of $H_0 - cI$ is positive definite and its Schur complement tends to

$$D - c - \beta^2 v^\top K^{-1} v = \beta \left(d - \frac{\beta p}{2p+m} \right) > 0,$$

where $v^\top K^{-1} v = p/(2p+m)$ and $d(2p+m) - \beta p = p^2 + 3p + m > 0$. Hence $H_0 - cI \succ 0$ for sufficiently large L, so c is not an eigenvalue. There is therefore no eigenvector orthogonal to b_0; as in Theorem 5.2, the spectrum is simple and b_0 is cyclic.

Give all rows positive uniform mass. Exact path preservation now forces exact matching of H and B, and independence of the moment rows forces all $D_{d,m}$ positive coefficients to remain. Because $x_0 = 1$, matching H_{00} already fixes their total mass. For $d = 1$, take the $m + 1$ rows $(x, Y) = (1, 0), (1, f_1), \ldots, (1, f_m)$ directly; their moment vectors are independent and their nonzero one-dimensional cross moment is cyclic. Thus the count is sharp for every $d, m \ge 1$ with a mandatory intercept, under both conic and normalized weights.

E.4 Generic minimum support

Proof of Theorem 5.4. Cyclicity holds generically, using the first response column and Corollary E.1. For the intercept model, choose the other feature columns mutually orthogonal and orthogonal to the constant column in the weighted data inner product, with distinct squared norms different from one, and choose labels giving a cross moment with every coordinate nonzero. This again makes the Krylov determinant nonzero.

Let c_i be the moment atoms for conic weights, or their mass-augmented atoms for normalized weights, and let M be their full function-space dimension: $D_{d,m}$, $D_{d,m} + 1$, or $D_{d,m}$ in the three cases. The sharp constructions supply M independent atoms in each model. For any fixed index set S of size $s < \min(N, M)$, the condition $t = \sum_i p_i c_i \in \operatorname{span}\{c_i : i \in S\}$ is confined to a proper algebraic set. To see properness, assign s independent atoms to S and an additional independent atom to every outside index; its coefficient in t is positive, so t is not in the selected span. Hence some $(s+1)$-minor of the augmented row matrix is a nonzero polynomial. The finite union over all such S, together with the noncyclic set, has measure zero. Outside it, every coreset preserving the path must match moments and needs at least $\min(N, M)$ rows. Theorem 5.2 and the intercept construction give the matching upper bounds. □

E.5 Finite checks and error propagation

Interpolation in Proposition 5.5. Write the nonzero spectral components of b as $b_j = P_j b$, with distinct active eigenvalues $\mu_1, \ldots, \mu_\kappa$. Then $w(\lambda) = \sum_j b_j/(\mu_j + \lambda)$. For distinct $\lambda_0, \ldots, \lambda_\kappa$, the differences $w(\lambda_i) - w(\lambda_0)$ span $\mathcal{K}$: after nonzero row and column scaling their coefficient matrix is the nonsingular Cauchy matrix $(1/(\mu_j + \lambda_i))_{j,i=1}^{\kappa}$. The selected normal equations at these points imply

$$(H_a - H)w(\lambda_i) = b_a - b.$$

Subtract the equation at $i = 0$ to obtain $(H_a - H)\mathcal{K} = 0$. Since $w(\lambda_0) \in \mathcal{K}$, also $b_a = b$. Apply Theorem 5.1. □

Let $w_i = w(\lambda_i)$ at $\kappa + 1$ distinct checkpoints and set

$$D = [w_1 - w_0, \dots, w_\kappa - w_0], \qquad e_i = (H_a + \lambda_i I)w_i - b_a.$$

The matrix D has full column rank κ and spans $\mathcal{K}$. For any $\lambda > 0$, define $c(\lambda) = D^+(w(\lambda) - w_0)$. Then

$$(H_a - H)w(\lambda) - (b_a - b) = \left(1 - \sum_i c_i(\lambda)\right) e_0 + \sum_i c_i(\lambda) e_i. \tag{31}$$

Indeed $e_i = (H_a - H)w_i - (b_a - b)$, and the affine representation of $w(\lambda)$ proves the identity. Consequently,

$$\|w_a(\lambda) - w(\lambda)\| \le \frac{|1 - \sum_i c_i(\lambda)|\,\|e_0\| + \sum_i |c_i(\lambda)|\|e_i\|}{\lambda + \lambda_{\min}(H_a)}. \tag{32}$$

If only coefficient discrepancies $\delta_i = w_a(\lambda_i) - w_i$ are available, then $e_i = -(H_a + \lambda_i I)\delta_i$. The resulting a posteriori bound depends on the interpolation coefficients of the chosen checkpoints and on the lower spectral bound of $H_a + \lambda I$.

F Experimental protocol

Observed data. Diamonds contains 53,940 rows. We predict log price from carat, depth, table, physical dimensions, and indicators for cut, color, and clarity [20]. Independent permutations with seeds 17, 29, and 43 give 43,152 training and 10,788 test rows. Linear and 64-feature Fourier designs use all three seeds; the 128- and 256-feature Fourier designs use seed 17; neural features use seeds 17 and 29.

The RAND table has 20,190 rows [22]. The target is $\log(1 + \text{mdvis})$, with its remaining nine variables as covariates. Seeds 17 and 29 give 16,152 training rows and 4,038 test rows for both linear and neural features. The Texas table has 8,602 rows before removal of missing observations [21]. We retain positive median prices with observed sales, listings, and inventory. The target is log median price; covariates are city indicators, year, log sales, log listings, inventory, and sine/cosine month. Years through 2011 train the model; later years test it. This leaves 5,164 training rows. Both linear and neural designs use this same chronological split.

Feature fitting. Numeric scales use training means and standard deviations. Category names define the indicator dictionary without using responses. Fourier frequencies are Gaussian, scaled by the inverse square root of the input dimension, with uniform phases. Neural features come from two tanh layers of width 64 for diamonds and 32 for the other sources, followed by a scalar training head. Adam uses learning rate 0.003, batches of 1,024, and 30 epochs. After training on the training rows, we freeze the last hidden layer, discard the scalar head, and fit a new least-squares head. The supplement retains the transformations and original float32 activations used for replay.

After feature construction we standardize each nonconstant feature, add an intercept, and remove parameter directions below 10^{-11} times the largest training singular value. This basis change is orthogonal. A QR factorization supplies orthonormal prediction coordinates. The separate replay reconstructs preprocessing and each weighted refit from the observations and saved transformations.

Selection and comparisons. Each construction begins with $2d$ rows: a pivoted QR basis from a leverage sample, followed by additional leverage-sampled rows without replacement. We compare corrections that minimize total added weight $\sum_i z_i$, added trace $\sum_i z_i \|v_i\|^2$, or the largest eigenvalue in (10). All satisfy the same residual equations and retain the initial rows. The eigenvalue formulation begins with a trace constraint and adds violated leading-eigenvector constraints. Cut limits are 96 for $d \le 32$, 64 for $d \le 80$, 16 for $d \le 150$, and eight otherwise; column generation uses at most 20 passes, or six when $d \ge 100$. The target relative gap is $(U - L)/U \le 0.05$, and the reported results retain the gap reached at termination.

For comparison at the support size s returned by the eigenvalue objective, the minimum-weight and trace-minimizing variants begin with $s - d$ initial rows and therefore retain at most s rows. Uniform and QR selection use exactly s rows. Leverage sampling draws s times with replacement, uses inverse-probability weights, and combines repeated indices.

All linear programs use HiGHS dual simplex through SciPy, with primal and dual feasibility tolerances 10^{-9}. Before a dual bound is accepted, every candidate row is checked. The separate replay recomputes the weighted refits, residual equations, covariance eigenvalues, retained initial rows, and dual inequalities without rerunning selection.

Condition refinement. Every reweighting run starts from the minimum-weight construction. It fixes that support, permits weights to become zero, and retains the best positive-definite covariance found. Both sides of the spectral interval start from the initial eigenbasis, and each iteration adds up to eight violated eigenvectors from each end. Limits are 80 iterations for $d \le 80$, 32 for $d \le 150$, and 16 otherwise, with target $D/t \le 1.05$.

The replay checks support restriction, residual cancellation, the full weighted spectrum, positive semidefiniteness and trace of both dual matrices, and every dual inequality on the initial support. It then refits the retained rows. The numerical tolerance is 2×10^{-7}.

Metrics and precision. Condition numbers use $U_S^\top A_S U_S$, so they are invariant to feature units. For full-rank supports, sensitivity is $\mathrm{tr}(TT^\top) - d$ in the same coordinates; for singular supports we use $\|T - U^\top\|_F^2$, extending T by zero on omitted rows. Test RMSE uses responses standardized by their training mean and standard deviation. The 32-bit and 64-bit least-squares calls use the same relative rank threshold, 10^{-12}, and prediction discrepancy is divided by $\max(1, \|X_{\mathrm{test}} w_\star\|)$.

Construction times include the correction solver and column generation but exclude the full fit, feature training, and initial row selection. Refit times are medians of five calls. Gradient descent counts evaluate the spectral recurrence from zero with step $2/(\lambda_{\max} + \lambda_{\min})$ until relative prediction-coordinate error is at most 10^{-8}; they are iteration counts rather than timing measurements.

Ridge reuse. The ridge experiments contain 21 configurations and use a 41-value logarithmic grid from 10^{-5} to 10^{3}. We compare weights matching H and B with weights matching only the fit at $\lambda = 0.1$. Nine configurations have two observed response columns and compare preservation of the first response with joint preservation under one fixed weight vector.

Rational checks. Nine rational instances cover orthogonal blocks, overlapping circuits, and a noiseless direction. The checker enumerates every full-rank support of size $d + 1$, verifies the rational Karush–Kuhn–Tucker inequalities, excludes the relevant singular row spans, and checks the four-row nonconvexity example. A separate rational calculation evaluates the recursion for all 5,050 parameter pairs through $d = 100$.

Source and features	Seed	Rows	κ_{sp}	κ_{mass}	Sensitivity	32-bit error	Test RMSE
Diamonds: linear	17	117	99.5	666.8	19044	2.7e-08	0.165
Diamonds: linear	29	116	140.5	910.2	21356	4.2e-08	0.193
Diamonds: linear	43	118	210.2	956.2	18529	4.5e-08	0.183
Diamonds: Fourier 64	17	228	70.9	65.7	15197	4.0e-08	0.296
Diamonds: Fourier 64	29	233	49.2	66.9	16340	4.7e-08	0.272
Diamonds: Fourier 64	43	228	40.5	42.5	17835	3.3e-08	0.280
Diamonds: Fourier 128	17	378	247.1	332.2	22035	4.6e-08	0.198
Diamonds: Fourier 256	17	746	610.1	591.0	21527	3.0e-08	0.146
Diamonds: neural 64	17	232	220.0	172.1	16710	2.5e-08	0.110
Diamonds: neural 64	29	232	258.2	241.6	16303	4.1e-08	0.111
Insurance: linear	17	49	41.7	20.3	4797	1.5e-07	0.951
Insurance: linear	29	49	26.7	38.7	4176	6.9e-08	0.957
Insurance: neural 32	17	133	151.6	65.5	5208	1.1e-07	0.946
Insurance: neural 32	29	131	65.4	44.1	4695	1.2e-07	0.951
Texas: linear	17	175	41.5	32.4	1991	3.4e-08	0.441
Texas: neural 32	17	122	22.7	12.9	1851	2.9e-08	0.494

Table 3: All covariance-constrained constructions. The minimum-weight comparison uses the same upper row budget. Condition numbers are measured in orthonormal prediction coordinates.

Source and features	Seed	Rows before	Rows after	κ before	κ after	Lower bound	Steps before	Steps after
Diamonds, linear	17	116	74	666.81	5.31	5.05	5653	47
Diamonds, linear	29	115	75	910.18	4.77	4.55	6806	42
Diamonds, linear	43	117	77	956.23	4.97	4.74	7506	44
Diamonds, Fourier 64	17	225	157	65.71	12.53	8.36	519	105
Diamonds, Fourier 64	29	232	166	66.87	12.85	8.38	540	104
Diamonds, Fourier 64	43	226	166	42.54	13.63	8.94	354	112
Diamonds, Fourier 128	17	373	216	332.22	92.90	14.65	2488	745
Diamonds, Fourier 256	17	724	380	590.96	312.63	9.80	4886	2352
Diamonds, neural 64	17	229	152	172.14	18.18	11.49	1457	152
Diamonds, neural 64	29	230	155	241.61	16.99	10.17	1792	140
RAND, linear	17	49	28	20.30	3.40	3.25	182	30
RAND, linear	29	49	28	38.66	5.56	5.30	325	49
RAND, neural 32	17	133	89	65.53	7.14	6.59	558	61
RAND, neural 32	29	131	92	44.14	5.79	5.38	381	50
Texas, linear	17	174	128	32.40	6.38	4.93	265	53
Texas, neural 32	17	122	90	12.86	4.98	4.60	108	43

Table 4: Every reweighting configuration. The lower bound $1/D$ applies to all exact nonnegative weightings on the initial support. Steps evaluate the fixed-step gradient descent recurrence to relative error 10^{-8}.