EXACT RECOVERY THRESHOLDS FOR WEIGHTED DATA SELECTION IN VECTOR-VALUED LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We resolve the threshold part of Question 4 of the COLT 2025 open problem "Data Selection for Regression Tasks" of Hanneke, Moran, Shlimovich and Yehudayoff. We study vector-valued linear regression with square loss $\ell_{(x,y)}(W) = \|Wx-y\|_2^2$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$. The learner returns the minimum-Frobenius-norm empirical risk minimizer. We prove that the minimal budget of weighted examples for recovering the full-data loss on *every* finite dataset is exactly $n^\star(d,m) = (m+1)d$. We determine the weighted selection profile $F_{\text{weighted}}(d,m,n)$ at the near-threshold budget: $F_{\text{weighted}}(d,m,(m+1)d-1) = 1 + 1/(dm^2)$. We recover the known spanning-budget value $F_{\text{weighted}}(d,m,d) = d+1$ for every m, and $F_{\text{weighted}}(d,m,n) = \infty$ for $n < d$. For the smallest open intermediate cell $(d,m) = (2,2)$ we prove $F_{\text{weighted}}(2,2,3) \in [13/8,\ 15/8]$ and $F_{\text{weighted}}(2,2,4) \in [5/4,\ 3/2]$. We reduce the conjectured exact values $13/8$ and $5/4$ to a finite moment problem on the circle with at most seven atoms and assemble structural evidence for it. The upper bounds use a fixed-basis conic compression lemma, a determinant–facet rigidity theorem for maximal certificates, and sharp sparsification lemmas for zero-mean weighted point systems. These tools may be of independent interest. We also exhibit an explicit six-point integer dataset with $d = m = 2$ on which *no* weighted selection of $2d$ points recovers the optimal loss. Thus the scalar sufficient budget $2d$ does not extend to vector-valued outputs. Our new regression-profile results for $m \geq 2$ extend the scalar theory for $m = 1$.

1 INTRODUCTION

Data curation and reusable representations (Gadre et al., 2023; Oquab et al., 2023) call for precise theoretical benchmarks for data selection.

How well can a fixed, natural learning rule learn from only n examples selected from a larger dataset? This question underlies coreset construction, data pruning and subsampling for regression. A typical multi-output instance is a linear head trained on a fixed representation, where all outputs must share one selected set of examples and one weight vector. Hanneke et al. (2025b) posed a worst-case version for basic regression in a COLT 2025 open-problem note, with a companion paper (Hanneke et al., 2025a). For scalar linear regression they determined the endpoint regimes under weighted selection and the minimum-norm empirical risk minimizer. The worst-case ratio between the selected model's loss and the optimal full-data loss is infinite for $n < d$, equals $d+1$ at $n = d$, and equals 1 for $n \geq 2d$ (Hanneke et al., 2025b, Theorem 1). The intermediate range $d < n < 2d$ was posed as their Question 2. Their note concludes with a unified *vector-valued* formulation using predictors $W\colon \mathbb{R}^d \to \mathbb{R}^m$, square loss and the minimum *Frobenius*-norm ERM. It asks, as Question 4:

> *Given d, m, n, what is the value of $F_{\text{weighted}}(d,m,n)$? In particular, what is the minimal $n = n(d,m)$ such that this ratio equals 1?*

This paper resolves the threshold question completely and determines several further values of the profile $F_{\text{weighted}}(d,m,\cdot)$. Throughout, a dataset $D = (z_i)_{i=1}^N$ is a finite sequence of points $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^m$. We write $\ell_z(W) = \|Wx - y\|_2^2$, $L_D(W) = \frac{1}{N}\sum_i \ell_{z_i}(W)$ and $L_D^\star = \min_W L_D(W)$. The learner A returns the empirical risk minimizer of minimal Frobenius norm. Weighted selection of budget n chooses n points of D (repetitions allowed) and a convex combination F of their losses. Its loss on the full dataset is $L_D(A(F))$. The profile $F_{\text{weighted}}(d,m,n)$

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.

is the supremum over datasets of $\inf_F L_D(A(F))/L_D^\star$, with degenerate-ratio conventions fixed in Section 2. The exact-recovery threshold is $n^\star(d,m) = \min\{n : F_{\text{weighted}}(d,m,n) = 1\}$.

Contributions. All statements below are for $d, m \geq 2$. They hold verbatim for $d, m \geq 1$ (Remark 2.4).

Exact threshold (Theorem 3.4): $n^\star(d,m) = (m+1)d$. The upper bound is data-dependent: every dataset of feature rank $r \geq 1$ is recovered exactly by $(m+1)r$ weighted points (Theorem 3.1). The lower bound is an explicit integer dataset on $(m+1)d$ points whose optimal ratio at budget $(m+1)d-1$ is exactly $1 + \frac{m+1}{2md(m^2+m-1)}$ (Theorem 3.3).

Near-threshold value (Theorem 5.2): $F_{\text{weighted}}(d,m,(m+1)d-1) = 1 + 1/(dm^2)$. The proof hinges on rigidity of maximal exact certificates (Theorem 4.1). Every dataset unrecoverable within budget $(m+1)d-1$ has a minimal spanning zero certificate whose support decomposes into d disjoint $(m+1)$-point residual circuits on independent feature lines. All nonzero features of such a dataset lie on these lines (Theorem 4.2). A sharp m-point mean lemma (Lemma 5.1) then finishes the argument.

Low budgets (Theorem 6.1): $F_{\text{weighted}}(d,m,n) = \infty$ for $n < d$ and $F_{\text{weighted}}(d,m,d) = d+1$, *independently of* m, as a direct consequence of Dereziński & Warmuth (2017) and the scalar lower bounds. A shared volume-sampling subset serves all output rows.

The smallest open cell $(d,m) = (2,2)$ (Section 6.2): $F_{\text{weighted}}(2,2,3) \in [13/8, 15/8]$ and $F_{\text{weighted}}(2,2,4) \in [5/4, 3/2]$ (Theorem 6.4). The conjectured values $13/8$ and $5/4$ are reduced to a finite-dimensional moment problem with at most seven atoms on the unit circle (Proposition 6.6). For this moment problem, we prove a sharp result for the two-direction class, stratified local maximality, two exact obstructions, and closure of several classes (Appendix F).

A byproduct: the scalar sufficient budget $2d$ fails for every $m \geq 2$. On an explicit six-point integer dataset with $d = m = 2$, every weighted selection of four points has loss at least $\frac{23}{20} L_D^\star$ (Proposition 5.3).

Table 1 summarizes the resulting picture of the profile.

Table 1: The weighted selection profile $F_{\text{weighted}}(d,m,n)$ for the minimum-Frobenius-norm ERM ($d, m \geq 2$; the scalar specializations $m = 1$ are due to Hanneke et al. (2025b), Dewasurendra (2026a) and Zhang (2026)).

budget n	$F_{\text{weighted}}(d,m,n)$
$n < d$	∞ (Theorem 6.1)
$n = d$	$d+1$ (Theorem 6.1)
$d < n < (m+1)d-1$	open; at $(2,2)$: $[13/8, 15/8]$, $[5/4, 3/2]$ for $n = 3, 4$ (Theorem 6.4)
$n = (m+1)d-1$	$1 + 1/(dm^2)$ (Theorem 5.2)
$n \geq (m+1)d$	1 (Theorem 3.4)

Techniques. The threshold upper bound uses a *fixed-basis conic compression* lemma (Lemma B.1). After we pin a feature basis of r points, the negated sum of their residual dyads lies in the cone of the remaining dyads. Conic Carathéodory represents it with at most mr of them, giving an exact certificate on $(m+1)r$ points. The argument is shorter than the Steinitz-based scalar route of Hanneke et al. (2025a) and is data-dependent. The near-threshold upper bound uses a *determinant–facet rigidity theorem* (Theorem 4.1). Consider a certificate of maximal size $(m+1)d$. Its weighted feature Gram determinant is a polynomial of degree at most d in the weights, although the dyads span an md-dimensional space. It vanishes on every facet of the normalized certificate polytope. This forces a simplex and decomposes the certificate into d disjoint $(m+1)$-point circuits on d independent feature lines. A sharp *m-point mean lemma* for zero-mean weighted point systems then finishes the assembly. For the $(2,2)$ cell, a complex-variable dictionary turns whitened instances into moment systems on the unit circle and the selection cost into a convex combination of *two-point interpolation coefficients* (Lemma 6.2).

1.1 Related and concurrent work

Prior work. The problem and the scalar endpoint results are due to Hanneke et al. (2025b;a). The scalar value at $n = 2d - 1$ is proved in Dewasurendra (2026a) and Zhang (2026), and further scalar intermediate values in Zhang (2026). The budget-d upper bound uses volume sampling and the exact expectation formulas of Dereziński & Warmuth (2017) (Theorem 5, used as an inequality; Section 6.1). The general k-point mean lemma (Theorem G.1) is proved in Appendix G by an elementary argument. For $k = 2$ and $M \geq 3$, its constant coincides with, and follows from, Theorem 2 of Dewasurendra (2026b). That theorem covers neither $M = 2$ nor general k. Tinarrage (2026) proves a radius-normalized, equal-weight analogue. Carathéodory-type compressions also underlie exact weighted coresets for least squares. For example, the summaries of Maalouf et al. (2019) preserve the covariance sums of the input. An exact coreset preserves the objective for every parameter. Here the summary must reproduce the model returned by a fixed learner retrained on the selected original points. We obtain matching lower bounds on the support and exact worst-case ratios below it.

Concurrent unrefereed preprints. Recent unrefereed preprints treat other questions of Hanneke et al. (2025b): Dewasurendra (2026b) determines the budget-2 column of Question 1, and Dewasurendra (2026a) proves the scalar value at budget $2d - 1$ (Question 2). Neither manuscript treats the weighted vector-valued Question 4. The relation of our sparsification results to Theorem 2 of Dewasurendra (2026b) is stated above. We make no priority claim on them.

Organization. Sections 2–6 give results and proof sketches. Appendices A–F give full proofs. Appendix G gives the general k-point mean lemma.

2 Recovery certificates and proof architecture

Throughout, $d, m, n \geq 1$ are integers. A dataset is a finite sequence $D = (z_i)_{i=1}^N$ of points $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^m$, $N \geq 1$, with repetitions allowed. A linear predictor $W \in \mathbb{R}^{m \times d}$ has loss $\ell_z(W) = \|Wx - y\|_2^2$ and average loss $L_D(W) = \frac{1}{N} \sum_{i=1}^N \ell_{z_i}(W)$. Write $L_D^\star = \min_W L_D(W)$. The learning rule A maps a nonnegatively weighted objective $F = \sum_i c_i \ell_{z_i}$ (with $c_i \geq 0$, $\sum_i c_i = 1$) to the minimum-Frobenius-norm minimizer of F. For $c \in \mathbb{R}^N_{\geq 0}$ write $F_c = \sum_i c_i \ell_{z_i}$. Certificates below are nonzero and unnormalized. For such c, we set $A(F_c) := A(F_{c/\sum_i c_i})$, the minimal-norm minimizer of F_c (positive rescaling preserves the minimizer set). The smallest full-data loss achievable by weighted selection with budget n is

$$L_D^\star(n; \text{weighted}) = \inf_{\substack{z_{j_1}, \ldots, z_{j_n} \in D \\ F \in \operatorname{conv}(\ell_{z_{j_1}}, \ldots, \ell_{z_{j_n}})}} L_D\big(A(F)\big),$$

$$R_D(n) = \frac{L_D^\star(n; \text{weighted})}{L_D^\star}, \qquad F_{\text{weighted}}(d, m, n) = \sup_D R_D(n),$$

where the supremum is over all datasets (all $N \geq 1$), and $n^\star(d, m) = \min\{n : F_{\text{weighted}}(d, m, n) = 1\}$. Following the convention of Hanneke et al. (2025a), $R_D(n)$ is 1 when numerator and denominator are both 0, and ∞ when only the denominator is 0. Since $L_D(A(F)) \geq L_D^\star$ for every objective F, $R_D(n) \geq 1$. We write the inner optimization as an infimum (as in Hanneke et al. (2025a)): all exact-recovery certificates below are attained, while some bounds in Section 6.2 are limits of admissible selections. The following normalization is immediate and used implicitly.

Lemma 2.1 (Selection semantics)**.** *Selections of budget n are in value-preserving correspondence with weight vectors $c \in \mathbb{R}^N_{\geq 0}$, $\sum_i c_i = 1$, $|\operatorname{supp} c| \leq n$. Repeated picks merge weights, and zero-weight picks can be discarded. Any c with $|\operatorname{supp} c| \leq n$ is realizable with exactly n slots by repeating a chosen point and splitting its weight. In particular the feasible family grows with n, so $L_D^\star(n; \text{weighted})$, $R_D(n)$ and $F_{\text{weighted}}(d, m, n)$ are nonincreasing in n.*

Geometry of the full-data minimizer. Let $T = \operatorname{span}\{x_1, \ldots, x_N\}$, $r = \dim T$ (the feature rank), and let P_T be the orthogonal projection onto T. Let W° be the minimal-Frobenius-norm full-data minimizer, and set $\rho_i = W^\circ x_i - y_i$ (residuals) and $M_i = \rho_i x_i^\top \in \mathbb{R}^{m \times d}$ (residual dyads). The objective and the squared Frobenius norm are additive across output rows, so $A(F_c)$ consists of the

minimal-ℓ_2-norm scalar solutions row by row, with the single shared weight vector c (Lemma A.1). Moreover, $W^\circ = W^\circ P_T$, the first-order condition $\sum_i M_i = 0$ holds, and

$$L_D(W^\circ + H) = L_D^\star + \tfrac{1}{N}\sum_{i=1}^N \|Hx_i\|^2 \qquad \text{for every } H \in \mathbb{R}^{m\times d} \tag{1}$$

(Lemma A.2). The two conditions in the next lemma have separate roles: the zero-dyad condition makes W° stationary for the selected objective, while the spanning condition makes the selected and full-data minimizer sets coincide.

Lemma 2.2 (Exact certificate). *Let $c \in \mathbb{R}^N_{\geq 0}$ be nonzero, with support S (so $c_i > 0$ on S). If*

$$\sum_{i\in S} c_i M_i = 0 \qquad \textit{and} \qquad \operatorname{span}\{x_i : i \in S\} = T, \tag{2}$$

then $A(F_c) = W^\circ$. In particular, the selection realizes the full-data optimal loss, and the ratio is 1 (also when $L_D^\star = 0$, by the $0/0$ convention).

We call a nonzero weight vector satisfying (2) a *spanning zero certificate*, and define the minimal certificate size

$$\tau(D) \;=\; \min\big\{|\operatorname{supp} c| :\; c \text{ is a spanning zero certificate}\big\}. \tag{3}$$

Certificates are built from positive circuits of the residual dyads.

Definition 2.3 (Positive circuit). For a family of vectors $(a_i)_{i\in J}$, a nonempty set $C \subseteq J$ is a *positive circuit* if there are coefficients $\lambda_i > 0$ ($i \in C$) with $\sum_{i\in C} \lambda_i a_i = 0$ and $\dim\ker[a_i]_{i\in C} = 1$. Equivalently, the positive dependence is unique up to scaling and no proper subset of C supports a nonzero nonnegative dependence. The positive circuits are exactly the supports of the extreme rays of the cone $\{\lambda \geq 0 : \sum_i \lambda_i a_i = 0\}$. Every nonzero point of this cone is a positive combination of extreme rays (conic Carathéodory). A singleton $\{i\}$ with $a_i = 0$ is a positive circuit.

Finally, scalar instances embed: if all m coordinates of every label agree, then every row of the vector problem is the same scalar problem, so $F_{\text{weighted}}(d, m, n) \geq F^{\text{sc}}_{\text{weighted}}(d, n)$, the scalar weighted profile of Hanneke et al. (2025b) (Lemma A.3).

Remark 2.4 (Range of validity). Nothing in the proofs of Theorems 3.4, 5.2 and 6.1 uses $d \geq 2$ or $m \geq 2$: they hold for all $d, m \geq 1$. At $m = 1$ they specialize to $n^\star(d, 1) = 2d$ and $F_{\text{weighted}}(d, 1, 2d-1) = 1 + \frac{1}{d}$, matching Hanneke et al. (2025b, Theorem 1), Dewasurendra (2026a) and Zhang (2026). At $d = 1$ the whole profile is Corollary G.3. We state the theorems for $d, m \geq 2$ only to delineate the new contribution: for $m = 1$ both conclusions of Theorem 6.1 and the threshold $2d$ are due to Hanneke et al. (2025b;a), and the value at $n = 2d - 1$, stated in Hanneke et al. (2025b), is proved in Dewasurendra (2026a) and Zhang (2026). **Our new regression-profile results concern the vector case $m \geq 2$.**

Proof routes. Selected reading links: *Threshold:* Lemmas A.2, 2.2, B.1; Theorems 3.1, 3.3, 3.4 (Appendices A–B). *Near threshold:* Definition 2.3; Theorems 4.1, 4.2; Lemma 5.1; Theorem 5.2 (Appendix D).

3 The exact threshold $(m + 1)d$

Upper bound: fixed-basis conic compression. The key combinatorial fact is that a zero-sum family of vectors admits a nonnegative zero-sum subfamily that contains any prescribed index set B and at most $\dim \operatorname{span}\{a_i : i \notin B\}$ further elements (Lemma B.1, a consequence of conic Carathéodory). Applied to the residual dyads with B a feature basis, it gives:

Theorem 3.1 (Data-dependent upper bound). *Every dataset with feature rank $r \geq 1$ admits a spanning zero certificate of support at most $(m + 1)r$. Hence $\tau(D) \leq (m + 1)r$ and $L_D^\star\big((m + 1)d; \text{weighted}\big) = L_D^\star$ for every dataset. If $r = 0$, a single point suffices.*

Proof sketch. Pick a basis B of r features. The dyads $M_i = \rho_i x_i^\top$ have all rows in T, so the dyads outside B span a space of dimension at most mr. Lemma B.1 therefore yields $c \geq 0$, positive on B, with $\sum_i c_i M_i = 0$ and support at most $r + mr$. The support contains the basis, hence spans T, and Lemma 2.2 gives $A(F_c) = W^\circ$. The full proof is in Appendix B. The same argument with $s = \dim \operatorname{span}\{\rho_i\}$ replacing m gives $\tau(D) \leq (s + 1)r$ for $r \geq 1$ (Remark B.2): the dataset's budget is governed by its effective residual output dimension. □

Lower bound: the axial simplex instance. Fix $d, m \geq 1$, let $\mathbf{1}_m$ be the all-ones vector, and let $v_j = e_j$ $(1 \leq j \leq m)$, $v_{m+1} = -\mathbf{1}_m$, so that $\sum_j v_j = 0$ and no proper subset of $\{v_j\}$ contains 0 in its convex hull.

Definition 3.2 (Integer axial instance)**.** $D_{d,m}$ consists of the $N = (m+1)d$ points $z_{ij} = (e_i,\ \mathbf{1}_m - v_j)$ for $i \in [d]$, $j \in [m+1]$. Explicitly, the label of z_{ij} is $\mathbf{1}_m - e_j$ for $j \leq m$ and $2 \cdot \mathbf{1}_m$ for $j = m+1$.

On $D_{d,m}$ the unique full-data minimizer is $W^\circ = \mathbf{1}_m \mathbf{1}_d^\top$ with residual v_j at z_{ij} and $L^\star_D = \frac{2m}{m+1}$. The weighted normal equations decouple across the d axes. The returned column on an axis is the conditional weighted mean of its labels (or 0 if the axis is unused). The returned model's loss exceeds $L^\star_D$ by $\frac{1}{d}\sum_i \|\delta_i\|^2$, where δ_i is the deviation of column i from $\mathbf{1}_m$ (Lemma B.3). With $Q_m = m^2 + m - 1$, the squared distance from the origin to the convex hull of any nonempty proper subset of $\{v_j\}$ is at least $1/Q_m$, attained on the facets omitting some e_k (Lemma B.4).

Theorem 3.3 (Exact deficit of the axial instance)**.** *For every $d, m \geq 1$,*

$$\frac{L^\star_{D_{d,m}}\big((m+1)d - 1; \text{weighted}\big)}{L^\star_{D_{d,m}}} \;=\; 1 + \frac{m+1}{2md\,(m^2+m-1)} \;>\; 1.$$

Proof sketch. A selection of support at most $N-1$ leaves some axis either empty (deviation $\|\mathbf{1}_m\|^2 = m$) or supported on a proper subset of its $m+1$ points (deviation at least $1/Q_m$ by Lemma B.4), so the excess loss is at least $\frac{1}{dQ_m}$. Omitting one point e_1 on one axis and using the facet-foot weights there attains this bound. Dividing by $L^\star_D$ gives the value. □

Theorem 3.4 (Threshold)**.** $n^\star(d, m) = (m+1)d$ *for all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4), and $F_{\text{weighted}}(d, m, n) = 1$ for all $n \geq (m+1)d$.*

Proof. Theorem 3.1 gives ratio 1 at budget $(m+1)d$ for every dataset. Theorem 3.3 exhibits a dataset with ratio > 1 at budget $(m+1)d - 1$, and $F_{\text{weighted}}(d, m, \cdot)$ is nonincreasing (Lemma 2.1). □

Corollary 3.5. $n^\star(2, 2) = 6$. *On the instance $D_{2,2}$ (six integer points, labels $(0,1), (1,0), (2,2)$ over each of e_1, e_2) the optimal five-point ratio is exactly $\frac{43}{40}$.*

4 Rigidity of maximal certificates

Theorem 3.1 gives $\tau(D) \leq \max\{1, (m+1)r\}$. The upper bound at budget $(m+1)d - 1$ requires understanding the datasets with $\tau(D) = (m+1)d$, which we call *certificate-maximal*. The key is a rigidity theorem for maximal certificates.

Theorem 4.1 (Determinant–facet rigidity)**.** *Let the features span $\mathbb{R}^d$ and suppose $\tau(D) = k := (m+1)d$. Let c be a spanning zero certificate with $|\operatorname{supp} c| = k$ and support S. Then S splits uniquely (up to permutation) into d pairwise disjoint blocks $S = C_1 \sqcup \cdots \sqcup C_d$ with $|C_j| = m+1$, where each C_j supports a positive circuit of the dyads whose features span a one-dimensional line U_j, and $U_1, \ldots, U_d$ are linearly independent.*

Proof sketch. Vectorize the dyads of S into a matrix A with md rows and k columns. The normalized nonnegative kernel $P = \{\theta \geq 0 : \mathsf{A}\theta = 0,\ \mathbf{1}^\top\theta = 1\}$ is a polytope of dimension $h-1 \geq d-1$ whose relative interior consists of the strictly positive points. The polynomial $f(\theta) = \det(\sum_i \theta_i x_i x_i^\top)$, of degree at most d, is positive on the relative interior and must vanish at every boundary point: a boundary point with $f > 0$ would be a spanning zero certificate of smaller support. A nonzero polynomial of degree $\leq d$ vanishing on every facet is divisible by the facets' affine forms, so P has at most d facets. A bounded $(h-1)$-polytope has at least h facets, so $h = d$ and P is a simplex (a point if $d = 1$). The d extreme rays give the blocks C_j. Dropping a ray gives a boundary point where the features cannot span. This forces the feature spans U_j of the blocks to be a minimal spanning family, hence lines. The bound $|C_j| \leq m+1$ then forces $|C_j| = m+1$ and disjointness. The full proof is in Appendix D. □

Using the maximality of $\tau(D)$ once more, the same line-circuit structure governs every nonzero-feature point:

Theorem 4.2 (Line structure). *Let the features span $\mathbb{R}^d$ and $\tau(D) = (m+1)d = k$. Then:*

1. *every data point with $x_i \neq 0$ belongs to some positive circuit of dyads supported on a single feature line and of size exactly $m+1$;*
2. *there is no positive circuit of dyads whose features span a subspace of dimension ≥ 2, and no line circuit of size $\leq m$; in particular no point has $x_i \neq 0$ and $\rho_i = 0$;*
3. *the nonzero features lie on exactly d linearly independent lines.*

Consequently, any dataset whose nonzero features are not *of this d-line form satisfies $\tau(D) \leq (m+1)d - 1$ and is recovered exactly within budget $(m+1)d - 1$.*

5 The near-threshold value $1 + 1/(dm^2)$

The assembly needs one more ingredient: a sharp sparsification bound for zero-mean weighted point systems. The general k-point version is Theorem G.1. The case below is elementary:

Lemma 5.1 (m-point mean lemma). *Let $z_i \in \mathbb{R}^m$ carry weights $p_i > 0$ with $\sum_i p_i = 1$ and $\sum_i p_i z_i = 0$, and let $E = \sum_i p_i \|z_i\|^2$. Then some convex combination δ of at most m of the points satisfies $\|\delta\|^2 \leq E/m^2$. The constant is attained by the uniformly weighted regular simplex.*

Proof sketch. Decompose p into positive circuits and choose one with second moment at most E. For a circuit $q_1, \ldots, q_{m+1} > 0$, deleting point j leaves the m-point mean $\mu_j = -\frac{q_j}{1-q_j} z_j$. The inequality $\sum_j (1-q_j)^2/q_j \geq m^2$ (Cauchy–Schwarz) shows that not all j can violate $\|\mu_j\|^2 \leq E/m^2$. The full proof is in Appendix D. □

Theorem 5.2 (Near-threshold value). *For all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4),* $F_{\text{weighted}}\big(d, m, (m+1)d-1\big) = 1 + \dfrac{1}{dm^2}$.

Proof sketch. Upper bound. If the feature rank is below d, or $L_D^\star = 0$, or $\tau(D) \leq (m+1)d - 1$, the ratio is 1. Otherwise Theorem 4.2 places all nonzero features on d independent lines $u_1, \ldots, u_d$, with $x_i = s_i u_{j(i)}$. The first-order condition splits into one zero-mean condition per line, $\sum_{i \in I_j} s_i \rho_i = 0$. Let $z_i = \rho_i / s_i$ and $p_i = s_i^2 / \sum_{I_j} s_{i'}^2$, so that $\sum_{I_j} p_i z_i = 0$ and $\sum_{I_j} p_i \|z_i\|^2 = R_j/g_j$ with $R_j = \sum_{I_j} \|\rho_i\|^2$ and $g_j = \sum_{I_j} s_i^2$. Choose the lightest line $j_\star$ ($R_{j_\star} \leq N L_D^\star / d$). Fit every other line exactly with a positive $(m+1)$-point circuit, and line $j_\star$ with the m-point mean δ of Lemma 5.1. The weighted normal equations decouple along the basis, giving $\widehat{W} = W^\circ + H$ with $Hu_j = 0$ for $j \neq j_\star$ and $Hu_{j_\star} = -\delta$. By (1), the excess loss is $g_{j_\star} \|\delta\|^2 / N \leq R_{j_\star}/(Nm^2) \leq L_D^\star/(dm^2)$ on a support of size $(d-1)(m+1) + m = (m+1)d - 1$. *Lower bound.* Take d axes, each carrying the vertices $w + r_j$ of a regular simplex $r_1, \ldots, r_{m+1}$ with $\|r_j\|^2 = \frac{m}{m+1}$. The simplex's facets are at squared distance $h^2 = \frac{1}{m(m+1)}$ from the origin, and we take $\|w\| = h$. Any selection of $N-1$ points leaves an axis empty or on a proper subset, costing at least h^2/d in excess loss. Deleting one vertex attains this bound. The ratio is $1 + \frac{h^2/d}{m/(m+1)} = 1 + \frac{1}{dm^2}$. □

Failure at $n = 2d$ for $m \geq 2$. The scalar rule "$n \geq 2d$ suffices" does not extend to vector outputs (Remark 5.4).

Proposition 5.3. *For all $d \geq 1, m \geq 2$, $F_{\text{weighted}}(d, m, 2d) \geq F_{\text{weighted}}\big(d, m, (m+1)d-1\big) = 1 + \frac{1}{dm^2} > 1$, since $2d \leq (m+1)d - 1$ and F_{weighted} is nonincreasing in n. Concretely, for $d = m = 2$ the instance $D_{2,2}$ of Corollary 3.5 satisfies: every weighted selection of at most $2d = 4$ points has*

$$L_D(A(F)) \;\geq\; \frac{23}{20} \cdot L_D^\star,$$

with equality attained (allocate two points per axis, at conditional weights realizing the nearest facet points).

The proof enumerates per-axis costs and allocations (Appendix B).

Remark 5.4 (Scalar and vector recovery budgets)**.** The weighted scalar trichotomy of Hanneke et al. (2025b, Theorem 1) gives ratio 1 for all $n \geq 2d$. This sufficient budget does not extend to general m: the scalar result applies to $m = 1$, whereas for every $m \geq 2$, Theorem 3.4 shows that worst-case exact recovery requires $(m+1)d > 2d$ points. Quantitatively, $F_{\text{weighted}}(d, m, 2d) \geq F_{\text{weighted}}(d, m, (m+1)d - 1) = 1 + \frac{1}{dm^2} > 1$ by monotonicity. Proposition 5.3 makes the failure concrete.

6 Further results on the selection profile

6.1 Low budgets $n \leq d$

Theorem 6.1. *For all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4),*

$$F_{\text{weighted}}(d, m, n) = \infty \text{ for every } n < d, \qquad \text{and} \qquad F_{\text{weighted}}(d, m, d) = d + 1.$$

The lower bounds are inherited from the scalar case through Lemma A.3: by Hanneke et al. (2025b, Theorem 1), $F^{\text{sc}}_{\text{weighted}}(d, n) = \infty$ for $n < d$ and $F^{\text{sc}}_{\text{weighted}}(d, d) = d+1$. The content of the theorem is the upper bound $F_{\text{weighted}}(d, m, d) \leq d + 1$ *uniformly in* m, which requires one subset serving all m output rows simultaneously.

Proof sketch. Let $X \in \mathbb{R}^{d \times N}$ have full row rank and $y \in \mathbb{R}^N$. Size-d volume sampling draws d columns with probability proportional to $\det(X_S)^2$, and $w(S)$ is the interpolant of the subproblem. Dereziński & Warmuth (2017) proved that $\mathbb{E}[L(w(S))] \leq (d+1)L(w^\star)$ for the total square loss, with equality when X is in general position (their Theorem 5). We use only the inequality, which holds without general position. Equality can fail for degenerate X; see Appendix C. The distribution over subsets depends only on the features. For a dataset of feature rank r (the cases $r = 0$ and $L^\star_D = 0$ are trivial), we run size-r volume sampling in coordinates of T, apply the inequality to each of the m output rows, and sum: $\mathbb{E}[L_D(W_S)] \leq (r+1)L^\star_D \leq (d+1)L^\star_D$ for the matrix W_S whose rows are the per-row interpolants. Some subset S achieves the bound, and W_S is exactly what the Frobenius rule returns on S with any positive weights, since the r selected features are linearly independent. □

6.2 The smallest intermediate cell $(d, m) = (2, 2)$

For $d, m \geq 2$ and $d < n < (m+1)d - 1$ the profile remains open (for $d = 1$ it is given by Corollary G.3). Apart from the cells settled by Zhang (2026), its scalar counterpart also remains open. We treat the smallest cell $(d, m) = (2, 2)$, where the open budgets are $n = 3, 4$.

A complex dictionary for the hard branches. Fix a $(2, 2)$ dataset of feature rank 2 with $L^\star_D > 0$ and $\tau(D) > n$. For selections whose features span $\mathbb{R}^2$, the weighted minimizer is unique. Its full-data loss is invariant under invertible linear changes of features, so we may assume the data are whitened, $\sum_i x_i x_i^\top = \gamma I_2$. Identify $\mathbb{R}^2 \cong \mathbb{C}$, write ξ_i for the feature and η_i for the residual ρ_i of point i, and for $\xi_i \neq 0$ set

$$p_i = \frac{|\xi_i|^2}{\sum_j |\xi_j|^2}, \qquad t_i = \frac{\bar{\xi}_i}{\xi_i} \in S^1, \qquad z_i = \frac{\eta_i}{\xi_i} \in \mathbb{C}, \qquad E = \sum_i p_i |z_i|^2 > 0.$$

Whitening and the first-order condition become the three moment conditions $\mathbb{E}_p t = \mathbb{E}_p z = \mathbb{E}_p(\bar{t}z) = 0$. Conversely, every finite *moment system* (p_i, t_i, z_i) satisfying them is realized by a whitened dataset. A selection with weights c corresponds to $q_i \propto c_i |\xi_i|^2$. With $\mu = \mathbb{E}_q t$, $\nu = \mathbb{E}_q z$, $\lambda = \mathbb{E}_q(\bar{t}z)$, the map $W^\circ - \widehat{W}$ is $\xi \mapsto a\xi + b\bar{\xi}$ where

$$\begin{pmatrix} 1 & \mu \\ \bar{\mu} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \nu \\ \lambda \end{pmatrix}, \qquad |\mu| < 1 \iff \text{selected features span } \mathbb{R}^2, \tag{4}$$

and the excess ratio of the selection is at most $C(q)/E$ with $C(q) = |a|^2 + |b|^2$ (with equality when no feature vanishes; Appendix E.2). For $k \in \{3, 4\}$ let Γ_k be the supremum over moment systems of $\frac{1}{E} \inf\{C(q) : |\operatorname{supp} q| \leq k,\ |\mu(q)| < 1\}$. The ill-conditioning factor $(1 - |\mu|^2)^{-1}$ in (4) disappears when (a, b) is written in terms of the two-point interpolation coefficients:

Lemma 6.2 (Interpolation convex combination). *For $t_i \neq t_j$ let $B_{ij} = \left(\frac{t_i z_j - t_j z_i}{t_i - t_j}, \frac{z_i - z_j}{t_i - t_j}\right) \in \mathbb{C}^2$ (the coefficients of the affine function of t through the two points) and $d_{ij} = |t_i - t_j|^2$. Then for every admissible q,*

$$\Delta(q) := 1 - |\mu|^2 = \sum_{i<j} q_i q_j d_{ij}, \qquad (a, b) = \sum_{\substack{i<j \\ t_i \neq t_j}} \frac{q_i q_j d_{ij}}{\Delta(q)} B_{ij}.$$

In particular, (a, b) is a convex combination of the pairwise interpolation coefficients. The conditioning factor $(1 - |\mu|^2)^{-1}$ is absorbed entirely.

Consequently, for a support S with $2 \leq |S| \leq 3$ containing two distinct directions, the infimum of $C(q)$ over admissible q supported in S equals the squared distance from the origin to $K_S = \operatorname{conv}\{B_{ij} : i, j \in S,\ t_i \neq t_j\}$ (Lemma E.1). Two further identities follow. The star of an atom i, i.e. the coefficients B_{ij} over j, lies in a real two-dimensional affine plane. The star has $P_{ij} = \frac{1}{2} p_j d_{ij}$-weighted mean β_i with $\|\beta_i\|^2 = \frac{1}{2}|z_i|^2$ and variance $E + \frac{|z_i|^2}{2} - L_i$ (Lemma E.4). Applying Lemma 5.1 in that plane and choosing i with $|z_i|^2 \leq E$ gives $\inf_{|\operatorname{supp} q| \leq 3, |\mu| < 1} C(q) \leq \frac{7E}{8}$. Letting the weight concentrate on atom i gives $\inf_{|\operatorname{supp} q| \leq 4, |\mu| < 1} C(q) \leq \frac{E}{2}$ (Lemma E.5). Hence $\Gamma_3 \leq \frac{7}{8}$ and $\Gamma_4 \leq \frac{1}{2}$.

Lower bounds.

Proposition 6.3. *There is an explicit dataset $D^{\triangle}$ with*

$$\frac{L^\star_{D^\triangle}(3; \text{weighted})}{L^\star_{D^\triangle}} = \frac{13}{8}, \qquad \frac{L^\star_{D^\triangle}(4; \text{weighted})}{L^\star_{D^\triangle}} = \frac{5}{4}.$$

Hence $F_{\text{weighted}}(2, 2, 3) \geq \frac{13}{8}$ and $F_{\text{weighted}}(2, 2, 4) \geq \frac{5}{4}$.

The instance is axial: two axes, each carrying the vertices $w + r_j$ of a centered equilateral triangle ($\|r_j\|^2 = \frac{2}{3}$, $\|w\|^2 = \frac{5}{6}$). Minimizing the per-axis costs over allocations gives the two values (Appendix E).

Theorem 6.4. *$F_{\text{weighted}}(2, 2, 3) \in \left[\frac{13}{8}, \frac{15}{8}\right]$ and $F_{\text{weighted}}(2, 2, 4) \in \left[\frac{5}{4}, \frac{3}{2}\right]$.*

Proof sketch. The lower bounds follow from Proposition 6.3. For the upper bounds, the branches of feature rank ≤ 1, $L^\star_D = 0$, or $\tau(D) \leq n$ have ratio 1 (Theorem 3.1). Otherwise, the dictionary applies and the ratio is at most $1 + \Gamma_3 \leq \frac{15}{8}$ for $n = 3$ and $1 + \Gamma_4 \leq \frac{3}{2}$ for $n = 4$. The closure points are limits of admissible selections within the same budget. This suffices for the inner infimum. The full proofs are in Appendix E. □

The conjecture and its reduction.

Conjecture 6.5. *$F_{\text{weighted}}(2, 2, 3) = \frac{13}{8}$ and $F_{\text{weighted}}(2, 2, 4) = \frac{5}{4}$. Equivalently, for every moment system,*

$$\inf_{|\operatorname{supp} q| \leq 3,\ |\mu| < 1} C(q) \leq \frac{5E}{8} \qquad \text{and} \qquad \inf_{|\operatorname{supp} q| \leq 4,\ |\mu| < 1} C(q) \leq \frac{E}{4}.$$

The two formulations are equivalent (Remark E.6), and the conjecture is a finite-dimensional problem:

Proposition 6.6. *For fixed atoms (t_i, z_i), the closed set $\mathcal{P} = \{p \in \mathbb{R}^N_{\geq 0} : \sum_i p_i = 1,\ \mathbb{E}_p t = \mathbb{E}_p z = \mathbb{E}_p(\bar{t}z) = 0\}$ is a polytope cut out by at most 7 linear equalities. Its points with full support are the admissible weight vectors. Minimizing the linear functional $E(p)$ over this polytope yields a basic feasible solution $p^\star$ of support at most 7 with $E(p^\star) \leq E(p)$. The subsystem on $\operatorname{supp} p^\star$ is again a moment system, and every selection admissible for it is admissible for the original system. Consequently each inequality of Conjecture 6.5 holds for all finite systems if and only if it holds for all systems with at most seven atoms. Since atoms may be split into identically placed copies without changing anything, it suffices to treat systems with* exactly *seven labelled atoms.*

Evidence. Appendix F proves four further statements. **(i)** *Two-direction systems* (all atoms on two antipodal directions) satisfy both conjectured inequalities. Both constants are attained simultaneously by six equally weighted atoms forming an equilateral z-triangle over each direction. For the three-point bound, this is, up to merging coincident atoms, the only equality configuration (Proposition F.1). **(ii)** The equality fiber is a stratified local maximum of the normalized three-point value $E^{-1}\inf C$: energy imbalance gives the cusp $\frac{1}{8}(5-3|\eta|)$ below its value $\frac{5}{8}$, and splitting a direction class drops it to $\limsup \leq \frac{1}{2}$ (Proposition F.2). **(iii)** Two exact obstructions: a proof cannot restrict to selections with $\mu = 0$ (for the seventh-roots system $t_j = \zeta^j$, $z_j = \zeta^{3j}$, Parseval forces $C \geq \frac{3}{8}, \frac{2}{3}$ at supports $4, 3$, while unconstrained selections reach $< \frac{1}{4}$ and $\frac{2}{7}$). It also cannot rest on comparing aggregate second moments of original-weight subsets (for the fifth-roots system the aggregate quotient is $\frac{15}{23} > \frac{5}{8}$) (Propositions F.3–F.4). **(iv)** Closed classes:

Proposition 6.7 (Classes on which Conjecture 6.5 holds)**.** *The three-point inequality of Conjecture 6.5 holds for all moment systems with at most four atoms. The four-point inequality holds for all systems with at most five atoms. Both inequalities hold for all two-direction systems, and the four-point inequality holds for all antipodally paired systems (disjoint unions of pairs $\{(t,z),(-t,z)\}$ with equal z-values inside each pair).*

Moreover, by Lemma E.4 any counterexample to the three-point inequality must satisfy $5|z_i|^2 - 2L_i > 3E$ for every atom. The regular hexagon third-harmonic system has three-point value exactly $\frac{1}{2}$ and four-point value 0.

7 Discussion and Open Problems

The threshold splits conceptually as $(m+1)d = d + md$: d points to pin a feature basis and md points to cancel the $m \times d$ residual gradient. The lower-bound instances realize this as d independent copies of weighted mean estimation in $\mathbb{R}^m$, one per feature direction. Each requires its full simplex of $m+1$ points. At $m = 1$ the two mechanisms coincide. Table 2 in Appendix E separates the values of explicit instances from the global profile values at $(2,2)$. Our theorems concern a fixed learner and a fixed finite dataset. Certificates use the full-data minimizer W°, with arbitrary nonnegative real weights. Some intermediate bounds are attained only as limits. The theorems describe what a small weighted subsample can represent and its exact budget limits. They provide neither a faster selection algorithm nor guarantees for unseen data.

Open problems. (1) *The intermediate curve.* Determine $F_{\text{weighted}}(d,m,n)$ for $d, m \geq 2$ and $d < n < (m+1)d - 1$. Even the scalar case is largely open (Hanneke et al. (2025b), Question 2; see Zhang (2026) for the cells resolved so far). Axial instances reduce to an allocation problem over per-line mean-estimation deficits (Theorem G.1 supplies the sharp per-line constants). (2) *The $(2,2)$ cell* (Conjecture 6.5): a question about at most seven atoms on the circle. A proof must use selections with $\mu \neq 0$ and cannot rest on comparing aggregate second moments of original-weight subsets. (3) *Unweighted vector selection.* Relating the unweighted profile $F(d,m,n)$ of Question 3 of Hanneke et al. (2025b) to the weighted threshold $(m+1)d$. (4) *Other learning rules.* How does the threshold depend on the ERM tie-breaking rule? Our upper bounds use it only through Lemma 2.2, while the lower bounds use it in their rank-deficient branches.

7.1 Conclusion

We establish the exact threshold $n^\star(d,m) = (m+1)d$ for weighted data selection with the minimum-Frobenius-norm ERM to recover the full-data loss of every dataset. This answers the threshold part of Question 4 posed at COLT 2025 by Hanneke et al. (2025b). We also determine the profile at two further budgets, $F_{\text{weighted}}(d,m,(m+1)d-1) = 1 + 1/(dm^2)$ and $F_{\text{weighted}}(d,m,d) = d+1$. The proofs rest on three reusable tools: fixed-basis conic compression, determinant–facet rigidity of maximal certificates, and sharp sparsification of zero-mean weighted point systems. The proofs also use a complex-variable dictionary that reduces the smallest open cell to a seven-atom moment problem. The intermediate curve $d < n < (m+1)d - 1$ for $d, m \geq 2$ remains open. For $d = 1$ the complete curve is given by Corollary G.3. Even on the open curve the inner infimum in F_{weighted} need not be attained (Remark E.2: an integer instance with $d = m = 2$, $n = 3$). The first target is the $(2,2)$ conjecture $F_{\text{weighted}}(2,2,3) = 13/8$, $F_{\text{weighted}}(2,2,4) = 5/4$.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the theoretical model and conceptual framework, to assist in the writing of proofs, to design and provide feedback on the research methodology, to implement methods, to assist with translation, and to construct and check the explicit finite instances that serve as lower-bound and non-attainment examples, which are the only synthetic data in this paper; we also used them to help interpret the results. We have not used generative AI tools to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, or to propose or refine hypotheses; cleaning and reformatting datasets and supporting qualitative and thematic data analysis are not applicable to this work, which contains no empirical data. Additionally, we used generative AI tools to create and modify scientific figures, to create and edit software code, to draft parts of the paper, to summarize and analyse existing literature, to discover research topics and identify gaps, for brainstorming, for sourcing and searching for information, to edit the paper to improve readability, to identify relevant literature, to format references, to suggest a structure for the paper, and to propose a title and keywords. We have reviewed all AI-assisted work: every derivation and every piece of code produced with AI assistance was checked by hand by three people and found to be correct. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This is a theoretical paper on the sample complexity of data selection for linear regression. It involves no human subjects, personal data, or deployment of a system, and we are not aware of ethical concerns raised by its content.

REPRODUCIBILITY STATEMENT

All results are mathematical. Complete proofs of every theorem, lemma and proposition of Sections 2–6.2 are given in Appendices A–F. The general sparsification theorem is proved in Appendix G. Conjecture 6.5 remains open. Appendix F provides partial results and evidence for it. The explicit instances in Definition 3.2 and Proposition 6.3, along with the moment systems of Appendix F, are given in closed form with rational or algebraic coordinates. Every numerical value in the paper can therefore be recomputed by exact arithmetic from the formulas in the text. These include $43/40$, $23/20$, $13/8$, $5/4$, $1/2$, $4/7$, $2/7$, $15/23$, the hexagon four-point value 0, and the constants of Lemma 5.1 and Theorem G.1. We performed these exact computations as a supplementary check. No numerical search is used in the proof of any theorem.

REFERENCES

Dimitri P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Pahan Dewasurendra. One short of Steinitz: Weighted regression selection. Unrefereed preprint, Zenodo, 2026a. URL https://zenodo.org/records/21881535. DOI: 10.5281/zenodo.21881535. Scalar case ($m = 1$) only.

Pahan Dewasurendra. Sharp pair selection for mean regression. Unrefereed preprint, Zenodo, 2026b. URL https://zenodo.org/records/22061100. August 22, 2026. DOI: 10.5281/zenodo.22061100 (the Zenodo record lists the creator as Anonymous; the byline of the PDF is used here).

Sergei Drozdov. Egan conjecture holds. *Discrete Applied Mathematics*, 377:562–572, 2025. doi: 10.1016/j.dam.2025.08.034. arXiv:2310.10816.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga

Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112. Curran Associates, Inc., 2023. doi: 10.52202/075280-1179. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/56332d41d55ad7ad8024aac625881be7-Paper-Datasets_and_Benchmarks.pdf.

Steve Hanneke, Shay Moran, Alexander Shlimovich, and Amir Yehudayoff. Data selection for ERMs. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2634–2665, 2025a. arXiv:2504.14572.

Steve Hanneke, Shay Moran, Alexander Shlimovich, and Amir Yehudayoff. Open problem: Data selection for regression tasks. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 6225–6229, 2025b.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. arXiv:1906.04705.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Raphaël Tinarrage. Sharp approximate Carathéodory theorem and application to iterated Delaunay refinement. arXiv:2606.25854, 2026.

Guangjian Zhang. Exact risk ratios for weighted data selection in linear regression. arXiv:2608.28007, 2026.

A Proofs for Section 2

This appendix proves the structural lemmas of Section 2.

A.1 Row decomposition and minimum-norm geometry

Lemma A.1 (Row decomposition). *Write $w_j^\top$ for the j-th row of W and y_{ij} for the j-th entry of y_i. Then $F_c(W) = \sum_{j=1}^m \sum_i c_i(\langle w_j, x_i\rangle - y_{ij})^2$ and $\|W\|_F^2 = \sum_j \|w_j\|^2$. Hence the minimizer set of F_c is the Cartesian product over rows of scalar weighted least-squares solution sets. The model $A(F_c)$ consists of the minimal-ℓ_2-norm scalar solutions row by row,* with the single shared weight vector *c.*

Proof. Both the objective and the squared Frobenius norm are additive across rows, and a product set is minimized in norm coordinate-wise. □

Let $T = \operatorname{span}\{x_1, \dots, x_N\}$, $r = \dim T$, and let P_T denote the orthogonal projection onto T. Let W° be the minimal-Frobenius-norm full-data minimizer, and set $\rho_i = W^\circ x_i - y_i$ (residuals) and $M_i = \rho_i x_i^\top \in \mathbb{R}^{m\times d}$ (residual dyads).

Lemma A.2 (Full-data geometry). *$W^\circ = W^\circ P_T$. The first-order condition $\sum_{i=1}^N M_i = 0$ holds, and for every $H \in \mathbb{R}^{m\times d}$,*

$$L_D(W^\circ + H) = L_D^\star + \frac{1}{N}\sum_{i=1}^N \|Hx_i\|^2. \tag{5}$$

Consequently, the full-data minimizer set is $\{W^\circ + H : HP_T = 0\}$. If $T = \mathbb{R}^d$, the minimizer is unique.

Proof. Predictions depend on W only through WP_T, and $\|W\|_F^2 = \|WP_T\|_F^2 + \|WP_{T^\perp}\|_F^2$, so minimality of the norm forces $W^\circ P_{T^\perp} = 0$. The condition $\sum_i M_i = 0$ is the vanishing gradient of the smooth convex L_D at W°. Expanding $L_D(W^\circ + H)$, the cross term is $\frac{2}{N}\langle H, \sum_i M_i\rangle_F = 0$, giving (5). The excess term in (5) vanishes iff $Hx_i = 0$ for all i, i.e. $HP_T = 0$, which identifies the minimizer set. □

The two conditions in the next lemma have separate roles: the zero-dyad condition makes W° stationary for the selected objective, while the spanning condition makes the selected and full-data minimizer sets coincide.

Lemma 2.2 (Exact certificate; restated). *Let $c \in \mathbb{R}^N_{\geq 0}$ be nonzero, with support S (so $c_i > 0$ on S). If*

$$\sum_{i \in S} c_i M_i = 0 \qquad \textit{and} \qquad \operatorname{span}\{x_i : i \in S\} = T,$$

then $A(F_c) = W^\circ$. In particular, the selection realizes the full-data optimal loss, and the ratio is 1 (also when $L_D^\star = 0$, by the $0/0$ convention).

Proof. Using (2), for any H, $F_c(W^\circ + H) - F_c(W^\circ) = 2\langle H, \sum_{i\in S} c_i M_i\rangle_F + \sum_{i\in S} c_i \|Hx_i\|^2 = \sum_{i\in S} c_i \|Hx_i\|^2$, which vanishes iff $Hx_i = 0$ on S, iff $HP_T = 0$ because the support spans T. Thus the weighted minimizer set is $W^\circ + \{H : HP_T = 0\}$, the same set as the full-data minimizer set. Since $W^\circ = W^\circ P_T$ and $H = HP_{T^\perp}$ are Frobenius-orthogonal, the unique minimal-norm element is W°. □

Lemma A.3 (Scalar embedding). *Let $D_0 = ((x_i, t_i))_{i=1}^N$ be a scalar dataset ($x_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$) and let $\widetilde{D} = ((x_i, t_i \mathbf{1}_m))_{i=1}^N$ be its copy in which all m coordinates of every label agree. For every weight vector c, all rows of $A(F_c)$ computed on $\widetilde{D}$ equal the scalar minimum-norm weighted ERM of D_0 with the same weights. For every W whose rows all equal a common $w \in \mathbb{R}^d$, one has $L_{\widetilde{D}}(W) = m\, L_{D_0}(w)$. Consequently, $F_{\text{weighted}}(d, m, n) \geq F^{\text{sc}}_{\text{weighted}}(d, n)$ for all n, where $F^{\text{sc}}_{\text{weighted}}(d, n)$ denotes the scalar weighted profile of (Hanneke et al., 2025b).*

Proof. By Lemma A.1, every row of $A(F_c)$ is the minimum-norm solution of one and the same scalar weighted least-squares problem, so the rows agree. The identity $L_{\widetilde{D}}(W) = mL_{D_0}(w)$ is immediate from the definition of the loss. Applying this to the selected model and to the full-data minimizer (uniform weights), numerator and denominator of the ratio both scale by m, so every scalar instance embeds with the same ratio. □

B Proofs for Section 3 and Proposition 5.3

B.1 Upper bound: fixed-basis conic compression

Lemma B.1 (Fixed-basis conic compression). *Let $a_1, \ldots, a_N$ be elements of a finite-dimensional real vector space with $\sum_{i=1}^N a_i = 0$, and let $B \subseteq [N]$. Put $q_B = \dim \operatorname{span}\{a_i : i \notin B\}$. Then there exist coefficients $c_i \geq 0$ with $c_b > 0$ for all $b \in B$, $\sum_i c_i a_i = 0$, and $|\operatorname{supp} c| \leq |B| + q_B$.*

Proof. The global zero sum gives $-\sum_{b\in B} a_b = \sum_{i\notin B} a_i$, which is a nonnegative combination of $\{a_i\}_{i\notin B}$ and hence lies in $\operatorname{cone}\{a_i : i \notin B\}$. By conic Carathéodory (see e.g. (Bertsekas, 2009, Prop. 1.2.1(a))), every element of this cone is a positive combination of a *linearly independent* subfamily, so $-\sum_{b\in B} a_b = \sum_{i\in C} \alpha_i a_i$ with $\alpha_i > 0$, $C \subseteq [N] \setminus B$ and $|C| \leq q_B$. If the target is 0, take $C = \varnothing$. Set $c_b = 1$ on B, $c_i = \alpha_i$ on C, and 0 elsewhere. □

Theorem 3.1 (Data-dependent upper bound; restated). *Every dataset with feature rank $r \geq 1$ admits a spanning zero certificate of support at most $(m+1)r$. Hence $\tau(D) \leq (m+1)r$ and $L_D^\star\big((m+1)d; \text{weighted}\big) = L_D^\star$ for every dataset. If $r = 0$, a single point suffices.*

Proof. Choose $B \subseteq [N]$ with $|B| = r$ such that $\{x_b\}_{b\in B}$ is a basis of T; this is possible because the features lie in and span T. Apply Lemma B.1 to the residual dyads $a_i = M_i$, whose global sum vanishes by Lemma A.2. Every $M_i = \rho_i x_i^\top$ has all rows proportional to $x_i^\top \in T$, so $\operatorname{span}\{M_i : i \notin B\} \subseteq \mathbb{R}^m \otimes T$, and thus $q_B \le mr$. The resulting c is strictly positive on a set $S \supseteq B$ of size at most $r + mr = (m+1)r$ and satisfies $\sum_{i\in S} c_i M_i = 0$ and $\operatorname{span}\{x_i : i \in S\} \supseteq \operatorname{span}\{x_b : b \in B\} = T$. Since all features lie in T, the span equals T. Normalize c to sum 1; its support is nonempty. Lemma 2.2 then gives $A(F_c) = W^\circ$, so the selected model attains the optimal full-data loss. By Lemma 2.1, the support fits in $(m+1)d \ge (m+1)r$ slots. If $r = 0$, then all $x_i = 0$. In this case, every W has the same loss on every objective, and A returns $W = 0$ in both cases, so one point suffices. □

Remark B.2 (Effective residual dimension)**.** The proof used only that the residual dyads lie in $\mathbb{R}^m \otimes T$. If $s = \dim \operatorname{span}\{\rho_1, \dots, \rho_N\}$, they lie in the sr-dimensional space $\operatorname{span}\{\rho_i\} \otimes T$, and the same argument gives $\tau(D) \le r + sr = (s+1)r$ for $r \ge 1$. The budget needed to recover a given dataset exactly is governed by its effective residual output dimension $s \le m$, not by the nominal output dimension.

B.2 Lower bound: the axial simplex instance

Fix $d, m \ge 1$ and let $\mathbf{1}_m \in \mathbb{R}^m$ denote the all-ones vector,

$$v_j = e_j \ (1 \le j \le m), \qquad v_{m+1} = -\mathbf{1}_m,$$

so $\sum_{j=1}^{m+1} v_j = 0$. The unique linear dependence among $v_1, \dots, v_{m+1}$ has all coefficients equal. In particular, no proper subset of $\{v_j\}$ admits a nonzero nonnegative zero-sum; equivalently, $0 \notin \operatorname{conv}$ of any proper subset.

Definition 3.2 (Integer axial instance; restated). *$D_{d,m}$ consists of the $N = (m+1)d$ points $z_{ij} = (e_i,\ \mathbf{1}_m - v_j)$ for $i \in [d]$, $j \in [m+1]$. Explicitly, the label of z_{ij} is $\mathbf{1}_m - e_j$ for $j \le m$ and $2 \cdot \mathbf{1}_m$ for $j = m+1$.*

Lemma B.3. *For $D_{d,m}$, the unique full-data minimizer is $W^\circ = \mathbf{1}_m \mathbf{1}_d^\top$, with residual v_j at z_{ij} and $L_D^\star = \frac{2m}{m+1}$. Moreover, for any weight vector c with per-axis totals $t_i = \sum_j c_{ij}$, the returned model $\widehat{W} = A(F_c)$ has the following form: column i of $\widehat{W}$ equals the conditional weighted label mean $\mathbf{1}_m - a_i$ with $a_i = \sum_j \frac{c_{ij}}{t_i} v_j$ when $t_i > 0$, and equals 0 when $t_i = 0$. Writing $\delta_i = \widehat{W} e_i - \mathbf{1}_m$, we have*

$$L_D(\widehat{W}) = L_D^\star + \frac{1}{d}\sum_{i=1}^{d} \|\delta_i\|^2. \tag{6}$$

Proof. The full Gram matrix is $\sum_{ij} e_i e_i^\top = (m+1)I \succ 0$ and $\sum_{ij} v_j e_i^\top = 0$, so W° is the unique minimizer, with $L_D^\star = \frac{d\sum_j \|v_j\|^2}{N} = \frac{2md}{(m+1)d}$. Since the features are standard basis vectors, the weighted normal equations decouple across columns. A positive-weight column returns the conditional mean, and a zero-weight column is unconstrained. The Frobenius rule sets the latter to zero because the norm decomposes column-wise. For (6), expand $\|\delta_i + v_j\|^2$ and use $\sum_j v_j = 0$: $\sum_j \|v_j + \delta_i\|^2 = 2m + (m+1)\|\delta_i\|^2$. □

Lemma B.4 (Facet distances)**.** *Let $Q_m = m^2 + m - 1$. The squared distance from the origin to the convex hull of any nonempty proper subset of $\{v_1, \dots, v_{m+1}\}$ is at least $1/Q_m$. The minimum is attained on the facets omitting some e_k, while the facet omitting v_{m+1} has squared distance $1/m$.*

Proof. Every proper subset is contained in a facet. For the facet $\operatorname{conv}\{e_1, \dots, e_m\}$, the nearest point is $\mathbf{1}_m/m$, of squared norm $1/m$. For the facet omitting e_k, let n_k have $-m$ in coordinate k and 1 elsewhere. The facet lies on the hyperplane $n_k^\top x = 1$ and $\|n_k\|^2 = Q_m$, so the distance to the hyperplane is $1/\sqrt{Q_m}$. The foot n_k/Q_m lies in the facet because $n_k/Q_m = \frac{m}{Q_m}(-\mathbf{1}_m) + \sum_{j\ne k} \frac{m+1}{Q_m} e_j$ is a convex combination. Finally, $Q_m \ge m$ with equality only at $m = 1$, where both distances equal $1 = 1/Q_1$. □

Theorem 3.3 (Exact deficit of the axial instance; restated). *For every $d, m \geq 1$,*

$$\frac{L^\star_{D_{d,m}}\big((m+1)d-1;\,\text{weighted}\big)}{L^\star_{D_{d,m}}} \;=\; 1+\frac{m+1}{2md\,(m^2+m-1)} \;>\; 1.$$

Proof. A selection of support at most $N-1$ either misses some axis entirely, giving $\|\delta_i\|^2 = \|\mathbf{1}_m\|^2 = m \geq 1/Q_m$, or uses a proper subset of the $m+1$ points of some axis, in which case $\delta_i = -a_i$ lies in minus the convex hull of a proper subset of $\{v_j\}$ and $\|\delta_i\|^2 \geq 1/Q_m$ by Lemma B.4. With (6), every such selection has $L_D \geq L^\star_D + \frac{1}{dQ_m}$. Conversely, omit $z_{1,1}$ and give axis 1 the conditional weights of the foot n_1/Q_m: put $\frac{m}{Q_m}$ on v_{m+1} and $\frac{m+1}{Q_m}$ on each v_j, $2 \leq j \leq m$, and scale by $t_1 = \frac{1}{d}$. Give every other axis uniform weights. The support is exactly $N-1$ and $L_D = L^\star_D + \frac{1}{dQ_m}$. Dividing by $L^\star_D = \frac{2m}{m+1}$ gives the claim. □

Theorem 3.4 (Threshold; restated). *$n^\star(d,m) = (m+1)d$ for all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4), and $F_{\text{weighted}}(d,m,n) = 1$ for all $n \geq (m+1)d$.*

Proof. Theorem 3.1 gives ratio 1 at budget $(m+1)d$ for every dataset, and Theorem 3.3 exhibits a dataset with ratio > 1 at budget $(m+1)d-1$. Since $F_{\text{weighted}}(d,m,\cdot)$ is nonincreasing (Lemma 2.1), the value is 1 at every budget $n \geq (m+1)d$ and exceeds 1 at every budget $n \leq (m+1)d-1$. □

Corollary 3.5 (restated). *$n^\star(2,2) = 6$. On the instance $D_{2,2}$ (six integer points, labels $(0,1),(1,0),(2,2)$ over each of e_1, e_2) the optimal five-point ratio is exactly $\frac{43}{40}$.*

We emphasize that $43/40$ is the exact optimum *of this instance*. The worst case over all datasets at budget 5 is $F_{\text{weighted}}(2,2,5) = 1 + \frac{1}{8} = \frac{9}{8}$ by Theorem 5.2.

B.3 Failure at $n = 2d$ for $m \geq 2$

The scalar rule "$n \geq 2d$ suffices" does not extend to vector outputs (Remark 5.4).

Proposition 5.3 (restated). *For all $d \geq 1, m \geq 2$, $F_{\text{weighted}}(d,m,2d) \geq F_{\text{weighted}}\big(d,m,(m+1)d-1\big) = 1+\frac{1}{dm^2} > 1$, since $2d \leq (m+1)d-1$ and F_{weighted} is nonincreasing in n. Concretely, for $d = m = 2$ the instance $D_{2,2}$ of Corollary 3.5 satisfies: every weighted selection of at most $2d = 4$ points has*

$$L_D(A(F)) \;\geq\; \frac{23}{20}\cdot L^\star_D,$$

with equality attained (allocate two points per axis, at conditional weights realizing the nearest facet points).

Proof. The monotonicity statement is immediate from the definitions. For $D_{2,2}$, (6) and per-axis decoupling show that a budget-4 selection allocates (k_1,k_2) points to the two axes with $k_1+k_2 \leq 4$. The minimal per-axis costs are 0 (all three points), $1/5$ (best two points, the facet $\{e_2, -\mathbf{1}_m\}$ or $\{e_1, -\mathbf{1}_m\}$ at squared distance $1/Q_2 = 1/5$), 1 (best single point), and 2 (empty axis). The minimum total cost over allocations is $2 \cdot \frac{1}{5}$ at $(2,2)$, giving $L_D = \frac{4}{3} + \frac{1}{2}\cdot\frac{2}{5} = \frac{23}{15}$ and ratio $\frac{23/15}{4/3} = \frac{23}{20}$. □

C Proof of Theorem 6.1

Theorem 6.1 (restated). *For all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4),*

$$F_{\text{weighted}}(d,m,n) = \infty \text{ for every } n < d, \qquad \text{and} \qquad F_{\text{weighted}}(d,m,d) = d+1.$$

The lower bounds are inherited from the scalar case through Lemma A.3: by (Hanneke et al., 2025b, Theorem 1), $F^{\mathrm{sc}}_{\mathrm{weighted}}(d,n)=\infty$ for $n<d$ and $F^{\mathrm{sc}}_{\mathrm{weighted}}(d,d)=d+1$, so $F_{\mathrm{weighted}}(d,m,n)\geq F^{\mathrm{sc}}_{\mathrm{weighted}}(d,n)$. The content of this section is the upper bound $F_{\mathrm{weighted}}(d,m,d)\leq d+1$ *uniformly in* m, which requires selecting one subset that serves all m output rows simultaneously.

Volume sampling. Let $X\in\mathbb{R}^{d\times N}$ have full row rank and let $y\in\mathbb{R}^N$. Size-d volume sampling draws a subset S of d column indices with probability proportional to $\det(X_S)^2$. Let $w(S)$ denote the least-squares solution of the subproblem (X_S,y_S) (the interpolant when X_S is invertible). Dereziński & Warmuth (2017) proved two results. The first is unbiasedness: $\mathbb{E}[w(S)]=w^\star$ for every label vector, requiring only full row rank (their Theorem 3 and Proposition 7). The second bounds the total square loss L on the full data:

$$\mathbb{E}\big[L(w(S))\big]\;\leq\;(d+1)\,L(w^\star),\tag{7}$$

with *equality* when X is in general position (every d-column submatrix nonsingular), as stated in their Theorem 5. We only use the inequality (7), which holds without general position. Equality in (7) can fail for degenerate X (e.g. $x_1=x_2=e_1$, $x_3=x_4=e_2$, $y=(1,-1,1,-1)$ has $\mathbb{E}[L(w(S))]=2L(w^\star)$). Dividing by N converts (7) to average losses.

Proof of Theorem 6.1. Only $F_{\mathrm{weighted}}(d,m,d)\leq d+1$ remains. Fix a dataset and let $r=\mathrm{rank}$ of the features. If $r=0$, the ratio is 1 because both rules return $W=0$. If $L^\star_D=0$, choose r points whose features form a basis of T with positive weights. The weighted problem then has minimum 0 and minimizer set $W^\circ+\{H:HP_T=0\}$. The Frobenius rule returns W°, so the ratio is 1 by the $0/0$ convention.

Otherwise, choose an orthonormal basis $U\in\mathbb{R}^{d\times r}$ of T and write $x_i=U\xi_i$ with $\Xi=[\xi_1\cdots\xi_N]\in\mathbb{R}^{r\times N}$ of full row rank. Run size-r volume sampling on Ξ. Crucially, *the distribution over subsets depends only on the features*. For the q-th output row, (7) (in dimension r) gives $\mathbb{E}\,L^{(q)}_D(w_{q,S})\leq(r+1)L^{(q),\star}_D$. Summing over the m rows and using Lemma A.1,

$$\mathbb{E}\big[L_D(W_S)\big]\;\leq\;(r+1)\,L^\star_D\;\leq\;(d+1)\,L^\star_D,$$

where $W_S=\Theta_S U^\top$ is the matrix whose rows are the per-row subproblem solutions in Ξ-coordinates with the $T^\perp$-component set to zero. Hence some subset S (of $r\leq d$ points) achieves $L_D(W_S)\leq(d+1)L^\star_D$. This W_S is realizable by a legal selection: the r chosen features are linearly independent in T. Thus, with any positive weights, the weighted ERM set consists of all interpolants of the r selected points. The Frobenius rule returns exactly W_S (the interpolant with vanishing $T^\perp$-component). Padding to d slots by Lemma 2.1 completes the proof. □

D Proofs for Sections 4 and 5

Throughout this section, fix a dataset with residual dyads $M_i=\rho_i x_i^\top$ and recall the minimal certificate size $\tau(D)$ from (3). Theorem 3.1 gives $\tau(D)\leq\max\{1,(m+1)r\}$. The upper bound at budget $(m+1)d-1$ requires understanding the datasets with $\tau(D)=(m+1)d$, which we call *certificate-maximal*. The key is a rigidity theorem for maximal certificates.

D.1 Rigidity of maximal certificates

Theorem 4.1 (Determinant–facet rigidity; restated). *Let the features span $\mathbb{R}^d$ and suppose $\tau(D)=k:=(m+1)d$. Let c be a spanning zero certificate with $|\operatorname{supp}c|=k$ and support S. Then S splits uniquely (up to permutation) into d pairwise disjoint blocks $S=C_1\sqcup\cdots\sqcup C_d$ with $|C_j|=m+1$, where each C_j supports a positive circuit of the dyads whose features span a one-dimensional line U_j, and $U_1,\ldots,U_d$ are linearly independent.*

Proof. The proof has four steps. We first identify the normalized nonnegative kernel as a polytope, then use the determinant polynomial to bound its number of facets. This shows that the polytope is a simplex. Finally, we read off line-supported circuits from the extreme rays. Vectorize the dyads into

$\mathsf{A} = [\mathrm{vec}(M_i)]_{i\in S} \in \mathbb{R}^{md\times k}$ and let $K = \ker \mathsf{A} \cap \mathbb{R}^k_{\geq 0}$, $h = \dim \ker \mathsf{A} \geq k - md = d$. Since $c > 0$ on S and lies in $\ker \mathsf{A}$, a neighborhood of c in $\ker \mathsf{A}$ remains nonnegative. Thus K spans $\ker \mathsf{A}$, and the compact section $P = K \cap \{\mathbf{1}^\top \theta = 1\}$ is a polytope of dimension $h - 1$ whose relative interior consists exactly of the strictly positive points.

Consider $f(\theta) = \det\big(\sum_{i\in S} \theta_i x_i x_i^\top\big)$, a polynomial of degree at most d. On $\mathrm{relint}\, P$, all weights are positive and the features span $\mathbb{R}^d$, so $f > 0$. At any boundary point, some coordinate vanishes. If $f > 0$ there, the positively weighted features would span $\mathbb{R}^d$ and the point would be a spanning zero certificate of support $\leq k - 1$, contradicting $\tau = k$. Hence $f \equiv 0$ on every facet of P. Parametrizing $\mathrm{aff}\, P$ by $\mathbb{R}^{h-1}$, the restriction g of f is a nonzero polynomial of degree $\leq d$. It vanishes on the relative interior of each facet and is therefore divisible by the pairwise non-associated affine linear forms of the distinct facet hyperplanes. Thus P has at most d facets. If $h \geq 2$, then P has positive dimension and, being bounded, at least h facets, so $h \leq d$. If $h = 1$, then $h \leq d$ trivially. Together with $h \geq d$, this gives $h = d$. Thus, for $d \geq 2$, P is a $(d-1)$-dimensional polytope with exactly d facets and hence a $(d-1)$-simplex. For $d = 1$, P is a single point. In both cases, K is a simplicial cone with extreme rays $v^{(1)}, \ldots, v^{(d)}$.

Let $U_j = \mathrm{span}\{x_i : v_i^{(j)} > 0\}$. As c is a strictly positive combination of all rays, the supports of the rays cover S, so $\sum_j U_j = \mathbb{R}^d$. If $d = 1$, there is a single ray and $U_1 = \mathbb{R}$, so $\dim U_1 = 1$. Now let $d \geq 2$. Dropping ray j and normalizing the sum of the remaining rays gives a boundary point of P, where the features cannot span, so $\sum_{\ell\neq j} U_\ell \neq \mathbb{R}^d$ for each j: the family $\{U_j\}$ is a minimal spanning family. Choose functionals ϕ_j vanishing on $\sum_{\ell\neq j} U_\ell$ and not on U_j, and normalize them against vectors $u_j \in U_j$ with $\phi_j(u_j) = 1$. Then (ϕ_j) is a dual basis, and $U_j \subseteq \bigcap_{i\neq j} \ker \phi_i$, a one-dimensional space, so $\dim U_j = 1$ in every dimension $d \geq 1$.

Each extreme ray of the cone $\{\theta \geq 0 : \mathsf{A}\theta = 0\}$ has inclusion-minimal support C_j, and $\dim \ker \mathsf{A}_{C_j} = 1$. Otherwise, a two-sided perturbation along a second kernel direction would split the ray. All dyads of C_j lie in $\mathbb{R}^m \otimes U_j$, a space of dimension m, so $|C_j| = \mathrm{rank}\, \mathsf{A}_{C_j} + 1 \leq m + 1$. Finally, $k = |S| \leq \sum_j |C_j| \leq d(m+1) = k$ forces every $|C_j| = m + 1$ and pairwise disjoint supports. The decomposition is unique because the extreme rays of K are. □

D.2 Certificate-maximal datasets live on d lines

Theorem 4.1 describes one maximal certificate. We now use the maximality of $\tau(D)$ to show that the same line-circuit structure governs every nonzero-feature point of the dataset.

Theorem 4.2 (Line structure; restated). *Let the features span $\mathbb{R}^d$ and $\tau(D) = (m+1)d = k$. Then:*

1. *every data point with $x_i \neq 0$ belongs to some positive circuit of dyads supported on a single feature line and of size exactly $m + 1$;*
2. *there is no positive circuit of dyads whose features span a subspace of dimension ≥ 2, and no line circuit of size $\leq m$; in particular no point has $x_i \neq 0$ and $\rho_i = 0$;*
3. *the nonzero features lie on exactly d linearly independent lines.*

Consequently, any dataset whose nonzero features are not *of this d-line form satisfies $\tau(D) \leq (m+1)d - 1$ and is recovered exactly within budget $(m+1)d - 1$.*

Proof. (1) Extend x_t, with t fixed and $x_t \neq 0$, to a feature basis $B \ni t$ and apply Lemma B.1: the resulting certificate has support at most k and, since $\tau = k$, exactly k. Theorem 4.1 splits it into $(m+1)$-point line circuits, one of which contains t.

(2) Let C be a positive circuit whose features span an s-dimensional space with $s \geq 2$. Then $|C| \leq ms + 1$ because its dyads lie in an ms-dimensional space. Extend the features of C by $d - s$ data points to a basis of $\mathbb{R}^d$, and attach to each added point its $(m+1)$-point line circuit from (1). Summing all these positive relations gives a spanning zero certificate of support at most $ms + 1 + (d-s)(m+1) = k - s + 1 \leq k - 1$, contradicting $\tau = k$. A line circuit of size $\leq m$

similarly completes with $d-1$ line circuits to support $\leq m+(d-1)(m+1) = k-1$. A point with $x_i \neq 0 = \rho_i$ is by itself a 1-point line "circuit" ($M_i = 0$) and is excluded the same way.

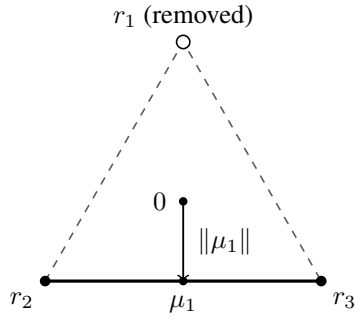
(3) Suppose $d+1$ distinct nonzero feature lines exist. Take an $(m+1)$-point line circuit on each using (1), and let J be their disjoint union, $|J| = (d+1)(m+1)$. Then $\dim \ker \mathsf{A}_J \geq |J| - md = d+m+1$. The sum of the circuits' positive kernel vectors is strictly positive on J, so the nonnegative kernel cone spans $\ker \mathsf{A}_J$. But each extreme ray of that cone is a positive circuit, hence by (2) supported on a single line and of size $m+1$. On each line, the kernel of the corresponding block is one-dimensional, so there are at most $d+1$ extreme rays, spanning at most $d+1$ dimensions. This contradicts $d+m+1 > d+1$. As the features span $\mathbb{R}^d$, there are exactly d independent lines. □

D.3 THE m-POINT MEAN LEMMA

For a positive $(m+1)$-point circuit, deleting one vertex leaves an m-point mean. The lemma shows that some deletion moves the mean by at most a $1/m$ fraction of the root-mean-square radius.

Lemma 5.1 (m-point mean lemma; restated). *Let $z_i \in \mathbb{R}^m$ carry weights $p_i > 0$ with $\sum_i p_i = 1$ and $\sum_i p_i z_i = 0$, and let $E = \sum_i p_i \|z_i\|^2$. Then some convex combination δ of at most m of the points satisfies $\|\delta\|^2 \leq E/m^2$. The constant is attained by the uniformly weighted regular simplex.*

Proof. By conic Carathéodory, decompose p inside the cone $\{q \geq 0 : \sum_i q_i z_i = 0\}$ as a positive combination of positive circuits (extreme rays), each of support at most $m+1$, and normalize each circuit to a probability vector $q^{(\alpha)}$. Thus $p = \sum_\alpha \lambda_\alpha q^{(\alpha)}$ with $\lambda_\alpha > 0$ and $\sum_\alpha \lambda_\alpha = \sum_i p_i = 1$. The second moments $E_{q^{(\alpha)}}$ then average to E, so some circuit q has $E_q \leq E$. If $E_q = 0$, all its points vanish and a single point gives $\delta = 0$. If its support has size $\leq m$, its own zero mean is the desired $\delta = 0$. Otherwise, the support has weights $q_1, \ldots, q_{m+1} > 0$ with $\sum_j q_j z_j = 0$. Deleting point j leaves the convex combination $\mu_j = -\frac{q_j}{1-q_j} z_j$ of the other m points. With $t_j = q_j \|z_j\|^2 / E_q$ (so $\sum_j t_j = 1$), $\|\mu_j\|^2 / E_q = q_j t_j / (1-q_j)^2$. The Cauchy–Schwarz inequality gives $\sum_j \frac{(1-q_j)^2}{q_j} = \sum_j \frac{1}{q_j} - 2(m+1) + 1 \geq (m+1)^2 - 2(m+1) + 1 = m^2$. Hence not all j can violate $\|\mu_j\|^2 \leq E_q/m^2$, for otherwise summing $t_j > \frac{(1-q_j)^2}{m^2 q_j}$ over j yields $1 > 1$. Finally, $E_q \leq E$. For tightness, consider the uniformly weighted regular simplex with vertices $r_1, \ldots, r_{m+1}$, so $E = \|r_j\|^2$ for every j. Every set of at most m vertices lies in a facet. The point of the facet opposite r_j closest to the origin is its centroid $-r_j/m$, and $\|r_j/m\|^2 = E/m^2$. Hence no convex combination of at most m vertices has smaller squared norm, and the deletions μ_j attain it. □



$m = 2$, uniform original weights $p_i = 1/3$
$r_1 + r_2 + r_3 = 0, \quad E = \|r_j\|^2$
Retain r_2, r_3 with weights $1/2, 1/2$:
$\mu_1 = \dfrac{r_2 + r_3}{2} = -\dfrac{r_1}{2}, \quad \|\mu_1\|^2 = \dfrac{E}{4}.$
Centered regular m-simplex,
uniformly weighted:
$\mu_j = -\dfrac{r_j}{m}, \quad \|\mu_j\|^2 = \dfrac{E}{m^2}.$

Figure 1: Regular-simplex equality in Lemma 5.1. The diagram lies in the mean-point space. The midpoint of the retained edge is its nearest point to the origin, giving the squared error $E/4$ for $m = 2$. The opposite-facet centroid gives E/m^2 for a centered, uniformly weighted regular m-simplex. The general upper bound is proved in Appendix D.3.

D.4 ASSEMBLY

The budget deficit is exactly one point. We therefore fit all but one feature line exactly and spend only m points on the line with the smallest residual energy.

Theorem 5.2 (Near-threshold value; restated). *For all $d, m \geq 2$ (and in fact for all $d, m \geq 1$, cf. Remark 2.4), $F_{\text{weighted}}\big(d, m, (m+1)d-1\big) = 1 + \dfrac{1}{dm^2}$.*

Proof. Upper bound. Let $n_0 = (m+1)d - 1$ and fix a dataset. If the feature rank is $r < d$, Theorem 3.1 gives a certificate of support at most $\max\{1, (m+1)r\} \leq n_0$, with one point when $r = 0$, and the ratio is 1. The same holds if $L_D^\star = 0$, by basis selection as in Section 6.1, or if $\tau(D) \leq n_0$. In the remaining certificate-maximal case $\tau(D) = (m+1)d$, Theorem 4.2 provides a basis $u_1, \ldots, u_d$ and scalars $s_i \neq 0$ with $x_i = s_i u_{j(i)}$ for every nonzero-feature point. Reading the first-order condition $\sum_i M_i = 0$ against the basis (u_j), each line satisfies $\sum_{i \in I_j} s_i \rho_i = 0$. Set, per line,

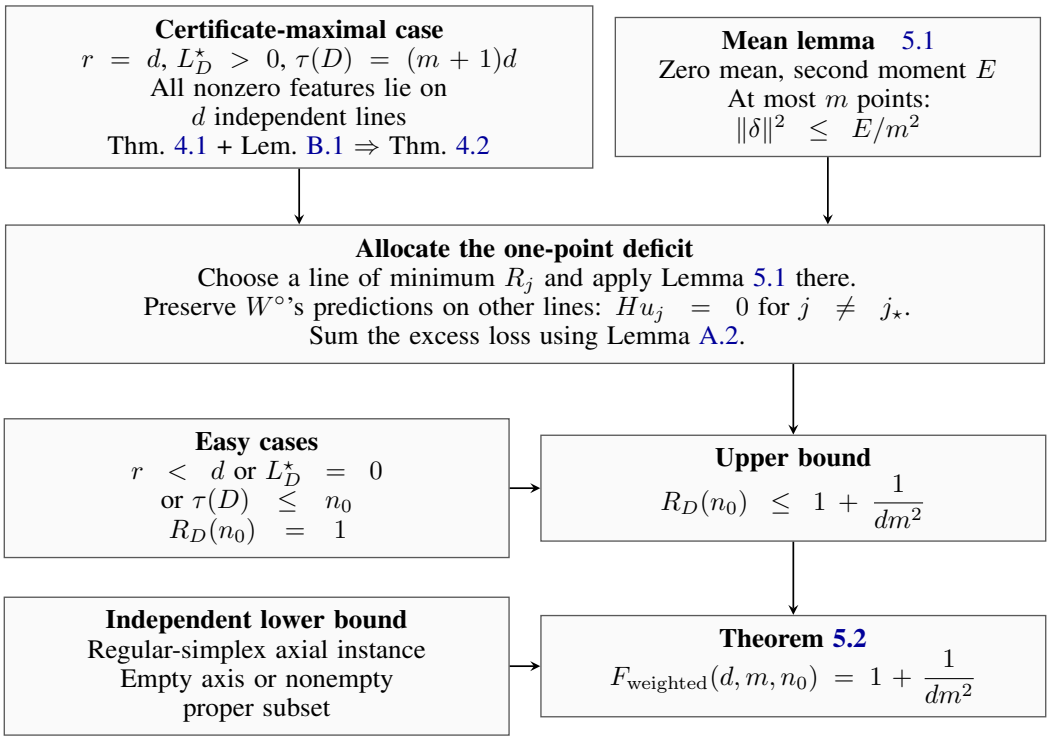
$$z_i = \rho_i / s_i, \qquad p_i = s_i^2 / g_j, \qquad g_j = \sum_{i \in I_j} s_i^2, \qquad R_j = \sum_{i \in I_j} \|\rho_i\|^2,$$

so that $\sum_{i \in I_j} p_i z_i = 0$ and $\sum_i p_i \|z_i\|^2 = R_j / g_j$. Choose the lightest line $j_\star$ with $R_{j_\star} \leq \frac{1}{d} \sum_j R_j \leq \frac{N L_D^\star}{d}$ (features equal to 0 only increase $L_D^\star$). On each line $j \neq j_\star$, the probability vector $(p_i)_{i \in I_j}$ is a nonzero element of the pointed cone $\{\theta \geq 0 : \sum_{i \in I_j} \theta_i z_i = 0\}$. Decompose it into extreme rays, take one of them, and normalize it to coordinate sum one. This gives a positive zero-mean combination of the z_i supported on a positive circuit, hence on at most $m+1$ points (Definition 2.3). On line $j_\star$ pick at most m points and a convex combination δ with $\|\delta\|^2 \leq \frac{R_{j_\star}}{m^2 g_{j_\star}}$ (Lemma 5.1). Translate the per-line convex coefficients α_i into selection weights $c_i = \alpha_i / s_i^2 > 0$, rescaling each line independently and normalizing globally. The weighted normal equations then decouple along the basis and give $\widehat{W} = W^\circ + H$ with $H u_j = 0$ for $j \neq j_\star$ and $H u_{j_\star} = -\delta$. The selected features contain all d lines, so the weighted Gram matrix is positive definite and the minimizer is unique. The support has size at most $(d-1)(m+1) + m = n_0$. Pad the selection by Lemma 2.1 if fewer slots are used. By (5),

$$L_D(\widehat{W}) - L_D^\star = \frac{g_{j_\star} \|\delta\|^2}{N} \leq \frac{R_{j_\star}}{N m^2} \leq \frac{L_D^\star}{dm^2}.$$

Lower bound. Take the regular-simplex axial instance: let $r_1, \ldots, r_{m+1} \in \mathbb{R}^m$ be regular simplex vertices with $\sum_j r_j = 0$ and $\|r_j\|^2 = \frac{m}{m+1}$, whose facets are at squared distance $h^2 = \frac{1}{m(m+1)}$ from the origin. Put $w = h e_1$ and $D = \{(e_i,\, w + r_j)\}_{i \in [d], j \in [m+1]}$. The unique full minimizer is $W^\circ = [w \;\cdots\; w]$ with $L_D^\star = \frac{m}{m+1}$. For any selection of support at most $N-1$, some axis is either empty, in which case the Frobenius rule sets its column to zero with deviation $\|w\|^2 = h^2$, or represented by a proper subset, with deviation at least the facet distance h^2. By the analogue of (6), the loss exceeds $L_D^\star$ by at least h^2/d, and deleting one vertex with uniform facet weights attains it. The ratio is $1 + \frac{h^2/d}{m/(m+1)} = 1 + \frac{1}{dm^2}$. □

Remark D.1 (General centered-simplex families)**.** The same computation applies to any centered simplex $r_1, \ldots, r_{m+1}$ ($\sum_j r_j = 0$, $S = \sum_j \|r_j\|^2$, $h = \min_j \operatorname{dist}(0, \operatorname{conv}\{r_k\}_{k \neq j})$), provided the columns of W° are chosen with $\|W^\circ e_i\| \geq h$: the exact $(N-1)$-point ratio is $1 + \frac{(m+1)h^2}{dS}$. Within this family, $h^2 \leq S/(m^2(m+1))$ with equality exactly for regular simplices. Indeed, the centroid of the facet opposite r_j is $-r_j/m$, so $h^2 \leq \min_j \|r_j\|^2/m^2 \leq S/(m^2(m+1))$. Equality forces all $\|r_j\|$ to be equal and each facet centroid to be the point of that facet nearest the origin, i.e. $\langle r_j, r_k \rangle$ independent of $k \neq j$. Thus, in the equality case, all pairwise inner products coincide and the simplex is regular. For $m \geq 2$, the integer instance of Definition 3.2 realizes the slightly smaller rational value of Theorem 3.3 with integer data. For $m = 1$, the two values coincide.



$$|\operatorname{supp} c| \le (d-1)(m+1) + m = n_0, \qquad n_0 = (m+1)d - 1.$$

$$L_D(\widehat{W}) - L_D^\star \le \frac{R_{j_\star}}{Nm^2} \le \frac{L_D^\star}{dm^2}.$$

Figure 2: Local proof assembly for Theorem 5.2. The upper bound combines the easy cases with the certificate-maximal case. The mean lemma is an independent input. The regular-simplex axial construction supplies the matching lower bound. The arrows summarize the assembly in Appendix D.4. The structural ingredients are proved in Appendices D.1–D.3.

E THE COMPLEX DICTIONARY AND PROOFS FOR SECTION 6.2

For $d, m \ge 2$ and $d < n < (m+1)d - 1$, the profile $F_{\text{weighted}}(d, m, n)$ remains open. For $d = 1$, it is given by Corollary G.3. The corresponding scalar regime also remains open, apart from the cells settled in (Zhang, 2026): it is the intermediate regime of Question 2 of (Hanneke et al., 2025b). In this section we treat the smallest cell $(d, m) = (2, 2)$, where the open budgets are $n = 3, 4$. We prove the interval theorem (Theorem 6.4) and reduce the conjectured exact values to a finite moment problem. The evidence for the conjecture is collected in Appendix F.

E.1 LOWER BOUNDS: AN EXPLICIT AXIAL INSTANCE

Proposition 6.3 (restated). *There is an explicit dataset $D^\triangle$ with*

$$\frac{L_{D^\triangle}^\star(3;\text{weighted})}{L_{D^\triangle}^\star} = \frac{13}{8}, \qquad \frac{L_{D^\triangle}^\star(4;\text{weighted})}{L_{D^\triangle}^\star} = \frac{5}{4}.$$

Hence $F_{\text{weighted}}(2,2,3) \ge \frac{13}{8}$ and $F_{\text{weighted}}(2,2,4) \ge \frac{5}{4}$.

Proof. Let $r_1, r_2, r_3 \in \mathbb{R}^2$ be the vertices of a centered equilateral triangle with $\|r_j\|^2 = \frac{2}{3}$. For instance, take $r_1 = \sqrt{\frac{2}{3}}\,(1,0)$ and $r_{2,3} = \sqrt{\frac{2}{3}}\,(-\frac{1}{2}, \pm\frac{\sqrt{3}}{2})$. The total is then $S = 2$ and the squared distance from the origin to an edge midpoint is $\frac{1}{6}$. Let $w = \sqrt{\frac{5}{6}}\,(1,0)$, so $\|w\|^2 = \frac{5}{6}$. Take

$D^{\triangle} = \{(e_i,\, w + r_j)\}_{i\in[2],j\in[3]}$. As in Lemma B.3, $W^{\circ} = [w\ w]$ and $L_D^{\star} = \frac{2}{3}$. The same calculation shows that a selection allocating k_i points to axis i incurs excess $\frac{1}{2}(c(k_1) + c(k_2))$ where $c(3) = 0$, $c(2) = \frac{1}{6}$ (best edge), $c(1) = \frac{2}{3}$ (best vertex), $c(0) = \|w\|^2 = \frac{5}{6}$ (empty axis). For budget 3 the minimum over allocations is $c(2) + c(1) = c(3) + c(0) = \frac{5}{6}$, giving ratio $1 + \frac{5/12}{2/3} = \frac{13}{8}$. For budget 4, the minimum is $2c(2) = \frac{1}{3}$, giving $1 + \frac{1/6}{2/3} = \frac{5}{4}$. □

E.2 A COMPLEX DICTIONARY FOR THE HARD BRANCHES

We use three kinds of weights below: selection weights c on data points (Lemma 2.1), full-data atom weights p determined by the features, and selected atom weights q. The correspondence between c and q is given after the definitions.

Fix a $(2, 2)$ dataset with feature rank 2 and $L_D^{\star} > 0$. We only consider selections whose features span $\mathbb{R}^2$. For these selections, the weighted minimizer is unique, so Frobenius tie-breaking plays no role. Their predictions and full-data loss are invariant under an invertible linear change of features $x \mapsto Gx$, with $W \mapsto WG^{-1}$. For a rank-deficient selected objective, such a change need not preserve the minimum-Frobenius-norm tie-breaking. This is why we restrict to spanning selections. Since our aim here is to exhibit good spanning selections, we may assume that the data are whitened: $\sum_i x_i x_i^{\top} = \gamma I_2$. Identify $\mathbb{R}^2 \cong \mathbb{C}$, and write ξ_i for the feature and η_i for the residual ρ_i of point i, viewed in $\mathbb{C}$. Let $I = \{i : \xi_i \neq 0\}$, $S = \sum_{j\in I} |\xi_j|^2$, $R_0 = \sum_{i\notin I} |\eta_i|^2$, and for $i \in I$ set

$$p_i = \frac{|\xi_i|^2}{S}, \qquad t_i = \frac{\bar{\xi}_i}{\xi_i} \in S^1, \qquad z_i = \frac{\eta_i}{\xi_i} \in \mathbb{C}, \qquad E = \sum_{i\in I} p_i |z_i|^2.$$

Whitening is equivalent to $\sum_i \xi_i^2 = 0$, that is, $\mathbb{E}_p t = 0$. The first-order condition $\sum_i \rho_i x_i^{\top} = 0$ is equivalent to the pair $\mathbb{E}_p z = 0$, $\mathbb{E}_p(\bar{t}z) = 0$. A selection with weights c corresponds to $q_i \propto c_i|\xi_i|^2$ on I; zero-feature points do not influence the selected minimizer. Writing $\mu = \mathbb{E}_q t$, $\nu = \mathbb{E}_q z$, $\lambda = \mathbb{E}_q(\bar{t}z)$, represent $W^{\circ} - \widehat{W}$ as the real-linear map $\xi \mapsto a\xi + b\bar{\xi}$, so that the residual of point i under $\widehat{W}$ is $\eta_i - (a\xi_i + b\bar{\xi}_i) = \xi_i(z_i - a - bt_i)$. The weighted normal equations then read

$$\begin{pmatrix} 1 & \mu \\ \bar{\mu} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \nu \\ \lambda \end{pmatrix}, \qquad |\mu| < 1 \iff \text{selected features span } \mathbb{R}^2, \tag{8}$$

and, with $C(q) = |a|^2 + |b|^2$, the excess loss satisfies the following bound, using $\mathbb{E}_p t = 0$:

$$\frac{L_D(\widehat{W}) - L_D^{\star}}{L_D^{\star}} = \frac{S\, C(q)}{S\, E + R_0} \leq \frac{C(q)}{E} \qquad (E > 0),$$

with equality when there are no zero-feature points. If $E = 0$, every residual dyad vanishes and any feature basis with positive weights is a spanning zero certificate, so the ratio is 1. Thus $E > 0$ in every nonexact branch. Conversely, every finite system (p_i, t_i, z_i) with the three moment conditions is realized by a whitened dataset without zero-feature points: take $\xi_i = \sqrt{p_i} s_i$ with $\bar{s}_i/s_i = t_i$ and $\eta_i = z_i\xi_i$. For this realization, the excess ratio of a spanning selection is exactly $C(q)/E$.

For $k \in \{3, 4\}$ define

$$\Gamma_k = \sup_{\substack{\text{moment systems} \\ E>0}} \frac{1}{E} \inf_{\substack{|\operatorname{supp} q| \leq k \\ |\mu(q)|<1}} C(q).$$

E.3 INTERPOLATION IDENTITIES

Lemma 6.2 (Interpolation convex combination; restated). *For $t_i \neq t_j$ let $B_{ij} = \left(\frac{t_i z_j - t_j z_i}{t_i - t_j},\ \frac{z_i - z_j}{t_i - t_j}\right) \in \mathbb{C}^2$ (the coefficients of the affine function of t through the two points) and $d_{ij} = |t_i - t_j|^2$. Then for every admissible q,*

$$\Delta(q) := 1 - |\mu|^2 = \sum_{i<j} q_i q_j d_{ij}, \qquad (a, b) = \sum_{\substack{i<j \\ t_i \neq t_j}} \frac{q_i q_j d_{ij}}{\Delta(q)}\, B_{ij}.$$

In particular (a, b) is a convex combination of the pairwise interpolation coefficients; the conditioning factor $(1 - |\mu|^2)^{-1}$ is absorbed entirely.

Proof. Expanding $1-|\mu|^2 = \frac{1}{2}\sum_{i,j} q_i q_j |t_i - t_j|^2$ gives the first identity. With $w_i = \bar{t}_i z_i$, one checks that $d_{ij}a_{ij} = (t_i - t_j)(w_i - w_j)$ and $d_{ij}b_{ij} = (\bar{t}_i - \bar{t}_j)(z_i - z_j)$. Summing against $q_i q_j$ and using $\sum_{i<j} q_i q_j (x_i - x_j)(y_i - y_j) = \sum_i q_i x_i y_i - (\sum_i q_i x_i)(\sum_i q_i y_i)$ yields $\sum_{i<j} q_i q_j d_{ij} a_{ij} = \nu - \mu\lambda$ and $\sum_{i<j} q_i q_j d_{ij} b_{ij} = \lambda - \bar{\mu}\nu$. The resulting pair equals $\Delta(q) \cdot (a, b)$ by (8). Same-direction pairs have $d_{ij} = 0$ and are omitted. □

Lemma E.1 (Three-point closure identity)**.** *For $2 \le |S| \le 3$ let $K_S = \operatorname{conv}\{B_{ij} : i, j \in S,\ t_i \ne t_j\}$ (discard S with no valid pair). Then*

$$\inf_{\substack{\operatorname{supp} q \subseteq S \\ |\mu(q)|<1}} C(q) \;=\; \operatorname{dist}^2\big(0, K_S\big), \qquad \inf_{\substack{|\operatorname{supp} q| \le 3 \\ |\mu(q)|<1}} C(q) \;=\; \min_S \operatorname{dist}^2(0, K_S).$$

For three pairwise distinct directions the interior of the edge-weight triangle is attained by positive weights (via $q_1{:}q_2{:}q_3 = \frac{d_{23}}{\alpha_{23}}{:}\frac{d_{13}}{\alpha_{13}}{:}\frac{d_{12}}{\alpha_{12}}$), the vertices by two-point supports, and non-degenerate edge interiors only as limits with $|\mu| \to 1$. If two of the three directions coincide, the corresponding segment of K_S is attained exactly. Moreover, for three distinct directions the three points B_{ij}, B_{ik}, B_{jk}, when distinct, are never collinear over $\mathbb{R}$.

Proof. By Lemma 6.2, for $\operatorname{supp} q \subseteq S$, the value (a, b) ranges over convex combinations of the valid B_{ij} with weights $r_{ij} \propto q_i q_j d_{ij}$. The parametrization above shows that all interior weight profiles are realized. The boundary cases can be checked directly: for a repeated direction, distributing weight inside the repeated pair moves along the segment while preserving $|\mu| < 1$. Distances to closures equal infima of continuous functions over the realized sets. To prove non-collinearity, note that $B_{ik} - B_{ij}$ is a complex multiple of $(-t_i, 1)$ and $B_{jk} - B_{ij}$ of $(-t_j, 1)$. Real proportionality of such nonzero vectors forces $t_i = t_j$. □

Remark E.2 (The infimum need not be attained)**.** The inner infimum in F_{weighted} cannot in general be replaced by a minimum, even for integer data. Let $d = m = 2$, $N = 5$, with rows $(x_i^\top \mid y_i^\top)$ equal to $(1, 0 \mid 1, 4)$, $(0, 1 \mid -1, 1)$, $(1, 1 \mid -2, 0)$, $(1, -1 \mid 1, 1)$, $(1, 2 \mid 1, 4)$. Then $L_D^\star = \frac{221}{60}$. By the interpolation identity of Lemma 6.2 (which uses only the normal equations, not the whitening), for a spanning selection the deviation $W^\circ - \widehat{W}$ is a convex combination of the two-point interpolation deviations H_{ij} with coefficients proportional to $c_i c_j \det(x_i, x_j)^2$. An exact enumeration of the ten three-point supports shows that the smallest excess loss over their closures is $\frac{11}{20N}$, attained only for the support $\{3, 4, 5\}$ at the point $\frac{11}{45}H_{34} + \frac{34}{45}H_{45}$ of the *open edge* joining H_{34} and H_{45}. Since the five feature directions are distinct, a three-point selection with positive weights has all three edge coefficients positive and a two-point selection is a vertex, so this point is not realized. Rank-deficient selections are single points and have larger loss. Hence $L_D^\star(3; \text{weighted}) = \frac{569}{150}$, i.e. $R_D(3) = \frac{1138}{1105}$, is not attained, although it is approached by the weights $(0, 0, 99\varepsilon, 36, 136\varepsilon)/(36+235\varepsilon)$ as $\varepsilon \downarrow 0$: the weight vectors converge to the singleton $\{4\}$, whose loss is much larger, reflecting the discontinuity of the minimum-norm rule at a rank drop.

Lemma E.3 (Global edge identities)**.** *Let $w_{ij} = p_i p_j d_{ij}$. Then*

$$\sum_{i<j} w_{ij} = 1, \qquad \sum_{\substack{i<j \\ t_i \ne t_j}} w_{ij} B_{ij} = 0,$$

and

$$\sum_{\substack{i<j \\ t_i \ne t_j}} w_{ij} \|B_{ij}\|^2 \;=\; 2E - 2\sum_G \Big(P_G \sum_{i \in G} p_i |z_i|^2 - \Big| \sum_{i \in G} p_i z_i \Big|^2 \Big) \;\le\; 2E,$$

where G ranges over the classes of equal direction and $P_G = \sum_{i \in G} p_i$. In particular some valid pair has $\|B_{ij}\|^2 \le 2E$.

Proof. The first identity follows from Lemma 6.2 at $q = p$ with $\mu = 0$. The second follows from Lemma 6.2 at $q = p$, where $(a, b) = (0, 0)$ by the moment conditions. For the third, expand $d_{ij}\|B_{ij}\|^2 = |t_i z_j - t_j z_i|^2 + |z_i - z_j|^2$. When summed over all pairs, the two expressions on the right telescope to $E - |\mathbb{E}_p(\bar{t}z)|^2 = E$ and $E - |\mathbb{E}_p z|^2 = E$, respectively. In these expanded sums, same-direction pairs, for which B_{ij} is undefined, contribute $2p_i p_j|z_i - z_j|^2$. Regrouping these contributions over classes gives the correction term. □

E.4 Star localization and the two unconditional bounds

Lemma E.4 (Star localization). *Fix an atom i and set $P_{ij} = \frac{1}{2}p_j d_{ij}$ for $j \neq i$. Then $\sum_j P_{ij} = 1$, the valid star coefficients B_{ij} lie on a real two-dimensional affine plane through β_i with $\|\beta_i\|^2 = \frac{1}{2}|z_i|^2$, and $\sum_j P_{ij}B_{ij} = \beta_i$. Their variance is $V_i = E + \frac{|z_i|^2}{2} - L_i$ where $L_i = \sum_{j:t_j=t_i} p_j|z_j - z_i|^2 \geq 0$. Consequently there is a convex combination δ of at most two valid star coefficients with*

$$\|\delta\|^2 \;\leq\; \frac{E}{4} + \frac{5|z_i|^2}{8} - \frac{L_i}{4},$$

and δ is a closure point of three-point selections (Lemma E.1). Choosing i with $|z_i|^2 \leq E$:

$$\inf_{\substack{|\operatorname{supp} q|\leq 3\\ |\mu|<1}} C(q) \;\leq\; \frac{7E}{8}, \qquad \textit{so} \qquad \Gamma_3 \leq \frac{7}{8}. \tag{9}$$

Proof. Put $\beta_i = \left(\frac{z_i}{2},\ \frac{\bar{t}_i z_i}{2}\right) \in \mathbb{C}^2$, so $\|\beta_i\|^2 = \frac{1}{2}|z_i|^2$. Every valid star coefficient $B_{ij} = (a_{ij}, b_{ij})$ interpolates atom i, so $a_{ij} + t_i b_{ij} = z_i$. Since β_i satisfies the same equation, $B_{ij} - \beta_i \in \{(-t_i\xi, \xi) : \xi \in \mathbb{C}\}$, a real two-dimensional space orthogonal to β_i. Indeed, $\langle(-t_i\xi, \xi), \beta_i\rangle = -t_i\xi\bar{z}_i/2 + \xi t_i\bar{z}_i/2 = 0$. Next, $P_{ij} = \frac{1}{2}p_j|t_i - t_j|^2 = p_j\big(1 - \Re(\bar{t}_i t_j)\big)$ sums over j to $1 - \Re(\bar{t}_i\mathbb{E}_p t) = 1$. Expanding $d_{ij}b_{ij} = (\bar{t}_i - \bar{t}_j)(z_i - z_j)$, $d_{ij}a_{ij} = (t_i - t_j)(\bar{t}_i z_i - \bar{t}_j z_j)$ and using the three moment conditions gives the mean and variance identities

$$\sum_j P_{ij}B_{ij} = \beta_i, \qquad \sum_j P_{ij}\|B_{ij} - \beta_i\|^2 = E + \frac{|z_i|^2}{2} - L_i,$$

where the sums run over the valid j, those with $t_j \neq t_i$. The indices with $t_j = t_i$ have $P_{ij} = 0$ and are the source of the term L_i. After discarding these indices, the remaining P_{ij} form a probability vector. The centered points $B_{ij} - \beta_i$ have P-weighted mean zero in a real two-dimensional space, so Lemma 5.1 with $m = 2$ produces a convex combination v of at most two of them with $\|v\|^2 \leq V_i/4$. The same convex combination of the corresponding B_{ij} is $\delta' = \beta_i + v$, and orthogonality of β_i to the centered directions gives $\|\delta'\|^2 = \|\beta_i\|^2 + \|v\|^2 \leq \frac{|z_i|^2}{2} + \frac{V_i}{4}$, which is the stated bound. The bound (9) follows from $\min_i |z_i|^2 \leq E$ and $L_i \geq 0$. □

Lemma E.5 (Four-point $E/2$ bound). *For every moment system, $\inf_{|\operatorname{supp} q|\leq 4,\ |\mu|<1} C(q) \leq \frac{E}{2}$; hence $\Gamma_4 \leq \frac{1}{2}$.*

Proof. Fix i with $|z_i|^2 \leq E$. By Lemma E.4 the valid star coefficients B_{ij} lie in a real two-dimensional affine plane and average to β_i under the weights P_{ij}. By Carathéodory in the plane, $\beta_i = \sum_{r=1}^{r_0} \alpha_r B_{ij_r}$ for some $r_0 \leq 3$ valid neighbours and convex weights α. Take $q_i = 1 - \varepsilon$, $q_{j_r} = \varepsilon c_r$ $(r \leq r_0)$ with $c_r \propto \alpha_r/d_{ij_r}$. By Lemma 6.2, the normalized edge weights of the star edges tend to α_r, while all leaf–leaf edges carry total weight $O(\varepsilon)$. Thus $(a, b) \to \beta_i$ as $\varepsilon \downarrow 0$, along selections with support ≤ 4 and $|\mu| < 1$. Hence the infimum is at most $\|\beta_i\|^2 = \frac{1}{2}|z_i|^2 \leq \frac{E}{2}$. □

E.5 The interval theorem

Theorem 6.4 (restated). $F_{\text{weighted}}(2, 2, 3) \in \left[\frac{13}{8}, \frac{15}{8}\right]$ *and* $F_{\text{weighted}}(2, 2, 4) \in \left[\frac{5}{4}, \frac{3}{2}\right]$.

Proof. The lower bounds follow from Proposition 6.3. For the upper bounds, fix a dataset and a budget $n \in \{3, 4\}$. If the feature rank is $r \le 1$, then $\tau(D) \le \max\{1, (m+1)r\} \le 3 \le n$ (Theorem 3.1; one point when $r = 0$) and the ratio is 1. The ratio is also one if $L_D^\star = 0$ or $\tau(D) \le n$. Otherwise, the dictionary of Appendix E.2 applies, and by (9) and Lemma E.5 (a three-point selection is also a four-point selection) the ratio is at most $1 + \Gamma_3 \le \frac{15}{8}$ for $n = 3$ and at most $1 + \Gamma_4 \le \frac{3}{2}$ for $n = 4$. The closure points used in Lemmas E.4–E.5 are limits of admissible selections within the same budget. This suffices because the inner optimization in F_{weighted} is an infimum. □

E.6 The conjecture

Conjecture 6.5 (restated). $F_{\text{weighted}}(2, 2, 3) = \frac{13}{8}$ *and* $F_{\text{weighted}}(2, 2, 4) = \frac{5}{4}$. *Equivalently, for every moment system,*

$$\inf_{|\operatorname{supp} q| \le 3,\ |\mu| < 1} C(q) \le \frac{5E}{8} \qquad \textit{and} \qquad \inf_{|\operatorname{supp} q| \le 4,\ |\mu| < 1} C(q) \le \frac{E}{4}.$$

Remark E.6 (The two formulations are equivalent)**.** The moment inequalities imply the equalities by the branch analysis in the proof of Theorem 6.4 together with the lower bounds of Proposition 6.3. Conversely, suppose a moment system violates one of the inequalities, say $\inf_{|\operatorname{supp} q| \le k} C(q) = vE$ with $v > c$. Realize it as a whitened dataset (x_i, ρ_i) as in Appendix E.2 and add a large linear signal: $y_i = LUx_i - \rho_i$ with U invertible, so that $W^\circ = LU$ and the residuals are the given ρ_i. For spanning selections the signal is recovered exactly and the excess ratio is unchanged, hence at least v. For a non-spanning selection with feature span P_S, the minimum-Frobenius-norm solution is $LUP_S + R_S P_S$, whose deviation from the full minimizer LU has squared Frobenius norm $L^2\|U(I - P_S)\|_F^2 + \|R_S P_S\|_F^2 \ge L^2\|U(I - P_S)\|_F^2$ because the two terms act on orthogonal feature subspaces. There are finitely many non-spanning supports, and U is invertible, so $\min_S \|U(I - P_S)\|_F^2 > 0$ over these supports. For L large, every non-spanning selection therefore has ratio above $1 + v$ as well. Thus $F_{\text{weighted}}(2, 2, n) \ge R_D(n) \ge 1 + v > 1 + c$ with $c \in \{\frac{5}{8}, \frac{1}{4}\}$, contradicting the conjectured equalities.

E.7 Instance values versus profile values at $(2, 2)$

Table 2: Explicit instances versus global profile values at $(d, m) = (2, 2)$.

instance / profile	n	value
$D_{2,2}$ (Corollary 3.5)	5	$43/40$
$F_{\text{weighted}}(2, 2, 5)$ (Theorem 5.2)	5	$9/8$
$D_{2,2}$ (Proposition 5.3)	4	$23/20$
$D^\triangle$ (Proposition 6.3)	4	$5/4$
$F_{\text{weighted}}(2, 2, 4)$ (Theorem 6.4)	4	$[5/4,\ 3/2]$
$D^\triangle$ (Proposition 6.3)	3	$13/8$
$F_{\text{weighted}}(2, 2, 3)$ (Theorem 6.4)	3	$[13/8,\ 15/8]$

F Evidence for Conjecture 6.5: full statements and proofs

This section records five independent pieces of structural evidence for Conjecture 6.5: a finite reduction, the sharp solution of the two-direction class, a stability statement for its equality fiber, two obstructions that rule out natural proof strategies, and the classes on which the conjecture is already proved.

F.1 Finite reduction: at most seven atoms

Proposition 6.6 (restated). *For fixed atoms (t_i, z_i), the closed set $\mathcal{P} = \{p \in \mathbb{R}^N_{\ge 0} : \sum_i p_i = 1,\ \mathbb{E}_p t = \mathbb{E}_p z = \mathbb{E}_p(\bar{t}z) = 0\}$ is a polytope cut out by at most 7 linear equalities, whose points*

with full support are the admissible weight vectors. Minimizing the linear functional $E(p)$ over this polytope yields a basic feasible solution $p^\star$ of support at most 7 with $E(p^\star) \le E(p)$. The subsystem on $\operatorname{supp} p^\star$ is again a moment system, and every selection admissible for it is admissible for the original system. Consequently each inequality of Conjecture 6.5 holds for all finite systems if and only if it holds for all systems with at most seven atoms. Since atoms may be split into identically placed copies without changing the problem, it suffices to treat systems with exactly *seven labelled atoms.*

Proof. $E(p)$ is linear in p, subject to normalization (1 equation) and the three complex moment constraints (6 real equations). A basic optimal solution has support at most the constraint rank ≤ 7. Its subsystem inherits the moment conditions, and its witness lifts verbatim to the original system. If the minimum is $E^\star = 0$, all its supported z_i vanish. In this case, $\mathbb{E}t = 0$ guarantees two distinct directions, and choosing one atom from each gives a two-point selection with $C = 0$. Splitting an atom (p_i, t_i, z_i) into $(\theta p_i, t_i, z_i), ((1-\theta)p_i, t_i, z_i)$ changes no moments. Merging the copies in a witness does not increase its support. The number seven cannot be lowered by this route: for the seventh-roots system of Proposition F.3 the seven real constraints prescribe the Fourier coefficients of p at frequencies $0, 1, 2, 3$ and hence, since p is real, at all frequencies. Thus $p = \frac{1}{7}\mathbf{1}$ is the only feasible weight vector. □

F.2 The two-direction class is solved sharply

Proposition F.1. *Suppose the atoms take exactly two directions. Then these directions are antipodal, and each class has total weight $\frac{1}{2}$ and zero conditional mean. Moreover, $C(q) = \frac{1}{2}(|m_+|^2 + |m_-|^2)$, where $m_\pm$ are the conditional means of the selected weights. Consequently*

$$\inf_{\substack{|\operatorname{supp} q| \le 4 \\ |\mu(q)|<1}} C(q) \le \frac{E}{4}, \qquad \inf_{\substack{|\operatorname{supp} q| \le 3 \\ |\mu(q)|<1}} C(q) \le \frac{5E}{8},$$

and both constants are attained: for six equally weighted atoms with z-values the cube roots of unity over each direction, the three-point optimum is exactly $\frac{5E}{8}$ and the four-point optimum exactly $\frac{E}{4}$. Modulo merging atoms with identical (t, z), and for $E > 0$, equality at $\frac{5E}{8}$ forces each class to be an equilateral z-triangle with conditional weight $\frac{1}{3}$ at each vertex, and the two classes to have equal energy. Furthermore, for five-atom two-direction systems (after merging, class sizes $3{+}2$ or $4{+}1$) the sharp constant at three points is $\frac{4}{7} < \frac{5}{8}$. In the $3{+}2$ case with conditional class energies e_3, e_2,

$$\frac{1}{E} \inf_{\substack{|\operatorname{supp} q| \le 3 \\ |\mu(q)|<1}} C(q) \;\le\; \frac{\min\{e_3,\; e_3/4 + e_2\}}{e_3 + e_2} \;\le\; \frac{4}{7},$$

with equality at $e_2 = \frac{3}{4}e_3$ for an equilateral triple against a balanced antipodal pair. The $4{+}1$ case gives at most $\frac{1}{4}$.

Proof. $\mathbb{E}_p t = 0$ with two unit directions forces antipodality and equal class weights. The two complex moments then force zero conditional class means. The regression fit at $\pm$ is the pair of selected conditional means, giving the formula for C. Call the smallest squared norm of a convex combination of two atoms of a class its *two-point cost*, and let e_G be the conditional energy of the class. Lemma 5.1 (with $m = 2$) bounds the two-point cost by $e_G/4$. For three atoms with conditional weights α_i, this is the deletion bound: deleting one atom leaves the two-point mean $-\frac{\alpha_i z_i}{1-\alpha_i}$, and $\sum_i \frac{(1-\alpha_i)^2}{\alpha_i} \ge 4$ (Cauchy–Schwarz). A single atom of a class costs at most e_G. Allocate $(2, 2)$ or $(2, 1)$ across the classes, both of which must appear in a spanning selection. Balancing the two allocations yields the bounds. For the six-atom system the two allocations are tied and give exactly $\frac{5E}{8}$ and $\frac{E}{4}$. For $E > 0$, equality at $\frac{5E}{8}$ forces equality in the balancing step, i.e. equal class energies, and in each class a two-point cost of exactly $e_G/4$. By the proof of Lemma 5.1, the latter forces every circuit in a decomposition of the class into positive circuits to have second moment e_G and two-point cost $e_G/4$. In this equality case, the three-atom analysis, with equality in Cauchy–Schwarz, therefore forces each circuit to be a uniformly weighted equilateral triangle of radius $\sqrt{e_G}$. Two distinct such triangles inscribed in one circle contain a pair of atoms at angular distance above $120°$, whose

segment passes at squared distance below $e_G/4$ from the origin. Hence, after merging coincident atoms, each class is a single uniformly weighted equilateral triangle, as stated. In the five-atom 3+2 case the pair has zero conditional mean, so two points of it cost 0 and one point costs at most e_2, while the triple costs at most $e_3/4$ with two points and at most e_3 with one. This gives the displayed bound. The expression $\min\{e_3, e_3/4 + e_2\}/(e_3 + e_2)$ is maximized at $e_2 = \frac{3}{4}e_3$ with value $\frac{4}{7}$, attained as stated. In the 4+1 case the singleton class has $z = 0$, and two points of the four-atom class cost at most $e_4/4$ by Lemma 5.1, giving at most $\frac{1}{4}$. □

F.3 Local stability of the reduced equality fiber

Proposition F.2. *Let $\mathcal{E}$ be the set of reduced 3+3 two-direction equality systems of Proposition F.1 (two uniformly weighted equilateral z-triangles of equal radius on antipodal directions). Then: (i) for a system consisting of two uniformly weighted equilateral z-triangles on antipodal directions with class energies $E(1 \pm \eta)$, the normalized three-point value $E^{-1} \inf_{|\operatorname{supp} q| \le 3,\, |\mu| < 1} C$ equals $\frac{1}{8}(5 - 3|\eta|)$, so within this stratum the equality fiber ($\eta = 0$, value $\frac{5}{8}$) sits on a downward cusp; (ii) for every sequence of moment systems with a fixed number of labelled atoms converging to a point of $\mathcal{E}$ in which some direction class genuinely splits (two atoms of one limiting direction class have distinct directions for all sufficiently large indices),*

$$\limsup \; \frac{1}{E} \inf_{|\operatorname{supp} q| \le 3,\, |\mu| < 1} C \;\le\; \frac{1}{2} \;<\; \frac{5}{8}.$$

Hence no nearby ascent direction exists: $\mathcal{E}$ is a local maximum of the normalized three-point value $E^{-1} \inf C$ in this stratified sense. (Only the normalized value is meaningful here: rescaling all z_i by a common factor multiplies C and E by the same constant.)

Proof. (i) A spanning three-point selection takes two atoms from one class and one from the other. In a uniformly weighted equilateral triangle of energy e the best two-point mean has squared norm $e/4$ (an edge midpoint) and the best single point has squared norm e. By the formula for C, the value is $\frac{1}{2}\min\{e_+/4 + e_-,\, e_+ + e_-/4\} = \frac{E}{8}(5 - 3|\eta|)$ for $e_\pm = E(1 \pm \eta)$. Dividing by E gives the normalized value $\frac{1}{8}(5 - 3|\eta|)$. For (ii), let atoms j, k of the $+$ class split: $\delta = t_j - t_k \to 0$, $\delta \neq 0$, with $t_j, t_k \to t_*$ and z_j, z_k tending to distinct points of the limit equilateral triangle (radius A, $E \to A^2$). Such a pair is available at every large index: if two atoms of distinct directions converge to the same vertex, at least one of them differs in direction from any fixed atom converging to another vertex of the class, and that pair may be used instead. Only finitely many pairs occur, so the estimates below are uniform in the choice. Use the prediction coordinates $U_\pm(a, b) = a \pm t_* b$, in which $C = \frac{1}{2}(|U_+|^2 + |U_-|^2)$ at the limit directions. Writing $t_j = t_* + \eta_j$, $t_k = t_* + \eta_k$, the same-class edge $V = B_{jk}$ satisfies the exact identities

$$U_+(V) = \frac{\eta_j z_k - \eta_k z_j}{\delta}, \qquad U_-(V) = \frac{(2t_* + \eta_j) z_k - (2t_* + \eta_k) z_j}{\delta},$$

so that $|\delta|\, U_+(V) \to 0$ while $R_\delta := |\delta|\, U_-(V)$ has $|R_\delta| \to 2|z_j - z_k| > 0$ (its phase need not converge). Pick an atom ℓ of the $-$ class. The two cross edges $B_{j\ell}, B_{k\ell}$ average, in the U-coordinates, to $\left(\frac{z_j + z_k}{2},\, z_\ell\right) + o(1)$, whose first coordinate has modulus $\frac{A}{2}$. Among the three equilateral vertices z_ℓ one always satisfies $\Re(\bar{z}_\ell R_\delta) \le -\frac{1}{2}A|R_\delta|$ (some vertex makes an angle $\le \pi/3$ with $-R_\delta$). Give the same-class edge the convex weight $s_\delta = c_\delta|\delta|$ with $c_\delta = -\Re(\bar{z}_\ell R_\delta)/|R_\delta|^2 = O(1)$, and the two cross edges weights $(1 - s_\delta)/2$ each. This is a valid point of the triangle closure $K_{\{j,k,\ell\}}$ (Lemma E.1). In the limit the first coordinate is unchanged ($s_\delta U_+(V) \to 0$) and the second is optimized along the ray: $|z_\ell + c_\delta R_\delta|^2 \le \frac{3}{4}A^2 + o(1)$. Hence $\limsup \inf C_3 \le \frac{1}{2}\left(\frac{A^2}{4} + \frac{3A^2}{4}\right) = \frac{A^2}{2}$. If neither class splits, the system stays in the two-direction stratum, where the value is at most $\frac{5E}{8}$ with the rigid equality characterization of Proposition F.1. □

F.4 Two obstructions: proof templates that cannot work

Proposition F.3 ($\mu = 0$ obstruction)**.** *A proof of Conjecture 6.5 cannot restrict to selections with $\mu = 0$. Concretely, for the seventh-roots system $t_j = \zeta^j$, $z_j = \zeta^{3j}$, $p_j = \frac{1}{7}$ ($\zeta = e^{2\pi i/7}$), every q with $\mu = 0$ satisfies $C \ge \frac{7/k - 1}{2}$ for support k (a discrete Parseval identity). For these constrained*

selections, $C \geq \frac{3}{8} > \frac{1}{4}$ at $k = 4$ and $C \geq \frac{2}{3} > \frac{5}{8}$ at $k = 3$. Yet unconstrained selections achieve $C = \frac{2}{7}$ at three points (the difference set $\{0, 1, 3\}$) and $C < \frac{1}{4}$ at four consecutive points (an exact algebraic computation modulo $x^3 + x^2 - 2x - 1$). Similarly, for the regular hexagon third-harmonic system ($t_j = e^{\pi ij/3}$, $z_j = (-1)^j$, $p_j = \frac{1}{6}$) all $\mu = 0$ selections of support ≤ 3 have $C = 1$, while $\{0, 1, 2\}$ at weights 2:3:2 attains the exact three-point optimum $C = \frac{1}{2}$.

Proposition F.4 (Ratio-of-averages obstruction)**.** *A second natural template also fails: retaining the original weights on the selected support and comparing only the aggregate second moments $\sum_S \|G_S\|^2$ and $\sum_S D_S^2$ over all k-subsets does not yield the constant $\frac{5}{8}$. For an edge $e = \{i, j\}$ put $w_e = p_i p_j d_{ij}$ and $y_e = w_e B_{ij}$ if $t_i \neq t_j$, $y_e = 0$ if $t_i = t_j$. Define $U = \sum_e w_e^2$, $Q = \sum_e \|y_e\|^2$, $R = 4\sum_i p_i^2$, $V = 2\sum_i p_i^2 |z_i|^2$, $D_S = \sum_{e \subset S} w_e$, $G_S = \sum_{e \subset S} y_e$. For $2 \leq k \leq N$, define the binomials $A = \binom{N-2}{k-2}$, $B = \binom{N-3}{k-3}$, $H = \binom{N-4}{k-4}$ (zero when the lower index is negative). Then the following aggregate identities over all k-subsets S hold:*

$$\sum_{|S|=k} \|G_S\|^2 = (A - 2B + H)Q + (B - H)V, \qquad \sum_{|S|=k} D_S^2 = H + (A - 2B + H)U + (B - H)R.$$

For the fifth-roots system $t_j = \zeta_5^j$, $z_j = t_j^2$, $p_j = \frac{1}{5}$ one gets $U = \frac{3}{25}, R = \frac{4}{5}, Q = \frac{1}{5}, V = \frac{2}{5}$. The ratio of the two aggregates is therefore $\frac{Q+V}{U+R} = \frac{15}{23} > \frac{5}{8}$, so the aggregate comparison cannot certify $\frac{5}{8}$ on this system. This does not say that every original-weight triangle is bad: on the same system the original-weight support $\{0, 1, 3\}$ has $C/E = \frac{195 - 57\sqrt{5}}{242} \approx 0.279 < \frac{5}{8}$. The obstruction is to the aggregate averaging argument, not to original-weight witnesses in general.

Proof of Propositions F.3 and F.4. For the heptagon, $\mu = m_1, \lambda = m_2, \nu = m_3$ where $m_r = \sum_j q_j \zeta^{rj}$, and Parseval gives $1 + 2\sum_{r=1}^3 |m_r|^2 = 7\sum_j q_j^2 \geq 7/k$. With $m_1 = 0$, $C = |m_2|^2 + |m_3|^2$. For the unconstrained witnesses: $q = \frac{1}{3}$ on $\{0, 1, 3\}$ gives, after reduction with $1 + \zeta + \cdots + \zeta^6 = 0$, the value $C = \frac{2}{7}$. For $q = \frac{1}{4}$ on $\{0, 1, 2, 3\}$, substituting $m_r = \frac{1}{4}\sum_{j=0}^3 \zeta^{rj}$ into $C = \big(|m_3 - m_1 m_2|^2 + |m_2 - \bar{m}_1 m_3|^2\big)/(1 - |m_1|^2)^2$ and expressing everything through $x = 2\cos\frac{2\pi}{7}$ (which satisfies $x^3 + x^2 - 2x - 1 = 0$) shows that $\frac{1}{4} - C$ has the sign of $33x^2 + 71x - 36 > 0$. For the hexagon, the $\mu = 0$ supports of size ≤ 3 are the antipodal pairs and the two alternating triangles, all with $C = 1$. The three-point optimum follows from Lemma E.1 by evaluating $\mathrm{dist}^2(0, K_S)$ on the three dihedral types of triangles: the cyclic gap types $(1, 1, 4)$ and $(1, 2, 3)$ give $\frac{1}{2}$, while $(2, 2, 2)$ gives 1. The value $\frac{1}{2}$ is attained on $\{0, 1, 2\}$ by the weights 2:3:2. The subset averages follow by counting occurrences of an edge, an adjacent edge pair, and a disjoint edge pair in k-subsets, together with the cross-term bookkeeping $\mathrm{Adj} = V - 2Q$, $\mathrm{Disj} = Q - V$ (from $\sum_e y_e = 0$) and their scalar analogues. The fifth-roots values use $\|B_{ij}\|^2 = 5 - d_{ij}$ and the two edge lengths $d_\pm = \frac{5 \pm \sqrt{5}}{2}$. □

F.5 Closed classes, a necessary condition, and records

Proposition 6.7 (Classes on which Conjecture 6.5 holds; restated). *The three-point inequality of Conjecture 6.5 holds for all moment systems with at most four atoms, and the four-point inequality for all systems with at most five atoms. Both inequalities hold for all two-direction systems, and the four-point inequality holds for all antipodally paired systems (disjoint unions of pairs $\{(t, z), (-t, z)\}$ with equal z-values inside each pair).*

Proof. Systems with fewer atoms than the budget are trivial: $q = p$ gives $C = 0$. If $E = 0$, then every $z_i = 0$. In this zero-energy case, $\mathbb{E}_p t = 0$ ensures that at least two distinct directions occur, and choosing one atom from each of two such directions gives $|\mu| < 1$ and $\nu = \lambda = 0$, hence $C = 0$. We may therefore assume $E > 0$. If some atom has $p_i \geq \frac{1}{2}$, then $\mathbb{E}_p t = 0$ forces $p_i = \frac{1}{2}$ and all other atoms antipodal to t_i, so the system is two-directional and Proposition F.1 applies. Otherwise every $p_i < \frac{1}{2}$, and deleting a single atom i and renormalizing the original weights on the others gives an admissible selection, since $|\mu| = p_i/(1 - p_i) < 1$. Realize the system by a whitened dataset without zero-feature points and with $\sum_i |\xi_i|^2 = 1$, so that $\sum_i x_i x_i^\top = \frac{1}{2} I_2$ (Appendix E.2). In

the full least-squares problem, atom i then has leverage $h_i = x_i^\top (\sum_j x_j x_j^\top)^{-1} x_i = 2p_i$, and the leave-one-out identity (Sherman–Morrison) gives the exact deletion cost

$$C_{-i} = p_i |z_i|^2 \, \frac{h_i}{(1-h_i)^2}.$$

Write $\theta_i = p_i|z_i|^2/E$, so $\sum_i \theta_i = 1$, and let s be the number of atoms. Since $\sum_i h_i = 2$, Cauchy–Schwarz gives $\sum_i \frac{(1-h_i)^2}{h_i} = \sum_i \frac{1}{h_i} - 2s + 2 \geq \frac{s^2}{2} - 2s + 2 = \frac{(s-2)^2}{2}$, so not every C_{-i} can exceed $\frac{2E}{(s-2)^2}$ (summing $\theta_i > \frac{2(1-h_i)^2}{(s-2)^2 h_i}$ over i would give $1 > 1$). For $s = 4$ this yields a three-atom selection with $C \leq \frac{E}{2} \leq \frac{5E}{8}$, and for $s = 5$ a four-atom selection with $C \leq \frac{2E}{9} \leq \frac{E}{4}$. Two-direction systems are covered by Proposition F.1. For an antipodally paired system, select two pairs and give the two atoms of each pair equal weight. Then $\mu = 0$ and $\lambda = 0$, so $C = |\nu|^2$ with ν a convex combination of the two pair values z. The pair values, weighted by the pair masses, form a zero-mean system in $\mathbb{C} \cong \mathbb{R}^2$ with second moment E, and Lemma 5.1 with $M = 2$, $k = 2$ provides two pairs with $|\nu|^2 \leq E/4$. □

A necessary condition and exact records. By Lemma E.4, any counterexample to the three-point inequality must satisfy $5|z_i|^2 - 2L_i > 3E$ for *every* atom i (for pairwise distinct directions: $|z_i|^2 > \frac{3E}{5}$ for all i). Exact records: the hexagon three-point optimum is $\frac{1}{2}$, while its four-point value is 0 (the weights $(1, 2, 2, 1)/6$ on four consecutive atoms give $\nu = \lambda = 0$ and $|\mu| = 1/\sqrt{3}$). The five-atom two-direction worst case is exactly $\frac{4}{7}$.

G THE GENERAL k-POINT MEAN LEMMA

Lemma 5.1 is the case $k = M$ of a sharp family of sparsification bounds. We record this family both for its own interest and for its applications: it supplies the sharp per-line constants for the allocation problems mentioned in Section 7 and settles the whole profile for $d = 1$ (Corollary G.3). The proof is elementary and self-contained. *None of Theorems 3.4, 6.1, 5.2 or 6.4 depends on this appendix.* The only case used elsewhere is $k = M$ (Lemma 5.1). In the boundary variance fact and in Sublemma G below, n denotes the dimension of a simplex. In Corollary G.3, it again denotes the budget.

Relation to other sparsification bounds. For $k = 2$ and $M \geq 3$, the constant $\frac{M-1}{2M}$ of Theorem G.1 coincides with, and follows from, the bound in Theorem 2 of (Dewasurendra, 2026b) after the circuit reduction used below. The cited theorem gives an equal-weight two-point bound for zero-mean distributions on at most s atoms, with constant $\max\{\frac{1}{3}, \frac{s-2}{2(s-1)}\}$. The reduction restricts the circuit support to $s \leq M + 1$. For $M = 2$, Lemma 5.1 gives $\frac{1}{4}$ for arbitrary convex weights on two points. This feasible set differs from the equal-weight pairs behind the constant $\frac{1}{3}$ in the cited result, which also does not cover general k. A radius-normalized, equal-weight analogue concerns the worst squared distance from a point of a simplex in the unit ball to a k-fold barycenter of its vertices. This is the sharp approximate Carathéodory theorem of (Tinarrage, 2026). That statement normalizes by the radius of a ball containing the simplex and uses equal weights, whereas Theorem G.1 normalizes by the variance E and allows arbitrary convex weights. Thus neither is a formal special case of the other. Theorem G.1 does imply a skeleton-distance statement with radius normalization and a weaker, face-distance conclusion. If the vertices v_i of an M-simplex lie in the unit ball and $y = \sum_i q_i v_i$ with $q_i > 0$, then $E = \sum_i q_i \|v_i - y\|^2 = \sum_i q_i \|v_i\|^2 - \|y\|^2 \leq 1 - \|y\|^2$, so some face F spanned by at most k vertices satisfies $\operatorname{dist}(y, F)^2 \leq \frac{M+1-k}{kM}(1 - \|y\|^2)$.

Theorem G.1 (k-point mean lemma)**.** *Let $M \geq 1$ and let $z_i \in \mathbb{R}^M$ carry weights $p_i > 0$, $\sum_i p_i = 1$, with $\sum_i p_i z_i = 0$, and $E = \sum_i p_i \|z_i\|^2$. For every $1 \leq k \leq M + 1$ there is a convex combination δ of at most k of the points with*

$$\|\delta\|^2 \;\leq\; E \cdot \frac{M+1-k}{kM}.$$

The constant is sharp: for the uniformly weighted regular M-simplex, every convex combination supported on k points has squared norm at least $E\frac{M+1-k}{kM}$, with equality for k vertices at equal weights.

The proof has two steps: (i) reduce to a single positive circuit of support $s \le M+1$ (extreme points of the zero-mean polytope, as in Lemma 5.1); (ii) apply the following sparsification lemma with $a_i = z_i$.

Lemma G.2 (Simplex skeleton sparsification)**.** *Let $s \ge 2$, let $a_1, \dots, a_s$ lie in a Hilbert space, let q be a probability vector with full support s, and let $V_q = \sum_i q_i \|a_i - \bar{a}_q\|^2$ where $\bar{a}_q = \sum_i q_i a_i$. For every $1 \le k \le s$ there is a probability vector λ with $|\operatorname{supp} \lambda| \le k$ and*

$$\Big\| \sum_i \lambda_i a_i - \bar{a}_q \Big\|^2 \;\le\; \frac{s-k}{k(s-1)}\, V_q.$$

Proof (the two nontrivial steps are stated in full). We construct a martingale of probability vectors that successively eliminates coordinates. The key step is the following fact.

Boundary variance fact. *Let q have $r \ge 2$ positive coordinates and let $H \succeq 0$ be a quadratic form. Then there is a random probability vector Q_H supported on the boundary $\partial\Delta_r = \{u \in \Delta_r : \min_i u_i = 0\}$ with $\mathbb{E}[Q_H] = q$ and*

$$\mathbb{E}\big[(Q_H - q)^\top H (Q_H - q)\big] \;\le\; \frac{\operatorname{tr}\big(H(\operatorname{Diag} q - qq^\top)\big)}{(r-1)^2}.$$

(One boundary variable per form H suffices for our purposes.)

Proof of the fact. The left-hand side, minimized over admissible laws, is a finite moment problem on the compact set $\partial\Delta_r$. An optimal law with finite support exists. Strong duality gives

$$\Phi_H(q) = \sup\Big\{ \ell(q) :\ \ell \text{ affine},\ \ell(u) \le (u-q)^\top H (u-q)\ \forall u \in \partial\Delta_r \Big\},$$

because the set of achievable pairs (mean, cost) is compact and convex and q is a relative interior point of Δ_r, so the supporting hyperplane of that set at the optimal value is non-vertical. Fix a feasible ℓ and write $a = \ell(q)$. If $a \le 0$, the desired bound $(r-1)^2 a \le \operatorname{tr}(H(\operatorname{Diag} q - qq^\top))$ holds because the right-hand side is nonnegative. Otherwise, replace H on the tangent space $T = \{\mathbf{1}^\top y = 0\}$ by $H_\varepsilon = H + \varepsilon P_T$, which is positive definite on T and keeps ℓ feasible. In this positive-value case, the bound for H_ε implies the bound for H as $\varepsilon \downarrow 0$. We may therefore assume $H \succ 0$ on T. Writing $\ell(q+y) = a + 2\langle c, y\rangle_H$ ($c \in T$) and $R^2 = a + \|c\|_H^2 > 0$, feasibility says that the open H-ball $B_H(c,R) \subseteq T$ misses $\partial\Delta_r - q$. This ball contains 0 (as $a > 0$), a relative interior point of the convex set $\Delta_r - q$. Since the ball is connected, $B_H(c,R) \subseteq \Delta_r - q$. The desired bound $(r-1)^2 \ell(q) \le \operatorname{tr}(H(\operatorname{Diag} q - qq^\top)) = \sum_i q_i (e_i - q)^\top H (e_i - q)$ then follows by applying the sublemma below in the Euclidean structure $\langle\cdot,\cdot\rangle_H$ on T, with $n = r-1$, the simplex $\Delta_r - q$, the ball $B_H(c,R)$, and the barycentric representation $0 = \sum_i q_i (e_i - q)$.

Sublemma G (simplex–ball interpolation). *Let $n \ge 1$, let an n-simplex $T \subset \mathbb{R}^n$ contain the ball $B(c,R)$, assume $\|c\| < R$, and let $0 = \sum_i q_i v_i$ be the barycentric representation of the origin in the vertices v_i of T. Then*

$$\sum_i q_i \|v_i\|^2 \;\ge\; n^2 \,(R^2 - \|c\|^2).$$

Proof of Sublemma G. Normalize the ball to the unit ball centered at the origin by translating by $-c$ and dividing by R. The vertices become V_i, and the original origin becomes $y = -c/R$ with $\beta = \|y\| < 1$. We have $y = \sum_i q_i V_i$ with the same barycentric weights, so the claim becomes $\mathcal{V}_T(y) := \sum_i q_i \|V_i - y\|^2 \ge n^2(1-\beta^2)$. Since T contains the unit ball, the facet of T opposite V_i lies on a hyperplane $\{x : \langle b_i, x\rangle = 1\}$ with $\|b_i\| \le 1$, and $T = \{x : \langle b_i, x\rangle \le 1\ \forall i\}$. Let λ_i be the i-th barycentric coordinate function of T and $p_i = \lambda_i(0) > 0$. As λ_i is affine and vanishes on the facet opposite V_i, $\lambda_i(x) = p_i(1 - \langle b_i, x\rangle)$. Comparing constant and linear terms in the identities $\sum_i \lambda_i(x) = 1$ and $\sum_i \lambda_i(x) V_i = x$ gives

$$\sum_i p_i = 1, \qquad \sum_i p_i b_i = 0, \qquad \sum_i p_i V_i = 0, \qquad \sum_i p_i V_i b_i^\top = -I_n.$$

Put $t_i = \langle b_i, y\rangle$, so $|t_i| \le \beta$ and $q_i = \lambda_i(y) = p_i(1 - t_i) > 0$. Taking the trace of the last identity and using the second gives $\sum_i p_i \langle V_i - y, b_i\rangle = -n$. Then Cauchy–Schwarz gives

$$n^2 = \Big| \sum_i p_i \langle V_i - y, b_i\rangle \Big|^2 \le \Big(\sum_i q_i \|V_i - y\|^2 \Big) \Big(\sum_i \frac{p_i^2}{q_i} \|b_i\|^2 \Big) \le \mathcal{V}_T(y) \sum_i \frac{p_i}{1 - t_i}.$$

Finally $\frac{1}{1-t_i} \le \frac{1+t_i}{1-\beta^2}$ (equivalent to $t_i^2 \le \beta^2$) and $\sum_i p_i t_i = \langle \sum_i p_i b_i, y\rangle = 0$ give $\sum_i \frac{p_i}{1-t_i} \le \frac{1}{1-\beta^2}$. Hence $\mathcal{V}_T(y) \ge n^2(1-\beta^2)$. This proves Sublemma G, hence the boundary variance fact. □

(For $n \ge 2$, Sublemma G can also be derived from the Grace–Danielsson inequality $(\mathcal{R} - nr)(\mathcal{R} + (n-2)r) \ge \omega^2$ between the circumradius $\mathcal{R}$, the inradius r and the distance ω of the centers of a simplex, proved in all dimensions by Drozdov (Drozdov, 2025). The argument above does not use this inequality.)

Now iterate: starting from q with support s, apply the fact with $H = \mathsf{A}^\top \mathsf{A}$ (where $\mathsf{A} = [a_1 \cdots a_s]$), obtaining a boundary law Q with $\mathbb{E}[\bar a_Q] = \bar a_q$ and $\mathbb{E}\|\bar a_Q - \bar a_q\|^2 \le V_q/(r-1)^2$. The total-variance identity $\mathbb{E}[V_Q] = V_q - \mathbb{E}\|\bar a_Q - \bar a_q\|^2$ gives $\mathbb{E}[V_Q] \ge \big(1 - \frac{1}{(r-1)^2}\big)V_q$. Stop as soon as the support has size at most k. Set $f_r = 1 - \frac{1}{(r-1)^2}$ and $\Pi_j = \prod_{t=k+1}^{j} f_t$ (with $\Pi_j = 1$ for $j \le k$), which is nonincreasing in j. By backward induction on the support size r, the terminal variance obeys $\mathbb{E}[V_\Lambda] \ge \Pi_r V$ for every starting law of support r and variance V. A step may eliminate several coordinates at once, but if the support drops from r to $t \le r-1$, then $\mathbb{E}[V_\Lambda \mid Q] \ge \Pi_t V_Q \ge \Pi_{r-1} V_Q$, whence $\mathbb{E}[V_\Lambda] \ge \Pi_{r-1}\mathbb{E}[V_Q] \ge \Pi_{r-1} f_r V_q = \Pi_r V_q$. Since the mean sequence is a martingale with orthogonal increments, $\mathbb{E}\|\bar a_\Lambda - \bar a_q\|^2 = V_q - \mathbb{E}[V_\Lambda] \le \big(1 - \Pi_s\big)V_q$, and the telescoping product $\Pi_s = \prod_{r=k+1}^{s} \frac{r(r-2)}{(r-1)^2} = \frac{s(k-1)}{k(s-1)}$ yields the bound $\frac{s-k}{k(s-1)} V_q$. Some realization λ of the terminal law achieves it. □

Proof of Theorem G.1. Reduce to a circuit q of support $s \le M+1$ and second moment $E_q \le E$ exactly as in Lemma 5.1, handling the degenerate branches $E_q = 0$ and $s \le k$ as in that argument. In the remaining case, $s > k \ge 1$, so $s \ge 2$. Apply Lemma G.2 with $a_i = z_i$, $\bar a_q = 0$: some k-sparse convex δ has $\|\delta\|^2 \le \frac{s-k}{k(s-1)} E_q$. Since $s \mapsto \frac{s-k}{s-1}$ is nondecreasing for $k \ge 1$, the constant is at most $\frac{M+1-k}{kM}$. To prove sharpness, note that for regular simplex vertices, $\|\sum_i \alpha_i v_i\|^2 = R^2\big(\frac{M+1}{M}\sum_i \alpha_i^2 - \frac{1}{M}\big) \ge R^2 \frac{M+1-k}{kM}$ whenever $|\operatorname{supp}\alpha| \le k$, using $\sum \alpha_i^2 \ge 1/k$. □

Corollary G.3 (The full profile for $d = 1$)**.** *For all $m \ge 1$ and $n \ge 1$,*

$$F_{\text{weighted}}(1,m,n) = 1 + \frac{m+1-n}{mn} \ \ (1 \le n \le m+1), \qquad F_{\text{weighted}}(1,m,n) = 1 \ \ (n \ge m+1).$$

Proof. Upper bound. Fix a dataset with $d = 1$. If all $x_i = 0$ the ratio is 1 (both rules return $W = 0$). Otherwise, let $I = \{i : x_i \ne 0\}$, $g = \sum_{i\in I} x_i^2$, $p_i = x_i^2/g$, $a_i = y_i/x_i \in \mathbb{R}^m$ and $\mu = \sum_{i\in I} p_i a_i$. The full-data minimizer is the column $W^\circ = \mu$. With $E = \sum_{i\in I} p_i \|a_i - \mu\|^2$, we have $gE \le NL_D^\star$ (the difference is the loss on zero-feature points). If $E = 0$, a single point of I recovers W°. Otherwise, for $n \le m+1$, Theorem G.1 applied to $z_i = a_i - \mu$ gives a convex α supported on at most n points of I with $\|\sum_i \alpha_i a_i - \mu\|^2 \le E\frac{m+1-n}{mn}$. The selection weights $c_i \propto \alpha_i/x_i^2$ have the unique minimizer $\widehat{W} = \sum_i \alpha_i a_i$. By (5), the excess loss is $g\|\widehat{W} - \mu\|^2/N \le L_D^\star \frac{m+1-n}{mn}$. *Lower bound.* Take $x_i \equiv 1$ and labels the vertices of a uniformly weighted regular m-simplex centered at 0. Then $W^\circ = 0$ and $L_D^\star = E$. Every selection of at most n points returns a convex combination of at most n vertices, of squared norm at least $E\frac{m+1-n}{mn}$ by the sharpness part of Theorem G.1. For $n \ge m+1$ the value 1 is Theorem 3.4. □

At $n = 1$ and $n = m$ the corollary reproduces $F_{\text{weighted}}(1, m, 1) = 2 = d + 1$ and $F_{\text{weighted}}(1, m, m) = 1 + \frac{1}{m^2} = 1 + \frac{1}{dm^2}$, in agreement with Theorems 6.1 and 5.2.

Results and proof locations

Object	Statement	Original proof location
Theorem 3.1	Section 3	Appendix B.1
Theorem 3.3	Section 3	Appendix B.2
Theorem 3.4	Section 3	Appendix B.2
Theorem 4.1	Section 4	Appendix D.1
Theorem 4.2	Section 4	Appendix D.2
Lemma 5.1	Section 5	Appendix D.3
Theorem 5.2	Section 5	Appendix D.4
Proposition 5.3	Section 5	Appendix B.3
Theorem 6.1	Section 6.1	Appendix C
Lemma 6.2	Section 6.2	Appendix E.3
Proposition 6.3	Section 6.2	Appendix E.1
Theorem 6.4	Section 6.2	Appendix E.5
Proposition 6.6	Section 6.2	Appendix F.1
Proposition 6.7	Section 6.2	Appendix F.5
Theorem G.1	Appendix G	Appendix G, theorem proof
Corollary G.3	Appendix G	Appendix G, following statement

Conjecture 6.5 is stated in Section 6.2. Details appear in Appendix E.6.