Sharp Pair Selection for Mean Regression

Pahan Dewasurendra
Johns Hopkins University

Abstract

Suppose a learner may train the empirical mean on two optimally selected examples, with repetition, and is evaluated by squared loss on the full dataset. How much worse can this be than the full mean? This is the first open column in a recent data-selection problem. We determine the exact answer in every dimension:

$$F(d,2) = 1 + \max\left\{\frac{1}{3}, \frac{d-1}{2d}\right\}.$$

Thus the sharp factor is $4/3$ for $d \leq 3$ and $(3d-1)/(2d)$ for $d \geq 4$. The two branches reveal distinct obstructions. A rare point on a line forces the dimension-free term, while a uniform regular simplex forces the dimension term.

The proof gives a stronger convex-geometric theorem for distributions on at most m atoms. Its key step is an explicit antithetic distribution over pairs. If one atom has mass at least $1/4$, a star coupling has exactly one third of the original second moment. Otherwise, a complete-graph coupling has a closed-form second moment, and a single Jensen inequality proves that uniform weights are worst. A variance-minimizing Carathéodory reduction then transfers the result to arbitrary finite datasets. This resolves all cases with selection budget two, supplies a short proof of the previously isolated planar case, and separates variance-normalized data selection from radius-normalized approximate Carathéodory bounds.

1 Introduction

Data selection asks what a fixed learning rule can achieve when an informed teacher chooses a small training set from a larger population. This viewpoint appears in machine teaching, sample compression, and coreset construction (Dasgupta et al., 2019; Feldman, 2020; Hanneke et al., 2025a). Even the most elementary regression rule leaves a sharp geometric question open.

Let $D = \{z_1, \ldots, z_N\}$ be a finite multiset in $\mathbb{R}^d$. For squared loss, write

$$L_D(h) = \frac{1}{N}\sum_{i=1}^{N} \|h - z_i\|^2, \qquad L_D^* = \min_h L_D(h).$$

The minimizer is the full mean $\mu_D = N^{-1}\sum_i z_i$. If the rule receives only n examples, it returns their equal-weight average. Repetitions are allowed, as in the original formulation. Define

$$L_D^*(n) = \min_{x_1,\ldots,x_n \in D} L_D\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right),$$
$$F(d,n) = \sup_{D \subset \mathbb{R}^d} \frac{L_D^*(n)}{L_D^*}.$$

The ratio is set to one for a constant dataset.

Hanneke et al. (2025a) proved

$$F(1,n) = \frac{2n}{2n-1}, \qquad \frac{2n}{2n-1} \leq F(d,n) \leq \frac{n+1}{n},$$

and showed that the upper bound is approached as d grows. Their COLT open problem records only one additional finite-dimensional value, $F(2,2) = 4/3$, obtained by a computational geometric argument (Hanneke et al., 2025b). Every other entry with $d, n > 1$ was left open.

We close the full $n = 2$ column.

Theorem 1 (Exact pair-selection factor). *For every integer $d \geq 1$,*

$$F(d,2) = \begin{cases} 4/3, & d \leq 3, \\ (3d-1)/(2d), & d \geq 4. \end{cases} \tag{1}$$

For reference, the first values are

d	1	2	3	4	5	$d \to \infty$
$F(d,2)$	4/3	4/3	4/3	11/8	7/5	3/2

The formula is the maximum of two qualitatively different lower bounds. Four points on a line, three at zero and one at one, force excess risk 1/3. The uniform distribution on a regular d-simplex forces excess risk $(d-1)/(2d)$. Their crossover at $d=3$ explains why extrapolating from the known planar answer misses the higher-dimensional behavior.

Our main technical result is stronger than Theorem 1. It gives the exact constant as a function of the number of atoms, independent of the ambient Hilbert space.

Theorem 2 (Sharp atomic theorem)**.** *Let $m \geq 2$. Let X be a finitely supported random vector in a real Hilbert space, with $\mathbb{E}X = 0$, $0 < \mathbb{E}\|X\|^2 < \infty$, and $|\operatorname{supp}(X)| \leq m$. Then there are $x, y \in \operatorname{supp}(X)$, possibly equal, such that*

$$\left\| \frac{x+y}{2} \right\|^2 \leq C_m \mathbb{E}\|X\|^2, \tag{2}$$

$$C_m = \max\left\{ \frac{1}{3}, \frac{m-2}{2(m-1)} \right\}. \tag{3}$$

The constant is optimal for every $m \geq 2$.

The proof is constructive at the level of a randomized certificate. It does not sample two independent observations. Instead it builds an antithetic pair whose average has the sharp second moment. This distinction is essential. Independent sampling gives only 1/2, which corresponds to the previously known dimension-free bound $F(d,2) \leq 3/2$.

2 Risk and Atomic Reduction

The squared-loss identity

$$L_D(h) = L_D^* + \|h - \mu_D\|^2 \tag{4}$$

turns the problem into equal-weight mean approximation. After translating by μ_D and normalizing $L_D^* = 1$, it is enough to find two data points whose midpoint has squared norm at most the claimed excess factor.

The following standard reduction makes support size, rather than dataset size, the relevant parameter. We include the argument because preserving the variance direction is important here.

Lemma 3 (Variance-minimizing reduction)**.** *Let $D \subset \mathbb{R}^d$ be a finite multiset with mean μ and variance V. There is a probability distribution P supported on at most $r+1$ points of D, where r is the affine dimension of D, such that*

$$\mathbb{E}_P X = \mu, \qquad \mathbb{E}_P \|X - \mu\|^2 \leq V.$$

Proof. Minimize $\sum_i p_i \|z_i - \mu\|^2$ subject to $p_i \geq 0$, $\sum_i p_i = 1$, and $\sum_i p_i z_i = \mu$. The uniform weights are feasible. An extreme optimal solution has at most $r+1$ positive coordinates because there are $r+1$ independent affine constraints. □

Any pair selected from the reduced support is also available in the original dataset. Therefore, a bound relative to the reduced variance is at least as strong as the same bound relative to V. Lemma 3 and Theorem 2 immediately give the upper bound in Theorem 1, since $m \leq d+1$ and C_m is nondecreasing in m.

2.1 A finite minimax formulation

The reduction exposes a small semidefinite game. Fix positive masses $p \in \mathbb{R}^m$ and let $\mathcal{Q}_2$ contain the empirical frequency vectors e_i and $(e_i + e_j)/2$. Put $D_p = \operatorname{diag}(p)$, $s = \sqrt{p}$, and

$$u_q = D_p^{-1/2}(q - p) \in s^\perp.$$

Every centered geometry with unit variance has a Gram matrix G satisfying $G \succeq 0$, $Gp = 0$, and $\operatorname{tr}(D_p G) = 1$. Equivalently, $B = D_p^{1/2} G D_p^{1/2}$ is positive semidefinite on $s^\perp$ and has trace one. Its squared mean error at q is $u_q^\top B u_q$.

Proposition 4 (Pair minimax duality)**.** *For a fixed positive mass vector p, the worst normalized pair error over all Hilbert-space configurations is*

$$R(p) = \max_{\substack{B \succeq 0,\ Bs=0 \\ \operatorname{tr} B = 1}} \min_{q \in \mathcal{Q}_2} u_q^\top B u_q \tag{5}$$

$$= \min_{\lambda \in \Delta(\mathcal{Q}_2)} \lambda_{\max}\left(\mathbb{E}_{q \sim \lambda}[u_q u_q^\top]\big|_{s^\perp} \right). \tag{6}$$

Proof. Replace the finite minimum by a minimum over distributions λ on $\mathcal{Q}_2$. The objective is bilinear in B and λ, so finite dimensional minimax duality swaps max and min. Maximizing $\operatorname{tr}(BM)$ over positive semidefinite trace-one matrices on $s^\perp$ returns the largest eigenvalue of M on that subspace. □

Independent sampling from p makes the second moment in (5) equal to $(1/2)I$ on $s^\perp$. This recovers excess 1/2. The proof below constructs a dependent distribution λ with a strictly smaller spectral second moment. The construction depends only on p, not on the geometry G.

3 The Antithetic Pair Certificate

We prove Theorem 2. Write the support as $x_1, \dots, x_m$, with positive masses $p_1, \dots, p_m$. Center and scale so that

$$\sum_i p_i x_i = 0, \qquad \sum_i p_i \|x_i\|^2 = 1. \tag{7}$$

It is enough to construct a distribution on pairs whose expected squared midpoint is at most C_m. Some pair then attains the same bound.

3.1 A dominant atom

Suppose $p_j \geq 1/4$ for some j. Consider the following distribution on unordered pairs:

$$\Pr\{(j,j)\} = \frac{4p_j - 1}{3}, \tag{8}$$

$$\Pr\{(i,j)\} = \frac{4p_i}{3}, \qquad i \neq j. \tag{9}$$

The probabilities are nonnegative and sum to one. If M denotes the midpoint of the sampled pair, then

$$\begin{aligned} 3\mathbb{E}\|M\|^2 &= (4p_j - 1)\|x_j\|^2 + \sum_{i\neq j} p_i \|x_i + x_j\|^2 \\ &= \sum_i p_i \|x_i\|^2 = 1. \end{aligned}$$

The second equality uses $\sum_{i\neq j} p_i x_i = -p_j x_j$. Thus a dominant atom always yields a pair with squared midpoint norm at most $1/3$. The threshold $1/4$ is exact because it is precisely where the loop probability in (8) becomes nonnegative.

3.2 Diffuse masses

Now assume $p_i < 1/4$ for every i. Necessarily $m \geq 5$. Define

$$S = \sum_{i=1}^{m} \frac{p_i}{1 - 4p_i}, \qquad C(p) = \frac{S+1}{3S+1}, \tag{10}$$

$$a_i = \frac{C(p)p_i}{(1+S)(1-4p_i)}. \tag{11}$$

For each distinct pair, set

$$w_{ij} = 4(p_i a_j + p_j a_i), \qquad i < j. \tag{12}$$

All weights are positive. They also sum to one. To see this, let $A = \sum_i a_i$ and $U = \sum_i p_i a_i$. The identities

$$A = \frac{C(p)S}{1+S}, \qquad U = \frac{C(p)(S-1)}{4(1+S)}$$

give $\sum_{i<j} w_{ij} = 4(A - U) = 1$.

Lemma 5 (Exact diffuse certificate)**.** *If a distinct pair (i,j) is drawn with probabilities* (12)*, then its midpoint M satisfies*

$$\mathbb{E}\|M\|^2 = C(p)\sum_i p_i\|x_i\|^2 = C(p). \tag{13}$$

Proof. Let $q = (e_i + e_j)/2$ encode the sampled midpoint, so $M = \sum_i q_i x_i$. Write $Q = \mathbb{E}[qq^\top]$. The off-diagonal entries from (12) are

$$Q_{ij} = p_i a_j + a_i p_j.$$

The degree identity

$$\frac{1}{4}\sum_{j\neq i} w_{ij} = C(p)p_i + 2p_i a_i$$

follows from $a_i(1-4p_i) = p_i(C(p) - A)$. Hence

$$Q = C(p)\operatorname{diag}(p) + pa^\top + ap^\top. \tag{14}$$

If $G_{ij} = \langle x_i, x_j\rangle$, centering gives $Gp = 0$. Therefore

$$\mathbb{E}\|M\|^2 = \operatorname{tr}(GQ) = C(p)\operatorname{tr}(G\operatorname{diag}(p)) = C(p).$$

□

It remains to maximize $C(p)$. The function $t \mapsto t/(1-4t)$ is strictly convex on $(0, 1/4)$. Jensen's inequality gives

$$S \geq \frac{m}{m-4}. \tag{15}$$

Since $(S+1)/(3S+1)$ decreases with S,

$$C(p) \leq \frac{m/(m-4)+1}{3m/(m-4)+1} = \frac{m-2}{2(m-1)}. \tag{16}$$

Equality in (15) holds exactly at $p_i = 1/m$. Combining the dominant and diffuse cases proves the upper bound in Theorem 2.

The diffuse certificate is not merely an upper bound. It solves the fixed mass-profile minimax problem exactly.

Theorem 6 (Exact diffuse mass profile)**.** *Fix positive masses $p_1, \ldots, p_m < 1/4$ that sum to one. Over all centered Hilbert-space configurations with unit variance,*

$$R(p) = C(p) = \frac{1 + \sum_i p_i/(1-4p_i)}{1 + 3\sum_i p_i/(1-4p_i)}. \tag{17}$$

The supremum is attained in dimension $m-1$ by an explicit simplex.

Proof. Lemma 5 proves $R(p) \leq C(p)$. For the reverse inequality, write $C = C(p)$ and define

$$b_i = C + \frac{3C-1}{1-4p_i}. \tag{18}$$

Let G have diagonal $G_{ii} = b_i$ and, for $i \neq j$,

$$G_{ij} = 2C - \frac{b_i + b_j}{2}. \tag{19}$$

First, $\sum_i p_i b_i = 1$. This follows from $3C - 1 = 2/(3S+1)$ and the definition of C. The identity

$$(1-4p_i)b_i = 4C(1-p_i) - 1$$

then gives $Gp = 0$ by a row-wise calculation.

The matrix is positive semidefinite. Indeed, for every y with $\sum_i y_i = 0$, Equations (18) and (19) give

$$y^\top G y = 2 \sum_i (b_i - C) y_i^2 > 0. \tag{20}$$

For an arbitrary z, set $y = z - (\sum_i z_i)p$. Then $\sum_i y_i = 0$, and $z^\top G z = y^\top G y$ because $Gp = 0$. Hence $G \succeq 0$, with rank $m-1$. It is therefore the Gram matrix of vectors $x_1, \dots, x_m$ with weighted mean zero and weighted variance one.

Every repeated pair has cost $b_i \geq C$. By (19), every distinct pair has cost exactly

$$\frac{G_{ii} + G_{jj} + 2G_{ij}}{4} = C.$$

Thus the best pair in this configuration has cost C, proving (17). □

Theorem 6 describes a continuum of least favorable nonregular simplices. As a mass approaches 1/4 from below, its squared vertex norm approaches 3 under our normalization and $C(p)$ approaches 1/3. At uniform masses, the Gram matrix becomes the regular-simplex matrix and $C(p)$ reaches the diffuse maximum in (16).

4 Sharpness and Extremizers

The lower bounds are elementary and expose why both branches are necessary.

Rare-point obstruction. Let X equal zero with probability 3/4 and one with probability 1/4. Its mean is 1/4 and its variance is 3/16. Every average of two support points lies in $\{0, 1/2, 1\}$. Its squared distance from the mean is at least 1/16, giving normalized excess 1/3. The finite multiset $\{0, 0, 0, 1\}$ realizes this example exactly.

Simplex obstruction. Let $x_1, \dots, x_m$ be the unit vertices of a regular $(m-1)$-simplex, so $\sum_i x_i = 0$ and $\langle x_i, x_j \rangle = -1/(m-1)$ for $i \neq j$. Under uniform masses the variance is one. Repeating a vertex gives squared norm one, while every distinct midpoint has squared norm

$$\frac{1}{4}\left(2 - \frac{2}{m-1}\right) = \frac{m-2}{2(m-1)}.$$

This proves sharpness of Theorem 2. With $m = d+1$, it gives the second lower bound in Theorem 1.

The diffuse proof also characterizes the high-dimensional extremizer.

Proposition 7 (Uniqueness on a minimal support)**.** *Fix $m \geq 5$. If a centered distribution on exactly m atoms attains the constant $(m-2)/(2(m-1))$ in Theorem 2, then its masses are uniform and its support is a regular $(m-1)$-simplex, up to scaling and an isometry on its span.*

Proof. The dominant certificate is strictly smaller than the claimed value. In the diffuse case, equality in (15) forces uniform masses. The weights w_{ij} then become uniform over distinct pairs. Since their average cost equals the minimum, all distinct midpoint costs are equal. Normalize the variance to one and write $b_i = \|x_i\|^2$. Centering, equal midpoint costs, and $\sum_i b_i = m$ imply

$$0 = \left\langle x_i, \sum_j x_j \right\rangle = \frac{m-4}{2}(1 - b_i).$$

Thus every $b_i = 1$, after which all off-diagonal inner products equal $-1/(m-1)$. □

5 Consequences

The proof provides more information than the worst-case constant. For any variance-minimizing representation with masses p, it gives an explicit certificate. A mass at least 1/4 guarantees excess 1/3. If all masses are smaller, the sharper data-dependent value is

$$C(p) = \frac{1 + \sum_i p_i/(1-4p_i)}{1 + 3\sum_i p_i/(1-4p_i)}. \tag{21}$$

This interpolates continuously between the dominant-atom boundary and the uniform-simplex maximum. It also supplies a direct witness distribution over good pairs, rather than only a minimax value.

Lemma 3 gives an intrinsic-dimension version. If a dataset has affine dimension r, regardless of its ambient dimension, two examples achieve

$$\frac{L_D^*(2)}{L_D^*} \leq 1 + \max\left\{\frac{1}{3}, \frac{r-1}{2r}\right\}. \tag{22}$$

The bound is sharp for every r. The exact factor approaches 3/2 from below, quantifying the finite-dimensional improvement over independent sampling.

The certificate has a simple semidefinite interpretation. For masses p, let $\Sigma_p = \mathrm{diag}(p) - pp^\top$. A distribution on empirical two-point frequencies q proves a uniform geometric bound C whenever

$$\mathbb{E}[(q-p)(q-p)^\top] \preceq C\Sigma_p.$$

The star construction attains equality with $C = 1/3$. In the diffuse case, Equation (14) is the same certificate after quotienting out the centered direction p.

This explains why the pair distribution can be chosen without knowing the geometry of the support.

The exact profile also gives a stability statement for the diffuse branch. For $m \geq 5$, strict convexity in (15) implies that any sequence of diffuse mass vectors whose minimax values approach $(m-2)/(2(m-1))$ must converge, up to permutation, to the uniform vector. Together with Proposition 7, near-worst minimal-support instances must approach the regular simplex in both masses and Gram geometry.

6 Relation to Approximate Carathéodory Theory

Approximate Carathéodory theorems replace a convex combination by a sparse or equal-weight combination. Maurey's empirical method gives the familiar $O(1/\sqrt{n})$ Euclidean error by independent sampling. Later work gives constructive algorithms, tight dimension-free rates, and discrepancy-based extensions (Mirrokni et al., 2017; Combettes and Pokutta, 2023; Reis and Rothvoss, 2022).

The normalization is decisive. Those results control error by a containing radius or diameter. Our denominator is the variance under the particular barycentric weights, which can be much smaller than the squared radius when a distant vertex has low mass. The rare-point obstruction in Section 4 is exactly this gap.

The closest result is the sharp dimension-dependent theorem of Tinarrage (2026). For a simplex in the unit ball, the worst squared distance to a two-fold barycenter is $(d-1)/(2d)$ for $d \geq 2$, attained by a regular simplex. That theorem settles the radius-normalized problem through a Gram-matrix optimization. It does not imply a variance-normalized bound. Our formula shows precisely what changes after variance normalization. The same regular simplex remains extremal for $d \geq 4$, while a one-dimensional rare point dominates in dimensions at most three.

Regular simplices also uniquely maximize variance among distributions with a fixed support diameter (Lim and McCann, 2022). That isodiametric result varies the distribution and support under a diameter constraint. Here the variance is fixed, pair midpoints are the admissible decisions, and the extremizer changes with dimension.

Regular-grid rounding on the standard simplex has also been solved for broad classes of coordinate norms (Bomze et al., 2014). In our formulation, the norm on barycentric error is induced by the unknown support Gram matrix, while the normalization is induced by its weighted trace. Proposition 4 makes this adversarial metric explicit.

7 Discussion

Theorem 1 resolves one complete column of the data-selection table and gives a short analytic proof of its previously known planar entry. The proof identifies a dominant-or-diffuse dichotomy. Dominant atoms are handled by a star coupling. Diffuse masses are handled by a complete-graph coupling, whose worst case follows from scalar convexity.

For budgets larger than two, the admissible averages form a finer simplex grid. The same second-moment viewpoint leads to distributions over count vectors, but pair weights are replaced by higher-order inclusion probabilities. Determining their sharp variance-normalized constants remains open. The present result suggests that any full formula must retain at least two competing families: rare lower-dimensional distributions and balanced regular simplices.

References

Bomze, I. M., Gollowitzer, S., and Yıldı rım, E. A. (2014). Rounding on the standard simplex: Regular grids for global optimization. *Journal of Global Optimization*, 59(2–3):243–258.

Combettes, C. W. and Pokutta, S. (2023). Revisiting the approximate Carathéodory problem via the Frank–Wolfe algorithm. *Mathematical Programming*, 197(1):191–214.

Dasgupta, S., Hsu, D., Poulis, S., and Zhu, X. (2019). Teaching a black-box learner. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1547–1555. PMLR.

Feldman, D. (2020). Core-sets: An updated survey. *arXiv preprint arXiv:2011.09384*.

Hanneke, S., Moran, S., Shlimovich, A., and Yehudayoff, A. (2025a). Data selection for ERMs. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 2634–2665. PMLR.

Hanneke, S., Moran, S., Shlimovich, A., and Yehudayoff, A. (2025b). Open problem: Data selection for regression tasks. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 6225–6229. PMLR.

Lim, T. and McCann, R. J. (2022). Geometrical bounds for variance and recentered moments. *Mathematics of Operations Research*, 47(1):286–296.

Mirrokni, V., Paes Leme, R., Vladu, A., and Wong, S. C.-w. (2017). Tight bounds for approximate Carathéodory and beyond. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2440–2448. PMLR.

Reis, V. and Rothvoss, T. (2022). Approximate Carathéodory bounds via discrepancy theory. *arXiv preprint arXiv:2207.03614*.

Tinarrage, R. (2026). Sharp approximate Carathéodory theorem and application to iterated Delaunay refinement. *arXiv preprint arXiv:2606.25854*.

A Pointers to the Supplement

The supplement gives a standalone proof of the atomic theorem, a matrix form of both antithetic certificates, the intrinsic-dimension reduction, and additional details on sharpness and equality.