hexagon:2611.00511v1

The planar data-selection constants $F(2,n) = 2n/(2n-1)$, including $F(2,3) = 6/5$ and $F(2,2) = 4/3$

Samuil Petkov

Department of Physics, École normale supérieure, Université PSL, Paris, France

samuil.petkov@phys.ens.psl.eu

Abstract

For a finite multiset $D \subset \mathbb{R}^2$ let L_D^* be the least mean squared distance from a point to D, and let $L_D^*(n)$ be the least such loss attained at the average of n points selected from D with repetition. Hanneke, Moran, Shlimovich and Yehudayoff asked for $F(d,n) = \sup_D L_D^*(n)/L_D^*$. We prove $F(2,n) = 2n/(2n-1)$ for every $n \geq 2$: every finite planar multiset satisfies $L_D^*(n) \leq \frac{2n}{2n-1} L_D^*$, with equality for $2n-1$ copies of $(0,0)$ and one copy of $(1,0)$. In particular $F(2,3) = 6/5$, and $F(2,2) = 4/3$ receives a short proof. The proof reduces to three support points by a mean-preserving Carathéodory argument and then exhibits an explicit random selection of n support points, supported on one small triangle of the selection grid, whose second moment about the population weights is dominated by $\frac{1}{2n-1}$ times the categorical covariance. For $n \in \{2,3\}$ we also give random selections attaining this second moment exactly.

Mathematics Subject Classification (2020). Primary 68Q32; Secondary 62H12, 52A40, 60E15.
Keywords. data selection, mean estimation, worst-case approximation ratio, Carathéodory theorem, selection certificates, COLT open problem.

1 Problem and context

Let $D = \{z_1, \dots, z_N\} \subset \mathbb{R}^d$ be a finite multiset and

$$L_D(h) = \frac{1}{N} \sum_{r=1}^{N} \|h - z_r\|^2, \qquad L_D^* = \min_{h \in \mathbb{R}^d} L_D(h), \qquad L_D^*(n) = \min_{x_1, \dots, x_n \in D} L_D\Big(\frac{1}{n} \sum_{k=1}^{n} x_k\Big).$$

The selected points $x_1, \dots, x_n$ range over D independently, so a point may be selected more often than it occurs in D; equivalently, $(x_1, \dots, x_n)$ is any sequence in $\operatorname{supp}(D)$, the set of distinct points of D (this is the convention of [2, §7.2]). Question 1 of the COLT 2025 open-problem note [1] asks for

$$F(d,n) = \sup_{D \subset \mathbb{R}^d} \frac{L_D^*(n)}{L_D^*},$$

where we use the conventions $0/0 = 1$ and that D ranges over nonempty finite multisets. The note summarizes the known values as follows; all entries except $d = n = 2$ are due to [2], and the authors describe their derivation of the $d = n = 2$ entry as "somewhat tedious and computational" [1, §2].

$d \backslash n$	1	2	3	n	$n \to \infty$
1	2	4/3	6/5	$2n/(2n-1)$	1
2	2	4/3	?	?	1
3	2	?	?	?	1
$d \to \infty$	2	3/2	4/3	$(n+1)/n$	undefined

For every d and n one has $2n/(2n-1) \leq F(d,n) \leq (n+1)/n$ [2, §7.5]; the lower bound comes from the one-dimensional example recalled in Proposition 13.

Theorem 1. *Let $n \geq 2$. Every finite multiset $D \subset \mathbb{R}^2$ satisfies $L_D^*(n) \leq \frac{2n}{2n-1} L_D^*$, with equality for $2n-1$ copies of $(0,0)$ and one copy of $(1,0)$. Hence $F(2,n) = 2n/(2n-1)$, and the supremum is attained.*

Together with $F(2,1) = 2$ from the table, Theorem 1 determines the row $d = 2$ of the table. Its cases $n = 3$ and $n = 2$ read as follows.

Theorem 2. *Every finite multiset $D \subset \mathbb{R}^2$ satisfies $L_D^*(3) \leq \frac{6}{5} L_D^*$, with equality for five copies of $(0,0)$ and one copy of $(1,0)$. Hence $F(2,3) = 6/5$, and the supremum is attained.*

Theorem 3. *Every finite multiset $D \subset \mathbb{R}^2$ satisfies $L_D^*(2) \leq \frac{4}{3} L_D^*$, with equality for three copies of $(0,0)$ and one copy of $(1,0)$. Hence $F(2,2) = 4/3$.*

The value $F(2,2) = 4/3$ is already stated in [1]; Theorem 3 gives a short proof of it. Other questions from [1], on weighted data selection for linear regression, are treated in [3, 4]; those works do not concern the mean-estimation constants $F(d,n)$ studied here. For $n \in \{2,3\}$ we give, besides the general argument, certificates that are exact in the sense of (5) below. Throughout, $n \geq 2$ and

$$r_n := \frac{1}{2n-1}.$$

2 Reduction to three support points

Write a finite multiset as a probability law $P = \{(z_i, p_i)\}_{i=1}^m$ on its distinct points, $p_i = n_i/N$, and set

$$\mu(P) = \sum_i p_i z_i, \qquad V(P) = \sum_i p_i \|z_i - \mu(P)\|^2, \qquad q_n(P) = \min_{x_1,\dots,x_n \in \operatorname{supp} P} \Big\| \frac{1}{n} \sum_{k=1}^n x_k - \mu(P) \Big\|^2.$$

These definitions make sense for any finitely supported law with real weights. Expanding $h - z_i = (h - \mu) - (z_i - \mu)$ and using $\sum_i p_i (z_i - \mu) = 0$ gives

$$L_D(h) = V(P) + \|h - \mu(P)\|^2 \qquad (h \in \mathbb{R}^2). \tag{1}$$

Hence $L_D^* = V(P)$, attained only at $h = \mu(P)$, and $L_D^*(n) = V(P) + q_n(P)$. If $V(P) = 0$ then D is a single repeated point and $L_D^*(n) = L_D^* = 0$. If $V(P) > 0$ then

$$\frac{L_D^*(n)}{L_D^*} \leq 1 + r_n = \frac{2n}{2n-1} \quad \Longleftrightarrow \quad q_n(P) \leq r_n V(P). \tag{2}$$

Lemma 4 (Mean-preserving support reduction)**.** *Let P be a finitely supported probability law on $\mathbb{R}^2$. There is a probability law P' with $\operatorname{supp} P' \subseteq \operatorname{supp} P$, $|\operatorname{supp} P'| \leq 3$, $\mu(P') = \mu(P)$ and $V(P') \leq V(P)$. Consequently $q_n(P) \leq q_n(P')$ for every n.*

Proof. Let $z_1, \dots, z_m$ be the support of P and $\mu = \mu(P)$. The set

$$\mathcal{P} = \Big\{ a \in \mathbb{R}^m : \ a \geq 0, \ \sum_i a_i = 1, \ \sum_i a_i z_i = \mu \Big\}$$

contains the weight vector p of P and is compact. The linear function $J(a) = \sum_i a_i \|z_i\|^2$ therefore attains its minimum on $\mathcal{P}$; choose a minimizer a^* with the fewest positive coordinates. Suppose a^* had at least four positive coordinates. The corresponding columns $(1, z_i) \in \mathbb{R}^3$ are linearly dependent, so there is $u \neq 0$ supported on $\{i : a_i^* > 0\}$ with $\sum_i u_i = 0$ and $\sum_i u_i z_i = 0$. For small $t > 0$ both $a^* \pm tu$ lie in $\mathcal{P}$, so minimality gives $\pm t J(u) \geq 0$, that is $J(u) = 0$. Since $u \neq 0$ and $\sum_i u_i = 0$, u has a negative coordinate; with $t_0 = \min_{u_i < 0} a_i^*/(-u_i) > 0$, the vector $a^* + t_0 u$ lies in $\mathcal{P}$, has $J(a^* + t_0 u) = J(a^*)$, and has fewer positive coordinates than a^*, a contradiction. So the law P' with weights a^* has at most three support points, all in $\operatorname{supp} P$, and mean μ. For every $a \in \mathcal{P}$,

$$\sum_i a_i \|z_i - \mu\|^2 = J(a) - 2\Big\langle \sum_i a_i z_i, \mu \Big\rangle + \|\mu\|^2 = J(a) - \|\mu\|^2,$$

so $V(P') = J(a^*) - \|\mu\|^2 \leq J(p) - \|\mu\|^2 = V(P)$. Finally, every selection legal for P' is legal for P and both laws have the same mean, so $q_n(P) \leq q_n(P')$. □

By Lemma 4, if $q_n(P') \le r_n V(P')$ holds for all laws with at most three support points, then for every finitely supported P

$$q_n(P) \le q_n(P') \le r_n V(P') \le r_n V(P). \tag{3}$$

This reduction is of Carathéodory type; compare the use of Carathéodory's theorem for convex functions in [2, §7.1].

3 Grid certificates

Let P have support contained in three labelled points v_1, v_2, v_3 with weights $p = (p_1, p_2, p_3)$, $p_i \ge 0$, $\sum_i p_i = 1$; labels with $p_i = 0$ are padding and may not be selected. Put $w_i = v_i - \mu(P)$, so that $\sum_i p_i w_i = 0$ and $V(P) = \sum_i p_i \|w_i\|^2$. A selection of n labels with multiplicities $k = (k_1, k_2, k_3)$, $\sum_i k_i = n$, is encoded by the grid point $s = k/n$ in

$$G_n = \Big\{ \tfrac{1}{n} k : \ k \in \mathbb{Z}^3_{\ge 0}, \ k_1 + k_2 + k_3 = n \Big\},$$

and its selected mean, centred at $\mu(P)$, is $\sum_i s_i w_i = \sum_i (s_i - p_i) w_i$. Let $C_p = \operatorname{diag}(p) - pp^{\mathsf T}$ be the covariance matrix of a categorical draw with probabilities p, and write $A \preceq B$ if $B - A$ is positive semidefinite.

Definition. A *certificate* for (n, p) is a probability distribution λ on G_n such that $\lambda_s = 0$ whenever $s_i > 0$ for some i with $p_i = 0$, and

$$M_\lambda := \sum_{s \in G_n} \lambda_s \, (s - p)(s - p)^{\mathsf T} \preceq r_n \, C_p. \tag{4}$$

It is *exact* if

$$M_\lambda = r_n \, C_p. \tag{5}$$

Lemma 5 (Extraction)**.** *If a certificate exists for (n, p), then $q_n(P) \le r_n V(P)$.*

Proof. Let S be a random grid point with law λ and $Y = \sum_i S_i w_i = \sum_i (S_i - p_i) w_i$. Let Γ be the Gram matrix $\Gamma_{ij} = \langle w_i, w_j \rangle$, which is positive semidefinite. Then

$$\mathbb{E}\|Y\|^2 = \sum_{i,j} \mathbb{E}[(S_i - p_i)(S_j - p_j)] \, \Gamma_{ij} = \operatorname{tr}(M_\lambda \Gamma), \qquad \operatorname{tr}(C_p \Gamma) = \sum_i p_i \|w_i\|^2 - \Big\| \sum_i p_i w_i \Big\|^2 = V(P).$$

The trace of a product of two positive semidefinite matrices is nonnegative, so (4) gives $\operatorname{tr}\big((r_n C_p - M_\lambda)\Gamma\big) \ge 0$, that is $\mathbb{E}\|Y\|^2 \le r_n V(P)$, with equality for exact certificates. Every s with $\lambda_s > 0$ encodes a legal selection of n support points, and $\mathbb{E}\|Y\|^2$ is a λ-average of their squared errors. Hence some legal selection has squared error at most $r_n V(P)$. □

No condition $\mathbb{E}S = p$ is required. The following identity reduces (5) to first and second moments.

Lemma 6. *Let λ be a probability distribution on G_n, $m = \mathbb{E}S$ and $Q = \mathbb{E}[SS^{\mathsf T}]$. If for some $c \in \mathbb{R}^3$*

$$m = \tfrac{1 + r_n}{2} \, p + c \qquad \text{and} \qquad Q = r_n \operatorname{diag}(p) + cp^{\mathsf T} + pc^{\mathsf T},$$

then λ satisfies (5)*.*

Proof. $\mathbb{E}[(S-p)(S-p)^{\mathsf T}] = Q - mp^{\mathsf T} - pm^{\mathsf T} + pp^{\mathsf T} = r_n \operatorname{diag}(p) + cp^{\mathsf T} + pc^{\mathsf T} - (1 + r_n) pp^{\mathsf T} - cp^{\mathsf T} - pc^{\mathsf T} + pp^{\mathsf T} = r_n C_p$. □

Write e_1, e_2, e_3 for the vertices of G_n and assume, after relabelling, $p_1 \ge p_2 \ge p_3$; a simultaneous permutation of labels maps G_n to itself and conjugates both sides of (5) by the same permutation matrix.

Lemma 7 (Corner certificate, any $n \ge 2$)**.** *Let $\kappa = n^2/(2n-1)$ and suppose $p_1 \ge \big(\frac{n-1}{n}\big)^2$. The weights*

$$\lambda\big(e_1 + \tfrac{1}{n}(e_2 - e_1)\big) = \kappa p_2, \qquad \lambda\big(e_1 + \tfrac{1}{n}(e_3 - e_1)\big) = \kappa p_3, \qquad \lambda(e_1) = 1 - \kappa(p_2 + p_3)$$

form an exact certificate for (n, p).

Proof. The first two weights are nonnegative, and $1-\kappa(1-p_1) = \big(n^2p_1-(n-1)^2\big)/(2n-1) \geq 0$ by hypothesis; the weights sum to one. A label with $p_i = 0$ $(i = 2, 3)$ receives weight zero. Using $\kappa/n = n/(2n-1) = (1+r_n)/2$ and $\kappa/n^2 = r_n$,

$$m_1 = 1 - \tfrac{\kappa}{n}(p_2+p_3) = \tfrac{np_1+n-1}{2n-1}, \qquad m_2 = \tfrac{n}{2n-1}p_2, \qquad m_3 = \tfrac{n}{2n-1}p_3,$$

so $m = \frac{1+r_n}{2}p + c$ with $c = \frac{n-1}{2n-1}e_1$. For the second moments,

$$Q_{22} = \tfrac{\kappa}{n^2}p_2 = r_np_2, \quad Q_{33} = r_np_3, \quad Q_{23} = 0, \quad Q_{12} = \tfrac{\kappa(n-1)}{n^2}p_2 = \tfrac{n-1}{2n-1}p_2, \quad Q_{13} = \tfrac{n-1}{2n-1}p_3,$$

$$Q_{11} = 1 - \kappa(p_2+p_3)\Big(1 - \big(\tfrac{n-1}{n}\big)^2\Big) = 1 - (1-p_1) = p_1.$$

These are exactly the entries of $r_n\operatorname{diag}(p) + cp^{\mathsf{T}} + pc^{\mathsf{T}} = r_n\operatorname{diag}(p) + \frac{n-1}{2n-1}(e_1p^{\mathsf{T}} + pe_1^{\mathsf{T}})$; for instance its $(1,1)$ entry is $r_np_1 + \frac{2(n-1)}{2n-1}p_1 = p_1$ and its $(2,3)$ entry is 0. Lemma 6 applies. □

Lemma 8 (Central certificate, $n = 3$)**.** *Let $n = 3$, $p_1 \geq p_2 \geq p_3$ and $p_2 \geq \frac{1}{6}$. Writing $s_{abc} = \frac{1}{3}(a,b,c)$, the weights*

$$\lambda(s_{111}) = \tfrac{9}{5}p_3, \qquad \lambda(s_{120}) = \tfrac{9}{5}p_2 - \tfrac{3}{10}, \qquad \lambda(s_{210}) = \tfrac{9}{5}p_1 - \tfrac{3}{5}, \qquad \lambda(s_{300}) = \tfrac{1}{10}$$

form an exact certificate for $(3, p)$.

Proof. Nonnegativity: $p_3 \geq 0$; $p_2 \geq \frac{1}{6}$; and $p_1 \geq \frac{1}{3}$ by sortedness. The sum is $\frac{9}{5} - \frac{3}{10} - \frac{3}{5} + \frac{1}{10} = 1$. Only s_{111} uses label 3, and its weight vanishes if $p_3 = 0$; labels $1, 2$ have positive weight since $p_1 \geq p_2 \geq \frac{1}{6}$. The first moments are

$$m_1 = \tfrac{1}{3}\lambda_{111} + \tfrac{1}{3}\lambda_{120} + \tfrac{2}{3}\lambda_{210} + \lambda_{300} = \tfrac{3p_1+1}{5}, \qquad m_2 = \tfrac{1}{3}\lambda_{111} + \tfrac{2}{3}\lambda_{120} + \tfrac{1}{3}\lambda_{210} = \tfrac{3p_2+1}{5},$$
$$m_3 = \tfrac{1}{3}\lambda_{111} = \tfrac{3}{5}p_3,$$

so $m = \frac{3}{5}p + c$ with $c = \frac{1}{5}(1,1,0)$, and $\frac{3}{5} = \frac{1+r_3}{2}$. The second moments are

$$Q_{11} = \tfrac{1}{9}\lambda_{111} + \tfrac{1}{9}\lambda_{120} + \tfrac{4}{9}\lambda_{210} + \lambda_{300} = \tfrac{3}{5}p_1, \qquad Q_{22} = \tfrac{1}{9}\lambda_{111} + \tfrac{4}{9}\lambda_{120} + \tfrac{1}{9}\lambda_{210} = \tfrac{3}{5}p_2,$$
$$Q_{12} = \tfrac{1}{9}\lambda_{111} + \tfrac{2}{9}\lambda_{120} + \tfrac{2}{9}\lambda_{210} = \tfrac{p_1+p_2}{5}, \qquad Q_{13} = Q_{23} = Q_{33} = \tfrac{1}{9}\lambda_{111} = \tfrac{1}{5}p_3.$$

With $c = \frac{1}{5}(1,1,0)$ and $r_3 = \frac{1}{5}$, the matrix $r_3\operatorname{diag}(p) + cp^{\mathsf{T}} + pc^{\mathsf{T}}$ has entries $(1,1)$: $\frac{3p_1}{5}$, $(2,2)$: $\frac{3p_2}{5}$, $(3,3)$: $\frac{p_3}{5}$, $(1,2)$: $\frac{p_1+p_2}{5}$, $(1,3)$: $\frac{p_3}{5}$, $(2,3)$: $\frac{p_3}{5}$, which agree with Q. Lemma 6 applies. □

Proposition 9. *For $n \in \{2, 3\}$ an exact certificate exists for every p in the simplex.*

Proof. Sort $p_1 \geq p_2 \geq p_3$. For $n = 2$ the hypothesis of Lemma 7 is $p_1 \geq \frac{1}{4}$, which holds because $p_1 \geq \frac{1}{3}$. For $n = 3$: if $p_2 \geq \frac{1}{6}$ use Lemma 8; otherwise $p_3 \leq p_2 < \frac{1}{6}$, so $p_1 > \frac{2}{3} \geq \frac{4}{9} = \big(\frac{2}{3}\big)^2$ and Lemma 7 applies. □

For $n = 2$ the certificate of Lemma 7 is explicit: weight $\frac{4p_1-1}{3}$ on e_1 and $\frac{4}{3}p_2$, $\frac{4}{3}p_3$ on the midpoints $\frac{1}{2}(e_1+e_2)$, $\frac{1}{2}(e_1+e_3)$. For $n = 3$ and $p_2 < \frac{1}{6}$ it is $\frac{9}{5}p_2$ on s_{210}, $\frac{9}{5}p_3$ on s_{201} and $\frac{9p_1-4}{5}$ on s_{300}.

4 Certificates for every n

For $k \in \mathbb{Z}^3_{\geq 0}$ with $k_1+k_2+k_3 = n-1$, the three grid points $(k+e_i)/n$, $i = 1, 2, 3$, are the vertices of a small triangle of G_n that is a translate of $\frac{1}{n}\{e_1, e_2, e_3\}$.

Lemma 10 (Triangle certificate, any $n \geq 2$)**.** *Let $k \in \mathbb{Z}^3_{\geq 0}$ with $k_1+k_2+k_3 = n-1$ and $(n-1)k_i \leq n^2p_i$ for $i = 1, 2, 3$. Put*

$$t_i = \tfrac{1}{n}(k+e_i) \in G_n, \qquad \beta_i = \frac{n^2p_i - (n-1)k_i}{2n-1} \qquad (i = 1, 2, 3),$$

and let $\lambda = \sum_i \beta_i\delta_{t_i}$. Then λ is a certificate for (n, p); more precisely

$$M_\lambda = \sum_{i=1}^{3} \beta_i(t_i-p)(t_i-p)^{\mathsf{T}} = r_nC_p - \frac{r_n}{n^2}R_k, \qquad R_k := (n-1)\operatorname{diag}(k) - kk^{\mathsf{T}} \succeq 0. \tag{6}$$

Proof. The weights are nonnegative by hypothesis and sum to $\big(n^2-(n-1)^2\big)/(2n-1)=1$. If $p_i=0$, the hypothesis forces $k_i=0$; then $\beta_i=0$, and t_j has i-th coordinate 0 for $j\neq i$, so the padding condition holds.

Put $u=k/n$ and $a=u-p$, so that $t_i-p=a+\frac{1}{n}e_i$. Since $\sum_i\beta_i=1$, $\sum_i\beta_i e_i=\beta$ and $\sum_i\beta_i e_ie_i^{\mathsf T}=\operatorname{diag}(\beta)$,

$$M_\lambda=aa^{\mathsf T}+\tfrac{1}{n}\big(a\beta^{\mathsf T}+\beta a^{\mathsf T}\big)+\tfrac{1}{n^2}\operatorname{diag}(\beta).$$

From $\beta=\big(n^2p-n(n-1)u\big)/(2n-1)$ and $p=u-a$ we get $\frac{1}{n}\beta=r_n\big(np-(n-1)u\big)=r_n(u-na)$ and $\frac{1}{n^2}\beta=r_np-r_n\frac{n-1}{n}u$. Hence, using $1-2nr_n=-r_n$,

$$M_\lambda=r_n\Big(\operatorname{diag}(p)-aa^{\mathsf T}+au^{\mathsf T}+ua^{\mathsf T}-\tfrac{n-1}{n}\operatorname{diag}(u)\Big).$$

On the other hand $pp^{\mathsf T}=(u-a)(u-a)^{\mathsf T}=uu^{\mathsf T}-au^{\mathsf T}-ua^{\mathsf T}+aa^{\mathsf T}$, so

$$r_nC_p=r_n\big(\operatorname{diag}(p)-uu^{\mathsf T}+au^{\mathsf T}+ua^{\mathsf T}-aa^{\mathsf T}\big).$$

Subtracting, $M_\lambda-r_nC_p=r_n\big(uu^{\mathsf T}-\frac{n-1}{n}\operatorname{diag}(u)\big)=-\frac{r_n}{n^2}R_k$. Finally, for $x\in\mathbb{R}^3$, $x^{\mathsf T}R_kx=\big(\sum_i k_i\big)\big(\sum_i k_ix_i^2\big)-\big(\sum_i k_ix_i\big)^2\geq 0$ by the Cauchy–Schwarz inequality, so $R_k\succeq 0$ and (4) holds. □

For $k=(n-1)e_1$ one has $R_k=0$, and Lemma 10 is exactly the corner certificate of Lemma 7; the same holds for $k=(n-1)e_2$ or $(n-1)e_3$ after relabelling. If two coordinates of k are positive, then $R_k\neq 0$ and the certificate is not exact.

Lemma 11. *For every $n\geq 2$ and every p in the simplex there is $k\in\mathbb{Z}^3_{\geq 0}$ with $k_1+k_2+k_3=n-1$ and $(n-1)k_i\leq n^2p_i$ for $i=1,2,3$.*

Proof. Put $x_i=n^2p_i/(n-1)$, so $\sum_i x_i=\frac{n^2}{n-1}=n+1+\frac{1}{n-1}$. Since $\lfloor x_i\rfloor>x_i-1$,

$$\sum_{i=1}^3\lfloor x_i\rfloor>\sum_i x_i-3=n-2+\tfrac{1}{n-1}>n-2,$$

and the left side is an integer, so it is at least $n-1$. Choose integers $0\leq k_i\leq\lfloor x_i\rfloor$ with $k_1+k_2+k_3=n-1$. □

Proposition 12. *For every $n\geq 2$ a certificate exists for every p in the simplex.*

Proof. Combine Lemmas 11 and 10. □

5 Proofs of the theorems

Proposition 13 (Extremal examples [2])**.** *Let D_n consist of $2n-1$ copies of $(0,0)$ and one copy of $(1,0)$. Then $L^*_{D_n}(n)/L^*_{D_n}=2n/(2n-1)$.*

Proof. With $\theta=\frac{1}{2n}$, the mean is $(\theta,0)$ and $L^*_{D_n}=V=\theta(1-\theta)=\frac{2n-1}{4n^2}$. A selection containing j copies of $(1,0)$ has mean $(j/n,0)$, at squared distance $(j/n-\theta)^2$ from the mean; the minimum, at $j\in\{0,1\}$, is $\theta^2=\frac{1}{4n^2}=r_nV$. By (1), $L^*_{D_n}(n)=(1+r_n)V$. □

For $n=3$ this is five copies of $(0,0)$ and one copy of $(1,0)$: mean $(\frac{1}{6},0)$, $V=\frac{5}{36}$, selected means $(j/3,0)$ at squared distances $\frac{1}{36},\frac{1}{36},\frac{1}{4},\frac{25}{36}$, so $L^*_D(3)=\frac{1}{6}=\frac{6}{5}\cdot\frac{5}{36}$. For $n=2$ it is three copies of $(0,0)$ and one copy of $(1,0)$, with $V=\frac{3}{16}$ and $q_2=\frac{1}{16}$.

Proof of Theorems 1, 2 and 3. Let $n\geq 2$ and let $D\subset\mathbb{R}^2$ be a finite multiset with law P. If $V(P)=0$ both losses vanish. Otherwise, Lemma 4 gives P' with at most three support points, all in $\operatorname{supp}D$, the same mean and $V(P')\leq V(P)$. Label the support of P' as v_1,v_2,v_3, padding with zero-weight labels if necessary. Proposition 12 (or, for $n\in\{2,3\}$, Proposition 9) and Lemma 5 give $q_n(P')\leq r_nV(P')$, and (3) gives $q_n(P)\leq r_nV(P)$. By (1) and (2), $L^*_D(n)\leq\frac{2n}{2n-1}L^*_D$, which is $\frac{6}{5}L^*_D$ for $n=3$ and $\frac{4}{3}L^*_D$ for $n=2$. Proposition 13 shows that equality is attained. □

Remark 14. Equality in Theorem 2 is not confined to collinear data. Take four copies of $a = (1, -1)$ and one copy each of $b = (1, 5)$ and $c = (-5, -1)$. The mean is 0 and $V = \frac{1}{6}(4 \cdot 2 + 26 + 26) = 10$. A selection with multiplicities $k_1 + k_2 + k_3 = 3$ has mean $\frac{1}{3}(k_1 a + k_2 b + k_3 c) = (1 - 2k_3,\ 2k_2 - 1)$, whose coordinates are odd integers; its squared norm is at least 2, with equality for instance for $k = (3, 0, 0)$. Thus $q_3 = 2 = V/5$.

Remark 15 (Higher dimensions). Lemmas 5 and 10 extend verbatim to $d+1$ labels, and in $\mathbb{R}^d$ the argument of Lemma 4 reduces to $d+1$ support points. The counting in Lemma 11 then only gives $\sum_i \lfloor x_i \rfloor > n - d + \frac{1}{n-1}$, which is not enough for $d \geq 3$. We make no claim about $F(d, n)$ for $d \geq 3$.

AI assistance disclosure

OpenAI GPT-5.6 Sol, accessed through ChatGPT and Codex, was used to develop and audit the original proof of $F(2, 3) = 6/5$ and its exact verification, and for drafting. Anthropic Claude Opus 5.5, used through Claude Code, condensed the manuscript, found the triangle certificate (Lemma 10, Lemma 11) that gives $F(2, n) = 2n/(2n - 1)$ for every n (Theorem 1), and checked the certificate identities with exact symbolic and rational computation. The author reviewed and takes responsibility for every statement and proof. AI systems are not authors of this work, and the work has not been peer reviewed.

References

[1] S. Hanneke, S. Moran, A. Shlimovich, and A. Yehudayoff, *Open problem: Data selection for regression tasks*, Proceedings of the Thirty Eighth Conference on Learning Theory, Proc. Mach. Learn. Res. **291** (2025), 6225–6229. `https://proceedings.mlr.press/v291/hanneke25e.html`

[2] S. Hanneke, S. Moran, A. Shlimovich, and A. Yehudayoff, *Data selection for ERMs*, Proceedings of the Thirty Eighth Conference on Learning Theory, Proc. Mach. Learn. Res. **291** (2025), 2634–2665; section numbers refer to arXiv:2504.14572v2. `https://proceedings.mlr.press/v291/hanneke25a.html`

[3] G. Zhang, *Exact risk ratios for weighted data selection in linear regression*, preprint (2026), arXiv:2608.28007.

[4] G. Zhang, *Exact recovery thresholds for weighted data selection in vector-valued linear regression*, preprint (2026), arXiv:2608.30254.