ROBUST Q-AGGREGATION UNDER COVARIATE SHIFT WITH PAIRWISE OVERLAP

Anonymous authors
Paper under double-blind review

ABSTRACT

We study squared-loss aggregation of a fixed dictionary $\mathcal{F}$ of M predictors under covariate shift: a labeled sample of size n comes from a source population with input law P, while the target risk is under another input law Q. With known density ratios, the difficulty of a comparison depends on the weights where its two predictors disagree. We capture this dependence through an original-pair overlap coefficient $\Gamma_{\mathcal{F}}$ and construct a robust Q-aggregate from median-of-means (MOM) estimates of risk differences and squared distances. Coordinatewise robust estimates can produce comparison cycles and non-Euclidean distance matrices. Our method repairs the distances by a root-distance ℓ_∞ projection onto the Euclidean distance matrices, then combines the predictors by solving a monotone variational inequality on the simplex. With exact population quantities, this rule is the classical uniform-prior Q-aggregate with mixing parameter $1/2$. With noisy pairwise quantities, the variational form also accepts nonzero comparison cycles. The inputs are the density-ratio weights, a failure probability δ, and a bound B on predictions such that, for a source input-response pair (X, Y), the source conditional second moment satisfies $\mathbb{E}[Y^2 \mid X] \le B^2$ wherever candidate predictors disagree. Without a supplied overlap bound, the resulting convex mixture achieves a leading-one oracle inequality of order $B^2 \max\{1, \Gamma_{\mathcal{F}}\} \log(M/\delta)/n$, including settings in which the global second moment of the density ratio is infinite. Matching lower bounds show the worst-case dependence on overlap, dictionary size, and confidence.

1 INTRODUCTION

Suppose several regression models have been trained, and labels are available from a source population. Which combination should we use for a different target population? Importance weighting adjusts validation for the change in input law, although a few large weights can make ordinary risk estimates unstable. Looking only at the largest weight, however, overlooks the models being compared. **A large weight is harmless for comparing two models at an input where they agree.** Their predictions, and therefore their squared losses, coincide at that input.

Let P and Q be the source and target input laws, let w be their density ratio on the candidate disagreement set, and let $\mathcal{F} = \{f_1, \dots, f_M\}$ be the candidate dictionary of M predictors. We use the pairwise coefficient

$$\Gamma_{\mathcal{F}} = \max_{i,j:\, \mathbb{E}_Q[(f_i - f_j)^2] > 0} \frac{\mathbb{E}_P[w^2 (f_i - f_j)^2]}{\mathbb{E}_Q[(f_i - f_j)^2]}. \tag{1}$$

The numerator measures variability from weighting a disagreement, and the denominator measures its size under the target law. Their ratio captures the sampling cost and can remain finite when $\mathbb{E}_P w^2$ diverges. We seek a sharp oracle inequality governed by this dictionary-dependent quantity, without supplying its bound to the algorithm.

Two features of squared-loss aggregation guide the construction. Comparing losses cancels the common term wY^2 (where Y denotes the response), which would otherwise require stronger response moments to control an individual weighted risk. The remaining pairwise difference has variance governed by (1). Robust estimation creates a separate difficulty: writing T_{ij} for the robust estimate of the target-risk difference between f_i and f_j, medians taken pair by pair can yield

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.

$T_{ij}+T_{jk}+T_{ki} \neq 0$ (i.e., the three estimated comparisons around a cycle need not sum to zero). Such a cycle prevents us from representing all comparisons as differences of one empirical-risk vector, as ordinary Q-aggregation would require.

Our construction works directly with the comparison matrix, including its cycles. We estimate target squared distances from the same observations and repair them to a nearby Euclidean distance matrix using semidefinite feasibility and bisection. The repaired Gram matrix supplies the curvature needed for an affine variational inequality (VI), which can be solved as a convex gap-minimization quadratic program (QP). The resulting weights combine the original predictors and may spread mass across the dictionary.

The link to Q-aggregation also explains the risk guarantee. With population risk differences and distances, the VI becomes the optimality condition of the uniform-prior criterion at mixing parameter $\nu = 1/2$ (Dai et al., 2012; Lecué & Rigollet, 2014). We extend this criterion to robust comparisons under covariate shift, using only statistics indexed by the original dictionary pairs. The combination of distance repair and the VI preserves a leading-one oracle bound even though the estimates can have cycles and the overlap scale is unknown. On one simultaneous event, with n labeled source observations, failure probability δ, and a bound B controlling predictions and conditional response moments, the excess target risk is at most

$$524B^2 \max\{1, \Gamma_{\mathcal{F}}\}/\lfloor n/k \rfloor, \qquad k \asymp \log(M/\delta).$$

The method therefore adapts to the overlap scale that appears in its risk bound. We also give examples where disagreement-based overlap is substantially smaller than global or enlarged-class measures, together with lower bounds showing the sharp worst-case rate dependence.

2 Related work

Our starting point is the curvature mechanism of Q-aggregation. Dai et al. (2012) introduced the criterion to obtain deviation-optimal, leading-one model-selection bounds, and Lecué & Rigollet (2014) established its random-design theory for bounded strongly convex losses. The uniform-prior optimizer can be independent of the noise or boundedness scale used in the proof. We retain that feature while replacing bounded-loss concentration and preserving curvature when robust comparisons fail to form one risk vector. The adaptation to doubly robust causal losses by Lan & Syrgkanis (2024) offers another example of using the Q criterion beyond ordinary supervised validation.

A close comparison is the weak-moment finite-dictionary result of Mourtada et al. (2022). Their estimator combines pairwise median-of-means (MOM) tests, survivor geometry, and robust selection over survivor midpoints, building on the preselection and midpoint methods of Mendelson (2017) and the tournament approach of Lugosi & Mendelson (2020). Their fast rates use distance information to control noisy comparisons. More recently, Høgsgaard & Paudice (2025) studied uniform MOM estimation under finite p-th moments, for $1 < p \leq 2$. Our pairwise estimates leave a further task: recovering geometry that supports aggregation. Applying the theorem of Mourtada et al. (2022) after the change of variables $Y' = \sqrt{w}Y$, $f_i' = \sqrt{w}f_i$ would, however, require a uniform bound on $w(x)\mathbb{E}[Y^2 \mid X = x]$. Our disagreement-based assumption permits that quantity to be unbounded. We instead estimate the geometry of the original predictors and use it to control the curvature of a convex mixture. Generic min–max MOM optimization (Lecué & Lerasle, 2020) provides a broader framework with multiplier and small-ball conditions over its optimization class. For a finite dictionary, norm equivalence supplies finite constants on its span. The quantitative issue is that those constants can grow arbitrarily relative to $\Gamma_{\mathcal{F}}$. Section 5 illustrates the distinction between original pairs and generated directions.

From the transfer-learning perspective, importance-weighted validation (Sugiyama et al., 2007) corrects the input-law change, while bounds based on a uniform weight or global divergence (Cortes et al., 2010) summarize the mismatch between the two distributions. The models themselves enter more directly through task-dependent quantities, such as the classification transfer exponents of Kpotufe & Martinet (2021). In regression, Amortila et al. (2024) study approximation-error guarantees under bounded density ratios. Bounded ratios are unnecessary in some other settings. For well-specified parametric models, Ge et al. (2024) obtain target-risk guarantees for source-only maximum likelihood estimation. In nonparametric quantile regression, Feng et al. (2024) analyze learned and truncated importance weights under global moment and smoothness assumptions. These results

highlight the role of model structure and the target loss in determining the cost of shift. We focus on the excess risk relative to the best member of a given dictionary, for which pairwise disagreements provide a natural way to localize the weighting cost.

The optimization tools have a separate, well-established history. Centering a Euclidean distance matrix produces a positive semidefinite (PSD) Gram matrix, a characterization that supports approximation by semidefinite programming (SDP) (Alfakih et al., 1999; Dokmanić et al., 2015). Our repair uses the root-ℓ_∞ objective because its error can be inserted directly into the pairwise risk calculation. Likewise, convex reformulations of affine VIs are available from Aghassi et al. (2006). The statistical construction couples these tools so that the optimization acts on robust original-pair information and the resulting bound depends on original-pair overlap.

3 Setup and pairwise overlap

Fix $B > 0$. Each fixed measurable predictor $f_i : \mathcal{X} \to [-B, B]$ (where $\mathcal{X}$ is the input space) is independent of the validation sample. We observe n iid (independent and identically distributed) source pairs (X_t, Y_t) with input marginal P. The target input law is Q. When an expectation contains Y, its subscript denotes the corresponding source or target joint law. In particular, $R_Q(f) = \mathbb{E}_Q[(f(X) - Y)^2]$.

Let $\mathcal{D} = \{x : f_i(x) \neq f_j(x) \text{ for some } i, j\}$. Assume $Q|_{\mathcal{D}} \ll P$ and that the algorithm knows $w = d(Q|_{\mathcal{D}})/dP$, extended by zero off $\mathcal{D}$. Source and target conditional response means agree Q-almost everywhere on $\mathcal{D}$ and call their common version $\mu(x)$. We require

$$\mathbb{E}_P[Y^2 \mid X = x] \le B^2 \quad (P\text{-a.e. } x \in \mathcal{D}), \qquad \mathbb{E}_Q Y^2 < \infty. \tag{2}$$

Ordinary covariate shift, with the same conditional response law in both populations, satisfies these assumptions. For squared-loss comparisons, equality of the conditional means is enough because the conditional variance cancels. Jensen's inequality then gives $|\mu| \le B$ on target disagreements.

Write $R_i = R_Q(f_i)$, $R_* = \min_i R_i$, and $d_{ij}^2 = \mathbb{E}_Q[(f_i - f_j)^2]$. We use the convention that the maximum in (1) is zero when every pair has zero target distance. The rate statement assumes $\Gamma_{\mathcal{F}} < \infty$, while the procedure uses the supplied weights and B. Restricting the density ratio to $\mathcal{D}$ gives

$$R_i - R_j = \mathbb{E}_P[w(f_i - f_j)(f_i + f_j - 2Y)], \qquad d_{ij}^2 = \mathbb{E}_P[w(f_i - f_j)^2]. \tag{3}$$

Appendix A.1 derives both identities, including pairs at zero target distance. Restricting the weights is useful when the target has mass outside the source support: on any such region where all predictors agree, every convex combination makes the same prediction.

4 Robust pairwise Q-aggregation

The algorithm has two stages. It first puts the robust distance estimates into a common Euclidean geometry, then uses that geometry to aggregate the robust loss comparisons. Both stages use the same validation sample, justified by simultaneous pairwise bounds. Algorithm 1 in Appendix C.3 collects the inputs, steps, and stopping conditions. For $M \ge 2$ and $0 < \delta < 1$, let k be the smallest odd integer at least $8 \log(2M^2/\delta)$. When $n \ge k$, put $s = \lfloor n/k \rfloor$ and split the first ks observations into k blocks of size s. For a scalar variable V, let $\mathrm{MOM}_k(V)$ be the median of its block means.

4.1 Pairwise statistics and Euclidean repair

For each unordered pair in the dictionary, estimate its risk difference and squared distance by

$$T_{ij} = \mathrm{MOM}_k(w(f_i - f_j)(f_i + f_j - 2Y)), \tag{4}$$

$$H_{ij} = \mathrm{MOM}_k\big(w(f_i - f_j)^2\big), \qquad \widehat{d}_{ij} = \sqrt{\min\{4B^2, \max\{0, H_{ij}\}\}}. \tag{5}$$

Let $T = (T_{ij})_{i,j=1}^M$ be the comparison matrix obtained by setting $T_{ji} = -T_{ij}$ and $T_{ii} = 0$. Let $\widehat{d} = (\widehat{d}_{ij})_{i,j=1}^M$ be the symmetric root-distance estimate matrix obtained by setting $\widehat{d}_{ji} = \widehat{d}_{ij}$ and $\widehat{d}_{ii} = 0$. Each pair therefore has consistent forward and reverse comparisons, while comparisons

around a cycle may still sum to a nonzero value. Clipping the squared-distance estimate to $[0, 4B^2]$ uses the prediction bound and cannot increase its error relative to the true squared distance.

Let $J = I - \mathbf{1}\mathbf{1}^\top/M$ where I is the $M \times M$ identity matrix and $\mathbf{1}$ is the all-ones vector. Define

$$\mathcal{E}_B = \left\{ D : D = D^\top,\ D_{ii} = 0,\ 0 \le D_{ij} \le 4B^2,\ -\tfrac{1}{2} JDJ \succeq 0 \right\}. \tag{6}$$

The set $\mathcal{E}_B$ consists of squared Euclidean distance matrices (EDMs) with diameter at most $2B$ and is compact and convex. To see why the true matrix belongs to this set, regard each f_i as a point in the Hilbert space $L_2(Q)$, where $\|g\|^2_{L_2(Q)} = \mathbb{E}_Q[g(X)^2]$: the finite collection has a Euclidean realization of dimension at most $M - 1$ (Appendix B.1). We measure the fit to the robust estimates in root distances,

$$\rho(D) = \max_{i,j} |\sqrt{D_{ij}} - \widehat{d}_{ij}|, \qquad \rho_* = \min_{D \in \mathcal{E}_B} \rho(D). \tag{7}$$

At a fixed radius r, the root-distance constraint is equivalent to the linear entrywise bounds

$$\max\{0, \widehat{d}_{ij} - r\}^2 \le D_{ij} \le \min\{2B, \widehat{d}_{ij} + r\}^2. \tag{8}$$

For a given r, these entrywise bounds and (6) form an SDP feasibility problem. Increasing r relaxes the bounds, which permits bisection on $[0, 2B]$. Appendix B.2 gives the existence and approximation argument. For a prescribed solver tolerance τ, we retain a feasible $\widetilde{D} \in \mathcal{E}_B$ whose achieved discrepancy satisfies

$$\rho(\widetilde{D}) \le \rho_* + \tau, \qquad 0 < \tau \le B/\sqrt{s}. \tag{9}$$

The root-distance objective is chosen to match the statistical error bound: on the event used below, every $\widehat{d}_{ij}$ is within the same additive error of d_{ij}. A uniform repair radius then transfers that control to every pair in the reconstructed geometry, without estimating the overlap coefficient.

4.2 A MONOTONE VI AND ITS CONVEX GAP PROGRAM

Let $\Delta_M = \{\beta \in \mathbb{R}^M : \beta \ge 0,\ \mathbf{1}^\top\beta = 1\}$ be the probability simplex, and put $A = T - \widetilde{D}/2$. We seek $\beta \in \Delta_M$ satisfying

$$\mathrm{gap}(\beta) := \beta^\top A\beta - \min_i (A\beta)_i = \max_{\gamma \in \Delta_M} \langle A\beta, \beta - \gamma \rangle \le \varepsilon, \qquad \varepsilon = B^2/s. \tag{10}$$

For any x with $\mathbf{1}^\top x = 0$, antisymmetry and the EDM property give $x^\top Ax = -x^\top \widetilde{D}x/2 \ge 0$. The affine map is consequently monotone along simplex displacements. To solve its VI, we express the gap as a convex objective using the centered Gram matrix

$$K = -\tfrac{1}{2} J\widetilde{D}J \succeq 0, \qquad v = \mathrm{diag}(K).$$

For simplex vectors, $\beta^\top \widetilde{D}\beta = 2v^\top\beta - 2\beta^\top K\beta$, while $\beta^\top T\beta = 0$. Substituting these identities turns gap minimization into

$$\begin{aligned} \text{minimize}_{\beta,z} \quad & \beta^\top K\beta - v^\top\beta - z \\ \text{subject to} \quad & \beta \ge 0, \quad \mathbf{1}^\top\beta = 1, \quad z \le (A\beta)_i \quad (1 \le i \le M). \end{aligned} \tag{11}$$

A VI solution exists and attains the minimum zero, as proved in Appendix C. The procedure takes measurable solver outputs and returns $\widehat{g} = f_\beta = \sum_i \beta_i f_i$ once the recomputed gap satisfies (10). The simplex formulation permits a single predictor, a sparse mixture, or weights spread across the dictionary.

The approximation requirements in (9) and (10) concern the returned matrix and weights. The implementation verifies these feasibility conditions and recomputes ρ and the gap. Appendix F.1 reports the numerical checks used for the experiments, alongside the exact feasibility conditions required by the theorem.

Theorem 1 (Overlap-bound-free robust aggregation)**.** *Under Section 3, let $G_0 = \max\{1, \Gamma_{\mathcal{F}}\}$. For $M \ge 2$ and $n \ge k$, the rule* (4)–(11)*, with the certified tolerances in* (9) *and* (10)*, satisfies with probability at least $1 - \delta$*

$$R_Q(\widehat{g}) - R_* \le \min\left\{ 4B^2, \frac{524B^2G_0}{\lfloor n/k \rfloor} \right\}. \tag{12}$$

The algorithm uses w, B, δ, n, and the dictionary, but not G_0 or any upper bound on it. If $n < k$, returning f_1 retains the $4B^2$ bound. If $M = 1$, return the sole predictor.

The factor B^2 sets the risk scale, while G_0 measures the sampling cost. Dictionary size and confidence enter through k: more candidates or a smaller failure probability leave fewer observations per block. For fixed M and δ, the remainder decreases at order $1/n$. Leading-one means that R_* has coefficient one, so the additive term measures the cost of aggregation even when the best candidate has substantial approximation error.

4.3 Deterministic risk calculation

The proof connects the optimization gap to excess risk through one dispersion term. We first condition on the simultaneous event from Appendix A.3, where the scalar MOM bound of Lemma 5 is applied to every original pair:

$$\begin{aligned} |T_{ij} - (R_i - R_j)| \le a d_{ij}, \qquad |\widehat{d}_{ij} - d_{ij}| \le c, \\ a = 8B\sqrt{G_0/s}, \qquad c = 4B\sqrt{G_0/s}. \end{aligned} \tag{13}$$

Because the true matrix $D^0_{ij} = d^2_{ij}$ is feasible, (13) bounds the optimal repair radius by c. Appendix B gives the geometric details. Writing

$$\ell = c + \rho(\widetilde{D}) \le 2c + \tau, \tag{14}$$

the triangle inequality and a difference of squares give, for all pairs,

$$|\widetilde{D}_{ij} - d^2_{ij}| \le 2\ell d_{ij} + \ell^2. \tag{15}$$

Fix $i_* \in \arg\min_i R_i$ and define

$$\begin{aligned} V &= \frac{1}{2}\sum_{i,j} \beta_i \beta_j d^2_{ij}, \qquad L = \|f_\beta - f_{i_*}\|^2_{L_2(Q)}, \\ W &= V + L = \sum_i \beta_i d^2_{i,i_*}. \end{aligned} \tag{16}$$

Here V measures how much the mixture spreads its mass across different predictors, while L is the squared distance from the mixture to an oracle predictor. Their sum W is the average squared distance of the selected components from that oracle. This last expression involves only original pairs, which is what allows their overlap coefficient to control the aggregate.

Comparing the gap with coordinate i_* gives

$$\sum_i \beta_i T_{i,i_*} + \frac{1}{2}\sum_i \beta_i \widetilde{D}_{i,i_*} - \frac{1}{2}\sum_{i,j} \beta_i \beta_j \widetilde{D}_{ij} \le \varepsilon. \tag{17}$$

Replacing T and $\widetilde{D}$ by their population values turns the left side into

$$R_Q(f_\beta) - R_* + \frac{1}{2}W. \tag{18}$$

To verify the identity, expand squared loss to obtain $\sum_i \beta_i R_i = R_Q(f_\beta) + V$. The average distance term in (17) then cancels V, leaving the oracle distance contribution $W/2$. The replacement error from (13) and (15) is at most

$$(a + \ell)\sum_i \beta_i d_{i,i_*} + \ell \sum_{i,j} \beta_i \beta_j d_{ij} + \ell^2.$$

By Cauchy–Schwarz, the two distance averages are bounded by $\sqrt{W}$ and $\sqrt{2V} \le \sqrt{2W}$, respectively. Combining the terms gives

$$R_Q(f_\beta) - R_* + \frac{1}{2}W \le \{a + (1 + \sqrt{2})\ell\}\sqrt{W} + \ell^2 + \varepsilon.$$

The negative quadratic contribution $-W/2$ absorbs the comparison error, which grows only as $\sqrt{W}$. Maximizing over $W \ge 0$ yields

$$R_Q(f_\beta) - R_* \le \frac{1}{2}\{a + (1 + \sqrt{2})\ell\}^2 + \ell^2 + \varepsilon. \tag{19}$$

Now put $q = B\sqrt{G_0/s}$. Since $G_0 \geq 1$, (9) gives $\ell \leq 9q$ and $\varepsilon \leq q^2$. Substitution gives

$$R_Q(f_\beta) - R_* \leq (307.5 + 153\sqrt{2})q^2 < 524q^2.$$

The conditional variance common to two squared risks cancels outside and inside $\mathcal{D}$. Bounded predictions and $|\mu| \leq B$ give the additional $4B^2$ cap. Together with the probability bound from Appendix A.3, this completes the proof of Theorem 1. Solving both programs exactly improves the numerical constant to $256 + 128\sqrt{2} < 438$ (Appendix C.1).

4.4 Exact-data equivalence to Q-aggregation

With exact population quantities, the VI recovers classical Q-aggregation. Let $r = (R_1, \ldots, R_M)^\top$, let $T^0_{ij} = R_i - R_j$, and use the exact squared-distance matrix D^0. Set $K^0 = -JD^0J/2$ and $v^0 = \mathrm{diag}(K^0)$. Squared loss gives

$$R_Q(f_\beta) = \sum_i \beta_i R_i - \frac{1}{2}\beta^\top D^0 \beta.$$

The unpenalized form of Q-aggregation with $\nu = 1/2$ minimizes the following criterion. Equivalently, with a uniform prior, the linear-prior penalty of Dai et al. (2012) is constant on the simplex and can be omitted:

$$Q_{1/2}(\beta) = \tfrac{1}{2}R_Q(f_\beta) + \tfrac{1}{2}\sum_i \beta_i R_i = r^\top\beta - \tfrac{1}{2}(v^0)^\top\beta + \tfrac{1}{2}\beta^\top K^0\beta. \tag{20}$$

For $A^0 = T^0 - D^0/2$,

$$A^0\beta = \nabla Q_{1/2}(\beta) - \left(r^\top\beta + \tfrac{1}{2}(v^0)^\top\beta\right)\mathbf{1}.$$

The multiple of $\mathbf{1}$ vanishes against every simplex displacement. As a result, $\langle A^0\beta, \gamma - \beta\rangle \geq 0$ for all $\gamma \in \Delta_M$ is precisely the first-order condition for (20). With MOM estimates, a cycle component can remain in T. The VI continues to use those comparisons, with monotonicity supplied by the repaired distances. Appendix C.2 explains how that component acts on a particular mixture.

5 What pairwise overlap permits

The same input laws can be easy or difficult for different dictionaries. The three settings illustrate how agreement near large weights controls the pairwise coefficient, how new mixture directions can enlarge a uniform bound, and why common predictions permit target-only regions.

Infinite global weight variance. Let Q be uniform on $(0, 1)$ and let P have density $2x$. Then $w(x) = 1/(2x)$ and $\mathbb{E}_P w^2 = \infty$. Take a fixed bounded function h_0, common to all predictors, and let $f_j(x) = h_0(x) + \theta_j x^\alpha$ for fixed $\alpha > 0$ and at least two distinct coefficients. Their common part cancels, and direct integration gives

$$\Gamma_{\mathcal{F}} = \frac{\int_0^1 x^{2\alpha-1}/2\,dx}{\int_0^1 x^{2\alpha}\,dx} = \frac{2\alpha+1}{4\alpha} < \infty. \tag{21}$$

Under (2), Theorem 1 therefore gives a finite overlap-dependent remainder despite the divergent global weight variance. Near $x = 0$, the weight grows while the differences between predictors shrink as x^α. The weighted squared disagreement remains integrable in the numerator of (21).

Original pairs versus generated directions. Write $\mathcal{F}_{\mathrm{mid}} = \{(f_i + f_j)/2 : 1 \leq i \leq j \leq M\}$, including repeated endpoints. For fixed finite $\mathcal{F}$, Cauchy–Schwarz gives finite weighted moments for every midpoint difference whenever the original-pair moments are finite. Consequently, $\Gamma_{\mathcal{F}} < \infty$ if and only if $\Gamma_{\mathcal{F}_{\mathrm{mid}}} < \infty$, although their magnitudes can be very different. A three-predictor example makes the gap explicit.

Proposition 2. *For every integer $\kappa \geq 2$, there are three bounded predictors with $\Gamma_{\mathcal{F}} = 2 - 1/\kappa \leq 2$, while their midpoint closure contains a pair with overlap coefficient κ.*

Proof. Use inputs $A, C, 0$ with $Q(A) = 1 - 1/\kappa$, $Q(C) = 1/\kappa$, $Q(0) = 0$ and $P(A) = 1 - 1/\kappa$, $P(C) = 1/\kappa^2$, $P(0) = 1/\kappa - 1/\kappa^2$. The weights are $1, \kappa, 0$. For $0 < b \leq B$, let

$$f_1 = (b, b, 0), \qquad f_2 = (-b, b, 0), \qquad f_3 = (0, 0, 0).$$

The ratio (1) is 1 for $f_1 - f_2$ and $2 - 1/\kappa$ for each difference from f_3. But $(f_1 + f_2)/2 - f_3 = (0, b, 0)$ has ratio κ. □

Averaging f_1 and f_2 cancels their predictions on the well-covered input A, leaving a direction concentrated on the poorly covered input C. The original-pair coefficient stays below two, whereas the uniform midpoint coefficient grows with κ. In our risk calculation, the averages in (16) retain the original-pair distances and avoid this increase.

Common unsupported regions. The source and target supports can also differ. Wherever all candidates agree, every convex combination has the same squared loss, so target-only mass there contributes equally to their risks. The requirement $Q|_{\mathcal{D}} \ll P$ places the support condition precisely on the comparisons that matter. Appendix A.1 develops the corresponding weighting identity.

6 Worst-case lower bounds

We next ask whether the overlap factor in the upper bound reflects an intrinsic sampling cost. The answer follows from families in which informative inputs are observed only a fraction $1/G$ of the time under the source law. The results apply to any measurable prediction rule, even with knowledge of $P, Q, \mathcal{F}$ and unlimited unlabeled data.

Theorem 3 (Dictionary-size lower bound)**.** *For every $M \geq 2$, $G \geq 1$, and $n \geq 1$, there is a dictionary of at most M predictors and a family satisfying Section 3, with $|Y| = B$ and $\Gamma_{\mathcal{F}} = G$, such that every estimator has, on some member of the family,*

$$\mathbb{P}\left\{ R_Q(\widehat{g}) - R_* \geq \frac{B^2}{256} \min\left(1, \frac{G\lfloor \log_2 M \rfloor}{n}\right)\right\} \geq \frac{2}{7}.$$

Theorem 4 (Confidence lower bound)**.** *For every $G \geq 1$, $n \geq 1$, and $0 < \delta \leq 1/16$, a two-predictor family satisfying the same assumptions and $\Gamma_{\mathcal{F}} = G$ makes every estimator obey, on some member,*

$$\mathbb{P}\left\{ R_Q(\widehat{g}) - R_* \geq \frac{B^2}{16} \min\left(1, \frac{G\log(1/\delta)}{n}\right)\right\} \geq \delta.$$

For Theorem 3, the dictionary encodes conditional-mean signs at $\lfloor \log_2 M \rfloor$ informative inputs. Distinguishing neighboring sign patterns requires a label where they differ, and source labels arrive there G times less often, producing the overlap factor. Appendix E.1 gives the hypercube construction and its probability calculation. A two-point version, detailed in Appendix E.2, yields the logarithmic confidence term in Theorem 4. Together the bounds match the upper rate's dependence on overlap, dictionary size, and confidence, up to numerical constants.

7 Experiments

We refer to the proposed robust pairwise Q-aggregation rule in Section 4 as MOM equilibrium. We evaluate this source-only rule with $B = 1$ and $\delta = 0.1$, using the weights and dictionary as inputs. The main comparisons are MOM dictionary selection, ordinary convex importance-weighted (IW) least squares (Convex IW), IW minimization over all original-model midpoints (IW midpoint), direct IW Q-aggregation at $\nu = 1/2$ (IW Q), and the reconstructed risk-score midpoint rule in Appendix D. Comparing with IW Q tests the robust pairwise construction against direct estimation of the same population objective. Convex IW and both Q methods allow dense outputs. The screening baselines also receive $G = \max\{1, \Gamma_{\mathcal{F}}\}$, as required by their thresholds. Every method works with loss differences or omits the common Y^2 term.

Design and numerical checks. The study uses 4,100 validation draws, each evaluated by ten methods, including supplementary selectors and reference variants. The primary grid has 100 draws per synthetic condition and 30 per real-data training split. We set repair accuracy $\tau = B/(8\sqrt{s})$ and gap target B^2/s independently of target risks. In these runs, the returned matrices satisfy $\rho(\widetilde{D}) \leq \tau$, and the largest recomputed VI gap is 6.02×10^{-11}. The raw MOM distance matrices all have a negative centered-Gram eigenvalue, and the comparison matrices all contain a nonzero cycle sum. Both consistency issues addressed by the construction therefore arise in the experiments. Appendix F.1

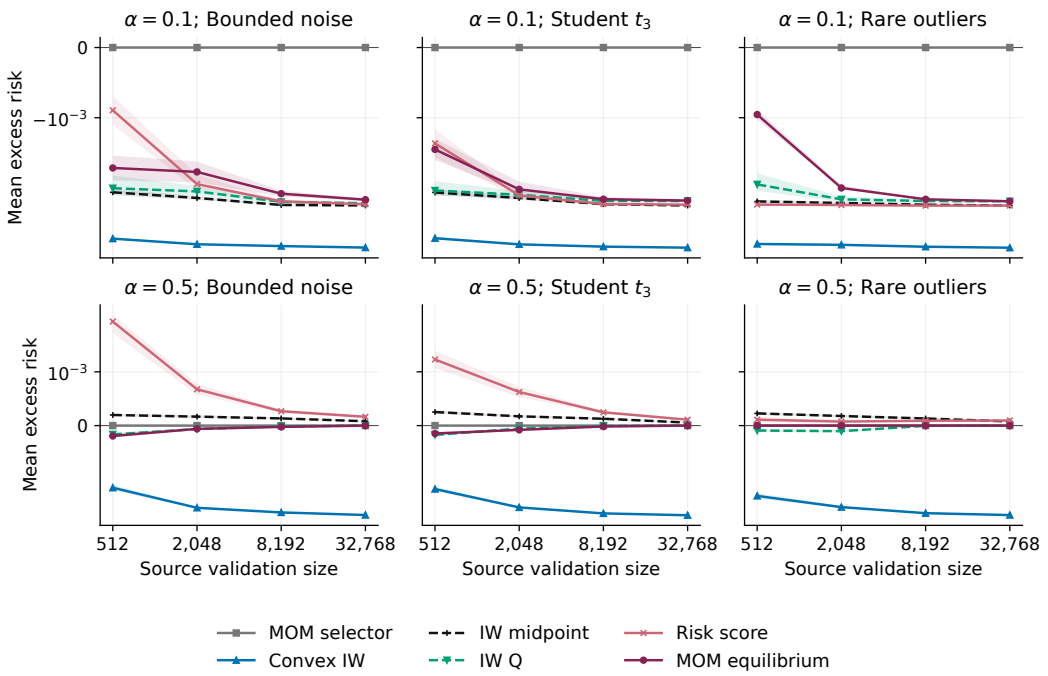

Figure 1: Synthetic excess risk relative to the best original predictor. Lines are means and bands are one standard error over 100 seeds. The symmetric logarithmic axis is linear in $[-0.002, 0.002]$. Both Q methods and Convex IW optimize over the same convex output class, which can outperform the best original predictor and yield negative excess risk.

gives the solver settings and residual checks. The following subsections describe the data designs and results.

Infinite global weight variance. The source density is $2x$ and the target is uniform on $(0, 1)$, so $w = 1/(2x)$. The 17 candidates are fixed linear combinations of $x^\alpha, x^{\alpha+1/2}, x^{\alpha+1}$ with $\alpha \in \{0.1, 0.5\}$ and coefficient ℓ_1 norms at most 0.8. The true coefficient, $(0.23, -0.17, 0.11)$, lies in their convex hull but differs from every candidate and candidate midpoint. Independent noise has standard deviation 0.3 and follows one of three laws: bounded, scaled Student t_3, or a rare symmetric outlier. All three satisfy the conditional second-moment assumption, and the Student law has infinite fourth moment. Analytical integration gives $\Gamma_{\mathcal{F}} \approx 26.15$ or 3.81 despite $\mathbb{E}_P w^2 = \infty$. It also lets us evaluate target risks without a test-sample approximation. The complete construction appears in Appendix F.2.

MOM equilibrium broadly follows IW Q in Figure 1. At $n = 32768$, both reach near-zero excess for $\alpha = 0.5$. For $\alpha = 0.1$, their mean excess is about -0.0025, while IW midpoint is slightly lower and Convex IW has the lowest observed risk. The 90th-percentile curves in Appendix F.2 show how the ordering changes with the noise law. These examples illustrate the method's behavior in a setting covered by the pairwise assumption even though global weight variance is infinite.

Two fixed real-data pools. Diabetes (Efron et al., 2004) and red Wine Quality (Cortez et al., 2009) each use three fixed splits. In each split, 40% of observations train 13 clipped ridge, nearest-neighbor, and tree candidates. The remainder form the target pool. We standardize responses using training statistics and clip them to $[-1, 1]$. The target law is uniform on the pool, while source probabilities are proportional to $\exp\{\lambda \operatorname{clip}(z, -2, 2)\}$ for $\lambda \in \{0, 1, 2\}$. Here z is training-standardized body mass index (BMI) or alcohol, and $\operatorname{clip}(z, -2, 2) = \min\{2, \max\{-2, z\}\}$. Sampling pool indices independently with replacement gives source observations with known weights. Aggregation sees a pool label only when that observation is sampled, and all pool labels are then used to evaluate risk under the specified finite target law. Appendix F.3 lists the candidates, sampling probabilities, and within-split results.

Pool	λ	MOM	Convex IW	IW mid.	IW Q	Risk score	MOM eq.
Diabetes	0	0.0 ± 0.0	-4.9 ± 0.5	-0.6 ± 1.2	-0.8 ± 1.4	-0.4 ± 1.3	-0.9 ± 1.5
Diabetes	1	0.0 ± 0.1	-4.7 ± 0.5	-0.4 ± 1.1	-0.8 ± 1.3	-0.3 ± 1.2	-0.8 ± 1.4
Diabetes	2	0.2 ± 0.1	-3.4 ± 0.4	1.1 ± 2.1	-0.9 ± 0.8	3.2 ± 4.6	-1.3 ± 0.8
Wine	0	0.1 ± 0.0	-16.9 ± 2.0	-6.4 ± 5.7	-6.8 ± 5.3	-6.1 ± 5.8	-6.9 ± 5.1
Wine	1	0.1 ± 0.0	-16.6 ± 2.0	-5.9 ± 5.6	-6.8 ± 4.0	-5.5 ± 5.9	-7.1 ± 3.8
Wine	2	0.5 ± 0.7	-13.9 ± 2.2	-2.7 ± 5.1	-5.8 ± 2.4	-0.4 ± 4.4	-8.7 ± 2.7

Table 1: Real pools at $n = 8192$. Entries are 10^3 times excess risk. We average 30 draws within each split and report mean $\pm$ standard deviation of the three split means. The spread describes variation across the three overlapping training splits.

MOM equilibrium has lower mean excess than IW Q in all six rows of Table 1. The difference is largest on wine at $\lambda = 2$: -0.00871 versus -0.00576, with smaller gains in the other rows. It also improves on the risk-score and IW-midpoint baselines in these comparisons, while Convex IW remains lower in every row. The paired split-level differences and standard errors in Appendix F.6 show where these gains occur. At larger synthetic sample sizes, MOM equilibrium again approaches IW Q. Appendix F.4 examines that scale and the screening behavior, and Appendix F.5 compares the reference-based rules.

8 Conclusion and limitations

Pairwise disagreement measures the sampling cost of covariate shift for a given dictionary. Robust loss comparisons and repaired Euclidean geometry yield a leading-one Q-aggregation bound, with the overlap scale appearing only in the analysis. The VI accommodates comparison cycles and convex mixtures, and its population counterpart is classical Q-aggregation at mixing parameter $1/2$. Our examples allow infinite global weight variance with bounded original-pair overlap, and the lower bounds establish the sharp worst-case dependence on overlap, dictionary size, and confidence.

The analysis assumes squared loss, a fixed bounded dictionary, known ratios on the disagreement set, and the conditional second-moment bound in (2). Estimated ratios and other losses are natural further problems. The coefficient examples distinguish sufficient uniform bounds. Comparing other localized analyses remains a separate question. Computing pair statistics costs at least quadratically in M, followed by the geometry and gap programs. Larger dictionaries and experiments that isolate each component's predictive effect would help extend the practical evidence beyond the studied $M = 13$ and 17 designs.

The real-pool means show gains over direct IW Q, especially at the strongest wine shift, while Convex IW remains lower. The evidence concerns the specified synthetic laws and finite target pools, with split variation reported descriptively. Conservative constants give worst-case guarantees and observed risk and solver cost also guide the choice of rule.

AI use statement

Generative AI tools were used to help with selected aspects of research development, implementation, experimentation, writing, and artifact preparation. During research development, they helped to refine—but did not originate—the proposed algorithm, identify and correct issues in the initial assumptions, formalize the mathematical statements initially described by the authors, check proofs and fill in omitted details, refine the experimental designs and provide feedback on early experiments based on figures and descriptions supplied by the authors. AI tools helped to refine the implementation of the proposed method, but not in constructing its initial core structure. They also helped with preparing supporting code for data processing, plotting templates, testing and execution scripts, debugging based on execution results, and organizing experimental outputs.

For writing and release preparation, AI tools helped with drafting certain sentences only after the authors had specified their intended content, improving the readability and clarity of some paragraphs, refining the presentation of contributions after the manuscript was completed, secondarily optimizing the title, and performing the final correction of typographical and grammatical errors. AI tools

were also used to conduct a comprehensive revision of the final artifact across the entire codebase, including modifying code across modules, improving comments and docstrings, standardizing code style, naming, and interfaces, and strengthening input validation, error handling, numerical diagnostics, tests, verification scripts, and reproducibility safeguards. They finally helped with organizing and minimizing the released artifact, preparing its README, and packaging it for distribution.

All AI-assisted outputs were reviewed and, where necessary, revised by the authors. For example: (1) three authors independently verified the final proofs, comparing every mathematical claim formalized by AI with the authors' original intended statement and making necessary adjustments; (2) Two authors checked that refined core implementation aligned with the algorithm and principles described in the paper. Supporting code and execution scripts were further adjusted based on our requirements by responsible authors; (3) Experimental modifications were repeatedly reviewed and revised by the authors responsible for the experimental design, while AI feedback on experiments was treated only as non-binding suggestions; (4) Two authors cross-checked the generated figures using the source data; (5) Text edited by AI (in both the paper and the code) was further revised by the authors, and the final artifact was verified through two end-to-end runs.

Generative AI was not used to originate research idea, foundational algorithm, paper structure, initial title, proof strategy, experimental directions or objectives, literature collection, paper typesetting, data analysis, overall interpretation of the results, or the basic visual design of figures and tables. The authors take responsibility for the final content, claims, code, results, and artifacts in this work.

Reproducibility statement

The appendix provides the proofs, optimization conditions, baseline formulas, and experimental settings. The code and data are available in an anonymous repository: `https://github.com/iclr2027-32406/pairwise-overlap-q-aggregation`. The artifact includes the estimator implementations, fixed configurations and real-data pools, per-draw results, and pinned dependencies. The verification script recomputes target risks from saved coefficients and target laws and checks method coverage and solver diagnostics. A separate plotting script regenerates all experimental figures and tables. The README gives commands for verification, plotting, and full experiment reruns, with seeds and data attribution. Candidate training uses observations separate from source validation.

References

Michele Aghassi, Dimitris Bertsimas, and Georgia Perakis. Solving asymmetric variational inequalities via convex optimization. *Operations Research Letters*, 34(5):481–490, 2006.

A. Y. Alfakih, A. Khandani, and Henry Wolkowicz. Solving euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12:13–30, 1999.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 130–160. PMLR, 2024.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547–553, 2009.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

Ivan Dokmanić, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Xingdong Feng, Xin He, Yuling Jiao, Lican Kang, and Caixing Wang. Deep nonparametric quantile regression under covariate shift. *Journal of Machine Learning Research*, 25(385):1–50, 2024.

Stéphane Gaïffas and Guillaume Lecué. Hyper-sparse optimal aggregation. *Journal of Machine Learning Research*, 12(50):1813–1833, 2011.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *International Conference on Learning Representations*, volume 2024, pp. 55558–55569, 2024.

Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives, 2024. URL https://arxiv.org/abs/2405.12762.

Mikael Møller Høgsgaard and Andrea Paudice. Uniform mean estimation for heavy-tailed distributions via median-of-means. In *Proceedings of the 42nd International Conference on Machine Learning*. JMLR.org, 2025.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Hui Lan and Vasilis Syrgkanis. Causal Q-aggregation for CATE model selection. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4366–4374. PMLR, 2024.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2020.

Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3–4):641–674, 2017.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 4(3):253–292, 2022.

Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. URL https://doi.org/10.1007/b13794.

A PAIRWISE MOM EVENT

A.1 RESTRICTED WEIGHTING AND ZERO-DISTANCE PAIRS

The derivative $w = d(Q|_{\mathcal{D}})/dP$ has total mass $\mathbb{E}_P w = Q(\mathcal{D})$, which can be smaller than one. If an integrable function h vanishes off $\mathcal{D}$, then

$$\mathbb{E}_P[wh] = \mathbb{E}_Q[h].$$

These observations give both equalities in (3). For the first, expand the difference of squared losses and replace the response by its conditional mean. Within each population, the conditional variance cancels between the two risks, while the source and target means agree on $\mathcal{D}$.

If $d_{ij} = 0$, then $\mathbb{E}_P[w(f_i - f_j)^2] = 0$. Its integrand is nonnegative, so $w(f_i - f_j)^2 = 0$, which forces $w(f_i - f_j) = 0$ almost surely. Both pairwise summands in (4)–(5) then vanish, including when the predictors differ only on a source-only region.

A.2 A SCALAR MEDIAN-OF-MEANS BOUND

Lemma 5. *Let $V_1, \dots, V_{ks}$ be iid with mean m and variance at most σ^2. If k is odd, the median $\widehat{m}$ of the means of k disjoint blocks of size $s \geq 1$ satisfies*

$$\mathbb{P}\{|\widehat{m} - m| > 2\sigma/\sqrt{s}\} \leq e^{-k/8}.$$

When $\sigma = 0$, $\widehat{m} = m$ almost surely.

Proof. Suppose first that $\sigma > 0$. For block b, let $\overline{V}_b$ be its mean and define the bad-block indicator

$$I_b = \mathbf{1}\{|\overline{V}_b - m| > 2\sigma/\sqrt{s}\}.$$

Since $\mathrm{Var}(\overline{V}_b) \leq \sigma^2/s$, Chebyshev's inequality gives $\mathbb{E}I_b \leq 1/4$. The indicators are independent because the blocks are disjoint and the observations are iid. If $|\widehat{m} - m| > 2\sigma/\sqrt{s}$, at least $(k+1)/2$ block means lie outside the interval $[m - 2\sigma/\sqrt{s}, m + 2\sigma/\sqrt{s}]$ on the same side. In particular, $\sum_{b=1}^{k} I_b \geq k/2$. Therefore

$$\begin{aligned}\mathbb{P}\{|\widehat{m} - m| > 2\sigma/\sqrt{s}\} &\leq \mathbb{P}\left\{\sum_{b=1}^{k} I_b - \mathbb{E}\sum_{b=1}^{k} I_b \geq k/4\right\} \\ &\leq \exp\left\{-\frac{2(k/4)^2}{k}\right\} = e^{-k/8},\end{aligned}$$

where the second line is Hoeffding's inequality. When $\sigma = 0$, each observation, and therefore each block mean, equals m almost surely. □

A.3 MOMENTS, CLIPPING, AND THE SIMULTANEOUS EVENT

Fix an original pair and write

$$Z_{ij} = w(f_i - f_j)(f_i + f_j - 2Y), \qquad U_{ij} = w(f_i - f_j)^2.$$

On $\mathcal{D}$, the prediction bound and (2) give

$$\begin{aligned}\mathbb{E}_P[(f_i + f_j - 2Y)^2 \mid X] &\leq 2(f_i + f_j)^2 + 8\mathbb{E}_P[Y^2 \mid X] \\ &\leq 16B^2.\end{aligned}$$

Both variables vanish off $\mathcal{D}$. With $G_0 = \max\{1, \Gamma_{\mathcal{F}}\}$,

$$\mathrm{Var}_P(Z_{ij}) \leq 16B^2\mathbb{E}_P[w^2(f_i - f_j)^2] \leq 16B^2 G_0 d_{ij}^2, \tag{22}$$

$$\mathrm{Var}_P(U_{ij}) \leq \mathbb{E}_P[U_{ij}^2] \leq 4B^2\mathbb{E}_P[w^2(f_i - f_j)^2] \leq 4B^2 G_0 d_{ij}^2. \tag{23}$$

Their means are $R_i - R_j$ and d_{ij}^2. Lemma 5 gives, for one pair, the two bounds in

$$|T_{ij} - (R_i - R_j)| \leq 8B\sqrt{G_0/s}\, d_{ij}, \qquad |H_{ij} - d_{ij}^2| \leq 4B\sqrt{G_0/s}\, d_{ij}. \tag{24}$$

There are $M(M-1)/2$ unordered pairs and two variables per pair. Thus the failure probability of their union is at most

$$M(M-1)e^{-k/8} \le M^2 e^{-k/8} \le \delta/2 \le \delta.$$

Different pair estimates may use the same observations and no independence between pairs is needed.

Every block mean of U_{ij} is nonnegative, so $H_{ij} \ge 0$. Projection of H_{ij} onto $[0, 4B^2]$ cannot increase its distance from $d_{ij}^2 \in [0, 4B^2]$. Therefore the clipped value $\widehat{d}_{ij}^2$ still satisfies the second inequality in (24). For $d_{ij} > 0$,

$$|\widehat{d}_{ij} - d_{ij}| = \frac{|\widehat{d}_{ij}^2 - d_{ij}^2|}{\widehat{d}_{ij} + d_{ij}} \le 4B\sqrt{G_0/s}.$$

Together with the zero-distance argument above, this establishes (13) with probability at least $1-\delta$.

Finally, for any $f, g \in \mathrm{conv}(\mathcal{F})$, the conditional squared-risk difference is zero off $\mathcal{D}$. On $\mathcal{D}$, after canceling conditional variance, it is $(f-\mu)^2 - (g-\mu)^2$. Both squares lie in $[0, 4B^2]$, so

$$|R_Q(f) - R_Q(g)| \le 4B^2. \tag{25}$$

The same calculation proves the cap in Theorem 1, including its fallback cases.

B Euclidean repair

B.1 The true distance matrix is feasible

The finite collection $\{f_1, \dots, f_M\}$ lies in the Hilbert space $L_2(Q)$, where $\|g\|_{L_2(Q)}^2 = \mathbb{E}_Q[g(X)^2]$. Its affine span has dimension at most $M-1$, so its distances have a Euclidean realization in that dimension. It follows that $D_{ij}^0 = d_{ij}^2$ is an EDM. Equivalently, the centered Gram matrix

$$K^0 = -\tfrac{1}{2} J D^0 J$$

is positive semidefinite. Symmetry, a zero diagonal, and $0 \le D_{ij}^0 \le 4B^2$ show $D^0 \in \mathcal{E}_B$.

For completeness, if D is symmetric and hollow and $K = -JDJ/2 \succeq 0$, choose vectors x_i with centered Gram matrix K. Let $v = \mathrm{diag}(K)$. Centered double multiplication determines the original hollow matrix and gives

$$D_{ij} = v_i + v_j - 2K_{ij} = \|x_i - x_j\|_2^2. \tag{26}$$

Thus the constraints in (6) are sufficient as well as necessary.

B.2 Existence and bisection

Because it contains the zero matrix, $\mathcal{E}_B$ is nonempty. Its constraints also make it closed, bounded, and convex. Continuity of $\rho(D)$ on this set then ensures that a minimizer exists. For fixed $r \ge 0$,

$$|\sqrt{D_{ij}} - \widehat{d}_{ij}| \le r$$

holds exactly when (8) holds. The lower and upper bounds in that display are constants at fixed r, so the sublevel set is an SDP-feasible set. It expands with r. Since the zero matrix is feasible at $r = 2B$, bisection on $[0, 2B]$ finds a feasible upper endpoint within any prescribed radius accuracy.

On the event (13), feasibility of D^0 gives

$$\rho_* \le \rho(D^0) = \max_{i,j} |d_{ij} - \widehat{d}_{ij}| \le c.$$

For a returned $\widetilde{D}$ satisfying (9), let $r_{ij} = \sqrt{\widetilde{D}_{ij}}$. Then

$$|r_{ij} - d_{ij}| \le |r_{ij} - \widehat{d}_{ij}| + |\widehat{d}_{ij} - d_{ij}| \le \rho(\widetilde{D}) + c = \ell.$$

Also $r_{ij} \le d_{ij} + \ell$, which proves

$$|\widetilde{D}_{ij} - d_{ij}^2| = |r_{ij} - d_{ij}|(r_{ij} + d_{ij}) \le \ell(2d_{ij} + \ell).$$

We have obtained (15).

C VI EXISTENCE AND CONVEX FORMULATION

For fixed A, the map

$$\beta \longmapsto \mathrm{Proj}_{\Delta_M}(\beta - A\beta)$$

is continuous from the compact convex simplex to itself. Brouwer's fixed-point theorem supplies a fixed point, at which the projection optimality condition is

$$\langle A\beta, \gamma - \beta \rangle \geq 0 \quad \text{for every } \gamma \in \Delta_M.$$

The gap in (10) is nonnegative because $\beta^\top A\beta$ is a convex combination of the coordinates of $A\beta$. At a VI point it is zero, and conversely zero gap implies the VI coordinate inequalities.

Now take $A = T - \widetilde{D}/2$. Equation (26) implies, on the simplex,

$$\begin{aligned}\beta^\top \widetilde{D}\beta &= \sum_{i,j} \beta_i \beta_j (v_i + v_j - 2K_{ij}) \\ &= 2v^\top \beta - 2\beta^\top K\beta.\end{aligned}$$

Exact antisymmetry gives $\beta^\top T\beta = 0$. Therefore

$$\beta^\top A\beta = \beta^\top K\beta - v^\top \beta.$$

For fixed β, minimizing the objective requires the largest feasible z, namely $z = \min_i (A\beta)_i$. Eliminating z leaves exactly the gap. The quadratic part is convex because $K \succeq 0$, and the minimum is zero by the existence argument. One may add the harmless bound $z \geq \min_{i,j} A_{ij}$ to make the displayed feasible set compact before eliminating z.

The operator is monotone on simplex displacements: for $\mathbf{1}^\top x = 0$,

$$x^\top (Ax) = x^\top Tx - \tfrac{1}{2} x^\top \widetilde{D}x = x^\top Kx \geq 0.$$

Here $x^\top Tx = 0$ and $x^\top \widetilde{D}x = -2x^\top Kx$.

C.1 EXACT OPTIMIZATION CONSTANT

If $\tau = \varepsilon = 0$, then $\ell \leq 2c$. Put $q = B\sqrt{G_0/s}$, so $a = 8q$, $c = 4q$, and $\ell \leq 8q$. The deterministic bound (19) becomes

$$\begin{aligned}R_Q(f_\beta) - R_* &\leq \frac{1}{2}\{8 + 8(1+\sqrt{2})\}^2 q^2 + 64q^2 \\ &= (256 + 128\sqrt{2})q^2 < 438q^2.\end{aligned}$$

Here exact optimization refers to solving both programs, with the repaired matrix still estimated from the sample.

C.2 HOW COMPARISON CYCLES ENTER THE GAP

For an antisymmetric comparison matrix T, put

$$u = T\mathbf{1}/M, \qquad T^{\mathrm{pot}} = u\mathbf{1}^\top - \mathbf{1}u^\top, \qquad C = T - T^{\mathrm{pot}}.$$

Then $C^\top = -C$ and $C\mathbf{1} = 0$. The matrix T^{pot} contains differences of one potential vector; C contains the remaining cycle component. With the repaired geometry fixed, write $g_T(\beta)$ for the gap in (10). Antisymmetry gives $\beta^\top C\beta = 0$, so

$$|g_T(\beta) - g_{T^{\mathrm{pot}}}(\beta)| \leq \|C\beta\|_\infty.$$

The quadratic term is unchanged, and each coordinate of the linear term changes by $(C\beta)_i$. The cycle component therefore acts through the current mixture. It vanishes at the uniform vector and may also vanish at other mixtures despite nonzero cycle sums. The identity characterizes how cycles affect the optimization at a given mixture.

C.3 ALGORITHM SUMMARY

Algorithm 1 collects the construction from Section 4. The tolerances refer to the returned distances and weights. Appendix F.1 describes the numerical feasibility checks used in the experiments.

Algorithm 1 Robust pairwise Q-aggregation

Inputs: source observations $(X_t, Y_t)_{t=1}^n$, fixed dictionary $\mathcal{F}$, weights w, bound $B > 0$, and failure probability $0 < \delta < 1$.
Output: a convex combination $\widehat{g}$ of the original predictors.

1. If $M = 1$, return f_1. Otherwise, take k to be the smallest odd integer at least $8\log(2M^2/\delta)$. If $n < k$, return f_1. Otherwise put $s = \lfloor n/k \rfloor$.
2. Divide the first ks observations into k blocks of size s. For each $i < j$, form T_{ij} and $\widehat{d}_{ij}$ by (4)–(5). Set $T_{ji} = -T_{ij}$, $\widehat{d}_{ji} = \widehat{d}_{ij}$, and both diagonals to zero.
3. Choose $0 < \tau \le B/\sqrt{s}$. Use the feasibility constraints (6) and (8) to obtain $\widetilde{D} \in \mathcal{E}_B$ with $\rho(\widetilde{D}) \le \rho_* + \tau$, as in Appendix B.2.
4. Set $A = T - \widetilde{D}/2$ and solve (11) for $\beta \in \Delta_M$. Once the recomputed gap is at most B^2/s, return $\widehat{g} = \sum_i \beta_i f_i$.

D SCALE-DEPENDENT MIDPOINT BASELINES

We compare with two MOM midpoint baselines that combine robust tests with standard diameter and midpoint geometry (Gaïffas & Lecué, 2011; Mourtada et al., 2022). They use the untruncated root distances $\bar{d}_{ij} = \sqrt{\max\{0, H_{ij}\}}$. Both require a supplied value $G \ge \max\{1, \Gamma_{\mathcal{F}}\}$ and use

$$a = 8B\sqrt{G/s}, \qquad c = 4B\sqrt{G/s}, \qquad S = \{i : T_{ij} \le a(\bar{d}_{ij} + c)\ \forall j\}.$$

The oracle $i_* \in \arg\min_i R_i$ belongs to S on the event (24), so S is nonempty.

Proposition 6 (Diameter midpoint and reconstructed risk score)**.** *On the simultaneous pairwise event, the following statements hold.*

1. *Choose $(u, v) \in S^2$ maximizing $\bar{d}_{uv}$ and return $(f_u + f_v)/2$. Its excess risk is at most*
$$384B^2G/s.$$
2. *Fix any measurable $r \in S$, minimize over $u \le v$ in S the score*
$$\widehat{L}_r(u, v) = \frac{T_{ur} + T_{vr}}{2} - \frac{\bar{d}_{uv}^2}{4},$$
and return the corresponding midpoint. Its excess risk is at most
$$1220B^2G/s.$$

Both outputs are supported on at most two original predictors.

Proof. The squared certificate implies $|\bar{d}_{ij} - d_{ij}| \le c$. For every survivor, comparison with i_* gives

$$R_i - R_* \le a(\bar{d}_{i,i_*} + c) + a d_{i,i_*} \le 2a(\bar{d}_{i,i_*} + c).$$

For the empirically farthest pair, put $\Delta = d_{uv}$. Empirical maximality and the distance error give $\bar{d}_{u,i_*}, \bar{d}_{v,i_*} \le \bar{d}_{uv} \le \Delta + c$. The squared-loss midpoint identity therefore yields

$$R_Q((f_u + f_v)/2) - R_* \le 2a(\Delta + 2c) - \Delta^2/4 \le 4a^2 + 4ac = 384B^2G/s.$$

For the score rule, let $\Delta = \max_{i,j \in S} d_{ij}$ and choose a population-diameter pair (p, q). Every survivor obeys $R_i - R_* \le 2a d_{i,i_*} + 2ac \le 2a\Delta + 2ac$. Thus

$$R_Q((f_p + f_q)/2) - R_* \le 2a\Delta + 2ac - \Delta^2/4.$$

The population counterpart

$$L_r(u,v) = \frac{(R_u - R_r) + (R_v - R_r)}{2} - \frac{d_{uv}^2}{4}$$

differs from $\widehat{L}_r(u,v)$ by at most $(a + c/4)\Delta$. Comparing the empirical minimizer with (p,q) and paying this error twice gives

$$R_Q(\widehat{g}) - R_* \le (4a + c/2)\Delta - \Delta^2/4 + 2ac \le (4a + c/2)^2 + 2ac.$$

Substitution of a and c gives $16a^2 + 6ac + c^2/4 = 1220B^2G/s$. □

The VI rule uses the same original-pair statistics without a supplied overlap threshold. The midpoint baselines use that threshold to screen candidates and guarantee two-model support. Because coordinatewise MOM comparisons can be cycle-inconsistent, the reconstructed score can depend on the reference.

E Complete lower-bound proofs

We use standard hypercube and two-point testing arguments (Tsybakov, 2009). The construction supplies the overlap dependence.

E.1 Dictionary-size dependence

Put $d = \lfloor \log_2 M \rfloor \ge 1$. On $\mathcal{X} = \{0, 1, \dots, d\}$ define

$$Q(j) = 1/d, \quad P(j) = 1/(Gd) \quad (1 \le j \le d), \qquad Q(0) = 0, \quad P(0) = 1 - 1/G.$$

Let $t = \min\{1/2, \sqrt{Gd/(32n)}\}$ and $r = Bt$. The dictionary has $2^d \le M$ predictors indexed by $v \in \{-1, 1\}^d$:

$$f_v(j) = rv_j \quad (1 \le j \le d), \qquad f_v(0) = 0.$$

For statistical parameter v, let $Y \in \{-B, B\}$ with conditional mean $f_v(X)$. At input zero, give both signs equal probability. Use the same conditional response law for source and target. All moment assumptions hold. The ratio is $w = G$ on informative inputs, and all predictor differences vanish at zero. For every distinct pair,

$$\frac{\mathbb{E}_P[w^2(f_v - f_{v'})^2]}{\mathbb{E}_Q[(f_v - f_{v'})^2]} = G.$$

The target regression function is in the dictionary, so every estimator satisfies

$$R_Q(\widehat{g}) - R_* = \frac{1}{d}\sum_{j=1}^{d} (\widehat{g}(j) - rv_j)^2. \tag{27}$$

Write $\mathbb{P}_v$ for the law of all n source observations. Adjacent parameters v, v' differ in one coordinate. Their Kullback–Leibler divergence is

$$\mathrm{KL}(\mathbb{P}_v, \mathbb{P}_{v'}) = \frac{n}{Gd} t \log \frac{1+t}{1-t} \le \frac{4nt^2}{Gd} \le \frac{1}{8}. \tag{28}$$

For $0 \le t \le 1/2$, the logarithmic inequality follows by integrating $2/(1-u^2) \le 8/3 \le 4$ from zero to t. Pinsker's inequality gives total variation, denoted by TV at most $1/4$. Extra unlabeled observations and known marginals have the same distribution under adjacent parameters and do not change the divergence.

Decode $\widehat{v}_j = 1$ when $\widehat{g}(j) \ge 0$ and -1 otherwise. For each adjacent pair, the sum of the two probabilities of a coordinate-j error is at least $1 - \mathrm{TV} \ge 3/4$. Average over the uniform hypercube and sum over coordinates. With $Z = \sum_j \mathbf{1}\{\widehat{v}_j \ne v_j\}$, this gives $\mathbb{E}Z \ge 3d/8$. Since $0 \le Z \le d$,

$$\mathbb{E}Z \le d/8 + (7d/8)\mathbb{P}\{Z \ge d/8\},$$

so the averaged probability is at least $2/7$. At least one parameter has that probability. Each sign error contributes at least r^2/d to (27). For this parameter, the excess is at least $r^2/8$ with probability at least $2/7$. Finally,

$$\frac{r^2}{8} = B^2 \min\left\{\frac{1}{32}, \frac{Gd}{256n}\right\} \ge \frac{B^2}{256} \min\left\{1, \frac{Gd}{n}\right\}.$$

The claimed dictionary-size lower bound in Theorem 3 follows.

E.2 Confidence dependence

Use the same construction with $d = 1$, two predictors, and

$$t = r/B = \min\left\{\frac{1}{2}, \sqrt{\frac{G\log(1/(4\delta))}{8n}}\right\}.$$

Equation (28) gives

$$\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-) \le \frac{4nt^2}{G} \le \frac{1}{2}\log\frac{1}{4\delta}.$$

The Bretagnolle–Huber bound says that the sum of the two testing errors is at least $\frac{1}{2}\exp(-\mathrm{KL})$. Consequently, the larger error is at least

$$\frac{1}{4}\exp(-\mathrm{KL}) \ge \frac{\sqrt{\delta}}{2} \ge \delta$$

for $\delta \le 1/16$. Apply this to the sign of $\widehat{g}(1)$. An erroneous sign has excess at least r^2. Also $\log(1/(4\delta)) \ge \frac{1}{2}\log(1/\delta)$, so

$$r^2 = B^2 \min\left\{\frac{1}{4}, \frac{G\log(1/(4\delta))}{8n}\right\} \ge \frac{B^2}{16}\min\left\{1, \frac{G\log(1/\delta)}{n}\right\}.$$

The confidence lower bound in Theorem 4 follows.

F Experimental details

F.1 Methods and common settings

Every validation draw is evaluated by ten methods: the importance-weighted (IW) dictionary selector, the MOM dictionary selector, convex IW least squares, IW empirical minimization over all original-model midpoints, direct IW Q-aggregation, the nominal diameter aggregate, the nominal risk-score aggregate with three survivor references, and MOM equilibrium. The main figures and tables show six scientifically distinct methods: MOM selector, Convex IW, IW midpoint, IW Q, the first-reference risk score, and MOM equilibrium. The IW selector, diameter rule, and two alternative risk-score references remain in the complete records and supplementary comparisons. All methods use identical validation draws, with the candidate dictionaries fixed before validation. No target labels tune a method or condition.

All runs use $B = 1$ and $\delta = 0.1$. The diameter and risk-score rules use threshold multiplier one and the exact condition-specific $G = \max\{1, \Gamma_{\mathcal{F}}\}$. No smaller practical threshold is included. The MOM equilibrium rule receives no G, screening threshold, or reference. All MOM methods use the same comparison matrix T and squared-distance statistics H_{ij} from each validation draw. The screening baselines use $\bar{d}_{ij} = \sqrt{H_{ij}}$. Only MOM equilibrium clips the root distances to $[0, 2B]$. For the risk-score comparator, the reference is the first survivor. Its two supplementary variants use the last survivor and the survivor minimizing $\max_j T_{rj}$, with lexicographic ties. Each variant minimizes the reconstructed score over all unordered survivor pairs, including repeated endpoints.

Every output is represented by coefficients on the original dictionary. The MOM statistics use the block count in Section 4 and the first $k\lfloor n/k \rfloor$ observations. We form block means of $w(f_i^2 - 2Yf_i)$ and then take medians of their blockwise differences, an algebraically equivalent form of the paired-loss statistic that avoids the common Y^2 term. Distances use the direct nonnegative values $w(f_i - f_j)^2$. The IW methods use all n observations, whereas the MOM methods discard the common remainder of fewer than k observations.

Write

$$\widehat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} w_t F_t F_t^\top, \qquad b = \frac{1}{n}\sum_{t=1}^{n} w_t Y_t F_t,$$

where $w_t = w(X_t)$ and $F_t = (f_1(X_t), \ldots, f_M(X_t))^\top$. Convex IW minimizes

$$\min_{\beta \in \Delta_M} \{\beta^\top \widehat{\Sigma}\beta - 2b^\top\beta\}.$$

It is an unpenalized empirical convex-hull benchmark. IW midpoint minimizes weighted empirical risk over every unordered original-model midpoint, including repeated endpoints. After dropping the common response-square term, its objective for $1 \le i \le j \le M$ is

$$\frac{\widehat{\Sigma}_{ii} + \widehat{\Sigma}_{jj}}{4} + \frac{\widehat{\Sigma}_{ij}}{2} - b_i - b_j.$$

Direct IW Q-aggregation uses the same dense output class as Convex IW and minimizes the $\nu = 1/2$ objective

$$\min_{\beta \in \Delta_M} \left\{ \frac{1}{2} \beta^\top \widehat{\Sigma} \beta + \frac{1}{2} \operatorname{diag}(\widehat{\Sigma})^\top \beta - 2 b^\top \beta \right\}.$$

The linear penalty associated with its uniform prior is constant on the simplex and is omitted. Both methods use all observations. IW midpoint breaks ties lexicographically, and IW Q uses the returned simplex minimizer of its convex program. Their inputs exclude G and target labels.

MOM equilibrium first divides T by B^2, divides $\widehat{d}$ by B, and clips the normalized root distances to $[0, 2]$. It repairs the resulting matrix to squared Euclidean geometry by minimizing maximum root discrepancy. The fixed physical radius accuracy is

$$\tau = \frac{B}{8\sqrt{s}}.$$

The implementation reconstructs distances from a centered PSD Gram embedding, scales any numerical diameter above $2B$ back to $2B$, and recomputes the actual root discrepancy. Radius bisection uses an always-feasible SDP slack problem. A lower endpoint advances only when nonnegative dual multipliers, corrected to make the reduced dual matrix PSD and checked afterward, give a positive dual certificate of infeasibility (a Farkas witness).

It then solves the convex variational-inequality gap program, clips negative numerical coefficients to zero, renormalizes them to the simplex, and requires the recomputed physical gap to be at most

$$\varepsilon = \frac{B^2}{s}.$$

The routine returns an output only after every numerical check passes. Otherwise, it raises an explicit error.

The exact theorem assumes a feasible Euclidean distance matrix and a simplex vector satisfying the displayed discrepancy and gap bounds. Floating-point runs therefore record the achieved radius and bracket, measured PSD and distance-reconstruction residuals, simplex residuals, solver choices, calls and runtimes, the corrected-dual minimum eigenvalue, and the recomputed gap. They also record two input-only normalized diagnostics:

$$\lambda_{\min}\left(-\frac{1}{2} J (\widehat{d}/B)^{\circ 2} J \right), \qquad \max_{i,j,\ell} \frac{|T_{ij} + T_{j\ell} - T_{i\ell}|}{B^2}.$$

Here $(\cdot)^{\circ 2}$ denotes entrywise squaring. The first quantity measures departure of the raw MOM distances from Euclidean geometry; the second measures failure of the coordinatewise MOM comparisons to be cycle-integrable. These quantities and the residuals are floating-point diagnostics instead of exact certificates. Clarabel (Goulart & Chen, 2024) solves the gap QP, whereas the Splitting Conic Solver (SCS) (O'Donoghue et al., 2016) is available only as a fallback if the SDP solve fails. The implementation rejects outputs that fail its fixed geometry, radius, simplex, dual, or gap checks. No target risk is used to choose a solver or tolerance.

Clarabel solved each of the $4{,}100$ equilibrium QPs, and every returned distance matrix satisfied $\rho(\widetilde{D}) \le \tau$. Because $\rho_* \ge 0$, the achieved bound $\rho(\widetilde{D}) \le \tau$ also meets the required accuracy $\rho(\widetilde{D}) \le \rho_* + \tau$. The largest recomputed gap was 6.02×10^{-11}, the largest PSD residual was 4.38×10^{-16}, all other geometry residuals were at most 2.14×10^{-15}, and the largest simplex-sum residual was 3.34×10^{-16}. The median number of SDP calls was zero, and no draw required more than two. No positive Farkas witness advanced a lower radius endpoint. Median equilibrium fit time was 0.0403 seconds over all synthetic draws, including the larger sizes, and 0.0135 seconds over real-data draws. The corresponding maxima were 4.96 and 0.0511 seconds.

The normalized centered Gram matrix had a negative minimum eigenvalue in every equilibrium draw, with even the least-negative value below -1.47×10^{-9}. Every comparison matrix also had a nonzero cycle sum. The smallest normalized maximum cycle sum was 0.000496 in the synthetic draws and 0.008247 in the real-data draws. The repair and equilibrium therefore address discrepancies present in all recorded samples, while the initial PSD repair met the fixed radius tolerance in most draws, so SDP bisection was rarely needed.

The primary summary is mean excess risk at every fixed condition and sample size, with medians and empirical 90th percentiles as secondary descriptive summaries. The primary grid contains $24{,}000$ synthetic and $16{,}200$ real-data records, for $40{,}200$ rows. The larger-sample grid contains 800 rows. The figures summarize this complete grid.

F.2 Noncollinear synthetic model

The target input law is uniform on $(0, 1)$, the source density is $2x$, and the known weight is $w(x) = 1/(2x)$. For $\alpha \in \{0.1, 0.5\}$, define

$$p = (\alpha, \alpha + 0.5, \alpha + 1), \qquad f_\vartheta(x) = \sum_{\ell=1}^{3} \vartheta_\ell x^{p_\ell}.$$

The 17 coefficient vectors are fixed independently of the validation draws. They consist of the zero vector, the six vectors $\pm 0.7 e_\ell$, where e_ℓ is the ℓ-th standard basis vector in $\mathbb{R}^3$, and ten standard-normal vectors drawn together with NumPy seed 20270909 and rescaled separately to have ℓ_1 norm 0.8. The regression function has coefficient

$$\vartheta_* = (0.23, -0.17, 0.11).$$

Its ℓ_1 norm is 0.51, so it lies in the convex hull of the six vectors $\pm 0.7 e_\ell$. The dense comparators can therefore attain it, whereas the original candidate set and its pairwise midpoints do not contain it.

For coefficient vectors ϑ and ϑ', exact integration uses

$$A^{(Q)}_{\ell m} = \frac{1}{p_\ell + p_m + 1}, \qquad C_{\ell m} = \frac{1}{2(p_\ell + p_m)}.$$

The target squared distance is $(\vartheta - \vartheta')^\top A^{(Q)} (\vartheta - \vartheta')$, and its weighted second moment is $(\vartheta - \vartheta')^\top C (\vartheta - \vartheta')$. Maximizing their ratio over the fixed dictionary gives

$$\Gamma_{\mathcal{F}} \approx 26.1531298240 \quad (\alpha = 0.1), \qquad \Gamma_{\mathcal{F}} \approx 3.8069647687 \quad (\alpha = 0.5).$$

The response is

$$Y = f_{\vartheta_*}(X) + \epsilon_0,$$

with independent mean-zero noise of variance 0.09. The three fixed noise laws are Rademacher noise $\epsilon_0 = \pm 0.3$ with equal probabilities. Student t_3 noise multiplied by $0.3/\sqrt{3}$, and rare outliers that equal zero with probability 0.995 and otherwise equal $\pm 0.3/\sqrt{0.005}$ with equal signs. Since $|f_{\vartheta_*}| \le \|\vartheta_*\|_1 = 0.51$, the conditional response second moment is at most $0.51^2 + 0.09 < 1$. The last two settings respectively test infinite fourth moment and infrequent large responses while retaining the theorem's conditional second-moment assumption.

The exact risk is

$$R_Q(f_\vartheta) = (\vartheta - \vartheta_*)^\top A^{(Q)} (\vartheta - \vartheta_*) + 0.09.$$

The common noise variance cancels from excess risk, which is evaluated relative to the smallest exact risk among the 17 original candidates. Target scoring therefore has no Monte Carlo error.

The source validation sizes are 512, 2048, 8192, and 32768. Each of the six (α, noise) conditions uses seeds $1000, \ldots, 1099$, giving 100 replicates within a condition. At each sample size the random generator is reinitialized with the same seed. Covariate samples are consequently coupled across sizes, including shared prefixes, and observations at different sizes are not independent replicates. Standard errors use the 100 seeds within each fixed condition and size. Conditions carrying the same numeric seed may also be dependent. Comparisons are made within their stated condition.

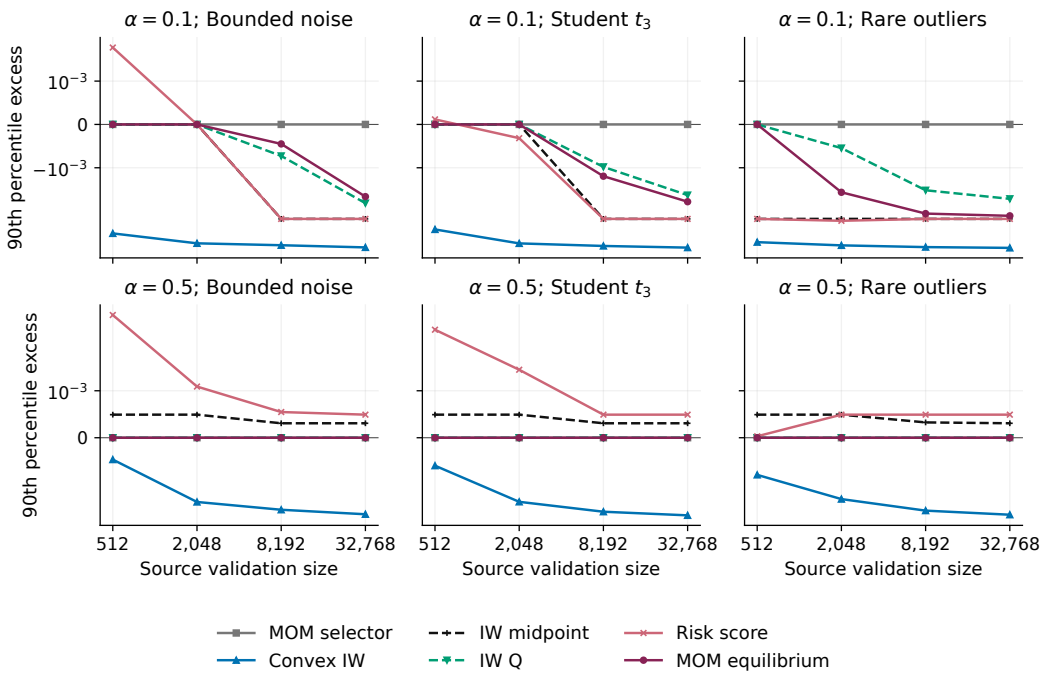

Figure 2: Empirical 90th percentile of synthetic excess risk over 100 seeds for the six main methods. Rows vary α and columns vary the noise law. The symmetric logarithmic axis is linear in $[-0.002, 0.002]$. The percentile is descriptive. Seeds at different sample sizes are coupled.

Figure 1 reports mean curves for the six main methods, and Figure 2 reports their empirical 90th percentiles. The latter describes upper-tail performance across the 100 seeds.

Across the synthetic grid, MOM equilibrium has mean excess similar to direct IW Q, while Convex IW has the lowest mean excess overall. At $\alpha = 0.1$ and $n = 32768$, MOM equilibrium has lower 90th-percentile excess than IW Q under Student-t_3 and rare-outlier noise. These comparisons cover two dictionaries and three fixed noise laws.

F.3 Real-data pools held out for evaluation

We use the 442-observation diabetes regression data distributed by scikit-learn (Pedregosa et al., 2011) with `scaled=False` (Efron et al., 2004), and the red subset of the UCI Wine Quality data, containing 1599 observations and 11 physicochemical inputs (Cortez et al., 2009).

Each dataset is processed independently at split seeds 2718, 31415, and 16180. Writing N for the number of observations in the dataset, a seeded permutation assigns the first $\lfloor 0.4N \rfloor$ observations to candidate training and holds out the rest as a held-out target pool. Diabetes therefore uses 176 training observations and a 266-observation pool. Wine uses 639 training observations and a 960-observation pool. The three splits reuse the same underlying dataset and their training samples overlap. Variation across these overlapping splits provides a sensitivity summary.

Feature means and population standard deviations are fitted on the training part only. The response mean and population standard deviation are also fitted there. All responses are then standardized with these quantities and clipped to $[-1, 1]$. The 13 candidate recipes are five ridge regressions with penalties $0.01, 0.1, 1, 10, 100$, four nearest-neighbor regressions with $3, 7, 15, 31$ neighbors, and four regression trees with maximum depths $2, 3, 4, 6$. Each tree uses the split seed as its random state. Models are fitted only on the training part, and their pool predictions are clipped to $[-1, 1]$.

For each split, Q is uniform on its fixed pool. The shift feature is training-standardized BMI for diabetes and training-standardized alcohol for wine. If z_x is that feature on pool point x, the source

probability and known weight are

$$P_\lambda(x) = \frac{\exp\{\lambda \operatorname{clip}(z_x, -2, 2)\}}{\sum_{x'} \exp\{\lambda \operatorname{clip}(z_{x'}, -2, 2)\}}, \qquad w_\lambda(x) = \frac{Q(x)}{P_\lambda(x)},$$

for $\lambda \in \{0, 1, 2\}$. The exact finite-pool coefficient is

$$\Gamma_{\mathcal{F}} = \max_{i,j:d_{ij}>0} \frac{\sum_x P_\lambda(x) w_\lambda(x)^2 (f_i(x) - f_j(x))^2}{\sum_x Q(x)(f_i(x) - f_j(x))^2}.$$

The finite-pool coefficient is computed from pool inputs and the candidates' pool predictions, without pool labels.

For each dataset, split, λ, and validation size $n \in \{512, 2048, 8192\}$, seeds $100, \ldots, 129$ generate 30 iid samples with replacement from P_λ. The generator is initialized by

```
SeedSequence([20270910, split_seed, seed, n, lambda_index]).
```

Sampled pool labels act as source validation responses. Full-pool labels are reserved for defining evaluation and oracle risks, while candidate fitting and aggregation exclude the full label arrays and their risk values. During aggregation, a pool label enters only when its point is sampled as a source response. Preprocessing, candidates, shift feature, source tilt, sample sizes, references, and numerical tolerances are all fixed without using pool labels.

As in Section 7, the estimand is uniform risk on each fixed pool. Repeated sampled indices are iid draws from that finite source law.

Table 1 gives the main real-data summary at $n = 8192$. For each dataset, shift, and method, we first compute the mean over 30 validation draws separately within each split. The displayed center is the mean of those three split-specific means, and the displayed spread is their sample standard deviation. The spread describes variation across the overlapping splits. Tables 2 and 3 give every split, size, and shift for the six main methods. Their entries are 10^3 times the split-specific mean excess risk plus or minus its standard error over 30 validation draws.

Split	n	λ	MOM	Convex IW	IW mid.	IW Q	Risk score	MOM eq.
2718	512	0	0.2 ± 0.2	-2.5 ± 0.3	2.0 ± 1.0	0.1 ± 0.3	4.5 ± 1.2	-0.1 ± 0.3
2718	512	1	0.9 ± 0.3	-0.9 ± 0.6	4.1 ± 1.0	-0.2 ± 0.3	12.3 ± 2.1	-0.1 ± 0.4
2718	512	2	47.1 ± 12.5	8.2 ± 2.2	14.5 ± 3.4	5.5 ± 1.6	41.6 ± 5.1	18.5 ± 3.3
2718	2,048	0	0.0 ± 0.0	-4.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
2718	2,048	1	0.0 ± 0.0	-3.5 ± 0.2	0.4 ± 0.4	0.0 ± 0.0	1.5 ± 0.7	-0.1 ± 0.1
2718	2,048	2	0.3 ± 0.2	-0.3 ± 0.6	4.3 ± 1.7	0.5 ± 0.4	13.3 ± 2.1	-0.9 ± 0.6
2718	8,192	0	0.0 ± 0.0	-4.4 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
2718	8,192	1	0.0 ± 0.0	-4.3 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
2718	8,192	2	0.0 ± 0.0	-3.0 ± 0.3	2.5 ± 0.9	-0.3 ± 0.1	8.1 ± 1.8	-1.0 ± 0.3
31415	512	0	0.5 ± 0.2	-4.2 ± 0.2	-1.4 ± 0.1	-2.8 ± 0.3	-0.6 ± 1.0	-3.2 ± 0.3
31415	512	1	0.4 ± 0.2	-2.9 ± 0.4	-1.2 ± 0.2	-1.5 ± 0.4	1.9 ± 2.0	-1.4 ± 0.4
31415	512	2	0.9 ± 0.2	11.2 ± 3.0	14.1 ± 3.7	9.2 ± 3.7	37.0 ± 2.9	7.2 ± 1.3
31415	2,048	0	0.8 ± 0.3	-5.0 ± 0.1	-1.7 ± 0.1	-2.3 ± 0.2	-1.6 ± 0.1	-2.4 ± 0.3
31415	2,048	1	0.1 ± 0.1	-4.5 ± 0.2	-1.6 ± 0.1	-2.2 ± 0.3	-1.4 ± 0.1	-2.0 ± 0.3
31415	2,048	2	0.4 ± 0.2	-1.3 ± 0.7	-1.2 ± 0.2	0.2 ± 0.8	4.9 ± 3.1	-1.5 ± 0.5
31415	8,192	0	0.0 ± 0.0	-5.5 ± 0.0	-2.0 ± 0.1	-2.5 ± 0.2	-1.9 ± 0.1	-2.6 ± 0.2
31415	8,192	1	0.0 ± 0.0	-5.3 ± 0.0	-1.7 ± 0.1	-2.3 ± 0.2	-1.7 ± 0.1	-2.4 ± 0.2
31415	8,192	2	0.2 ± 0.1	-3.9 ± 0.3	-1.3 ± 0.2	-1.8 ± 0.3	-1.1 ± 0.2	-2.1 ± 0.4
16180	512	0	0.3 ± 0.0	-3.1 ± 0.2	2.9 ± 0.7	-0.7 ± 0.3	5.9 ± 0.8	-2.0 ± 0.3
16180	512	1	5.1 ± 3.4	-1.0 ± 0.7	3.0 ± 0.6	0.2 ± 1.0	6.9 ± 1.9	0.2 ± 0.9
16180	512	2	2.4 ± 1.8	11.9 ± 2.6	15.7 ± 4.2	10.0 ± 3.1	32.6 ± 4.6	2.9 ± 1.4
16180	2,048	0	0.1 ± 0.0	-4.3 ± 0.1	0.7 ± 0.3	0.0 ± 0.1	2.5 ± 0.5	-0.3 ± 0.2
16180	2,048	1	0.6 ± 0.3	-3.7 ± 0.2	1.3 ± 0.4	-0.1 ± 0.2	1.6 ± 0.3	-0.4 ± 0.3
16180	2,048	2	0.3 ± 0.0	1.1 ± 0.8	5.3 ± 1.3	1.8 ± 1.0	8.8 ± 1.7	-2.1 ± 0.3
16180	8,192	0	0.0 ± 0.0	-4.7 ± 0.0	0.2 ± 0.1	0.0 ± 0.0	0.6 ± 0.2	0.0 ± 0.0
16180	8,192	1	0.1 ± 0.0	-4.6 ± 0.0	0.4 ± 0.2	-0.1 ± 0.1	0.6 ± 0.2	-0.1 ± 0.1
16180	8,192	2	0.3 ± 0.0	-3.4 ± 0.3	2.2 ± 0.5	-0.6 ± 0.2	2.7 ± 0.6	-0.7 ± 0.3

Table 2: Complete diabetes fixed-pool results for the six main methods. Each entry is 10^3 times mean excess risk plus or minus its standard error over 30 validation draws, conditional on the displayed training split and held-out pool.

Split	n	λ	MOM	Convex IW	IW mid.	IW Q	Risk score	MOM eq.
2718	512	0	5.4 ± 2.4	-16.2 ± 0.4	-5.2 ± 1.1	-9.6 ± 0.8	-5.2 ± 1.1	-9.7 ± 0.9
2718	512	1	13.0 ± 3.9	-14.3 ± 0.7	-4.1 ± 1.1	-6.5 ± 0.9	0.5 ± 1.2	-11.4 ± 0.9
2718	512	2	33.5 ± 6.1	-0.3 ± 2.6	7.6 ± 3.2	9.5 ± 3.6	9.8 ± 1.8	-2.0 ± 1.2
2718	2,048	0	1.2 ± 0.5	-18.5 ± 0.1	-9.5 ± 0.2	-9.5 ± 0.5	-9.1 ± 0.3	-9.6 ± 0.6
2718	2,048	1	1.0 ± 0.4	-17.8 ± 0.2	-8.4 ± 0.6	-8.8 ± 0.6	-6.2 ± 1.0	-10.6 ± 0.7
2718	2,048	2	5.1 ± 2.7	-9.9 ± 0.8	-0.7 ± 1.2	-1.3 ± 1.1	2.3 ± 1.1	-11.6 ± 0.9
2718	8,192	0	0.0 ± 0.0	-19.2 ± 0.0	-10.1 ± 0.1	-10.5 ± 0.3	-9.7 ± 0.2	-10.5 ± 0.4
2718	8,192	1	0.1 ± 0.1	-18.9 ± 0.1	-9.3 ± 0.5	-9.7 ± 0.4	-9.6 ± 0.2	-10.0 ± 0.5
2718	8,192	2	1.3 ± 0.5	-16.3 ± 0.3	-6.1 ± 1.0	-6.8 ± 0.9	-4.2 ± 1.2	-11.7 ± 0.8
31415	512	0	1.6 ± 1.0	-13.3 ± 0.3	2.9 ± 0.8	-3.9 ± 0.5	5.8 ± 1.1	-6.9 ± 0.8
31415	512	1	11.1 ± 3.9	-9.7 ± 0.7	5.6 ± 1.2	-2.8 ± 0.9	11.4 ± 1.4	-5.0 ± 1.3
31415	512	2	35.4 ± 5.6	1.0 ± 2.6	11.8 ± 3.1	8.2 ± 3.4	15.1 ± 1.3	-2.2 ± 1.4
31415	2,048	0	0.1 ± 0.1	-14.8 ± 0.1	0.8 ± 0.3	-2.3 ± 0.4	1.8 ± 0.5	-3.5 ± 0.5
31415	2,048	1	0.2 ± 0.1	-14.0 ± 0.2	2.0 ± 0.5	-3.8 ± 0.5	3.7 ± 0.6	-5.9 ± 0.6
31415	2,048	2	5.3 ± 3.2	-7.0 ± 0.9	5.4 ± 1.2	-0.5 ± 1.2	12.5 ± 1.6	-9.0 ± 0.6
31415	8,192	0	0.0 ± 0.0	-15.4 ± 0.0	0.2 ± 0.2	-0.7 ± 0.2	0.5 ± 0.2	-1.1 ± 0.2
31415	8,192	1	0.0 ± 0.0	-15.1 ± 0.1	0.5 ± 0.2	-2.2 ± 0.3	1.3 ± 0.4	-2.9 ± 0.4
31415	8,192	2	0.1 ± 0.0	-12.0 ± 0.5	3.1 ± 0.9	-3.0 ± 0.7	4.4 ± 0.8	-6.6 ± 0.7
16180	512	0	2.3 ± 0.8	-13.4 ± 0.3	-2.7 ± 0.9	-6.9 ± 0.6	-1.1 ± 1.2	-8.7 ± 0.7
16180	512	1	8.8 ± 3.1	-11.1 ± 0.5	-3.0 ± 1.0	-4.9 ± 0.8	4.5 ± 1.2	-7.4 ± 1.0
16180	512	2	51.0 ± 8.2	2.8 ± 2.8	8.4 ± 2.4	10.0 ± 3.7	12.0 ± 1.5	2.3 ± 2.3
16180	2,048	0	0.4 ± 0.3	-15.5 ± 0.1	-8.3 ± 0.5	-8.0 ± 0.4	-6.1 ± 0.8	-7.5 ± 0.6
16180	2,048	1	0.2 ± 0.1	-14.7 ± 0.1	-6.2 ± 0.8	-7.7 ± 0.6	-4.4 ± 1.0	-7.3 ± 0.7
16180	2,048	2	1.8 ± 0.7	-8.6 ± 0.9	0.3 ± 1.2	-3.5 ± 1.0	8.1 ± 1.5	-11.4 ± 0.5
16180	8,192	0	0.1 ± 0.0	-16.2 ± 0.0	-9.2 ± 0.0	-9.2 ± 0.1	-9.2 ± 0.0	-9.1 ± 0.2
16180	8,192	1	0.1 ± 0.0	-15.9 ± 0.1	-9.0 ± 0.3	-8.5 ± 0.3	-8.2 ± 0.5	-8.4 ± 0.4
16180	8,192	2	0.2 ± 0.1	-13.4 ± 0.4	-5.2 ± 0.8	-7.5 ± 0.8	-1.5 ± 1.1	-7.8 ± 0.8

Table 3: Complete red-wine fixed-pool results for the six main methods. Each entry is 10^3 times mean excess risk plus or minus its standard error over 30 validation draws, conditional on the displayed training split and held-out pool.

F.4 LARGER-SAMPLE DIAGNOSTIC

The two synthetic dictionaries with Student t_3 noise are also evaluated at $n = 262{,}144$ and $1{,}048{,}576$, using seeds $2000, \ldots, 2019$. All ten methods are evaluated, giving 800 records. These sizes place the theorem remainder $524G/\lfloor n/71 \rfloor$ below the universal cap for both dictionaries and show a further fourfold increase in n. Equal seeds couple the two sizes, and uncertainty is computed within each fixed condition and size. The table includes the theorem bound as a worst-case reference.

The screening comparators clarify the effect of their additional machinery. Every one of the 17 synthetic candidates and 13 real-data candidates survives in every primary-grid draw, so primary-grid risk scoring and IW midpoint use the same midpoint output class, while the diameter output is conditionally independent of the sampled responses. On that grid, IW midpoint has similar or lower mean excess than risk scoring. At $n = 1{,}048{,}576$, screening retains 7 and 4 candidates in the two Student dictionaries, and risk scoring matches IW midpoint in every draw. Across these experiments, the reconstructed risk-score comparator shows no empirical advantage.

At $n = 1{,}048{,}576$, the displayed equilibrium bounds are 0.928 and 0.1351 for $\alpha = 0.1$ and 0.5. Observed MOM-equilibrium excess is close to IW Q, while Convex IW remains best. Observed MOM-equilibrium excess is far below these worst-case bounds.

Table 4 reports MOM equilibrium and four competing methods in all four larger-sample conditions. Numerical geometry, comparison-cycle, and solver diagnostics are recorded independently of target performance.

α	n	Bound	Convex IW	IW mid.	IW Q	Risk score	MOM eq.
0.1	262,144	3711.9	-5.98 ± 0.00	-2.84 ± 0.02	-2.58 ± 0.04	-2.72 ± 0.03	-2.56 ± 0.05
0.1	1,048,576	928.0	-5.98 ± 0.00	-2.88 ± 0.00	-2.71 ± 0.03	-2.88 ± 0.00	-2.70 ± 0.03
0.5	262,144	540.3	-1.68 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.5	1,048,576	135.1	-1.68 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 4: Larger-sample Student-t_3 diagnostic. Bound and excess-risk entries are multiplied by 10^3. The bound is $\min\{4, 524G/\lfloor n/71 \rfloor\}$. Method entries are mean plus or minus standard error over 20 seeds, and a displayed zero spread may reflect rounding.

F.5 REFERENCE SENSITIVITY

The risk-score theorem permits any measurable reference in the survivor set, while coordinatewise MOM estimates can be cycle-inconsistent. Table 5 compares the first survivor, last survivor, and survivor MOM-selector references while holding pair statistics, screening threshold, and score rule fixed. No target labels select among them. Synthetic entries at $n = 32768$ are 10^3 times mean excess risk plus or minus its standard error over 100 seeds. Real-data entries at $n = 8192$ use the summary in Appendix F.3, scaled by 10^3: the mean of three split-specific means plus or minus their sample standard deviation. The spread describes variation across overlapping splits. The observed range across reference means reached 1.10×10^{-3} for diabetes and 1.55×10^{-3} for wine.

Problem	Setting	First	Last	MOM reference
Synthetic $\alpha = 0.1$	rademacher	-2.69 ± 0.01	-2.70 ± 0.01	-2.70 ± 0.01
Synthetic $\alpha = 0.1$	t_3	-2.71 ± 0.01	-2.70 ± 0.01	-2.70 ± 0.01
Synthetic $\alpha = 0.1$	rare	-2.76 ± 0.01	-2.77 ± 0.01	-2.77 ± 0.01
Synthetic $\alpha = 0.5$	rademacher	0.16 ± 0.02	0.17 ± 0.02	0.17 ± 0.02
Synthetic $\alpha = 0.5$	t_3	0.11 ± 0.02	0.12 ± 0.02	0.12 ± 0.02
Synthetic $\alpha = 0.5$	rare	0.09 ± 0.02	0.06 ± 0.01	0.06 ± 0.01
Diabetes	$\lambda = 0$	-0.43 ± 1.27	-0.15 ± 1.19	-0.40 ± 1.23
Diabetes	$\lambda = 1$	-0.35 ± 1.17	0.75 ± 1.57	-0.29 ± 1.22
Diabetes	$\lambda = 2$	3.23 ± 4.62	2.83 ± 2.87	2.56 ± 3.87
Wine	$\lambda = 0$	-6.13 ± 5.78	-5.74 ± 5.61	-6.15 ± 5.75
Wine	$\lambda = 1$	-5.51 ± 5.91	-3.97 ± 4.83	-5.45 ± 5.77
Wine	$\lambda = 2$	-0.40 ± 4.40	-0.16 ± 4.30	-0.65 ± 4.57

Table 5: Sensitivity of the reconstructed risk-score comparator to its survivor reference. "MOM reference" is the survivor minimizing $\max_j T_{rj}$. Synthetic entries report the mean plus or minus its standard error over 100 seeds. Real-data entries report the mean of three split-specific means plus or minus their sample standard deviation. The spread describes variation across overlapping splits.

F.6 PAIRED COMPARISON WITH DIRECT IW Q

For each real-data condition at $n = 8192$, we compare the two Q methods on the same 30 validation draws. Let Δ_t be MOM equilibrium's target risk minus IW Q's target risk on draw t. Table 6 reports the mean $\bar{\Delta}$ and its paired standard error, $\mathrm{sd}(\Delta_1, \ldots, \Delta_{30})/\sqrt{30}$, where the sample standard deviation uses denominator 29. Pairing accounts for the dependence between the methods' risks on a shared draw. The reported uncertainty describes source-sampling variation conditional on the indicated training split and target pool.

At $\lambda = 2$, all three split means favor MOM equilibrium for both datasets. On wine, the gains are largest for splits 2718 and 31415, while the difference is smaller for split 16180. At the two weaker wine shifts, split 16180 has small positive differences and the other two splits favor MOM equilibrium. The paired summaries show where the average gains in Table 1 occur.

Dataset	λ	Split seed	$10^3\bar{\Delta} \pm 10^3\mathrm{SE}_{\mathrm{paired}}$
Diabetes	0	2718	0.000 ± 0.000
		31415	-0.129 ± 0.119
		16180	-0.019 ± 0.034
Diabetes	1	2718	0.000 ± 0.000
		31415	-0.054 ± 0.127
		16180	-0.006 ± 0.059
Diabetes	2	2718	-0.705 ± 0.242
		31415	-0.295 ± 0.407
		16180	-0.149 ± 0.302
Wine	0	2718	-0.080 ± 0.216
		31415	-0.354 ± 0.141
		16180	0.023 ± 0.094
Wine	1	2718	-0.359 ± 0.369
		31415	-0.719 ± 0.229
		16180	0.072 ± 0.291
Wine	2	2718	-4.881 ± 0.787
		31415	-3.594 ± 0.730
		16180	-0.381 ± 0.808

Table 6: Paired target-risk difference, MOM equilibrium minus IW Q, at $n = 8192$. Entries are 10^3 times mean difference plus or minus its paired standard error over 30 source draws, conditional on each fixed training split and target pool. Negative values favor MOM equilibrium.