

Five Days Are Enough: A Sharp Finite-Sample Gap in Calibrated Forecast Aggregation

Arya Somu

Abstract

Exact empirical calibration constrains integer outcome counts. Replacing those counts by arbitrary conditional means can change a robust decision problem. We give a three-expert, five-observation example in which exact calibration determines every binary outcome, yet the continuous relaxation guarantees only three correct classifications instead of five. The example is minimal both in the number of experts and in the number of observations: with at most two experts, or at most four observations and any number of experts, the exact-calibration count polytope is integral. The five-observation gap is also largest possible at that horizon. We prove stability under calibration error, an exact parity law under replication, and a fixed-instance bound showing how replication reduces the normalized gap. The example separates two benchmarks identified in a recent forecast-aggregation formulation; it does not challenge the conditional-mean model itself.

1 Finite counts and conditional means

Fix a horizon $T \geq 1$ and $n \geq 1$ experts. On day t , expert i issues a fixed report $p_{it} \in [0, 1]$ about a binary outcome $y_t \in \{0, 1\}$. For each report p actually used by expert i , let

$$B_{ip} = \{t : p_{it} = p\}, \quad N_{ip} = |B_{ip}|.$$

Exact empirical calibration means

$$\sum_{t \in B_{ip}} y_t = p N_{ip} \quad \text{for every } (i, p). \quad (1)$$

Throughout the exact-calibration results, assume at least one binary sequence satisfies these equations.

Let S be the set of observed joint forecast profiles and N_s the multiplicity of profile s . Write k_s for its number of positive outcomes. Define the incidence matrix and bin totals by

$$A_{(i,p),s} = \mathbf{1}\{s_i = p\}, \quad b_{ip} = p N_{ip}.$$

The continuous count polytope and finite feasible count set are

$$P = \{k \in \mathbb{R}^S : Ak = b, 0 \leq k_s \leq N_s\}, \quad Z = P \cap \mathbb{Z}^S. \quad (2)$$

Every element of Z is realized by a binary sequence, by assigning k_s successes within each profile. Feasibility implies that every component of b is an integer.

A decision rule assigns a mixed action $x_s \in \Delta(\mathcal{A})$ to each observed profile, where $\mathcal{A}$ is a finite action set and $u(a, y)$ is a bounded payoff. Set $u_y(x_s) = \sum_a x_s(a) u(a, y)$ and

$$F(x, k) = \sum_{s \in S} [(N_s - k_s) u_0(x_s) + k_s u_1(x_s)]. \quad (3)$$

The two robust values are

$$V_{\text{fin}} = \max_x \min_{k \in Z} F(x, k), \quad V_{\text{mean}} = \max_x \min_{k \in P} F(x, k). \quad (4)$$

Thus $V_{\text{mean}} \leq V_{\text{fin}}$. Since $F(x, \cdot)$ is affine, the finite inner minimum may equivalently use $\text{conv}(Z)$. It need not use all of P . In particular, randomizing over admissible binary sequences does not remove this distinction.

This issue arises in Guo, Li, and Mu [1]: Definition 2, equation (3), uses binary sequences satisfying empirical calibration, with expectation only over the decision rule. Equation (7) identifies that value with optimization over real profile frequencies. Their equation (10) separately defines a population conditional-mean benchmark, for which a continuous ambiguity set is appropriate. Our counterexample concerns the finite-sequence identification. It neither disputes continuous LP duality nor, by itself, refutes the population online results.

2 A minimal five-observation obstruction

Consider the profiles in Table 1, each observed once. Each expert reports $1/2$ twice and $1/3$ three times. Consequently every report bin must contain exactly one positive outcome.

Profile	Expert 1	Expert 2	Expert 3	Unique feasible outcome
A	$1/2$	$1/2$	$1/2$	1
B	$1/2$	$1/3$	$1/3$	0
C	$1/3$	$1/2$	$1/3$	0
D	$1/3$	$1/3$	$1/2$	0
E	$1/3$	$1/3$	$1/3$	1

Table 1: Three individually calibrated experts determine all five binary outcomes.

Theorem 1. *For Table 1 with correct-classification payoff $u(a, y) = \mathbf{1}\{a = y\}$, the finite robust value is $V_{\text{fin}} = 5$, while the conditional-mean value is $V_{\text{mean}} = 3$.*

Proof. Write (a, b, c, d, e) for the five counts in table order. Calibration gives

$$a + b = a + c = a + d = 1, \quad c + d + e = b + d + e = b + c + e = 1.$$

The entire real feasible set is therefore

$$(a, 1 - a, 1 - a, 1 - a, 2a - 1), \quad \frac{1}{2} \leq a \leq 1. \quad (5)$$

Binary feasibility forces $a = 1$, leaving only $(1, 0, 0, 0, 1)$. Predicting this sequence by profile gives five correct classifications.

For the continuous upper bound, choose

$$q^* = (1/2, 1/2, 1/2, 1/2, 0) \in P.$$

Every mixed action earns $1/2$ on each of the first four profiles and at most 1 on the last, so $V_{\text{mean}} \leq 3$. Conversely, all points of P have total positive mass 2. Always predicting zero therefore earns exactly 3, proving the matching lower bound. $\square$

The missing restriction is particularly visible here: a profile observed once cannot have empirical positive frequency $1/2$. Moreover, q^* cannot be a mixture of calibrated binary sequences, because only one such sequence exists.

Proposition 2. *For any exactly calibrated five-observation instance with nonempty binary feasibility and correct-classification payoff, $V_{\text{fin}} - V_{\text{mean}} \leq 2$. Table 1 attains this bound.*

Proof. Summing one expert's bin equations shows that every feasible real count vector has the same integer total K . The constant policies guarantee K and $5 - K$ correct classifications, respectively. Hence $V_{\text{mean}} \geq \max(K, 5 - K) \geq 3$, while $V_{\text{fin}} \leq 5$. $\square$

3 Why fewer experts or fewer observations cannot work

Theorem 3 (Two-expert exactness). *For one or two experts under exact calibration and nonempty binary feasibility, $P = \text{conv}(Z)$. Consequently $V_{\text{fin}} = V_{\text{mean}}$ for every finite action set and bounded payoff.*

Proof. For two experts, form a bipartite graph whose vertices on each side are that expert's distinct reports. Each observed profile is an edge s , with variable k_s and integer capacity N_s . Calibration fixes the incident sum at each vertex to the integer b_{ip} .

Suppose a feasible point has a nonintegral edge. In the subgraph formed by all nonintegral edges, no vertex can have degree one: the required incident sum and all other incident integral values are integers. This finite subgraph therefore contains a cycle. The cycle has even length. Add and subtract a sufficiently small $\delta > 0$ alternately around it. Both signs of the perturbation preserve all incident sums and capacities, because each affected edge is strictly between its integer bounds. The original point is the midpoint of two distinct feasible points, so it is not a vertex. All vertices of P are integral, proving the assertion.

With one expert, an extreme point cannot have two fractional variables in the same bin, since opposite perturbations preserve the bin sum. It cannot have exactly one fractional variable there because the bin sum is integer. Thus all vertices are again integral. Affine minimization over P and Z agrees for each fixed decision rule, and hence so do the robust values. $\square$

This proof is the classical integrality argument for capacitated bipartite transportation polytopes; see also the total-unimodularity treatment in [2]. The application to empirical counts is what matters here.

Theorem 4 (Four-observation exactness). *If $T \leq 4$, then under exact calibration and nonempty binary feasibility the count polytope P is integral, regardless of the number of experts.*

Proof. Call a report bin mixed if its report lies strictly between 0 and 1. Its required count is an integer strictly between 0 and the bin size, so it contains at least two observations. An expert with two mixed bins at $T \leq 4$ would have two bins of size two. Both would have to report $1/2$ and would consequently be the same bin, a contradiction. Thus each expert has at most one mixed bin.

Work first with day-level variables $q_t \in [0, 1]$. Let F_0 and F_1 be the days on which at least one expert reports 0 and 1, respectively. Those coordinates are fixed to 0 and 1 throughout the relaxation. Binary feasibility ensures that $F_0 \cap F_1 = \emptyset$. Let F be the remaining days.

If F is empty, every coordinate is fixed. Otherwise all days of F lie in the unique mixed bin of every expert. After subtracting the fixed coordinates, each expert's sole nontrivial equation has the form

$$\sum_{t \in F} q_t = c_i.$$

Each c_i is an integer, and a common feasible binary vector shows that these constants coincide. Therefore the remaining feasible set is a cube intersected with a single integer-sum equality. A fractional feasible point must have at least two fractional coordinates, which admit opposite small perturbations. Hence all its vertices are binary.

Finally, the count polytope is the linear image of this day-level polytope under summation within profiles. The image of the convex hull of binary feasible vectors is the convex hull of their integer count vectors. Thus $P = \text{conv}(Z)$. $\square$

Together, Theorems 1, 3, and 4 show that three experts and five observations are simultaneously necessary and sufficient for a strict gap under exact calibration.

4 The obstruction survives calibration error

Permit the common error tolerance $\varepsilon \geq 0$ in every empirical report bin:

$$\left| \frac{1}{N_{ip}} \sum_{t \in B_{ip}} y_t - p \right| \leq \varepsilon. \quad (6)$$

Use the same five forecasts and let $V_{\text{fin}}(\varepsilon)$ and $V_{\text{mean}}(\varepsilon)$ denote the resulting values.

Theorem 5. *For every $0 \leq \varepsilon < 1/3$,*

$$V_{\text{fin}}(\varepsilon) = 5, \quad V_{\text{mean}}(\varepsilon) = \max\{5/2, 3 - 3\varepsilon\}.$$

In particular, for $1/6 \leq \varepsilon < 1/3$ the gap is $5/2$, the largest possible classification gap at this horizon.

Proof. In a bin of size two, changing the target integer count 1 changes its frequency by at least $1/2$. In a bin of size three, the corresponding change is at least $1/3$. Thus for $\varepsilon < 1/3$ the feasible binary set remains the singleton in Table 1.

For $0 \leq \varepsilon \leq 1/6$, the real vector

$$q_\varepsilon = (1/2, 1/2, 1/2, 1/2, 3\varepsilon)$$

satisfies all calibration intervals. Its best classification payoff is $3 - 3\varepsilon$, giving an upper bound. For the lower bound, let the probabilities of predicting 1 on profiles A, B, C, D, E be

$$x = (1/2, 1/6, 1/6, 1/6, 0).$$

Against a real vector (a, b, c, d, e) its payoff is

$$4 - \frac{2}{3}(b + c + d) - e.$$

The three size-three upper calibration bounds imply

$$(c + d + e) + (b + d + e) + (b + c + e) \leq 3 + 9\varepsilon.$$

Dividing by three proves the payoff is at least $3 - 3\varepsilon$.

For $\varepsilon \geq 1/6$, the all- $1/2$ vector is feasible, capping every rule at $5/2$. Independent fair-coin predictions attain $5/2$ against every vector. Finally, every classification instance has continuous value at least $T/2$ and finite value at most T , so $5/2$ is the largest possible gap for $T = 5$. $\square$

At $\varepsilon = 1/3$, the complement $(0, 1, 1, 1, 0)$ of the original binary sequence becomes feasible. The two complementary sequences give an upper bound of $5/2$ against any rule by averaging their payoffs, and fair-coin predictions attain that bound. Therefore $V_{\text{fin}}(\varepsilon) = V_{\text{mean}}(\varepsilon) = 5/2$ for all $\varepsilon \geq 1/3$. The finite value has an exact downward jump at $1/3$, whereas the continuous value has already reached $5/2$ at $1/6$.

The exact-calibration qualification in Theorems 3 and 4 is essential. Already for one day and one expert, report $p = 2/5$ with tolerance $\varepsilon = 2/5$ permits only the binary outcome zero, but the real interval $[0, 4/5]$ contains $1/2$. The finite classification value is 1, and the continuous value is $1/2$.

5 Replicating the forecasts changes the ambiguity set

Repeat each profile in Table 1 exactly r times and impose exact calibration on the resulting $5r$ observations. This operation permits new outcome sequences; it does not require each five-day block to be calibrated separately.

Proposition 6 (Exact replication law). *The continuous value is $3r$, and the finite value is*

$$V_{\text{fin}}(r) = \begin{cases} 5, & r = 1, \\ 3r, & r \text{ even}, \\ 3r + 1, & r \geq 3 \text{ odd}. \end{cases}$$

Proof. The real count vectors are

$$(a, r - a, r - a, r - a, 2a - r), \quad r/2 \leq a \leq r.$$

For finite counts, a must be an integer. Always predicting zero guarantees $3r$ in the real model; the point $a = r/2$ caps all rules at $3r$. For even r this point is also integral, proving equality of both values.

For odd $r \geq 3$, the feasible count vector at $a = (r + 1)/2$ caps any rule at

$$4\frac{r+1}{2} + (r-1) = 3r + 1.$$

The last term uses the majority-zero action on the final profile, whose positive count is 1. Conversely, the rule that predicts 1 only on profile A earns $2a + 2r$ against every feasible count vector, at least $3r + 1$. The case $r = 1$ is Theorem 1. $\square$

Thus simply doubling the observed forecast multiset changes the finite guarantee per day from 1 to $3/5$. This is a consequence of the changed set of admissible outcomes, not a loss from an inferior decision rule.

6 Repairs and a general replication bound

There are two direct ways to maintain a consistent benchmark: retain Z (or its integer hull) for empirical calibration, or use P as the definition when calibration is required only in expectation. For at most two experts an additional constructive repair handles approximate calibration.

Proposition 7 (Rounded bounds with two experts). *Allow a tolerance ε_{ip} for each report bin. Define*

$$L_{ip} = \lceil N_{ip}(p - \varepsilon_{ip}) \rceil, \quad H_{ip} = \lfloor N_{ip}(p + \varepsilon_{ip}) \rfloor.$$

For at most two experts, the rounded polytope

$$P_{\text{round}} = \{k \in \mathbb{R}^S : L \leq Ak \leq H, 0 \leq k \leq N\}$$

is integral. If nonempty, it gives exactly the finite robust value for every bounded finite-action payoff.

Proof. Integer bin counts satisfy the original intervals exactly when they satisfy the rounded ones. The bipartite incidence matrix A is totally unimodular. One elementary verification considers a square submatrix: if a column has at most one nonzero, expand its determinant and induct; otherwise every column has one entry in each side of the bipartition, and the signed sum of its rows is zero. All square subdeterminants are therefore 0 or ± 1 .

Appending negative copies of rows and signed identity rows preserves this property. The inequality matrix $[A; -A; I; -I]$ is thus totally unimodular, with integer right-hand side. At any vertex, a nonsingular active subsystem has determinant ± 1 , so Cramer's rule gives integer coordinates. The one-expert incidence matrix has the same property. Affine minimization now agrees with minimization over integer counts. $\square$

A rational-polytope argument also explains why the fixed-instance normalized discrepancy vanishes under uniform replication. The bound below does not assert a dimension-independent constant or an efficient denominator bound.

Theorem 8 (Fixed-instance replication bound). *Let the base exact-calibration polytope P have an integer feasible point k^0 , and put $T_0 = \sum_s N_s$. Choose a positive integer D such that Dv is integral for every vertex v of P . Let V_{mean} be its base continuous value, and let $V_{\text{fin}}(r)$ be the finite value after multiplying every profile multiplicity by r . Define*

$$C = \max_{a \in \mathcal{A}} |u(a, 1) - u(a, 0)|.$$

Then

$$0 \leq V_{\text{fin}}(r) - rV_{\text{mean}} \leq CT_0(r \bmod D). \quad (7)$$

Consequently equality holds for $D \mid r$, and the difference per observation is at most $C(D - 1)/r$.

Proof. The data defining P are integer, so its vertices are rational and such a finite D exists. Replication changes the real count polytope to rP . Fix a decision rule x and choose a vertex v minimizing its base affine payoff. Write $r = qD + s$, where $0 \leq s < D$. The vector

$$k = qDv + sk^0$$

is integral and belongs to rP , since k/r is a convex combination of v and k^0 . Moreover,

$$\|k - rv\|_1 = s\|k^0 - v\|_1 \leq sT_0.$$

Each count coefficient in the payoff has absolute value at most C . Therefore the finite worst-case payoff for this rule is at most its continuous worst-case payoff plus CsT_0 . Taking the maximum over rules proves the upper bound. The lower bound follows from inclusion of integer counts in real counts. $\square$

For payoffs bounded by $|u| \leq U$, one may use $C \leq 2U$. For correct classification, $C = 1$. Our five-observation polytope has $D = 2$; Proposition 6 gives its sharper exact values.

7 Context, verification, and status

Continuous calibrated posterior feasibility is well established. Arieli, Babichenko, Sandomirskiy, and Tamuz [3] characterize feasible joint posterior beliefs; their dominated-measure formulation is closely related to bounded marginal constraints. Wang and Yang [4] study integrality gaps for calibration distance, where outcomes are fixed and forecasts are repaired. Here forecasts are fixed and binary outcomes vary subject to simultaneous calibration. Bipartite-flow integrality and rational denominator clearing are classical; they are used here as tools.

The accompanying standard-library script `verify_calibration.py` uses integers and rational arithmetic. It enumerates all 32 binary outcome vectors, independently enumerates the relaxed vertices using active box constraints, and verifies upper and lower payoff certificates. It also checks the replication formula for $r = 1, \dots, 100$ and rational certificates for the approximate-calibration result. These finite checks supplement the proofs, rather than replace them.

This is an AI-assisted research draft, with three agents used for scouting and independent mathematical/source audits. It has not received external peer review or author feedback. The literature search checked the cited formulations and related work, but cannot establish priority. The claimed mathematical contributions of this note are the explicit obstruction, its minimality and extremality, and the stated aggregation-specific extensions. No general hardness result or failure of the continuous population theory is asserted.

References

- [1] X. Guo, Y. Li, and Y. Mu. *Robust Aggregation of Calibrated Forecasts*. arXiv:2606.31020v1, 2026. <https://arxiv.org/html/2606.31020v1>.
- [2] M. X. Goemans. *Linear Programming and Polyhedral Combinatorics*. MIT 18.433 lecture notes, February 28, 2011, Section 3.5. <https://math.mit.edu/~goemans/18433S11/polyhedral.pdf>.
- [3] I. Arieli, Y. Babichenko, F. Sandomirskiy, and O. Tamuz. Feasible joint posterior beliefs. *Journal of Political Economy*, 129(9):2546–2594, 2021. <https://doi.org/10.1086/714993>.
- [4] Z. Wang and X. Yang. *Sharp Integrality Gaps in Calibration Distance*. arXiv:2610.05679v1, 2026. <https://arxiv.org/abs/2610.05679>.