NO PERFECTLY TRUTHFUL SEQUENTIAL CALIBRATION MEASURE

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

A calibration measure should assign sublinear error to correct probabilistic forecasts, linear error to a fixed incorrect forecast, and give a forecaster no incentive to misreport its beliefs. Whether these three requirements can coexist under perfect truthfulness has remained open for sequential prediction. We prove that they cannot. Our main structural result shows that every bounded perfectly truthful measure computed from the outcomes and the forecasts observed along the realized path has a tree-additive Bayes risk. This representation requires no continuity, differentiability, strict truthfulness, or downstream decision guarantee. Although the measure itself may be nonsmooth, one fixed report probability avoids all nondifferentiability points across every horizon. At that report, the measure is forced to be a sum of nonnegative one-step proper losses. Concavity then implies that, under fair independent outcomes, this fixed incorrect report has expected error at most twice the truthful error. Completeness makes the latter sublinear, contradicting soundness. This resolves the perfect-truthfulness question posed by Haghtalab et al. and separates sequential evaluation from recent positive results for batch calibration. The boundary is sharp: if the evaluator can inspect the full forecast tree rather than one realized path, a simple squared-count score satisfies perfect truthfulness, completeness, and product soundness simultaneously.

1 INTRODUCTION

Probabilistic forecasts are useful only when their numerical values can be trusted. Calibration captures a basic form of this requirement: among events assigned probability p, approximately a p fraction should occur. Calibration has classical roots in forecasting and statistics (Brier, 1950; Dawid, 1982), supports calibrated learning against arbitrary outcome sequences (Foster & Vohra, 1998), and is now a standard criterion for uncertainty estimates in machine learning.

A numerical calibration measure affects incentives as well as evaluation. If a forecaster knows the true conditional probability of the next outcome, minimizing the reported calibration error should not reward a lie. Haghtalab et al. (2024) formalized this requirement in sequential prediction. A measure is perfectly truthful when the strategy that reports every true conditional probability minimizes expected error for every joint distribution of the outcome sequence. They paired truthfulness with two minimal usefulness conditions. Completeness requires sublinear expected error for truthful i.i.d. Bernoulli forecasts. Soundness requires linear error when the same data are evaluated against a fixed incorrect probability.

Known measures expose a sharp tension. Standard calibration errors are complete and sound but can reward strategic forecasts. Subsampled smooth calibration error reduces the incentive gap to a constant multiplicative factor, but it is not perfectly truthful (Haghtalab et al., 2024). Standard additive proper scoring rules are perfectly truthful, but their one-step noise accumulates to linear error and violates completeness. Qiao & Zhao (2025) ruled out perfect truthfulness for measures that also control every downstream decision problem. Their theorem relies on this decision-theoretic condition and does not address the three basic requirements alone. In the batch setting, the answer is positive: Hartline et al. (2026) constructed a perfectly truthful, complete, and sound measure using variance additivity. They explicitly retained the sequential case as an open problem.

We close the sequential question with a negative answer. The result holds under exactly the standard information model. At time t, the evaluator eventually sees the outcome x_t and the forecast p_t made after observing $x_{<t}$. It does not see forecasts at histories that did not occur.

Theorem 1 (Informal)**.** *No bounded sequential calibration measure is simultaneously perfectly truthful, complete, and sound. In fact, deterministic completeness and completeness only for fair i.i.d. outcomes already imply that one fixed incorrect constant forecast has sublinear expected error.*

The proof is not a lower bound for a particular calibration statistic. It classifies every perfectly truthful realized-path score. A forecast strategy is a probability tree. In the interior, it defines a joint distribution on the 2^T leaves. Perfect truthfulness turns the calibration measure into a proper scoring rule for that leaf distribution. The special information structure is crucial: the score of a realized leaf can inspect only conditional probabilities on its root-to-leaf path.

Our main representation theorem shows that the Bayes risk, or entropy, of every bounded path-local proper score decomposes over the internal nodes of the probability tree. Each node contributes its visitation probability times a concave binary entropy. This is a converse to the familiar construction that sums one-step proper losses. The representation holds even when the original measure is discontinuous and chooses different supporting hyperplanes at nonsmooth points.

The remaining obstacle is precisely that nonsmooth freedom. Each one-dimensional concave node entropy is nondifferentiable at only countably many probabilities. There are countably many nodes over all finite horizons, so one fixed $\beta \in (1/3, 2/3)$, distinct from $1/2$, avoids every exceptional point. At the constant report β, the supporting score is unique and equals the sum of the nodewise tangent losses. Deterministic completeness and boundedness let us normalize every node entropy to vanish at 0 and 1. The tangent losses are then nonnegative. A short concavity argument gives, at every node,

$$\mathbb{E}_{Y\sim\mathrm{Ber}(1/2)}\ell_\beta(Y) \le 2h(1/2).$$

Summing with the true history probabilities proves that the wrong constant report costs at most twice the truthful error. Product completeness and product soundness now contradict each other.

The theorem pinpoints the information boundary. Suppose the evaluator receives the entire probability tree. It can compute the tree's reported expectation μ of $\sum_t X_t$ and use $T^{-1}(\sum_t x_t - \mu)^2$. This score is perfectly truthful. Its truthful risk under i.i.d. Bernoulli data is $O(1)$, while any fixed incorrect product tree has $\Omega(T)$ risk. The construction is unavailable from one realized path because μ depends on forecasts in unrealized branches.

Contributions. Our results are:

- A no-regularity representation theorem for bounded perfectly truthful realized-path scores. Their entropy is a sum of history-dependent binary entropies over the forecast tree.
- An impossibility theorem resolving the open coexistence of perfect truthfulness, completeness, and soundness in sequential calibration. It removes the decision-theoretic assumption from the previous impossibility result.
- A quantitative witness. For one fixed $\beta \neq 1/2$, the expected error of reporting β on fair coins is at most twice the truthful expected error at every horizon.
- A matching information separation. Full-tree access admits a perfectly truthful, deterministically complete, product-complete, and product-sound score.

2 Related work

Truthful calibration. Haghtalab et al. (2024) introduced truthfulness for sequential calibration measures, documented large incentive gaps for standard measures, and constructed subsampled smooth calibration error with constant-factor truthfulness. Their discussion asks whether perfect truthfulness can coexist with completeness and soundness. Qiao & Zhao (2025) studied measures that additionally control downstream regret. They constructed subsampled step calibration with strong approximate guarantees and showed that every complete decision-theoretic measure is non-truthful in the unsmoothed setting. Their lower bound uses domination of U-calibration. Our result

Result	Setting	Complete	Sound	Perfect truth
Haghtalab et al. (2024), SSCE	sequential	yes	yes	no, constant factor
Qiao & Zhao (2025)	seq., decision-theoretic	yes	yes	impossible
Hartline et al. (2026), ATB	independent batch	yes	yes	yes
This work	sequential path	yes	yes	impossible
This work, full-tree score	full forecast tree	yes	product	yes

Table 1: The information model separates the positive batch and full-tree results from the sequential impossibility.

neither assumes nor proves any downstream decision guarantee. It instead derives a representation from perfect truthfulness itself.

Hartline et al. (2026) gave the first perfectly truthful batch calibration measure. Their averaged two-bin error exploits variance additivity for independent batch outcomes. It is complete, sound, continuous, and sample efficient. The same work states the sound, complete, perfectly truthful sequential case as open. A multiclass batch extension was subsequently developed by Lu et al. (2025). Our theorem explains why batch constructions do not transfer to adaptive sequential reports.

Proper scores and locality. Proper scoring rules elicit probabilistic beliefs and correspond to supporting hyperplanes of concave entropies (Gneiting & Raftery, 2007; Waghmare & Ziegel, 2026). Dawid et al. (2012) characterize graph-local proper scores on finite outcome spaces under regularity conditions. Smith & Bickel (2020) study strict proper scores that satisfy additivity requirements across marginal and conditional distributions. Prequential forecasting commonly constructs a proper sequence score by summing one-step proper scores (Pacchiardi et al., 2024). Our representation runs in the converse direction. It assumes only realized-path observability and weak propriety, then derives a history-dependent tree decomposition. The proof uses bounded binary properness rather than differentiability or a graph-clique expansion.

Calibration measures. Calibration against arbitrary outcome sequences goes back to Foster & Vohra (1998), with later work studying smooth calibration and strategic hedging (Kakade & Foster, 2008; Foster & Hart, 2021). The systematic comparison of calibration measures includes distance, continuity, statistical estimation, and decision consequences (Błasiok et al., 2023). Our result concerns a different axis: exact incentive compatibility when forecasts can adapt to observed outcomes.

3 Setup

Write $\{0,1\}^{<T} = \bigcup_{t=0}^{T-1}\{0,1\}^t$, and let $x_{<t} = (x_1, \dots, x_{t-1})$. A deterministic forecaster is a binary probability tree $A = (a_u)_{u\in\{0,1\}^{<T}}$, where $a_u \in [0,1]$ is the prediction after history u. On outcome sequence $x \in \{0,1\}^T$, its realized prediction vector is

$$A(x) = (a_{x_{<1}}, \dots, a_{x_{<T}}).$$

A calibration measure is a family $M = (M_T)_{T\geq 1}$ with $M_T : \{0,1\}^T \times [0,1]^T \to [0,T]$. Its expected error under outcome distribution P and strategy A is

$$\mathrm{err}_{M_T}(P, A) = \mathbb{E}_{X\sim P} M_T(X, A(X)).$$

For a full-support $P \in \Delta(\{0,1\}^T)$, its truthful tree has node reports

$$q_u = P(X_{|u|+1} = 1 \mid X_{1:|u|} = u).$$

At zero-probability histories the truthful report can be chosen arbitrarily.

Definition 2 (Truthfulness, completeness, and soundness)**.** The measure M is perfectly truthful if the truthful tree minimizes expected error for every T and every P. It is deterministically complete if $M_T(x,x) = 0$ for all T, x. It is product-complete if, for every fixed $\alpha \in [0,1]$,

$$\mathbb{E}_{X\sim \mathrm{Ber}(\alpha)^{\otimes T}} M_T(X, \alpha\mathbf{1}) = o_\alpha(T).$$

It is product-sound if, for every fixed $\alpha \neq \beta$,

$$\mathbb{E}_{X\sim \mathrm{Ber}(\alpha)^{\otimes T}} M_T(X, \beta\mathbf{1}) = \Omega_{\alpha,\beta}(T).$$

Deterministic and product completeness are exactly the two completeness clauses of Haghtalab et al. (2024). Product soundness is one of their two soundness clauses. Their other soundness clause requires linear error on the pointwise opposite prediction. The later definition of Qiao & Zhao (2025) adds sublinear error on every empirically calibrated pair. Our impossibility does not need either addition.

Every interior report tree A, with $a_u \in (0,1)$, induces a unique full-support distribution Q_A on the leaves:

$$Q_A(x) = \prod_{t=1}^{T} a_{x_{<t}}^{x_t} (1 - a_{x_{<t}})^{1-x_t}.$$

Conversely, the truthful tree of Q_A is A. Define the induced leaf score and its Bayes risk by

$$S_T(Q, x) = M_T(x, A_Q(x)), \qquad H_T(P) = \mathbb{E}_{X\sim P} S_T(P, X). \tag{1}$$

Perfect truthfulness says that S_T is a proper scoring rule:

$$H_T(P) \le \mathbb{E}_{X\sim P} S_T(Q, X) \tag{2}$$

for all full-support P, Q. It follows that H_T is concave. Crucially, $S_T(Q, x)$ depends only on the conditional probabilities of Q along the realized path to x.

4 The tree-additive representation

The representation theorem is stated for binary trees because that is the calibration framework of Definition 2. For a history u, write $P(u) = P(X_{1:|u|} = u)$.

Theorem 3 (Tree-additive entropy)**.** *Fix T and a bounded perfectly truthful realized-path score. There are bounded concave functions $h_{T,u} : [0,1] \to \mathbb{R}$, one for every $u \in \{0,1\}^{<T}$, such that every full-support distribution P satisfies*

$$H_T(P) = \sum_{u\in\{0,1\}^{<T}} P(u) h_{T,u}(q_u). \tag{3}$$

If the score is nonnegative and deterministically complete, the functions can be chosen so that

$$h_{T,u}(0) = h_{T,u}(1) = 0, \qquad h_{T,u}(q) \ge 0. \tag{4}$$

At any report distribution Q where every $h_{T,u}$ is differentiable at the corresponding conditional q_u, the actual realized score is

$$S_T(Q, x) = \sum_{u \prec x} \left[h_{T,u}(q_u) + (x_{|u|+1} - q_u) h'_{T,u}(q_u) \right]. \tag{5}$$

Equation (3) says that the expected truthful score is assembled from binary uncertainty at each visited history. Equation (5) is stronger at smooth reports. It identifies the actual measure, not only its expectation.

The proof starts with a binary rigidity fact. It is useful because the conditional score below a branch can depend on the reported probability above that branch.

Lemma 4 (Shared-side uniqueness)**.** *Let (ℓ_0, ℓ_1) and $(\widetilde{\ell}_0, \ell_1)$ be bounded proper binary losses on $(0,1)$. Then $\ell_0 - \widetilde{\ell}_0$ is constant. The symmetric statement holds when the zero-outcome component is shared.*

For $0 < p < q < 1$, properness at true parameters p and q traps the increment of ℓ_0:

$$\frac{p}{1-p}\big(\ell_1(p) - \ell_1(q)\big) \le \ell_0(q) - \ell_0(p) \le \frac{q}{1-q}\big(\ell_1(p) - \ell_1(q)\big). \tag{6}$$

The same interval traps the increment of $\widetilde{\ell}_0$. On a refining partition, the sum of the interval widths tends to zero because ℓ_1 is bounded and monotone. Thus the two increments agree. Appendix B gives the details.

To see how Lemma 4 yields Theorem 3, split a leaf distribution at the root as

$$P = ((1-a)P_0, aP_1).$$

Path locality writes the score on root branch b as $S_b(p, Q_b, x)$, with no dependence on Q_{1-b}. For fixed root report p, this is a proper score for the conditional leaf distribution in branch b. Let its entropy be $K_b^p(R)$. Varying only the root report shows that

$$p \mapsto \big(K_0^p(R_0), K_1^p(R_1)\big)$$

is a proper binary loss for every R_0, R_1. Holding one conditional fixed and applying Lemma 4 gives

$$K_b^p(R) = a_b(p) + K_b(R). \tag{7}$$

The pair (a_0, a_1) is proper at the root, and each K_b is the entropy of a bounded path-local proper score on a depth-$(T-1)$ tree. Recursion proves Equation (3).

The endpoint normalization needs no continuity. If P approaches a deterministic leaf x, perfect truthfulness allows comparison with the strategy hard-coded to predict x. It loses zero on x and at most T elsewhere, so

$$0 \le H_T(P) \le T(1 - P(x)) \longrightarrow 0. \tag{8}$$

The endpoint values of the node entropies therefore sum to zero on every root-to-leaf path. Affine node potentials set both endpoints of every $h_{T,u}$ to zero while telescoping out of Equation (3). Concavity then gives nonnegativity. Finally, the score vector in Equation (2) is a supporting hyperplane of H_T. When all node entropies are differentiable, this hyperplane is unique. Differentiating the perspective term $P(u)h_{T,u}(q_u)$ gives Equation (5).

5 Perfect truthfulness contradicts soundness

The local comparison behind the impossibility is elementary.

Lemma 5 (Tangent comparison)**.** *Let $h : [0,1] \to \mathbb{R}_+$ be concave with $h(0) = h(1) = 0$, and suppose h is differentiable at $p \in (0,1)$. Define*

$$\ell_p(b) = h(p) + (b-p)h'(p), \qquad b \in \{0,1\}.$$

For every $q \in (0,1)$,

$$(1-q)\ell_p(0) + q\ell_p(1) \le C_{p,q}h(q), \tag{9}$$

where

$$C_{p,q} = \max\left\{\frac{q}{p}, \frac{1-q}{1-p}\right\} \max\left\{\frac{p}{q}, \frac{1-p}{1-q}\right\}.$$

The tangent line lies above the zero endpoint values, hence $\ell_p(0), \ell_p(1) \ge 0$. Reweighting these two nonnegative numbers gives

$$(1-q)\ell_p(0) + q\ell_p(1) \le \max\left\{\frac{q}{p}, \frac{1-q}{1-p}\right\} h(p).$$

Concavity between p, q, and the appropriate endpoint gives the second factor in Equation (9).

Theorem 6 (Sequential impossibility)**.** *No calibration measure $M_T : \{0,1\}^T \times [0,1]^T \to [0,T]$ is simultaneously perfectly truthful, deterministically complete, product-complete, and product-sound.*

Proof. Apply Theorem 3 at every horizon. Each concave node entropy is nondifferentiable at at most countably many points. The union of these exceptional sets over all nodes and horizons is countable. Choose one fixed

$$\beta \in (1/3, 2/3) \setminus \{1/2\}$$

outside this union, and set $\alpha = 1/2$.

At the product report $Q_\beta = \mathrm{Ber}(\beta)^{\otimes T}$, every node conditional is β. Equation (5) and Lemma 5 give

$$\begin{aligned}
\mathbb{E}_{X\sim\mathrm{Ber}(1/2)^{\otimes T}} M_T(X, \beta\mathbf{1}) &\le C_{\beta,1/2} \sum_u P_{1/2}(u) h_{T,u}(1/2) \\
&= C_{\beta,1/2} H_T(P_{1/2}) \\
&= C_{\beta,1/2} \mathbb{E}_{X\sim\mathrm{Ber}(1/2)^{\otimes T}} M_T(X, \tfrac{1}{2}\mathbf{1}).
\end{aligned}$$

For $\beta \in (1/3, 2/3)$, the constant is

$$C_{\beta,1/2} = \max\left\{\frac{\beta}{1-\beta}, \frac{1-\beta}{\beta}\right\} < 2.$$

Product completeness makes the last expectation $o(T)$. Product soundness requires the first expectation to be $\Omega(T)$, a contradiction. □

The same history law $P_{1/2}(u)$ weights both sides of the nodewise comparison. No likelihood-ratio comparison between the truthful and incorrect product distributions is used. This point rules out an apparent escape in which a measure concentrates penalties on histories that are rare under the reported model.

Corollary 7 (Quantitative failure)**.** *For every bounded, perfectly truthful, deterministically complete family M, there is a fixed $\beta \in (1/3, 2/3) \setminus \{1/2\}$ such that, for every T,*

$$\mathbb{E}_{P_{1/2}} M_T(X, \beta\mathbf{1}) \le 2\mathbb{E}_{P_{1/2}} M_T(X, \tfrac{1}{2}\mathbf{1}).$$

Thus the obstruction is stronger than the failure of a particular asymptotic rate. A fixed wrong report cannot even be separated from truth by more than a universal factor.

6 Full-tree access restores possibility

The standard evaluator observes $A(X)$, not the reports at histories outside the realized path. We now show that this restriction is exactly where the impossibility enters.

Suppose an evaluator receives the full probability tree A. Let Q_A be its induced leaf distribution and define

$$\mu(A) = \mathbb{E}_{Y\sim Q_A} \sum_{t=1}^{T} Y_t, \qquad M_T^{\mathrm{tree}}(x, A) = \frac{1}{T}\left(\sum_{t=1}^{T} x_t - \mu(A)\right)^2. \tag{10}$$

Proposition 8 (Full-tree possibility)**.** *The score in Equation (10) takes values in $[0, T]$, is perfectly truthful for every joint outcome distribution, is zero on deterministic truth, and is product-complete and product-sound.*

Proof. For true distribution P, write $Z = \sum_t X_t$. Then

$$\mathbb{E}_P M_T^{\mathrm{tree}}(X, A) = \frac{\mathrm{Var}_P(Z)}{T} + \frac{(\mathbb{E}_P Z - \mu(A))^2}{T}.$$

The truthful tree induces P and reports $\mu(A_P) = \mathbb{E}_P Z$, so it is optimal. For deterministic $P = \delta_x$, its score is zero. Under $P_\alpha = \mathrm{Ber}(\alpha)^{\otimes T}$, the truthful expected score is $\alpha(1-\alpha)$. The constant tree $A \equiv \beta$ has expected score

$$\alpha(1-\alpha) + T(\alpha - \beta)^2.$$

This is linear for every fixed $\alpha \neq \beta$. □

The score only elicits the expected total count, so truth need not be its unique minimizer. Perfect truthfulness in the calibration definition also does not require uniqueness. Computing $\mu(A)$ requires integrating over unrealized branches. One path of scalar conditional forecasts contains insufficient information.

7 Discussion

Perfect truthfulness in sequential calibration is impossible for a structural reason. A measure that sees only realized forecasts becomes a path-local proper score. Its uncertainty must accumulate through conditional binary entropies. Completeness asks this accumulated truthful uncertainty to be sublinear. Concavity then prevents the same node losses from assigning linear error to a fixed incorrect probability.

The proof leaves several directions open. Approximate truthfulness remains possible and quantitatively important. Existing constant-factor constructions show that an exact-to-approximate transition changes the problem sharply. It would be useful to determine the best universal factor under the original completeness and soundness axioms. The full-tree result also suggests studying limited counterfactual access, such as a bounded number of queried unrealized histories or auxiliary forecasts of aggregate outcomes. Finally, a multiclass extension should replace binary node entropies by concave functions on a simplex. The tangent comparison extends directly, while a no-regularity shared-component factorization for categorical nodes requires additional care.

REFERENCES

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1727–1740, 2023. doi: 10.1145/3564246.3585182.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

A. Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

A. Philip Dawid, Steffen Lauritzen, and Matthew Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012. doi: 10.1214/12-AOS972.

Dean P. Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12):3447–3490, 2021.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117237–117290, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4cbcae8cfc8aa3ae897a1296e4e0cac-Abstract-Conference.html.

Jason Hartline, Lunjia Hu, and Yifan Wu. A perfectly truthful calibration measure. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3185–3223. PMLR, 2026. URL https://proceedings.mlr.press/v336/hartline26a.html.

Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

Yuxuan Lu, Yifan Wu, Jason Hartline, and Lunjia Hu. Making and evaluating calibrated forecasts. *arXiv preprint arXiv:2510.06388*, 2025.

Lorenzo Pacchiardi, Rilwan A. Adewoyin, Peter Dueben, and Ritabrata Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024.

Mingda Qiao and Eric Zhao. Truthfulness of decision-theoretic calibration measures. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4686–4739. PMLR, 2025. URL https://proceedings.mlr.press/v291/qiao25a.html.

Zachary J. Smith and J. Eric Bickel. Additive scoring rules for discrete sample spaces. *Decision Analysis*, 17(2):115–133, 2020. doi: 10.1287/deca.2019.0398.

Kartik Waghmare and Johanna F. Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*, 13:271–296, 2026. doi: 10.1146/annurev-statistics-042424-050626.

A PROPER-SCORE PRELIMINARIES

We record the finite-dimensional facts used in the proof. Let $\mathcal{Y}$ be finite. A loss-oriented score $S(Q, y)$ is proper if

$$\mathbb{E}_{Y\sim P} S(P, Y) \le \mathbb{E}_{Y\sim P} S(Q, Y)$$

for all $P, Q \in \Delta(\mathcal{Y})$. Its entropy is $H(P) = \mathbb{E}_P S(P, Y)$. Since

$$H(P) = \inf_Q \mathbb{E}_P S(Q, Y),$$

it is an infimum of affine functions of P and is therefore concave. For each Q, the vector $S(Q, \cdot)$ defines an affine majorant of H:

$$H(P) \le \sum_y P(y) S(Q, y), \qquad H(Q) = \sum_y Q(y) S(Q, y). \tag{11}$$

Conversely, a choice of supporting affine majorant at every Q defines a proper score. This is the finite-outcome Savage representation (Gneiting & Raftery, 2007; Dawid et al., 2012).

If H is differentiable at an interior Q, the supporting affine function in Equation (11) is unique. To see the normalization explicitly, extend H homogeneously to the positive cone by

$$\overline{H}(z) = \left(\sum_y z_y\right) H\left(\frac{z}{\sum_y z_y}\right).$$

The score vector is a supergradient of the concave one-homogeneous function $\overline{H}$. At differentiability points it equals $\nabla \overline{H}(Q)$.

B PROOF OF SHARED-SIDE UNIQUENESS

We prove Lemma 4. A bounded pair (ℓ_0, ℓ_1) is proper when, for all $p, q \in (0, 1)$,

$$(1-q)\ell_0(q) + q\ell_1(q) \le (1-q)\ell_0(p) + q\ell_1(p). \tag{12}$$

Fix $0 < p < q < 1$. Apply Equation (12) first with true parameter p and report q, then with true parameter q and report p. With

$$\Delta_0 = \ell_0(q) - \ell_0(p), \qquad \Delta_1 = \ell_1(p) - \ell_1(q),$$

the two inequalities become

$$\frac{p}{1-p}\Delta_1 \le \Delta_0 \le \frac{q}{1-q}\Delta_1. \tag{13}$$

The interval is nonempty only if $\Delta_1 \ge 0$, and then $\Delta_0 \ge 0$. Thus ℓ_0 is nondecreasing and ℓ_1 is nonincreasing.

Let $(\widetilde{\ell}_0, \ell_1)$ also be proper. Its increment $\widetilde{\Delta}_0$ lies in the same interval in Equation (13). Set $g(r) = r/(1-r)$. On any compact interval $[a, b] \subset (0, 1)$, take a partition $a = x_0 < \cdots < x_n = b$. Then

$$\begin{aligned}
&\left|[\ell_0(b) - \ell_0(a)] - [\widetilde{\ell}_0(b) - \widetilde{\ell}_0(a)]\right| \\
&\quad \le \sum_{i=1}^{n} [g(x_i) - g(x_{i-1})][\ell_1(x_{i-1}) - \ell_1(x_i)] \\
&\quad \le \max_i [g(x_i) - g(x_{i-1})][\ell_1(a) - \ell_1(b)].
\end{aligned}$$

The last expression tends to zero along partitions whose mesh tends to zero. Therefore the two increments agree on every compact interval, and $\ell_0 - \widetilde{\ell}_0$ is constant. Swapping outcomes proves the symmetric statement.

C Proof of the entropy factorization

We prove Equation (3) of Theorem 3 by induction on T. The case $T = 1$ is the ordinary binary entropy of a bounded proper score.

For $T \geq 2$, split the leaf space into $\mathcal{Y}_0 = \{x : x_1 = 0\}$ and $\mathcal{Y}_1 = \{x : x_1 = 1\}$. Parameterize a full-support report distribution as

$$Q = ((1-p)Q_0, pQ_1).$$

For $x \in \mathcal{Y}_b$, realized-path locality implies

$$S_T(Q, x) = S_b(p, Q_b, x), \tag{14}$$

with no dependence on Q_{1-b}.

Fix $p \in (0, 1)$. For every true conditional R_b, global propriety with root probability p and all other reports fixed at truth implies

$$\mathbb{E}_{R_b} S_b(p, R_b, Y) \leq \mathbb{E}_{R_b} S_b(p, Q_b, Y).$$

Thus $S_b(p, \cdot, \cdot)$ is a bounded proper score on $\mathcal{Y}_b$. Denote its entropy by

$$K_b^p(R_b) = \mathbb{E}_{Y \sim R_b} S_b(p, R_b, Y).$$

Now fix true conditionals R_0, R_1 and vary only the root report. Global propriety gives, for every true root probability $a \in (0, 1)$,

$$(1-a)K_0^a(R_0) + aK_1^a(R_1) \leq (1-a)K_0^p(R_0) + aK_1^p(R_1). \tag{15}$$

Hence $p \mapsto (K_0^p(R_0), K_1^p(R_1))$ is a bounded proper binary loss for every pair R_0, R_1.

Fix a reference R_0°. Holding R_1 fixed in Equation (15), Lemma 4 shows that

$$K_0^p(R_0) - K_0^p(R_0^\circ)$$

is independent of p. Thus $K_0^p(R_0) = a_0(p) + K_0(R_0)$. The symmetric argument gives $K_1^p(R_1) = a_1(p) + K_1(R_1)$. Substituting these expressions into Equation (15) shows that (a_0, a_1) is a proper binary loss because the $K_b(R_b)$ terms are independent of the root report. Its entropy

$$h_{T,\varnothing}(a) = (1-a)a_0(a) + aa_1(a)$$

is concave.

Choose any $p_0 \in (0, 1)$. The branch score

$$\widetilde{S}_b(Q_b, y) = S_b(p_0, Q_b, y) - a_b(p_0)$$

is bounded, proper, and path-local, with entropy K_b. The induction hypothesis decomposes K_b over the internal nodes below branch b. Finally,

$$\begin{aligned} H_T(P) &= (1-a)K_0^a(P_0) + aK_1^a(P_1) \\ &= h_{T,\varnothing}(a) + (1-a)K_0(P_0) + aK_1(P_1), \end{aligned}$$

which yields Equation (3).

D Boundary normalization and score identification

We first prove Equation (8) carefully. For a leaf $x \in \{0, 1\}^T$, define a valid hard-coded strategy A^x by $A^x(u) = x_{|u|+1}$ at every history of the appropriate depth. On outcome x, its realized prediction vector equals x, so deterministic completeness gives $M_T(x, A^x(x)) = 0$. Since every loss is at most T, perfect truthfulness gives

$$0 \leq H_T(P) \leq \mathbb{E}_P M_T(X, A^x(X)) \leq T(1 - P(x)). \tag{16}$$

Each bounded concave node entropy in Equation (3) has finite one-sided limits at 0 and 1. Write these limits as $e_{u,0}$ and $e_{u,1}$. Approach a deterministic leaf x through interior trees whose conditional

probability at every node on the path tends to the corresponding bit of x. Terms at off-path nodes vanish because their visitation probabilities tend to zero. Equations (3) and (16) imply

$$\sum_{u \prec x} e_{u,x_{|u|+1}} = 0 \tag{17}$$

for every leaf x.

Set $c_\varnothing = 0$ and recursively define $c_{ub} = c_u - e_{u,b}$. Equation (17) gives $c_x = 0$ at every leaf. Replace

$$h_{T,u}(q) \quad \text{by} \quad \widetilde{h}_{T,u}(q) = h_{T,u}(q) + (1-q)c_{u0} + qc_{u1} - c_u. \tag{18}$$

The change in the entropy sum is

$$\begin{aligned}\sum_u P(u)[(1-q_u)c_{u0} + q_u c_{u1} - c_u] &= \sum_u [P(u0)c_{u0} + P(u1)c_{u1} - P(u)c_u] \\ &= \sum_x P(x)c_x - c_\varnothing = 0.\end{aligned}$$

Both endpoint values in Equation (18) are zero. The transformed functions remain concave, so they are nonnegative. We drop the tilde below.

Suppose every $h_{T,u}$ is differentiable at the conditional q_u of an interior report Q. In leaf-probability coordinates, the node contribution is the perspective

$$[Q(u0) + Q(u1)]h_{T,u}\left(\frac{Q(u1)}{Q(u0)+Q(u1)}\right).$$

Its derivative with respect to the probability of a descendant leaf in child b is

$$h_{T,u}(q_u) + (b - q_u)h'_{T,u}(q_u).$$

It has derivative zero for leaves outside the subtree. Summing over nodes gives the right side of Equation (5). By Appendix A, this derivative is the unique supporting score at Q, so it equals the actual $S_T(Q, x)$.

E Proof of the tangent comparison

Let

$$\ell_p(0) = h(p) - ph'(p), \qquad \ell_p(1) = h(p) + (1-p)h'(p).$$

The tangent inequality for a concave function gives

$$0 = h(0) \le h(p) - ph'(p), \qquad 0 = h(1) \le h(p) + (1-p)h'(p).$$

Therefore both losses are nonnegative, and

$$\begin{aligned}(1-q)\ell_p(0) + q\ell_p(1) &\le \max\left\{\frac{1-q}{1-p}, \frac{q}{p}\right\}[(1-p)\ell_p(0) + p\ell_p(1)] \\ &= \max\left\{\frac{1-q}{1-p}, \frac{q}{p}\right\} h(p).\end{aligned} \tag{19}$$

If $p \le q$, write $q = \lambda p + (1-\lambda)1$, where $\lambda = (1-q)/(1-p)$. Concavity gives

$$h(q) \ge \frac{1-q}{1-p}h(p).$$

If $p \ge q$, write $q = \lambda p + (1-\lambda)0$, where $\lambda = q/p$, to obtain $h(q) \ge (q/p)h(p)$. In both cases,

$$h(p) \le \max\left\{\frac{1-p}{1-q}, \frac{p}{q}\right\} h(q). \tag{20}$$

Equations (19) and (20) prove Lemma 5.

F A TWO-STEP ILLUSTRATION

For $T = 2$, write a report tree as (a, b_0, b_1), where a is the first forecast and b_i is the second forecast after observing $X_1 = i$. A realized-path score has the form

$$S_{00}(a, b_0), \quad S_{01}(a, b_0), \quad S_{10}(a, b_1), \quad S_{11}(a, b_1).$$

Theorem 3 says that its entropy must be

$$H(a, c_0, c_1) = h_{\varnothing}(a) + (1 - a)h_0(c_0) + ah_1(c_1).$$

At a differentiability point, the four losses are

$$S_{0j} = h_{\varnothing}(a) - ah'_{\varnothing}(a) + h_0(b_0) + (j - b_0)h'_0(b_0),$$
$$S_{1j} = h_{\varnothing}(a) + (1 - a)h'_{\varnothing}(a) + h_1(b_1) + (j - b_1)h'_1(b_1).$$

The score may use a different binary entropy after each first-round history. It cannot include an entropy interaction between b_0 and b_1, because the realized score sees only one of them. This small case displays the general causal factorization.

G WHY A PRODUCT-FAMILY SCORE IS NOT ENOUGH

If forecasts were restricted to one constant parameter p, the score

$$\frac{1}{T}\left(\sum_{t=1}^{T}(x_t - p)\right)^2 = T(\overline{x} - p)^2$$

would be proper for the Bernoulli product family. Its truthful expected value is $\alpha(1 - \alpha)$, while a fixed wrong report has an additional $T(\alpha - p)^2$ cost. This does not contradict Theorem 6. For a general sequential distribution, the forecasts p_t can adapt to earlier residuals. Squaring the terminal sum creates cross terms through which a strategic forecaster can hedge past noise. The full-tree score in Equation (10) avoids this problem by using the ex ante expectation of the total count under the entire reported tree.