Likelihood versus CRPS: A New Perspective

Composition, chained elicitation, and the score under which density refinements add up

Peter Cotton

Working draft v0.1 · 2026

Abstract

The choice between the logarithmic score and the continuous ranked probability score (CRPS) is usually argued on four grounds: locality, propriety, robustness, and interpretability. This note adds a fifth, one the debate rarely weighs. When the elicited object is a reusable predictive *density* and forecasts are *chained*, each stage refining the residual left by the last, the logarithmic score is the natural default. The log-likelihood of a composed forecast is the base score plus the log-density of each residual refinement, so credit is additive and each stage's increment is itself the proper score of what that stage added. This is the prequential and logarithmic-market-scoring structure; CRPS shares no comparable per-stage identity. None of that is new mathematics: the decomposition is the probability integral transform and the prequential chain rule. The contribution is a framing. It makes composability the organizing axis of the log-versus-CRPS choice, which is conventionally settled on locality, propriety, robustness, and interpretability, and ties it to chained elicitation, the setting a prediction supply chain lives in. We give the compositional accounting, a population-level worked example in which the two scores prefer measurably different forecasts, and an honest account of where CRPS is the right target, its tail-insensitivity there often a virtue. Conformal prediction appears as the limiting case: a method that need not produce a density at all, scored on the one metric that does not notice.

1 The debate and the axis it usually turns on

A scoring rule is a settlement rule. It fixes what a forecaster is paid to produce, and forecast evaluation is coherent only when the score is matched to the object being forecast (Gneiting 2011; Gneiting and Raftery 2007). For a full predictive distribution the two standard proper scores are the logarithmic score (Good 1952), which reads the density at the realized outcome, and CRPS (Matheson and Winkler 1976), which integrates squared distribution-function error across the line. Both are proper, so in expectation each is optimized by the truth; the argument between them is about which imperfect forecasts they reward.

That argument is usually conducted on four axes. *Locality*: the log score depends only on the density assigned to what happened, and among scores with that property it is essentially unique up to affine transformation (Bernardo 1979), though broader derivative-local scores exist once the restriction is relaxed (Parry et al. 2012). *Propriety*: both qualify, so this settles nothing between them. *Robustness*: the log score is unbounded below and CRPS is not, which is read for and against depending on whether an extreme outcome signals a fixable model or an unmodelable tail (Machete 2013; Bolin and Wallin 2023). *Interpretability*: CRPS lives in the units of the

outcome and the log score in nats. Reviews weigh these and reasonably conclude that the right score depends on the use (Gneiting and Raftery 2007; Dawid and Musio 2014).

This note presses a fifth consideration, absent from that list. It concerns not a single forecast but a *chain* of them, and it is decisive precisely when distributional prediction is organized as a supply chain (Cotton 2019, 2022).

2 Composition: the score under which refinements add up

Take a base forecast with density p_1 and distribution function F_1. A second stage forecasts the residualized variable

$$z = \Phi^{-1}(F_1(y)),$$

the base forecast's own probability integral transform pushed to the Gaussian scale (Rosenblatt 1952; Diebold et al. 1998). If the second stage assigns density g to z, the change of variables $dz/dy = p_1(y)/\varphi(z)$ gives the composed density $p(y) = p_1(y)\, g(z)/\varphi(z)$, so

$$\log p(y) \;=\; \log p_1(y) \;+\; \big(\log g(z) - \log \varphi(z)\big).$$

The log score of the composed forecast is the base score plus the incremental score of the refinement. A stage that reports the crowd's own residual law, $g = \varphi$, adds nothing and is paid nothing; a stage that finds structure the base missed is paid exactly its contribution. Credit is additive, and skill is attributable stage by stage. This is the accounting a prediction supply chain needs, and it is developed for chained pools and stacked lotteries in (Cotton 2026c, 2020) under the algebra of (Cotton 2026a). The dependence case factors the same way: by Sklar's theorem (Sklar 1959) a joint log-density splits into marginal terms plus a copula term, one stage each.

What makes this matter for the choice of score is not that the log-likelihood telescopes, which any additive bookkeeping of any score's differences would. It is that each increment $\log g(z) - \log \varphi(z)$ is itself the logarithmic score of the stage's residual refinement, strictly proper for the law of z: the stage is paid exactly, and only, for the density it added. The market form is Hanson's logarithmic market scoring rule, in which a trader who moves the report from p to p' is paid $S(p', y) - S(p, y)$, a chain of them telescopes to $S(p_{\text{final}}, y) - S(p_0, y)$, and the logarithmic rule is singled out because it alone yields a path-independent, additive cost function (Hanson 2007). Composability under chaining is, in that setting, a defining property of the log score, not a new observation.

Three further constructions are the same identity. A residual chain is gradient boosting under log loss: each stage multiplies the running density by a likelihood ratio fitted to what the chain so far gets wrong, the greedy stagewise step of boosting (Mason et al. 1999; Friedman 2001) applied to density estimation with the log-likelihood as the loss (Rosset and Segal 2002), with participants in place of weak learners and staked wealth in place of the learning rate. Sequential prediction scored by cumulative log loss is the prequential view (Dawid 1984), under which a probability model and a sequential code are the same object, the connection minimum description length makes exact (Grünwald and Roos 2019). And an autoregressive or normalizing-flow model is already such a chain, its log-likelihood a sum of a base log-density and per-stage change-of-variables terms (Papamakarios et al. 2021).

CRPS has no matching per-stage identity. Its values along a chain telescope, as any score's differences do, but the increment is not the CRPS of the stage's residual refinement, and there is no path-independent cost of which the chain is the sum. This is distinct from the reliability–resolution decomposition of a single CRPS value (Hersbach 2000), which splits one score into

calibration components rather than a chain into per-stage credit. The point is not that CRPS is improper or uninformative, only that it is not the accounting system a chain runs on.

3 Two geometries of error

The two scores measure different things, and the cleanest way to see it is through their regret against the truth p under a forecast q:

$$\text{log-score regret} = D_{\mathrm{KL}}(p \,\|\, q), \qquad \text{CRPS regret} = \int_{-\infty}^{\infty} \big(F_q(t) - F_p(t)\big)^2 \, dt.$$

Both are Bregman divergences of the respective proper score (Gneiting and Raftery 2007). The log score penalizes relative density error where the truth puts its mass; CRPS penalizes integrated distribution-function error across thresholds. The first is the divergence a likelihood ratio, a Bayesian update, or a growth-optimal bet consumes; the value a conditioning forecaster extracts over a crowd that ignores a covariate is exactly the mutual information $I(R; X)$ (Cotton 2026b). The second is gentler in the tails: a distant outcome costs CRPS on the order of its distance, but costs the log score quadratically, or without bound where the density vanishes.

There is a market reading of the first. In a complete, arbitrage-free market normalized Arrow–Debreu state prices define a risk-neutral state-price density; in an incomplete market there may be many, and none need equal the physical law (Cochrane 2005). Either way prices behave like a density-ratio settlement, not an arbitrary point forecast, and the growth-optimal investor stakes her physical density to maximize expected log wealth (Kelly 1956). A logarithmic market scoring rule is the mechanism that pays exactly this (Hanson 2007). CRPS has no such reading.

The gentleness of CRPS in the tails is not abstract. Let the truth be Student-t with three degrees of freedom, heavy-tailed, and fit a single Gaussian $N(0, \sigma^2)$ to it under each score. Likelihood insists on covering the tail: its optimum is variance-matching, $\sigma = \sqrt{3} \approx 1.73$. CRPS prefers a sharper central fit and shrugs off the occasional outlier: its optimum is $\sigma \approx 1.25$, a 28% narrower forecast with the tail squished in.

The narrower forecast wins CRPS and loses log-likelihood:

Gaussian fit to t_3	mean CRPS	mean NLL
$\sigma = 1.73$, likelihood-optimal	0.848	1.968
$\sigma = 1.25$, CRPS-optimal	0.829	2.102

Two honest qualifications. This does not beat a correct model: CRPS is proper, so at the population level the true t_3 scores best on both (CRPS 0.827, NLL 1.773), and the trade-off lives only among imperfect forecasts. And it runs the other way when the tail is genuinely unmodelable, where the narrower forecast is the sensible one and CRPS is right to prefer it; that same tail-insensitivity is what the robustness literature counts in CRPS's favor (Bolin and Wallin 2023). The point is only that the two proper scores rank the same imperfect forecasts differently, and that the log score is the one whose ranking a chain can add up.

Conformal prediction is this taken to its limit. A standard split-conformal predictor is not a density forecast at all; it targets finite-sample marginal coverage (Lei et al. 2018). Coerce its calibration residuals into an unsmoothed empirical predictive law and it assigns zero density

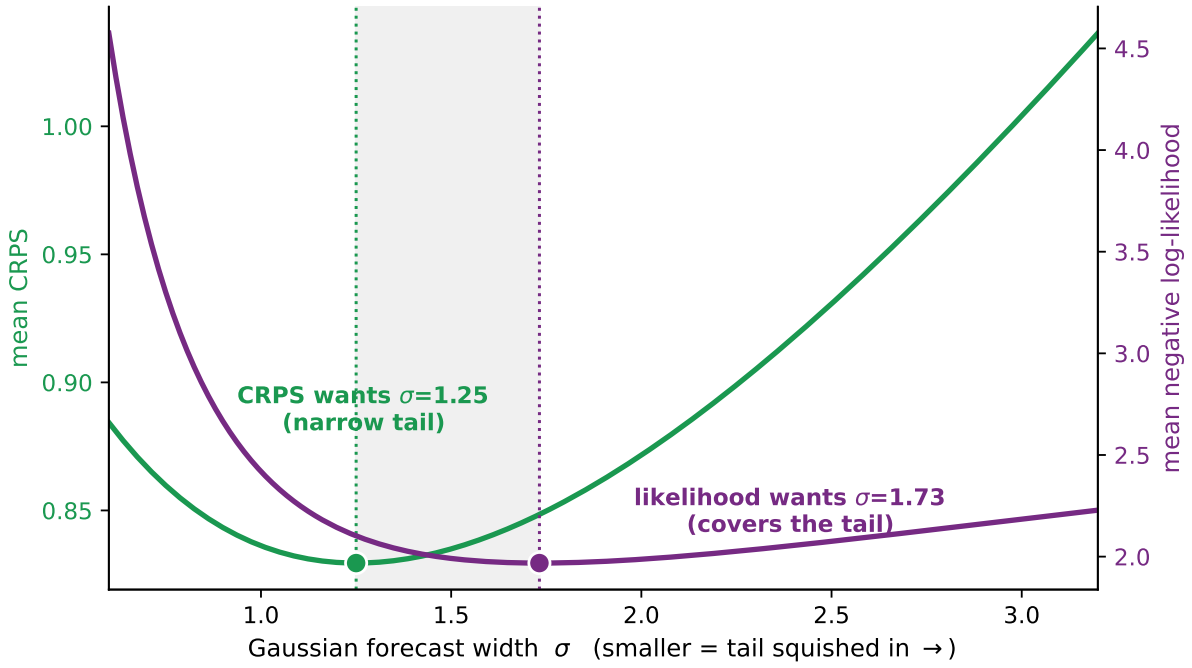

Figure 1: Fitting a Gaussian to Student-t_3 truth. Mean CRPS is minimized by a narrow forecast ($\sigma \approx 1.25$) that squeezes the tail in; mean negative log-likelihood is minimized by the wider, tail-covering forecast ($\sigma = \sqrt{3} \approx 1.73$). The two proper scores prefer measurably different approximations. Population values by numerical integration.

outside the observed range, hence $-\infty$ log-likelihood on any outcome beyond it, while its CRPS stays finite and competitive. A method that need not produce a density, yet scores well on CRPS, is the metric's tail-insensitivity turned into a feature of the approach, and the conditional information $I(R; X)$ it forfeits is exactly what a log-scored competitor collects against it (Cotton 2026b).

4 Where CRPS is the right target

None of this makes CRPS a bad score. It answers a different question, and it is well matched to broad distribution-function accuracy, threshold and quantile behavior, robustness under tails one cannot model, and interpretability in the outcome's own units. When the decision is known and coarse, a threshold, a quantile, an interval, an inventory limit, a smooth payoff integral, the decision-relevant functional should be scored directly, with a threshold- or quantile-weighted rule aimed at the region that matters (Gneiting and Ranjan 2011; Gneiting 2011). CRPS and its relatives are the right tools there, and as a secondary diagnostic CRPS is a useful check that a log-score winner has not won by pathological sharpness (Gneiting et al. 2007). The claim of this note is narrower and, we think, robust: when the elicited object is a reusable density and the future use is unknown, and above all when forecasts are to be chained, the log score is the default because it is the score under which refinements compose.

5 Related work and what is and is not new

Each ingredient is old. The residual identity is the probability integral transform (Rosenblatt 1952) and Sklar's theorem (Sklar 1959); probability-integral-transform diagnostics are standard for density forecasts (Diebold et al. 1998; Gneiting et al. 2007), and the same change-of-variables log-additivity is what autoregressive and normalizing-flow models exploit (Papamakarios et al. 2021). That the log score is additive under sequential prediction is the prequential principle (Dawid 1984) and the compression view of minimum description length (Grünwald and Roos 2019). That it composes under chaining is, in the mechanism-design setting, the defining rationale of Hanson's logarithmic market scoring rule (Hanson 2007), and log-score optimization is likewise the natural objective when predictive densities are stacked (Yao et al. 2018). Residual refinement under log loss is gradient boosting (Mason et al. 1999; Friedman 2001), applied to density estimation in (Rosset and Segal 2002). The conventional log-versus-CRPS axes, locality, propriety, sensitivity, robustness, interpretability, are set out in (Gneiting and Raftery 2007; Machete 2013; Bolin and Wallin 2023), with locality's role for the log score going back to (Bernardo 1979). The convex-duality background is consolidated for these mechanisms in (Cotton 2026a).

What is new is not a theorem but a framing, and a modest one. Each fact above lives in a separate literature, market scoring, prequential analysis, boosting, flows, and the log-versus-CRPS *evaluation* debate has drawn on almost none of them. It is conducted as if the object were a single forecast graded once. Once the object is a *reusable density* and the process is a *chain of elicitations*, a consideration those reviews do not weigh becomes the deciding one: only the log score's stage increment is itself the proper score of what the stage added. That is the natural reading of a distributional prediction supply chain (Cotton 2019, 2022, 2026c), and it recasts conformal prediction, often defended on CRPS and coverage, as the case where the tail-insensitivity of those metrics is doing the defending.

One question the literature appears not to have asked would sharpen any reading of it. Is an author's choice of evaluation metric correlated with the comparative advantage of the method being proposed? If methods tend to be reported on the score they happen to win, the field's stated preferences between log-likelihood and CRPS would partly reflect selection rather than merit. We are not aware of a study measuring this correlation, and it seems worth one.

6 Summary, for now

The right score matches the target. When the target is a threshold, a quantile, an interval, or an action, score it directly, and CRPS or a weighted relative may be the better tool. When the target is a reusable density under unknown later use, and when densities are to be composed, held-out log-likelihood is the default: it is local, its regret is a divergence one can read as information, it is the settlement rule of a density-ratio market, and, the consideration this note adds, it is the score under which refinements add up. CRPS remains a useful secondary diagnostic and a legitimate primary target for coarser questions. It is not wrong. It is a different contract, and it is not the one a chain signs.

Bernardo, José M. 1979. "Expected Information as Expected Utility." *The Annals of Statistics* 7 (3): 686–90. https://doi.org/10.1214/aos/1176344689.

Bolin, David, and Jonas Wallin. 2023. "Local Scale Invariance and Robustness of Proper Scoring Rules." *Statistical Science* 38 (1): 140–59. https://doi.org/10.1214/22-STS864.

Cochrane, John H. 2005. *Asset Pricing*. Revised. Princeton University Press.

Cotton, Peter. 2019. *Self-Organizing Supply Chains for Micro-Prediction: Present and Future Uses of the ROAR Protocol*. https://arxiv.org/abs/1907.07514.

Cotton, Peter. 2020. *The Lottery Paradox: A New Use*. https://www.slideshare.net/slideshow/lottery-paradox-csaildec2020/242173597.

Cotton, Peter. 2022. *Microprediction: Building an Open AI Network*. MIT Press.

Cotton, Peter. 2026a. *An Algebra of Prediction-Rewarding Mechanisms*. https://mechanisms.microprediction.org/papers/composition-and-the-algebra-of-mechanisms.html.

Cotton, Peter. 2026b. *Betting Against a Conformal Predictor: A Parimutuel Account of the Information Gap*. https://conformalprediction.net/papers/parimutuel/.

Cotton, Peter. 2026c. *Multi-Stage Solicitation of Probability Distributions: Experiments, Theory and Perspective on Conformal Prediction*. https://mechanisms.microprediction.org/papers/multi-stage-solicitation.html.

Dawid, A. P. 1984. "Present Position and Potential Developments: Some Personal Views. Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A (General)* 147 (2): 278–92. https://doi.org/10.2307/2981683.

Dawid, A. Philip, and Monica Musio. 2014. "Theory and Applications of Proper Scoring Rules." *METRON* 72 (2): 169–83. https://doi.org/10.1007/s40300-014-0039-y.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39 (4): 863–83. https://doi.org/10.2307/2527342.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–232. https://doi.org/10.1214/aos/1013203451.

Gneiting, Tilmann. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106 (494): 746–62. https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69 (2): 243–68. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78. https://doi.org/10.1198/016214506000001437.

Gneiting, Tilmann, and Roopesh Ranjan. 2011. "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules." *Journal of Business & Economic Statistics* 29 (3): 411–22. https://doi.org/10.1198/jbes.2010.08110.

Good, I. J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society. Series B (Methodological)* 14 (1): 107–14. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x.

Grünwald, Peter, and Teemu Roos. 2019. "Minimum Description Length Revisited." *International Journal of Mathematics for Industry* 11 (1): 1930001. https://doi.org/10.1142/S2661335219300018.

Hanson, Robin. 2007. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation." *The Journal of Prediction Markets* 1 (1): 3–15. https://doi.org/10.5750/jpm.v1i1.417.

Hersbach, Hans. 2000. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems." *Weather and Forecasting* 15 (5): 559–70. https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2.

Kelly, Jr., J. L. 1956. "A New Interpretation of Information Rate." *The Bell System Technical Journal* 35 (4): 917–26. https://doi.org/10.1002/j.1538-7305.1956.tb03809.x.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113 (523): 1094–111. https://doi.org/10.1080/01621459.2017.1307116.

Machete, Reason L. 2013. "Contrasting Probabilistic Scoring Rules." *Journal of Statistical Planning and Inference* 143 (10): 1781–90. https://doi.org/10.1016/j.jspi.2013.05.012.

Mason, Llew, Jonathan Baxter, Peter L. Bartlett, and Marcus Frean. 1999. "Boosting Algorithms as Gradient Descent." *Advances in Neural Information Processing Systems (NeurIPS)*, 512–18.

Matheson, James E., and Robert L. Winkler. 1976. "Scoring Rules for Continuous Probability Distributions." *Management Science* 22 (10): 1087–96. https://doi.org/10.1287/mnsc.22.10.1087.

Papamakarios, George, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. "Normalizing Flows for Probabilistic Modeling and Inference." *Journal of Machine Learning Research* 22. https://jmlr.org/papers/v22/19-1028.html.

Parry, Matthew, A. Philip Dawid, and Steffen Lauritzen. 2012. "Proper Local Scoring Rules." *The Annals of Statistics* 40 (1): 561–92. https://doi.org/10.1214/12-AOS971.

Rosenblatt, Murray. 1952. "Remarks on a Multivariate Transformation." *The Annals of Mathematical Statistics* 23 (3): 470–72. https://doi.org/10.1214/aoms/1177729394.

Rosset, Saharon, and Eran Segal. 2002. "Boosting Density Estimation." *Advances in Neural Information Processing Systems 15 (NeurIPS)*. https://proceedings.neurips.cc/paper_files/paper/2002/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Sklar, Abe. 1959. "Fonctions de répartition à n Dimensions Et Leurs Marges." *Publications de l'Institut de Statistique de l'Université de Paris* 8: 229–31.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis* 13 (3): 917–1007. https://doi.org/10.1214/17-BA1091.