On Sharpness Diagrams

Alexander I. Jordan
Heidelberg Institute for
Theoretical Studies

Abstract

The sharpness principle states that the goal in probabilistic forecasting is to "maximize the sharpness of the predictive distributions subject to calibration". For that statement to be consistent with the principle of the best-informed forecast, the goal should be refined to "minimize the expected entropy subject to autocalibration". Not all popular sharpness measures follow this principle – the average length of the predictive interquartile range being a prime example. Valid alternatives include the expected interval score entropy and the expected tail probability with respect to the modal interval.

1 INTRODUCTION

Fundamental to the evaluation of probabilistic forecasts is the idea that we have a joint distribution of multiple competing probabilistic forecasts and an observation. For the evaluation of, say, two competing forecasts, we take a tuple (F_1, F_2, Y) with joint distribution $\mathbb{Q}$ of two random variables F_1 and F_2 that map to cumulative distribution functions (CDFs), and a real-valued random variable Y. Then we determine which of F_1 and F_2 better describes the outcome Y. In the sharpness principle (Murphy and Winkler, 1987; Gneiting et al., 2007), calibration refers to the statistical consistency between a forecast and the outcome – we should not be able to statistically distinguish between a set of observed values and a sample drawn from the corresponding probabilistic forecasts. Sharpness is a property of the forecast alone – the more concentrated a probabilistic forecast is, the sharper it is. Only calibrated forecasts can be meaningfully compared in terms of sharpness.

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

Among the many different modes of calibration, the most important for us are probabilistic calibration due to its popularity and autocalibration for its technical properties. A continuous forecast F is probabilistically calibrated when the probability integral transform (PIT) $F(Y)$ follows a uniform distribution on $[0, 1]$ – checked, in practice, either via a PIT histogram or the empirical distribution function of the PIT values, depending on the discipline or application. The advantage of this mode of calibration is that it is particularly easy to assess when F is given as a CDF. The reliance on the unconditional distribution of $F(Y)$ is the main drawback, as biases or other types of miscalibration of the conditional PITs may cancel out. In contrast, one of the strongest modes of calibration is autocalibration, where the predictive distribution F equals the conditional law of Y given F, formally $F = \mathcal{L}(Y|F)$ almost surely. That is, an autocalibrated forecast is absolutely reliable in all scenarios, whether predicting typical conditions or a higher probability of an extreme outcome. Unfortunately, that property is notoriously difficult to check in practice, meaning that autocalibration can be disproven but hardly confirmed. Since autocalibration implies most other modes of calibration, a violation of any of these weaker notions also indicates a violation of autocalibration – an exception is the case of a binary outcome, where many modes of calibration, including probabilistic calibration, coincide with autocalibration. For a detailed discussion of hierarchies of notions of calibration see Gneiting and Resin (2023).

Gneiting et al. (2007) originally kept the calibration requirement for the sharpness principle intentionally vague. An attempted proof of the conjecture relying on probabilistic calibration did not prove successful (Pal, 2009, 2010). Shortly after, Tsyplakov (2013) argued that for the sharpness principle to hold the required mode of calibration is autocalibration, confirmed by Holzmann and Eulert (2014).

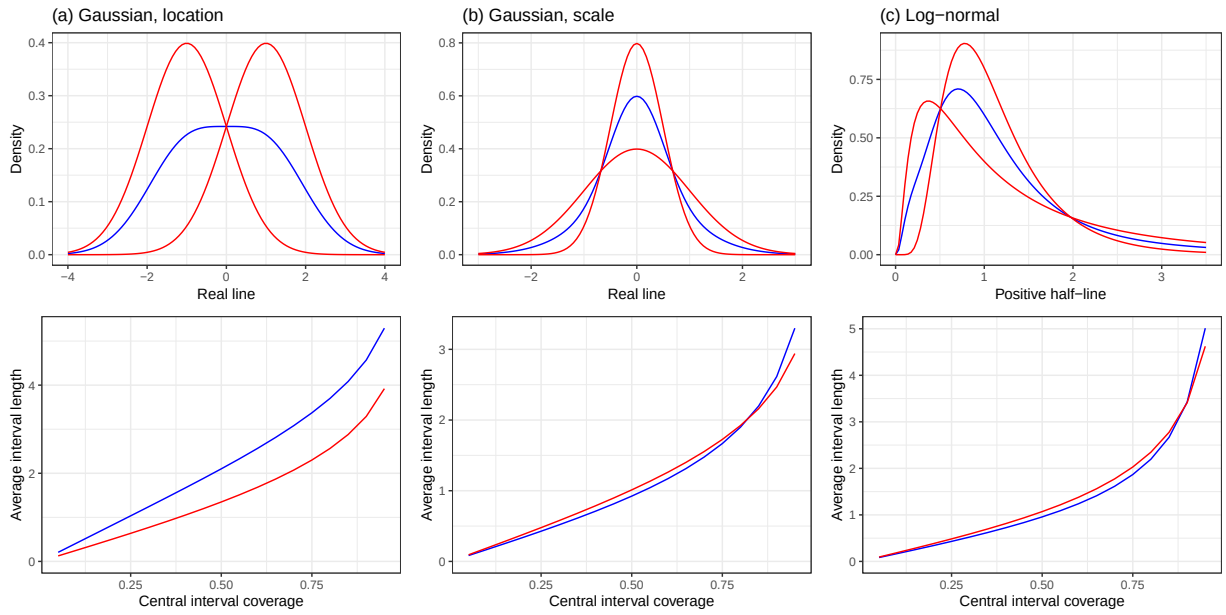

Figure 1: Density plots (upper row) and sharpness diagrams (lower row) for a translational Gaussian example (left column), a heteroskedastic Gaussian example (middle column), and a corresponding example based on log-normal distributions (right column). The two red lines in the respective density plot indicate the autocalibrated conditional distributions (occuring with equal probability) and the blue line indicates the corresponding autocalibrated unconditional distribution. The sharpness diagrams show the average length of the $(1-\alpha)\times 100\%$ central prediction interval for given nominal levels $\alpha \in (0,1)$.

Turning to sharpness, a common diagnostic tool for evaluation is the sharpness diagram (Gneiting et al., 2007; Pinson et al., 2007; Lauret et al., 2019; Yang and Kleissl, 2024), which shows some empirical summary of the lengths of central prediction intervals at multiple nominal levels for each forecasting method. This summary can be the average length or entire box plots. Naturally, a sharp forecast has a narrow central prediction interval containing the requested probability mass. For an illustration, consider a data-generating process where the outcome Y is drawn from one of two distinct conditional distributions. Then, take two autocalibrated forecasts F_1 and F_2, where F_1 has access to a covariate that determines from which conditional distribution the outcome is drawn, and F_2 is simply the unconditional distribution. According to the principle of the best-informed forecast (Holzmann and Eulert, 2014), F_1 is preferable to F_2 as both forecasts are autocalibrated and F_1 contains more information. Specifically, consider the following examples based on Gaussian distributions:

(a) Let $Y \,|\, \mu \sim \mathcal{N}(\mu, 1)$, where μ takes a value of $+1$ or -1 with equal probability. Then forecast $F_1(x) = \Phi(x-\mu)$ is the conditional CDF of Y, whereas $F_2(x) = \frac{1}{2}\left(\Phi(x-1)+\Phi(x+1)\right)$ is the unconditional CDF of Y.

(b) Let $Y \,|\, \sigma \sim \mathcal{N}(0, \sigma^2)$, where σ takes a value of 1.0 or 0.5 with equal probability. Then forecast $F_1(x) = \Phi(x/\sigma)$ is the conditional CDF of Y, whereas $F_2(x) = \frac{1}{2}\left(\Phi(x)+\Phi(x/0.5)\right)$ is the unconditional CDF of Y.

(c) Let $Z \,|\, \sigma \sim \mathcal{N}(0, \sigma^2)$ and $Y = \exp(Z)$, where σ takes a value of 1 or 0.5 with equal probability. Then, on the positive half-line, forecast $F_1(x) = \Phi(\log(x)/\sigma)$ is the conditional CDF of Y, whereas $F_2(x) = \frac{1}{2}\left(\Phi(\log(x))+\Phi(\log(x)/0.5)\right)$ is the unconditional CDF of Y.

Figure 1 illustrates these three setups in density plots and shows the resulting sharpness diagrams based on the expected length of the central $(1-\alpha)\times 100\%$ prediction intervals, $\alpha \in (0,1)$. Immediately, we observe an issue in this traditional sharpness diagram for the toy examples (b) and (c). The uninformed forecast F_2 is designated as sharper than the informative forecast F_1 for all low and moderate coverage levels, including when measuring sharpness by the expected length of the interquartile range ($\alpha = 0.50$). Technically, by counterexample, we have shown that the map from a distribution to the length of its central prediction interval is not concave. In section 4, we instead propose sharpness diagrams based on entropies for the central prediction interval and the modal prediction interval.

2 SHARPNESS AND FORECAST UNCERTAINTY

According to the sharpness principle (Gneiting et al., 2007), the sharpness of a forecast should not depend on the outcome but solely on the forecast. By definition, an autocalibrated forecast can only be improved by incorporating information beyond the forecast itself.

Definition 1 (Autocalibrated probabilistic forecast)**.** *Let $\mathcal{F}$ be a convex class of CDFs and let (F, Y) be a random tuple taking values in $\mathcal{F} \times \mathbb{R}$. We say F is a probabilistic forecast for Y and we call it autocalibrated if*

$$F = \mathcal{L}(Y \mid F) \quad \textit{almost surely.}$$

Note that the convexity of $\mathcal{F}$ ensures that the uninformative version of F also lies in that class, that is, $\mathbb{E}[F] =: F^0 \in \mathcal{F}$, where the expectation of $F(\cdot)$ is understood to be taken pointwise in $x \in \mathbb{R}$. Importantly, F^0 is also autocalibrated, as is any other less informative version $F' \in \mathcal{F}$ of F that satisfies $F' = \mathbb{E}[F \mid F']$ almost surely.

In order to compare these autocalibrated forecasts, any measure of forecast uncertainty needs to be defined on the convex class of distributions $\mathcal{F}$ and map to real number, say, a measure of (negatively-oriented) forecast uncertainty is a function $H \colon \mathcal{F} \to \mathbb{R}$. Now, we have a decision to make: With F being a random probabilistic forecast taking values in $\mathcal{F}$, its uncertainty is a random variable $H(F)$ taking values in $\mathbb{R}$. How should $H(F)$ be summarized for a comparison with the forecast uncertainty of the expected CDF, $H(F^0)$?

Arguably, the sharpness principle needs to be consistent with the evaluation of forecast performance via proper scoring rules – quoting the informative probabilistic forecast F should be better, in expectation, than quoting its unconditional version F^0. Therefore, the essential requirement is that Jensen's inequality holds,

$$\mathbb{E}[H(F)] \leq H(F^0) = H(\mathbb{E}[F]).$$

In summary, a measure of (negatively-oriented) forecast uncertainty needs to be a concave function.

Definition 2 (Sharpness)**.** *Let F_1 and F_2 be autocalibrated probabilistic forecasts for Y and let $H \colon \mathcal{F} \to \mathbb{R}$ be a concave function.*

(a) The (positively-oriented) sharpness of F_1 (relative to H) is measured by $H(\mathbb{E}[F_1]) - \mathbb{E}[H(F_1)]$.

(b) We say F_1 is at least as sharp as F_2 (relative to H) if $\mathbb{E}[H(F_1)] \leq \mathbb{E}[H(F_2)]$.

3 ENTROPIES ARE CONCAVE FUNCTIONS

It has long been suggested to use entropies as measures of sharpness (DeGroot, 1962; Angulo and Ruiz-Medina, 2008; Bröcker, 2009; Tsyplakov, 2013). An entropy is a concave function that is the result of minimizing the expectation of a loss function (DeGroot, 1962; Grünwald and Dawid, 2004). In particular, we consider proper scoring rules which are loss functions that evaluate a probabilistic forecast against an observation and that encourage honest predictions. A proper scoring rule $S \colon \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ is minimized in expectation by a probabilistic forecast $G \in \mathcal{F}$ when the observable Y taking values in $\mathbb{R}$ can be understood as drawn from G, that is,

$$\mathbb{E}_{Y \sim G}[S(G, Y)] \leq \mathbb{E}_{Y \sim G}[S(F, Y)]$$

for all $F \in \mathcal{F}$, where $\mathcal{F}$ is a class of distributions relative to the given scoring rule S. The left-hand side of the inequality is the entropy which we will write as the function

$$H \colon \mathcal{F} \to \mathbb{R}, \quad G \mapsto \mathbb{E}_{Y \sim G}[S(G, Y)].$$

Gneiting and Raftery (2007, Theorem 1) showed a one-to-one correspondence between proper scoring rules and supertangents of concave functions.

A powerful and general principle is the Bayes act construction studied by Grünwald and Dawid (2004) and quoted here in the following form:

Theorem 1. *Let A be a general action domain and let $L \colon A \times \mathbb{R} \to \mathbb{R}$ be a loss function that has a non-empty set of expectation minimizers (or Bayes acts)*

$$A_G = \arg\min_{a \in A} \mathbb{E}_{Y \sim G}[L(a, Y)]$$

for all $G \in \mathcal{F}$. Then the scoring rule $S \colon (F, y) \mapsto L(a_F, y)$, where $a_F \in A_F$, is proper relative to the class $\mathcal{F}$, and the corresponding concave function $H \colon \mathcal{F} \to \mathbb{R}$ is given by

$$H(G) = \mathbb{E}_{Y \sim G}[L(a_G, Y)].$$

We now paraphrase a result by Holzmann and Eulert (2014) on the role of the information set as encoded by the generated σ-algebra:

Theorem 2. *Let F_1 and F_2 be autocalibrated probabilistic forecasts for Y and let $H \colon \mathcal{F} \to \mathbb{R}$ be the entropy of a proper scoring rule. If F_1 is better-informed than F_2, that is, $\sigma(F_1) \supseteq \sigma(F_2)$, then F_1 is at least as sharp as F_2 (relative to H).*

4 SHARPNESS DIAGRAMS

For an action domain $A \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, the set of Bayes acts may include acts which can be described in a simple closed form by a functional $T\colon \mathcal{F} \to \mathbb{R}^d$. Any associated loss function is also known as a scoring function that is consistent for the functional T (Gneiting, 2011).

Using a parameterized functional, such as the central $(1-\alpha)\times 100\%$ prediction interval where $\alpha \in (0,1)$, allows the resolution of the entire distribution. However, based on the average length of the central prediction interval, the traditional sharpness diagram can be misleading. We need to use the expected score of a consistent scoring function, where the natural choice for a loss function in that case is the interval score,

$$L_\alpha(l,u,y) = u - l + \frac{2}{\alpha}\left|y - y_{[l,u]}\right|,$$

defined on the action domain $A = \{(l,u) \in \mathbb{R}^2 : l \le u\}$ and where $y_{[l,u]} = \max\{l, \min\{u,y\}\}$. The corresponding interval score entropy is

$$H_\alpha(F) = \min_{(l,u)\in A}\left(u - l + \frac{2}{\alpha}\mathbb{E}_{Y\sim F}\left|Y - Y_{[l,u]}\right|\right),$$

where the Bayes act is the pair of two quantiles of F, at levels $\alpha/2$ and $1-\alpha/2$, respectively.

Another option is the modal prediction interval with length $\theta > 0$ with its corresponding zero-one loss

$$L_\theta(m,y) = \mathbb{1}\left(|y-m| \ge \frac{\theta}{2}\right),$$

defined on the action domain $A = \mathbb{R}$. The corresponding entropy is the tail probability,

$$H_\theta(F) = \min_{m\in\mathbb{R}} \mathbb{P}_{Y\sim F}\left(|Y-m| \ge \frac{\theta}{2}\right).$$

In Figure 2, we revisit the example from the introduction, using the (positively-oriented) sharpness $H(\mathbb{E}[F]) - \mathbb{E}[H(F)]$ based on the interval score entropy and the modal interval tail probability. The informative forecast now has a non-negative sharpness (lower average forecast uncertainty than the uninformative forecast, cf. Figure 1) for all parameter values. The interval score sharpness is less intuitive compared to the average length of the central prediction interval due to the additional term for tail expectations. While the tail probability sharpness has a great interpretability, the bottom left panel of Figure 2 shows a weakness of this sharpness measure in environments where the midpoint of the modal interval is constant across realizations of the probabilistic forecast – we cannot distinguish between informative and uninformative versions of a given forecast.

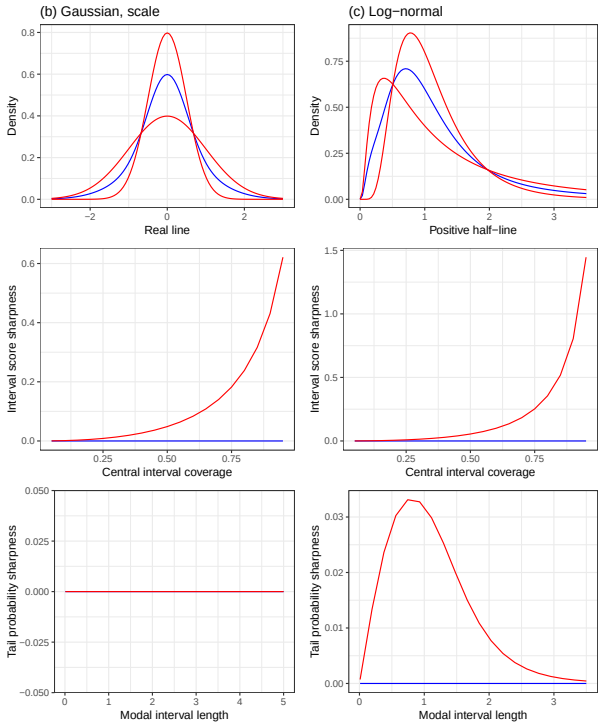

Figure 2: Density plots (upper row) and sharpness diagrams based the interval score entropy and the modal interval tail probability (in the middle and lower row, respectively) for examples (b) and (c) from the introduction (in the left and right column, respectively). The red lines indicate the autocalibrated conditional distributions (occuring with equal probability) and the blue lines correspond to the autocalibrated unconditional distributions.

5 DISCUSSION

As shown by Tsyplakov (2013), the sharpness principle holds when we assume autocalibrated forecasts and measure sharpness via entropies or expectations of proper scoring rules. As the length of the central prediction interval is not a concave function, it cannot be an entropy, and therefore it may lead to unexpected results despite its intuitive appeal.

We define sharpness as a positively-oriented summary measure of the entire distribution of a CDF-valued random quantity. In particular, sharpness of an informative forecast can be interpreted as the reduction in average forecast uncertainty relative to the forecast uncertainty of its uninformative version.

We propose sharpness diagrams based on two forecast uncertainties that are close in spirit to the length of the central prediction interval used in traditional sharpness diagrams - the interval score entropy and the modal interval tail probability.

Acknowledgements

The author thanks Tilmann Gneiting for insightful comments and discussion. The author is grateful for the generous support of the Klaus Tschira Foundation.

Code availability

Replication code is available at https://github.com/aijordan/sharpness_diagrams.

References

Angulo, J. M. and Ruiz-Medina, M. D. (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:236–237.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics*, 33:404–419.

Gneiting, T. (2011). Making and evaluating point forecast. *Journal of the American Statistical Association*, 106:746–762.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102:359–378.

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17:3226–3286.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32:1367–1433.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting – with applications to risk management. *Annals of Applied Statistics*, 8:595–621.

Lauret, P., David, M., and Pinson, P. (2019). Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254–271.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Pal, S. (2009). A note on a conjectured sharpness principle for probabilistic forecasting with calibration. *Biometrika*, 96:1019–1023.

Pal, S. (2010). Amendments and corrections: 'on a conjectured sharpness principle for probabilistic forecasting with calibration'. *Biometrika*, 97:1013–1013.

Pinson, P., Nielsen, H. A., Møller, J. K., and Madsen, H. (2007). Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy*, 10:497–516.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. https://mpra.ub.uni-muenchen.de/45186/, last accessed on 24 Feb 2026.

Yang, D. and Kleissl, J. (2024). *Solar Irradiance and Photovoltaic Power Forecasting*. Energy Analytics. Boca Raton: CRC Press.