Beat the House: Probably a Better Method for Rewarding Forecasters

Nuño Sempere*

April 15, 2022

Motivation

In Alignment Problems With Current Forecasting Platforms, Sempere and Lawsen outline a variety of problems with current forecasting platforms, whose scoring rules are found to either not be proper—as in the case of Good Judgment Open or CSET-Foretell (now INFER)—or incentivize distorting one's true probabilities to maximize the chances of placing in the top few positions which earn a monetary reward—as in the case of Metaculus. In addition, in almost all cases, forecasting platforms—or, for that matter, prediction markets—disincentivize collaboration.

In this working paper, we outline an incentivization method, "paying for bits", which combines features of prediction markets and forecasting platforms to produce a scoring rule that pays forecasters well, is proper, and nonetheless incentivizes collaboration. This scoring rule has both a discrete form—where resolution is either a yes or a no, and a continuous form—where resolution can be any probability between 0 and 1.

In essence, we start out with Hansons' logarithmic market scoring rule, and we add a few bells and whistles to make it collaborative. Although interacting with a prediction market run by an automatic market maker is known to be equivalent to interacting with a scoring rule, we prefer the former interpretation, which is a better abstraction because it deals more elegantly with forecasters not being able to input arbitrarily high or low probabilities.

In particular, the practical innovations we suggest are:

- Instead of comparing forecasters against other forecasters, compare them against against the initial probability of the stakeholder who sponsors the question

*Quantified Uncertainty Research Institute.

- Create a prediction market for each forecaster with a logarithmic market scoring rule, and allow each forecaster to move only the probability of their own prediction market.
- Require forecasters to put their money where their mouth is by betting their own money, so that if their forecast is worse that the houses' own forecast, they loose money. Optionally, sponsor forecasters once who start out without many fund themselves.

In two accompanying papers, *Amplified Oracle* and *Amplify a Bayesian*, we use the continuous form of this scoring rule like a lego-block: we combine it with other methods to build more powerful and general incentive schemes.

Some rough notes on limitations of previous approaches, and their solutions

Market scoring rules overcome the limitations of Brier and logarithmic scoring rules

The Brier score has limitations stemming from its deep theoretical inelegance. It is indeed proper, but it also doesn't have desirable properties such as:

- Composability: The Brier score on "A" and the Brier score on "B given A" can't be straightforwardly related to the Brier score on "A and B"
- Comparability: The cumulative relative Brier score doesn't correspond to any particular meaning.
- Comparability II: The Brier score of the average guess will always be at least as good as the average of the Brier scores, and usually better. This incentivizes forecasters to forecast the average crowd guess.

In contrast, the logarithmic scoring rule has an array of desirable properties.

- Composability: The logarithmic score on "A&B" is the sum of the log scores on "A" and "B given A"
- Comparability: The cumulative relative log score corresponds to bits of information which would be added or removed
- Comparability II: The arithmetic mean of the log scores is the log score of the geometric mean of the guesses. If forecasters are rewarded in proportion to their log score vs the crowd, there is no advantage to forecasting the current crowd.

More generally, for many nice theoretical properties, the log score will have them, whereas the Brier score will not. And yet, the Brier score offers bounded payoffs, whereas the logarithmic score does not. That is, when distributing a fixed pot of money as rewards to forecasters, the natural way to do this would be in proportion to their scores. But the logarithmic scoring rule is not bounded, which complicates payoff calculations.[1]

[1]Readers may think that normalization by the worst logarithmic score may solve this. But

This is why platforms such as Cultivate Labs still use the Brier score to reward their members. In contrast, platforms such as Metaculus solve this problem by forbidding their members from inputting probabilities lower than 1% or higher than 99%.

However, a logarithmic market scoring rule can approximate[2] the logarithmic score, while bounding payoffs by only allowing players to loose as much as they have previously accumulated. Conversely, the more money—or points in the case of a play-money prediction market—a participant has accumulated, the more extreme they are allowed to make the market. This preserves some of the desirable properties of the logarithmic scoring rule in approximate form, while making payoffs bounded without having to institute a sharp cut-off.

From a theoretical perspective and pointers to previous literature, Hanson's Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation is a dry if comprehensive introduction. From the perspective of a programmer seeking to implement a logarithmic market scoring rule, Gnosis' LMSR Primer is particularly readable.

To summarize, a logarithmic scoring rule is given by $s_i = a_i + b \cdot log(p_i)$, where p_i is the probability the player assigns to an event. We can rewrite this as $s_i = b \cdot log(p_i/c_i)$, so that it becomes clear that b determines how much the automatic market maker pays or receives per some amount of improvement of information over its prior, c_i. Intuitively, b should be higher the more the market sponsor cares about the event.

Then, from that scoring rule, we can create an equivalent market scoring rule. A market scoring rule is determined by a cost function, C, such that the cost of buying an additional n shares of outcome i is given by $C(q_i + n) - C(q_i)$, where q_i are the number of shares for that outcome already issued. In the case of the logarithmic market marker, the cost function is given by:

$$C(\vec{q}) = b \cdot log\left(\sum_i exp\left(\frac{q_i + a_i}{b}\right)\right)$$

where a_i is given by the original credences of the market creator[3]:

$$a_i = b \cdot log\left(\frac{1}{c_i}\right)$$

because the worst score can be arbitrarily low, preserving incentive compatibility is not trivial.

[2] In theory, it could be equivalent. In practice, this is less clear, because forecasters could, for instance, have an unboundedly low score on a forecasting platform such as Metaculus, but not be able to loose an unbounded amount of funds

[3] Most discussions of the LMSR omit the a_i parameter. This might be because the only ones who bothered to do the math were [Pennock](http://blog.oddhead.com/2006/10/30/implementing-hansons-market-maker/) and [Chenn and Pennock](https://arxiv.org/abs/1206.5252), and they likewise omit this parameter.

Then the marginal price is given by

$$p_i = \frac{\partial C}{\partial q_i} = \frac{exp\left(\frac{q_i + a_i}{b}\right)}{\sum_j exp\left(\frac{q_j + a_j}{b}\right)}$$

Note that from a platform design perspective, $\vec{a}$ can be abstracted away by setting $\vec{q} = \vec{a}$, i.e., by initializing a market maker with a number of shares already issued. Note also that players cannot loose more than the account in their balance, or, conversely, that the more they are willing to risk, the more extreme their probabilities can be. This seems more elegant than discrete cutoffs at $< 1\%$ or $> 99\%$, as in Metaculus.

Competing against the house provides an incentive for collaboration in a way which competing against other forecasters doesn't.

To determine the relative quality of forecasters, forecasting platforms currently score and compare forecasters in either of two ways:

- Against the aggregate of other forecasters. For instance, a forecaster's score for a question could be difference between his Brier score and the average Brier score of other forecasters who participated in the same question. And his total score could be the sum of those differences.
- In absolute terms, e.g., a forecaster's score could be their Brier score averaged across all questions he has participated in.

Both comparison methods do indeed produce a ranking. But they also introduce a clear zero sum component: forecasters are not incentivized to share information because that would hurt their relative positioning. The second method further introduces the distortion that forecasters are incentivized to choose easier questions, i.e., questions which are easier to predict on.

The alternative suggested in this paper is to reward forecasters monetarily in proportion to their improvement over a question's prior weighted by importance. By the question prior, I mean an initial guess provided by the question creator or sponsor. And by weighing by importance, we incentivize forecasters to predict on more important questions. So overall, we incentivize forecasters to predict on important questions for which the sponsoring stakeholder's initial guess is worse.

This creates a more meaningful metric for comparison: importance-adjusted bits of information added to the decision-maker's initial forecast. From Bayesian probability theory, we get the strong hint that probabilities without priors don't really make sense, so that might be a hint that we want comparisons vis a vis a prior.

There is still a residual incentive to compete with other forecasters to produce the most importance-adjusted bits of information. However, if monetary incentives

are large enough, the efficiencies through collaboration and some lower degree of latent altruism might outweigh those incentives. Conversely, relative positional results could be made anonymous, by rewarding forecasters using a privacy-preserving crypto-currency, such as Monero.

Making forecasters risk their own funds makes them have skin in the game

For a treatment of why this is beneficial, see Nassim Taleb's Incerto. But note that this step is optional. That is, forecasting platforms could reward points to forecasters in according to the market scoring rule specified above. As long as forecaster reward is eventually proportional to points acquired, this scoring rule becomes proper.

Description of the method.

In the interest of brevity, we shall outline our proposed method by means of an example, and the example shall be the question "Will the People's Republic of China have annexed at least half of Taiwan by 2050?", as operationalized by Metaculus. Parts which justify or explain the use of other parts are explained first, even if they would come later in terms of time.

The patron determines a rough prior

Taiwan has been independent of mainland China since the 25th of October 1945, i.e., 76 years into the past. Per Laplace's law, the chances that this will change by 2050 is $1 - (1 - \frac{1}{(2021-1945)+1})^{2050-2021} \approx 32\%$. Lets take this 32% as the market's initial probability. Note that per the reference class problem, other reference classes might have been chosen, so the point of this prior is not to be definitive, but rather to provide a starting point less arbitrary than 50% from which forecaster reward might be computed in the next steps.

By sharing this initial step, the question creator saves the time of the many forecasters which may participate in the question.

In the case of a patron aiming to learn from sponsoring a forecasting tournament, the prior might represent the patron's initial probability. If the patron is unsophisticated and the question is not amenable to base-rate analysis, we may use a 50%, the percent of questions which resolve positively on the forecasting site which hosts a given question, or some other such abstract reference class. Alternatively, the patron may sub-contract the creation of the prior, for instance by paying other forecasters to quickly do so or by creating low-liquidity (and thus low-potential reward) prediction markets.

If there is no stakeholder or proxy-stakeholder willing and able to provide a prior, this also provides a useful signal to forecasters that there might be no

stakeholders interested in the question, and thus that forecasts for the question may not be as valuable.

Forecasters predict on the question

Forecasters each[4] get access to a prediction market governed by a logarithmic market scoring rule, and whose initial probability is the prior determined in the previous step. This prediction market is also characterized by a further parameter: liquidity. Liquidity is the amount of money which is available to take bets on both sides. The more liquidity there is, the more money it takes to moves a market's odds, and the more money that can be made if those odds are wrong. For this reason, question creators would put more liquidity in markets they care more about. Above, liquidity corresponds to the parameter b.

If an individual forecaster thinks that the initial probability is wrong, he can move it to a probability which he thinks is more correct by betting some of his funds. Further, we might hope that the different forecasters might each research different aspects of the question and then reveal their information to each other, so that they can collectively correct their own markets more efficiently. Because each forecaster interacts only with his own market, we may hope that forecasters are in fact incentivized to do this.[5]

When contemplating this scheme, two issues become apparent after some amount of reflection. First, in a prediction market in which the outcome becomes more and more apparent as time goes on, the question creator loses all the funds he put up as liquidity. Secondly, forecasters have an incentive to predict as late as possible.

Within the automatic market-maker abstraction, both of these issues have a clear solution. First, close the market before the outcome is known. One particularly advisable point would be to cut trading when the stakeholder interested in the question makes the decision that the market was created to influence. Secondly, slowly reduce the amount of liquidity there is in the markets as time goes on, so that forecasters have a small incentive to predict sooner rather than later. This is easy to do by making the parameter $b = b(t)$, i.e., dependent on time, and slowly decreasing.

The question gets a binary resolution and forecasters get rewarded.

Taiwan either gets invaded or doesn't. Forecasters get rewarded in proportion to the number of shares of the winning outcome that they accumulated while interacting with their prediction market.

[4]Note that although we are speaking about individual forecasters, this scheme could also be applied to teams of forecasters. That is, we could give access to one prediction market to each team, and allow team members to distribute winnings or losses as they see fit.

[5]This may stop holding if there is a more liquid outside market in which all forecasters can participate. In that case, the situation is more complex to analyze.

The question gets a probabilistic resolution

Sometimes, the resolution of a question is uncertain. In cases of uncertainty about how a question should resolve, current forecasting platforms tend to resolve questions as ambiguous. Prediction markets tend to break to one side (as 0% or as 100%), but they also have the option of resolving questions probabilistically, so that shares pay out at X cts and $(100 - X)$ cts rather than \$0 or \$1.

Some common reasons for uncertain resolutions might be:

Because of uncertain ontologies

The Metaculus question as currently written resolves only if the People's Republic of China takes over at least half of Taiwan. However, this is somewhat arbitrary. If the question had been on a prediction market, it could also have been structured such that an invasion of n% of Taiwan pays out n cents for each share.

Because of probabilistic knowledge

At the time of question resolution, Taiwan could be enmeshed in a civil war, such that it's unclear how much territory each side controls. If the question had to be resolved, one way to do so would be to pay shares out in proportion to the likelihood that the PRC controls more than 50% of the territory.

Because the question hasn't resolved yet

The original question asked about an invasion of Taiwan by 2050, which is fairly far in the future. So forecasters might not be motivated to predict in question for which the payoff might be very far away. In the two accompanying papers, *Amplified Oracle* and *Amplify a Bayesian*, we outline two methods for providing forecasters with a reward now for questions whose resolution is far away. But that reward will be probabilistic.

That is, shares for the prediction market are paid out according to the probability estimated by a different forecasting system. If that forecasting system estimates an X% probability, shares of yes are paid out X cts.

Conclusion

We outlined several improvements to scoring rules as they are currently implemented in forecasting platforms:

- Instead of comparing forecasters against each other, compare them to the initial probability of the stakeholder who is interested in the question
- Assign a prediction market governed by an automatic market maker to each forecaster, rather than a scoring rule, because this better governs how extreme the probabilities which forecasters input can be.
- Make forecasters risk their own funds, optionally sponsoring them once.

These improvements solve the majority of the problems identified in Sempere and Lawsen's Alignment Problems With Current Forecasting Platforms. Still, the hard work of implementation is yet to be done. For instance, although we favor the logarithmic market scoring rule, platforms may find others market scoring rules more suitable in practice. In addition, our method's cost increases roughly linearly with the number of forecasters, so platforms should figure out how to invite only the good ones, or a scheme to reward later forecasters less.

Note also that the three improvements are modular, and their combination fairly opinionated. But there is nothing preventing platforms from comparing forecasters' Brier scores against the Brier scores of a stakeholder, from using a scoring rule or an automatic market-maker different from the logarithmic one[6], or from using play-money rather than real-money.

[6]For instance, practitioners such as Gnosis, Augur or [Manifold Markets](https://manifoldmarkets.substack.com/p/above-the-fold-market-mechanics?s=r) have historically prefered the constant product market maker due to ease of implementation