Incentive-Compatible Forecasting Competitions

Jens Witkowski
ETH Zurich
jensw@inf.ethz.ch

Rupert Freeman
Duke University
rupert@cs.duke.edu

Jennifer Wortman Vaughan
Microsoft Research
jenn@microsoft.com

David M. Pennock
Microsoft Research
dpennock@microsoft.com

Andreas Krause
ETH Zurich
krausea@ethz.ch

Abstract

We consider the design of forecasting competitions in which multiple forecasters make predictions about one or more independent events and compete for a single prize. We have two objectives: (1) to award the prize to the most accurate forecaster, and (2) to incentivize forecasters to report truthfully, so that forecasts are informative and forecasters need not spend any cognitive effort strategizing about reports. Proper scoring rules incentivize truthful reporting if all forecasters are paid according to their scores. However, incentives become distorted if only the best-scoring forecaster wins a prize, since forecasters can often increase their probability of having the highest score by reporting extreme beliefs. Even if forecasters do report truthfully, awarding the prize to the forecaster with highest score does not guarantee that high-accuracy forecasters are likely to win; in extreme cases, it can result in a perfect forecaster having zero probability of winning. In this paper, we introduce a truthful forecaster selection mechanism. We lower-bound the probability that our mechanism selects the most accurate forecaster, and give rates for how quickly this bound approaches 1 as the number of events grows. Our techniques can be generalized to the related problems of outputting a ranking over forecasters and hiring a forecaster with high accuracy on future events.

1 Introduction

The study of probabilistic forecasting dates back to the 1950s when meteorologists developed proper scoring rules as a way to both incentivize truthful predictions about future events and compare the relative accuracy of different forecasters (Brier 1950; Good 1952). Brier's original quadratic scoring rule is still widely used to motivate and measure forecasting accuracy (e. g., Atanasov et al. 2016). When forecasters are paid proportional to their quadratic scores, they maximize expected payment by truthfully reporting their beliefs.

However, in typical forecasting competitions, forecasters care not about maximizing expected score, but about whether their forecasts are judged to be better than others'. For example, in the Good Judgment Project, a recent geopolitical forecasting tournament, the top 2% of forecasters were awarded so-called "superforecaster" status (Tetlock and Gardner 2015), which (on top of bragging rights) gave them full travel reimbursement to a superforecaster conference. On ProbabilitySports,[1] participants predict the outcomes of NFL games, competing for prizes that are awarded to the highest-scoring forecaster in a given week or month. In play-money prediction markets, forecasters often compete for a place at the top of a leaderboard (e. g., Servan-Schreiber et al. 2004). And the same phenomenon holds for algorithmic forecasters; Netflix offered $1,000,000 to the team whose machine learning algorithm could best predict how users would rate movies based on their past preferences,[2] and the machine learning competitions run by Kaggle[3] rank submitted algorithms based on how well they predict the labels of data points from an undisclosed test set. One of Kaggle's main uses today is for recruiters to hire the developers of the best-performing algorithms (Harris 2013).

Unless they are designed with care, these winner-take-all competitions can distort incentives, encouraging forecasters to take big risks as opposed to truthfully reporting their beliefs. Lichtendahl and Winkler (2007) study a strategic game between two forecasters reporting on a single event. In their model, each forecaster wishes to maximize her utility, which is assumed to be a mixture of a proper scoring rule payment and an (explicit or implicit) bonus for being the best forecaster, with a parameter trading off these two components. They show that when forecasters optimize for their relative rank, they typically want to report more extreme probabilities than those corresponding to their true beliefs. Even putting truthfulness aside, we show that awarding a prize to the forecaster with highest score does not guarantee that high-accuracy forecasters are likely to win, and in fact can lead to situations in which a perfect forecaster has zero probability of winning.

In this paper, we present the **E**vent-**L**otteries **F**orecaster Selection Mechanism (ELF). ELF borrows a trick from the competitive scoring rule of Kilgour and Gerchak (2004), a self-financed betting mechanism that truthfully elicits probabilistic forecasts for single events. Under Kilgour and Gerchak's mechanism, a forecaster's payment depends on her relative performance (measured by a proper scoring rule)

[1] www.probabilitysports.com
[2] www.netflixprize.com
[3] www.kaggle.com

compared with other forecasters. Specifically, her total payment is the difference between her own score and the average score of all other forecasters. For a single event, ELF uses a similar idea to compute scores for all forecasters that are non-negative and sum up to 1. Treating these scores as a probability distribution over forecasters, ELF then runs a lottery to determine the winner of the prize. With multiple events, ELF runs one such lottery for each individual event, eventually awarding the prize to the forecaster who has won the most event lotteries.

In this way, ELF probabilistically selects a single winning forecaster while incentivizing truthful forecasts for any sequence of independent events, regardless of the number of events being predicted or the specific risk preferences of the forecasters. We lower-bound the probability that ELF selects the most accurate forecaster, and show that this bound approaches 1 as the number of events grows. Our techniques generalize to other natural settings, such as the truthful ranking of forecasters and hiring a forecaster with high accuracy on future events.

We emphasize that using ELF as an incentive scheme does not restrict the choice of whether and how to aggregate forecasts once they have been elicited. Indeed, ELF is not a substitute for a forecast aggregation algorithm, but a complement. The question of how to aggregate forecasts has been studied extensively. Lichtendahl et al. (2013) show that under a commonly-known public-private signal model, a simple average of "gamed" forecasts is more accurate than a simple average of truthful forecasts, but state-of-the-art aggregation algorithms, such as the "extremized mean," consistently outperform simple averaging in practice (Atanasov et al. 2016) and can take advantage of truthful reports.

2 Model

We consider a group of $n \geq 2$ forecasters, indexed by $i \in [n] = \{1, \ldots, n\}$, and m independent events, indexed by $k \in [m] = \{1, \ldots, m\}$. We model these as m independent random variables X_k that take values in $\{0, 1\}$, and we say that "event k occurred" if $X_k = 1$ and that "event k did not occur" if $X_k = 0$. In each of these cases, we say that "event k materialized," and we denote the vector of all materialized outcomes with $\boldsymbol{x} = \big(x_1, \ldots, x_k, \ldots, x_m\big)$. The true, unknown probability that event k occurs is θ_k with $\theta_k \in (0, 1)$ for all $k \in [m]$. Every forecaster i has a subjective belief $p_{i,k}$ of the probability that event k will occur with $p_{i,k} \in (0, 1)$ for all $i \in [n]$ and all $k \in [m]$. Throughout the paper we assume that it is common knowledge that the m events are independent. (We discuss the problems that arise when events can be correlated in Section 6.5.) All forecasters report their beliefs about event k at the same time, before event k materializes. When reporting on event k, we allow forecasters to know the outcomes of all past events. The reported forecast of forecaster i for event k is denoted by $y_{i,k} \in [0, 1]$. A forecaster's report can be equal to her true belief (i.e., $y_{i,k} = p_{i,k}$) but does not have to be, and we denote the vector of i's reported forecasts as $\boldsymbol{y_i} = \big(y_{i,1}, \ldots, y_{i,k}, \ldots, y_{i,m}\big)$.

Once all m events have materialized, the mechanism selects one of the n forecasters as the "winner." The selection is based on the event outcomes and all forecasters' reports on all events. We allow this selection to be randomized.

Definition 1. *A* forecaster selection mechanism $M : \boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x} \to [n]$ *takes all forecasters' reports on all events and the materialized outcomes of all events, and outputs a single forecaster.*

In contrast to standard proper scoring rules, forecasters only care about being selected. Every forecaster thus seeks to maximize the probability of being selected. Incorporating forecaster i's subjective beliefs over event outcomes and the mechanism's randomization (if any), we obtain the following definition for strict truthfulness of a mechanism.

Definition 2. *Forecaster selection mechanism* $M(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x})$ *is* strictly truthful *if and only if for all* $i \in [n]$, *all* $\boldsymbol{p_i}$, *all* $\boldsymbol{y_i'} \neq \boldsymbol{p_i}$, *and all* $\boldsymbol{y_j}$ *for* $j \neq i$, $\Pr_{\boldsymbol{x} \sim \boldsymbol{p_i}}\big(M(\boldsymbol{y_1}, \ldots, \boldsymbol{p_i}, \ldots, \boldsymbol{y_n}, \boldsymbol{x}) = i\big) > \Pr_{\boldsymbol{x} \sim \boldsymbol{p_i}}\big(M(\boldsymbol{y_1}, \ldots, \boldsymbol{y_i'}, \ldots, \boldsymbol{y_n}, \boldsymbol{x}) = i\big)$.

Observe that we do not require the typical assumption that forecasters are risk neutral: every forecaster strictly prefers being selected over not being selected, so that the higher the probability of being selected, the better. This idea is not new; previous work used lotteries to address unknown risk preferences of forecasters (Karni 2009; Lambert 2011). While we also reward forecasters probabilistically (and obtain robustness to unknown risk preferences as a bonus), the primary reason we use lotteries is because we have many forecasters but only a single prize to award. To the best of our knowledge, we are the first to study this competitive lottery setting in the context of forecasting.

We compare forecasters by their accuracy, which is defined as 1 minus the squared distance between their reports and the true (unknown) probabilities, averaged over all m events. The accuracy a_i of forecaster i is thus

$$a_i = 1 - \frac{1}{m} \sum_{k=1}^{m} \big(y_{i,k} - \theta_k\big)^2$$

with higher a_i being better. Observe that $a_i \in (0, 1]$ for all $i \in [n]$. Of course, other definitions of accuracy are thinkable, but squared loss is commonly used in practice. We will see in Section 3 how the definition of accuracy dictates the choice of which proper scoring rule to use.

The objective in this work is to select the forecaster with the highest accuracy with as high a probability as possible, and ideally with probability approaching 1 as m grows. Of course, one could imagine other objectives, such as maximizing the expected accuracy of the selected forecaster or minimizing the accuracy gap between the selected and the best forecaster. We briefly discuss alternatives in Section 6.

3 Forecaster Selection Using Standard Proper Scoring Rules

Consider a single forecaster and a single event. A scoring rule computes a payment that depends on the event outcome x and the forecaster's report y regarding the probability that $x = 1$, paying the forecaster some amount $R(y, x)$.

Definition 3 (Strictly Proper Scoring Rule)**.** *A* scoring rule $R(y,x) \in \mathbb{R} \cup \{-\infty\}$ *is a mapping from reports* $y \in [0,1]$ *and outcomes* $x \in \{0,1\}$ *to scores. A scoring rule* R *is* proper *if for all* $p, y \in [0,1]$, $\mathbf{E}_{x\sim p}\big[R(p,x)\big] \geq \mathbf{E}_{x\sim p}\big[R(y,x)\big]$*, and* strictly proper *if the inequality is strict whenever* $y \neq p$*.*

There exist infinitely many proper scoring rules since any (strictly) convex function corresponds to a (strictly) proper scoring rule (Gneiting and Raftery 2007, Theorem 1). In this paper, we focus on the *quadratic scoring rule* (Brier 1950), the most widely used scoring rule both in the literature (e.g., in Lichtendahl et al. (2013)) and in practice (e.g., in the Good Judgment Project and ProbabilitySports).

Proposition 1. *(Brier 1950) The* quadratic scoring rule $R_q(y,x) = 1-(y-x)^2$ *is strictly proper.*

Observe that Definition 3 is phrased in an incentive spirit, where the expectation is taken with respect to a forecaster's subjective belief p. Proper scoring rules also have an accuracy interpretation. If the expectation is taken with respect to the true probability θ of the event occurring, then properness implies that reporting the true probability obtains a higher expected score than any other report. Less accurate reports lead to lower expected scores. Different proper scoring rules interact particularly nicely with different notions of accuracy. As shown in the second statement of Proposition 2, the quadratic score has a nice connection with our definition of accuracy. Specifically, the expected difference in quadratic score between two forecasters is equal to the difference between their accuracies. Alternative definitions of accuracy would suggest the use of alternative scoring rules.

Proposition 2. *The quadratic scoring rule has expected score* $\mathbf{E}_{x\sim\theta}\big[R_q(y,x)\big] = \theta^2-\theta+1-(\theta-y)^2$*. Further,* $\mathbf{E}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big[\sum_{k=1}^m \big(R_q(y_{i,k},x_k)-R_q(y_{j,k},x_k)\big)\Big] = m(a_i-a_j)$.

Proof. The first statement is easily derived by expanding out and rearranging the terms in

$$\mathop{\mathbf{E}}_{x\sim\theta}\big[R_q(y,x)\big] = \theta\big(1-(y-1)^2\big)+(1-\theta)\big(1-(y-0)^2\big).$$

For the second statement,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\left[\sum_{k=1}^m\Big(R_q(y_{i,k},x_k)-R_q(y_{j,k},x_k)\Big)\right]\\
&=\sum_{k=1}^m\Big(\big(\theta_k^2-\theta_k+1-(\theta_k-y_{i,k})^2\big)\\
&\qquad-\big(\theta_k^2-\theta_k+1-(\theta_k-y_{j,k})^2\big)\Big)\\
&=\left(m-\sum_{k=1}^m\big(y_{i,k}-\theta_k\big)^2\right)-\left(m-\sum_{k=1}^m\big(y_{j,k}-\theta_k\big)^2\right)\\
&=m(a_i-a_j).
\end{aligned}$$

□

3.1 Mechanism

A natural way to extend a proper scoring rule R to a forecaster selection mechanism is to output the forecaster with highest score according to R, summed across all m events. This mechanism is commonly used in practice to choose top forecasters, including by the Good Judgment Project and ProbabilitySports. Let M_q denote the mechanism derived in this way from the quadratic score. That is, M_q selects the forecasters with highest quadratic score,

$$M_q(\boldsymbol{y_1},\ldots,\boldsymbol{y_n},\boldsymbol{x}) \in \operatorname*{argmax}_{i\in[n]}\sum_{k=1}^m R_q(y_{i,k},x_k),$$

with ties broken uniformly at random. In the remainder of this section, we illustrate that, despite its common use, M_q may select an arbitrarily bad forecaster, while not incentivizing forecasters to report their beliefs truthfully.

3.2 Incentives

It is well known that selecting a forecaster according to highest quadratic score may produce perverse incentives. In general, forecasters are incentivized to make over-confident reports to increase their chance of being judged the best forecaster *ex post* for at least some outcomes. In this section we present a representative example; for a thorough analysis of the strategic behavior of competitive forecasters when ranked by standard proper scoring rules, we defer to Lichtendahl and Winkler (2007).

Example 1. *Consider* $m \geq 1$*,* $n \geq 2$*, and* $p_{i,k} = 0.9$ *for all forecasters* i *and events* k*. If all forecasters report truthfully, then they achieve the same quadratic score regardless of the outcome. Each is therefore chosen as the winner with probability* $1/n$*. Suppose however that forecaster 1 misreports* $y_{1,1} = 0.95$*. Then, forecaster 1 has the highest quadratic score whenever* $x_1 = 1$*, which occurs with probability 0.9, so this is an advantageous misreport.*

3.3 Accuracy

We now show that M_q may fail to choose even a perfect forecaster with any positive probability.

Proposition 3. *For arbitrary* $m \geq 1$*,* $n \geq 3$*, and true event probabilities* $\theta_1,\ldots,\theta_m \in (0,1)$*, there exist reports* $\boldsymbol{y_1},\ldots,\boldsymbol{y_n}$ *such that the best forecaster has accuracy 1 but is selected with probability 0 by* M_q*.*

Proof. Let forecaster 1 report perfectly on all events, (i.e., $y_{1,k} = \theta_k$ for all $k \in [m]$). For the first event, let there be at least one other forecaster with more weight on each possible outcome. For example, let forecaster 2 report $y_{2,1} = 1$ and forecaster 3 report $y_{3,1} = 0$. For all following events, let all n forecasters report perfectly (i.e., $y_{i,k} = \theta_k$ for all i and all $k \geq 2$). Then, for any outcome vector $\boldsymbol{x}$, either forecaster 2 or forecaster 3 will have a higher quadratic score than forecaster 1, despite forecaster 1 having accuracy 1. □

Note that the example used in the proof of Proposition 3 shows that the best forecaster may be selected with low probability, but does not say anything about the quality of the selected forecaster. Indeed, for large m, all forecasters

have very similar quality, so the selected forecaster will be close to best in this particular example.

4 Truthful Forecaster Selection with a Single Event

In this section and the next, we present a forecaster selection mechanism that avoids the shortcomings of M_q. To build intuition, we begin by considering a single-event setting ($m = 1$). In Section 5, we show how to extend our mechanism to handle multiple events.

What needs to hold in order for a forecaster selection mechanism to be truthful? First note that truthfulness requires that, holding the reports of everyone but forecaster i fixed, the probability f_i of choosing forecaster i must behave like a proper scoring rule for i. If this is not the case, then i could increase her probability of being selected by misreporting. Thus we need proper scoring rules for each forecaster that are non-negative and always sum to 1 to form a valid probability distribution. A natural first attempt to achieve this would be to use a standard scoring rule, like the quadratic score, and renormalize by dividing by the sum of all forecasters' scores. However, as Example 2 shows, this renormalized scheme is not truthful.

Example 2. *Let $n = 2$, and suppose $p_1 = p_2 = 0.9$. If both forecasters report truthfully, each is chosen with probability 0.5, regardless of the outcome. However, suppose that forecaster 1 reports $y_1 = 0.8$. Now, if $x = 1$, the probability that she is chosen is $R_q(0.8, 1)/\big(R_q(0.8, 1)+R_q(0.9, 1)\big) = 0.96/1.95$, and if $x = 0$, the probability that she is chosen is $R_q(0.8, 0)/\big(R_q(0.8, 0) + R_q(0.9, 0)\big) = 0.36/0.55$. Thus, her probability of being chosen is $0.9 \cdot 0.96/1.95 + 0.1 \cdot 0.36/0.55 \approx 0.509 > 0.5$, which means that $y_1 = 0.8$ is an advantageous misreport.*

The reason that such a renormalization of proper scores breaks truthfulness is that the probability of choosing forecaster i depends *multiplicatively* on a function of other forecasters' reports and, crucially, the outcome x. To get around this, we borrow a trick from the competitive scoring rule mechanism of Kilgour and Gerchak (2004), which takes advantage of the fact that truthfulness is preserved when adding or subtracting a function of other reports and the outcome. Using their mechanism, each forecaster's payment is a standard proper score (such as quadratic) minus the average standard score of all other forecasters. Our Event-Lotteries Forecaster Selection Mechanism (ELF) uses a similar idea to "normalize" all forecasters' scores *additively*, so that they are non-negative and sum up to 1. ELF then runs a lottery based on these scores to determine the winner of the prize. Alternatively, one can think of ELF as giving each forecaster a 1/n probability to start with and adjusting this up or down depending on how their performance compares with that of other forecasters. As we will see, ELF is strictly truthful and selects high-accuracy forecasters with higher probability than low-accuracy forecasters.

4.1 Mechanism

For a single event, the *Event-Lotteries Forecaster Selection Mechanism (ELF)* $M_l(y_1, \ldots, y_n, x)$ selects forecaster $i \in [n]$ with probability

$$f_i = \frac{1}{n} + \frac{1}{n}\left(R_q(y_i, x) - \frac{1}{n-1}\sum_{j \neq i} R_q(y_j, x)\right).$$

It is easy to see that $\big(f_1, \ldots, f_n\big)$ is a valid probability distribution. That each f_i is non-negative follows immediately from R_q being bounded in $[0, 1]$. And $\sum_{i=1}^n f_i = 1$ since

$$\sum_{i=1}^{n}\left(R_q(y_i, x) - \frac{1}{n-1}\sum_{j \neq i} R_q(y_j, x)\right) = 0.$$

4.2 Incentives

Using a similar argument to that of Kilgour and Gerchak (2004), we can show that ELF is truthful.

Theorem 4. *ELF is strictly truthful for a single event.*

Proof. Note that the only term in f_i that depends on y_i is $(1/n)R_q(y_i, x)$. By linearity of expectation, maximizing $\mathbf{E}_{x\sim p_i}[f_i]$ is therefore equivalent to maximizing $R_q(y_i, x)$, and truthfulness follows from the fact that R_q is a strictly proper scoring rule. □

4.3 Accuracy

We now show that ELF chooses forecasters with higher accuracy more often than those with lower accuracy.

Proposition 5. *The probability that ELF chooses forecaster i given true probability θ is $\frac{1}{n} + \frac{1}{n}\left(a_i - \frac{1}{n-1}\sum_{j\neq i} a_j\right)$.*

Proof. Using the second half of Proposition 2,

$$\begin{aligned}
\Pr_{x\sim\theta}\Big(M_l(y_1, \ldots, y_n, x) = i\Big) &= \mathop{\mathbf{E}}_{x\sim\theta}\big[f_i\big] \\
&= \mathop{\mathbf{E}}_{x\sim\theta}\left[\frac{1}{n} + \frac{1}{n}\cdot\frac{1}{n-1}\sum_{j\neq i}\Big(R_q(y_i, x) - R_q(y_j, x)\Big)\right] \\
&= \mathop{\mathbf{E}}_{x\sim\theta}\left[\frac{1}{n} + \frac{1}{n}\cdot\frac{1}{n-1}\sum_{j\neq i}(a_i - a_j)\right] \\
&= \frac{1}{n} + \frac{1}{n}\left(a_i - \frac{1}{n-1}\sum_{j\neq i} a_j\right).
\end{aligned}$$

□

5 Truthful Forecaster Selection with Multiple Events

In this section, we show how to generalize our single-event selection mechanism to handle multiple independent events. One seemingly natural generalization would be to run ELF on each of the m events independently, and then choose each forecaster i with probability $\sum_{k=1}^m f_{i,k}/m$, where $f_{i,k} = \Pr\big(M_l(y_{1,k}, \ldots, y_{n,k}, x_k) = i\big)$. Unfortunately, doing this would not satisfy one of our key desiderata: that our mechanism chooses the most accurate forecaster with probability tending to 1 as the number of events grows. This failure is illustrated by Example 3.

Example 3. *Let $n = 2$, and suppose that there are m events, with $\theta_k = 0.5$ for all $k \in [m]$. Suppose that forecaster 1 reports $y_{1,k} = 0.5$ for all k, while forecaster 2 reports $y_{2,k} = 1$ for all k. As m grows large, approximately half the events will have outcome $x_k = 0$, so that $f_{1,k} = \frac{1}{2} + \frac{1}{2}(0.75 - 0) = 0.875$ and $f_{2,k} = \frac{1}{2} + \frac{1}{2}(0 - 0.75) = 0.125$. The other half of the events will have outcome $x_k = 1$, so that $f_{1,k} = \frac{1}{2} + \frac{1}{2}(0.75 - 1) = 0.375$ and $f_{2,k} = \frac{1}{2} + \frac{1}{2}(1 - 0.75) = 0.625$. Therefore, after all m outcomes are observed, forecaster 1 is chosen with probability $(0.875 + 0.375)/2 = 0.625$, and forecaster 2 is chosen with probability $(0.125 + 0.625)/2 = 0.375$. Despite the fact that forecaster 1 is a much better forecaster, we still choose forecaster 2 with 37.5% probability, even with a large number of events.*

In the following, we propose and analyze an alternative generalization that is guaranteed to select the best forecaster with probability tending to 1 as the number of events grows. It does this by running an independent lottery on each round and selecting the forecaster who wins the most lotteries. Having independent lotteries retains truthfulness while picking up on forecasters' accuracy differences better than a single lottery, for the same underlying reason that statistical concentration inequalities hold (e. g., Hoeffding 1963).

5.1 Mechanism

For multiple events, the Event-Lotteries Forecaster Selection Mechanism (ELF) $M_l(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x})$ is defined as:

1. For each event k, pick forecaster i to be the event winner w_k with probability

$$f_{i,k} = \frac{1}{n} + \frac{1}{n}\Big(R_q\big(y_{i,k}, x_k\big) - \frac{1}{n-1}\sum_{j \neq i} R_q\big(y_{j,k}, x_k\big)\Big).$$

2. Select the forecaster who won the most events, $\operatorname{argmax}_i \sum_{k=1}^{m} \mathbb{1}(w_k = i)$, breaking ties uniformly at random. Here $\mathbb{1}$ denotes the 0/1 indicator function.

5.2 Incentives

We now show that ELF remains truthful with multiple events. This proof relies heavily on our assumption that the independence of the m events is common knowledge. In particular, if a forecaster believes that events are correlated, she could have an incentive to misreport.

Theorem 6. *ELF is strictly truthful for $m \geq 1$ events.*

Proof. Without loss of generality, order the events by the time at which the forecasters report on them. So event 1 is reported on first, and event m is reported on last.

Fix all forecasters' reports $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$. Consider forecaster i and suppose that i's report on at least one event does not equal her true belief. Of all such events, let k' be the one with lowest index. That is, $y_{i,k'} \neq p_{i,k'}$, and $y_{i,k} = p_{i,k}$ for all events $k < k'$. We will argue that i can improve her probability of being selected as the best forecaster by instead reporting $y_{i,k'} = p_{i,k'}$. The intuition is that doing so increases i's probability of winning event k', by truthfulness on a single event, while not affecting her probability of winning any other event. This shows truthfulness, since we can repeat the argument as long as there remains an event on which i does not report truthfully.

Formally, consider the $m - 1$ events other than k'. Since we have assumed that the outcomes for these events are independent of the outcome of event k', we can reason about them independently. There are three possible cases that we distinguish, depending on the winners of these events:

1. There exists some forecaster $j \neq i$ that wins at least two more events than forecaster i, or i wins at least two more events than all other forecasters.
2. There exists some forecaster $j \neq i$ that wins exactly one more event than forecaster i, but no forecaster that wins two or more events more than i.
3. No forecaster wins more events than i, but there exists some forecaster $j \neq i$ that wins either the same number of events as i, or one event less than i.

In Case 1, i's probability of being selected is either 0 or 1, regardless of who wins event k'. Therefore, her utility is unaffected by her report $y_{i,k'}$.

In Case 2, forecaster i wants to maximize $f_{i,k}$, since winning event k is the only scenario in which she gets selected with any non-zero probability. By strict truthfulness of ELF for a single event (Theorem 4), she accomplishes this only by truthfully reporting $y_{i,k'} = p_{i,k'}$.

In Case 3, forecaster i has two (potentially conflicting) objectives: to maximize $f_{i,k}$, and to minimize $f_{j,k}$ for all forecasters j from some subset of the other forecasters. However, it is easy to check that these objectives are not actually in conflict; truthfully reporting $y_{i,k'} = p_{i,k'}$ simultaneously uniquely maximizes $f_{i,k}$, and uniquely minimizes $f_{j,k}$, for any $j \neq i$.

Since the three cases are exhaustive, and i is weakly incentivized to report truthfully in all of them, we have already shown that ELF is weakly truthful. To complete the argument that i is strictly incentivized to report $y_{i,k'} = p_{i,k'}$, we argue that i believes that at least one of Cases 2 and 3 occurs with positive probability (in fact, she will believe that both occur with positive probability but one is sufficient). To see this, note that for all events k that have already materialized at the time the forecasters report on event k' it holds that

$$f_{j,k} > 0 \qquad \forall j \in [n], \tag{1}$$

and for all events k that have not yet materialized at the time the forecasters report on event k' it holds that

$$\mathop{\mathbf{E}}_{x_k \sim p_{i,k}}\big[f_{j,k} > 0\big] \qquad \forall j \in [n]. \tag{2}$$

Equation 1 holds because if k has materialized before the forecasters report on k', then the forecasters reported on k before they reported on k' (i.e., $k < k'$). Therefore, by minimality of k', $y_{i,k} = p_{i,k} \in (0, 1)$. From this, it can easily be verified that $f_{j,k} > 0$ for all j, regardless of the outcome x_k. To verify Equation 2, note that $p_{i,k} \in (0, 1)$, which means that i has strict uncertainty about outcome x_k. This implies that $\mathbf{E}_{x_k \sim p_{i,k}}\big[R_q(y_{j,k}, x_k) > 0\big]$ for any $y_{j,k} \in [0, 1]$, and it is easy to check that any forecaster j with non-zero quadratic

score for event k has $f_{j,k} > 0$. Therefore, since (the expected value of) $f_{j,k} > 0$ for all forecasters $j \in [n]$ and all events $k \neq k'$, it is possible that all forecasters win the same number of events (up to one, due to indivisibility of events) from these $m-1$ events, which corresponds to Case 3.

Thus, i improves her probability of being selected by reporting $y_{i,k} = p_{i,k}$, and ELF is strictly truthful. □

5.3 Accuracy

Finally, we bound the probability that ELF chooses the most accurate forecaster. The proof uses Hoeffding's inequality (Hoeffding 1963), which we state here for convenience.

Theorem 7 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_m$ be independent random variables bounded by the interval $[0,1]$. Define $S_m = X_1 + \ldots + X_m$. Then*

$$\Pr\Big(\big|S_m - \mathbf{E}\big[S_m\big]\big| \geq t\Big) \leq 2e^{-\frac{2t^2}{m}}.$$

Theorem 8. *Suppose that $a_i \geq a_j + \epsilon$ for all $j \neq i$. Then the probability that ELF chooses forecaster i is*

$$\Pr_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big(M_l\big(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x}\big) = i\Big) \geq 1 - 4(n-1)e^{-\frac{m\epsilon^2}{2(n-1)^2}}.$$

That is, for fixed n and 'accuracy gap' ϵ, for any $\delta > 0$, ELF chooses the best forecaster with probability at least $1-\delta$ if

$$m \geq \frac{2(n-1)^2}{\epsilon^2} \ln\left(\frac{4(n-1)}{\delta}\right).$$

Proof. We first bound the difference between the expected number of events won by i and the expected number of events won by some other forecaster $j \neq i$:

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\left[\sum_{k=1}^{m} f_{i,k}\right] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\left[\sum_{k=1}^{m} f_{j,k}\right] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\left[\sum_{k=1}^{m}\big(f_{i,k} - f_{j,k}\big)\right]$$

$$= \frac{\mathbf{E}_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big[\sum_{k=1}^{m}\big(R_q(y_{i,k}, x_k) - R_q(y_{j,k}, x_k)\big)\Big]}{n-1} \geq \frac{m\epsilon}{n-1}.$$

The second equality follows from substituting the definition of $f_{i,k}$ and simplifying, and the inequality follows from the second part of Proposition 2 and the assumed difference in accuracy between i and all other forecasters.

We now upper bound the probability that forecaster j wins more events than forecaster i. Let F_i be a random variable for the number of events won by forecaster i, so that $\mathbf{E}[F_i] = \mathbf{E}\big[\sum_{k=1}^{m} f_{i,k}\big]$, where the latter expectation is taken over the outcomes, and the former is taken over the outcomes and the randomness of the lotteries. Likewise, let F_j be the number of events won by forecaster j. From the equation above, if $F_j \geq F_i$, then it holds that either $\mathbf{E}[F_i] - F_i \geq \frac{m\epsilon}{2(n-1)}$ or $F_j - \mathbf{E}[F_j] \geq \frac{m\epsilon}{2(n-1)}$. By Hoeffding's inequality,

$$\Pr\left(\Big|F_i - \mathbf{E}\big[F_i\big]\Big| \geq \frac{m\epsilon}{2(n-1)}\right) \leq 2e^{-\frac{m\epsilon^2}{2(n-1)^2}},$$

with the analogous inequality holding for F_j. Putting these together, we have

$$\begin{aligned}
&\Pr(F_j \geq F_i)\\
&\leq \Pr\left(\left(\mathbf{E}\big[F_i\big] - F_i \geq \frac{m\epsilon}{2(n-1)}\right) \cup \left(F_j - \mathbf{E}\big[F_j\big] \geq \frac{m\epsilon}{2(n-1)}\right)\right)\\
&\leq \Pr\left(\mathbf{E}\big[F_i\big] - F_i \geq \frac{m\epsilon}{2(n-1)}\right) + \Pr\left(F_j - \mathbf{E}\big[F_j\big] \geq \frac{m\epsilon}{2(n-1)}\right)\\
&\leq 4e^{-\frac{m\epsilon^2}{2(n-1)^2}}.
\end{aligned}$$

Finally, we lower bound the probability that ELF selects forecaster i.

$$\begin{aligned}
&\Pr_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big(M_l\big(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x}\big) = i\Big)\\
=&1 - \sum_{j\neq i} \Pr_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big(M_l\big(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}, \boldsymbol{x}\big) = j\Big)\\
\geq&1 - \sum_{j\neq i} \Pr_{\boldsymbol{x}\sim\boldsymbol{\theta}}\Big(F_j \geq F_i\Big)\\
\geq&1 - 4(n-1)e^{-\frac{m\epsilon^2}{2(n-1)^2}},
\end{aligned}$$

where the first transition holds because exactly one forecaster is selected and the second because $F_j \geq F_i$ is a necessary condition for forecaster j to be selected by ELF. The final transition holds by plugging in the earlier inequality. □

6 Discussion

In this section, we discuss connections between our work and related research areas, describe extensions to our model, and discuss the open problem of handling correlations.

6.1 Correspondence with Wagering Mechanisms

As discussed in Section 4, ELF is very closely related to Kilgour and Gerchak's competitive scoring rule. Competitive scoring rules are a special case of *wagering mechanisms* (Lambert et al. 2008; Chen et al. 2014), in which each forecaster reports both a probability p_i of an event occurring and a monetary wager $w_i \in \mathbb{R}^+$. Once the event has materialized, wagers are redistributed to the forecasters in such a way that the redistributed amounts depend on each forecaster's relative performance. A competitive scoring rule is simply a wagering mechanism in which all forecasters are required to wager the same amount.

There is a one-to-one correspondence between (truthful) budget-balanced competitive scoring rules and (truthful) single-event forecaster selection mechanisms. In particular, we can view a forecaster selection mechanism as a mechanism in which each forecaster starts out with an initial probability and "wagers" this probability against other forecasters. These initial probabilities are then redistributed among the n forecasters according to their relative performance. Using this interpretation, this paper provides a general framework for constructing forecaster selection mechanisms: first, fix a particular competitive scoring rule (or wagering mechanism) and define the corresponding single-event forecaster

selection mechanism. Second, extend to multiple events by picking a winner for each event and selecting the forecaster who won the most events.

Because of this strong correspondence, existing work on wagering mechanisms is informative about limitations for the design of forecaster selection mechanisms. A general problem with wagering mechanisms is that of low stakes: even for fairly different reports, forecasters generally stand to lose only a small fraction of their wager, regardless of the outcome. When a low-stakes wagering mechanism is used as a building block for a forecaster selection mechanism, this means that even a bad forecaster will be chosen with relatively high probability. Lambert et al. (2008) show that, under fairly mild assumptions on the behavior of the wagering mechanism (all of which are satisfied by the Kilgour-Gerchak competitive scoring rules), it is impossible for any forecaster to more than double her wager. In our forecaster selection setting, this directly implies that no forecaster wins an event with probability higher than $2/n$, even if she reports perfectly and all other forecasters have the worst-possible reports. In practice, this means that any forecaster selection mechanism built on a wagering mechanism that satisfies the assumptions of Lambert et al. (including ELF) will only slowly converge to selecting the best forecaster with probability 1 as the number of events grows.

To circumvent this, some assumptions would need to be relaxed. One promising direction is to drop the requirement of *strict budget balance*, meaning that the mechanism must pay out exactly what it takes in, and instead require *weak* budget balance, which allows the mechanism to profit. For forecaster selection, this would mean that probabilities could sum to less than one, which one can interpret as allowing the possibility of abstaining from choosing a forecaster (with some appropriate, application-dependent, penalty for abstaining). Using a wagering mechanism that gives up strict budget balance for higher stakes, such as the Double Clinching Auction of Freeman, Pennock, and Vaughan (2017), may produce faster convergence in practice.

6.2 Forecaster Hiring

Forecasting competitions are often used as a method of choosing a forecaster to hire when future predictions are needed. In this setting, the goal of the selection mechanism is to choose the forecaster who will be (approximately) the most accurate on future events. There is an implicit assumption here that good performance on the observed events translates into good performance in the future, a well-established fact in practice (e. g., Mellers et al. 2014).

Our methods and results can be extended to this setting. Instead of defining accuracy as a function of the m events being predicted, we could instead assume a joint distribution D over event probabilities θ and the beliefs p_i of each forecaster i. We could then define the accuracy of forecaster i in terms of the expected quadratic score of her forecasts with respect to D.

Under this model, mechanism M_q discussed in Section 3 can be viewed as performing an analog of empirical risk minimization. Similar to how basic empirical risk minimization bounds are proved for PAC learning (Kearns and Vazirani 1994), we could then argue that, with high probability, the forecaster with the highest accuracy on any observed sample of events has expected accuracy close to that of the best forecaster in the set. Therefore, as the number of events grows large, the forecaster selected by M_q would be guaranteed to have accuracy arbitrarily close to that of the most accurate forecaster. However, the incentive issues remain. The advantage of ELF is that it obtains truthful reports for any m while achieving similar accuracy guarantees as m grows large. In this sense, ELF can be viewed as a mechanism for learning in the presence of strategic agents.

6.3 Beyond Binary Outcomes

So far, we have restricted our analysis to events with binary outcomes. In practice, we are also interested in events with non-binary (categorical) outcomes, to which the definition of forecaster accuracy can be naturally extended. Unsurprisingly, selecting the forecaster with highest quadratic score (using the generalization introduced by Brier, 1950) inherits all the problems exhibited in Section 3.

ELF readily extends to categorical outcomes. The competitive scoring rule of Kilgour and Gerchak (2004) is truthful for categorical outcomes when the generalized quadratic scoring rule is used, and ELF inherits this truthfulness for a single event. With multiple events, truthfulness follows from the same arguments used in the proof of Theorem 6. Moreover, in terms of accuracy, it still holds that more accurate forecasters obtain higher quadratic scores in expectation, so the most accurate forecaster still wins the most events in expectation. Hence, by a qualitatively identical argument as the one in Theorem 8, we can lower bound the probability of selecting the most accurate forecaster with a bound that approaches 1 as the number of events grows.

A similar extension would allow us to truthfully elicit a sequence of expectations of bounded continuous random variables and select the highest accuracy forecaster.

6.4 Outputting a Forecaster Ranking

In some practical applications, it may be more appropriate to output a ranking rather than a single forecaster. For example, most play-money prediction markets maintain a ranking of contestants. Ranking forecasters in order of quadratic score again inherits all of the problems described in Section 3.

ELF can be adapted to produce a ranking by simply ordering forecasters according to the number of events that they win. As long as forecasters strictly prefer higher positions in the ranking, ELF remains truthful, since forecasters maximize their probability of winning an event (and potentially moving up in the ranking) by reporting truthfully.

Moreover, the same style of accuracy results from Section 5.3 hold, at least qualitatively, when the objective is to maximize the probability of outputting the correct ranking. In expectation, more accurate forecasters achieve higher quadratic scores, leading to higher expected values of $f_{i,k}$. Thus, more accurate forecasters win more events in the long run, and the true ranking is faithfully revealed.

6.5 Correlated Events

When $m > 1$, a forecaster who believes that two or more events are correlated may have an incentive to misreport under ELF, as illustrated in the following example.

Example 4. *Let $n = 2$ and let m be large. Suppose that all events are perfectly correlated (so that either all $x_k = 1$ or all $x_k = 0$), with $\theta_1 = \ldots = \theta_m = 0.8$. If $\boldsymbol{y_1} = \boldsymbol{y_2} = (0.8, \ldots, 0.8)$, then ELF chooses each forecaster with probability 0.5. Suppose instead that forecaster 1 reports $\boldsymbol{y_1} = (1, \ldots, 1)$. If $x_k = 1$ for all k, then for each k, $\Pr(w_k = 1) = 0.5 + 0.5(1 - 0.96) = 0.52$. If m is sufficiently large and forecaster 1 has probability 0.52 of winning each event, she is selected by ELF over forecaster 2 with probability close to 1. This happens with probability 0.8, so forecaster 1 is selected with probability close to 0.8, much higher than if she had reported truthfully.*

For similar reasons, ELF may fail to identify the best forecaster with high probability if events are correlated.

It is an open question whether the performance of ELF (in terms of both truthfulness and accuracy) degrades gracefully under mild correlations. As one piece of evidence suggesting that it might, if a particular event is not correlated with any others, a forecaster has incentive to report truthfully for that event. It would also be interesting to investigate whether forecaster selection mechanisms can be designed to be (theoretically or empirically) robust to some level of correlation, perhaps by allowing more expressive reports.

7 Conclusion

We examined a setting in which forecasters compete for a single prize, and have shown that choosing the forecaster with highest score according to a standard proper scoring rule can lead to selecting a sub-optimal forecaster with high probability, as well as creating incentives for forecasters to lie about their beliefs. To overcome these drawbacks, we designed the Event-Lotteries Forecaster Selection Mechanism (ELF). ELF both incentivizes truthful reporting and yields provable guarantees on the probability that the most accurate forecaster is selected. As the number of events predicted grows large, the probability that ELF selects the best forecaster approaches 1. Beyond the future research directions outlined in Section 6, another important next step will be to evaluate ELF experimentally against other truthful and non-truthful mechanisms.

Acknowledgments

We thank Rafael Frongillo for pointing us to the work on using lotteries to address unknown risk preferences and the anonymous reviewers for helpful feedback. This work is supported by ERC StG 307036, and was completed in part while R. Freeman was an intern at Microsoft Research.

References

Atanasov, P.; Rescober, P.; Stone, E.; Servan-Schreiber, E.; Tetlock, P. E.; Ungar, L.; and Mellers, B. 2016. Distilling the Wisdom of Crowds: Prediction Markets versus Prediction Polls. *Management Science*. Forthcoming.

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78(1):1–3.

Chen, Y.; Devanur, N. R.; Pennock, D. M.; and Vaughan, J. W. 2014. Removing Arbitrage from Wagering Mechanisms. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC'14)*, 377–394. ACM.

Freeman, R.; Pennock, D. M.; and Vaughan, J. W. 2017. The Double Clinching Auction for Wagering. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC'17)*, 43–60. ACM.

Gneiting, T., and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102:359–378.

Good, I. J. 1952. Rational Decisions. *Journal of the Royal Statistical Society. Series B* 14(1):107–114.

Harris, D. 2013. Facebook is hiring a data scientist, but you'll have to fight for the job. `https://gigaom.com/2013/08/30/facebook-is-hiring-a-data-scientist-but-youll-have-to-fight-for-the-job/`. [Online; accessed 10-September-2017].

Hoeffding, W. 1963. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58(301):13–30.

Karni, E. 2009. A Mechanism for Eliciting Probabilities. *Econometrica* 77(2):603–606.

Kearns, M. J., and Vazirani, U. V. 1994. *An Introduction to Computational Learning Theory*. MIT press.

Kilgour, D. M., and Gerchak, Y. 2004. Elicitation of Probabilities Using Competitive Scoring Rules. *Decision Analysis* 1(2):108–113.

Lambert, N.; Langford, J.; Wortman, J.; Chen, Y.; Reeves, D.; Shoham, Y.; and Pennock, D. M. 2008. Self-Financed Wagering Mechanisms for Forecasting. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*, 170–179. ACM.

Lambert, N. S. 2011. Probability Elicitation for Agents with Arbitrary Risk Preferences. Working Paper.

Lichtendahl, K. C. J., and Winkler, R. L. 2007. Probability Elicitation, Scoring Rules, and Competition Among Forecasters. *Management Science* 53(11):1745–1755.

Lichtendahl, K. C.; Grushka-Cockayne, Y.; and Pfeifer, P. E. 2013. The Wisdom of Competitive Crowds. *Operations Research* 61(6):1383–1398.

Mellers, B.; Ungar, L.; Baron, J.; Ramos, J.; Gurcay, B.; Fincher, K.; Scott, S. E.; Moore, D.; Atanasov, P.; Swift, S. A.; Murray, T.; Stone, E.; and Tetlock, P. E. 2014. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* 25(5):1106–1115.

Servan-Schreiber, E.; Wolfers, J.; Pennock, D. M.; and Galebach, B. 2004. Prediction Markets: Does Money Matter? *Electronic Markets* 14(3):243–251.

Tetlock, P. E., and Gardner, D. 2015. *Superforecasting: The Art and Science of Prediction*. New York, NY, USA: Crown Publishing Group.