Incentive-Compatible Forecasting Competitions

Author(s):
Witkowski, Jens; Freeman, Rupert; Wortman Vaughan, Jennifer; Pennock, David M.; Krause, Andreas

Publication date:
2023-03

Permanent link:
https://doi.org/10.3929/ethz-b-000559971

Originally published in:
Management Science 63(3), https://doi.org/10.1287/mnsc.2022.4410

Funding acknowledgement:
307036 - Large-scale Adaptive Sensing, Learning and Decision Making: Theory and Applications (EC)

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 3, March 2023, pp. 1354–1374
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Incentive-Compatible Forecasting Competitions

Jens Witkowski,[a,*] Rupert Freeman,[b] Jennifer Wortman Vaughan,[c] David M. Pennock,[d] Andreas Krause[e]

[a] Frankfurt School of Finance & Management, Frankfurt 60322, Germany; [b] Darden School of Business, University of Virginia, Charlottesville, Virginia 22903; [c] Microsoft Research, New York, New York 10012; [d] Department of Computer Science, Rutgers University, New Brunswick, New Jersey 08854; [e] Department of Computer Science, ETH Zurich, Zurich 8092, Switzerland
*Corresponding author

Contact: j.witkowski@fs.de, https://orcid.org/0000-0002-4748-9099 (JW); freemanr@darden.virginia.edu, https://orcid.org/0000-0003-4744-9449 (RF); jenn@microsoft.com, https://orcid.org/0000-0002-7807-2018 (JWV); dpennock@dimacs.rutgers.edu, https://orcid.org/0000-0003-0522-4815 (DMP); krausea@ethz.ch, https://orcid.org/0000-0001-7260-9673 (AK)

Received: December 24, 2020
Revised: August 28, 2021; December 6, 2021
Accepted: December 11, 2021
Published Online in Articles in Advance: May 17, 2022

https://doi.org/10.1287/mnsc.2022.4410

Abstract. We initiate the study of incentive-compatible forecasting competitions in which multiple forecasters make predictions about one or more events and compete for a single prize. We have two objectives: (1) to incentivize forecasters to report truthfully and (2) to award the prize to the most accurate forecaster. 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 more extreme beliefs. In this paper, we introduce two novel forecasting competition mechanisms. Our first mechanism is incentive compatible and guaranteed to select the most accurate forecaster with probability higher than any other forecaster. Moreover, we show that in the standard single-event, two-forecaster setting and under mild technical conditions, no other incentive-compatible mechanism selects the most accurate forecaster with higher probability. Our second mechanism is incentive compatible when forecasters' beliefs are such that information about one event does not lead to belief updates on other events, and it selects the best forecaster with probability approaching one as the number of events grows. Our notion of incentive compatibility is more general than previous definitions of dominant strategy incentive compatibility in that it allows for reports to be correlated with the event outcomes. Moreover, our mechanisms are easy to implement and can be generalized to the related problems of outputting a ranking over forecasters and hiring a forecaster with high accuracy on future events.

History: Accepted by Yan Chen, behavioral economics and decision analysis.

Funding: This work was supported by the European Research Council [Grant ERC StG 307036] and the National Science Foundation [Grant CCF-1445755].

Keywords: artificial intelligence • forecasting • economics: game theory and bargaining

1. Introduction

The study of probabilistic predictions dates back to at least the 1950s when meteorologists developed proper scoring rules as a way to both incentivize truthful forecasts about future events and compare the relative accuracy of different forecasters (Brier 1950, Good 1952). Proper scoring rules are still widely used today to motivate and measure forecasting accuracy (e.g., Atanasov et al. 2017) and an active area of research in decision analysis (e.g., Grushka-Cockayne et al. 2017, Jose 2017).

When forecasters are paid their proper scores, they maximize expected payment by truthfully reporting their beliefs. However, it is rare to see proper scoring rule payments outside of experimental labs. Instead, most real-world forecasting settings are competitions, where forecasters are ranked according to their score and where prizes are given only to the highest-ranked forecasters. Hence, forecasters do not care about maximizing their expected score but about whether their forecasts are judged to be better than others' forecasts. 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. In play-money prediction markets, forecasters often compete for a place at the top of a leaderboard (e.g., Servan-Schreiber et al. 2004). 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,[1] and the machine learning competitions run by Kaggle[2] rank submitted models 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, Chakraborty 2016).

There are good reasons for organizations to run forecasting competitions as opposed to directly paying each forecaster her proper score. First, from a marketing perspective, awarding a single, large prize to the winner is more enticing than offering small payments to everyone. For example, it is unlikely that the Netflix Prize would have created the same media buzz without offering participants the prospect of winning $1,000,000. Second, organizations significantly reduce transaction costs when only a single or small number of prizes are awarded. In addition to the literal transaction costs involved in transferring payments from the organization to the forecasters, there are sometimes legal reasons that are facilitated by having only a single transaction.

However, 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.

This kind of misreporting is not a purely academic possibility but is also observed in real-world forecasting competitions. An example is Kaggle's annual machine learning competition to predict the game outcomes of the NCAA March Madness college basketball tournament, where every participant submits up to two statistical models predicting the outcomes of each possible team pairing. At the end of the 2017 competition, Andrew Landgraf, the creator of that year's winning model, was interviewed by the Kaggle team about his approach, saying (Kaggle 2017) "My idea was to model not only the probability of each team winning each game, but also the competitors' submissions. Combining these models, I searched for the submission with the highest chance of finishing with a prize (top 5 on the leaderboard).[…] The three main processes are[…]: (1) A model of the probability of winning each game, (2) a model of what the competitors are likely to submit, and (3) an optimization of my submission based on these two models." While rational from a forecaster's point of view, this strategic behavior creates two problems for organizations that run forecasting competitions to obtain accurate forecasts: first, the reported forecasts are not truthful and hence not optimized for accuracy but for "winning the game." Second, each forecaster responding to the gaming incentives spends significant effort on strategizing and predicting other forecasters' behavior instead of investing full effort into acquiring the most accurate prediction for the event in question.

In this paper, we initiate the study of incentive-compatible forecasting competitions. After showing that the failure to provide strict truthfulness incentives is inherent to any deterministic forecasting competition mechanism, we present the Event Lotteries Forecasting Competition Mechanism (ELF). ELF borrows a trick from the competitive scoring rule of Kilgour and Gerchak (2004), which truthfully elicits probabilistic forecasts for single events. Under the mechanism of Kilgour and Gerchak, a forecaster's payment depends on her relative performance (measured by a proper scoring rule) 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 nonnegative and sum up to one. Treating these scores as a probability distribution over forecasters, ELF then runs a lottery to determine the winner of the prize. For the prominent single-event, two-forecaster setting, as also studied by Lichtendahl and Winkler (2007), we prove that, under mild technical conditions, there exists no other incentive-compatible mechanism that selects the more accurate forecaster with higher probability.

Our second mechanism is the Independent-Event Lotteries Forecasting Competition Mechanism (I-ELF), which is specifically designed for multiple, independent events and strictly incentive compatible when forecasters' beliefs are such that information about one event does not lead to a belief update on the other events. I-ELF runs one ELF lottery for each individual event, eventually awarding the prize to the forecaster who has won the most event lotteries. As the number of events grows, I-ELF selects the most accurate forecaster with probability approaching one. Moreover, both ELF and I-ELF are robust toward unknown risk preferences and our techniques generalize to other natural settings, such as the incentive-compatible ranking of forecasters and hiring a forecaster with high accuracy on future events.

Forecasting competitions are different from the usual contest settings studied in the literature, such as innovation contests modeled as all-pay auctions (e.g., Konrad 2009). In those models, although there is also a prize to

be awarded, a participant's strategic choice is the effort they invest, determining the quality of their provided solution. In contrast, participants in forecasting competitions strategize about what they should report given their private information. Moreover, the mechanism designer's objective is different in the two settings. Whereas classical contest models seek to maximize the quality of the provided solutions, the primary objective of forecasting competitions such as the Good Judgment Project is to truthfully elicit accurate information from participants.

The question of how to aggregate forecasts has been studied extensively in the decision analysis community (e.g., Satopää et al. 2014, Palley and Soll 2019). We emphasize that using ELF or I-ELF as incentive schemes does not restrict the choice of whether and how to aggregate forecasts once they have been elicited. Indeed, a forecasting competition mechanism is not a substitute for a forecast aggregation algorithm but a complement. 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. However, state-of-the-art aggregation algorithms, such as the extremized mean (Atanasov et al. 2017) and the logit aggregator (Satopää et al. 2014), consistently outperform simple averaging in practice 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 events, indexed by $k \in [m] = \{1, \ldots, m\}$. We model these as m 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 $X_k = 0$. Independent of the event's outcome, we say that "event k materialized." Let $\boldsymbol{X}$ denote the vector-valued random variable of event outcomes and $\boldsymbol{x} = (x_1, \ldots, x_k, \ldots, x_m)$ its realization. Every forecaster i has a subjective belief $p_{i,k} \in [0,1]$ of the probability that event k will occur. We denote the vector of forecaster i's subjective beliefs over all m events as $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,k}, \ldots, p_{i,m}) \in [0,1]^m$. All forecasters report their probabilistic forecasts for all events at the same time, before the first event materializes. (In Section 6.5, we discuss how this assumption can be relaxed for practical purposes.) 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 $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,k}, \ldots, y_{i,m}) \in [0,1]^m$. In settings with only a single event, that is, $m = 1$, we drop the subscript k denoting the event from the event outcomes and from the forecasters' reports and beliefs. 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 *forecasting competition mechanism* M takes all forecasters' reports on all events $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in [0,1]^m \times \ldots \times [0,1]^m$ and the materialized outcomes of all events $\boldsymbol{x} \in \{0,1\}^m$, and selects a forecaster $M(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{x}) \in [n]$.

In contrast to standard proper scoring rules, forecasters only care about being selected. Every forecaster thus seeks to maximize the probability of being selected. The primary objective is to incentivize forecasters to report their true beliefs about the expectation of $\boldsymbol{X}$. Incorporating forecaster i's subjective beliefs, the uncertainty about other forecasters' reports, and the mechanism's randomization (if any), we obtain the following definition for strict incentive compatibility of a mechanism.[3]

Definition 2. Forecasting competition mechanism M is *(robust) strictly incentive compatible* if and only if for all forecasters $i \in [n]$, all belief vectors $\boldsymbol{p}_i$, all joint distributions D over outcomes $\boldsymbol{X}$ and reports $\boldsymbol{Y}_{-i}$ such that the marginal distribution of $\boldsymbol{X}$ is $\mathbf{E}_{\boldsymbol{X} \sim D}[\boldsymbol{X}] = \boldsymbol{p}_i$, and all alternative report vectors $\boldsymbol{y}_i' \neq \boldsymbol{p}_i$,

$$\Pr_{\boldsymbol{X}, \boldsymbol{Y}_{-i} \sim D}(M(\boldsymbol{Y}_1, \ldots, \boldsymbol{p}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i) > \Pr_{\boldsymbol{X}, \boldsymbol{Y}_{-i} \sim D}(M(\boldsymbol{Y}_1, \ldots, \boldsymbol{y}_i', \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i).$$

Observe that this definition of incentive compatibility is very general, allowing us to capture, for instance, that forecaster i believes that $j \neq i$ perfectly forecasts the correct outcome while i herself does not. More generally, it allows for settings in which forecaster i would update her belief upon learning forecaster j's report. In particular, our definition of incentive compatibility applies to standard Bayesian models, where the participants' beliefs stem from noisy observations of some ground truth (e.g., Lichtendahl and Winkler 2007). This is in contrast to previous work that defined *immutable-belief incentive compatibility* (Kilgour and Gerchak 2004, Lambert et al. 2008), which only requires truthful reporting to be optimal when the reports of other forecasters are constant (i.e., with no dependence on each other or the event outcomes). We refer the reader to Appendix A for an extensive discussion of this distinction, which also includes a concrete numerical example showing that immutable-belief incentive compatible mechanisms suggested in the literature tend to incentivize misreports in Bayesian contexts. In contrast to previously studied competitive forecasting settings, most notably those by Lichtendahl and Winkler (2007) and Lichtendahl et al. (2013), we do not require $\boldsymbol{p}_i$ to come from any particular parametric belief model. Moreover, and crucially, we do not restrict our analysis

to Bayes' Nash equilibrium play. Instead, and in line with the literature on (single-forecaster) proper scoring rules (e.g., Gneiting and Raftery 2007), the mechanisms we design obtain strict incentive compatibility in dominant strategies. That is, our objective is to provide strict incentives for truthful reports independent of the reports of other forecasters.

Also 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 2018, Hossain and Okui 2013). Although we also reward forecasters probabilistically (and obtain robustness to unknown risk preferences as a bonus[4]), 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 lotteries in this context of competitive forecasters.

3. Forecasting Competitions Using Standard Proper Scoring Rules

Consider a single forecaster and a single event X. A *scoring rule* R computes a payment that depends on the materialized event outcome x and the forecaster's report $y \in [0,1]$ regarding the probability that $X = 1$, paying the forecaster some amount $R(y,x)$.

Definition 3 (Scoring Rules). A *scoring rule* R is a mapping from reports $y \in [0,1]$ and outcomes $x \in \{0,1\}$ to scores $R(y,x) \in \mathbb{R} \cup \{-\infty\}$. A *scoring rule* R is *(strictly) proper* if, for all $y,p \in [0,1]$ with $y \neq p$, it holds that $\mathbf{E}_{X\sim p}[R(p,X)] > \mathbf{E}_{X\sim p}[R(y,X)]$. R is *bounded* if there exist $\underline{R},\overline{R} \in \mathbb{R}$ such that $R(y,x) \in [\underline{R},\overline{R}]$ for all $y \in [0,1]$, $x \in \{0,1\}$. Proper scoring rule R is *normalized* if it is bounded between zero and one, and if $R(0,0) = R(1,1) = 1$ and $R(y,x) = 0$ for some $y \in [0,1]$ and $x \in \{0,1\}$.

When clear from context, we will write $R \in [0,1]$ to refer to a scoring rule bounded between zero and one. There exist infinitely many proper scoring rules because any (strictly) convex function yields a (strictly) proper scoring rule (Gneiting and Raftery 2007, theorem 1). A widely used bounded scoring rule is the *quadratic scoring rule* (Brier 1950), which we will regularly refer to throughout the paper and give here in its standard, normalized form.

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

Bounded proper scoring rules used in practice are often already normalized. For example, both the quadratic scoring rule and the spherical rule (e.g., Jose 2009) already are. We note that any bounded strictly proper scoring rule R can be transformed into a normalized proper scoring rule $\tilde{R}$ and refer the reader to Appendix B for details.

3.1. Mechanism

A natural way to extend a strictly proper scoring rule R to a forecasting competition 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 select top forecasters, including by the Good Judgment Project (Tetlock and Gardner 2015) and FiveThirtyEight's NFL Forecasting Game.[5] Let M_{PSR^R} denote the mechanism derived in this way from proper scoring rule R. That is, M_{PSR^R} selects the forecasters with highest score,

$$M_{\mathrm{PSR}^R}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{x}) \in \arg\max_{i\in[n]} \sum_{k=1}^{m} R(y_{i,k},x_k),$$

with ties broken by forecaster index.[6]

3.2. Incentive Analysis

It is well known that selecting a forecaster according to highest proper scoring rule score may produce perverse incentives. In general, forecasters are incentivized to make overconfident reports to increase their chance of being judged the best forecaster ex post for at least some outcomes. To see this, consider an event X and two forecasters who believe that X occurs with probability 0.8 and 0.9, respectively. If both report their beliefs truthfully, the forecaster who reports 0.8 achieves the highest score—and is therefore selected by the mechanism—whenever $X = 0$, which she believes to occur with probability 0.2. However, if she instead reports some $y > 0.9$, she is selected by the mechanism whenever $X = 1$, which she believes to occur with probability 0.8. We present a more general example illustrating the failure of incentive compatibility of proper scoring rule selection for any $n \geq 2$ and $m \geq 1$ in Appendix C. For a thorough analysis of the (nontruthful) strategic behavior of competitive forecasters when ranked by standard proper scoring rules, we defer to Lichtendahl and Winkler (2007). Moreover, as shown by Theorem 1, failure to provide strict incentive compatibility is inherent to any deterministic forecasting competition mechanism. For intuition, the proof proceeds by showing that any deterministic mechanism only has finitely many possible outputs, whereas each agent has an infinite reporting space, and hence, forecasters cannot always strictly prefer truthful reporting.

Theorem 1. *No deterministic forecasting competition mechanism is strictly incentive compatible.*

4. Incentive-Compatible Forecasting Competitions

Theorem 1 motivates the study of randomized forecasting competition mechanisms. In Section 4.1, to build intuition, we begin by considering the single-event setting ($m = 1$) and introduce ELF, a strictly incentive-compatible forecasting competition mechanism. In Section 4.2, we then show how to extend ELF to handle multiple, arbitrarily correlated events.

What needs to hold for a forecasting competition to be strictly incentive compatible? First note that strict incentive compatibility requires that, for any beliefs over outcomes X and reports Y_{-i}, the probability f_i of selecting forecaster i must behave like a strictly 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 strictly proper scoring rules for each forecaster that are nonnegative and always sum to one so that they form a valid probability distribution. A natural first attempt to achieve this would be to use any strictly proper scoring rule, such as the quadratic scoring rule R_q, and "normalize" by dividing by the sum of all forecasters' scores. However, such a *multiplicative* normalization violates incentive compatibility because the factor by which scores are normalized is 1/(sum of forecasters' scores), which may differ between outcomes, causing forecasters to bias their predictions toward less likely outcomes. For an example illustrating this phenomenon, see Appendix E.

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 incentive compatibility is preserved when adding or subtracting a function of other reports and the outcome. Using their mechanism, each forecaster's payment is her score according to a proper scoring rule minus the average score of all other forecasters. ELF uses a similar idea to normalize all forecasters' scores *additively*, so that they are nonnegative and sum up to one. ELF then runs a lottery based on these scores to determine the winner of the prize.

4.1. Single-Event Mechanism

For a single event, *ELF*, more formally $M_{\mathrm{ELF}^R}(y_1, \ldots, y_n, x)$ selects forecaster $i \in [n]$ with probability

$$f_i(y_1, \ldots, y_n, x) = \frac{1}{n} + \frac{1}{n}\left(R(y_i, x) - \frac{1}{n-1}\sum_{j \neq i} R(y_j, x)\right), \quad (1)$$

where $R \in [0, 1]$ is a bounded strictly proper scoring rule.[7]

One can think of ELF as giving each forecaster a $1/n$ probability to start with, adjusting this up or down depending on how their performance compares to that of other forecasters. It is easy to see that the vector[8] $(f_1, \ldots, f_n)$ is a valid probability distribution: that each f_i is nonnegative follows immediately from R being bounded in $[0, 1]$, and $\sum_{i=1}^n f_i = 1$ because

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

Generalizing the result of Kilgour and Gerchak (2004) to incorporate Bayesian reasoning about other forecasters, we can show that ELF is incentive compatible.

Theorem 2. *ELF is strictly incentive compatible for $m = 1$.*

4.2. Multiple-Event Mechanism

We now consider a natural generalization of single-event ELF to multiple events. For multiple events, ELF proceeds as follows after all events have materialized. $M_{\mathrm{ELF}^R}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{x})$ selects forecaster $i \in [n]$ with probability

$$g_i(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^m f_{i,k},$$
$$\text{where } f_{i,k} = \frac{1}{n} + \frac{1}{n}\left(R(y_{i,k}, x_k) - \frac{1}{n-1}\sum_{j \neq i} R(y_{j,k}, x_k)\right), \quad (2)$$

and where $R \in [0, 1]$ is a bounded strictly proper scoring rule.

This corresponds to running single-event ELF for every event and selecting each forecaster with probability equal to the average probability assigned to her across all events. Note that this procedure can equivalently be interpreted as sampling a single event uniformly at random, and awarding the prize to the forecaster selected by single-event ELF on that event. Strict incentive compatibility of ELF then follows directly from strict incentive compatibility of single-event ELF.

Theorem 3. *ELF is strictly incentive compatible for $m \geq 1$ events.*

5. Incentive-Compatible and Accurate Forecasting Competitions

The ELF mechanism from Section 4.2 is strictly incentive compatible for arbitrarily correlated events. If (strict) incentive compatibility is the only objective, ELF provides a definitive solution. In many settings, however, the system designer strives for an additional objective, namely that the prize is awarded to the most accurate forecaster. In the Good Judgment Project, for example, the 2% of forecasters with highest quadratic scores were awarded "superforecaster" status (Tetlock and Gardner 2015). It is implicit in the

term that these individuals should be the most accurate forecasters. Similarly, recruiters on Kaggle seek to make job offers to the data scientists who create the most accurate models (Harris 2013). Hence, in addition to incentive compatibility, 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 tending to one as the number of events 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.

In judging accuracy, one needs to have a model for ground truth. Here, we borrow from statistical learning theory and assume that event outcomes are drawn from an unknown joint probability distribution θ over $X_1, \ldots, X_m$. We emphasize that θ is latent and hence never observed by either the forecasters or the mechanism. The marginal probability that event k will occur is denoted by $\theta_k \in [0, 1]$. This is strictly more general than defining outcomes as ground truth since, in particular, it allows for $\theta_k = x_k$. In Definition 3, proper scoring rules are defined in an incentive spirit, as a tool for the incentive-compatible elicitation of subjective beliefs. In particular, 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_k of event k occurring, then (strict) properness implies that reporting the true probability obtains a higher expected score than any other report. Reports that do not coincide with the true probability lead to lower expected scores, and different proper scoring rules correspond to different accuracy measures in that they punish reports diverging from the true probability differently. For example, with true probability θ_k, the quadratic scoring rule (Proposition 1) punishes a report y by $\mathbf{E}_{X_k \sim \theta_k}[R_q(\theta_k, X_k) - R_q(y, X_k)] = (y - \theta_k)^2$.

Importantly, the choice of proper scoring rule has implications for the relative rank of forecasters. For example, let $\theta_k = 0.7$ and let two forecasters report $y_{1,k} = 0.9$ and $y_{2,k} = 0.51$, respectively. Then, under the quadratic scoring rule, forecaster 2 obtains a higher expected score than forecaster 1 (less punishment), whereas under the spherical scoring rule,[9] forecaster 1 obtains a higher expected score than forecaster 2. That is, the system designer's choice of proper scoring rule in a forecasting competition determines the (relative) accuracy measure that forecasters are judged by.[10] For the incentive-compatible mechanisms in this paper, the proper scoring rules need to be bounded. In particular, the accuracy measure implied by the unbounded logarithmic scoring rule (Good 1952) cannot be used. Note that this restriction to bounded scoring rules (such as the quadratic or spherical scoring rule) is also present outside of competition settings when forecasters are simply paid their score as one cannot ensure nonnegative payments for unbounded scoring rules. Moreover, we will later show in Theorem 5 that no other incentive-compatible forecasting competition mechanism can implement accuracy measures corresponding to unbounded scoring rules under mild technical assumptions. Hence, for the remainder of the paper (with the exception of Theorem 5), the accuracy measure that is used will be given by a particular bounded proper scoring rule. The objective will be to select the forecaster with highest expected score according to that scoring rule while ensuring that the mechanism is strictly incentive compatible even in the competition setting. For this, it is helpful to overload notation of proper scoring rule R and define

$$R(\boldsymbol{y}_i, \theta) := \mathop{\mathbf{E}}_{\boldsymbol{X} \sim \theta} \frac{1}{m} \sum_{k=1}^{m} R(y_{i,k}, X_k)$$

as the expected score of report $\boldsymbol{y}_i$ using R and given joint probability θ. This allows us to make statements about the relative accuracy of forecasters with respect to R and θ. In particular, forecaster i is more accurate than forecaster j on the $m \geq 1$ events if and only if $R(\boldsymbol{y}_i, \theta) > R(\boldsymbol{y}_j, \theta)$.

5.1. Accuracy of ELF

We first observe that ELF selects forecasters with higher accuracy more often than those with lower accuracy.

Definition 4. Forecasting competition mechanism M is *rank accurate* with respect to proper scoring rule R if and only if it holds that $R(\boldsymbol{y}_i, \theta) > R(\boldsymbol{y}_j, \theta) \Longleftrightarrow \Pr_{\boldsymbol{X} \sim \theta}(M(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{X}) = i) > \Pr_{\boldsymbol{X} \sim \theta}(M(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{X}) = j)$ for all joint distributions θ over $X_1, \ldots, X_m$, all $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in [0,1]^m$, and all $i, j \in [n]$.

The next statement follows immediately from taking expectation over $\boldsymbol{X}$ in Equation (2).

Proposition 2. *The probability that ELF selects forecaster i given joint probability θ is* $\Pr_{\boldsymbol{X} \sim \theta}(M_{\mathrm{ELF}^R}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{X}) = i) = \frac{1}{n} + \frac{1}{n}\left(R(\boldsymbol{y}_i, \theta) - \frac{1}{n-1}\sum_{j \neq i} R(\boldsymbol{y}_j, \theta)\right)$.

Corollary 1. *ELF is rank accurate with respect to scoring rule R used in its definition. In particular, it selects the most accurate forecaster with higher probability than any other forecaster.*

One may wonder if there exist incentive-compatible forecasting competition mechanisms that select the most accurate forecaster with higher probability than ELF. In Theorem 4, we rule out this possibility for the standard two-forecaster, single-event setting (e.g., Lichtendahl and Winkler 2007), subject to mild conditions on the form of the forecasting competition mechanism.

Definition 5. Forecasting competition mechanism M is *anonymous* if the selected forecaster does not depend on the identities of the forecasters. That is, M is anonymous if for any permutation σ of $[n]$, any forecaster i,

any reports $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$, and any outcome vector $\boldsymbol{x}$, it holds that $\Pr\left(M(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{x})=i\right)=\Pr\left(M(\boldsymbol{y}_{\sigma^{-1}(1)},\ldots,\boldsymbol{y}_{\sigma^{-1}(n)},\boldsymbol{x})=\sigma(i)\right)$.

To exploit existing characterization theorems of competitive scoring rules (Lambert et al. 2008), we restrict attention to *smooth* forecasting competition mechanisms in Theorem 4.

Definition 6. A forecasting competition mechanism M is *smooth* if the corresponding function that outputs a probability distribution over forecasters, $\Pr(M(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{x}))$, is twice continuously differentiable with respect to each $\boldsymbol{y}_i$.

Theorem 4 shows that if a strictly incentive-compatible mechanism M ever selects the more accurate forecaster from a single-event, two-forecaster competition with higher probability than ELF with normalized R, then M is not rank accurate with respect to R, i.e., there must exist another instance in which M selects the less accurate forecaster with higher probability than the more accurate one. Recall that we denote by $\tilde{R}$ the proper scoring rule that results from normalizing R as described in Appendix B.

Theorem 4. *Let M be a smooth and anonymous forecasting competition mechanism that is rank accurate with respect to R and for which there exist $y_1,y_2\in[0,1]$ and distribution θ such that $R(y_1,\theta)>R(y_2,\theta)$ and $\Pr_{X\sim\theta}(M(y_1,y_2,X)=1)>\Pr_{X\sim\theta}(M_{ELF^{\tilde{R}}}(y_1,y_2,X)=1)$. Then M is not strictly incentive compatible.*

By adapting elements of the proof of Theorem 4, we obtain an impossibility result for unbounded scoring rules.

Theorem 5. *Let R be an unbounded scoring rule. No smooth, anonymous, and strictly incentive compatible forecasting competition mechanism is rank accurate with respect to R.*

A notable consequence of Theorem 5 concerns the logarithmic scoring rule, which is the proper scoring rule most grounded in classical information theory (e.g., Gneiting and Raftery 2007, section 2.2). In particular, the theorem implies that no incentive compatible forecasting competition mechanism is rank accurate with respect to the logarithmic rule.

5.2. Accuracy in the Limit

Theorem 4 shows that we cannot do better than ELF for the standard single-event, two-forecaster setting in terms of maximizing the probability of selecting the most accurate forecaster. However, what if there is more than just a single event? Let $\Delta:=\min_{j\neq i}(\max_i R(\boldsymbol{y}_i,\theta)-R(\boldsymbol{y}_j,\theta))$ denote the difference between the expected scores of the most accurate forecaster and the second-most accurate forecaster. Ideally, one would like to guarantee that for any "accuracy gap" Δ and any probability π arbitrarily close to one, there exists some minimal number of events after which it is guaranteed that the forecasting competition mechanism selects the most accurate forecaster with probability at least π. This intuition is formally captured in the definition of *limit accuracy*.

Definition 7. Forecasting competition mechanism M is *limit accurate* with respect to proper scoring rule R and set of joint distributions Θ if and only if, for any n, any $\underline{\Delta}>0$, and any $\pi\in[0,1)$, there exists a finite number of events $\underline{m}\in\mathbb{N}$ such that for all joint distributions $\theta\in\Theta$ and all $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\in[0,1]^m$ with $m\geq\underline{m}$ and $\Delta>\underline{\Delta}$, it holds that $\Pr_{X\sim\theta}\left(M(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{X})=\arg\max_i R(\boldsymbol{y}_i,\theta)\right)>\pi$.

Proposition 3 shows that some restriction on θ is necessary as limit accuracy cannot be achieved for all joint distributions. In particular, consider the extreme case in which events are "identical copies" of one another, such that whenever $X_1=1$, it holds that $X_k=1$ for all $k\in\{2,\ldots,m\}$ and whenever $X_1=0$, we have $X_k=0$ with $k\in\{2,\ldots,m\}$. In that case, all information contained in events $2,\ldots,m$ is already contained in the first event, and so increasing m is not helpful for identifying the most accurate forecaster.

Proposition 3. *No forecasting competition mechanism is limit accurate for all distributions θ over $X_1,\ldots,X_m$.*

In the remainder of this section, we design a forecasting competition mechanism that is limit accurate when the events are independent and strictly incentive compatible when this independence is also reflected in the uncertainty about others' reports. The restriction on forecasters' beliefs is referred to as *belief independence*.

Definition 8. For joint distribution D over outcomes $\boldsymbol{X}$ and reports $\boldsymbol{Y}_{-i}$, let D_k be the corresponding joint distribution over outcome X_k and reports $\boldsymbol{Y}_{-i,k}$. D is *belief independent* if and only if all D_k for $k\in[m]$ are independent.

Note that under belief independence, forecaster i can still believe that other forecasters are more accurate than herself and also that others' reports are more accurate on some events than others.

Definition 9. Forecasting competition mechanism M is *strictly incentive compatible under belief independence* if and only if for all forecasters $i\in[n]$, all belief vectors $\boldsymbol{p}_i$, all belief independent joint distributions D over outcomes $\boldsymbol{X}$ and reports $\boldsymbol{Y}_{-i}$ such that $\mathbf{E}_{X\sim D}[\boldsymbol{X}]=\boldsymbol{p}_i$, and all alternative report vectors $\boldsymbol{y}_i'\neq\boldsymbol{p}_i$,

$$\Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left(M(\boldsymbol{Y}_1,\ldots,\boldsymbol{p}_i,\ldots,\boldsymbol{Y}_n,\boldsymbol{X})=i\right)>\Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left(M(\boldsymbol{Y}_1,\ldots,\boldsymbol{y}_i',\ldots,\boldsymbol{Y}_n,\boldsymbol{X})=i\right).$$

5.3. Incentive-Compatible and Limit-Accurate Mechanism for Independent Events

The Independent-Event Lotteries Forecasting Competition Mechanism (I-ELF), more formally $M_{\text{I-ELF}^R}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{x})$ is defined as follows:

1. For each event k, pick forecaster i to be the event winner $w_k \in [n]$ with probability

$$f_{i,k}(y_{1,k},\ldots,y_{n,k},x_k) = \frac{1}{n}+\frac{1}{n}\left(R(y_{i,k},x_k)-\frac{1}{n-1}\sum_{j\neq i}R(y_{j,k},x_k)\right),$$

where $R \in [0,1]$ if $m = 1$ and $R \in [0,1)$ if $m \geq 2$ is a bounded strictly proper scoring rule.[11]

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

In essence, I-ELF runs a single ELF lottery for each event and awards the prize to the forecaster who won the most lotteries.

Theorem 6. *I-ELF is strictly incentive compatible under belief independence for $m \geq 1$ events.*

Take the perspective of any forecaster $i \in [n]$ seeking to maximize the probability of being selected. The proof proceeds by showing that she can reason about each event independently because of belief independence and, in a second step, that increasing her probability of winning event k strictly increases her probability of winning overall.

To conclude this section, we show that I-ELF is limit accurate when events are independent.

Theorem 7. *For all θ such that event outcomes $X_1,\ldots,X_m$ are independent, I-ELF is limit accurate with respect to scoring rule R used in its definition.*

For intuition, note that more accurate forecasters have a higher probability of winning each event (by Proposition 2). Hence, by standard concentration inequality arguments, the most accurate forecaster wins the most events with high probability when events are independent and the number of events is large.

As an alternative to I-ELF, one could collapse the m binary random variables into a single categorical random variable with 2^m outcomes and apply ELF to the joint distribution implied by the forecasters' (marginal) reports. (As we discuss in Section 6.1, ELF readily extends to the categorical case.) The problem with this mechanism is that it is not limit accurate. In particular, it will not select the most accurate forecaster with probability higher than $2/n$. To see this, observe that, in Equation (1), the first term in the parentheses is at most one and the second term at least zero, resulting in at most $2/n$.

6. Discussion

In this section, we describe extensions to our model and discuss the practical implementation of our methods.

6.1. Categorical Outcomes

Thus far, we have restricted our analysis to events with binary outcomes. In practice, we are also interested in events with nonbinary (categorical) outcomes. Unsurprisingly, selecting the forecaster with highest average proper score (e.g., using the categorical generalization of the quadratic scoring rule of (Brier (1950)) inherits the violation of incentive compatibility exhibited in Section 3.

ELF readily extends to categorical outcomes. The competitive scoring rule of Kilgour and Gerchak (2004) is incentive compatible for categorical outcomes when used in conjunction with any proper multioutcome scoring rule, and ELF inherits this incentive compatibility for all such rules that are bounded. Under belief independence, incentive compatibility of I-ELF follows from the same arguments used in the proof of Theorem 6. Moreover, it still holds that more accurate forecasters obtain higher scores in expectation, so the most accurate forecaster still wins the most events in expectation. Hence, we can prove limit accuracy by a qualitatively identical argument to the one in the proof of Theorem 7.

6.2. Real-Valued Outcomes and Reports

In many business contexts, we are interested in forecasting events that take real-valued outcomes instead of categorical values. For instance, events could be the monthly demand of particular items, the cost of infrastructure projects, or the annual inflation rate. Both ELF and I-ELF readily extend to handle these cases. In contrast to events with categorical outcomes, where one typically seeks to elicit the forecaster's entire subjective probability distribution over the outcomes, this is cumbersome with infinitely many outcomes on the real line. Instead, practitioners typically choose to only elicit properties of the underlying distribution, such as the mean or the median, which summarize the underlying distribution in ways meaningful for the application at hand. There exist many proper scoring rules for the elicitation of these properties. For example, it is well known that the quadratic scoring rule $R_q(y,x) = 1-(y-x)^2$, which was introduced in Section 3, generalizes to real valued outcomes $x \in [0,1]$. More precisely, if random variable X denotes the real-valued outcome, the forecaster maximizes her expected score by reporting $y = \mathbf{E}[X]$, that is, her subjective estimate of the mean of X. Meanwhile, the absolute scoring rule $R_a(y,x) = 1-|y-x|$ is strictly proper when used to elicit subjective estimates of the median of X (e.g., Jose 2017). Note that these scoring rules can be scaled to incorporate any bounded interval $[a,b]$ with $b > a$. Moreover, although it is easy to obtain upper and lower bounds on the variable of interest for almost any conceivable application, tighter bounds yield better discrimination in score between more and less accurate reports.

Although the quadratic and absolute scoring rules are strictly proper when used as payments to elicit subjective estimates of the mean and median, respectively, misreporting remains an issue when they are naively applied to forecasting competitions. Consider random variable X commonly known to be uniformly distributed on $[0, 1]$. If $n = 3$ forecasters all report a subjective estimate of the mean, that is, $y_i = 0.5$ for all i, and the forecaster with highest quadratic score is selected as the prize winner, then each forecaster wins with probability $1/3$ (assuming ties are broken uniformly at random). However, if forecaster 1 instead reports $y_1 = 0.5 - \epsilon$ for some small ϵ, then she achieves the highest score whenever $X < 0.5 - \epsilon$, which occurs with probability $0.5 - \epsilon > 1/3$.[12] The same example continues to break incentive compatibility when the absolute scoring rule is used to elicit estimates of the median.

To overcome the issue of misreporting, we can define ELF and I-ELF as in Sections 4 and 5, just with an appropriately chosen scoring rule R that is strictly proper for the property being elicited. Strict incentive compatibility of ELF and I-ELF (under the belief independence restriction) follows by reasoning analogous to the binary case. The accuracy guarantee provided by I-ELF carries over as well, with the accuracy implied by the scoring rule R used to define the mechanism. As for the binary-outcome setting, both ELF and I-ELF work in conjunction with any bounded R. Observe that this is analogous to using proper scoring rules as payments, where R needs to be bounded to guarantee nonnegative payments.

6.3. 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. Similarly, many Kaggle competitions award prizes to the highest-ranked forecasters with prizes decreasing in value as the forecasters' ranks increase. Ranking forecasters in order of any proper score again inherits all of the problems described in Section 3.

I-ELF can be adapted to produce a ranking by simply ordering forecasters according to the number of events that the forecasters win. As long as forecasters strictly prefer higher positions in the ranking (e.g., because higher rankings correspond to higher-valued prizes), I-ELF remains strictly incentive compatible, because 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 proper 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.

6.4. Forecaster Hiring and Connections to Learning

Forecasting competitions are often used as a method of selecting a forecaster to hire when future predictions are needed. In this setting, the goal of the competition mechanism is to select 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 determining accuracy through the m events being predicted, we could instead assume a joint distribution D_θ 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 proper score of her truthful forecasts with respect to D_θ, that is, $\mathbf{E}_{p_i,\theta\sim D_\theta}[R(p_i, \theta)]$.

Under this model, mechanism M_{PSR^R} 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 score 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_{PSR^R} would be guaranteed to have accuracy arbitrarily close to that of the most accurate forecaster. However, the incentive issues remain. The advantage of I-ELF is that it obtains truthful reports for any m while achieving similar accuracy guarantees as m grows large. In this sense, I-ELF can be viewed as a mechanism for learning in the presence of strategic agents, where the objective is to select a forecaster that will perform well on future events.

6.5. Practical Implementation

In Section 2, we require that all forecasters report their predictions for all events before the first event materializes. With an appropriate generalization of the definition of incentive compatibility, this requirement can be relaxed without sacrificing the properties of ELF. In particular, when reporting on event k, we can allow forecasters to update joint distribution D conditioned on the outcomes of past events and the reports on these events. Our results continue to hold if incentive compatibility requires that forecasters truthfully report their updated beliefs.

For I-ELF, suppose that a forecaster reports on event k after some subset of the other events have materialized. Given belief independence, the reports of other forecasters on any other event, as well as the

corresponding outcomes for any events already materialized, do not lead to a belief update. Therefore, the competition organizer does not need to protect or withhold any information from the forecasters as long as the randomness involved in selecting event winners w_k from probabilities $f_{i,k}$ is not realized until all predictions have been reported.

More speculatively, one could imagine applying our techniques to other elicitation methods. In particular, prediction markets are often implemented using play money, with monetary prizes for top-ranked traders or simply high positions on public leaderboards used as incentive (e.g., Jia et al. 2017). Directly awarding prizes to participants with the highest play money account balances, however, leads to incentive problems analogous to those in the forecasting competition model we consider in this paper: maximizing the probability of having the highest account balance is not the same as maximizing expected account balance. Variations on this idea induce similar gaming incentives; for example, Chakraborty et al. (2013) award prizes uniformly at random among participants placed sufficiently high on the leaderboard. Although it is not clear how to directly translate (I-)ELF to this setting, it is easy to see that awarding a single prize randomly with probability proportional to account balance does lead to forecasters maximizing their expected account balance[13] (e.g., Cowgill and Zitzewitz 2015, section 1.2.2). Further exploring applications to prediction markets and other elicitation methods is a compelling direction for future work.

Note that both ELF and I-ELF are easy to implement. Indeed, even for very large competitions, both mechanisms can be implemented in standard spreadsheet software. Each value $f_{i,k}$ is computed by a simple formula, after which the only remaining step is to implement 1 or m lotteries for ELF and I-ELF, respectively.

7. Conclusion

In real-world forecasting settings, forecasters typically compete for a single prize. Motivated by the prevalence of these forecasting competitions and their poor incentive properties, we initiate the study of incentive-compatible forecasting competitions. Despite a rich literature on incentive-compatible forecast elicitation in the noncompetitive setting, the mechanisms in this work are the first to solve the incentive challenge in the competition setting. The forecasting competition mechanism most widely used in practice is to simply select the forecaster with highest score according to some proper scoring rule. Not only does this particular mechanism fail to elicit truthful forecasts, but, as we show, any deterministic forecasting competition mechanism must violate incentive compatibility. We therefore turn to randomized forecasting competitions, which can be thought of as rewarding forecasters with a lottery ticket that has a higher chance of winning the more accurate the forecaster was relative to the other forecasters in the competition. This intuitive principle is behind both mechanisms we design.

We first define the Event Lotteries Forecasting Competition Mechanism (ELF). Because of its randomized nature, ELF may not always select the most accurate forecaster, but it does select more accurate forecasters with higher probability than less accurate ones. For the special case of one event and two forecasters, we show that, under mild technical conditions, no incentive-compatible mechanism can select the most accurate forecaster with higher probability than ELF does.

Our second mechanism, I-ELF, is strictly incentive compatible when forecasters' beliefs satisfy belief independence, which, intuitively, requires that information about one event does not inform forecasters' beliefs about other events. I-ELF uses ELF as a building block, first selecting a winner for each event using ELF, and then selecting the competition winner as the forecaster who won the most individual events. In addition to being incentive compatible under belief independence, I-ELF also selects the most accurate forecaster with a probability that tends to one as the number of events grows.

Our results have significant implications for organizations that use groups of forecasters to inform managerial decision making under uncertainty. Previous studies on forecasters' competitive incentives encouraged the fostering of collaboration and cooperation to mitigate the distorted incentives at play (Lichtendahl and Winkler 2007). Our work yields a different perspective. By cleverly exploiting randomization, the decision maker can embrace competitive stakes when eliciting predictions without having to sacrifice the quality of the information received.

Acknowledgments

The authors thank Sebastian Ebert, Rafael Frongillo, Mirko Kremer, Jochen Schlapp, Bo Waggoner, and the anonymous reviewers for helpful feedback. This paper is a significantly extended version of Witkowski et al. (2018). This work was completed in part while R. Freeman and D. M. Pennock worked for Microsoft Research.

Appendix A. Generalizing Immutable-Belief Incentive Compatibility to Robust Incentive Compatibility

In this section, we are going to unpack how the (standard) immutable-belief model—while appropriate for the wagering setting, where different forecasters, by definition, agree to disagree and seek to bet on their individual convictions (Lambert et al. 2008)—is too limited for forecasting settings, where forecasters believe that other forecasters' reports contain information that they themselves do not already have. This includes but is not limited to the competition setting that is the focus of this paper. The section is organized as follows. First, we provide the definition of

immutable-belief incentive compatibility from Kilgour and Gerchak (2004) and Lambert et al. (2008) and show that it is a special case of the definition used in the main text of this paper. Second, we provide an example of a Bayesian belief model along the lines of standard models in the literature, which is incompatible with the assumption of immutable beliefs because forecasters update their beliefs about the outcome when learning the beliefs of other forecasters. Finally, using a particular numerical example, we demonstrate that an immutable-belief incentive compatible mechanism that has been suggested in the literature incentivizes misreports under this Bayesian model.

We emphasize here that neither ELF nor I-ELF assume that beliefs are formed using this particular model. We also emphasize that our robust incentive compatibility generalizes both Bayesian and immutable-belief models. In particular, truthful reporting is a dominant strategy in both ELF and I-ELF, regardless of whether agents would update their beliefs knowing the reports of other agents or not. Hence, this Bayesian model is given here for illustrative purposes only, showing that mechanisms that are incentive compatible only for immutable beliefs are not sufficient when forecasters believe that other forecasters' reports contain information that they themselves do not already have. We now state the incentive compatibility definition, applied to the competition setting, that was used in the work of Kilgour and Gerchak (2004) and Lambert et al. (2008).

Definition A.1. Forecasting competition mechanism M is *strictly incentive compatible for immutable beliefs* if and only if for all forecasters $i \in [n]$, all belief vectors $\boldsymbol{p}_i$, all others' reports $\boldsymbol{y}_{-i}$, and all alternative report vectors $\boldsymbol{y}'_i \neq \boldsymbol{p}_i$, $\Pr_{\boldsymbol{X}\sim p_i}(M(\boldsymbol{y}_1,\ldots,\boldsymbol{p}_i,\ldots,\boldsymbol{y}_n,\boldsymbol{X})=i) > \Pr_{\boldsymbol{X}\sim p_i}(M(\boldsymbol{y}_1,\ldots,\boldsymbol{y}'_i,\ldots,\boldsymbol{y}_n,\boldsymbol{X})=i)$.

Observe that Definition A.1 coincides with the (robust) incentive compatibility definition used in this work (Definition 2) when joint distribution D is restricted such that $\boldsymbol{y}_{-i}$ only takes a single value, regardless of the realization of X. Hence, every mechanism that is robust incentive compatible (Definition 2) is also incentive compatible for immutable beliefs (Definition A.1). However, the reverse is not true. For intuition as to how robust incentive compatibility is different from immutable-belief incentive compatibility and to understand why one wants forecasting competition mechanisms to satisfy the stronger robust incentive compatibility, ignore for a moment that, in competitions, "payments" (selection probabilities) need to add up to one. Consider then a forecaster i who is paid $y_j \cdot R_q(y_i, x)$, where y_j is the report of another forecaster $j \neq i$. In the immutable-belief model, y_j is assumed to be a constant from forecaster i's perspective, so that she should report truthfully because linear transformations of proper scoring rules preserve properness. However, if forecaster i believes j's report to be correlated with outcome X, then j's report Y_j is in fact a random variable and not a constant. This typically leads to misreports. In the extreme case, if forecaster i believes that forecaster j reports all probability mass on the eventually materialized outcome, that is, $Y_j = X$, then, if $X = 0$, she receives payment 0, and if $X = 1$, she receives $R_q(y_i, 1)$. Thus, forecaster i strategizes by conditioning on $X = 1$, maximizing her payment by reporting $y_i = 1$ regardless of her true belief. As we will see later in this section, this intuition also applies to competition settings, including settings where forecaster i believes that she is more accurate than all other forecasters.

In contrast to immutable-belief incentive compatibility, robust incentive compatibility does guarantee truthful reporting incentives even in settings in which forecaster i would update her belief on learning forecaster j's report. Note that such conditional belief updating is implied by standard Bayesian models in the forecasting literature where individual forecasters' beliefs stem from noisy observations of some ground truth (e.g., Lichtendahl and Winkler 2007, Lichtendahl et al. 2013, Palley and Soll 2019). To make this concrete, consider the following simple Bayesian model along those lines. (An example of this model is depicted in Figure A.1; multiple-event models can be defined analogously.) The event outcome is given by random variable X, which takes values in $\{0, 1\}$. All n forecasters share a common prior $\Pr(X = 1)$ that the event outcome is 1 (e.g., a commonly known base rate). Before the event outcome materializes, each forecaster $i \in [n]$ observes a binary, noisy signal S_i, taking values in $\{l, h\}$. Each forecaster is of one of two types: "expert" types have a noise level (error rate) of ϵ_e and forecasters of "rookie" types have a noise level of ϵ_r with $0.5 > \epsilon_r > \epsilon_e > 0$. If $X = 1$, the probability of observing h is $1-\epsilon$, and if $X = 0$, the probability of observing l is $1-\epsilon$, where the ϵ value depends on the forecaster type. The belief model as well as which forecaster is of which type is common knowledge. After observing her signal $S_i = s_i$, forecaster i updates her belief about X. Moreover, she also updates her (meta) beliefs about the beliefs of the other forecasters conditional on that X. The updated belief on the outcome is given by

$$\Pr(X = 1 \mid S = s) = \frac{\Pr(S = s \mid X = 1) \cdot \Pr(X = 1)}{\Pr(S = s)}, \tag{A.1}$$

where $\Pr(S = s \mid X = 1)$ depends on the forecaster's noise level as determined by her type and

$$\Pr(S = s) = \Pr(S = s \mid X = 1) \cdot \Pr(X = 1) + \Pr(S = s \mid X = 0) \cdot \Pr(X = 0).$$

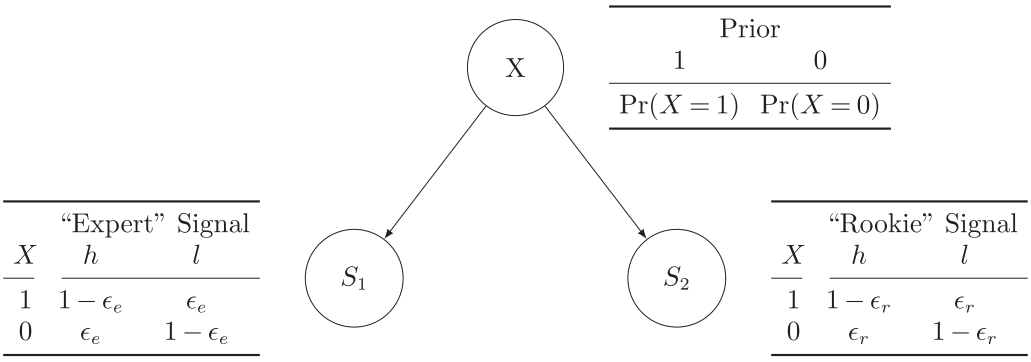
For the meta beliefs of forecaster i about the belief of forecaster j given X, first observe that forecaster j can only hold one of two possible beliefs, namely $\Pr(X = 1 \mid S_j)$ for $S_j = h$ and $S_j = l$, respectively. Which of these two beliefs forecaster j holds is thus determined by her signal, which itself is influenced by X. In particular, the expected value of forecaster j's belief given each possible instantiation of X is calculated by

$$\mathbf{E}[P_j \mid X = x] = \Pr(X = 1 \mid S_j = h) \cdot \Pr(S_j = h \mid X = x) + \Pr(X = 1 \mid S_j = l) \cdot \Pr(S_j = l \mid X = x), \tag{A.3}$$

where P_j denotes the random variable for forecaster j's belief p_j.

To see that immutable-belief incentive compatibility (Definition A.1) is inappropriate for this kind of Bayesian model from a technical perspective, it is sufficient to observe that truthful reports of the other forecasters (i.e.,

Figure A.1. Example of a Bayesian Forecaster Model with $n = 2$



beliefs) do indeed depend on the realization of X. In particular, the forecasters' beliefs are correlated with the outcome. Unfortunately, this observation is not just a technical nuisance but has immediate implications on mechanisms suggested in the literature. In the remainder of this section, we will show that immutable-belief incentive compatible mechanisms suggested in the literature do indeed lead to misreports in Bayesian models such as the one exemplified here.

More precisely, we consider a member of the *adaptive weighted score* mechanism family suggested in section 6.1 of Lambert et al. (2008), which we present here applied to the forecasting competition setting. This family of mechanisms is parameterized by a choice of scoring rule R. We emphasize that the particular choice we make is not an edge case. For simplicity, we present the example mechanism for $n = 4$. Intuitively, the mechanism repeatedly partitions the four forecasters into two groups A and $\overline{A}$ of two forecasters each and scores the forecasters in the first group "against" each other using a scheme similar to that of Kilgour and Gerchak (2004). The interesting part is that the proper scoring rule that is used to score forecasters in the first group is defined by the reports of the second group. As we will see, this mechanism is immutable-belief incentive compatible but leads to misreports in the Bayesian model we just introduced.

The mechanism proceeds as follows:

1. Given the set of four players $\{1,2,3,4\}$, we consider the set of forecaster groups of size 2, which we denote by $\mathcal{A}$. Furthermore, let $\mathcal{A}_i \subset \mathcal{A}$ denote the set of forecaster groups that contain forecaster i.

2. Let $R^{z,z'}(y,x) = \frac{z+z'+\varepsilon}{2+\varepsilon} \cdot R_q(y,x)$ with $\varepsilon > 0$ be a strictly proper scoring rule, whose form is parameterized by $z,z' \in [0,1]$. For any constants z,z', $R^{z,z'}$ is a (weakly) scaled-down version of the (normalized) quadratic scoring rule R_q. Further note that the role of ε is to ensure that the scaling factor is positive even if $z,z' = 0$, so that $R^{z,z'}$ remains strictly proper in that case. Observe that $R^{z,z'}$ is bounded between zero and one.

3. For a single event and $n = 4$, the *Adaptive-Score Forecasting Competition Mechanism* $M_{\mathrm{ASF}^{R^{z,z'}}}(y_1,\ldots,y_n,x)$ selects forecaster $i \in [n]$ with probability

$$f_i(y_1,\ldots,y_n,x) = \frac{1}{4} + \sum_{A \in \mathcal{A}_i} \frac{1}{12}\left(R^{z,z'}(y_i,x) - R^{z,z'}(y_j,x)\right),$$

where $j \in A$ refers to the other forecaster $j \neq i$ in each $A \in \mathcal{A}_i$, and z,z' are the two reports from forecaster group $\overline{A}$ not containing i, that is, $\overline{A} := [n] \backslash A$.

It is easy to see that $M_{\mathrm{ASF}^{R^{z,z'}}}$ is strictly incentive compatible for immutable beliefs (Lambert et al. 2008): if forecaster i believes that the other forecasters' reports y_j,z,z' are constants, which are uninformative about X, then, for each $A \in \mathcal{A}_i$, forecaster i believes that she is scored by a scaled-down R_q, and hence should report truthfully. Alternatively, one can think of the immutable-belief setup as though the reports of all forecasters are known beforehand, which is explicit in a wagering setting, where, by definition, forecasters agree to disagree. That is, there is no uncertainty about the reports of the other forecasters and hence also no uncertainty about the scoring rule that will be used. The only uncertainty that remains is about the outcome.

In the remainder of this section, we will show that despite $M_{\mathrm{ASF}^{R^{z,z'}}}$ being strictly incentive compatible for immutable beliefs, forecasters can have incentives to misreport in the Bayesian model described earlier in this section. It is important to emphasize that none of this is an edge case: other families of immutable-belief incentive compatible mechanisms, other choices of scoring rules for this family, and other numbers for this particular choice of scoring rule would also lead to misreporting incentives in a Bayesian context.

The numerical example setting we consider has $n = 4$ forecasters and a uniform prior of $\Pr(X = 1) = \Pr(X = 0) = 0.5$. Furthermore, forecaster 1 is of the expert type with $\epsilon_e = 0.2$, and forecasters 2, 3, and 4 are of the rookie type with $\epsilon_r = 0.3$. For scoring rule $R^{z,z'}$, we use $\varepsilon = 0.1$.

We now consider the situation of forecaster 1 and show that she has an incentive to misreport in the special case of all other forecasters reporting truthfully, that is, $y_j = p_j$ for all $j \neq 1$. (Remember that both definitions of incentive compatibility are with respect to dominant strategies, which require that truthful reporting is maximizing forecaster 1's selection probability for *any* reports of the other forecasters.) Forecaster 1's (expected) selection probability is

$$\mathbf{E}[f_i(Y_1,\ldots,y_i,\ldots,Y_n,X)] = \frac{1}{4} + \sum_{A \in \mathcal{A}_i} \frac{1}{12}\mathbf{E}\left[R^{Z,Z'}(y_i,X) - R^{Z,Z'}(Y_j,X)\right], \tag{A.4}$$

where the expectation is taken over the randomness of the Bayesian model. $\mathcal{A}_1$ contains forecaster groups {1, 2}, {1, 3}, and {1, 4}. Since forecasters 2, 3, and 4 are of the rookie type, forecaster j from $A \in \mathcal{A}_1$ is always a rookie and reports z and z' from $\overline{A}$ are also from rookies. Hence, forecaster i's expected score for forecaster groups {1, 2}, {1, 3}, and {1, 4} are the same, and so we first consider only forecaster group $A = \{1,2\}$ and later multiply the expected score for that group by three.

Rookie types have one of two possible beliefs about the outcome, depending on which signal they observed (Equation (A.1)). For forecaster 2, this is either $\Pr(X = 1 \mid S_2 = h) = \frac{\Pr(S_2=h \mid X=1)\cdot\Pr(X=1)}{\Pr(S_2=h)} = \frac{0.7\cdot 0.5}{0.5} = 0.7$ or $\Pr(X = 1 \mid S_2 = l) = 0.3$. Since $\overline{A} = \{3,4\}$ also contains only rookies, their possible beliefs are the same as for forecaster 2. Thus, scoring rule $R^{z,z'}$ has three possible scaling factors $\frac{z+z'+\varepsilon}{2+\varepsilon}$ for R_q, which depend on the reports of the forecasters in $\overline{A} = \{3,4\}$, namely $\frac{0.7+0.7+0.1}{2+0.1} = \frac{5}{7}$, $\frac{0.7+0.3+0.1}{2+0.1} = \frac{11}{21}$, and $\frac{0.3+0.3+0.1}{2+0.1} = \frac{1}{3}$. Using notation $S_{\overline{A}}$ to denote the signals observed by the forecasters in $\overline{A}$, the probabilities for the first and second scaling given $X = x$ can, because of conditional independence of S_3 and S_4, be calculated by (the third is calculated analogously to the first)

$$\Pr(S_{\overline{A}} = \{h,h\} \mid X = x) = \Pr(S_3 = h \mid X = x) \cdot \Pr(S_4 = h \mid X = x)$$

and

$$\Pr(S_{\overline{A}} = \{l,h\} \mid X = x) = \Pr(S_3 = h \mid X = x) \cdot \Pr(S_4 = l \mid X = x) + \Pr(S_3 = l \mid X = x) \cdot \Pr(S_4 = h \mid X = x).$$

This results in $\Pr(S_{\overline{A}} = \{h,h\} \mid X = 1) = \Pr(S_{\overline{A}} = \{l,l\} \mid X = 0) = 0.49$, $\Pr(S_{\overline{A}} = \{l,h\} \mid X = 1) = \Pr(S_{\overline{A}} = \{l,h\} \mid X = 0) = 0.42$, and $\Pr(S_{\overline{A}} = \{l,l\} \mid X = 1) = \Pr(S_{\overline{A}} = \{h,h\} \mid X = 0) = 0.09$. With this, forecaster 1 can now reason about the scoring rule she expects for each event outcome. If $X = 1$, forecaster 1 expects to be scored by scoring rule $(0.49 \cdot \frac{5}{7} + 0.42 \cdot \frac{11}{21} + 0.09 \cdot \frac{1}{3})\, R_q(y_1, 1) = 0.6 \cdot R_q(y_1, 1)$. Analogously, if $X = 0$, she expects scoring rule $(0.09 \cdot \frac{5}{7} + 0.42 \cdot \frac{11}{21} + 0.49 \cdot \frac{1}{3})\, R_q(y_1, 0) = \frac{47}{105} \cdot R_q(y_1, 0) = 0.448 \cdot R_q(y_1, 0)$.

Forecaster 1's belief about forecaster 2's report given X is calculated by Equation (A.3) and results in $\mathbf{E}[Y_2 \mid X = 1] = 0.7 \cdot 0.7 + 0.3 \cdot 0.3 = 0.58$ and $\mathbf{E}[Y_2 \mid X = 0] = 0.7 \cdot 0.3 + 0.3 \cdot 0.7 = 0.42$. If $X = 1$, the expectation in the right-hand side of Equation (A.4) for $A = \{1,2\}$ is then

$$\begin{aligned}&\mathbf{E}\left[R^{Y_3,Y_4}(y_1, 1) - R^{Y_3,Y_4}(Y_2, 1) \mid X = 1\right] \\ &= 0.6 \cdot \left(1 - (y_1 - 1)^2 - (1 - (0.58 - 1)^2)\right) \\ &= 0.106 - 0.6\,(y_1 - 1)^2,\end{aligned}$$

where the expectation is again taken over the randomness of the Bayesian model. Analogously, if $X = 0$, her expectation for that part is

$$\mathbf{E}\left[R^{Y_3,Y_4}(y_1, 0) - R^{Y_3,Y_4}(Y_2, 0) \mid X = 0\right]$$

$$= 0.448 \cdot (1 - y_1^2 - (1 - 0.42^2)) = 0.079 - 0.448\, y_1^2.$$

Observe that in each outcome, forecaster 1 is scored using a positive-affine transformed R_q. Crucially however, the scaling factor is higher for $X = 1$ than for $X = 0$. As we will see, this has the effect that forecaster 1 has an incentive to shift her report toward the $X = 1$ outcome as it carries more weight. To obtain forecaster 1's overall expected scores for each X, we multiply the expected scores for $A = \{1,2\}$ by three (to account for the symmetric cases of $A = \{1,3\}$ and $A = \{1,4\}$), divide the result by 12 (resulting in a division by four), and add $\frac{1}{4}$.

To complete the example, suppose that forecaster 1 observes $S_1 = h$. Using Equation (A.1), she updates her belief about the outcome to $\Pr(X = 1 \mid S_1 = h) = 0.8$. Putting this all together, forecaster 1's expected score reporting y_1 is

$$\begin{aligned}\mathbf{E}[f_i(Y_1, \ldots, y_i, \ldots, Y_n, X)] = \frac{1}{4} + \frac{1}{4}\Big(&0.8 \cdot (0.106 - 0.6\,(y_1 - 1)^2) \\ &+ 0.2 \cdot (0.079 - 0.448\, y_1^2)\Big),\end{aligned}$$

which is uniquely maximized for $y_1 = \frac{75}{89} = 0.843$. Forecaster 1 thus has an incentive to misreport her true belief of 0.8. It is important to note here that, although the exact calculations are rather extensive, forecasters in this Bayesian setting faced with this mechanism do not need to compute their conditional beliefs precisely but can simply make a report that is slightly higher than their belief.

We emphasize that this example also shows that even if a forecaster believes that she is the most accurate forecaster, she may still have an incentive to misreport under immutable-belief incentive compatibility. The key advantage of robust incentive compatibility over immutable-belief incentive compatibility is that it allows for the possibility that forecasters may believe that other forecasters' reports contain some information they do not already have. Or, phrased differently, in contrast to immutable-belief incentive compatibility, robust incentive compatibility allows for the possibility that forecasters would update their beliefs upon learning the reports of other forecasters.

Appendix B. Procedure to Normalize a Bounded Strictly Proper Scoring Rule

Let R be a bounded strictly proper scoring rule with $\underline{R} = \min_{y,x} R(y,x)$ and $\overline{R} = \max_{y,x} R(y,x)$ for $y \in [0,1]$, $x \in \{0,1\}$. Then R can be transformed into a normalized proper scoring rule $\tilde{R}$ as follows. As an intermediate step, define $R'(y,x) = R(y,x) + \beta'(x)$ with $\beta'(0) = -R(0,0)$ and $\beta'(1) = -R(1,1)$. Since R is strictly proper, so is R', and both the maximum and the minimum must be taken for $y \in \{0,1\}$. In particular, it must hold that both $0 = R'(0,0) > R'(1,0)$ and $0 = R'(1,1) > R'(0,1)$. Let $r_0 := R'(0,0) - R'(1,0)$ and $r_1 := R'(1,1) - R'(0,1)$ be the intervals ("ranges") of R' for $X = 0$ and $X = 1$, respectively. Then $\tilde{R}(y,x) := \frac{1}{\max(r_0, r_1)} R'(y,x) + 1$ is a normalized scoring rule.

Appendix C. Proper Scoring Rule Selection Violates Incentive Compatibility

Let R be any strictly proper scoring rule. Consider an instance with $m \geq 1$, and $n \geq 2$. Suppose that $\boldsymbol{p}_i = (0.5, \ldots, 0.5, 0.8)$[14] and consider joint distribution D over $\boldsymbol{X}$ and $\boldsymbol{Y}_{-i}$ defined as follows.

- With probability 0.4, $\boldsymbol{X} = (0, \ldots, 0, 1)$ and $\boldsymbol{Y}_j = (0.5, \ldots, 0.5, 0.8 + \frac{j}{10n})$ for all $j \neq i$.

- With probability 0.4, $\boldsymbol{X} = (1, \ldots, 1, 1)$ and $\boldsymbol{Y}_j = (0.5, \ldots, 0.5, 0.8 + \frac{j}{10n})$ for all $j \neq i$.
- With probability 0.1, $\boldsymbol{X} = (0, \ldots, 0, 0)$ and $\boldsymbol{Y}_j = (0.5, \ldots, 0.5, 0.8 + \frac{j}{10n})$ for all $j \neq i$.
- With probability 0.1, $\boldsymbol{X} = (1, \ldots, 1, 0)$ and $\boldsymbol{Y}_j = (0.5, \ldots, 0.5, 0.8 + \frac{j}{10n})$ for all $j \neq i$.

Note in particular that $\mathbf{E}_{\boldsymbol{X}\sim D}[\boldsymbol{X}] = \boldsymbol{p}_i$, and that $0.8 < Y_{j,m} \leq 0.9$ with probability one for all $j \neq i$.

If forecaster i reports $\boldsymbol{p}_i$, then all forecasters receive the same score on all events except event m. Forecaster i receives the highest score, and is therefore selected by M_{PSR^R}, whenever $X_m = 0$, which occurs with probability 0.2. That is, $\Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}(M_{\mathrm{PSR}^R}(\boldsymbol{Y}_1, \ldots, \boldsymbol{p}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i) = 0.2$. However, if forecaster i reports $\boldsymbol{y}_i' = (0.5, \ldots, 0.5, 1)$, then she is selected by M_{PSR^R} whenever $X_m = 1$, which occurs with probability 0.8. That is, $\Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}(M_{\mathrm{PSR}^R}(\boldsymbol{Y}_1, \ldots, \boldsymbol{y}_i', \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i) = 0.8 > 0.2 = \Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}(M_{\mathrm{PSR}^R}(\boldsymbol{Y}_1, \ldots, \boldsymbol{p}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i)$, violating incentive compatibility.

Appendix D. Proof of Theorem 1

Proof. Let M be a deterministic and strictly incentive compatible forecasting competition mechanism. Furthermore, let $m \geq 1$ and $n \geq 2$, and observe that there are $|\mathcal{P}([m])| = 2^m$ possible values of the outcome vector $\boldsymbol{x}$. Consider forecaster i, and suppose that every forecaster $j \neq i$ reports a probability $y_{j,k} = 0.5$ for every event k. We first use these fixed reports of agents $j \neq i$ to derive candidate misreports for agent i, and then again to define an appropriate joint distribution D that yields a violation of strict incentive compatibility.

For any report $\boldsymbol{y}_i$, forecaster i is selected as the winner for some subset of possible event outcomes $\mathcal{X} \subseteq \{0,1\}^m$. Since there are 2^m possible values of $\boldsymbol{x}$, there are $|\mathcal{P}(\{0,1\}^m)| = 2^{2^m}$ possible subsets $\mathcal{X}$. Consider then $2^{2^m} + 1$ different possible reports of forecaster i, denoted $\boldsymbol{y}_i^0, \boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{2^{2^m}}$, and the corresponding subsets $\mathcal{X}^0, \mathcal{X}^1, \ldots, \mathcal{X}^{2^{2^m}}$ of event outcomes for which she is selected given these reports. By the pigeonhole principle, there must exist $r, s \in \{0, \ldots, 2^{2^m}\}$ with $r \neq s$ such that $\mathcal{X}^r = \mathcal{X}^s$. That is, forecaster i is selected for exactly the same set of possible event outcomes regardless of whether she reports $\boldsymbol{y}_i^r$ or $\boldsymbol{y}_i^s$.

We use this fact to illustrate a violation of strict incentive compatibility. Define D as follows: each event k occurs with probability equal to $y_{i,k}^r$ independent of other events, and every forecaster $j \neq i$ reports a probability of 0.5 for every event. Note that $\boldsymbol{p}_i = \boldsymbol{y}_i^r$. Then we have that $\Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}(M(\boldsymbol{Y}_1, \ldots, \boldsymbol{p}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i) = \Pr_{\boldsymbol{X}\sim D}(\boldsymbol{X} \in \mathcal{X}^r) = \Pr_{\boldsymbol{X}\sim D}(\boldsymbol{X} \in \mathcal{X}^s) = \Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}(M(\boldsymbol{Y}_1, \ldots, \boldsymbol{y}_i^s, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i)$, violating strict incentive compatibility. □

Appendix E. Multiplicatively Normalizing Scores from Proper Scoring Rules Violates Truthfulness

Let $n = 2$ and $m = 1$, and suppose $p_1 = 0.5$. Let distribution D over X and Y_2 be defined as follows. With probability 0.5, $Y_2 = 1$ and $X = 0$, and with probability 0.5, $Y_2 = 1$ and $X = 1$. Observe that $\mathbf{E}_{X\sim D}[X] = p_1$. If forecaster 1 reports p_1, then she is selected with probability $R_q(0.5, 1)/(R_q(0.5, 1) + R_q(1, 1)) = 0.75/1.75 = 3/7$ when $X = 1$, and $R_q(0.5, 0)/(R_q(0.5, 0) + R_q(1, 0)) = 1$ when $X = 0$. That is, $\Pr_{X,Y_2\sim D}(M(p_1, Y_2, X) = 1) = 5/7 \approx 0.71$. If forecaster 1 instead reports $y_1' = 0.8$, then she is selected with probability $R_q(0.8, 1)/(R_q(0.8, 1) + R_q(1, 1)) = 0.96/1.96 = 24/49$ when $X = 1$, and $R_q(0.8, 0)/(R_q(0.8, 0) + R_q(1, 0)) = 1$ when $X = 0$. Her probability of being selected has increased to $\Pr_{X,Y_2\sim D}(M(y_1', Y_2, X) = 1) = 73/98 \approx 0.74$, violating truthfulness.

Appendix F. Proof of Theorem 2

Proof. To show strict truthfulness of M_{ELF^R} for $m = 1$, we show that reporting $y_i = p_i$ maximizes forecaster i's probability of being selected for any joint distribution over outcomes X and reports Y_{-i}:

$$\begin{aligned}
&\arg\max_{y_i} \Pr_{X,Y_{-i}\sim D}\left(M_{\mathrm{ELF}^R}(Y_1, \ldots, y_i, \ldots, Y_n, X) = i\right) \\
&= \arg\max_{y_i} \mathbf{E}_{X,Y_{-i}\sim D}[f_i(Y_1, \ldots, y_i, \ldots, Y_n, X)] \\
&= \arg\max_{y_i} \mathbf{E}_{X,Y_{-i}\sim D}\left[\frac{1}{n} + \frac{1}{n}\left(R(y_i, X) - \frac{1}{n-1}\sum_{j\neq i} R(Y_j, X)\right)\right] \\
&= \arg\max_{y_i} \mathbf{E}_{X,Y_{-i}\sim D}[R(y_i, X)] = p_i.
\end{aligned}$$

The last line follows from linearity of expectation and from R being a strictly proper scoring rule. □

Appendix G. Proof of Theorem 3

Proof. To show strict truthfulness of M_{ELF^R} for $m \neq 1$, we show that reporting $\boldsymbol{y}_i = \boldsymbol{p}_i$ maximizes forecaster i's probability of being selected for any joint distribution over outcomes X and reports Y_{-i}:

$$\begin{aligned}
&\arg\max_{\boldsymbol{y}_i} \Pr_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left(M_{\mathrm{ELF}^R}(\boldsymbol{Y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X}) = i\right) \\
&= \arg\max_{\boldsymbol{y}_i} \mathbf{E}_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}[g_i(\boldsymbol{Y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{Y}_n, \boldsymbol{X})] \\
&= \arg\max_{\boldsymbol{y}_i} \mathbf{E}_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left[\frac{1}{m}\sum_{k=1}^{m}\left(\frac{1}{n} + \frac{1}{n}\left(R(y_{i,k}, X_k) - \frac{1}{n-1}\sum_{j\neq i} R(Y_{j,k}, X_k)\right)\right)\right] \\
&= \arg\max_{\boldsymbol{y}_i} \mathbf{E}_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left[\sum_{k=1}^{m}\left(R(y_{i,k}, X_k) - \frac{1}{n-1}\sum_{j\neq i} R(Y_{j,k}, X_k)\right)\right] \\
&= \arg\max_{\boldsymbol{y}_i} \mathbf{E}_{\boldsymbol{X},\boldsymbol{Y}_{-i}\sim D}\left[\sum_{k=1}^{m} R(y_{i,k}, X_k)\right] = \boldsymbol{p}_i \quad \square
\end{aligned}$$

Appendix H. Proof of Proposition 2

Proof. The statement follows directly from the definition of M_{ELF^R}.

$$\begin{aligned}
&\Pr_{\boldsymbol{X}\sim\theta}(M_{\mathrm{ELF}^R}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{X}) = i) \\
&= \mathbf{E}_{\boldsymbol{X}\sim\theta}\left[\frac{1}{m}\sum_{k=1}^{m}\left(\frac{1}{n} + \frac{1}{n}\left(R(y_{i,k}, X_k) - \frac{1}{n-1}\sum_{j\neq i} R(y_{j,k}, X_k)\right)\right)\right] \\
&= \frac{1}{n} + \frac{1}{n}\left(\mathbf{E}_{\boldsymbol{X}\sim\theta}\frac{1}{m}\sum_{k=1}^{m}[R(y_{i,k}, X_k)] - \mathbf{E}_{\boldsymbol{X}\sim\theta}\frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{n-1}\sum_{j\neq i} R(y_{j,k}, X_k)\right]\right) \\
&= \frac{1}{n} + \frac{1}{n}\left(R(\boldsymbol{y}_i, \theta) - \frac{1}{n-1}\sum_{j\neq i} R(\boldsymbol{y}_j, \theta)\right) \quad \square
\end{aligned}$$

Appendix I. Proof of Theorem 4

Our proof of Theorem 4 proceeds in two parts. In the first part, we exploit the connection between wagering mechanisms and forecasting competition mechanisms to narrow down the particular form that any smooth, anonymous, strictly truthful forecasting competition mechanism must take. This form is parameterized by the choice of strictly proper scoring rule R. In the second part of the proof, we show that using any normalized proper scoring rule different from the one used to define accuracy must violate rank accuracy. Since we are considering only a single event X, for this proof we will slightly abuse notation and use θ to denote a single probability rather than a joint distribution.

Part 1. We begin by formally introducing wagering mechanisms. A wagering mechanism $\Pi = (\Pi_i)_{i\in[n]}$ is a set of functions Π_i, each of which takes as input the forecasters' reports $\boldsymbol{y} = (y_1, \ldots, y_n) \in [0,1]^n$, a vector of wagers $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n) \in \mathbb{R}^n_{\geq 0}$, and the event outcome $x \in \{0, 1\}$, and outputs a payment to forecaster i, $\Pi_i(\boldsymbol{y}, \boldsymbol{\omega}, x) \geq 0$. For our analysis, it will be sufficient to restrict ourselves to wagering mechanisms that only accept the vector of wagers $\boldsymbol{\omega} = (1/n, \ldots, 1/n)$. We refer to the resulting mechanisms as *equal-wager wagering mechanisms*[15] and denote the payments $\Pi_i(y_1, \ldots, y_n, x)$, omitting the (non-)dependence on $\boldsymbol{\omega}$.

The following definitions are standard in the wagering mechanism literature.

Definition I.1. An equal-wager wagering mechanism Π is *budget balanced* if, for all reports $y_1, \ldots, y_n \in [0, 1]$ and outcomes $x \in \{0, 1\}$, it holds that $\sum_{i=1}^n \Pi_i(y_1, \ldots, y_n, x) = 1$. That is, the sum of payments from the mechanism equals the sum of agents' wagers.

Definition I.2. An equal-wager wagering mechanism Π is *strictly incentive compatible under immutable beliefs* if, for all p_i, all reports $y_i \neq p_i$, and all $y_j \in [0, 1]$ for $j \neq i$, it holds that $\mathbf{E}_{X\sim p_i}\Pi_i(y_1, \ldots, y_i, \ldots, y_n, X) < \mathbf{E}_{X\sim p_i}\Pi_i(y_1, \ldots, p_i, \ldots, y_n, X)$. That is, truthfully reporting their subjective probability maximizes a forecaster's expected payment, given the reports of the other forecasters.

Definition I.3. An equal-wager wagering mechanism Π is *normal* if, for all probabilities $\theta \in [0, 1]$, all reports $y_1, \ldots, y_n \in [0, 1]$ and all $y_i' \in [0, 1]$, if $\mathbf{E}_{X\sim\theta}\Pi_i(y_1, \ldots, y_i, \ldots, y_n, X) < \mathbf{E}_{X\sim\theta}\Pi_i(y_1, \ldots, y_i', \ldots, y_n, X)$, then $\mathbf{E}_{X\sim\theta}\Pi_j(y_1, \ldots, y_i, \ldots, y_n, X) \geq \mathbf{E}_{X\sim\theta}\Pi_j(y_1, \ldots, y_i', \ldots, y_n, X)$ for all $j \neq i$. That is, if a forecaster i changes her report yielding a change ϵ_i in her expected payment, the change in expected payments of all other forecasters ϵ_j is null or has the opposite sign of ϵ_i.

Definition I.4. An equal-wager wagering mechanism Π is *anonymous* if for any permutation σ of $[n]$, any forecaster i, and any outcome x, it holds that $\Pi_i(y_1, \ldots, y_n, x) = \Pi_{\sigma(i)}(y_{\sigma^{-1}(1)}, \ldots, y_{\sigma^{-1}(n)}, x)$. That is, the payouts do not depend on the identities of the agents.

It will be useful to define smoothness for wagering mechanisms and proper scoring rules.

Definition I.5. An equal-wager wagering mechanism is *smooth* if, for all $i \in [n]$, Π_i is twice continuously differentiable with respect to each report y_j, $j \in [n]$. A proper scoring rule R is smooth if it is twice continuously differentiable with respect to the report y.

Our first lemma provides a formal relationship between budget-balanced equal-wager wagering mechanisms and forecasting competition mechanisms.

Definition I.6. Given a forecasting competition mechanism M, define the *corresponding* equal-wager wagering mechanism by $\Pi_i^M(y_1, \ldots, y_n, x) = \Pr(M(y_1, \ldots, y_n, x) = i) \geq 0$ for all $i \in [n]$.

Lemma I.1. *If a forecasting competition mechanism M is strictly incentive compatible, anonymous, and smooth, then the corresponding equal-wager wagering mechanism Π^M is budget-balanced, strictly incentive compatible for immutable beliefs, anonymous, and smooth.*

Proof. Consider a strictly incentive compatible and anonymous forecasting competition mechanism M and the corresponding equal-wager wagering mechanism Π^M.

For budget balance, note that $\sum_{i=1}^n \Pi_i^M(y_1, \ldots, y_n, x) = \sum_{i=1}^n \Pr(M(y_1, \ldots, y_n, x) = i) = 1$, where the latter equality follows from the fact that M outputs a probability distribution over forecasters.

For anonymity, we have $\Pi_i^M(y_1, \ldots, y_n, x) = \Pr(M(y_1, \ldots, y_n, x) = i) = \Pr(M(y_{\sigma^{-1}(1)}, \ldots, y_{\sigma^{-1}(n)}, x) = \sigma(i)) = \Pi_{\sigma(i)}^M(y_{\sigma^{-1}(1)}, \ldots, y_{\sigma^{-1}(n)}, x)$.

For strict incentive compatibility under immutable beliefs, for any p_i, reports $y_i \neq p_i$, and any $y_j \in [0, 1]$ for $j \neq i$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim p_i} \Pi_i^M(y_1, \ldots, y_i, \ldots, y_n, X) &= \Pr_{X\sim p_i}(M(y_1, \ldots, y_i, \ldots, y_n, X) = i) \\ &< \Pr_{X\sim p_i}(M(y_1, \ldots, p_i, \ldots, y_n, X) = i) \\ &= \mathop{\mathbf{E}}_{X\sim p_i} \Pi_i^M(y_1, \ldots, p_i, \ldots, y_n, X),\end{aligned}$$

where the inequality follows from strict incentive compatibility of M, taking joint distribution D to be such that $Y_j = y_j$ with probability one, and $\mathbf{E}_{X\sim D}[X] = p_i$.

Finally, smoothness of Π^M follows directly from smoothness of M and Definition I.6. □

Lambert et al. (2008) show that any smooth equal-wager wagering mechanism that is budget balanced, strictly incentive compatible for immutable beliefs, normal, and anonymous must have a particular form. We note that the versions of normality and incentive compatibility for immutable beliefs that Lambert et al. (2008) define are slightly weaker than the ones we use. In particular, Lambert et al. do not require that incentive compatibility holds for forecasters with belief $p_i = 0$ or $p_i = 1$ and normality is required only to hold for $\theta \in (0, 1)$. The following statement still holds for our versions of these properties since the mechanisms that satisfy our conditions are a subset of the mechanisms that satisfy theirs.

Lemma I.2 (Lambert et al. 2008, lemma 4 (Restated)). *For any $n \geq 2$, if a smooth[16] equal-wager wagering mechanism Π is budget balanced, strictly incentive compatible for immutable beliefs, anonymous and normal then there exists a smooth strictly proper scoring rule R such that*

$$\Pi_i(y_1, \ldots, y_n, x) = \frac{1}{n} + R(y_i, x) - \frac{1}{n-1}\sum_{j\neq i} R(y_j, x). \tag{I.1}$$

The following lemma incorporates two observations about Lemma I.2. First, R must be bounded to guarantee

nonnegative payouts as required by the definition of a wagering mechanism. Second, when restricted to $n = 2$, normality is implied by budget balance.

Lemma I.3. *For $n = 2$, if an equal-wager wagering mechanism is budget balanced, strictly incentive compatible for immutable beliefs, anonymous, and smooth, then there exists a smooth strictly proper scoring rule $R \in [0,1]$ such that*

$$\Pi_i(y_1, y_2, x) = \frac{1}{2} + \frac{1}{2}\Big(R(y_i, x) - R(y_{3-i}, x)\Big). \tag{I.2}$$

Proof. When $n = 2$, budget balance implies that $\Pi_1(y_1, y_2, x) = 1 - \Pi_2(y_1, y_2, x)$ for all $y_1, y_2 \in [0,1]$ and all $x \in \{0,1\}$. Taking the expectation over possible outcomes yields $\mathbf{E}_{X\sim\theta}\Pi_1(y_1, y_2, X) = 1 - \mathbf{E}_{X\sim\theta}\Pi_2(y_1, y_2, X)$. In particular, any change in the expected payment to forecaster i is exactly offset by the change in expected payment to forecaster $3 - i$. Therefore, normality is implied by budget balance.

Boundedness of R follows from Lemma I.2 and the definition of a wagering mechanism. By the constraint that $0 \le \Pi_i(y_1, y_2, x) \le 1$, where Π_i is defined as in Lemma I.2, it must be the case that $|R(y_i, x) - R(y_{3-i}, x)| \le 0.5$ for all y_1, y_2, x. We can therefore define $R'(y, x)$ by $R'(y, x) = 2(R(y, x) + \beta(x))$, where $\beta(0) = -R(1,0)$ and $\beta(1) = -R(0,1)$. Scoring rule R' is derived from R by a positive affine transformation and therefore inherits strict properness from R. Note that the minimum value of R' is $R'(0,1) = R'(1,0) = 0$ and the maximum value is either $R'(0,0) = 2(R(0,0) - R(1,0)) \le 1$ or $R'(1,1) = 2(R(1,1) - R(0,1)) \le 1$, and so R' is bounded in $[0,1]$. Furthermore, plugging R' into Equation (I.2) yields exactly Equation (I.1). □

We can now characterize the form that any strictly incentive-compatible, anonymous, and smooth forecasting competition mechanism must have.

Lemma I.4. *For $n = 2$, if a forecasting competition mechanism M is strictly incentive compatible, anonymous, and smooth, then there exists a smooth strictly proper scoring rule $R(y, x) \in [0,1]$ such that for all $i \in \{1,2\}$*

$$\Pr(M(y_1, y_2, x) = i) = \frac{1}{2} + \frac{1}{2}\Big(R(y_i, x) - R(y_{3-i}, x)\Big).$$

Proof. Let M be a strictly incentive compatible, anonymous, and smooth forecasting competition mechanism. Then, by Lemma I.1, the corresponding equal-wager wagering mechanism Π^M is budget balanced, strictly incentive compatible for immutable beliefs, anonymous, and smooth. Therefore, by Lemma I.3, there must exist a smooth strictly proper scoring rule $R \in [0,1]$ such that for all $i \in \{1,2\}$

$$\Pi_i^M(y_1, y_2, x) = \frac{1}{2} + \frac{1}{2}\Big(R(y_i, x) - R(y_{3-i}, x)\Big).$$

By Definition I.6, this implies that for all $i \in \{1,2\}$

$$\Pr\Big(M(y_1, y_2, x) = i\Big) = \frac{1}{2} + \frac{1}{2}\Big(R(y_i, x) - R(y_{3-i}, x)\Big),$$

which is the desired result. □

We have now established the form that any strictly incentive compatible, smooth, and anonymous forecasting competition mechanism M must have for $n = 2$. In particular, M is equivalent to M_{ELF^R} for some smooth, bounded, strictly proper scoring rule R. Next, we show that R can always be represented by a differentiable strictly convex function G.

Lemma I.5. *Let R be a smooth strictly proper scoring rule. There exists a strictly convex, differentiable function $G : [0,1] \to \mathbb{R}$ with*

$$R(y, \theta) = G(y) + dG(y) \cdot (\theta - y),$$

where $\theta \in [0,1]$ and $dG(y)$ is the derivative of G at y. Furthermore, $G(y)$ is the expected score for reporting $y = \theta$. Every R defines a unique G and every G defines a unique R.

Proof. It is well known that every strictly proper scoring rule can be expressed as $R(y, \theta) = G(y) + dG(y) \cdot (\theta - y)$ for some strictly convex function G, where $dG(y)$ is a subgradient of G at y (McCarthy 1956, Savage 1971, Schervish 1989, Gneiting and Raftery 2007). Observe that setting $y = \theta$ yields expected score $G(y)$, and it immediately follows that every R defines a unique G.

Let R be smooth (and, in particular, continuous). Suppose for the sake of contradiction that the convex function G associated with R is not differentiable at some $y' \in [0,1]$. That is, the left and right derivatives of G at y' ($d_-G(y')$ and $d_+G(y')$, respectively) are not equal. Note that convexity implies that $d_-G(y') \le d_+G(y')$, so the fact that the left and right derivatives are not equal yields $d_-G(y') < d_+G(y')$. We therefore have $\lim_{\epsilon\to 0^+} R(y' - \epsilon, 1) = G(y') + d_-G(y') \cdot (1 - y') < G(y') + d_+G(y') \cdot (1 - y') = \lim_{\epsilon\to 0^+} R(y' + \epsilon, 1)$, violating continuity of R at y' for $\theta = 1$, a contradiction to smoothness of R. Furthermore, note that differentiability of G implies a unique scoring rule R. □

Part 2. The remainder of the proof is devoted to comparing the behavior of M_{ELF^R} for different choices of smooth proper scoring rule R. We will require the notion of equivalent scoring rules. A proper scoring rule R is equivalent to another proper scoring rule R' if R can be obtained from R' by a positive affine transformation.

Definition I.7. Proper scoring rules R and R' are *equivalent* if and only if $R'(y, x) = \alpha R(y, x) + \beta(x)$ for some $\alpha > 0$ and $\beta(x) \in \mathbb{R}$ for $x \in \{0,1\}$.

This definition partitions the space of proper scoring rules into equivalence classes. It will be useful to define the *canonical form* of a scoring rule R as a convenient representative of each class. In particular, the canonical form ensures that every perfect forecast of a sure event obtains a score of one and that the minimum expected score of a perfect forecast is zero.

Definition I.8. Let R and R' be strictly proper scoring rules. We say that R' is the *canonical form* of R if R' and R are equivalent, and $R'(0,0) = R'(1,1) = 1$ and $\min_\theta R'(\theta, \theta) = 0$ for some $\theta \in (0,1)$.

Lemma I.6. *For any strictly proper scoring rule R, there exists a canonical form R'.*

Proof. It is sufficient to show that any strictly proper scoring rule R can be brought into canonical form through

one particular positive-affine transformation. To transform any strictly proper scoring rule R into its canonical form, we first define linear function $f(x)$ for $x \in \{0,1\}$ such that, when added to $R(y,x)$, every perfect forecast of a sure event obtains a score of zero. That is, $f(0) := -R(0,0)$ and $f(1) := -R(1,1)$. In a second step, we are multiplying $R(y,x)+f(x)$ by $\alpha := \frac{1}{-\min_\theta \mathbf{E}_{X\sim\theta}[R(\theta,X)+f(X)]}$ such that its minimum expected score of a perfect forecast is -1. Note that $\alpha > 0$ since $\min_\theta \mathbf{E}_{X\sim\theta}[R(\theta,X)+f(X)] < 0$ because $R(0,0)+f(0)=0$ and $R(1,1)+f(1)=0$, and because of strict convexity of the expected score function. Finally, we add a constant 1 to R, resulting in $R'(y,x) := \alpha(R(y,x)+f(x))+1$. □

It immediately follows from Definition I.8 and Lemma I.6 that if two strictly proper scoring rules have the same canonical form, then they are equivalent. In order to prove our key result, we require a technical lemma.

Lemma I.7. *Let $f,g:[0,1]\to\mathbb{R}$ be differentiable, strictly convex functions. Additionally, suppose that f is strictly decreasing, $f(0)=g(0)=1$ and that there exists a $\bar{t}\in(0,1]$ for which $f(\bar{t})<g(\bar{t})$. Then there must exist a $t'\in(0,\bar{t}]$ for which $f(t')<g(t')$ and $d(f(t'))<d(g(t'))$.*

Proof. Let $t^* = \sup\{t\in[0,\bar{t}] : f(t)\geq g(t)\}$. We are guaranteed that t^* is well defined because $f(0)=g(0)$ so we are taking a supremum over a nonempty set. Furthermore, it is easy to see that $f(t^*)=g(t^*)$ and that $f(t)<g(t)$ for all $t\in(t^*,\bar{t}]$. Suppose for contradiction that $d(f(t))\geq d(g(t))$ for all $t\in(t^*,\bar{t}]$. This would imply that $f(\bar{t})\geq g(\bar{t})$, contradicting the assumption of the lemma. Therefore, there must exist a $t'\in(t^*,\bar{t}]$ with $d(f(t'))<d(g(t'))$. □

Finally, we show that two smooth strictly proper scoring rules R and R' are equivalent if and only if they always agree on the relative accuracy of forecasters.

Lemma I.8. *Smooth strictly proper scoring rules R and R' are equivalent if and only if $R'(y_1,\theta)>R'(y_2,\theta)\iff R(y_1,\theta)>R(y_2,\theta)$ for all $y_1,y_2,\theta\in[0,1]$.*

Proof. We first prove the forward direction. Suppose that R and R' are equivalent, that is, $R'(y,x)=\alpha R(y,x)+\beta(x)$ for some $\alpha>0$ and $\beta(x)\in\mathbb{R}$. Then, $R'(y_1,\theta)>R'(y_2,\theta)\iff \mathbf{E}_{X\sim\theta}[\alpha R(y_1,X)+\beta(x)] > \mathbf{E}_{X\sim\theta}[\alpha R(y_2,X)+\beta(X)]\iff \mathbf{E}_{X\sim\theta}[\alpha R(y_1,X)]+\mathbf{E}_{X\sim\theta}[\beta(X)]>\mathbf{E}_{X\sim\theta}[\alpha R(y_2,X)]+\mathbf{E}_{X\sim\theta}[\beta(X)]\iff \mathbf{E}_{X\sim\theta}[\alpha R(y_1,X)]>\mathbf{E}_{X\sim\theta}[\alpha R(y_2,X)]\iff R(y_1,\theta)>R(y_2,\theta)$.

For the backward direction, suppose that R and R' are not equivalent. Assume that R and R' are in their respective canonical forms (if not, we can convert them to canonical form without changing the way they rank forecasters). Note that smoothness of R and R' implies the existence of associated differentiable convex functions G and G', as per Lemma I.5. Since R and R' are in canonical form, $\min_\theta G(\theta)=\min_\theta G'(\theta)=0$, and $G(0)=G(1)=G'(0)=G'(1)=1$. Furthermore, since R and R' are not equivalent, we know that $G\neq G'$. We treat two cases.

Case 1: Suppose that $\arg\min_\theta G(\theta)=\arg\min_\theta G'(\theta)$. However, since $G\neq G'$, there must exist a y at which $G(y)\neq G'(y)$. Without loss of generality, suppose $G(y)<G'(y)$. For mathematical convenience, suppose that $y<\arg\min_\theta G(\theta)$; the case in which $y>\arg\min_\theta G(\theta)$ follows similarly.

By Lemma I.7, taking $f=G$ and $g=G'$, there must exist a point $y_1<y$ for which $0<G(y_1)<G'(y_1)$ and $d(G(y_1))<d(G'(y_1))<0$. Set $y_2=\arg\min_\theta G(\theta)$ equal to the point at which $G(y_2)=G'(y_2)=0$. Since G and G' are both differentiable, $d(G(y_2))=d(G'(y_2))=0$. Finally, set θ so that $R(y_1,\theta)=0$. That is,

$$G(y_1)+d(G(y_1))(\theta-y_1)=0.$$

Note that, since $G(y_1)>0$ and $d(G(y_1))<0$, we have $\theta>y_1$. Then,

$$\begin{aligned}R(y_1,\theta)&=G(y_1)+d(G(y_1))\cdot(\theta-y_1)\\&=0\\&=G(y_2)+d(G(y_2))\cdot(\theta-y_2)\\&=R(y_2,\theta).\end{aligned}$$

However,

$$\begin{aligned}R'(y_1,\theta)&=G'(y_1)+d(G'(y_1))\cdot(\theta-y_1)\\&>G(y_1)+d(G(y_1))\cdot(\theta-y_1)\\&=0\\&=G'(y_2)+d(G'(y_2))\cdot(\theta-y_2)\\&=R'(y_2,\theta),\end{aligned}$$

so that forecasters 1 and 2 obtain the same expected score according to R, but forecaster 1 obtains higher expected score according to R'. In particular, R and R' disagree on the relative accuracy.

Case 2: Suppose that, without loss of generality, $\theta_{\min}:=\arg\min_\theta G(\theta)<\arg\min_\theta G'(\theta):=\theta'_{\min}$. In particular, $G(\theta_{\min})=0<G'(\theta_{\min})$, and $G(\theta'_{\min})>0=G'(\theta'_{\min})$. By Lemma I.7, there must exist a $y_1<\theta_{\min}$ for which $G(y_1)<G'(y_1)$ and $d(G(y_1))<d(G'(y_1))<0$. Similarly, there must exist a $y_2>\theta'_{\min}$ for which $G(y_2)>G'(y_2)$ and $0<d(G(y_2))<d(G'(y_2))$. Let θ be such that R gives the same expected score to both reports. That is,

$$\begin{aligned}R(y_1,\theta)&=G(y_1)+d(G(y_1))\cdot(\theta-y_1)\\&=G(y_2)+d(G(y_2))\cdot(\theta-y_2)=R(y_2,\theta).\end{aligned}$$

Note that, by strict convexity of G, it needs to hold that $\theta\in(y_1,y_2)$. For R', we have

$$\begin{aligned}R'(y_1,\theta)&=G'(y_1)+d(G'(y_1))\cdot(\theta-y_1)\\&>G(y_1)+d(G(y_1))\cdot(\theta-y_1)\\&=G(y_2)+d(G(y_2))\cdot(\theta-y_2)\\&>G'(y_2)+d(G'(y_2))\cdot(\theta-y_2)\\&=R'(y_2,\theta),\end{aligned}$$

where the first and last equalities follow from Lemma I.5, the inequalities hold because $\theta\in(y_1,y_2)$, and the second equality follows from the definition of θ. Again, forecasters 1 and 2 obtain the same expected score according to R, but forecaster 1 obtains higher expected score according to R'. This completes the backward direction. □

We can now complete the proof of Theorem 4.

Proof of Theorem 4. By Lemmas I.4 and I.5, when $n=2$, any smooth, anonymous, strictly incentive-compatible

forecasting competition mechanism M must take the form of $M_{\text{ELF}^{R'}}$ for some smooth, bounded, strictly proper scoring rule $R' \in [0,1]$ with associated differentiable convex function G'. We complete the proof by showing that every forecasting competition mechanism of this form either fails to be rank accurate with respect to R, or has $\Pr_{X\sim\theta}(M(y_1,y_2,X)=1) \le \Pr_{X\sim\theta}(M_{ELF^{\tilde{R}}}(y_1,y_2,X)=1)$ for every $y_1,\ y_2, \theta \in [0,1]$ for which $R(y_1,\theta) > R(y_2,\theta)$.

If R' is not equivalent to R, then $M_{\text{ELF}^{R'}}$ is not rank accurate with respect to R by Corollary 1 and Lemma I.8. If R' is equivalent to R, then we have that $R'(y,x) = \alpha R(y,x) + \beta(x)$. We also know that $\tilde{R}(y,x) = \tilde{\alpha}R(y,x) + \tilde{\beta}(x)$, where $\tilde{\alpha} \ge \alpha$ (if $\tilde{\alpha} < \alpha$ then R' is not bounded in $[0,1]$). Let $y_1, y_2, \theta \in [0,1]$ such that $R(y_1,\theta) > R(y_2,\theta)$. Then

$$\begin{aligned}
&\Pr_{X\sim\theta}\left(M_{ELF^{\tilde{R}}}(y_1,y_2,X)=1\right)\\
&=\frac{1}{2}+\frac{1}{2}\left(\tilde{R}(y_1,\theta)-\tilde{R}(y_2,\theta)\right)\\
&=\frac{1}{2}+\frac{1}{2}\left(\tilde{\alpha}R(y_1,\theta)+\mathop{\mathbf{E}}_{X\sim\theta}[\tilde{\beta}(X)]-\tilde{\alpha}R(y_2,\theta)-\mathop{\mathbf{E}}_{X\sim\theta}[\tilde{\beta}(X)]\right)\\
&=\frac{1}{2}+\frac{1}{2}\left(\tilde{\alpha}R(y_1,\theta)-\tilde{\alpha}R(y_2,\theta)\right)\\
&\ge\frac{1}{2}+\frac{1}{2}\left(\alpha R(y_1,\theta)-\alpha R(y_2,\theta)\right)\\
&=\frac{1}{2}+\frac{1}{2}\left(\alpha R(y_1,\theta)+\mathop{\mathbf{E}}_{X\sim\theta}[\beta(X)]-\alpha R(y_2,\theta)-\mathop{\mathbf{E}}_{X\sim\theta}[\beta(X)]\right)\\
&=\Pr_{X\sim\theta}\left(M_{\text{ELF}^{R'}}(y_1,y_2,X)=1\right),
\end{aligned}$$

where the inequality follows from $\tilde{\alpha} \ge \alpha$ and $R(y_1,\theta) > R(y_2,\theta)$. □

Appendix J. Proof of Theorem 5

Proof. We first make a basic observation about unbounded proper scoring rules. The proof then proceeds by leveraging Lemma I.4, which characterizes the form that any strictly incentive-compatible, anonymous, and smooth forecasting competition mechanism must take. Finally, it shows that no mechanism of that form can be rank accurate with respect to an unbounded proper scoring rule.

Let R be an unbounded strictly proper scoring rule. First note that since R is strictly proper, it must be the case that $R(0,1) < R(y,1)$ for any $y > 0$ and, analogously, $R(1,0) < R(y,0)$ for any $y < 1$. Therefore, since R is unbounded (i.e., $R(y,x) = -\infty$ for some $y \in [0,1]$ and $x \in \{0,1\}$), it must be the case that $R(0,1) = -\infty$ and/or $R(1,0) = -\infty$, and $R(y,x) \in \mathbb{R}$ for all $y \in (0,1)$ and $x \in \{0,1\}$. Suppose now that $R(0,1) = -\infty$. (The case with $R(1,0) = -\infty$ can be proven identically.)

Let M be a strictly incentive-compatible, anonymous, and smooth forecasting competition mechanism. By Lemma I.4, we know the form that M must take for $n = 2$. In particular, there must exist a smooth, bounded strictly proper scoring rule $R' \in [0,1]$ such that $M = M_{\text{ELF}^{R'}}$. We now show that $M_{\text{ELF}^{R'}}$ is not rank accurate with respect to R. Fix $y \in (0,1)$ and let θ be such that R' gives the same expected score to reports zero and y. That is,

$$\begin{aligned}
R'(0,\theta) &= G(0) + d(G(0))\cdot\theta = G(y) + d(G(y))\cdot(\theta - y)\\
&= R'(y,\theta),
\end{aligned}$$

where G is the convex function associated with R' (Savage 1971). Note that $\theta \in (0,y)$ by strict convexity of G and the fact that $d(G(0)) \in \mathbb{R}$ and $d(G(y)) \in \mathbb{R}$ (which is implied by boundedness of R'). Furthermore, since R' gives the same expected score to reports 0 and y, if forecaster 1 reports $y_1 = 0$ and forecaster 2 reports $y_2 = y$, we have

$$\Pr_{X\sim\theta}\left(M_{\text{ELF}^{R'}}(y_1,y_2,X)=1\right) = \Pr_{X\sim\theta}\left(M_{\text{ELF}^{R'}}(y_1,y_2,X)=2\right).$$

However, $R(y_1,\theta) = R(0,\theta) = \theta\cdot R(0,1) + (1-\theta)\cdot R(0,0) = \theta\cdot(-\infty) + (1-\theta)\cdot R(0,0) = -\infty$ and $R(y_2,\theta) = R(y,\theta) \in \mathbb{R}$. Therefore, $M_{\text{ELF}^{R'}}$ is not rank accurate with respect to R. □

Appendix K. Proof of Proposition 3

Proof. Let $n = 2$ with $\boldsymbol{y}_1 = (0.4,\ldots,0.4)$ and $\boldsymbol{y}_2 = (0.6,\ldots,0.6)$. Let R be the strictly proper scoring rule that defines accuracy. Now suppose M is a limit accurate forecasting competition mechanism and consider the following two cases with two different "perfectly correlated" joint distributions θ for which all m outcomes are the same, that is, either $X_k = 0$ for all k or $X_k = 1$ for all k:

1. Suppose $\theta_k = 0.4$ for all k. Since $y_{1,k} = \theta_k$ and $y_{2,k} \ne \theta_k$ for all k, strict properness of R implies that forecaster 1 is strictly more accurate. Hence, limit accuracy implies that there exists an $\underline{m_1}$ such that for all $m \ge m_1$, M selects forecaster 1 with probability at least $\pi = 0.7$.

2. Suppose $\theta_k = 0.6$ for all k. Since $y_{2,k} = \theta_k$ and $y_{1,k} \ne \theta_k$ for all k, strict properness of R implies that forecaster 2 is strictly more accurate. Hence, limit accuracy implies that there exists an $\underline{m_2}$ such that for all $m \ge m_2$, M selects forecaster 2 with probability at least $\pi = 0.7$.

Now let $m = \max(\underline{m_1},\underline{m_2})$ be the number of events. Since both θ are "perfectly correlated," the outcome vector is either $\boldsymbol{x} = (0,\ldots,0)$ or $\boldsymbol{x} = (1,\ldots,1)$, and so it is sufficient to consider whom M selects given each of these. Let $q_{1|\mathbf{0}}$ and $q_{1|\mathbf{1}}$ be the probabilities that M selects forecaster 1 given $\mathbf{x} = (0,\ldots,0)$ and $\mathbf{x} = (1,\ldots,1)$, respectively. From Case 1, it needs to hold that $0.4\cdot q_{1|\mathbf{1}} + 0.6\cdot q_{1|\mathbf{0}} > 0.7$ and from Case 2, it needs to hold that $0.6\cdot(1-q_{1|\mathbf{1}}) + 0.4\cdot(1-q_{1|\mathbf{0}}) > 0.7$. However, this is impossible because the former implies that $q_{1|\mathbf{1}} > \frac{7}{4} - \frac{3}{2}q_{1|\mathbf{0}}$ and the latter implies that $q_{1|\mathbf{1}} < \frac{1}{2} - \frac{2}{3}q_{1|\mathbf{0}}$, with no $q_{1|\mathbf{0}}, q_{1|\mathbf{1}} \in [0,1]$ satisfying both, a contradiction that M is limit accurate. □

Appendix L. Proof of Theorem 6

Proof. Without loss of generality, take the perspective of any forecaster $i \in [n]$ seeking to maximize the probability of being selected. In reasoning about forecaster i's probability of winning, she needs to reason about the joint probability of the event winners vector $(w_1,\ldots,w_m)$, which is given by the vector of probability distributions $(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_m)$, where each $\boldsymbol{f}_k$ is the distribution over forecasters for event k. From forecaster i's perspective, each $\boldsymbol{f}_k$ is an instantiation of a random variable $\boldsymbol{F}_k$, depending on her belief about $\boldsymbol{Y}_{-i}$ and $\boldsymbol{X}$. Without any restrictions on $\boldsymbol{Y}_{-i}$ and $\boldsymbol{X}$, these $\boldsymbol{F}_k$ can be dependent even if—given instantiated $(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_m)$—the draws of the event winners themselves are independent by definition of the mechanism. For belief independent joint distributions D over outcomes $\boldsymbol{X}$ and

reports $\boldsymbol{Y}_{-i}$, however, all random vectors $(Y_{1,k},\ldots,Y_{i-1,k},Y_{i+1,k},\ldots,Y_{n,k},X_k)$ indexed by k are independent, so that all $\boldsymbol{F}_k$ are independent as well. Consider now event k and let $K' \in \mathcal{P}([m])$ be any subset of event indices with $k \notin K'$. By independence of $\boldsymbol{F}_k$ for all k, changing forecaster i's report on event k does not affect the (joint) distribution of $\boldsymbol{F}_{K'}$.

It is easy to see that increasing forecaster i's expected (subjective) winning probability for event k, $\mathbf{E}[F_{i,k}]$, simultaneously decreases the expected winning probability $\mathbf{E}[F_{j,k}]$ of every $j \neq i$. To see this, first observe that, if $\mathbf{E}[F_{i,k}]$ increases, the sum of all other forecasters' event winning probabilities needs to decrease by the same amount because $\mathbf{E}[F_{i,k}] + \sum_{j\neq i}\mathbf{E}[F_{j,k}] = 1$ for all k. Second, by definition of $f_{i,k}$, any increase of $\epsilon > 0$ in $\mathbf{E}[F_{i,k}]$ leads to a uniform decrease of $\frac{\epsilon}{n-1}$ in each $\mathbf{E}[F_{j,k}]$ with $j \neq i$. This means that, since the $\boldsymbol{F}_k$ are independent, increasing $\mathbf{E}[F_{i,k}]$ on event k cannot decrease your probability of winning overall.

It remains to be shown that increasing $\mathbf{E}[F_{i,k}]$ strictly increases forecaster i's probability of winning overall. To show this, we need to show that there are situations, where event k is pivotal for winning overall and that these situations occur with positive probability. First, there exist event win outcomes $w_1,\ldots,w_{k-1},w_{k+1},\ldots,w_m$ on the other $m-1$ events such that k is pivotal; that is, winning or losing event k changes the probability of winning the prize. This is the case if and only if, without event k, (1) some forecaster $j \neq i$ won most events with forecaster i winning one fewer, or (2) forecaster i won most events with at least one other forecaster $j \neq i$ having won exactly the same number, or one event less than forecaster i. For example, with m odd, $m-1$ is even, and forecasters i and $j \neq i$ can each win half of those events. Similarly, with m even, $m-1$ is odd, and it can be the case that forecaster i wins $\lfloor\frac{m-1}{2}\rfloor$ and j wins $\lceil\frac{m-1}{2}\rceil$. Second, these cases occur with positive probability because we know that every $\mathbf{E}[F_{j,k}]$ for all j and all k is strictly in between zero and one by definition of $f_{i,k}$ and $R \in [0,1)$. Hence, event k is pivotal for forecaster i with positive probability and reporting truthfully on event k strictly increases the probability of winning the prize. □

Appendix M. Proof of Theorem 7

The proof uses the one-sided version of Hoeffding's inequality (Hoeffding 1963), which we state here for convenience.

Theorem M.1. (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*

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

and

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

Proof of Theorem 7. Let $w_{i,k} := \mathbb{1}(w_k = i)$ indicate whether forecaster i is the event winner for event k, and let $W_{i,k}$ be the corresponding random variable. Note that the reports $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are fixed, so that the uncertainty is only about the event outcomes $\boldsymbol{X}$. In particular, with $X_1,\ldots,X_m$ independent, $W_{i,1},\ldots,W_{i,m}$ are independent conditional on $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$.

Let $z_i = \sum_{k=1}^m w_{i,k}$ be the number of events won by forecaster i. Furthermore, let Z_i be the corresponding random variable, so that

$$\mathop{\mathbf{E}}_{X\sim\theta}[Z_i] = \mathop{\mathbf{E}}_{X\sim\theta}\left[\sum_{k=1}^m f_{i,k}\right],$$

where the latter expectation is taken over the outcomes, and the former is taken over the outcomes and the randomness of the lotteries.

To show limit accuracy, let i be the most accurate forecaster with $\Delta := \min_{j\neq i}(R(\boldsymbol{y}_i,\theta) - R(\boldsymbol{y}_j,\theta)) > 0$ denoting the difference between the expected scores of i and the second-most accurate forecaster. 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 $j \neq i$:

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim\theta}[Z_i] - \mathop{\mathbf{E}}_{X\sim\theta}[Z_j] &= \mathop{\mathbf{E}}_{X\sim\theta}\left[\sum_{k=1}^m (f_{i,k} - f_{j,k})\right] \\ &= \frac{\mathop{\mathbf{E}}_{X\sim\theta}\left[\sum_{k=1}^m \Big(R(y_{i,k},X_k) - R(y_{j,k},X_k)\Big)\right]}{n-1} \\ &= \frac{m\Big(R(\boldsymbol{y}_i,\theta) - R(\boldsymbol{y}_j,\theta)\Big)}{n-1} \geq \frac{m\Delta}{n-1}. \end{aligned} \tag{M.1}$$

The second equality follows from substituting the definition of $f_{i,k}$ and simplifying, the third equality follows from rewriting in terms of expected average score, and the inequality follows from the definition of Δ.

We now upper bound the probability that forecaster j wins more events than forecaster i. From Equation (M.1), if $z_j \geq z_i$, then it holds that $\mathbf{E}[Z_i] - z_i \geq \frac{m\Delta}{2(n-1)}$ or $z_j - \mathbf{E}[Z_j] \geq \frac{m\Delta}{2(n-1)}$ (both may apply simultaneously). By Hoeffding's inequality,

$$\Pr\left(\mathbf{E}[Z_i] - z_i \geq \frac{m\Delta}{2(n-1)}\right) \leq e^{-\frac{m\Delta^2}{2(n-1)^2}},$$

and

$$\Pr\left(z_j - \mathbf{E}[Z_j] \geq \frac{m\Delta}{2(n-1)}\right) \leq e^{-\frac{m\Delta^2}{2(n-1)^2}}.$$

Putting these together, we have

$$\begin{aligned}&\Pr(z_j \geq z_i) \\ &\leq \Pr\left(\left(\mathbf{E}[Z_i] - z_i \geq \frac{m\Delta}{2(n-1)}\right) \bigcup \left(z_j - \mathbf{E}[Z_j] \geq \frac{m\Delta}{2(n-1)}\right)\right) \\ &\leq \Pr\left(\mathbf{E}[Z_i] - z_i \geq \frac{m\Delta}{2(n-1)}\right) + \Pr\left(z_j - \mathbf{E}[Z_j] \geq \frac{m\Delta}{2(n-1)}\right) \\ &\leq 2e^{-\frac{m\Delta^2}{2(n-1)^2}}.\end{aligned}$$

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

$$\begin{aligned}&\Pr_{X\sim\theta}(M_{\mathrm{ELF}^R}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{X}) = i) \\ &\quad = 1 - \sum_{j\neq i}\Pr_{X\sim\theta}\Big(M_{\mathrm{ELF}^R}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{X}) = j\Big) \\ &\quad \geq 1 - \sum_{j\neq i}\Pr_{X\sim\theta}\Big(z_j \geq z_i\Big) \\ &\quad \geq 1 - 2(n-1)e^{-\frac{m\Delta^2}{2(n-1)^2}},\end{aligned}$$

where the first transition holds because exactly one forecaster is selected, and the second because $z_j \geq z_i$ is a necessary condition for forecaster j to be selected by ELF. The final transition holds by plugging in the earlier inequality. In particular, for fixed n and "accuracy gap" Δ, for any $\pi \in [0, 1)$, I-ELF selects the best forecaster with probability at least π if

$$m \geq \frac{2(n-1)^2}{\Delta^2} \ln\left(\frac{2(n-1)}{1-\pi}\right),$$

which yields limit accuracy. □

Endnotes

[1] See www.netflixprize.com.

[2] See www.kaggle.com.

[3] In Section 5, we will introduce a restricted definition that assumes that the events $\boldsymbol{X}$ are known to be independent and that this independence of events is reflected in the uncertainty about others' reports.

[4] In fact, we do not even require that forecasters are expected utility maximizers but only require that they are "probabilistically sophisticated" (Machina and Schmeidler 1992). We thank an anonymous reviewer for this observation.

[5] See https://projects.fivethirtyeight.com/2019-nfl-forecasting-game.

[6] Other tie-breaking procedures are possible, and our results do not rely on any particular one.

[7] Although our definition allows for any bounded R, we will see in Section 5 that the optimal accuracy guarantees are achieved for normalized R.

[8] We drop the dependencies of each f_i for clarity.

[9] The spherical scoring rule (Jose 2009) is defined as $R_s(y,x) := \frac{yx+(1-y)(1-x)}{\sqrt{y^2+(1-y)^2}}$. Forecaster 1 obtains an expected score of 0.73 and forecaster 2 obtains an expected score of only 0.71.

[10] We emphasize here that this choice needs to be made in any application of proper scoring rules, including but not limited to forecasting competitions. If, for example, one considers two reports to be of the same accuracy when they are "equally far away" from θ_k, then this implies that one would want to use the quadratic scoring rule since it is known to be the only proper scoring rule that punishes forecasters according to their Euclidean distance from θ_k (Selten 1998). Similarly, the spherical scoring rule is the only proper scoring rule to satisfy proportionality (Jose 2009). We allow the designer to choose other proper scoring rules that satisfy different properties.

[11] If used in conjunction with a normalized R for $m \geq 2$, $M_{\text{I-ELF}^R}$ may fail to be strictly incentive compatible (it is still weakly incentive compatible) when there exists an event for which a forecaster believes that she is a perfect forecaster reporting 100% for the eventually materialized outcome and every other forecaster is doing the opposite, that is, reporting 0% for the eventually materialized outcome. We do not expect this to be an issue in practical application.

[12] Observe that this misreport is somewhat different from those in the categorical setting, where rational forecasters will generally "extremize" their reports toward an outcome. In contrast, in the previous example, a forecaster who unilaterally deviates to reporting an extreme value of 0 or 1 would only be selected with probability 1/4.

[13] This assumes that no money leaves the system in the form of fees or withdrawals, a reasonable assumption for play money markets.

[14] We instantiate a particular $\boldsymbol{p}_i$, but the example is not sensitive to this choice.

[15] Equal-wager wagering mechanisms can be equivalently expressed as *Competitive Scoring Rules* (Kilgour and Gerchak 2004).

[16] Lambert et al. (2008) restrict attention to smooth wagering mechanisms, so this condition does not explicitly appear in their lemma statement.

References

Atanasov P, Rescober P, Stone E, Servan-Schreiber E, Tetlock PE, Ungar L, Mellers B (2017) Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Sci.* 63(3):691–706.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Chakraborty A (2016) How companies are using Kaggle to find the best machine learning talent. Accessed December 24, 2020, https://blog.udacity.com/2016/07/companies-kaggle-machine-learning-talent.html.

Chakraborty M, Das S, Lavoie A, Magdon-Ismail M, Naamad Y (2013) Instructor rating markets. DesJardins M, Littman M, eds. *Proc. 27th AAAI Conf. on Artificial Intelligence* (AAAI Press, Palo Alto), 159–165.

Cowgill B, Zitzewitz E (2015) Corporate prediction markets: Evidence from Google, Ford, and Firm X. *Rev. Econom. Stud.* 82(4): 1309–1341.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102:359–378.

Good IJ (1952) Rational decisions. *J. Roy. Statist. Soc. B* 14(1):107–114.

Grushka-Cockayne Y, Lichtendahl KC, Jose VR, Winkler RL (2017) Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Oper. Res.* 65(3):557–836.

Harris D (2013) Facebook is hiring a data scientist, but you'll have to fight for the job. Accessed December 24, 2020, https://gigaom.com/2013/08/30/facebook-is-hiring-a-data-scientist-but-youll-have-to-fight-for-the-job/.

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58(301):13–30.

Hossain T, Okui R (2013) The binarized scoring rule. *Rev. Econom. Stud.* 80(3):984–1001.

Jia Y, Liu Y, Yu X, Voida S (2017) Designing leaderboards for gamification: Perceived differences based on user ranking, application domain, and personality traits. Mark G, Fussell S, Lampe C, Schraefel MC, eds. *Proc. CHI Conf. on Human Factors in Comput. Systems* (ACM, New York), 1949–1960.

Jose VR (2009) A characterization for the spherical scoring rule. *Theory Decision* 66(3):263–281.

Jose VR (2017) Percentage and relative error measures in forecast evaluation. *Oper. Res.* 65(1):200–211.

Kaggle (2017) March machine learning mania, 1st place winner's interview: Andrew Landgraf. Accessed December 24, 2020, https://medium.com/kaggle-blog/march-machine-learning-mania-1st-place-winners-interview-andrew-landgraf-f18214efc659.

Karni E (2009) A mechanism for eliciting probabilities. *Econometrica* 77(2):603–606.

Kearns MJ, Vazirani UV (1994) *An Introduction to Computational Learning Theory* (MIT Press, Cambridge, MA).

Kilgour DM, Gerchak Y (2004) Elicitation of probabilities using competitive scoring rules. *Decision Anal.* 1(2):108–113.

Konrad KA (2009) *Strategy and Dynamics in Contests* (Oxford University Press, Oxford, UK).

Lambert NS (2018) Probability elicitation for agents with arbitrary risk preferences. Working paper, Stanford University, Stanford, CA.

Lambert N, Langford J, Wortman J, Chen Y, Reeves D, Shoham Y, Pennock DM (2008) Self-financed wagering mechanisms for forecasting. Sandholm T, Riedl J, Fortnow L, eds. *Proc. 9th ACM Conf. on Electronic Commerce* (ACM, New York), 170–179.

Lichtendahl KCJ, Winkler RL (2007) Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.* 53(11):1745–1755.

Lichtendahl KC, Grushka-Cockayne Y, Pfeifer PE (2013) The wisdom of competitive crowds. *Oper. Res.* 61(6):1383–1398.

Machina MJ, Schmeidler D (1992) A more robust definition of subjective probability. *Econometrica* 60(4):745–780.

McCarthy J (1956) Measures of the value of information. *Proc. National Acad. Sci. USA* 42(9):654–655.

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psych. Sci.* 25(5):1106–1115.

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. *Management Sci.* 65(5):1949–2443.

Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH (2014) Combining multiple probability predictions using a simple logit model. *Internat. J. Forecasting* 30(2):344–356.

Savage LJ (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66:783–801.

Schervish MJ (1989) A general method for comparing probability assessors. *Ann. Statist.* 17(4):1856–1879.

Selten R (1998) Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* 1:43–61.

Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B (2004) Prediction markets: Does money matter? *Electronic Marketing* 14(3):243–251.

Tetlock PE, Gardner D (2015) *Superforecasting: The Art and Science of Prediction* (Crown Publishing Group, New York).

Witkowski J, Freeman R, Wortman Vaughan J, Pennock DM, Krause A (2018) Incentive-compatible forecasting competitions. McIlraith S, Weinberger K, eds. *Proc. 32nd AAAI Conf. on Artificial Intelligence* (AAAI Press, Palo Alto), 1282–1289.