Hedging and Approximate Truthfulness in Traditional Forecasting Competitions

Mary Monroe, Anish Thilagar, Melody Hsu, Rafael Frongillo

University of Colorado Boulder

Abstract

In forecasting competitions, the traditional mechanism scores the predictions of each contestant against the outcome of each event, and the contestant with the highest total score wins. While it is well-known that this traditional mechanism can suffer from incentive issues, it is folklore that contestants will still be roughly truthful as the number of events grows. Yet thus far the literature lacks a formal analysis of this traditional mechanism. This paper gives the first such analysis. We first demonstrate that the "long-run truthfulness" folklore is false: even for arbitrary numbers of events, the best forecaster can have an incentive to hedge, reporting more moderate beliefs to increase their win probability. On the positive side, however, we show that two contestants will be approximately truthful when they have sufficient uncertainty over the relative quality of their opponent and the outcomes of the events, a case which may arise in practice.

Extended version — https://arxiv.org/abs/2409.19477

1 Introduction

In a forecasting competition, multiple forecasters submit predictions on multiple events or held-out data points. Prominent examples include geopolitical forecasting tournaments like the Good Judgment Project and the Hybrid Forecasting Competition, and data science competitions such as those hosted on Kaggle. The most popular mechanism platforms use to select a winner is the traditional mechanism we call Simple Max: after scoring each forecaster's predictions against the realized event outcomes, the forecaster with the highest total score wins.

While intuitive, it is well-established that Simple Max is not a truthful mechanism (Lichtendahl and Winkler 2007; Witkowski et al. 2023). To maximize their chance of winning the competition, forecasters have incentives to submit predictions which differ substantially from their true beliefs. Several mechanisms have been proposed to address this problem, such as Event Lotteries Forecasting (Witkowski et al. 2023) and Follow the Regularized Leader (Frongillo et al. 2021). Yet implementation of these proposals remains rare, while Simple Max continues to enjoy widespread popularity despite its incentive issues.

The ubiquity of Simple Max may be partly due to folklore that the advantage of deviating from reporting truthfully is attenuated by the large number of events typically used, relative to the number of forecasters. For example, Aldous (2019) states that "in the long run it is best to be 'honest.'" However, the community lacks any theoretical analysis backing up this claim. Indeed, we lack *any* strategic analysis of the Simple Max mechanism beyond a single event—a major omission given its ubiquity.

In this paper, we give the first strategic analysis of Simple Max for multiple events, with a goal of understanding when the long-run-truthfulness folklore does and does not hold. We first identify a natural example where folklore fails: when the best forecaster is much better than the rest. Here forecasters will not be truthful in equilibrium, even for an arbitrarily large number of events. In particular, it is strictly better for the best forecaster to *hedge* their report towards the others' reports to reduce the variance of their scores.

We next study the two-forecaster case and identify a regime where contestants *will* be approximately truthful: when they both believe they have a shot at winning but have sufficient uncertainty about the report of their opponent. These conditions can be difficult to achieve if the number of events is large, as the scores of forecasters with fixed skill will separate and the worse forecaster will no longer have a shot at winning. Therefore we expect our result to apply in settings where the number of events is not too large, such as geopolitical forecasting competitions, or in data science competitions with sufficient noise in the leaderboard.

Taken together, our results show that behavior under the Simple Max mechanism is not as straightforward as folklore claims. If platforms care about eliciting true beliefs from contestants, our counterexample shows that there are settings they need to avoid. Meanwhile, our approximate truthfulness result shows that, at least in some regimes, forecasters will behave in accordance with their beliefs. We conclude with several other implications and future work.

1.1 Related Work

Non-truthful behavior in forecasting competitions under Simple Max is well-documented in previous work. Lichtendahl and Winkler (2007) first showed that forecasters extremize their reports when they want to maximize the chance they win under Simple Max. Witkowski et al. (2018, 2023)

note that no matter the total number of events, strategic forecasters may still extremize on some subset of them. These works demonstrate the existence of non-truthfulness, but fall short of directly challenging folklore, as the results make no claims about the significance of these deviations in the long run. For example, even if forecasters extremize on a small subset of events, their average deviation from truthful may still be small. By contrast, we demonstrate consistent non-truthful behavior in arbitrarily large competitions. We also show regimes where approximate truthfulness still holds.

2 Model

In a forecasting competition, there are m independent binary events $Y_1, \ldots, Y_m \in \{0, 1\}$, each with bias $\theta_t = \mathbb{E}[Y_t]$. Let $\mathcal{I} = [0, 1]$ be the unit interval, and $\mathcal{I}^m$ be the m-dimensional unit cube. There are n forecasters; each forecaster i submits report $r_{it} \in \mathcal{I}$ for event t. Let $\boldsymbol{r}_i \in \mathcal{I}^m$ be the vector of forecaster i's reports across events. We define $\boldsymbol{r}_{i,-t}$ as the vector of forecaster i's reports for all events except t, and $\boldsymbol{r}_{-i}$ as the set of report vectors for all forecasters $j \neq i$. Each event Y_t is sampled to obtain the outcome vector $\boldsymbol{y} \in \{0, 1\}^m = \mathcal{Y}$. Finally, the competition uses some mechanism to declare a singular winner using the set of all reports $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ and outcomes $\boldsymbol{y}$. Forecasters may play pure strategies, always choosing the same report $\boldsymbol{r}_i$, or mixed strategies, where they sample $\boldsymbol{r}_i$ from some distribution σ_i.

Definition 1 (Strategy)**.** *A strategy for player i is a probability distribution $\sigma_i \in \Delta(\mathcal{I}^m)$ over the set of possible report vectors from which $\boldsymbol{r}_i$ is sampled. A strategy profile $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is the vector of strategies for all players.*

2.1 The Simple Max Mechanism

The traditional forecasting competition mechanism, which we call *Simple Max*, scores each forecaster on every event and selects the forecaster with the highest aggregate score, breaking ties uniformly.

Definition 2 (Simple Max)**.** *Given reports $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n) \in (\mathcal{I}^m)^n$ and an outcome vector $\boldsymbol{y} \in \mathcal{Y}$, define the* winning set *to be the set*

$$\textit{winner}(\boldsymbol{r_1}, \ldots, \boldsymbol{r}_n, \boldsymbol{y}) = \operatorname*{arg\,max}_{i \in \{1, \ldots, n\}} \sum_{t=1}^{m} S(r_{it}, y_t)\,. \quad (1)$$

The Simple Max *mechanism is the randomized mechanism $\textit{SimpleMax}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{y}) \in \Delta_n$ that chooses a forecaster i as the winner with probability*

$$\textit{SimpleMax}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{y})_i = \frac{\mathbf{1}\{i \in \textit{winner}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{y})\}}{|\textit{winner}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{y})|}. \quad (2)$$

When considering strategic players, we will define their utility to be exactly this win probability $U_i(\boldsymbol{r}_i; \boldsymbol{r}_{-i}, \boldsymbol{y}) = \text{SimpleMax}(\boldsymbol{r}, \boldsymbol{y})_i$. We will exclusively focus on the case of the quadratic score, $S(r, y) = 1 - (r - y)^2$, as is standard in the literature. Because of this choice of S, the winning set is equivalent to the set of reports closest in Euclidean distance to $\boldsymbol{y}$, i.e., $\text{winner}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{y}) = \arg\min_i \|\boldsymbol{r}_i - \boldsymbol{y}\|_2$. We will use this geometric perspective often.

2.2 Approximate Truthfulness

To measure the degree of truthfulness of the mechanism, we quantify how far the forecasters' reports deviate from their beliefs. In particular, we define two notions of the *approximate truthfulness* of agents' reports, based on ℓ^2 and ℓ^∞ distances.[1]

Definition 3 (Approximately truthful)**.** *A report vector $\boldsymbol{r}_i$ for forecaster i is γ-ℓ^2 approximately truthful if $\frac{1}{\sqrt{m}}\|\boldsymbol{r}_i - \boldsymbol{p}_i\|_2 \leq \gamma$. A strategy σ_i is a γ-ℓ^2 truthful strategy for forecaster i when it is γ-ℓ^2 truthful with probability 1, i.e., a strategy profile is γ-ℓ^2 approximately truthful when each of its elements is a γ-ℓ^2 approximately truthful strategy. A report $\boldsymbol{r}_i$ is γ-ℓ^∞* approximately truthful *if $\|\boldsymbol{r}_i - \boldsymbol{p}_i\|_\infty \leq \gamma$.*

Note that ℓ^∞ is a much stronger notion of approximate truthfulness than ℓ^2, as it tightly bounds the maximum deviation of a report on any single event, whereas the ℓ^2 notion only requires that the total deviation across all reports is within some ℓ^2 ball.

2.3 Bayesian Forecasters

We consider a Bayesian belief model where forecasters receive some information about each event via a signal. Formally, each forecaster i receives a signal $X_{it} \in \mathcal{X}_{it}$ for event t. We denote all of forecaster i's signals by $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im})$, and the set of signals for event t by $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{nt})$. The X_{it} are independent across all i and t. Furthermore, all forecasters share a common prior $\mathbb{E}[Y_t]$ of the outcome Y_t and the set of all X_{it}. When they receive their signal $x_{it} \sim X_{it}$, each forecaster Bayesian updates to their posterior $p_{it}(X_{it}) = \mathbb{E}[Y_t | X_{it} = x_{it}]$. We define the ground truth probability as $\theta_t(\boldsymbol{x}_t) = \mathbb{E}[Y_t | \boldsymbol{X}_t = \boldsymbol{x}_t]$, the posterior of a Bayesian who has seen all signals. We suppress the arguments of p_{it} and θ_t when they are clear.

3 Hedging: a Counterexample to Folklore

The community has long understood the lack of incentive compatibility under Simple Max for small numbers of events. Moreover, winning forecasters in real competitions have repeatedly reported more extreme probabilities to increase their chance of winning (Kaggle 2017; Alexander 2023). Fundamentally, reporting a more extreme report increases the variance of one's score, which can increase the likelihood of winning despite a decrease in the expected score (Figure 1). Folklore, such as Aldous (2019), posits that such behavior is an artifact of having too few events, and forecasters should converge to truthful reporting as the number of events increases. Many forecasting competitions run in practice use Simple Max under the assumption that with enough events the best forecasters will still be truthful.

We show that this folklore is not generally the case. In particular, while bad forecasters may extremize, even good forecasters who perfectly know $\vec{\theta}$ may intentionally report less extreme probabilities to increase their chance of winning, a practice we call *hedging*. The tradeoff of hedging is

[1] These definitions of approximate truthfulness, as in Frongillo et al. (2021), are stronger than the typical one which requires that *utility* is approximately optimized by truthful reporting.

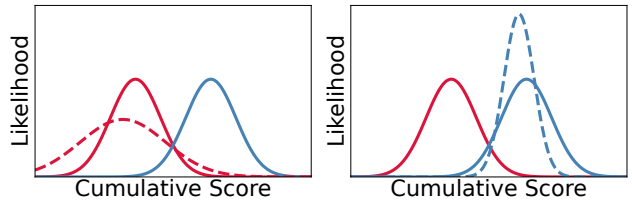

Figure 1: The score distributions of a good (blue, solid curve on the right in each plot) and bad (red, solid curve on the left in each plot) forecaster. Roughly speaking, the bad forecaster's chance of winning is proportional to the region where the distributions overlap; a larger overlap region increases the bad forecaster's win share and decreases the good forecaster's win share. The left plot shows the bad forecaster can extremize to the dashed red distribution: despite lowering their mean, this *increases* their variance, and thus their win share. The right plot shows the good forecaster can similarly increase their win share by hedging to the blue dashed distribution, which despite lowering their mean, also *decreases* their variance. Intuitively, the good forecaster benefits by hedging because it decreases the variance of their score, "locking in" their lead, even while decreasing their expected score.

similar to that of extremizing, except now a forecaster *decreases* the variance of their final score at the cost of lowering their expected score. When a forecaster's expected score is already far ahead of the competition, hedging increases their win probability. To draw an analogy from sports: near the end of a game, the team which is behind will opt for high-risk, high-reward plays (e.g. more 3-point attempts in basketball) to increase their variance, while the team which is ahead will play conservatively (e.g. run down the clock) to decrease their variance and lock in their lead.

To our knowledge, our result is the first to show that such hedging deviations can be beneficial to knowledgeable players in this setting. In particular, we will show that when they are far ahead, Bayesian forecasters' equilibrium reports will be bounded away from their true beliefs. Contrary to folklore, this bound grows as the number of events increases. The formal proof is in the extended version of the paper.

3.1 Setting

We use a specific instance of the Bayesian model with p-biased coins for some $p \in (0, 1/2)$. We can always observe the same behavior without the Bayesian assumption, but it is less clear that all agents will play equilibrium strategies. Each event Y_t is a random coin flip whose bias is $\theta_t \in \{p, 1-p\}$ chosen independently and uniformly at random. The prior $\mathbb{E}[Y_t] = 1/2$ for all t. We fix some i to be the *informed* forecaster, while all other forecasters $j \neq i$ remain uninformed. Each such forecaster j always receives *no* signal for all events, so their belief remains $\boldsymbol{p}_j = \boldsymbol{c} := (1/2, \ldots, 1/2)$, the center of the m-dimensional hypercube. However, forecaster i receives a signal that is exactly the true probabilities $\mathcal{X}_i = \boldsymbol{\theta} \in \{p, 1-p\}^m$. This gap in information, and knowledge of it, is key to the hedging behavior we observe.

For mathematical convenience and without loss of generality, we invert the outcome of all events t satisfying $\theta_t = 1 - p$ and let $\boldsymbol{\theta} = \boldsymbol{p} = (p, \ldots, p)$. Any specific set of reports $\boldsymbol{r}_{-i}$ is then equivalent to the mixed strategy that randomly applies a reflection of the m-hypercube[2] to $\boldsymbol{r}_{-i}$, since all such reflections are equally likely. We consider mixed strategy equilibria as pure strategy equilibria do not generally exist, even for simple cases like $m = 2, n = 3$. We are particularly interested in when players play strategies that are close to their beliefs, which we measure using the notions of approximate truthfulness in Definition 2.

3.2 Exact Equilibria

For sufficiently small choices of m and n, we can exactly characterize the equilibria. We start with an extremely simple case: that of a single event ($m = 1$). Formal derivations are deferred to the extended version of the paper.

We first consider just $n = 2$ forecasters. Here, the equilibrium is for i to report $r_i = 0$, and j to maximally extremize and choose $r_j \in \{0, 1\}$ uniformly at random. This gives the forecasters win probabilities of $U_i = 3/4 - p/2$ and $U_j = 1/4 + p/2$ respectively.

As we increase the number of forecasters, the incentive to extremize only increases. i will continue to play $r_i = 0$ and win with probability

$$\frac{1}{2^{n-1}}\left[\sum_{k=0}^{n-1}\frac{\binom{n-1}{k}}{k+1} + \frac{p}{n+1}\right] = \frac{2}{n} - \frac{1}{2^{n-1}}\left[\frac{1}{n} - \frac{p}{n+1}\right],$$

while forecasters $j \neq i$ report a vertex at random and evenly split the remaining winshare.

When $m = 2$, the equilibrium gets much more complex. We describe it for $n = 2$ forecasters when $p \in (1/3, 1/2)$, as shown in Figure 2. In this case, i will play $\boldsymbol{r}_i = \boldsymbol{c} = (1/2, 1/2)$, the center of the unit square, with probability $\frac{1-p}{2-p}$, and choose some $\boldsymbol{r}_i \in \{(0, 1/2), (1/2, 0)\}$ with probability $\frac{1}{2(2-p)}$ each. Forecaster j will play $\boldsymbol{r}_j = \boldsymbol{c}$ with probability $3 \cdot \frac{p}{2-p}$, and each of $\boldsymbol{r}_j \in \{1/2 \pm 1/4\}^2$ with probability $\frac{1/2-p}{2-p}$. While forecaster j is extremizing with some non-zero probability, their average report is still truthful. On the other hand, forecaster i alternates between hedging to the center and extremizing to the midpoints. Their average report is $\left(\frac{3-2p}{4(2-p)}, \frac{3-2p}{4(2-p)}\right)$. So when $p < \frac{5-\sqrt{13}}{4}$, i's average report is hedged towards the center. Otherwise, their average report is slightly extremized.

3.3 Hedging

As the number of events and players grow larger, it remains an open question to find an equilibrium even in this specific setting. Our focus instead will be on determining whether an approximately truthful equilibrium can exist. In Corollary 1 we answer this question negatively: an equilibrium cannot be ℓ^2-approximately truthful even in the limit of large m.

[2]Formally, any involution in the hyperoctahedral group B_m.

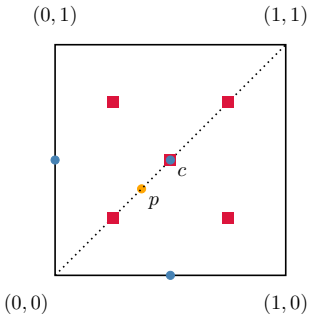

Figure 2: The equilibrium strategies of i and j for $m = n = 2$ and $p \in (1/3, 1/2)$. The blue circles denote points i plays and the red squares denote j's. Notably both players play in the center $\boldsymbol{c}$ some of the time but i always plays bounded away from their belief $\boldsymbol{p}$ (orange point).

In particular, Theorem 1 shows that by hedging and reporting $\boldsymbol{r}^* = (p^*, \ldots, p^*)$, where $p^* = \frac{1/2+p}{2}$, forecaster i can only increase their win probability. Theorem 1 holds even for relatively large values of ϵ. For example, when $p = 0.9$ and $m \geq 32$, deviating by 0.2 in every coordinate is strictly better than *any* 0.04 ℓ^2-approximately truthful strategy.

Our proof is geometric, leveraging the fact that a forecaster wins when their report is the closest to the outcome in Euclidean distance, as discussed in § 2.1. Given any ϵ-ℓ^2 approximately truthful reports $\boldsymbol{r}$ for all forecasters, we show that hedging to $\boldsymbol{r}^*$ is a strictly dominant strategy for i. Specifically, given any vertex $\boldsymbol{y}$, we show that either $\boldsymbol{r}^*$ is closer to $\boldsymbol{y}$ than any $\boldsymbol{r}_j$ for $j \neq i$, or $\boldsymbol{r}^*$ is closer to $\boldsymbol{y}$ than $\boldsymbol{r}_i$ is to $\boldsymbol{y}$. See Figure 3. The implication is that $\boldsymbol{r}^*$ wins on a strict superset of outcomes that $\boldsymbol{r}_i$ wins.

As shorthand, let us define the following distances, all from a report vector to some vertex $\boldsymbol{y}$: $d_p(\boldsymbol{y}) = \|\boldsymbol{p} - \boldsymbol{y}\|_2$, $d^*(\boldsymbol{y}) = \|\boldsymbol{r}^* - \boldsymbol{y}\|_2$, $d_i(\boldsymbol{y}) = \|\boldsymbol{r}_i - \boldsymbol{y}\|_2$, $d_j(\boldsymbol{y}) = \|\boldsymbol{r}_j - \boldsymbol{y}\|_2$.

To bound these distances away from each other, we need ϵ such that the balls of radius $\epsilon\sqrt{m}$ are bounded away from the arc of radius $d^*(\boldsymbol{y})$ about $\|\boldsymbol{y}\|_1 = p^*m$ (the arc about the red point in Figure 3). Therefore, we choose m large enough such that there is enough space in the circle of radius $\sqrt{m}/2$ for the arc to lie above $\boldsymbol{c}$ and $\boldsymbol{p}$ (Condition 1(a)). Furthermore, to show strict dominance the arc must be at least distance 2 from the balls. (Note that all the distances scale as $\sqrt{m}$, so the gap between the arc and the balls gets proportionally smaller for large m). Therefore, we must choose p such that there is enough space between the arc and the main diagonal for a gap of distance 2 to exist (Condition 1(b)). The following condition suffices for all of these relationships to hold.

Condition 1.

(a) $m \geq 21$

(b) $0 < p < \frac{1}{2} - 2\sqrt{\frac{2}{\sqrt{m}}(1 - \frac{2}{\sqrt{m}})}$

(c) $0 < \epsilon < \frac{1}{2} - \sqrt{p^*(1-p^*)} - \frac{2}{\sqrt{m}}$.

Note that such an ϵ will always exist when the first two inequalities hold. To show that $\boldsymbol{r}^*$ dominates $\boldsymbol{r}_i$ we consider two cases. When $\boldsymbol{y}$ lies to the left of $\boldsymbol{r}^*$, $\boldsymbol{r}^*$ wins each $\boldsymbol{r}_j$, as shown by the blue lines in Figure 3, so $\boldsymbol{r}_i$ cannot strictly dominate $\boldsymbol{r}^*$.

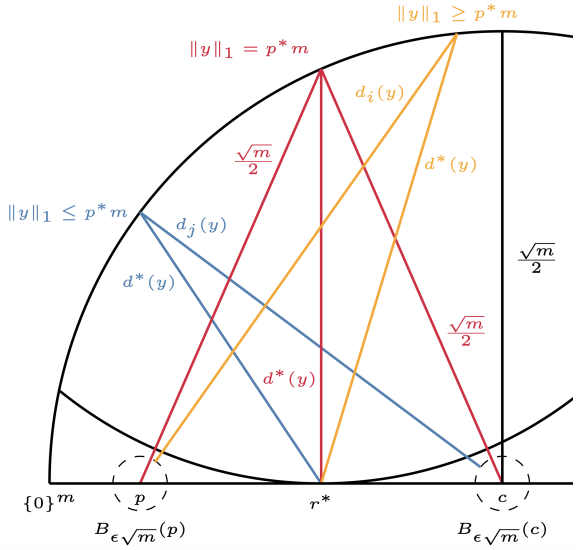

Figure 3: The geometry of the plane containing $\boldsymbol{p}$, $\boldsymbol{c}$, $\boldsymbol{r}^*$ and some $\boldsymbol{y}$. The bottom line is the main diagonal of the m-dimensional hypercube that goes from $\{0\}^m$ to $\{1\}^m$. $\boldsymbol{r}^*$ is the midpoint of i's belief $\boldsymbol{p}$ and the belief of the other forecasters $\boldsymbol{c}$, so i is moving to the middle of the "information gap." The radius of the balls around $\boldsymbol{p}$ and $\boldsymbol{c}$ are chosen so that they are bounded away from the ball of radius $d^*(\boldsymbol{y})$ around a $\boldsymbol{y}$ that lies perpendicular to the diagonal at $\boldsymbol{r}^*$ (red point). Lemma 1 considers $\|\boldsymbol{y}\|_1 \leq p^*m$ (blue point), while Lemma 2 considers $\|\boldsymbol{y}\|_1 \geq p^*m$ (orange point). The high level idea is that forecaster i can shift from winning roughly just the $\boldsymbol{y}$ to the left of $\boldsymbol{r}^*$ when reporting $\boldsymbol{r}_i$, to additionally winning some fraction of the $\boldsymbol{y}$ between $\boldsymbol{r}^*$ and $\boldsymbol{c}$ by hedging to $\boldsymbol{r}^*$.

Lemma 1. *Suppose m, p and ϵ satisfy Condition 1. Then for any $\boldsymbol{y} \in \mathcal{Y}$ with $\|\boldsymbol{y}\|_1 \leq p^*m$, if $\|\boldsymbol{r}_j - \boldsymbol{c}\|_2 < \epsilon\sqrt{m}$ then $d^*(\boldsymbol{y}) < d_j(\boldsymbol{y}) - 2$.*

When $\boldsymbol{y}$ lies to the right of $\boldsymbol{r}^*$, $\boldsymbol{r}^*$ dominates $\boldsymbol{r}_i$, as shown by the orange lines in Figure 3.

Lemma 2. *Suppose m, p and ϵ satisfy Condition 1. Then for any $\boldsymbol{y} \in \mathcal{Y}$ with $\|\boldsymbol{y}\|_1 \geq p^*m$, if $\|\boldsymbol{r}_i - \boldsymbol{p}\|_2 < \epsilon\sqrt{m}$ then $d^*(\boldsymbol{y}) < d_i(\boldsymbol{y}) - 2$.*

Jointly, they show that when all players report ϵ-ℓ^2 truthfully, $\boldsymbol{r}^*$ dominates $\boldsymbol{r}_i$.

Lemma 3. *Suppose m, p and ϵ satisfy Condition 1 and each forecaster chooses an ϵ-ℓ^2 approximately truthful report. Then for every $\boldsymbol{y} \in \mathcal{Y}$, $U_i(\boldsymbol{r}_i, \boldsymbol{r}_{-i}, \boldsymbol{y}) \leq U_i(\boldsymbol{r}^*, \boldsymbol{r}_{-i}, \boldsymbol{y})$.*

Furthermore, because of the strict gap of distance 2 between the balls, there must be some $\boldsymbol{y}$ where $\boldsymbol{r}^*$ strictly dominates $\boldsymbol{r}_i$.

Lemma 4. *Suppose m, p and ϵ satisfy Condition 1 and each forecaster chooses an ϵ-ℓ^2 truthful report. Then there exists some $\boldsymbol{y} \in \mathcal{Y}$ such that $U_i(\boldsymbol{r}_i, \boldsymbol{r}_{-i}, \boldsymbol{y}) < U_i(\boldsymbol{r}^*, \boldsymbol{r}_{-i}, \boldsymbol{y})$.*

Taken together, the previous results imply that hedging to $\boldsymbol{r}^*$ strictly dominates $\boldsymbol{r}_i$ when any ϵ-ℓ^2 approximately truthful strategy is played.

Theorem 1. *Suppose m, p and ϵ satisfy Condition 1. Let $\boldsymbol{r}_i$ be any ϵ-ℓ^2 approximately truthful report for forecaster i and σ_{-i} be an ϵ-ℓ^2 approximately truthful strategy profile for all $j \neq i$. Then $\boldsymbol{r}_i$ is strictly dominated by hedging and reporting $\boldsymbol{r}^*$.*

In particular, Condition 1 implies $p^* > p + \epsilon$, so i will always strictly benefit by hedging to $\boldsymbol{r}^*$ from any ϵ-ℓ^2 approximately truthful strategy. Therefore, no ϵ-ℓ^2 approximately truthful equilibrium exists.

Corollary 1. *Suppose m, p and ϵ satisfy Condition 1. Then, every equilibrium strategy profile $\boldsymbol{\sigma}$ is not ϵ-ℓ^2 approximately truthful.*

4 Achieving Approximate Truthfulness

Though we have shown regimes where forecasters deviate from their beliefs, we now prove Simple Max is still approximately truthful in settings which may arise in practice. These results do not identify specific equilibria, and instead place conditions on the beliefs about the reports of their opponents. Our results are practical in actual competitions, in the sense that forecasters do form beliefs about the reports of their opponents (either by reasoning about equilibrium behavior or looking at historical data (Kaggle 2017)) and those beliefs could plausibly satisfy these conditions.

Our main result, Theorem 2, considers two forecasters i and j, and shows that both will report approximately truthfully when they have sufficient uncertainty over each others' reports. Specifically, we require that (1) the expected score difference is smooth, (2) any report vector induces enough variance in the score, and (3) forecaster i does not think their score will be much larger than j's, or vice versa. In real-world competitions, all 3 conditions are quite reasonable. We use this uncertainty to show that each player's utilities are approximately affine in their score. Therefore, to maximize their utility, they will roughly aim to maximize their score. This leads to *leave-one-out approximate truthfulness*: fixing event t, the optimal report satisfies $r_{it} \approx p_{it}$ for *any* fixed report vector $\boldsymbol{r}_{i,-t}$ across events $t' \neq t$. The strength of that guarantee leads to the strong ℓ^∞ version of approximate truthfulness. However, significant groundwork is required to quantify "enough" uncertainty, and demonstrate how it propagates from beliefs to utilities to approximately truthful reports. The remainder of this section outlines our strategy for proving Theorem 2; a complete technical proof is available in the extended version of the paper.

4.1 Notation

In this section, we use σ to denote variance. Let the random variable $R_{jt} \in [0,1]$ be j's report for event t, and let $\boldsymbol{R}_j = (R_{j1}, R_{j2}, \ldots, R_{jm})$ be j's random report vector. We model forecaster i's belief as a joint distribution $\mathcal{D}_i$ over event outcomes $\boldsymbol{Y}$ and forecaster j's reports $\boldsymbol{R}_j$, as in Witkowski et al. (2023). We denote $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{im})$ as the marginal distribution under $\mathcal{D}_i$ over outcomes $\boldsymbol{Y}$. We use this belief model partly as a shorthand in notation, but it is also a more general version of the Bayesian setting introduced earlier and thus encapsulates a broader range of scenarios.

4.2 Utilities

To distinguish between different utility functions, we will use subscripts to denote the variable of interest and overlines to denote expectations. Let $\overline{U}_i(\boldsymbol{r}_i) = \mathbb{E}_{\mathcal{D}_i} U_i(\boldsymbol{r}_i; \boldsymbol{R}_j, \boldsymbol{Y})$ be forecaster i's expected utility under their belief distribution as a function of their report vector $\boldsymbol{r}_i$. We also define $U_{it}(r_{it}; \boldsymbol{r}_{i,-t}, \boldsymbol{R}_j, \boldsymbol{Y}) = U_i(\boldsymbol{r}_i; \boldsymbol{R}_j, \boldsymbol{Y})$ as a function of just r_{it}, fixing $\boldsymbol{r}_{i,-t}$. Let

$$\overline{U}_{it}(r_{it}; \boldsymbol{r}_{i,-t}, R_j, Y) = \mathop{\mathbb{E}}_{(\boldsymbol{R}_j, \boldsymbol{Y}) \sim \mathcal{D}_i} U_{it}(r_{it}; \boldsymbol{r}_{i,-t}, \boldsymbol{R}_j, \boldsymbol{Y})$$

be i's expected utility as a function of r_{it}, over all outcomes and reports of forecaster j. When the right hand terms are clear we will simply use $U_{it}(r_{it})$ and $\overline{U}_{it}(r_{it})$.

For any t, let $S(r_{it}, R_{jt}, Y_t) = S(R_{jt}, Y_t) - S(r_{it}, Y_t)$ denote the difference between j and i's score on event t. We impose a smoothness assumption on the sum of score differences $\sum_t S(r_{it}, R_{jt}, Y_t)$ to ensure that the probability of a tie is zero.

Assumption 1. *For every forecaster i and any report $\boldsymbol{r}_i \in \mathcal{I}^m$, their expected score difference CDF $G_i(x; \boldsymbol{r}_i) = \Pr_{D_i}\left[\sum_t S(r_{it}, R_{jt}, Y_t) \leq x\right]$ is absolutely continuous.*

As long as one event's score difference is absolutely continuous, the assumption is satisfied. Though it may seem unrealistic to assume a distribution over reports is continuous, note that our analysis would easily extend to settings without this assumption since ties occur with vanishingly small probability in practice; a full discussion can be found in the extended version of the paper.

Under Assumption 1, $\Pr_{\mathcal{D}_i}[|\text{winner}(\boldsymbol{r}_i, \boldsymbol{R}_j, \boldsymbol{y})| > 1] = 0$. It follows that each forecaster's expected utility is simply their probability of being in the winning set under Simple Max. That is, the expected utility satisfies $\overline{U}_i(\boldsymbol{r}_i) = G_i(0; \boldsymbol{r}_i)$. Moreover, $\overline{U}_{it}(r_{it})$ corresponds to the expected probability that $-S(r_{it}, R_{jt}, Y_t)$ is greater than the cumulative score difference across events $t' \neq t$. Specifically, if

$$G_{it}(x) = \Pr_{(\boldsymbol{R}_{j,-t}, \boldsymbol{Y}_{-t}) \sim \mathcal{D}_i}\left[\sum_{t' \neq t} S(r_{it'}, R_{jt'}, Y_{t'}) \leq x\right]$$

is the CDF of the score difference distribution for $t' \neq t$,

$$\overline{U}_{it}(r_{it}) = \mathop{\mathbb{E}}_{(R_{jt}, Y_t) \sim \mathcal{D}_i} G_{it}\left(-S(r_{it}, R_{jt}, Y_t)\right) . \quad (3)$$

4.3 Approximate Affineness

We first derive sufficient conditions for forecaster i to be γ-ℓ^∞ approximately truthful on event t, given a fixed report vector $\boldsymbol{r}_{i,-t}$. In particular, we show that forecaster i's utility function G_{it} is *approximately affine*.

Definition 4. *Function F is $(\beta, \alpha, \epsilon)$-*approximately affine *when*

$$\sup_{x \in [-1,1]} |F(x) - (\beta x + \alpha)| \leq \epsilon.$$

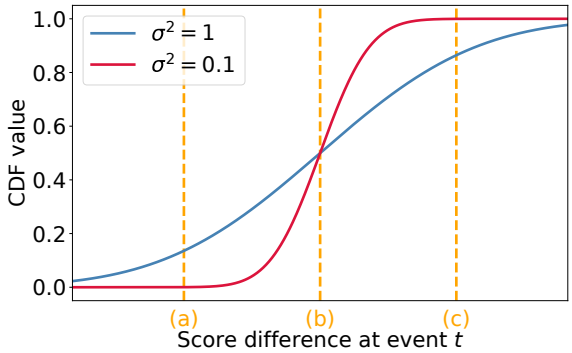

Figure 4: Two examples of the function G_{it}, the CDF of the score difference distribution for events $t' \neq t$. The approximate affine error is large if the score difference at event t (in $[-1, 1]$) occurs at (a) or (c) since the slope flattens out. We observe small error around (b), the mean, but note the area where the CDF does not flatten out is narrower for the distribution with lower variance (red). Thus the approximation error is lower when the magnitude of the mean is small relative to the variance.

Note that we are evaluating G_{it} at the score difference $S(r_{it}, Y_t) - S(R_{jt}, Y_t)$. Since G_{it} is approximately affine, forecaster i is (roughly) maximizing their score. By properties of proper scoring rules, they will report (roughly) truthfully. Specifically, we can quantify the worst-case deviation of their report to their belief as a function of approximate affineness.

Lemma 5. *If G_{it} is $(\beta, \alpha, \epsilon)$-approximately affine for each event t, agent i is γ-ℓ^∞ approximately truthful for $\gamma = \sqrt{2\epsilon/\beta}$.*

Figure 4 gives intuition for when this result is useful: specifically, approximating G_{it} by a Normal CDF, we achieve small γ when the mean of the cumulative score difference has small magnitude relative to the variance.

Edgeworth expansions. We next aim to characterize forecaster i's utility function over their report r_{it}. Since G_{it} is the CDF of a sum of independent random variables, we can apply the Central Limit Theorem (CLT). Unfortunately, the convergence rate of G_{it} to a normal distribution is not tight enough to control the approximate affine error ϵ in Lemma 5 (Esseen 1942). We thus turn to the theory of *Edgeworth expansions*. An Edgeworth expansion can be thought of as a tighter version of the CLT. While we know G_{it} eventually approaches a normal distribution, we can achieve a better convergence rate by adding on more terms (Petrov 1975). These terms depend on higher cumulants of the distribution; for example, the first-order Edgeworth expansion includes a function depending on the third moment. We can bound the ℓ^∞ distance between G_{it} and its second-order Edgeworth expansion E_{it} under some conditions on the variance and smoothness of score differences:

Condition 2. *For forecaster i's strategic report $\boldsymbol{r}_i$ and all t,*

1. *(Uncertainty) the score differences $S(r_{it}, R_{jt}, Y_t)$ have variance uniformly bounded below by some constant.*
2. *(Smoothness) the probability density of each score difference $S(r_{it}, R_{jt}, Y_t)$ is a function of bounded variation.*[3]

We use a weaker condition which only requires smoothness of some score differences in the extended version of the paper.

Leave-one-out approximate truthfulness. We now have the tools to characterize forecaster i's utility function G_{it} as $(\beta, \alpha, \epsilon)$-approximately affine. In particular, we derive the values of β, α and ϵ for the Edgeworth expansion E_{it} of G_{it}. To do so, we take the first-order Taylor expansion around 0 to derive the slope β and intercept α. We can then bound the error ϵ_1 of the affine Taylor function by the Lagrange remainder. We also have the convergence rate of G_{it} to its Edgeworth expansion, i.e. $\|G_{it} - E_{it}\|_\infty \leq \epsilon_2$. By the triangle inequality, then, G_{it} is $(\beta, \alpha, \epsilon_1 + \epsilon_2)$-approximately affine. It follows by Lemma 5 that forecaster i will be γ-ℓ^∞ approximately truthful for $\gamma = \sqrt{\frac{2(\epsilon_1+\epsilon_2)}{\beta}}$ on event t (under Condition 2). Note that γ depends on the mean and variance of score differences $\{S(r_{it'}, R_{jt'}, Y_{t'})\}_{t' \neq t}$, which in turn depends on the report vector $\boldsymbol{r}_{i,-t}$.

4.4 Approximate Truthfulness

Our leave-one-out result gives a sufficient condition on the value of γ for approximate truthfulness on event t, under *any* fixed report vector $\boldsymbol{r}_{i,-t}$. In order to reach our main result, then, we impose conditions on the entire vector $\boldsymbol{r}_i$ such that the condition for each event t is met. We define $\sigma_i^2(\boldsymbol{r}_i)$ and $P_i(\boldsymbol{r}_i)$ as the variance and absolute third moment of the aggregate score difference $\sum_t (S(r_{it}, R_{jt}, Y_t))$ induced by $\boldsymbol{r}_i$. Let $P_i = \max_{\boldsymbol{r}_i} P_i(\boldsymbol{r}_i)$ and $\sigma_i = \min_{\boldsymbol{r}_i} \sigma_i(\boldsymbol{r}_i)$.

The first condition implies there is enough randomness (variance) in scores from forecaster i's perspective, which ensures leave-one-out approximate truthfulness for each event and also prevents γ from growing arbitrarily large. Meanwhile, the latter two conditions imply that forecaster i thinks they are *good*, as the expected utility of their belief is bounded below, but not *too* good, as the expected utility of any report vector is bounded above.

Condition 3. *1.* $\sigma_i \geq 4$.

2. *For some $\delta \in (0, 1)$, $\overline{U}(\boldsymbol{p}_i) \geq \frac{1}{2} - \delta$; and for any report vector $\boldsymbol{r}_i$, $\overline{U}(\boldsymbol{r}_i) \leq \frac{1}{2} + \delta$.*
3. $\frac{P_i}{\sigma_i^3} + \delta \leq 0.33$.

Note that $\frac{P_i}{\sigma_i^3} = O(1/\sqrt{m})$, so we expect Condition 3 is satisfied more easily as m grows. With that bound, and assuming $\frac{P_i}{\sigma_i^3}$ is small, the latter two items imply forecaster i's expected utility is roughly bounded to the interval $[0.2, 0.8]$.

We can now state our approximate truthfulness result.

Theorem 2. *When Conditions 2 and 3 hold, forecaster i is γ-ℓ^∞ approximately truthful with $\gamma = O\left(1/\sqrt{\sigma_i}\right)$.*

[3] Absolute continuity or continuous differentiability are both sufficient conditions for bounded variation.

Note that $\sigma_i^2 \in O(m)$, so that we expect γ to decrease like $O(m^{-1/4})$. At first glance, then, Theorem 2 seems to imply that the ℓ^∞ distance of forecasters' reports converge to their beliefs as the number of events grows larger. This is true to an extent, but recall that Condition 3 requires that the forecasters believe they are competitive with each other. In a typical setting where both forecasters have fixed skills, e.g. with forecaster i being slightly better at predicting the weather than forecaster j, for large m forecaster j can no longer be competitive: each forecaster's aggregate score would concentrate about its mean and the distribution of forecaster i's score would far surpass j's. Thus, the conditions for approximate truthfulness in practice may require m to be small enough for the best forecaster to have peers.

We expect that similar results would hold for $n > 2$ forecasters, though analysis becomes much more difficult since utilities now correspond to beating the *maximum* cumulative score difference across forecasters. Extending our current approach to this setting would require extremely restrictive assumptions about the correlation and statistics of score difference distributions. We expect the value of γ to grow quickly in n, as even when all forecasters have the same skill, when n is larger, they each expect to perform much worse than the maximum of the rest. This extension would thus be consistent with the fact that good forecasters still extremize in practice; see § 5.

5 Discussion

This paper presents the first strategic analysis of traditional forecasting competitions beyond small numbers of events. We first refute folklore claims about long-run truthful behavior in forecasting competitions, via a counterexample with no approximately truthful equilibrium even for an arbitrarily large number of events. This example suggests that the best forecaster benefits by hedging, moving their predictions closer to the prior. We then show that two forecasters will be approximately truthful when close together in skill and given sufficient uncertainty about the other's reports.

Implications for practice. Let us summarize our results with a practical conjecture of how strategic behavior will play out in real competitions. It seems rare that a single forecaster will be far ahead of the rest; a more typical scenario might have a pack of good forecasters with a long tail of lower-skilled forecasters. It might be tempting to assume that something similar to Theorem 2 will hold for the pack of good forecasters, yielding approximate truthfulness within the leaders, with increasing incentive to extremize for lower-skilled forecasters. Yet as alluded to in § 4.4, when there are more than 2 good forecasters, each must face the maximum score of the rest, again giving an incentive to extremize (consider Figure 1 (left) where now the blue distribution is the maximum score of the competition, which necessarily has a larger mean than any individual skilled opponent). This rough conjecture matches behavior seen in practice, where skilled (and indeed, winning) forecasters extremize their true beliefs (Kaggle 2017; Alexander 2023).

That said, our results suggest ways to alleviate these bad incentives. Even when the number of events is quite large, contestants may deviate significantly from being truthful when they know they are far ahead or far behind. In data science and machine learning competitions, this phenomenon could mean limiting the amount of information revealed in the leaderboard—not just for the usual information-theoretic reasons of preventing contestants from implicitly learning the data set, but also because of the strategic properties of Simple Max. Finally, to reduce the extremizing among the most skilled forecasters, it could be beneficial to run a truncated version of the Soft Max mechanism of Frongillo et al. (2021) which drops low scoring forecasters, with the analysis of § 4 allowing a higher parameter η than suggested by that paper.

Robustness of forecaster selection. The designer of a forecasting competition may naturally have two goals: (1) collect accurate predictions, and (2) select (one of) the best forecasters as the winner. Thus far we have mainly focused on (1), which may be the most important when the forecasts are aggregated or used directly to make timely predictions, as in geopolitical competitions. Yet in some settings goal (2) may be the primary objective. Interestingly, while we have seen that Simple Max can fail wildly in goal (1), even for arbitrarily large numbers of events, our results are consistent with it still succeeding in goal (2).[4] That is, even as forecasters deviate substantially from their beliefs, it has remained true that a best forecaster wins with high probability. Specifically, we conjecture that with probability $1 - \delta$, when the number of events is $m = \Omega\left(\log(n/\delta)/\epsilon^2\right)$, Simple Max selects the best forecaster in some Bayes-Nash equilibrium.

Future work. Aside from addressing any of the specific conjectures posited above, perhaps the most pressing need is to understand the equilibrium of Simple Max for $m > 2$. A full equilibrium analysis is missing even for the specific example in § 3; while our results suggest that the bad forecasters will extremize while the good forecaster hedges, it is unknown if both behaviors are stable in equilibrium. For much larger values of m, numerical experiments have thus far failed to yield a viable extremizing strategy; thus, we conjecture that extremizing is not a good response for the bad forecaster in the limit of the number of events. Perhaps the next step to fully characterize equilibria would be to understand the equilibrium in a "full information" setting, where the true probabilities are common knowledge.

Extensions of this paper's results to forecasting competitions that employ different scoring rules, such as log loss, may also hold interesting insights.

Acknowledgements

We are grateful to Jens Witkowski and Siddarth Srinivasan for their insights and feedback. We would also like to thank Adam Bloniarz, Gülce Kardes, Ezra Karger, Bo Waggoner, and Brian Zaharatos for helpful discussions and comments. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2045347.

[4] A more realistic model of effort may paint a worse picture even for (2), as time is spent strategizing instead of improving forecasts.

References

Aldous, D. J. 2019. A Prediction Tournament Paradox. *The American Statistician*, 1–6.

Alexander, S. 2023. Who Predicted 2022? https://www.astralcodexten.com/p/who-predicted-2022. Accessed: 2024-08-14.

Esseen, C.-G. 1942. On the Liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28: 1–19.

Frongillo, R.; Gomez, R.; Thilagar, A.; and Waggoner, B. 2021. Efficient Competitions and Online Learning with Strategic Forecasters. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, 479–496.

Kaggle. 2017. March Machine Learning Mania, 1st Place Winner's Interview: Andrew Landgraf. http://blog.kaggle.com/2017/05/19/march-machine-learning-mania-1st-place-winners-interview-andrew-landgraf/. Accessed: 6/29/2019.

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

Petrov, V. V. 1975. *Sums of Independent Random Variables*. Springer.

Witkowski, J.; Freeman, R.; Vaughan, J. W.; Pennock, D. M.; and Krause, A. 2018. Incentive-Compatible Forecasting Competitions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.

Witkowski, J.; Freeman, R.; Vaughan, J. W.; Pennock, D. M.; and Krause, A. 2023. Incentive-compatible forecasting competitions. *Management Science*, 69(3): 1354–1374.