Efficient and Truthful Forecasting Competitions

Rafael Frongillo, Robert Gomez, Anish Thilagar, Bo Waggoner

November 24, 2020

Abstract

Witkowski et al. [2018] studied competitions among forecasters who predict a set of common events. When such a competition picks a single winner, incentives can be distorted, so that paper introduced the ELF mechanism which incentivizes truthful forecasts. In this paper, we give a tighter analysis of the ELF mechanism, then give a mechanism that improves upon it exponentially, in the sense of also being truthful but requiring only logarithmically many events in order to select the best forecaster with high probability. These results have potential applications to machine-learning competitions such as Kaggle, where winners are currently selected in non-truthful ways.

1 Introduction

- SHELF
- GNOME
- ???

keywords: event, forecast, voting

EFV

2 Model and Definitions

Our model will consist of n forecasters s.t. $n \geq 2$ where $i \in [1, n]$ denotes the ith forecaster, m independent events where $k \in [1, m]$ denotes the kth event. Each event $k \in [1, m]$ will be associated with a random variable X_k. $X_k = 1$ will denote that event k occurred, and, likewise, $X_k = 0$ will denote event k did not occur. $\forall k \in [m] \exists \theta_k$ s.t. θ_k denotes the true, unknown probability of event k occurring; $\theta_k = P(X_k = 1) = 1 - P(X_k = 0)$. We will denote each outcome drawn as $x_i \sim X_i$. For each forecaster i, we denote their true belief that event k will occur as $\hat{\theta}_{i,k} \in (0, 1)$. We then denote $y_{i,k} \in [0, 1]$ as the actual report forecaster i makes for event k. Note that $y_{i,k}$ need not equal $\hat{\theta}_{i,k}$.

After every event $k \in [1, m]$ has occurred, and thus each forecaster has reported $y_{i,k} \forall i f \forall k$, a mechanism will select exactly one forecaster from the set of $[n]$ as the 'best forecaster'.

Definition 1. *A **selection mechanism** $M(y, x)$ is a map from the forecasts $y = (y_1, ..., y_n)$ and outcomes of the events $x = (x_1, ..., x_k)$ to an index $w \in [1, n]$.*

We say that forecaster w wins the competition if $M(y, x) = w$. We call a forecaster selection mechanism truthful iff, for all forecasters, their probability of winning is maximized when they reports their beliefs as their forecasts.

Definition 2. *M is **truthful** iff $\forall i \in [n]$, $\forall \hat{\theta}_i$, $\forall y_i' \neq \hat{\theta}_i$, and $\forall y_j s.t. j \neq i$,*

$$P(M((y_1, ..., \hat{\theta}_i, .., y_n), x) = i) \geq P(M((y_1, ..., \hat{y}_i, .., y_n), x) = i)$$

*M is **strictly truthful** if the strict inequality inequality holds.*

Additionally, we would like our mechanism to choose the forecaster with the best predictions. We define the accuracy of a single forecaster i as 1 minus the average squared loss of their forecasts compared to the true probabilities.

Definition 3. *A forecaster's **accuracy** is $a_i = 1 - \frac{1}{m} \sum_{k=1}^{m} (y_{i,k} - \theta_k)^2$.*

An ideal mechanism would choose forecasters with larger accuracies with higher probabilities. Let $i^* = \arg\max_i a_i$ be the highest accuracy forecaster, and let $a^* = a_{i^*}$.

Definition 4. *The **accuracy gap** $\epsilon = a^* - \max_{i \neq i^*} a_i$ is the difference in accuracy of the first and second best forecasters.*

For a given mechanism M and n forecasters with accuracy gap ϵ, we let m_δ be the minimum number of events such that M necessarily chooses i^* with probability at least $1 - \delta$. Specifically, for any $m > m_\delta(M, n, \epsilon)$, $P(M(y, x) \neq i^*) \leq \delta$. m_δ gives us a

measure of how many events are needed for a mechanism to pick the best forecaster. We will use it as a metric to compare mechanisms, where we say M is more efficient than M' if $m_\delta(M, n, \epsilon) < m_\delta(M', n, \epsilon)$, since M needs less events to consistently pick the best forecaster.

Similar to work in Witkowski et al. [2018], we will focus on using the proper scoring rule, the quadratic scoring rule in our forecaster selection mechanism.

Definition 5. *A **scoring rule**, R is proper if* $\forall p, y \in 0, 1, E_{x\ p}[R(p, x)] \geq E_{x\ p}[R(y, x)]$

Definition 6. *The **quadratic scoring rule**, is $R_q = 1 - (y - x)^2$, where x denotes the observed outcome, and y denotes the report.*

2.1 Useful facts

Bo: TODO name this subsection better.

The following facts, easy to derive and used by Witkowski et al. [2018], will be useful to us as well.

Lemma 7. *The expected score, from the ground truth point of view, of forecaster i on event k is $1 - \theta_k(1 - \theta_k) - (y_{i,k} - \theta_k)^2$.*

Proof. By definition, the expected score is

$$\begin{aligned}
&\theta_k \left[1 - (1 - y_{i,k})^2\right] + (1 - \theta_k) \left[1 - y_{i,k}^2\right] \\
&= \theta_k \left[2y_{i,k} - y_{i,k}^2\right] + (1 - \theta_k) \left[1 - y_{i,k}^2\right] \\
&= 1 - \theta_k + 2\theta_k y_{i,k} - y_{i,k}^2 \\
&= 1 - \theta_k + \theta_k^2 - (y_{i,k} - \theta_k)^2.
\end{aligned}$$

□

Averaging over events, we get that i's score on average over the m events, in expectation, is exactly her accuracy minus some penalty term that only depends on how extreme the ground truth events are.

Corollary 8. *The expectation of forecaster i's average score on all events, from the ground truth point of view, is $a_i - \frac{1}{m} \sum_{k=1}^{m} \theta_k(1 - \theta_k)$, where a_i is i's accuracy.*

We also immediately obtain:

Corollary 9. *The expected difference in average scores of forecaster i and j, from a ground truth point of view, is $a_i - a_j$.*

3 Simple average mechanism

A straightforward selection mechanism often used in practice is what we call the *simple average mechanism* M_s. For this mechanism, we assign each forecaster a score $f_{i,k} = R_q(y_{i,k}, x_k)$ for each event. Then, we assign their final score by summing these over all events, $F_i = \sum_{j=1}^{m} f_{i,j}$. Finally, we choose the forecaster with the highest cumulative score $M_s(y, x) = \arg\max_i F_i$.

3.1 Truthfulness

As observed by Witkowski et al. [2018] and discussed in depth by Lichtendahl and Winkler [2007], this mechanism is generally not truthful. We can consider a simple example with three forecasters and one event, whose true probability is 0.5. Alice predicts 0.5, but Bob predicts 1 and Charlie predicts 0. Observe that Alice *cannot win*, despite being the best forecaster by far: either the event occurs (Bob wins) or it doesn't (Charlie wins). She can only win be predicting either 0 or 1, raising her probability of winning from 0 to 0.25, assuming a random tiebreaking rule.

3.2 Efficiency

Although the simple average mechanism is not truthful, it is worth studying its efficiency because, intuitively, it seems unlikely to do better using a mechanism that relies on the quadratic scores of the participants. So as a benchmark for truthful quadratic-score-based mechanisms, let us consider the efficiency of M_s if forecasters were to report their true beliefs $\hat{\theta}_i$.

For a single event, the most extreme forecasters will always get selected, as in the example above. However, for multiple events, the averaging starts to favor more accurate forecasters. In fact, one can show that with n forecasters and an accuracy gap of ϵ, only

$$m = \frac{2}{\epsilon^2} \ln\left(\frac{n}{\delta}\right)$$

events suffice to select the best forecaster with probability $1 - \delta$.

Sketch (similar to the Witkowski et al. [2018] analysis of ELF): Let i be the best forecaster. She beats some competitor j if she has a higher average score in the

competition. The expected difference in average scores is the difference in accuracies (Corollary 9), which is $a_i - a_j \geq \epsilon$. On any given event, the difference in scores is an independent random variable bounded in $[-1, 1]$. A standard Hoeffding bound **?** shows that the average of $m = \frac{2}{\epsilon^2} \ln \frac{n}{\delta}$ such variables is at least ϵ below its mean with probability $1 - \frac{\delta}{n}$. Complete the argument by taking a union bound over all $n - 1 \leq n$ opponents j.

This shows that it is possible to select the best forecaster using a number of events only logarithmic in n, the number of forecasters – if truthfulness is not required. We next consider the number of events used by the ELF mechanism.

4 ELF

The ELF mechanism M_{ELF} [Witkowski et al., 2018], improves upon M_s, by incentivizing forecasters to be truthful. For each event k, we hold a lottery to choose a winner. Forecaster i is chosen with probability

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

This score is the additively normalized quadratic score among all forecasters. A forecaster increasing their own quadratic score increases the chances they win an event, while decreasing the chances others win that event. Let w_k be the winner of event k. Then we choose the forecaster $M_{\text{ELF}}(y, x) = \arg\max_i \sum_{k=1}^{m} \mathbb{1}(w_k = i)$ as the winner of the competition. Essentially, we give the winner of each event lottery a point, and then pick the forecaster who got the most points as the overall winner.

In Theorem 6 of Witkowski et al. [2018], it is proven that M_{ELF} is strictly truthful. In Theorem 8 of the same paper, they also upper bound m_δ by showing

$$m_\delta(M_{\text{ELF}}, n, \epsilon) \leq \frac{2(n-1)^2}{\epsilon^2} \ln\left(\frac{4(n-1)}{\delta}\right)$$

For a fixed forecaster accuracy gap and desired mechanism accuracy, the number of events ELF needs to consistently choose the right forecaster is proportional to $\frac{1}{\epsilon^2} n^2 \ln n$. However, this upper bound is not tight.

Let $F_{i,k}$ be the indicator function for forecaster i winning the lottery for event k. Then, $F_i = \sum_k F_{i,k}$ is the random variable that is the number of events forecaster i

wins. The proof of Theorem 8 Witkowski et al. [2018] utilized the Hoeffding bound applied to F_i, which assumes they have high variance. However, this is not the case, and utilizing that fact gives an improvement in the lower bound.

Theorem 10 (Bernstein's Inequality)**.** *Given random variables X_i for $1 \leq i \leq n$ such that $0 \leq X_i \leq 1$ almost surely, let $S = \sum_i X_i$. Then,*

$$P\left(|S - \mathbb{E}[S]| < t\right) < 2\, e^{\frac{-t^2}{2\left(\mathrm{Var}(S) + \frac{t}{3}\right)}}$$

Theorem 11. $m_\delta(M_{\mathrm{ELF}}, n, \epsilon) \leq \frac{5(n-1)}{\epsilon^2} \ln\left(\frac{4(n-1)}{\delta}\right)$

Proof: Note that for any single event k, every forecaster wins with probability at best $\frac{2}{n}$, and at worst $\frac{1}{n} - \frac{1}{n(n-1)}$ (achieved when $y_i = 1$, $y_{j\neq i} = 0$, and $x_i = 1$). Since $F_{i,k}$ is just a binary random variable, it will have highest variance when it's expectation is as close to $\frac{1}{2}$ as possible. For nontrivially small n, this will mean its expectation is $\frac{2}{n}$ (or equivalently $P(F_{i,k} = 1) = \frac{2}{n}$), so its variance will be at most $\frac{2(n-2)}{n^2}$. So we have $\mathrm{Var}(F_{i,k}) \leq \frac{2(n-2)}{n^2} < \frac{2}{n}$.

Since $F_i = \sum_k F_{i,k}$, we have $\mathrm{Var}(F_i) = \sum_k \mathrm{Var}(F_{i,k}) < \sum_k \frac{2}{n} = \frac{2m}{n}$. As shown in Witkowski et al. [2018], for $j \neq i$ we have

$$\mathbb{E}[F_i] - \mathbb{E}[F_j] \geq \frac{m\epsilon}{n-1}$$

Therefore, if $F_j \geq F_i$, (forecaster j beats forecaster i), then either forecaster j overperformed their expectation by at least $\frac{m\epsilon}{2(n-1)}$, or forecaster i underperformed their expectation by $\frac{m\epsilon}{n-1}$. Specifically, either $\mathbb{E}[F_i] - F_i \geq \frac{m\epsilon}{2(n-1)}$ or $F_j - \mathbb{E}[F_j] \geq \frac{m\epsilon}{2(n-1)}$

Applying Theorem 10, we have

$$\begin{aligned}
P\left(|F_i - \mathbb{E}[F_i]| < \frac{m\epsilon}{2(n-1)}\right) &< 2\, e^{\frac{-\left(\frac{m\epsilon}{2(n-1)}\right)^2}{2\left(\frac{2m}{n} + \frac{m\epsilon}{6(n-1)}\right)}} \\
&< 2\, e^{\frac{-\left(\frac{m\epsilon}{2(n-1)}\right)^2}{2\left(\frac{m}{n-1}(2+\frac{\epsilon}{6})\right)}} \\
&= 2\, e^{\frac{-\frac{m\epsilon^2}{2(n-1)}}{(2+\frac{\epsilon}{6})}} \\
&< 2\, e^{-\frac{m\epsilon^2}{5(n-1)}} \qquad (1)
\end{aligned}$$

And similarly for F_j. Putting both of those cases together, we have by the union bound

$$\begin{aligned} P(F_j \geq F_i) &= P\left(\left(F_j - \mathbb{E}[F_j] < \frac{m\epsilon^2}{2(n-1)}\right) \cup \left(\mathbb{E}[F_i] - F_i < \frac{m\epsilon^2}{2(n-1)}\right)\right) \\ &\leq P\left(F_j - \mathbb{E}[F_j] < \frac{m\epsilon^2}{2(n-1)}\right) + P\left(\mathbb{E}[F_i] - F_i < \frac{m\epsilon^2}{2(n-1)}\right) \\ &\leq 4\, e^{-\frac{m\epsilon^2}{5(n-1)}} \end{aligned}$$

Using the union bound over all j, we have

$$\begin{aligned} P\left(M_l(y_1, ..., y_n, x) = i\right) &= 1 - \sum_{j \neq i} P\left(M_l(y_1, ..., y_n, x) = j\right) \\ &\geq 1 - \sum_{j \neq i} P\left(F_j \geq F_i\right) \\ &\geq 1 - 4(n-1)e^{-\frac{m\epsilon^2}{5(n-1)}} \end{aligned}$$

Assigning this probability to $1 - \delta$, we see that for a fixed n, ϵ, ELF will choose the correct forecaster with probability at least $1 - \delta$ if m is large enough for the above inequality to hold. Solving it for m, we get the minimum such satisfying value

$$m_\delta \leq \frac{5(n-1)}{\epsilon^2} \ln\left(\frac{4(n-1)}{\delta}\right)$$

□

Theorem 12. *For $\delta < \frac{1}{2}$, $m_\delta(M_{\text{ELF}}, n, \epsilon) > n \log n$*

Proof: Pick some $\delta < \frac{1}{2}$ and consider the case where $m = n \log n$ and $\theta_k = 1$ for every event. Let $y_{1,k} = 1$ and $y_{i,k} = 0$ for every $i > 1$. Therefore, forecaster 1 always forecasts correctly and has a quadratic score $R_q(1, 1) = 1$ for each event, while everyone else is wrong and has a quadratic score of $R_q(0, 1) = 0$ for each event. Then, $a_1 = 1$, and $a_i = 0$ for $i > 1$, so we have the worst possible accuracy gap $\epsilon = 1$.

For every event, forecaster 1's score will be

$$\begin{aligned} f_1 &= \frac{1}{n}\left(1 + R_q(y_1, 1) - \frac{1}{n-1}\sum_{j\neq i} R_q(y_j, x)\right) \\ &= \frac{1}{n}\left(1 + 1 - \frac{1}{n-1}\sum_{j\neq i} 0\right) \\ &= \frac{2}{n} \end{aligned}$$

Since all the other forecasters are symmetric, their f_i must all be the same, and since they sum to 1, we must have $f_i = \frac{n-2}{n(n-1)} < \frac{1}{n}$.

We can reframe the way winners are chosen for each event. Instead of holding a lottery, we first flip a coin that gives forecaster 1 a win with probability $\frac{2}{n}$. If it does not give them the win, we run a normal lottery for the remaining forecasters, where they each have probability $\frac{1}{n-1}$ of winning. This gives them an overall probability of $\frac{n-2}{n(n-1)}$ to win each event, so it is the same the original lottery.

Since all the events are uniform, the expected number of lotteries forecaster 1 will win is $mf_1 = \frac{2m}{n} = 2\log n$. Specifically, their wins will follow the Bernoulli distribution $B(n, \frac{2}{n})$. With some probability $C_1 = P(B(n, \frac{2}{n}) < 2\log n) \approx \frac{1}{2}$, forecaster 1 will win less events than expected.

The distribution of wins for the remaining forecasters will follow a uniform multinomial distribution. Specifically, we can model it with Balls and Bins. Let M be the maximum number of wins of any of the remaining forecasters.

Lemma 13. $P(M > 2\log n) = 1 - o(1)$.

Proof: From Theorem 1 of Raab and Steger [1998], we know that for $m = n\log n$, $P(M > k_\alpha) = 1 - o(1) > C_2$ for some constant C_2, where $c = \frac{m}{n\log n}$, $k_\alpha = (d_c - 1 + \alpha)\log n$, and $0 < 1 < \alpha$, such that d_c is the larger solution of $1 + x(\log c - \log x + 1) - c = 0$.

In our case, $c = 1$, so d_c is the solution to $x(1 - \log x) = 0$. Therefore, $d_c = e$. Since this inequality holds for any α, we let $\alpha = 3 - e$. Plugging in both values gives us $k_\alpha = 2\log n$, so $P(M > 2\log n) = 1 - o(1)$. □

Using the lemma, for any $C_2 < 1$ and sufficiently large n, $P(M > 2\log n) > C_2$. We choose some $C_2 > 2\delta$. Now, the probability that forecaster 1 is not chosen is at least

$$P(B(n, \frac{2}{n}) < 2\log n) \times P(M > 2\log n) > \frac{1}{2} \times 2\delta = \delta$$

Therefore, it is not true that for any n, ϵ, $P(M_{\text{ELF}}(y, x) \neq i^*) \leq \delta$ when $m = n \log n$. Therefore, $m_\delta(M_{\text{ELF}}, n, \epsilon) > n \log n$. □

5 SHELF

For this mechanism, we let the binary random variable $X_{i,k}$ be 1 with probability $R_q(y_{i,k}, x_i)$, and 0 otherwise. Then, for each event k we assign forecaster i the score $f_{i,k} = X_{i,k}$ obtained by sampling this variable. Finally, we assign their final score by summing these over all events, $F_i = \sum_{j=1}^{m} f_{i,j}$, then chose the forecaster with the highest score $M_{\text{SHELF}}(y, x) = \arg\max_i F_i$, breaking ties uniformly at random.

5.1 Truthfulness

Note: The following analysis is inspired by and follows the work done in Witkowski et al. [2018].

(This is wrong because we can't just look at their probability of winning a point. We need to also factor in the tiebreaker weighting some wins differently)

For a single event with true probability θ, every forecaster will either win 1 or 0 events. The probability that their final score is 1 will just be their expected score for the single event. Therefore, to maximize their final score, each forecaster must maximize their expected score for the single event. Their score is just $\mathbb{E}_{x\sim\theta}(R_q(y_i, x)) = \mathbb{E}_{x\sim\theta}(1 - (y_i - x)^2)$. This is maximized when $y_i = \theta$, so the mechanism is truthful. We can also see this directly by noting that R_q is a strictly proper scoring rule, so its expectation is maximized when $y_i = \theta$.

Theorem 14. *M_{SHELF} is truthful for $m \geq 1$ events.*

Proof: For each forecaster $j \in [1, n]$ fix their report.

Suppose that for some forecaster i, i's report on some subset of events, where $k \in [1, m]$ denotes the size of this subset, does not report their true belief for any of the k events.
Let k' be one of those events.

We will now show that i will maximize their probability of being chosen as the best forecaster in the SHELF mechanism iff they alter their k reports to report their true belief on each of the k events. A guiding intuition here would be as follows: The

higher the expected score a forecaster has on a given event, the higher the probability that SHELF awards them a point on this specific event, and the only way to maximize one's probability of being selected as the best forecaster is to increase the number of points they are awarded. Therefore, for a single independent event, increasing one's expected score will either increase the probability of being selected as the best forecaster or leave it unchanged, and it does so without altering results of any other event. In this way, we will show that being truthful for a singular event will always increase one's probability of being selected as the best forecaster, because, using a quadratic scoring rule, a forecaster's expected score is maximized when reporting their true belief. We can then repeat this process for each untruthful report to maximize the probability of being selected as the best forecaster.

Formally, we will analyze three cases relating to the context provided by the $m-1$ events other than k'.

Let $V(i)$ denote the number of points forecaster i has accrued.
Let $f_{i,k}$ denote the probability forecaster i scores a point on event k.

1. Some forecaster j scores at least two more points than i, OR i scores at least two more points than any other forecaster.

$$\exists j \in [1,m] \text{ s.t. } j \neq i \text{ and } V(j) - V(i) \geq 2$$
$$OR$$
$$\forall j \in [1,m] \text{ s.t. } j \neq i; V(i) - V(j) \geq 2$$

   For this case, i's probability of being selected as the best forecaster is either exactly 1 or 0. It is also unaffected by being truthful or not in k', because gaining, or failing to gain, 1 point in this context is not enough to alter the outcome as only the forecaster with the most number of points is chosen as the best forecaster with ties broken uniformly.

2. Some forecaster j scores exactly one more point than i, AND no other forecasters score strictly more points than i.

$$\exists j \in [1,m] \text{ s.t. } j \neq i \text{ and } V(j) - V(i) = 1$$
$$AND$$
$$\forall l \in [1,m] \text{ s.t. } l \neq i \text{ and } l \neq j; V(i) - V(l) \geq 0$$

For this case, i must seek to maximize $f_{i,k'}$ to have any non-zero probability of being selected as the best forecaster under the SHELF mechanism. In the SHELF mechanism, the maximum payout a forecaster can receive for a single event is 1, and that point is awarded to them with probability exactly equal to their quadratic score. Thus, we know a forecaster's expected value for a singular event is exactly their quadratic score. Since the quadratic scoring rule is proper, we know that the expectation of a forecaster's quadratic score is maximized when the forecaster is reporting their true beliefs. As a result, we can conclude that the only way for i to maximize their probability of receiving a point in k' is to report their true beliefs. By reporting their true beliefs, which maximizes the probability that they score another point, i also maximizes the probability of being selected as the best forecaster which, in this context, occurs when only i scores a point and the tie between j and i is broken uniformly via the rules of SHELF.

3. No forecasters score more points than i, but there is at least one forecaster j that scores the same number of points as i or scores exactly one less point than y_i.

$$\exists j \in [1, m] \text{ s.t. } j \neq i \text{ and } V(i) - V(j) \in \{0, 1\}$$

In any given configuration, a forecaster's probability of being selected as the best forecaster is maximized when the minimal number of other forecaster have as many points as them. In the SHELF mechanism, any forecaster's probability of scoring a point on a given event is independent of all others. Therefore, there is no action that can be taken to minimize the probability of any competitor's probability of being selected as the best forecaster except strictly increasing one's own point total. In this final case, the only way for i to increase their probability of being chosen as the best forecaster would be to increase the probability of scoring a point in k'. The same logic as case 2 holds here that i would then prefer to be truthful and maximize their expected score in k' to increase their probability of being selected as the best forecaster.

Since any configuration of the $m - 1$ events falls into one of these three cases, we have shown that any forecaster is weakly incentivized to report truthfully. To show that any forecaster is strongly incentivized to report truthfully, we simply show that i believes cases 2 or 3 will happen with positive probability. Since $y_{j,k} = p_{j,k} \in (0, 1) \forall j \forall k$, we know that every forecaster has some uncertainty on each event. As a result, we can conclude that $f_{j,k} > 0 \forall j \forall k$. We can think of all possible outcomes as the set of all possible binary strings of length n, and since $f_{j,k} > 0 \forall j \forall k$ there exists at least one sequence of outcomes strings picked from our set that constitutes cases

2 or 3. As a result, there is a positive probability that cases 2 or 3 occur, and thus forecasters are strongly incentivized to report truthfully.

5.2 Accuracy

Theorem 15. $m_\delta(M_{\text{SHELF}}, n, \epsilon) \leq \frac{2}{\epsilon^2} \ln \frac{4(n-1)}{\delta}$

Proof: This proof is a slight modification of the proof of Theorem 8 in Witkowski et al. [2018].

We first bound the difference in the expected number of wins of i^* and j, for any $j \neq i^*$.

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^{m} f_{i^*,k}\right) - \mathbb{E}\left(\sum_{k=1}^{m} f_{j,k}\right) &= \left(\sum_{k=1}^{m} \mathbb{E} f_{i^*,k}\right) - \left(\sum_{k=1}^{m} \mathbb{E} f_{j,k}\right) \\
&= \sum_{k=1}^{m} \left(\mathbb{E}(f_{i^*,k}) - \mathbb{E}(f_{j,k})\right) \\
&= \sum_{k=1}^{m} \left((a_{i^*} + \theta_k^2 - \theta_k) - (a_j + \theta_k^2 - \theta_k)\right) \\
&= \sum_{k=1}^{m} (a_{i^*} - a_j) \\
&\geq \sum_{k=1}^{m} \epsilon \\
&= m\epsilon
\end{aligned}$$

If $F_j \geq F_{i^*}$, then either $\mathbb{E}(F_{i^*}) - F_{i^*} \geq \frac{m\epsilon}{2}$ or $F_j - \mathbb{E}(F_j) \geq \frac{m\epsilon}{2}$. By Hoeffding's, we can upper bound the probability of each of these events by $2e^{-\frac{2(\frac{m\epsilon}{2})^2}{m}} = 2e^{\frac{-m\epsilon^2}{2}}$. Using the union bound, the probability that either of those events occurs is twice this, or $4e^{\frac{-m\epsilon^2}{2}}$.

Using the union bound again over all $j \neq i$, we have

$$\begin{aligned} P(M_{\text{SHELF}}(y,x) = i) &= P(F_{i^*} > F_j \quad \forall j \neq i^*) \\ &\geq 1 - \sum_{j \neq i^*} P(F_{i^*} \leq F_j) \\ &= 1 - 4(n-1)e^{\frac{-m\epsilon^2}{2}} \end{aligned}$$

Letting this quantity be $1-\delta$ and solving for m, we see that this mechanism correctly chooses forecaster i with probability $1-\delta$ when m is large enough to satisfy the inequality. Therefore

$$m_\delta(M_{\text{SHELF}}, n, \epsilon) \leq \frac{2}{\epsilon^2} \ln \frac{4(n-1)}{\delta}$$

□

Comparing this to the result of Theorem 11, we see that M_{SHELF} is much more efficient than M_{ELF}. We can also directly apply the proof of Theorem 15 to M_s, since the mean and variance of F_i is the same in both cases. Therefore, $m_\delta(M_{\text{SHELF}}, n, \epsilon) = m_\delta(M_s, n, \epsilon)$.

References

Kenneth C. Lichtendahl, Jr. and Robert L. Winkler. Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745—1755, 2007.

Martin Raab and Angelika Steger. "balls into bins" - a simple and tight analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, RANDOM 1998, pages 159–170. Springer-Verlag, 1998. ISBN 3-540-65142-X.

Jens Witkowski, Rupert Freeman, Jennifer Wortman Vaughan, David M. Pennock, and Andreas Krause. Incentive-compatible forecasting competitions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI 2018, 2018.