Information Theory for Prediction Markets

Notes and Extension by Ian Joffe
Based on *False Consensus, Information Theory, and Prediction Markets*
by Yuqing Kong and Grant Schoenebeck, 2022

Submitted March 2024 for CMSC 33581

1 Round Robin Reaches Consensus

In an agreement game, agents aim to accurately predict the probability of event W based on incomplete information. Each agent starts with a prior credence for W and additional private information X. K&S assume that agents are rational Bayesian aggregators of information. They credit the results of this section originally to Aaronson (2005) and Geanakoplos & Polemarchakis (1982), but K&S uniquely apply the information theoretic framework reviewed here.

The round robin protocol is a format where n agents, indexed by i, take turns sharing announcements related to W. Agent's i's announcement in round t is called h_i^t, and we call the history of all announcements up to but excluding that one H_i^t. Then before the next announcement h_i^{t+1}, each agent uses the previous announcement to update their credence in W to p_i^{t+1}

K&S consider the setting where all agents have a common prior for W. Let p_i^t denote the credence of agent i as the agent announces their h_i^t. For any random variables X and Y, let $H(X)$ denote the entropy of X, $H(X|Y)$ denote the entropy of X conditional on Y, and $I(X;Y) = \sum_{x,y} P(x,y)\log(\frac{P(x,y)}{P(x)P(y)})$ denote the mutual information of X and Y. One can verify that $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ and $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. The multivariable extension of mutual information is called co-information, and has analogous properties. The authors also adopt a convenient convention where ";" represents intersection of events, "," represents union, and "+" represents strictly disjoint union.

Round t achieves ε-MI ("ε-Mutual Information") if, for all i,

$$I(X_i; W|H^t) \leq \varepsilon$$

meaning no one's private information is any more informative about W than the public information (approximately). In other words, the marginal value of each agent's information is no greater than ε. We say that a protocol reaches consensus if, for some finite t, it guaranteed to achieve ε-MI by the of round t.

The authors make a couple observations about the public information on W, namely $I(H_i^t|W)$, at the end of each round.

- $I(H_i^t; W)$ in monotonically increasing with t
- The growth of the mutual information between H and W with each announcement is the mutual information between the agent's announcement and the and W conditional on the previous history - the announcement's "marginal mutual information"

$$I(H_{i+1}^t; W) - I(H_i^t; W) = I(h_i^t; W|H_i^t) \tag{1}$$

We can use this to show that certain agreement games do reach consensus. Take the round robin protocol. First, consider the game where each announcement is the agent's true credence, so $p_i^t = h_i^t$. Denoting the history of announcements *after* round t as H^t, Then the information gained in *each round* before consensus can be decomposed into the marginal information of the announcement of agent i, and the marginal information of the round's other announcements:

$$\begin{aligned}
I(H^{t+1}; W) - I(H^t; W) &\geq I(H_{i+1}^{t+1}; W) - I(H^t; W) && (H_{i+1}^{t+1} \text{ is before } H^{t+1}) \\
&= I(h_i^{t+1}, H_i^{t+1}; W) - I(H^t; W) \\
&= I(h_i^{t+1}, H_i^{t+1}; W|H^t) \\
&= I(H_i^{t+1}; W|H^t) + I(h_i^{t+1}; W|H_i^{t+1}) && (2)
\end{aligned}$$

Specifically, the authors aim to prove that **consensus is reached in less than $\frac{2}{\varepsilon}$ rounds**. To do this, we can show that in each round before consensus, at least $\frac{\varepsilon}{2}$ information is made public. Without loss of generality, say that before round $t+1$, agent i's marginal information is still at least ε. Then, either agent i's announcement will contribute $\frac{\varepsilon}{2}$ information, or the announcements in round $t+1$ before agent i's announcement will contribute $\frac{\varepsilon}{2}$ information. Mathematically, this is an application of mutual information's chain rule

$$\begin{aligned} I(X_i, H_i^{t+1}; W|H^t) &= I(H_i^{t+1}; W|H^t) + I(X_i; W|H_i^{t+1}) \\ &= \varepsilon \qquad \text{(so one of the terms above must be at least } \frac{\varepsilon}{2}) \end{aligned} \tag{3}$$

Clearly, if the first term is $\geq \frac{\varepsilon}{2}$, then at least $\frac{\varepsilon}{2}$ information is contributed in the round. To see that the second term being at least at least $\frac{\varepsilon}{2}$ leads to the same conclusion, we must understand that $I(X_i; W|H_i^{t+1}) \geq \frac{\varepsilon}{2} \implies I(h_i^{t+1}; W|H_i^{t+1}) \geq \frac{\varepsilon}{2}$, i.e. that marginally informative private information implies a marginally informative announcement. Since in round robin, agents simply announce their posteriors p_i^{t+1}, this is true if an agent's posterior has the same marginal information as their private X_i, which is in turn true due to the information processing inequality - the posterior is a function of X_i, and given the history we can also solve for X_i and a function of the posterior.

2 Extension to Random Participation

Kong and Schoenebeck's work concerns games where agents take turns making announcement, but what if instead agent's made announcements at random times? To avoid confusion, I'll index time with τ. We are interested in information content of all announcements made before time $\mathcal{T}$. Using the marginal information of an announcement,

$$\begin{aligned} I(H^{\mathcal{T}}; W) &= \sum_{\text{all } h_i^\tau : \tau < \mathcal{T}} I(h_i^\tau; W|H^\tau) \\ &= \sum_i \sum_{h_i^\tau : \tau < \mathcal{T}} I(h_i^\tau; W|H^\tau) \end{aligned} \tag{4}$$

To calculate the expected information value of the history at $\mathcal{T}$,

$$\begin{aligned} \mathbb{E}(I(H^{\mathcal{T}}; W)) &= \mathbb{E}\Big(\int_0^{\mathcal{T}} \sum_i \mathbb{P}(i \text{ makes announcement at } \tau) I(h_i^\tau; W|H^\tau) d\tau\Big) \\ &= \int_0^{\mathcal{T}} \sum_i \mathbb{P}(i \text{ makes announcement at } \tau) \mathbb{E}\Big(I(h_i^\tau; W|H^\tau)\Big) d\tau \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{(I hope idk Fubini's Thm)} \\ &= \int_0^{\mathcal{T}} \sum_i \mathbb{P}(i \text{ makes announcement at } \tau)\big(\sum_H I(h_i^\tau; W|H)\mathbb{P}(H^\tau = H)\big) d\tau \end{aligned} \tag{5}$$

Let's assume that agents are equally likely to make announcements at any time, and we expect agent i to make λ_i announcements over each unit of time. Note this is equivalent to saying that the waiting times between i's announcements are memoryless and $Exponential(\lambda_i)$ distributed, and that i's total number of announcements is $Poisson(\lambda_i)$ distributed. Now,

$$\mathbb{E}(I(H^{\mathcal{T}}; W)) = \int_0^{\mathcal{T}} \sum_i \frac{\lambda_i}{\mathcal{T}} \big(\sum_H I(h_i^\tau; W|H)\mathbb{P}(H^\tau = H)\big) d\tau \tag{6}$$

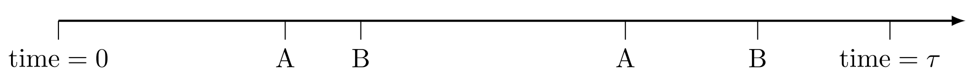
Looking at the inner summation, we are only interested in the history since agent i's last announcement, because for Bayesian agents i's last announcement contains all the information from announcements before it. Still, an infinite number of histories are possible, and even with truncated histories for a medium number of agents, the combinatorics quickly become intractable.

So to simplify, we'll assume very few agents. The case of two agents A and B is especially basic, since there are only two effective histories to account to account for. Say A is making an announcement at time τ. If A made the previous announcement, then the new one will not increase public information. If B made the previous announcement, then A's announcement will increase public information, but by the same amount as if B had made the last several announcements. Therefore, **for two agents**,

$$
\begin{aligned}
\mathbb{E}(I(H^{\mathcal{T}};W)) = \int_0^{\mathcal{T}} & \Bigg(\frac{\lambda_A}{\mathcal{T}}(0+\mathbb{E}\Big(I(h_A^{\tau};W|H^{\tau},\text{B made last announcement})\Big)P(\text{B made the last announcement})\Bigg) \\
&+\Bigg(\frac{\lambda_A}{\mathcal{T}}(0+\mathbb{E}\Big(I(h_B^{\tau};W|H^{\tau},\text{A made last announcement})\Big)P(\text{A made the last announcement})\Bigg)\,d\tau \\
=& \int_0^{\mathcal{T}} \frac{\lambda_A}{\mathcal{T}}\mathbb{E}\Big(I(h_A^{\tau};W|H^{\tau},\text{B made last announcement})\Big)\Big(\frac{\lambda_B}{\lambda_A+\lambda_B}\Big) \\
&+\frac{\lambda_B}{\mathcal{T}}\mathbb{E}\Big(I(h_B^{\tau};W|H^{\tau},\text{B made last announcement})\Big)\Big(\frac{\lambda_A}{\lambda_A+\lambda_B}\Big)\,d\tau \\
=& \frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}\cdot\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} \mathbb{E}\Big(I(h_B^{\tau};W|H^{\tau},\text{A made last announcement})\Big) \\
&\qquad +\mathbb{E}\Big(I(h_A^{\tau};W|H^{\tau},\text{B made last announcement})\Big)\,d\tau
\end{aligned}
\tag{7}
$$

One thing we already observe here is the dependence on $\frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}$. This suggests that one can expect higher information at a given τ if λ_A and λ_B are large (of course), but more interestingly if λ_A and λ_B are close together. This makes intuitive sense, since if either agent rarely makes announcements, the other agent making frequent announcements cannot make up for it. It is possible for this effect to be mitigated based on the agents strategies in announcing information (the term inside the integral), for example if the agent who makes less frequent announcements gives away a lot of information when they do share.

In equation (7), the randomness is in H^{τ}. Precisely, we are concerned with the length of H^{τ}, excluding double-announcements – the length of the longest "alternating subsequence" of announcements. The distribution of this quantity is very tricky, and I'll make an approximation. Say events occur in some order, for example



where in the figure, I've removed all non-alternating events. Now considering only events in the figure, the expected waiting time until the next event is P(Waiting on A)(Expected Waiting Time for A) + P(Waiting on B)(Expected Waiting Time for B). The probability that we're waiting for A is all the time we spend on the number line waiting for A over all the total time on the number line, and analogously for B. This total is expected to be $\frac{1/\lambda_A}{1/\lambda_A+1/\lambda_B}\cdot 1/\lambda_A+\frac{1/\lambda_B}{1/\lambda_A+1/\lambda_B}\cdot 1/\lambda_B=\frac{1/\lambda_A^2+1/\lambda_B^2}{1/\lambda_A+1/\lambda_B}$. If we let this distribution be memoryless, the calculated expected waiting time should be the inverse of the rate parameter of a Poisson distribution, namely $Pois(\frac{\lambda_B\lambda_A^2+\lambda_A\lambda_B^2}{\lambda_A^2+\lambda_B^2})$. This is an approximation for two reasons. First, the model doesn't quite match the timeline. The model assumes necessarily that an even number of events occurs, and it neglects time on the timeline after the last event. Second, the process in which events occur is not exactly memoryless, since there is a different probability of an event occurring depending on the previous event. This approximation is least impactful when λ_1 and λ_2 are close together.

Regardless, the approximation can sometimes be used to simplify our integral. Let $\lambda^*=\frac{\lambda_B\lambda_A^2+\lambda_A\lambda_B^2}{\lambda_A^2+\lambda_B^2}$ and let ℓ represent the length of a max length alternating subsequence. Looking at each term inside (7)'s integral,

$$
\mathbb{E}\Big(I(h_A^{\tau};W|H^{\tau},\text{B made last announcement})\Big)=\sum_{\ell=0}^{\infty} I(h_A^{\tau};W|\ell,\text{B made last announcement})\ e^{-\tau\lambda^*}\frac{(\tau\lambda^*)^{\ell}}{\ell!}
\tag{8}
$$

The first lesson from this is that the information disclosed in an announcement must depend on ℓ. If the information that A discloses in their announcement does not depend on ℓ, then the exponential's Maclaurin series sends the summation of the probability terms to 1 (which would occur for any probability distribution, since all densities integrate to 1) and leaves an infinite sum of constant terms, which blows up. This can't blow up, because W has a finite entropy, which bounds its mutual information with any other variable.

So the next interesting thing to do is determine what possible expressions for the information of each announcement prevent the summation from blowing up. For example, if the amount of information revealed increases exponentially with $\tau\lambda^*$, then the Maclaurin series shows that the infinite sum itself will come out to an exponential. However, this is unrealistic because the information content of each announcement is also bounded by W's entropy.

A better idea would be to assign a function of ℓ for the info revealed with each announcement that decreases quickly enough. This would depend on the information dynamics of each agent's strategy. But considering that (assuming honest agents) once information is revealed, it can't be taken back, and agents have finite amounts of information, it makes sense that such quickly enough decreasing functions exist. But, their interaction with the probability term makes the sum very complicated, and I'll cut my analysis off here for this brief project.

3 Application to Prediction Markets

In a prediction market, traders attempt to profit by predicting the binary outcomes of future events. A traditional implementation is for individuals to trade shares on the event, which pay out if the event is observed to occur, and go to junk if the event does not occur. Such a market turns out to be theoretically equivalent to a system where any participant may simply set the price, and receives a payout after the event does or does not occur according to a specific scoring rule. Similar to section (1), the authors attribute the foundational findings to Hanson (2002 & 2003), but include their own information theoretic twist, and will show how the payout under Hanson's setup relates to information.

Consider a market for event W, which resolves as 0 or 1. The prior probability, or initial price, for $W = 1$ is preset to p_0, and in this setup we consider the order of agents bids to be fixed and round-robin style. During an agent's turn, they may choose to change the price from p_i to p_{i+1}. Agents continue as long as they'd like, and after betting is complete, W is observed. Agents then receive a payment L according to Hanson's logarithmic scoring rule, where the agent who changed the price from p to p_{i+1} is paid/charged

$$L(W, p_i, p_{i+1}) = \begin{cases} \log p_{i+1} - \log p_i \text{ if } W = 1 \\ \log p_i - \log p_{i+1} \text{ if } W = 0 \end{cases} \tag{9}$$

The choice of scoring rule is interesting in itself. The logarithmic scoring rule is a proper scoring rule, meaning an agent's payout increases as they get closer to correctly setting the price to 0 or 1, and has other unique and intricate advantages. One inefficiency (whether this is an advantage or disadvantage is a matter of perspective) is that under the logarithmic scoring rule, traders are not necessarily incentivized to share all their information immediately. Such dynamics, however, are a bit out of scope for this review.

K&S make the simple but fascinating observation that under the log scoring rule, a player who changes the belief of the market to their own true credence in W (akin to the round robin consensus protocol) receives an expected payout of $I(X_i; W|H)$. We look at the expectation with respect to W and X_i for national purposes, but in our exmaple X_i can be considered fixed. Combining equation (8) with the pact that $p_i = P(W = 1) = 1 - P(W = 0)$,

$$\begin{aligned} \mathbb{E}_{X_i,W|H}\Big(L(W, p_i, p_{i-1})\Big) &= \sum_{X=X_i, W=\{0,1\}} P(W, X_i|H)\Big(\log(P(W|H, X_i)) - \log(P(W|H))\Big) \\ &= \sum_{X=X_i, W=\{0,1\}} P(W, X_i|H)\Big(\log \frac{P(W|H, X_i))}{(P(W|H)})\Big) \\ &= I(W, X_i|H) \end{aligned} \tag{10}$$

While this result is theoretically impressive, there is interest in making it more realistic to real players on prediction markets. One alternative examined in the paper is a consensus algorithm where, on an agent's turn, they simply reveal whether the current public information indicates a credence that's too high, or too low for them, given their private information (as opposed to giving away a specific number). This is more realistic because even if it may be optimal, real humans on prediction markets often don't buy or sell on events until the market price matches their personal credence.

Random participation is another practical extension of the results in the paper. I would argue this is an especially important extension given the way that humans (or even bots) play markets today, as I have never seen a real prediction market where agents take turns. In a many-agent game, how does each agent's λ impact their expected earnings, and how much additional earning does an agent achieve by increasing their rate of participation λ?

How quickly do the marginal returns of increasing λ diminish, and if we set a cost to agents increasing their λ (for example in compute or in a human's time), what its optimal value? The work in Section 2 is not yet complete enough to answer these questions, but I hope to continue to study the subject.