Senior Thesis in Mathematics

Prediction Markets: A Mechanism for Information Aggregation

Author:
Siddharth Namachivayam

Advisor:
Dr. Ami Radunskaya

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 23, 2022

Abstract

In general, markets compress collective information about the world (such as the scarcity of goods, shifting cultural trends, and new political events) into prices. Prediction markets attempt to leverage this general property to make forecasts about the future (e.g. the weather). The mechanism designer in a prediction market uses a scoring rule to elicit traders' beliefs about a particular event. If traders have common priors, pay attention to each other's reports, and employ a myopically optimal strategy of truthful betting (i.e. ignore profits from future trades) then the prediction market's prices will settle upon a distribution all traders agree with. This thesis suggests three new areas of research into prediction markets. First, we propose adding an additional option to renege on reports to discourage bluffing and reticence in prediction markets. Second, we apply rational inattention models to derive conditions under which traders will pay attention to each other's reports and incorporate outside information. Finally, we draw a series of thermodynamic analogies between the energy landscape of a physical system and prediction markets, yielding an elegant interpretation of the mechanism.

Contents

Chapter 1

Introduction and Background

1.1 Introduction

Prediction markets attempt to answer a seemingly straight forward question: given a group of forecasters with varying beliefs, how do we combine their opinions into a single coherent forecast? Upon first glance, we might think to form an aggregate forecast by simply taking the average of each individual's beliefs. However, this averaging approach has some serious flaws. To illustrate, suppose we are trying to create aggregate forecasts for the following events: (a) Trump will run for re-election and (b) It will rain tomorrow. Furthermore suppose we have three individuals who assign probabilities of 25%/75%, 50%/50%, and 75%/25% respectively to each event. If we take the average probability assigned to each event, then our aggregate forecast will predict there is a 50% chance that Trump runs for re-election and a 50% chance that it rains tomorrow. This ought to imply that the aggregate forecast for the joint event that it rains tomorrow and that Trump runs for re-election is $50\% \cdot 50\% = 25\%$. However, the probabilities assigned to the joint event by each individual are $25\% \cdot 75\% = 18.75\%$, $50\% \cdot 50\% = 25\%$, and $25\% \cdot 75\% = 18.75\%$, meaning our aggregate forecast is actually $20.8\overline{3}\%$.

Alternatively, another aggregation approach is to let our agents communicate with one another and agree upon an aggregate probability distribution collectively. A heuristic justification for this approach is given by Aumann's agreement theorem which asserts that if two agents have identical priors, each of them receive different information causing their posteriors to diverge,

and they proceed to iteratively report their true posteriors back and forth to one another, then they will eventually come to an agreement (or 'communication equilibrium') [GP82]. Prediction markets facilitate this process by incentivizing traders to buy and sell securities of the form "pays \$1 if event A occurs." The key idea behind prediction markets is that the prices set for such securities after an agent's transaction communicates that agent's posterior belief about event A. For instance, if the price of a "pays \$1 if it rains tomorrow" security is \$0.50 and a trader believes the chance it will rain tomorrow is 75%, then they will buy shares of the security until the price goes up to \$0.75. Thus once all profitable transactions are carried out, the market's prices will settle upon a communication equilibrium between the traders, yielding a single aggregate forecast.

Our journey towards formally defining and analyzing prediction markets as information aggregation mechanisms will begin with [Hay45] to provide historical background and motivation. We will then follow [Yu12]'s proof of Arrow's impossibility theorem and introduce a probabilistic Arrow's theorem with [Mcc81] and [LW83]. Next, we will examine a generalized version of Aumann's agreement theorem from [Aum76] and [GP82]. From there, [Han03] will detail the mechanics of prediction markets and illustrate the logarithmic market scoring rule. Rational inattention models will then be introduced using [Sim71] and [Sim10]. [Tsa20] and [Fol12] will then allow us to apply rational inattention to market scoring rules. Finally, we will explore how thermodynamics can be used to understand prediction markets, namely in their implementation as automated market makers as described in [CV10]. Since many of our results depend on various optimality conditions and bounds from information theory, we provide a primer on the subject in the appendix.

1.2 Background

In "The Use of Knowledge in Society," Hayek argues that the central issue facing contemporary economics does not simply consist in figuring out how to optimally allocate a 'given' set of resources. Such problems are relatively tractable from a mathematical point of view, and the marginalists more or less demonstrated the bliss point to be where the "marginal rates of substitution between any two commodities" [Hay45] are the same for all consumers.

While Hayek acknowledges the significance of this insight, he also believes answering the question of effective allocation is functionally useless without concrete knowledge of the scarcity which constrains economic welfare in the first place. The truly important function of economics then involves determining what resources actually are 'given' to society. The difficulty in making such a determination stems from the distributed and contradictory nature of relevant economic data amongst millions of agents. Thus, for Hayek, the question of how to best aggregate this decentralized patchwork of information lies at the heart of 'the dismal science.'

Hayek proceeds to identify three primary modes of information aggregation: central planning, competition, and monopoly. Central planning involves "direction of the whole economic system according to one unified plan," [Hay45] competition consists of "decentralized planning by many persons," [Hay45] and monopoly aims at the "delegation of planning to organized industries" [Hay45]. Importantly, for Hayek, all economic activity consists of planning. But who ought to be doing the planning? The answer, says Hayek, depends on the kind of knowledge necessary to carry out the planning in question. Far too often, people assume that "scientific knowledge...is the sum of all knowledge" [Hay45]. This erroneously suggests that the direction of all economic activity ought to be left to a small group of technical experts versed in operations research who make decisions via a central committee. To the contrary, scientific knowledge of 'general rules' cannot account for "knowledge of the particular circumstances of time and place" [Hay45]. This latter form of knowledge belongs to virtually every economic agent, and in particular, the agents whose occupation demands rapid adaptation to local changes in their environment. However, agents cannot rely exclusively on local signals to make optimal decisions and require "further information...of the larger economic system" [Hay45]. Crucially, since agents do not require knowledge of why a particular resource is more or less difficult to acquire, and only to what extent a particular resource is more or less difficult to acquire, it is sufficient to summarize all the data which comprise the totality of macroeconomic trends into prices which the agent can use to gauge in what manner they must alter their activities. In other words, Hayek paints a picture of the market as, first and foremost, a mechanism to compress information into prices.

But if one were to design a market tabula rasa to fit the aim of informa-

tion compression, exactly what form should such a price system take? The answer: prediction markets. Inspired directly by Hayek's argument in *The Use of Knowledge in Society*, prediction markets attempt to solve the age old problem of aggregating opinions by using the mathematics of information theory to incentivize participants to reveal their beliefs to one another. Namely, the opinions which prediction markets aggregate are opinions on the odds of a specific event occurring (e.g. a candidate winning an election or a startup delivering on its goals). We might reasonably wonder "why use prediction markets as opposed to other aggregation methods?" As was the case with the averaging approach we saw in the introduction, it turns out that most naïve aggregation methods (namely, ones which are non-dictatorial and only depend on static probabilities assigned to the events in question) will fail to maintain independence between events. This is because there is a probabilistic analogue of Arrow's impossibility theorem at work here! We now review the original Arrow's impossibility theorem and its proof before proceeding.

1.3 Arrow's Impossibility Theorem

We follow the proof of Arrow's impossibility theorem laid out by Yu 2012 in "A one-shot proof of Arrow's impossibility theorem" [Yu12].

1.3.1 3 Reasonable Demands

Let's start with some definitions. Given n voters $\{x_1, x_2..., x_n\}$ and m candidates $\{y_1, y_2..., y_m\}$, we define $\mathcal{O}$ to be the set of all $m!$ *order preferences* over our candidates.

A *preference profile* is an element $\theta \in \mathcal{O}^n$, where we let θ_i be the order preference of the ith voter. In particular, we write $\theta_i(a) \succ \theta_i(b)$ if voter x_i prefers candidate y_a to candidate y_b.

Finally, a *social aggregation function* is any function $f : \mathcal{O}^n \to \mathcal{O}$ which maps a given preference profile to a unique order preference (you can think of this as our 'voting system').

Now, we are ready to describe the 3 demands we want our social aggre-

gation function f to satisfy. I will define each demand formally and then use plain English to make each one seem reasonable.

First, f satisfies *unanimity* iff $\forall \theta \in \mathcal{O}^n$:

$$(\forall i \in [n],\ \theta_i(a) \succ \theta_i(b)) \implies f(\theta)(a) \succ f(\theta)(b)$$

This means if each of our voters prefers y_a to y_b, then the aggregate preference must also favor y_a to y_b.

Second, f satisfies *binary independence* iff $\forall \theta, \tau \in \mathcal{O}^n$:

$$(\forall i \in [n],\ \theta_i(a) \succ \theta_i(b) \leftrightarrow \tau_i(a) \succ \tau_j(b)) \implies (f(\theta)(a) \succ f(\theta)(b) \leftrightarrow f(\tau)(a) \succ f(\tau)(b))$$

In other words, the aggregate preference between y_a and y_b is uniquely determined by each voter's preference between y_a and y_b.

Third, f satisfies *non-dictatorship* iff:

$$\neg(\exists i \in [n] \text{ s.t. } \forall \theta \in \mathcal{O},\ f(\theta) = \theta_i)$$

That is to say, there does not exist a voter whose order preference is *always* the aggregate preference.

Hopefully, these demands seem like sensible things to ask from a voting system. If they don't, maybe try convincing yourself that a simple majority rule procedure between two candidates *does* satisfy them. Unfortunately, when there are more than 2 candidates, it turns out we're asking for too much.

1.3.2 Proof of the Theorem

Suppose toward a contradiction there exists a social aggregation function f satisfying unanimity, binary independence, and non-dictatorship in an election with more than 2 candidates.

Defining Pivots

Given two candidates y_a, y_b, we define the *pivot* from a to b as follows.

Consider a preference profile $\theta^0 \in \mathcal{O}^n$ where $\forall i \in [n]$:

$$\theta_i^0(a) \succ \theta_i^0(b)$$

By unanimity, we must have:

$$f(\theta^0)(a) \succ f(\theta^0)(b)$$

Now pick $\theta^k \in \mathcal{O}^n$ so that $\forall i \in [n] \backslash [k]$:

$$\theta_i^k(a) \succ \theta_i^k(b)$$

but $\forall i \in [k]$:

$$\theta_i^k(a) \prec \theta_i^k(b)$$

In particular, the last preference profile θ^n will have $\forall i \in [n]$:

$$\theta_i^n(a) \prec \theta_i^n(b)$$

Thus by unanimity, we must have:

$$f(\theta^n)(a) \prec f(\theta^n)(b)$$

Moreover, we are guarenteed the existence of an integer $b|a \in [n]$ satisfying:

$$b|a = \min\{k \mid f(\theta^k)(a) \prec f(\theta^k)(b)\}$$

We call the candidate $x_{b|a}$ our "pivot" from a to b. Note $b|a$ is *unique* despite our particular choice of $\theta^0, \theta^1 ..., \theta^n$ since f satisfies binary independence.

Additionally, if $\theta \in \mathcal{O}^n$ is a preference profile so that $\forall i \in [1, b|a)$:

$$\theta_i(a) \succ \theta_i(b)$$

and $\forall i \in (b|a, n]$:

$$\theta_i(a) \prec \theta_i(b)$$

then by construction:

$$f(\theta)(b) \succ f(\theta)(a) \iff \theta_{b|a}(b) \succ \theta_{b|a}(a)$$

The Power of Pivots

Let's introduce a third candidate y_c and suppose we have a preference profile $\tau \in \mathcal{O}^n$ so that $\forall i \in [1, b|a]$:

$$\tau_i(a) \succ \tau_i(b) \succ \tau_i(c)$$

and $\forall i \in (b|a, n]$:

$$\tau_i(a) \prec \tau_i(c) \prec \tau_i(b)$$

Since $b|a$ is the pivot from a to b, we know that:

$$f(\tau)(a) \succ f(\tau)(b)$$

Also by unanimity:

$$f(\tau)(b) \succ f(\tau)(c)$$

Thus:

$$f(\tau)(a) \succ f(\tau)(b) \succ f(\tau)(c)$$

Now, if voter $x_{b|a}$ switches their order preference so:

$$\tau_{b|a}(b) \succ \tau_{b|a}(a) \succ \tau_{b|a}(c)$$

and the rest of the voters arbitrarily shuffle their preferences between y_b and y_c, then binary independence implies we still have:

$$f(\tau)(a) \succ f(\tau)(c)$$

Moreover, since $b|a$ is the pivot from a to b:

$$f(\tau)(b) \succ f(\tau)(a)$$

Thus:

$$f(\tau)(b) \succ f(\tau)(a) \succ f(\tau)(c)$$

Finally, if all the voters arbitrarily change their ranking of y_a, then binary independence implies we still have:

$$f(\tau)(b) \succ f(\tau)(c)$$

Thus, so long as voter $x_{b|a}$ prefers candidate y_b to y_c, this dictates $f(\tau)(b) \succ f(\tau)(c)$, *regardless of everyone else's preferences.*

In other words the pivot from a to b gets to force all candidates (except for y_a) below y_b in the aggregate preference if they so chose. Interesting, n'est ce pas?

All Pivots are the Same

Consider a preference profile $\theta \in \mathcal{O}^n$ such that $\forall i \in [1, b|a]$:

$$\theta_i(b) \succ \theta_i(c)$$

and $\forall i \in (b|a, n]$:

$$\theta_i(c) \succ \theta_i(b)$$

As we saw at the end of the last section, since $\theta_{b|a}(b) \succ \theta_{b|a}(c)$ we must have:

$$f(\theta)(b) \succ f(\theta)(c)$$

Thus $b|c \leq b|a$.

Now consider a preference profile $\theta \in \mathcal{O}^n$ such that $\forall i \in [1, b|a)$:

$$\theta_i(c) \succ \theta_i(b)$$

and $\forall i \in [b|a, n]$:

$$\theta_i(b) \succ \theta_i(c)$$

As we saw at the end of the last section, since $\theta_{b|a}(b) \succ \theta_{b|a}(c)$ we must have:

$$f(\theta)(b) \succ f(\theta)(c)$$

Thus $c|b \geq b|a$. Putting this together with the previous inequality:

$$b|c \leq b|a \leq c|b$$

By symmetry we have:

$$c|b \leq c|a \leq b|c$$

Thus:

$$b|c = c|b = a|b = b|a = c|a = a|c$$

Since all pivots are the same, there exists a voter who gets to force each candidate below any other candidate in the aggregate preference. Thus this voter is a dictator, contradicting our assumption that f is non-dictatorial!

1.4 Probabilistic Arrow's Theorem

Now that we've seen Arrow's theorem proper, we are ready to examine its probabilistic analogue. Our proof will follow/combine the results in Mcconway 1981 [Mcc81] and Lehrer & Wagner 1983 [LW83].

1.4.1 3 Reasonable Demands

Instead of aggregating a group of individuals' preference orders over of set of candidates, we are now going to aggregate their *probability distributions* over a measurable space $(\Omega, \mathcal{F})$.

We require that Ω contains more than 2 measurable outcomes (this is analogous to having more than 2 candidates). Also, we will denote the set of all possible measures over $(\Omega, \mathcal{F})$ by $\mathcal{M}$.

In particular, the probability distributions of our n individuals will be characterized by a *opinion pool* $\{P_1, P_2, ...P_n\} \in \mathcal{M}^n$.

A *consensus function* will be a function $C : \mathcal{M}^n \to \mathcal{M}$ which maps a given opinion pool to a unique probability distribution over $(\Omega, \mathcal{F})$.

Once again, we have three seemingly reasonable demands:

First, C satisfies the *independence preservation property* (IPP) iff :

$$(\forall i \in [n],\ P_i(A \cap B) = P_i(A) \cdot P_i(B))$$
$$\implies C(P_1, P_2, ...P_n)(A \cap B) = C(P_1, P_2, ...P_n)(A) \cdot C(P_1, P_2, ...P_n)(B)$$

IPP parallels our previous notion of *unanimity* in that if each individual thinks events A and B are independent, then the aggregate distribution also treats A and B as independent.

Second, C satisfies the *strong setwise function property* (SSFP) iff $\exists G$: $[0,1]^n \to [0,1]$ such that for any given opinion pool $\{P_1, P_2, ..., P_n\} \in \mathcal{M}^n$:

$$\forall A \in \mathcal{F},\ C(P_1, P_2, ...P_n)(A) = G(P_1(A), P_2(A), ...P_n(A))$$

SSFP is analogous to *binary independence* since it implies the aggregate probability of a given event is uniquely determined by the *static* probabilities each individual assigns to that event.

Third, C satisfies *non-dictatorship* iff:

$$\neg(\exists i \in [n] \text{ s.t. } \forall \{P_1, P_2, ..., P_n\} \in \mathcal{M}^n,\ \forall A \in \mathcal{F},\ C(P_1, P_2, ...P_n)(A) = P_i(A))$$

1.4.2 Proof of the Theorem

As before, we will proceed by showing IPP + SSFP necessarily violate non-dictatorship. However, we need to first establish a crucial and somewhat surprising lemma.

SSFP $\iff$ Consensus is Linear

We claim C satisfies SSFP iff $\exists w_1, w_2, ...w_n \in [0,1]$ such that $\sum_{i=1}^{n} w_i = 1$ and for any given opinion pool $\{P_1, P_2, ...P_n\} \in \mathcal{M}^n$, $\forall A \in \mathcal{F}$:

$$C(P_1, P_2, ...P_n)(A) = G(P_1(A), P_2(A), ...P_n(A)) = \sum_{i=1}^{n} w_i \cdot P_i(A)$$

Proof:
It trivially follows that linear consensus $\implies$ SSFP so we will only show the other direction.

Consider three disjoint non-empty subsets $A, B, C \in \mathcal{F}$ which partition Ω (we are guarenteed these exist by the assumption Ω has more than 2 measurable outcomes).

Let $(a_1, b_1), (a_2, b_2), ...(a_n, b_n) \in \mathbb{R}^2_{\geq 0}$ so that $\forall i \in [n]$, $a_i + b_i \leq 1$. We construct an opinion pool $\{P'_1, P'_2, ...P'_n\} \in \mathcal{M}^n$ so that:

$$\begin{aligned} P'_i(A) &= a_i \\ P'_i(B) &= b_i \\ P'_i(C) &= 1 - (a_i + b_i) \end{aligned}$$

By SSFP:

$$\begin{aligned} G(a_1 + b_1, a_2 + b_2, ...a_n + b_n) &= G(P'_1(A) + P'_1(B), P'_2(A) + P'_2(B), ...P'_n(A) + P'_n(B)) \\ &= G(P'_1(A \cup B), P'_2(A \cup B), ...P'_n(A \cup B)) \\ &= C(P'_1, P'_2, ...P'_n)(A \cup B) \\ &= C(P'_1, P'_2, ...P'_n)(A) + C(P'_1, P'_2, ...P'_n)(B) \\ &= G(P'_1(A), P'_2(A), ...P'_n(A)) + G(P'_1(B), P'_2(B), ...P'_n(B)) \\ &= G(a_1, a_2, ...a_n) + G(b_1, b_2, ...b_n) \end{aligned}$$

In particular, this implies $\forall x = \{x_1, x_2, ...x_n\} \in [0,1]^n$:

$$G(x_1, x_2, ...x_n) = G(x_1, 0, ...0) + G(0, x_2, ...x_n) = \sum_{i=1}^{n} G(0, ...0, x_i, 0, ...0)$$

More concisely, if we write $G_i(a) = G(\underbrace{0, ...0}_{i-1\ 0's}, a, 0, ...0)$ then:

$$G(x) = \sum_{i=1}^{n} G_i(x_i)$$

Now if $a, b \in [0,1]$ so that $a + b \in [0,1]$, then from before:

$$\begin{aligned} G_i(a+b) &= G(0, ...0, a+b, 0, ...0) \\ &= G(0, ...0, a, 0, ...0) + G(0, ...0, b, 0, ...0) \\ &= G_i(a) + G_i(b), \end{aligned}$$

i.e. each G_i satisfies the Cauchy functional equation on the unit interval. Note $\forall n \in \mathbb{N}^+$:

$$G_i(x) = G_i\left(\frac{n}{n} \cdot x\right) = n \cdot G_i\left(\frac{x}{n}\right),$$

i.e.:

$$\frac{1}{n} \cdot G_i(x) = G_i\left(\frac{x}{n}\right)$$

Now if $m \in \mathbb{N}_{\geq 0}$ so that $\frac{m}{n} \in \mathbb{Q}_{\geq 0} \cap [0,1]$ then:

$$G_i\left(\frac{m}{n} \cdot x\right) = m \cdot G_i\left(\frac{x}{n}\right) = \frac{m}{n} \cdot G_i(x)$$

Finally $\forall r \in [0,1]$ can pick a rational $q_r \in [0,1]$ arbitrarily close so that $d_r = r - q_r \in [0, \frac{1}{N}]$ for any given $N \in \mathbb{N}$. Note:

$$\begin{aligned} G_i(r) - G_i(1) \cdot r &= G_i(r - q_r + q_r) - G_i(1) \cdot r \\ &= G_i(r - q_r) + G_i(q_r) - G_i(1) \cdot r \\ &= G_i(r - q_r) + (q_r - r) \cdot G_i(1) \\ &= G_i(d_r) - d_r \cdot G_i(1) \end{aligned}$$

We know $\mathrm{Im}(G_i) \subseteq [0,1]$ as its values must be probabilities. Thus $G_i(d_r) \leq \frac{1}{N}$ (otherwise $G_i(N \cdot d_r) = N \cdot G_i(d_r) > 1$) and $G_i(1) \leq 1$. Hence:

$$|G_i(r) - G_i(1) \cdot r| = |G_i(d_r) - d_r \cdot G_i(1)| \leq |G_i(d_r)| + d_r \cdot |G_i(1)| \leq \frac{2}{N}$$

Taking the limit as $N \to \infty$ yields:

$$G_i(r) = G_i(1) \cdot r$$

Thus $\forall i \in [n]$, G_i is of the form:

$$G_i(x) = w_i \cdot x$$

where $w_i \in [0, 1]$.

Putting things together, this implies for any given opinion pool $\{P_1, P_2, ...P_n\} \in \mathcal{M}^n$, $\forall A \in \mathcal{A}$:

$$\begin{aligned} C(P_1, P_2, ...P_n)(A) &= G(P_1(A), P_2(A), ...P_n(A)) \\ &= \sum_{i=1}^{n} G_i(P_i(A)) \\ &= \sum_{i=1}^{n} w_i \cdot P_i(A) \end{aligned}$$

In particular, when $A = \Omega$ we have:

$$1 = C(P_1, P_2, ...P_n)(\Omega) = \sum_{i=1}^{n} w_i \cdot P_i(\Omega) = \sum_{i=1}^{n} w_i$$

as desired. Phew! $\square$

IPP + SSFP $\implies$ Dictatorship

As before, consider three disjoint non-empty subsets $A, B, C \in \mathcal{F}$ which partition Ω. By the above, SSFP implies $\exists w_1, w_2...w_n \in [0, 1]$ summing to 1 so that for any given opinion pool $\{P_1, P_2, ...P_n\} \in \mathcal{M}^n$:

$$\begin{aligned} C(P_1, P_2, ...P_n)(A) &= \sum_{i=1}^{n} w_i P_i(A) \\ C(P_1, P_2, ...P_n)(B) &= \sum_{i=1}^{n} w_i P_i(B) \\ C(P_1, P_2, ...P_n)(C) &= \sum_{i=1}^{n} w_i P_i(C) \end{aligned}$$

Since not all w_i can be zero we know $\exists k \in [n]$ so that $w_k > 0$. We construct an opinion pool $\{P'_1, P'_2, ...P'_n\}$ so that:

$$(P'_k(A),\ P'_k(B), P'_k(C)) = \left(0, \frac{1}{2}, \frac{1}{2}\right)$$

and $\forall i \in [n]\backslash\{k\}$:

$$(P'_i(A),\ P'_i(B), P'_i(C)) = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$$

Note:

$$C(P'_1, P'_2, ...P'_n)(A) = \frac{1 - w_k}{2}$$
$$C(P'_1, P'_2, ...P'_n)(B) = \frac{1}{2}$$
$$C(P'_1, P'_2, ...P'_n)(C) = \frac{w_k}{2}$$

Consider the events $S = A \cup B$ and $T = B \cup C$. By the above:

$$C(P'_1, P'_2, ...P'_n)(S) = \frac{1 - w_k}{2} + \frac{1}{2} = 1 - \frac{w_k}{2}$$
$$C(P'_1, P'_2, ...P'_n)(T) = \frac{1}{2} + \frac{w_k}{2}$$

Since $P'_k(S) = P'_k(A) + P'_k(B) = \frac{1}{2}$ and $P'_k(T) = P'_k(B) + P'_k(C) = 1$:

$$P'_k(S \cap T) = P'_k(B) = \frac{1}{2} = \frac{1}{2} \cdot 1 = P'_k(S) \cdot P'_k(T)$$

Additionally, since $\forall i \in [n]\backslash\{k\}$, $P'_i(S) = P'_i(A) + P'_i(B) = 1$ and $P'_i(T) = P'_i(B) + P'_i(C) = \frac{1}{2}$:

$$P'_i(S \cap T) = P'_i(B) = \frac{1}{2} = 1 \cdot \frac{1}{2} = P'_i(S) \cdot P'_i(T)$$

Thus by IPP, we must have:

$$C(P'_1, P'_2, ...P'_n)(B) = C(P'_1, P'_2, ...P'_n)(S \cap T) = C(P'_1, P'_2, ...P'_n)(S) \cdot C(P'_1, P'_2, ...P'_n)(T)$$

Hence:

$$\frac{1}{2} = \left(1 - \frac{w_k}{2}\right) \cdot \left(\frac{1}{2} + \frac{w_k}{2}\right)$$

The only solutions to the above quadratic are $w_k = 0$ and $w_k = 1$. However, $w_k > 0$ by assumption, implying $w_k = 1$ and all other $w_i = 0$. This means for any given opinion pool $\{P_1, P_2, ...P_n\} \in \mathcal{M}^n$, $\forall A \in \mathcal{F}$:

$$C(P_1, P_2, ...P_n)(A) = \sum_{i=1}^{n} w_i \cdot P_i(A) = P_k(A)$$

In other words, we have found our dictator! Thus, given more than 2 measurable outcomes, *no* consensus function can satisfy IPP, SSFP, and non-dictatorship simultaneously.

Given the above impossibility result, how are we supposed to aggregate beliefs surrounding the odds of an event? Note that we have assumed our agents' beliefs are *static* so far. Hence, a clever way to side step our trilemma might involve letting our agents interact with one another, causing their beliefs to become *dynamic*. Even better, maybe everyone will agree with each other after communicating, thus producing a single collective probability distribution. Indeed, a generalization of Aumann's agreement theorem (which we will now explore) strongly suggests that this will occur when two given agents are interacting.

1.5 Aumann's Agreement Theorem

1.5.1 Introduction

First articulated by Robert Aumann in a paper titled "Agreeing to Disagree," the agreement theorem asserts that "people with the same priors cannot agree to disagree" [Aum76]. To see what this exactly means, we'll have to jump into some definitions again.

Let $(\Omega, \mathcal{F})$ be a measurable space of states of the world.
Let μ be a prior common to agent a and agent b.
Let $Q^a = \{Q_1^a, Q_2^a, ...Q_K^a\}$ be the information partition of agent a.
Similarly, let $Q^b = \{Q_1^b, Q_2^b, ...Q_L^b\}$ be the information partition of agent b.

If $\omega \in \Omega$ is the true state of the world then agent i is informed of $Q^i(\omega)$, which is the element of Q^i that contains ω. For an event $E \in \mathcal{F}$, if we have $E \subseteq Q^i(\omega)$, then we say agent i knows E. Additionally, if E includes that

member of the meet $Q^a \wedge Q^b$ (the finest common coarsening) which contains ω then we call E common knowledge. More informally, an event E is common knowledge if a knows E, b knows E, a knows that b knows E, b knows that a knows E etc.

Let A be an event and let $q^i(\omega) = \frac{P(A \cap Q^i(\omega))}{P(Q^i(\omega))}$ be agent i's posterior. If it is common knowledge at ω that $q^i(\omega) = q_i$ then let $P(\omega)$ be that member of the meet which contains ω. Note we can write:

$$P(\omega) = \cup_j \ Q_j^a$$

where the $Q_j^a \in Q^a$ partition $P(\omega)$. Hence:

$$\mu(A \mid P(\omega)) = \frac{\mu(\cup_j \ A \cap Q_j^a)}{\mu(\cup_j \ Q_j^a)} = \frac{\sum_j \mu(Q_j^a) \cdot \mu(A \mid Q_j^a)}{\sum_j \mu(Q_j^a)} = \frac{\sum_j \mu(Q_j^a) \cdot q_a}{\sum_j \mu(Q_j^a)} = q_a$$

by symmetry it also follows that:

$$\mu(A \mid P(\omega)) = q_b$$

thus:

$$q_a = q_b$$

Hence two agents cannot agree (have common knowledge of posteriors) to disagree (which are unequal).

1.5.2 Bringing Things to Life

We will now examine a generalization of Aumann's theorem by Geanakoplos and Polemarchakis 1982 [GP82] which does not assume a and b have common knowledge of each others' posteriors. Instead a and b "can't disagree forever" if they interact dynamically as follows:

Step 1- Agent a announces their posterior:

$$q_1^a = \frac{\mu(A \cap Q^a(\omega))}{\mu(Q^a(\omega))}$$

meaning a could be informed of partitions:

$$a_1 = \{k \mid \frac{\mu(A \cap Q_k^a)}{\mu(Q_k^a)} = q_1^a\}$$

Agent b will then adjust their posterior and announce it as:

$$q_1^b = \frac{\mu(A \cap (Q^b(\omega) \cap (\cup_{k \in a_1} Q_k^a)))}{\mu(Q^b(\omega) \cap (\cup_{k \in a_1} Q_k^a))}$$

meaning b could be informed of partitions:

$$b_1 = \{l \mid \frac{\mu(A \cap (Q_l^b \cap (\cup_{k \in a_1} Q_k^a)))}{\mu(Q_l^b \cap (\cup_{k \in a_1} Q_k^a))} = q_1^b\}$$

Step t- Agent a will then adjust their posterior and announce it as:

$$q_t^a = \frac{\mu(A \cap (Q^a(\omega) \cap (\cup_{l \in b_{t-1}} Q_l^b)))}{\mu(Q^a(\omega) \cap (\cup_{l \in b_{t-1}} Q_l^b))}$$

meaning a could be informed of partitions:

$$a_t = \{k \in a_{t-1} \mid \frac{\mu(A \cap (Q_k^a \cap (\cup_{l \in b_{t-1}} Q_l^b)))}{\mu(Q_k^a \cap (\cup_{l \in b_{t-1}} Q_l^b))} = q_t^a\}$$

Agent b will then adjust their posterior and announce it as:

$$q_t^b = \frac{\mu(A \cap (Q^b(\omega) \cap (\cup_{k \in a_t} Q_k^a)))}{\mu(Q^b(\omega) \cap (\cup_{k \in a_t} Q_k^a))}$$

meaning b could be informed of partitions:

$$b_t = \{l \in b_{t-1} \mid \frac{\mu(A \cap (Q_l^b \cap (\cup_{k \in a_t} Q_k^a)))}{\mu(Q_l^b \cap (\cup_{k \in a_t} Q_k^a))} = q_t^b\}$$

Importantly note that q_t^a is a function of b_{t-1}, a_t is a function of a_{t-1}, b_{t-1} and q_t^a, and q_t^b is a function of a_t. Thus if $a_t = a_{t-1}$ and $b_t = b_{t-1}$ then $q_{t+1}^a = q_t^a$, $a_{t+1} = a_t$ and $q_{t+1}^b = q_t^b$. Since b_t is a function of b_{t-1}, a_t, and q_t^b, it also follows that $b_{t+1} = b_t$. In other words, if $a_t = a_{t-1}$ and $b_t = b_{t-1}$ then our agents have reached a communication equilibrium.

Moreover, since $\forall i \in \mathbb{N}$, $(a_i \supseteq a_{i+1}) \wedge (b_i \supseteq b_{i+1})$ and $\max |a_1| = K$ and $\max |b_1| = L$, we know $\exists j \in [K + L]$ s.t.:

$$(a_j = a_{j+1}) \wedge (b_j = b_{j+1})$$

Thus we are guaranteed convergence.

To illustrate the point consider the following example where $(\Omega, \mathcal{F}, \mu) = \left([8], 2^{[8]}, \frac{1}{8} \cdot \#\right)$, $\omega = 1$ and:

$$Q^a = \{\{1,2,3,4,6\},\{5,7,8\}\}$$
$$Q^b = \{\{1,3,5,7\},\{2,4,6,8\}\}$$

Additionally, suppose the event being communicated about is $A = \{3,4\}$.

Initially, a will report a posterior of $\frac{2}{5}$. Then, b will realize a must be informed of the partition $\{1,2,3,4,5,6\}$ and the only possible true states of the world are $\{1,3\}$. Thus b will report a posterior of $\frac{1}{2}$. Next, a will realize b must have been informed of the partition $\{1,3,5,7\}$ since otherwise b would have reported a posterior of $\frac{1}{3}$. Subsequently, a will also report a posterior of $\frac{1}{2}$ and a communication equilibrium will be reached since each agent knows the partition of the other.

Just for fun, what would have occurred if b started? Well the initial posterior reported would be $\frac{1}{4}$. Since to a, this probability is consistent with both of b's partitions, a will still think $\{1,2,3,4,6\}$ are the possible true states of the world. Thus a will report a posterior of $\frac{2}{5}$ and the process proceeds as outlined above.

Odd that our agents posteriors converged to the same thing no? Suppose at some step we have $a_{t+1} = a_t$ and $b_{t+1} = b_t$ (as we are guaranteed). Since $\forall k \in a_{t+1} = a_t$:

$$\mu(A \cap \left(Q^a_k \cap \left(\cup_{l \in b_t} Q^b_l\right)\right) = q^a_{t+1} \cdot \mu(Q^a_k \cap \left(\cup_{l \in b_t} Q^b_l\right))$$

and $\forall l \in b_{t+1} = b_t$:

$$\mu(A \cap \left(Q^b_l \cap (\cup_{k \in a_t} Q^a_k)\right) = q^b_{t+1} \cdot \mu(Q^b_l \cap \left(\cup_{k \in a_t} Q^a_k\right))$$

we have:

$$\sum_{k \in a_t} \mu(A \cap \left(Q^a_k \cap \left(\cup_{l \in b_t} Q^b_l\right)\right) = q^a_{t+1} \cdot \sum_{k \in a_t} \mu(Q^a_k \cap \left(\cup_{l \in b_t} Q^b_l\right))$$

and:

$$\sum_{l \in b_t} \mu(A \cap \left(Q^b_l \cap (\cup_{k \in a_t} Q^a_k)\right) = q^b_{t+1} \cdot \sum_{l \in b_t} \mu(Q^b_l \cap \left(\cup_{k \in a_t} Q^a_k\right))$$

Hence:

$$\frac{\mu(A \cap \left(\cup_{k \in a_t} Q_k^a \cap \left(\cup_{l \in b_t} Q_l^b\right)\right)}{\mu(\cup_{k \in a_t} Q_k^a \cap \left(\cup_{l \in b_t} Q_l^b\right))} = \frac{\mu(A \cap (\cup_{k \in a_t} \cup_{l \in b_t} Q_k^a \cap Q_l^b))}{\mu(\cup_{k \in a_t} \cup_{l \in b_t} Q_k^a \cap Q_l^b)} = q_{t+1}^a$$

and:

$$\frac{\mu(A \cap \left(\cup_{l \in b_t} Q_l^b \cap (\cup_{k \in a_t} Q_k^a)\right)}{\mu(\cup_{l \in b_t} Q_l^b \cap (\cup_{k \in a_t} Q_k^a))} = \frac{\mu(A \cap (\cup_{l \in b_t} \cup_{k \in a_t} Q_l^b \cap Q_k^a))}{\mu(\cup_{l \in b_t} \cup_{k \in a_t} Q_l^b \cap Q_k^a)} = q_{t+1}^b,$$

i.e. $q_{t+1}^a = q_{t+1}^b$

1.5.3 Common Knowledge Will Equilibriate

To see why Aumann's result is a special case of Geanakoplos' and Polemarchakis' result suppose the posteriors of agents a and b are common knowledge in world ω.

Since agent b knows a's initial posteriors, when a announces their posteriors, b already knows this posterior from their current partition. Thus b will just report their initial posteriors.

Since a knows b knows a's posteriors, every element of a's partition already indicated the announcement would not eliminate anything from b's initial partition.

Additionally, since a knows b's initial posteriors, a's current partition already contains all relevant information. Thus a will just report their initial posteriors.

Since b knows a knows b knows a's posteriors and b knows a knows b's posteriors, every element of b's partition already indicated the announcement would not eliminate anything from a's initial partition.

Additionally, since b knows a's initial posteriors, b's current partition already contains all relevant information. Thus b will just report their initial posteriors.

Through symmetry, we can continue this process identically ad infinitum.

Thus, if the agents' posteriors are common knowledge at ω, then agents a and b can't disagree forever and must in fact agree!

Chapter 2

Prediction Markets

2.1 Prediction Markets

Now that we are heuristically justified in adopting a dynamic approach to aggregation, we are finally ready to see the mechanism which drives this dynamic: prediction markets! All we need is a way to incentivize traders to report their true beliefs and we are good to go since the traders will accordingly change their minds upon listening to one another and submit new reports. Going forward we will use many key results and terms from information theory, so the unfamiliar reader may want to read the primer in the appendix before proceeding.

2.1.1 Scoring Rules

Our description of prediction markets follows that laid out by Hanson 2003 [Han03].

The common approach used in eliciting the beliefs of a single individual employs something called a *scoring rule*. Let our subject's probability distribution over a disjoint partition of states $\{1, 2, ...N\}$ be represented by the vector $\vec{p}$. If their report is represented by $\vec{r}$ then a scoring rule s awards the subject $s_i(\vec{r})$ dollars in case of event i. A *proper* scoring rule satisfies:

$$\vec{p} = \text{argmax } \mathbb{E}[s(\vec{r})] = \text{argmax} \sum_{i=1}^{N} p_i \cdot s_i(\vec{r})$$

and:

$$\mathbb{E}[s(\vec{p})] = \sum p_i \cdot s_i(\vec{p}) \geq 0$$

for any $\vec{p}$. These constraints together imply that a subject has incentive to report their true probability distribution when paid out according to the scoring rule s.

The most natural proper scoring rule is the logarithmic scoring rule of the form:

$$a_i + b \cdot \ln(r_i); a_i, b > 0$$

To see why this scoring rule is proper, note that a_i and b can be set so that the agent's expected reward for telling the truth (an affine transformation of the reported distribution's entropy) is always non-negative and:

$$\begin{aligned}
\operatorname{argmax} \sum_{i=1}^{N} p_i \cdot s_i(\vec{r}) &= \operatorname{argmax} \sum_{i=1}^{N} p_i \cdot (a_i + b \ln(r_i)) \\
&= \operatorname{argmax} \sum_{i=1}^{N} p_i \cdot a_i + \sum_{i=1}^{N} p_i \cdot b \ln(r_i) \\
&= \operatorname{argmax}\ b \sum_{i=1}^{N} p_i \cdot \ln(r_i) \\
&= \operatorname{argmin}\ -b \sum_{i=1}^{N} p_i \cdot \ln(r_i) \\
&= \operatorname{argmin}\ b \sum_{i=1}^{N} p_i \cdot \ln(p_i) - b \sum_{i=1}^{N} p_i \cdot \ln(r_i) \\
&= \operatorname{argmin}\ b \sum_{i=1}^{N} p_i \cdot \ln\left(\frac{p_i}{r_i}\right) \\
&= \operatorname{argmin}\ b \cdot KL(\vec{p} \mid\mid \vec{r}) \\
&= \vec{p}
\end{aligned}$$

2.1.2 Market Scoring Rules

How do we go about eliciting the beliefs of *multiple traders*? Hypothetically, we could simply use a separate scoring rule for each individual. However, this

is expensive. Alternatively, in a market scoring rule, the market makes some initial default report and whenever a future trader submits a new report, they always agree to pay the award of the report last submitted. Thus, given that the nth report was r^n the person who submits the $n+1$st report expects to make:

$$s_i(r^{n+1}) - s_i(r^n)$$

in case of event i. This person also has incentive to report their true beliefs since $\mathbb{E}[s(r^n)]$ is given and guaranteed to be less than or equal to the maximum value of $\mathbb{E}[s(r^{n+1})]$.

The 0th trader is the market, so if there are N reports submitted, the market only need pay out:

$$\sum_{i=1}^{N} s_i(r^i) - s_i(r^{i-1}) = s_i(r^N) - s_i(r^0)$$

in case of event i.

If we use a *logarithmic* market scoring rule (LMSR) each person expects to make:

$$b \cdot \ln\left(\frac{r_i^{n+1}}{r_i^n}\right)$$

in case of event i and:

$$\sum_{i=1}^{k} b \cdot r_i^{n+1} \ln\left(\frac{r_i^{n+1}}{r_i^n}\right) = b \cdot KL(r^{n+1} \mid\mid r^n)$$

overall. That is to say, each trader in a prediction market using a LMSR expects to profit proportional to how many bits of information they improved the old report by.

Moreover, if the market starts with an initial report that is uniform over all k outcomes, then in the worst case scenario it will need to pay out:

$$\max_{r^f}[s_i(r^f) - s_i(r^0)] = \max_{r^f} b \cdot \left(\ln(r_i^f) - \ln\left(\frac{1}{N}\right)\right) = \max_{r^f} b \cdot \ln\left(\frac{r_i^f}{1/N}\right) = b \cdot \ln(N)$$

which can be thought of as the subsidy necessary to run the market.

2.1.3 Reneging

The logarithmic market scoring rule provides a tool which creates incentives for agents to tell the truth during each step of the process in Aumann's agreement theorem. However, this incentive is only myopic. Intuitively, agents could lie to trick others into making worse reports and profit at their expense. Agents could also wait for others to report first so they have better info. These are known as the *bluffing* and *reticence* problems.

For a concrete example, consider two agents interacting via a LMSR using the example where $\omega = 1$,:

$$Q^a = \{\{1, 2, 3, 4, 6\}, \{5, 7, 8\}\}$$
$$Q^b = \{\{1, 3, 5, 7\}, \{2, 4, 6, 8\}\}$$

and $A = \{3, 4\}$. Suppose the market's default estimate is $\mu(A) = \frac{1}{2}$ and $\mu(\neg A) = \frac{1}{2}$

If agent a were to start then $q_1^a = \frac{2}{5}$, $q_1^b = \frac{1}{2}$, and $q_2^a = \frac{1}{2}$. In this case, after adjusting their initial prior, agent a would expect to make an amount proportional to:

$$\frac{1}{2} \cdot \left(\ln\left(\frac{2/5}{1/2}\right) + \ln\left(\frac{3/5}{1/2}\right) \right) \approx -0.0204$$

on their first report (i.e. lose money)!

If agent b were to start then $q_1^b = \frac{1}{4}$, $q_1^a = \frac{2}{5}$, and $q_1^b = \frac{1}{2}$. Hence, agent b will eventually expect to lose money on their first report as well since:

$$\frac{1}{2} \cdot \left(\ln\left(\frac{1/4}{1/2}\right) + \ln\left(\frac{3/4}{1/2}\right) \right) \approx -0.1438$$

This is no good! Both agents might indefinitely wait for the other to report first so they can improve their prediction.

Alternatively, consider the scenario where $\omega = 1$,:

$$Q^a = \{\{1, 2, 3, 4, 6\}, \{5, 7, 8, 9\}\}$$
$$Q^b = \{\{1, 2, 5, 7\}, \{3, 4, 6, 8, 9\}\}$$

and $A = \{3, 6, 9\}$ with the market's default estimate set to $\mu(A) = \frac{1}{2}$ as before.

If agent b chooses to start and plays truthfully then they will announce a posterior of 0, causing agent a to also revise their posterior to 0. Thus in this case, agent a would expect to make nothing, while agent b would expect to make an amount proportional to:

$$1 \cdot \ln\left(\frac{1}{1/2}\right) = \ln(2)$$

However, agent b could make more by lying at agent a's expense! For instance, if agent b reports a posterior of $\frac{3}{5}$ and agent a assumes agent b is playing truthfully, then a will think b is informed of the partition $\{3, 4, 6, 8, 9\}$ and report a posterior of $\frac{2}{3}$. Next, agent b could reveal their lie and report a posterior of 0 causing agent a to also revise their posterior to 0. Overall then, agent b would expect to net an amount proportional to:

$$1 \cdot \ln\left(\frac{2/5}{1/2}\right) + 1 \cdot \ln\left(\frac{1}{1/3}\right) = \ln\left(\frac{12}{5}\right)$$

While agent a would expect to net:

$$\ln\left(\frac{1/3}{2/5}\right) = \ln\left(\frac{5}{6}\right)$$

Theoretically, we can remedy the myopic nature of the LMSR's incentives if we allow agents who change their beliefs to safely *renege* (or withdraw) prior reports. If someone agrees to pay the award of a prediction that later gets reneged, we simply make them pay the award of the preceding prediction instead! Of course, this means when an agent submits a report they no longer agree to pay the reward of only the prediction before them. Rather, they agree to pay the reward of *some* prediction submitted before them. This maintains the incentive to submit reports since an agent expects to make money so long as the prediction they pay is different from theirs and lose nothing if it is identical. Thus, we can quell agents' fears about retroactively expecting to lose money and also protect them from being lied to. However, this comes at the cost of agents not knowing the exact structure of their payoff until the market settles.

2.2 Reneging Experiment

To experimentally test whether adding a reneging option actually helps participants reach communication equilibria more often, I have written a Python script to generate scenarios where at least one agent retroactively expects to lose money on a report at some point in the agreement process. The script can be found here– Aumann Scenario Generator. In future work, these scenarios will be used to conduct 75 rounds of 2 different two-player games on Amazon Mechanical Turk (150 total). A single round in both games will consist of players attempting to make as much money as possible across 10 separate scenarios. For a given scenario the game will:

1. Display both information partitions
2. Based on past reports
   - Tell players
     - What their posterior is/which worlds are still plausible
     - What the expected payoff of their past reports are
     - What they expect to make on a new report
   - If game 1 (control): Give players options to report posterior or to leave
   - If game 2 (experimental): Give players options to report posterior, renege past reports, or to leave
3. The scenario ends and players are paid out (before being given one last chance to renege in game 2) if
   - One of the two players chooses to leave at any point
   - Agreement is reached

2.3 Rational Inattention

If we are designing a prediction market with the intent of agents incorporating each others beliefs into their reports (as they do in Aumann's model of interaction), then we might wonder how to incentivize such behavior. Alternatively, if we have a group of experts who can be negatively influenced by outside information, we might wonder how to prevent the acquisition of extra information. If a scoring rule disincentivizes the acquisition of outside information we call it *robust* [Tsa20]. Alternatively, if a scoring rule incentivizes information acquisition we call it *frail*. Luckily, there is a straight forward

way to control whether a LMSR is robust or frail, but before visiting this we will need to briefly discuss *rational inattention*.

Rational inattention models attempt to formalize what economist Herbert A. Simon called 'bounded rationality.' Simon argued that our decision making capabilities are limited by a finite capacity to process information and that "a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently" [Sim71]. By simply assigning a cost to information processing, rational inattention models are able to explain the emergent properties of bounded rationality across a wide range of empirical phenomena. In the words of the rational inattention's pioneer, Christopher Sims, the theory's approach is "to construct optimizing-agent models that are consistent with people not using freely available information" so as to "explain why some freely available information is not used, or imperfectly used" [Sim10].

2.3.1 Mutual Information Costs

Consider an event S which takes on some value $s \in \{s_1, s_2, ...s_N\}$. Suppose an agent's beliefs about S assigns probability $\mathbb{P}(S = s_n)$ to each outcome. Additionally suppose the agent can run an experiment X which takes on some value $x \in \{x_1, x_2, ...x_M\}$ to gain information about S and revise their probabilities to $\mathbb{P}(S = s_i \mid X = x)$. By running the experiment X, the agent expects to gain:

$$\begin{aligned} H(S) - H(S \mid X) &= \sum_n \mathbb{P}(S = s_n) \cdot \ln\left(\frac{1}{\mathbb{P}(S = s_n)}\right) \\ &\quad - \mathbb{E}\left[\sum_n \mathbb{P}(S = s_n \mid X = x) \cdot \ln\left(\frac{1}{\mathbb{P}(S = s_n \mid X = x)}\right)\right] \\ &= I(S; X) \end{aligned}$$

bits of information about S. In other words, the amount of information X reveals about S is simply the mutual information between S and X. Rational inattention models assign the cost of an agent running such an experiment to be:

$$c \cdot I(S; X),$$

i.e. the cost of running of an experiment is $\$c$ per bit of information that the agent pays attention to.

2.3.2 Mutual Information Benefits

Continuing the above, suppose our agent is being paid out according to an LMSR for reporting their beliefs and the most recent report made assigns $\mathbb{P}'(S = s_n)$ to each outcome. Then if our agent reports their beliefs without running the experiment X, they will expect to make:

$$b \cdot KL(\mathbb{P}_S, \mathbb{P}'_S) = b \cdot H(\mathbb{P}_S, \mathbb{P}'_S) - b \cdot H(S)$$

where $H(p, q)$ is the cross entropy of q relative to p.

However, if our agent runs the experiment and discovers $X = x$ and reports their new probabilities, they will expect to make:

$$b \cdot KL(\mathbb{P}_{S \mid X=x}, \mathbb{P}'_S) = \sum_n b \cdot \mathbb{P}(S = s_n \mid X = x) \ln\left(\frac{\mathbb{P}(S = s_n \mid X = x)}{\mathbb{P}'(S = s_n)}\right)$$

Thus, by running the experiment, on average they will expect to make:

$$\begin{aligned}
\mathbb{E}\left[b \cdot KL(\mathbb{P}_{S \mid X}, \mathbb{P}'_S)\right] &= \sum_m \mathbb{P}(X = x_m) \sum_n b \cdot \mathbb{P}(S = s_n \mid X = x_m) \ln\left(\frac{\mathbb{P}(S = s_n \mid X = x_m)}{\mathbb{P}'(S = s_n)}\right) \\
&= \sum_m \sum_n b \cdot \mathbb{P}(S = s_n, X = x_m) \ln\left(\frac{\mathbb{P}(S = s_n \mid X = x_m)}{\mathbb{P}'(S = s_n)}\right) \\
&= \sum_m \sum_n b \cdot \mathbb{P}(S = s_n, X = x_m) \ln\left(\frac{1}{\mathbb{P}'(S = s_n)}\right) \\
&\quad - \sum_m \sum_n b \cdot \mathbb{P}(S = s_n, X = x_m) \ln\left(\frac{1}{\mathbb{P}(S = s_n \mid X = x_m)}\right) \\
&= \sum_n \sum_m b \cdot \mathbb{P}(S = s_n, X = x_m) \ln\left(\frac{1}{\mathbb{P}'(S = s_n)}\right) - b \cdot H(S \mid X) \\
&= \sum_n b \cdot \mathbb{P}(S = s_n) \ln\left(\frac{1}{\mathbb{P}'(S = s_n)}\right) - b \cdot H(S \mid X) \\
&= b \cdot H(\mathbb{P}_S, \mathbb{P}'_S) - b \cdot H(S \mid X)
\end{aligned}$$

Hence on average, running the experiment will increase the amount the agent expects to make by:

$$\mathbb{E}\left[b \cdot KL(\mathbb{P}_{S \mid X}, \mathbb{P}'_S)\right] - b \cdot KL(\mathbb{P}_S, \mathbb{P}'_S) = b \cdot H(S) - b \cdot H(S \mid X) = b \cdot I(S; X),$$

i.e. they will benefit $\$b$ per bit of information gained from the experiment.

2.3.3 Cost Benefit Analysis

When an agent assesses whether or not to run an experiment, they will weigh whether the expected benefits of running the experiment outweigh the expected costs. In other words, they will evaluate whether:

$$b \cdot I(S;X) - c \cdot I(S;X) = (b-c) \cdot I(S;X) > 0$$

Since $I(S;X) \geq 0$ (as mutual information is always non-negative), this boils down to evaluating whether $b > c$. Thus an individual agent will choose to acquire outside information iff the market's award of $\$b$ per bit of information is greater than their personal cost of $\$c$ per bit of information.

Thus if we are running a prediction market with N participants whose information costs are $c_1, c_2, ...c_N$, a *robust* LMSR will set:

$$b < \min_{i \in [N]} c_i$$

so that all agents are disinectivized from acquiring additional information.

Similarly, a *frail* LMSR will set:

$$b > \max_{i \in [N]} c_i$$

so that all agents are incentivized to acquire additional information.

Naturally, we might wonder how to find the information costs associated with each agent. As suggested in Foley 2012 [Fol12], we can deduce the c_i by running lotteries to find the slope of the agents' logistic quantal response curves. If we assume rationally inattentive behavior, these slopes will directly correspond to the each agent's respective information cost. Another possible solution is to increase b very slowly from 0 and see at what points the market's prices move/re-equilibrate. Firstly, this would prove that the parameter b controls how much information is incorporated into our market. Secondly, this would also yield an upper bound for each agent's information cost equal to the point at which each agent revises their report.

Chapter 3

Thermodynamics

Intuitively, it isn't all too surprising that there is a deep connection between prediction markets and thermodynamics given the logarithmic market scoring rule's relationship to information theory. However, to see the full extent of the relationship we will need discuss an alternative way of running prediction markets.

3.1 Automated Market Makers

Automated market makers (AMM's) are an elegant technology that establish an equivalence between prices and probabilities to run prediction markets [CV10].

Given a disjoint partition of outcomes $\{1, 2, ..., N\}$ for some event, an AMM buys and sells securities corresponding to each outcome i of the form "pays \$1 if i occurs." Let $\vec{q}$ be the vector whose ith entry represents the number of shares of security i held by the traders. Since the point of an AMM is to easily interpret prices as probabilities, we want security i's price to increase if q_i increases (the traders purchase some amount of security i) and decrease if q_i decreases (the traders sell some amount of security i). However, this means that the price of a security must change for each infinitesimal amount of its shares that are bought or sold. Thus, prediction markets must use a 'cost function' C to dictate how much traders will have to pay for a specific transaction given the market's current state. For instance, if the market's participants currently hold q_i of each security collectively and a trader wishes

to now transact r_i shares of each security, then this action would cost them:

$$C(\vec{q}+\vec{r}) - C(\vec{q})$$

Note if we parameterize the path from $\vec{q}$ to $\vec{q}+\vec{r}$ by some function $a(t)$, then the price of security i at each point along the transaction is given by:

$$p_i(a(t)) = \frac{\partial C}{\partial a_i}$$

There are many different cost functions which can be used to run a prediction market, however the cost function:

$$C(\vec{q}) = b \cdot \ln \sum_{i=1}^{N} e^{q_i/b}$$

is particularly nice since its prices are of the form:

$$p_i(\vec{q}) = \frac{e^{q_i/b}}{\sum_{i=1}^{N} e^{q_i/b}},$$

i.e. the Boltzmann distribution!

The Boltzmann AMM is equivalent to a LMSR since if the market's current probabilities are given by the above vector of prices and a trader wishes to change them to:

$$p_i(\vec{q}^*) = \frac{e^{q_i^*/b}}{\sum_{i=1}^{N} e^{q_i^*/b}},$$

then in case of event i they will make:

$$\begin{aligned}
(q_i^* - q_i) - (C(\vec{q}^*) - C(\vec{q})) &= b \cdot \left(\frac{q_i^*}{b} - \ln \sum_{i=1}^{N} e^{q_i^*/b} \right) - b \cdot \left(\frac{q_i}{b} - \ln \sum_{i=1}^{N} e^{q_i/b} \right) \\
&= b \cdot \ln \left(\frac{e^{q_i^*/b}}{\sum_{i=1}^{N} e^{q_i^*/b}} \right) - b \cdot \ln \left(\frac{e^{q_i}/b}{\sum_{i=1}^{N} e^{q_i/b}} \right) \\
&= b \cdot \ln \left(\frac{p_i(\vec{q}^*)}{p_i(\vec{q})} \right)
\end{aligned}$$

That is to say, their expected payoff is:

$$b \sum_{i=1}^{N} p_i(\vec{q}^*) \cdot \ln \left(\frac{p_i(\vec{q}^*)}{p_i(\vec{q})} \right) = b \cdot KL(p(\vec{q}^*) \mid\mid p(\vec{q}))$$

3.2 The Second Law

3.2.1 Jarzynski's Equality

Now that we have established the Boltzmann AMM is equivalent to a LMSR, it follows the maximum amount of money the market must pay out to its participants is:

$$b \cdot \ln(N)$$

which can be thought of as the subsidy the market uses for its own funding pool.

Additionally since:

$$C(\vec{0}) = b \cdot \ln\left(\sum_{i=1}^{N} e^{0/b}\right) = b \cdot \ln(N)$$

and $C(\vec{q}) - C(\vec{0})$ represents how much money agents must give the market for it to issue $\vec{q}$ of each security, the cost function of a Boltzmann AMM indicates how much money agents have given the market in the course of making bets in addition to the market's own funding pool. Therefore, the amount of money the market has given out at any point in time is:

$$F_{\text{AMM}}(\vec{q}) = -(C(\vec{q}) - b \cdot \ln(N)) = -b \cdot \ln \frac{\sum_{i=1}^{N} e^{q_i/b}}{N} = -b \cdot \ln\langle e^{q_i/b}\rangle$$

This exactly parallels Jarzynski's equality and suggests we should start drawing some analogies! Note that if we interpret the amount of security i held by the traders as the work performed by the AMM in state i, then F_{AMM} (or the amount of money the market has given out) is identical to the equilibrium free energy in a thermodynamic process. This makes sense since the more money the market gives out, the more its capacity to perform work (have securities bought from it).

This also allows us to immediately conclude that the money given out by the market at any given point in time is bounded above by the average num-

ber of shares sold to the market for each security i since by Jensen's:

$$\begin{aligned}
&\langle e^{q_i/b} \rangle \geq e^{\frac{\sum_{i=1}^{N} q_i/b}{N}} \\
&\rightarrow \ln\langle e^{q_i/b} \rangle \geq \frac{\sum_{i=1}^{N} q_i/b}{N} \\
&\rightarrow -b \cdot \ln\langle e^{q_i/b} \rangle \leq -\frac{\sum_{i=1}^{N} q_i}{N} \\
&\rightarrow F_{\text{AMM}}(\vec{q}) \leq -\frac{\sum_{i=1}^{N} q_i}{N},
\end{aligned}$$

giving a version of the second law of thermodynamics!

3.2.2 The Second Law for any Transaction

The above version of the second law only applies for a transaction in which there is intially 0 of each security issued.

In general, if the amount of each security sold changes from q_i to q_i^* then the average work performed on the market (or the expected number of paying shares the market bought) is:

$$\overline{W} = \sum_{i=1}^{N} -(q_i^* - q_i) \cdot p_i(\vec{q}) = \sum_{i=1}^{N} W_i \cdot p_i(\vec{q})$$

and the change in free energy of our market (or the amount of money it gives out in the transaction) is:

$$\begin{aligned}
\Delta F_{\text{AMM}} &= \left(-b \cdot \ln\langle e^{q_i/b} \rangle + (C(\vec{q}) - C(\vec{q}^*))\right) - \left(-b \cdot \ln\langle e^{q_i/b} \rangle\right) \\
&= b \cdot \ln\left(\frac{\sum_{i=1}^{N} e^{q_i/b}}{\sum_{i=1}^{N} e^{q_i^*/b}}\right) \\
&= -b \cdot \ln\left(\sum_{i=1}^{N} e^{(q_i^* - q_i)/b} \cdot \frac{e^{q_i/b}}{\sum_{j=1}^{N} e^{q_j/b}}\right) \\
&= -b \cdot \ln\left(\sum_{i=1}^{N} e^{-W_i/b} \cdot p_i(\vec{q})\right)
\end{aligned}$$

We wish to show:

$$\Delta F_{\text{AMM}} \leq \overline{W},$$

i.e. the amount of money given out by the market in a transaction is bounded above by the expected number of paying shares the market bought.

Note by Jensen's:

$$\begin{aligned}
&\sum_{i=1}^{N} e^{-W_i/b} \cdot p_i(\vec{q}) \geq e^{\sum_{i=1}^{N} -\frac{W_i}{b} \cdot p_i(\vec{q})} \\
&\rightarrow \ln\left(\sum_{i=1}^{N} e^{-W_i/b} \cdot p_i(\vec{q})\right) \geq \sum_{i=1}^{N} -\frac{W_i}{b} \cdot p_i(\vec{q}) \\
&\rightarrow -b \cdot \ln\left(\sum_{i=1}^{N} e^{-W_i/b} \cdot p_i(\vec{q})\right) \leq \sum_{i=1}^{N} W_i \cdot p_i(\vec{q})
\end{aligned}$$

as desired.

Hence we have recovered the second law of thermodynamics for any arbitrary transaction that takes $\vec{q}$ to $\vec{q}^*$!

3.3 Thermodynamic Operations

We can also interpret the quantities of each security i issued at any given point in time as determining the energy landscape of our system.

If a trader's beliefs $\text{Pr}_{\text{trader}}$ disagrees with the current probability distribution given by the AMM's prices Pr_{AMM} then according to their coarse graining the free energy is:

$$F_{\text{trader}} = -b \cdot \ln\langle e^{q_i/b}\rangle + b \cdot KL(\text{Pr}_{\text{trader}} \mid\mid \text{Pr}_{\text{AMM}}),$$

i.e. they believe the market has given out $b \cdot KL(\text{Pr}_{\text{trader}} \mid\mid \text{Pr}_{\text{AMM}})$ more dollars than the market believes. This strengthens our analogy since this expression is identical to that of the non-equilibrium free energy in a thermodynamic system.

If the trader carries out the appropriate transaction to adjust the market's

prices to $\Pr_{\text{trader}}$ (by changing the amount of each security sold from $\vec{q}$ to $\vec{q}^*$), then the AMM will believe the change in free energy is:

$$\begin{aligned}\Delta F_{\text{AMM}} &= \left(-b \cdot \ln\langle e^{q_i/b}\rangle + (C(\vec{q}) - C(\vec{q}^*))\right) - \left(-b \cdot \ln\langle e^{q_i/b}\rangle\right) \\ &= b \cdot \ln\left(\frac{\sum_{i=1}^{N} e^{q_i/b}}{\sum_{i=1}^{N} e^{q_i^*/b}}\right)\end{aligned}$$

and the trader will believe the change in free energy is:

$$\begin{aligned}\Delta F_{\text{trader}} &= \left(-b \cdot \ln\langle e^{q_i/b}\rangle + (C(\vec{q}) - C(\vec{q}^*))\right) - \left(-b \cdot \ln\langle e^{q_i/b}\rangle + b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{AMM}})\right) \\ &= b \cdot \ln\left(\frac{\sum_{i=1}^{N} e^{q_i/b}}{\sum_{i=1}^{N} e^{q_i^*/b}}\right) - b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{AMM}})\end{aligned}$$

This is nice because the first quantity tracks with the change in free energy for a system that starts and ends in equilibrium and the second quantity tracks with the change in free energy for a system that starts out of and ends in equilibrium.

Moreover, we can interpret the discrepancy between the free energies as due to the 'arbitrage opportunity' arising from the system being out of equilibrium no longer existing for the trader and never having existed for the market.

Alternatively, suppose our trader waits to adjust the market's prices to their beliefs and another participant comes along in the mean time and adjusts the market's prices to $\Pr_{\text{participant}}$ (by changing the amount of each security sold from $\vec{q}$ to $\vec{q}^*$). Then the trader will now believe the change in free energy is:

$$\begin{aligned}\Delta F_{\text{trader}} &= \left(-b \cdot \ln\langle e^{q_i/b}\rangle + (C(\vec{q}) - C(\vec{q}^*)) + b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{participant}})\right) \\ &\quad - \left(-b \cdot \ln\langle e^{q_i/b}\rangle + b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{AMM}})\right) \\ &= b \cdot \ln\left(\frac{\sum_{i=1}^{N} e^{q_i/b}}{\sum_{i=1}^{N} e^{q_i^*/b}}\right) + (b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{participant}}) - b \cdot KL(\Pr_{\text{trader}} \,||\, \Pr_{\text{AMM}}))\end{aligned}$$

which tracks with the change in free energy for a system that starts and ends out of equilibrium.

Thus, we can summarize the basic process underlying prediction markets as below:

$\rightarrow$ [Initial market/landscape has certain amount of free energy corresponding to $\vec{q}$]

$\rightarrow$ [Trader carries out transaction/performs work to profit from extra non-equilibrium free energy]

$\rightarrow$ [Adjustment shifts market/landscape so that free energy corresponds to new $\vec{q}^{*}$]

$\rightarrow$ [Rinse and repeat]

Chapter 4

Conclusion

Prediction markets are a mechanism for aggregating a group of agents' beliefs about an event. By dropping the assumption of a static consensus function, prediction markets bypass a probabilistic analogue of Arrow's impossibility trilemma. However, in order for traders to settle on a final set of prices they all agree on, three conditions need to be met:

(1) Traders must have common priors as per Aumann's agreement theorem

(2) Traders must pay attention to each other's reports

(3) Traders must act myopically since truth-telling isn't necessarily optimal if one considers profits from future trades

Theoretically, we can remove the third condition and overcome the bluffing and reticence problems by adding an option to renege on reports. Allowing this behavior removes the incentive to wait for better information before reporting and helps traders protect themselves in case they feel they've been deceived. The best way to empirically gauge whether allowing reneging will actually improve market performance will be through experiments on Amazon Mechanical Turk which I plan on carrying out over these next few years.

Applying rational inattention models to prediction markets sheds light on when the second condition will be met and how we can control whether agents incorporate outside information into their forecasts. Namely, these models predict an LMSR will be frail iff the market's award per bit of information is higher than the maximum of the agents' personal costs per bit of information. We propose a few methods for computing these information

costs as well as experimentally proving that increasing the market's award increases the amount of information aggregated.

Finally, prediction markets have deep connections to information theory and non-equilibrium statistical mechanics, namely in their AMM implementation. If we interpret selling securities as the market performing work and the money given out by the market as the free energy, we can derive a version of the second law of thermodynamics! When a trader disagrees with the market's prices, they perceive it to be 'out of equilibrium' and believe they can profit from the additional free energy. These analogies lay the preliminary groundwork for future research at the intersection of thermodynamics and prediction markets.

To collaborate or convey any questions, comments, or concerns please contact me at my Pomona email: snab2018@mymail.pomona.edu.

Appendix A

Information Theory Primer

A.1 A Game

Suppose your friend, Alex, is thinking of a random integer:

$$R \sim \text{Unif}(\{0, 1...1023\})$$

Furthermore, suppose you can ask Alex yes or no questions to deduce his number. What strategy should you employ to find R using the fewest number of expected questions?

One way of approaching the problem is to figure out the binary representation of Alex's number as follows:

"Is the first bit 1?"

"Is the second bit 1?"

...

"Is the tenth bit 1?"

Each question is guaranteed to eliminate half of the possible values for R, and after asking all 10 questions we are guaranteed to know Alex's number. Thus, we might say knowing R requires "10 bits of information."

If Alex was instead thinking of a more random $R' \sim \text{Unif}(\{0, 1...2047\})$, then knowing R' would require 11 bits of information. In general, the more randomness or 'entropy' a random variable has, the more bits of information

one requires to know its outcome.

Turns out we can actually *define* entropy this way: the fewest number of expected yes or no questions necessary to deduce the outcome of a discrete random variable. More specifically, if X is a discrete random variable whose possible values are indexed by $i \in I$ and occur with probability p_i, then the number n_i of yes or no questions necessary to deduce event i will satisfy:

$$\left(\frac{1}{2}\right)^{n_i} = p_i \to n_i = \log_2\left(\frac{1}{p_i}\right)$$

since each question in the optimal strategy approximately reduces the measure of our possible outcomes by $\frac{1}{2}$. Thus the entropy of X is:

$$H(X) = \sum_{i \in I} p_i \log_2\left(\frac{1}{p_i}\right)$$

For instance:

$$H(R) = \sum_{i \in \{0,1 \ldots 1023\}} \frac{1}{1024} \cdot \log_2(1024) = \log_2(1024) = 10$$

as we saw in our game with Alex.

A.2 A Deeper Look at H

We can think of our function H as mapping vectors $\vec{p} \in \Delta^{|I|}$ (the standard $|I|$-simplex) to $\mathbb{R}_{\geq 0}$. H is pretty special in that it satisfies the following properties:

1. Continuity
2. Symmetry
3. Maximized when $\vec{p}$ is uniform
4. Coarse-Graining

Continuity means the pre-image of every open set under H is open. Symmetry means if the entries of $\vec{p}_1$ and $\vec{p}_2$ are permutations of each other then $H(\vec{p}_1) = H(\vec{p}_2)$. Uniformity maximizes information since H attains its maximum when $\vec{p}$ is uniform. Finally, coarse-graining means if we collapse m and

n into a 'super'-event $m \vee n = k$ and call our new coarse-grained probability vector $\vec{q}$ then:

$$H(\vec{p}) - H(\vec{q}) = p_k \cdot H\left(\left[\frac{p_m}{p_k}, \frac{p_n}{p_k}\right]\right)$$

To understand coarse-graining intuitively, consider the following: how many extra yes or no questions do you need on average to deduce $\vec{p}$ if you're given $\vec{q}$?

If $k = m \vee n$ is not true then $\vec{p}$ and $\vec{q}$ are identical. So in this case, 0 extra bits of information are necessary to deduce $\vec{p}$.

If $k = m \vee n$ is true then you think $\vec{p}$'s true state is m with probability $\frac{p_m}{p_k}$ and n with probability $\frac{p_n}{p_k}$. So in this case, $H\left(\left[\frac{p_m}{p_k}, \frac{p_n}{p_k}\right]\right)$ extra bits of information are necessary to deduce $\vec{p}$.

Since k is not true $1 - p_k$ of the time and true p_k of the time, on average you need $p_k \cdot H\left(\left[\frac{p_m}{p_k}, \frac{p_n}{p_k}\right]\right)$ extra bits of information to deduce $\vec{p}$. Thus, the coarse-graining property simply demands that the number of extra yes or no questions necessary to deduce $\vec{p}$ given a coarse-graining $\vec{q}$ is equal to the difference in entropy between $\vec{p}$ and $\vec{q}$.

Remarkably, *all* continuous, symmetric functions which simultaneously maximize information under uniformity and satisfy the coarse-graining property are of the form:

$$H(\vec{p}) = \sum_{i \in 1}^{\dim \vec{p}} p_i \log_B \left(\frac{1}{p_i}\right),$$

i.e. all such functions are identical up to a constant factor.

Throughout this thesis, we set $B = e$ for convenience and still call the units of H 'bits' even though they are technically 'nats'. To convert between bits and nats, note:

$$\ln(2) \cdot \log_2 \left(\frac{1}{p_i}\right) = \ln\left(\frac{1}{p_i}\right) \rightarrow 1 \text{ bit} = \ln(2) \text{ nats}$$

A.3 Joint and Conditional Entropy + Mutual Information

Given two discrete random variables X and Y whose outcomes are indexed by I and J respectively, the number of bits required to deduce (X, Y) is:

$$H(X,Y) = \sum_{(i,j)\in I\times J} p_{ij} \log_2 \left(\frac{1}{p_{ij}}\right)$$

where p_{ij} is the probability of $i \wedge j$. For this reason, $H(X,Y)$ is called *joint entropy* between X and Y.

Via a series of collapses which create super-events $\vee_{j\in J}(i \wedge j) = i$ we can coarse-grain (X,Y) as X. Thus, given X, we require $H(X,Y) - H(X)$ additional bits of information of information to deduce Y. This allows us to define the *conditional entropy* of $Y|X$ as:

$$H(Y|X) = H(X,Y) - H(X)$$

Since knowing X on average falls $H(Y|X)$ bits short of knowing Y, knowing X on average reveals:

$$I(X;Y) = H(Y) - H(Y|X)$$

bits about Y. We call $I(X;Y)$ the *mutual information* between X and Y.

Note:

$$I(X;Y) = H(Y) - [H(X,Y) - H(X)] = [H(X) + H(Y)] - H(X,Y)$$

is symmetric in X and Y. Thus:

$$H(Y) - H(Y|X) = H(X) - H(X|Y),$$

i.e. the amount of information X reveals about Y is equal to the amount of information Y reveals about X.

Let's go back to our initial game with Alex. Suppose you only care about knowing Alex's mystery number but your partner, Charlie, only cares about

knowing another friend Bob's mystery number. Alex and Bob team up and choose their mystery numbers in the following manner: First, Alex generates $R_a \sim \text{Unif}(\{0, 1...1023\})$ and Bob generates $R_b \sim \text{Unif}(\{0, 1...1023\})$. Next, Alex sets their mystery number to $X = R_a - R_b$ and Bob sets their mystery number to $Y = R_a + R_b$. After you play Alex or Charlie plays Bob, the two of you can communicate to help the other person. Will you on average help Charlie more by playing first than Charlie on average will help you by playing first? Since mutual information is symmetric, you and Charlie will help each other equally!

A.4 Kullback–Leibler Divergence

How much will it hurt you if Alex lies? For instance, if Alex actually draws R from the distribution where events $i \in [0, 1023]$ occur with probability:

$$p_i = \frac{i+1}{512 \cdot 1025}$$

but tells you R is instead drawn from the distribution:

$$q_i = \frac{1}{1024}$$

then how many extra bits of information will you on average use to guess Alex's number?

Since you think R is drawn according to $\vec{q}$, you believe halving the measure of possible outcomes with each question will result in asking:

$$\log_2\left(\frac{1}{q_i}\right) = 10$$

questions to deduce event i. Hence, you will on average require:

$$\sum_{i\in[0,1023]} p_i \log_2\left(\frac{1}{q_i}\right) = \sum_{i\in[0,1023]} 10p_i = 10$$

bits to guess R.

However, if you were to *actually* halve the measure of possible outcomes with each question you would require:

$$\log_2\left(\frac{1}{p_i}\right) = 9 + \log_2(1025) - \log_2(i+1)$$

bits to deduce event i. As a result, you would ideally only ask:

$$\sum_{i\in[0,1023]} p_i \log_2\left(\frac{1}{p_i}\right) = \sum_{i\in[0,1023]} p_i \cdot (9 + \log_2(1025) - \log_2(i+1)) \approx 9.722$$

questions on average to guess R.

Thus you end up using about:

$$10 - 9.722 = 0.278$$

extra bits to guess Alex's number.

In general, since our optimal 'halving' strategy relies on knowing the actual distribution $\vec{p}$ we should expect any other strategy (namely ones that are misinformed and designed for $\vec{q}$) to perform worse, i.e.:

$$\sum_{i\in I} p_i \log_2\left(\frac{1}{q_i}\right) \geq \sum_{i\in I} p_i \log_2\left(\frac{1}{p_i}\right)$$

We call the term on the left the *cross-entropy* of $\vec{q}$ relative to $\vec{p}$ and denote it $H(\vec{p}, \vec{q})$. We can quantify the inefficiency of our misinformed strategy by subtracting $H(\vec{p})$ from the cross-entropy above to attain:

$$\sum_{i\in I} p_i \log_2\left(\frac{p_i}{q_i}\right) \geq 0$$

We call this the *Kullback-Leibler divergence* from $\vec{q}$ to $\vec{p}$ and denote it as $KL(\vec{p}\,||\,\vec{q})$. To prove that knowing the true distribution $\vec{p}$ will result in using the fewest number of bits on average we need to prove:

$$\vec{p} = \text{argmin } KL(\vec{p}\,||\,\vec{q})$$

Note since $e^{x-1} \geq x \to x - 1 \geq \ln(x)$ we have:

$$\begin{aligned} KL(\vec{p} \,||\, \vec{q}) &= \sum_{i \in I} p_i \log_2 \left(\frac{p_i}{q_i} \right) \\ &= -\log_2(e) \cdot \sum_{i \in I} p_i \ln \left(\frac{q_i}{p_i} \right) \\ &\geq -\log_2(e) \cdot \sum_{i \in I} p_i \left(\frac{q_i}{p_i} - 1 \right) \\ &= \log_2(e) \cdot \sum_{i \in I} (p_i - q_i) \\ &= 0 \end{aligned}$$

with equality holding iff $\forall i \in I$:

$$\ln \left(\frac{q_i}{p_i} \right) = \frac{q_i}{p_i} - 1$$

This is occurs iff $\forall i \in I$:

$$\frac{q_i}{p_i} = 1,$$

i.e. $\vec{q} = \vec{p}$ minimizes $KL(\vec{p} \,||\, \vec{q})$.

The non-negativity of KL-divergence also allows us to prove that the uniform distribution maximizes H! Note:

$$KL(\vec{p} \,||\, \vec{q}) \geq 0 \to \sum_{i \in I} p_i \log_2 \left(\frac{1}{q_i} \right) \geq \sum_{i \in I} p_i \log_2 \left(\frac{1}{p_i} \right)$$

If $\forall i \in I$ we set $q_i = \frac{1}{|I|}$ then:

$$\ln(|I|) \geq H(\vec{p})$$

with equality holding iff $\forall i \in I,\ p_i = \frac{1}{|I|}$.

A.5 Information Never Hurts

For two discrete random variables X and Y:

$$I(X;Y) \geq 0,$$

i.e. we don't expect learning X to make knowing Y harder since we at most need as many bits on average as before. Proving this fact however requires us to establish a crucial relationship between KL divergence and mutual information, which we now investigate.

Suppose X and Y are indexed by $i \in I$ and $j \in J$ respectively. If one assumes X and Y are independent (i.e. X reveals no information about Y and vice versa) then by linearity of expectation:

$$H(X)+H(Y) = \sum_{i \in I} p_i \log_2 \left(\frac{1}{p_i}\right) + \sum_{j \in J} p_j \log_2 \left(\frac{1}{p_j}\right) = \sum_{(i,j) \in I \times J} p_{ij} \log_2 \left(\frac{1}{p_i p_j}\right)$$

bits of information are necessary to deduce (X, Y).

Since:

$$I(X;Y) = [H(X) + H(Y)] - H(X,Y)$$

, mutual information quantifies how many extra bits of information one asks to deduce (X, Y) if they assume X and Y are independent.

In the language of KL divergence, this translates to:

$$\begin{aligned} I(X;Y) &= \sum_{(i,j) \in I \times J} p_{ij} \log_2 \left(\frac{1}{p_i p_j}\right) - \sum_{(i,j) \in I \times J} p_{ij} \log_2 \left(\frac{1}{p_{ij}}\right) \\ &= \sum_{(i,j) \in I \times J} p_{ij} \log_2 \left(\frac{p_{ij}}{p_i p_j}\right) \\ &= KL(p_{(X,Y)} || p_X \otimes p_Y) \end{aligned}$$

where $p_{(X,Y)}$ corresponds to the joint measure of (X, Y) and $p_X \otimes p_Y$ corresponds to the product measure of the marginals p_X and p_Y. As we saw in the previous section, KL divergence is always non-negative so it follows:

$$I(X;Y) = KL(p_{(X,Y)} || p_X \otimes p_Y) \geq 0$$

as desired.

Acknowledgements

This thesis was largely possible due the support and encouragement of Dr. Ami Radunskaya. Her advice and guidance have been invaluable to me over the past 4 years. I also would not have undertaken this specific project without Dr. Alexander Boyd kindling my passion for information theory and statistical mechanics. The beauty of these subjects has permanently changed the way I view the world. Finally, I am forever grateful to my mom Aruna, my dad Shiva, and my brother Suki for always believing in me and pushing me to follow my dreams. Unlike human rationality, their love knows no bounds.

References

[Hay45] F. Hayek. "The Use of Knowledge in Society". In: *The American Economic Review* 35.4 (1945).

[Sim71] H. Simon. "Designing Organizations for an Information-Rich World". In: (1971), pp. 6–7.

[Aum76] R. Aumann. "Agreeing to Disagree". In: *The Annals of Statistics* 4.6 (1976).

[Mcc81] K. Mcconway. "Marginalization and Linear Opinion Pools". In: *Journal of the American Statistical Association* 76.374 (1981).

[GP82] J. Geanakoplos and H. Polemarchakis. "We Can't Disagree Forever". In: *Journal of Economic Theory* 28 (1982).

[LW83] K. Lehrer and C. Wagner. "Probabililty Amalgamation and the Independence Issue: A Reply to Laddaga". In: *Synthese* 55.3 (1983).

[Han03] R. Hanson. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation". In: *The Journal of Prediction Markets* 1.1 (2003).

[CV10] Y. Chen and J. Vaughan. "A New Understanding of Prediction Markets Via No-Regret Learning". In: *Eleventh ACM Conference on Electronic Commerce* (2010).

[Sim10] C. Sims. "Rational inattention and monetary economics". In: *B. M. Friedman & M. Woodford (Eds.), Handbook of Monetary Economics* 3 (2010).

[Fol12] D. Foley. "Information theory and behavior". In: (2012), pp. 1–21.

[Yu12] N. Yu. "A one-shot proof of Arrow's impossibility theorem". In: *Economic theory* 50.2 (2012).

[Tsa20] E. Tsakas. "Robust Scoring Rules". In: *Theoretical Economics* 15 (2020).