ALTER PRIZE WRITEUP

PAUL "LORXUS" RAPOPORT

Intro

The following results and writing constitute my entry to the ALTER prize. Apart from instances where I draw on existing literature, which are clearly marked, they represent my own original work; I received no help in my research, thinking, or writing past using friends and acquaintances as sounding boards and beta readers. My work takes the form of several smaller results and motivating writeups, so for ease of comprehension, I have broken down the work I have done into sections, each sharing an overarching topic, followed by a final section detailing looser ideas and possible nearby further work. In the ancient tradition of mathematics and research papers, I will use "we" in an authorial sense, rather than "I", for the rest of this paper; this should not be taken as indication that anyone did this thinking or writing but me. The most significant new results in this writeup are the asymptotics analysis of IB agents playing Generalized Antelope/Boar Hunt - a Generalized Stag Hunt-like game - and a proposed method of addressing the problem of non-monotonic loss functions that retains compatibility with refinements of bridge hypotheses.

1. IB Agents Play Population Games

In this section, we build off of work in [4]. Accordingly, $\mathcal{H}_i$ is the hypothesis class of player i. $\mathcal{S}$ is the set of states a player can be in, with initial state s_0. Player i also has action-set $\mathcal{A}_i$ and observation-set $\mathcal{O}_i \subseteq \prod_{-i} \mathcal{A}_i$. Its observation kernel $B : \mathcal{S} \times \mathcal{A} \to \Box\mathcal{O}$ is a map taking current state and chosen action to an infradistribution over observations (ie, what does the agent think it might observe if it's in state s and takes action a) and its transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{O}$ is a map updating the state, given starting state, action, and observation. We also take the setup of SIB-UCB learning and Generalized Stag Hunt from that draft.

We begin with a review of the description of Generalized Stag Hunt as stated in [4]. In this case, we have a singleton $\mathbf{a} \in \prod_i \mathcal{A}_i$ for a Pareto outcome, and observation-sets given by $\mathcal{O}_i = \prod_{j \neq i} \mathcal{A}_j$. Every agent begins by considering only those hypotheses which allow them to force $\mathbf{a}$, if there are any. Following the optimal policy for such a hypothesis h leads to an infinite sequence of repeating $\mathbf{a}$ after some bounded number of rounds, assuming that h is plausible enough not to have been ruled out by previous observations. If all agents reach the start of this infinite sequence at the same time with plausible hypotheses intact, they then get to keep collectively playing $\mathbf{a}$ forever, receiving the Pareto payoff. Because the agents would thus immediately attain Stag Hunt if the "true" hypothesis - that the players could all attempt to attain Stag Hunt immediately - were in the hypothesis space, by assumption it is not.

Date: September 30, 2023.

This brings us to the "grain of truth" problem. In its classical form, we recall that if a bayesian agent has the true hypothesis or correct answer in its hypothesis space, and it assigns any prior probability mass to it at all - the titular grain of truth - then it will eventually learn the true hypothesis. The grain of truth *problem* is that if the agent assigns prior probability 0 to the true hypothesis - possibly because the true hypothesis isn't even in the agent's hypothesis space - then trivially it will never learn the true hypothesis, and will begin to display pathological or undefined behavior once all hypotheses in the space are falsified. Thankfully, infrabayesian agents are more robust in this regard, as we see throughout the infrabayesian sequence.
Kosoy gives us a relatively asymptotically-small set of sharp infra-RDPs for hypotheses. These hypotheses, which are of the rough prose form "I have to enter this specific (meaningless) password in the form of moves in the game. Once I have, the time might be right for all players to attempt to reach Stag Hunt, and if we do, we can stay there forever.", are thus called password-hypotheses; we denote them by h_r if the relevant password is $r \in \mathcal{A}_i^*$, an unbounded finite word in the action-space of player i. Notably, these hypotheses are *not* actually justified, but if everyone has nondogmatic priors on all such hypotheses above some choice of minimal password length, and tries them out some meaningful fraction of the time, eventually - by sheer persistence and luck - the players all sync up when they finish entering their meaningless passwords. More formally, the password-hypothesis h_r is given by

- States: $\mathcal{S} = \{s_0, s_1, s_2, ..., s_{|r|}\}$
- Observation and transition kernels (for $k < |r|$): $B(s_k, r_k) = \mathcal{O}_i$ (maximally unconstrained); $T(s_k, q, o) = s_{k+1}$ if $q = r_k$ (advance) and s_0 (reset) otherwise.
- Special observation and transition kernels: $B(s_{|r|}, \mathbf{a}_i) = \mathbf{a}_{-i}$; $B(s_{|r|}, \neg\mathbf{a}_i) = \mathcal{O}_i$; $T(s_{|r|}, \mathbf{a}_i, o) = s_{|r|}$; $T(s_{|r|}, \neg\mathbf{a}_i, o) = s_0$.

For each player i, we specify that its hypothesis space $\mathcal{H}_i$ consists of all infra-RDPs of diameter $D_i >> 0$, and we let each prior $\zeta_i \in \Delta\mathcal{H}_i$ be given by some notion of descriptive complexity. Then the following sufficiency requirements are also important: that the total probability mass of all password hypotheses $\sum_p \zeta_i(h_r)$ is bounded below by some constant α in the limit we consider, and that the probability distribution over password hypotheses has non-negligible variance in password length. The former makes sure that the agents consider password hypotheses a meaningful proportion of the time, and the latter makes sure that we don't end up with anything like a uniform distribution over password length, which fails in the limit as maximal password lengths are allowed to grow large.
We can model the learning process of player i as a random walk in $\mathbb{Z}/\left\lceil \frac{1}{\alpha}\max_j D_j \right\rceil$, given that an agent will enter one full password at least once per $\tau = \left\lceil \frac{1}{\alpha}\max_j D_j \right\rceil$. With $O(2^{D_i})$ password hypotheses, we can continue the random walk for time $\Omega(2^{D_i}) >> \tau^2$, so that the walks of all such players reach the end state at the same time with probability 1, converging to the Pareto outcome in the process.

Now for our own work, a small but meaningful extension of Generalized Stag Hunt. For Generalized Antelope/Boar Hunt, we posit instead that the Pareto outcome set $\mathbb{P} \subseteq \prod_i \mathcal{A}_i$ is instead given by exactly two points, namely $\mathbb{P} = \{\mathbf{a}, \mathbf{b}\} \subseteq \prod_i \mathcal{A}_i$. We'll also introduce two flavors of agent: antelope-hungry players $p|_a \in \mathcal{P}_A$ and boar-hungry ones $p|_b \in \mathcal{P}_B$, $\mathcal{P}_A \cup \mathcal{P}_B = \mathcal{P}$, whose only distinction will be a reward difference such that for some parameter $k \geq 1$, the reward functions are related as $r(p|_a, o = \mathbf{a}) = k \cdot r(p|_a, o = \mathbf{b})$ and similarly $r(p|_b, o = \mathbf{b}) = k \cdot r(p|_b, o = \mathbf{a})$. Players will be drawn from the player pool randomly between the two as given by another parameter $\rho = \frac{|\mathcal{P}_A|}{|\mathcal{P}|}$

- we now need such a parameter to describe the population, and it's also important in determining the final equilibrium outcome.
The password hypotheses h_r will need to be expanded as well, if we want to extend Generalized Stag Hunt. Now they come in two flavors: $h_{r,\mathbf{a}}$ and $h_{r,\mathbf{b}}$. As the notation suggests, these are password hypotheses "seeking" to achieve Pareto outcomes $\mathbf{a}, \mathbf{b}$.

- States: $\mathcal{S} = \{s_0, s_1, s_2, ..., s_{|p|}\}$
- Observation and transition kernels (for $k < |p|$): $B(s_k, p_k) = \mathcal{O}_i$ (maximally unconstrained); $T(s_k, q, o) = s_{k+1}$ if $q = p_k$ (advance) and s_0 (reset) otherwise.
- Observation and transition kernels (special cases for outcome $\mathbf{a}$, only in $h_{a,p}$): $B(s_{|p|}, a_i) = a_{-i}$; $B(s_{|p|}, \neg a_i) = \mathcal{O}_i$; $T(s_{|p|}, a_i, o) = s_{|p|}$; $T(s_{|p|}, \neg a_i, o) = s_0$.
- Observation and transition kernels (special cases for outcome $\mathbf{b}$, only in $h_{b,p}$): $B(s_{|p|}, b_i) = b_{-i}$; $B(s_{|p|}, \neg b_i) = \mathcal{O}_i$; $T(s_{|p|}, b_i, o) = s_{|p|}$; $T(s_{|p|}, \neg b_i, o) = s_0$.

As before, we let the hypothesis space for each player i contain in its hypothesis space $\mathcal{H}_i$ all infra-RDPs of some radius $D_i >> 0$, with a starting distribution $\zeta_i \in \Delta\mathcal{H}_i$ given by some reasonable notion of descriptive complexity. We note the sufficiency of the two conditions described earlier: namely, that the total probability mass of all password hypotheses $\sum_p \zeta_i(h_{a,p})$ is lower-bounded in the limit we consider, and also that the probability distribution over different passwords has non-negligible variance in password length. Like before, the agents spend a lower-bounded fraction α of time trying password hypotheses; generalizing from before, we take this to mean that $\frac{k}{k+1} \cdot \alpha$ of the time, an agent will test hypotheses drawn from the class matching its agent type, and the other $\frac{1}{k+1} \cdot \alpha$ of the time, the agent will try hypotheses from the "wrong" class.

Proposition 1.1. *Agents using the infra-RDPs specified to learn to play Generalized Antelope/Boar Hunt will eventually converge to one of the Pareto outcomes with probability* 1.

Similar to before, we can model the learning process of player i as a random walk in $\mathbb{Z}/\left\lceil \frac{1}{\alpha}\max_j D_j \right\rceil$, given that an agent will enter one full password at least once per $\tau = \left\lceil \frac{1}{\alpha}\max_j D_j \right\rceil$. With $O(2^{D_i+1})$ password hypotheses, we can continue the random walk for time $\Omega(2^{D_i}) >> \tau^2$ just like before, which still results in - on expectation - all players reaching the end state at the same time well before the walk ends. This time, though, we need to consider what happens each time all players are in their respective states $s_{|p|}$ - the question then remains of whether that player is trying for outcome a or b. Assume that we have N players, and recall that we use ρ for the proportion of type-a players and k for the payout ratio, which also determines hypothesis-test share.

Proposition 1.2. *If $N = 2, k = 1, \rho = \frac{1}{2}$, the probability of success per universal $s_{|p|}$ occurrence is $\frac{1}{2}$, and the resulting outcome is equally likely to be $\boldsymbol{a}$ or $\boldsymbol{b}$.*

In the minimal symmetric case where $N = 2, k = 1, \rho = \frac{1}{2}$, we have a success probability per universal $s_{|p|}$ incidence of $[\rho^2 + (1-\rho)^2] \cdot [{\frac{k}{k+1}}^2 + {\frac{1}{k+1}}^2] + [2\rho(1-\rho) \cdot \frac{2k}{(k+1)^2}] = \frac{1}{2} \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{2}$; a quick analysis of subcases gives us equal chances of settling in outcomes a and b both.
This is both good and bad news: the good news is that the success probability, at least for small N, is decidedly not immediately negligible and in fact quite good, and will contribute an expected constant-factor slowdown in any case of concern; the bad news is that this still scales quickly in N especially. We may take this as indication that while carefully defined IB agents can beat more prosaic agents, they still fall prey to some extent to Buridan's ass and similar such failures of

symmetry-breaking.

Proposition 1.3. *If $N = 2, k >> 1, \rho \simeq \frac{1}{k}$, the probability of success per universal $s_{|p|}$ occurrence is $1 - 2\rho - \frac{2}{k+1}$, and the resulting outcome is **b** with probability ~ 1.*

Brighter news is obtained if we approximate a different 2-player case, where type-**a** players form a small minority of the players but have commensurately bigger payouts. To first order, $\rho^2, \frac{1}{k^2}, \frac{\rho}{k} \simeq 0$ and $\rho \cdot k \simeq 1$, so that this time, the success probability per universal $s_{|p|}$ incidence is given by

$$\sum_{i=0}^{2} \binom{2}{i}^2 \rho^i (1-\rho)^{2-i} \frac{k^i}{(k+1)^2} \simeq$$
$$[\rho^2 + (1-\rho)^2] \cdot [\frac{k}{k+1}^2 + \frac{1}{k+1}^2] + [2\rho(1-\rho) \cdot \frac{k}{(k+1)^2}] \simeq$$
$$[0 + 1 - 2\rho + 0] \cdot [1 - \frac{2}{k+1} + 0 + 0] + (2\rho - 0) \cdot \frac{k}{(k+1)^2} \simeq$$
$$(1 - 2\rho) \cdot (1 - \frac{2}{k+1}) + \frac{2\rho k}{(k+1)^2} \simeq$$
$$1 - 2\rho - \frac{2}{k+1} + \frac{4\rho}{k+1} + 0 \simeq$$
$$1 - 2\rho - \frac{2}{k+1},$$

that is, as long as $\rho, \frac{1}{k} << \frac{1}{2}$, the agents should still generally converge almost as quickly as in the single-optimum case, and will generally ($p \sim 1$) converge to the population-favored **b**-outcome. This time, a more detailed analysis of how often each case happens is startlingly easy thanks to the realization that the only term contributing to an a-outcome is the case where for the two agents, we draw an **a**-agent and a **b**-agent, and they both happen to seek to bring about outcome **a**; this is represented by $2\rho(1-\rho) \cdot \frac{k}{(k+1)^2} \simeq \frac{2\rho k}{(k+1)^2} \simeq 0$, which is thus approximately the expected proportion of the time the agents settle on **a** instead of **b**.
If instead we have $2 < N < \frac{1}{\rho} << \frac{1}{\rho}^2$, that is, only a few players, the $(\rho + (1-\rho))^N$ factor in our corresponding first-order approximation collapses to $\simeq 1 - N\rho + N\rho - (N^2 - N)\rho^2 + \frac{N^2\rho^2}{2} = 1 - \rho^2 \frac{N^2 - 2N}{2} \simeq 1$; that is, taking larger draws from the populations of players even more strongly tilts the resulting equilibrium towards the majority equilibrium, with the minority equilibrium occurring only negligibly often.

Taken together, and contrasting against instances where there exists no Pareto outcome, as in Chicken, or where the Pareto outcome is not a Nash equilibrium, as in Prisoner's Dilemma and its variants, we begin to see more clearly how the use of an IB agent mitigates the "grain of truth" problem. In the Stag Hunt-based examples above, the fact that the true optimal hypothesis - that all the agents can immediately cooperate and achieve the Pareto outcome - is absent from the hypothesis space is utterly irrelevant to the final outcome. As the asymptotics analysis shows, the agents can still make use of hypotheses that are unjustified and which only dimly resemble the true hypothesis; all the same, they still eventually result in a Pareto outcome, just the same as if a true hypothesis had been available in the hypothesis space.

It might then occur to us to wonder: can a "grain of truth" be subdivided? We might posit a discrete valuation on the IB agents' provided hypothesis space, based somehow on a suitably defined distance metric on the larger "true" hypothesis space, which contains both the provided hypothesis space - one which might already be known to be sufficient to induce Pareto outcomes in groups of agents using it - and the true hypothesis, which would cause those same agents to immediately settle on the Pareto outcome. To be useful as a discrete valuation, the valuation would need in particular to be exponentially decaying in terms of bits of relative entropy between the (very simple) "true" hypothesis and (for instance) a password hypothesis.
In the specific case of password hypotheses, the exponent should be dependent on password strength measured in bits - the weaker, the better. In the case of a hypothesis which never predicts arrival at the Pareto outcome, this value would always need to be 0. Using "grains of truth" as a unit of account, we can then sum up all these valuations under the distribution ζ over the hypothesis space to see how many grains' worth of truth reside in the (smaller) hypothesis space. Having done this, we might observe that for a Pareto outcome to be achievable, agents using it must always have priors over the hypothesis space such that the priors-weighted sum of the grain-values is always positive; this holds true for both the classical and IB cases, the major difference being that a classically Bayesian agent can't usefully assign grain-valuations outside $\{0, 1\}$.
Using this framework, we can get some more concrete results for more specific cases. In the case of Generalized Stag Hunt above, suppose we assume that $|\mathcal{A}_i| = 2$, that for any given password p, $\forall i, D_i = O(|p|)$, and that $\zeta \in \Delta\mathcal{H}_i$ is given by the appropriately normalized version of $\zeta(h_p) = 2^{-|p|}$ whenever p is long enough for h_p to have diameter $\geq D_i$ and 0 otherwise, with N representing the minimal password length for which h_p has sufficiently large diameter. Then a lower bound on our grain-valuation is given by $\sum_p \zeta_i(h_p) \cdot 2^{-|p|} \leq \sum_{i \geq N} 2^i \cdot 2^{-i} \cdot 2^{-i} = \frac{1}{2^{N-1}}$, that is, the smaller our N, the greater the total strength of the "subgrains of truth", but the value is always positive. Importantly, in the Generalized Stag Hunt cases examined here, the two hypotheses from before - a lower bound on total password-hypothesis probability mass, and non-negligible variance in probability mass by password length - ensure that this value remains positive. It never gets anywhere near 1 if $N >> 0$, but this is what we should expect - after all, in the classical case of a single indivisible "grain of truth" in the hypothesis space, ζ will (by nontriviality assumption) not assign full probability mass to that hypothesis. Because IB agents can tentatively believe in hypotheses that are in strict sense false or unsupported, they can make use of any such hypotheses close enough to the (nonrealizable) true hypothesis to have nonzero valuation; classical agents, founded on more rigid inferential rules, cannot.

2. Interpreting the Monotonicity Criterion

This section builds on the ideas set forth in [3]. Recall Theorem 3.1 from IBP[3], which says that the expected value of a downward-closed HUC Θ on any functional $f : X \to \mathbb{R}$, where X is a poset, is entirely determined by f^{max}, which is given by $f^{max}(x) = max_{y \geq x} f(y)$, which is monotone nondecreasing. Applying this to our favorite case, that of loss functions, we begin by assuming that we have for a hypothesis a downward-closed HUC $\Theta \in \Box^c el^\Gamma$. We then construct $L^{max} : el^\Gamma \to \mathbb{R}$, which is nondecreasing in 2^Γ, so that for $g \in \alpha \subsetneq \beta \in 2^\Gamma$, $L^{max}(g, \alpha) \leq L^{max}(g, \beta)$. Then by Theorem 3.1, $\Theta(L(g, \alpha)) \equiv \Theta(L^{max}(g, \alpha))$ for all $g \in \alpha \in 2^\Gamma$. But this seems strange - surely non-monotone loss functions should give rise to different valuations? And if they don't, surely it would make more sense to simply require that loss functions be monotone? And having done that, what then to say of ordinary human loss functions?

The case of ordinary human loss functions gives rise to two major philosophical issues. First, a physicalist agent as described in this way would feel nothing at all about anything happening in a causally disjoint universe from it. By contrast, human loss functions can care quite a lot about what might occur in causally disjoint universes from it. Second, and more troublingly, for a physicalist agent, (g, Γ) must be an argmax of $\Theta \circ L$; that is, the physicalist agent must perceive a universe where no computations can be said to run at all as a worst possible outcome[1]. On the other hand, human loss functions are perfectly happy to posit the existence of worlds that they disprefer to literally nothing existing at all.

The first of these conundra is easier to address. Regarding this problem, the physicalist agent is normatively correct: by assumption it has no causal influence over any other causally disjoint universe, and they have no causal influence over it, and if this were ever to change, so would the physicalist agent's attitude. Until such time, though, it does not and should not care about things it has no chance at all of influencing, let alone controlling to any good extent.
The second conundrum is thornier, and the answer lies in the use of HUCs. A homogeneous ultracontribution, just like a cohomogeneous infradistribution, is best understood as being pessimistic yet nondogmatically hopeless - not in the mathematical sense, but rather in emotional valence. Interpreting its expected values on loss functions, we see that it implicitly assumes that the worst that can happen over the possible futures of the agent using it, will be what it assumes will happen, and all the same does not totally rule out possible futures it likes better. Since this also extends to assuming that the grimmest possible manifest fact that could explain the element of el^{Γ} the agent finds itself in, a HUC will spit out identical expected valuations for L, L^{max}. But clearly this need not actually correspond to the observed outcome, even if a cautious agent of the kind that relies on such pessimism knows better than to rely on miraculous decreases in loss function - such eventualities can still occur, to say nothing of actions with uncertain but much less miraculous outcomes. This nuance is not something that a HUC will capture; we might cash out the monotonicity criterion in the fact that when we pass from α to its superset β, Θ doesn't care about unusually pleasant new outcomes - only about any new worst outcomes β has past α that facts corresponding to α and not to β would otherwise rule out. In other words, if there are facts that could explain elements of β but not α which the agent would find more pleasant, the HUC is not interested in those facts; passing to a larger set of indistinguishable possibilities is never good.

Consequently, an IBP agent assigns loss-values to universes where it dies or where nothing exists at all which are exactly the same as the loss-value it assigns to world-states that human loss functions would consider to be much worse than nothing at all. Even in such universes, it nondogmatically holds out hope that there's something that it can do to move the world-state to one it prefers on expectation. The same cannot be said of a totally empty universe, or one in which it doesn't exist. In light of this, the "afterlife" posited by alternative 3 in the subsequent discussion need not be anything nonphysical - it represents the difference between the true worst possible outcome - that of the void - and the agent's death, which may well still occur in a universe whose suitably-restricted "final" outcome is something that the agent would still consider to be a good outcome, were it alive to do so. We might thus reinterpret "non-survivalist" preferences not as a disbelief in death, but a disbelief that death is the absolute end, that it means that the agent's preferences must necessarily be pessimized; a disbelief that its death is meaningfully and certainly

[1]This in turn because only the empty fact can explain any arbitrary subset of Γ being the subset corresponding to which computations appear to be run.

the end for its preferences. Certainly dying is generally bad, by such an agent's lights - it severely limits how much an agent can affect the universe around it, no matter where it finds itself inside el^Γ - but even if an agent knows that it will soon die, it may well be counting on plan-oriented actions it took earlier in its history to bear fruit after it's gone. In the extreme case of this, the agent's death might be critical to a plan which will pay off immediately after it dies, and the successful execution of which it considers to be of such extreme value that it is willing to sacrifice itself to enact it. We may therefore think of L^{max} as the agent's loss function after it has been corrected for the background fact that less computation being run - and thus less existing meaningful stuff - in the universe is generally worse for agents with reasonably complex goals.

All the same, we may *want* to pass to the hull of maximal points of the bridge transform. This will likely be in pursuit of computational efficiency or the ability to disambiguate L from L^{max}; additionally, we consider there to be an outside chance that we might decide to reject the monotonicity principle for loss functions for philosophical reasons instead. In any case, the approach by Harfe and Yegreg as found in [2] seems part of a promising solution to us. Briefly, it involves modifying the beforeness relation on contributions to only sample from 1-Lipschitz functions rather than more generally continuous or measurable functions; we may motivate this through recourse to the definition of the space of signed measures $\mathcal{M}^\pm$ - in particular, its use of the Kantorovich-Rubenstein metric to impose desirable metric structure - as in [1]. Finally, this would require modifying the definition of a HUC to use this new partial ordering. Although the approach as described there doesn't quite work, the fact that a stricter notion of homogeneity results in e.g. a $\mathcal{M}(\wedge)$ which is continuous (although not 1-Lipschitz-continuous) seems hopeful.

We weren't able to satisfyingly extend this in time[2], but we have some initial thoughts and notes that look promising.

Definition 2.1. *Let $\Theta \in \Box^c(\Gamma \times \Phi)$ be a physicalist hypothesis. The **bridge space** of Θ, denoted by $\mathbb{B}_\Theta$, is the set of all **candidate bridges** $\theta \in \Box^c(\Gamma \times 2^\Gamma \times \Phi)$ such that* $\text{supp}\, \theta \subseteq \text{el}^\Gamma \times \Phi$.

More plainly: a hypothesis $\theta \in \Box^c(\Gamma \times 2^\Gamma \times \Phi)$ is a candidate bridge for a physicalist hypothesis Θ if it comes from the same space as any choice of $Br(\Theta)$ would have to come from, and it also already satisfies the first condition for what it means to be a bridge transform - containment of support within $\text{el}^\Gamma \times \Phi$. Elements of $\mathbb{B}_\Theta$ which aren't in $Br(\Theta)$ look like elements $\beta \in \Box^c(\Gamma \times 2^\Gamma \times \Phi)$ for which some map $s : \Gamma \to \Gamma$, the projection $\text{pr}_{\Gamma\times\Phi}(\chi_{\text{el}^\Gamma\times\Phi} \cdot (s \times \text{id}_{2^\Gamma\times\Phi})_* \beta$ doesn't look like any such projection of an element from Θ.

In particular, we propose the following method for addressing the way in which "[passing] to the hull of maximal points of the bridge transform... spoils compatibility with refinements" [3]. Let $\Theta \in \Box^c(\Gamma \times \Phi)$ be a physicalist hypothesis with bridge transform $Br(\Theta)$, and let $\widehat{Br}(\Theta)$ be the hull of maximal points of $Br(\Theta)$, which will no longer be a bridge transform, but will have identical expected-valuations on monotonic loss functions. Monotonic loss function or not, we can then regenerate $Br(\Theta)$ from $\widehat{Br}(\Theta)$ by adding back in all $\eta \in \mathbb{B}_\Theta$ such that for some $\theta \in Br(\Theta), \eta \preceq \theta$ in the sense of Harfe and Yegreg [2]. We clean up by trimming away any elements β of $\mathbb{B}_\Theta$ for which there exists even one $s : \Gamma \to \Gamma$ for which we ever have $\text{pr}_{\Gamma\times\Phi}(\chi_{\text{el}^\Gamma\times\Phi} \cdot (s \times \text{id}_{2^\Gamma\times\Phi})_* \beta \notin \Theta$. At this point, we have a bridge transform again, and we can now refine freely.

Proposition 2.2. *Bridge transforms $Br(\Theta)$ are returned effectively unchanged by the above process.*

[2]We first heard of the ALTER prize in June 2023, and started work on this writeup in late August.

3. Directions for Further Work

The analysis we've written at the end of the section about IB agents playing population games after [4] is thus far limited to the case of a discrete set of Pareto outcomes which are also Nash equilibria; the clear direction of further research would be in figuring out what happens in the more general case of Pareto outcomes which fail to be equilibria. We have some unfinished notes on using the SIB-UCB framework to model a tit-for-N-tats + password-guessing strategy for (I)PD, but suspect that it will be necessary to either give the agents access to the other players' histories (or something equivalent, like their source codes, for use in logical-time pseudo-iterated games) or else require gains from cooperation in order to survive - this latter possibility drives the success of tit-for-N-tats strategies in trembling-hand setups. More prosaically, there's some more likely-easy work to be done in further analysis of asymptotics of suitably-defined population games.

Another major direction is that of physically-manifested facts. Contrasting with [3], we don't think that physically-manifested facts are - or can be - truly objective; the major obstructions for IBP agents G, G' would be in first knowing for certain that each other are rational physicalist agents, and then in making sure that their respective physicalist hypotheses and bridge transforms on those hypotheses line up perfectly. Despite these problems, we can relax our requirements slightly to get suggestive initial results: if G, G' have bridge transforms that overlap heavily, the corresponding overlap on agreement about physically-manifested facts means that they can at least achieve a joint co-subjectivity on those facts, such that for the shared purposes of G, G', those facts might as well be objective. However, it remains to be seen whether (or how badly) this fails in the case of more agents agreeing and disagreeing about more possible facts, especially in ways that might not be linearly orderable or which might reflect irreconcilable differences in bridge transform or physicalist hypothesis - more plainly, the agents can agree on what Φ and Γ look like, and even have broad pairwise agreement (large-measure overlap) on physicalist hypotheses and bridge functions on those hypotheses, and even then any triplet of agents could possibly disagree arbitrarily badly on what that means for (for example) the same physicalist hypotheses they pairwise agree broadly on. Less controversially, and perhaps less originally, we point out explicitly that admissible facthood is an inclusion-reversing operation on subsets of Γ; that is, for $\alpha \subseteq \beta \subseteq \Gamma$, if some fact B admits β, then it also admits α, and there exist facts that admit α and not β. We thus consider such facts as admit more computations - and fewer possible universes - to be sharper or stronger facts; in the lower limit, the empty fact admits all of Γ, and perfect knowledge of exactly which facts are physically manifest admits a singleton set inside Γ.

Finally, from a larger-scale view, we observe that HUCs currently exist in tension with other IB primitives as they are presently defined, and thus believe that some part of the solution to the philosophical problem with monotonicity as described above will involve a refactoring of the definition, or a different choice of primitive somewhere. Its snags with IB logic - which mostly have to do with continuity - and its issues with evaluation on reasonable-seeming loss functions - especially with respect to the mutual incompatibility between passing to maximal points of the bridge transform and subsequent refinement of the bridge transform - both point towards some technical issue with the definition of HUCs; the current best formulation of IB logic, with $\mathcal{M}(\vee)$ being given by the intersection but $\mathcal{M}(\wedge)$ being given by the *convex hull* of the union, looks to fall in the same vein. More fortunately for the approach in [2], our experience with model theory as found in sections 4 and 5.1 of [5] also leads us to strongly suspect that passing to a first-order infrabayesian logic on finite sets will prove to be entirely sufficient, without the need - for most practical purposes - for

the use of higher-order logics.

References

[1] Diffractor. Basic inframeasure theory. https://www.greaterwrong.com/posts/nEFAno6PsCKnNgkd5/infra-bayesian-logicl, 2020.
[2] Harfe and Yegreg. Infra-bayesian logic. https://www.greaterwrong.com/posts/nEFAno6PsCKnNgkd5/infra-bayesian-logic, 2023.
[3] V. Kosoy. Infra-bayesian physicalism: a formal theory of naturalized induction. https://www.lesswrong.com/posts/gHgs2e2J5azvGFatb/infra-bayesian-physicalism-a-formal-theory-of-naturalized, 2021.
[4] V. Kosoy. Infra-bayesian ucb seems pareto efficient in repeated games. *unpublished draft*, 2023.
[5] P. Rapoport. On the profinite distinguishability of hyperbolic dehn fillings of finite-volume 3-manifolds. *arXiv preprint arXiv:2102.10445*, 2021.